SCALABLE MATCHING OF DEFORMED IMAGES

Rohit Kumar Jena

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2026

Supervisor of Dissertation	Co-Supervisor of Dissertation
James C. Gee	Pratik Chaudhari
Director, Penn Image Computing and Science Laboratory, Department of Radiology	Assistant Professor of Electrical and Systems Engineering

Graduate Group Chairperson

Anindya De, Associate Professor, Computer and Information Science

Dissertation Committee

Kostas Daniilidis (Chair), Ruth Yalom Stone Professor, Department of Computer and Information Science
Walter R. Witschey, Associate Professor, Department of Radiology, Perelman School of Medicine
Jianbo Shi, Professor, Department of Computer Science, University of Pennsylvania
Polina Golland, Sunlin and Priscilla Chou Professor, Electrical Engineering and Computer Science, MIT

SCALABLE MATCHING OF DEFORMED IMAGES

To my parents, Hemalata and Rajib.

कर्मण्येवाधिकारस्ते मा फलेषु कदाचन।
मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि॥

— Bhagavad Gītā, 2.47

''You have a right to perform your duty,
but never to the fruits thereof.
Let not the fruits of action be your motive,
nor let your attachment be to inaction.''

ACKNOWLEDGEMENT

I would like to thank my advisors James Gee and Pratik Chaudhari. I have been extremely fortunate to work with them for the past four and a half years and learn immensely from their wisdom and guidance. Their mentorship has permeated into not only how I think about research, but my outlook on life itself. I'd like to thank my thesis committee members - Kostas Daniilidis, Walter Witschey, Jianbo Shi, and Polina Golland for their time and effort in reading my thesis and providing valuable feedback. I also thank Katia Sycara and Kayhan Batmanghelich who have been instrumental in crystallizing my research interests and career path during my time at Carnegie Mellon, and imbued me with the confidence to pursue a PhD. I'd like to thank my undergraduate advisor Suyash Awate for being an incredible mentor, and introducing me to research in AI for medical imaging.

Thank you to all my friends and colleagues both in the United States and back home in India. I've been fortunate to work with incredibly smart and kind people at Amazon and NVIDIA. I extend my gratitude to Brandon Smith and Sid Chaudhary from Amazon, and Nima Tajbakhsh, Ali Taghibakshi, Ashwath Aithal, from NVIDIA for their confidence in my work and their mentorship. I thank my collaborators Bailiang, Chenyang, Yue, Sumedha, Yifan and mentees Vedant, Vasanth, and Deeksha for their enthusiasm in partaking in the ideas, discussions, experiments, and writing that shaped this work. I thank my friends Shubham, Arpita, Tarun, Bhavya, Aniket, Sid, Shaunak, and Jovina for engaging in deep discussions and debates about everything and anything.

I also thank my parents, Hemalata and Rajib, for their endless support and encouragement. No amount of words will be enough to express my gratitude for the unending stream of love, hard work, sacrifices, and support they have provided me through every challenge life has thrown at me; this acknowledgement is only a humble attempt at that expression. I thank my younger brother Rohan for his constant support and for cheering me on through the long stretches of this degree, and I hope that in turn I have been able to inspire him - so that he can aim just as high.

I pay my respects to the divine cosmic force that harmonizes the universe and guides us all to fulfill our *dharma* (orderly duty). I am grateful that these years of research and growth were not separate from my *dharma*, but a way of walking it.

ABSTRACT

SCALABLE MATCHING OF DEFORMED IMAGES

Rohit Kumar Jena

James C. Gee

Pratik Chaudhari

Image registration is a fundamental inverse problem ubiquitous across virtually all biomedical and life science applications. Over the past three decades, significant advances in imaging technology have democratized access to unprecedented spatial and temporal detail, unveiling novel biological structure and dynamics at microscopic and mesoscopic scales. An equally staggering growth in GPU hardware capabilities has transformed major high-performance computing domains including machine learning, computational fluid dynamics, molecular dynamics, and quantum physics. Despite several methodological advances in image registration, the algorithms themselves have not scaled in tandem with advances in imaging and computing capabilities, constraining researchers to work with highly downsampled versions of the rich data they acquire.

This dissertation presents a comprehensive investigation into developing a Pareto-efficient image registration framework that simultaneously optimizes for accuracy, robustness, and scalability. First, we conduct a systematic empirical study comparing classical optimization-based and deep learning paradigms for image registration. This work addresses instrumentation bias in the literature, discusses the strengths and weaknesses of major image registration paradigms, and provides practitioners with a principled framework for choosing appropriate registration paradigms.

Second, we introduce FireANTs, a GPU-accelerated framework for adaptive Riemannian optimization on the space of diffeomorphisms. We formally quantify the severe ill-conditioning of deformable registration and develop a novel Eulerian descent formulation that enables powerful adaptive optimization directly on time-dependent diffeomorphic flows. FireANTs achieves state-of-the-art zero-shot performance across multiple imaging modalities, anatomies, and benchmark datasets while providing orders of magnitude speedups over existing methods.

Third, we develop Deep Implicit Optimization (DIO), a framework that transforms FireANTs into a fully differentiable layer within deep networks. By explicitly decoupling feature learning from optimization, DIO inherits task-specific appearance invariances from the optimization objective while enabling end-to-end learning of dense multi-scale features from weak supervision signals such as anatomical landmarks and label maps. This paradigm provides superior generalization to domain shifts, allows zero-cost plug-and-play of arbitrary transformation representations and regularizations at test time, and enables interactive registration with additional prompts available during inference.

Fourth, we develop a suite of system-level innovations to scale registration algorithms to gigavoxel-scale imaging problems. We propose IO-aware fused CUDA kernels that reduce memory overheads for several bottleneck operations in image registration, enabling multimodal registration on datasets that were previously out of reach – in minutes. Our innovations also accelerate both deep learning training and classical registration workflows, empowering researchers and practitioners to conduct experiments faster and with less compute.

Finally, we showcase the potential impact of the proposed innovations on real-world applications including

multimodal in-vivo brain to ex-vivo hemisphere registration (MRI and histopathology), restricted deformations for echo-planar MRI images for geometry distortion correction, and gradient-free algorithms for sparse landmark-guided registration for lung CT scans as three ‘in-the-wild’ applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

In medical imaging, different subjects or the same subject across various time points present subtle to drastic anatomical variability due to inter-subject variations, differences in image acquisition, pathology, and other factors. This variability prevents accurate and consistent quantitative morphometric analysis and multi-modal fusion of complementary modalities, hindering downstream tasks like anomaly detection, disease progression tracking, and population-level analysis. Therefore, images must be brought onto a common coordinate frame prior to performing any downstream analysis. The ability to quantitatively transform a set of images onto a common coordinate frame - known as *image registration* is a fundamental inverse problem ubiquitous across virtually all of biomedical imaging and life sciences. The computational task is to find a transformation that maps a 'source' configuration of pixels (i.e. also called the *moving image*) to an 'target' configuration of pixels (i.e. also called the *fixed image*). Since the optimization variable is to find a transformation that maps the source to the target, it is an inverse problem. This computational task is defined as an inverse problem that operates on images defined on a spatial domain Ω, typically a compact subset of $\mathbb{R}^2$ or $\mathbb{R}^3$. Images are defined as functions $I : \Omega \to \mathbb{R}^K$ that map the spatial coordinates into a K-dimensional feature vector. The task is to find a coordinate transformation $\varphi : \Omega \to \Omega$ that establishes voxelwise correspondences between a *moving* image I_m, and a *fixed* image I_f, i.e.

$$I_m(x) \approx I_f(\varphi(x)) \quad \forall x \in \Omega$$

where the $\approx$ sign is overloaded to represent 'anatomical similarity' rather than intensity similarity, to subsume multimodal or geometric-matching based registration paradigms. The iterative nature of a standard registration method is illustrated in Fig. 1.1.

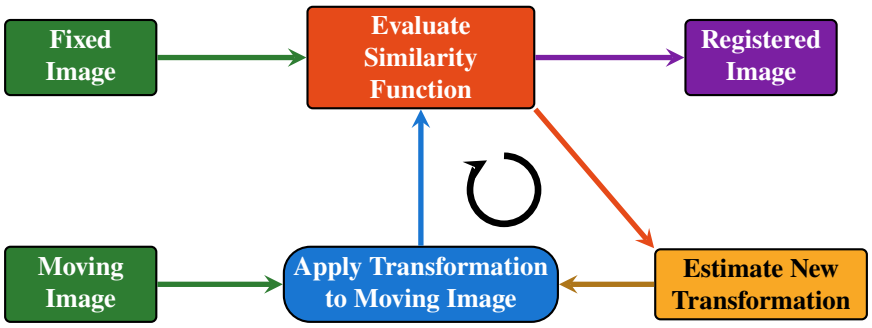


Figure 1.1: Overview of the iterative image registration pipeline. The moving image is warped by the current transformation and compared to the fixed image via a similarity function; a new transformation estimate is computed and the loop repeats until convergence, at which point the registered image is produced.

1.1. Applications of Image Registration

1.1.1. Clinical-grade datasets

In biomedical imaging, image registration is used for a wide range of quantitative studies. Since image registration algorithms establish dense voxelwise correspondences, the transformations are used for voxel-based morphometry (VBM) for detecting structural differences in populations. A common instance of

quantitative VBM in neuroimaging is the cortical thickness estimation and shape analysis across cohorts. One of the most common algorithms - DIRECT (Das et al., 2009) is built on a registration algorithm to estimate gray matter thickness using a gray+white matter (GM+WM) binary mask as the starting configuration and the WM mask as the ending configuration that essentially gives the displacement of the cortical surface, and consequently a per-voxel cortical thickness estimate. VBM has been widely applied to study structural brain changes associated with aging, neurodegenerative disorders, psychiatric conditions, and developmental differences. Population-wide atlas construction and normalization techniques use image registration to define a common coordinate frame (i.e. the atlas) and register subjects to the atlas. The atlas provides an anatomically consistent reference for group analyses, enabling statistical comparisons of structure, function, or connectivity. Popular atlases in neuroimaging include the MNI152, Talairach and ICBM atlases (Lancaster et al., 2007). More specific atlases (disease-specific (Ravikumar et al., 2020), pediatric (Avants et al., 2015; Xu et al., 2022), elderly (Wu et al., 2022b)) can be built with the same image registration algorithms. Image registration is used in building and understanding biomechanical models of organ deformation and anatomical variation modeling. In pulmonary imaging, image registration of inspiration-expiration scans are used to quantify how biomechanics are affected in populations sufffering from pulmonary conditions like COPD (Heinrich et al., 2015). In cardiac imaging, biomechanics-informed modeling uses image registration to introduce physical constraints such as tissue elasticity, myocardial fiber orientation, and boundary conditions to infer quantities that cannot be directly imaged, such as local stress, strain, or pressure fields (Qin et al., 2020, 2022). On 4D imaging sequences, image registration can be used to capture the cardiac motion across cardiac cycles, and finite element models or continuum mechanics frameworks could be used to simulate cardiac deformation and assess parameters such as contractility, torsion, and wall thickening.

1.1.2. *Ex-vivo* neuroimaging and histology for neuroanatomical and pathological studies.

A large body of neuroanatomical studies are performed in conjunction with ex-vivo and blockface imaging and histology to create detailed, multi-scale anatomical references by integrating structural, molecular, and cytoarchitectural information across imaging modalities (Casamitjana et al., 2025; Ravikumar et al., 2024). In-vivo MRI scans are typically limited by resolution due to constraints on scan time and motion artifacts associated with longer scan times. This makes in-vivo MRI scans unsuitable for studying the microstructural changes associated with neurodegenerative disease progression. Registration of fine anatomical details like cortical layers, axonal projections, or individual nuclei are useful to understand neuropathology, and such analyses are not possible at macroscopic clinical scales. Therefore, high-resolution ex-vivo scans and blockface imaging are used as a bridge between in-vivo and histology, with the latter used as a gold standard for ground-truth microscopic tissue characterization and pathology. Many complementary stains are used to visualize neuropathological features, including protein aggregates, neuronal loss, gliosis, and myelin integrity. Accurate registration of these structures is important to improve our understanding of morphological effects of pathology. For example, Alzheimer's Disease (AD) is characterized by cortical atrophy in the medial temporal lobes, particularly hippocampus, entorhinal cortex, and parahippocampal gyrus (Ravikumar et al., 2024; Echávarri et al., 2011). Accurate atrophy quantification of these structures can only be reliably performed at $\sim$0.5 mm or better resolution MRI or ex vivo imaging, necessitating high resolution registration. Parkinson's Disease (PD) is characterized by degeneration of DA neurons in the substantia nigra (Triarhou, 2013) and subthalamic nucleus that are small ($\sim$5-10 mm), requiring $<$0.7 mm isotropic or ex vivo imaging for volumetry or susceptibility mapping for accurate delineation (Welton et al., 2023). Multiple Schelosis is characterized by cortical lesions (Madsen et al., 2021; Beck et al., 2018) that cannot be delineated at the in-vivo resolution and typically requires high resolution ex-vivo imaging and histopathology integration. Except in-vivo imaging,

all other modalities are very high resolution typically ranging from $500\mu m$ up to $100\mu m$ (Ravikumar et al., 2024; Echávarri et al., 2011; Welton et al., 2023; Madsen et al., 2021) for ex vivo imaging and $\sim 10\mu m$ for histology sections. High-resolution imaging and registration are essential in these contexts because they enable accurate cross-modal alignment and preservation of fine anatomical detail that would otherwise be lost through downsampling. Most of these studies, however, limit their analyses to localized effects due to the significant computational cost of registering the entire brain at high resolution. Recently, projects like Allen Brain Atlas (Allen Institute for Brain Science) and multiple large scale consortia including Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium and the Human Mouse Brain Atlas (HMBA) consortium are aimed at creating detailed, multimodal brain atlases linking cellular, molecular, and anatomical organization across species and disease states - combining individual efforts from multiple institutions together into a unified resource. Achieving this multimodal organization at the whole brain level requires high-resolution registration tools to accurately align diverse imaging modalities while preserving fine-scale cytoarchitectural detail. Most in-vivo to histology registration workflows require registration of an in-vivo image to its ex-vivo counterpart. The $100\mu m$ ex-vivo and $250\mu m$ in-vivo images released in Edlow et al. (2019); Lüsebrink et al. (2017) are intended to be used as high-resolution templates to enable accurate studies, but lack of computationally efficient methods restricts their broad usage in the neuroimaging community.

1.1.3. Large scale registration in model organisms.

Over the past decade, imaging across the life sciences and biomedical domains has progressed from mesoscale surveys to organ- and organism-wide acquisitions at cellular or even subcellular resolution. These span transparent organisms and small animal models (e.g., *C. elegans*, zebrafish, adult *Drosophila*) (Varol et al., 2020; Venkatachalam et al., 2016; Marquart et al., 2017; Gupta et al., 2018a; Peng et al., 2011; Brezovec et al., 2024), whole-rodent brains imaged at micron or submicron sampling (Gong et al., 2016b; Wang et al., 2020a), and non-human primate (NHP) and human ex vivo MRI at hundreds of microns (Skibbe et al., 2023; Milham et al., 2018a; Edlow et al., 2019; Lüsebrink et al., 2017). Such modalities routinely generate giga- to teravoxel volumes (Kutten et al., 2016; Nazib et al., 2018). Their scientific utility, however, hinges on the ability to perform registration at the native resolution of acquisition, i.e. aligning specimens (or modalities) in a common coordinate system without sacrificing the fine-scale morphologies-cell bodies, layers, axon bundles, synaptic neighborhoods– that motivate high-resolution acquisition in the first place (Nazib et al., 2018; Goubran et al., 2013).

Cellular-resolution atlases in model organisms. In *C. elegans*, statistical atlases of neuron positions require aligning whole-animal volumes to preserve the fidelity of closely apposed cells (Varol et al., 2020; Venkatachalam et al., 2016). In zebrafish, deformable registration with cellular-level precision and minimal perturbation of tissue morphology enables pooling of gene expression, single-neuron morphologies, and brain-wide activity (Marquart et al., 2017; Gupta et al., 2018a). In adult *Drosophila*, whole-brain registration underpins large-scale databases and enables structure–function integration (for example, aligning two-photon functional volumes to EM-derived connectomes) (Peng et al., 2011; Brezovec et al., 2024).

Whole-brain rodent imaging Large scale efforts like NIH's Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network (BICCN) aims to provide researchers and the public with a comprehensive reference of the diverse cell types in human, mouse, and non-human primate brain, and researchers collect a wide range of multimodal data including MRI, sectioning tomography, microscopy, antibody stains (e.g. calbindin), and spatial transcriptomics. In rodents, fMOST pipelines yield whole-brain images at micron sampling (e.g., $0.32\mu m$ voxels generating $>$10TB datasets) for tracing long-range axons and quantifying cytoarchitecture (Gong et al., 2016b). Constructing stereotaxic spaces such as the Allen CCFv3

and Waxholm rat atlas requires deformable registration that preserves layers and boundaries (Wang et al., 2020a; Kleven et al., 2023; Kronman et al., 2024). Currently, there is a huge gap between the resolution at which data is acquired and the resolution at which templates are created. For example, STPT images can be collected at less than $1\mu m$ resolution (Liwang et al., 2023), but the Allen CCFv3 template is generated at $10\mu m$ by upsampling the registrations from $25\mu m$ due to compute constraints. Developmental atlases also register the CCFv3 at resolutions significantly downsampled from the original $10\mu m$ template (Kronman et al., 2024; Liwang et al., 2025) citing lack of computational resources as one of the primary reasons. Certain phenomena of interest like cellular organization and brain-wide connectomes are emergent only at very high resolutions, necessitating computational tools that can scale with the data.

Zebrafish Initially adopted as a developmental biology model because of its ease of domestication, high fecundity, and transparent early life stages, the zebrafish has gained broader prominence with advances in brain imaging, molecular genetic tools, and behavioral assays (Kenney et al., 2021b). For the AZBA template (Kenney et al., 2021b), the raw images are collected at $4\mu m$ but was resampled to $8\mu m$ ($8\times$ downsampling) due to system constraints. The tools used for the registration (Friedel et al., 2014) do not recommend running locally and only on a distributed cluster. Brain-wide cellular resolution imaging of transgenic zebrafish lines (Tabor et al., 2019) is performed on large clusters like Biowulf Linux cluster at the National Institutes of Health, significantly reducing accessibility of these imaging resources to researchers, signifying an unmet need for efficient and distributed multimodal registration frameworks.

Non-human primates (NHP), human ex vivo MRI, and biomedical imaging. At larger scales, NHP and human ex vivo MRI achieve a resolution of few hundreds of microns (Milham et al., 2018a; Skibbe et al., 2023; Edlow et al., 2019; Lüsebrink et al., 2017). Registration is essential for fusing these volumes with histology or in vivo MRI, enabling alignment of cytoarchitectural detail and correction of distortions (Goubran et al., 2013). Without such alignment to a stereotaxic space, the cellular and laminar motifs motivating ultra-high-resolution imaging cannot be meaningfully compared or aligned across specimens or modalities.

Across these diverse domains, the unifying requirement demands access to scalable multimodal registration algorithms.

1.2. Types of Image Registration

Image registration methods can be categorized by the nature of the transformation model $\varphi(x)$ and its extent used to solve the problem. Broadly speaking, there are two types of desired transformations - (i) global transformations, and (ii) local transformations.

1.2.1. Global transformations

Global transformations are characterized by a small number of parameters, typically a linear or heavily regularized nonlinear (rigid, affine, polyrigid (Gopalakrishnan et al., 2025) or polyaffine (Legouhy et al., 2023)) transformations. These transformations are typically used to align the overall orientation and position of the images before fine-tuning with local (deformable) transformations. Sometimes they can be used with gradient-based methods if the transformation of the moving image is not too complex relative to the fixed image, i.e. if the images are obtained in the same scanner, or technicians have standard orientation protocols for histology, etc. Fig. 1.2 illustrates some applications of global registration. In histology, global registation is deployed often to stack consecutive sections of the same subject to create a 3D volume, or align MRI casettes or histology to blockface images since there is no standardized image acquisition protocol. Mancini et al. (2020)

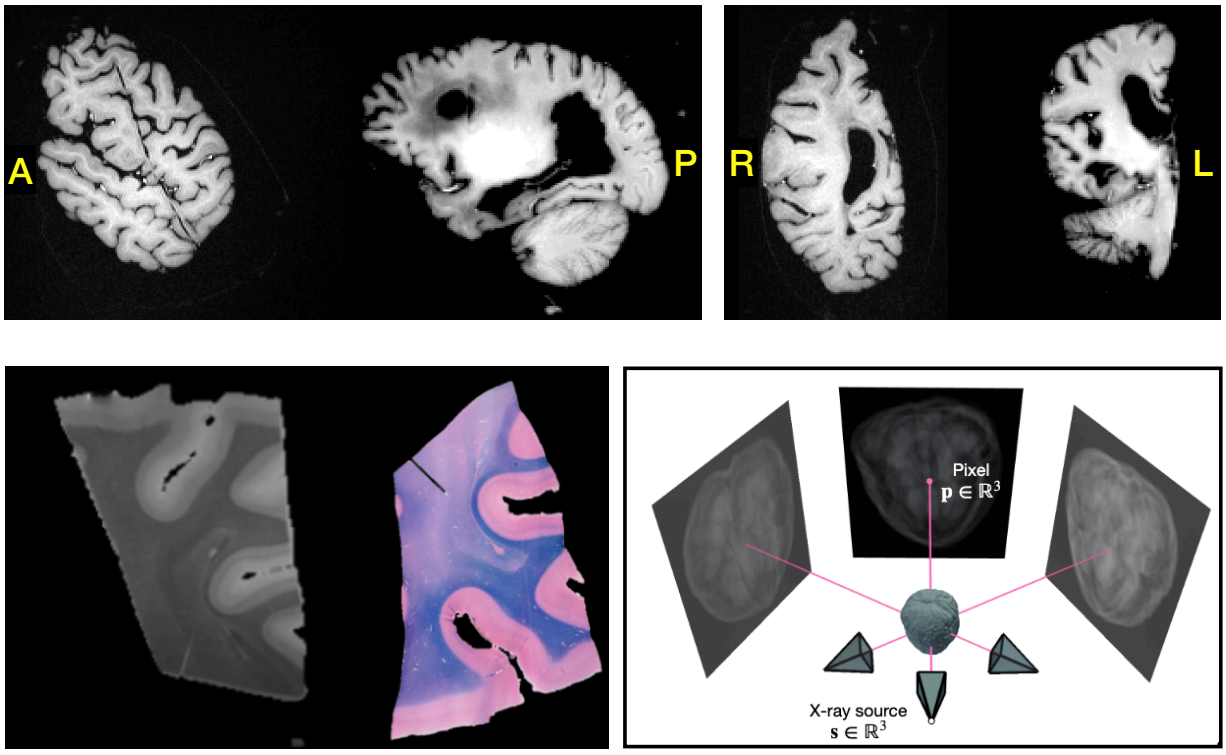


Figure 1.2: Applications of global registration. Top row shows *ex-vivo* hemisphere (left) and *in-vivo* MRI (right) images along sagittal and coronal views (with A,P,R,L labels denoting anterior, posterior, right, and left respectively). The ex-vivo hemisphere has a large oblique rotation axis and direction relative to the *in-vivo* MRI, necessitating a global registration step. **Bottom row, left** shows an MRI casette chunk from the superior frontal lobe and its corresponding histology section. There are large rotations, flipping, and is multimodal by nature. **Bottom row, right** shows a 3D CT scan and three different X-ray projections. Finding accurate camera parameters of the X-ray projection without any prior knowledge is a challenging task requiring global optimization. This picture is taken from Momeni et al. (2024).

use SURF features and RANSAC to find global alignments between microtome block sections and whole slice blockface. Pichat et al. (2017) use shape elements to find global alignments between MRI sections and histology slides. Other approaches (Gangolli et al., 2017) use manually defined landmarks on Myelin Block Gold II stained sections to find initial alignments. Huszar et al. (2023) formulate a multi-stage pipeline for histology to blockface MRI registration by leveraging a set of intermediate images obtained after performing every cutting step of the coronal slab. This set of images is used to form an XOR map of candidate locations which is used as initialization for the subsequent rigid / affine registration. Deep learning methods like Wang et al. (2023) and Momeni et al. (2024) use neural networks and differentiable rendering techniques to perform global registration to initialize the optimization for a subsequent local transformation step. Other methods (Carey et al., 2023) parameterize the global transformation as the set of corner points of the 2D section in the 3D image coordinates to perform global registration using supervised learning. This thesis will primarily focus on local transformations, although we will utilize a few global transformation optimization techniques in Chapter 6 for 'in-the-wild' scenarios. Algebraic second-order moments is also a common approach to find global alignments between images using the intensities as 'masses' (Jaklic and Solina, 2003; Yang and Cohen, 1999; Taubin and Cooper, 1991; Yushkevich et al., 2016).

1.2.2. A walk down the 'deformable transform' memory lane

In contrast to global transformations, local transformations are characterized by a large number of parameters, typically a deformable transformation model. A d-dimensional ($d = \{2, 3\}$) deformable model $\varphi(x) = x + u(x)$ where x can represent upto N voxels, the vector field $u(x)$ requires storing dN parameters, compared to just $d(d+1)$ parameters for an affine transform. This necessitates spatial regularization $R(u)$ to ensure that the deformation field is smooth and physically plausible. These transformations are typically used to align the fine-grained details of the images after the global transformation has been applied. Despite what the name might suggest, local does not imply ''small'' - in fact, many local deformation models are expected to capture 'large deformations' (when compared to the scale of the anatomical structure being deformed). The primary challenge to enable standardization of subsequent morphometric analysis in neuro, pulmonary, cardiac, and virtually all other organ systems is the high amount of anatomical variability across subjects, or within a subject across time points. Therefore, the computational challenge of most of the applications of deformable registration lies in optimizing an expressive yet regularized high-dimensional transformation spaces.

Elastic models Early works in deformable image matching proposed using geometric transformations derived from physical models like elastic bodies. Elastic body registration treats the moving image as a flexible, solid material that deforms under the influence of internal forces generated by image dissimilarities. In a linear elastic framework, the displacement field u is governed by the Navier-Cauchy equations:

$$\mu \nabla^2 u + (\lambda + \mu) \nabla (\nabla \cdot u) + f = 0 \tag{1.1}$$

where μ and λ are the Lamé constants representing the material's shear modulus and bulk modulus, and f is the external body force derived from the gradient of the image similarity term. Early works on elastic body registration included Broit (1981); Bajcsy and Kovačič (1989); Gee and Bajcsy (1998); Davatzikos (1997); Rabbitt et al. (1995); Pennec et al. (2005); Burger et al. (2013) employing a multitude of techniques atop the Navier-Cauchy equations Eq. (1.1) including multi-resolution optimization, probabilistic formulations, and hyperelastic models (i.e. where the stress-strain relationship is a differential equation capturing non-linear instead of a linear relationship). Elastic models are highly effective for capturing small, smooth deformations and possess a 'rest state' or 'memory,' meaning the energy increases with the magnitude of displacement. This property is advantageous for maintaining anatomical topology but can become a limitation when attempting to register images with significant morphological differences, as the restorative forces may prevent the model from reaching a global optimum.

Viscous fluid models To address the limitations of elastic models in scenarios involving large deformations, viscous fluid models were introduced. In this paradigm, the transformation is not viewed as a static displacement but as the result of a velocity field $v(x, t)$ acting over a virtual time interval. By penalizing the velocity rather than the displacement directly, fluid models allow the 'material' to flow and undergo substantial shape changes without a buildup of restorative energy. The seminal work of Christensen et al. (1996, 1997) used a modified Navier-Stokes equation

$$\mu \nabla^2 \vec{v} + (\lambda + \mu) \nabla (\nabla \cdot \vec{v}) + \vec{b}(\vec{u}) = \nabla p + \rho \left(\frac{d\vec{v}}{dt} \right) + \vec{v}\eta \tag{1.2}$$

and made a number of simplifying asumptions for very low Reynold's number flow neglecting the pressure

gradient and the inertial terms such that the equation becomes

$$\mu\nabla^2\vec{v} + (\lambda+\mu)\nabla(\nabla\cdot\vec{v}) + \vec{b}(\vec{u}) = 0 \tag{1.3}$$

This is similar to Eq. (1.1) but with the velocity field v replacing the displacement field u. The displacement field u is then obtained by integrating the velocity field v over time, i.e. $u(x,t) = \int_0^t v(x,t')dt'$. Other works (Bro-Nielsen and Gramkow, 1996; Crum et al., 2005; Cahill et al., 2007; Wang and Staib, 2000; D'agostino et al., 2003) used similar derivations but made different simplifying assumptions, like usage of a convolution filter in scale-space, multi-grid approach to handle anisotropy, using Fourier methods to solve linear Navier-Stokes equations, inverse consistency, using other regularization terms under Dirichlet, Neumann, or periodic boundary conditions. This approach is particularly effective for registering images with significant morphological differences, as the fluid flow can adapt to the shape of the anatomical structure.

Diffusion models Diffusion-based registration, pioneered by the ''Demons'' algorithm (Thirion, 1998), offers a computationally efficient alternative to full physical simulations. For diffusion models, the deformation is modeled by the diffusion equation:

$$\nabla\vec{u} + \vec{F} = 0 \tag{1.4}$$

Note that unlike other formulations, Eq. (1.4) is not explicitly modeled in the objective. However, this equation allows for modeling the Gaussian kernel as the Green's function of the diffusion equation. Most diffusion-based methods are inspired by the 'Demons' algorithm, an analog to Maxwell's demons (Thirion, 1998) wherein the Demon forces were computed for each particle (which becomes the displacement field at that location) , followed by an update of the displacement field using the calculated forces. A gaussian filter is applied to the nonparametric displacement field after every iteration for regularization. Several methods utilized Maxwell's demons as a building block for computing the deformation field, proposed diffeomorphic updates, and accurate numerics (Thirion, 1998; Pennec et al., 1999; Vercauteren et al., 2007a,b, 2009; Dru et al., 2010). However, a major limitation is that the optical flow equation and demons algorithm in the aforementioned formulations are well-defined only for monomodal images, significantly limiting their applicability to multimodal registration. Follow-up studies have attempted to use normalized mutual information or image translation as criteria for defining the demons forces either in the multi-modal or pseudo-monomodal settings (Modat et al., 2010b; Guimond et al., 2002).

Flows of diffeomorphisms Flows of diffeomorphisms have also been proposed to model the deformation. These methods have been by far the most popular and successful models used for both iterative and non-iterative formulations. In this case, the deformation is modeled by considering its velocity over time according to the transport equation:

$$\frac{d\vec{u}}{dt} = \vec{v_t}\circ(id+\vec{u}) \tag{1.5}$$

where $\vec{v_t}$ is a smooth velocity field at time t. The transport equation describes the dynamics of a set of particles in the Langrangian frame of reference, which are typically considered to be the lattice points of the image, or a set of control points. The velocity field is then integrated to obtain the deformation field. To ensure that the velocity field is smooth, a regularization term is added: $R(\vec{v_t}) = \int_\Omega \|\vec{v_t}\|_V^2 d\Omega$, where $\|\cdot\|_V^2$ is a suitable norm on the velocity field, typically a second order differential operator. A remarkable property of this formulation is that the final deformation field is a diffeomorphism by construction, i.e. it is invertible and its inverse is also a diffeomorphism. This is a highly sought after property for registration, as it ensures that the transformation is physically plausible and anatomically consistent and does not introduce any tearing or folding of the anatomy. Seminal works on the variational formulation (Dupuis et al., 1998) and

landmark matching via large deformation diffeomorphisms (Joshi and Miller, 2000; Beg et al., 2005) provided a rigorous variational and geometric formulation of registration. This was followed by groupwise analysis (Marsland and Twining, 2004), sparse LDDMM optimization (Sommer et al., 2011), and simultaneous multi-scale registration (Risser et al., 2011). A limitation of the vanilla LDDMM formulation is that it is not scalable to large datasets, and is computationally expensive due to storage of a 4D velocity field for a 3D image. To address this, several methods proposed to drop the time dependence and optimize either a *time-independent* velocity field directly or solved the geodesic equation which provides a *shortest length* time-dependent velocity field that depends only on an initial momentum (Goos et al.; Ashburner and Friston, 2011; Ashburner, 2007; Marsland and McLachlan, 2007; Cotter and Holm, 2006; Arsigny et al., 2006).

Discrete optimization approaches While continuous variational methods have dominated the field, discrete optimization has emerged as a powerful alternative that addresses some of the inherent limitations of gradient-based optimization. These methods reformulate deformable registration as a labeling problem on a graph, often utilizing Markov Random Field (MRF) theory (Glocker et al., 2011). Most formulations like Zikic et al. (2010) proposed a discrete optimization of an approximated energy function in an iterative loop, or used graph-cuts (So et al., 2011) to find *globally optimized* solutions to the underlying energy function. Linear programming formulations are also proposed with a multi-scale incremental approach to account for large deformations and high-resolution images (Glocker et al., 2008). In these methods, image matching is formulated as a graph matching problem, where nodes correspond to the voxels in the lattice, neighborhoods define the edges, and the labels are the deformation fields. Other methods use coupled convex optimization with brute-force discrete optimization with hand-engineered or learned image features (Modersitzki, 2009; Siebert et al., 2024; Li et al., 2023b).

Interpolation based transformations In contrast to geometric motivations for representing deformation fields, another line of work considers interpolation theory to formulate optimizable transformation models, among which thin-spline-based transformations (Bookstein, 2002, 1991; Johnson and Christensen, 2001; Donato and Belongie, 2002) are one of the most popular models. Thin-plate splines are also studied in the context of decomposition of deformations (Bookstein, 2002) into a set of principal warps i.e. orthogonal components of the deformation. Comparative study of transformation functions (Zagorchev and Goshtasby, 2006) compare the characteristics of thin-plate splines, multiquadric, piecewise linear, and weighted mean transformations and the effectiveness of these transformations depending on the spacing between control points.

However, most of these methods were not built for contemporary applications which require high computational throughput, and are not scalable to large datasets. For instance, one of the major feature requests for the ANTs framework (Pellman et al., 2016; NirutaDhimal and contributors, 2023) is to have GPU support for fast image registration to enable large-scale registration pipelines. This is especially timely, owing to the dramatic adoption of GPU hardware both in academia and industry, significantly lowering the cost of parallel computing and democratizing access to high-performance computing. Moreover, there have been significant innovations in adaptive optimization for deep learning (Kingma and Ba, 2014; Tieleman et al., 2012; Yao et al., 2021; Gupta et al., 2018b) that are not leveraged for spatially regularized and ill-conditioned problems like deformable image registration, motivating the need for a stronger class of optimizers deployed on fast SIMD-friendly GPU hardware to enable large-scale registration pipelines.

Deep Learning for Image Registration Contrary to iterative registration, which is formulated as a variational optimization problem, deep learning for image registration is almost universally modeled as a statistical learning problem using feedforward inference. In contrast to most classical methods, earliest Deep Learning

for Image Registration (DLIR) methods employed supervised learning for registration tasks (Cao et al., 2017; Krebs et al., 2017; Rohé et al., 2017; Sokooti et al., 2017) where the deformation field is obtained either manually or from a classical method. Voxelmorph (Balakrishnan et al., 2019) was one of the first approaches that introduced unsupervised learning for registration of in-vivo brain MRI images. Subsequent research expanded upon this paradigm, exploring diverse architectural designs (Chen et al., 2022b; Lebrat et al., 2021; Jia et al., 2022; Mok and Chung, 2022), loss functions (Zhao et al., 2019b,a; Joshi and Hong; De Vos et al., 2019; Mok and Chung, 2020a; Zhang et al., 2021b; Qiu et al., 2021; Chen et al., 2022a), and formulations based on incorporating inverse-consistency or symmetric transforms (Mok and Chung, 2020b; Kim et al., 2021, 2019; Tian et al., 2023a; Zhao et al., 2019b). However, hyperparameter tuning became a challenge for DLIR methods since the methods had to be retrained for every new value of the regularization parameter. This motivated techniques such as conditional hyperparameter injection which addressed hyperparameter tuning (Mok and Chung, 2021; Hoopes et al., 2021), while domain randomization and fine-tuning (Hoffmann et al., 2021; Uzunova et al., 2017; Pérez de Frutos et al., 2023; Fu et al., 2020b) aimed to addressed generalizability of DLIR methods across domains. Recently, pretrained or foundation models are also proposed to address the generalizability of DLIR methods across different imaging and anatomy (Liu et al., 2021a; Tian et al., 2023b). However, these methods perform a monolithic prediction of the warp field from the input images, losing feedback from the intermediate stages of the registration process as done in classical methods. To refine the warp fields, recurrent or cascade-based architectures were proposed (Zhao et al., 2019a,b; Zhang et al., 2021b; Chen et al., 2022a). Cascade-based methods create a substantial memory overhead due to backpropagation through the entire sequential registration operations and storage of intermediate volumes (Bai et al., 2022). Another line of work considers using iterative optimization methods but using neural networks to impose an implicit structural or functional prior on the deformation field. This leverages the idea of deep implicit priors (Ulyanov et al., 2020) within optimization frameworks to improve the performance of optimization methods or incorporate implicit constraints of the optimized warp field (Wu et al., 2022a; Wolterink et al.; Joshi and Hong; Hu et al., 2024). Another promising avenue in deep learning for registration methods to ensure good domain generalization is the identification of registration-specific designs (Jian et al., 2025, 2024; Liu et al., 2025a) playing a more prominent role than 'trend-driven architectural designs' in domain generalization. Newer designs (Honkamaa and Marttinen, 2023; Liu et al., 2024c) utilize learnable versions of correlation volumes or multi-scale deformation composition to achieve large deformations. This is an active area of research, which can effectively bring the best of both worlds of iterative and deep learning methods.

However, all these methods have a very high memory footprint even at inference. For example, a clinical MRI scan of size 34MB can consume upto 10GB of GPU memory at inference. In contrast, the resolution of images discussed in Section 1.1.2 are typically an order or two magnitudes larger than the resolution of clinical MRI scans, requiring significantly more memory. Although methods like Jia et al. (2023) aim to produce a bandlimited deformation field requiring less memory, its scalability to a high-resolution image is not well studied. These modern frameworks have also not been widely adopted in clinical settings, which could be due to multifactorial issues like lack of interpretability of the deformation field, brittleness to out-of-distribution datasets, high computational requirements, and lack of distributed computing capabilities - highlighting an unmet need for a robust and accurate (theory), interpretable and steerable (theory and empirics), and a distributed image registration inference framework (systems).

1.3. Statement of Contributions

A holistic view of performance of various registration frameworks In the following (second) chapter, we perform a systematic study of the strengths and weaknesses for both iterative optimization and deep learning methods for unsupervised image registration (i.e. using *only* image based similarity loss functions). We show that several deep learning methods in the literature suffer from instrumentation bias for running iterative methods, leading to the performance of these methods being misrepresented in the literature. Our study performs a careful and thorough re-evaluation and finds that deep learning methods are *no better* than iterative methods when recommended parameters are used for the latter. Furthermore, we show that the optimization of deep learning and iterative methods are trained essentially with the same training gradient with different Jacobian projections, and we conjecture that the training gradient is the bottleneck for accurate registration. This is combined with the observation that the mutual information of the intensity and labelmaps are highly correlated with the Dice score overlap of registered labelmaps using unsupervised registration. These observations imply that no method can provide consistently superior performance in the unsupervised case. This is followed by the behavior of deep learning methods in the supervised case, and whether the improvements in performance lead to gains for out-of-distribution datasets. The results imply that deep learning methods provide a significant benefit only when a lot of auxiliary labelmap data is available, and if the test data distribution does not deviate from the training data distribution - assumptions that are very unlikely to occur in many clinical and translational scenarios. Moreover, other work shows that deep learning for registration trained with labelmap supervision learns implausible deformations to maximize the labelmap score - which makes the weakly supervised registration less attractive. This is followed by rigorous independent evaluation of similar claims made for out-of-distribution generalization across T2w, T2*, FLAIR, 9.4T MRI images in the LUMIR challenge (Chen et al., 2025). We show that the principal conclusions still hold.

FireANTs In the third chapter we introduce FireANTs, a powerful framework for registration of the long-tail of modalities. FireANTs is motivated as a framework that unlike deep learning-based methods, is not limited to certain bespoke registration tasks, and can be applied to a wide range of registration tasks. In most of the registration literature, ill-conditioning is mentioned as a potential problem but never formally addressed. Our work directly addresses and quantifies some of the theoretical properties of the image registration problem - most notably the ill-conditioned nature of deformable registration, which has mostly been speculated but never been explicitly quantified. This strongly motivates the use of first-order adaptive optimization on the space of diffeomorphisms that are represented as integrals of time-dependent flows. Since the space of diffeomorphisms is a Lie group and therefore lies on a non-Eucliean Riemannian manifold, applying adaptive optimization is not trivial. We explore a few alternate designs - specifically using a stationary velocity field to represent the diffeomorphism, and using Riemannian Adam on diffeomorphisms. We identify computational and numerical challenges with each alternative, and propose a novel Eulerian descent formulation to derive adaptive optimization on diffeomorphisms directly. This leads to a state-of-the-art framework that is more accurate than ANTs due to its powerful optimizer, and is highly efficient compared to other iterative and deep learning methods both in terms ofmemory footprint and runtime. FireANTs shows state-of-the-art zero-shot performance on multiple systems, challenge datasets and benchmarks, often beating specialized methods that are designed for bespoke applications across both 2D and 3D registration tasks. FireANTs allows much faster hyperparameter grid searches, template building, and distributed optimization, and its efficiency allows running registration on high-resolution datasets like rodent and zebrafish images.

Deep Implicit Optimization for backpropagation through registration solvers The fourth chapter builds upon the robust and efficient nature of FireANTs to convert FireANTs from an isolated blackbox optimizer

to a fully differentiable solver. We formulate the mathematical framework necessary for enabling a fully differentiable multi-scale registration solver that can be used to learn multi-scale feature extractors using neural networks. These neural networks produce features that are provided to the solver, and the final warp is used to minimize task-specific losses, such as labelmap overlap or landmark distance. Our method, named DIO, allows FireANTs or any iterative solver to be used "in-the-loop" to learn task-specific image features, allowing the best of both worlds - higher fidelity of learned features, with robust performance and convergence guarantees of the iterative solver. This paradigm presents a few more benefits compared to a typical deep learning registration framework - namely, interpretability of image features, hotfixing arbitrary solvers and regularizations at test time, the ability to use auxiliary labelmaps or landmarks available at test time, and superb robustness to out-of-distribution datasets. This work alleviates one limitation of iterative solvers (i.e. the ability to be integrated end-to-end with learned features) by making the solver a *differentiable plug-and-play module*.

Scalable registration for arbitrarily large images The fifth chapter introduces core systems level contributions to scale registration methods to arbitrarily large images. Over the past decade, there has been a tremendous growth in image acquisition capabilities for various biomedical and life science applications, including MRI, CT, PET, microscopy (Balchandani and Naidich, 2015; Esquivel et al., 2022; Badawi et al., 2019; Gambarotto et al., 2019; Wassie et al., 2019). Ultra-high resolution imaging technology has enabled acquisition of images beyond three orders of magnitude larger than macroscopic biomedical domains (Kleven et al., 2023; Wang et al., 2020b; Mansour et al., 2025; Kleinfeld et al., 2011). For instance, a typical clinical scan registration requires solving $\sim$ 20M parameters, while a high-resolution ex-vivo human brain scan requires solving upto 11B parameters. However, current approaches work reliably only at the scale of *macroscopic* biomedical domains ($\sim$ 50M warp parameters) and quickly run out of memory on larger problems due to high computational and memory requirements. This leads to a significant gap in the ability to accurately register images at high resolutions. Under the current status quo, several life science applications use heavily downsampled versions of their acquired image datasets and perform registration on these downsampled datasets, losing many key insights. Our work identifies key computational bottlenecks in a typical image registration pipeline - a composite interpolator, the local normalized cross correlation and mutual information loss functions, and proposes hardware-aware CUDA kernels that reduce the additional memory requirement from $O(n)$ to $O(1)$ for an image with n voxels. This leads to a significant improvement in the capability to register images upto two orders of magnitude larger than existing methods on a single GPU. However, for larger datasets that do not fit on a single GPU, we need additional considerations. Inspired by Model Parallel techniques proposed in the Large Language Models training literature, we propose Grid Parallel, a distributed primitive that allows sharding all variables associated with the optimization problem (moving and fixed images, displacement grid) uniformly across G GPUs. The Grid Parallel allows optimizing the sharded problem without any approximation as if registration were performed on a single GPU. To perform deformable interpolation of the moving image without performing an allgather operation (that would not scale for large problems), we propose a Ring Sampler that uses a principle of superposition to interleave communication of image shards and computation and accumulation of partial bilinear interpolation terms to perform mathematically correct interpolation without imposing any assumptions or constraints on the deformation fields. Our novel operational contributions allows us to perform multimodal registration of a $100\mu m$ human *ex-vivo* MRI scan in a minute, an unprecedented result in large scale registration. Due to the lack of evaluation criteria, we also design a synthetic high resolution dataset and show that our method shows monotonic improvement in performance with increasing resolution, while deep learning methods degrade in performance.

Applications on "in-the-wild" registration tasks The sixth chapter ties all the algorithmic and system-level contributions together to showcase complex end-to-end "in-the-wild" workflows. The first of these workflows is a multi-modal, multi-scale registration of *in-vivo* MRI to histology sections. Such registration workflows require extensive access to intermediate modalities like in-vivo MRI, ex-vivo MRI, blockface photographs, binarized and progressively subtracted blockface image sequences after performing block cuts (Huszar et al., 2023) for initialization of blockface registration, and histology sections. Registration of the histology sections that rely on neuropathology information pertinent to the spatial distribution of protein aggregates-specifically amyloid-beta plaques and tau tangles in AD, or TDP-43 and tau inclusions in FTD — to morphological 3D volumes requires a complex end-to-end workflow that leverages the algorithmic and system-level contributions of the thesis. Our work provides a robust and flexible framework that can be used to perform this registration without the need for manual intervention or annotation-based bookkeeping, allowing researchers to prevent spending enormous labor and extensive resources on sophisticated *ex-vivo* workflows, and instead build flexible workflows that do not compromise on accuracy or performance. A second application is the registration of distorted echo-planar MRI images (EPI) only along the phase-encoding direction, to recover the "true geometry" of the image. Since EPI imaging is fast compared to conventional MRI images at the cost of distortion, this allows rapid acquisition of imaging and recovery of the "true geometry" of the image. A third application is the registration of lung CT scans using a shallow learnable model trained using evolutionary algorithms to minimize landmark error of a sparse set of landmarks, which is hard to achieve using gradient-based methods. Sparse landmark-guided registration is also increasingly popular in highly hetereogenous datasets like mouse brains (Tustison et al., 2024) but integrating this information in intensity-based registration is relatively underexplored. These range of applications show the potential impact of using our contributions to real-world applications.

The material in this thesis has been disseminated in the following articles:

1. **Rohit Jena**, Pratik Chaudhari, and James C. Gee. "FireANTs: Adaptive riemannian optimization for multi-scale diffeomorphic registration." Accepted at **Nature Communications**.

2. **Rohit Jena**, Pratik Chaudhari, and James C. Gee. "Deep implicit optimization enables robust learnable features for deformable image registration." **Medical Image Analysis (2025)**: 103577.

3. **Rohit Jena**, Vedant Zope, Pratik Chaudhari, and James C. Gee. "A Scalable Distributed Framework for Multimodal GigaVoxel Image Registration." In International Conference on Learning Representations **(ICLR 2026 Oral)**.

4. **Rohit Jena**, Deeksha Sethi, Pratik Chaudhari, and James C. Gee. "Deep learning in medical image registration: Magic or mirage?." Advances in Neural Information Processing Systems 37 **(NeurIPS 2024)**.

5. **Rohit Jena**, Pratik Chaudhari, and James C. Gee. "The LU-Mirage - An independent evaluation of the zero-shot claims in the LUMIR challenge", Medical Imaging for Deep Learning **(MIDL 2026)**.

6. Chenyang Li, Huize Pang, Tianyu Gao, **Rohit Jena**, Dominique Leitner, Thomas Wisniewski, Arline Faustin, Henrieta Scholtzova, Laura Gould, Orrin Devinsky, James Gee, Youssef Z Wadghiri, Jiangyang Zhang, Yulin Ge, "Whole-Brain Mapping of Formaldehyde Fixation Using Multiparametric MRI", "International Society for Magnetic Resonance in Medicine" **(ISMRM 2026)**.

7. **Rohit Jena**, Pratik Chaudhari, and James C. Gee. "Factored Levenberg-Marquardt for Diffeomorphic Image Registration: An efficient optimizer for FireANTs", preprint.

Other works that are published and/or disseminated during the PhD, but not part of the thesis are:

1. **Rohit Jena**, Zhornyak, L., Doiphode, N., Chaudhari, P., Buch, V., Gee, J. and Shi, J., 2023. Beyond mAP: Towards better evaluation of instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11309-11318). **(CVPR 2023 Oral)**

2. **Rohit Jena**, Yifan Wu, John C. Greenwood, Pratik Chaudhari, and James Gee. "Automated estimation of microcirculation capillary density using relative perfusion maps." In Medical Imaging 2024: Image Processing, vol. 12926, pp. 188-194. **(SPIE 2024 Oral)**

3. **Rohit Jena**, Ali Taghibakhshi, Sahil Jain, Gerald Shen, Nima Tajbakhsh, and Arash Vahdat. "Elucidating optimal reward-diversity tradeoffs in text-to-image diffusion models." In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (**WACV 2025 Oral**)

4. Min Chen, Nicholas J. Tustison, **Rohit Jena**, and James C. Gee. "Image registration: Fundamentals and recent advances based on deep learning." Machine Learning for Brain Disorders (2023): 435-458. **(book chapter)**

5. Yifan Wu, **Rohit Jena**, Mehmet Gulsun, Vivek Singh, Puneet Sharma, and James C. Gee. "Towards Establishing Dense Correspondence on Multiview Coronary Angiography: From Point-to-Point to Curve-to-Curve Query Matching.", **Data Curation and Augmentation Workshop at CVPR 2024 (Oral)**

6. Changhyun Lee, Juergen Biederer, Yoshiharu Ohno, Joon Beom Seo, Grace Parraga, David L Levin, James C Gee, **Rohit Jena**, Yoshiyuki Ozawa, Mark O Wielpuetz, Eric A Hoffman, Edwin JR van Beek "Functional Lung Imaging Using CT: An Update", **Radiology: Cardiothoracic Imaging**

7. Bailiang Jian, Jiazhen Pan, **Rohit Jena**, Morteza Ghahremani, Hongwei Bran Li, Daniel Rueckert, Christian Wachinger, Benedikt Wiestler. "Disentangling progress in medical image registration: Beyond trend-driven architectures towards domain-specific strategies" **submitted to Medical Image Analysis.**

8. Yue Li, Pulkit Khandelwal, **Rohit Jena**, Long Xie, Michael Duong, Amanda E Denning, Christopher A Brown, Laura EM Wisse, Sandhitsu R Das, David A Wolk, Paul A Yushkevich. "Achieving detailed medial temporal lobe segmentation with upsampled isotropic training from implicit neural representation". **Submitted to NeuroImage.**

9. Rachel Blue, Nehal Doiphode, Rohit Jena, Peter Madsen, John YK Lee, Jianbo Shi, Vivek Buch "Designing and Developing a Novel Deep Computer Vision Platform for Intraoperative Prediction and Analytics in Skull Base Surgery", **Journal of Neurological Surgery Part B: Skull Base**

10. **Rohit Jena**, Ganesh Subramanian Iyer, Siddharth Choudhary, Brandon Smith, Pratik Chaudhari, James Gee. "Splatarmor: Articulated gaussian splatting for animatable humans from monocular rgb videos". Preprint.

CHAPTER 2

An Empirical Study and Evaluation of Image Registration paradigms

In this chapter, we go beyond a comprehensive literature review and provide a unified formulation to view the two prevailing paradigms of image registration. In deformable image registration, the two prevailing paradigms are iterative optimization and deep learning. While classical optimization-based methods are based on solving a variational optimization problem, deep learning-based methods are based on learning a deep network to formulate the inverse optimization problem as a statistical learning problem. While deep learning methods provide good performance on in-distribution datasets, their large scale clinical adoption is limited due to their brittle performance on out-of-distribution datasets. We first examine the performance of classical and deep learning methods on a variety of datasets, with and without labelmap supervision, and whether the performance transfers to out-of-distribution datasets. The performance of *supervised* deep learning methods performing poorly on OOD datasets compared to their own unsupervised counterparts has spurred a lot of research to develop unsupervised image registration benchmark datasets (Chen et al., 2025) and to develop a new class of image registration networks (Liu et al., 2024c; Honkamaa and Marttinen, 2023; Jian et al., 2025) that improve domain generalization capabilities. We study some of the recent claims made in the literature, and provide a holistic view of the image registration evaluation landscape.

2.1. The two prevailing paradigms of Deformable Image Registration

Classical optimization-based and learning-based methods are the two reigning paradigms in deformable image registration (DIR). Classical image registration methods are based on solving a variational optimization problem, where a similarity metric is optimized to find the best transformation that aligns the images. Most classical methods are formulated without any particular domain knowledge encoded in the optimization problem, and are therefore general and applicable to a wide range of problems. For instance, the popularly known registration toolkit ANTs (Avants et al., 2009) has been successfully applied to structural *and* functional neuroimaging data (Klein et al., 2009; Yassa et al., 2010; Jiang et al., 2013), CT lung imaging (Murphy et al., 2011), cardiac motion modeling (Likhite et al., 2015), developmental mouse brain atlases utilizing MRI and light sheet fluorescence microscopy (Kronman et al., 2023) with virtually no change in the optimization algorithm. Due to their generality, flexibility, and robustness to varying degrees of imaging, these methods are well poised to tackle a wide range of clinical and research problems alike. Most clinical workflows routinely use tools like SPM, Elastix, and ANTs for image registration (Lüsebrink et al., 2017; Tustison et al., 2019).

Most classical iterative methods were developed before the widespread adoption of deep learning, and therefore, have not been adapted to leverage the benefits of deep learning. Specifically, their multi-threaded CPU implementations have slow convergence which has been a major bottleneck for their adoption in larger-scale studies. Since classical methods rely on photogrammetric loss functions for similarity matching, their performance is limited by the fidelity of image intensities, and they cannot incorporate learning to leverage a training set containing weak supervision such as anatomical landmarks, label maps or expert annotations. Deep Learning for Image Registration (DLIR) is an interesting paradigm to overcome these challenges. DLIR methods take a pair of images as input to a neural network and outputs a warp field that aligns the images, and their associated anatomical landmarks. The neural network parameters are trained to minimize the alignment loss over image pairs, label maps, and landmarks from a training set. During inference, an image pair is provided and the network predicts a warp field, with the parameters of the network encoding the transform to apply to align the image and labelmaps. A primary benefit of this method is the ability to incorporate

weak supervision like anatomical landmarks or expert annotations during training, which performs better landmark alignment without access to landmarks at inference time. This is a notable paradigm shift where image registration requires task-awareness via explicit correspondence matching of labelmaps and landmarks that have semantically meaningful information.

However, several recurring patterns in recent work complicate direct comparison between these paradigms. First, since most publications in recent literature propose to employ deep networks for this task, they typically report the performance of iterative optimization methods by using suboptimal parameters. For example, VoxelMorph (Balakrishnan et al., 2019) compares performance with ANTS using regularization parameters and number of iterations that are different than the recommended ANTs parameters by an order of magnitude. Even so, VoxelMorph reports performance that is only slightly worse than ANTs, and presents *amortized optimization* (i.e. spend a lot of time and compute to predict the warp field for a dataset simultaneously and perform quick inference) as the primary benefit of deep learning methods over classical methods. Other methods unlocked the ability of deep learning methods to maximize ROI overlap as part of the amortized optimization process (Mok and Chung, 2020c), outperforming than classical methods that operated on photometric similarity alone. However, this trend started to emerge in other *unsupervised* deep methods (Jia et al., 2022; Mok and Chung, 2020b; Chen et al., 2022b) without clear theoretical justification. The proposition in these methods is that progressively "advanced computational blocks" are the primary driving factor for the performance improvements over an iterative solver with convergence and minimization guarantees. We re-examine these claims in this chapter. Deep learning methods also claim to provide amortized optimization since classical methods are extremely slow to run, however, modern GPU implementations (Mang et al., 2019b; Siebert et al., 2024; Modat et al., 2010a; Jena et al., 2026) have overcome this shortcoming of classical methods while providing state-of-the-art performance. Recent studies have even claimed that simply training a deep network on T1-weighted images can zero-shot generalize to other modalities such as T2-weighted or ultra high-field (UHF)images (Chen et al., 2025), a claim that defies established statistical learning theory and is grounded neither in theory nor in practice. In this chapter, we show that almost all empirical justifications for the superiority of DLIR methods over classical methods suffer from instrumentation bias (Tustison et al., 2013), and are therefore based on a misrepresentation of the performance of classical methods. We also provide a unified formulation to view the two paradigms of image registration, and recipes for fairly evaluating both deep learning and iterative methods that are more aligned with real-world clinical practice.

2.2. A unified formulation of optimization and deep learning image registration algorithms

We rehash the image registration problem statement to unify both classical and deep learning methods. Consider a dataset of image pairs $\mathcal{D} = \{(I_f, I_m) \mid f \in \{1, 2, \ldots, N_f\},\ m \in \{1, 2, \ldots, N_m\}\}$, where I_f and I_m are the fixed and moving images defined over a spatial domain $\Omega \in \mathbb{R}^d$, and f and m are the indices of the fixed and moving images in the dataset. Also consider segmentation maps S_f and S_m for the fixed and moving images, respectively, defined over Ω. Given a family of transformations $T(\Omega)$, the goal of image registration is to estimate transformations $\varphi_\theta(f, m) \in T(\Omega)$ parameterized by θ that minimize the following objective:

$$\arg\min_{\theta} \sum_{f,m} \mathcal{L}(I_f, I_m \circ \varphi_\theta(f, m)) + \mathcal{R}(\varphi_\theta(f, m)) \tag{2.1}$$

where $\mathcal{L}$ is a dissimilarity function such as mean squared error, or negative local cross correlation, and $\mathcal{R}$ is a regularization term that encourages desirable properties of the transformation, such as smoothness or elasticity. We call Eq. (2.1) the *image matching* objective, since the transformations only need to align the

intensity images. We can also call this the *unsupervised* objective, since it does not require any labeled data. If a suitably chosen label alignment loss $\mathcal{D}$ is added as well, the optimization problem becomes:

$$\arg\min_{\theta} \sum_{f,m} \mathcal{L}(I_f, I_m \circ \varphi_\theta(f,m)) + \mathcal{D}(S_f, S_m \circ \varphi_\theta(f,m)) + \mathcal{R}(\varphi_\theta(f,m)) \tag{2.2}$$

We call Eq. (2.2) the *label matching* objective, or a *weakly-supervised* objective. The image matching objective can subsume both DLIR and classical methods by choosing

$$\varphi_\theta(f,m) = \begin{cases} f_\theta(I_f, I_m), & \text{for deep networks,} \\ \varphi_{(f,m)}, & \text{for classical methods.} \end{cases} \tag{2.3}$$

where f_θ is a deep network parameterized by θ and $\varphi_{(f,m)}$ are optimizable free parameters that are indexed by the 2-tuple (f,m), i.e. $\theta = \bigcup_{f,m}\{\varphi_{(f,m)}\}$.

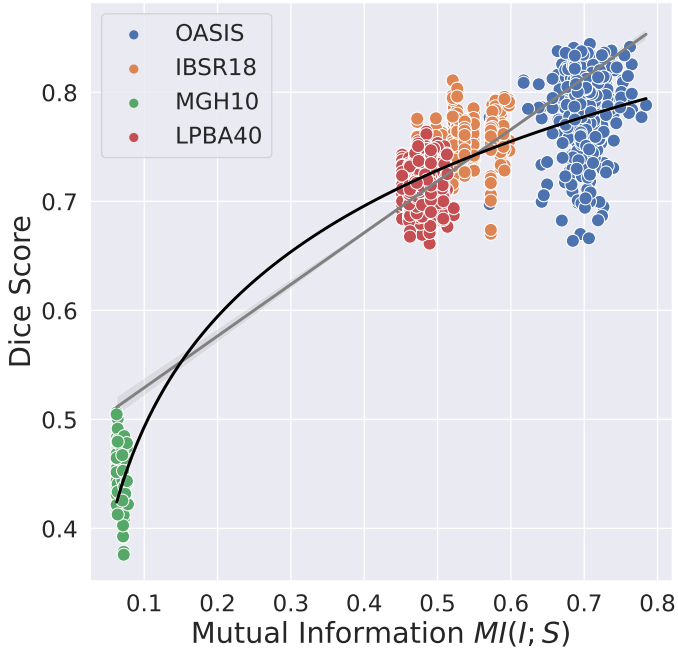


Figure 2.1: Correlation between Dice Score and Mutual Information. Classical registration methods like ANTs show a strong correlation between the Dice Score of registered pairs, and the mutual information between the corresponding image and label across 4 brain datasets.

In this chapter, we consider methods that solve Eq. (2.1) using gradient-based methods. The gradient of Eq. (2.1) with respect to θ is given by (we remove the $\mathcal{R}$ term for simplicity):

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{f,m} \frac{\partial \mathcal{L}}{\partial \varphi_\theta(f,m)} \frac{\partial \varphi_\theta(f,m)}{\partial \theta} \tag{2.4}$$

The first term $\frac{\partial \mathcal{L}}{\partial \varphi_\theta(f,m)}$ is the training signal from the dissimilarity function which does not depend on the parameters θ for a given value of $\varphi_\theta(f,m)$ and choice of $\mathcal{L}$. The second term $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$ is the Jacobian of the

transformation with respect to the parameters, which is a projection of the gradient from the space of warp fields to the space of arbitrary parameters. For classical methods, the Jacobian is the identity matrix, for deep networks it is determined by the functional relationship of the output with respect to network parameters. Therefore, the difference in training dynamics and overall performance gap between classical and deep learning methods is likely to be attributed to the choice of $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$.

Mutual Information between image intensity and label maps is a good proxy for Label Matching Performance. Image matching objectives ensure that intensities from the moving image are displaced to locations in the fixed image where they are most similar, without regard for alignment for any higher order structures. Intuitively, this will ensure label matching only to the extent that the intensity is predictive of the label. If an intensity value strongly corresponds to a particular label, then image matching will lead to label matching. Similarly, if a given intensity value corresponds to multiple possible labels, then image matching does not tell us which labels are matched via the image matching objective. More formally, considering the per-pixel intensity i and labels s as random variables, one can compute the mutual information between the intensity and label maps, denoted as $MI(i;s)$ to determine the predictability of one from the other. We now show that the label matching performance of classical methods is highly correlated with $MI(i;s)$. We consider a widely used classical method, ANTs (Avants et al., 2008a, 2009), to eliminate the effect of any Jacobian term. We consider four brain datasets - OASIS, LPBA40, MGH10, and IBSR18, which are acquired under different scanners, under different resolutions, and have different preprocessing, labelling and postprocessing protocols (Marcus et al., 2007b; Klein et al., 2009). For each dataset, we use ANTs for registering all pairs within the dataset and then evaluate the Dice score as an indicator of label matching performance. For each image I and its corresponding label map S, we compute the probability maps $p(i), p(s), p(i,s)$ using histogram binning, followed by the mutual information $MI(i;s) = H(s) - H(s|i)$. A Pearson's correlation coefficient between the Dice scores and the mutual information of the image and label (Fig. 2.1) reveals a strong linear ($\mathbf{r = 0.886}$) and logarithmic ($\mathbf{r = 0.933}$) relationship between the two quantities, shown by the gray and black lines respectively. Image matching improves label matching performance *only to the extent of the information about the label obtained from the image* (i.e. $MI(i;s)$). At a first glance, the Jacobian term $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$ seemingly does not have a role in improving this mutual information further.

2.3. Instrumentation Bias in Image Registration

If the performance of a method is highly correlated with $MI(i;s)$, then the Jacobian projection term does not presumably provide any additional benefit beyond operating on the assumption that the task-relatedness of different image pairs will lead to a Gestalt effect during inference. However, checking this hypothesis requires careful and fair evaluation of the performance of iterative methods. We demonstrate that most state-of-the-art deep learning methods exhibit significant instrumentation bias in their reported performance of tools like ANTs and NiftyReg. Acknowledging instrumentation bias is important (Tustison et al., 2013) because the evaluation in challenge datasets may be significantly different than how a practitioner uses the methods in clinical and research workflows. Primary sources of instrumentation bias include:

- **Choosing suboptimal or non-recommended hyperparameters**: Most iterative optimization methods provide recommended hyperparameters for common tasks like in-vivo neuroimaging MRI, pulmonary CT, and cardiac sine MRI. However, several DLIR works choose to use suboptimal or non-recommended hyperparameters, which may be done to tradeoff accuracy for speed, or due to lack of expertise. However, this leads to subtle to significant misrepresentation of the performance of classical baselines. For instance, (Balakrishnan et al., 2019) mention that the default parameters of ANTs are not optimal, and choose a very

different set of parameters (a Gaussian smoothing of 9 pixels, followed by an extremely small 0.4 pixels at the next scale). By stark contrast, we found the recommended parameters to work extremely well for all datasets considered in this chapter (see Table 2.1).

- **Running multimodal registration with unimodal similarity functions**: Iterative solvers will catastrophically fail if multimodal images are attempted to be registered using unimodal losses. For example, the Ultracortex dataset (Mahler et al., 2024) contains a mix of MP-RAGE and MP2RAGE sequences for different subjects, which are qualitatively and quantitatively distinct in terms of contrast and resolution. Our independent evaluation on this dataset for iterative optimization methods considers the effect of choosing different similarity functions for multimodal registration.

- **Evaluating registration algorithms on low resolution images**: Most modern registration challenges downsample the data into a standard isotropic resolution and attempt to fit their training data, due to high memory requirements. Two major benefits of using iterative optimization methods is their very low memory footprint at inference, and that they perform *better* at high resolutions. In almost every practical scenario, a practitioner would desire the images to be registered at the highest resolution possible to obtain high fidelity warp fields. Running optimization solvers on downsampled resolutions therefore constitutes *intentional weakening* of the baseline. Instead, DLIR proponents must focus on improving the capabilities of deep learning to register large scale images, rather than weakening optimization solvers.

- **Labelmap bias due to non-existent intensity boundaries**: In the LUMIR challenge, SLANT (Huo et al., 2019) is used for labelling, which uses the BrainCOLOR protocol to obtain a comprehensive segmentation of the brain. However, cortical parcellation is performed by lifting the atlas to the subject coordinate frame; the cortical boundaries do not exist as intensity features in the in-vivo MRI. This leads to spurious results when Dice score is computed due to the low mutual information between the image and label maps, already discussed in Section 2.2. For in-vivo intensity images, cortical and subcortical structures that *can* be delineated must be included in evaluation.

On several state-of-the-art DLIR methods, we observed markedly better results (Fig. 2.3) for classical baselines than reported in the literature simply by using their recommended scripts. We compare the discrepancy in performance between the baselines reported in the literature and the ones we obtained in Table 2.1. We follow the guidelines in (Tustison et al., 2013) to evaluate all methods. To ensure our work does not introduce its own instrumentation bias for DLIR baselines, we compare the performance of our trained/pretrained models to the ones reported in the literature (Table 2.1). In the LUMIR dataset, we observe a few sources of instrumentation bias related to selection of test datasets and generated label maps for evaluation. The LUMIR challenge claims to perform evaluation on a variety of resolutions, but the text mentions that all datasets are resampled to the 1mm MNI space, essentially discarding the effect on performance due to the varying resolution of the datasets. Moreover, we find that registering the PRIME-DE dataset to the MNI template is somewhat unjustified. Consequently, our evaluation adopts a different preprocessing protocol which leads to improved performance of the deep learning method and a smaller discrepancy in performance between deep and iterative methods for this particular dataset. Furthermore, the LUMIR challenge uses SLANT (Huo et al., 2019) that uses a deep learning model to segment a T1 MRI scan into 133 labels based on the BrainCOLOR protocol (Klein et al., 2010). The SLANT network is trained only on 45 T1-weighted MRI scans from the OASIS dataset, which is already markedly different than many other T1w datasets. This training set bias can lead to a significant lack of generalization to other modalities like T2w, T2*, FLAIR, and Ultra High Field (UHF) MRI. While label fusion can help, systematic bias in the multi-atlas fusion step can propagate into the

Evaluation of classical methods reported by baselines					
Method	**Evaluated Baseline**	**Statistic**	**Reported value**	**Our eval**	**Difference**
SymNet	ANTs	Mean	0.680	0.787	0.107
PIRATE	ANTs	Mean	0.699	0.787	0.088
LapIRN	Demons	Mean	0.715	0.802	0.087
LapIRN	ANTs	Mean	0.723	0.787	0.064
NODEO	Demons	Mean	0.764	0.802	0.038
NODEO	ANTs	Mean	0.729	0.787	0.058
Voxelmorph	ANTs	Mean	0.749	0.787	0.038
Voxelmorph	NiftyReg	Mean	0.755	0.776	0.021
SynthMorph	ANTs	Median	0.770	0.797	0.027
Evaluation of DLIR baselines reported by us					
Method	**Dice supervision**	**Statistic**	**Reported value**	**Our eval**	**Difference**
SynthMorph	-	Median	0.780	0.785	0.005
TransMorph-Regular	✓	Mean	0.858	0.855	-0.003
LKU-Net	✓	Mean	0.886	0.904	0.018
LapIRN	✗	Mean	0.808	0.788	-0.020
SymNet	✗	Mean	0.743	0.748	0.005

Table 2.1: Instrumentation bias in evaluation of image registration algorithms. We highlight a significant difference in evaluation metrics reported by baselines and our evaluation on the OASIS validation dataset. This difference can be attributed to deviation in hyperparameters from the recommended parameters or early stopping to save time. In either case, this misrepresentation leads to incorrect conclusions about the performance of the algorithm. The reported dice scores are anywhere from 2 to 10 Dice points lower than our evaluation, showing a non-trivial instrumentation bias. We report our own evaluation of DLIR algorithms and compare them with reported values to avoid introducing instrumentation bias in our evaluation.

learned UNet. We observe degradation in performance of the SLANT algorithm on the Ultracortex dataset. Therefore, a different labelling protocol must be chosen that is (a) robust to variety of contrasts, (b) represent intensity-based boundaries such as gray/white matter boundaries and subcortical structures, and (c) is not biased towards a particular dataset.

In the next section, we validate some behaviors of both unsupervised and supervised DLIR methods, address the instrumentation bias in the LUMIR challenge, and highlight their implications in the context of realistic and practical clinical and research scenarios.

2.4. Supervised DLIR methods do not lead to better domain generalization

One of the benefits of DLIR methods is that their representation allows learning to leverage label maps as extra supervision to improve performance on unseen data. This is a key strength over using photometric similarity alone, since most studies typically perform morphometric analysis on certain anatomical regions of interest. When label matching is introduced as an objective in Eq. (2.2), DLIR methods show superior performance than classical methods. Unlike the previous discussion, where only a pixelwise definition of $MI(i; s)$ was used to quantify the coaction of image intensities and label maps, we consider the entire image I and label volume S as high-dimensional random variables. Label maps are now a deterministic function of the image, i.e. $S = f(I)$, where f is the labelling protocol. In addition to image intensity, label maps are a function of morphological features, location, contrast, and the labelling protocol itself – properties that are beyond the

scope of photometric similarity alone. When trained with the label maps as extra supervision, the network can infer these deterministic relationships to output a warp field that maximize both image similarity and label overlap. Classical intensity-based methods, on the other hand, do not have any mechanism to encode this additional relationship. Aligning intensities or intensity patches discards any functional relationship between high-level image features and labels.

Empirical Validation. We verify this claim empirically on the OASIS dataset, by minimizing Eq. (2.1) in both DLIR and classical methods. We split the OASIS dataset into a training set of 364 images and a validation set of 50 images. We choose 50 instead of 20 images as in the original split (Hering et al., 2022) to compute statistical significance. Dice score over 35 subcortical structures is used as the label matching metric. We choose SynthMorph (Hoffmann et al., 2021), LapIRN (Mok and Chung, 2020c), SymNet (Mok and Chung, 2020b), LKU-Net (Jia et al., 2022) and TransMorph (Chen et al., 2022b) as state-of-the-art DLIR baselines and ANTs (Avants et al., 2009), NiftyReg (Modat et al., 2010a), Symmetric Log Demons (Vercauteren et al., 2009), Greedy (Yushkevich et al., 2016), FireANTs (Jena et al., 2026) as state-of-the-art classical baselines. For all DLIR methods, we use pretrained models if they are trained with Eq. (2.1), or train them with the architecture and hyperparameters provided in their original source code. The only exception is SynthMorph, which is trained on synthetically generated data and Dice loss of its corresponding synthetic labels (`shapes-sm` model). To compare SynthMorph's domain generalization capabilities with only the image matching objective, we add another model, dubbed '`shapes-sm-ncc`' that is trained on synthetically generated data as in the original pretrained model, but with the normalized cross-correlation of the synthetic images. For all classical methods, we follow their recommended hyperparameters and run till convergence. All experiments are run on a cluster with 2 AMD EPYC 7713 CPUs and 8 NVIDIA A6000 GPUs.

Results. Fig. 2.2(top) shows the Dice scores for supervised classical and DLIR methods trained on the OASIS dataset, sorted by median validation performance. In this case, state-of-the-art DLIR methods outperform classical methods by a large margin, with notably higher Dice score on the *trainval* set than the *val* set, due to overfitting to the label matching for the training set. These differences are statistically significant, with the exception of SymNet, which diverged under many training settings with the Dice loss, and only works marginally better than its unsupervised counterpart. SynthMorph is not trained on real data, and is added only as a reference for domain-agnostic performance.

This is an unsurprising result – the label matching objective provides additional training signal to the registration task, which is a highly ill-posed problem. Classical methods cannot incorporate this additional signal from a training dataset, and learning-based methods exploit this to achieve better registration on unseen data. Classical methods are, however, agnostic to modalties, intensity distributions, voxel resolutions, and anisotropy. The same registration algorithm (with possibly modified parameters) is applied to datasets with different characteristics, and they still retain their state-of-the-art performance. A related question arises for DLIR methods trained with label matching – does label matching performance transfer to other datasets?

Supervised DLIR methods do not generalize across datasets A key strength of classical optimization registration algorithms is their agnostic nature to the image modality, physical resolution, voxel sizes, and preprocessing protocols. Most DLIR methods, on the contrary, have been evaluated extensively on the same distribution of validation datasets as the training data, it is unclear if the performance improvements transfer to other datasets of the same anatomy. To this end, we evaluate the performance of both the classical and DLIR methods on four brain datasets – CUMC12, LPBA40, MGH10, and IBSR18. These datasets represent community-standard brain mapping challenge data (Klein et al., 2009) for a comprehensive evaluation of 14 nonlinear classical registration methods, across various acquisition, preprocessing and labelling protocols.

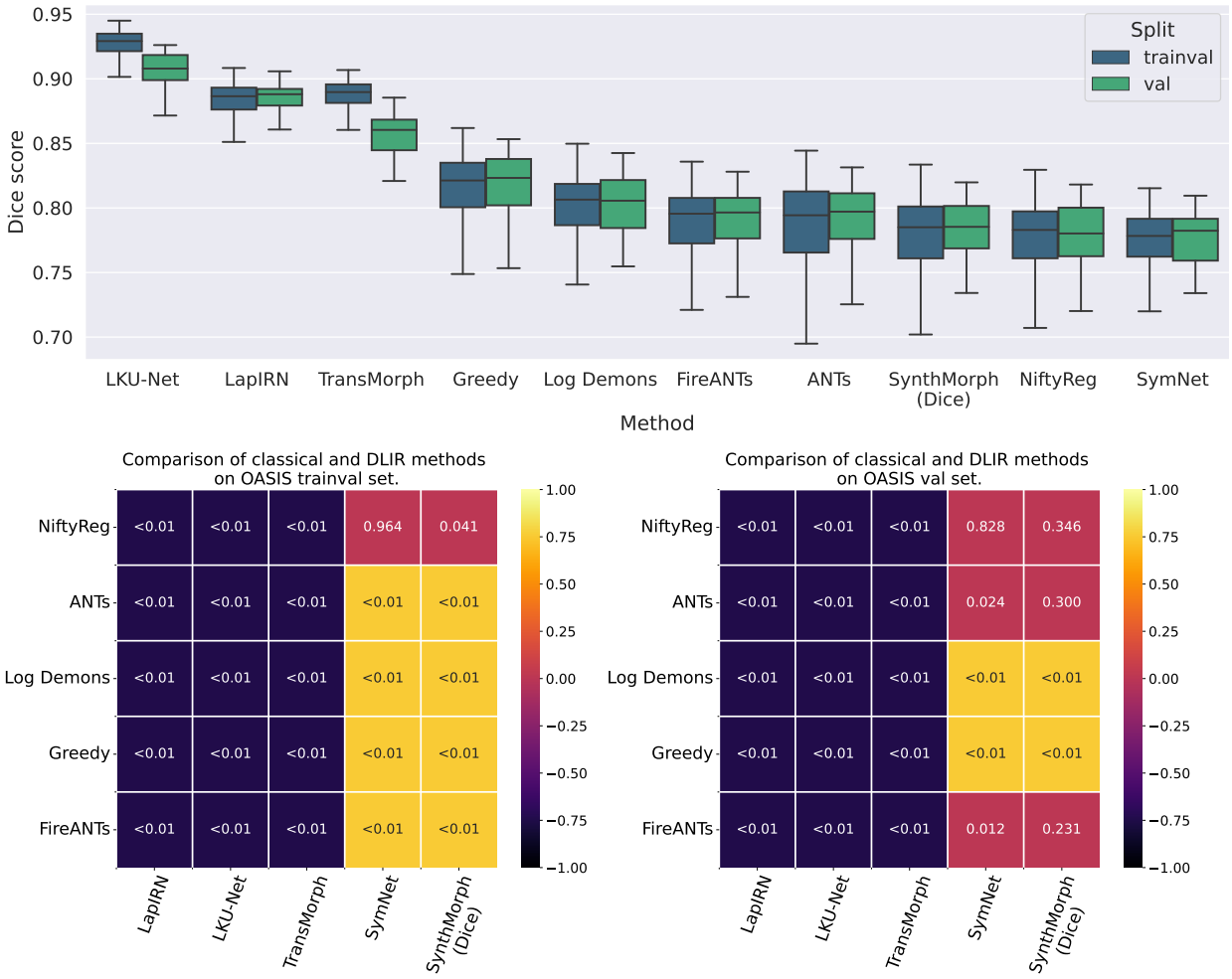


Figure 2.2: Performance of classical and supervised DLIR methods on OASIS data. Boxplots (**top**) show that DLIR methods show superior performance compared to classical methods. Unlike the unsupervised case, the effect of overfitting is clearly visible in the gap between the *trainval* and *val* splits. Tables (**bottom**) of p-values show the results of a pairwise two-sided t-test between the performance of classical and DLIR methods on the *trainval* and *val* splits. ■ denotes a cell where the classical method is significantly better than the DLIR method ($p < 0.01$), a ■ denotes the opposite, ■ denotes no significant difference. State-of-the-art DLIR methods show significantly better performance than classical methods when label supervision is added.

Each dataset contains a different set of labeled regions acquired manually using different labeling protocols. For each dataset, all previously considered registration algorithms are run on all image pairs, and the mean Dice score over all labeled regions is computed. The methods are then sorted by median validation performance in Fig. A.1. For DLIR methods, we plot the performance with models trained with and without the label matching loss in the OASIS dataset, shown as blue and green boxplots respectively. Across all datasets, FireANTs, Greedy, ANTs and NiftyReg consistently perform better than DLIR methods, regardless of whether they are trained with or without the label matching loss. Among the DLIR methods, SynthMorph performs consistently better due to its domain-agnostic training paradigm. Remarkably, even though DLIR methods outperform classical methods on the OASIS dataset with label matching objective, the performance does not transfer to other datasets, even compared to its own unsupervised variant. This is a negative result – implying that to improve performance on a new dataset, one must collect label maps from that dataset and retrain the model – existing collections of label maps are not sufficient to improve performance on new datasets.

2.5. Properties of unsupervised DLIR methods

In the previous section, we showed that although DLIR methods show superior performance on the in-distribution dataset, they do not generalize to other datasets. This is problematic because the method might be trained on brains from a particular scanner, resolution, and preprocessing protocol, and might be deployed elsewhere with different imaging characteristics. Moreover, convolutional networks like Jia et al. (2022); Mok and Chung (2020c,b) showed increased propensity to produce implausible deformations with supervised training (Jian et al., 2024; Liu et al., 2025a), which are not observed in classical methods. This makes convolutional networks unsuitable for clinical deployment. A new generation of DLIR methods has emerged that moves away from ‘‘advanced computational deep learning blocks’’ and towards more registration-specific designs (Jian et al., 2024; Liu et al., 2025a; Jian et al., 2025). The limitation of supervised deep learning methods has also prompted the development of large-scale unsupervised registation challenges like **LUMIR** (Chen et al., 2025) focusing on zero-shot domain generalization which are highly beneficial for clinical deployment.

While these developments are highly promising, we show that the claims of the current state-of-the-art methods are often overstated and are not supported by empirical evidence. We explore some of these properties in this section.

2.5.1. Unsupervised DLIR does not improve label matching performance over iterative optimization

Deep learning methods posit that they can provide better label matching performance on a given dataset by training a network to minimize Eq. (2.1) in an unsupervised setting. Such improvements are claimed to come from architectural designs, which correspond to choice of Jacobian $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$. A variety of architectures and parameterizations (Chen et al., 2022b; Mok and Chung, 2020c, 2021, 2022; Heinrich et al., 2015; Teshima et al., 2020; Wu et al., 2022a) have been proposed to this effect. First, we show that for convolutional networks, this is indeed not the case. To show this, we use the same experiment setup as in Section 2.4 on the same splits, but only with the image matching objective. Next, we show that the zero-shot domain generalization capabilities of the top performing methods in the LUMIR challenge (Vector Field Attention (Liu et al., 2024c)) are not *superior* to classical methods.

Results on OASIS dataset.

For all methods, we compute the Dice score of all 35 subcortical regions on images in the validation set (denoted as *val*), and all images (denoted as *trainval*). These Dice scores are sorted by median validation performance in Fig. 2.3(top). Moreover, we perform a two-sided t-test for each *(classical, DLIR)* pair, both on the trainval and validation sets, shown in Fig. 2.3(bottom). Fig. 2.3 shows the following conclusions: (a) the top performing classical method (Greedy) and the top performing DLIR method (TransMorph) achieve similar label matching performance on the val and the trainval set, i.e. the differences are *not* statistically significant ($p = 0.161$), (b) classical methods almost always perform better than DLIR methods, even on the training set showing that the Jacobian term does not improve label matching more than the mutual information between the image and label, and (c) for unsupervised DLIR methods, there is no improvement label matching performance in the training set compared to val set. The only role of the Jacobian term is to perform amortized learning, provide implicit regularization through the parameterization and architecture choice, and possibly leverage any potential task-relatedness of the image pairs in the dataset. However, without supervised objectives, this does not guarantee any additional boost in label matching.

Results on NIMH T1w dataset.

The LUMIR challenge shows zero-shot evaluation on a variety of datasets spanning different contrasts, two species, and three tasks (inter-subject, atlas-to-subject, and subject-to-atlas registration). However, the labeled data generation and evaluation are not discussed in sufficient detail to ensure reproducibility. There are also few oversights in the dataset descriptions and evaluation criteria that we discuss, and consider their effect in our evaluation. For each dataset, we also consider the primary sources of instrumentation bias that can affect evaluation, and how we control for these conditions.

Baselines. Chen et al. (2025) predicate that the new generation of deep learning architectures surpass optimization solvers on all registration tasks. To carefully evaluate this claim, we consider independently evaluating the top performing methods ranked in Table 1 of the LUMIR challenge paper, with FireANTs (Jena et al., 2026) which is reported as the best performing iterative solver. We will discuss more about FireANTs in the next chapter. However, at the time of writing this chapter, out of the top eight best performing methods, only *two* implementations are available in the public domain: SITReg (Honkamaa and Marttinen, 2023) (Rank 1) and Vector Field Attention (Rank 4) (Liu et al., 2024c). Despite SITReg providing an open-source implementation, it does not provide user friendly interfaces for evaluation on arbitrary datasets and despite our best efforts at modifying the codebase, we could not run the trained model on our evaluation setup. VFA on the other hand provided highly customizable configurations that allowed us to seamlessly run evaluations with minimal changes to the original codebase. FireANTs (Rank 12) provides both CLI-based and Python-based scripts for evaluation of arbitrary datasets, and we use the Python-based script for consistency. Therefore, we use VFA as the primary deep learning method for comparison with FireANTs.

Dataset Description. The National Institute of Mental Health (NIMH) Data Archive uses human subject data collected from hundreds of research projects across many scientific domains. We use the Research Volunteer Dataset that characterizes healthy adult research volunteers in clinical assessments using mood-related psychometrics, cognitive function neurophysiological tests, structural and functional MRI, DRI, and MEG. We use a subset of the T1w MRI dataset for inter-subject registration.

T1w MRI provides excellent gray/white matter contrast, and is routinely used for structural segmentation, and

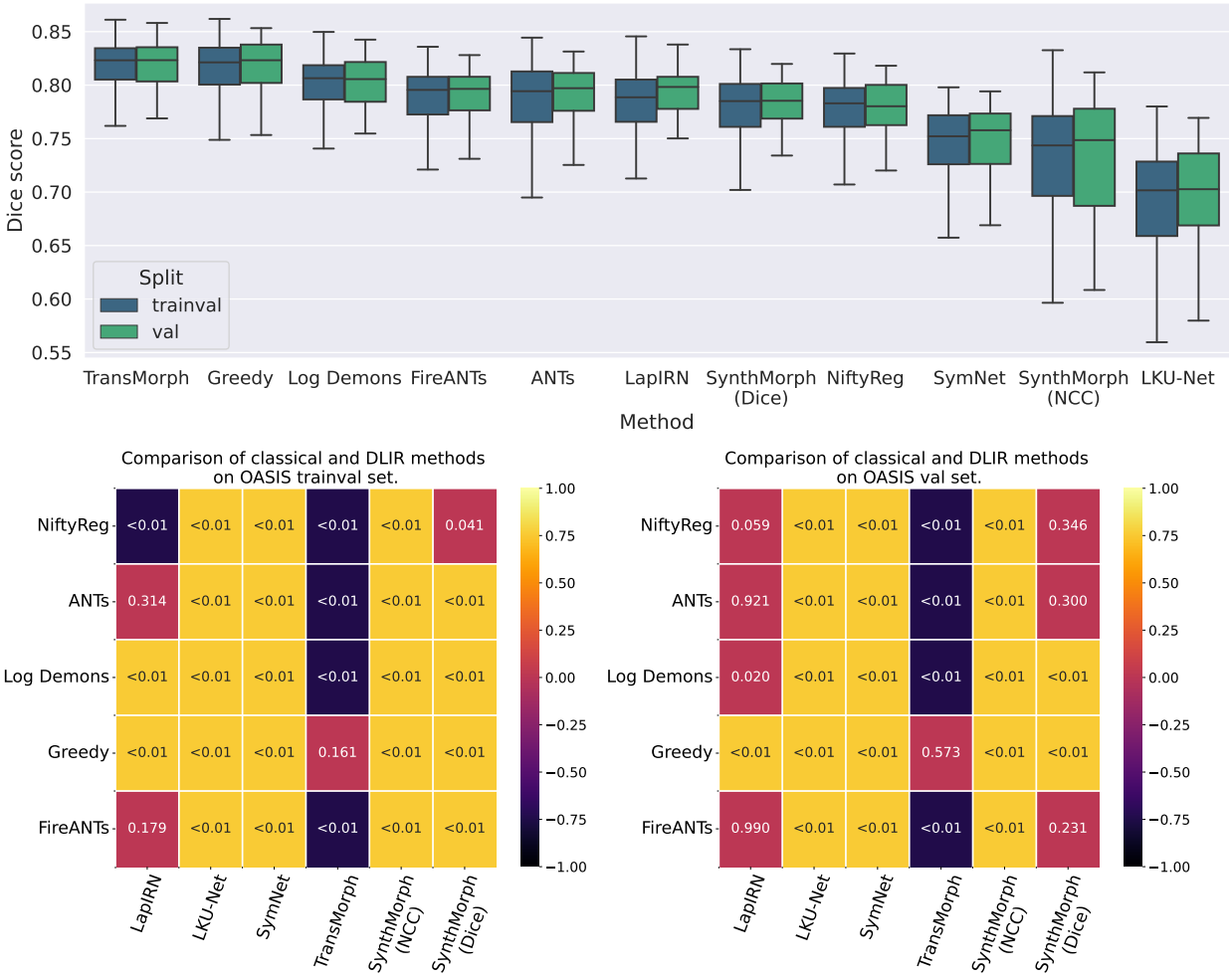


Figure 2.3: Performance of classical and unsupervised DLIR methods on OASIS data. Boxplots (**top**) show that classical methods on average are ranked higher than DLIR methods, both on the *trainval* and *val* splits. Interestingly, the performance of unsupervised DLIR methods does not improve on the *trainval* split compared to *val* split – showing that deep learning does not have an intrinsic advantage in label alignment. Tables (**bottom**) of p-values show the results of a pairwise two-sided t-test between the performance of classical and DLIR methods on the *trainval* and *val* splits. denotes a cell where the classical method is significantly better than the DLIR method ($p < 0.01$), a denotes the opposite, denotes no significant difference. Most of the cells are , indicating that classical methods are significantly better than DLIR methods.

morphometry. Similar to most methods, we use overlap of cortical and subcortical structures for evaluation. The LUMIR challenge's evaluation protocol uses a tool called SLANT (Huo et al., 2019) that uses a deep learning model to segment a T1 MRI scan into 133 labels based on the BrainCOLOR protocol (Klein et al., 2010). However, there are a few issues with using SLANT for evaluation. *First*, SLANT produces 133 labels, but is trained only on 45 T1-weighted MRI scans from the OASIS dataset. This can lead to a significant lack of generalization to other modalities like T2w, T2*, FLAIR, and Ultra High Field (UHF) MRI. While label fusion can help, systematic bias in the multi-atlas fusion step can propagate into the learned UNet. We observe degradation in performance of the SLANT algorithm on the Ultracortex dataset. *Second*, measures like Dice Scores are highly sensitive to the volume of the structures, and consequently the choice of interpolation method can significantly affect the score. For example, in our experiments, changing the interpolation method from trilinear to nearest neighbor in VFA leads to a drop of about 10 points in Dice score for the SLANT labelmaps. To ensure fair comparison, we fix the label interpolation scheme wherein we first convert each labelmap to a binary mask, perform trilinear interpolation to obtain probability maps for each label, and for each voxel select the label with the highest probability. This interpolation scheme avoids blocky artifacts introduced by nearest neighbor interpolation, considers partial volume effects of the probability maps, and assigns a single label to each voxel. *Third*, our previous work Section 2.2 shows that the mutual information between images and label maps is correlated with Dice Score of registration. The BrainCOLOR protocol used in SLANT provides extensive fine grained structures including sulcal/gyri boundaries, and various lobe boundaries, which cannot be delineated by intensity features alone. This can lead to Dice scores of registration methods capturing spurious associations rather than anatomical relationships that can be delineated by intensity features since we are interested in evaluating intensity-based registration methods.

Evaluation. To address all these issues, we choose three labelling protocols with varying degrees of granularity and anatomical coverage: (1) SLANT as used in the original LUMIR challenge, (2) SynthSeg (Billot et al., 2023) for a comprehensive segmentation of various subcortical structures while segmenting the cerebral cortex as a single label for each hemisphere, and (3) DeepAtropos (Tustison et al., 2021) that provides a coarse six label segmentation of CSF, GM, WM, deep GM, brainstem, and cerebellum. We randomly choose 100 subjects from the dataset, resample to 1mm isotropic resolution, and apply all three segmentation protocols to obtain labelmaps. This provides us a total of 9900 image pairs for evaluation. To provide robust estimates, we crop the bottom five percentile of the Dice scores for each registration method. We provide common statistical measures (mean, median, standard deviation) for the Dice scores of the three registration methods in Table 2.2, and violin plots in Figure 2.4.

Significance Tests. To evaluate the practical impact of the differences in labelling protocols, we perform a paired t-test and a Wilcoxon signed rank test between the Dice scores of the three registration methods. For such a high sample size, statistical significance ($p < 0.05$) is almost guaranteed for any difference, and we observed p-values lower than 10^{-4} for all method pairs. To report statistical significance, we take inspiration from Klein et al. (2009) and perform permutation tests (Menke and Martinez, 2004) to determine if the means of a small set of independent overlap values obtained by each of the registration methods are the same. The subset of brain pairs was selected so that each brain was used only once, and we fixed the number of permutations to 1024, and calculate 10,000 p-values for each method pair. We report the fraction of p-values less than 0.05 (represented as μ) for each method pair as a proxy for statistical significance, as suggested by Klein et al. (2009), with higher values indicating greater statistical significance. To measure practical impact, we measure Cohen's d (that represents effect sizes) for practical significance ($d > 0.2$) for each pair of registration methods.

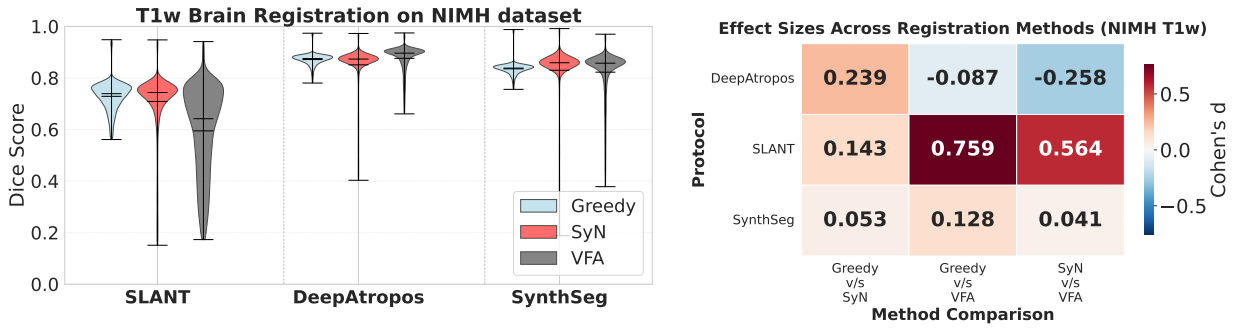


Figure 2.4: Comparison of the three registration methods on the NIMH T1w dataset. Left shows violin plots of the Dice scores of the top iterative and deep learning registration methods on the NIMH T1w dataset. **Right** shows Cohen's d scores for all method pairs, quantifying the practical significance of the differences in Dice scores between the three registration methods.

Method	SLANT			DeepAtropos			SynthSeg		
	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
Greedy	0.7289	0.7393	0.0547	0.8717	0.8755	0.0219	0.8356	0.8384	0.0232
SyN	0.7090	0.7437	0.1178	0.8511	0.8735	0.0844	0.8300	0.8593	0.1183
VFA	0.5950	0.6421	0.1700	0.8764	0.8964	0.0507	0.8227	0.8575	0.0933

Table 2.2: Registration method performance across different labelling protocols on the NIMH T1w dataset. Table shows the mean, median, and standard deviation of the Dice scores of the top three registration methods on the NIMH T1w dataset.

Discussion. Interestingly, VFA performs significantly worse on the SLANT labelmaps (Table 2.2), in direct contrast to the results in the LUMIR challenge. Since the conditions for evaluation of the original challenge are unspecified, we can only speculate that the differences are due to preprocessing conditions and label interpolation schemes. On the SynthSeg and DeepAtropos labelmaps, the performance of VFA is comparable to (but still lower than) Greedy and SyN, with minor differences in the Cohen's d scores (Figure 2.4). Permutation tests in Table 2.5 show that VFA significantly underperforms Greedy and SyN on the SLANT labelmaps indicated by high μ values, while lower μ values for DeepAtropos and SynthSeg labels suggest that the differences in Dice scores for these labelmaps are not statistically significant. This is an indicator that modern deep methods like VFA are able to register coarse anatomical structures well, but may struggle with highly parcellated structures. However, deep methods are comparable to iterative methods on inter-subject registration of in-distribution contrast, showing maturity of deep learning methods in terms of task understanding for image registration, compared to the previous generation of methods that performed well on the training data but failed to generalize to modest and practical amounts of domain shift (Jena et al., 2024; Jian et al., 2024; Jena et al., 2025; Jian et al., 2025; Liu et al., 2025a).

Results on PRIME-DE dataset.

A natural extension of zero-shot evaluation from the T1w human MRI is to evaluate registration performance on a human-adjacent mammalian species like the Macaque. To that end, the PRIME-DE dataset provides a collection of T1w MRI images of the Macaque brain, with original resolutions varying from 0.3 to 0.8mm. This is in contrast to Chen et al. (2025) which incorrectly claims that the brain images are originally acquired

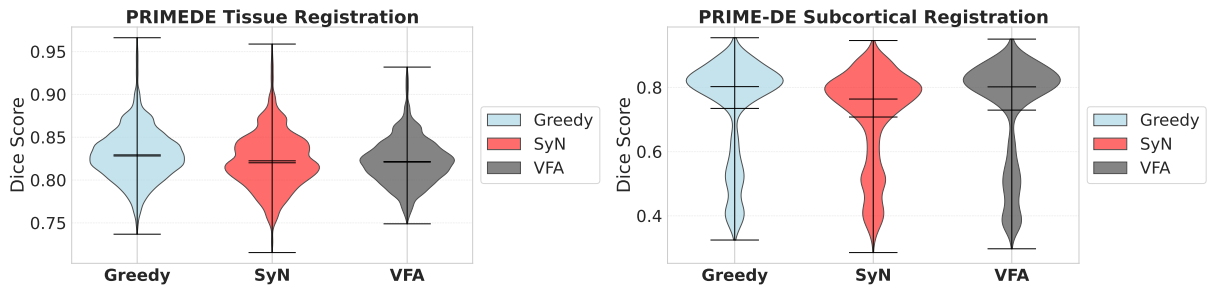


Figure 2.5: Comparison of the three registration methods on the PRIME-DE dataset. Left shows violin plots of the Dice scores of tissue overlap (GM, WM, CSF), **Right** shows violin plots of the Dice scores of subcortical overlap between the registered and reference labelmaps.

Method	Cortical	Subcortical
Greedy v.s. SyN	1.0	1.0
Greedy v.s. VFA	1.0	0.6873
SyN v.s. VFA	0.0512	0.9990

Table 2.3: Statistical significance on the PRIME-DE dataset represented as fraction of p-values less than 0.05 for each method pair using permutation tests. Higher values represent greater statistical significance.

at 1mm isotropic resolution, indicating a potential lack of quality control in the original evaluation. We download data from the five different sites mentioned in the original challenge, followed by brain extraction and segmentation using the nBEST (Zhong et al., 2024) tool. All subjects are affinely registered using FireANTs to a manually chosen subject with 0.3mm resolution, to maximize the field of view and resolution for subjects with lower resolution. We obtain 116 brain images, resulting in 13,340 (= 116 $\times$ 115) image pairs for evaluation. We include preprocessing and affine alignment scripts in our provided code. The nBEST tool provides two segmentations: (1) segmentation of three cortical labels (GM, WM, CSF) and (2) segmentation of six subcortical labels, including thalamus, caudate, putamen, pallidum, hippocampus, and amygdala.

Discussion. We include violin plots, summary statistics, and Cohen's d scores of Dice score overlap between inter-subject registered labelmaps for both the cortical and subcortical segmentations in Figure 2.5 and Figure 2.6. We note a small gap between the performance of Greedy and VFA for both cortical and subcortical segmentations, showing that modern deep learning methods demonstrate improved task understanding on a familiar modality but unseen anatomy (i.e. T1w MRI of the macaque cerebrum). The Cohen's d scores in Figure 2.6 show that although small, the performance difference between Greedy and VFA is of practical significance for both cortical and subcortical segmentations, with d scores of 0.308 and 1.016, significantly outside the accepted standard for ''small effects'' ($d < 0.2$). Permutation tests in Table 2.3 also indicate that the difference in Dice scores between Greedy and VFA *are* statistically significant for independent subsets of image pairs for cortical segmentations, and slightly less but still significant for subcortical segmentations. However, SyN underperforms VFA *significantly* for the subcortical segmentations, indicating that Greedy is a better overall choice for iterative registration. This modest performance gap is in contrast to the results in Chen et al. (2025) where VFA underperformed FireANTsGreedy substantially, which could be attributed to poorly designed preprocessing conditions in the original evaluation. This underscores the importance of careful design of preprocessing conditions for zero-shot evaluation, and the need for a standardized evaluation protocol for inter-subject registration.

Method	Tissue	Subcortical
Greedy	0.829 ± 0.030	0.735 ± 0.158
SyN	0.823 ± 0.032	0.708 ± 0.151
VFA	0.822 ± 0.026	0.729 ± 0.165

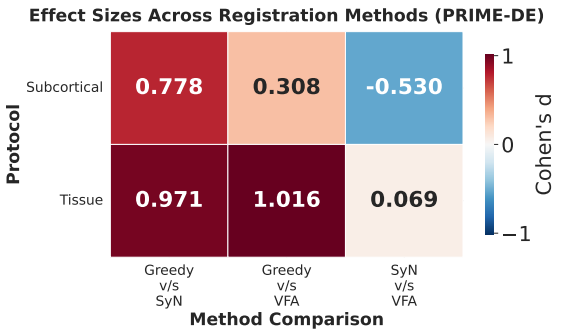


Figure 2.6: Quantitative comparison of the three registration methods on the PRIME-DE dataset. Left shows the mean, median, and standard deviation of the Dice scores of the top three registration methods on the PRIME-DE dataset. **Right** shows Cohen's d scores for all method pairs.

2.5.2. Unsupervised DLIR does not generalize to novel contrasts

While T1-weighted imaging provides excellent anatomical detail with superior gray-white matter contrast ideal for morphometric analysis and structural segmentation, it exhibits limited sensitivity to many pathological processes. T2-weighted and FLAIR sequences offer complementary contrast mechanisms that are essential for detecting white matter lesions, edema, inflammation, and demyelination—pathologies that often have subtler appearance on T1w scans. For example, T2* imaging is sensitive to magnetic susceptibility effects, making it useful for detecting hemorrhages, iron buildup, and other magnetic substances that are not visible in T1w scans. T2 weighted images are particularly useful for visualizing fluid-filled structures, such as cerebrospinal fluid (CSF) and white matter lesions that appear isointense with the background on T1w scans. FLAIR images on the other hand suppresses signal from CSF, highlighting abnormalities like lesions, tumors, and stroke against a more homogeneous background. In these modalities, the GM-WM boundary is often less distinct than in a T1w scan. In clinical practice, multimodal protocols combining these contrasts are standard precisely because no single sequence provides comprehensive tissue characterization: T1w reveals anatomy while T2/FLAIR/T2* reveal pathophysiology.

The LUMIR challenge uses the NIMH dataset with T1w, T2w, T2*, and FLAIR sequences for zero-shot evaluation of out-of-distribution contrasts, where they use the SLANT segmentation from the co-registered T1w images to the T2w, T2*, and FLAIR images to obtain labelmaps. However, since the majority of labels in the SLANT segmentation are cortical parcellations and the T2w, T2*, and FLAIR images do not provide sufficient GM-WM contrast compared to T1w images, we argue that this parcellation is not representative of the registration task. Moreover, our experiments in subsection 2.5.1 show that the performance of deep learning methods (VFA) on the SLANT labelmaps is significantly worse than iterative methods (Greedy and SyN). Therefore, we consider SynthSeg for labelmap generation on the T2w, T2*, and FLAIR images. SynthSeg is a general purpose segmentation model that is trained on a wide range of contrasts, and is suitable for accurate segmetnation for all three contrasts. Moreover, VFA performs comparably to iterative methods on the SynthSeg labelmap on T1w images, setting a benchmark for performance comparison between in-distribution and out-of-distribution contrasts. Initially, we segmented 438 images from each contrast, leading to a total of 191,406 (= 438×437) image pairs for evaluation of each contrast. We evaluate the average Dice Score overlap between the registered and reference labelmaps on a randomly chosen (and fixed) subset of 5,000 pairs for each contrast to reduce computational cost.

Discussion. Violin plots and pairwise Cohen's d scores are reported in Figure 2.7 and statistical summaries are shown in Table 2.4. Compared to T1w images, the performance of the top deep learning method VFA drops significantly for unseen contrasts. The largest difference in performance is observed in T2w images, followed by T2* and FLAIR images. The Cohen's d scores in Figure 2.7 underscore that the practical impact of the performance difference is significant for all three contrasts, contrary to the results in the T1w dataset where the differences are minor and do not have significant practical impact. For example, on the T2 and T2* images, the Cohen's d scores are in the range of 0.70-1.51, significantly outside the accepted standard for ''small effects'' ($d < 0.2$). Permutation tests (Table 2.5) further confirm that these performance gaps are statistically significant. The consistently large mean test statistics (μ) for T2 and T2* reinforce that the observed differences reflect systematic performance degradation rather than sampling variability. In contrast, both Cohen's d and permutation test results indicate smaller effect sizes and weaker statistical significance for FLAIR. This aligns with the long-tailed Dice distribution observed in the violin plots (Figure 2.7), suggesting higher variability but less consistent separation between methods. We hypothesize that this behavior stems from the contrast properties of FLAIR imaging: FLAIR emphasizes T2-hyperintense pathology (e.g., white matter lesions, edema, periventricular abnormalities) while providing comparatively weak contrast at morphometric tissue boundaries such as the GM-WM interface and deep subcortical structures. The resulting boundary ambiguity likely increases registration variability of morphometric boundaries without producing a consistent directional performance gap between methods.

These results are in stark contrast to the results in the LUMIR challenge, where VFA performed *better* than iterative methods on out-of-distribution contrasts, with seemingly no empirical consideration or theoretical justification for the observed difference. Domain shift is an established problem in deep learning, and is a well-studied phenomena that is pervasive in a broad range of application areas, and has garnered significant resources to systematically study and mitigate it (Beery et al., 2020; Zech et al., 2018; AlBadawy et al., 2018; Jadon et al., 2025). The LUMIR challenge asserts that a variety of deep learning architectures are robust to domain shift just by training on a large set of T1w images, a direct contradiction to the extensive literature on domain shift and domain adaptation in deep learning. A crucial question therefore emerges: can this seemingly absurd conclusion from the original challenge be extended to other tasks like lung CT or abdomen registration? Moreover, the LUMIR challenge did not explicitly discuss or account for the interaction between registration performance and image contrast. In particular, T1w and FLAIR images emphasize qualitatively different structures—T1w highlights morphometric boundaries, whereas FLAIR emphasizes T2-hyperintense pathology. Because these contrasts are complementary rather than equivalent, registration difficulty and evaluation metrics may reflect modality-specific contrast properties rather than purely methodological differences. Our results are consistent with the expectation that deep methods learn a distribution of *T1w (task) specific* features that is then used in conjunction with registration-aware modules (Jian et al., 2025, 2024) to achieve generalization to T1w images. Moreover, VFA exhibits significantly higher variance (a proxy for predictive variance) than Greedy and SyN for all three contrasts, which is consistent with the literature on predictive uncertainty and entropy estimation of deep learning methods on out-of-distribution data (Lakshminarayanan et al., 2017; Maddox et al., 2019; Malinin and Gales, 2018).

2.5.3. Unsupervised DLIR does not scale with increasing resolution

The Ultracortex dataset (Mahler et al., 2024) includes a collection of 9.4T ultra-high field MRI images of the human brain, with resolutions varying from 0.6 to 0.8mm. The images are acquired using a 9.4T MRI scanner, and the data consists of both MP-RAGE and MP2RAGE sequences. The dataset includes high-quality manual

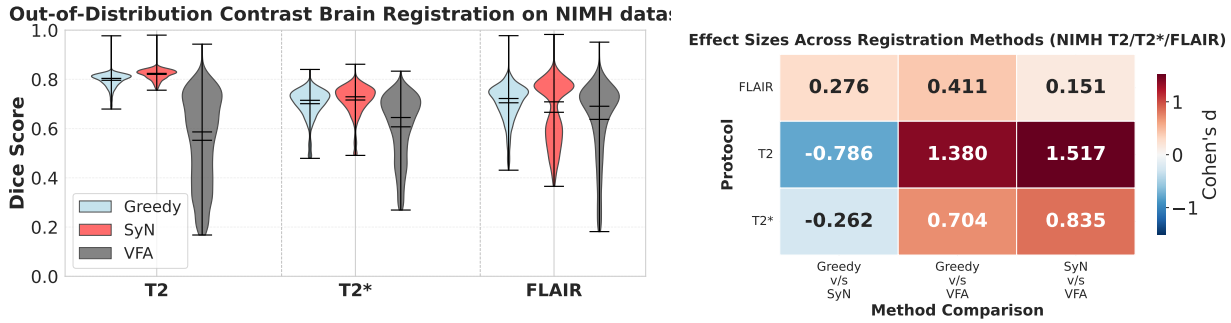


Figure 2.7: Comparison of the three registration methods on out-of-distribution contrasts on the NIMH dataset with labels generated by SynthSeg.

Method	T2			T2*			FLAIR		
	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
Greedy	0.7961	0.8038	0.0292	0.7012	0.7148	0.0608	0.7049	0.7222	0.0732
SyN	0.8203	0.8241	0.0213	0.7161	0.7292	0.0606	0.6662	0.7087	0.1245
VFA	0.5524	0.5865	0.1792	0.6072	0.6450	0.1287	0.6373	0.6907	0.1509

Table 2.4: Registration method performance across different out-of-distribution contrasts on the NIMH dataset with labels generated by SynthSeg.

segmentations for 12 subjects - which include both gray and white matter segmentations for each hemisphere - leading to 4 labels.

The LUMIR challenge uses SLANT to obtain labelmaps for the Ultracortex dataset, and downsamples the images to 1mm isotropic. However, this preprocessing has two undesirable effects. *First*, submillimeter resolution images can provide additional cytoarchitectural detail and act as a bridge between low resolution *in-vivo* scans and high resolution histology images. Lowering the resolution can lead to loss of this information and defeats the purpose of using high-resolution scans in the first place. Moreover, this is not representative of clinical and research workflows where high-resolution blockface scans are used as an intermediate modality between in-vivo scans and histology slides (Puonti et al., 2025; Alegro et al., 2016). *Second*, the MP2RAGE sequences in the dataset are both qualitatively and quantitatively different compared to the MP-RAGE sequences seen in the OASIS or LUMIR datasets. This constitutes a significant source of domain shift that leads to poor performance of SLANT on the Ultracortex dataset, making it unsuitable for robust evaluation. This aspect is not discussed and possibly unaccounted for in the original evaluation. We examine the volumes

Method	SLANT	DeepAtropos	SynthSeg			
	T1	T1	T1	T2	T2*	FLAIR
Greedy v.s. SyN	0.1870	0.6383	0.0031	0.5930	0.4292	0.0443
Greedy v.s. VFA	1.0000	0.0222	0.1258	0.4244	0.2407	0.0280
SyN v.s. VFA	0.9781	0.2143	0.0025	0.6035	0.4629	0.0292

Table 2.5: Statistical significance on the NIMH dataset represented as fraction of p-values less than 0.05 for each method pair using permutation tests. Higher values represent greater statistical significance.

and histograms of the subjects and show that the MP-RAGE sequences (corresponding to subjects `sub-37`, `sub-45`, `sub-57`) indeed look qualitatively different than the MP2RAGE sequences in Figure 2.8. Specifically, histograms of the MP2RAGE sequences are characterized by two or three peaks, close to the extreme values of the intensity range, while the MP-RAGE sequences have a more unimodal distribution with a single dominant peak.

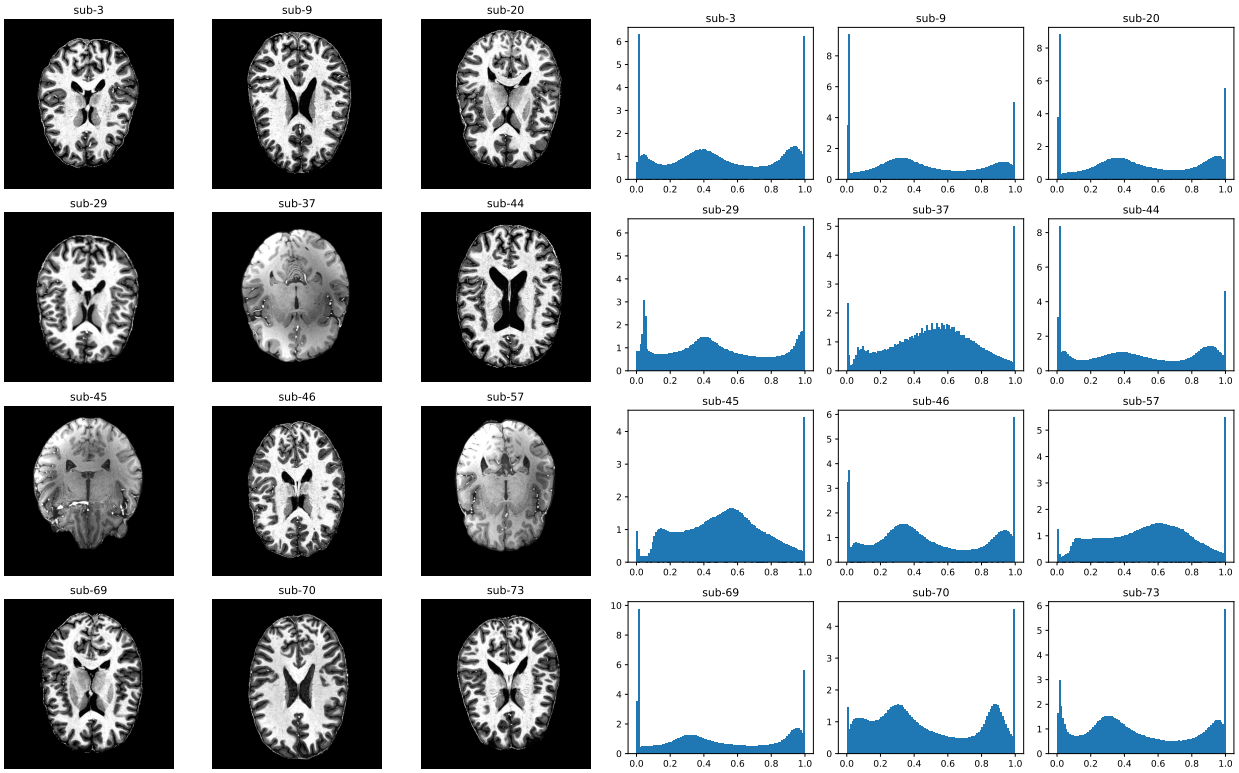


Figure 2.8: Multimodal characterization of the Ultracortex dataset. Left shows axial slices of subjects from the Ultracortex dataset. Out of 12 subjects with labeled segmentations, 3 subjects have MP-RAGE sequence data, and 9 subjects have MP2RAGE sequence data. **Right** shows histograms of the intensity values of the subjects. The MP2RAGE sequences are characterized by two or three peaks close to the extreme values of the intensity range, while the MP-RAGE sequences have a more unimodal distribution with a single dominant peak. The qualitative differences in both the intensity values and histograms are indicative of the multimodal nature of the dataset.

Evaluation. To account for the possible effect of domain shift on the performance of SLANT, we perform an alternative evaluation that leverages the high-quality manual segmentations already provided as part of the dataset. MP2RAGE sequences in the dataset provide excellent gray/white matter contrast, making it a practical testbed for evaluating performance of registration algorithms. **Resolution:** First, we affinely register all images to the `sub-3` subject's MP2RAGE image. This brings all images to a 0.6mm isotropic resolution. Initially, we proposed evaluation of both the Greedy and SyN modes in FireANTs, and VFA on the 0.6mm isotropic resolution images. Unfortunately, VFA runs out of memory for 0.6mm isotropic registration on a GPU with 48GB of memory, highlighting the limitations of deep learning methods for high-resolution image registration. Therefore, we further resample the dataset and labels to 1mm isotropic resolution, and evaluate the performance of the same methods on the 1mm isotropic resolution images. **Multimodality:** Moreover, in the dataset, 3 out of the 12 subjects have MP-RAGE sequences, while the other 9 subjects have MP2RAGE

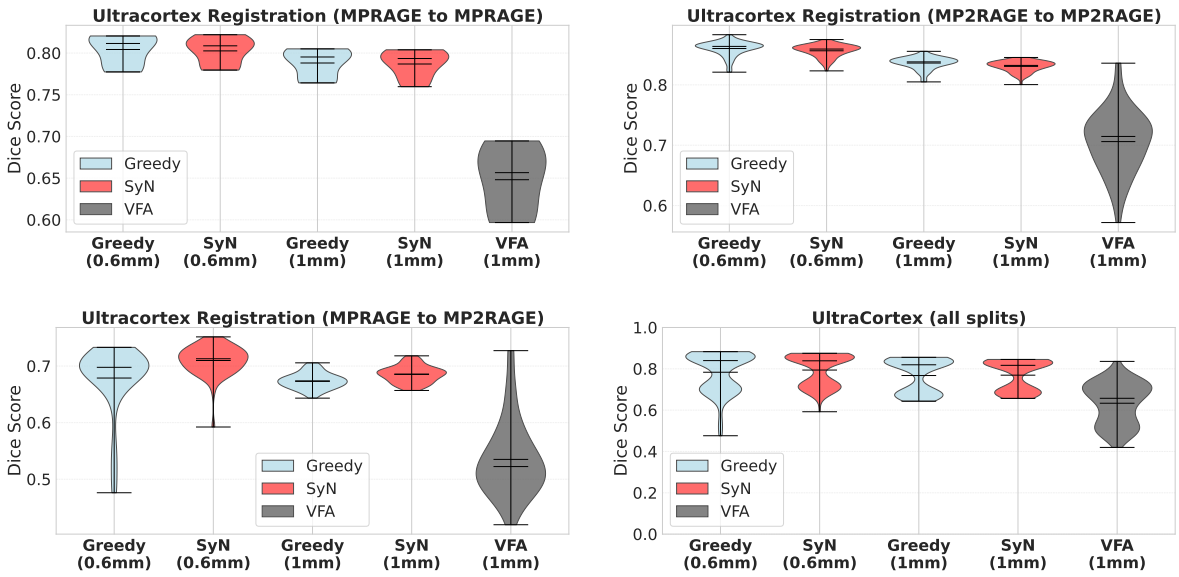


Figure 2.9: Comparison of the three registration methods on the Ultracortex dataset.

sequences. Registration of and MP-RAGE to an MP2RAGE sequence constitutes a multimodal task, and we use MIND features for FireANTs for every pair of images that have different modalities. VFA does not support any other feature images other than intensities as input, and the evaluation for VFA remains unchanged. We report the Dice scores on four splits of the dataset - (a) MP-RAGE to MP-RAGE ($n = 6$), (b) MP2RAGE to MP2RAGE ($n = 72$), and (c) MP-RAGE $\longleftrightarrow$ MP2RAGE ($n = 54$), and (d) all subjects ($n = 132$).

Table 2.6: Registration performance on Ultracortex dataset across different split types and methods

Split	Greedy (0.6mm)	SyN (0.6mm)	Greedy (1mm)	SyN (1mm)	VFA (1mm)
All splits	0.784 ± 0.096	0.794 ± 0.073	0.768 ± 0.080	0.769 ± 0.071	0.633 ± 0.100
MPRAGE to MPRAGE	0.804 ± 0.017	0.803 ± 0.015	0.788 ± 0.016	0.787 ± 0.015	0.648 ± 0.037
MPRAGE to MP2RAGE	0.679 ± 0.060	0.710 ± 0.026	0.674 ± 0.014	0.685 ± 0.015	0.535 ± 0.065
MP2RAGE to MP2RAGE	0.860 ± 0.012	0.857 ± 0.010	0.836 ± 0.010	0.831 ± 0.009	0.706 ± 0.051

Discussion. The results in Table 2.6 and Figure 2.9 highlight three key insights. First, MPRAGE to MP2RAGE registration is a significantly harder task than either MPRAGE to MPRAGE or MP2RAGE to MP2RAGE registration, illustrated by about an 18 point drop in Dice score compared to the MP2RAGE-MP2RAGE split. The MP2RAGE images are well poised to register gray and white matter boundaries due to the excellent contrast, reaching an average Dice score of upto 0.86 for Greedy version of FireANTs. Second, high-resolution registration leads to around a 2 point increase in Dice score for both Greedy and SyN versions of FireANTs essentially obtained for 'free' without any additional domain-specific considerations. Third, the results show that beyond the poor generalization of a representative top performing deep learning method on out-of-distribution contrasts, the methods cannot accomodate multimodal images out-of-the-box. Furthermore, these methods do not scale beyond 1mm isotropic resolution, limiting their applicability to the broad range of high-resolution images and the insights provided by advanced high resolution scanners, ex-vivo imaging studies, and multimodal integration. With improved efficiency of iterative optimization methods, able to register 0.6mm isotropic images in seconds, they are well positioned to tackle the scale of high resolution

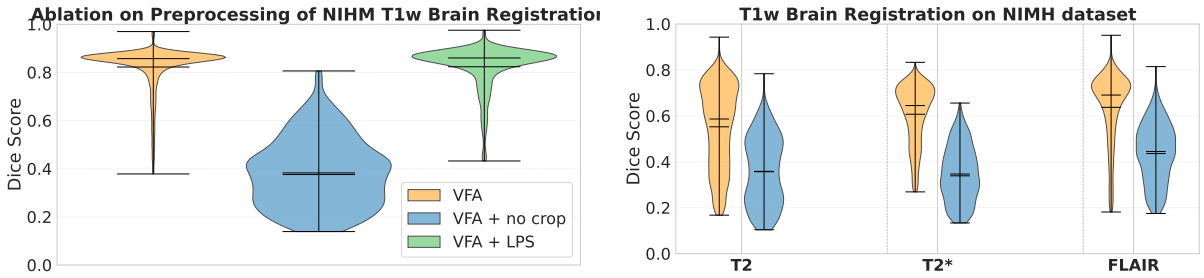


Figure 2.10: Ablation study on the NIMH dataset showing the effect of preprocessing choices on the performance of the model. **Left** shows the performance of VFA on the cropped images, on the original images (denoted as *no crop*), and on images in the LPS orientation (denoted as *LPS*) on the T1w modality. **Right** shows the performance of VFA on the cropped and original images on the T2w, T2*, and FLAIR modalities.

imaging workflows pertinent in MRI to histology workflows.

2.5.4. Unsupervised DLIR methods are sensitive to preprocessing choices

An often overlooked limitation of deep learning methods is their sensitivity to preprocessing choices. Most deep learning methods are trained on a fixed set of preprocessing steps, and may perform poorly if the preprocessing steps are not the same as the ones used during training. In contrast to highly controlled evaluation environments like registration challenges, real-world data such as *ex-vivo* hemispheres, histology, and blockface images are rarely standardized to the same preprocessing steps or stereotaxic coordinates. Other domains like MRA imaging can have limited field of view and are highly anisotropic, making it difficult to standardize the preprocessing steps across modalities. Sensitivity to preprocessing choices shifts the burden of preprocessing from the model to the practitioner, who may not be familiar with the preprocessing protocol used during training and might produce suboptimal results.

Evaluation. To demonstrate the sensitivity of state-of-the-art deep learning method VFA to preprocessing choices, we perform an ablation study on the NIMH dataset using the SynthSeg segmentation protocol. The NIMH dataset originally contains $208 \times 256 \times 256$ voxels when resampled to 1mm isotropic resolution. Our preliminary experiments with VFA on the original 1mm isotropic T1w images resulted in significantly worse performance than expected. Upon further investigation, we found that VFA performs registration well only if the images are cropped to $192 \times 160 \times 224$ voxels. Therefore, we cropped the images to a smaller region of interest (ROI) of $192 \times 160 \times 224$ voxels and evaluated the performance of VFA on these cropped images, upon which we obtained significantly better performance. We also note that VFA was trained on images oriented in RAS frame, and therefore evaluated its performance on the DICOM-standard LPS orientation.

Discussion. Our results in Figure 2.10 show that VFA performs significantly better on the cropped images across all modalities, and that the performance is significantly worse on the original images, implying that the model is 'locked in' to a particular voxel size. This poses a practical limitation wherein the practitioner may not be able to use the model if the anatomy of interest does not fit inside the field of view of the cropped image. In contrast, iterative methods suffer from no such limitation, and can be readily used by the practitioner without worrying about esoteric preprocessing choices. Fortunately, there is no significant difference in performance by changing the orientation of the images, demonstrating some task understanding and generalization by the model.

2.6. Summary and Conclusions

OASIS and Klein et al. (2009) data. Preceding experiments show that classical methods provide an unprecedented level of robustness and generalizability across datasets, but are limited by the fidelity of the image matching objective. DLIR methods provide a promising step towards improving registration performance of anatomical regions by implicitly discovering these structures and predicting appropriate warp fields. However, this anatomical-awareness on the training dataset does not help in generalizing to other datasets, limiting the practical utility of these methods. The usability of anatomical landmarks and labelmaps to obtain domain-invariant registration performance still remains an open research problem. At the current state, a practitioner should choose DLIR methods only if they have access to a large labeled dataset, and their application is limited to the same dataset distribution. In all other cases, classical optimization-based methods are the more accurate and reliable choice.

LUMIR challenge. Our conclusions are rather unsurprising, but strikingly different than in Chen et al. (2025). *First*, we observe that the performance of SOTA deep learning methods on T1 weighted MRI imaging is indeed comparable with iterative optimization methods, even on a human-adjacent species like Macaque - showing that the next generation of deep learning algorithms for registration demonstrate substantially better task understanding. However, deep learning methods can show inferior performance on highly parcellated regions like the SLANT segmentation, compared to other segmentation labels like DeepAtropos or SynthSeg. *Second*, the task understanding does not translate to better performance on out-of-distribution contrasts, contrary to the results shown in Chen et al. (2025). *Third*, scalability remains an issue with deep learning methods as demonstrated on the high-resolution Ultracortex dataset, while iterative optimization methods enjoy improved performance by registering high resolution brains due to their low memory footprint. The scalability makes iterative optimization method a more practical choice for high-resolution image registration pertinent in histopathological workflows and life sciences research. *Fourth*, we show that deep learning methods are highly sensitive to even trivial changes in image preprocessing, including retaining padding from the original dataset. These modes of sensitivity puts a burden on the practitioner to ensure the data conforms to the emulated preprocessing standards during training, potentially limiting its applications to high variability pertinent in real clinical scenarios.

CHAPTER 3

FireANTs: Adaptive Riemannian Optimization for Multi-Scale Diffeomorphic Matching

In the previous chapter, we performed a holistic empirical evaluation of the strengths and weaknesses of deep learning-based image registration methods in comparison to optimization-based methods. The supervised learning methods show high propensity to overfit to the training data, even in relation to its in-distribution validation set. Although this has steered researchers to improve unsupervised deep learning registration, such methods typically show no real benefit over optimization-based methods - their performance on in-distribution data does not significantly surpass optimization-based methods, they do not reliably generalize to novel contrasts, they do not scale well with resolution, and are highly sensitive to preprocessing choices, which are less than ideal factors for real-world clinical deployment. Although iterative solvers seem tempting, they are slow and utilize first-order gradient descent approaches, and do not leverage differentiablity to enable end-to-end learning of multi-scale features.

In this chapter, we introduce FireANTs, a multi-scale Adaptive Riemannian Optimization framework for diffeomorphic registration. Our method captures the full expressivity of time-dependent velocity fields for modelling diffeomorphisms, and enables adaptive optimization on this space directly, while maintaining high computational efficiency and low memory overhead for inference. We formally introduce the registration problem and its mathematical challenges including high-dimensionality and ill-conditioning, discuss properties of dense diffeomorphisms, and quantify the extent of ill-conditioning of image registration. This sets the stage for adaptive optimization on the space of diffeomorphisms, where we first discuss a few alternate designs for adaptive optimization on diffeomorphisms and identify their limitations. This is followed by a novel Eulerian descent formulation to derive adaptive optimization on diffeomorphisms directly. The chapter concludes with a detailed empirical evaluation of FireANTs on multiple benchmarks and datasets, showing its state-of-the-art performance, scalability, and robustness to hyperparameters.

3.1. Preliminaries of *Diffeomorphic* Image Registration

Recall from the previous chapter that given d-dimensional images $I_f : \Omega \to \mathbb{R}^K$ and $I_m : \Omega \to \mathbb{R}^K$ where the domain Ω is a compact subset of $\mathbb{R}^2$ or $\mathbb{R}^3$, image registration is formulated as an optimization problem to find a transformation φ that warps I_m to I_f. The transformation can belong to an algebraic group, say G, whose elements $g \in G$ act on the image by transforming the domain as $(I_m \circ g)(x) = I_m(g(x))$ for all $x \in \Omega$. The registration problem solves for

$$\varphi^* = \arg\min_{\varphi \in G} L(\varphi) \doteq C(I_f, I_m \circ \varphi) + R(\varphi) \tag{3.1}$$

where C is a cost function, e.g., that matches the pixel intensities of the warped image with those of the fixed image, or local normalized cross-correlation or mutual information of image patches. There are many types of regularizers R used in practice, e.g., total variation, elastic regularization (Gee et al., 1993), enforcing the transformation to be invertible (Christensen and Johnson, 2001), or volume-preserving (Haber and Modersitzki, 2004) using constraints on the determinant of the Jacobian of φ, etc. If, in addition to the pixel intensities, one also has access to label maps or different anatomical regions marked with correspondences across the two images, the cost C can be modified to ensure that φ transforms these label maps or landmarks appropriately.

Properties of the considered Transformation Group A diffeomorphism is a smooth and invertible map with a corresponding differentiable inverse map (Banyaga, 2013; Leslie, 1967; Younes, 2010). We denote the set of all diffeomorphisms on Ω as $\text{Diff}(\Omega; \mathbb{R}^d)$. It is useful to note that unlike rigid or affine transforms that have a fixed number of parameters, diffeomorphisms require dense and variable parameterization, typically proportional to the size of the image. When groups of transformations on continuous domains are endowed with a differentiable structure, they are called Lie groups. Diffeomorphisms are also examples of Riemannian manifolds, and are amenable to Riemannian optimization. This property is explored in Section 3.3.3 in the context of Riemannian gradient descent.

In this chapter, we only consider a subgroup of diffeomorphisms. Consider the set of continuously differentiable functions $u \in C_0^1(\Omega, \mathbb{R}^d)$ such that $u, J(u) = 0$ on $\partial\Omega$, where $J(u)$ is the Jacobian of u, such that $[J(u)(x)]_{ij} = \frac{\partial u(x)_i}{\partial x_j}$. These functions can be extended to have $u \equiv 0$ outside Ω. Then, for a small enough $\epsilon > 0$, $x + \epsilon u(x)$ is a diffeomorphism. We refer the reader to Proposition 8.6 in (Younes, 2010) for a succinct proof. Although these diffeomorphisms are close to identity, diffeomorphisms with larger deviations from the identity can be constructed by composing these 'small diffeomorphisms'. Therefore, we study the subgroup of diffeomorphisms of the form

$$\phi_n = (id + \epsilon_1 u_1) \circ \ldots \circ (id + \epsilon_n u_n) \tag{3.2}$$

where u_is are defined as before. We denote this subgroup as $G(\Omega, \mathbb{R}^d)$. This subgroup retains the group structure with identity element id, the composition operation $\circ$ induced from $\text{Diff}(\Omega, \mathbb{R}^d)$, and the inverse group element: $\phi_n^{(-1)} = (id + \epsilon_n u_n)^{(-1)} \circ \ldots \circ (id + \epsilon_1 u_1)^{(-1)}$ (as each individual $id + \epsilon_n u_n$ is shown to have an inverse (Younes, 2010)). The elements of this subgroup can be thought of as diffeomorphisms arising from time-varying continuously differentiable flows.

However, the rate of convergence of these algorithms are contingent on the severity of ill-conditioning of Eq. (3.1). In the following text, we first show the extent of ill-conditioning for diffeomorphic registration which subsequently warrants adaptive optimization over this subgroup of diffeomorphisms.

3.2. The Ill-Conditioned Nature of Image Registration objectives

The ill-conditioned nature of image registration represents a comparatively neglected domain of inquiry within the extant literature. As of writing this chapter, recent works in the literature (Mang and Ruthotto, 2017b; Mang and Biros, 2017) speculate the ill-conditioned nature of registration, especially for high-resolution heterogeneous datasets but do not quantify it. Computing the ill-conditioning requires us to analyze the Hessian of the registration cost function. This is infeasible in general due to the high dimensionality of the problem; the full Hessian of a MRI brain registration problem requires more than **15 petabytes** of memory to store. However, we consider a typical scenario of T1-weighted 3D MRI image registration with the L2 loss (Balakrishnan et al., 2019; Avants and Gee, 2004; Beg et al., 2005): i.e. $C(I_f, I_m, \varphi) = \sum_i (I_f(x_i) - I_m(\varphi(x_i)))^2$. In this case, the gradient of C w.r.t. $\varphi(x_i)$ is $(I_m(\varphi(x_i)) - I_f(x_i))\nabla I_m(\varphi(x_i))$, which does not depend on $\varphi(x_j), j \neq i$. Therefore, the full Hessian is simply a block-diagonal matrix containing pixelwise Hessians $H_i = \nabla^2_{\varphi(x_i)} C$ with eigenvalues $\{\lambda_i; i = \{1, 2, 3\}\}$. This makes the conditioning analysis tractable. We calculate the per-pixel condition number, defined as $\kappa_i = |\lambda_i^{\max}|/|\lambda_i^{\min}|$; and investigate the relationship between the fraction of foreground pixels and κ_i across multiple spatial resolutions of the images. The study considers three downsampling factors: 1x (original resolution), 2x, and 4x, in accordance with existing multi-scale optimization techniques. Fig. B.1 shows that across all resolutions, more than 60% of foreground pixels have a condition number greater than 10. To elucidate the impact of poor conditioning on optimization, we construct a simplified example of an ill-conditioned two-dimensional convex optimization problem, detailed in the Appendix Section B.1. The toy example shows the difficulty of optimization of ill-conditioned problems using first-order methods like SGD. A typical MRI registration task with L2 loss has a $\kappa > 10^8$, making it extremely ill-conditioned, and strongly motivating the need for first-order adaptive optimization.

3.3. Adaptive Optimization for Diffeomorphisms

We provide a brief overview of the mathematical frameworks employed to optimize parameters that reside on Riemannian manifolds like diffeomorphisms, followed by a novel algorithm that exploits the group action to define a gradient descent algorithm that eliminates computationally expensive steps. This novel formulation of the 'gradient descent' algorithm can then be formulated to incorporate adaptive algorithms such as Adam (Kingma and Ba, 2014) to optimize diffeomorphisms.

3.3.1. Euclidean gradient descent using the Lie algebra in shooting methods

Each Lie group has a corresponding Lie algebra $\mathfrak{g}$ which is the tangent space at identity. This creates a locally one-to-one correspondence between elements of the group $g \in G$ and elements of its Lie algebra $v \in \mathfrak{g}$ given by the exponential map $\exp : \mathfrak{g} \to G$; effectively to reach $g = \exp(v)id$ from identity $id \in G$, the exponential map dictates that the group element has to move along v for unit time along the manifold. Exponential maps for many groups can be computed analytically, e.g., Rodrigues transformation for rotations, Jordan-Chevalley decomposition (Chevalley, 1951), or the Cayley Hamilton theorem (Mertzios and Christodoulou, 1986) for matrices. For diffeomorphisms, the Lie algebra is the space of all smooth velocity fields $v : \Omega \to \mathbb{R}^d$. There exist iterative methods to approximate the exponential map called the scaling-and-squaring approach (Ashburner, 2007; Balakrishnan et al., 2019) which uses the identity

$$\varphi = \exp(v) = \lim_{N \to \infty} \left(\mathrm{id} + \frac{v}{N}\right)^N \tag{3.3}$$

to define a recursion by choosing N to be a large power of 2, i.e. $N = 2^M$ as

$$\varphi^{(1/2^M)} = id + v/2^M \tag{3.4}$$

$$\varphi^{(1/2^k)} = \varphi^{(1/2^{(k+1)})} \circ \varphi^{(1/2^{(k+1)})} \qquad \forall k \in \{0, 1 \ldots, M-1\}; \tag{3.5}$$

This can be thought of as a special case of Eq. (3.2) with $n = 2^M, \epsilon = \frac{1}{n}$ and $u_1 = \ldots = u_n = v$.

By virtue of the exponential map, we can solve the registration problem of finding $\varphi \in G$ by directly optimizing over the Lie algebra v. This is because the Lie algebra is a vector space and we can perform, for example, standard Euclidean gradient descent for registration (Moler and Van Loan, 2003; Hall and Hall, 2013; Hall, 2000). Such methods are called stationary velocity field or shooting methods. At each iteration, one uses the exponential map to get the transformation φ from the velocity field v, computes the gradient of the registration objective with respect to φ, pulls back this gradient into the tangent space where v lies

$$\nabla_v L = \frac{\partial \varphi}{\partial v} \nabla_\varphi L \tag{3.6}$$

and finally makes an update to v. Traditional methods like DARTEL (Ashburner, 2007) implement this approach. This is also very commonly used by deep learning methods for registration (Balakrishnan et al., 2019; Krebs et al., 2019; Niethammer et al., 2019) due to its simplicity. Deep learning methods typically produce outputs that lie on some Euclidean vector space, and intrepreting the output as a Lie algebra element and using the exponential map to obtain the diffeomorphism is therefore common practice. Geodesic shooting methods are more sophisticated implementations of this approach where φ is the solution of a time-dependent velocity that follows the geodesic equation; the geodesic is completely determined by the initial velocity $v_0 \in \mathfrak{g}$. Therefore, Eq. (3.6) is a valid gradient computation (with a different form of $\frac{\partial \varphi}{\partial v}$) for optimizing geodesic or momentum based representations of diffeomorphisms as well.

Adaptive optimization algorithms can be applied to the Lie algebra since it a Euclidean vector space $\mathfrak{g}$. However, there are a number of challenges with this method. First, this method requires computing the exponential map and its derivative, both of which need to be iteratively evaluated at each step of gradient descent. This is evident in Fig. 3.4 where direct optimization with ANTs runs much faster than the Lie-algebra counterpart. Moreover, the exponential map is only *locally* diffeomorphic, meaning it is suitable for modelling deviations close to the identity but not for large deformations – this leads to less expressivity and poor performance. In Fig. 3.4, the greedy SyN method which employs direct optimization significantly outperforms the Lie algebra-based DARTEL. In Fig. 3.1 we observed that across a large variety of hyper-parameters evaluated via grid search, direct optimization consistently led to better target overlap compared to its Lie algebra counterpart on the LPBA40 dataset. Therefore, Lie algebra based methods are typically empirically observed to be suboptimal in representational power.

3.3.2. Limitations of Stationary Velocity Fields

A common approach in diffeomorphic registration is to use stationary velocity fields, i.e. velocity fields that are constant in time. This velocity field is also an element of the Lie algebra that can be used to generate a diffeomorphism using the exponential map. Since the velocity field itself resides in Euclidean space, adaptive optimization algorithm like Adam can be applied to optimize the velocity field. Many deep learning methods employ this approach since it is hard to produce valid diffeomorphic transforms using a network but it is easy

to produce a valid stationary velocity field. CLAIRE (Mang, 2024) mentions the limitations of this approach without further elaboration. We discuss three limitations of the SVF based optimization approach:

Computational Cost Optimizing the velocity field requires computing the exponential map using the scaling-and-squaring approach (Eq. (3.5)) to obtain the diffeomorphism. Typical registration pipelines run scaling-and-squaring 6-8 times to obtain the diffeomorphism, and the backprop requires another 6-8 steps of iterative backward calls to compute the gradient of the velocity field. This is a significantly expensive operation performed *every iteration* of the optimization, leading to a substantial slowdown in runtime. In contrast, direct optimization requires only one warp composition to perform the diffeomorphic update $\varphi_{t+1} = \varphi_t \circ (id + \eta_t v_t)$. The significant difference in runtime is observed for ANTs and DARTEL in Fig. 3.6b. SVFs also cannot represent diffeomorphisms that are integrals of time-dependent velocity fields, which are more flexible and can represent a wider range of deformations (Wu et al., 2022a, 2024; Mang, 2024).

Tradeoff between Numerical Accuracy and Computational Cost Exponential maps of SVFs are mathematically diffeomorphic in nature, but it is observed that numerically SVFs may not be diffeomorphic. Other works (Wu et al., 2022a) show that SVF based baselines like Log Demons have significantly more non-diffeomorphic voxels than time-dependent velocity fields like SyN and NODEO. We see similar trends in Fig. 3.4c, where the SVF based DARTEL has significantly more non-diffeomorphic voxels than the direct optimization in ANTs and FireANTs. This numerical inaccuracy traces back to the discretization error in the scaling and squaring approach, which is essentially Euler integration of the velocity field. In the base case of the scaling-and-squaring approach (Eq. (3.4)), $\varphi^{(1/2^M)} = id + v_0/2^M$ is not guaranteed to be a diffeomorphism unless the Lipschitz constant of v_0, denoted as $LP(v_0)$ is less than 2^M. The inductive recursion step (Eq. (3.5)) only preserves the diffeomorphic property if the base case is a diffeomorphism, otherwise the non-diffeomorphism can propagate throughout the subsequent steps to the final diffeomorphism. Here, we provide a proof that the base case is a diffeomorphism only for v_0 with Lipschitz constant less than 2^M, showing that velocity fields with large deformations or highly variable deformations require finer Euler integration steps (i.e. larger M) to ensure numerical diffeomorphisms.

Theorem 1. For a $C^\infty(\Omega, \mathbb{R}^d)$ velocity field v_0 with compact support on Ω such that $v_0(x) = 0$ on $x \in \partial\Omega$, the transform $\varphi = id + \epsilon v_0$ is a diffeomorphism for $|\epsilon| < 1/LP(v_0)$, where $LP(v_0)$ is the Lipschitz constant of v_0.

Proof. Since v_0 is a $C^\infty(\Omega, \mathbb{R}^d)$ (is continuously differentiable and is compact on Ω) velocity field, the Jacobian of the velocity exists, and is denoted as $J(v_0)$. We invoke the Hadamard's global inverse function theorem (Hadamard, 1906) (HGIF theorem) to show that φ is a diffeomorphism for $|\epsilon| < 1/LP(v_0)$.

The HGIF theorem requires that φ is smooth (true by our definition), and the Jacobian of the transformation is non-singular for all $x \in \Omega$, and that $\|J\varphi(x)^{-1}\|$ is bounded for all $x \in \Omega$. Since v_0 is defined only on a compact domain Ω, we use the Whitney extension theorem (Whitney, 1992) to extend v_0 to a $C^\infty(\mathbb{R}^d)$ velocity field by simply setting $v_0(x) = 0$ for $x \in \mathbb{R}^d \setminus \Omega$.

For $x \in \mathbb{R}^d \setminus \Omega$, we have $\varphi(x) = x$, and therefore $J\varphi(x) = I$ which is invertible, and $\|J\varphi(x)^{-1}\| = 1$ which is bounded.

For $x \in \Omega$, we have

$$J\varphi(x) = I + \epsilon J v_0(x) \tag{3.7}$$

$$\Rightarrow \|J\varphi(x) - I\| = |\epsilon| \|Jv_0(x)\| \leq |\epsilon| LP(v_0) < 1 \tag{3.8}$$

since $|\epsilon| < 1/LP(v_0)$. Since $\|J\varphi(x) - I\| < 1$, $J\varphi(x)$ is non-singular for all $x \in \Omega$ from the Neumann convergent series of matrix $(I - A)$ (i.e. $A^{-1} = \sum_{k=0}^{\infty}(I - A)^k$).

For $A = J\varphi(x)$, we have $\|I - A\| < 1$ from the above inequality. Let $\|I - A\| \leq \delta$ for some $\delta < 1$. Using the Neumann convergent series of matrix $(I - A)$ (i.e. $A^{-1} = \sum_{k=0}^{\infty}(I - A)^k$ for $\|I - A\| < 1$), we have

$$\|A^{-1}\| \leq \sum_{k=0}^{\infty} \|(I - A)^k\| \tag{3.9}$$

$$\leq \sum_{k=0}^{\infty} \delta^k = \frac{1}{1 - \delta} \tag{3.10}$$

This shows that $\|J\varphi(x)^{-1}\| \leq \frac{1}{1-\delta}$ for all $x \in \Omega$ and is bounded.

Since φ is $C^{\infty}(\mathbb{R}^d, \mathbb{R}^d)$, and the Jacobian is non-singular and its inverse is bounded for all $x \in \mathbb{R}^d$, we have that φ is a diffeomorphism for all $x \in \mathbb{R}^d$. □

If $|\epsilon| \geq 1/LP(v_0)$, then $\|J\varphi(x) - I\| < 1$ may not hold and the Jacobian may be singular for some $x \in \Omega$, breaking local invertibility. When scaling-and-squaring is employed, the velocity field might have large magnitudes to capture large deformations during optimization, leading to a large Lipschitz constant. Fixing the number of integration steps M can lead to numerical non-diffeomorphisms if the Lipschitz constant exceeds 2^M. In principle, M should adaptively chosen to the lowest value such that $2^M > LP(v_0)$. We validate this empirically in Section B.2.

Sensitivity to Perturbations and limits of Expressivity Another potential source of numerical instability arises from the sensitivity of diffeomorphisms to perturbations in their underlying velocity fields. While perturbation and sensitivity analyses are well established for matrix exponentials, often showing that output deviations grow exponentially with the norm of input perturbations (Van Loan, 1977; Zhu et al., 2008). Several works have also investigated the singularities of the Euler equation (Drivas and Elgindi, 2023; Preston, 2004; Lee, 2018) that leads to blowups in geodesic flows that represent diffeomorphisms. Unlike finite-dimensional Lie groups, the derivative of the exponential can fail to be surjective, possibly producing ill-conditioning and numerical instability near certain vector fields representing conjugate directions (Ebin et al., 2006). SVFs have a few limitations in regards to expressivity. For example, there are diffeomorphisms arbitrarily close to the identity that are not contained in flows (1-parameter subgroups or SVFs) (Milnor, 1984), showing that the exponential map is not surjective to the group of diffeomorphisms even locally. The strong dependence on initial conditions and parameters observed in such systems suggests that analogous sensitivities may contribute to numerical instability and inexpressivity in SVF-based optimization methods. This is empirically observed in Fig. 3.1, where the *exp* representation underperforms for the same cost function and dataset, potentially due to instability and inexpressivity being factors in the small performance degradation since the other parameters of the optimization (loss function, regularization) are kept constant or determined using cross-validation (optimal learning rate, for example).

Direct optimization of diffeomorphisms do not suffer from these limitations, since the perturbations of the diffeomorphism (outputs) are controlled directly by the magnitude of the velocity field in the update rule $\varphi_{t+1} = \varphi_t \circ (id + \epsilon_t v_t)$, and any diffeomorphism close to the identity can be obtained trivially by controlling v_t and ϵ_t appropriately.

3.3.3. Riemannian Gradient Descent

Solving the registration problem directly on the space of diffeomorphisms avoids repeated computations to and fro via the exponential map. The downside however is that one now has to explicitly account for the curvature and tangent spaces of the manifold. The updates for Riemannian gradient descent (Boumal et al., 2014) at the t^{th} iteration are

$$\begin{aligned} \varphi_{t+1} &= \exp_{\varphi_t}\left(-\eta \operatorname{Proj}_{\varphi_t}(\nabla_\varphi L)\right) \\ \text{where } \nabla_\varphi L &= g_{\varphi_t}^{-1} \frac{\partial L}{\partial \varphi}, \end{aligned} \tag{3.11}$$

where one pulls back the Euclidean gradient $\frac{\partial L}{\partial \varphi}$ onto the manifold using the inverse metric tensor g (which makes the gradient invariant to the parameterization of the manifold of diffeomorphisms) before projecting it to the tangent space using Proj_{φ_t}. Since the tangent space is a local first-order approximation of the manifold's surface, we can move along this descent direction by a step-size η and compute the updated diffeomorphism φ_{t+1}, represented as the exponential map from φ_t computed in the direction of $-\text{Proj}_{\varphi_t}(\nabla_\varphi L)$.

However, there are a few challenges in optimizing diffeomorphisms using Riemannian gradient descent. First, adaptive optimization algorithms such as RMSProp (Tieleman et al., 2012), Adagrad (Duchi et al., 2011) and Adam (Kingma and Ba, 2014) have become popular because they can handle poorly conditioned optimization problems in deep learning. Variants for optimization on low-dimensional Riemannian manifold exist (Bonnabel, 2013; Zhang et al., 2016; Bécigneul and Ganea, 2018; Kochurov et al., 2020). In contrast to these manifolds, diffeomorphisms are a high-dimensional variable-sized group (e.g., the parameterization of the warp field scales with that of the image size). Therefore, operations like computing the Riemannian metric tensor, and parallel transport of the optimization state variables (momentum and curvature) are very computationally expensive. For diffeomorphisms, computing the parallel transport requires solving a system of partial differential equations, which is computationally expensive. For these reasons, we do not consider direct Riemannian optimization for diffeomorphisms in our work.

3.4. Exploiting the group structure of diffeomorphisms

Diffeomorphisms are imbued with additional structure compared to a Riemannian manifold – they are a Lie group as well. Not all Riemannian manifolds are Lie groups - notable examples of non-Lie group Riemannanian manifolds include the sphere $\mathbb{S}^n$, fixed-rank matrices, and the Stiefel and Oblique manifolds (Boumal et al., 2014). The additional Lie group structure of $G(\Omega, \mathbb{R}^d)$ allows us to exploit the group action to define a gradient descent algorithm that eliminates computationally expensive steps. In the following text, we provide a novel method to compute a descent direction in the group of diffeomorphisms that is computationally efficient and can be used with adaptive optimization algorithms.

Minimizing the Eulerian differential. Consider a function $U : G \to \mathbb{R}$ that we aim to minimize. Let V be an admissible Hilbert space of vector fields on Ω embedded in $C_0^1(\Omega, \mathbb{R}^d)$. We define an *Eulerian differential*

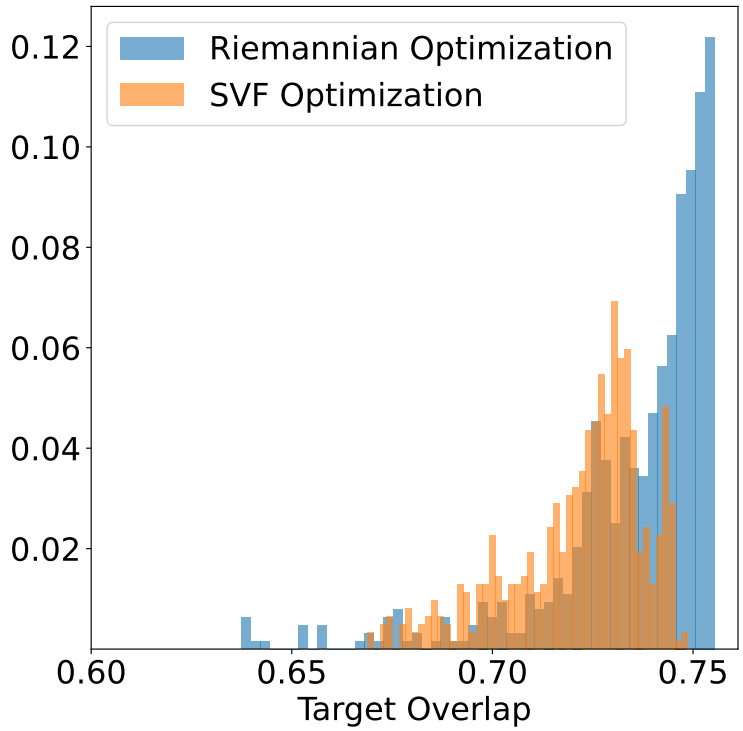


Figure 3.1: Comparison of exponential versus direct optimization on LPBA40 dataset: We run the hyperparameter grid search on the LPBA40 dataset using direct Eulerian updates with Adam optimizer (denoted as *rgd*), and optimizing the velocity field by computing the exponential map to represent the diffeomorphism (denoted as *exp*) across all the configurations shown in Fig. 3.7(a). The average target overlap for each configuration is then stored, and a histogram of target overlap values of the dataset is constructed. Note that the *rgd* variant has a significantly more number of configurations near the optimal value, and the average performance and the overall distribution of our optimization is better for the *rgd* variant than *exp*. Similar trends can be observed for the EMPIRE10 lung challenge in Fig. 3.4, where the *exp* representation underperforms for the same cost function, data, etc. Therefore, we recommend direct Eulerian optimization for diffeomorphisms.

in V if there exists a linear form $\partial\bar{U} \in V^*$ such that for all $v \in V$:

$$\left(\partial\bar{U}(\varphi)|v\right)_E = \partial_t U(\varphi \circ \varphi_{0t}^v)\Big|_{t=0} \tag{3.12}$$

and $\varphi_{0t}^v(x) = \exp_{id}(tv)(x) = x + \int_0^t v(\varphi_{0s}^v(x))ds$ is the flow of the vector field v starting from the identity. This definition of Eulerian differential is different from the one in (Younes, 2010) to perform all updates (v) in the tangent space at identity and leverage Jacobian-free descent (see later). The goal is to choose a suitable v such that the directional change of the Eulerian differential along v is negative, making v a descent direction. A more familiar rate of change of U along a curve v is given by the *Gateaux derivative*:

$$\left(\frac{\partial U}{\partial \varphi}\bigg|v\right)_G = \partial_t U(\varphi + tv)\Big|_{t=0} \tag{3.13}$$

The Eulerian differential is closely related to the Gateaux derivative of U at φ as:

$$\left(\partial \bar{U}(\varphi)|v\right)_E = \left(\frac{\partial U}{\partial \varphi}\middle| J(\varphi)v\right)_G$$

using chain rule. The right side is further expanded as:

$$\left(\partial \bar{U}(\varphi)|v\right)_E = \int_\Omega \left(\frac{\partial U}{\partial \varphi}(\varphi)(x)\right)^\top J(\varphi(x))v(x)dx$$

where $J(\varphi)(x) = J(\varphi(x))$ with slight abuse of notation. We introduce the Gateaux derivative and relate it to the Eulerian derivative because we typically have access to the Gateaux derivative using automatic differentiation tools like PyTorch, but to perform optimization on the group of diffeomorphisms, we need to compute the Eulerian differential. Choosing

$$v_d(x) = -J(\varphi(x))^\top \frac{\partial U}{\partial \varphi}(\varphi)(x)$$

gives us:

$$\left(\partial \bar{U}(\varphi)|v_d\right)_E = -\int_\Omega \left\| J(\varphi)^\top \frac{\partial U}{\partial \varphi}(\varphi)(x)\right\|^2 dx < 0$$

This choice of $v_d(x)$ is therefore a descent direction for the Eulerian differential of U at φ. To perform gradient descent on the Eulerian differential at φ, we need to compute the descent direction v_d, perform the exponential map with a small learning rate η_t, and perform the update:

$$\varphi_{t+1} = \varphi_t \circ \exp_{id}(\eta_t v_d)$$

For small enough η_t, the exponential map can be approximated with a retraction map (i.e. $\exp_{id}(\eta_t v_d) \approx id + \eta_t v_d$), which is quick to compute.

We quickly contextualize the key differences between Gateaux gradient descent and our proposed Eulerian descent. First, the steepest descent direction in Gateaux gradient descent is $-\frac{\partial U}{\partial \varphi}(\varphi)$, whereas it is $-J(\varphi)^\top \frac{\partial U}{\partial \varphi}(\varphi)$ in Eulerian descent. Second, the update rule in Gateaux gradient descent is $\varphi_{t+1} = \varphi_t - \eta_t \frac{\partial U}{\partial \varphi}(\varphi)$, whereas it is $\varphi_{t+1} = \varphi_t \circ \exp_{id}(\eta_t v_d)$ in Eulerian descent. These two differences capture the essence of performing optimization on the group of diffeomorphisms in contrast to optimizing on the (Euclidean) ambient space directly.

Adaptive optimization on diffeomorphisms. Note that for small enough t, the descent direction $v_d(x)$ can also be interpreted as a vector in the tangent space at identity, with $\varphi^v_{0t} = \exp_{id}(tv)$ since $\varphi^v_{00} = id$, and $\partial_t \varphi^v_{0t}|_{t=0} = v$. Descent directions over gradient descent iterations i denoted as $v_d^{(i)}$ all lie on the same vector space, i.e. the tangent space at identity. Therefore, first order algorithms like Adam can be applied on the sequence of descent directions $v_d^{(i)}$ which now lie in the same vector space, without requiring computing the metric tensor, parallel transport or change of coordinates (charts) throughout the optimization process. This framework leveraging the group structure forms the core of our adaptive optimization algorithm for diffeomorphisms. Our framework is therefore a significant advantage over Riemannian optimization methods

which require parallel transport of the momentum and curvature vectors at each iteration.

Jacobian-Free Eulerian Descent

We provided an obvious choice of descent direction $v_d(x)$ for the Eulerian differential of U at φ. The Gateaux derivative $\frac{\partial U}{\partial \varphi}$ is readily obtained using automatic differentiation tools like PyTorch. However, the descent direction requires us to multiply this derivative with the Jacobian of the diffeomorphism $J(\varphi)$, which may be computationally expensive. However, in most diffeomorphic image registration applications, the role of the diffeomorphism is warp the image by performing local translations, scaling and shearing without introducing large local rotations. Mathematically, we consider the polar form of the Jacobian $J(\varphi)(x) = U(x)P(x)$ where $U(x)$ is a unitary matrix, and $P(x)$ is a positive definite matrix. We assume that for most applications, $U(x) \approx I_{d\times d}$, making $J(\varphi)(x)$ positive definite. With this assumption, we can choose the modified descent direction

$$v'_d(x) = -\frac{\partial U}{\partial \varphi}(\varphi)(x)$$

and the Eulerian differential at φ is

$$\left(\partial \bar{U} | v'_d\right)_E = -\int_\Omega \left(\frac{\partial U}{\partial \varphi}(\varphi)(x)\right)^\top J(\varphi(x)) \frac{\partial U}{\partial \varphi}(\varphi)(x) dx < 0$$

since $v'_d(x)^\top J(\varphi(x)) v'_d(x) \geq 0$ for all $x \in \Omega$, owing to the (assumed) positive definiteness of $J(\varphi)(x)$. For all experiments, Jacobian-free descent directions $v'_d(x)$ are used, and they provide faster runtime and with same accuracy. Adaptive first-order optimization can now be performed on this modified sequence on descent directions $v'^{(i)}_d(x)$, saving significant computational and memory overhead by avoiding computation of $J(\varphi)$.

Note that this algorithm using the Eulerian differential is only possible due to the group structure of diffeomorphisms. For an arbitrary Riemannanian manifold $\mathcal{M}$ and points $\varphi, \varphi^v_{0t} \in \mathcal{M}$, the operation $\varphi \circ \varphi^v_{0t}$ does not make sense. The additional group structure of $G(\Omega, \mathbb{R}^d)$ allows us to propose a novel Eulerian descent algorithm without performing Lie algebra optimization, or Riemannian gradient descent, both of which are computationally expensive for diffeomorphisms.

Alternative formulations for Eulerian differential

Our definition of Eq. (3.12) is different from the one in (Younes, 2010) in two subtle but important ways. First, we do not define the registration objective in terms of the group action or pullback image $\varphi.I = I(\varphi^{-1}(x))$, and instead define the objective in terms of the pushforward image $I(\varphi(x))$. This is to avoid computing and storing both φ for autodifferentiation and φ^{-1} for computing the objective, implementational simplicity, and consistency with more modern registration framework formulation. FireANTs provides additional functionality to compute φ^{-1} *post hoc* using a multi-scale objective function similar to the image matching objective: $\varphi^{-1} = \arg\min_{\psi \in G} \sum_{x\in\Omega} \|\psi(\varphi(x)) - x\|_2^2 + \|\varphi(\psi(x)) - x\|_2^2$. This allows researchers to obtain an inverse transform on a *post hoc* basis without computing φ^{-1} *during* optimization. This subroutine is also used in the symmetric registration objective to compute the final transformation $\varphi = \varphi_M \circ \varphi_F^{-1}$. The second difference is the definition of the Eulerian differential itself - note that we use the composition $U(\varphi \circ \varphi^v_{0t})$ in Eq. (3.12) instead of $U(\varphi^v_{0t} \circ \varphi)$. Defining the Eulerian differential using the second formulation without using the group action $\varphi.I$ implies that we will compute the velocity field in the Langrangian frame, i.e.

$V(y) = v(\varphi(x)) = -\frac{\partial U}{\partial \varphi}$. To compute adaptive optimization updates and Lipschitz constant to scale the learning rate (see Section 3.3.2), we need to compute the corresponding velocity field in the Eulerian frame, i.e. $v(x) = V(\varphi^{-1}(y))$ which requires computing φ^{-1}. Therefore, we choose a definition that avoids computing φ^{-1} and allows inexpensive adaptive optimization updates.

3.5. Interpolation strategies for multi-scale registration

Classical approaches to deformable image registration is performed in a multi-scale manner. Specifically, an image pyramid is constructed from the fixed and moving images by downsampling them at different scales, usually in increasing powers of two. Optimization is performed at the coarsest scale first, and the resulting transformation at each level is used to initialize the optimization at the next finer scale. Specifically, for the fixed image I and the moving image I' and K levels, let the downsampled versions be $\{I_k\}_{k=1}^K$ and $\{I'_k\}_{k=1}^K$, where k is the scale index from coarsest to finest. At the k-th scale, the transformation φ_k is optimized as

$$\varphi_k^* = \arg\min_{\varphi_k \in G} L(I_k, I'_k \circ \varphi_k)$$

where φ_k is initialized as

$$\varphi_k = \begin{cases} id & \text{if } k = 1 \\ \text{Upsample}(\varphi_{k-1}) & \text{otherwise} \end{cases}$$

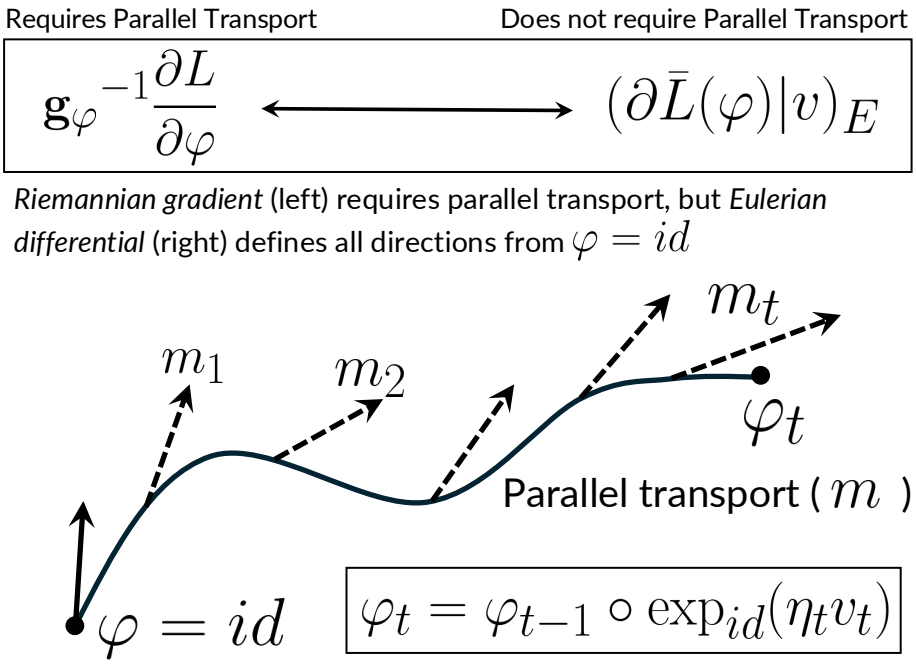
Unlike existing gradient descent based approaches, our Riemannian adaptive optimizer also contains state variables m_k corresponding to the momentum and ν_k corresponding to the EMA of squared gradient, at the same scale as φ_k, which require upsampling as well.

Unlike upsampling images, upsampling warp fields and their corresponding optimizer state variables requires careful consideration of the interpolation strategy. Bicubic interpolation is a commonly used strategy for upsampling images to preserve smoothness and avoid aliasing. However, bicubic interpolation of the warp field can lead to overshooting, leading to introducing singularities in the upsampled displacement field when there existed none in the original displacement field. In contrast, bilinear or trilinear interpolation does not lead to overshooting, and therefore diffeomorphism of the upsampled displacement is guaranteed, if the original displacement is diffeomorphic. We demonstrate this using a simple 2D warp field in Fig. 3.2(b). On the left, we consider a warp field created by nonlinear shear forces. This warp field does not contain any tears or folds - and is diffeomorphic. We upsample this warp field using bicubic interpolation (top) and bilinear interpolation (bottom). We also plot a heatmap of the negative of the determinant of the Jacobian of the upsampled warp, with a white contour representing the zero level set. Qualitatively, bicubic interpolation introduces noticable folds in the warping field, leading to non-diffeomorphisms in the upsampled warp field. The heatmap shows a significant portion of the upsampled warp field has a negative determinant, indicating non-invertibility. On the other hand, bilinear interpolation looks blocky but preserves diffeomorphism everywhere, as also quantitatively verified by the absence of a zero level set in the heatmap. The complete algorithm is described in Algorithm 5.

3.6. Results

FireANTs represents the next generation of frameworks superseding the widely established and successful adoption of the ANTs ecosystem spanning the gamut of biomedical and life sciences research. We evaluate our method on fourteen datasets spanning more than 15,000 image pairs, three organs (brain, lung, abdomen),

(a) Trick to avoid parallel transport in Riemannian Adaptive Optimization using Eulerian differentials



(b) Bicubic interpolation of diffeomorphic map does not preserve diffeomorphism

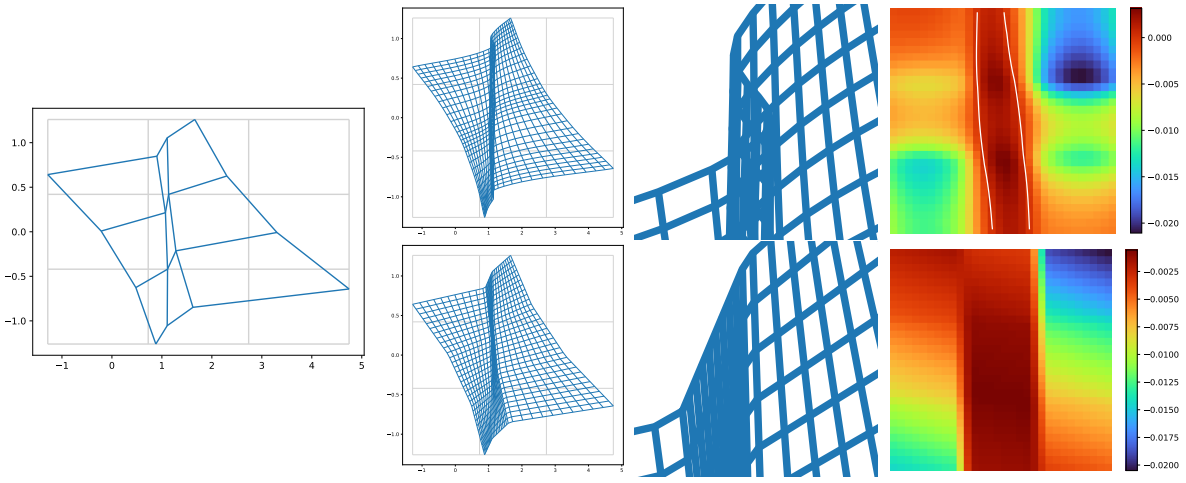


Figure 3.2: Overview of tricks for multi-scale adaptive optimization for diffeomorphisms: (a) We exploit the group structure of diffeomorphisms to define an Eulerian differential that avoids the need for parallel transport in adaptive optimization algorithms. (b) We show the effect of downsampling on the warp and determinant of the Jacobian for a single image pair. The first column shows the initial warp, and the second and third columns show the warp and determinant of the Jacobian for the cubic and bilinear interpolations, respectively.

seven modalities (T1w MRI, T2*w MRI, CT, expansion microscopy, LSFM, fMOST, 9.4T MRI), six species (human, ovine, zebrafish, mouse, rat, non-human primates), and show remarkable generalization across all datasets without any domain-specific enhancements. Since our method is a methodological improvement over ANTs, we evaluate FireANTs against established benchmarks where ANTs is one of the top performing methods, among other winning methods for the respective challenges.

FireANTs demonstrates remarkable runtime efficiency compared to ANTs on both CPU and GPU, while also outperforming most deep learning methods at inference runtime and consuming up to a tenth of the GPU memory, setting a new standard for runtime and memory efficiency. This unprecedented efficiency allows a multitude of novel capabilities including registering 50× larger volumes in minutes on a single GPU, faster amortized runtimes over large batches enabling scalable registration of large datasets, efficient hyperparameter grid search studies, and high-resolution atlas building in under 25 minutes.

3.6.1. Experiment Setup

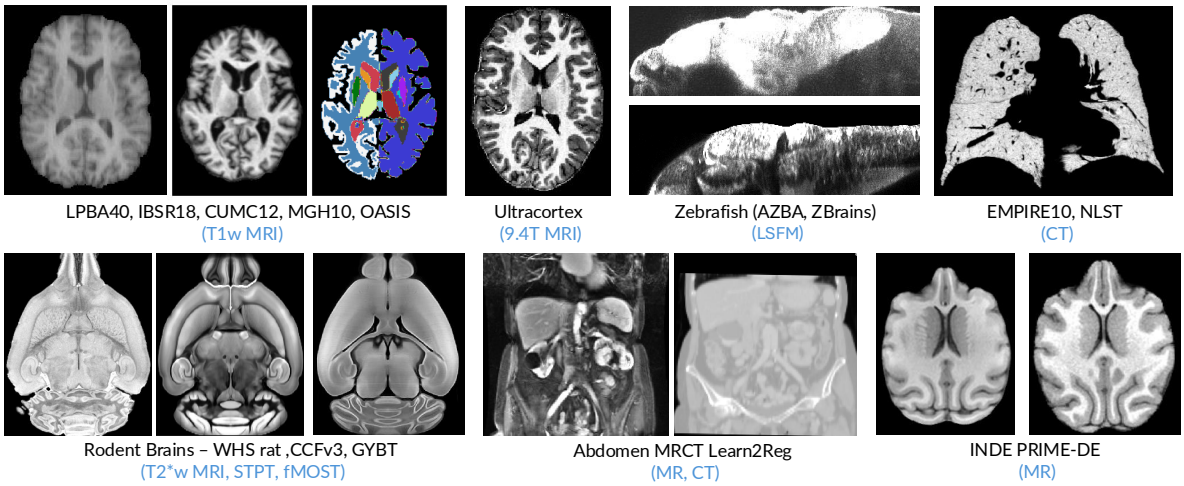
We briefly describe the significance, existing state-of-the-art and challenges associated with the chosen benchmarks to demonstrate the efficacy of FireANTs. More details about the datasets and evaluation metrics are outlined in Section B.3.

In-vivo brain mapping challenges (Klein et al., 2009; Marcus et al., 2007b) Klein *et al.* (Klein et al., 2009) in their landmark paper reported an extensive evaluation of fourteen state-of-the-art registration algorithms on four neuroimaging datasets. The four neuroimaging datasets (IBSR18, CUMC12, MGH10, LPBA40) comprise different whole-brain labelling protocols, eight different evaluation measures and three independent analysis methods of over 2000 brain volume pairs. The Learn2Reg (Hering et al., 2022) version of the OASIS dataset (Marcus et al., 2007b) is another large scale dataset with 414 subjects for inter-subject brain MRI registration, routinely used as a training dataset for deep learning algorithms (Balakrishnan et al., 2019; Jia et al., 2022; Tian et al., 2024). Evaluating on these challenges is therefore imperative to establish FireANTs as an effective, versatile and robust algorithm for neuroimaging applications. In total, we compare with state-of-the-art baselines on over *2500 brain volume pairs*, with varying number of labeled anatomical regions and resolutions.

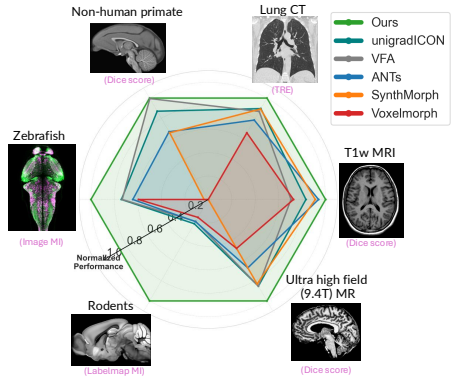
PRIMatE Data Exchange (PRIME-DE) (Milham et al., 2018b) The overarching goal of PRIMatE Data Exchange (PRIME-DE) is to create an open science resource for the neuroimaging community to facilitate the mapping of the non-human primate connectome. The dataset features a familiar modality and anatomy (T1w MRI brain) but different structural organization (non-human primate). This presents a challenge to compare the generalization capabilities of domain-agnostic or foundational registration algorithms with FireANTs.

Ultracortex (Mahler et al., 2024) The Ultracortex dataset hosts a unique collection of ultra-high field (9.4 Tesla) MRI data of the human brain. This challenge provides a complementary problem to PRIME-DE - familiar anatomy and structural organization (human brain) but different modality (9.4T MRI) and resolution (sub-millimeters).

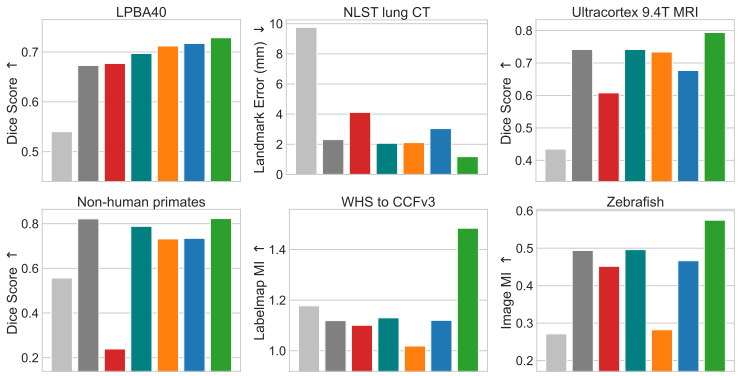
Waxholm Rat Brain and Allen CCFv3 mouse brain datasets (Kleven et al., 2023; Wang et al., 2020c) The datasets feature high-resolution atlases of the rat and mouse brain with different modalities (T2*w MRI and STPT) respectively. The motivation for using these datasets is to provide a benchmark for *cross-species, multimodal* registration. This addresses the growing need to map neuroanatomy across species (Mezias et al., 2024; Beauchamp et al., 2022), which is central to revealing the core evolutionary computational motifs and unique adaptations to handle specific ecological and behavioral demands.



(a) Overview of the datasets used in the work



(b) Normalized performance of state-of-the-art registration algorithms across a wide range of datasets and benchmarks. FireANTs achieves asymptotically best normalized performance across various datasets and evaluation criteria.



(c) Raw performance (measured by Dice Score, Mutual Information of Labelmap and Intensity, Landmark Distance) of state-of-the-art registration algorithms across various datasets. In all datasets except NLST, higher scores are better. Colorbars are shown in Section 3.6.1 with lightgray denoting zero displacement (baseline).

Figure 3.3: FireANTs can generalize to a large variety of modalities and datasets: Registration quality is validated by measuring either the labelmap overlap, Mutual Information between aligned labelmap for different labelmaps across datasets, or anatomical landmark distance between the fixed and warped coordinate frames. We consider two community standard challenges where ANTs was the winner, two analogous contemporary challenges to enable broader comparison with deep learning methods, and five other scenarios spanning a broad set of challenges. Across six datasets spanning a spectrum of anatomical systems, species, and modalities, FireANTs achieves the best performance across all evaluation criteria, showcasing its generalization capabilities.

Lung CT mapping challenges (Murphy et al., 2011; Team, 2011) Pulmonary registration has significant clinical applications, including aligning breath-hold scans for visual comparison, modeling lung expansion, and tracking disease progression. Murphy et al. introduced the EMPIRE10 challenge (Murphy et al., 2011) to facilitate the evaluation of CT lung registration algorithms, including inspiration-expiration, breath-hold over time, 4D, ovine, contrast-noncontrast, and artificially warped scans. EMPIRE10 provides only scan pairs and binary lung masks, withholding fissures and landmarks for *private* evaluation. The scans vary in spatial and physical resolution, necessitating a registration algorithm agnostic to anisotropy in both voxel and physical space. The National Lung Screening Trial (NLST) (Team, 2011) subset curated by Learn2Reg challenge is another widely used community-standard dataset. It consists of 210 intra-subject lung pairs, with low-dose helical CT scans with limited field of view and high-dose scans with full field of view, supplemented with more than a thousand keypoints per subject pair. This challenge provides a benchmark for comparison of methods beyond the neuroanatomical domain.

RnR ExM Mouse Isocortex Dataset (rnr) The RnR-ExM challenge evaluates the ability to perform non linear deformable registration on ultra-high-resolution images. Out of the three species (mouse brain, C. elegans, zebrafish), the mouse isocortex dataset is the only dataset with non-trivial non-linear deformations. Registration of high-resolution sub-micron volumes is imperative to creating and understanding the comprehensive cell atlas of the mammalian brain at scale. The voxel size of each image volume is $2048 \times 2048 \times 81$ and the voxel spacing is 0.1625μm $\times$ 0.1625μm $\times$ 0.4μm. These volume sizes are about two orders of magnitude larger compared to existing biomedical datasets, representing a significant challenge in quick and scalable registration.

AZBA and ZBrain datasets (Kenney et al., 2021a; Randlett et al., 2015) The ZBrain atlas is an anatomical and functional reference constructed from high-resolution confocal microscopy images of larval zebrafish (6 days post fertilization) expressing nuclear and cytoplasmic fluorescent markers. The AZBA atlas, in contrast, represents the adult zebrafish brain at cellular resolution. Registration of these templates enables a powerful cross-developmental comparison between the larval and adult zebrafish brains. Conceptually, this tsak establishes spatial correspondence between larval and adult brain regions, providing a foundation for developmental neuroanatomy. We use this dataset to conduct preliminary experiments to access the generalization capabilities of registration algorithms to cross-developmental data on an unseen species and modality.

BICCN Mouse Dataset The high-throughput and high-resolution fluorescence micro-optical sectioning tomography (fMOST) platform (Zheng et al., 2013; Gong et al., 2016a) was used to image 55 mouse brains containing gene-defined neuron populations. The brains are imaged at a resolution of $0.35 \times 0.35 \times 1.0\mu m^3$. The dataset is used to generate a $25\mu m$-resolution atlas of the mouse brain in under 25 minutes. This unprecedented scale, enabled by FireANTs, will advance multimodal integration, standardize cross-species comparisons, and drive scalable, reproducible neuroscience research highly pertinent to large-scale collaborative efforts such as BICCN and BICAN.

Learn2Reg Abdomen MRCT registration (Hering et al., 2022) The dataset features intra-patient multimodal abdominal MRI and CT registration (122 scans in total) for diagnostic and follow-up. We use this dataset as a testbed to ablate the effect of Jacobian-free optimization on abdominal MRCT registration.

3.6.2. Results on generalization to long-tail of modalities

Generalization to unseen modalities, species, resolutions, and anatomical organization is a central requirement for accessible and scalable registration algorithms. Model-free optimization algorithms generalize well on

clinical datasets, but the increased heterogenity on large-scale datasets presents a significant challenge in terms of convergence and runtime. Deep Learning methods typically do not generalize well beyond the data distribution seen during training (Balakrishnan et al., 2019), although domain-agnostic or foundational methods (Hoffmann et al., 2021; Tian et al., 2024) have shown promising results. Other methods (Liu et al., 2024c) claim generalization due to architectural design encoding inductive biases for the task. Therefore, we compare FireANTs against ANTs, VoxelMorph as a DL baseline, and SynthMorph, unigradICON, and VFA as other methods that claim generalization to unseen data.

Fig. 3.3 shows the performance on six datasets - LPBA40, NLST, Ultracortex, PRIME-DE, Zebrafish (ZBrain and AZBA), and Rodents (Waxholm and CCFv3) encompassing four evaluation criteria (Anatomical Label overlap, Landmark Distance, Mutual information of registered Intensity and Labelmap volumes). The normalized performance is obtained by rescaling the performance of the baseline to 0 and the best performing method to 1. The radar chart shows the generalization of FireANTs across all datasets, and individual plots show unnormalized performance on each dataset. SynthMorph, unigradICON, and VFA perform at par with ANTs on the brain datasets (Ultracortex, PRIME-DE), but severely underperform on the Zebrafish and Rodent datasets. Surprisingly, VFA also underperforms compared to other methods on the LPBA40 dataset, showing a potential weakness in registering highly parcelled anatomical regions. FireANTs achieves the best performance on *all* datasets, and performs significantly better on the multimodal cross-species registration task. This establishes FireANTs as a general-purpose registration algorithm that can be used across a wide range of modalities, species, and resolutions.

3.6.3. Results on state-of-the-art biomedical benchmarks

FireANTs proposes an algorithmic improvement over the state-of-the-art ANTs toolkit. As such, we compare FireANTs with ANTs on community standard benchmarks where ANTs is established as one of the top performing methods. This include Klein *et al.* (Klein et al., 2009) neuroimaging challenge, EMPIRE10 (Murphy et al., 2011) pulmonary challenge. We also add comparisons on two contemporary benchmarks for neuroanatomy (OASIS) and pulmonary (NLST) datasets from the Learn2Reg (Hering et al., 2022) challenge, which is of high relevance to the neuroimaging and pulmonary communities.

In-vivo brain MRI mapping We compare FireANTs with two state-of-the-art optimization algorithms on Klein *et al.* (Klein et al., 2009): ANTs - which won the original challenge, and Symmetric Log Demons (Vercauteren et al., 2007b), and two widely used deep learning algorithms: VoxelMorph (Balakrishnan et al., 2019) and SynthMorph (Hoffmann et al., 2021) using their provided pretrained models. Since no methods utilize label maps, we run registration on all 414 image pairs prescribed in the dataset. Results for the brain datasets are shown in Fig. 3.4a, Fig. B.8. Our algorithm outperforms all baselines on four out of five datasets, with an improvement in *all* metrics evaluating the volume overlap of the fixed and warped label maps. The improvements are consistent across varying parcellations and relative sizes of anatomical label maps. In the IBSR18 and CUMC12 datasets, the median target overlap of our method is better than the third-quantile of ANTs. Fig. B.8 also highlights the improvement in label overlap per labeled brain region across all datasets. For deep learning methods, a noticable performance drop is observed when the anisotropic volumes are fed into the network, which is undesirable as the trained model is essentially 'locked' to a single physical resolution - which limits the generalizability of the model to various modalities with different physical resolutions. For Demons, ANTs, and FireANTs (Ours), we do not perform any additional normalization or resampling. On the OASIS dataset, all methods perform at par with each other with no significant differences. SynthMorph is more robust to the domain gap than VoxelMorph due to its training strategy with synthetic images, but still underperforms

optimization baselines when their recommended hyperparameters are chosen.

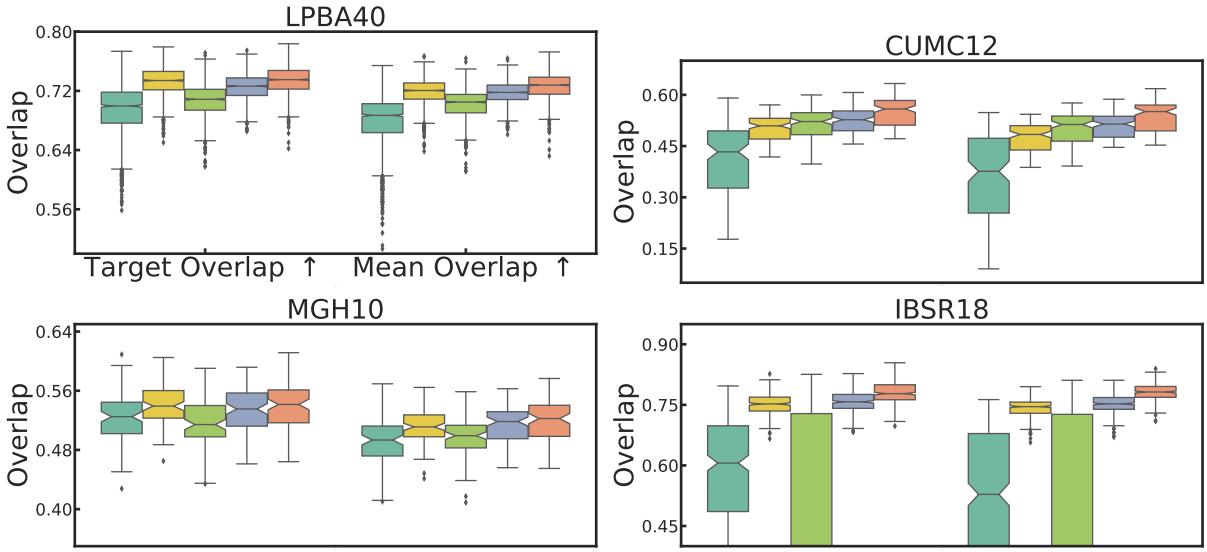
Lung CT mapping challenges The EMPIRE10 lung dataset(Murphy et al., 2011) consists of volumes that are about 10× larger than the brain dataset, thereby presenting a scaling challenge for deformable registration algorithms. Evaluation is done with privately withheld labels; we use the provided results from the leaderboard to compare with other methods. We evaluate three criteria: (1) fissure alignment errors (%)—the fraction of misaligned fissure voxels (Figs. 3.4b and 3.4e), (2) landmark distance in mm (Fig. 3.4d), and (3) singularity errors—the fraction of non-diffeomorphic voxels (Fig. 3.4c). Fig. 3.4 highlights the impact of representation choice in modeling diffeomorphisms. DARTEL, using an exponential map, performs significantly worse than ANTs across all metrics by three orders of magnitude. In contrast, our method reduces fissure alignment error by 5× compared to ANTs and outperforms it in 5 out of 6 landmark subregions. While all methods theoretically ensure diffeomorphism, SVF-based approaches introduce singularity errors due to non-adaptive scaling-and-squaring. We discuss the numerical limitations of SVF-based approaches in Section 3.3.2. ANTs also introduces some singularities, whereas our method computes numerically perfect diffeomorphic transforms. Finally, Fig. 3.4e compares fissure alignment errors among EMPIRE10 submissions, showing FireANTs achieves the lowest landmark errors and the fastest runtime among the top 10 methods, setting new benchmarks in computational efficiency and accuracy. **NLST**: For the NLST dataset(Team, 2011), we compare with representative state-of-the-art optimization and deep-learning baselines. We use the evaluation criteria provided by the challenge, and measure results on the Robust Target Registration Error (TRE30) in millimeters between the registered keypoints. Results in Fig. 3.4f show that FireANTs outperforms all baselines on the NLST dataset, with improvements of upto 51.6% in robust target registration error (TRE30) of provided keypoints compared to state-of-the-art deep learning benchmarks including Im2Grid, Vector-Field Attention, RWC-Net, and a 50.8% improvement in TRE30 over foundation models like unigradICON. This demonstrates the broad applicability of FireANTs beyond neuroimaging applications.

3.6.4. Evaluation on high-resolution mouse isocortex registration

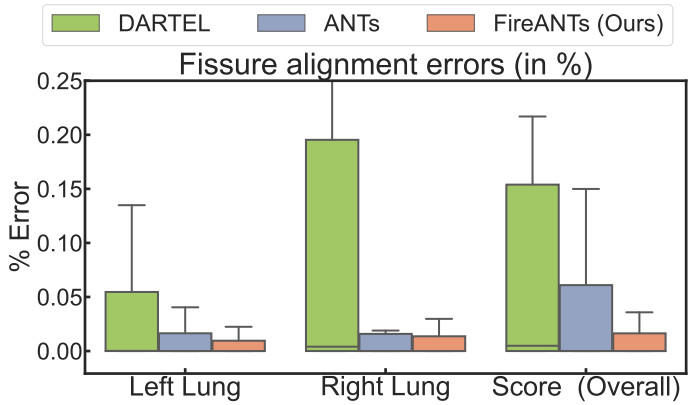
Expansion Microscopy (ExM) is an emerging super-resolution fluorescence imaging technique that enables 3D nanoscale visualization of cellular and molecular structures (Chen et al., 2015). While ExM provides rich structural data, its large-scale images remain challenging for existing registration algorithms due to repetitive textures, highly non-linear hydrogel deformations, imaging noise, and size constraints. The Robust Non-rigid Registration Challenge for Expansion Microscopy (RnR-ExM) (rnr) offers a benchmark dataset, where we focus on registering mouse isocortex images, characterized by hydrogel-induced deformations and staining intensity loss. Each volume ($2048 \times 2048 \times 81$ voxels) has a voxel spacing of $0.1625\mu m \times 0.1625\mu m \times 0.4\mu m$ and is 40.5 times larger than brain imaging datasets. Current state-of-the-art methods either register small independent chunks (Fleishman, 2023), losing inter-chunk information, or process highly downsampled images (Jia et al., 2022), significantly reducing resolution (by 64× in-plane).

In contrast, FireANTs is able to register the volume at native resolution. We perform an affine registration followed by a diffeomorphic registration step. The entire method takes about 2-3 minutes on a single A6000 GPU. As shown in Fig. 3.5, our method secures the first place on the leaderboard, with a considerable improvement in the Dice score and a 4.42× reduction in the standard deviation of the Dice scores compared to the next best method. Fig. 3.5 also shows qualitative comparison of our method compared to Bigstream (Fleishman, 2023), the winner of the RnR-ExM challenge. Bigstream performs only an affine registration, leading to inaccurate registration in one of three test volumes, leading to a lower average Dice score and higher variance. Moreover, the affine registration leads to boundary in-plane slices being knocked out of the volume, leading

(a) Following the evaluation setup of Klein *et al.* paper, we validate registration performance using the average volume overlap of all anatomical label maps between the fixed and warped label maps. We consider ANTs (the winner of the challenge), and Diffeomorphic Demons as state-of-the-art optimization algorithms, and Voxelmorph and Synthmorph as state-of-the-art unsupervised deep learning baselines. Evaluation is shown for five metrics with ↑ denoting a higher score is better, and ↓ signifying a lower score is better. For deep learning baselines, appropriate preprocessing (intensity normalization, alignment, and resampling to 1mm isotropic) is performed to ensure a fair comparison, whereas no such preprocessing is required for optimization methods, including FireANTs. FireANTs shows significant gains in performance that are consistent across all four datasets, with the median overlap scores outperforming the third quartile of all other methods for IBSR18 and CUMC12 datasets. Comparison of overlap metrics by specific anatomical regions are in Fig. B.8. For the overlap aggregation mentioned in Klein et al. (2009), results are shown in Fig. B.7.

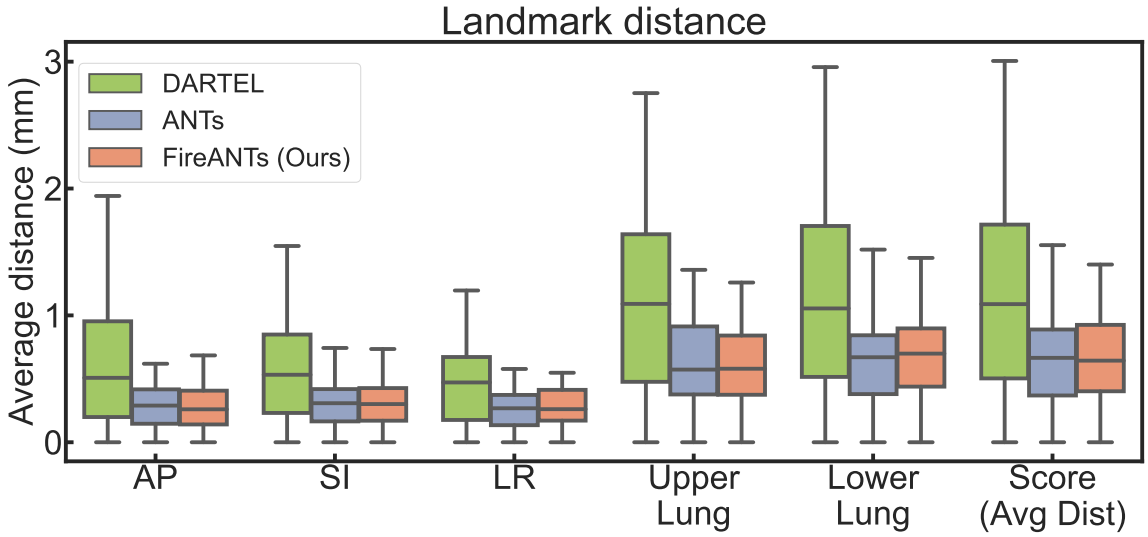


(b) FireANTs achieves substantially lower inter-quartile range of fissure errors, defined as the percentage of marked pixels that are registered to points on the opposite side of the fissure boundary.



% Error	DARTEL	ANTs	Ours
Left Lung	3.9983	0.0069	**0.0000**
Lower Lung	2.7514	0.0177	**0.0000**
Right Lung	2.4930	0.0107	**0.0000**
Upper Lung	5.2037	**0.0000**	**0.0000**
Score (Overall)	3.0681	0.0088	**0.0000**

(c) Singularity errors are defined as fraction of voxels that define a non-invertible deformation. Singularity quantifies the percentage of implausible deformations. FireANTs achieves *zero percent* singularity errors.



(d) Landmark distance is the Euclidean distance between well-dispersed landmark points between the fixed and warped images. FireANTs has a lower median and narrower interquartile range than baselines on five out of six subregions.

Method	Left Lung	Right Lung	Score (% Error Overall)
FireANTs (Ours)	0.0185	0.0254	0.0227
MRF Correspondence Fields	0.0824	0.0211	0.0485
ANTs	0.0249	0.1016	0.0747
Dense Displacement Sampling	0.0578	0.0919	0.0826
ANTs + BSpline	0.0821	0.0848	0.0861
DISCO	0.1256	0.0499	0.0882
VIRNet	0.0834	0.0934	0.0890
Feature-constrained nonlinear registration	0.1210	0.0758	0.1032
Explicit Boundary Alignment	0.1063	0.1246	0.1209
MetaReg	0.1049	0.2224	0.1791

(e) Fissure alignment error on top 10 algorithms in the challenge sorted by fissure alignment error, averaged on all scan pairs. FireANTs outperforms a wide array of baselines, including direct optimization (ANTs, ANTs+BSpline), neural networks (VIRNet), and explicit correlation volumes (MRF, Disco).

Validation metrics on NLST	
Method	**TRE30 (in mm)**
Zero displacement (Baseline)	9.76
VoxelMorph Balakrishnan et al. (2019)	4.12
Im2Grid Liu et al. (2022)	3.05
ANTs	3.04
Vector-Field Attention Liu et al. (2024c)	2.31
RWC-Net Sivan et al. (2023)	2.11
unigradICON Tian et al. (2024)	2.07
unigradICON + instance optimization	1.77
FireANTs (Ours)	**1.18**

(f) Robust landmark distance (TRE30) comparison of state-of-the-art algorithms highlights the effective performance of FireANTs on the NLST dataset, outperforming a plethora of state-of-the-art optimization and deep learning baselines.

Figure 3.4: FireANTs demonstrates state-of-the-art performance on community-standard neuroimaging and pulmonary challenges: **(a) EMPIRE10**: Lung fissure plates are an important anatomical landmark demarcating lobes within the lung. Fissure alignment errors (in %) denote the percentage of locations near the lung fissure plates that are on the wrong side of the fissure post-registration. Over all 30 scan pairs, our method performs 5× better than ANTs. **(b) EMPIRE10**: Singularity errors defined as percentage of voxels that have a non-diffeomorphic deformation, a proxy for physically implausible deformations. In the DARTEL baseline, singularities can be introduced for larger deformations due to numerical approximations of the integration. Even for ANTs, the solutions (deformations) returned are not entirely diffeomorphic. Our method shows much better fissure and landmark alignment (Fig. 3.4**(a,c)**, Fig. B.9, Fig. B.10) with guaranteed diffeomorphic transforms. **(c) EMPIRE10**: Landmark distance in mm for selected landmarks. Across different lung subregions, our method shows results at least at par with ANTs, with slightly better average and median results across all regions. **(d) EMPIRE10**: Shows the top 10 algorithms for average fissure alignment error in % in the EMPIRE10 challenge. Error metrics are taken from the evaluation server. Other methods perform well on one lung (MRF for right, ANTs for left) but comparatively poorly on the other lung, compared to our method showing both accurate and robustness to both the left and right lung. = First, = Second, = Third best result. **(e) NLST**: Landmark distance in mm for provided landmarks. Our method outperforms a variety of state-of-the-art optimization and deep learning algorithms.

to poor registration (Fig. 3.5). FireANTs preserves the boundary in-plane slices during its affine step, and subsequently performs an accurate diffeomorphic registration at submicron resolution leading to accurate registration with substantially lower variance. This experiment demonstrates the versatility and applicability of FireANTs for high-resolution microscopy registration.

3.6.5. Runtime and Memory Efficiency Analysis

One of the critical bottlenecks for scalable registration with ANTs is the prohibitively large runtimes for single-threaded CPU registration (Balakrishnan et al., 2019). Deep learning methods aim to reduce the runtime by performing feedforward inference, but these methods in turn have steep memory requirements due to activation overheads (Nouamane et al., 2025; Korthikanti et al., 2023), making them infeasible for high-resolution registration. FireANTs circumvents both these issues using a lightweight implementation on GPU.

Runtime compared to ANTs We evaluate the efficiency of FireANTs compared to ANTs by running both algorithms on the CPU with 32 threads, with identical multi-scale optimization settings. Furthermore, we run FireANTs on the GPU with a batch size of 1 to avoid amortizing runtimes over larger batches. The runtimes on all five brain datasets (IBSR18, CUMC12, MGH10, LPBA40, and OASIS) are shown in Fig. 3.6a. FireANTs is upto $7\times$ faster than ANTs on CPU, and $442\times$ faster on GPU. The runtime improvement on the CPU can be attributed to faster convergence and better implementation of the optimization since both methods are run with identical multi-scale optimization settings and capped at the same CPU resources. Fig. 3.6b shows the runtime of FireANTs compared to ANTs on the EMPIRE10 dataset. Since the runtime for ANTs and DARTEL are provided in the submission writeup without details on hardware or number of threads, it is not possible to reproduce the same results, and use their provided numbers as expected runtime for the dataset. However, FireANTs runs an average of $560\times$ faster than ANTs on the GPU, which is at par with the runtime improvements on neuroimaging datasets.

Runtime and memory requirements compared to deep learning methods FireANTs is highly efficient compared to deep learning methods, both in terms of runtime and memory usage. We highlight the efficiency using three experiments:

- **Accuracy-Runtime tradeoff**: We compare the performance, runtime, and memory usage of FireANTs with SOTA deep learning methods averaged over three neuroimaging datasets (LPBA40, Ultracortex, PRIME-DE).
- **Runtime and Memory requirements with increasing problem sizes**: We take an OASIS MRI image pair and progressively upsample it by factors of $2\times, 4\times, 6\times, 8\times$ larger and measure the runtime and memory usage of each method.
- **Amortized runtime with increasing batch sizes**: Amortized runtime over larger batches can improve GPU utilization and hide kernel launch and activation cache loading overheads. We perform batched inference using an OASIS MRI volume for increasing batch sizes, and plot the amortized runtime (i.e. runtime divided by batch size) as a function of the batch size, and keep increasing the batch size until we run out of memory.

For all methods, we measure the runtime only for the registration call function, not including image loading, preprocessing and postprocessing steps, model loading, etc. We note that if these steps were to be included in

(a) Snapshot of the RnR-ExM leaderboard

#	User (Team)	Created	DSC
1st	rohit.rango	11 Aug. 2023	0.92049045 ± 0.00996840
2nd	rohit.rango	17 Aug. 2023	0.91875541 ± 0.01803930
3rd	cwmokab (Orange)	15 March 2023	0.91688563 ± 0.04410269
4th	cwmokab (Orange)	12 March 2023	0.91544257 ± 0.04463970
5th	NLi10Me (bigstream)	14 March 2023	0.91426871 ± 0.03391914
6th	acasamitjana	23 June 2023	0.91321484 ± 0.02358535
7th	NLi10Me (bigstream)	14 March 2023	0.91209382 ± 0.03194342
8th	cwmokab (Orange)	15 March 2023	0.91111042 ± 0.04326616
9th	NLi10Me (bigstream)	14 March 2023	0.90968117 ± 0.02988877
10th	xi	15 March 2023	0.90895331 ± 0.03555638

(b) Qualitative comparison of registration of Bigstream and FireANTs

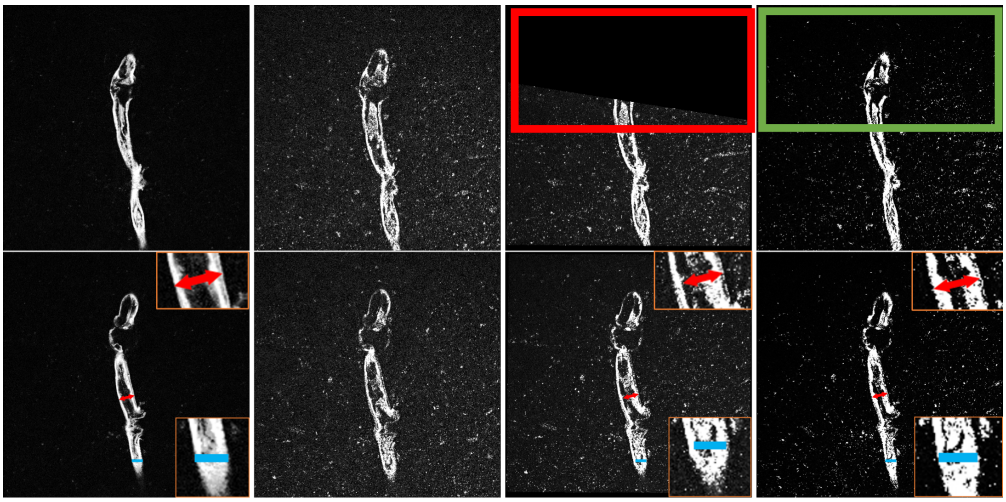


Figure 3.5: FireANTs secures first rank in the RnR-ExM mouse dataset: **(a)**: As of March 1, 2025, our method ranks first in the mouse brain registration task, which is the only task in the challenge requiring deformable registration. Our top two successful submissions secure the first and second position, with a 0.361 improvement in Dice score compared to the 3rd ranked submission, which is 0.261 better than the 5th ranked submission (bigstream). Note that among the top 10 submissions, our method has the lowest standard deviation (4.42× lower than the second best submission) showing the robustness of our model across different microscopy volumes. **(b)** shows a qualitative comparison of FireANTs with Bigstream (Fleishman, 2023), the other top leading method in the challenge. The moving image volumes have substantially more noise than the fixed image volumes, making intensity-based registration difficult. The non-rigid deformation dynamics of the hydrogel are clearly visible, as the moving volume has a thicker boundary than the fixed volume. Bigstream does not capture these dynamics very well – this is illustrated by comparing the thickness of the cortex at various points (zoomed orange crops in bottom row), where Bigstream does not deform the cortex enough to match the fixed image. FireANTs deforms and accurately depicts these morphological changes, which can be crucial for downstream morphometric studies. Moreover, the affine registration in Bigstream knocks the boundary slices out of the volume (red highlight in top row), leading to drop in registration performance. On contrary, our method's affine and deformable stages are more stable, leading to better registration and avoiding spurious out-of-bound artifacts at the boundary slices.

the runtime comparison, the efficiency gains of FireANTs would be even more significant.

These results in Fig. 3.6c show that FireANTs is upto $10\times$ more memory efficient than SOTA deep learning methods, while performing faster than most of them at inference. This is a result that challenges the common belief that deep learning methods are faster than iterative optimization methods at inference, similar to results shown in Hering *et al.* (Hering et al., 2022). Deep learning feedforward inference can be slow due to convolutions of large activations, skip connections requiring repeated memory accesses. In contrast, the iterative optimization updates in FireANTs are lightweight and can be run for more iterations. Since FireANTs does not generate any feature activations beyond that in the loss function, it is very memory efficient. Over three brain datasets, FireANTs achieves the best accuracy-runtime performance, showing that a tradeoff is not necessary for good performance. Amortized inference can further improve efficiency - on the OASIS dataset, FireANTs can register a batch of 32 image pairs in less than 0.25 seconds per pair, . This unprecedented efficiency paves the way for rapid prototyping, hyperparameter tuning, and scaling to high resolution datasets. This sets a new standard for high-throughput image registration on GPUs.

FireANTs enables rapid prototyping and hyperparameter tuning In optimization toolkits such as ANTs, correct choice of hyperparameters are key to high quality registration. Some of these hyperparameters are the window size for the similarity metric Cross-Correlation or bin size for Mutual Information. In our experience, the Gaussian smoothing kernel $\sigma_{\text{grad}}, \sigma_{\text{warp}}$ for the gradient and the warp field are two of the most important parameters for accurate diffeomorphic registration. The optimal values of these hyperparameters vary by image modality, intensity profile, noise and resolution. Typically, these values are provided by some combination of expertise of domain experts and trial-and-error. However, non experts may not be able to adopt these parameters in different domains or novel acquisition settings. Recently, techniques such as hyperparameter tuning have become popular, especially in deep learning (Hoopes et al., 2021).

In the case of registration, hyperparameter search can be performed by considering some form of label/landmark overlap measure between images in a validation set. We demonstrate the stability and runtime efficiency of our method using two experiments : (1) Owing to the fast runtimes of our implementation, we show that hyperparameter tuning is now feasible for different datasets. The optimal set of hyperparameters is dependent on the dataset and image statistics, as shown in the LPBA40 and EMPIRE10 datasets; (2) within a particular dataset, the sensitivity of our method around the optimal hyperparameters is very low, demonstrating the robustness and reliability of our method. We choose the LPBA40 dataset among the 4 brain datasets due to its larger size ($40\times39 = 1560$ pairs). We choose three parameters to search over : the learning rate (η), and the gaussian smoothing parameters $\sigma_{\text{warp}}, \sigma_{\text{grad}}$. We use the Ray library (`https://docs.ray.io/`) to perform a hyperparameter tuning using grid search. For the LPBA40 dataset, a grid search over three parameters (shown in Fig. 3.7a) takes about 40.4 hours with 8 parallel jobs. On the contrary, ANTs would require around 3.6 years to complete the same grid search, with 8 threads allocated to each job and 8 parallel jobs. This makes hyperparameter search for an unknown modality computationally tractable. A deep learning solution like HyperMorph (Hoopes et al., 2021) can perform amortized training over a predefined hyperparameter space, but still requires significant GPU hours for training and inference of 1560 pairs for each configuration to generate a plot like Fig. 3.7a. Fig. 3.7b shows the runtime and memory usage of FireANTs, ANTs, and HyperMorph on the LPBA40 dataset, showing that even a brute force grid search with FireANTs is about $4\times$ faster than state-of-the-art amortized hyperparameter learning.

FireANTs is robust to a wide range of hyperparameters Fig. 3.7a shows a dense red region suggesting the final target overlap is not sensitive to the choice of hyperparameters. Specifically, the maximum target overlap is 0.7565 and 58.4% of these configurations have an average target overlap of ≥ 0.74. This is demonstrated in

(a) Runtime analysis and comparison with ANTs on CPU and GPU on five brain datasets

Dataset	ANTs	Ours (asymmetric)			Ours (symmetric)		
		CPU (s)	GPU (s)	Speedup	CPU (s)	GPU (s)	Speedup
IBSR18	324.39 ± 16.67	119.65 ± 9.31	1.06 ± 0.04	$2.71/305.76$	399.64 ± 2.58	2.04 ± 0.02	$0.81/159.30$
CUMC12	484.29 ± 64.14	83.85 ± 4.08	1.09 ± 0.07	$5.78/442.43$	176.64 ± 7.06	2.09 ± 0.03	$2.74/231.29$
MGH10	329.95 ± 119.37	46.75 ± 0.84	1.13 ± 0.06	$7.06/291.84$	98.17 ± 1.76	1.19 ± 0.44	$3.36/278.03$
LPBA40	227.30 ± 37.37	153.76 ± 10.04	1.22 ± 0.06	$1.48/186.61$	272.98 ± 20.33	2.30 ± 0.12	$0.83/98.74$
OASIS	131.06 ± 9.01	54.47 ± 1.44	0.75 ± 0.01	$2.41/174.75$	94.92 ± 1.018	1.62 ± 0.11	$1.38/80.90$

(b) Runtime analysis and summary on EMPIRE10 dataset. ANTs and DARTEL take significantly longer to run due to the high-resolution nature of the CT lung dataset. FireANTs runs a *minimum* of 320 times faster than ANTs, saving a substantial amount of time, at no loss (in fact, with substantial gains) in registration quality.

	ANTs	DARTEL	Ours	Speedup (ANTs)	Speedup (DARTEL)
Avg	6hr 14m	7hr 16m	0m 39s	562.67	663.77
Min	0h 55m	1h 8m	0m 9s	320.74	315.23
Max	12h 41m	10h 11m	1m 5s	1231.27	796.51

(c) Tradeoff between accuracy, runtime, and memory requirements with state-of-the-art deep learning methods.

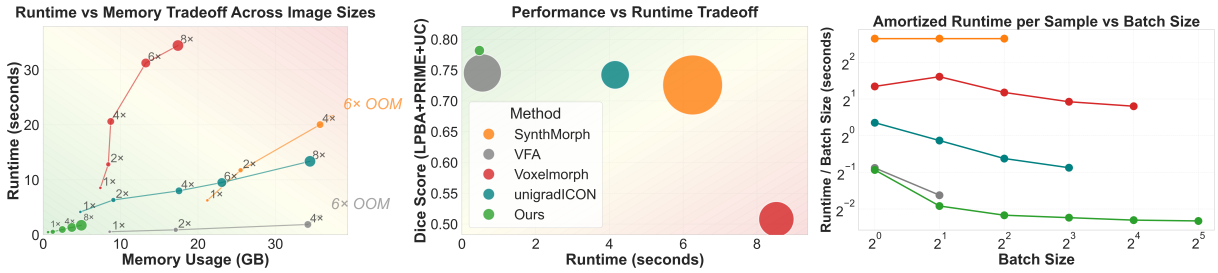
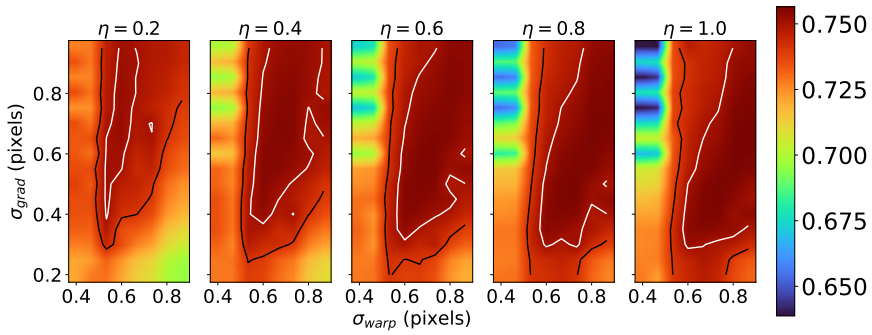


Figure 3.6: FireANTs facilitates quick and scalable registrations. We compare the runtime of our implementation with the ANTs library. **(a)** shows histogram of speedup (runtime of ANTs divided by runtime of our method) and statistics of runtimes (in seconds) for the four brain MRI datasets. For all datasets, our implementation runs a *minimum* of two orders of magnitudes faster, making it suitable for hyperparameter search algorithms, and larger datasets. Table **(b)** shows the runtime of ANTs, DARTEL and our implementation on the EMPIRE10 challenge data. The first three colums show the actual runtime of the methods, followed by the speedup obtained by our method when compared to ANTs and DARTEL. Note that our method runs a *minimum* of 320 times faster than ANTs, saving a substantial amount of time, at no loss in registration quality. **(c)** shows the runtime and memory requirements of our method compared to deep learning methods. **Left** shows the runtime and memory requirements of our method compared to deep learning methods for increasing problem sizes. FireANTs is upto $10\times$ more memory efficient than SOTA deep learning methods, while performing faster than almost all of them at inference. **Middle** shows the plot of average performance over three brain datasets compared with average runtime, with the size of the bubble indicating average memory usage. FireANTs performs *better* while being faster and more memory efficient than all deep learning methods, indicating that a tradeoff is not necessary for good performance. **Right** shows that further gains in amortized runtime are possible by increasing the batch size at inference. FireANTs achieves less than 0.25 seconds per image pair and runs more than double the number of image pairs copmared to all other deep methods, showing unprecedented efficiency for high-throughput registration.

Fig. 3.7a by the white contour line denoting the level set for target overlap = 0.75, and the black contour line denoting the level set for target overlap of 0.74. The target overlap is quite insensitive to the learning rate (≥ 0.4) showing that our algorithm achieves fast convergence with a smaller learning rate. On the EMPIRE10

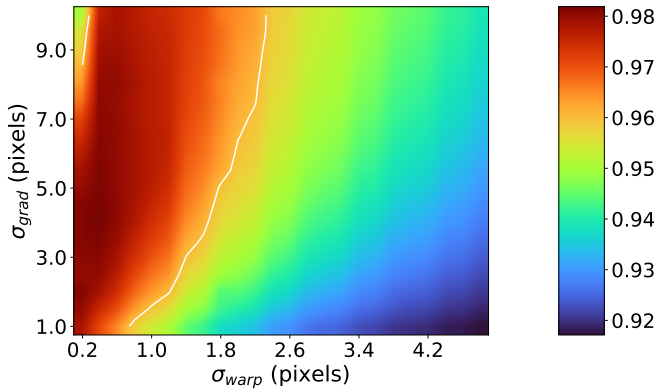
(a) Grid search on LPBA40 dataset



(b) Hyperparameter search efficiency on LPBA40 dataset

Method	GPU Hours
ANTs	7884.00
HyperMorph(Hoopes et al., 2021)	704.67
FireANTs	161.60

(c) Grid search on the EMPIRE10 dataset, with Dice score



(d) Ablation on using Jacobian-free optimization on OASIS, NLST, and AbdomenMRCT-L2R datasets, spanning three organ types and modalities.

OASIS	**Dice Score ↑**	**Runtime (s) ↓**
no Jacobian	0.783 ± 0.032	0.77 ± 0.06
with Jacobian	0.784 ± 0.032	1.44 ± 0.04
NLST	**TRE (mm) ↓**	**Runtime (s) ↓**
no Jacobian	1.181 ± 0.220	1.80 ± 0.12
with Jacobian	1.160 ± 0.206	4.95 ± 0.49
AbdomenMRCT-L2R	**Dice Score ↑**	**Runtime (s) ↓**
no Jacobian	0.762 ± 0.196	3.03 ± 0.08
with Jacobian	0.764 ± 0.195	3.87 ± 0.13

Figure 3.7: FireANTs facilitates feasibilility of extensive hyperparameter search in registration The speed of FireANTs makes hyperparameter studies like these feasible, which ANTs would take years to complete. **(a)**: We perform a hyperparameter grid search on three hyperparameters of interest - smoothing kernel for the warp field (σ_{warp}) in pixels, smoothing kernel for the gradient of warp field (σ_{grad}) in pixels and learning rate η. The metric to optimize in this case is the target overlap. For the LPBA40 dataset, we perform a hyperparameter sweep over 640 configurations in 40 hours with 8 A6000 GPUs. A corresponding hyperparameter sweep with 8 concurrent jobs with each job consuming 8 CPUs would take $\sim$3.6 years to complete. The white contour representing the level set for target overlap = 0.75, and the black contour representing the level set for target overlap of 0.74 show the robustness of our method to hyperparameters - performance is not brittle or sensitive to choice of hyperparameters. **(b)**: Hyperparameter grid search on the EMPIRE10 dataset over σ_{warp} and σ_{grad} parameters (456 configurations), with a fixed learning rate of $\eta = 0.25$. The metric to optimize is the Dice score of the provided binary lung mask. This sweep takes about 12.37 hours on 8 GPUs, whereas a corresponding sweep would take 296 days for ANTs and 345 days for DARTEL (more in Fig. 3.6). The white contour corresponds to the level set for Dice score = 0.96, showing both a huge spectrum of parameters that achieve high Dice scores, and low sensitivity of the method to choice of hyperparameters.

dataset, we fix the learning rate and perform a similar hyperparameter search over two parameters, the Gaussian smoothing parameters $\sigma_{\text{warp}}, \sigma_{\text{grad}}$, shown in Fig. 3.7c. We use the average Dice score between the fixed and moving lung mask to choose the optimal hyperparameters. FireANTs can perform a full grid search over 456 configurations on the EMPIRE10 dataset in 12.37 hours with 8 A6000 GPUs, while it takes SyN 10.031 days to run over a single configuration. Normalizing for 8 concurrent jobs and 456 configurations, it would take ANTs about 296 days, and DARTEL about 345 days. This shows that our method and accompanying implementation can now make hyperparameter search for high-resolution 3D image registration studies feasible.

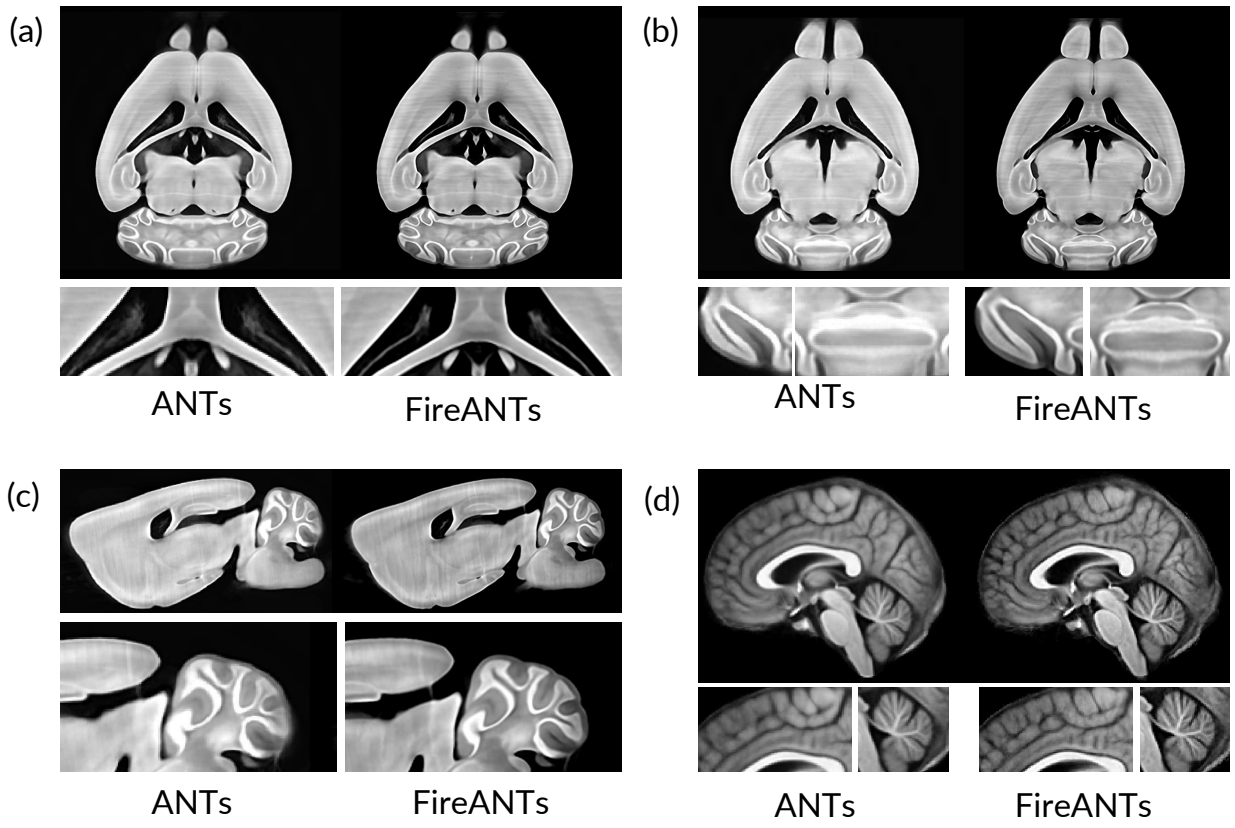


Figure 3.8: Comparison of brain templates (atlases) constructed using ANTs (left) and FireANTs (right). **(a–c)** Coronal and sagittal sections of the $25\mu m$ fMOST mouse brain template illustrate the improved structural fidelity of FireANTs. In the ANTs template, the internal regions of the lateral ventricles appear blurred (a), and the cerebellar architecture exhibits intensity bleeding (b, c), whereas FireANTs yields crisper delineation of these anatomical structures. **(d)** The in vivo human brain atlas further demonstrates the advantages of FireANTs, with sharper cortical folding and improved contrast and realistic intensity features in the cerebellum compared to ANTs. FireANTs generates multiple high-fidelity templates while being 200-400 times faster than ANTs.

Runtime efficiency due to Jacobian-free optimization A key methodological approach proposed in this work is the Jacobian-free approximation of the gradient field for faster diffeomorphic optimization (Section 3.4). A concern that may arise is the effect of this approximation on the accuracy of the registration, specifically due to the assumption that the Jacobian is positive-definite. We ablate on the effect of this approximation on the accuracy of the registration on three datasets - OASIS, NLST, and AbdomenMRCT-L2R, spanning three organ types and modalities. The heterogenity and variable biomechanics and deformation dynamics across multiple

organ systems is a challenging testbed for measuring the effect of Jacobian-free optimization on registration accuracy. Fig. 3.7d shows that the Jacobian-free approximation does not significantly affect the accuracy of the registration, with a maximum Dice score difference of 0.002 and a maximum TRE difference of 0.021 mm. However, avoiding computation of the Jacobian-augmented descent direction significantly reduces the runtime of the registration, making convergence upto 2.75× faster. Turning on the Jacobian-augmented descent direction is easy in our implementation by only changing a flag during runtime, but we recommend turning it off for most applications. In all our experiments except this ablation study, we do not compute the Jacobian-augmented descent direction.

3.6.6. FireANTs facilitates scalable atlas generation

Atlas generation is an important component of integrating large-scale imaging data – including gene expression, connectivity patterns, and functional properties -— onto a common spatial coordinate system facilitating multimodal data alignment and comparison. This requires atlas (or template) generation capabilities that scale with the unprecedented scale of acquired data. In this section, we showcase the efficiency of atlas generation by reproducing the fMOST atlas proposed in the ANTsX ecosystem (Tustison et al., 2024) for the mouse brain. We follow the steps outlined in the ANTsX Ecosystem (Tustison et al., 2024) for generating an fMOST atlas of the mouse brain, including preprocessing steps like downsampling to $25\mu m$ resolution, destriping, flipping along the sagittal plane for left-right symmetry, bias field correction, and affine preregistration to a common template. Since no parcellations are available for the dataset, we qualitatively compare the atlases generated by ANTs and FireANTs, and compare their runtimes. We also generate an in-vivo atlas for the OASIS dataset, to show scalability on smaller datasets and for quantitative evaluation. Fig. 3.8 shows that the atlas generated by both methods is similar in terms of quality. On a 64-thread core machine, ANTs takes *141.5 hours* to generate the atlas with 6 epochs of template refinement. With the identical number of iterations and configuration, FireANTs runs in *22 minutes* with a distributed setup on an 8-GPU workstation, showing a significant improvement in runtime efficiency. On a much lower-resolution OASIS dataset, ANTs takes *2 hours and 16 minutes* to generate an atlas with 16 subjects, while FireANTs runs in *32 seconds*. To quantify atlas fidelity, we evaluate Dice Score overlap of image pairs after registering them to the atlas. While pairwise Dice score overlap of subjects is 0.704 ± 0.163 with the ANTs template, the FireANTs template improves the Dice score to 0.722 ± 0.161. This demonstrates that FireANTs can be used to generate high-fidelity atlases two orders of magnitude faster than ANTs at no loss in image quality, making it a powerful tool for large-scale atlas generation.

3.6.7. Independent Evaluation

Since the release of our code and documentation, FireANTs has been independently adopted by researchers in the field. Few anectodal examples include the registration of high-resolution histology slides, and non-human primate data (GitHub, 2024a,b). A compelling independent application and evaluation is performed by NextBrain (Puonti et al., 2025), a tool that utilizes FireANTs to perform Bayesian segmentation of in-vivo and ex-vivo brain MRI scans. NextBrain uses FireANTs to register input brain MRI scans to an augmented template with a resolution of $200\mu m$ and 333 regions of interest, providing a comprehensive structural analysis of the subject. Quantitatively, incorporating FireANTs in the pipeline leads to no loss in performance as measured by Dice overlap. The utilization of GPU-based toolkits including FireANTs reduces the runtime from 2-3 days / one week for $1mm$ in-vivo / $300\mu m$ ex-vivo scans on a multi-core workstation, to *less than 5 minutes* on a GPU. Owing to the robustness and efficiency of FireANTs, it is set as the default registration method

in NextBrain. Another independent evaluation is performed on the registration of high-resolution X-ray images (Gopalakrishnan et al., 2025), where FireANTs establishes itself as a strong baseline, outperforming several baselines specialized for X-ray image registration. This demonstrates the accessibility, scalability and efficiency of FireANTs in real-world applications.

3.7. Discussion

In this chapter, we presented FireANTs, a powerful and general-purpose multi-scale registration algorithm. This chapter goes through a deep dive into the numerical challenges of diffemorphic registration, and the impact of the choice of representation of diffeomorphisms on the performance of the registration algorithm. Our novel Eulerian descent formulation enables powerful adaptive optimization on the space of diffeomorphisms directly, while maintaining high computational efficiency and low memory overhead for inference. This framework brings state-of-the-art performance on a wide range of datasets and modalities, while addressing the many limitations of deep learning-based registration methods discussed in Chapter 2.

CHAPTER 4

Deep Implicit Optimization enables Robust Learnable Features for Deformable Image Registration

In the previous chapter, we introduced FireANTs, a fast and efficient iterative solver for robust image registration. FireANTs provides real-time runtime on many clinical neuroimaging datasets, performs at-par or better than unsupervised deep learning counterparts in terms of labelmap overlap, optimizes directly over time-dependent flows of diffeomorphisms yielding higher flexibility of representation, and scales very gracefully to high-resolution volumes. This addresses a major limitation of iterative optimization methods - the slow runtime of the optimization process. FireANTs being a solver cannot learn from auxiliary labelmaps or landmarks directly. However, attempting to use a deep network to predict end-to-end transformation fields does not translate well to producing plausible or accurate deformation fields for out-of-distribution datasets. A framework that combines the best of both worlds is therefore highly desirable.

In this chapter, we tackle the second big limitation of iterative solvers, namely their inability to learn *task-specific* features from auxiliary labelmaps or landmarks. To this end, we introduce *Deep Implicit Optimization* (DIO) [1], a framework that transforms FireANTs into a fully differentiable layer within deep networks. This allows us to learn task-specific image features from the intensity images directly by backpropagating the supervised loss directly through the optimization solver. We show that this paradigm allows for a number of benefits, including significant out-of-distribution robustness across scanners and resolutions, the ability to switch between different transformation representations at test time, and the ability to use arbitrary regularization and hyperparameter search or tuning at test time.

[1] Source Code is available at https://github.com/rohitrango/DIO

4.1. The ingredients of extensible deep image registration

The success of image acquisition and processing technologies over the past decade has radically transformed various disciplines including in biomedical and biological sciences, including neuroimaging and neuropathology, pulmonary imaging, spatial transcriptomics and expansion microscopy. This growth has been complemented by the success of deep learning methods in their ability to learn from large amounts of data and perform complex tasks with high task-specificity. In the field of image registration, these advances call for an extensible class of methods that can be used to learn task-specific image features from the intensity images directly, while retaining a few key properties that we list below.

- **Controllable or steerable transformations:** The first key property is the ability for the user to be able to control or steer the transformation family, representation, and parameters at test time. This is to adapt to the downstream application and the data at hand, for example, a global alignment probably requires using an affine registration, but a non-linear registration might be required for a local alignment. For time-series registration, the user might prefer to use LDDMM (Beg et al., 2005), but for static image registration, the user might prefer to use a diffeomorphic registration like FireANTs (Jena et al., 2026).
- **Sublinear memory footprint for optimization iterations:** Harder examples might require more iterations to converge, and therefore a model whose memory footprint does not grow linearly (ideally $O(1)$) with the number of iterations is desirable.
- **Ability to integrate weak supervisory signals during training:** Large amounts of data available for training deep learning methods are annotated by humans or algorithms to serve as silver standard for downstream tasks like biomarker tracking, morphometric analysis, and segmentation. These landmarks or ROIs can be used to guide the optimization process to perform better registration, with the understanding that they may not be available for a test datum.
- **Robustness to domain shift:** While we desire the model to *internalize* task-aware registration, we do not want it to overfit to the training data and cause catastrophic failure on out-of-distribution data.
- **Flexibility in hyperparameter tuning:** Users might desire choosing hyperparameters to better fit their data from bespoke domains and modalities without expensive retraining of the model.

4.2. Implications of different designs for learnable image registration

Recall that Deformable Image Registration (DIR) is typically formulated as a variational optimization problem:

$$\varphi^* = \arg\min_{\varphi} L(I_f, I_m \circ \varphi) + R(\varphi) = \arg\min_{\varphi} C(\varphi, I_f, I_m) \tag{4.1}$$

We call this the *image matching* objective. If the images I_f and I_m are supplemented with anatomical label maps S_f and S_m, we call this the *label matching* objective. Classical methods perform image matching on the intensity images, but the label matching performance is bottlenecked by the fidelity of image gradients with respect to the label matching objective.

Deep learning methods mitigate this shortcoming by injecting label matching objectives (for example, Dice score or landmark distances) into the objective Eq. (4.1) and using a deep network with parameters θ to predict

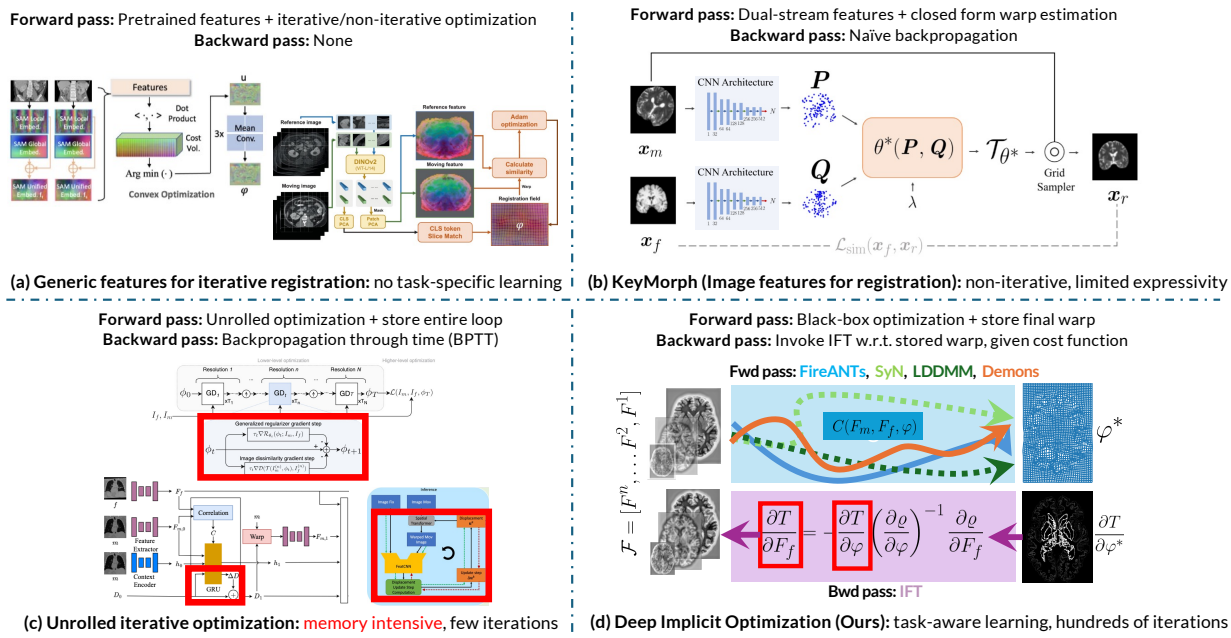


Figure 4.1: An illustrative comparison of existing methods and our method. **(a)** Generic features for image registration leverage the expressiveness and robustness of iterative optimization but do not incorporate task-specific learning, leading to suboptimal asymptotic performance on the in-distribution task. **(b)** Feature learning for closed-form parametric warp representations enable task-aware image features for registration, but are limited in expressiveness due to limited families of closed-form transforms and lack of error-correcting nature intrinsic to iterative optimization. **(c)** Unrolled iterative optimization using recurrent modules mimic the flavor of traditional optimization and enable task-aware image features. However, they are limited in expressivity because they can run only for a few number of iterations due to infeasible computational requirements. **(d)** DIO (our method) synergizes the expressivity of advanced iterative solvers and task-aware image feature learning by defining a custom backward pass that does not require unrolling or iteration. DIO provides the best of both worlds by inheriting the accuracy, expressivity, and robustness of iterative solvers, and asymptotic performance of learnable features.

φ for every image pair as input. In essence, learning-based problems solve the following objective:

$$\theta^* = \arg\min_{\theta} \sum_{f,m} L(I_f, I_m \circ \varphi_\theta) + D(S_f, S_m \circ \varphi_\theta) + R(\varphi_\theta) \tag{4.2}$$

$$= \arg\min_{\theta} \sum_{f,m} T(\varphi_\theta, I_f, I_m, S_f, S_m) \tag{4.3}$$

where $\varphi_\theta(I_f, I_m)$ is abbreviated to φ_θ. This leads to learned transformations φ_θ that perform both good image and label matching. However, the feature learning and optimization are coupled, and features are learned implicitly to produce deformation fields. Moreover, this formulation does not explicitly imbue any task-specific invariance into the learning framework, and the learned features are optimized only for a specific training domain, leading to poor generalization to domain shift.

Here, we discuss three dominant designs for learnable image registration, and how their key inductive biases play a role in the benefits and limitations of the methods.

Monolithic end-to-end transformation prediction This paradigm was dominant in the first generation of deep learning methods for image registration. The idea is to learn a deep network with parameters θ to predict the transformation φ for every image pair as input. This is typically done by concatenating the fixed and moving images and passing them to the network to estimate a parameteric warp representation. The progress in these family of methods was driven by "computationally advanced blocks" that were adapted for image registration (Balakrishnan et al., 2019; Jia et al., 2022; Mok and Chung, 2020c, 2021; Guo et al., 2024). Although this paradigm was very successful in learning from training data, these methods did not generalize well to out-of-distribution data, or even data from the same contrast (T1w MRI) but acquired from a different scanner or protocol. Moreover, these designs do not admit steerable transformations at inference, and do not allow arbitrary hyperparameter tuning at test time, although designs to perform "amortized optimization" over a pre-defined set of hyperparameters have been proposed (Hoopes et al., 2021; Mok and Chung, 2021). Other works like Jian et al. (2024); Liu et al. (2025a) and our own previous analysis from Chapter 2 show that these computational blocks provide no clear justification or motivation, and do not consider the characteristics of the registration task. Jian et al. (2024) instead advocate for registration-specific architectural designs that are inspired and battle-tested from iterative solvers in image registration. We discuss a few of these designs in the following sections. We also refer the reader to a visual summary of these designs provided in Fig. 4.1.

Generic features for iterative optimization One limitation of intensity images is that it doesn't encode anatomical details like shape, texture, discernable biomarkers directly. Therefore, a common practice is to learn features that encode higher-level representations of the images that are more task-specific. Typically, these features are trained using contrastive learning between synthesized views representing multimodal contrasts of the same anatomical configuration (Dey et al., 2025), using pretrained networks like DINOv2 followed by feature lifting and projection to a smaller subspace (Song et al., 2024), using self-supervised anatomical embeddings (Li et al., 2023b; Liu et al., 2021b), hand-engineered features (Heinrich et al., 2012), or using segmentation-network features (Siebert et al., 2024; Huang et al., 2024). Although these features are promising alternatives to intensity images, they are not inherently task-specific, and might not be able to capture the full range of anatomical variations present in the data. For example, a segmentation-based or contrastive learning-based network focusing on large objects might ignore subtle white-matter lesions or smaller subcortical structures that are washed out due to noise of partial volume effects.

Closed-form parametric warp representations with dual-stream feature extractor Recent works including Jian et al. (2024, 2025); Liu et al. (2025a) have shown that networks that utilize a dual-stream feature extractor tend to perform better than monolithic models. Dual stream feature extractors allow for decoupling the feature learning from the optimization, avoiding straightforward memorization of the training data, and allows for more flexible representations and multi-task learning (Kang et al., 2022; Honkamaa and Marttinen, 2023). However, decoder-based warp field prediction are prone to overfitting to the training data similar to their monolithic cousins, and are not flexible to different warp representations. A potential solution is to couple the feature learning from the dual-stream extractor with a closed-form parametric warp representation. A notable example is KeyMorph (Wang et al., 2023) where the feature extractor returns a set of K keypoints that represent semantically meaningful landmarks in the image, and have trivial correspondences indexed by their ordering. Linear fields like affine transformations can be obtained using least-squares fitting of the keypoints to the target keypoints, which admits a closed-form solution. However, this representation is mostly limited to linear transformations (thin-plate spline (TPS) transforms can be reduced to a system of linear equations in the parameters), and cannot represent more complex deformations like non-rigid deformations, large deformations, or diffeomorphic transformations.

Unrolled learnable optimization The limitations of the previous two paradigms motivates the use of an "optimization decoder". Unlike the dual-stream feature extractors with parametric decoder designs, the optimization decoder uses a "local search" strategy to iteratively obtain a warp field. This design is very popular in the optical flow literature (Teed and Deng, 2020; Jia et al., 2021; Lipson et al., 2021; Sivan et al., 2023). The decoder utilizes a learned "correlation volume" design that explicitly weighs features by their similarity to update the warp field, providing a way to perform iterative optimization on a feature space. However, these designs are limited to 2D images due to the $O(TN^4)$ complexity of the backpropagation through time (BPTT) and correlation volume computation. For 3D images, the complexity is $O(TN^6)$, which is infeasible for large images. This limits their expressivity to a few iterations, and learned correlation volume might learn associative patterns limited to training data. The primary limitation of these designs is therefore the high computational cost, but it provides a highly flexible representation.

These designs motivate us to combine the flexibility of the dual-stream feature extractor with a "tried and tested" optimization solver (i.e. FireANTs). To avoid the $O(T)$ complexity of the optimization, we propose using the *implicit function theorem* to backpropagate gradients through the optimization solver directly using a custom backprop rule. The following section discusses the technical details of our method.

4.3. Our Method

In this section, we introduce the methodological details of our framework. Our framework consists of three main components: (1) a dual-stream feature extractor network that processes fixed and moving images independently to produce dense multi-scale feature maps, (2) an implicit differentiation approach that enables backpropagation through arbitrary iterative optimization solvers without unrolling or storing intermediate optimization steps, and (3) a multi-scale optimization strategy that leverages pyramidal features from the network to improve convergence. We detail how implicit differentiation allows us to compute gradients through the optimization solver efficiently by exploiting the block-diagonal structure of the Hessian for mean squared error objectives, enabling efficient computation of feature gradients via vector-Jacobian products. We also describe implementation details including Jacobian-free backpropagation for stable training and handling double-backward passes through the grid sampling operator.

4.3.1. Dual-stream Feature Extractor Network

The first component of our framework is a dual stream feature network that extracts dense features from the intensity images. This network is parameterized by θ, and takes an image $I \in \mathbb{R}^{H\times W\times D\times C_{in}}$ as input and outputs a feature map $F \in \mathbb{R}^{H\times W\times D\times C}$, where C is the number of feature channels, i.e. $F = g_\theta(I)$. Unlike parameteric DLIR methods where moving and fixed images are concatenated and passed to the network to estimate a parameteric warp representation, our feature network processes the images *independently*. This allows the fixed and moving images to be of different voxel sizes. The feature network can also output multi-scale feature maps $\mathcal{F} = g_\theta(I) = [F^0, F^1, \dots, F^N]$, where $F^k \in \mathbb{R}^{H/2^k\times W/2^k\times D/2^k\times C_k}$, which can be used by multi-scale optimization solvers. The overall framework does not dictate a particular choice of architecture, and we ablate on different popular architectures in the experiments.

Note that in contrast to Eq. (4.1) that has fixed dynamics because of fixed I_f and I_m, the learned features

induce a modified learned optimization described as follows:

$$\arg\min_{\varphi} C(\varphi, F_f, F_m) \tag{4.4}$$

Since F_f and F_m are learned using a deep network, we can now *explicitly* imbue the task-specific inductive biases into arbitrary learned features.

In this work, we focus on the iterative refinement stage for end-to-end optimization of the learned image features from a task-specific training dataset.

4.4. Implicit Differentiation through Optimization

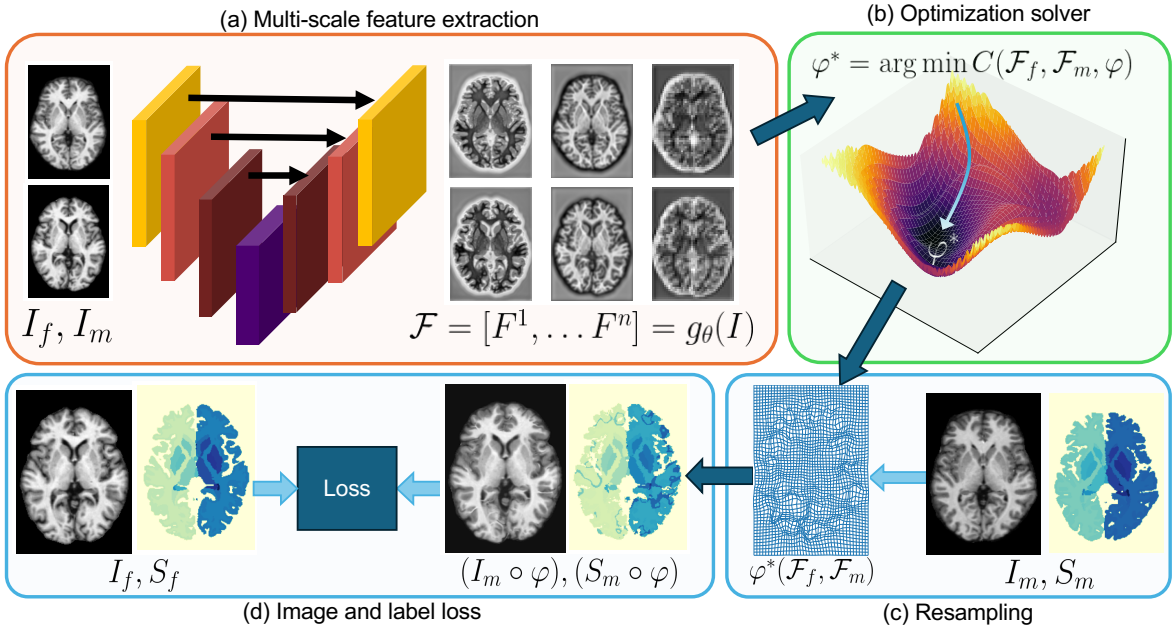


Figure 4.2: Overview of our framework. **(a)** A neural network extracts *dense* multi-scale features from the input images. **(b)** These features are used to optimize warp fields using a multi-scale differentiable optimization solver. **(c)** The optimized transform is used to warp the moving image and labels. **(d)** The warped image/label are compared with the fixed image/label using a similarity metric.

Due to the inherent limitations of parameteric warp field representation, prior works like (Qiu et al., 2022; Sivan et al., 2023; Blendowski et al., 2021) have proposed to use recurrent architectural designs to mimic the flavor of traiditonal iterative optimization algorithms. The iterative optimization is designed to minimize the image and label matching dissimilarity from the training data. However, these methods are still different from traditional optimization in a few fundamental ways. First, traditional optimization algorithms have a well-defined stopping criteria (i.e. convergence to a local minima). This usually requires hundreds of iterative optimization steps over multiple resolutions. In contrast, existing works employing recurrent architectures have a few number of fixed iterations at each step. Second, an instance optimization solver does not require the entire optimization path, in contrast to recurrent architectures where the entire unrolled path must be stored to perform backpropagation-through-time (BPTT). These reasons limit the expressive capacity of iterative optimization while allowing backpropagation through the solver.

We propose to close the gap using an implicit differentiation approach to leverage powerful image registration solver toolkits. Specifically, given the feature maps F_f and F_m extracted from the fixed and moving images using a neural network, a gradient-based iterative solver optimizes Eq. (4.4) to obtain the optimal transformation φ^*. The minimization objective converges when the gradient of the dissimilarity is zero:

$$\varrho(\varphi^*, F_f, F_m) = \nabla_\varphi C = \frac{\partial C}{\partial \varphi}\bigg|_{\varphi^*} = 0 \tag{4.5}$$

At this point, subsequent iterations of the optimization do not change the value of φ^*. Therefore, φ^* can be thought of as the fixed point of an 'infinite-layer' iterative optimization solver. This value of φ^* is then used to compute the loss Eq. (4.2) to minimize image and label matching objective.

Note that the analytical form of the vector-valued function ϱ is induced by the choice of scalar-valued loss function C used to run the optimization in Eq. (4.4). For example, choosing to minimize the sum of squared distance loss $C(F_f, F_m \circ \varphi) = \|F_m \circ \varphi - F_f\|_2^2$ induces $\varrho(\varphi, F_f, F_m) = (F_m \circ \varphi - F_f)(\nabla F_m \circ \varphi)$. To propagate derivatives from φ^* to the feature images F_f, F_m, we invoke the Implicit Function Theorem (Krantz and Parks, 2002):

Theorem 2. For a function $\varrho : \mathbb{R}^n \times \mathbb{R}^{m_1+m_2} \to \mathbb{R}^n$ that is continuously differentiable, if $\varrho(\varphi^*, F_f, F_m) = 0$ and $\left|\frac{\partial \varrho}{\partial \varphi}\right|_{\varphi^*} \neq 0$, then there exist open sets U, V_f, V_m containing φ^*, F_f, F_m, and a function $\varphi^*(F_f, F_m)$ defined on these open sets such that $\varrho(\varphi^*(F_f, F_m), F_f, F_m) = 0$.

Given the Implicit Function Theorem (IFT), we write $\varrho(\varphi^*(F_f, F_m), F_f, F_m) = 0$ and differentiate with respect to F_f to obtain:

$$\frac{d\varrho}{dF_f} = \frac{\partial \varrho}{\partial \varphi}\frac{\partial \varphi}{\partial F_f} + \frac{\partial \varrho}{\partial F_f} = 0 \tag{4.6}$$

$$\Rightarrow \frac{\partial \varphi}{\partial F_f} = -\left(\frac{\partial \varrho}{\partial \varphi}\right)^{-1}\frac{\partial \varrho}{\partial F_f} \tag{4.7}$$

$$\Rightarrow \frac{\partial T}{\partial F_f} = \frac{\partial T}{\partial \varphi}\frac{\partial \varphi}{\partial F_f} \tag{4.8}$$

$$= -\frac{\partial T}{\partial \varphi}\left(\frac{\partial \varrho}{\partial \varphi}\right)^{-1}\frac{\partial \varrho}{\partial F_f} \tag{4.9}$$

The forward pass of the layer is simply the iterative solver run without any unrolling or storing of any intermediate steps. During the backward pass, Eq. (4.9) provides the analytical form for computing the derivative of φ with respect to the feature images. We explain how to use the result of Eq. (4.9) to compute the gradients of the network with respect to the training loss in Section 4.4.1.

This design allows maximal expressivity of the iterative optimization solver by allowing hundreds of iterations until convergence, while being agnostic to the nature of the solver. Moreover, there are no additional memory overheads for optimization. In contrast, explicit T-step unrolling in prior work requires an $O(T)$ memory overhead for BPTT, rendering it infeasible for 3D image registration.

To summarize, the implicit optimization's forward pass directly runs the optimization without additional

overhead. During the backward pass, the optimal φ^* is used to compute the gradients with respect to the feature images F_f, F_m. The gradients are subsequently passed back to the weights of the neural network.

4.4.1. Computing the Implicit Gradient

There are two parts to computing the feature gradients:

- Computing the modified gradient $v^T = \frac{\partial T}{\partial \varphi}\left(\frac{\partial \varrho}{\partial \varphi}\right)^{-1}$
- Computing the gradient w.r.t. feature image $v^T \frac{\partial \varrho}{\partial F_f}$

We describe how to compute these gradients in the following sections.

Computing the Inverse Jacobian

An important component of the implicit differentiation is the computation of the inverse Jacobian $\left(\frac{\partial \varrho}{\partial \varphi}\right)^{-1}$. (Bai et al., 2019) propose using a quasi-Newton approach to solve the linear system $\left(\frac{\partial \varrho}{\partial \varphi}\right) v = -\frac{\partial T}{\partial \varphi}$. This requires solving another iterative optimization in the backward pass, that can be slow. For general problems, there are typically no alternatives to performing iterative optimization, since the Jacobian does not have a reduced form.

However, we exploit a special structure of the Jacobian that allows us to compute the inverse Jacobian efficiently without any iterative methods. First, we note that since $\varrho = \frac{\partial C}{\partial \varphi}$, the Jacobian $\frac{\partial \varrho}{\partial \varphi}$ is the Hessian of the loss function $\nabla^2_\varphi C(\varphi)$. This quantity is a $(n_v \cdot d) \times (n_v \cdot d)$ matrix, where n_v is the number of voxels in φ, and d is the spatial dimension, In a typical 3D registration scenario, n_v is of the order of 10^7, making this quantity hard to compute in general. However, for the mean squared error, i.e. $C(\varphi, F_f, F_m) = \|F_f - F_m \circ \varphi\|_2^2$, the Hessian $\frac{\partial \varrho}{\partial \varphi}$ is a block-diagonal matrix, since there are no terms in C containing both $\varphi(x_p)$ and $\varphi(x_q)$ for voxel indices $p \neq q$. Specifically, we have

$$
\begin{aligned}
(\varrho)(\varphi(x_p)) &= \nabla_{\varphi(x_p)} C(\varphi, F_f, F_m) && (4.10)\\
&= (F_m(\varphi(x_p)) - F_f(x_p)) \nabla F_m(\varphi(x_p)) && (4.11)
\end{aligned}
$$

This quantity is a vector of size d due to the term $\nabla F_m(\varphi(x_p))$, and has no terms involving $\varphi(x_q)$ for $q \neq p$. We now consider the scalar

$$
g_i = \sum_p (\varrho)(\varphi(x_p))[i] \tag{4.12}
$$

, where $[i]$ is the i^{th} index of a vector. The gradient of g_i with respect to $\varphi(x_q)$ is therefore

$$
\begin{aligned}
\nabla_{\varphi(x_q)}(g_i) &= \nabla_{\varphi(x_q)} \sum_p \left(\varrho(\varphi(x_p))\right)[i] && (4.13)\\
&= \nabla_{\varphi(x_q)} \left(\varrho(\varphi(x_q))[i]\right) && (4.14)\\
&= \left(\nabla^2_{\varphi(x_q)} C(\varphi)\right)[i] && (4.15)
\end{aligned}
$$

which is the i^{th} row of the Hessian block corresponding to the voxel x_q.

All the aforementioned operations can be performed efficiently using automatic differentiation libraries. We compute the gradients of each g_i $(i = 1, 2 \dots d)$ and stack them to obtain the full blockwise Hessian of size $n_v \times d \times d$. Next, we can solve the following $d \times d$ system of equations for v_p for each voxel p independently:

$$\frac{\partial T}{\partial \varphi(x_p)} = [\nabla_{\varphi(x_p)}(g_1); \dots; \nabla_{\varphi(x_p)}(g_d)]v_p \tag{4.16}$$

Since d is 2 or 3, Eq. (4.16) can be solved efficiently using standard linear algebra methods. Eq. (4.16) allows us to compute the modified gradient $v^T = \frac{\partial T}{\partial \varphi}\left(\frac{\partial \varrho}{\partial \varphi}\right)^{-1}$.

Computing the Feature Gradients

Note that $\frac{\partial \varrho}{\partial F_f}$ is a matrix of size $n \times m_1$. Here, $n = n_v \cdot d$ is the number of parameters in φ and $m_1 = n \cdot C$ is the number of parameters in F_f. Similar to $\frac{\partial \varrho}{\partial \varphi}$, this matrix is infeasible to compute in general.

Fortunately, similar to most automatic differentiation libraries, this quantity is not computed explicitly. Instead, Vector-Jacobian Products (JAX) are used to compute the quantity $\frac{\partial T}{\partial F_f} = -v^T \frac{\partial \varrho}{\partial F_f}$ directly. The quantity φ^* is used during the forward pass to compute the training loss Eq. (4.2) and the backward gradient $\frac{\partial T}{\partial \varphi}$. This gradient is modified using Eq. (4.16) to obtain the modified gradient v. The backward pass for feature gradients is then obtained by first computing the scalar quantity $h = v^T \cdot \varrho(F_f, F_m, \varphi)$. The derivative of the scalar h with respect to F_f is $\frac{\partial h}{\partial F_f} = v^T \frac{\partial \varrho}{\partial F_f}$. This is an application of the chain rule to compute vector-Jacobian product $\frac{\partial h}{\partial F_f}$ without explicitly computing the full matrix $\frac{\partial \varrho}{\partial F_f}$. The gradients of F_m are obtained similarly. C.4 outlines the pseudocode for computing the gradients of the feature images with respect to the training loss. These features can then be propagated back to the network to update the weights.

4.4.2. Multi-scale optimization

Iterative optimization based methods typically use a multi-scale approach to improve convergence and avoid local minima with the image matching objective (Avants et al., 2006, 2008a; Ashburner, 2007; Beg et al., 2005). However, the downsampling of intensity images leads to indiscriminate blurring and loss of details at the coarser scales. We adopt a multi-scale approach by using pyramidal features from the network, which are naturally built into many UNet-like architectures. We consider two network designs for extracting multi-scale features:

- **Shared decoder**: We consider UNet-like architectures, and use the features from the decoder layers at each resolution as multi-scale features. The decoder features from each resolution are fed to an additional convolutional layer to obtain multi-scale feature maps for the fixed and moving images. This allows propagation of dense gradients to multiple decoder layers and feature sharing within the network. This architecture is illustrated in Fig. C.10(a).

- **Independent decoders**: We consider a cascade of networks with separate decoders, where each decoder processes the images at different resolutions. We hypothesize that for multi-scale optimization, independent consideration of different scales may be necessary to extract relevant features at each scale.

This architecture is illustrated in Fig. C.10(b).

We perform an ablation study on the choice of network architecture in Section 4.5.4.

Given these multi-scale features $\mathcal{F}_f = [F_f^n \dots F_f^2, F_f^1]$ and $\mathcal{F}_m = [F_m^n \dots F_m^2, F_m^1]$, we first perform optimization at the coarsest scale n, and store the result $\varphi^{*(n)}$. For each subsequent level $k < n$, we first upsample $\varphi^{*(k+1)}$ to the resolution of F_f^k, and use this as initialization for the optimization at the next finer scale. Finally, all the upsampled $\varphi^{*(k)}$ are used to compute the training loss Eq. (4.2). This mimics traditional multi-scale optimization methods while storing the result of the optimization at each scale for backpropagating to all feature maps. This asymmetry of the multi-scale features allows the network to learn different features at different scales, for example, large ventricles at coarser scales and small sulci structure at finer scales. A comparison of classical registration algorithm and our algorithm is highlighted in Algorithms 6 and 7. More details about the implementation of our method are discussed in Section C.2.

4.5. Experiments

We show the efficacy of DIO on a comprehensive experiment setup. First, we show that our method can synthesize dense feature maps from sparse intensity images, facilitating sparse or dense registration. We illustrate this on a toy dataset where classical optimization methods fail due to the lack of gradients in the loss landscape. This is especially relevant for incorporating sparse anatomical landmark losses into registration, where classical methods typically do not provide meaningful gradients. Second, we compare the in-distribution performance and flexibility of our learned representations with existing methods that aim to leverage either (a) pretrained features or intensity images for iterative optimization, (b) end-to-end or learned image features for parametric warp field regression, and (c) learning-based explicit unrolled iterative methods. We choose two community-standard datasets for this comparison – the OASIS dataset for inter-subject brain MRI registration, and the NLST dataset for intra-subject lung CT registration. Qualitatively, we show our multi-scale features are task-aware, interpretable and agnostic to choice of solver, and the implicit differentiation framework allows high expressive capacity for optimization than baselines. Third, to substantiate the robustness of DIO, we evaluate its performance on three out-of-distribution (OOD) neuroimaging datasets. Our method demonstrates remarkable robustness to domain shift, outperforming other prediction-based methods. This robustness is important in the context for DLIR since domain-shift leads to a shift in the distribution of warps, subseqeuntly resulting in poor generalization (Fu et al., 2020a; Wolterink et al.; Mok and Chung, 2022; Bigalke et al., 2022; Hansen and Heinrich, 2021) , limiting deployment in clinical settings. Furthermore, we show that our method allows *zero-shot* test-time switching of optimizers and efficacy across architectures, enabling arbitrary transformation representations and constraints at test time. We also evaluate the inference time of our method and compare it to explicit recurrent architectures that emulate iterative optimization, and show that our method is fast, compute-efficient and amenable to rapid experimentation and hyperparameter tuning. Finally, we examine the effect of choosing different implicit differentiation backends, and show that Jacobian-free backprop is the most well-conditioned and efficient for our task.

4.5.1. DIO learns flatter loss landscapes from sparse images

A key strength of DIO is the ability to learn interpretable dense features from sparse intensity images for accurate and robust image matching. This is particularly pertinent for medical image registration, where intensity images often exhibit significant heterogeneity in their gradient profiles, making registration difficult.

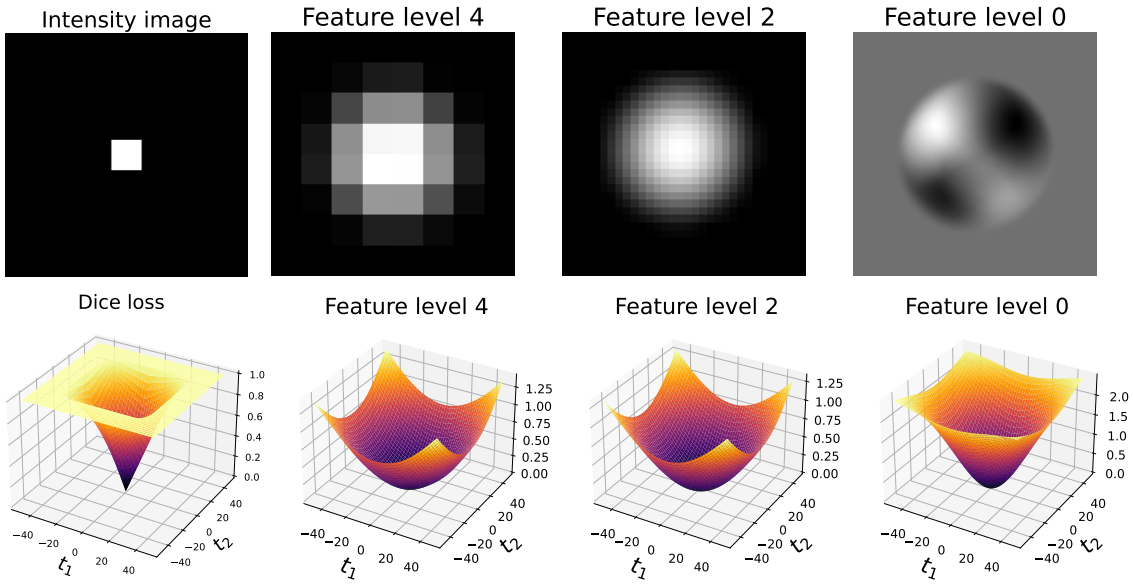


Figure 4.3: Dense feature learning leads to flatter loss landscapes. *Top row* shows the intensity image with the corresponding multi-scale features predicted by the deep network, where the L^{th} level denotes a feature of size $H/2^k \times W/2^k \times C_k$. *Bottom row* shows the loss landscape as a function of the relative translation between the squares in the fixed and moving image. Note the flat maxima which occurs when there is no overlap between the fixed and moving image, making optimization impossible if there is no overlap of the squares at initialization. On the contrary, the loss landscape for learned features is smooth, even at the finest scale, leading to much faster convergence even when there is no overlap between the intensity images. This allows registration without any centroid or moment-based preprocessing.

We design a toy task to isolate and demonstrate this behavior. In this task, the fixed and moving images are generated by placing a square of size 32×32 pixels on an empty canvas of 128×128 pixels. The probability of the squares in the fixed and moving images having non-zero overlap is set to 50%. The objective is to find an affine transformation to align the two images. However, classical optimization methods will fail this task 50% of the time, since there is no gradient of the loss function when the squares do not overlap, illustrated by the flat loss landscape in Fig. 4.3. In contrast, deep networks learn features that significantly flatten this loss landscape in the feature space. To demonstrate this, we train a network to output multi-scale feature maps that is used to iteratively optimize Eq. (4.1) to recover an affine transform. We choose a 2D UNet architecture, and the multi-scale feature maps are recovered from different layers of the decoder path of the UNet. Since the features are trained to maximize dice overlap, the loss landscape is much flatter, and the network is able to recover the affine transform with $> 99\%$ overlap regardless of whether there is any initial overlap or not (Section C.5). This allows registration of labelmaps with sparse gradients without any centroid or moment-based preprocessing (Legouhy et al., 2023; Yushkevich et al., 2016), which is typically done to offset the lack of gradients in the loss landscape. Moreover, end-to-end learning also enables learning of features that are most conducive to registration, unlike existing work (Wu et al., 2015; Ma et al., 2021; Wu et al., 2013; Quan et al., 2022) that may not contain discriminative registration-aware features about anatomical labels due to stagewise training.

Table 4.1: Quantitative performance on OASIS and NLST validation sets. DIO learns high-fidelity features incorporating both image and label matching into iterative optimization, showing superior performance compared to a variety of baselines.

Validation metrics on OASIS		
Method	**Dice**	**HD95**
Affine (Baseline)	0.572 ± 0.051	3.831 ± 0.718
ANTs Avants et al. (2008a)	0.786 ± 0.033	2.209 ± 0.534
NiftyReg Modat et al. (2010a)	0.775 ± 0.029	2.382 ± 0.723
LogDemons Vercauteren et al. (2008)	0.804 ± 0.022	2.068 ± 0.448
FireANTs Jena et al. (2026)	0.791 ± 0.028	2.793 ± 0.602
SynthMorph Hoffmann et al. (2021)	0.785 ± 0.023	2.311 ± 0.452
ConvexAdam + intensity Siebert et al. (2024)	0.792 ± 0.030	2.710 ± 0.555
DINO-reg Song et al. (2024)	0.509 ± 0.031	5.667 ± 0.638
Cyclic-Reg Bigalke et al. (2023)	0.763 ± 0.033	2.539 ± 0.723
GradIRN Qiu et al. (2022)	0.746 ± 0.016	8.232 ± 0.715
SUITs Blendowski et al. (2021)	0.615 ± 0.047	3.923 ± 0.498
KeyMorph (MSE)	0.608 ± 0.039	3.886 ± 0.458
KeyMorph (Dice)	0.642 ± 0.021	3.560 ± 0.394
Ours (UNet backbone)	$\mathbf{0.853 \pm 0.018}$	$\mathbf{1.675 \pm 0.379}$
Ours (LKU backbone)	$\mathbf{0.862 \pm 0.017}$	$\mathbf{1.584 \pm 0.351}$

Validation metrics on NLST	
Method	**TRE30 (in mm)**
Zero displacement (Baseline)	9.76
VoxelMorph Balakrishnan et al. (2019)	4.12
Im2Grid Liu et al. (2022)	3.05
SyN	3.04
Vector-Field Attention Liu et al. (2024c)	2.31
RWC-Net Sivan et al. (2023)	2.11
unigradICON Tian et al. (2024)	2.07
unigradICON + instance optimization	1.77
FireANTs	1.28
FireANTs + MIND	1.18
ConvexAdam + MIND	1.17
Ours + MIND	**1.02**

4.5.2. Comparison of in-distribution performance

Datasets We evaluate our method on two datasets - the OASIS dataset for inter-subject brain MRI registration, and the NLST dataset for intra-subject lung CT registration.

OASIS: The OASIS dataset (Marcus et al., 2007b) contains 414 T1-weighted MRI scans of the brain with label maps containing 35 subcortical structures extracted from automatic segmentation with FreeSurfer and SAMSEG. We use the preprocessed version and train-val split from the Learn2Reg challenge (Hering et al., 2022) where all the volumes are skull-stripped, intensity-corrected and center-cropped to $160 \times 192 \times 224$. We

evaluate the Dice score and the 95th percentile of the Hausdorff distance (HD95) between the warped and fixed label maps.

NLST: The National Lung Screening Trial (NLST) dataset (Team, 2011) consists of intra-subject inspiration-expiration pairs. The preprocessed version consists of 200 training pairs and 10 validation pairs, with corresponding keypoints obtained using automatic landmark detection using the Foerstner operator. Owing to large variability in lung volume due to inspiration and expiration, the NLST dataset requires large deformation fields to align the two volumes reliably. We evaluate the 70th percentile of target registration error (TRE30) of the keypoints between the warped and fixed volumes from the validation dataset.

Baselines We consider a variety of baselines for this comparison. Our primary contribution is enabling the synergy of task-aware feature learning and powerful black-box solvers end-to-end. Therefore, the baselines are categorized into three relevant groups – **(a)** using intensity images or generic pretrained features combined with iterative optimization-based methods, **(b)** parametric regression of warp fields using neural network, and **(c)** learning-based explicit unrolled iterative methods.

For the OASIS daaset, we consider **(a)** SyN (Avants et al., 2008a), NiftyReg (Modat et al., 2010a), Log Demons (Vercauteren et al., 2008), FireANTs (Jena et al., 2026), ConvexAdam (Siebert et al., 2024), DINO-Reg (Song et al., 2024), **(b)** SynthMorph (Hoffmann et al., 2021), KeyMorph (Wang et al., 2023), Cyclical Self-Training (Bigalke et al., 2023), **(c)** multimodal SUITS (Blendowski et al., 2021) and GradIRN (Qiu et al., 2022). For the NLST dataset, we consider (a) ConvexAdam, SyN, FireANTs (b) VoxelMorph (Heinrich and Hansen, 2022), unigradICON (Tian et al., 2024) (with and without instance optimization), Vector-Field Attention (Liu et al., 2024c), Im2grid (Liu et al., 2022), and (c) RWC-Net (Sivan et al., 2023). All methods are trained with a combination of intensity and label or keypoint matching losses, wherever applicable.

Results Table 4.1 summarizes the results. On the OASIS dataset, we observe that all iterative optimization methods perform in the same ballpark without supervision. We run ConvexAdam with the intensity images and do not observe any improvement over the unsupervised baselines. We swap the intensity images with DINO features (DINO-reg without ensembling) and observe no improvement in performance - bolstering our claim that generic features do not guarantee task-specific performance. Iterative methods like GradIRN and SUITs are modified and trained on both the intensity images and label maps, but do not show significant improvement either, due to the limited expressivity of unrolled optimizations. Supervised parametric baselines like LapIRN and LKU-Net show much better performance for in-distribution datasets, but completely breakdown for out-of-distribution datasets (Section 4.5.3). Our method shows a significant improvement over unsupervised iterative methods, generic features, and explicit unrolling of optimization.

On the NLST dataset, we see a similar trend where unsupervised optimization methods like ConvexAdam and FireANTs show solid performance, while parametric methods like VoxelMorph, unigradICON, Vector-Field Attention, and Im2grid show relatively poor performance. unigradICON substantially improves with instance optimization, indicating the necessity of instance optimization for robust registration. RWC-Net being an iterative method also reports a poorer performance compared to ConvexAdam and FireANTs, showing that powerful optimization solvers with handcrafted features can surpass learned features with limited expressivity of unrolled optimization. DIO improves over the unsupervised baselines and parametric warp field estimators, showing robust performance to multiple anatomical structures and large deformations.

Key differences with closely related methods The closest works to our method are **(a)** KeyMorph that emulates feature learning from images for (non-iterative) registration, and **(b)** GradIRN, RWC-Net, and SUITs that emulate iterative optimization with explicit recurrent modules. Using a framework like KeyMorph limits

the warp representation that can be computed using closed-form solutions like affine, or thin-plate splines (TPS). TPS represents a very limited class of warps, cannot be guaranteed to be diffeomorphic, and a vast majority of widely used parameterizations (free-form, SVF, geodesic, LDDMM, SyN) do not admit closed form solutions rendering KeyMorph unsuitable for many advanced registration applications. We compare the qualitative expressivity of the warp field and transformed images generated by KeyMorph with that of our method in Figs. C.7 and 4.4. Explicit recurrent modules, on the other hand, are stateful and are limited to few iterations due to memory constraints. This also limits the expressivity of the generated warp fields despite not being limited to closed-form solutions. Moreover, we note that KeyMorph is highly compute-intensive, quickly running out of memory on an A6000 GPU with 512 keypoints, even with a truncated UNet backbone and float16 mixed precision training. GradIRN, RWC-Net, and SUITs face memory contraints because of their explicit recurrent modules, and are limited to a few iterations. On the other hand, DIO produces dense multi-scale image features, which would equivalently correspond to about $192 * 224 * 160 * (1 + 1/8 + 1/64) * 16/3 \sim 41$ million keypoints for a standard MRI image across multiple scales, and can be run for a hundreds of iterations without memory constraints. This allows us to express maximal expressivity both in the feature representation and the capacity of the optimization solver.

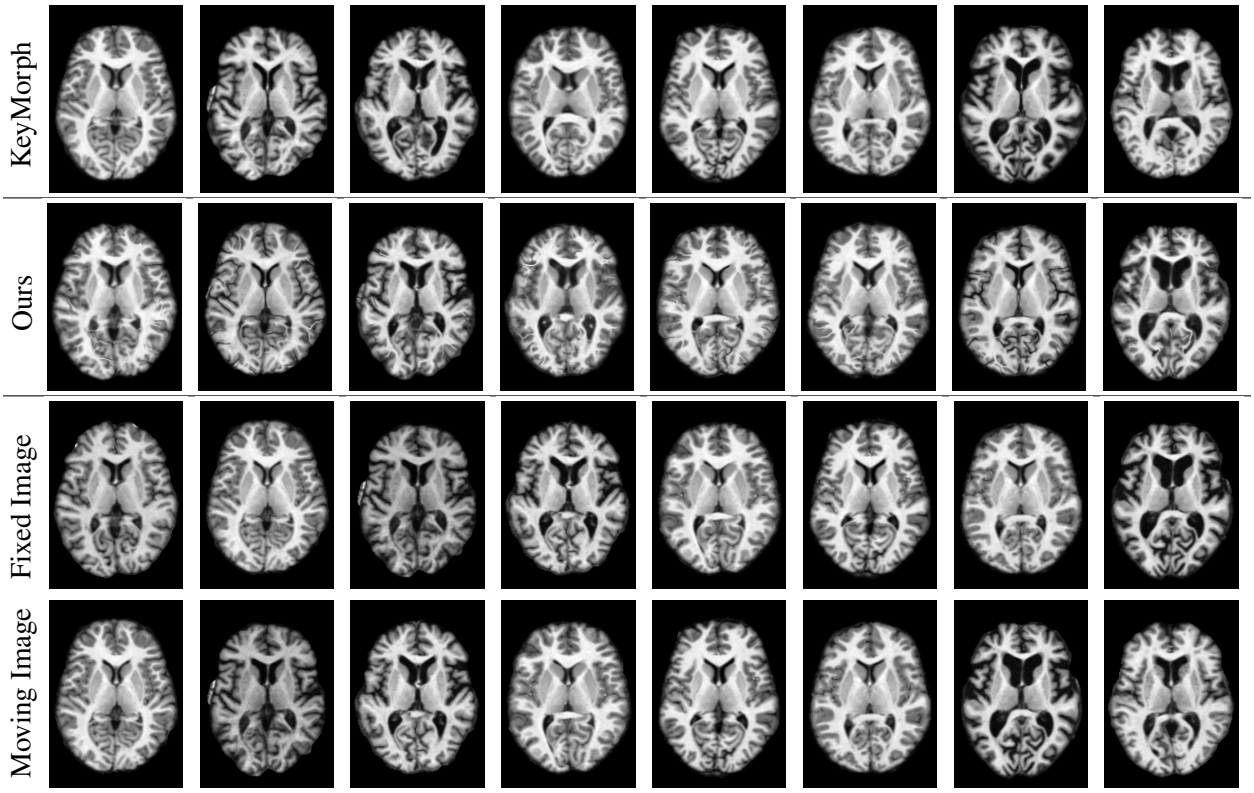


Figure 4.4: Qualitative comparison of KeyMorph and our method on OASIS dataset. The first row shows the warped images using KeyMorph and the second row shows the warped images using our method. The third and fourth rows show the fixed and moving images, respectively. The OASIS dataset consists of skull-stripped T1-MRI brains that are affinely registered to the Talairach space, consequently we focus on deformable registration. KeyMorph uses 512 keypoints to parameterize a thin-plate spline transformation, while our method uses an optimizer to predict a dense deformation field. Our method demonstrates high fidelity registration, compared to KeyMorph that only partially warps large differences in ventricles (last two columns). More qualitative comparisons, including segmenation maps, and predicted warp fields are shown in Figs. C.7 to C.9.

4.5.3. DIO inherits robustness to domain shift from iterative optimization

A key requirement of registration algorithms is to be robust to a spectrum of scanner configurations, acquisition, preprocessing and labelling protocols, since there are different standards across institutions. Existing prediction-based DLIR methods are very sensitive to domain shift (Mok et al., 2023; Jena et al., 2024; Jian et al., 2024), and catastrophically fail on other brain datasets. On the contrary, DIO inherits the domain agnosticism of the optimization solver, and is robust under feature distortions introduced by domain shift.

Datasets We evaluate the robustness of the trained models on three brain datasets: LPBA40, IBSR18, and CUMC12 datasets (Shattuck et al., 2008; ibs; Klein et al., 2009). Contrary to the OASIS dataset, these datasets were obtained on different scanners, affinely pre-aligned to different atlases (MNI305, Talairach) with varying algorithms used for skull-stripping, bias correction (BrainSuite, autoseg), and different *manual* labelling protocols for different anatomical regions (as opposed to automatically generated Freesurfer labels in OASIS). Unlike the OASIS dataset, these datasets also have different voxel sizes for different brain scans, and IBSR18 and CUMC12 datasets have non-uniform anisotropic volumes. More details about the datasets are provided in Section C.7. We note that increasing label map overlap with automatically generated labels during training is easier for DL-based parametric registration methods. Therefore, the performance of DL-based methods on *unseen, manually generated* parcellations is crucial for clinical translation. The aforementioned aspects of the chosen community-standard datasets make them challenging for DLIR methods, and highlight the crucial shortcoming of these methods, i.e. lack of generalization to domain shift.

Baselines We employ a variety of deep learning baselines for this experiment. We consider the original VoxelMorph (Balakrishnan et al., 2019) pretrained model that is trained using an unsupervised objective function, SynthMorph (Hoffmann et al., 2021) that is trained on procedurally generated synthetic data using upsampled Perlin noise. Cyclical-Reg (Bigalke et al., 2023) is similar to SynthMorph in that it is trained on a self-supervised objective without any label or image supervision. The training framework emulates a few consistencies of the predicted warp field like inverse-consistency and matching the results of iterative optimization. Furthermore, two pyramidal architectures that mimic multi-scale prediction - LapIRN (Mok and Chung, 2020c) and its conditional counterpart named Conditional LapIRN (Mok and Chung, 2021) are also suitable prediction-based baselines. A symmetric normalization network dubbed SymNet (Mok and Chung, 2020b) that performs symmetric predictions from the fixed and moving images is also used to compare with their non-symmetric counterparts. The pretrained models in SymNet and LapIRN are trained without dice loss; we also train models that include dice loss for comparison. We also include a large kernel UNet (LKU) (Jia et al., 2022) which has showed high accuracy in the OASIS dataset, albeit with implausible deformations (Jian et al., 2024). We also consider three variants of transformer-based TransMorph for registration (Chen et al., 2022b). Specifically, we use the provided pretrained model for *TransMorph-large* and two variants of *TransMorph-regular* trained with and without Dice loss. Finally, we consider ConvexAdam, DINO-reg, multimodalSUITs, and GradIRN as baselines employing iterative optimization.

This assortment of baselines represent a spectrum of design choices in deep learning for registration, and are representative of the state-of-the-art in DLIR. These methods show excellent performance on in-distribution datasets with automatically generated parcellations. To evaluate the generalization to out-of-distribution datasets, we train all models on the OASIS training split and evaluate on all pairs of the LPBA40, IBSR18, and CUMC12 datasets.

Owing to the predictive paradigm of most baselines, we also evaluate their performance with and with-

out instance optimization. Following VoxelMorph++ (Heinrich and Hansen, 2022), we finetune the output representation for 100 iterations with the normalized cross-correlation (NCC) loss, and Adam optimizer with a learning rate of 10^{-3}. Note that almost none of these baselines come with instance optimization postprocessing, therefore we manually implement, evaluate and validate the performance of the instance optimization solver for each baseline, requiring significant effort.

Evaluation We evaluate across a variety of configurations – (i) preserving the anisotropy of the volumes or resampling them to 1mm isotropic (denoted as *anisotropic* or *isotropic* respectively), and (ii) center-cropping the volumes to match the size of the OASIS dataset (denoted as *Crop* and *No Crop*). The results for all three datasets are shown in Fig. 4.5 sorted by mean Dice score; quantitative comparison is also shown in Appendix Table C.1. Fig. 4.5 shows boxplots with each color representing a different method, and a more translucent shade for the baseline without instance optimization. Note that TransMorph, VoxelMorph, and SynthMorph do not work for sizes that are different than the OASIS dataset due to design decisions and implementation constraints, therefore they only work in the *Crop* setting. The IBSR18 dataset consists of different volumes with different levels of anisotropy; consequently resampling them to 1mm isotropic leads to different voxel sizes. These volumes cannot be concatenated along the channel dimension, consequently every DLIR method cannot run under this configuration (Fig. 4.5(a)). In contrast, similar to KeyMorph, our method employs a dual-stream-like architecture that processes one volume at a time. Since our method utilizes a dual-stream-like convolutional architecture processing one volume at a time, the fixed and moving images can have different voxel sizes, i.e. **feature extraction is not contingent on the voxel sizes of the moving and fixed images being equal**. The optimization solver can also handle different voxel sizes for the fixed and moving volumes – which is useful in applications like multimodal registration (in-vivo to ex-vivo, histology to 3D, pre-operative to intra-operative, microscopy to MRI). This unprecedented flexibility brings forth a new operational paradigm in deep learning for registration combining feature learning to incorporate label fidelity with optimization-as-a-layer to be robust, widening the scope of applications for registration with deep features. This experiment provides a few key insights about existing DLIR methods.

Predictive registration methods do not generalize their performance under domain shift

Image registration is a highly ill-defined and non-convex problem, which is NP-hard to solve in general. Learning a parametric statistical model to amortize optimization can learn a distribution of warps that are specific to the training dataset. However, there is no explicit mechanism to ensure that the predicted warp field indeed performs correspondence in *any* space of feature maps. For domain shift in the input images, the warp fields predicted by the model need not be the local minima of *any* optimization function. This implies that predictive methods for registration would not easily generalize outside the training domain. Moreover, this lack of generalization is not mitigated by label supervision during the training phase, as evident by baselines with supervised label losses underperforming their unsupervised counterparts. This behavior is not noticed by us alone; (Mok and Chung, 2022) observe that the supervised models are inferior to their unsupervised models in the LPBA dataset, indicating anatomical knowledge injected to the model with supervision may not generalize well to unseen data beyond the training data. The need for instance optimization (IO) for improved performance is shown to be necessary for foundational models as well (Tian et al., 2024). The benefit of amortized optimization does not hold anymore since IO becomes a necessity and consequently a bottleneck for generalization to domain shift. In fact, most of the inference time is now dominated by the (sequential) IO routine. However, instance optimization routines have become fast, motivating a shift towards robust feature learning paradigm instead.

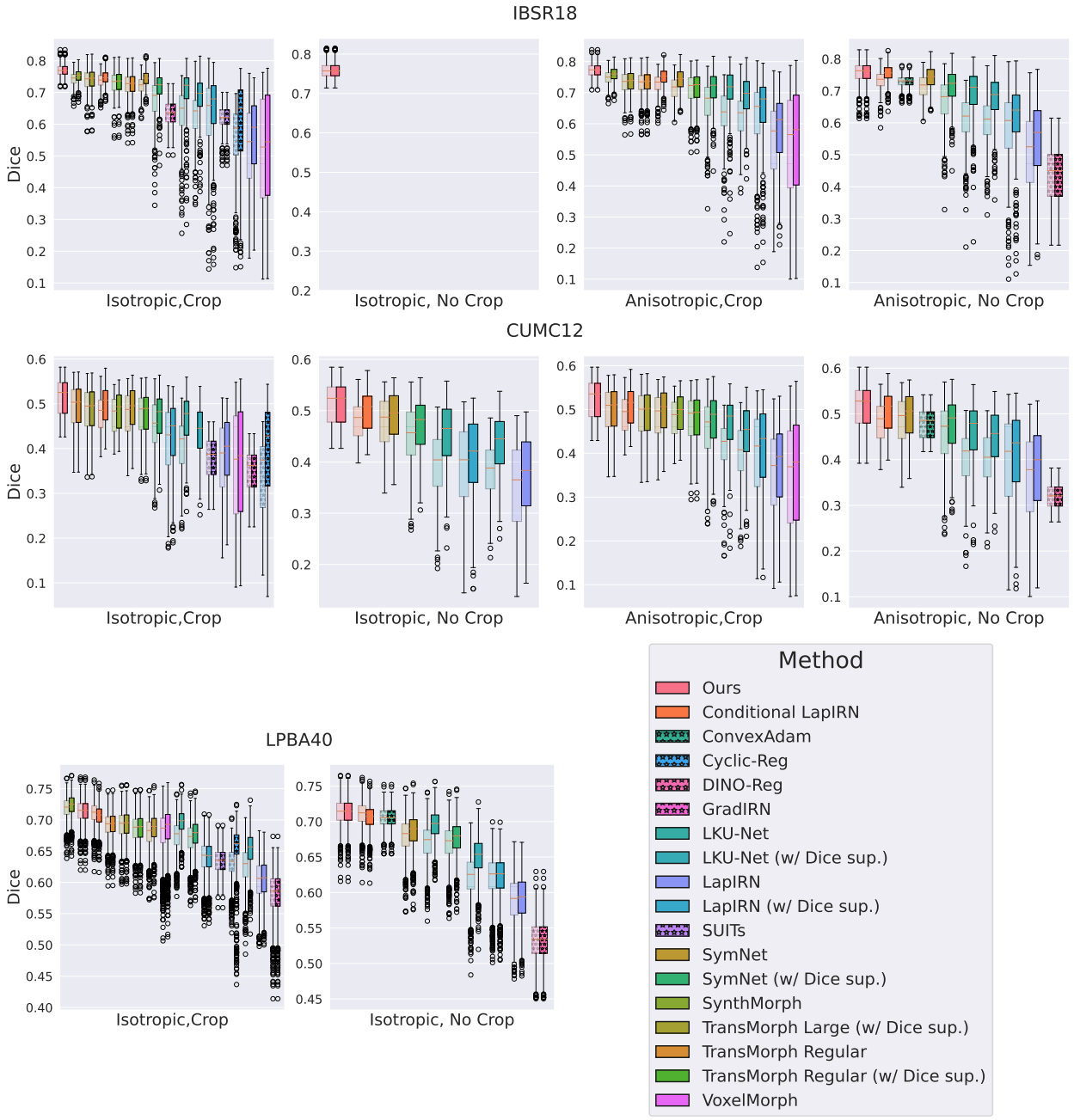


Figure 4.5: Boxplots of Dice scores for three out-of-distribution datasets. DIO performs significantly better across three datasets without additional finetuning. Contrary to other baselines that output warp fields considering 1mm isotropic data, leading to a performance drop with anisotropic volumes, DIO performs better with anisotropic data due to the optimization's resolution-agnostic nature. Even with image-space instance optimization, almost all baselines underperform compared to DIO.

DLIR methods do not provide good initialization for out-of-distribution data

Despite the need for instance optimization, one may want to use predictive registration methods for initialization to reduce the number of iterations required for the subsequent instance optimization. However, predictive methods do not provide good initialization either, as the performance of baselines does not surpass our method even with 100 iterations at the finest scale, compared to only 20 iterations at the finest scale for our method in Fig. 4.5. If the initialization is downsampled to perform multi-scale instance optimization, most of the initialization information is lost during downsampling. For example, if a multi-scale instance optimization is performed with the coarsest scale at $1/4^{\text{th}}$ resolution, around **98.4%** $(= 63/64)$ of the initialization is discarded. This kind of instance optimization then closely resembles classical intensity-based optimization instead, rendering the initialization from predictive methods redundant. Another limitation of instance optimiztion is also observed in (Mok et al., 2023) wherein instance optimization typically achieves minimal improvements on solidly trained neural networks. For out-of-distribution data, our experiments also corrobate the fact that initialization from learned coarser feature maps (ours) is consistently robust compared to initialization from predictive methods.

DIO remedies both these issues using high-fidelity multi-scale features

Under our feature learning paradigm, we are able to circumvent the bad initialization problem by not predicting any warps at all, and instead performing a multi-stage instance optimization with learned features. Figs. 4.6 and 4.7 show that our learned feature maps provide higher-fidelity warps compared to intensity images at all levels, while being interpretable. Since most of the iterative computation is performed at the coarser scales, this leads to fast runtimes than baselines with instance optimization. DIO also provides robust performance and low variance across different datasets, as shown in Fig. 4.5. Our novel methodology sidesteps initialization using prediction altogether.

4.5.4. Robust feature learning enables zero-shot performance by switching optimizers at test-time

Another major advantage of our framework is that we can switch the optimizer *at test time* without any retraining. This is useful when the registration constraints evolve over time (i.e. initially diffeomorphic transforms were required but now non-diffeomorphic transforms are acceptable), or when the registration is used in a pipeline where different parameterizations (freeform, diffeomorphic, geodesic, B-spline) may be compared. Since our framework decouples the feature learning from the optimization, we can switch the optimizer arbitrarily at test time, at no additional cost. A crucial requirement is that learned features should not be too sensitive to the instance optimization routine.

To demonstrate this functionality, we use the validation set of the OASIS dataset and four network architectures. We consider the vanilla UNet (Ronneberger et al., 2015) and Large Kernel UNet (Jia et al., 2022) networks, and Encoder-only and Encoder-Decoder architectures for each network. The difference in architectures are visualized in Fig. C.10. These networks were initially trained using the SGD optimizer without any additional constraints on the warp field. At test time, we switch the optimizer to the FireANTs optimizer (Jena et al., 2026), that uses a Riemannian Adam optimizer for multi-scale diffeomorphisms. If the features had overfit to the training dynamics of the SGD optimizer, we would expect a significant drop in performance at test time. Unlike explicit iterative unrolling, implicit optimization theoretically ensures that the gradient of the inputs to

Optimizer	SGD			FireANTs (diffeomorphic)		
Architecture	**DSC**	**HD95**	$\%(\|\mathbf{J}\| < \mathbf{0})$	**DSC**	**HD95**	$\%(\|\mathbf{J}\| < \mathbf{0})$
UNet Encoder	0.845 ± 0.018	1.790 ± 0.433	0.7866 ± 0.1371	0.834 ± 0.018	1.847 ± 0.410	0.0000 ± 0.0000
LKU Encoder	0.849 ± 0.018	1.733 ± 0.401	0.8079 ± 0.1308	0.838 ± 0.018	1.806 ± 0.373	0.0000 ± 0.0000
UNet	0.853 ± 0.018	1.675 ± 0.379	1.0718 ± 0.1662	0.842 ± 0.018	1.748 ± 0.397	0.0000 ± 0.0000
LKU	0.862 ± 0.017	1.584 ± 0.351	0.8646 ± 0.1429	0.849 ± 0.017	1.740 ± 0.345	0.0000 ± 0.0000

Table 4.2: Zero shot performance by switching optimizers at test-time. Our method is trained on the OASIS dataset with the SGD optimizer to obtain the warp field. At inference time, we use an SGD optimizer for no constraint on the warp field, and the FireANTs optimizer to ensure diffeomorphic warps. Across all architectures, the Dice Score remains robust, with only a slight dip attributed to the constraints introduced by diffeomorphic mappings. The SGD optimization introduces $\sim$1% singularities, while FireANTs shows no singularities.

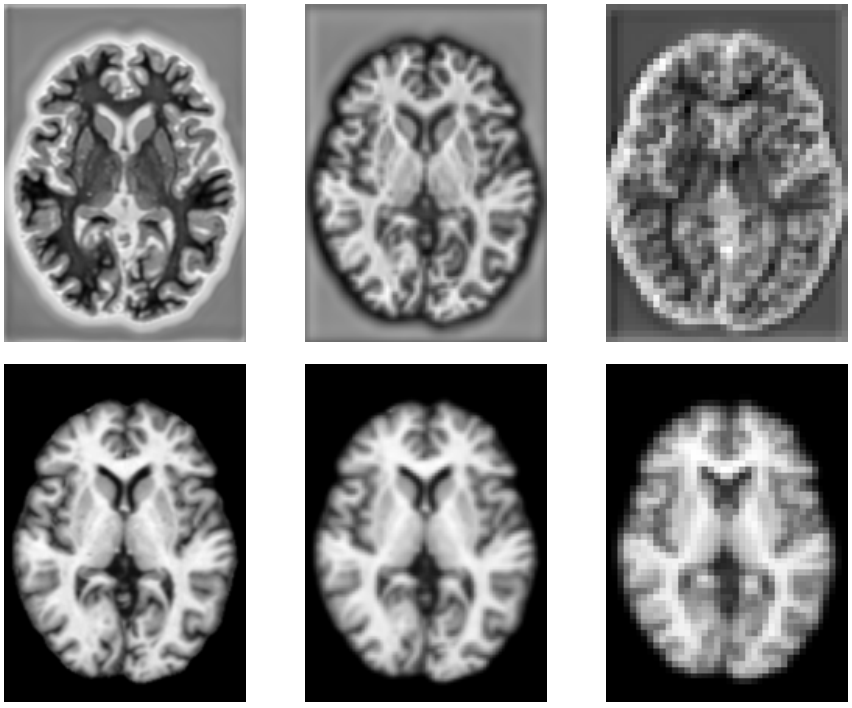


Figure 4.6: Examples of multi-scale features learned by the feature extractor. Scale-space features (*bottom row*) obtained by downsampling the image downsample all image features indiscriminately. Our features (*top row*) preserve necessary anatomical information at all scales, and introduce inhomogenity in the feature space for better optimization (watershed effect and enhanced contrast near gyri and a halo around the outer surface to delineaate background from gray matter).

the solver is *independent* of the optimization path, and is only dependent on the final result of the solver.

Results in Table 4.2 compare the Dice score, 95th percentile of the Haussdorf distance (denoted as *HD95*) and percentage of volume with negative Jacobians (denoted as $\%(\|J\| < 0)$) for the two optimizers. The SGD optimizer introduces anywhere from 0.79% to 1.1% of singularities in the registration, while the FireANTs optimizer does not introduce any singularities. A slight drop in performance can be attributed to the additional implicit constraints imposed by diffeomorphic transforms. However, the high-fidelity features lead to a much better label overlap than FireANTs run with image features (Table 4.1 and Fig. 4.7). Our framework introduces an unprecedented amount of flexibility at test time that is an indispensible feature in deep learning

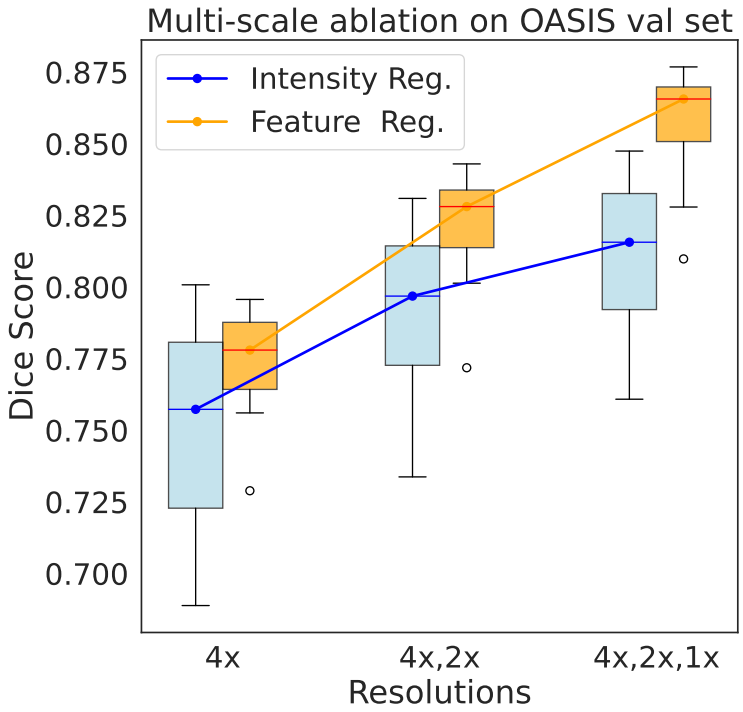
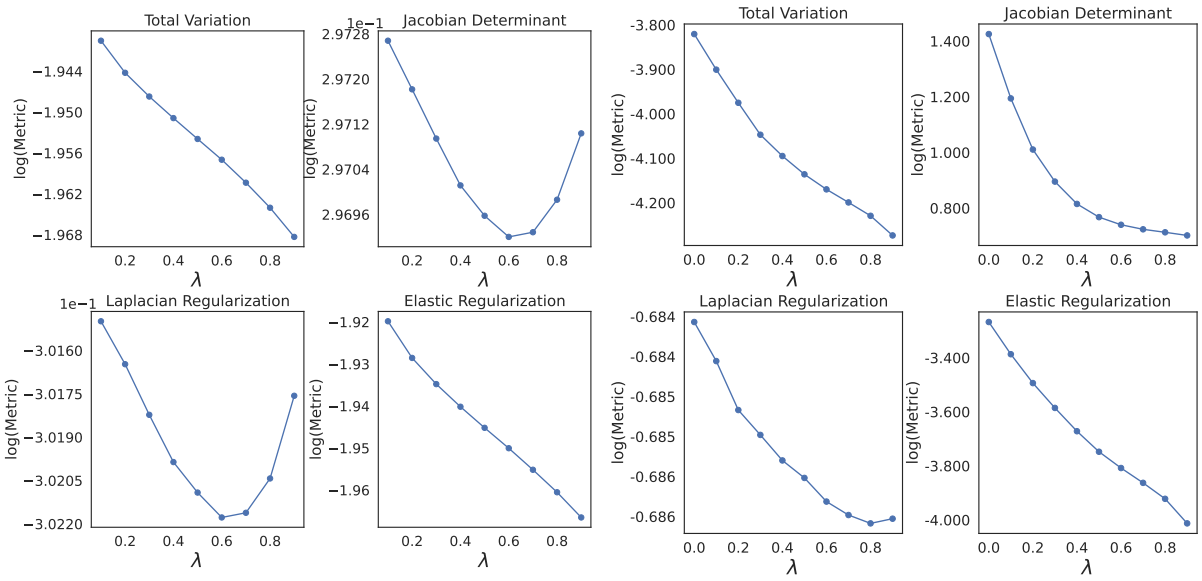


Figure 4.7: Ablation on fidelity of multi-scale features compared to multi-scale intensity images. To show that multi-scale features provide more label-aware information than intensity images alone, we perform registration on the OASIS validation set using multi-scale features and intensity images. For intensity-based multi-scale registration, the intensity images are smoothed and downsampled at each level. x-axis shows the resolutions at which optimization is performed, and y-axis shows the distribution of Dice scores. For identical multi-scale optimization routines, feature-based registration provides better label alignment than intensity images at all resolutions. This demonstrates the efficacy of task-awareness in features learned using our framework.

for registration, and can be useful in a variety of applications where the registration requirements change over time, without expensive retraining.

4.5.5. Interpretability of features

Decoupling of feature learning and optimization allows us to examine the feature images obtained at each scale to understand what feature help in the registration task. Classical methods use scale-space images (smoothened and downsampled versions of the original image) to avoid local minima, but lose discriminative image features at lower resolutions. Moreover, intensity images may not provide sufficient details to perform label-aware registration. Since our method learns dense features to minimize label matching losses, we can observe which features are necessary to enable label-aware registration. Fig. 4.6 highlights differences between scale-space images and features learned by our network. At all scales, the features introduces heterogeneity using a watershed effect and enhanced contrast to improve label matching performance.



(a) Effect of λ on different R_u with HyperMorph.

(b) Effect of λ on different R_u with DIO.

Figure 4.8: Comparison of regularization at inference time. With HyperMorph, regularizations like Volume Preservation and Laplacian Registration are not monotonic with the training hyperparameter λ, and have to be considered during training. In contrast, due to the decoupled feature learning and optimization, DIO can be run with arbitrary regularization families at test time without any retraining, and monotonic trends with λ are observed.

4.5.6. DIO provides flexible Regularization Tuning

DLIR methods are typically trained with a *fixed* loss function and regularization, leading to inflexible regularization for novel image contrasts, resolutions, or anatomy. HyperMorph (Hoopes et al., 2021) introduced a method to amortize optimization over different hyperparameters in a deep network by providing the regularization parameter λ as an input to the network. The HyperMorph network is trained with the following loss function conditioned on λ:

$$\mathcal{C}_\theta(\varphi, \lambda) = (1-\lambda)L(I_f, I_m \circ \varphi_\theta) + \lambda R_v(\varphi_\theta) \tag{4.17}$$

where $R_v(\varphi)$ is the total variation on the velocity field of the diffeomorphic transform.

$$R_v(\varphi_\theta) = \|\nabla v_\theta\|_2^2 \quad , \quad \varphi_\theta = \exp_{\text{Id}}(v_\theta) \tag{4.18}$$

However, the regularization is fixed during training, and a model trained to minimize the total variation may not have similar regularization effects on other unseen regularization families, like Jacobian regularization, curvature, or Laplacian regularizations. Incorporating n different regularization families would require a combinatorial amount of conditional inputs to capture the full hyperparameter space. This will require significant training time, and will still be inflexible for other unseen hyperparameter families. In contrast, our method can work with *arbitrary* unseen regularization families and hyperparameters at test time without any retraining.

To demonstrate this, we consider the pretrained HyperMorph model. For our method, we perform feature

training on Eq. (4.1) without any regularization, and at inference time, we add a regularization term to the optimization loss as follows:

$$\mathcal{C}_\theta(\varphi, \lambda) = (1 - \lambda)L(F_f, F_m \circ \varphi) + \lambda R_u(\varphi) \tag{4.19}$$

We consider four families of R_u:

- **Total variation of the warp field**: $R_u(\varphi) = \int_\Omega \|\nabla\varphi\|_2^2 d\Omega$. We hypothesize that this term will be directly affected by the total variation of the velocity field in HyperMorph, as the exponential map of a smooth velocity field is likely to be smooth, due to the smoothness of the exponential map itself.

- **Elastic reg**: $R_u(\varphi) = \int_\Omega \left(\alpha\|\nabla\varphi\| + \beta\|\nabla^2\varphi\|\right) d\Omega$. This term is performed implicitly in the popular SyN algorithm, and is likely to be affected by the total variation energy in HyperMorph as well. We set $\alpha = \beta = 1$ for this experiment.

- **Jacobian det**: $R_u(\varphi) = \int_\Omega \left(|\det(\nabla\varphi) - 1|_2^2\right) d\Omega$. This term is used in diffeomorphic registration to ensure volume preservation, and this term is less likely to have a monotonic relationship with the total variation of the velocity field.

- **Laplacian reg**: $R_u(\varphi) = \int_\Omega \|\Delta\varphi\|_2^2 d\Omega$. The effects on this regularization are not monotonic with the total variation of the velocity field.

For the HyperMorph model, we evaluate the regularization losses for each λ to see the effect of R_v on other regularization families R_u. Results in Fig. 4.8a show that total variation and elastic regularization follow monotonic trends with λ since reducing $\|\nabla v\|_2^2$ will induce smoothness to the velocity field, and consequently smoothness to the warp field due to the smoothness of the exponential map. However, the Laplacian and Jacobian regularization do not follow monotonic trends with λ, indicating that additional training would be required to incorporate these regularizations. In contrast, Fig. 4.8b shows that DIO can work with arbitrary regularization families at test time without any retraining, providing immense flexibility to arbitrary registration constraints at test time.

4.5.7. Ablation on choice of implicit gradient

In all our experiments, we use the Jacobian-free Backprop ((Fung et al., 2021)) approximation for approximating the gradient of the feature image. We ablate on the following choices of implicit gradient approximations: (a) full Hessian, (b) unrolled phantom gradients (Geng and Kolter, 2023; Geng et al., 2021) (UPG), and (c) Jacobian-free Backprop, on the OASIS dataset. Note that phantom gradients simply correspond to BPTT-like unrolling over k steps. We train the network with the same architecture and hyperparameters for 100 epochs, and evaluate the performance on the validation set.

Table 4.3 shows that Jacobian-free Backprop provides the best performance, followed by unrolled phantom gradients with $k = 3$. For the UPG variants, we run out of memory with $k = 10$ at the finest resolution due to the computational demands of explicit unrolling. The results for UPG also show that explicit unrolling is both computationally demanding and unstable compared to cheaper variants like JFB. For the full Hessian IFT, we observe poor training performance due to the ill-conditioning of the Hessian $\nabla_\varphi^2 C(\varphi, F_f, F_m) = \frac{\partial g}{\partial \varphi}$. Since this ill conditioned Hessian's inverse is multiplied with the incoming warp gradient $\frac{\partial T}{\partial \varphi}$, the feature gradient

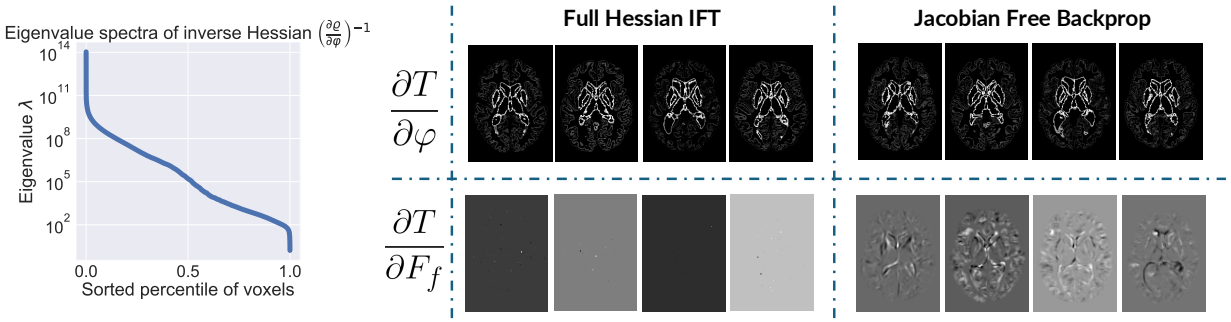


Figure 4.9: Qualitative comparison of different backward passes for DIO. **(Left)** Top eigenvalues of the inverse Hessian skew the feature gradient due to their large magnitude compared to the rest of the eigenspectra, **(Right)** qualitatively demonstrates the effect of the Hessian on the gradient of the training loss with respect to the transformation field φ and the fixed feature F_f using different instantiations of the backward pass at the beginning of training. Gradients w.r.t. feature images from Hessian-based IFT are very sparse and do not facilitate network learning. On the contrary, gradients obtained using JFB are dense and the network quickly convergences to low training loss.

Method	Dice
Full Hessian IFT	0.688
Unrolled Phantom Gradients (UPG) ($k = 10^*$)	0.782
Unrolled Phantom Gradients (UPG) ($k = 5$)	0.841
Unrolled Phantom Gradients (UPG) ($k = 3$)	0.842
Jacobian-free Backprop	**0.862**

Table 4.3: Ablation on choice of implicit gradient approximation. On the OASIS dataset, Jacobian-free Backprop achieves highest validation score while being computationally efficient. The full Hessian IFT suffers from the ill-conditioned Hessian of the registration problem, leading to poor convergence. We also observe monotonic decrease in validation performance with increasing k for UPG. * indicates that the model runs out of memory at finest resolution.

$\frac{\partial T}{\partial F}$ is sparse and noisy. This severe ill-conditioning of the inverse Hessian is also observed in (Jena et al., 2026).

We also examine the eigenspectra of the inverse Hessian matrix and its effect on the feature gradient computation in Fig. 4.9. We observe in Fig. 4.9 that although both full Hessian IFT and JFB receive the same gradients $\frac{\partial T}{\partial \varphi}$, they both provide significantly different feature gradients. The top eigenvalues of the Hessian matrix skew the gradient due to their large magnitude compared to the rest of the eigenspectra. This leads to sparse gradients with respect to the feature images (visualized as bright and dark speckles), consequently leading to poor training performance. This ablation provides motivation for future work to precondition the Hessian while addressing its ill-conditioned nature.

4.6. Discussion

DLIR methods provide several benefits such as amortized optimization, ability to leverage weak supervision and learn from large (labeled) datasets. However, coupling of the feature learning and optimization steps in

DLIR methods limits the flexibility and robustness of the deep networks. Existing attempts to synergize optimization and feature learning in deep networks have been limited due to two reasons. First, storing the entire computational graph of iterative optimization of 3D images will require an excessive (and often infeasible) memory footprint. Second, existing methods have limited mathematical formulation to enable the ability to backpropagate features from a generic iterative optimization-based solver to learnable, *task-aware* features of images. We highlight the shortcoming of the existing classes of methods that aim to mitigate this issue, and propose a novel paradigm that incorporates optimization-as-a-layer for learning-based frameworks. This paradigm allows the use of advanced black-box optimization toolkits in the forward pass, and a mathematically sound formulation to backpropagate features from the optimization solver to the feature learning network, without any additional compute or memory overhead.

Our comprehensive experimental setup on multiple datasets measuring both in-distribution and out-of-distribution performance demonstrates two solid empirical conclusions. First, task-aware learning is required for task-aware performance. Using generic feature extractors or emulating iterative solvers only for a few steps cannot achieve asymptotically optimal in-distribution performance. Second, iterative optimization is *necessary* for robustness to out-of-distribution data. Regardless of the performance of parameteric deep learning methods on in-distribution data, most methods fail to generalize on out-of-distribution data. Multi-scale iterative optimization is therefore necessary for robustness to unseen image characteristics typically encountered in real-world clinical scenarios. Since DIO combines task-specific feature learning and black-box iterative optimization end-to-end, our method achieves state-of-the-art performance on the in-distribution setting, and is robust to out-of-distribution data. Densification of features from our method also leads to better optimization landscapes, and our method is robust to unseen anisotropy and domain shift. To our knowledge, our method is the first to switch between transformation representations (free-form to diffeomorphic) at *test time* without any retraining.

CHAPTER 5

A Scalable and Distributed Framework for Multimodal GigaVoxel Image Registration

In chapter Chapter 3, we introduced FireANTs as a real-time registration framework that is robust to a long tail of modalities and datasets, while enabling the full flexibility of representing time-dependent flows of diffeomorphisms. This was followed by a framework introduced in Chapter 4 for using FireANTs as a fully differentiable layer to learn task-specific image features from auxiliary labelmaps or landmarks. Together, they address two major limitations of iterative optimization methods while preserving all their advantages. Over the past three decades, there has been a tremendous growth in image acquisition capabilities for various biomedical and life science applications, including MRI, CT, PET, microscopy (Balchandani and Naidich, 2015; Esquivel et al., 2022; Badawi et al., 2019; Gambarotto et al., 2019; Wassie et al., 2019). Ultra-high resolution imaging technology has enabled acquisition of images beyond three orders of magnitude larger than macroscopic biomedical domains (Kleven et al., 2023; Wang et al., 2020b; Mansour et al., 2025; Kleinfeld et al., 2011). For instance, a typical clinical 1mm brain MRI scan registration requires solving for $\sim$ 20M parameters, while a high-resolution *ex-vivo* human brain scan requires solving upto 11B parameters (more than 500 times larger). However, current approaches work reliably only at the scale of *macroscopic* biomedical domains ($\sim$ 50M warp parameters) and quickly run out of memory on larger problems due to high computational and memory requirements. We observed this behavior both in Chapter 2 where VFA ran out of memory for the Ultracortex dataset at 0.6mm isotropic resolution, and in Chapter 3 where deep learning methods typically run out of memory at $\sim 1.8\times$ higher resolution per dimension than 1mm isotropic images. In context, neuropathology typically operates at few microns per pixel resolution for histology (Mancini et al., 2020), and at $\sim 100 - 300\mu m$ for *ex-vivo* brain scans (Edlow et al., 2019; Khandelwal et al., 2025). This leads to a significant gap in the ability to accurately register images at high resolutions.

Surge in distributed frameworks for large model training Recent years have also witnessed tremendous innovations in large-scale transformer model training, including IO-aware fused operations to minimize latency and large memory overheads (Dao et al., 2022; Dao, 2023; Spector et al., 2025), 5D parallelism to distribute larger-than-memory models and inputs into multi-GPU/node setups (Shoeybi et al., 2019; Li et al., 2023a; Jacobs et al., 2024; Li et al., 2024; Zhao et al., 2023; Ansel et al., 2024). Although most existing techniques are specialized for generalized matrix multiplication (GEMM) like operations only, the fundamental concepts utilized by these methods (IO-awareness, recomputing and aggregating intermediates on shared memory to minimize high bandwidth memory (HBM) storage, identifying partial aggregates across hosts to minimize communication overheads for distributed optimization) are broadly applicable to a wide class of problems of the non-GEMM nature.

In this chapter, we transfer these concepts to scale image registration algorithms to match parity with the developments in both increasing resolution of image acquisition *and* compute capabilities. To that end, the contributions in this chapter are twofold. First, we identify key compute and memory bottlenecks in image registration algorithms, and propose novel components that fit problems upto $64\times$ larger than existing algorithms on a single GPU. Second, we propose **F**lash **F**used **D**istributed **P**rimitives (FFDP), a distributed framework to scale registration to an arbitrary number of GPUs, thereby scaling to ultra high-resolution problems. We present a first-of-its-kind demonstration: aligning a $250\mu m$ in-vivo MRI (Lüsebrink et al., 2017) to a $100\mu m$ *ex-vivo* human brain FLASH volume (Edlow et al., 2019) – a multimodal registration problem more than **570**$\times$ larger than a standard clinical datum (Marcus et al., 2007a), with over 11.8B transform parameters – completed in *one minute* using only 8 A6000 GPUs. FFDP accelerates existing traditional registration pipelines by upto $7.48\times$ while reducing memory consumption by upto 59%, and deep

learning pipelines by upto $6.14\times$ while consuming upto 24% less memory. We highlight the necessity of performing high-resolution registration by comparing our method with various SOTA optimization and deep learning baselines on a $250\mu m$ T1-weighted MRI dataset, showing unprecedented performance and gains in efficiency.

5.1. Fused Kernels for Memory Efficient Registration on a Single GPU

We first propose efficient designs to fit larger problems on a single GPU, and then extend the framework to distributed registration.

Bottlenecks of a deformable image registration pipeline Our primary objective is to identify compute and memory bottlenecks in *large-scale* image matching tasks. In identifying these bottlenecks, training-free optimization methods are better suited than deep networks since the latter has a much larger activation memory footprint, which forms the primary memory bottleneck (Nouamane et al., 2025). For instance, for a $250\mu m$ image pair, a standard deep learning method (Hoffmann et al., 2021) generates an activation map of size 27GB only after the first layer. Extrapolating memory usage for clinical data, existing deep networks will require upto 1.2TB of GPU memory at inference to process these image volumes at native resolution. In contrast, a training-free optimizer can fit this problem in less than 45GB of GPU memory. We use FireANTs (Jena et al., 2026) as our base framework to identify compute and memory bottlenecks in a typical image registration problem.

Memory hierarchy of a GPU A GPU's memory hierarchy spans multiple tiers: registers (per-thread, single-cycle), shared memory/L1 cache (on-chip, tens of KB, low latency within a block), L2 cache (MBs, shared across SMs, moderate latency), and global memory (**HBM**). Our work focuses on reducing HBM usage for key non-GEMM operations used in image registration, by maximizing register and shared memory usage while minimizing global memory traffic.

Memory overhead analysis of a typical clinical MRI registration task Flamegraphs are visualization tools used to represent the memory or compute profile of an algorithm over time, revealing which functions or operations consume the most resources. Flamegraphs are typically useful to track large memory allocations, lifecycle of important variables, debug memory leaks and OOMs. We analyze the flamegraph of a typical clinical MRI registration task from the *OASIS* brain dataset (Marcus et al., 2007a) in Fig. 5.1. We identify three key memory bottlenecks in image matching pipelines (1) deformable interpolation and warp composition (2) cross-correlation loss, and (3) mutual information loss (see Fig. 5.2(right)).

5.2. Composite Implicit Grid Sampler

A fundamental operation used in image registration is the *grid sampler*. This operator allows us to warp an image M using a deformation field $\varphi : \Omega \to \Omega$ and computes the image $M' : M'(x) = M(\varphi(x))$. Virtually every image registration pipeline uses this operation to warp the moving image using an affine, deformable, or composite transform. For affine and composite transforms, the operator initializes a regular grid $[x]_\Omega$, a grid of size $3N$. The affine grid $A[x]_\Omega + t$ is another grid of size $3N$. If a deformable grid $[u]_\Omega$ is optimized, then a third grid $A[x]_\Omega + t + [u]_\Omega$ is materialized, costing a total of $9N$ overhead for an image of size N. To consolidate these memory overheads, we propose a composite implicit grid sampler. This is a fused CUDA kernel that performs the following operation:

$$\texttt{fused\_grid\_sampler}(I; A, t, [u], S, x_{\text{bounds}})(x) = I(Ax + t + Su(x))$$

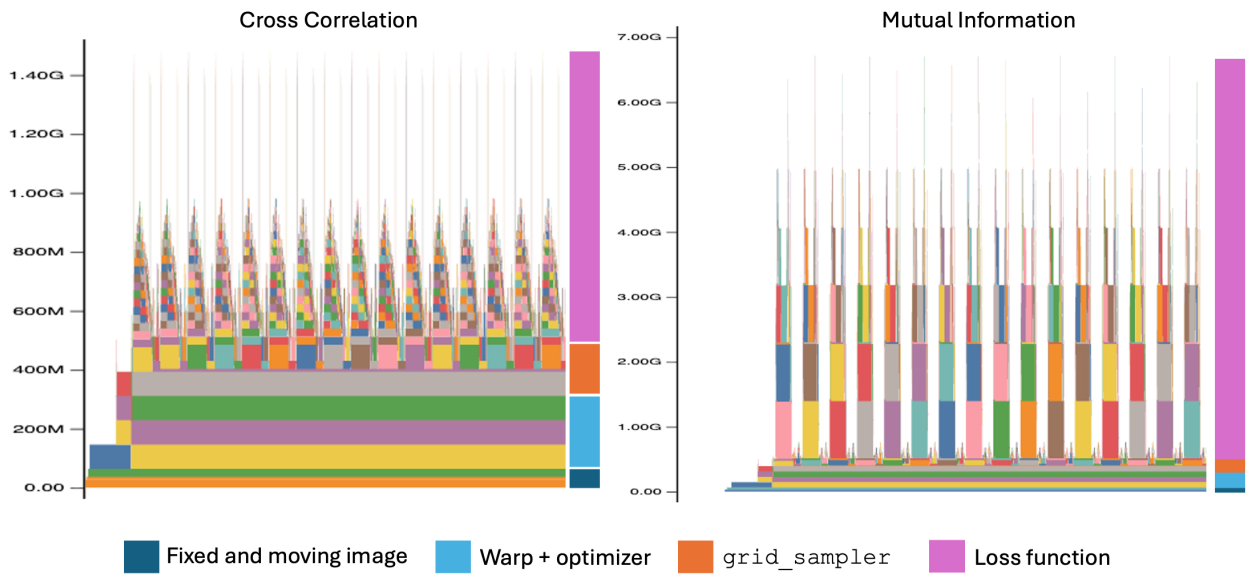


Figure 5.1: Flamegraph of FireANTs for Cross Correlation (left) and Mutual Information (right) losses on the OASIS dataset. The flamegraph is annotated on the right with colored blocks denoting the memory overheads for the fixed and moving images, the warp field and its optimizer state, the `grid_sampler` operation, and the loss function. Most of the computational overhead is due to the loss function, followed by the `grid_sampler` operation. This motivates the use of fused kernels to eliminate intermediate memory overheads.

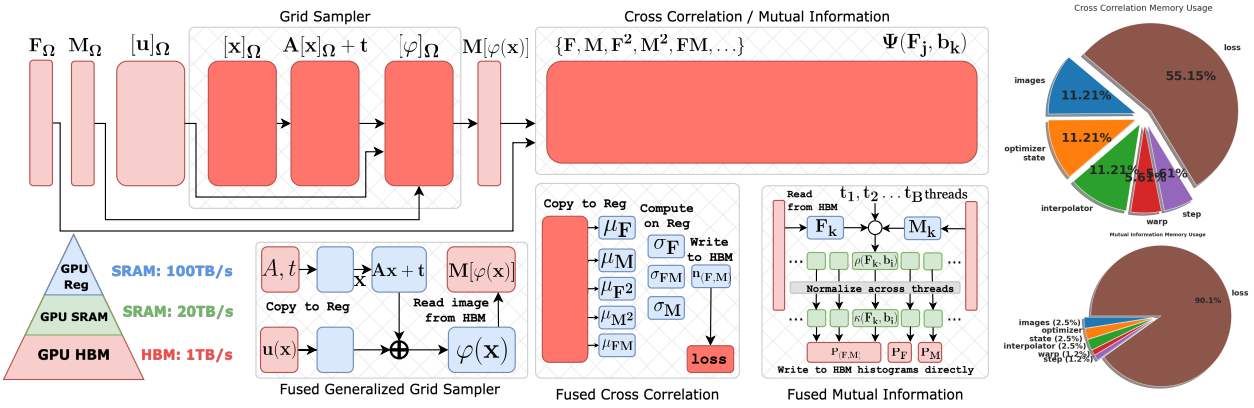


Figure 5.2: Left: FFDP uses fused kernels to eliminate intermediate HBM memory usage (in **dark red**) for memory-bound workhorse operations (`grid_sampler`, LNCC, MI) for large-scale image registration. For `grid_sampler` and LNCC, additional intermediate per-pixel variables (warp coordinates, patchwise statistics) are computed per-pixel in registers (blue). For MI, the Parzen Windowing and histogram aggregation is performed using shared memory (green), avoiding large HBM overheads. **Right**: Pie charts show the breakdown of memory overheads for storing the image, grid, optimizer state, and intermediate variables for MI and LNCC losses.

where $A, S \in GL(d, \mathbb{R})$ are affine matrices, t is a translation vector, $[u]$ is the deformation grid, and x_{bounds} are the bounds of the (implicit) identity grid $[x]_\Omega$. There are three benefits of this approach. First, the kernel avoids materializing any additional grids in HBM, reducing the memory overhead of the kernel from $O(n)$ to $O(1)$ with no loss in runtime or accuracy. Second, when the warp $[u]_\Omega$ is sharded across hosts in a distributed

Algorithm 1 Grid Sampler Implementation

Require: J_h (moving image shard), $[u]_j$ (warp field shard), A_h (rescaled affine), t_h (rescaled translation), S_h (diag. scale)

function FORWARD($J_h,\ A_h,\ t_h,\ S_h,\ [u]_j$)
  `out` $\leftarrow$ `zeros_like`($[u]_j[0]$)
  for all target voxels (z, y, x) **in parallel (one thread per voxel) do**
    $X \leftarrow (x, y, z)$
    $X_{\text{aff}} \leftarrow A_h X + t_h$ ▷ *affine transform only*
    $X_{\text{disp}} \leftarrow S_h\,[u]_j[:, z, y, x]$ ▷ *add scaled displacement*
    $X_{\text{src}} \leftarrow X_{\text{aff}} + X_{\text{disp}}$
    `out`$[z, y, x] \leftarrow$ `trilinear_interpolate`$(J_h,\ X_{\text{src}})$ *zero padding at bounds*
  end for
  return `out`
end function

function BACKWARD($g = \frac{\partial \mathcal{L}}{\partial \texttt{out}}, J_h,\ A_h,\ t_h,\ S_h,\ [u]_j$)
  Initialize $g_{J_h} = 0,\ g_{[u]_j} = 0,\ g_{A_h} = 0,\ g_{t_h} = 0$
  for all target voxels (z, y, x) **in parallel (one thread per voxel) do**
    Recompute $X,\ X_{\text{aff}},\ X_{\text{disp}},\ X_{\text{src}}$
    Compute tri-linear weights $w_{b_x b_y b_z}$ and $\frac{\partial v}{\partial X_{\text{src}}}$
    Accumulate g_{J_h} into 8 neighbors using $w_{***} \cdot g[z, y, x]$ *(bounds-checked, zero-padded)*
    $g_{[u]_j}[:, z, y, x]$ += $S_h\ \frac{\partial v}{\partial X_{\text{src}}}\ g[z, y, x]$
    g_{A_h} += $\left(\frac{\partial v}{\partial X_{\text{src}}}\ g[z, y, x]\right) X^\top$
    g_{t_h} += $\frac{\partial v}{\partial X_{\text{src}}}\ g[z, y, x]$
  end for
  return $g_{J_h},\ g_{[u]_j},\ g_{A_h},\ g_{t_h}$
end function

setting, the identity grid $[x]_\Omega$ needs to be sharded correctly too. Since the identity grid is implicitly defined by its bounds $x_{\text{bounds}} = (x_{\min}, x_{\max}) \in \mathbb{R}^{2d}$, our implementation can be easily used in a distributed optimization setting without instantiating partial shards $[x]_{\Omega_h}$. Finally, the matrix S is used to rescale the deformation field to sample from the coordinates of the sharded images I_h which lie on the grid Ω_h instead of Ω (see Section 5.7.2) without initializing additional memory. The backward pass is very similar to the existing PyTorch implementation, with the exception of the gradient of the affine matrix. We discuss the derivation and pseudocode of the forward and backward pass in Algorithm 1 below.

5.3. Implicit Parzen Windowing for Mutual Information

Mattes Mutual Information (MI) is one of the most commonly used loss functions for *multimodal* image matching (Chen et al., 2022b; Avants et al., 2009; Mattes et al., 2001). For random variables X and Y, MI is the KL divergence between the joint distribution $P(X, Y)$ and product of marginal distributions $P(X)P(Y)$ of the intensities of the two images. For image matching, X and Y are the pixel intensities for the images

I, J. The distributions are estimated using a kernel density estimator:

$$P_I(v) = \frac{1}{N}\sum_k \kappa(v - I_k), \quad P_{(I,J)}(v, w) = \frac{1}{N}\sum_k \kappa(v - I_k)\kappa(w - J_k) \tag{5.1}$$

where κ is a kernel function of choice. Common choices of κ are the Gaussian (Guo, 2019) and 3rd order B-Spline kernels (Thévenaz and Unser, 2000).

Vanilla MI implementation To empirically compute the KL divergence, the distributions Eq. (5.1) are discretized over B equally spaced bins on the domain of $u \in I, v \in J$. To compute the discrete PMF with autodifferentiation, we compute a Parzen Block $\Psi_I \in \mathbb{R}^{B\times N}$, s.t.$\Psi_I(i, k) = \kappa(b_i - I_k)$. This forms the memory bottleneck in computing the Mattes MI similarity criteria. In the following, we provide a fused implementation that avoids the $O(NB)$ cost of the Parzen Block, making our implementation only $O(1)$ additional HBM overhead. However, to compute the joint histogram of size B^2, this method requires materializing the entire Parzen Block $\Psi_I(j, k) = \kappa(b_j - I_k)$ of size $2k_PBN$, where k_P is a kernel-dependent constant. Since $N >> B$ (B is typically chosen to be 32), this operation becomes a significant memory bottleneck for large N. For instance, a typical clinical image volume ($N \approx$ 30MB) with 32 bins will consume **7.5GB** of HBM - a significantly huge cost that grows much faster for larger problems.

5.3.1. Implicit MI implementation

We implement custom forward and backward passes to compute the joint and marginal histograms p_{IJ}, p_I, p_J from I and J directly, avoiding the $O(NB)$ cost of the Parzen Block. We derive the backward pass first, followed by the forward pass followed by an efficient approximate estimator of the histograms leading to a faster forward pass.

5.3.2. Backward pass

We are interested in computing the gradients $\frac{\partial L}{\partial I}, \frac{\partial L}{\partial J}$ given $\frac{\partial L}{\partial p_{IJ}}, \frac{\partial L}{\partial p_I}, \frac{\partial L}{\partial p_J}$. We denote $\omega(b_i - I_k) = \frac{\partial \kappa(b_i - I_k)}{\partial I_k}$.

$$\frac{\partial L}{\partial I_k} = \sum_{m,n} \frac{\partial L}{\partial p_{IJ}[m, n]} \frac{\partial p_{IJ}[m, n]}{\partial I_k} + \sum_n \frac{\partial L}{\partial p_I[n]} \frac{\partial p_I[n]}{\partial I_k} \tag{5.2}$$

$$= \sum_{m,n} g_{IJ}[m, n] \left(\omega(b_m - I_k)\kappa(b_n - J_k)\right) + \sum_n g_I[n] \left(\omega(b_n - I_k)\right) \tag{5.3}$$

$$= \sum_n \left[g_I[n]\omega(b_n - I_k) + \sum_m g_{IJ}[m, n]\omega(b_m - I_k) \right] = \sum_n \zeta_1[n] + \zeta_2[n] \tag{5.4}$$

where $\zeta_1[n] = g_I[n]\omega(b_n - I_k)$ and $\zeta_2[n] = \sum_m g_{IJ}[m, n]\omega(b_m - I_k)$.

To compute this backward pass efficiently, we launch $\lceil N/B \rceil$ threadblocks and partition each threadblock in groups of B threads, and compute the partial gradients $\zeta_1[n], \zeta_2[n]$ on each thread. Each group loads the values of I_k, J_k into register memory. we first compute the quantities $\kappa(b_n - I_k), \kappa(b_n - J_k), \omega(b_n - I_k), \omega(b_n - J_k)$ on thread n and use four shared memory arrays to store them. On thread n, we compute the partial gradient $\zeta_1[n] = g_I[n]\omega(b_n - I_k)$ and $\zeta_2[n] = \sum_m g_{IJ}[m, n]\omega(b_m - I_k)$ using a for-loop over the index $m \in$

$\{1, \ldots, B\}$. Finally, on each thread we store the value $\zeta_1[n] + \zeta_2[n]$ on shared memory indexed at n, followed by a $O(\log(n))$ parallel sum over partitioned threads to compute the gradient $\frac{\partial L}{\partial I_k} = \sum_n \zeta_1[n] + \zeta_2[n]$. A similar argument is used to compute the gradient over $\frac{\partial L}{\partial J_k}$. This leads to a faster backward pass than the vanilla PyTorch implementation using no additional HBM overhead Fig. 5.10(b).

Generalization to novel kernels Note that unlike the vanilla implementation, where some choices of κ are more memory intensive than others (for example, the BSpline kernel has $k_P = 14$ versus $k_P = 4$ for the Gaussian kernel), the memory overhead of our implementation does not depend on the analytical form of κ. To generalize the Implicit MI implementation to novel kernels, the user can specify the form of κ and its derivative ω in the forward and backward passes without any additional considerations.

5.3.3. Forward Pass

The forward pass is computed similarly. Note that the individual contributions from I_k, J_k to the joint histogram $p_{IJ}[m, n]$ are $p_{IJ}[m, n] = \kappa(b_m - I_k)\kappa(b_n - J_k)$ for all $m, n \in \{1, \ldots, B\}$. The marginal histograms $p_I[n], p_J[n]$ are computed as $p_I[n] = \kappa(b_n - I_k)$ and $p_J[n] = \kappa(b_n - J_k)$ for all $n \in \{1, \ldots, B\}$. Similar to the backward pass, we launch $\lceil N/B \rceil$ threadblocks and partition the threadblock in groups of B threads. Each group of B threads loads the values of I_k, J_k into register memory. On thread n, we compute the quantities $\kappa(b_n - I_k), \kappa(b_n - J_k)$ and store them in shared memory. Thread n can add these quantities into the HBM for histogram entries $p_I[n], p_J[n]$ directly. For computing the joint histogram $p_{IJ}[m, n]$, thread n loops over $m \in \{1, \ldots, B\}$ and adds the quantities $\kappa(b_m - I_k)\kappa(b_n - J_k)$ into the HBM for histogram entries $p_{IJ}[m, n]$. Since all values of $\kappa(b_m - I_k), \kappa(b_n - J_k)$ are stored in shared memory, this operation is not bottlenecked by slow HBM reads. To avoid HBM write contentions, we write these values into intermediate histogram buffers of sizes $C \times B \times B, C \times B$ (where C is a constant of choice), and sum along the C dimension. However, this is still a relatively slow operation due to computation of $\kappa(b_m - I_k), \kappa(b_n - J_k)$ and making NB^2 HBM writes. We propose an efficient approximate forward pass that launches only N instead of NB threads, and makes only $3N$ HBM writes.

An approximate histogram estimator Given a kernel κ, we can write $\kappa(b_m - I_k) = \int_t \delta(b_m - I_k - t)\kappa(t)dt = \delta(b_m - I_k) * \kappa$, where δ is the Dirac delta function with the property $\int_{x=-\infty}^{\infty} \delta(x)f(x)dx = f(0)$ for any function f. Using the principle of superposition, we can write $p_I[m] = \frac{1}{N}\sum_k \kappa(b_m - I_k) = \frac{1}{N}\sum_k \kappa * \delta(b_m - I_k) = \kappa * \left(\frac{1}{N}\sum_k \delta(b_m - I_k)\right)$.

In the continuous case, p_I can be obtained *exactly* by calculating the Dirac delta distribution $p_I^{\delta}(b) = \frac{1}{N}\sum_k \delta(b - I_k)$ and convolving it with the kernel κ. However, in the discrete case, this value is inexact. To see this, consider a value I_k that is in bin m, i.e. $\|I_k - b_m\| < \frac{1}{2B}$. The exact value of the PMF due to this sample is $\kappa(b_m - I_k)$. However, the approximate value of the PMF is $\kappa(0)$ since $\delta(b_m - I_k) = 1$ for all $I_k : \|I_k - b_m\| < \frac{1}{2B}$ due to binning, and convolving with κ returns $\kappa(0)$. Since $\|I_k - b_m\| < \frac{1}{2B}$, we can assume that $\|\kappa(0) - \kappa(b_m - I_k)\|$ is small.

To implement this histogram computation efficiently, we launch N threads and in each thread k, compute the bin indices $m^* = \lfloor I_k B \rfloor, n^* = \lfloor J_k B \rfloor$ for each thread, avoiding computation of *soft entries* $\kappa(b_m - I_k), \kappa(b_n - J_k)$ altogether. We simply add 1 to the histogram entries $p_{IJ}[m^*, n^*], p_I[m^*], p_J[n^*]$ in the aggregated histogram buffers, avoiding writing into HBM entries for all $(m, n) \in \{1, \ldots, B\}^2$. This reduces the number of HBM writes from $NB^2 + 2NB$ to $3N$. For $B = 32$, this represents $362\times$ less HBM writes. After performing the average, we convolve this histogram with the kernel κ to get the approximate PMF. Since the convolution is done on a B and $B \times B$ sized histograms, this operation is cheap. This implementation leads to faster

runtime, consistent performance for both TransMorph and FireANTs (see Table 5.2).

5.4. Efficient Implicit Fused Cross-Correlation

Local Normalized Cross-Correlation (LNCC) is used ubiquitously in signal and image processing as a similarity metric. In deformable image registration, it is used as a robust similarity function to compare anatomical similarities (Chen et al., 2022b; Hoffmann et al., 2021; Avants et al., 2008b; Wu et al., 2024).

Definition of LNCC loss. Given two images F and M, and a radially symmetric averaging convolution filter W such that $\sum_k w_k = 1$, we define the Local Normalized Cross Correlation (LNCC) loss as:

$$\mathcal{L} = \frac{1}{N}\sum_i n_i \quad , \quad n_i = \frac{A_i^2}{B_i C_i + \epsilon} \tag{5.5}$$

where

$$\mu_i^F, \mu_i^M = \sum_k w_{ik} F_k, \sum_k w_{ik} M_k \tag{5.6}$$

$$A_i = \sum_k w_{ik}(F_k - \mu_i^F)(M_k - \mu_i^M) \tag{5.7}$$

$$B_i = \sum_k w_{ik}(F_k - \mu_i^F)^2 \tag{5.8}$$

$$C_i = \sum_k w_{ik}(M_k - \mu_i^M)^2 \tag{5.9}$$

Here, we use overloaded notation $w_{ik} = w_{(i-k)} = w_{(k-i)} = w_{ki}$ due to radial symmetry of w. We can expand Eqs. (5.7) to (5.9) as follows:

$$A_i = \left(\sum_k w_{ik} F_k M_k\right) - \mu_i^F \mu_i^M = \mu_i^{FM} - \mu_i^F \mu_i^M \tag{5.10}$$

$$B_i = \left(\sum_k w_{ik} F_k^2\right) - (\mu_i^F)^2 = \mu_i^{F^2} - (\mu_i^F)^2 \tag{5.11}$$

$$C_i = \left(\sum_k w_{ik} M_k^2\right) - (\mu_i^M)^2 = \mu_i^{M^2} - (\mu_i^M)^2 \tag{5.12}$$

Algorithm 2 outlines a vanilla PyTorch implementation of the LNCC loss function. The computational overhead of the algorithm arises due to many intermediates stored in high-bandwidth memory (HBM). Specifically, the quantities $W * \texttt{state}, I^2, J^2, IJ, \sigma_I^2, \sigma_J^2, \sigma_{IJ}, \mu_I^2, \mu_J^2, \mu_I\mu_J, \sigma_I^2\sigma_J^2, (\sigma_I^2\sigma_J^2 + \epsilon), \sigma_{IJ}^2, \sigma_{IJ}^2/(\sigma_I^2\sigma_J^2 + \epsilon)$ are all stored as intermediate tensors, each of size N, totalling a $16N$ memory overhead in addition to storing `state`. The computational graph of the vanilla PyTorch implementation is shown in Fig. 5.3. During the backward pass, the backprop algorithm computes the gradient with respect to each of these variables costing an additional $16N$ memory overhead. A `torch.compile` implementation fuses some of the arithmetic, but

Algorithm 2 Vanilla PyTorch LNCC implementation

Require: F (input image), M (reference image), w (window size), $reduction$ (reduction type)

Define a radially symmetric convolution filter W of size $w \times w \times w$ with $\sum W[i] = 1$
Define $\texttt{state} = (F, M, F^2, M^2, FM)$ ▷ Elementwise operations
Compute $\texttt{state} = W * \texttt{state}$ ▷ Convolution
Get $\mu_F = \texttt{state}[0]$ ▷ Local mean of F
Get $\mu_M = \texttt{state}[1]$ ▷ Local mean of M
Compute $\sigma_F^2 = \texttt{state}[2] - \mu_F^2$ ▷ Local variance of F
Compute $\sigma_M^2 = \texttt{state}[3] - \mu_M^2$ ▷ Local variance of M
Compute $\sigma_{FM} = \texttt{state}[4] - \mu_F \cdot \mu_M$ ▷ Local covariance of F and M
Compute LNCC $= \frac{\sigma_{FM}^2}{\sigma_F^2 \sigma_M^2 + \epsilon}$ ▷ Add small ϵ to avoid division by zero
if `reduction` == `NONE` **then**
    return LNCC
else
    Compute loss: Loss $= 1 -$ mean(LNCC)
    return Loss
end if

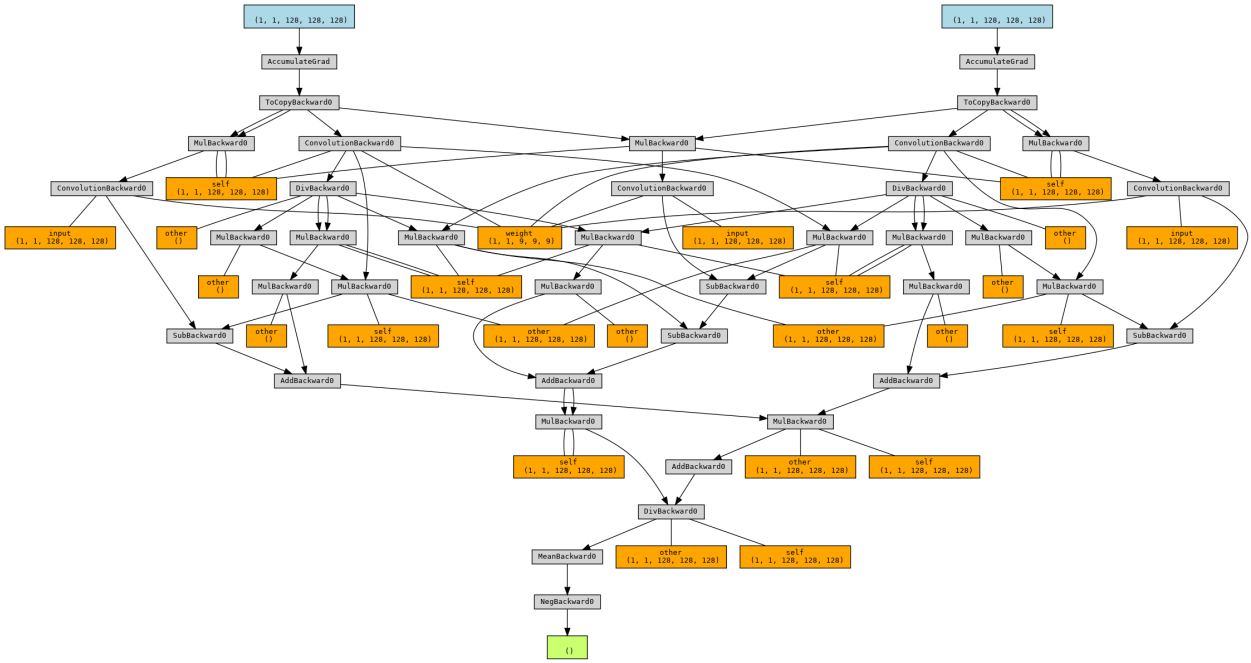


Figure 5.3: Computational graph of the vanilla PyTorch implementation of the LNCC loss function. Blue nodes denote the input images, Orange nodes denote intermediate tensors that are stored in HBM, Gray nodes denote operations on the computational graph, and Green node denotes the final loss. Orange nodes are the primary memory bottleneck.

leaves a lot of room for improvement (see Fig. 5.10). We present an algorithm that only requires an additional intermediate variable `state` of size $5N$, saving upto $27N$ memory.

5.4.1. An efficient fused LNCC implementation

During the forward pass, we initialize a `state` variable of size $5N$. To minimize HBM reads from F and M, we write a fused kernel to initialize the `state` variable using only one HBM read from F and M. The code in Line 4-9 are elementwise operations, and can be fused into another kernel. The forward pass therefore consumes only $5N$ additional memory. The pseudocode for the efficient fused LNCC implementation is shown in Algorithm 3.

Efficient Backward Pass In a vanilla PyTorch implementation, the gradients are computed for each intermediate variables in the reverse order in the computational DAG shown in Fig. 5.3. Typically, our implementation would also require defining the backward pass by computing the gradients with respect to the intermediate variables, and then propagating them to the input images. However, we derive the backpropagation with respect to I and J, given the gradient $g_i = \frac{\partial L}{\partial n_i}$ to avoid calculating intermediate gradients. Using the chain rule, we have:

$$\frac{\partial L}{\partial F_k} = \sum_i \frac{\partial L}{\partial n_i}\frac{\partial n_i}{\partial F_k} \tag{5.13}$$

$$= \sum_i g_i \left(\frac{2A_i}{B_i C_i}\frac{\partial A_i}{\partial F_k} - \frac{A_i}{B_i^2 C_i}\frac{\partial B_i}{\partial F_k} \right) \tag{5.14}$$

$$\tag{5.15}$$

which can be simplified to:

$$\frac{\partial \mu_i^F}{\partial F_k} = \frac{\partial \mu_i^M}{\partial M_k} = w_{ik} \tag{5.16}$$

$$\frac{\partial A_i}{\partial F_k} = \frac{\partial \left(\sum_k w_{ik} F_k M_k - \mu_i^F \mu_i^M \right)}{\partial F_k} = w_{ik}\left(M_k - \mu_i^M \right) \tag{5.17}$$

and

$$\frac{\partial B_i}{\partial F_k} = \frac{\partial \left(\sum_k w_{ik} F_k^2 - (\mu_i^F)^2 \right)}{\partial F_k} = 2w_{ik}(F_k - \mu_i^F) \tag{5.18}$$

Substituting these results to Eq. (5.14) we have:

$$= \sum_i g_i \left(\frac{2A_i}{B_i C_i}\left(w_{ik}\left(M_k - \mu_i^M \right) \right) - \frac{A_i^2}{B_i^2 C_i} 2w_{ik}(F_k - \mu_i^F) \right) \tag{5.19}$$

$$= \sum_i \frac{2g_i A_i}{B_i C_i} w_{ik} \left[M_k - \frac{F_k A_i}{B_i} + \mu_i^F \frac{A_i}{B_i} - \mu_i^M \right] \tag{5.20}$$

Using the property $w_{ik} = w_{ki}$, and letting $\gamma_i = \frac{2g_iA_i}{B_iC_i}$, we rewrite the previous equation as:

$$= M_k \cdot \left(\sum_i w_{ki}\gamma_i\right) - F_k \cdot \left(\sum_i w_{ki}\frac{\gamma_i A_i}{B_i}\right) + \sum_i w_{ki}\gamma_i\left(\frac{\mu_i^F A_i}{B_i} - \mu_i^M\right) \tag{5.21}$$

$$= M_k \cdot (w * \gamma)_k - F_k \cdot (w * \gamma_{AB})_k + (w * \gamma_{FM})_k \tag{5.22}$$

where $\gamma_{AB} = \gamma_i\frac{\mu_i^F A_i}{B_i}$, $\gamma_{FM} = \gamma_i \cdot \left(\frac{\mu_i^F A_i}{B_i} - \mu_i^M\right)$ - and $*$ is the convolution operation. Similarly, the gradient with respect to the moving image M_k is:

$$\frac{\partial L}{\partial M_k} = F_k\left(\sum_i w_{ki}\gamma_i\right) - M_k\left(\sum_i w_{ki}\frac{\gamma_i A_i}{C_i}\right) + \sum_i w_{ki}\gamma_i\left(\frac{\mu_i^M A_i}{C_i} - \mu_i^F\right) \tag{5.23}$$

$$= F_k \cdot (w * \gamma)_k - M_k \cdot (w * \gamma_{AC})_k + (w * \gamma_{MF})_k \tag{5.24}$$

where $\gamma_{AC} = \gamma_i\frac{\mu_i^M A_i}{C_i}$, $\gamma_{MF} = \gamma_i \cdot \left(\frac{\mu_i^M A_i}{C_i} - \mu_i^F\right)$. To compute the gradients with respect to F and M, we need to compute five tensors of the γ family, namely $\gamma, \gamma_{AB}, \gamma_{AC}, \gamma_{FM}$, and γ_{MF}. This is followed by performing a convolution with all the tensors, and computing elementwise operations given by Eq. (5.22) and Eq. (5.24). The γ family of tensors are simple elementwise operations on the `state` variable, and therefore can be computed by modifying the `state` variable *inplace* to avoid initializing additional HBM memory.

ANTs gradient approximation. In the ANTs implementation, the gradient computation skips performing the convolution of the γ family of tensors. We implement this as an additional flag that the user can toggle as an option for faster backward passes. All our experiments use this approximation.

Algorithm 3 Fused LNCC Implementation

Require: F (fixed image), M (moving image), w (window size), ϵ (smoothing term)

function FORWARD(F, M, w, ϵ)
  Define convolution filter W of size $w \times w \times w$ with $\sum W[i] = 1$
  `state` $\leftarrow$ `fused_create_interm`(F, M) ▷ Single HBM read: (F, M, F^2, M^2, FM)
  `state` $\leftarrow W *$ `state` ▷ Convolution on all channels
  LNCC $\leftarrow$ `fusedcc_kernel`(`state`, ϵ) ▷ Computes Eqs. (5.10) to (5.12) followed by Eq. (5.5)
  return LNCC
end function

function BACKWARD($g = \frac{\partial \mathcal{L}}{\partial n}$, `state`, F, M, W, use_ants_approximation)
  `state` $\leftarrow$ `fused_compute_gamma`$(g,$ `state`$)$ ▷ Computes γ family of tensors *inplace*
  if use_ants_approximation **then**
    `no-op` ▷ ANTs approximation: skip convolutions
  else
    `state` $\leftarrow W *$ `state` ▷ Convolution on all intermediates
  end if
  $\frac{\partial L}{\partial F} \leftarrow M \odot \gamma - F \odot \gamma_{AB} + \gamma_{FM}$ ▷ Eq. (5.22) computed in fused kernel
  $\frac{\partial L}{\partial M} \leftarrow F \odot \gamma - M \odot \gamma_{AC} + \gamma_{MF}$ ▷ Eq. (5.24) computed in fused kernel
  return $\frac{\partial L}{\partial F}, \frac{\partial L}{\partial M}$
end function

5.4.2. Performance

We compare the performance of our fused implementation to various backend implementations. Fig. 5.10 shows the speedup and memory usage over different image sizes; we tabulate the results here. For this experiment, we initialize two random images of size $N_v \times N_v \times N_v$ and compute the runtime and memory usage for the forward and backward passes. Results are in Table 5.1. Our implementation consistently achieves upto $6\times$ forward time speedup and $\sim 98\times$ backward time speedup compared to (Jia et al., 2025) and consumes upto 76% less memory than a compiled PyTorch implementation and 61.9% less than a groupwise convolution implementation (Jia et al., 2025).

5.5. Extending image registration to multiple GPUs

Our composite implicit grid sampler and improved loss functions allows optimizing problems with image sizes that are upto two magnitudes larger than other baselines on a single A6000 GPU (Fig. 5.11a). However, many applications using mesoscopic and microscopic data require registration of images that do not fit on a single GPU. Inspired by distributed frameworks for LLM training (Shoeybi et al., 2019; Rajbhandari et al., 2020) and initial work on distributed image registration (Mang et al., 2019a), we propose a distributed framework that allows sharding large images across multiple GPUs to efficiently scale to arbitrarily large problem sizes with any similarity loss function.

Distributed Setting. For distributed registration with H hosts or GPUs, we partition the domain $P(\Omega) = \{\Omega_1, \Omega_2, \ldots \Omega_H\}$ such that $|\Omega_i| = N/H$, $\Omega_i \cap \Omega_j = \phi \quad \forall i \neq j$ and $\cup_i \Omega_i = \Omega$. We use $[x]_{\Omega_h}$, $A[x]_{\Omega_h} + t$, and $[u]_{\Omega_h}$ to denote the sharded tensors defined on domain Ω_h.

5.6. Grid Parallel for Boundary-Synchronized Image Sharding

Techniques like Tensor/Sequence/Expert/Context Parallel have been tremendously successful in distributed optimization by sharding large models and sequences across multiple GPUs (Shoeybi et al., 2019; Li et al., 2023a; Liu et al., 2024b,a). However, these techniques work for transformer-like architectures and input sequences where the model parameters and activations do not require boundary synchronization. In contrast, image registration contains operations that require boundary synchronization between image and grid shards to perform mathematically correct convolutions. Examples of such operations include convolutions for calculating LNCC, total variation loss, Sobolev norm of the gradient and warp fields (Mang et al., 2019a; Avants et al., 2008b; Beg et al., 2005).

To enable these functionalities and complement existing parallelism techniques, we propose '*Grid Parallel*' (GP) as an abstraction on a tensor. GP shards a tensor across hosts, stores the sharded dimension and bounds as metadata, and provides synchronization operations to augment the tensor with sufficient boundary padding from neighboring shards prior to performing a convolution operation. GP allows us to partition the fixed images, $[u]$, and the optimizer state $[m_1], [m_2]$ – essentially sharding the entire problem across H hosts while allowing the user to apply convolutional operations seamlessly. We compare the performance of GP with naive DTensor sharding in Fig. D.1.

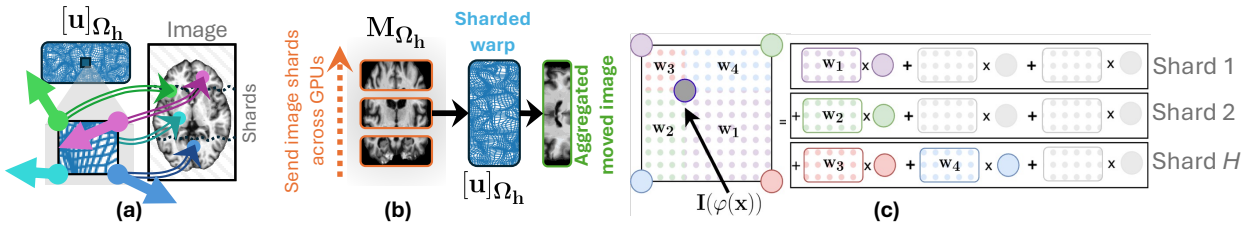


Figure 5.4: **(a)** Neighboring coordinates in the warp field may refer to pixel locations on arbitrary image shards due to the deformable nature of the warp field, making distributed interpolation non-trivial. **(b)** Ring Sampler interleaves fetching of image shards and aggregating the partial sums of interpolated values, avoiding a memory-expensive allgather. **(c)** Bilinear Interpolation is decomposed into partial sums over image shards, which are accumulated with a ring topology communication, similar to Liu et al. (2024b).

5.7. Distributed Ring Sampler

Despite the sharding in GP, the moving image M cannot be naively sharded across GPUs due to the random-access nature of the `grid_sample` operation applied on M. In general, the warp vector $\varphi(x)$ residing on GPU i can point to coordinates that reside on the sharded image on GPU j for any $j \neq i$. Even for neighboring coordinates $x_s, x_u \in [x]_i$, the coordinates $\varphi(x_s)$ and $\varphi(x_u)$ can point to different shards $j_1 \neq j_2 \neq i$. This is illustrated in Fig. 5.4(a). Keeping the entire moving image in memory limits the maximum problem size to $N \leq V$, where V is the memory per GPU, regardless of the number of hosts H. However, we want the maximum problem size to scale with H. Therefore, we propose a distributed `grid_sampler` that allows us to *correctly* interpolate the moving image with sharded images scattered across multiple hosts without performing an `allgather` operation on the moving image.

Our approach leverages the key observation that (bi/tri)linear interpolation can be decomposed as an aggregate of partial sums of interpolated values on individual image shards. Fig. 5.4(b) illustrates this example. These individual image shards are sent across hosts in a ring topology, similar to Liu et al. (2024b), and the partial sum is aggregated inplace. This operation only incurs an additional N/H HBM overhead for fetching the sharded image from other hosts, scaling efficiently to arbitrary large problem sizes for sufficiently large H.

5.7.1. Derivation

Consider a d-linear interpolation of an image I defined on Ω using warp coordinates $[u]_\Omega$ defined on Ω.

$$I = \sum_{b \in \{0,1\}^n} \left(\prod_{k=1}^{n} (1-\alpha_k)^{1-b_k} \, \alpha_k^{b_k} \right) I\big[i_1 + b_1,\ i_2 + b_2,\ \ldots,\ i_n + b_n\big] \tag{5.25}$$

where $i_k = \lfloor \varphi(x)_k \rfloor, \qquad \alpha_k = \varphi(x)_k - i_k, \qquad$ for $k = 1, \ldots, d$. Let the individual pixels $I\big[i_1 + b_1,\ i_2 + b_2,\ \ldots,\ i_n + b_n\big]$ be partitioned across H hosts. Since each pixel belongs to exactly one host, we can write $\sum_{h=1}^{H} \mathbb{I}(i + b \in [x]_h) = 1$ and multiply with $I[i + b]$ to get:

$$I = \sum_{b\in\{0,1\}^n} \left(\prod_{k=1}^{n} (1-\alpha_k)^{1-b_k}\, \alpha_k^{b_k}\right) \left(I\big[i+b\big] * \left(\sum_{h=1}^{H} \mathbb{I}(i+b \in [x]_h)\right)\right) \tag{5.26}$$

$$= \sum_{h=1}^{H} \sum_{b\in\{0,1\}^n} \left(\prod_{k=1}^{n} (1-\alpha_k)^{1-b_k}\, \alpha_k^{b_k}\right) \left(I\big[i+b\big] * \mathbb{I}(i+b \in [x]_h)\right) \tag{5.27}$$

$$= \sum_{h=1}^{H} I_h \tag{5.28}$$

where

$$I_h = \sum_{b\in\{0,1\}^n} \left(\prod_{k=1}^{n} (1-\alpha_k)^{1-b_k}\, \alpha_k^{b_k}\right) I\big[i+b\big] * \mathbb{I}(i+b \in [x]_h) \tag{5.29}$$

$$= \sum_{b\in\{0,1\}^n} \left(\prod_{k=1}^{n} (1-\alpha_k)^{1-b_k}\, \alpha_k^{b_k}\right) J_h\big[i+b\big] \tag{5.30}$$

where $J_h\big[x\big] = I[x]$ if $x \in [x]_h$ else 0. Image J_h is therefore *identical* to the sharded image I on host h. Eq. (5.30) refers to performing trilinear interpolation on the shard I_h (with zero padding) since the sum is only over coordinates that reside in $[x]_h$. This means the warped image in Eq. (5.25) can be obtained by performing interpolation over the shards individually and adding the warped images together. This is illustrated in Fig. 5.4(c). Coordinates residing between multiple shards will accumulate partial sums from each sharded image, and no additional consideration is needed for boundary conditions. The communication protocol in this algorithm is similar to Ring Attention (Liu et al., 2024b), where image shards are passed across hosts, and partial results are accumulated into the final result. Our algorithm requires a memory overhead of only N/H to store the sharded image from host $j \neq i$. Our pseudocode is provided in Algorithm 4.

5.7.2. Implementation Considerations

Rescaling the warp function to sample sharded images Interpolating from sharded images requires one additional consideration. The grid sampler interpolates an image I defined on Ω using warp coordinates $[u]_\Omega$ defined on Ω. However, the sharded image J_h is defined on the domain Ω_h, and therefore any warped coordinate $\varphi(x) \in \Omega$ must be rescaled to the corresponding coordinates in $\varphi_h(x) \in \Omega_h$. From the implementation standpoint, the leftmost coordinate of J_h is x^h_{min} when the entire image I is passed to `grid_sampler`. However, when J_h is provided as input to `grid_sampler`, the leftmost pixel of J_h is located at $[-1, -1, \ldots, -1]$ according to PyTorch convention. Since our optimization variables $A, t, [u]$ refer to locations on Ω, and not Ω_h, we need to rescale these variables appropriately when sampling from J_h.

The rescaling corresponds to a diagonal scaling matrix S_h and translation t_h such that $S_h x^h_{\text{min}} + t_h = x^\Omega_{\text{min}}$ and $S_h x^h_{\text{max}} + t_h = x^\Omega_{\text{max}}$. The resampled warp function to sample from J_h becomes $\varphi_h(x) = S_h(Ax + t + u(x)) + t_h = (A'_h x + t'_h) + S_h u(x)$. where $A'_h, t'_h = S_h A, (S_h t + t_h)$. Therefore, we must sample J_h using the transform $A'_h[x]_{\Omega_h} + t'_h + S_h[u]_{\Omega_h}$. In the vanilla grid sampler implementation, the intermediate grid $S_h[u]_{\Omega_h}$ and its gradient consume another $6N/H$ memory. Combined with the N/H overhead for storing the

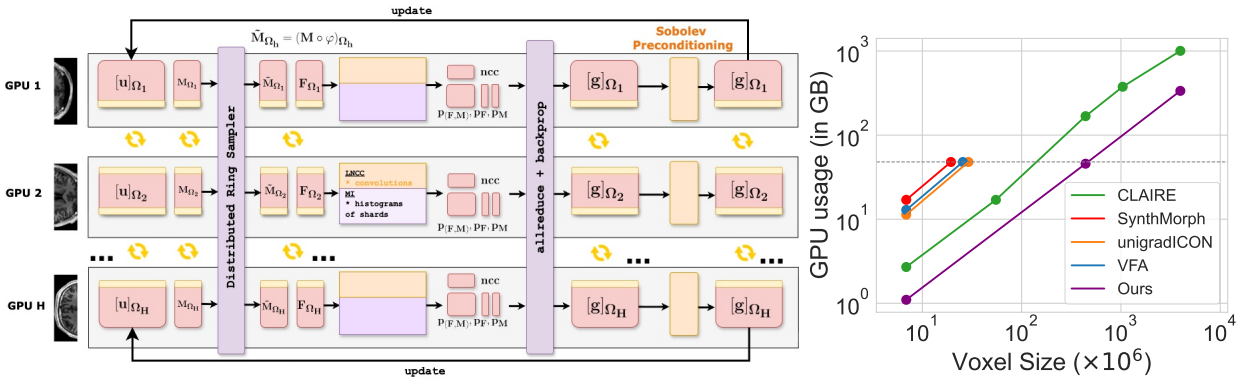


Figure 5.5: Left: Overview of our distributed framework. GridParallel (GP) shards the fixed and moving images (F, M) and the warp field $[u]$ across multiple GPUs. Yellow blocks and arrows denote synchronized halo boundaries between GPUs, enabling smoothing on images and warp fields without an allgather. The ring sampler (violet) computes interpolated image shards on the fly, avoiding materialization of the full moving image. We then compute losses (MSE, LNCC, MI), compute gradients w.r.t. each warp shard, apply Sobolev regularization with GP, and update shards by gradient descent. **Right**: Scaling efficiency compared to deep methods and CLAIRE (Mang et al., 2019a), a distributed registration method. Most SOTA deep learning baselines require orders-of-magnitude more memory for the same problem size and scalability is limited to a single GPU (dotted line). Our framework scales to arbitrarily large problem sizes while using about $5\times$ less memory than CLAIRE.

received image shard, we add a total of $7N/H$ memory overhead, which is less than N for $H \geq 8$, making the algorithm impractical for fewer GPUs (say $H = 4$).

To prevent this $6N/H$ additional overhead, we extend the generalized grid sampler as mentioned Section 5.2 to sample from a transform of the form $A[x] + t + S[u]$ directly. This computes the value $Su(x)$ directly inside the CUDA kernel, and the backward pass also computes and accumulates the gradient w.r.t. $u(x)$ directly, avoiding the $6N/H$ overhead.

Interleaved communication An important implementation detail is the interleaving of communication and computation in the ring sampler. While we compute the partial moved image aggregate, the next image shard can be fetched asynchronously in the background. This is illustrated in Fig. 5.6.

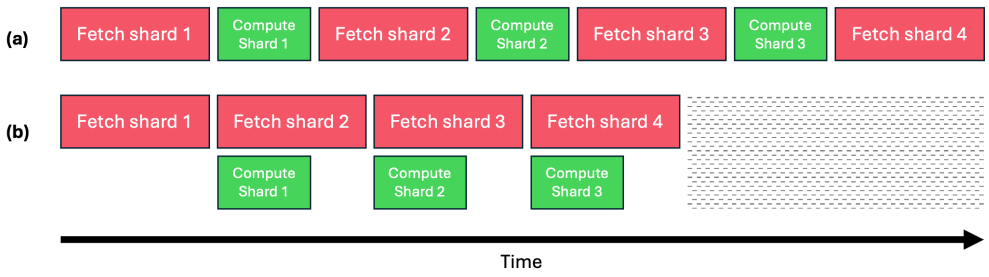


Figure 5.6: Interleaved communication (red) and computation (green) in the ring sampler. gray denotes time saved by interleaving communication and computation.

Algorithm 4 Ring Sampler Implementation

Require: M_j (moving image shard), $[u]_j$ (warp field shard), (A, t) (affine transform)

function FORWARD(M_j, $[u]_j$, (A, t))
  Define $\texttt{moved}_j = 0$
  for $h = 1$ **to** H **do**
    $J_h \leftarrow \texttt{send\_and\_recv}(M_j, h)$ ▷ Send and receive the image shard from offset h
    Compute diagonal S_h, t_h such that $S_h x^h_{\text{min}} + t_h = x^\Omega_{\text{min}}$ and $S_h x^h_{\text{max}} + t_h = x^\Omega_{\text{max}}$
    Rescale affine transform $A_h \leftarrow S_h A, t_h \leftarrow S_h t + t_h$
    $\texttt{moved}_j \leftarrow \texttt{moved}_j + \texttt{grid\_sampler}(J_h; A_h, t_h, S_h, [u]_j)$ ▷ Avoid computing $S_h[u]_j$ explicitly
  end for
  return $\texttt{moved}_j$
end function

function BACKWARD($g = \frac{\partial \mathcal{L}}{\partial \texttt{moved}_j}$, $\texttt{moved}_j$, M_j, $[u]_j$, (A, t))
  Define $g_{[u]_j} = 0, g_A = 0, g_t = 0, g_{M_j} = 0$
  for $h = 1$ **to** H **do**
    $J_h \leftarrow \texttt{send\_and\_recv}(M_j, h)$ ▷ Send and receive the image shard from offset h
    Compute diagonal S_h, t_h such that $S_h x^h_{\text{min}} + t_h = x^\Omega_{\text{min}}$ and $S_h x^h_{\text{max}} + t_h = x^\Omega_{\text{max}}$
    Rescale affine transform $A_h \leftarrow S_h A, t_h \leftarrow S_h t + t_h$
    if $\texttt{requires\_grad}(M_j)$ **then**
      $g_{inp} \leftarrow \texttt{zeros\_like}(M_j)$
    else
      $g_{inp} \leftarrow \texttt{None}$
    end if
    Compute $\texttt{backward\_grid\_sampler}(g, J_h, A_h, t_h, S_h, [u]_j, g_{[u]_j}, g_A, g_t, g_{inp})$
    if $\texttt{requires\_grad}(M_j)$ **then**
      $g'_{M_j} = \texttt{send\_and\_recv}(g_{inp}, -h)$
      $g_{M_j} \leftarrow g_{M_j} + g'_{M_j}$
    end if
  end for
  return $g_{[u]_j}, g_A, g_t, g_{M_j}$
end function

5.7.3. Alternative Designs for Distributed Interpolation

A naive approach can be to route the coordinate $\varphi(x_i) \in [x]_j$ to GPU j and retrieve the image coordinate, similar to routing tokens using expert parallelism (EP) used for Mixture-of-Experts (MoEs) (Shazeer et al., 2017; Jordan and Jacobs, 1994). However, this approach has two major drawbacks in our setting. First, due to the deformable nature of φ, the partitioning of coordinates across hosts is generally uneven. In the worst case, a single GPU can receive all $3N$ coordinates leading to an indirect `allgather` operation resulting in OOMs or uneven GPU utilization across hosts. Second, coordinates that point to regions between two multiple image boundaries need to be sent to variable number of hosts, which is non-trivial to implement. These two factors make both the forward and backward pass implementations cumbersome. Inspired by (Liu et al., 2024b), we proposed the distributed ring sampler in the previous section, that decomposes the computation into partial sums, leading to a simple implementation without degraded scaling performance Fig. 5.9.

5.8. Distributed Loss Functions

Since the moved image and fixed image are sharded cross H hosts, the loss function must take this into account to compute the loss function correctly.

Mean Squared Error (MSE). Since MSE is a per-pixel loss, we compute the individual MSE on host h and perform an `allreduce` operation.

Localized Normalized Cross Correlation (LNCC). The LNCC computes per-pixel patch similarities for each pixel, using a convolution over its neighbors. For sharded images, the patch statistics at the boundary requires a boundary synchronization with its neighboring shards which is provided by our GP implementation. After computing the LNCC for all pixels in each shard, we perform another `allreduce` to compute the LNCC over the entire image.

Mutual Information (MI). The MI loss computes the joint histograms $p_{(I,J)}(x, y)$ and marginals $p_I(x), p_J(y)$. However, these distributions are partial aggregates from the sharded images on each GPU. Eq. (5.1) can be rewritten as $p_I(v) = \sum_h \frac{N_h}{N} \left(\frac{1}{N_h} \sum_{k \in \Omega_h} \kappa(v - I_k) \right), p_{IJ}(v, w) = \sum_h \frac{N_h}{N} \left(\frac{1}{N_h} \sum_{k \in \Omega_h} \kappa(v - I_k)\kappa(w - J_k) \right)$, where the red terms correspond to the per-host histogram computation. Performing an `allreduce` to compute the weighted average of these histograms (with weights N_h/N) results in a valid and correct joint and marginal distributions over all hosts. This also leads to only a $B^2 + 2B$ communication overhead regardless of N, making a distributed implementation highly practical.

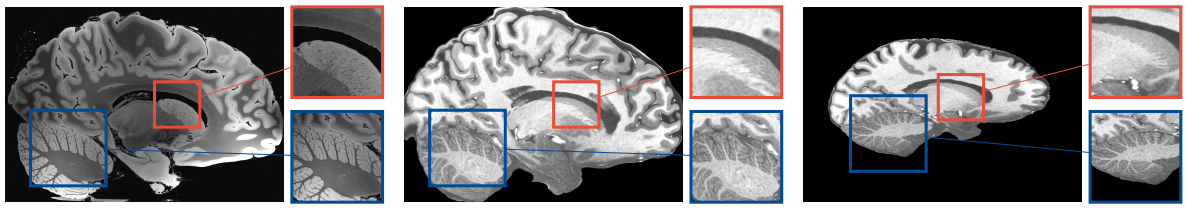


Figure 5.7: Qualitative comparison on registration of 100μm ex-vivo brain MRI (T1 $\rightarrow$ FLASH) image. Fine details like cerebellar white matter are not visible at macroscopic scales, but are aligned at 100μm. Fixed image is of size $1760 \times 1760 \times 1278$. Best viewed zoomed in.

5.9. Experiments

Our primary goals are to (a) accelerate both optimization and neural network based registration workflows, and (b) solve significantly larger image registration problems. We show the efficacy of our method by accelerating existing registration workflows on standard clinical data. This is followed by optimizing a multimodal registration task with more than 11.8B optimizable parameters, an unprecedented result in large-scale registration. We compare the performance and computational efficiency of our method with various state-of-the-art baselines on a simulated 250μm ex-vivo brain MRI dataset, followed by ablations on various components of our framework.

Baselines. To accelerate existing registration workflows, we compare against TransMorph (Chen et al., 2022b) and FireANTs (Jena et al., 2026), which are state-of-the-art deep learning and optimization based registration frameworks respectively. In addition, we perform comparative evaluation with two methods

Table 5.1: Speedup and memory usage of different LNCC backends

N	Method	Forward Time (s)	Forward Speedup	Backward Time (s)	Backward Speedup	Memory (MB)	Memory Reduction (%)
64	Fast LNCC	0.001	2.95	0.003	4.86	21	61.9
	FireANTs	0.003	7.18	0.002	3.07	25	68
	VoxelMorph	0.06	158.76	0.016	24.10	17	52.9
	`torch.compile`	0.003	6.83	0.002	2.30	24	66.7
	Ours	< 0.001	1.00	0.001	1.00	8	0
128	Fast LNCC	0.008	5.88	0.026	34.09	168	61.9
	FireANTs	0.013	9.04	0.008	10.73	200	68
	VoxelMorph	0.482	341.65	0.126	168.33	136	52.9
	`torch.compile`	0.012	8.67	0.007	8.95	192	66.7
	Ours	0.001	1.00	0.001	1.00	64	0
256	Fast LNCC	0.069	6.19	0.204	82.52	1344	61.9
	FireANTs	0.103	9.25	0.294	118.80	2176	76.5
	VoxelMorph	3.905	351.54	3.903	1577.37	1536.2	66.7
	`torch.compile`	0.1	9.02	0.284	114.74	2176	76.5
	Ours	0.011	1.00x	0.002	1.00x	512	0
512	Fast LNCC	0.627	6.56	1.657	98.75	10752	61.9
	FireANTs	0.856	8.95	2.396	142.77	17408	76.5
	VoxelMorph	31.335	327.71	31.665	1887.14	12288.2	66.7
	`torch.compile`	0.829	8.67	2.312	137.80	17408	76.5
	Ours	0.096	1.00	0.017	1.00	4096	0

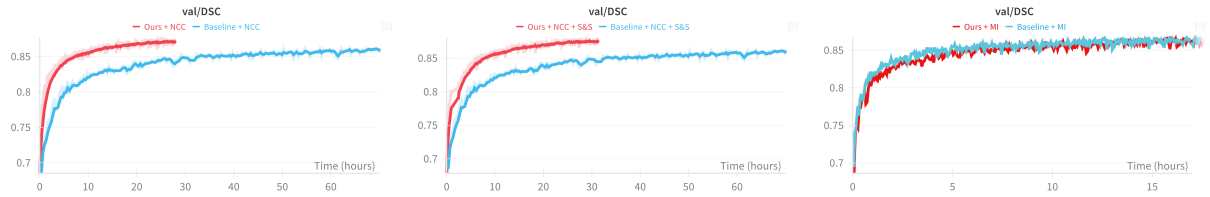


Figure 5.8: Ablation on TransMorph training runtime with and without our fused operations. For LNCC, our method converges in about 30 hours, while the baseline converges in about a week.

explicitly designed for large-scale registration: ITK-DReg (itk) (CPU-based) and CLAIRE (Mang et al., 2019a) (multi-GPU), and several SOTA learning-based approaches for clinical data - SynthMorph (Hoffmann et al., 2021), Vector-Field Attention (Liu et al., 2024c), unigradICON (Tian et al., 2024) (with/without instance optimization), anatomix+ConvexAdam (Dey et al., 2025).

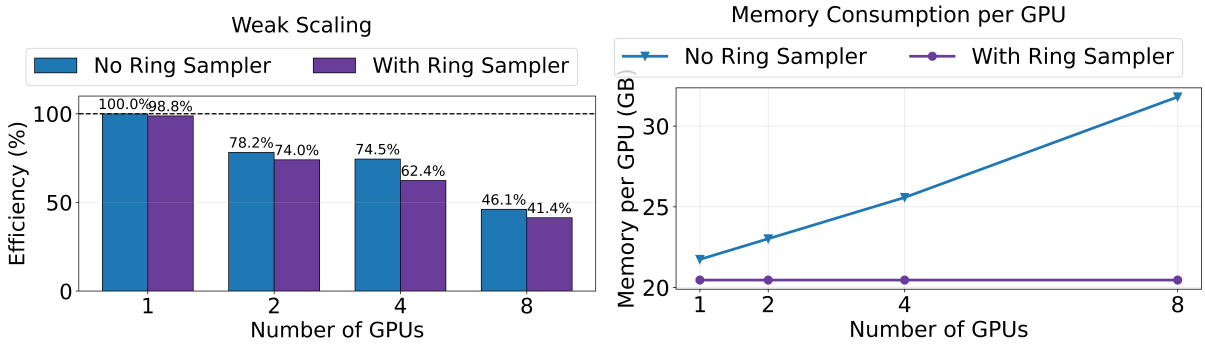
Table 5.2: Accelerating TransMorph (**Top**) and FireANTs (**Bottom**) training with various computation backends.

Variant	Loss	Diffeomorphic	Training Time (h)	GPU Mem (GB)	Val DSC
Baseline	LNCC	✗	171.20	20.01	86.74
Ours	LNCC	✗	27.84	16.95	87.23
Baseline	LNCC	✓	171.42	21.28	86.55
Ours	LNCC	✓	27.93	17.34	87.09
Baseline	MI	✗	26.09	22.34	86.74
Ours	MI	✗	24.94	16.80	86.80

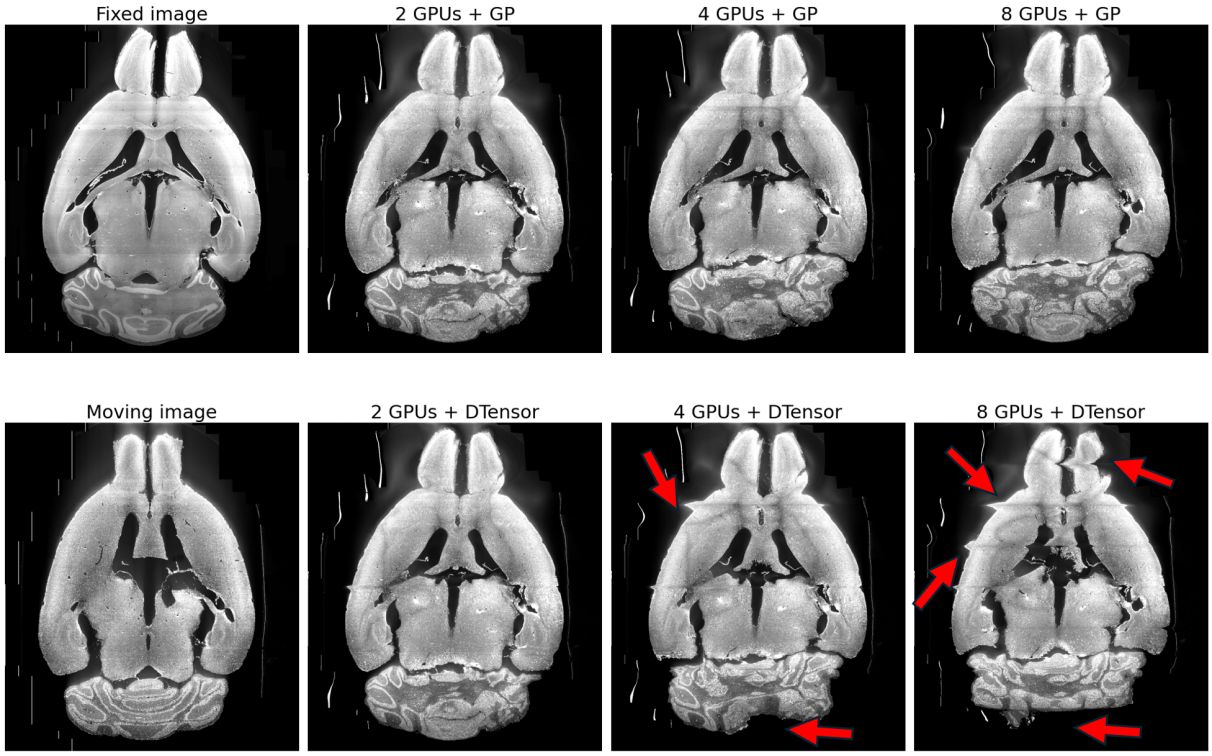
Loss	Backend	Dice Score ↑	Runtime (s) ↓	Memory (MB) ↓
LNCC	FireANTs	78.81 ± 3.87	1.44 ± 0.08	1044.5 ± 0.0
LNCC	FastLNCC	76.96 ± 3.60	3.76 ± 0.16	1026.3 ± 0.0
LNCC	VXM/TM	76.96 ± 3.60	57.08 ± 2.45	1418.5 ± 0.0
LNCC	`torch.compile`	69.35 ± 4.09	0.82 ± 0.04	860.7 ± 0.0
LNCC	Ours	78.67 ± 3.04	0.50 ± 0.01	577.5 ± 0.0
MI	PyTorch	75.88 ± 3.45	7.51 ± 0.37	12206.3 ± 0.0
MI	`torch.compile`	75.88 ± 3.45	1.05 ± 0.05	3865.5 ± 0.0
MI	Ours	75.88 ± 3.44	2.90 ± 0.16	577.5 ± 0.0
MI	`torch.compile`+Ours	75.93 ± 3.47	2.95 ± 0.16	657.3 ± 0.0

Table 5.3: Extended Results on accelerated registration on FireANTs: Accelerating FireANTs registration with various computation backends and registration algorithms (Greedy and SyN). Our implementations maintain accuracy while substantially reducing runtime and peak memory usage. (Green)/ (Yellow) = best/second; Speedup and memory reduction are computed with respect to our kernels. Our fused kernels maintain accuracy while substantially reducing runtime and peak memory usage.

Algorithm	Method	Backend	Dice Score ↑	Runtime (s) ↓	Memory (MB) ↓	Speedup ↑	Mem. Reduction (%) ↑
Greedy	LNCC	VXM/TM	76.96 ± 3.60	57.08 ± 2.45	1418.5 ± 0.0	113.47	59.29
	LNCC	FastLNCC	76.96 ± 3.60	3.76 ± 0.16	1026.3 ± 0.0	7.48	43.73
	LNCC	FireANTs	72.81 ± 3.87	1.44 ± 0.08	1044.5 ± 0.0	2.87	44.71
	LNCC	`torch.compile`	69.35 ± 4.09	0.82 ± 0.04	860.7 ± 0.0	1.63	32.90
	LNCC	Ours	78.67 ± 3.04	0.50 ± 0.01	577.5 ± 0.0	1.00	0.00
Greedy	MI	PyTorch	75.88 ± 3.45	7.51 ± 0.37	12206.3 ± 0.0	2.59	95.27
	MI	`torch.compile`	75.88 ± 3.45	1.05 ± 0.05	3865.5 ± 0.0	0.36	85.06
	MI	Ours	75.87 ± 3.44	2.90 ± 0.16	577.5 ± 0.0	1.00	0.00
	MI	Ours + `torch.compile`	75.93 ± 3.47	2.95 ± 0.16	657.3 ± 0.0	1.02	12.13
SyN	LNCC	VXM/TM	76.69 ± 2.88	63.57 ± 0.58	1892.0 ± 0.0	65.92	50.05
	LNCC	FastLNCC	76.70 ± 2.88	4.27 ± 0.05	1486.7 ± 0.0	4.43	36.43
	LNCC	FireANTs	74.70 ± 2.93	2.55 ± 0.10	1616.4 ± 0.0	2.65	41.54
	LNCC	`torch.compile`	71.65 ± 3.41	1.46 ± 0.04	1472.0 ± 0.0	1.51	35.80
	LNCC	Ours	78.79 ± 2.82	0.96 ± 0.08	945.0 ± 0.0	1.00	0.00
SyN	MI	PyTorch	76.74 ± 2.58	12.84 ± 0.66	17720.8 ± 0.0	2.96	94.67
	MI	`torch.compile`	76.76 ± 2.58	2.40 ± 0.13	7758.9 ± 0.0	0.55	87.82
	MI	Ours	76.86 ± 2.59	4.34 ± 0.28	945.0 ± 0.0	1.00	0.00
	MI	Ours + `torch.compile`	77.00 ± 2.57	4.56 ± 0.24	1104.5 ± 0.0	1.05	14.44



(a) Weak scaling and Per-GPU memory consumption of FFDP.



(b) Qualitative ablation of GP synchronization in FFDP on the fMOST mouse brain dataset (Tustison et al., 2024). Red arrows highlight regions affected by incorrect boundary effects due to no GP. See Fig. D.1 for more examples.

Figure 5.9: Scaling and GP ablations.

5.9.1. Accelerating existing registration workflows and ablations

For deep networks, we train `TransMorph-large` under three loss configurations: (a) LNCC+Dice, (b) MI+Dice, and (c) LNCC+scaling-and-squaring (Ashburner, 2007) +Dice. For classical optimization, we benchmark runtime and memory against multiple LNCC backends (FireANTs, VoxelMorph/TransMorph, Fast LNCC, `torch.compile`, and Ours) and MI backends (PyTorch and Ours with and without `torch.compile`). Tables 5.2 and 5.3 and Fig. 5.8 show that during network training our kernels converge 6.1× faster with LNCC while using 16.5% less memory, and reduce MI memory usage by 24.7%. Despite being designed

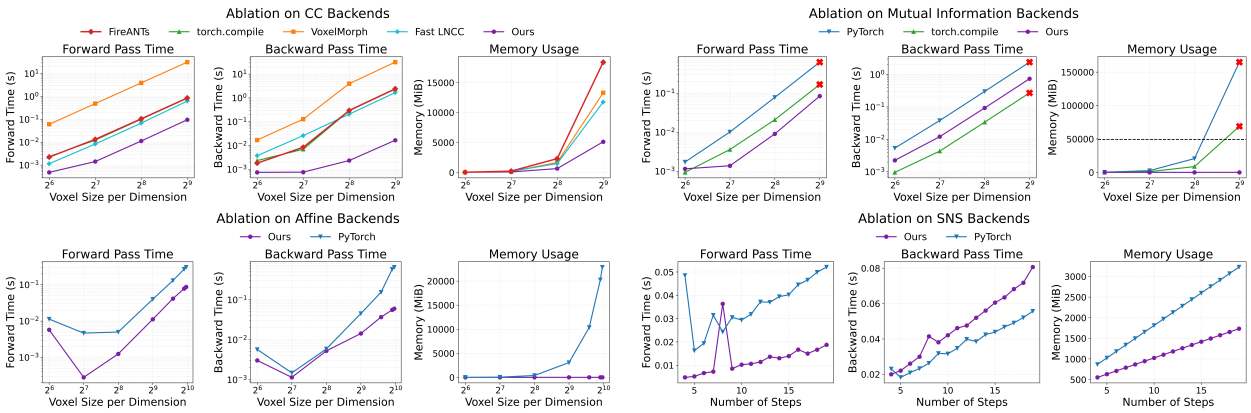


Figure 5.10: Ablations on key workhorse operations: LNCC, MI, `grid_sampler`, and scaling-and-squaring operations. Our fused kernels consume significantly less HBM and runtime.

for very large images, the runtime and memory benefits are significant for clinical-scale data (i.e., 30MB for OASIS). Optimization frameworks see larger gains: FireANTs achieves up to 95.2% memory savings and 2.6× speedup with MI, and a 7.5× speedup over FastLNCC (Jia et al., 2025) (and 2.9× over FireANTs' LNCC backend which applies separable convolutions on FastLNCC), with 44-59% lower memory usage overall.

5.9.2. Registration to a 100 micron ex-vivo brain MRI volume

To showcase the efficacy of our method on real large scale images, we register a $250\mu m$ in-vivo MRI image (Lüsebrink et al., 2017) to a 100 μm *ex-vivo* FLASH human brain volume (Edlow et al., 2019). This represents an inverse problem with more than 11.2B optimizable parameters (compared to ∼20M for clinical datasets), or 44.8GB of GPU memory. The entire problem does not fit on most GPUs, necessitating distributed multimodal registration. We optimize a composite transform - affine followed by a diffeomorphic mapping. Multimodal deformable registration took ∼58 seconds on 8 NVIDIA A6000 GPUs, which is unprecedented at this resolution. Fig. 5.7 shows qualitative results, highlighting the ability to register highly detailed structures such as cerebellar white matter; these structures are not visible at macroscopic scales. The resultant advantages of performing registration at this scale can allow researchers to characterize the neuroanatomy at microscopic resolutions and allow morphometric analysis of cortical layers and subcortical nuclei among other structures.

Registration accuracy in these studies is measured using privately annotated fiducial markers, hindering reproducibility and comparability of methodological advances. Due to lack of scalable frameworks, most high-resolution studies simply run ANTs at a significantly downsampled resolution (Kleven et al., 2023; Mansour et al., 2025; Wang et al., 2020b; Kronman et al., 2024; Bogovic et al., 2020; Edlow et al., 2019) and upsample the warp field to the native resolution.

5.9.3. Comparative Analysis on a Simulated ex-vivo Brain MRI Dataset

The faux-OASIS dataset To compare registration performance at high resolutions and leverage existing methods as baselines, we synthesize the *faux-OASIS* dataset, which mimics the anatomical distribution of an MRI dataset at $250\mu m$ isotropic resolution. At $250\mu m$, the deformation field has 1.32B degrees of freedom per image pair, compared to ∼20M for OASIS.

Baselines and evaluation. All methods (including CLAIRE and FireANTs without FFDP) run out of memory at $250\mu m$ resolution. We proposed two modifications to deep learning based methods to enable them to work on this dataset: (a) inspired by several high-resolution studies (Wang et al., 2020b; Mansour et al., 2025; Edlow et al., 2019), we register the images at a downsampled resolution, and then upsample the deformation field (b) inspired by several histology registration methods (Wodzinski et al., 2024; Lotz et al., 2015; Liang et al., 2021), we perform patchwise registration and mosaicing of the final deformation. We compare the methods at three resolutions: 1mm, $500\mu m$, and $250\mu m$. At 1mm, the full image fits within a patch, providing a baseline reference comparable to reported OASIS performance. At higher resolutions, patches are defined by each method's default input size with stride equal to 50% of the patch size. FireANTs augmented with FFDP is denoted as *Ours*. We report Dice, inverse-weighted Dice (InvDice; Mang et al. (2019a)), and average Haussdorf distance capped at 90 percentile (AvgHD90). To compare efficiency, we measure both wall-clock time and GPU-hours.

Results. Fig. 5.11a summarizes performance metrics. At 1mm, most methods achieve performance consistent with their reported performance on OASIS, including VFA and TransMorph which were trained on the OASIS dataset with label supervision. At higher resolutions, nearly all methods degrade, especially for InvDice and HD90, which emphasize alignment of fine structures. In contrast, our method improves in accuracy: at $250\mu m$, we improve Dice by 18.1 points, InvDice by 31.6 points, and reduce AvgHD90 by 62.1%. The correlation between resolution and performance is also observed in (Mang et al., 2019a; Mang and Ruthotto, 2017a; Nazib et al., 2018); in addition we verify that patch-based methods *degrade* in performance at higher resolutions.

This degradation among patchwise methods is expected; histology-style pipelines typically register consecutive slides with small deformations after affine alignment. At high resolution, patching reduces anatomical context and the patches become progressively more out-of-distribution. Patchwise or downsampling strategies are therefore insufficient for ultra-high resolution large-scale registration, and existing deep methods cannot be repurposed to work at higher resolutions efficiently. Accuracy-efficiency tradeoffs in Figs. 5.11b and 5.11c show that our method is Pareto-efficient compared to all other methods (CPU, deep learning, and distributed GPU methods), requiring up to $500\times$ fewer GPU-hours compared to alternatives at $250\mu m$.

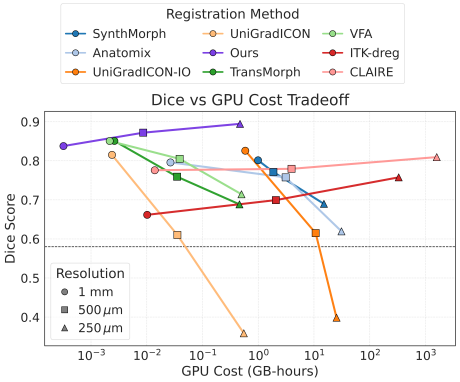
5.9.4. Ablation Studies

We ablate on the efficiency of various workhorse operations used in image registration in Fig. 5.10 and Table 5.1. We compare our implementations to community-standard PyTorch implementation (Jia et al., 2025; Chen et al., 2022b) and `torch.compile` versions. For grid sampler and MI kernels, our kernels have $O(1)$ extra HBM overhead instead of $O(N)$ in the PyTorch implementation. For LNCC, our implementation achieves an average speedup in the forward pass by $5.22\times$ and $56.98\times$ in the backward pass. Our `grid_sampler` also leads to an efficient scaling-and-squaring operation, commonly used in deep learning registration pipelines (Chen et al., 2022b), with a memory reduction of 50% compared to the baseline implementation.

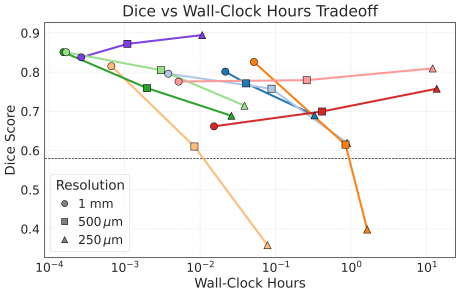
Scalability Analysis. We test the weak scaling of our distributed framework by registering synthetic images with increasing voxel sizes. For H GPUs, we instantiate an image pair of size $700 \times 700 \times 700H$ and shard the images, warp, and optimizer state across H GPUs. Fig. 5.9a shows weak scaling of FFDP with and without ring sampler. Without the ring sampler, the `grid_sample` operation requires storing the moving image of size $700 \times 700 \times 700H$ on each GPU, leading to peak HBM memory increasing linearly with H. This implies the framework would not scale to arbitrarily large problem sizes, regardless of cluster size H. Peak Memory

(a) Performance comparison across methods and resolutions.

Resolution	Method	AvgDice Score ↑	InvDice Score ↑	AvgHD90^{cum} (mm) ↓
1 mm	Baseline	0.579 ± 0.055	0.141 ± 0.142	1.587 ± 0.908
	Anatomix	0.796 ± 0.035	0.386 ± 0.138	0.468 ± 0.137
	CLAIRE	0.776 ± 0.044	0.344 ± 0.120	0.554 ± 0.150
	ITK-dreg	0.662 ± 0.055	0.199 ± 0.125	1.002 ± 0.277
	SynthMorph	0.801 ± 0.022	0.378 ± 0.133	0.455 ± 0.098
	TransMorph	0.851 ± 0.016	0.468 ± 0.161	0.310 ± 0.064
	UniGradICON (IO)	0.826 ± 0.022	0.391 ± 0.155	0.384 ± 0.095
	UniGradICON	0.815 ± 0.026	0.393 ± 0.156	0.419 ± 0.113
	VFA	0.851 ± 0.023	0.494 ± 0.169	0.323 ± 0.096
	Ours	0.838 ± 0.028	0.436 ± 0.148	0.341 ± 0.109
500 µm	Baseline	0.580 ± 0.055	0.138 ± 0.143	1.357 ± 0.326
	Anatomix†	0.758 ± 0.040	0.325 ± 0.159	0.619 ± 0.169
	CLAIRE	0.779 ± 0.051	0.275 ± 0.210	0.570 ± 0.211
	ITK-dreg	0.699 ± 0.056	0.240 ± 0.130	0.834 ± 0.254
	SynthMorph†	0.771 ± 0.035	0.337 ± 0.133	0.557 ± 0.144
	TransMorph†	0.759 ± 0.028	0.300 ± 0.175	0.624 ± 0.127
	UniGradICON†	0.610 ± 0.044	0.133 ± 0.122	1.231 ± 0.262
	UniGradICON (IO)†	0.615 ± 0.047	0.149 ± 0.136	1.527 ± 1.495
	VFA†	0.805 ± 0.044	0.419 ± 0.181	0.462 ± 0.163
	Ours	0.872 ± 0.028	0.528 ± 0.180	0.258 ± 0.099
250 µm	Baseline	0.580 ± 0.055	0.136 ± 0.141	1.409 ± 0.322
	Anatomix†	0.620 ± 0.031	0.161 ± 0.115	1.179 ± 0.190
	CLAIRE	0.809 ± 0.054	0.378 ± 0.133	0.570 ± 0.211
	ITK-dreg	0.758 ± 0.048	0.299 ± 0.125	0.613 ± 0.191
	SynthMorph†	0.690 ± 0.052	0.243 ± 0.164	0.882 ± 0.239
	TransMorph†	0.689 ± 0.044	0.191 ± 0.132	0.973 ± 0.245
	UniGradICON (IO)†	0.398 ± 0.062	0.063 ± 0.071	3.491 ± 3.198
	UniGradICON†	0.359 ± 0.044	0.045 ± 0.056	2.992 ± 0.670
	VFA†	0.714 ± 0.066	0.281 ± 0.216	0.821 ± 0.300
	Ours	0.895 ± 0.029	0.597 ± 0.204	0.216 ± 0.098



(b) Accuracy vs. GPU Compute Cost.



(c) Accuracy vs. Wall-clock Time.

Figure 5.11: Registration performance on Faux-OASIS dataset at 1 mm, 500 µm, and 250 µm (native 250 µm); mean ± std over pairs. ↑ higher is better; ↓ lower is better. (Green)/ (Yellow) = best/second; †= patch-based

consumption is independent of H with the Ring Sampler, and scaling efficiency is only minimally affected.

Ablation on GP. We ablate the effect of GP by replacing it with DTensor sharding (no boundary sync). Figs. D.1 and 5.9b show that incorrect boundary synchronization leads to undesirable artifacts in the moved images, and reduces labelmap overlap.

5.10. Related Work

This chapter identifies the key workhorse operations in image registration, and builds an extensive systems design framework for both single GPU and distributed registration. For the average image registration practitioner, we include related works that span key literature on memory efficient and large scale optimization, and the growth of modern applications in life sciences and biomedical imaging that demand an urgent need for such methods.

Table 5.4: Extended Efficiency Results on faux-OASIS-dataset: Comparison of registration methods across multiple resolutions. Reported metrics include average Dice similarity coefficient (higher is better), wall-clock runtime, GPU cost (measured in GB-hours), relative speedup, and GPU cost reduction with respect to FireANTs + FFDP (Ours). GPU usage (e.g., single GPU, multi-GPU, or CPU) is annotated alongside the cost values.

Resolution	Method	Avg Dice Score ↑	Wall Clock ↓ (10^{-2} Hours)	GPU Cost ↓ (10^{-2} GB-Hours)	Speedup	GPU Cost Reduction (%)
1 mm	TransMorph	0.851 ± 0.016	0.015	0.262^{1}	0.56×	87.81
	VFA	0.851 ± 0.023	0.017	0.216^{1}	0.63×	85.18
	Ours	0.838 ± 0.028	0.027	0.032^{1}	1.00×	0.00
	UniGradICON-IO	0.826 ± 0.022	5.167	58.498^{1}	194.07×	99.95
	UniGradICON-noIO	0.815 ± 0.026	0.067	0.238^{1}	2.50×	86.55
	SynthMorph	0.801 ± 0.022	2.155	99.061^{1}	80.93×	99.97
	Anatomix	0.796 ± 0.035	0.379	2.656^{1}	14.24×	98.80
	CLAIRE	0.776 ± 0.044	0.518	1.389^{1}	19.47×	97.70
	ITK-dreg	0.662 ± 0.055	1.527	1.017^{CPU}	57.37×	–
500 µm	Ours	0.872 ± 0.028	0.109	0.862^{1}	1.00×	0.00
	VFA	0.805 ± 0.044	0.302	3.896^{1}	2.78×	77.87
	CLAIRE	0.779 ± 0.051	25.903	396.169^{1}	238.04×	99.78
	SynthMorph	0.771 ± 0.035	4.068	187.049^{1}	37.39×	99.54
	TransMorph	0.759 ± 0.028	0.198	3.501^{1}	1.82×	75.38
	Anatomix	0.758 ± 0.040	8.837	310.818^{1}	81.21×	99.72
	ITK-dreg	0.699 ± 0.056	41.259	207.466^{CPU}	379.17×	–
	UniGradICON-IO	0.615 ± 0.047	84.538	1072.657^{1}	776.89×	99.92
	UniGradICON	0.610 ± 0.044	0.842	3.545^{1}	7.73×	75.69
250 µm	Ours	0.895 ± 0.029	1.065	47.059^{1}	1.00×	0.00
	CLAIRE	0.809 ± 0.054	1207.536	$159\,046.981^{4}$	1133.84×	99.97
	VFA	0.714 ± 0.066	3.872	49.939^{1}	3.64×	5.77
	SynthMorph	0.690 ± 0.052	32.808	1507.133^{1}	30.80×	96.88
	TransMorph	0.689 ± 0.044	2.597	45.965^{1}	2.44×	−2.38
	Anatomix	0.620 ± 0.031	88.480	3112.015^{1}	83.07×	98.49
	UniGradICON-IO	0.398 ± 0.062	163.812	2539.721^{1}	153.80×	98.15
	UniGradICON	0.359 ± 0.044	7.811	55.057^{1}	7.33×	14.53
	ITK-dreg	0.758 ± 0.046	1363.868	$33\,065.677^{\text{CPU}}$	1280.63×	–

5.10.1. Memory Efficient and Large Scale Optimization

Recent advances in large scale transformer-based model training has amassed significant attention and efforts to alleviate key bottlenecks in both memory and compute efficiency. Activation memory forms a key bottleneck in many deep learning training pipelines, and recent advances propose fused operations (Dao et al., 2022; Dao, 2023; Shah et al., 2024; PyTorch, 2023; Bikshandi and Shah, 2023; Dong et al., 2024) to significantly reduce HBM usage without approximations. Other techniques propose sub-quadratic approximations to the quadratic complexity of the attention operation and propose highly efficient and IO-aware fused kernels (Yuan et al., 2025; Dong et al., 2024; Wang et al., 2024). However, as these models and their inputs get increasingly larger in size, they do not fit on a single GPU. Various distributed techniques like Tensor Parallel (Shoeybi et al., 2019), Sequence Parallel (Li et al., 2023a; Jacobs et al., 2024; Li et al., 2024), pipeline parallel

(Qi et al., 2023; Lamy-Poirier, 2023; Liu et al., 2024a), fully-sharded data parallel (FSDP2) (Ansel et al., 2024; Zhao et al., 2023; Rajbhandari et al., 2020) have been proposed that distribute (shard) the model and its inputs across multiple GPUs for transformer-like models. Another research area approaches the problem of scaling large models by building compilers and intermediate representations to enable writing optimized kernels at runtime OpenAI (2021); Ansel et al. (2024); Spector et al. (2025); Chen et al. (2018); Abadi et al. (2016). To our knowledge, most of these techniques are tailored to transformer-specific architectures and GEMM-like operations (self attention, feedforward, batchnorm, etc.) only, and a Tensor/Model Parallel variant for convolution-aware sharding is not available. However, other disciplines including biomedical and clinical imaging, life sciences, climate modeling, drug discovery, genomics, geosciences, robotics leverage other key components that do not fit in the transformer-specific framework, or are GEMM-like in nature. We focus on the compute and memory bottlenecks in the image registration problem, that is a key component in a variety of biomedical and life science applications.

CHAPTER 6

The FireANTs Ecosystem for In-the-Wild Image Registration

Precision mapping techniques coupled with high resolution image acquisition of multiple modalities including MRI, high-resolution sectioning microscopy and histology, spatial transcriptomics permit the study of spatial organization of macroscopic tissues, functional connectome, microscopic cytoarchitecture, gene expression, and proteomics. These techniques in multi-modal imaging advance our understanding of biological organization and function at the macroscopic, mesoscopic, and microscopic scales. Quantitative research in these fields is facilitated by either standard anatomical coordinate frames, or subject specific frames, and the ability to spatially map such standardized spaces. Many of the scenarios collect data with a wide variety of protocols and constraints, and there is no one-size-fits-all workflow for collecting, processing, and registering the data.

In this chapter, we illustrate the versatility of the FireANTs ecosystem for generating precision spatial mappings while adhering to workflow or task-specific constraints. We will look at three different pipelines catering to a wide range of applications:

1. **End-to-end Multimodal Pipeline for histology to MRI registration workflows for Alzheimer's Disease (AD)**: To interpret clinical Alzheimer's Disease (AD) pathology, a multitude of imaging modalities are used to integrate crucial insights into brain morphology and pathology. Postmortem MRI provides a 3D reference of the morphological structure of the brain, while histology provides a detailed view of the hallmark accumulation of extracellular beta-amyloid (Aβ) plaques and intracellular neurofibrillary tangles (NFTs) composed of hyperphosphorylated tau protein. Moreover, due to microtome sizing constraints, postmortem MRI brains undergo a sequence of physical sectioning steps, first into coronal slabs, followed by block sections, each of which are used to section a stack of 2D histology sections that are stained with a variety of stains to visualize different pathological features. Registering a single 2D histology section to a 3D MRI scan is a severely underconstrained problem, and therefore requires a multimodal pipeline to register the histology sections to the MRI scan.

2. **Restricted deformations for distortion correction in spin-echo echo-planar MRI images**: Spin-echo echo-planar imaging (EPI) is widely used for its rapid acquisition, but it is inherently prone to geometric distortions due to magnetic field inhomogeneities, susceptibility differences, and the low bandwidth in the phase-encoding direction. These distortions manifest as geometric or spatial warping along the phase-encode axis. Interestingly, when k-space is traversed in the opposite direction, the same structures are distorted in mirror-opposite directions while the magnitude of the distortions remains the same. By acquiring images with opposite phase-encoding directions, one can leverage image registration techniques to estimate and correct the distortions, effectively generating an intermediate image that approximates the true undistorted anatomy. This approach enables accurate alignment across images and modalities, providing a more reliable basis for downstream analyses in applications ranging from functional MRI to postmortem structural studies.

3. **Gradient-Free approaches for feature-based registration with sparse landmark supervision**: In Chapter 4, we showed that FireANTs can be used as a fully differentiable layer when dense anatomical labelmaps are available. However, other applications exist where the data and labels are sparse. For example, the Learn2Reg CT lung challenge (Hering et al., 2022) provide only 20 intra subject inspiration-expiration lung CT pairs with a sparsely annotated set of automatically detected landmarks. These landmarks are of Lebesgue measure zero, and training a dense deformation field to minimize landmark

error will provide very sparse deformations. Nevertheless, we can use evolutionary algorithms to find a sparse set of deformations that minimize the landmark error, by using the negative landmark error as a fitness function.

6.1. End-to-end Multimodal Pipeline for histology to MRI registration workflows for Alzheimer's Disease (AD)

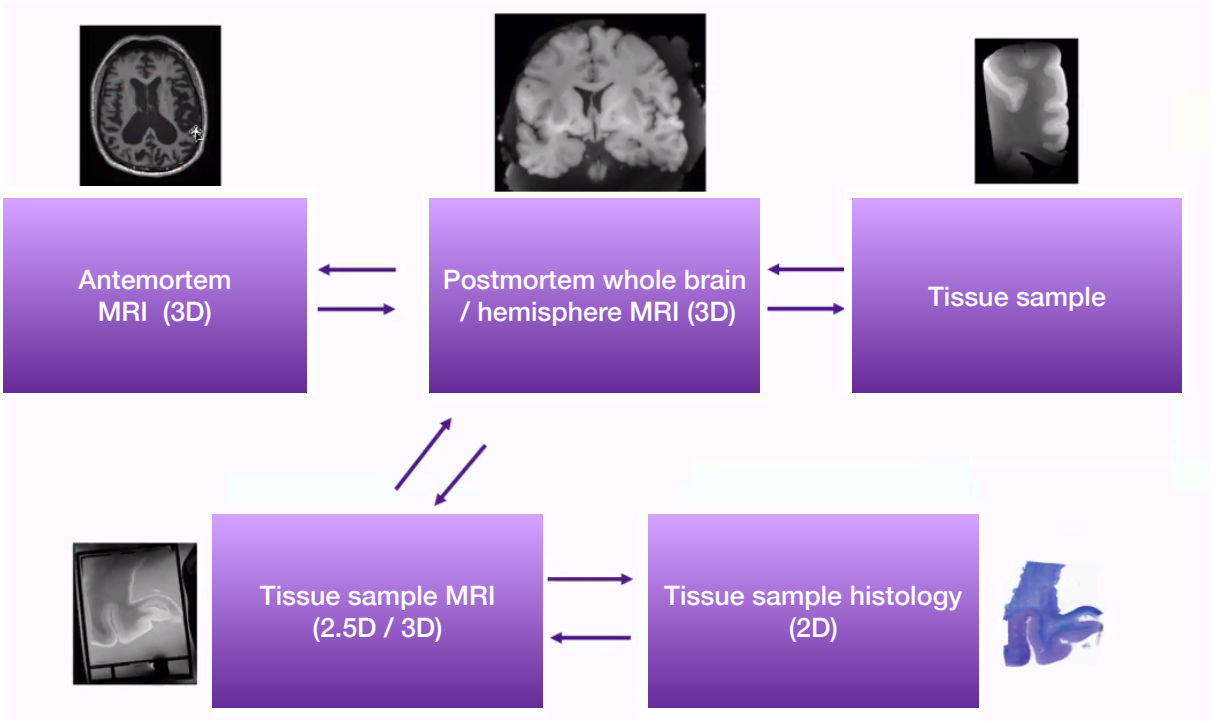


Figure 6.1: A visual overview of the end-to-end multimodal pipeline for histology to MRI registration workflows for Alzheimer's Disease (AD)

Fig. 6.1 shows a visual overview of the end-to-end multimodal pipeline for histology to MRI registration workflows for Alzheimer's Disease (AD). This pipeline uses multiple imaging stages to bridge the large gap in scale, modality, and distortion between in vivo MRI and microscopic histology. The antemortem MRI provides the reference representation of the brain in its natural, in vivo state, preserving clinically relevant anatomy and serving as the final target space for mapping microscopic findings. The in vivo MRI itself is very low resolution and is non invasive, therefore cannot provide microscopic details that are present in the histology. However, direct alignment between antemortem MRI and histology is infeasible due to severe geometric distortions and tissue deformation introduced during extraction and processing. Moreover, the acquisition of the histology images requires extracting and processing of the brain tissue after death, making postmortem MRI an intermediate modality that can be used in the pipeline. Postmortem whole-brain or hemisphere MRI is acquired to capture the brain after fixation but before sectioning, providing an intermediate that reflects postmortem tissue properties while preserving global anatomy. Postmortem MRI can also be acquired with significantly higher spatial resolution than the in vivo MRI because it is not subject to the same constraints on scan time and motion artifacts. The physical tissue sample is then extracted from the postmortem brain, and tissue sample MRI is performed to image the specimen in its block form, capturing its

three-dimensional structure and deformation prior to sectioning at a much higher spatial resolution. Finally, histology provides the microscopic, two-dimensional cellular detail that is the ultimate source of biological information but contains distortions from cutting, mounting, and staining. By registering each stage to its adjacent representation—histology to tissue sample MRI, tissue sample MRI to postmortem MRI, and postmortem MRI to antemortem MRI—the pipeline enables accurate transfer of microscopic information into the in vivo coordinate system while minimizing errors that would arise from attempting direct cross-scale registration.

In the following sections, we will describe each of the stages in the pipeline.

6.1.1. Stage 1: Antemortem to Postmortem MRI registration

The initial stage of the pipeline involves the spatial alignment of the *in vivo* (antemortem) MRI with the *ex vivo* (postmortem) scan. This step is critical as it establishes the foundational link between the clinical 'living' state of Alzheimer's Disease (AD) and the high-resolution anatomical detail captured after death. The primary objective of this stage is to account for the significant geometric and intensity transformations that occur during the peri-mortal period. Key challenges addressed in this registration step include:

- **Gross anatomical misalignment:** The antemortem MRI is acquired in a standardized 'radiological world' orientation due to standardized and calibrated imaging protocols. Typical visualization software like ITK-snap show radiological axes as L-R, A-P, and I-S, which is standard for virtually all MRI scanners for antemortem MRI. However, the postmortem MRI is acquired after the brain has been fixed and sectioned, and therefore the orientation of the brain is arbitrary depending on the containing environment, the technician's placement of the brain, and the coil geometry. In radiological world coordinates, the postmortem MRI is therefore in a random orientation with respect to the antemortem MRI which cannot be rectified with gradient based methods due to multiple local minima. Sometimes, only a single hemisphere is imaged, adding to the complexity of the alignment. Therefore, we employ a global transformation to align the two scans.

- **Morphological Deformation:** Postmortem tissue undergoes global changes due to the loss of intracranial pressure and the chemical fixation process. These factors typically lead to volumetric shrinkage and deformation. To reconcile these differences, we employ FireANTs that can warp the postmortem volume into the antemortem coordinate space while preserving anatomical topology.

- **Contrast Divergence:** Postmortem tissues exhibit significantly shortened $T1$ and $T2$ relaxation times compared to living tissue, caused by temperature drops and formaldehyde-induced protein cross-linking. For example, GM-WM contrast is significantly reduced in the postmortem MRI compared to the in vivo MRI, making it difficult to align the two scans. This requires robust registration similarity metrics.

By establishing this spatial correspondence, any microscopic pathology identified in later stages can be projected back into the original clinical scan. This allows for the direct correlation of cellular AD markers with the neuroimaging signatures observed during the patient's lifetime.

Global alignment using moments matching

The brain can be approximated as an ellipsoid with a major axis along the A-P (front to back) axis, a middle axis along the L-R (mediolateral) axis, and a minor axis along the I-S (dorsoventral) axis. Rigid, affine, and deformable registration assume that the fixed and moving images already reside in similar physical coordinates; when the postmortem volume is in an arbitrary orientation and possibly different scale, gradient-based methods can get stuck in poor local minima. Moment matching brings the images into a common physical space by aligning their statistical moments in physical coordinates, providing a robust initialization for subsequent registration.

Note that moment matching works only for intra-modal registration since the intensities are used as "masses" to compute second-order moments. Multimodal images will have different "mass distributions" for the same anatomy, and therefore moment matching is not applicable. A modality-agnostic feature extractor (Dey et al., 2025) can be used to extract features from the images that are invariant to the modality.

Moments in physical coordinates. Let x_i denote physical coordinates and $I(x_i)$ the image intensity. The first-order moment is the intensity-weighted center of mass:

$$m_1 = \frac{\sum_{i=1}^{n} x_i I(x_i)}{\sum_{i=1}^{n} I(x_i)}. \tag{6.1}$$

The second-order moment is the intensity-weighted covariance matrix:

$$M_2 = \frac{\sum_{i=1}^{n} (x_i - m_1)(x_i - m_1)^\top I(x_i)}{\sum_{i=1}^{n} I(x_i)}. \tag{6.2}$$

First- and second-order matching. For first-order (translation-only) matching, the transformation is the difference of the centers of mass:

$$T = m_1^{\text{moving}} - m_1^{\text{fixed}}. \tag{6.3}$$

This can be used to align the center of the two images (for example, when working with different radiological origin coordinates for the same orientation).

For second-order matching, the optimal rotation is obtained by minimizing $\min_{R \in \text{SO}(n)} \| M_2^{\text{moving}} - R M_2^{\text{fixed}} R^\top \|$, which is solved via the SVD of M_2^{fixed} and M_2^{moving}; the rotation is $R = UDV^\top$ where D is a diagonal matrix of ± 1 chosen so that $\det(R) = 1$ (or -1 if flips are allowed), typically by minimizing a similarity loss between the fixed image and the rotated moving image. There are only 4 to 8 choices for D depending on the orientations the user wants to iterate over. The translation is then

$$T = m_1^{\text{moving}} - R m_1^{\text{fixed}}. \tag{6.4}$$

Scale correction. When fixed and moving images have different scales (e.g., ex vivo stretching or inconsistent voxel spacing), second-order moment matching can be extended with a scaling matrix. Let λ_{fixed} and λ_{moving} denote the eigenvalues of the second-order moment matrices. Assuming that the major, middle, minor axes of the ellipsoids still correspond to consistent axes, the scaling matrix is computed as: $S = \sqrt{\text{diag}(\lambda_{\text{moving}} / \lambda_{\text{fixed}})}$, and the final transformation combines rotation and scaling as $R = U_{\text{fixed}} D S U_{\text{moving}}^\top$

Hemisphere detection In our data, the hemisphere is not mentioned in the metadata, and the goal is to detect the hemisphere from the image without manual intervention. To do this, we use SynthSeg (Billot et al., 2023), a segmentation model to generate a cortical and subcortical segmentation of the brain. We fit an SVM classifier to the left and right gray matter, white matter, and ventricles to detect the separating plane between the two hemispheres. This gives us two candidate hemispheres, and we register the ex-vivo hemisphere to both in-vivo hemispheres using moments matching, ensuring $\det(R) = 1$. We choose the hemisphere that minimizes the normalized cross-correlation of the registered images.

Nonlinear registration

The next stage is to register the postmortem MRI to the antemortem MRI using a nonlinear registration algorithm. We use FireANTs with greedy (asymmetric) optimization with the diffeomorphic Adam optimizer.

Evaluation

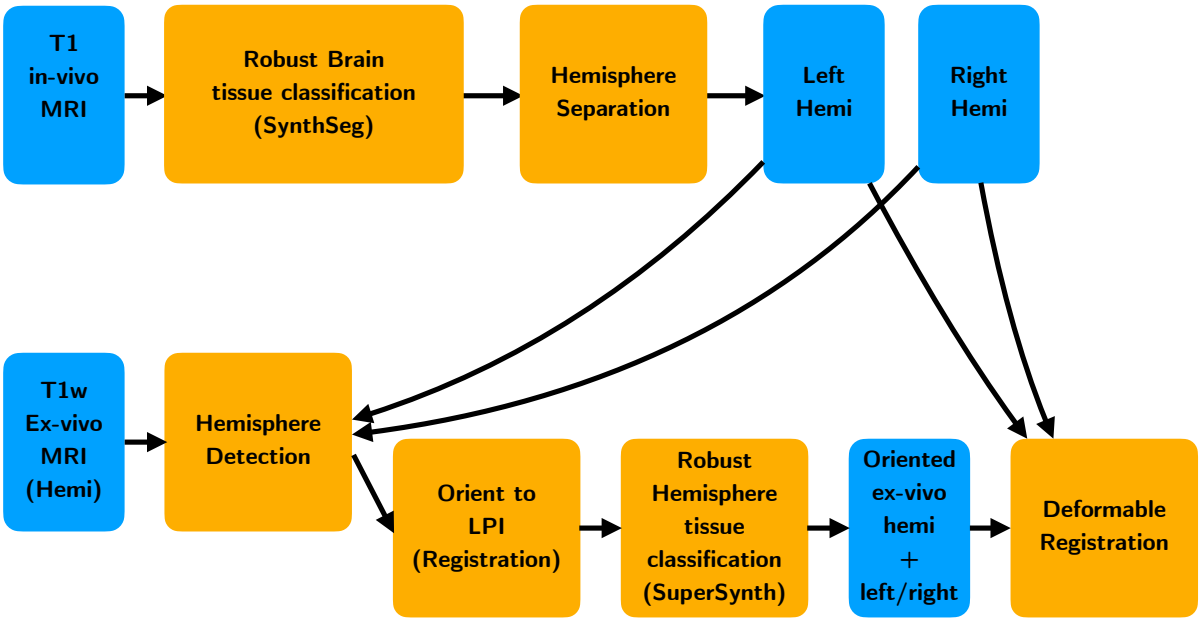


Figure 6.2: End-to-end pipeline for label generation and ex-vivo to in-vivo registration

Evaluating quality of registration is challenging due to a multitude of factors. First, unlike in-vivo MRI images with robust segmentation protocols (Billot et al., 2023), segmentations are scarce for postmortem MRI images, since brain collection protocols can typically result in hemispheres or whole brains with sometimes the cerebellum and / or brainstem removed. Fortunately, SuperSynth (Liu et al., 2025b) provides a modality-agnostic framework for generating segmentations from MRI images and supports postmortem MRI hemispheres, and has semantic labels that are consistent with SynthSeg. However, SuperSynth expects images to be aligned with the LPI orientation, which entangles the accuracy of moment matching and the accuracy of the subsequent segmentation which is used for evaluation of non-linear registration. Nevertheless, we use a straightforward strategy outlined in Fig. 6.2 to generate segmentations for the postmortem MRI. Specifically, SynthSeg provides robust segmentation maps of 34 subcortical structures from the in-vivo MRI.

This is used to separate the invivo brain into left and right hemispheres. For the ex-vivo brain, we perform hemisphere detection using moments matching on both in-vivo hemispheres to find the best match. This orients the ex-vivo brain to a standard LPI orientation which is in-distribution for SuperSynth to generate segmentations reliably. From both the SynthSeg and SuperSynth segmentations, three labels are extracted: gray matter, white matter, and CSF. To account for minor inaccuracies in the labelmaps, we use a finite mixture modelling approach based on the initial segmentations and an MRF prior to enforce spatial consistency and smoothing of the labels. We report the Dice Score overlap between the registered and reference labelmaps after non-linear registration.

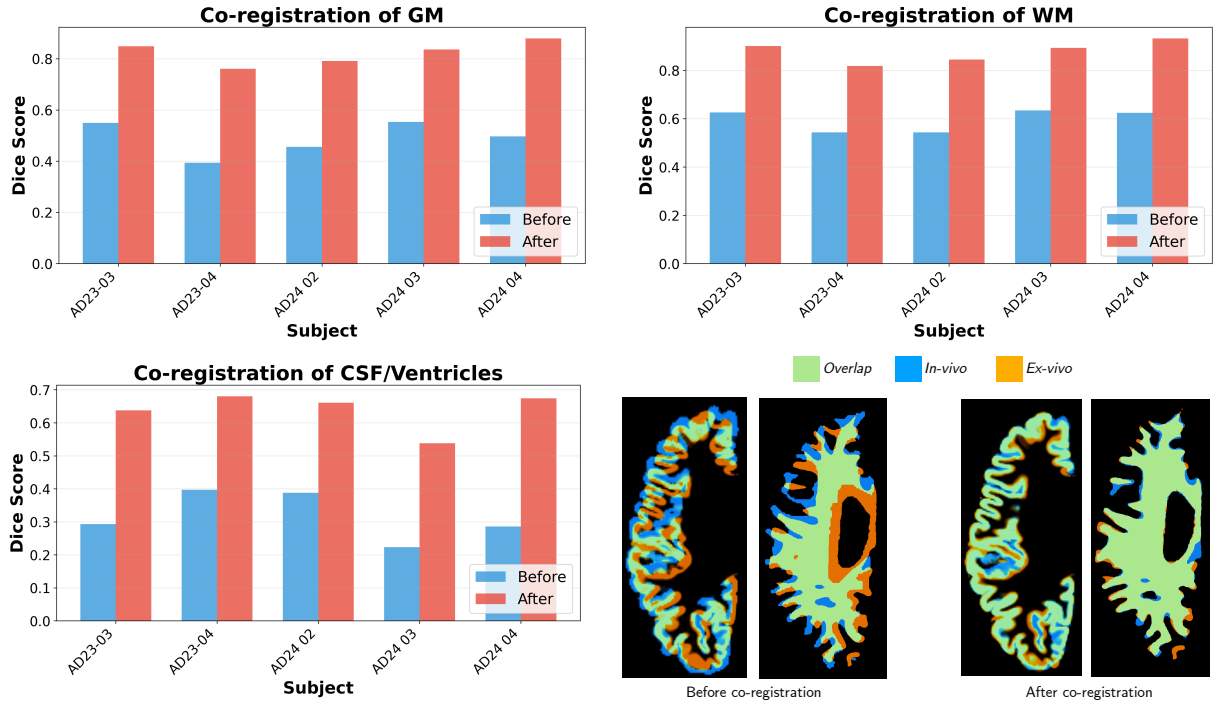


Figure 6.3: **(a, b, c)** Dice Score overlap between the registered and reference labelmaps after non-linear registration for gray matter, white matter, and CSF, **(d)** Qualitative comparison of the in-vivo and ex-vivo MRI scans

As of writing this chapter, we acquired in-vivo and ex-vivo MRI scans for five subjects that we use for evaluation. Fig. 6.3 shows the Dice Score overlap between the registered and reference labelmaps after non-linear registration for gray matter, white matter, and CSF. Dice score for gray matter improved from 0.5 to 0.8, which is a significant improvement considering the highly folded nature of the cortical surface. White matter Dice score improved from 0.6 to 0.8, consistent with the noticed improvement in the literature for OASIS and LPBA40 datasets. CSF segmentation performance is crucial due to the reduced contrast of the pial surface in the postmortem MRI, and the enlarged and deformed ventricles in the ex-vivo MRI. Notably, the Dice score before registration was 0.3 for CSF, and improved to 0.6 after registration. Improvements are consistent across all five subjects, indicating that the pipeline is robust to subject-specific variations.

6.1.2. Stage 2: Postmortem MRI chunk to Histology registration

Method

The method used for this stage is loosely based on a subspace approach for matching 2D shape contours under affine distortions (Mai et al., 2011). The core idea is that a pair of shapes characterized by their pointsets $X, Y \in \mathbb{R}^{2\times N}$ are affinely related if $Y(\tau) = AX$ where τ is a circular shift operation on the columns of Y. Mai et al. (2011) show that the pointsets are related by an affine transformation if and only if their canonical forms V_X and $V_Y(\tau)$ are related by a rotation matrix. The canonical form is defined by the reduced SVD decomposition $X = U_X \Sigma_X V_X^\top$. However, in real settings X and Y may not have the same number of points. Mai et al. (2011) propose resampling the pointsets to the same number of points using a uniform sampling over the arc-length along the canonical curves V_X and V_Y and optimize $\min_{\tau,R} \|V_Y(\tau) - RV_X\|_2^2$.

Instead, we normalize the pointset $X = (U_X \Sigma_X \rho_X)\frac{V_X^\top}{\rho_X}$ where $\rho_X = \sum_{i=1}^{N} \|V_X(:,i)\|_2$ is the scale of the pointset. For brevity, we define $A_X = (U_X \Sigma_X \rho_X)$ and $\tilde{V}_X = \frac{V_X^\top}{\rho_X}$. This normalizes both pointsets to have the same scale to allow for Procrustes analysis. Instead of optimizing over τ and R, we optimize only over R by computing the Chamfer distance between $\tilde{V}_Y$ and $R\tilde{V}_X$. Chamfer distance works with pointsets of different sizes avoiding any resampling step. Finally, we compute the affine transformation $A = A_Y R A_X^{-1}$ as the optimal affine transformation. Note that finding the optimial rotation is a fast operation in 2D because it has only one parameter, compared to directly optimizing a six-parameter affine transformation. This approach mirrors the global moment matching approach for in-vivo to ex-vivo MRI registration by correcting for gross anatomical misalignment, but uses contour matching instead of intensity matching.

Results

We obtain a T2 block from the superior frontal lobe of a single subject, along with corresponding DWI and FA volumes. The DWI image is thresholded to obtain a mask of the chunk which is resampled to the same resolution as the T2 block. For the block, several stains are applied to visualize different pathological features. Specifically, the histology stack contains vascular (CD31, SMA), immune/inflammatory (CD3, CD68, IBA1), glial (GFAP, MBP, PDGFR), neuronal (NFL), myelin (LFB, MBP), amyloid (4G8), and fibrosis/stroma (FIB). A custom script is used to extract the foreground of the stained histology slides by thresholding and selecting the largest connected component from each stain. Our FireANTs-based pipeline consists of three stages: (i) global alignment using subspace matching of silhouette of the mask, (ii) affine refinement using LNCC loss, and (iii) nonlinear registration using FireANTs with greedy (asymmetric) optimization with the diffeomorphic Adam optimizer. Results are shown in Fig. 6.4, which are qualitative due to the lack of quantitative evaluation metrics for this task.

Note that in all cases, the histology slides are aligned differently from each other, there is substantial difference in the characteristics of the tissue in the different slides, and have substantial deformations compared to the T2 block. Nevertheless, our pipeline is able to register the histology slides to the T2 block with high accuracy. Notably, the initial affine transformation is recovered correctly without any manual intervention, which is typically performed in various MRI-to-histology registration pipelines.

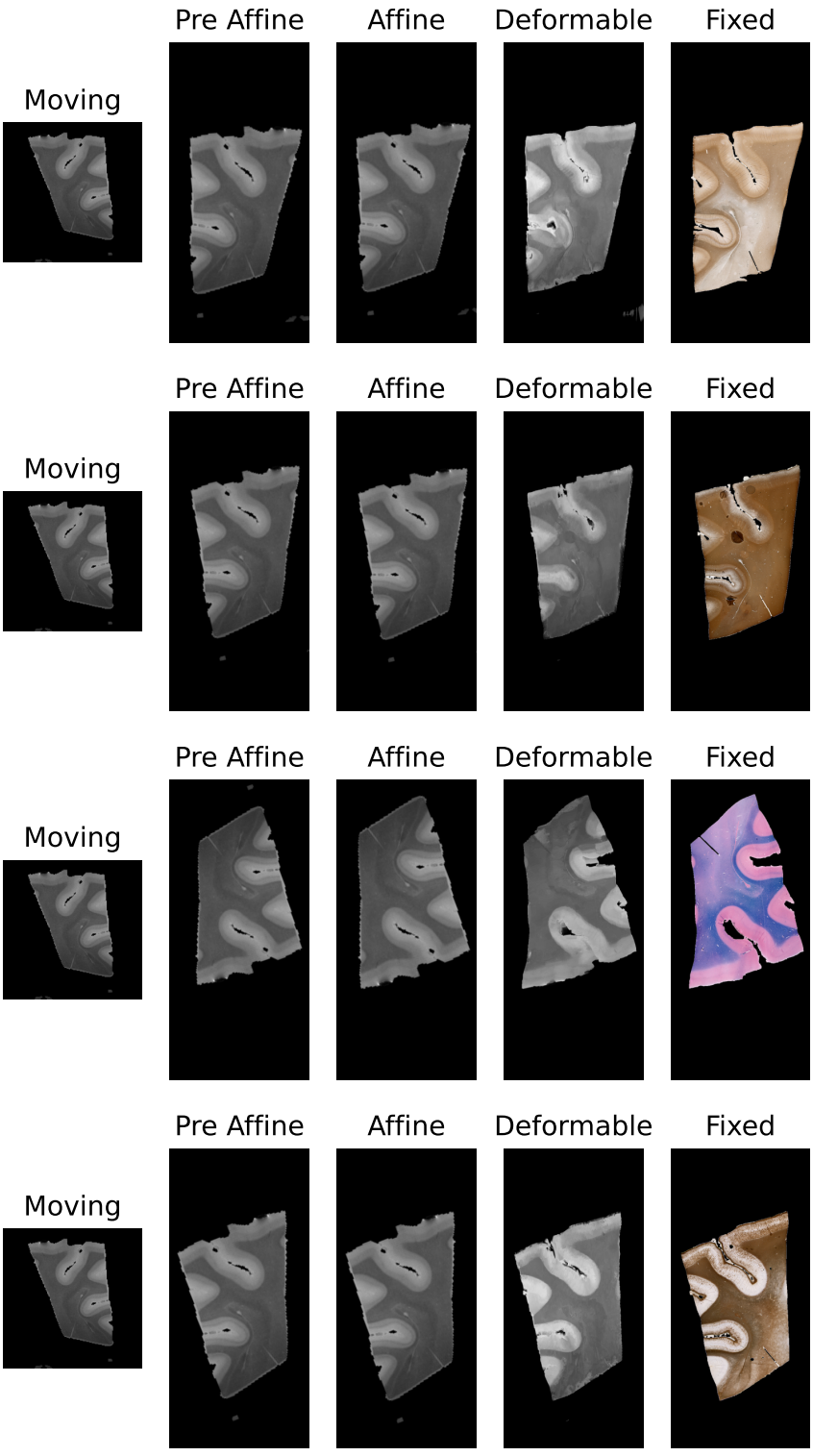


Figure 6.4: Qualitative comparison of the postmortem MRI chunk to histology registration

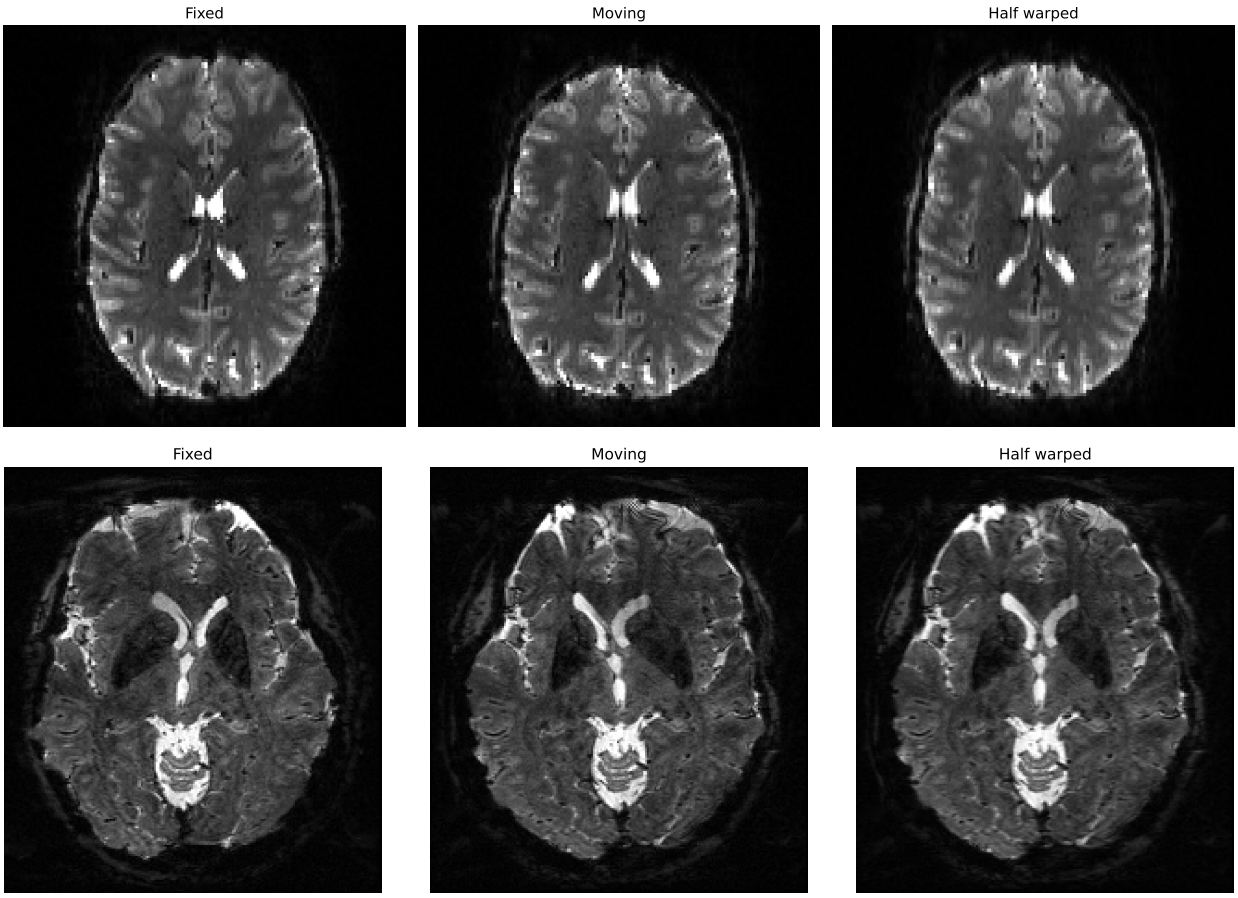


Figure 6.5: Qualitative comparison of the spin-echo echo-planar MRI distortion correction for AP and LR phase-encoding directions

6.2. Restricted deformations for distortion correction in spin-echo echo-planar MRI images

Echo Planar Imaging (EPI) is one of the fastest acquisition techniques in the MRI world, capable of capturing a complete 2D image slice in a fraction of a second—often under 100 milliseconds. Instead of the traditional step-by-step approach to filling data, EPI uses a rapid, ”zigzag” gradient switching pattern to collect all the necessary spatial information from a single radiofrequency pulse. This incredible speed makes it the backbone of dynamic imaging studies where timing is everything, such as Functional MRI (fMRI) for mapping brain activity or Diffusion-Weighted Imaging (DWI) for detecting early-stage strokes.

However, EPI is highly sensitive to ”susceptibility artifacts”. Because the data is collected so quickly, even minor magnetic field irregularities, like those caused by air-tissue interfaces near the sinuses or metal implants can lead to significant image warping or blurring. Therefore, EPI do not provide faithful representations of the brain anatomy. However, the geometric distortions usually manifest as the image looking ”stretched” or ”shrunken” along the phase-encoding direction. Interestingly, if another image is acquired with the phase-encoding direction reversed, the same structures are distorted in mirror-opposite directions while the magnitude of the distortions remains the same (Andersson et al., 2003).

This allows a simple modification to FireANTs to recover the ”true geometry” of the image by performing ‘‘restricted deformations’’ Specifically, the user can provide partial or binary direction of restrictions to the update field. This is done regardless of the optimizer used, since only the final update step is modified. Once a deformation $u(x)$ is obtained that transforms $F(x) \approx M(x + u(x))$, the user can compute the ”half deformation” that moves the particle at a fixed coordinate half way along its location in the second image, which is the supposed location of the particle in the ”true geometry” image.

6.2.1. Results

We evaluate the performance of the proposed method on a dataset of two subjects in Fig. 6.5, where distortions are present in AP and LR phase-encoding directions respectively. Specifically, in the first image, the

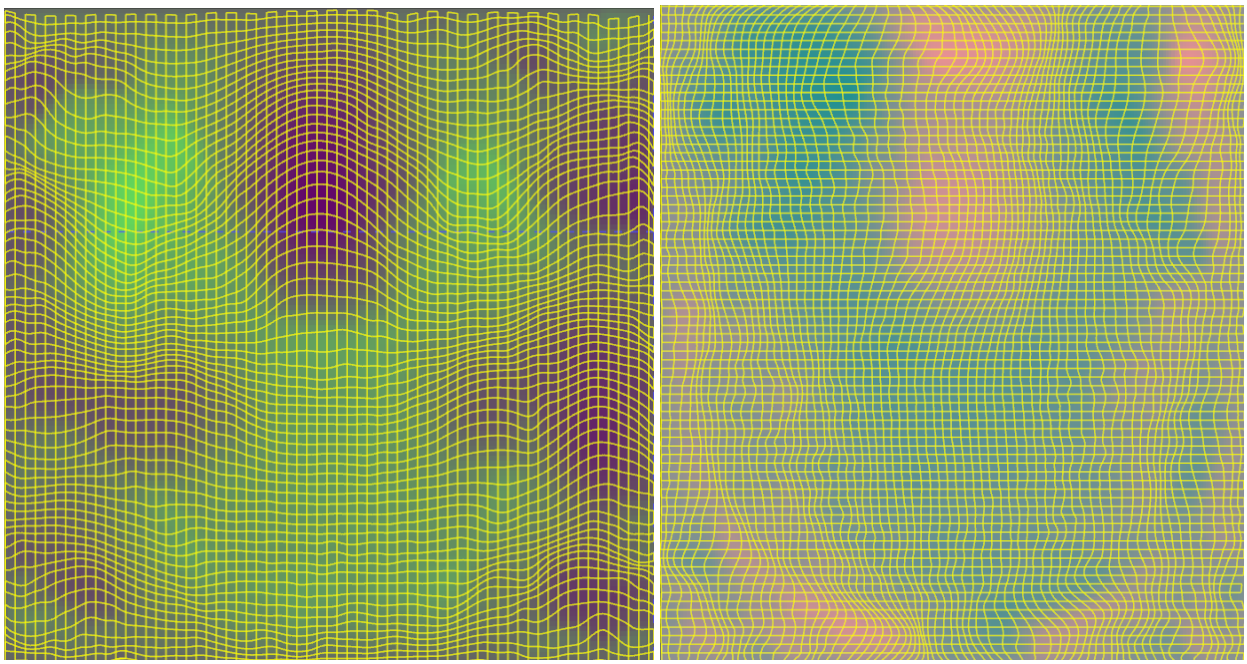


Figure 6.6: Warp fields visualized in Fig. 6.5 for AP and LR phase-encoding directions respectively. Note that in the first case, all deformations are diffeomorphic and restricted along the A-P axis, while in the second case, the deformations are diffeomorphic and restricted along the L-R axis.

Warp fields visualized in Fig. 6.6 show that FireANTs is able to accurately restrict the deformation to the specified directions while using Adam optimizer and preserving diffeomorphisms everywhere. A comparable ANTs registration takes about 12 hours to complete the same task that FireANTs completes in under a minute.

6.3. Gradient-free hybrid approaches for feature-based registration with ultra sparse landmark supervision

FireANTs provides a powerful and robust optimizer for registration, which can be further augmented with an end-to-end feature extractor to learn task-specific features. However, feature learning is typically performed in scnarios where dense supervision (a.k.a parcellated labelmaps) are available for a large number of images. However, in several applications, only a handful of images might be available for training, with a small number of labeled landmarks or keypoints. An example of this scenario is the LungCT (Hering et al., 2020) and OncoReg (Heyer et al., 2025) challenges in the Learn2Reg challenge, where only 20 image pairs with labeled landmarks are available for training and validation. Such scenarios present two major challenges: (i) the amount available for training is too small to train deep feature extractors, and (ii) the supervision signal (e.g. landmark errors) is measure-zero when optimizing over a dense deformation field, providing extremely sparse gradients for learning. This has led to many winning challenge entries for the OncoReg challenge leveraging pretrained models like DINOv2 Song et al. (2024), handcrafted features like MIND Heinrich et al. (2012), and traditional optimizers like Deeds Heinrich et al. (2014). To alleviate the challenge associated with sparse keypoint supervision, methods like Jia et al. (2023) convert keypoints into a $3 \times 3 \times 3$ heatmap and compute Dice loss over the fixed and moving heatmaps. The OncoReg challenge (Heyer et al., 2025) provides valuable insights and highlights that hybrid approaches that combine classifcal registration techniques with deep learning can yield improvements.

We take a different approach towards training a hybrid approach using a shallow network feature extractor that is trained using gradient-free methods instead. Most gradient-free methods in the literature are typically used to find optimal linear transformations under extreme initial misalignment where gradient-based methods would be stuck in a local minima (Brunnstrom and Stoddart, 1996; Rouet et al., 2002; Chalermwat et al., 2001; Chow et al., 2004; Cordón et al., 2006). In contrast, we propsoe using gradient-free methods to learn a shallow convolutional feature extractor for registration with sparse landmark supervision.

6.3.1. Method

We take inspiration from the MIND descriptor that has been shown to be effective for pulmonary image registration (Heinrich et al., 2012; Heyer et al., 2025). The MIND descriptor can be formulated as a convolutional layer which computes the patch difference at a given location i and its neighbor $i + r$, followed by a nonlinearity, i.e. squaring the difference, followed by a normalization step where the values are divided by the average of the squared values of the patch differences as a measure of 'variance'. This is followed by yet another nonlinearity, i.e. an exponential function. Equivalently, one can think of the exponential function in conjunction with the variance as a softmax layer on the patch difference feature. In a learnable analog to this feature extractor, we can train a network that performs convolution to compute patch features, followed by a non-linearity (GeLU), and a normalization layer (GroupNorm) to compute a learnable analog of the MIND descriptor. A residual stream is added to the output of the feature extractor to provide an inductive bias of the features towards the identity transformation.

The small network augments the image into a multi-channel feature that can be used to register using FireANTs. We can train this network using DIO Chapter 4 by backpropagating the loss through the optimizer. However, landmark supervision is very sparse, and is generally considered to be badly conditioned for gradient-based methods, which can steer optimization away from other candidate solutions that generalize better. Therefore, we use gradient-free methods to train the network. Specifically, we use a Genetic Algorithm (GA) to train

Method	Feature Extractor	Masking	Landmark Error (Train)	Landmark Error (Validation)
ConvexAdam	MIND	yes	-	-
		no	2.618	3.553
FireANTs	Intensity	yes	3.099	5.441
		no	3.095	5.427
FireANTs	MIND	yes	2.107	2.989
		no	2.721	4.198
FireANTs	Proposed (1×)	yes	**2.142**	**2.979**
		no	3.039	4.027
FireANTs	Proposed (8×)	yes	**1.759**	**2.558**
		no	2.619	3.707

Table 6.1: Performance of the proposed method on the LungCT challenge. Results are shown for the proposed method with and without masking, and for the baseline methods with and without MIND features.

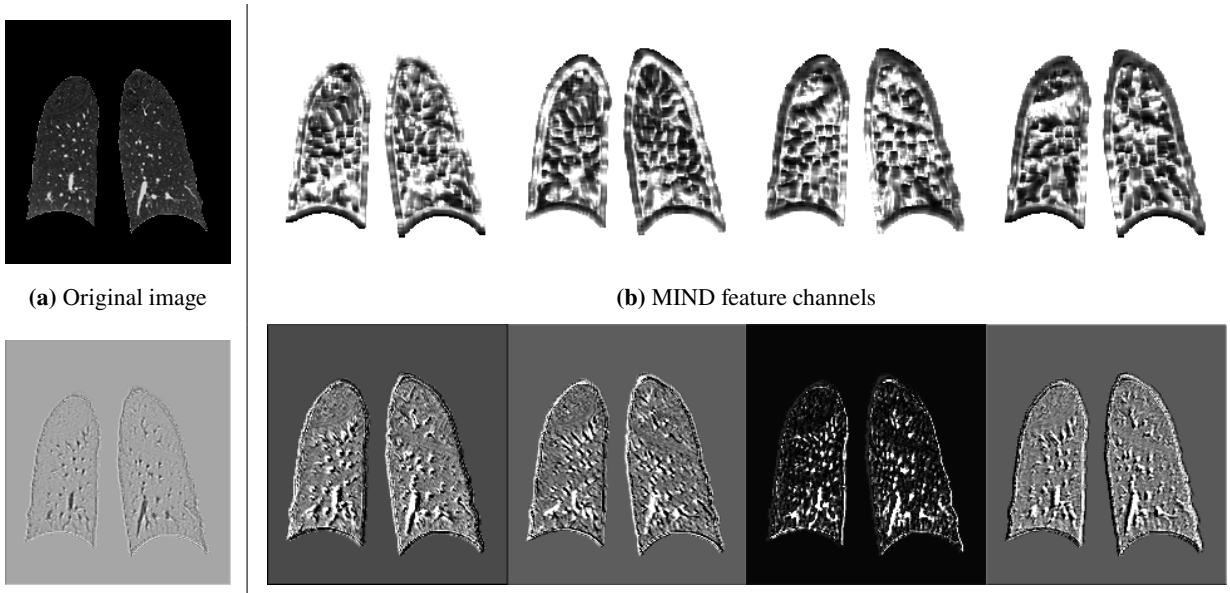
the network, although other paradigms like reinforcement learning or derivative-free trust-region methods like BOBYQA (Powell et al., 2009) are possible. Genetic algorithms are used in neural architecture search (Lu et al., 2019; Yang et al., 2020), and to finetune large language models (Qiu et al., 2025). We use genetic algorithms to search over the parameters of the network instead. Specifically, a genetic algorithm maintains a population of candidate solutions, (neural network parameters in our case)), and repeatedly performs the following steps: (i) select good solutions measured by the fitness function, (ii) combine good solutions, (iii) randomly mutate the solutions, and (iv) replace the worst solutions with the new solutions. Maintaining a population of candidates is expensive for large networks, but is completely tractable for small networks like in our case. For combining good solutions, we use the One Point Crossover operator with a tournament size of 2 (meaning we select 2 candidates). For mutation, we apply Gaussian noise to the parameters of the network with a standard deviation of 0.01. The fitness function is the landmark error between the predicted and ground truth landmarks after registration using FireANTs and the feature extractor. We run the GA for 40 generations.

6.3.2. Results

The performance of the proposed method is evaluated on the LungCT challenge. Since the Lung CT challenges typically provide binary lung masks, we test each baseline with and without masking, whereever possible. We normalize the image inside the mask to be the range $[0, 1]$, and the image outside the mask to be 0. We consider the following baselines with and without masking: (a) FireANTs with intensity images, (b) FireANTs with MIND features, (c) FireANTs with our proposed feature extractor, (d) ConvexAdam with MIND features. For the proposed feature extractor, we test with 1 and 8 feature maps in the output of the feature extractor to examine if additional feature maps improve performance. Results are shown in Table 6.1.

6.3.3. Interpretability of features

Another key benefit of the proposed method is the interpretability of the features learned by the network, analogous to that observed in DIO Chapter 4. Specifically, Figs. 6.7 and 6.8 illustrate the learned feature



(a) Original image

(b) MIND feature channels

(c) Single channel learned feature map

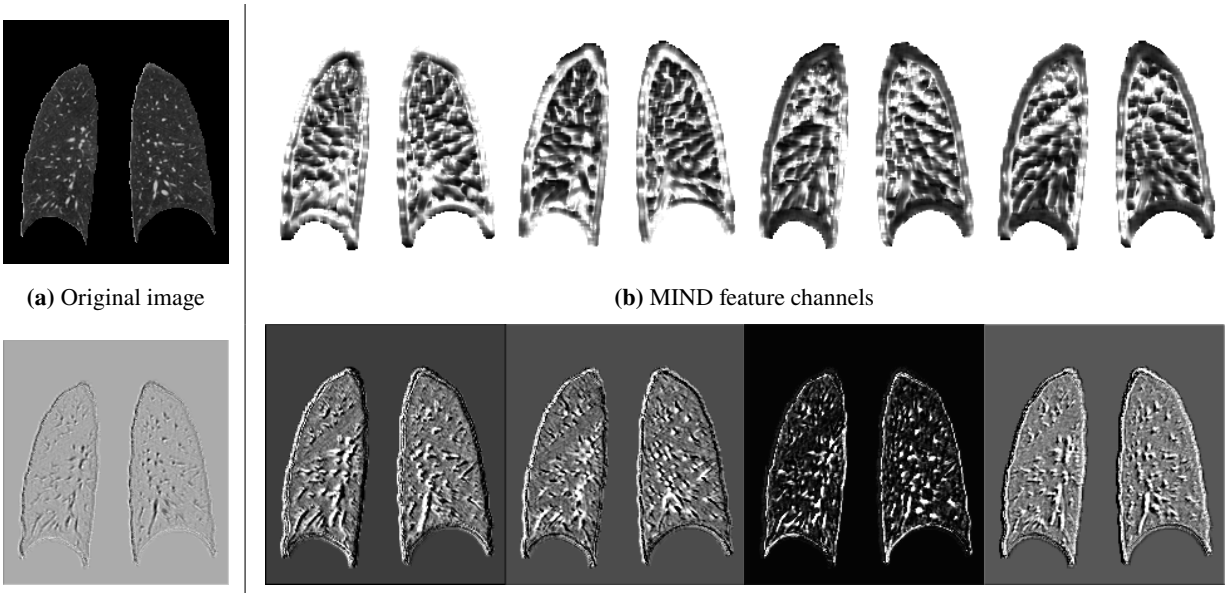
(d) Multi-channel learned feature maps

Figure 6.7: Interpretability of features learned by the proposed method. MIND feature channels enable higher contrast than the original feature image, but has blocky artifacts and is not immediately interpretable. The single-channel learned feature map resembles a high-pass filter applied to the original image, which may be important for accurate delineation and registration of vascular structures that show higher contrast in the high-pass image. The multi-channel learned feature maps are sharper than MIND features, and highlight different regions of interest for different channels.

representations for the three subjects in the LungCT challenge. Quantitatively, the MIND features achieve better performance than raw intensity features, but they exhibit blocky artifacts arising from patchwise normalization, which can amplify noise in low-variance regions. In contrast, both the single-channel and multi-channel learned features not only outperform MIND features, but are also qualitatively more interpretable. Because contrast (which leads to spatial gradients) is the primary driver of intensity-based registration, the single-channel learned feature map enhances vascular structures, thereby facilitating their alignment. The multi-channel representation further increases flexibility through the weighted gradient term $\sum_c r_c(\varphi(x))\nabla M_c(\varphi(x))$, which enables particles to move in different directions depending on the distribution of channelwise residuals. This, in turn, permits a more ‘‘piecewise’’ registration of distinct anatomical structures as a function of the local channelwise residuals. For example, the lung boundary is strongly highlighted in channels 2 and 4 but not in channel 3, which instead emphasizes the vascular structures in the lower fissure of the lung. Moreover, the vasculature exhibits laterally opposing contrast in channels 1 and 4, enabling the registration to drive vessels in opposite lateral directions according to the relative weighting of the channelwise residuals. This is in contrast to single-channel registration, where the residual must change sign in order to drive the deformation in the opposite direction.

6.4. Conclusion

In this chapter, we presented three applications of FireANTs and associated tools for varied applications in medical image registration.



(a) Original image

(b) MIND feature channels

(c) Single channel learned feature map

(d) Multi-channel learned feature maps

Figure 6.8: Another example of the proposed method with a different subject.

The first application is the end-to-end multimodal pipeline for histology to MRI registration workflows for Alzheimer's Disease (AD) where we demonstrated its performance on an internal dataset. The pipeline enables fully automated detection and registration of ex-vivo brain hemisphere scans to an in-vivo subject-specific reference image. This is a challenging task that requires global alignment of the ex-vivo brain hemisphere scan to the in-vivo reference image since the former is typically not aligned to the standard DICOM orientations. This is followed by a nonlinear registration using FireANTs to align the ex-vivo brain hemisphere scan to the in-vivo reference image. Moreover, histology slides are registered to the corresponding MRI casette chunk by performing a global affine alignment using a subspace based shape-matching approach, followed by FireANTs-based nonlinear registration. Future work will complete the chain by registering the MRI casette chunks to the ex-vivo hemisphere scan automatically, a step that is currently performed manually.

The second application is the restricted deformations for distortion correction in spin-echo echo-planar MRI images where we demonstrated its performance on a dataset of two subjects. Spin-echo echo-planar MRI is a fast acquisition technique that is prone to geometric distortions due to magnetic field inhomogeneities, susceptibility differences, and the low bandwidth in the phase-encoding direction. These distortions manifest as geometric or spatial warping along the phase-encode axis. Interestingly, when k-space is traversed in the opposite direction, the same structures are distorted in mirror-opposite directions while the magnitude of the distortions remains the same. This allows a simple modification to FireANTs to recover the "true geometry" of the image by performing ''restricted deformations'' Specifically, the user can provide partial or binary direction of restrictions to the update field. This is done regardless of the optimizer used, since only the final update step is modified. Once a deformation $u(x)$ is obtained that transforms $F(x) \approx M(x + u(x))$, the user can compute the "half deformation" that moves the particle at a fixed coordinate half way along its location in the second image, which is the supposed location of the particle in the "true geometry" image. This is an application that is hard to perform using deep learning methods since it requires large real world datasets, or

an accurate EPI distortion model for synthetic training.

The third application is the proposed masking-based registration method and a feature extractor trained usng gradient-free genetic algorithms where we demonstrated its performance on the LungCT challenge, comparing with both intensity-based and feature-based baselines, including MIND features and intensity images. Our results showed that the proposed method, even with a single learned feature channel, outperforms traditional intensity and MIND-based approaches, and that further improvement is possible with multi-channel learned representations. We also highlighted the interpretability of these learned features, showing how different channels emphasize anatomical structures such as vasculature and lung boundaries in ways that provide meaningful guidance for the registration process. The use of a flexible weighted gradient term (from multi-channel learned features) allows for ‘‘piecewise’’ or ‘‘curricular’’ registration, enabling adaptive and anatomically informed deformations that better capture the complexity of lung anatomy compared to fixed hand-crafted features. Quantitative and qualitative analyses demonstrate not only improved registration accuracy but also enhanced feature interpretability and flexibility. The learned feature maps resemble domain-relevant image filters, such as high-pass filtering to emphasize vessel structure, and multi-channel maps uniquely target different anatomical regions, aiding the registration algorithm in handling diverse local tissue properties.

CHAPTER 7

Conclusion

In this dissertation, we study the existing state of the art, and develop new mathematical foundations and systems for deformable image registration across diverse biomedical modalities and scales. Our work addresses contemporary limitations of registration methods, the first of which is robust coverage across a long tail of modalities and datasets in biomedical and life science applications; where image registration is a key workhorse component for data analysis, multimodal fusion, quantitative morphometrics, and scientific discovery. This is addressed by first studying and validating the limitations of state-of-the-art deep learning methods in achieving coverage even across different modalities within neuroimaging, and establishing the robustness of classical optimization-based methods while noting their shortcomings in terms of runtime, memory efficiency, lack of end-to-end support with deep learning models, and optimization dynamics.

Building on these insights, we developed FireANTs, a scalable and general-purpose multi-scale registration algorithm that leverages adaptive Riemannian optimization. FireANTs achieves real-time registration on clinical datasets and demonstrates unprecedented memory and compute efficiency, enabling broad applicability in clinical and research settings.

To further bridge the gap between classical optimization and data-driven approaches, we introduced Deep Implicit Optimization (DIO), a framework that transforms FireANTs and other iterative solvers into fully differentiable modules. This allows end-to-end learning of task-specific features within neural networks, uniting the reliability of numerical optimization with the expressivity of deep feature extractors. DIO preserves the convergence and interpretability guarantees of classical methods while enabling plug-and-play integration with modern machine learning workflows.

Recognizing that ever-growing image resolutions and volumes pose a new set of computational bottlenecks, we designed a scalable, hardware-aware, and distributed systems framework for image registration. By re-thinking the computational graphs of memory-intensive loss functions and introducing the Grid Parallel paradigm and Ring Sampler, we enable the registration of gigavoxel biomedical images on multi-GPU systems without approximations or sharding artifacts. Our approach demonstrated the first mathematically exact, end-to-end registration of previously intractable large-scale and multimodal datasets, with performance validated on both real and synthetic benchmarks.

Lastly, we demonstrated these advances in real-world neuro and pulmonary imaging, enabling robust multi-scale, multi-modal registration from MRI to histology. We also addressed specialized tasks like distortion correction in MRI and feature-based registration with sparse landmarks, applicable to developmental atlas building and scenarios with limited guidance Kronman et al. (2023).

Collectively, the contributions of the dissertation chart a path towards a new generation of registration tools that are robust, scalable, and adaptable. By unifying novel mathematical results with modern systems design and implementation, this thesis paves the way for accessible, high-fidelity registration workflows that can keep pace with the rapid expansion of biomedical imaging technologies. These advances will not only accelerate scientific discovery but also provide practical, open frameworks for the community to build upon and extend to new modalities, scales, and applications.

BIBLIOGRAPHY

Internet brain segmentation repository (IBSR). http://www.cma.mgh.harvard.edu/ibsr/.

Itk-dreg: A framework for distributed, large-scale image registration. URL https://itk-dreg.readthedocs.io/en/latest/.

Rnr-exm grand challenge. URL https://rnr-exm.grand-challenge.org/.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

Maryana Alegro, Edson Amaro-Jr, Burlen Loring, Helmut Heinsen, Eduardo Alho, Lilla Zollei, Daniela Ushizima, and Lea T Grinberg. Multimodal whole brain registration: Mri and high resolution histology. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 194–202, 2016.

Allen Institute for Brain Science. Allen brain atlas. URL https://atlas.brain-map.org/.

Jesper LR Andersson, Stefan Skare, and John Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, 20(2):870–888, 2003.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 929–947, 2024.

Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A Log-Euclidean Framework for Statistics on Diffeomorphisms. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Rasmus Larsen, Mads Nielsen, and Jon Sporring, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, volume 4190, pages 924–931. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-44707-8 978-3-540-44708-5. doi: 10.1007/11866565_113. URL http://link.springer.com/10.1007/11866565_113. Series Title: Lecture Notes in Computer Science.

John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

John Ashburner and Karl J Friston. Diffeomorphic registration using geodesic shooting and gauss–newton

optimisation. *Neuroimage*, 55(3):954–967, 2011.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, February 2008a. ISSN 1361-8423. doi: 10.1016/j.media.2007.06.004.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, February 2008b. ISSN 1361-8415. doi: 10.1016/j.media.2007.06.004. URL https://www.sciencedirect.com/science/article/pii/S1361841507000606.

Brian Avants and James C. Gee. Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage*, 23:S139–S150, January 2004. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.07.010. URL https://www.sciencedirect.com/science/article/pii/S1053811904003751.

Brian B. Avants, P. Thomas Schoenemann, and James C. Gee. Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. *Medical Image Analysis*, 10(3):397–412, June 2006. ISSN 13618415. doi: 10.1016/j.media.2005.03.005. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841505000411.

Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.

Brian B Avants, Jeffrey T Duda, Emily Kilroy, Kate Krasileva, Kay Jann, Benjamin T Kandel, Nicholas J Tustison, Lirong Yan, Mayank Jog, Robert Smith, Yi Wang, Mirella Dapretto, and Danny J J Wang. The pediatric template of brain perfusion. *Sci Data*, 2:150003, 2015. doi: 10.1038/sdata.2015.3.

Ramsey D Badawi, Hongcheng Shi, Pengcheng Hu, Shuguang Chen, Tianyi Xu, Patricia M Price, Yu Ding, Benjamin A Spencer, Lorenzo Nardo, Weiping Liu, et al. First human imaging studies with the explorer total-body pet scanner. *Journal of Nuclear Medicine*, 60(3):299–303, 2019.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in neural information processing systems*, 32, 2019.

Shaojie Bai, Zhengyang Geng, Yash Savani, and J. Zico Kolter. Deep Equilibrium Optical Flow Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 610–620, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.00070. URL https://ieeexplore.ieee.org/document/9880309/.

Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer vision, graphics, and image processing*, 46(1):1–21, 1989.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

P Balchandani and TP Naidich. Ultra-high-field mr neuroimaging. *American Journal of Neuroradiology*, 36 (7):1204–1215, 2015.

Augustin Banyaga. *The structure of classical diffeomorphism groups*, volume 400. Springer Science & Business Media, 2013.

Antoine Beauchamp, Yohan Yee, Ben C Darwin, Armin Raznahan, Rogier B Mars, and Jason P Lerch. Whole-brain comparison of rodent and human brains using spatial transcriptomics. *elife*, 11:e79418, 2022.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

Erin S Beck, Pascal Sati, Varun Sethi, Tobias Kober, Blake Dewey, Pavan Bhargava, Govind Nair, Irene C Cortese, and Daniel Salo Reich. Improved visualization of cortical lesions in multiple sclerosis using 7t mp2rage. *American Journal of Neuroradiology*, 39(3):459–466, 2018.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61:139–157, 2005.

Alexander Bigalke, Lasse Hansen, and Mattias P Heinrich. Adapting the mean teacher for keypoint-based lung registration under geometric domain shifts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 280–290. Springer, 2022.

Alexander Bigalke, Lasse Hansen, Tony CW Mok, and Mattias P Heinrich. Unsupervised 3d registration through optimization-guided cyclical self-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 677–687. Springer, 2023.

Ganesh Bikshandi and Jay Shah. A case study in cuda kernel fusion: Implementing flashattention-2 on nvidia hopper architecture using the cutlass library. *arXiv preprint arXiv:2312.11918*, 2023.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

Max Blendowski, Lasse Hansen, and Mattias P Heinrich. Weakly-supervised learning of multi-modal features for regularised iterative descent in 3d image registration. *Medical image analysis*, 67:101822, 2021.

John A Bogovic, Hideo Otsuna, Larissa Heinrich, Masayoshi Ito, Jennifer Jeter, Geoffrey Meissner, Aljoscha Nern, Jennifer Colonell, Oz Malkesman, Kei Ito, et al. An unbiased template of the drosophila brain and ventral nerve cord. *Plos one*, 15(12):e0236495, 2020.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Fred L Bookstein. Thin-plate splines and the atlas problem for biomedical images. In *Biennial international conference on information processing in medical imaging*, pages 326–342. Springer, 1991.

Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 2002.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014. URL https://www.manopt.org.

Bella E Brezovec, Andrew B Berger, Yukun A Hao, Feng Chen, Shaul Druckmann, and Thomas R Clandinin. Mapping the neural dynamics of locomotion across the drosophila brain. *Current Biology*, 34(4):710–726, 2024.

Morten Bro-Nielsen and Claus Gramkow. Fast fluid registration of medical images. In *International conference on visualization in biomedical computing*, pages 265–276. Springer, 1996.

Chaim Broit. *Optimal registration of deformed images*. University of Pennsylvania, 1981.

K Brunnstrom and Andrew J Stoddart. Genetic algorithms for free-form surface matching. In *Proceedings of 13th international conference on pattern recognition*, volume 4, pages 689–693. IEEE, 1996.

Martin Burger, Jan Modersitzki, and Lars Ruthotto. A hyperelastic regularization energy for image registration. *SIAM Journal on Scientific Computing*, 35(1):B132–B148, 2013.

Nathan D Cahill, J Alison Noble, and David J Hawkes. Fourier methods for nonparametric image registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 300–308. Springer, 2017.

Harry Carey, Michael Pegios, Lewis Martin, Chris Saleeba, Anita J Turner, Nicholas A Everett, Ingvild E Bjerke, Maja A Puchades, Jan G Bjaalie, and Simon McMullan. Deepslice: rapid fully automatic registration of mouse brain imaging to a volumetric atlas. *Nature Communications*, 14(1):5884, 2023.

Adrià Casamitjana, Matteo Mancini, Eleanor Robinson, Loïc Peter, Roberto Annunziata, Juri Althonayan, Shauna Crampsie, Emily Blackburn, Benjamin Billot, Alessia Atzeni, et al. A probabilistic histological atlas of the human brain for mri segmentation. *Nature*, pages 1–8, 2025.

Prachya Chalermwat, Tarek El-Ghazawi, and Jacqueline LeMoigne. 2-phase ga-based image registration on parallel clusters. *Future Generation Computer Systems*, 17(4):467–476, 2001.

Fei Chen, Paul W Tillberg, and Edward S Boyden. Expansion microscopy. *Science*, 347(6221):543–548, 2015.

Junyu Chen, Eric C Frey, and Yong Du. Unsupervised learning of diffeomorphic image registration via transmorph. In *International Workshop on Biomedical Image Registration*, pages 96–102. Springer, 2022a.

Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, November 2022b. ISSN 13618415. doi: 10.1016/j.media.2022.102615. URL http://arxiv.org/abs/2111.10480. arXiv:2111.10480 [cs, eess].

Junyu Chen, Shuwen Wei, Joel Honkamaa, Pekka Marttinen, Hang Zhang, Min Liu, Yichao Zhou, Zuopeng Tan, Zhuoyuan Wang, Yi Wang, et al. Beyond the lumir challenge: The pathway to foundational registration models. *arXiv preprint arXiv:2505.24160*, 2025.

Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. TVM: An automated End-to-End optimizing compiler for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 578–594, Carlsbad, CA, October 2018. USENIX Association. ISBN 978-1-939133-08-3. URL https://www.usenix.org/conference/osdi18/presentation/chen.

Claude Chevalley. Théorie des groupes de lie. 1951.

Chi Kin Chow, Hung Tat Tsui, and Tong Lee. Surface registration using a dynamic genetic algorithm. *Pattern recognition*, 37(1):105–117, 2004.

Gary E Christensen and Hans J Johnson. Consistent image registration. *IEEE transactions on medical imaging*, 20(7):568–582, 2001.

Gary E Christensen, Richard D Rabbitt, and Michael I Miller. Deformable templates using large deformation kinematics. *IEEE transactions on image processing*, 5(10):1435–1447, 1996.

G.E. Christensen, S.C. Joshi, and M.I. Miller. Volumetric transformation of brain anatomy. *IEEE Transactions on Medical Imaging*, 16(6):864–877, December 1997. ISSN 1558-254X. doi: 10.1109/42.650882. Conference Name: IEEE Transactions on Medical Imaging.

Oscar Cordón, Sergio Damas, and José Santamaría. Feature-based image registration by means of the chc evolutionary algorithm. *Image and Vision Computing*, 24(5):525–533, 2006.

Colin J Cotter and Darryl D Holm. Singular solutions, momentum maps and computational anatomy. *arXiv preprint nlin/0605020*, 2006.

WR Crum, C Tanner, and DJ Hawkes. Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging. *Physics in Medicine & Biology*, 50(21):5153, 2005.

Emiliano D'agostino, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. A viscous fluid model for multimodal non-rigid image registration using mutual information. *Medical image analysis*, 7(4):565–575, 2003.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Sandhitsu R Das, Brian B Avants, Murray Grossman, and James C Gee. Registration based cortical thickness measurement. *Neuroimage*, 45(3):867–879, 2009.

Christos Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Computer Vision and Image Understanding*, 66(2):207–222, 1997.

Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

Neel Dey, Benjamin Billot, Hallee E. Wong, Clinton Wang, Mengwei Ren, Ellen Grant, Adrian V Dalca, and Polina Golland. Learning general-purpose biomedical volume representations using randomized synthesis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xOmC5LiVuN.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. ISSN 00129658, 19399170. URL http://www.jstor.org/stable/1932409.

Gianluca Donato and Serge Belongie. Approximate thin plate spline mappings. In *European conference on computer vision*, pages 21–31. Springer, 2002.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2024.

Theodore D Drivas and Tarek M Elgindi. Singularity formation in the incompressible euler equation in finite and infinite time. *EMS Surveys in Mathematical Sciences*, 10(1):1–100, 2023.

Florence Dru, Pierre Fillard, and Tom Vercauteren. An ITK Implementation of the Symmetric Log-Domain Diffeomorphic Demons Algorithm. *The Insight Journal*, September 2010. ISSN 2327-770X. doi: 10.54294/8vm9t2. URL https://www.insight-journal.org/browse/publication/644.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Paul Dupuis, Ulf Grenander, and Michael I. Miller. Variational Problems on Flows of Diffeomorphisms for Image Matching. *Quarterly of Applied Mathematics*, 56(3):587–600, 1998. ISSN 0033-569X. URL https://www.jstor.org/stable/43638248. Publisher: Brown University.

DAVID G Ebin, G Misio? ek, and STEPHEN C Preston. Singularities of the exponential map on the volume-preserving diffeomorphism group. *Geometric and Functional Analysis*, 16(4):850–868, 2006.

Carmen Echávarri, P Aalten, Harry BM Uylings, HIL Jacobs, Pieter Jelle Visser, EHBM Gronenschild, FRJ Verhey, and S Burgmans. Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease. *Brain Structure and Function*, 215(3):265–271, 2011.

Brian L Edlow, Azma Mareyam, Andreas Horn, Jonathan R Polimeni, Thomas Witzel, M Dylan Tisdall, Jean C Augustinack, Jason P Stockmann, Bram R Diamond, Allison Stevens, et al. 7 tesla mri of the ex vivo human brain at 100 micron resolution. *Scientific data*, 6(1):244, 2019.

Andrea Esquivel, Andrea Ferrero, Achille Mileto, Francis Baffour, Kelly Horst, Prabhakar Shantha Rajiah, Akitoshi Inoue, Shuai Leng, Cynthia McCollough, and Joel G Fletcher. Photon-counting detector ct: key points radiologists should know. *Korean journal of radiology*, 23(9):854, 2022.

Alan C Evans, D Louis Collins, SR Mills, Edward D Brown, Ryan L Kelly, and Terry M Peters. 3d statistical neuroanatomical models from 305 mri volumes. pages 1813–1817, 1993.

Greg M. Fleishman. Bigstream. https://github.com/GFleishman/bigstream, 2023. GitHub repository.

Miriam Friedel, Matthijs C van Eede, Jon Pipitone, M Mallar Chakravarty, and Jason P Lerch. Pydpiper: a flexible toolkit for constructing novel registration pipelines. *Frontiers in neuroinformatics*, 8:67, 2014.

Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D Bradley, Walter J Curran, Tian Liu, and Xiaofeng Yang. Lungregnet: an unsupervised deformable image registration method for 4d-ct lung. *Medical physics*, 47(4):1763–1774, 2020a.

Yabo Fu, Yang Lei, Jun Zhou, Tonghe Wang, S Yu David, Jonathan J Beitler, Walter J Curran, Tian Liu, and Xiaofeng Yang. Synthetic ct-aided mri-ct image registration for head and neck radiotherapy. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11317, pages 572–578. SPIE, 2020b.

Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. JFB: Jacobian-Free Backpropagation for Implicit Networks, December 2021. URL http://arxiv.org/abs/2103.12803. arXiv:2103.12803 [cs].

Davide Gambarotto, Fabian U Zwettler, Maeva Le Guennec, Marketa Schmidt-Cernohorska, Denis Fortun, Susanne Borgers, Jörn Heine, Jan-Gero Schloetel, Matthias Reuss, Michael Unser, et al. Imaging cellular ultrastructures using expansion microscopy (u-exm). *Nature methods*, 16(1):71–74, 2019.

Mihika Gangolli, Laurena Holleran, Joong Hee Kim, Thor D Stein, Victor Alvarez, Ann C McKee, and

David L Brody. Quantitative validation of a nonlinear histology-mri coregistration method using generalized q-sampling imaging in complex human cortical white matter. *Neuroimage*, 153:152–167, 2017.

James C Gee and Ruzena K Bajcsy. Elastic matching: Continuum mechanical and probabilistic analysis. *Brain warping*, 2:183–197, 1998.

James C Gee, Martin Reivich, and Ruzena Bajcsy. Elastically deforming a three-dimensional atlas to match anatomical brain images. 1993.

Zhengyang Geng and J. Zico Kolter. TorchDEQ: A Library for Deep Equilibrium Models, October 2023. URL http://arxiv.org/abs/2310.18605. arXiv:2310.18605 [cs].

Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, and Zhouchen Lin. On training implicit models. *Advances in Neural Information Processing Systems*, 34:24247–24260, 2021.

GitHub. Fireants github issues. https://github.com/rohitrango/FireANTs/issues/15, 2024a. GitHub repository issues.

GitHub. Fireants github issues. https://github.com/rohitrango/FireANTs/pull/10, 2024b. GitHub repository issues.

Ben Glocker, Nikos Komodakis, Georgios Tziritas, Nassir Navab, and Nikos Paragios. Dense image registration through mrfs and efficient linear programming. *Medical image analysis*, 12(6):731–741, 2008.

Ben Glocker, Aristeidis Sotiras, Nikos Komodakis, and Nikos Paragios. Deformable medical image registration: setting the state of the art with discrete methods. *Annual review of biomedical engineering*, 13(1):219–244, 2011.

Hui Gong, Dongli Xu, Jing Yuan, Xiangning Li, Congdi Guo, Jie Peng, Yuxin Li, Lindsay A Schwarz, Anan Li, Bihe Hu, et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nature communications*, 7(1):12142, 2016a.

Hui Gong, Dongli Xu, Jing Yuan, Xiangning Li, Congdi Guo, Jie Peng, Yuxin Li, Lindsay A Schwarz, Anan Li, Bihe Hu, et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nature communications*, 7(1):12142, 2016b.

Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, David Hutchison, Takeo Kanade, Josef Kittler, Jon M Kleinberg, Friedemann Mattern, John C Mitchell, Moni Naor, Oscar Nierstrasz, C Pandu Rangan, and Bernhard Steffen. Geodesic Shooting and Diffeomorphic Matching Via Textured Meshes. page 671.

Vivek Gopalakrishnan, Neel Dey, and Polina Golland. Polypose: Localizing deformable anatomy in 3d from sparse 2d x-ray images using polyrigid transforms. *ArXiv*, pages arXiv–2505, 2025.

Maged Goubran, Cathie Crukley, Sandrine De Ribaupierre, Terence M Peters, and Ali R Khan. Image registration of ex-vivo mri to sparsely sectioned histology of hippocampal and neocortical temporal lobe

specimens. *Neuroimage*, 83:770–781, 2013.

Alexandre Guimond, Alexis Roche, Nicholas Ayache, and Jean Meunier. Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections. *IEEE transactions on medical imaging*, 20(1):58–69, 2002.

Courtney K Guo. *Multi-modal image registration with unsupervised deep learning*. PhD thesis, Massachusetts Institute of Technology, 2019.

Tao Guo, Yinuo Wang, Shihao Shu, Weimin Yuan, Diansheng Chen, Zhouping Tang, Cai Meng, and Xiangzhi Bai. Mambamorph: a mamba-based framework for medical mr-ct deformable registration. *arXiv preprint arXiv:2401.13934*, 2024.

Tripti Gupta, Gregory D Marquart, Eric J Horstick, Kathryn M Tabor, Sinisa Pajevic, and Harold A Burgess. Morphometric analysis and neuroanatomical mapping of the zebrafish brain. *Methods*, 150:49–62, 2018a.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018b.

Eldad Haber and Jan Modersitzki. Numerical methods for volume preserving image registration. *Inverse problems*, 20(5):1621, 2004.

Hadamard. Sur les transformations ponctuelles. *Bulletin de la Société Mathématique de France*, 34:71–84, 1906. URL http://eudml.org/doc/86165.

Brian C Hall. An elementary introduction to groups and representations. *arXiv preprint math-ph/0005032*, 2000.

Brian C Hall and Brian C Hall. *Lie groups, Lie algebras, and representations*. Springer, 2013.

Lasse Hansen and Mattias P. Heinrich. GraphRegNet: Deep Graph Regularisation Networks on Sparse Keypoints for Dense Registration of 3D Lung CTs. *IEEE Transactions on Medical Imaging*, 40(9): 2246–2257, September 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3073986. Conference Name: IEEE Transactions on Medical Imaging.

Mattias P Heinrich and Lasse Hansen. Voxelmorph++ going beyond the cranial vault with keypoint supervision and multi-channel instance optimisation. In *International Workshop on Biomedical Image Registration*, pages 85–95. Springer, 2022.

Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16:1423–1435, October 2012. ISSN 1361-8415.

Mattias P Heinrich, Bartlomiej W Papież, Julia A Schnabel, and Heinz Handels. Non-parametric discrete registration with convex optimisation. In *International Workshop on Biomedical Image Registration*, pages

51–61. Springer, 2014.

Mattias P. Heinrich, Heinz Handels, and Ivor J. A. Simpson. Estimating Large Lung Motion in COPD Patients by Symmetric Regularised Correspondence Fields. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 338–345, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24571-3. doi: 10.1007/978-3-319-24571-3_41.

Alessa Hering, Keelin Murphy, and Bram van Ginneken. Learn2reg challenge: Ct lung registration - training data, May 2020. URL https://doi.org/10.5281/zenodo.3835682.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

Wiebke Heyer, Yannic Elser, Lennart Berkel, Xinrui Song, Xuanang Xu, Pingkun Yan, Xi Jia, Jinming Duan, Zi Li, Tony CW Mok, et al. Oncoreg: Medical image registration for oncological challenges. *arXiv preprint arXiv:2503.23179*, 2025.

Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

Joel Honkamaa and Pekka Marttinen. Sitreg: Multi-resolution architecture for symmetric, inverse consistent, and topology preserving image registration. *arXiv preprint arXiv:2303.10211*, 2023.

Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pages 3–17. Springer, 2021.

Junhao Hu, Weijie Gan, Zhixin Sun, Hongyu An, and Ulugbek S. Kamilov. A Plug-and-Play Image Registration Network, March 2024. URL http://arxiv.org/abs/2310.04297. arXiv:2310.04297 [eess].

Shiqi Huang, Tingfa Xu, Ziyi Shen, Shaheer Ullah Saeed, Wen Yan, Dean Barratt, and Yipeng Hu. Samreg: Sam-enabled image registration with roi-based correspondence. *arXiv preprint arXiv:2410.14083*, 2024.

Yuankai Huo, Zhoubing Xu, Yunxi Xiong, Katherine Aboud, Prasanna Parvathaneni, Shunxing Bao, Camilo Bermudez, Susan M Resnick, Laurie E Cutting, and Bennett A Landman. 3d whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119, 2019.

Istvan N Huszar, Menuka Pallebage-Gamarallage, Sarah Bangerter-Christensen, Hannah Brooks, Sean Fitzgibbon, Sean Foxley, Marlies Hiemstra, Amy FD Howard, Saad Jbabdi, Daniel ZL Kor, et al. Tensor image registration library: Deformable registration of stand-alone histology images to whole-brain post-

mortem mri data. *Neuroimage*, 265:119792, 2023.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Reza Yazdani Aminadabi, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. System optimizations for enabling training of extreme long sequence transformer models. In *Proceedings of the 43rd ACM Symposium on Principles of Distributed Computing*, PODC '24, page 121–130, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706684. doi: 10.1145/3662158.3662806. URL https://doi.org/10.1145/3662158.3662806.

Arpit Jadon, Haoran Wang, Phillip Thomas, Michael Stanley, S Nathaniel Cibik, Rachel Laurat, Omar Maher, Lukas Hoyer, Ozan Unal, and Dengxin Dai. Realdrivesim: A realistic multi-modal multi-task synthetic dataset for autonomous driving. *arXiv preprint arXiv:2506.16319*, 2025.

Ales Jaklic and Franc Solina. Moments of superellipsoids and their application to range image registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(4):648–657, 2003.

JAX. Autodiff cookbook. https://jax.readthedocs.io/en/latest/notebooks/autodiff_cookbook.html#vector-jacobian-products-vjps-aka-reverse-mode-autodiff.

Rohit Jena, Deeksha Sethi, Pratik Chaudhari, and James C Gee. Deep learning in medical image registration: Magic or mirage? *arXiv preprint arXiv:2408.05839*, 2024.

Rohit Jena, Pratik Chaudhari, and James C. Gee. Deep implicit optimization enables robust learnable features for deformable image registration. *Medical Image Analysis*, 103:103577, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103577. URL https://www.sciencedirect.com/science/article/pii/S1361841525001240.

Rohit Jena, Pratik Chaudhari, and James C Gee. Fireants: Adaptive riemannian optimization for multi-scale diffeomorphic registration. *Nature Communications*, 2026.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Di Jia, Kai Wang, ShunLi Luo, TianYu Liu, and Ying Liu. Braft: Recurrent all-pairs field transforms for optical flow based on correlation blocks. *IEEE Signal Processing Letters*, 28:1575–1579, 2021.

Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? *arXiv preprint arXiv:2208.04939*, 2022.

Xi Jia, Joseph Bartlett, Wei Chen, Siyang Song, Tianyang Zhang, Xinxing Cheng, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. Fourier-net: Fast image registration with band-limited deformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1015–1023, 2023.

Xi Jia et al. A naive trick to accelerate training of lncc-based deep image registration models. *Preprints*, February 2025. doi: 10.20944/preprints202502.2200.v1.

Bailiang Jian, Jiazhen Pan, Morteza Ghahremani, Daniel Rueckert, Christian Wachinger, and Benedikt Wiestler. Mamba? catch the hype or rethink what really helps for image registration. *arXiv preprint arXiv:2407.19274*, 2024.

Bailiang Jian, Jiazhen Pan, Rohit Jena, Morteza Ghahremani, Hongwei Bran Li, Daniel Rueckert, Christian Wachinger, and Benedikt Wiestler. Disentangling progress in medical image registration: Beyond trend-driven architectures towards domain-specific strategies. *arXiv preprint arXiv:2512.01913*, 2025.

Di Jiang, Yuhui Du, Hewei Cheng, Tianzi Jiang, and Yong Fan. Groupwise spatial normalization of fmri data based on multi-range functional connectivity patterns. *Neuroimage*, 82:355–372, 2013.

Hans J Johnson and Gary E Christensen. Landmark and intensity-based, consistent thin-plate spline image registration. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 329–343. Springer, 2001.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Ankita Joshi and Yi Hong. Diffeomorphic Image Registration using Lipschitz Continuous Residual Networks. page 13.

Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

Miao Kang, Xiaojun Hu, Weilin Huang, Matthew R Scott, and Mauricio Reyes. Dual-stream pyramid registration network. *Medical image analysis*, 78:102379, 2022.

Justin W Kenney, Patrick E Steadman, Olivia Young, Meng Ting Shi, Maris Polanco, Saba Dubaishi, Kristopher Covert, Thomas Mueller, and Paul W Frankland. A 3d adult zebrafish brain atlas (azba) for the digital age. *Elife*, 10:e69988, 2021a.

Justin W Kenney, Patrick E Steadman, Olivia Young, Meng Ting Shi, Maris Polanco, Saba Dubaishi, Kristopher Covert, Thomas Mueller, and Paul W Frankland. A 3d adult zebrafish brain atlas (azba) for the digital age. *Elife*, 10:e69988, 2021b.

Pulkit Khandelwal, Michael Tran Duong, Lisa M Levorse, Sydney A Lim, Amanda E Denning, Nathaniel Gauthier, Ved Shenoy, Winifred Trotman, Ranjit Ittyerah, Alejandra Bahena, et al. High-resolution postmortem 7 tesla mri yields localized atrophy measures that are more sensitive to tau pathology and neuronal loss in alzheimer's disease than corresponding measures on antemortem 3 tesla mri. *Alzheimer's & Dementia*, 21:e098995, 2025.

Boah Kim, Jieun Kim, June-Goo Lee, Dong Hwan Kim, Seong Ho Park, and Jong Chul Ye. Unsupervised deformable image registration using cycle-consistent cnn. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 166–174. Springer, 2019.

Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis*, 71:102036, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson, Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, July 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.12.037. URL https://www.sciencedirect.com/science/article/pii/S1053811908012974.

Arno Klein, Tito Dal Canton, Satrajit S Ghosh, Bennett Landman, Joel Lee, and Andrew Worth. Open labels: online feedback for a public resource of manually labeled brain images. In *16th annual meeting for the organization of human brain mapping*, volume 84358, page 6, 2010.

David Kleinfeld, Arjun Bharioke, Pablo Blinder, Davi D Bock, Kevin L Briggman, Dmitri B Chklovskii, Winfried Denk, Moritz Helmstaedter, John P Kaufhold, Wei-Chung Allen Lee, et al. Large-scale automated histology in the pursuit of connectomes. *Journal of Neuroscience*, 31(45):16125–16138, 2011.

Heidi Kleven, Ingvild E Bjerke, Francisco Clascá, Henk J Groenewegen, Jan G Bjaalie, and Trygve B Leergaard. Waxholm space atlas of the rat brain: a 3d atlas supporting data analysis and integration. *Nature methods*, 20(11):1822–1829, 2023.

Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in pytorch, 2020.

Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5:341–353, 2023.

Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 344–352. Springer, 2017.

Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019.

Fae A Kronman, Josephine K Liwang, Rebecca Betty, Daniel J Vanselow, Yuan-Ting Wu, Nicholas J Tustison, Ashwin Bhandiwad, Steffy B Manjila, Jennifer A Minteer, Donghui Shin, et al. Developmental mouse

brain common coordinate framework. *bioRxiv*, 2023.

Fae N Kronman, Josephine K Liwang, Rebecca Betty, Daniel J Vanselow, Yuan-Ting Wu, Nicholas J Tustison, Ashwin Bhandiwad, Steffy B Manjila, Jennifer A Minteer, Donghui Shin, et al. Developmental mouse brain common coordinate framework. *Nature communications*, 15(1):9072, 2024.

Kwame S. Kutten, Joshua T. Vogelstein, Nicolas Charon, Li Ye, Karl Deisseroth M.D., and Michael I. Miller. Deformably registering and annotating whole CLARITY brains to an atlas via masked LDDMM. In Peter Schelkens, Touradj Ebrahimi, Gabriel Cristóbal, Frédéric Truchetet, and Pasi Saarikko, editors, *Optics, Photonics and Digital Technologies for Imaging Applications IV*, volume 9896, page 989616. International Society for Optics and Photonics, SPIE, 2016. doi: 10.1117/12.2227444. URL https://doi.org/10.1117/12.2227444.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Joel Lamy-Poirier. Breadth-first pipeline parallelism. *Proceedings of Machine Learning and Systems*, 5: 48–67, 2023.

Jack L Lancaster, Diana Tordesillas-Gutiérrez, Michael Martinez, Felipe Salinas, Alan Evans, Karl Zilles, John C Mazziotta, and Peter T Fox. Bias between mni and talairach coordinates analyzed using the icbm-152 brain template. *Human brain mapping*, 28(11):1194–1205, 2007.

Leo Lebrat, Rodrigo Santa Cruz, Frederic de Gournay, Darren Fu, Pierrick Bourgeat, Jurgen Fripp, Clinton Fookes, and Olivier Salvado. CorticalFlow: A Diffeomorphic Mesh Transformer Network for Cortical Surface Reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pages 29491–29505. Curran Associates, Inc., 2021. URL https://papers.nips.cc/paper/2021/hash/f6b5f8c32c65fee991049a55dc97d1ce-Abstract.html.

Jae Min Lee. *Geometry and analysis of some Euler-Arnold equations*. City University of New York, 2018.

Antoine Legouhy, Ross Callaghan, Hojjat Azadbakht, and Hui Zhang. Polaffini: Efficient feature-based polyaffine initialization for improved non-linear image registration. In *International Conference on Information Processing in Medical Imaging*, pages 614–625. Springer, 2023.

JA Leslie. On a differential structure for the group of diffeomorphisms. *Topology*, 6(2):263–271, 1967.

Dacheng Li, Rulin Shao, Anze Xie, Eric P. Xing, Xuezhe Ma, Ion Stoica, Joseph E. Gonzalez, and Hao Zhang. DISTFLASHATTN: Distributed memory-efficient attention for long-context LLMs training. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=pUEDkZyPDl.

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/

v1/2023.acl-long.134. URL https://aclanthology.org/2023.acl-long.134/.

Zi Li, Lin Tian, Tony CW Mok, Xiaoyu Bai, Puyang Wang, Jia Ge, Jingren Zhou, Le Lu, Xianghua Ye, Ke Yan, et al. Samconvex: Fast discrete optimization for ct registration using self-supervised anatomical embedding and correlation pyramid. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–569. Springer, 2023b.

Cher-Wei Liang, Ruey-Feng Chang, Pei-Wei Fang, and Chiao-Min Chen. Improving algorithm for the alignment of consecutive, whole-slide, immunohistochemical section images. *Journal of Pathology Informatics*, 12(1):29, 2021.

Devavrat Likhite, Ganesh Adluru, and Edward DiBella. Deformable and rigid model-based image registration for quantitative cardiac perfusion. In *Statistical Atlases and Computational Models of the Heart-Imaging and Modelling Challenges: 5th International Workshop, STACOM 2014, Held in Conjunction with MICCAI 2014, Boston, MA, USA, September 18, 2014, Revised Selected Papers 5*, pages 41–50. Springer, 2015.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Fengze Liu, Ke Yan, Adam P. Harrison, Dazhou Guo, Le Lu, Alan L. Yuille, Lingyun Huang, Guotong Xie, Jing Xiao, Xianghua Ye, and Dakai Jin. SAME: Deformable Image Registration Based on Self-supervised Anatomical Embeddings. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 87–97, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-87202-1. doi: 10.1007/978-3-030-87202-1_9.

Fengze Liu, Ke Yan, Adam P Harrison, Dazhou Guo, Le Lu, Alan L Yuille, Lingyun Huang, Guotong Xie, Jing Xiao, Xianghua Ye, et al. Same: Deformable image registration based on self-supervised anatomical embeddings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2021b.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=WsRHpHH4s0.

Hengjie Liu, Dan Ruan, and Ke Sheng. Unsupervised deformable image registration revisited: Enhancing performance with registration-specific designs. In *Medical Imaging with Deep Learning-Short Papers*, 2025a.

Peirong Liu, Oula Puonti, Xiaoling Hu, Karthik Gopinath, Annabel Sorby-Adams, Daniel C Alexander, W Taylor Kimberly, and Juan E Iglesias. A modality-agnostic multi-task foundation model for human brain imaging. *arXiv preprint arXiv:2509.00549*, 2025b.

Yihao Liu, Lianrui Zuo, Shuo Han, Yuan Xue, Jerry L Prince, and Aaron Carass. Coordinate translator for learning deformable medical image registration. In *International workshop on multiscale multimodal medical imaging*, pages 98–109. Springer, 2022.

Yihao Liu, Junyu Chen, Lianrui Zuo, Aaron Carass, and Jerry L Prince. Vector field attention for deformable image registration. *Journal of Medical Imaging*, 11(6):064001–064001, 2024c.

Josephine K Liwang, Hannah C Bennett, Hyun-Jae Pi, and Yongsoo Kim. Protocol for using serial two-photon tomography to map cell types and cerebrovasculature at single-cell resolution in the whole adult mouse brain. *STAR protocols*, 4(1):102048, 2023.

Josephine K Liwang, Fae N Kronman, Hyun-Jae Pi, Yuan-Ting Wu, Daniel J Vanselow, Steffy B Manjila, Deniz Parmaksiz, Donghui Shin, Yoav Ben-Simon, Michael Taormina, et al. epdevatlas: mapping gabaergic cells and microglia in the early postnatal mouse brain. *Nature Communications*, 16(1):9538, 2025.

Johannes Lotz, Janine Olesch, Benedikt Müller, Thomas Polzin, P Galuschka, JM Lotz, Stefan Heldmann, Hendrik Laue, Margarita González-Vallinas, Arne Warth, et al. Patch-based nonlinear image registration for gigapixel whole slide images. *IEEE Transactions on Biomedical Engineering*, 63(9):1812–1819, 2015.

Zhichao Lu, Ian Whalen, Vishnu Boddeti, Yashesh Dhebar, Kalyanmoy Deb, Erik Goodman, and Wolfgang Banzhaf. Nsga-net: neural architecture search using multi-objective genetic algorithm. In *Proceedings of the genetic and evolutionary computation conference*, pages 419–427, 2019.

Falk Lüsebrink, Alessandro Sciarra, Hendrik Mattern, Renat Yakupov, and Oliver Speck. T1-weighted in vivo human whole brain mri dataset with an ultrahigh isotropic resolution of 250 μm. *Scientific data*, 4(1): 1–12, 2017.

Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

Mads AJ Madsen, Vanessa Wiggermann, Stephan Bramow, Jeppe Romme Christensen, Finn Sellebjerg, and Hartwig R Siebner. Imaging cortical multiple sclerosis lesions with ultra-high field mri. *NeuroImage: Clinical*, 32:102847, 2021.

Lucas Mahler, Julius Steiglechner, Benjamin Bender, Tobias Lindig, Dana Ramadan, Jonas Bause, Florian Birk, Rahel Heule, Edyta Charyasz, Michael Erb, Vinod Jangir Kumar, Gisela E Hagberg, Pascal Martin, Gabriele Lohmann, and Klaus Scheffler. "ultracortex: Submillimeter ultra-high field 9.4t brain mr image collection and manual cortical segmentations". 2024. doi: doi:10.18112/openneuro.ds005216.v1.1.0.

Fei Mai, CQ Chang, and YS Hung. A subspace approach for matching 2d shapes under affine distortions. *Pattern Recognition*, 44(2):210–221, 2011.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Matteo Mancini, Adrià Casamitjana, Loic Peter, Eleanor Robinson, Shauna Crampsie, David L Thomas, Janice L Holton, Zane Jaunmuktane, and Juan Eugenio Iglesias. A multimodal computational pipeline for 3d histology of the human brain. *Scientific reports*, 10(1):13839, 2020.

Andreas Mang. Claire: Scalable gpu-accelerated algorithms for diffeomorphic image registration in 3d. In *Explorations in the Mathematics of Data Science: The Inaugural Volume of the Center for Approximation and Mathematical Data Analytics*, pages 167–215. Springer, 2024.

Andreas Mang and George Biros. A semi-lagrangian two-level preconditioned newton–krylov solver for constrained diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 39(6):B1064–B1101, 2017.

Andreas Mang and Lars Ruthotto. A lagrangian gauss–newton–krylov solver for mass-and intensity-preserving diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 39(5):B860–B885, 2017a.

Andreas Mang and Lars Ruthotto. A lagrangian gauss–newton–krylov solver for mass-and intensity-preserving diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 39(5):B860–B885, 2017b.

Andreas Mang, Amir Gholami, Christos Davatzikos, and George Biros. CLAIRE: A distributed-memory solver for constrained large deformation diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 41(5):C548–C584, January 2019a. ISSN 1064-8275, 1095-7197. doi: 10.1137/18M1207818. URL http://arxiv.org/abs/1808.04487. arXiv:1808.04487 [cs, math].

Andreas Mang, Amir Gholami, Christos Davatzikos, and George Biros. CLAIRE: A distributed-memory solver for constrained large deformation diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 41(5):C548–C584, January 2019b. ISSN 1064-8275, 1095-7197. doi: 10.1137/18M1207818. URL http://arxiv.org/abs/1808.04487. arXiv:1808.04487 [cs, math].

Harrison Mansour, Ryan Azrak, James J Cook, Kathryn J Hornburg, Yi Qi, Yuqi Tian, Robert W Williams, Fang-Cheng Yeh, Leonard E White, and G Allan Johnson. The duke mouse brain atlas: Mri and light sheet microscopy stereotaxic atlas of the mouse brain. *Science Advances*, 11(18):eadq8089, 2025.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007a.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007b.

Gregory D Marquart, Kathryn M Tabor, Eric J Horstick, Mary Brown, Alexandra K Geoca, Nicholas F Polys, Damian Dalle Nogare, and Harold A Burgess. High-precision registration between zebrafish brain atlases

using symmetric diffeomorphic normalization. *GigaScience*, 6(8):gix056, 2017.

Stephen Marsland and Robert McLachlan. A hamiltonian particle method for diffeomorphic image registration. In *Biennial international conference on information processing in medical imaging*, pages 396–407. Springer, 2007.

Stephen Marsland and Carole J Twining. Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images. *IEEE transactions on medical imaging*, 23(8): 1006–1020, 2004.

David Mattes, David R Haynor, Hubert Vesselle, Thomas K Lewellyn, and William Eubank. Nonrigid multimodality image registration. In *Medical imaging 2001: image processing*, volume 4322, pages 1609–1620. Spie, 2001.

Joshua Menke and Tony R Martinez. Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 1331–1335. IEEE, 2004.

B Mertzios and M Christodoulou. On the generalized cayley-hamilton theorem. *IEEE transactions on automatic control*, 31(2):156–157, 1986.

Christopher Mezias, Bingxing Huo, Mihail Bota, Jaikishan Jayakumar, and Partha P Mitra. Establishing neuroanatomical correspondences across mouse and marmoset brain structures. *Research Square*, pages rs–3, 2024.

Michael P Milham, Lei Ai, Bonhwang Koo, Ting Xu, Céline Amiez, Fabien Balezeau, Mark G Baxter, Erwin LA Blezer, Thomas Brochier, Aihua Chen, et al. An open resource for non-human primate imaging. *Neuron*, 100(1):61–74, 2018a.

Michael P Milham, Lei Ai, Bonhwang Koo, Ting Xu, Céline Amiez, Fabien Balezeau, Mark G Baxter, Erwin LA Blezer, Thomas Brochier, Aihua Chen, et al. An open resource for non-human primate imaging. *Neuron*, 100(1):61–74, 2018b.

J. Milnor. *Remarks on infinite-dimensional Lie groups*. North-Holland., 1984.

Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010a.

Marc Modat, Tom Vercauteren, Gerard R Ridgway, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Diffeomorphic demons using normalized mutual information, evaluation on multimodal brain mr images. In *Medical Imaging 2010: Image Processing*, volume 7623, pages 800–807. SPIE, 2010b.

Jan Modersitzki. *FAIR: flexible algorithms for image registration*. SIAM, 2009.

Tony C. W. Mok and Albert C. S. Chung. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks, June 2020a. URL http://arxiv.org/abs/2006.16148. arXiv:2006.16148 [cs, eess].

Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020b.

Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022.

Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. pages 211–221, 2020c.

Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural network. pages 35–45, 2021.

Tony CW Mok, Zi Li, Yingda Xia, Jiawen Yao, Ling Zhang, Jingren Zhou, and Le Lu. Deformable medical image registration under distribution shifts with neural instance optimization. In *International Workshop on Machine Learning in Medical Imaging*, pages 126–136. Springer, 2023.

Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003.

Mohammadhossein Momeni, Vivek Gopalakrishnan, Neel Dey, Polina Golland, and Sarah Frisken. Differentiable voxel-based x-ray rendering improves sparse-view 3d cbct reconstruction. *arXiv preprint arXiv:2411.19224*, 2024.

Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. Evaluation of registration methods on thoracic ct: the empire10 challenge. *IEEE transactions on medical imaging*, 30(11):1901–1920, 2011.

Abdullah Nazib, James Galloway, Clinton Fookes, and Dimitri Perrin. Performance of registration tools on high-resolution 3d brain images. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 566–569, 2018. doi: 10.1109/EMBC.2018.8512403.

Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. Metric learning for image registration. June 2019.

NirutaDhimal and ANTsX/ANTsPy contributors. Gpu support (how to make registration faster) — issue #441. GitHub issue, ANTsX/ANTsPy repository, March 2023. URL https://github.com/ANTsX/ANTsPy/issues/441. Closed issue discussing feature request for GPU support in registration.

Tazi Nouamane, Mom Ferdinand, Zhao Haojun, Nguyen Phuc, Mekkouri Mohamed, Werra Leandro, and

Wolf Thomas. The ultra-scale playbook: Training llms on gpu clusters, 2025.

OpenAI. Triton. https://openai.com/index/triton/, 2021.

John Pellman, Nick Tustison, Catherine Tallman, and ANTs Discussion Participants. Gpu support? SourceForge discussion, Advanced Normalization Tools (ANTs), July 2016. URL https://sourceforge.net/p/advants/discussion/840260/thread/4b134259/. Thread updated 2019-08-09.

Hanchuan Peng, Phuong Chung, Fuhui Long, Lei Qu, Arnim Jenett, Andrew M Seeds, Eugene W Myers, and Julie H Simpson. Brainaligner: 3d registration atlases of drosophila brains. *Nature methods*, 8(6):493–498, 2011.

Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 597–605. Springer, 1999.

Xavier Pennec, Radu Stefanescu, Vincent Arsigny, Pierre Fillard, and Nicholas Ayache. Riemannian elasticity: A statistical regularization framework for non-linear registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 943–950. Springer, 2005.

Javier Pérez de Frutos, André Pedersen, Egidijus Pelanis, David Bouget, Shanmugapriya Survarachakan, Thomas Langø, Ole-Jakob Elle, and Frank Lindseth. Learning deep abdominal ct registration through adaptive loss weighting and synthetic data generation. *Plos one*, 18(2):e0282110, 2023.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34: 29218–29230, 2021.

Jonas Pichat, Eugenio Iglesias, Sotiris Nousias, Tarek Yousry, Sébastien Ourselin, and Marc Modat. Part-to-whole registration of histology and mri using shape elements. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 107–115, 2017.

Michael JD Powell et al. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26(26-46):1, 2009.

Stephen C Preston. For ideal fluids, eulerian and lagrangian instabilities are equivalent. *Geometric and Functional Analysis*, 14(5):1044–1062, 2004.

Oula Puonti, Jackson Nolan, Robert Dicamillo, Yael Balbastre, Adria Casamitjana, MatteoMancini, Eleanor Robinson, Loic Peter, Roberto Annunziata, Juri Althonayan, ShaunaCrampsie, Emily Blackburn, Benjamin Billot, Alessia Atzeni, Peter Schmidt, James Hughes, Jean Augustinack, Brian Edlow, Lilla Zöllei, David Thomas, Dorit Kliemann, MartinaBocchetta, Catherine Strand, Janice Holton, Zane Jaunmuktane, and Juan Eugenio Iglesias. An open-source tool for fast segmentation of any brain mr scan with the nextbrain histological atlas. In *31th Annual Meeting of the Organization for Human Brain Mapping (OHBM 2025)*, 2025.

PyTorch. Fusing convolution and batch norm using custom function. https://docs.pytorch.org/tutorials/intermediate/custom_function_conv_bn_tutorial.html, 2023. Created July 22, 2021; Last updated April 18, 2023; Last verified November 5, 2024.

Penghui Qi, Xinyi Wan, Guangxing Huang, and Min Lin. Zero bubble pipeline parallelism. *arXiv preprint arXiv:2401.10241*, 2023.

Chen Qin, Shuo Wang, Chen Chen, Huaqi Qiu, Wenjia Bai, and Daniel Rueckert. Biomechanics-informed Neural Networks for Myocardial Motion Tracking in MRI, July 2020. URL http://arxiv.org/abs/2006.04725. arXiv:2006.04725 [cs, eess].

Chen Qin, Shuo Wang, Chen Chen, Wenjia Bai, and Daniel Rueckert. Generative Myocardial Motion Tracking via Latent Space Exploration with Biomechanics-informed Prior, June 2022. URL http://arxiv.org/abs/2206.03830. arXiv:2206.03830 [cs, eess].

Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. 2021.

Huaqi Qiu, Kerstin Hammernik, Chen Qin, Chen Chen, and Daniel Rueckert. Embedding gradient-based optimization in image registration networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2022.

Xin Qiu, Yulu Gan, Conor F Hayes, Qiyao Liang, Yinggan Xu, Roberto Dailey, Elliot Meyerson, Babak Hodjat, and Risto Miikkulainen. Evolution strategies at scale: Llm fine-tuning beyond reinforcement learning. *arXiv preprint arXiv:2509.24372*, 2025.

Dou Quan, Huiyuan Wei, Shuang Wang, Ruiqi Lei, Baorui Duan, Yi Li, Biao Hou, and Licheng Jiao. Self-distillation feature learning network for optical and sar image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

Richard D Rabbitt, Jeffrey A Weiss, Gary E Christensen, and Michael I Miller. Mapping of hyperelastic deformable templates using the finite element method. In *Vision Geometry IV*, volume 2573, pages 252–265. SPIE, 1995.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16, 2020. doi: 10.1109/SC41405.2020.00024.

Owen Randlett, Caroline L Wee, Eva A Naumann, Onyeka Nnaemeka, David Schoppik, James E Fitzgerald, Ruben Portugues, Alix MB Lacoste, Clemens Riegler, Florian Engert, et al. Whole-brain activity mapping onto a zebrafish brain atlas. *Nature methods*, 12(11):1039–1046, 2015.

S. Ravikumar, L. E. M. Wisse, R. Ittyerah, S. Lim, M. Lavery, L. Xie, J. L. Robinson, T. Schuck, M. Grossman, E. B. Lee, M. D. Tisdall, K. Prabhakaran, J. A. Detre, S. R. Das, G. Mizsei, E. Artacho-Pérula, M. M. I. de Onzono Martin, M. del Mar Arroyo Jiménez, M. Muñoz, F. J. M. Romero, M. del Pilar Marcos Rabal,

D. J. Irwin, J. Q. Trojanowski, D. A. Wolk, R. Insausti, and P. A. Yushkevich. Building an ex vivo atlas of the earliest brain regions affected by alzheimer's disease pathology. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 113–117, 2020.

Sadhana Ravikumar, Amanda E Denning, Sydney Lim, Eunice Chung, Niyousha Sadeghpour, Ranjit Ittyerah, Laura EM Wisse, Sandhitsu R Das, Long Xie, John L Robinson, et al. Postmortem imaging reveals patterns of medial temporal lobe vulnerability to tau pathology in alzheimer's disease. *Nature Communications*, 15 (1):4803, 2024.

Laurent Risser, François-Xavier Vialard, Robin Wolz, Maria Murgasova, Darryl D Holm, and Daniel Rueckert. Simultaneous multi-scale registration using large deformation diffeomorphic metric mapping. *IEEE transactions on medical imaging*, 30(10):1746–1759, 2011.

Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: learning deformable image registration using shape matching. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 266–274. Springer, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

J-M Rouet, J-J Jacq, and Christian Roux. Genetic algorithms for a robust 3-d mr-ct registration. *IEEE transactions on information technology in biomedicine*, 4(2):126–136, 2002.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Aliaksandr Siarohin. cuda-gridsample-grad2. GitHub Repository, 2023. URL https://github.com/AliaksandrSiarohin/cuda-gridsample-grad2.

Hanna Siebert, Christoph Großbröhmer, Lasse Hansen, and Mattias P Heinrich. Convexadam: Self-configuring dual-optimisation-based 3d multitask medical image registration. *IEEE Transactions on Medical Imaging*, 2024.

Vignesh Sivan, Teodora Vujovic, Raj Ranabhat, Alexander Wong, Stewart Mclachlin, and Michael Hardisty. Recurrence with correlation network for medical image registration. *arXiv preprint arXiv:2302.02283*, 2023.

Henrik Skibbe, Muhammad Febrian Rachmadi, Ken Nakae, Carlos Enrique Gutierrez, Junichi Hata, Hiromichi Tsukada, Charissa Poon, Matthias Schlachter, Kenji Doya, Piotr Majka, et al. The brain/minds marmoset connectivity resource: An open-access platform for cellular-level tracing and tractography in the primate brain. *PLoS biology*, 21(6):e3002158, 2023.

Ronald WK So, Tommy WH Tang, and Albert CS Chung. Non-rigid image registration of brain magnetic resonance images using graph-cuts. *Pattern Recognition*, 44(10-11):2450–2467, 2011.

Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 232–239. Springer, 2017.

Stefan Sommer, Mads Nielsen, François Lauze, and Xavier Pennec. A multi-scale kernel bundle for lddmm: Towards sparse deformation description across space and scales. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 624–635. Springer, 2011.

Xinrui Song, Xuanang Xu, and Pingkun Yan. Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 608–617. Springer, 2024.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Benjamin Frederick Spector, Simran Arora, Aaryan Singhal, Arjun Parthasarathy, Daniel Y Fu, and Christopher Re. Thunderkittens: Simple, fast, and $\textit{Adorable}$ kernels. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0fJfVOSUra.

Kathryn M Tabor, Gregory D Marquart, Christopher Hurt, Trevor S Smith, Alexandra K Geoca, Ashwin A Bhandiwad, Abhignya Subedi, Jennifer L Sinclair, Hannah M Rose, Nicholas F Polys, et al. Brain-wide cellular resolution imaging of cre transgenic zebrafish lines for functional circuit-mapping. *Elife*, 8:e42687, 2019.

Gabriel Taubin and David B Cooper. Recognition and positioning of rigid objects using algebraic moment invariants. In *Geometric Methods in Computer Vision*, volume 1570, pages 175–186. SPIE, 1991.

National Lung Screening Trial Research Team. The national lung screening trial: overview and study design.

Radiology, 258(1):243–253, 2011.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, August 2020. URL http://arxiv.org/abs/2003.12039. arXiv:2003.12039 [cs].

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. In *Advances in Neural Information Processing Systems*, volume 33, pages 3362–3373. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2290a7385ed77cc5592dc2153229f082-Abstract.html.

Philippe Thévenaz and Michael Unser. Optimization of mutual information for multiresolution image registration. *IEEE transactions on image processing*, 9(12):2083–2099, 2000.

J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, Richard Jarrett Rushmore, Nikolaos Makris, Sylvain Bouix, and Marc Niethammer. Gradicon: Approximate diffeomorphisms via gradient inverse consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18084–18094, 2023a.

Lin Tian, Zi Li, Fengze Liu, Xiaoyu Bai, Jia Ge, Le Lu, Marc Niethammer, Xianghua Ye, Ke Yan, and Daikai Jin. SAME++: A Self-supervised Anatomical eMbeddings Enhanced medical image registration framework using stable sampling and regularized transformation, November 2023b. URL http://arxiv.org/abs/2311.14986. arXiv:2311.14986 [cs].

Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, Sylvain Bouix, Richard Rushmore, and Marc Niethammer. unigradicon: A foundation model for medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 749–760. Springer, 2024.

Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Lazaros C Triarhou. Dopamine and parkinson's disease. In *Madame curie bioscience database [internet]*. Landes Bioscience, 2013.

Nicholas J Tustison, Hans J Johnson, Torsten Rohlfing, Arno Klein, Satrajit S Ghosh, Luis Ibanez, and Brian B Avants. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences, 2013.

Nicholas J Tustison, Andrew J Holbrook, Brian B Avants, Jared M Roberts, Philip A Cook, Zachariah M Reagh, Jeffrey T Duda, James R Stone, Daniel L Gillen, Michael A Yassa, et al. Longitudinal mapping of cortical thickness measurements: An alzheimer's disease neuroimaging initiative-based evaluation study. *Journal of Alzheimer's Disease*, 71(1):165–183, 2019.

Nicholas J Tustison, Philip A Cook, Andrew J Holbrook, Hans J Johnson, John Muschelli, Gabriel A Devenyi, Jeffrey T Duda, Sandhitsu R Das, Nicholas C Cullen, Daniel L Gillen, et al. The antsx ecosystem for quantitative biological and medical imaging. *Scientific reports*, 11(1):9068, 2021.

Nicholas J Tustison, Min Chen, Fae N Kronman, Jeffrey T Duda, Clare Gamlin, Mia G Tustison, Michael Kunst, Rachel Dalley, Staci Sorenson, Quanxin Wang, et al. The antsx ecosystem for mapping the mouse brain. *bioRxiv*, pages 2024–05, 2024.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. *International Journal of Computer Vision*, 128(7):1867–1888, July 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-020-01303-4. URL http://arxiv.org/abs/1711.10925. arXiv:1711.10925 [cs, stat].

Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 223–231. Springer, 2017.

Charles Van Loan. The sensitivity of the matrix exponential. *SIAM Journal on Numerical Analysis*, 14(6): 971–981, 1977.

Erdem Varol, Amin Nejatbakhsh, Ruoxi Sun, Gonzalo Mena, Eviatar Yemini, Oliver Hobert, and Liam Paninski. Statistical atlas of c. elegans neurons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 119–129. Springer, 2020.

Vivek Venkatachalam, Ni Ji, Xian Wang, Christopher Clark, James Kameron Mitchell, Mason Klein, Christopher J Tabone, Jeremy Florman, Hongfei Ji, Joel Greenwood, et al. Pan-neuronal imaging in roaming caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8):E1082–E1088, 2016.

Tom Vercauteren, Xavier Pennec, Ezio Malis, Aymeric Perchant, and Nicholas Ayache. Insight into efficient image registration techniques and the demons algorithm. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 495–506. Springer, 2007a.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, Nicholas Ayache, et al. Diffeomorphic demons using itk's finite difference solver hierarchy. *The Insight Journal*, 1, 2007b.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Symmetric Log-Domain Diffeomorphic Registration: A Demons-Based Approach. In Dimitris Metaxas, Leon Axel, Gabor Fichtinger, and Gábor Székely, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, Lecture Notes in Computer Science, pages 754–761, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-85988-8. doi: 10.1007/978-3-540-85988-8_90.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, March 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2008.10.040. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811908011683.

Alan Q Wang, M Yu Evan, Adrian V Dalca, and Mert R Sabuncu. A robust and interpretable deep learning framework for multi-modal registration via keypoints. *Medical Image Analysis*, 90:102962, 2023.

Guoxia Wang, Jinle Zeng, Xiyuan Xiao, Siming Wu, Jiabin Yang, Lujing Zheng, Zeyu Chen, Jiang Bian, Dianhai Yu, and Haifeng Wang. Flashmask: Efficient and rich mask extension of flashattention. *arXiv preprint arXiv:2410.01359*, 2024.

Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, Tim Dolbeare, Brandon Blanchard, Nick Dee, Wayne Wakeman, Karla E. Hirokawa, Aaron Szafer, Susan M. Sunkin, Seung Wook Oh, Amy Bernard, John W. Phillips, Michael Hawrylycz, Christof Koch, Hongkui Zeng, Julie A. Harris, and Lydia Ng. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, 181(4):936–953.e20, May 2020a. ISSN 00928674. doi: 10.1016/j.cell.2020.04.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420304025.

Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell*, 181(4):936–953, 2020b.

Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell*, 181(4):936–953, 2020c.

Yongmei Wang and Lawrence H Staib. Physical model-based non-rigid registration incorporating statistical shape information. *Medical image analysis*, 4(1):7–20, 2000.

Asmamaw T Wassie, Yongxin Zhao, and Edward S Boyden. Expansion microscopy: principles and uses in biological research. *Nature methods*, 16(1):33–41, 2019.

Thomas Welton, Septian Hartono, Yao-Chia Shih, Stefan T Schwarz, Yue Xing, Eng-King Tan, Dorothee P Auer, Noam Harel, and Ling-Ling Chan. Ultra-high-field 7t mri in parkinson's disease: ready for clinical use?—a narrative review. *Quantitative Imaging in Medicine and Surgery*, 13(11):7607, 2023.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. In *Hassler Whitney Collected Papers*, pages 228–254. Springer, 1992.

Marek Wodzinski, Niccolo Marini, Manfredo Atzori, and Henning Müller. Deeperhistreg: robust whole slide images registration framework. *arXiv preprint arXiv:2404.14434*, 2024.

Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit Neural Representations for Deformable Image Registration. page 11.

Guorong Wu, Minjeong Kim, Qian Wang, Yaozong Gao, Shu Liao, and Dinggang Shen. Unsupervised deep feature learning for deformable registration of mr brain images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 649–656. Springer, 2013.

Guorong Wu, Minjeong Kim, Qian Wang, Brent C Munsell, and Dinggang Shen. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE transactions on biomedical engineering*, 63(7):1505–1516, 2015.

Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. *arXiv preprint arXiv:2011.02538*, 2020.

Yifan Wu, Tom Z. Jiahao, Jiancong Wang, Paul A. Yushkevich, M. Ani Hsieh, and James C. Gee. NODEO: A Neural Ordinary Differential Equation Based Optimization Framework for Deformable Image Registration. *arXiv:2108.03443 [cs]*, February 2022a. URL http://arxiv.org/abs/2108.03443. arXiv: 2108.03443.

Yifan Wu, Mengjin Dong, Rohit Jena, Chen Qin, and James C Gee. Neural ordinary differential equation based sequential image registration for dynamic characterization. *arXiv preprint arXiv:2404.02106*, 2024.

Yingjuan Wu, Abdur Raquib Ridwan, Mohammad Rakeen Niaz, Xiaoxiao Qi, Shengwei Zhang, null Alzheimer's Disease Neuroimaging Initiative, David A. Bennett, and Konstantinos Arfanakis. Development of high quality T1-weighted and diffusion tensor templates of the older adult brain in a common space. *NeuroImage*, 260:119417, October 2022b. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2022.119417.

Xinyi Xu, Cong Sun, Jiwei Sun, Wen Shi, Yao Shen, Ruoke Zhao, Wanrong Luo, Mingyang Li, Guangbin Wang, and Dan Wu. Spatiotemporal Atlas of the Fetal Brain Depicts Cortical Developmental Gradient. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 42(50):9435–9449, December 2022. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1285-22.2022.

Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. Cars: Continuous evolution for efficient neural architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1829–1838, 2020.

Zhengwei Yang and Fernand S Cohen. Cross-weighted moments and affine invariants for image registration and matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):804–814, 1999.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10665–10673, 2021.

Michael A Yassa, Shauna M Stark, Arnold Bakker, Marilyn S Albert, Michela Gallagher, and Craig EL Stark. High-resolution structural and functional mri of hippocampal ca3 and dentate gyrus in patients with amnestic mild cognitive impairment. *Neuroimage*, 51(3):1242–1252, 2010.

Laurent Younes. *Shapes and diffeomorphisms*, volume 171. Springer, 2010.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*, 2025.

Paul A Yushkevich, John Pluta, Hongzhi Wang, Laura EM Wisse, Sandhitsu Das, and David Wolk. Ic-p-174: fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla t2-weighted mri. *Alzheimer's & Dementia*, 12:P126–P127, 2016. GitHub repository: https://github.com/pyushkevich/greedy.

Lyubomir Zagorchev and Ardeshir Goshtasby. A comparative study of transformation functions for nonrigid image registration. *IEEE transactions on image processing*, 15(3):529–538, 2006.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 29, 2016.

Liutong Zhang, Lei Zhou, Ruiyang Li, Xianyu Wang, Boxuan Han, and Hongen Liao. Cascaded feature warping network for unsupervised medical image registration. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 913–916. IEEE, 2021b.

Shengyu Zhao, Yue Dong, Eric I-Chao Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019a.

Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24 (5):1394–1404, 2019b.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Ting Zheng, Zhongqing Yang, Anan Li, Xiaohua Lv, Zhenqiao Zhou, Xiaojun Wang, Xiaoli Qi, Shiwei Li, Qingming Luo, Hui Gong, et al. Visualization of brain circuits using two-photon fluorescence micro-optical sectioning tomography. *Optics express*, 21(8):9839–9850, 2013.

Tao Zhong, Xueyang Wu, Shujun Liang, Zhenyuan Ning, Li Wang, Yuyu Niu, Shihua Yang, Zhuang Kang, Qianjin Feng, Gang Li, et al. nbest: Deep-learning-based non-human primates brain extraction and segmentation toolbox across ages, sites and species. *NeuroImage*, 295:120652, 2024.

Weifang Zhu, Jungong Xue, and Weiguo Gao. The sensitivity of the exponential of an essentially nonnegative matrix. *Journal of Computational Mathematics*, pages 250–258, 2008.

Darko Zikic, Ben Glocker, Oliver Kutter, Martin Groher, Nikos Komodakis, Ali Kamen, Nikos Paragios, and Nassir Navab. Linear intensity-based image registration by markov random fields and discrete optimization. *Medical image analysis*, 14(4):550–562, 2010.

APPENDIX A

Supplementary details for An Empirical Study and Evaluation of Image Registration paradigms

Fig. A.1 shows the boxplots of the performance of classical and DLIR methods trained on the OASIS dataset, on four T1-brain datasets.

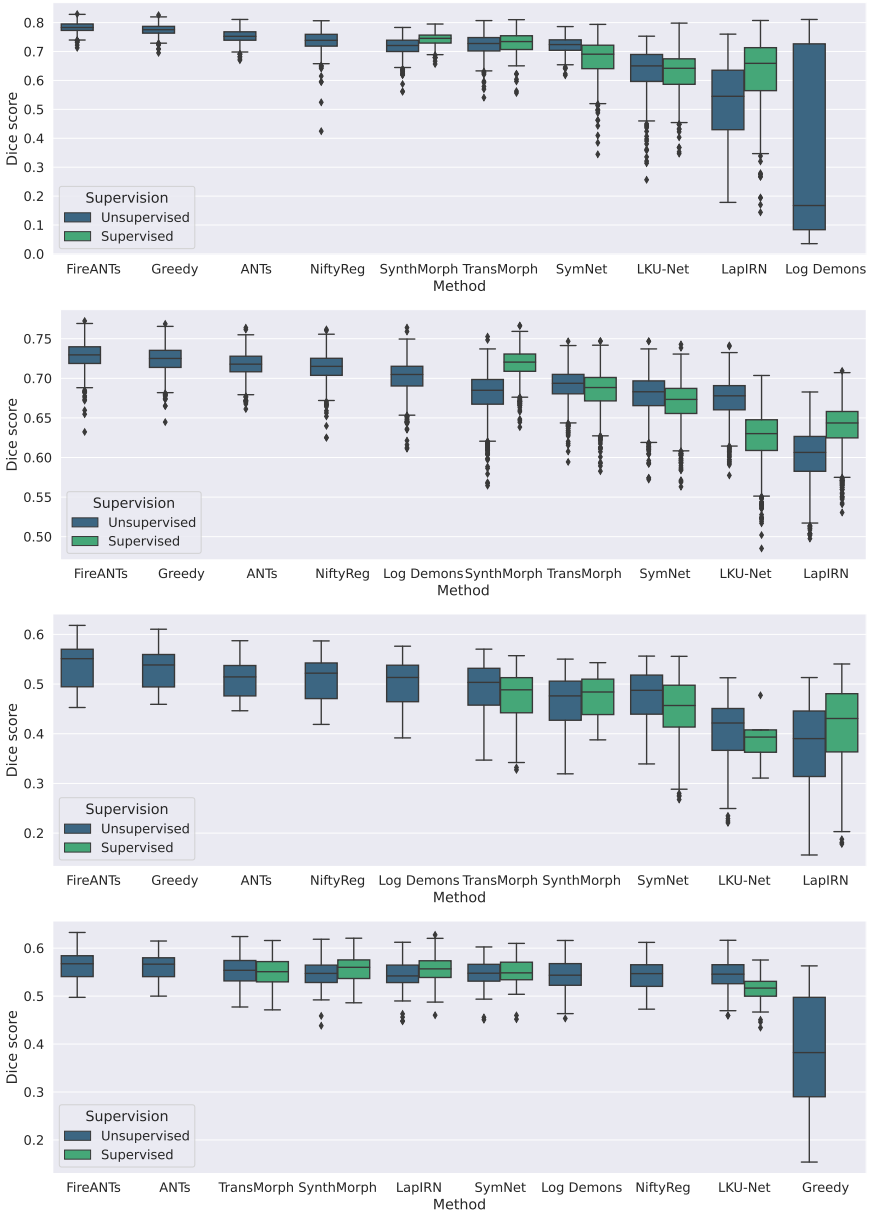


Figure A.1: Classical methods retain robustness across different datasets. Boxplots show the performance of classical and DLIR methods trained on the OASIS dataset, on four T1-brain datasets. For DLIR methods, we plot the performance of the supervised and unsupervised models. Across all datasets, FireANTs and ANTs consistently outperform DLIR methods, showing robustness to domain shift. Among DLIR methods, SynthMorph and TransMorph show robust performance, and training with label matching objective does not lead to significant improvement.

APPENDIX B

Supplementary details for FireANTs: Adaptive Riemannian Optimization for Multi-Scale Diffeomorphic Matching

B.1. A toy scenario to visualize the effect of κ on optimizers

To provide more intuition on the effect of κ on convergence of the SGD algorithm, we consider a toy example of a 2D optimization problem. Specfically, we consider a loss function $f_\kappa(x, y) = x^2 + \kappa y^2$ where $\kappa > 1$ becomes the condition number of the problem. Qualitatively, the effect of the first term diminishes exponentially fast with κ (Fig. B.2a). Quantitatively, we run both SGD and Adam optimization for a 1000 iterations starting from the point $(x, y) = (5, 5)$. Fig. B.2c shows that SGD works extremely well for $\kappa = 1$ which is the best-conditioned loss function, but quickly gets stuck for $\kappa \geq 100$. On the contrary, Adam is invariant to the condition number and converges to the minima for all values of κ. This is because for a diagonal Hessian (as in this case), the second-order adaptive terms are proportional to the diagonal elements of the Hessian. These condition numbers are vanishingly small compared to those in typical image registration tasks, which can exceed 10^5, making them extremely ill-conditioned.

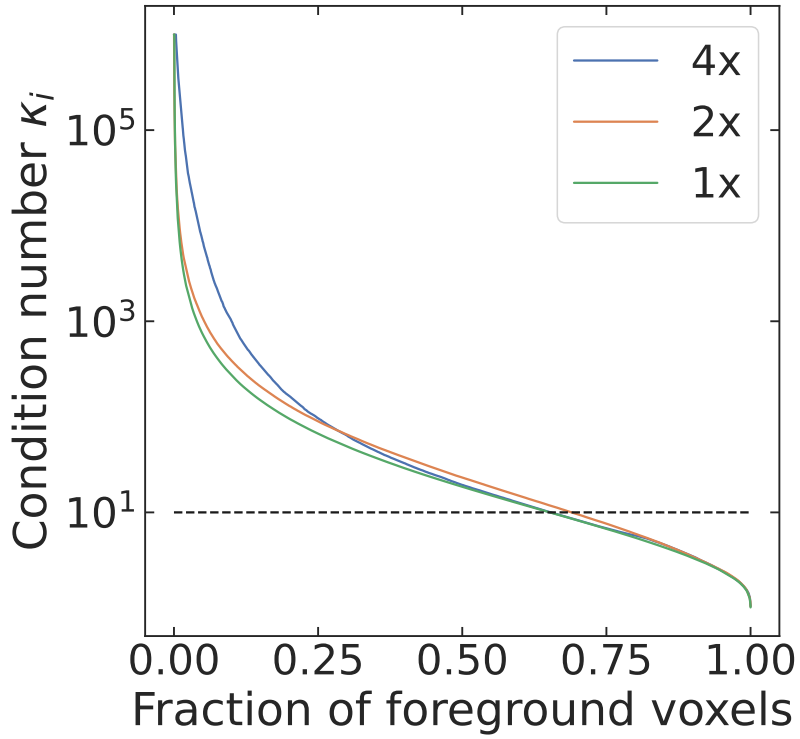
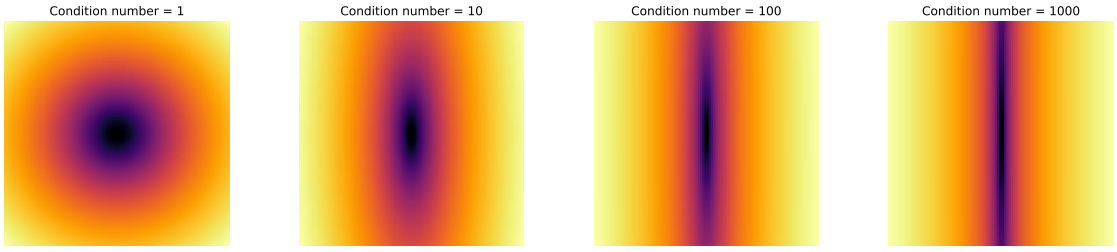


Figure B.1: Deformable image registration is ill-conditioned. To quantitatively examine ill-conditioning in registration, we compute the distribution of per-pixel condition number for a MRI registration task, at different image downsampling factors (denoted as 1x, 2x, and 4x). A high condition number signifies exacerbated ill conditioning and requires higher-order optimization. A horizontal dashed line denoting $\kappa = 10$ is drawn as a reference for substantial ill conditioning. Across all scales, a substantial fraction of foreground voxels are ill-conditioned ($\kappa > 10$), necessitating adaptive first-order optimization for faster convergence and accurate registration.

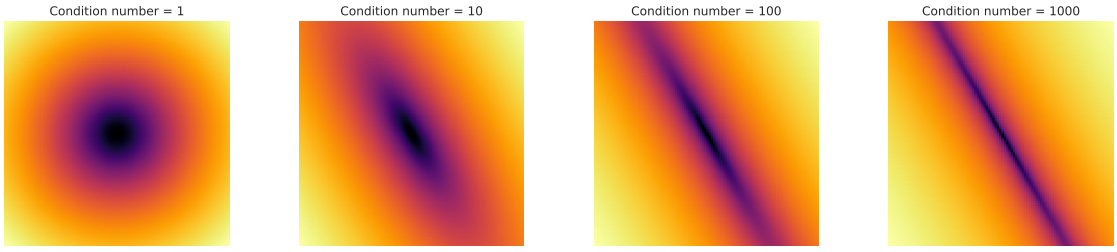
We also consider a more realistic, but tractable scenario of the convex loss function $f_{\kappa,\theta}(x, y) = x_\theta^2 + \kappa y_\theta^2$, where

$$\begin{bmatrix} x_\theta \\ y_\theta \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

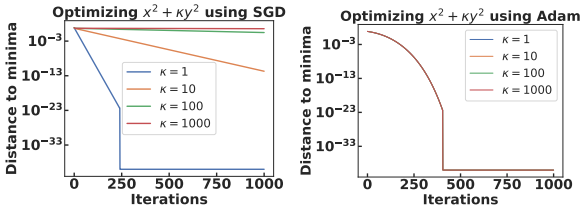
We choose $\theta = \pi/3$ for this experiment. This is simply a rotated version of the previous family of loss functions, as shown in Fig. B.2b. The trajectories obtained from optimization using SGD (Fig. B.2d) are virtually identical to that in Fig. B.2c since the new gradients are simply rotated versions of the previous gradients, and the distance from the minima is invariant to the rotation. However, the trajectories from Adam optimization are qualitatively very different, owing to the increasing difference between the true Hessian and its diagonal approximation. Even so, the final point is at a distance of less than 10^{-3} units to the minima for $\kappa = 1000$, showing the effectiveness of adaptive optimization even for ill-conditioned, non-diagonal Hessians. This is a strong motivation to extend adaptive optimization for non-Euclidean diffeomorphic registration, which is very high-dimensional and ill-conditioned.



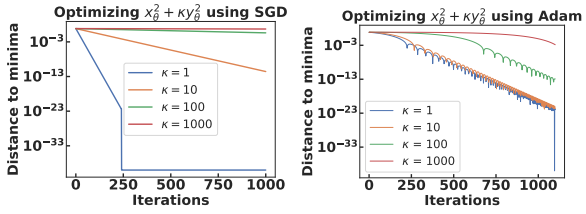
(a) Log-Loss landscape of the toy problem $f_\kappa(x, y) = x^2 + \kappa y^2$ for $\kappa = 1, 10, 100, 1000$. The log-loss becomes increasingly sharp along the y-direction as κ increases.



(b) Log-Loss landscape of the toy problem $f_\kappa(x, y) = x_\theta^2 + \kappa y_\theta^2$ for $\kappa = 1, 10, 100, 1000$, where (x_θ, y_θ) is the coordinate (x, y) rotated by an angle θ about the origin.



(c) Optimization of Fig. B.2a using SGD and Adam shows that SGD fails to recover the minima for $\kappa \geq 100$ while Adam is *invariant* to the condition number for diagonal Hessian matrices. This is a strong motivation to use first order adaptive optimization for registration where the condition number can exceed 10^5.



(d) Optimization of Fig. B.2b using SGD shows identical optimization trajectories as Fig. B.2c. Adam, however, is not invariant to the condition number because the difference between the true Hessian and its diagonal approximation increases with κ. Even so, the final point is at a distance of less than 10^{-3} units to the minima, showing the mitigating effect of adaptive optimization even for non-diagonal Hessians.

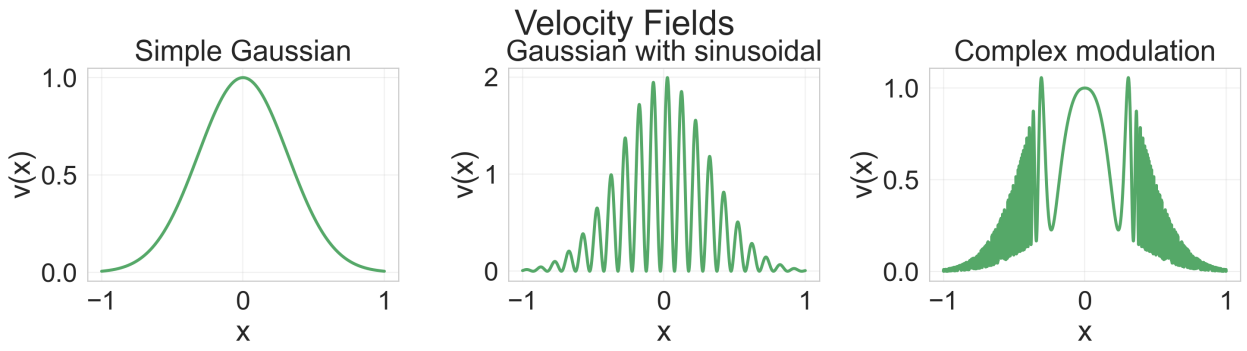


Figure B.3: Three 1D velocity fields with increasing Lipschitz constants to illustrate the dependence of the number of integration steps M on ensuring numerically accurate diffeomorphisms.

B.2. Empirical Verification of the dependence of the number of integration steps M on ensuring numerically accurate diffeomorphisms

To demonstrate this empirically, we choose three 1D velocity fields over the interval $\Omega = [-1, 1]$ with increasing Lipschitz constants (illustrated in Fig. B.3), and plot the amount of non-diffeomorphic voxels as a function of M. We choose 1D velocity fields for simplicity and easy visualization but all results generalize to higher dimensions. The three velocity fields are named and defined as follows:

- Simple gaussian: $v(x) = \exp(-5x^2)$
- Gaussian with sinusoidal: $v(x) = \exp(-5x^2)(1 + \sin(20\pi x))$
- Complex modulation: $v(x) = \exp(-5x^2)(1 + 0.7\sin(\pi \exp(|\pi x|^3)))$

The fraction of non-diffeomorphic voxels as a function of M is shown in Fig. B.4, along with the Lipschitz constant of the velocity fields. This plot shows that the Simple Gaussian velocity field required only 1 integration step to ensure a diffeomorphism over Ω but the sinusoidal velocity field required 6 steps and the complex modulation velocity field required 10 steps. Note that all three velocity fields are bounded by 1, but their Lipschitz constants are different by orders of magnitude. The result of the exponential map computed with increasing number of integration steps M for each velocity field is also visualized in Fig. B.5. Velocity fields with larger Lipschitz constants require a larger number of integration steps regardless of the actual magnitude of the velocity field. Since the Lipschitz constant of the velocity field is not known a priori, this creates a tradeoff between numerical accuracy and computational cost.

In contrast, when optimizing diffeomorphisms directly using the update rule $\varphi_{t+1} = \varphi_t \circ (id + \epsilon_t v_t)$, the scaling factor ϵ_t is chosen to be $\eta / LP(v_t)$ where η is the learning rate. $LP(v_t)$ can be bounded by the half of the norm of the velocity field divided by the resolution of the image, since we have

$$LP(v_t) = \sup_{x,y \in \Omega} \frac{\|v_t(x) - v_t(y)\|}{\|x - y\|} \leq \sup_{x' \in \Omega} 2 \frac{\|v_t(x')\|}{\|\delta x'\|} \tag{B.1}$$

avoiding the need to compute the Lipschitz constant directly.

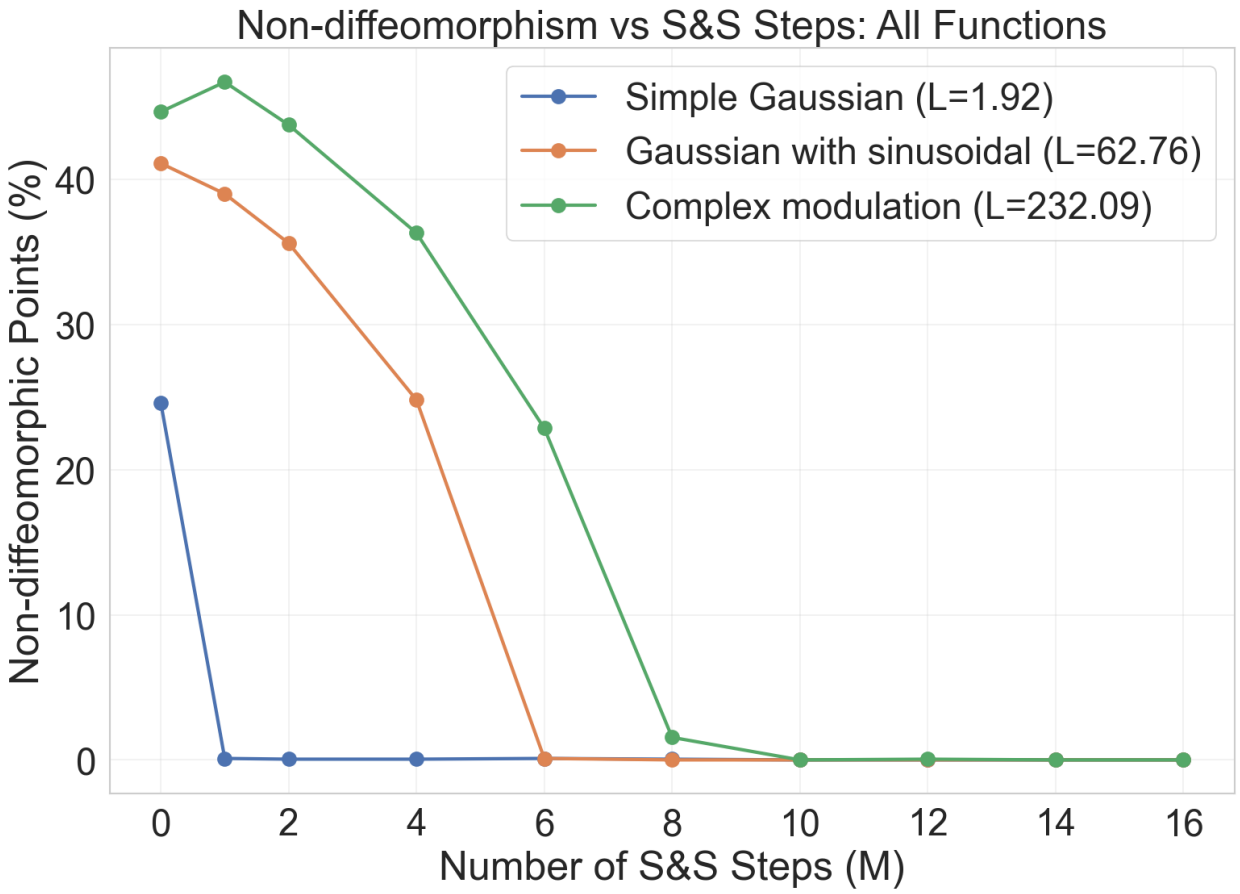


Figure B.4: Fraction of non-diffeomorphic voxels as a function of M for the three velocity fields.

B.3. Datasets and Evaluation Metrics

We provide details about the datasets and evaluation metrics used in the work.

B.3.1. In-vivo brain MRI mapping challenges (Klein *et al.*and OASIS)

Klein *et al.*neuromapping challenge Brain image data and their corresponding labels for 80 normal subjects were acquired from four different datasets. The *LPBA40* dataset contains 40 brain images and their labels to construct the LONI Probabilistic Brain Atlas (LPBA40). All volumes were skull-stripped, and aligned to the MNI305 atlas (Evans et al., 1993) using rigid-body transformation to correct for head tilt. For all these subjects, 56 structures were manually labelled and bias-corrected using the BrainSuite software. The *IBSR18* dataset contains brain images acquired at different laboraties through the Internet Brain Segmentation Repository. The T1-weighted images were rotated to be in Talairach alignment and bias-corrected. Manual labelling is performed resulting in 84 labeled regions. For the *CUMC12* dataset, 12 subjects were scanned at Columbia University Medical Center on a 1.5T GE scanner. Images were resliced, rotated, segmented and manually labeled, leading to 128 labeled regions. Finally, the *MGH10* dataset contains 10 subjects who were scanned at the MGH/MIT/HMS Athinoula A. Martinos Center using a 3T Siemens scanner. The data is bias-corrected, affine-registered to the MNI152 template, and segmented. Finally the images were manually

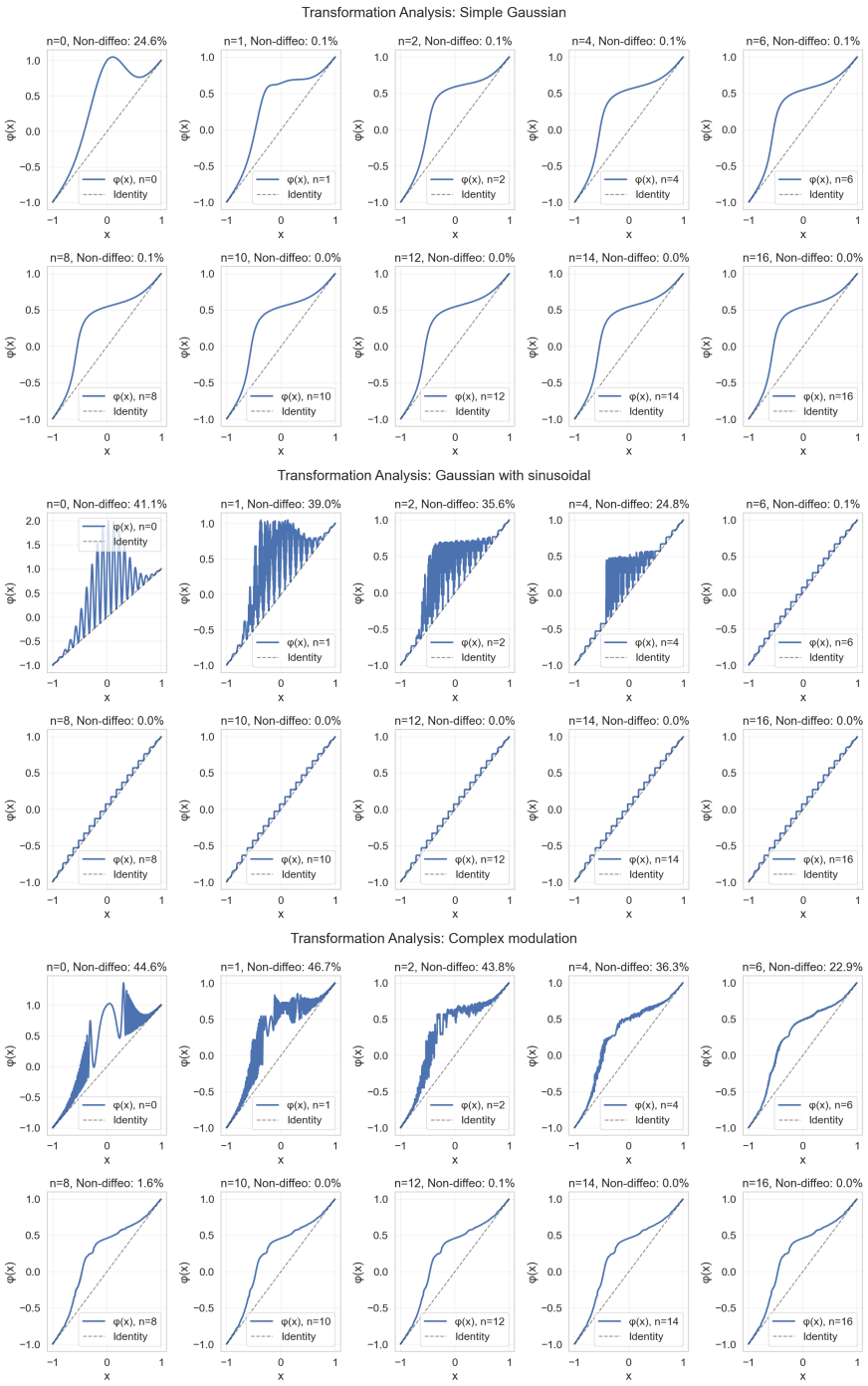


Figure B.5: Illustrative example of the effect of the number of integration steps M for scaling-and-squaring and the final deformation obtained for the three velocity fields. Larger Lipschitz constants require a larger number of integration steps to ensure numerical diffeomorphisms with the scaling-and-squaring approach.

labeled, leading to 74 labeled regions. All datasets have a volume of $256 \times 256 \times \{128, 124\}$ voxels with varying amounts of anisotropic voxel spacing, ranging from $0.84 \times 0.84 \times 1.5$mm to $1 \times 1 \times 1.33$mm. ANTs was one of the top performing methods for this challenge, performing well robustly across all four datasets.

Algorithm 5 FireANTs

Input: Fixed image I_f, Moving image I_m
Scales $[s_1, s_2, \ldots, s_n]$, Iterations $[T_1, T_2, \ldots T_n]$, n scales
`optstate` optimizer state (for Adam, RMSProp, etc.)
`use_jac` boolean specifying whether to use Jacobian in descent direction

Initialize $\varphi \leftarrow id_{s_1}$. ▷ Initialize warp to identity at first scale
Initialize $l \leftarrow 1$. ▷ Initialize current scale
while $l \leq n$ **do**
    Initialize $i \leftarrow 0$
    Initialize $I_f^l, I_m^l \leftarrow \text{downsample}(I_f, s_l), \text{downsample}(I_m, s_l)$
    while $i < T_l$ **do**
        $U_i \leftarrow C(I_f^l, I_m^l \circ \varphi^i) + R(\varphi)$
        Compute $v_d'(x) \leftarrow \frac{\partial U_i}{\partial \varphi}(\varphi^{(i)})(x)$ ▷ Jacobian-free Eulerian descent direction
        if `use_jac` **then**
            Compute $v_d'(x) \leftarrow J^\top(\varphi^{(i)}(x))v_d'(x)$ ▷ Eulerian descent direction
        end if
        Update $(v_d'(x), \texttt{optstate}) \leftarrow \texttt{optstate}(v_d'(x))$ ▷ Apply and update optimizer state
        Update $\varphi^{(i+1)} \leftarrow \varphi^{(i)} \circ \exp_{id}(\epsilon_i v_d') \approx \varphi^{(i)} \circ (id + \epsilon_i v_d')$
        $i \leftarrow i + 1$
    end while
    if $l < n$ **then**
        $\varphi \leftarrow \text{Upsample}(\varphi, s_{(l+1)})$ ▷ Upsample warp to next scale using bilinear/trilinear interpolation
    end if
    $l \leftarrow l + 1$
end while

Figure B.6: Algorithm for FireANTs Algorithm 5 outlines the key steps in FireANTs - computing the Jacobian-free Eulerian descent direction which is simply the Gateaux derivative. If the boolean `use_jac` is specified, then use the steepest Eulerian descent direction instead. This descent direction is then modified using any adaptive optimization algorithm denoted as `optstate`. The warp field is then updated using the exponential map or retraction map for small ϵ_i. After optimization at a given scale, the warp field is upsampled using bilinear or trilinear interpolation to the next scale until optimization is complete for all steps.

A natural way to evaluate whether two images are in a common coordinate frame is to evaluate the accuracy of overlap of gross morphological anatomical structures. The method considers measures of volume and surface overlap, volume similarity, and distance measures to evaluate the alignment of anatomical regions. Given a source label map S_r and target label map T_r and a cardinality operator $|.|$, we consider the following overlap measures. We consider 'target overlap' and 'mean overlap' (also known as Dice score) as the primary measures of agreement between the source and target label maps.

$$TO_r = \frac{|S_r \cap T_r|}{|T_r|}, MO_r = 2\frac{|S_r \cap T_r|}{|S_r| + |T_r|} \tag{B.2}$$

The aggregates over all regions are given by:

$$TO = \frac{1}{N_r}\sum_r TO_r \quad , \quad MO = \frac{1}{N_r}\sum_r MO_r \tag{B.3}$$

Klein *et al.*(Klein et al., 2009) also propose a 'Union Overlap' metric which is a monotonic function of the Dice score. Therefore, we do not use this in our evaluation. To complement the above agreement measures, we also compute false negatives (FN), false positives (FP), and volume similarity (VS) coefficient for anatomical region r:

$$FN_r = \frac{|T_r \backslash S_r|}{|T_r|}, \quad FP_r = \frac{|S_r \backslash T_r|}{|S_r|}, \quad VS_r = 2\frac{|S_r| - |T_r|}{|S_r| + |T_r|} \tag{B.4}$$

Comparison on other metrics proposed in (Klein et al., 2009) and regionwise analysis are shown in Figs. B.7 and B.8. Similar to the overlap metrics, we compute the aggregates as in the original evaluation denoted by $FN_{Klein}, FP_{Klein}, VS_{Klein}$ and average over regions denoted simply by FN, FP, VS.

OASIS dataset On the OASIS dataset, we use same the evaluation criteria as in the Learn2Reg challenge (Hering et al., 2022), i.e. Dice score overlap and 95th percentile of the Haussdorf distance computed for 35 subcortical structures. This leads to a total of 12 evaluation metrics that we use to compare our method with 4 baselines - ANTs, Demons (Vercauteren et al., 2007b), VoxelMorph (Balakrishnan et al., 2019) and SynthMorph (Hoffmann et al., 2021), representing established classical and deep learning registration algorithms.

In total, we compare with state-of-the-art baselines on over *2000 brain volume pairs*, with varying number of labeled anatomical regions and resolutions.

B.3.2. Lung CT mapping challenges (EMPIRE10 and NLST)

Registration of temporally spaced breathhold scans can help in tracking disease progression, or registration between inspirion and expiration scans can enable improved monitoring of airflow and pulmonary function. The EMPIRE10 challenge (Murphy et al., 2011) aims to provide a platform for in-depth evaluation and fair comparison of available registration algorithms for this application. The dataset consists of 30 pairs of chest CT scans, with intra-subject registration across a variety of healthy or diseased subjects. The scan pairs consist of inspiration-expiration scans, breathhold scans over time, scans from 4D data, ovine data, contrast-noncontrast, and artificially warped scan pairs. The ovine data was acquired where breathing was controlled, and metallic markers were surgically implanted to provide landmark annotations, followed by a hole-filling algorithm to disguise the markers so that registration algorithms cannot use this artificial information. Artificially warped scan pairs also provide ground truth correspondences for landmarks and lung boundaries. The challenge provides a broad range of data complexity, voxel sizes and image acquisition differences. ANTs is, again one of the top performing methods in this challenge. Unlike the brain datasets, ground truth labels for fissure and landmarks are not provided for validation. Therefore, we rely on the evaluation metrics computed privately by the challenge organizers in the evaluation server. We compare our method with two powerful baselines (i) ANTs, which optimizes the diffeomorphism directly, and (ii) the *DARTEL* (Ashburner, 2007) formulation optimizing a stationary velocity field (SVF), where the diffeomorphism is obtained using an exponential map of the SVF. We first affinely align the binary lung masks of the moving and fixed images using Dice loss (Dice, 1945). This is followed by a diffeomorphic registration using the intensity images.

We use the Adam optimizer with learning rate of 3e-3, and a multi-scale optimization with downsampling rates of 6,4,2,1 for 200, 100, 50, 20 iterations. This is followed by a diffeomorphic registration step with the same multi-scale resolutions and 200,150,75,25 iterations and a learning rate of 0.25. We use a Gaussian kernel for gradient smoothing with $\sigma_{\text{grad}} = 6.0$ and warp smoothing with $\sigma_{\text{warp}} = 0.4$. The optimal values for $\sigma_{\text{grad}}, \sigma_{\text{warp}}$ are found by a hyperparameter grid search, and are strikingly close to the parameters used in the

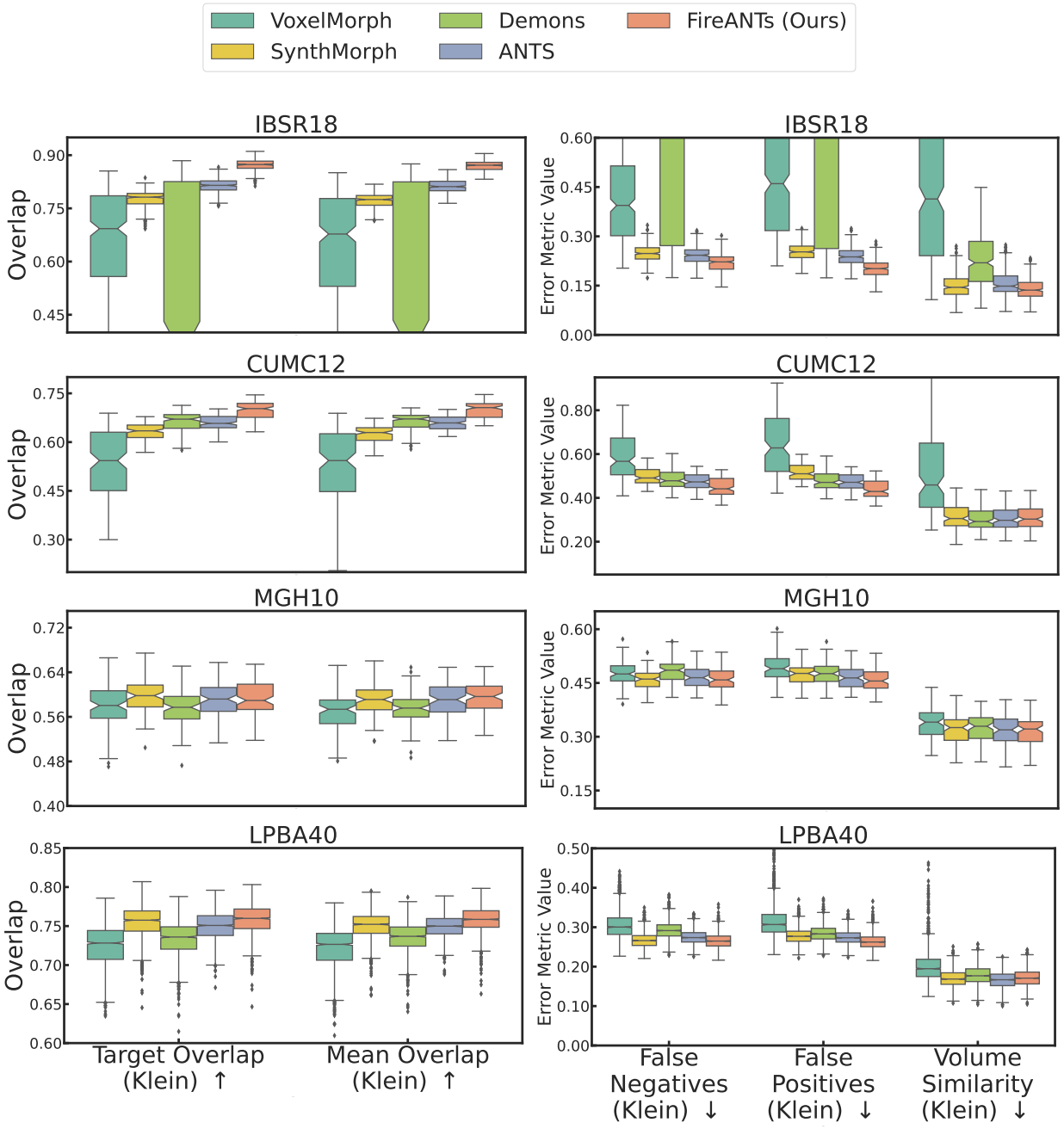


Figure B.7: Comparison of our method with ANTs on 4 MRI brain datasets: Registration quality is validated by measuring volume overlap of label maps between the fixed and warped label maps. **(a)**: For anatomical region r, warped (binary) label map S_r and fixed label map T_r, target and mean overlap are defined as $|S_r \cap T_r|/|T_r|$ and $2|S_r \cap T_r|/(|S_r| + |T_r|)$. We define the aggregate target overlap over all anatomical regions as $\sum_r(|S_r \cap T_r|/|T_r|)$ and Klein *et al.* (Klein et al., 2009) define it as $(\sum_r |S_r \cap T_r|)/(\sum_r |T_r|)$, likewise for other metrics. The latter aggregation is denoted with the suffix (Klein) in the figure. In all four datasets, the boxplots show a narrower interquartile range and substantially higher median than ANTs (higher is better), underscoring the stability and accuracy of our algorithm. **(b)**: Other measures of anatomical label overlap used in (Klein et al., 2009) are false positives ($|T_r \backslash S_r|/|T_r|$), false negatives ($|S_r \backslash T_r|/|S_r|$), and volume similarity ($2(|S_r| - |T_r|)/(|S_r| + |T_r|)$) (lower is better). We observe similar trends as in *(a)*, with a narrower interquartile range and substantially lower median values. Results of per region overlap metrics are in the Fig. B.8.

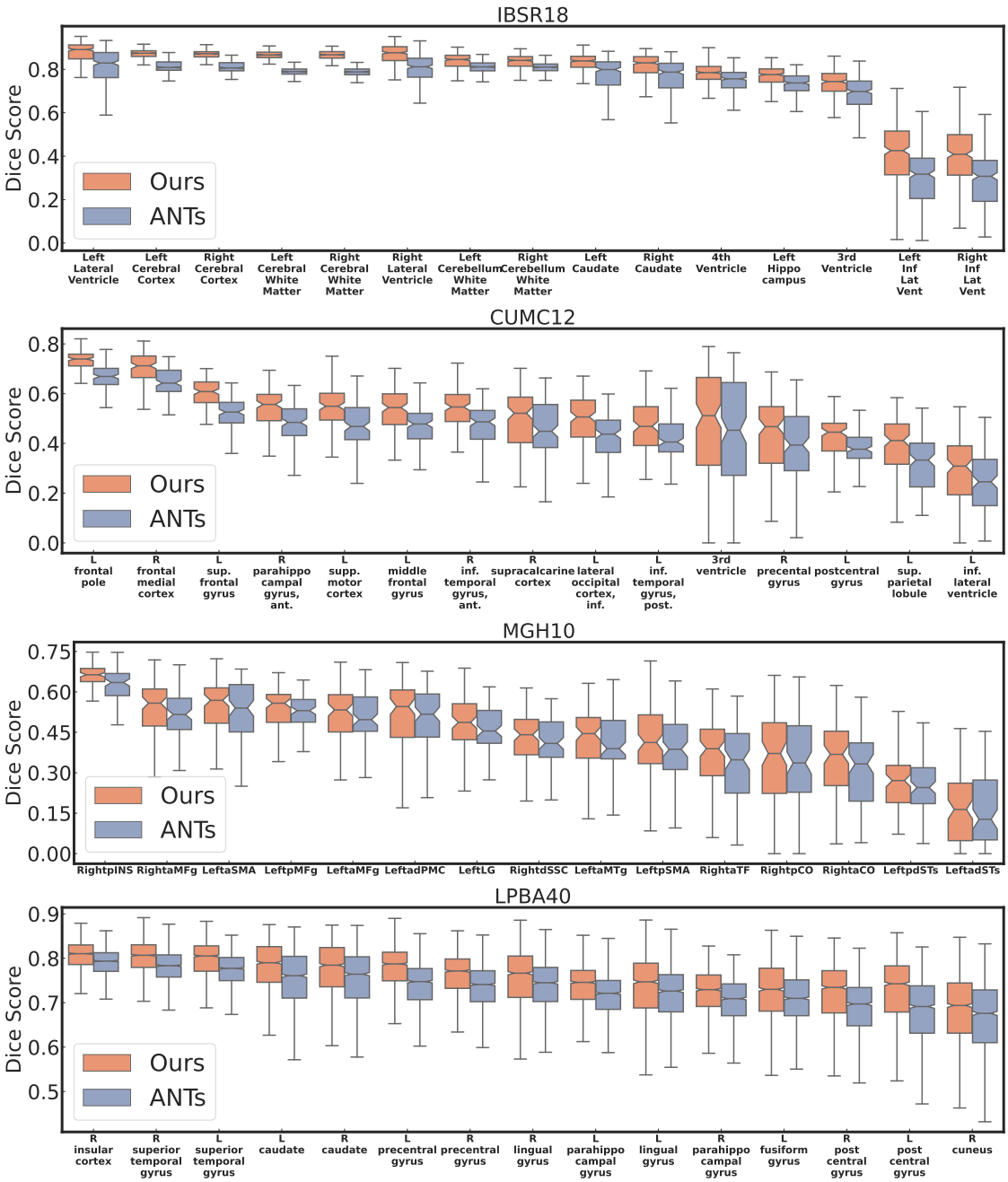


Figure B.8: Regionwise target overlap on the brain MRI datasets: We further evaluate regionwise overlap scores by sampling 15 regions from each dataset, and comparing their distribution using our method and ANTs. Our method has a much higher median score, and better interquartile ranges across regions, demonstrating both accuracy and robustness.

ANTs submission.

We evaluate three criteria: (1) fissure alignment errors (%)—the fraction of misaligned fissure voxels (Figs. 3.4b and 3.4e), (2) landmark distance in mm (Fig. 3.4d), and (3) singularity errors—the fraction of non-diffeomorphic voxels (Fig. 3.4c). Fig. 3.4 highlights the impact of representation choice in modeling diffeomorphisms. DARTEL, using an exponential map, performs significantly worse than ANTs across all metrics by three orders of magnitude. In contrast, our method reduces fissure alignment error by 5× compared to ANTs and outperforms it in 5 out of 6 landmark subregions. While all methods theoretically ensure diffeomorphism, SVF-based approaches introduce singularity errors due to non-adaptive scaling-and-squaring. We discuss the limitations of SVF-based approaches in Section 3.3.2. ANTs also introduces some singularities, whereas our method computes numerically perfect diffeomorphic transforms. Finally, Fig. 3.4e compares fissure alignment errors among EMPIRE10 submissions, showing FireANTs achieves the lowest landmark errors and the fastest runtime among the top 10 methods, setting new benchmarks in computational efficiency and accuracy. Our method, on the other hand computes numerically perfect diffeomorphic transforms. Finally, we compare the fissure alignment error of all submissions in the EMPIRE10 challenge, and show the top 10 algorithms in Fig. 3.4e. Results demonstrate that FireANTs attains the lowest landmark alignment errors compared to an array of contemporary state-of-the-art algorithms.

NLST dataset For the NLST dataset(Team, 2011), we compare with representative state-of-the-art optimization and deep-learning baselines. We use the evaluation criteria provided by the challenge, and measure results on the Robust Target Registration Error (TRE30) in millimeters between the registered keypoints. Results in Fig. 3.4f show that FireANTs outperforms all baselines on the NLST dataset, with improvements of upto 51.6% in robust target registration error (TRE30) of provided keypoints compared to state-of-the-art deep learning benchmarks including Im2Grid, Vector-Field Attention, RWC-Net, and a 50.8% improvement in TRE30 over foundation models like unigradICON. This demonstrates the broad applicability of FireANTs beyond neuroimaging applications.

B.3.3. Other Datasets and Metrics

PRIMatE Data Exchange (PRIME-DE) A growing body of research has documented the utility of MRI data to study neuroanatomical organization and function of non-human primates. The PRIMatE Data Exchange (PRIME-DE) resource (Milham et al., 2018b) provides a platform for the neuroimaging community to facilitate the mapping of the non-human primate connectome. We use a subset of this dataset collected from five different sources: Aix-Marseille Université, Mount Sinai School of Medicine, McGill University, Stem Cell and Brain Research Institute, and the University of California, Davis, resulting in 116 subjects, and subsequently 13340 subject pairs for registration. We use the nBEST deep learning framework to perform cerebrum extraction, followed by tissue segmentation. Since the images are markedly different than human brains, we affinely register all of the extracted cerebrum volumes to the first subject sorted by name to bring them to a common coordinate space and metadata. This is followed by a diffeomorphic registration using the intensity images. We use the Dice score of the registered tissue segmentations to evaluate the quality of registration.

Ultracortex The Ultracortex dataset (Mahler et al., 2024) hosts a unique collection of ultra-high field (9.4 Tesla) MRI data of the human brain. This dataset includes detailed structural images and high-quality manual segmentations, making it an invaluable resource for researchers in neuroimaging and computational neuroscience. Out of the 86 T1-weighted images with resolutions spanning from 0.6 to 0.8mm, precise manual segmentation of the gray and white matter for each hemisphere is provided for 12 subjects. We use the dataset

and manually provided segmentations to evaluate the quality of registration of cortical surface mapping using Dice score. Note that all deep learning methods run out of memory at 0.6 to 0.8mm resolutions, therefore we resample the images to 1.0mm isotropic resolution for evaluation of deep learning methods.

Rodent Datasets We use four rodent datasets in this study: Waxholm Rat Brain, Allen CCFv3 mouse brain, RnR-ExM mouse isocortex, and BICCN mouse dataset. The datasets feature high-resolution atlases of the rat and mouse brain with four different modalities (T2*w MRI, STPT, ExM, fMOST) respectively. The motivation for using the datasets is to provide a benchmark for *cross-species, multimodal* registration (Waxholm $\rightarrow$ Allen CCFv3), perform well on a high-resolution registration task and leaderboard (RnR-ExM), and to faithfully reproduce high-resolution rodent atlases (BICCN). Similar to the Ultracortex dataset, most deep learning methods run out of memory at 25μm resolution for these images, therefore we resample the images to 50μm resolution for evaluation of deep learning methods. To handle the multimodal nature of the cross-species registration task, we use Anatomix (Dey et al., 2025) as a modality-agnostic feature extractor as feature images to perform registration. Since the cross-species templates have different labelmaps, comparing Dice scores directly is not possible. The Waxholm template comprises 95 labeled regions, while the Allen CCFv3 template includes over 300 regions defined in the complete ARA ontology. To enable a comparable level of anatomical granularity, we coarsened the Allen CCFv3 parcellation to 34 regions by collapsing all subregions beyond depth 3 in the ontology hierarchy into their corresponding parent nodes. We compute the Mutual Information (MI) between the registered label map corresponding to the Waxholm template image to that of the Allen CCFv3 template image to evaluate the quality of registration. For the fMOST and RnR-ExM datasets, we use the images to perform pairwise registration and atlas generation respectively. The performance on the RnR-ExM dataset is evaluated using the Dice score of the registered label map corresponding to the ExM image pairs on a private evalaution server. For the BICCN dataset, we only provide qualitative results due to the lack of an evaluation criteria.

Zebrafish Datasets Analysis of the zebrafish is a growing field of research due to its unique advantages as a vertebrate model organism. The zebrafish brain is small yet structurally complex, offering a tractable system for studying whole-brain organization, development, and function at cellular resolution. High-quality atlases such as AZBA and Z-Brains provide detailed anatomical and gene expression reference templates, enabling cross-modality and cross-sample comparison. These datasets present a valuable testbed for registration algorithms, as they involve significant structural variability, diverse imaging modalities (including confocal, light-sheet, and two-photon microscopy), and finely detailed neuroanatomical annotations. Accurate registration in this setting is critical for integrating large-scale imaging data and mapping functional or genetic information onto common anatomical frameworks. The adult and larval zebrafish brains have very different structural organization, and show very different characteristics than human brains. This is a challenging registration task to truly access the out-of-distribution generalization capabilities of registration algorithms. Due to the lack of a consensus on appropriate evaluation criteria beyond qualitative comparison for these datasets, we use the mutual information between the registered image and template image to evaluate the quality of registration.

Learn2Reg Abdomen MRCT registration This dataset is used as a testbed to ablate the effect of Jacobian-free optimization on abdominal MRCT registration. Abdomen CT-MR registration is a conceptually different registration task copmared to neuroanatomical or pulmonary registration with completely unrelated anatomical structures, organization, and biomechanical dynamics. We use the validation split provided by the challenge to evaluate the performance of FireANTs with and without Jacobian-free optimization.

B.4. Modular software implementation to enable effective experimentation

Registration is a key part of many data processing pipelines in the clinical literature. Our software implementation is designed to be extremely flexible, e.g., it implements a number of existing registration methods using our techniques, modular, e.g., the user can choose different group representations (rigid or affine transforms, diffeomorphisms), objective functions, optimization algorithms, loss functions, and regularizers. Users can also stack the same class of transformations, but with different cost functions. For example, they can fit an affine transform using label maps and Dice loss, and use the resultant affine matrix as initialization to fit another affine transform using the cross-correlation registration objective. This enables seamless tinkering and real-time investigation of the data. Deformations can also be composed in increasing order of complexity (rigid $\rightarrow$ affine $\rightarrow$ diffeomorphisms), thereby avoiding multiple resampling and subsequent resampling artifacts. We have developed a simple interface to implement custom cost functions, which may be required for different problem domains, with ease; these custom cost functions can be used for any of the registration algorithms out-of-the-box. Our implementation can handle images of different sizes, anisotropic spacing, without the need for resampling into a consistent physical spacing or voxel sizes. All algorithms also support multi-scale optimization (even with fractional scales) and convergence monitors for early-stopping.

Our software is implemented completely using default primitives in PyTorch. All code and example scripts is available at https://github.com/rohitrango/fireants.

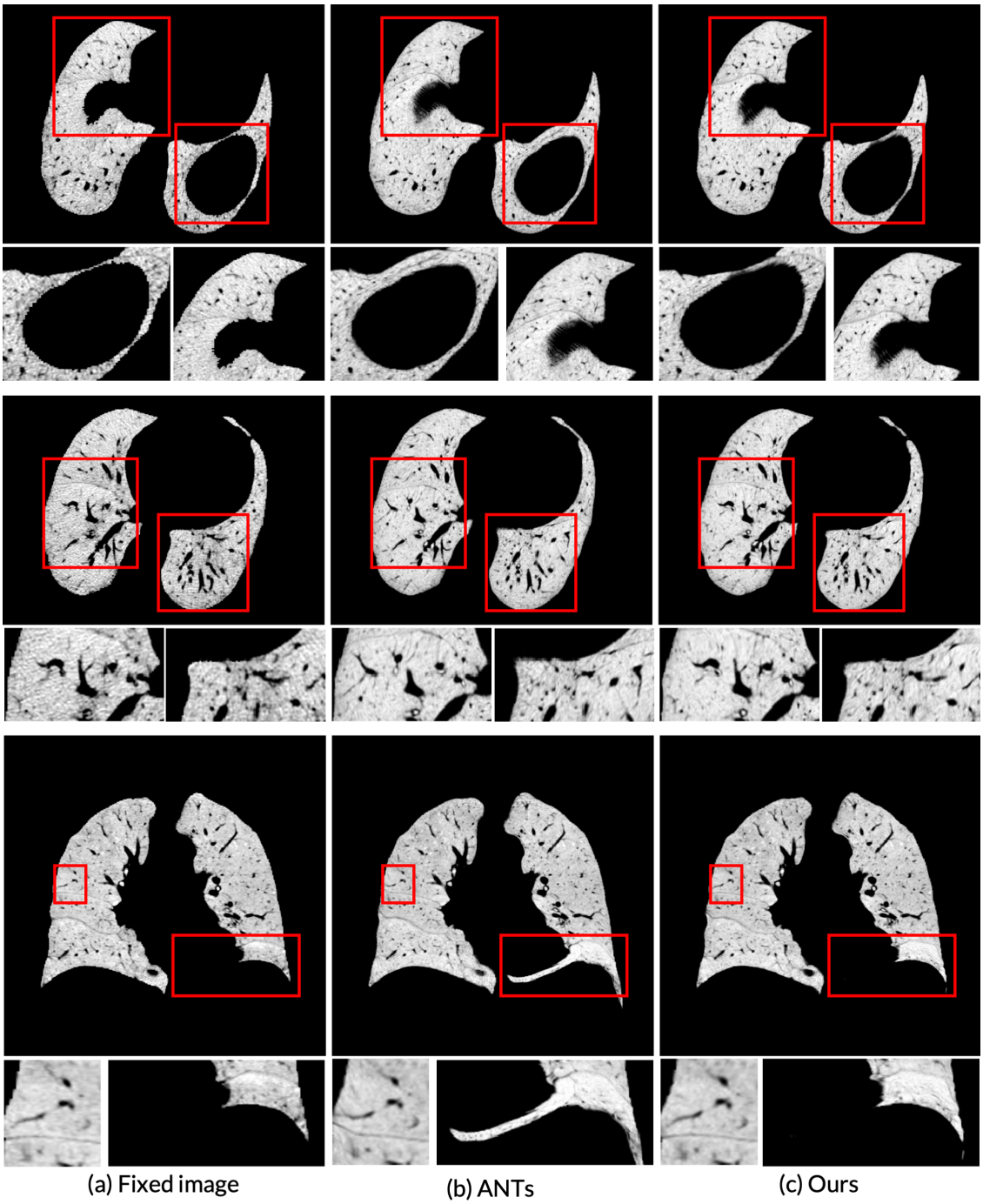


Figure B.9: Qualitative results on EMPIRE10 challenge: (a) shows the fixed image, (b) shows the registration performed by ANTs, and (c) our method, all with zoomed in regions. ANTs performs a coarse registration with ease, but still leaves out critical alignment of lung boundary and airways by not utilizing adaptive optimization. Our method performs *perfectly* diffeomorphic registration by construction, and does not lead to any registration errors, both in the lung boundaries or internal features.

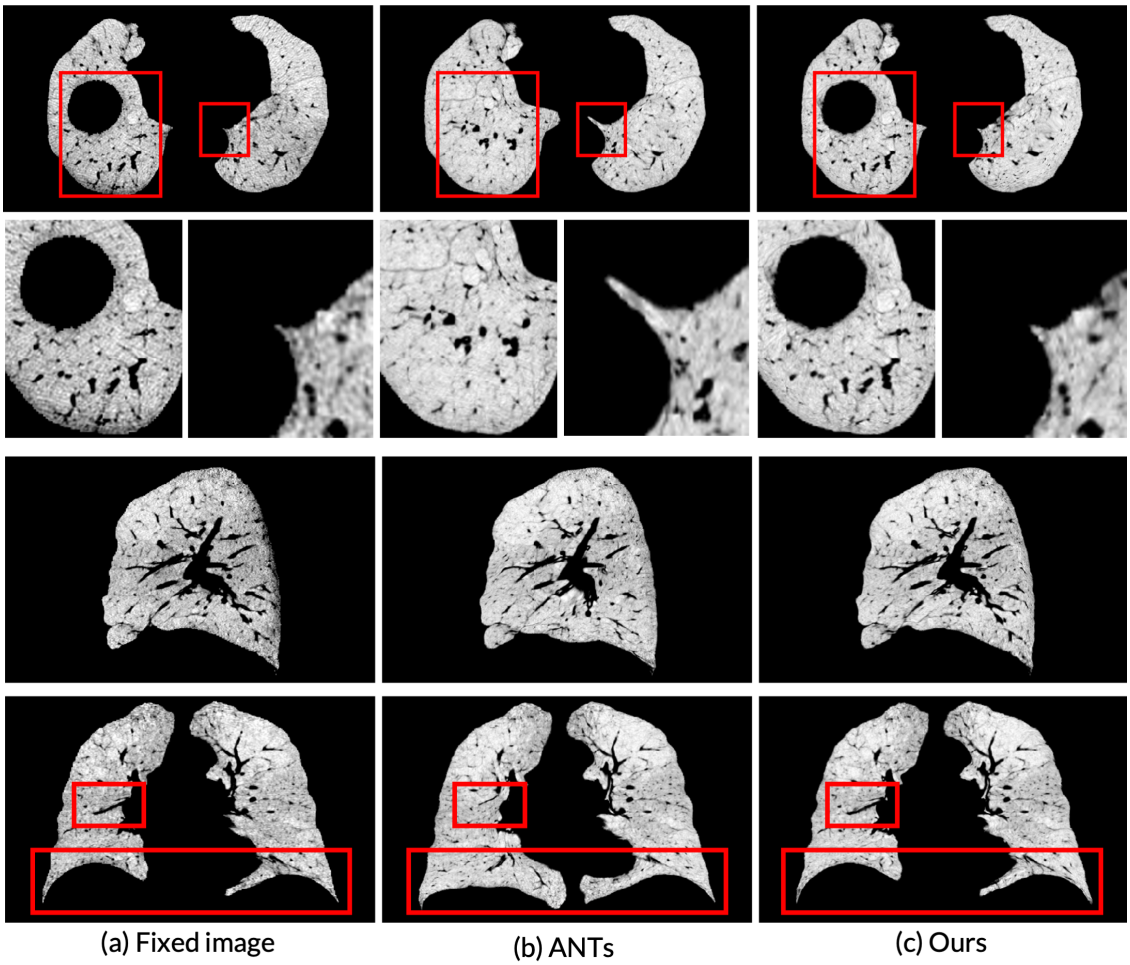


Figure B.10: More Qualitative results on EMPIRE10 challenge: **(a)** shows the fixed image, **(b)** shows the registration performed by ANTs, and **(c)** our method, all with zoomed in regions. ANTs performs a coarse registration with ease, but still leaves out critical alignment of lung boundary and airways by not utilizing adaptive optimization. Our method performs *perfectly* diffeomorphic registration by construction, and does not lead to any registration errors, both in the lung boundaries or internal features.

APPENDIX C

Supplementary details for Deep Implicit Optimization enables Robust Learnable Features for Deformable Image Registration

C.1. Inference time

DLIR methods have been very popular due to their fast inference time by performing amortized optimization (Balakrishnan et al., 2019). Classical methods generally focus on robustness and reproducubility, and do have GPU implementations for fast inference. However, modern optimization toolkits (Mang et al., 2019b; Jena et al., 2026) utilize massively parallel GPU computing to register images in seconds, and scale very well to ultrahigh resolution imaging. A concern with optimization-in-the-loop methods is the inference time. Fig. C.1 shows the inference time for our method for all four architectures. These inference times are fast for a lot of applications, and the plug-and-play nature of our framework makes DIO amenable to rapid experimentation and hyperparameter tuning.

Architecture	**Neural net** (sec)	**Optimization** (sec)
UNet	0.444	1.693
UNet-E	0.433	1.555
LKU	0.795	1.463
LKU-E	2.281	1.457

Method	**Iterations**	**Time (sec)**	**Avg. throughput (it/s)**
SUITS	15	255.211	0.058
GradIRN	9	0.351	25.641
Ours	350	**1.463**	**239.23**

Figure C.1: Inference time for various architectures. A multi-scale optimization takes only ~ 1.5 seconds to run all iterations (no early stopping) making it suitable for most applications. This is compared to the time for neural network's feature extraction which is architecture dependent.

C.2. Implementation Details

Formulating arbitrary iterative solvers using implicit differentiation allows full expressivity of powerful solvers for learning-based image registration. We elaborate on the implementation details that make this framework practical and scalable.

C.2.1. Jacobian-Free Backprop

In practice, the ill-conditioned nature of the inverse Hessian leads to poor training performance. To avoid the ill-conditioning, we follow (Fung et al., 2021) and substitute the Jacobian to identity, to compute $\frac{\partial \hat{T}}{\partial F_f} \approx -\frac{\partial T}{\partial \varphi}\frac{\partial \varrho}{\partial F_f}$. This leads to lesser memory and compute requirements during the backward pass, and stable training dynamics compared to other estimates of Jacobian like phantom gradients, damped unrolling, or Neumann series (Geng et al., 2021; Geng and Kolter, 2023). We perform an ablation on using full blockwise Hessian and unrolling-based phantom gradient (Geng et al., 2021) in Section 4.5.7.

C.2.2. Double Backward through `grid_sample`

Note that in Eq. (4.5), ϱ contains a $\nabla F_m \circ \varphi$ term, and the quantity $\frac{\partial \varrho}{\partial F_m}$ will require the double-backward pass of the `grid_sample` operator in PyTorch. Since this operation is not implemented in the PyTorch C backend, a backward pass for the gradient operation does not exist in PyTorch. We use the `gridsample_grad2` library (Siarohin, 2023) to compute the double-backward pass of the `grid_sample` operator in Eq. (4.5).

C.2.3. Other details

For all experiments, we use multi-scale features with $4\times, 2\times, 1\times$ downsampling for multi-scale optimization, unless otherwise mentioned. We run the solver for a maximum of $200, 100, 50$ iterations for each scale respectively, with an early stopping criteria if the relative loss does not change by more than 10^{-4} for 5 iterations. We choose the MSE loss for the feature matching objective within the solver. For the non-diffeomorphic iterative optimizer, we use a simple nonparameteric displacement field representation and an SGD-based solver with a learning rate of 0.003. For the diffeomorphic optimizer, we use the FireANTs library with Adam optimizer and a learning rate of 0.5. For learning the parameters of the feature network, we use the AdamW optimizer with a learning rate of 0.0003. All methods are implemented in PyTorch, and all experiments are performed on a single NVIDIA A6000 GPU.

C.3. Implicit bias of optimization for registration

Model based systems, such as deep networks are not immune to inductive biases due to architecture, loss functions, and optimization algorithms used to train them. Functional forms of the deep network induce constraints on the solution space, but optimization algorithms are not excluded from such biases either. The implicit bias for Gradient Descent is a well-studied phenomena for overparameterized linear and shallow networks. Gradient Descent for linear systems leads to an optimum that is in the span of the input data starting from the initialization (Zhang et al., 2021a; Soudry et al., 2018; Ji and Telgarsky, 2018; Pesme et al., 2021; Wu et al., 2020). This bias is also dependent on the chosen representation, since that defines the functional relationship of the gradients with the parameters and inputs. This limits the reachable set of solutions by the optimization algorithm when multiple local minima exist.

In the case of image registration, the optimization limits the space of solutions (warps) that can be obtained by the SGD algorithm. To show this, we consider the transformation φ as a set of particles in a Langrangian frame that are displaced by the optimization algorithm to align the moving image to the fixed image. Consider a regular grid of particles, whose locations specify the warp field. Let the location of i-th particle at iteration t be $\varphi^{(t)}(x_i)$. For a fixed feature image F_f, moving image F_m and current iterate $\varphi^{(t)}$, the gradient of the registration loss with respect to particle i at iteration t is given by

$$\frac{\partial C(F_f, F_m \circ \varphi^{(t)})}{\partial \varphi^{(t)}(x_i)} = C_i'(F_f, F_m \circ \varphi^{(t)}) \nabla F_m(\varphi^{(t)}(x_i)) \tag{C.1}$$

where

$$C_i'(F_f, F_m \circ \varphi^{(t)}) = \frac{\partial C(F_f, F_m \circ \varphi^{(t)})}{\partial M(\varphi^{(t)}(x_i))}$$

is the (scalar) derivative of scalar loss C with respect to the intensity of i-th particle computed at the current

iterate, and $\nabla F_m(\varphi^{(t)}(x_i))$ is the spatial gradient of the moving image at the location of the particle. Note that the **direction** of the gradient of particle i is *independent* of the fixed image, loss function, and location of other particles – it only depends on the spatial gradient of the moving image at the location of the particle. This restricts the movement of a particle located at any given location along a 1D line whose direction is the spatial gradient of the moving image at that location. Since F_f and F_m are computed independently of each other (and therefore no information of F_f and F_m is contained in each other), the space of solutions of φ is restricted by this implicit bias. This is restrictive because the similarity function and fixed image do not influence the direction of the gradient, and the optimization algorithm is biased towards solutions that are in the direction of the gradient of the moving image.

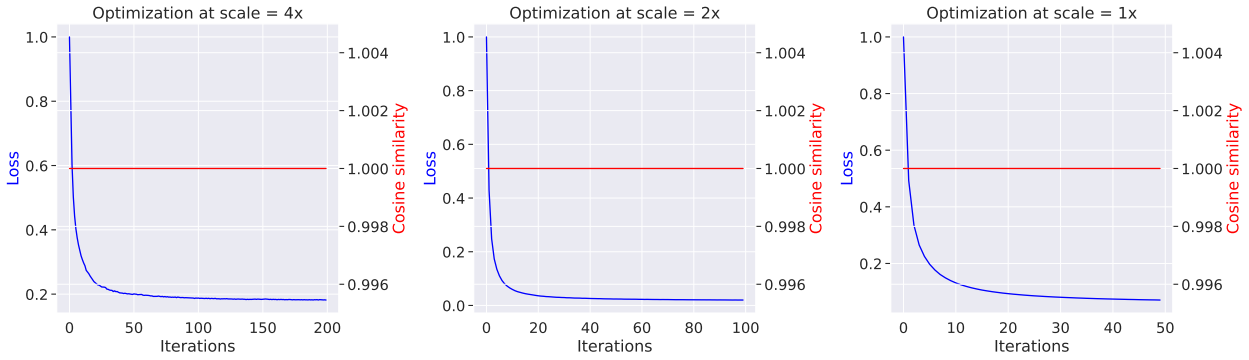


Figure C.2: Implicit bias in SGD for image registration. The plot shows the loss curves for a multi-scale optimization of two feature images. Each plot also shows the absolute cosine similarity of per-pixel gradients obtained by C and $C_{\text{surrogate}}$ at each iteration. Note that over the course of optimization, the cosine similarity is always 1 – demonstrating the implicit bias of the optimization for registration.

We show this bias empirically – we perform multi-scale optimization algorithm using feature maps obtained from the network. We keep track of two gradients, one obtained by the loss function, and another obtained by the gradient of a surrogate loss $C_{\text{surrogate}}(F_m, \varphi^{(t)}) = \sum_i F_m(\varphi^{(t)}(x_i))$. Note that $C_{\text{surrogate}}$ does not depend on the fixed image or the loss function. The gradient of $C_{\text{surrogate}}$ with respect to the i-th particle is given by $\nabla F_m(\varphi^{(t)}(x_i))$. At each iteration, we compute the magnitude of cosine similarly between the gradients of C and $C_{\text{surrogate}}$. Fig. C.2 shows that the loss converges, and the per-pixel gradients can be predicted by $C_{\text{surrogate}}$ alone, as depicted by the magnitude and standard deviation of cosine similarity between C and $C_{\text{surrogate}}$. This limits the movement of each particle along a 1D line in an N-D space, and limits the degrees of freedom of the optimization by N-fold for N-D images. Future work will aim at alleviating this implicit bias to allow for more flexible solutions.

C.4. Algorithm details

DIO is a learnable framework that leverages *implicit differentiation* of an arbitrary black-box optimization solver to learn features such that registration in this feature space corresponds to good registration of the images and additional label maps. This additional indirection leads to learnable features that are registration-aware, interpretable, and the framework inherits the optimization solver's versatility to variability in the data like difference in contrast, anisotropy, and difference in sizes of the fixed and moving images. We contrast our approach with a typical classical optimization-based registration algorithm in Fig. C.3. A classical multi-scale optimization routine *indiscriminately* downsamples the intensity images, and does not retain discriminative information that is useful for registration. Since our method is trained to maximize label alignment from all

scales, multi-scale features obtained from our method are more discriminative and registration-aware. We also compare DIO with a typical DLIR method in Fig. C.5. Note that the fixed end-to-end architecture and functional form of a deep network subsumes the representation choice into the architecture as well, limiting its ability to switch to arbitrary transformation representations at inference time without additional retraining. Our framework therefore combines the benefits of both classical (robustness to out-of-distribution datasets, and zero-shot transfer to other optimization routines) and learning-based methods (high-fidelity, label-aware, and registration-aware).

C.5. Toy example

Fig. C.6 shows the loss curves for the toy dataset described in Section 4.5.1. An image-based optimization algorithm would correspond to the green curve being a flat line at 1 due to the flat landscape of the intensity-based loss function.

C.6. Quantitative Results

Table C.1 shows the quantitative results of our method for out-of-distribution performance on the IBSR18, CUMC12, and LPBA40 datasets. In 9 out of 10 cases, DIO demonstrates the best accuracy with fairly lower standard deviations, highlighting the robustness of the model. DIO therefore serves as a strong candidate for out-of-distribution performance, and can be used in a variety of settings where the training and test distributions differ.

C.7. Datasets

We consider four brain MRI datasets in this work: OASIS dataset for in-distribution performance, and LPBA40, IBSR18, and CUMC12 datasets for out-of-distribution performance (Shattuck et al., 2008; ibs; Klein et al., 2009; Marcus et al., 2007b). More details about the datasets are provided below.

- **OASIS**. The Open Access Series of Imaging Studies (OASIS) dataset contains 414 T1-weighted brain images in Young, Middle Aged, Nondemented, and Demented Older adults. The images are skull-stripped and bias-corrected, followed by a resampling and afine alignment to the FreeSurfer's Talairach atlas. Label segmentations of 35 subcortical structures were obtained using automatic segmentation using Freesurfer software.

- **LPBA40**. 40 brain images and their labels are used to construct the LONI Probabilistic Brain Atlas (LPBA40) dataset at the Laboratory of Neuroimaging (LONI) at UCLA (Shattuck et al., 2008). All volumes are preprocessed according to LONI protocols to produce skull-stripped volumes. These volumes are aligned to the MNI305 atlas – this is relevant since existing DLIR methods may be biased towards images that are aligned to the Talairach and Tournoux (1988) atlas which is used to align the images in the OASIS dataset. This is followed by a custom manual labelling protocol of 56 structures from each of the volumes. Bias correction is perfrmed using the BrainSuite's Bias Field Corrector.

- **IBSR18**. the Internet Brain Segmentation Repository contains 18 different brain images acquired at different laboratories as IBSRv2.0. The dataset consists of T1-weighted brains aligned to the Talairach and Tournoux (1988) atlas, and manually segmented into 84 labelled regions. Bias correction of the images are performed using the 'autoseg' bias field correction algorithm.

Method	Dice supervision	Isotropic		Anisotropic	
		Crop	No Crop	Crop	No Crop
Conditional LapIRN	✗	0.7367 ± 0.0237	✗	0.7269 ± 0.0328	0.7317 ± 0.0303
LapIRN	✗	0.5257 ± 0.1316	✗	0.5435 ± 0.1266	0.5001 ± 0.1271
LapIRN	✓	0.6259 ± 0.1238	✗	0.6209 ± 0.1163	0.5759 ± 0.1207
LKU-Net	✗	0.6309 ± 0.0839	✗	0.6276 ± 0.0838	0.6072 ± 0.0787
LKU-Net	✓	0.6267 ± 0.0776	✗	0.6231 ± 0.0730	0.5992 ± 0.0757
SymNet	✗	0.7213 ± 0.0273	✗	0.7116 ± 0.0398	0.7117 ± 0.0398
SymNet	✓	0.6731 ± 0.0688	✗	0.6672 ± 0.0731	0.6674 ± 0.0728
TransMorph Large	✓	0.7383 ± 0.0353	✗	0.7312 ± 0.0405	✗
TransMorph Regular	✗	0.7221 ± 0.0400	✗	0.7289 ± 0.0417	✗
TransMorph Regular	✓	0.7293 ± 0.0370	✗	0.7113 ± 0.0520	✗
VoxelMorph	✗	0.5118 ± 0.1774	✗	0.5233 ± 0.1693	✗
SynthMorph	✓	0.7423 ± 0.0225	✗	0.7476 ± 0.0238	✗
Ours (LKU)	✓	0.7698 ± 0.0193	0.7587 ± 0.0208	0.7728 ± 0.0219	0.7572 ± 0.0369
Conditional LapIRN	✗	0.4793 ± 0.0373	0.4804 ± 0.0368	0.4880 ± 0.0416	0.4827 ± 0.0408
LapIRN	✗	0.3719 ± 0.0897	0.3491 ± 0.0895	0.3524 ± 0.1001	0.3556 ± 0.0989
LapIRN	✓	0.4121 ± 0.0907	0.3838 ± 0.0929	0.3911 ± 0.1060	0.3896 ± 0.1063
LKU-Net	✗	0.4054 ± 0.0641	0.3922 ± 0.0679	0.4086 ± 0.0732	0.3999 ± 0.0697
LKU-Net	✓	0.3904 ± 0.0547	0.3827 ± 0.0574	0.3967 ± 0.0745	0.3960 ± 0.0678
SymNet	✗	0.4761 ± 0.0524	0.4761 ± 0.0524	0.4822 ± 0.0565	0.4820 ± 0.0565
SymNet	✓	0.4457 ± 0.0675	0.4457 ± 0.0675	0.4518 ± 0.0787	0.4521 ± 0.0786
TransMorph Large	✓	0.4827 ± 0.0531	✗	0.4858 ± 0.0587	✗
TransMorph Regular	✗	0.4929 ± 0.0502	✗	0.4967 ± 0.0540	✗
TransMorph Regular	✓	0.4737 ± 0.0549	✗	0.4741 ± 0.0628	✗
VoxelMorph	✗	0.3519 ± 0.1271	✗	0.3469 ± 0.1308	✗
SynthMorph	✓	0.4761 ± 0.0397	✗	0.4797 ± 0.0426	✗
Ours (LKU)	✓	0.5137 ± 0.0410	0.5126 ± 0.0412	0.5237 ± 0.0433	0.5162 ± 0.0448
Conditional LapIRN	✗	0.7113 ± 0.0178	0.7109 ± 0.0178	-	-
LapIRN	✗	0.6026 ± 0.0317	0.5878 ± 0.0325	-	-
LapIRN	✓	0.6395 ± 0.0269	0.6211 ± 0.0294	-	-
LKU-Net	✗	0.6746 ± 0.0230	0.6708 ± 0.0249	-	-
LKU-Net	✓	0.6266 ± 0.0299	0.6220 ± 0.0296	-	-
SymNet	✗	0.6797 ± 0.0239	0.6797 ± 0.0238	-	-
SymNet	✓	0.6700 ± 0.0248	0.6698 ± 0.0248	-	-
TransMorph Large	✓	0.6918 ± 0.0219	✗	-	-
TransMorph Regular	✗	0.6919 ± 0.0191	✗	-	-
TransMorph Regular	✓	0.6855 ± 0.0225	✗	-	-
VoxelMorph	✗	0.6776 ± 0.0365	✗	-	-
SynthMorph	✓	0.7189 ± 0.0172	✗	-	-
Ours (LKU)	✓	0.7139 ± 0.0181	0.7131 ± 0.0181	-	-

Table C.1: Quantitative evaluation on out-of-distribution performance on IBSR18, CUMC12, and LPBA40 datasets. We compare DIO with other state-of-the-art DLIR methods. The '**Dice supervision**' column shows if the method is trained with label matching on the OASIS dataset. We evaluate the performance of the methods with and without isotropic and anisotropic data resampling. The results are reported as mean ± standard deviation. = First, = Second, = Third best result.

- **CUMC12**. The Columbia University Medical Center dataset contains 12 T1-weighted brain images with manual segmentation of 128 regions. The images were scanned on a 1.5T GE scanner, and the images were resliced coronally to a slice thickness of 3mm, rotated into cardinal orientation, and segmented by a technician trained according to the Cardviews labelling scheme.

C.8. Convergence of KeyMorph on OASIS

We run KeyMorph (Wang et al., 2023) on the OASIS dataset for 2000 epochs. We plot the Soft Dice ($= 1 -$ diceloss) and Mean Squared error between the fixed and moving images in Fig. C.11. Note that the soft Dice loss starts to plateau at ~ 0.70, and the hard dice loss on the validation set is even lower (~ 0.64). This represents a huge gap in performance compared to unsupervised baselines and our method. These numbers are also consistent with those reported in (Wang et al., 2023) for deformable registration. Note that although KeyMorph works in the contrived scenario of arbitrary rotations and translations (most MRI datasets are acquired in standard coordinate systems like RAS), it is not designed to handle the more complex deformations that are present in the brain MRI datasets.

Algorithm 6 Classical registration pipeline

Input: Fixed image I_f, Moving image I_m
Scales $[s_1, s_2, \ldots, s_n]$, Iterations $[T_1, T_2, \ldots T_n]$, n levels.
Initialize $\varphi = Id_{s_1}$. ▷ Initialize warp to identity at first scale
Initialize $l = 1$. ▷ Initialize current scale
while $l \leq n$ **do**
    Initialize $i = 0$
    Initialize $I_f^l, I_m^l = \text{downsample}(I_f, s_l), \text{downsample}(I_m, s_l)$
    while $i < T_l$ **do**
        $L_i = C(I_f^l, I_m^l \circ \varphi^i)$
        Compute $\nabla_\varphi L$
        Update $\varphi^{(i+1)} = \text{Optimize}(\varphi^i, \nabla_\varphi L_i)$ ▷ Optimization algorithm
        $i = i + 1$
    end while
    if $l < n$ **then**
        $\varphi = \text{Upsample}(\varphi, s_{(l+1)})$ ▷ Upsample warp to next level
    end if
    $l = l + 1$
end while

Algorithm 7 Differentiable Implicit Optimization for Registration (Our algorithm)

Input: Fixed features $\mathcal{F}_f = [F_f^1, F_f^2 \ldots F_f^n]$, Moving features $\mathcal{F}_f = [F_f^1, F_f^2 \ldots F_f^n]$
Scales $[s_1, s_2, \ldots, s_n]$, Iterations $[T_1, T_2, \ldots T_n]$, n levels.
Initialize $\varphi = Id_{s_1}$. ▷ Initialize warp to identity at first scale
Initialize $l = 1$. ▷ Initialize current scale
Outputs $= []$. ▷ Save intermediate outputs for backpropagation
while $l \leq n$ **do**
    Initialize $i = 0$
    Initialize $I_f^l, I_m^l = F_f^l, F_m^l$
    while $i < T_l$ **do**
        $L_i = C(I_f^l, I_m^l \circ \varphi^i)$
        Compute $\nabla_\varphi L$
        Update $\varphi^{(i+1)} = \text{Optimize}(\varphi^i, \nabla_\varphi L_i)$ ▷ Optimization algorithm
        $i = i + 1$
    end while
    Outputs.append $\left(\varphi^{(T_l)}\right)$ ▷ Save final warp at this level for backpropagation
    if $l < n$ **then**
        $\varphi = \text{Upsample}(\varphi, s_{(l+1)})$ ▷ Upsample warp for next level
    end if
    $l = l + 1$
end while

Figure C.3: Comparison of a typical classical registration algorithm and DIO: Algorithm 6 shows a typical classical registration algorithm that uses a multi-scale optimization routine to register the fixed and moving images. At each level l, the fixed and moving images are downsampled by a factor of s_l, therefore trading off between discriminative information and vulnerability to local minima. Algorithm 7 shows our algorithm (red text highlights differences compared to Algorithm 6) that uses a separate scale-space feature at each level. Unlike classical methods, the scale-space feature can capture different discriminative features at each level to maximize label alignment and the multi-scale nature helps avoid local minima.

Algorithm 8 Backward pass for DIO

Input: Fixed features $\mathcal{F}_f = [F_f^1, F_f^2 \dots F_f^n]$, Moving features $\mathcal{F}_f = [F_f^1, F_f^2 \dots F_f^n]$, Backend $backend$
Stored outputs $[\varphi^1, \varphi^2 \dots \varphi^n]$, gradients $\left[\frac{\partial T}{\partial \varphi^1}, \frac{\partial T}{\partial \varphi^2}, \dots \frac{\partial T}{\partial \varphi^n}\right]$
Backend $backend$, n levels.
Initialize $l = 1$. ▷ Initialize current scale
while $l \leq n$ **do**
    if $backend$ ==Hessian **then**
        Compute $H = \frac{\partial \varrho}{\partial \varphi^l}$
        Update $v = \texttt{linalg.lstsq}\left(H, \frac{\partial T}{\partial \varphi^l}\right)$ ▷ Full Hessian IFT
    else
        Update $v = \frac{\partial T}{\partial \varphi^l}$ ▷ Jacobian-Free Backprop
    end if
    Compute $h = v^T \cdot \varrho(\varphi^l, F_f^l, F_m^l)$
    Set $\texttt{grad}(F_f^l) = \texttt{autograd.grad}(h, F_f^l)$
    Set $\texttt{grad}(F_m^l) = \texttt{autograd.grad}(h, F_m^l)$
end while

Figure C.4: Pseudocode for backward pass with DIO: Given the stored features and outputs from the forward pass, and the gradients w.r.t. final warp from the backward pass, we compute the gradients of the loss function with respect to the fixed and moving features at each level. The gradients are analytically computed depending on the specified backend.

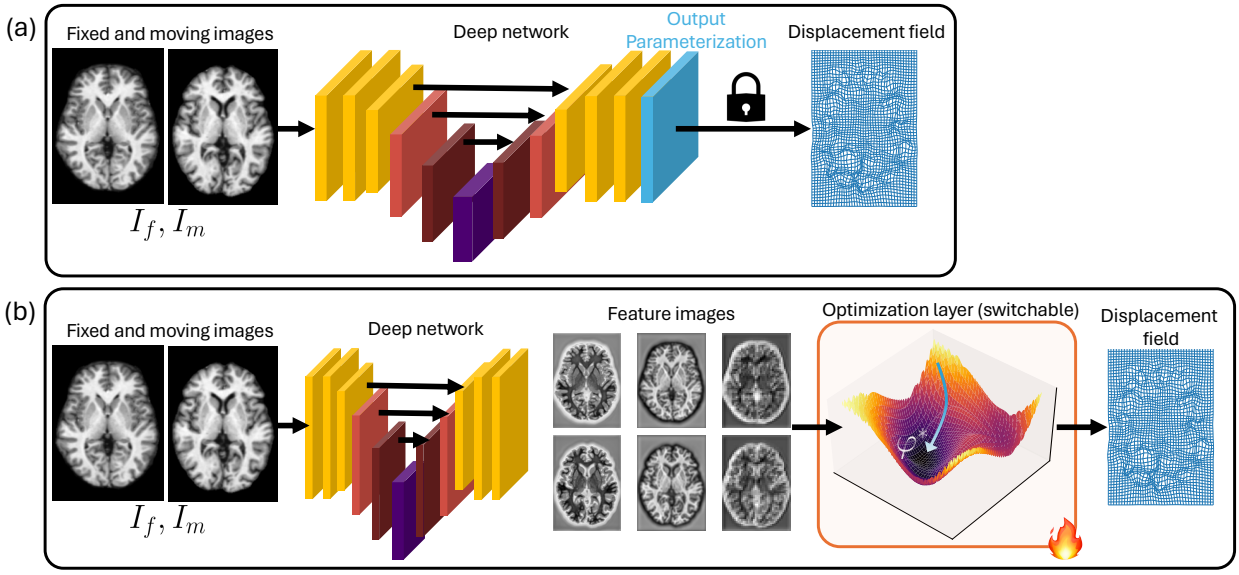


Figure C.5: Comparison of typical DLIR method and our method. (a) shows the pipeline of a typical deep network. The neural network architecture takes the channelwise concatenation of the fixed and moving images as input, and outputs a warp field, which has a *fixed* transformation representation (SVF, free-form, B-splines, affine, etc. denoted as the blue locked layer). This representation is fixed throughout training and cannot be switched at test-time, without additional finetuning of the network. **(b)** shows our framework wherein the fixed and moving images are input *separately* into a feature extraction network that outputs multi-scale features. These features are then passed onto an iterative black-box solver than can be *implicitly differentiated* to backpropagate the gradients from the optimized warp field back to the feature network. This allows for a more flexible transformation representation, and the optimization solver can be switched at test-time with zero finetuning.

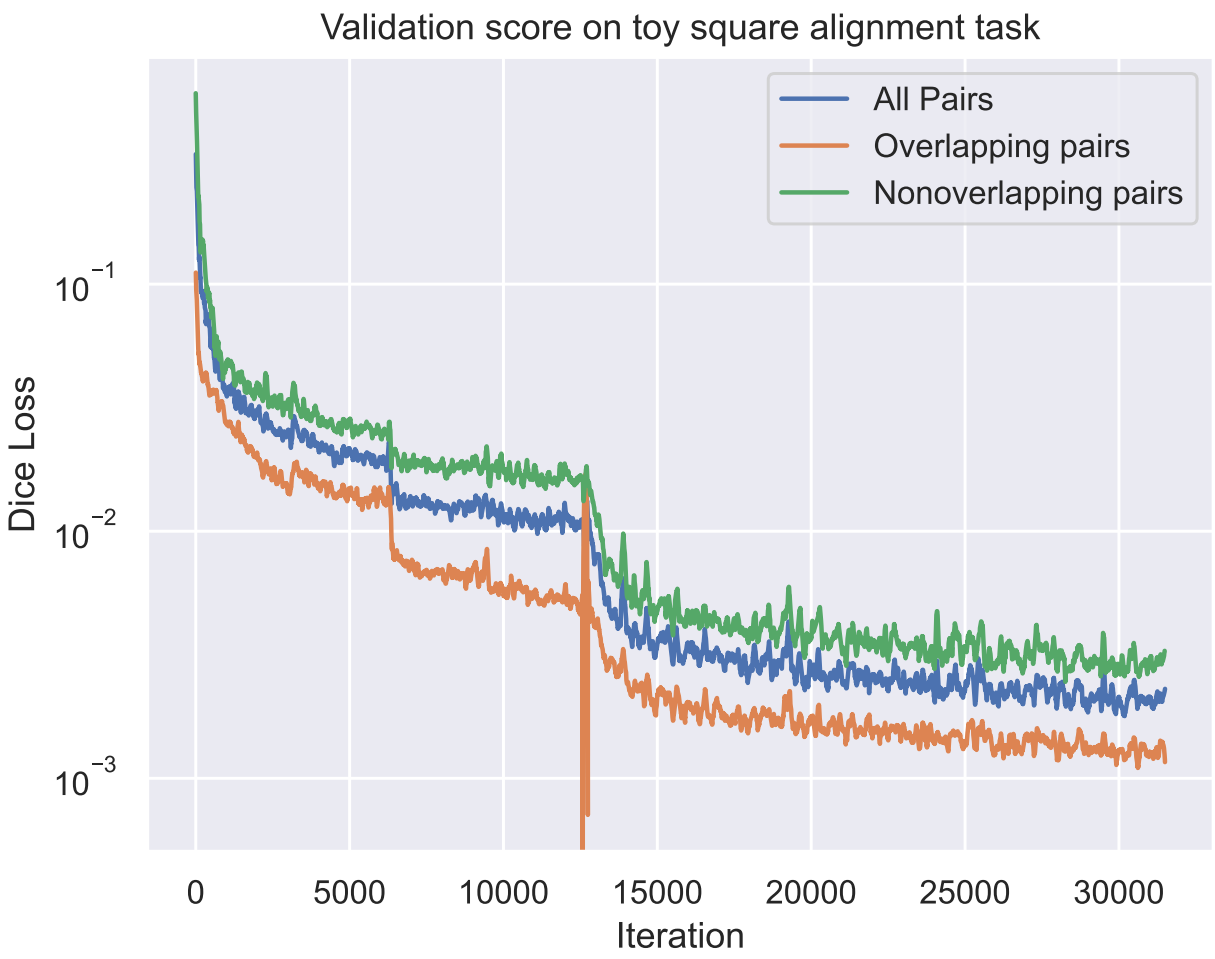


Figure C.6: Loss curves for toy dataset. Plot shows three curves - the Dice score for (a) all validation image pairs, (b) image pairs that have non-zero overlap in the image space (therefore a gradient-based affine solver will recover a transform from intensity images), and (c) image pairs that have zero overlap in the image space (therefore any gradient-based solver using intensity images will fail). Our feature network recovers dense multi-scale features (see Fig. 4.3) which allows all subsets to be registered with >0.99 Dice score.

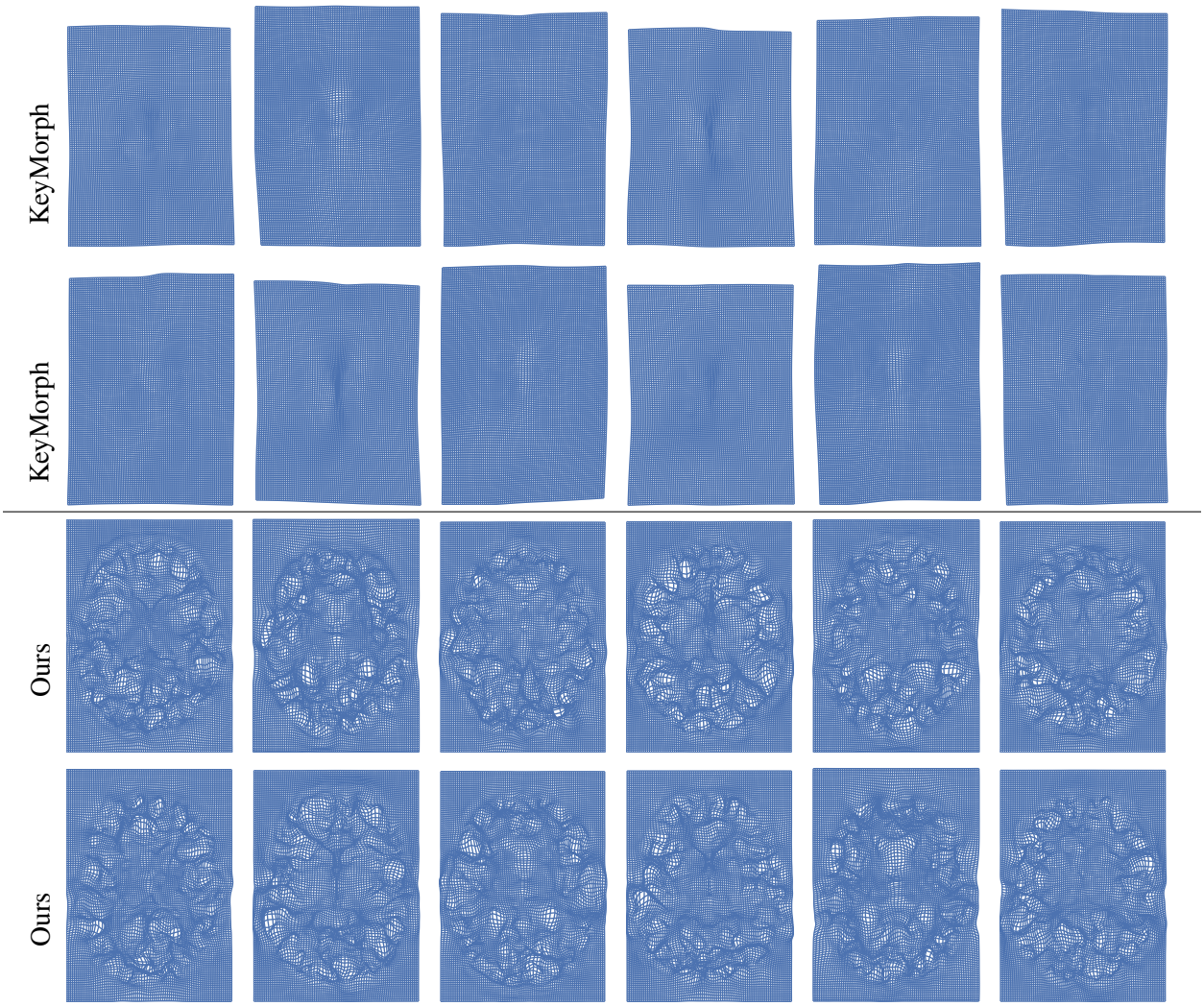


Figure C.7: Qualitative comparison of warp fields. Top two rows show the warp fields produced by thin plate spline using keypoints predicted by KeyMorph, bottom two rows show the warp fields produced by a diffeomorphic optimization routine from dense feature maps predicted by our method. Compared to the thin plate spline representation, our method is able to produce complex deformation fields to accurately capture subtle anatomical differences in inter-subject MRI registration.

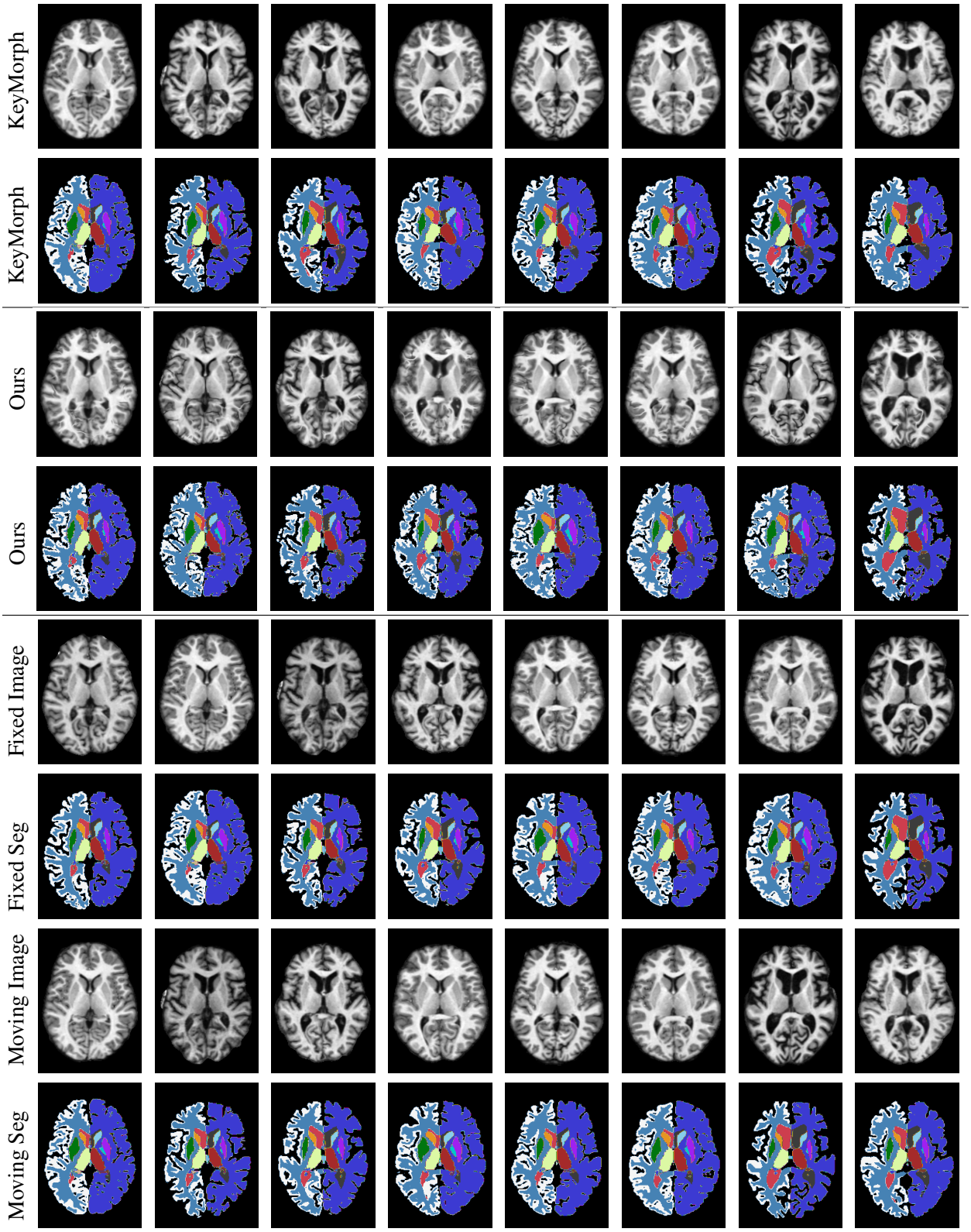


Figure C.8: Qualitative comparison of KeyMorph and our method on OASIS dataset. Qualitative evaluation of both labelmaps and intensity images shows that dense features from our method are instrumental in being robust and accurately registering complex deformable structures compared to sparse keypoints.

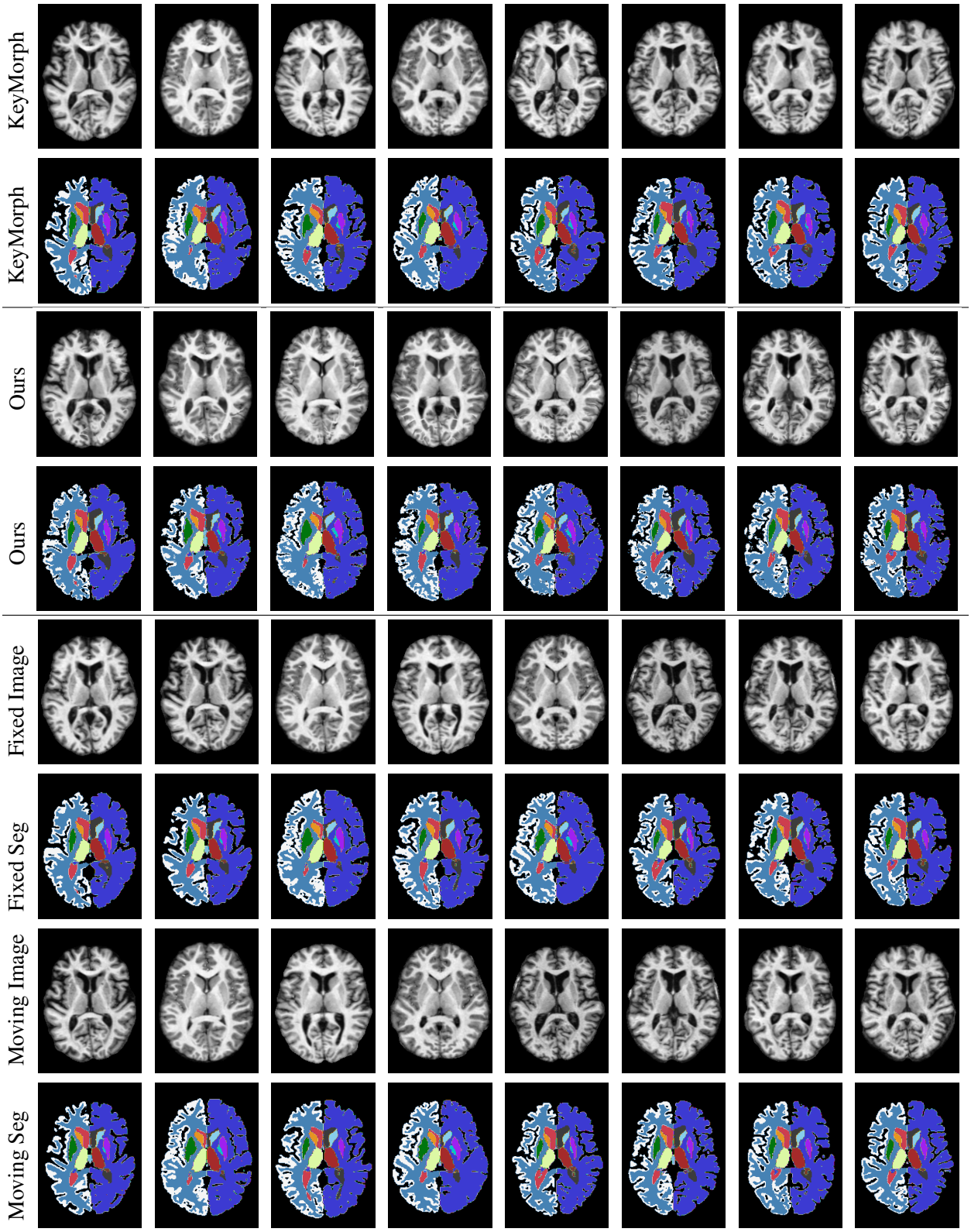


Figure C.9: Qualitative comparison of KeyMorph and our method on OASIS dataset. Qualitative evaluation of both labelmaps and intensity images shows that dense features from our method are instrumental in being robust and accurately registering complex deformable structures compared to sparse keypoints.

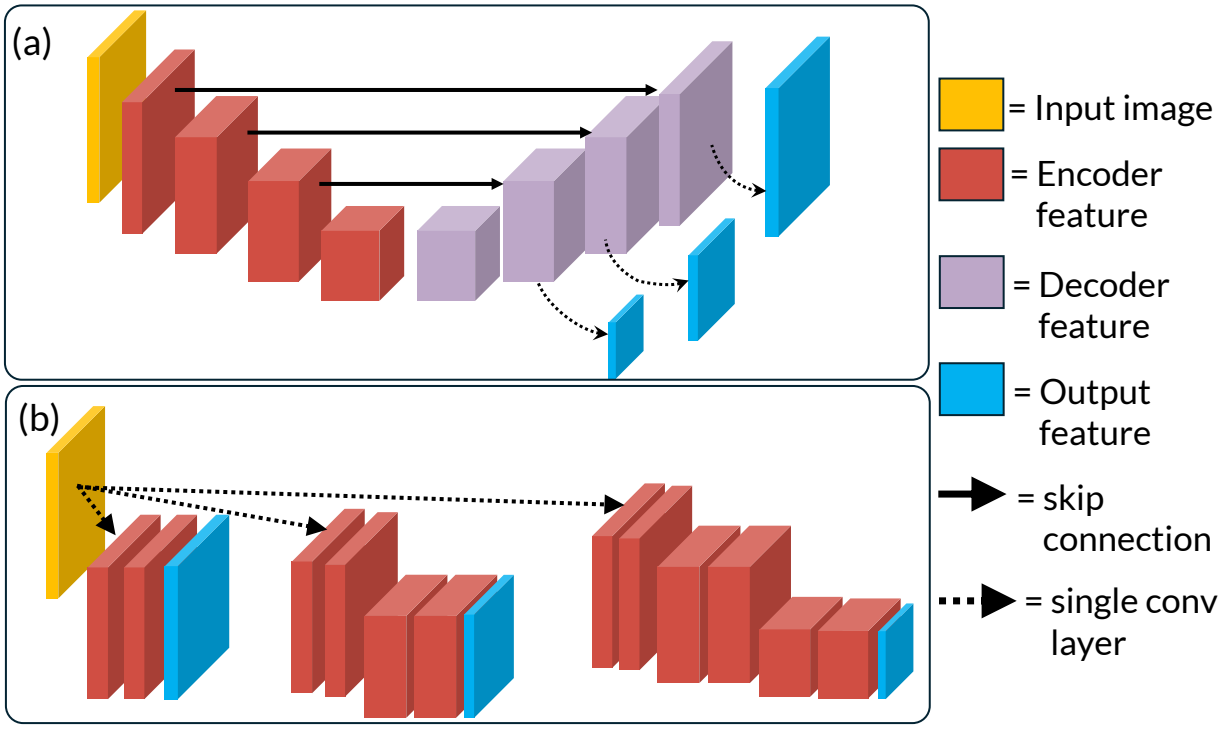


Figure C.10: Architecture details. **(a)** illustrates the UNet and Large Kernel U-Net (LKUNet) architecture designs, which consists of encoder blocks (red) and decoder blocks (purple) linked using skip connections. Multi-scale features are extracted from the intermediate decoder layers using a single convolutional layer. This design leads to shared features across multiple scales. UNet and LKUNet differ in the kernel parameters within each encoder and decoder blocks. **(b)** illustrates the 'Encoder-Only' versions of the same networks. The decoder path is entirely discarded, and each feature image is extracted using a separate encoder. This design enables independent learning of each multi-scale feature.

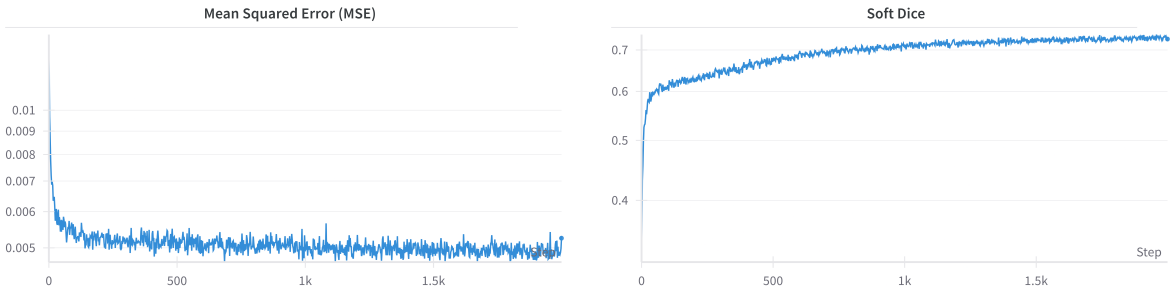


Figure C.11: Verifying convergence of KeyMorph. We verify the convergence of KeyMorph (with dice loss) on the OASIS dataset by plotting the Mean Squared Error (left) and Soft Dice (right) on the training set.

APPENDIX D

Supplementary details for A Scalable and Distributed Framework for Multimodal GigaVoxel Image Registration

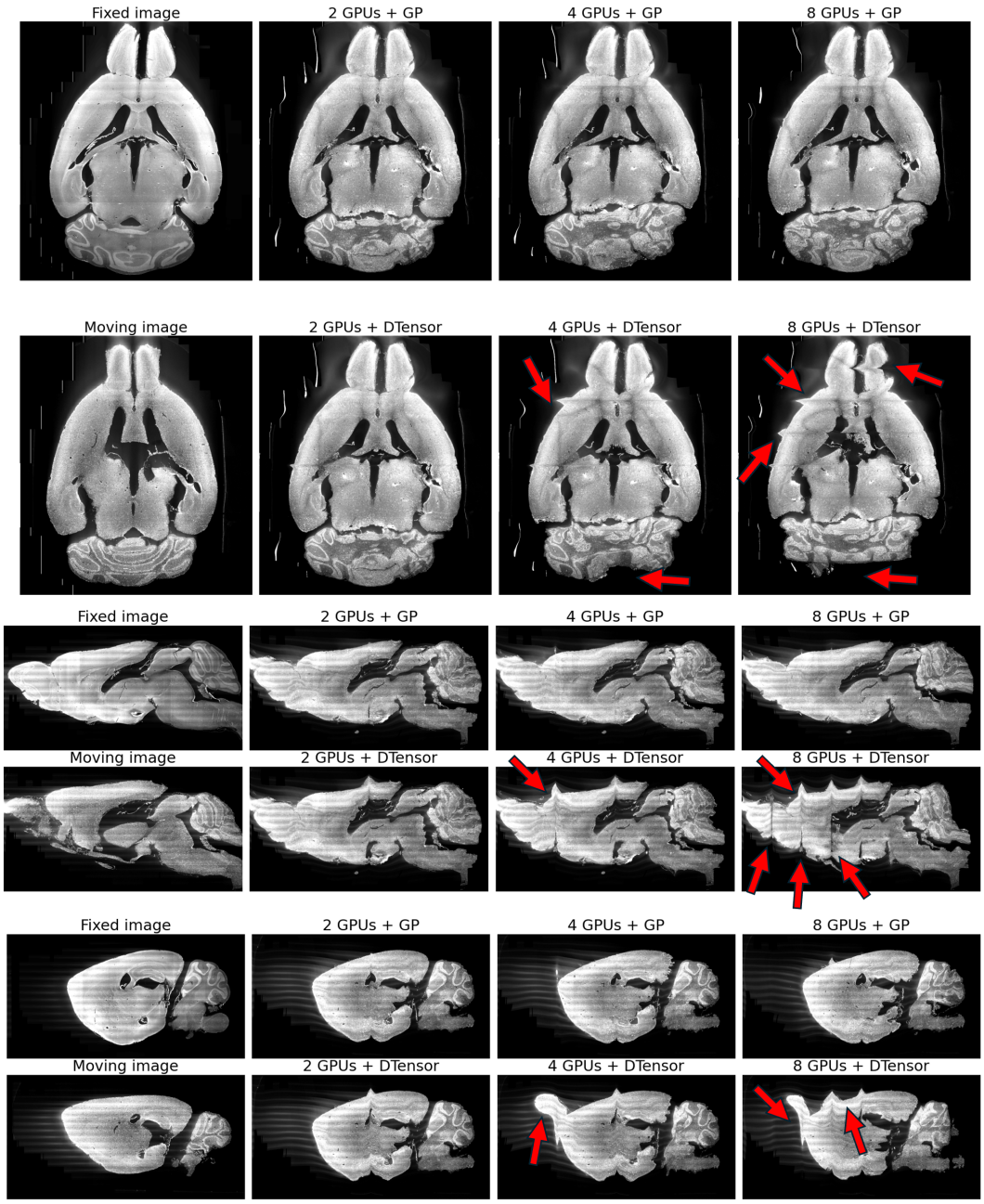


Figure D.1: Qualitative ablation of GP synchronization in FFDP on the fMOST mouse brain dataset. Red arrows highlight regions affected by incorrect boundary effects due to no synchronization.