

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Advancing 3D Medical Image Analysis: Registration, Representation, and Generation

Permalink

<https://escholarship.org/uc/item/80h8h6w3>

ISBN

9798270213367

Author

Han, Kun

Publication Date

2025-12-13

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Advancing 3D Medical Image Analysis: Registration, Representation, and Generation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Kun Han

Dissertation Committee:
Professor Xiaohui Xie, Chair
Professor Charless C. Fowlkes
Professor Alexander C. Berg

2025

DEDICATION

To my family.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	vi
LIST OF TABLES	x
LIST OF ALGORITHMS	xii
ACKNOWLEDGMENTS	xiii
VITA	xiv
ABSTRACT OF THE DISSERTATION	xvi
1 Introduction	1
2 Background	4
2.1 Image Registration	4
2.1.1 Traditional Methods	5
2.1.2 Learning-Based Methods	6
2.2 Neural Field Representation	7
2.2.1 Periodic Activation Functions	8
2.2.2 Deep Signed Distance Functions	8
2.3 Generative Models	9
2.3.1 Generative Adversarial Networks	10
2.3.2 Diffusion Models	10
3 Image Registration	12
3.1 Introduction	12
3.1.1 Deformable Registration	13
3.1.2 Diffeomorphic Registration	14
3.2 Neural Field Representation boosted Optimization Algorithm	14
3.2.1 Introduction	14
3.2.2 Method	16
3.2.2.1 Neural Velocity Field Representation	18
3.2.2.2 Diffeomorphic Deformation as Integration	19
3.2.2.3 Cascaded Registration	19
3.2.2.4 Optimization	20

3.2.3	Experiment	21
3.2.3.1	Datasets and Preprocessing	21
3.2.3.2	Experimental Setup	22
3.2.3.3	Evaluation Metrics	23
3.2.3.4	Results and Discussion	23
3.2.3.5	Ablation Studies	26
3.2.4	Conclusion	27
3.3	Learning Registration with Pseudo Groundtruth Generation	27
3.3.1	Introduction	27
3.3.2	Method	30
3.3.2.1	Overall Structure	30
3.3.2.2	Training with Pseudo Ground Truth	31
3.3.2.3	On-the-Fly Guidance	32
3.3.2.4	Online Optimizer	32
3.3.2.5	Implementation Details	33
3.3.3	Experiments	34
3.3.3.1	Experimental Setup	34
3.3.3.2	Image Registration Results	37
3.3.3.3	Training with Pseudo Ground Truth	38
3.3.3.4	On-the-Fly Guidance Strategy	38
3.3.3.5	Instance-Specific Optimization	38
3.3.3.6	Ablation Studies	39
3.3.4	Conclusion	40
3.4	Deformable Coordinate Attention for Multi-scale Registration	41
3.4.1	Introduction	41
3.4.2	Methodology	43
3.4.2.1	Problem Formulation	43
3.4.2.2	Architecture Overview	44
3.4.2.3	Multi-Scale Feature Extraction	44
3.4.2.4	Attention-Guided Deformation Propagation	45
3.4.2.5	Deformable Coordinate Attention	46
3.4.2.6	Correlation-guided Cross-scale Offset Prediction	46
3.4.2.7	Multi-Head Deformable Aggregation	49
3.4.2.8	Unsupervised Training Loss Function	50
3.4.3	Experiments	52
3.4.3.1	Datasets	52
3.4.3.2	Implementation Details	53
3.4.3.3	Evaluation Metrics	53
3.4.3.4	Comparison Methods	54
3.4.3.5	Brain Registration Results	54
3.4.3.6	AbdomenCT Registration Results	55
3.4.3.7	Ablation Study	56
3.4.3.8	Limitations	58
3.4.4	Conclusion	60

4	Shape Representation	61
4.1	Introduction	61
4.2	Neural Diffeomorphic Flow in Shape Modeling	62
4.2.1	Introduction	62
4.2.2	Method	64
4.2.2.1	Preliminaries	65
4.2.2.2	Review of NDF	66
4.2.2.3	Hybrid Shape Representation via Triplane	67
4.2.2.4	Topology-preserving Shape Generation	71
4.2.3	Experiments	72
4.2.3.1	Shape Representation	74
4.2.3.2	Point Correspondence and Shape Registration	75
4.2.3.3	Shape Generation	76
4.2.3.4	Ablation Study	76
4.2.4	Conclusion	77
5	Data Generation	79
5.1	Introduction	79
5.2	Enhancing OCT to OCTA Translation	80
5.2.1	Introduction	80
5.2.2	Method	82
5.2.2.1	Cross-Dimensional Supervision	82
5.2.2.2	Multi-Scale Feature Fusion	84
5.2.2.3	Overall Framework	85
5.2.3	Experiments	86
5.2.3.1	Datasets and Experimental Setup	86
5.2.3.2	Experimental Results	87
5.2.3.3	Ablation Study	89
5.2.4	Conclusion	89
5.3	Paired 3D Image and Mask Generation with Diffusion Model	90
5.3.1	Introduction	90
5.3.2	Method	92
5.3.2.1	Preliminary	92
5.3.2.2	3D Mask Generator	93
5.3.2.3	Conditional Image Generator	95
5.3.3	Experiments	96
5.3.3.1	Datasets and Setups	96
5.3.3.2	Evaluation of Synthetic Image Quality	98
5.3.3.3	Evaluation for Segmentation Tasks	98
5.3.4	Conclusion	100
6	Conclusion	102
	Bibliography	105

LIST OF FIGURES

	Page
3.1 Overall framework of the proposed DNVF and Cas-DNVF. I_m , I_f , and I_w denote the moving image, fixed image, and warped image, respectively. DNVF models the continuous velocity field v_θ using neural representations, where a 3D position $\mathbf{p} \in \mathbb{R}^3$ is mapped to a velocity vector $\mathbf{v} \in \mathbb{R}^3$ through an MLP with sinusoidal activation functions. The velocity integration is performed via scaling and squaring, as detailed in Sec. 3.2.2.2, implicitly capturing dense correspondences between two input images. Cas-DNVF integrates learning-based initialization with DNVF: an FCN is pretrained to predict an initial deformation, narrowing the search space so that DNVF can focus on refining local deformations with high accuracy and efficiency. While designed for 3D registration, we illustrate the framework using 2D images for simplicity.	17
3.2 Structure of the fully convolutional network (FCN) in Cas-DNVF. Input is the concatenation of moving and fixed images; output is the velocity field.	20
3.3 Qualitative registration results. Top: warped Mindboggle scans. Bottom: OASIS anatomical segmentations. DNVF and Cas-DNVF capture fine-scale cortical and subcortical deformations missed by other methods.	25
3.4 Accuracy–efficiency trade-off: DSC versus runtime across different iteration budgets.	25
3.5 Overview of the proposed OFG framework. It consists of a prediction stage (any registration model predicting a deformation field) and an optimization stage. During training, the optimizer refines the predicted deformation field for a small number of iterations, and the refined field is used as pseudo ground truth to supervise the registration model.	30
3.6 Illustration of on-the-fly guidance. At each iteration, the optimizer refines the predicted deformation field into a stronger pseudo label. Better predictions lead to better optimizations, which in turn provide stronger supervision, forming a positive feedback loop.	32
3.7 Loss landscape comparison between TransMorph with and without OFG across two datasets. OFG produces flatter and smoother optimization surfaces, improving trainability and convergence.	34
3.8 Qualitative registration results on IXI [88]. Each row shows comparison between baseline (row 2) and its OFG-enhanced counterpart (row 1) for TransMorph and ViT-V-Net. OFG produces sharper alignment while avoiding irregular deformations.	36

3.9	Training dynamics on LPBA40. We compare self-training (using fixed pre-trained deformations), optimized self-training (refining these fields offline), and OFG. OFG provides more stable convergence and higher validation DSC, particularly under limited data.	36
3.10	Ablation study on IXI (left) and LPBA40 (right). Increasing reliance on unsupervised losses degrades performance, highlighting the importance of supervision from OFG’s pseudo labels.	39
3.11	Deformable Coordinate Attention (DCA). DCA computes attention between a fixed-image feature and a set of sampled moving-image features guided by predicted correspondence offsets. These offsets adaptively determine candidate locations, and the resulting attention weights produce the final correspondence for deformation propagation.	41
3.12	Overview of the Proposed Method. The shared encoder extracts multi-scale representations from moving and fixed images, which are then used in DCA (Deformable Coordinate Attention) blocks in a coarse-to-fine registration manner. At each scale, DCA combines Correlation-guided Cross-scale Offset Prediction (CCOP) with Multi-Head Deformable Aggregation (MHDA) to iteratively refine the deformation field through attention-guided correspondence discovery and propagation.	43
3.13	Detailed illustration of Deformable Coordinate Attention (DCA). DCA consists of two modules: (1) CCOP predicts multi-head unconstrained spatial offsets by combining projected features, local-window correlations, and propagated features from the previous scale, while also generating features for the next scale. (2) MHDA uses these offsets to dynamically place attention keys, where queries come from fixed features, keys from warped moving features, and values from sampled positions. Attention weights highlight correspondence reliability, and the multi-head design captures diverse deformation patterns for robust registration.	46
3.14	Network Structure for the Encoder layers and Offset Predictor.	48
3.15	Dice score distributions on LPBA40 (54 ROIs).	55
3.16	Dice score distributions on AbdomenCT (13 ROIs).	57
3.17	Qualitative comparisons on LPBA40 (top) and AbdomenCT (bottom). . . .	58
3.18	Failure cases. Left: unnecessary cortical folding (red box). Right: high image similarity but misaligned structures.	59
4.1	Shape Representation and Shape Generation. Left half presents the diffeomorphic deformation from the learned template ($t=0$) to instance shapes ($t=1$), with color highlighting the dense correspondence captured by triplane features. Right half presents the denoising process for shape generation. The shapes are generated as deformed templates and the 3D deformation is controlled by the generated triplane features from diffusion.	65

4.2	Shape Representation. The framework consists of a deformation module $\mathcal{D}$, a template module $\mathcal{T}$, and per-object triplane features X_i . Given a point p in the instance space, we compute its corresponding canonical point p' in the template space using Eq. 4.1. The template module then evaluates the signed distance value s' . Training optimizes the entire framework by minimizing the L_1 difference between s' and the ground-truth s , along with regularization terms.	67
4.3	Reconstruction comparison. Left: our method with hybrid supervision. Middle: ground truth. Right: purely local supervision, which fails to recover correct global correspondence.	69
4.4	Shape Generation. Triplane features represented as multi-channel images. We train a 2D diffusion model to generate triplane features, which encode diffeomorphic deformations that warp the template into new shapes.	71
4.5	Reconstruction results on unseen data.	75
4.6	Generated 3D shapes across different methods.	76
5.1	OCT/OCTA volumes with retinal layer segmentation and en-face projections. Differences between Proj _{ILM-OPL} and Proj _{OPL-BM} highlight the heterogeneous imaging properties arising from distinct cellular and vascular distributions across retinal layers.	81
5.2	Overview of XOCT. The framework combines volumetric OCT-to-OCTA translation with Cross-Dimensional Supervision (CDS) and Multi-Scale Feature Fusion (MSFF).	83
5.3	Qualitative comparison of OCT-to-OCTA translation methods across en-face projections (Proj_{full} , Proj_{mean}) and a representative 3D z -slice. Red arrows indicate regions where XOCT better preserves vascular continuity, reconstructs vessels missed by other models, and avoids over-smoothing of subtle vessel dropouts.	88
5.4	Overview of MedGen3D. The framework first autoregressively generates a 3D mask volume using MC-DPM, then synthesizes realistic images conditioned on the generated masks via a sequence-to-sequence generator and a semantic diffusion refiner.	93
5.5	3D mask generator with MC-DPM. For a target position z , MC-DPM generates subsequences (length m) either unconditionally or conditioned on the first/last n slices, chosen with probabilities $p^C \in \{p_F, p_B, p_U\}$. Red boxes: conditional slices; green boxes: generated slices.	94
5.6	Conditional Image Generation. Vid2Vid synthesizes slices sequentially given masks. A semantic diffusion refiner further enhances image fidelity by adding small noise and denoising with SDMs trained on three orthogonal views. The final result averages the three refined reconstructions.	95
5.7	Qualitative comparison of synthetic data. Our method produces anatomically consistent images with sharper organ boundaries and more realistic textures than 3D- α -WGAN and Vid2Vid. Left: comparison across generative models. Right: visualization of synthetic MRI slices at different anatomical positions.	97

5.8	Experiment 2: Qualitative comparison of training strategies with Swin-UNETR. Improvements from synthetic data are more pronounced when training and testing are cross-dataset (e.g., train on OASIS, test on ADNI) compared to in-domain (OASIS*).	101
5.9	Experiment 3: Transfer learning strategies with Swin-UNETR. Synthetic pre-training enhances finetuning performance on unseen datasets (see Table 5.6).	101

LIST OF TABLES

	Page
3.1 The registration methods are evaluated using anatomical structures (31 for Mindboggle and 35 for OASIS) in atlas-based fashion. FCN is the model used in C2F-DNVF which provides better topology preservation compared with other learning-based methods. Table 3.5 shows the results of replacing FCN with other SOTA methods in C2F-DNVF. For instance-specific optimization of learning-based methods, we finetune the pretrained model for each given pair of test images.	24
3.2 Following NODEO, we use the same data split and evaluation metric. The DSC for OASIS is averaged over 28 structures excluding the very small ones and the DSC for CANDI is averaged over all 32 structures and 28 large structures as stated in NODEO’s setting. We limit the optimization of DNVF to 100 iteration (60s) compared with NODEO (80s)	24
3.3 Results using sinusoidal and ReLU activation functions.	26
3.4 Results with different velocity field representations.	26
3.5 Results with different pretrained FCN in Cas-DNVF.	26
3.6 Evaluation results for different methods on various datasets. The OFG architecture provides significant and substantial improvement on the unsupervised learning-based methods. These results validate OFG’s effectiveness and generalizability without compromising on inference efficiency.	35
3.7 Comparison between the 3 models’ convergence losses on 3 datasets. Note that the loss function settings were the same. For TransMorph on OASIS, the convergence loss is clearly the lowest.	37
3.8 Comparison between instance-specific optimizer (OFG) and network-based optimizer on IXI dataset[88]. The outcomes presented are derived from the initial 200 epochs.	38
3.9 Ablation results of the OFG Optimizer on LPBA40 Dataset. This evaluation focuses over the initial 30 epochs.	40
3.10 Quantitative comparisons on MindBoggle, LPBA, and OASIS datasets. Red , Green and Blue denote the 1st , 2nd and 3rd best.	55
3.11 Quantitative comparison on AbdomenCT dataset. Red , Green and Blue denote the 1st , 2nd and 3rd best results.	56

3.12	Ablation study on AbdomenCT dataset. Each component’s contribution is evaluated by adding modules to the baseline. CG : Correlation-Guidance; CS : Cross-Scale Feature Propagation; DA : Single-Head Deformation Aggregation; MH : Multi-Head Mechanism	57
3.13	Hyperparameter Tuning for Negative Jacobian Loss	58
3.14	Deformable Attention Size	58
4.1	Shape Reconstruction and Shape Registration results on Unseen Shapes. The chamfer distance results shown above are multiplied by 10^3 . DeepSDF and NFD cannot model the deformation between shapes, therefore, they are unable to conduct the shape registration task.	74
4.2	Shape generation results. Our method achieve better performance according to the FID, precision and recall. $\triangle$ denotes sampling from Gaussian distribution while $\star$ denotes sampling from PCA.	75
4.3	Shape Reconstruction with various supervision.	77
4.4	Shape Reconstruction with various feature representations.	77
5.1	OCT to OCTA translation results on the OCTA-3M and OCTA-6M datasets. $\downarrow$ indicates lower is better, $\uparrow$ indicates higher is better.	87
5.2	Ablation Study on Proposed Modules. $\downarrow$: lower is better, $\uparrow$: higher is better.	89
5.3	Synthetic image quality comparison between baselines and ours.	98
5.4	Experiment 2: DSC of different thoracic segmentation models. There are 5 training strategies, namely: E2-1 : Training with real SegTHOR training data; E2-2 : Training with synthetic data; E2-3 : Training with both synthetic and real data; E2-4 : Finetuning model from E2-2 using real training data; and E2-5 : finetuning model from E2-3 using real training data. (* denotes the training data source.)	99
5.5	Experiment 2: DSC of brain segmentation models. Please refer to Table 5.4 for detailed training strategies. (* denotes the training data source.)	99
5.6	Experiment 3: DSC of Swin-UNETR finetuned with real dataset. There are 4 training strategies: E3-1 : Training from scratch for each dataset using limited finetuning data; E3-2 Finetuning the model E2-1 from experiment 2; E3-3 Finetuning the model E2-4 from experiment 2; and E3-4 Finetuning the model E2-5 from experiment 2. (* denotes the finetuning data source.) . .	100

LIST OF ALGORITHMS

	Page
1 3D Mask Generation (Inference)	96

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Xiaohui Xie. His unwavering support, invaluable guidance, and profound expertise were instrumental in the completion of this dissertation. His passion for research and commitment to academic excellence have been a constant source of inspiration, shaping me into the researcher I am today.

I am deeply thankful to my family for their endless love and encouragement. To my parents, thank you for your unconditional support and for instilling in me the values of hard work and perseverance. To my wife, your patience, sacrifice, and unwavering belief in me have been my greatest source of strength. To my dog, Guaiguai, your boundless energy and unconditional affection brought comfort and joy through countless long nights of research and writing.

I am also grateful to my remarkable lab members, whose camaraderie, collaboration, and intellectual curiosity enriched my Ph.D. experience. A special thank you goes to Shanlin Sun, Haoyu Ma, Xiangyi Yan, Thanh-Tung Le, Yifeng Xiong, Yuelin Xin, Hao Tang, Deying Kong, Pooya Khosravi, Arghavan Rezvani, and Anthony Wu. Our shared challenges, discussions, and mutual support made the lab a true community.

I would like to thank my dissertation committee members, Prof. Charless C. Fowlkes and Prof. Alexander C. Berg, for their valuable feedback and insightful comments that greatly improved this dissertation.

Finally, I wish to acknowledge the University of California, Irvine, for providing the resources and supportive environment that made this research possible.

VITA

Kun Han

EDUCATION

Doctor of Philosophy in Computer Science University of California, Irvine	2025 <i>Irvine, California</i>
Master of Science in Computer Engineering University of California, Irvine	2020 <i>Irvine, California</i>
Bachelor of Science in Electric Engineering Southeast University	2018 <i>Nanjing, Jiangsu</i>

RESEARCH EXPERIENCE

Engineering Intern Apple	2024 <i>Cupertino, California</i>
Graduate Research Assistant University of California, Irvine	2023 <i>Irvine, California</i>
Research Intern United Imaging Intelligence	2022 <i>Cambridge, Massachusetts</i>

TEACHING EXPERIENCE

Teaching Assistant University of California, Irvine	2021–2025 <i>Irvine, California</i>
---	---

Selected Papers

- First authored papers:

Diffeomorphic Image Registration With Neural Velocity Field	2023
Winter Conference on Applications of Computer Vision (WACV)	
MedGen3D: A Deep Generative Framework for Paired 3D Image and Mask Generation	2023
Medical Image Computing and Computer Assisted Intervention(MICCAI)	
Hybrid Neural Diffeomorphic Flow for Shape Representation and Generation via Triplane	2024
Winter Conference on Applications of Computer Vision (WACV)	
On-the-Fly Guidance Training for Medical Image Registration	2024
Medical Image Computing and Computer Assisted Intervention(MICCAI)	
XOCT: Enhancing OCT to OCTA Translation via Cross-Dimensional Supervised Multi-Scale Feature Learning	2025
Medical Image Computing and Computer Assisted Intervention(MICCAI)	
Deformable Coordinate Attention for Medical Image Registration	2025
Under Review	

- Co-authored papers:

Topology-Preserving Shape Reconstruction and Registration via Neural Diffeomorphic Flow	2022
Conference on Computer Vision and Pattern Recognition (CVPR)	
Localized Region Contrast for Enhancing Self-supervised Learning in Medical Image Segmentation	2023
Medical Image Computing and Computer Assisted Intervention(MICCAI)	
Medical Image Registration via Neural Fields	2024
Medical Image Analysis	
AFTer-SAM: Adapting SAM With Axial Fusion Transformer for Medical Imaging Segmentation	2024
Winter Conference on Applications of Computer Vision (WACV)	
Self-Sampling Meta SAM: Enhancing Few-Shot Medical Image Segmentation With Meta-Learning	2024
Winter Conference on Applications of Computer Vision (WACV)	

ABSTRACT OF THE DISSERTATION

Advancing 3D Medical Image Analysis: Registration, Representation, and Generation

By

Kun Han

Doctor of Philosophy in Computer Science

University of California, Irvine, 2025

Professor Xiaohui Xie, Chair

The rapid proliferation of advanced medical imaging modalities, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), has generated vast volumes of three-dimensional data that are central to modern clinical practice. While these datasets offer unprecedented opportunities for diagnosis, treatment planning, and disease monitoring, they also pose major computational challenges: robust methods are required to capture anatomical variability, ensure topological plausibility, and generalize across patient populations. Traditional approaches, relying on hand-crafted features and iterative optimization, are often limited in efficiency and scalability. More recently, deep learning has transformed medical image analysis by enabling end-to-end frameworks that automatically extract hierarchical features and model complex anatomical structures.

This dissertation advances the state of 3D medical image analysis across three key research areas. First, in **image registration**, it introduces neural velocity field representations for accurate diffeomorphic transformations, an on-the-fly guidance framework that generates pseudo ground truth during training, and an adaptive attention-based model for coarse-to-fine correspondence discovery. Together, these contributions seek to overcome limitations of both traditional and early learning-based methods. Second, in **3D shape representation**, it presents a hybrid neural diffeomorphic flow that learns a common anatomical

template and generates new, biologically plausible instances while preserving topological consistency. Third, in **data generation**, it develops advanced generative models to address dataset scarcity: a cross-dimensional supervision framework for OCT-to-OCTA translation that reconstructs fine vascular details, and a paired 3D image–mask generation model that synthesizes realistic volumetric data for downstream tasks.

Collectively, this work demonstrates how deep learning can deliver accurate, efficient, and generalizable solutions for 3D medical imaging. By advancing registration, representation, and data generation, the dissertation lays the foundation for intelligent computational tools that enhance clinical diagnosis and biomedical research.

Chapter 1

Introduction

Medical image analysis is a cornerstone of modern clinical practice, empowering clinicians with non-invasive tools for diagnosis, surgical planning, and longitudinal monitoring of disease. The advent of advanced imaging modalities, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), has shifted attention toward three-dimensional (3D) volumetric data, offering unprecedented insights into human anatomy and pathology. However, analyzing these high-dimensional datasets poses substantial computational and methodological challenges, including data scarcity, large inter-subject anatomical variability, and the complexity of non-linear spatial transformations.

For decades, progress relied on conventional image processing techniques that emphasized hand-crafted feature extraction and iterative optimization. While these methods provided important foundations, they were computationally expensive, sensitive to initialization, and often failed to generalize across diverse patient populations. The emergence of deep learning has fundamentally transformed this paradigm. By leveraging hierarchical feature learning and modern computational resources, deep neural networks have enabled a transition from feature engineering to end-to-end learning. This shift has unlocked new capabilities: learning

complex, non-linear mappings directly from data, capturing subtle anatomical deformations, and inferring underlying structures with a precision previously unattainable. This dissertation is motivated by the conviction that deep learning can transcend the limitations of traditional, rule-based methods to deliver more accurate, efficient, and generalizable solutions for clinical applications. To this end, the research is organized around three pillars—image registration, shape representation, and data generation—that together establish a new paradigm for automated 3D medical image analysis.

Despite substantial progress, several fundamental challenges remain. In image registration, which aligns multiple images into spatial correspondence, a persistent tension exists between the computational efficiency of learning-based methods and their generalization to capture fine-grained deformations while preserving topology. For 3D shape representation, existing approaches often fail to create compact and continuous models that simultaneously establish dense correspondences between anatomical instances and preserve topological consistency, limiting their utility for disease progression modeling and surgical simulation. Finally, the development of robust data-driven methods is hampered by the scarcity of high-quality labeled datasets, which are costly and time-consuming to acquire. This bottleneck significantly constrains the development of generalizable models for core tasks such as segmentation and classification.

This dissertation addresses these challenges through three major contributions. First, in **image registration**, it introduces novel frameworks that integrate the strengths of optimization-based and learning-based methods. A neural velocity field representation enables highly accurate, diffeomorphic transformations; an on-the-fly guidance strategy generates pseudo ground truth during training, bridging the gap between supervised and unsupervised learning; and an adaptive attention-based framework improves coarse-to-fine correspondence discovery. Together, these contributions push registration toward higher accuracy, robustness, and clinical applicability. Second, in **shape representation**, the dissertation presents a

hybrid neural diffeomorphic flow method that learns a common anatomical template and generates new, biologically plausible shapes as deformations of this template. This design preserves topology while establishing dense correspondences, offering a powerful tool for analyzing anatomical variability and building realistic digital phantoms. Third, in **data generation**, advanced generative models are proposed to address dataset scarcity. These include a cross-dimensional supervision framework for OCT-to-OCTA translation that reconstructs fine vascular structures with high fidelity, and a paired 3D image–mask generation pipeline that produces realistic volumetric images with corresponding segmentations for downstream tasks.

The remainder of this dissertation is organized as follows. Chapter 2 reviews relevant background in image registration, neural field representations, and generative models. Chapter 3 introduces novel frameworks for 3D image registration, including neural velocity fields, on-the-fly pseudo ground truth generation, and deformable coordinate attention for coarse-to-fine alignment. Chapter 4 investigates 3D shape representation and generation, presenting a hybrid neural diffeomorphic flow method that learns a template and generates new instances as its deformations. Chapter 5 addresses the challenge of data scarcity with generative approaches, including OCT-to-OCTA translation and paired image–mask synthesis. Finally, Chapter 6 concludes with a summary of contributions and discusses promising directions for future research.

Chapter 2

Background

This chapter provides the theoretical and methodological foundations that underpin the contributions of this dissertation. We review three interrelated domains that form the basis for our work in 3D medical image analysis. First, we discuss image registration, a fundamental task for establishing spatial correspondence across images. We then introduce neural field representations, which offer continuous and differentiable models of anatomical structures and transformations. Finally, we survey generative models, particularly adversarial and diffusion-based approaches, which address the persistent challenge of data scarcity in medical imaging. Together, these areas provide the conceptual framework that motivates and supports the research presented in subsequent chapters.

2.1 Image Registration

Image registration is a cornerstone of medical image analysis, aiming to align two or more images such that corresponding anatomical structures are brought into correspondence. This task underpins numerous applications including longitudinal disease monitoring, multi-

modal image fusion, atlas-based analysis, and surgical planning. Conceptually, registration estimates a spatial transformation that warps a moving image I_m onto a fixed image I_f while preserving anatomical plausibility. Methods can be broadly divided into *traditional optimization-based* approaches and more recent *learning-based* frameworks.

2.1.1 Traditional Methods

Classical methods formulate registration as a variational optimization problem. Given $I_f : \Omega \rightarrow \mathbb{R}$ as the fixed image and $I_m : \Omega \rightarrow \mathbb{R}$ as the moving image over domain $\Omega \subset \mathbb{R}^3$, the goal is to find a transformation $\phi : \Omega \rightarrow \Omega$ that minimizes

$$\phi^* = \arg \min_{\phi} \mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi), \quad (2.1)$$

where $\mathcal{L}_{\text{sim}}$ measures alignment (e.g., sum of squared differences, normalized cross-correlation, mutual information) and $\mathcal{L}_{\text{reg}}$ enforces smoothness or physical plausibility.

Different parameterizations of ϕ gave rise to distinct methods. Elastic and fluid models drew on biomechanical analogies, free-form deformations with B-splines offered local flexibility with global smoothness, and optical-flow-based formulations estimated voxel-wise correspondences under brightness constancy assumptions. Among the most influential contributions are diffeomorphic frameworks such as Large Deformation Diffeomorphic Metric Mapping (LDDMM) and Symmetric Normalization (SyN), which model ϕ as the integration of a smooth velocity field v_t :

$$\frac{d\phi_t}{dt} = v_t(\phi_t), \quad \phi_0 = \text{id}, \quad \phi = \phi_1, \quad (2.2)$$

ensuring smooth, invertible mappings that preserve topology. These guarantees are critical in clinical contexts, where foldings or overlaps would be biologically implausible. Despite

high accuracy, such methods are computationally demanding, initialization-sensitive, and dependent on carefully designed similarity metrics.

2.1.2 Learning-Based Methods

Deep learning reframed registration as prediction rather than iterative optimization. Instead of solving anew for each image pair, a neural network f_θ predicts the deformation directly:

$$\phi_\theta = f_\theta(I_f, I_m). \quad (2.3)$$

Training objectives mirror the classical variational formulation:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi_\theta) + \lambda \mathcal{L}_{\text{reg}}(\phi_\theta). \quad (2.4)$$

This paradigm was popularized by VoxelMorph [12], which showed that unsupervised training on large image sets could yield accurate deformations without ground-truth labels. Once trained, such models perform inference in a single forward pass, enabling near real-time applications. Subsequent work expanded this foundation: coarse-to-fine hierarchies improved alignment across scales, attention and transformer blocks incorporated long-range dependencies, and probabilistic formulations modeled $p(\phi \mid I_f, I_m)$ for uncertainty quantification. Other innovations—correlation-based similarity, recursive refinement, and hybrid convolution-attention designs—further advanced robustness and flexibility.

Nevertheless, challenges remain. Ensuring topology preservation without explicit diffeomorphic guarantees is difficult. Trade-offs between computational efficiency and fine-grained accuracy persist, and many high-performing approaches require extensive annotated data. Some of these limitations can be alleviated through interleaving pseudo-ground truth su-

pervision or adopting more advanced architectures. At the same time, exploring alternative ways to represent deformations and anatomical structures offers another promising direction, motivating our investigation into neural field-based representations.

2.2 Neural Field Representation

Neural fields, also referred to as implicit neural representations, have emerged as a powerful paradigm for modeling continuous signals. Instead of explicitly storing discrete elements such as voxels in a 3D grid or vertices in a mesh, neural fields parameterize a function with a neural network that can be queried at arbitrary spatial coordinates. Formally, for a point $\mathbf{x} \in \mathbb{R}^n$, a neural field outputs a signal value $\mathbf{s}$:

$$f_{\theta}(\mathbf{x}) = \mathbf{s}, \tag{2.5}$$

where θ denotes the network parameters. The output $\mathbf{s}$ may correspond to occupancy, signed distance, color, density, or other attributes depending on the application. By defining the signal continuously rather than discretely, neural fields enable rendering at arbitrary resolution, smooth interpolation, and compact storage. This is especially advantageous in high-resolution 3D modeling: while a voxel grid with resolution N^3 scales cubically in memory, neural field representations avoid explicit grid storage and offer more flexible data access. Their memory footprint primarily depends on the network architecture and input encoding, while allowing queries at arbitrary resolutions without being tied to a fixed discretization.

The appeal of neural fields lies in their generality. With a single formulation, they can represent 2D signals, 3D geometries, or even dynamic scenes. For medical imaging, this flexibility is particularly valuable: neural fields can capture fine structural details while mitigating the memory bottlenecks of volumetric data, enabling compact yet high-fidelity

models of anatomy.

2.2.1 Periodic Activation Functions

A key limitation of early neural field models was the *spectral bias* of standard neural networks. Architectures with ReLU or tanh activations tend to favor low-frequency solutions, making them effective at modeling smooth signals but poor at representing sharp details such as edges or fine anatomical boundaries.

Sitzmann et al. [124] addressed this challenge with Sinusoidal Representation Networks (SIRENs), which replace standard activations with periodic functions:

$$y = \sin(W\mathbf{x} + b). \quad (2.6)$$

The oscillatory nature of sine activations aligns naturally with high-frequency components of signals, enabling networks to reconstruct sharp surfaces and intricate details.

2.2.2 Deep Signed Distance Functions

Another major line of research focused on geometric representation through the *Signed Distance Function* (SDF), which compactly describes surfaces. For a point $\mathbf{x} \in \mathbb{R}^3$,

$$\text{SDF}(\mathbf{x}) = \begin{cases} -d(\mathbf{x}, \partial\Omega), & \mathbf{x} \in \Omega, \\ +d(\mathbf{x}, \partial\Omega), & \mathbf{x} \notin \Omega, \end{cases} \quad (2.7)$$

where $\partial\Omega$ is the boundary of a shape Ω . The surface itself corresponds to the zero-level set.

Park et al. [109] introduced DeepSDF, which learns SDFs using neural fields. A shared

network f_θ is conditioned on a latent shape code $\mathbf{z}$ to approximate the SDF:

$$f_\theta(\mathbf{x}, \mathbf{z}) \approx \text{SDF}(\mathbf{x}). \quad (2.8)$$

Here, $\mathbf{z}$ encodes the geometry of an individual instance, while f_θ captures shared structure across a class of shapes. This enables interpolation between shapes, completion from partial data, and generative modeling of novel anatomies.

In summary, Neural field representations provide a compact, continuous, and flexible framework for modeling complex signals. By parameterizing functions with neural networks, they overcome the memory inefficiency of voxel grids while supporting arbitrary-resolution queries and smooth interpolation. Advances such as periodic activation functions mitigate spectral bias to recover fine details, while signed distance function-based formulations enable accurate and generative shape modeling. Together, these developments make neural fields a versatile representation that balances efficiency, expressiveness, and adaptability—qualities that are particularly advantageous for high-resolution 3D medical imaging and deformation modeling.

2.3 Generative Models

Generative models aim to capture the underlying data distribution $p_{\text{data}}(\mathbf{x})$ so that new samples can be synthesized that are statistically consistent with observed data. In medical imaging, this ability directly addresses challenges of limited annotated data and the need for realistic, high-fidelity images of anatomy and pathology. By generating synthetic yet plausible data, these models enable dataset augmentation, support cross-modality translation, and enhance downstream tasks such as segmentation, registration, and diagnosis.

Among the diverse generative approaches, two families stand out: *Generative Adversarial*

Networks (GANs) and Diffusion Models. GANs generate images through adversarial training, excelling at sharp, high-fidelity synthesis, while diffusion models adopt a probabilistic denoising formulation that ensures stable training and principled likelihood estimation. Both have been transformative, offering complementary strengths.

2.3.1 Generative Adversarial Networks

GANs, introduced by Goodfellow et al. [50], train two networks in competition: a generator G that maps noise $\mathbf{z} \sim p(\mathbf{z})$ to synthetic samples, and a discriminator D that distinguishes real from generated data:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (2.9)$$

GANs rapidly established themselves as powerful image synthesizers, producing sharper and more realistic samples than earlier likelihood-based models. Variants such as the Wasserstein GAN (WGAN) improved stability by replacing the Jensen–Shannon divergence with the Earth Mover’s distance, while conditional GANs (cGANs) enabled targeted synthesis for tasks like CT-to-MRI translation and lesion simulation. Despite these advances, GANs remain prone to instability and mode collapse, and their lack of explicit likelihood complicates evaluation.

2.3.2 Diffusion Models

Diffusion models adopt a contrasting philosophy by framing generation as the reversal of a gradual noising process. Starting from a clean sample $\mathbf{x}_0 \sim p_{\text{data}}$, Gaussian noise is added

over T steps until the signal becomes indistinguishable from pure noise $\mathbf{x}_T$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t \mid \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t I\right). \quad (2.10)$$

A neural network $\epsilon_\theta(\mathbf{x}_t, t)$ is trained to predict the added noise, with objective

$$\mathcal{L}_{\text{DM}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \right\|_2^2 \right]. \quad (2.11)$$

Sampling begins from pure noise and iteratively applies the learned denoising steps until a clean image emerges.

Diffusion models avoid mode collapse, allow likelihood estimation, and offer stable training. Variants such as Denoising Diffusion Implicit Models (DDIM) accelerate sampling, while latent diffusion reduces computation by operating in compressed spaces. Their robustness and flexibility have made them particularly impactful in medical imaging, powering advances in super-resolution, denoising, cross-modality translation, and dataset augmentation.

In summary, GANs and diffusion models provide two complementary paradigms. GANs produce high-fidelity images efficiently and excel in conditional settings, whereas diffusion models offer stability, diversity, and a principled probabilistic foundation. Together, they form the backbone of modern generative modeling and underpin many of the strategies developed in this dissertation.

Chapter 3

Image Registration

3.1 Introduction

Image registration is a cornerstone of medical image analysis, underpinning applications such as longitudinal disease monitoring, multi-modal fusion, population-based studies, and surgical planning. The task involves estimating a spatial transformation $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ that aligns a moving image I_m with a fixed image I_f so that corresponding anatomical structures are brought into correspondence:

$$I_f(\mathbf{x}) \approx I_m(\phi(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^3. \quad (3.1)$$

For clinical utility, ϕ must be smooth, invertible, and topology-preserving, ensuring anatomically meaningful mappings.

Classical registration methods cast this as an optimization problem, refining ϕ to maximize similarity subject to regularization. Diffeomorphic frameworks such as LDDMM and SyN provide strong guarantees of anatomical plausibility but are computationally intensive and

initialization-sensitive. Deep learning has transformed the field by shifting from pair-specific optimization to end-to-end prediction, enabling near real-time registration and improved robustness. Yet, challenges remain: enforcing diffeomorphic guarantees, handling limited supervision, and effectively capturing deformations across scales.

This dissertation tackles these issues through three contributions. First, we introduce continuous **Neural Velocity Fields** for diffeomorphic registration as an upgrade of the traditional iterative optimization methods. Second, we propose an **On-the-Fly Guidance** strategy that bridges unsupervised and supervised learning by dynamically generating pseudo-ground truth deformations. Third, we develop a **Deformable Coordinate Attention** mechanism that enforces reliable correspondence propagation across scales. Together, these advances improve registration accuracy, robustness, and clinical applicability.

3.1.1 Deformable Registration

Deformable registration estimates a dense displacement field ϕ that warps the moving image I_m to the fixed image I_f , expressed as $(I_m \circ \phi)(\mathbf{x}) = I_m(\phi(\mathbf{x}))$. The problem is typically formulated as minimization of an energy functional:

$$\phi^* = \arg \min_{\phi} \mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi), \quad (3.2)$$

where $\mathcal{L}_{\text{sim}}$ measures image similarity (e.g., mean squared error, normalized cross-correlation, or mutual information), $\mathcal{L}_{\text{reg}}$ enforces smoothness or topology preservation, and λ balances the terms.

3.1.2 Diffeomorphic Registration

While deformable registration is flexible, it does not guarantee invertibility or topology preservation. Diffeomorphic registration enforces ϕ to be a smooth, bijective mapping with a smooth inverse, ensuring anatomically consistent transformations without folding or tearing.

This is achieved by parameterizing ϕ as the flow of a velocity field v_t , governed by:

$$\frac{d\phi_t}{dt} = v_t(\phi_t), \quad \phi_0 = \text{id}, \quad \phi = \phi_1, \quad (3.3)$$

where ϕ_t is the transformation at time t , and id is the identity map. Under mild smoothness assumptions, the solution ϕ is guaranteed to be diffeomorphic, preserving anatomical topology while allowing flexible nonlinear deformations.

3.2 Neural Field Representation boosted Optimization Algorithm

3.2.1 Introduction

Image registration is a cornerstone of medical image analysis, enabling applications such as longitudinal disease monitoring, multi-modal data fusion, and shape analysis [63, 113]. The objective is to estimate a spatial transformation that aligns a moving image I_m with a fixed image I_f , thereby establishing dense anatomical correspondences. Among the various formulations, diffeomorphic registration has been especially influential, as it guarantees smoothness, invertibility, and topology preservation—properties that are critical for anatomically meaningful mappings.

Classical approaches addressed this problem by casting registration as pairwise optimization. Elastic-type models [10, 120], statistical parametric mapping [7], and free-form B-splines [117] directly optimized displacement fields, while diffeomorphic methods such as LDDMM [15, 147] and SyN [8] enforced topological validity through velocity-based formulations. Although highly accurate, these methods were computationally intensive and sensitive to initialization, limiting their scalability in large cohorts or time-critical clinical settings [126].

The advent of deep learning shifted registration toward learning-based paradigms, where models trained across datasets predict deformation fields directly. Pioneering work such as VoxelMorph [11] demonstrated that convolutional networks can efficiently regress deformations, and subsequent studies introduced multi-resolution refinement [104], symmetric warping [103], cascaded frameworks [150], and transformer-based architectures [148, 122]. While effective, these models are constrained by architectural design and training distribution, and often struggle to fully capture the complexity of diffeomorphic deformations, particularly in out-of-distribution anatomies.

In parallel, implicit neural representations (INRs), or neural fields, have emerged as a powerful tool for modeling continuous signals. By parameterizing functions directly over spatial coordinates, they provide compact and expressive representations of geometry, scenes, and dynamics [109, 124, 107]. Recent work has extended INRs to medical imaging tasks such as reconstruction [141] and deformation modeling [140, 142], highlighting their ability to capture fine-scale anatomical detail. These advances naturally motivate the question: can neural fields be leveraged to represent diffeomorphic registration maps in a continuous, data-efficient manner?

To this end, we introduce the **Diffeomorphic Neural Velocity Field (DNVF)**, a framework that formulates diffeomorphic registration as the optimization of an implicit neural representation. Specifically, we parameterize a continuous velocity field with a multilayer perceptron (MLP) using sinusoidal activations [124], mapping each 3D coordinate (x, y, z) to

a velocity vector (v_x, v_y, v_z) . The final diffeomorphic transformation is obtained via scaling-and-squaring integration [5], ensuring smooth and invertible mappings. Unlike traditional optimization, DNVF places no restrictive assumptions on deformation dynamics, while benefiting from the flexibility and differentiability of neural fields.

To balance accuracy and efficiency, we further propose a cascaded extension, **Cas-DNVF**. A convolutional network first predicts a coarse deformation, narrowing the search space and enabling rapid inference. DNVF then refines this initialization through pair-specific optimization of a residual neural velocity field, recovering fine-scale deformations without sacrificing computational practicality.

In summary, this work makes three main contributions: (1) a neural field formulation of diffeomorphic registration, (2) a cascaded framework that integrates data-driven initialization with per-case refinement, and (3) extensive evaluation on 3D brain MR datasets demonstrating state-of-the-art accuracy while preserving the theoretical guarantees of diffeomorphisms.

3.2.2 Method

Let I_f and I_m denote the fixed and moving images to be aligned. We consider 3D image registration, where $I_f, I_m : \Omega \subset \mathbb{R}^3 \rightarrow \mathbb{R}$. After affine preprocessing, only non-linear displacements remain to be modeled.

Figure 3.1 presents an overview of our approach. DNVF is an optimization-based model that employs an MLP to represent the neural velocity field v_θ , where θ are network parameters. Unlike conventional registration methods, DNVF takes 3D spatial coordinates as input rather than image intensities. For each spatial location $\mathbf{p} \in \Omega$, v_θ outputs the velocity vector $\mathbf{v} = v_\theta(\mathbf{p})$. The diffeomorphic deformation ϕ_θ is then obtained by integrating v_θ as described in Sec. 3.1.2, with scaling-and-squaring used for efficient computation (Sec. 3.2.2.2). The

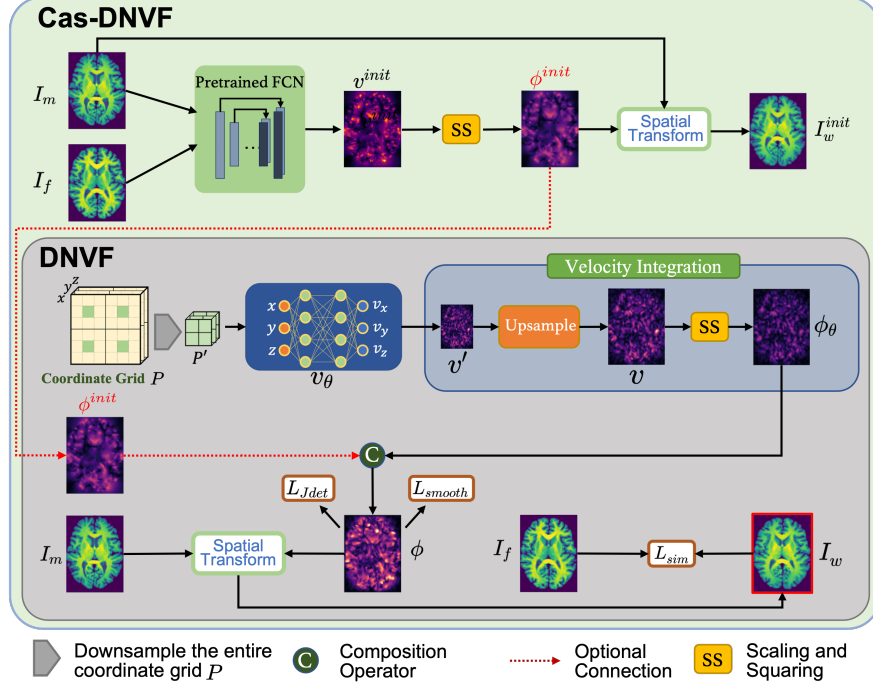


Figure 3.1: **Overall framework of the proposed DNVF and Cas-DNVF.** I_m , I_f , and I_w denote the moving image, fixed image, and warped image, respectively. DNVF models the continuous velocity field v_θ using neural representations, where a 3D position $\mathbf{p} \in \mathbb{R}^3$ is mapped to a velocity vector $\mathbf{v} \in \mathbb{R}^3$ through an MLP with sinusoidal activation functions. The velocity integration is performed via scaling and squaring, as detailed in Sec. 3.2.2.2, implicitly capturing dense correspondences between two input images. Cas-DNVF integrates learning-based initialization with DNVF: an FCN is pretrained to predict an initial deformation, narrowing the search space so that DNVF can focus on refining local deformations with high accuracy and efficiency. While designed for 3D registration, we illustrate the framework using 2D images for simplicity.

optimal parameters $\hat{\theta}$ are estimated by minimizing:

$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi_\theta) + \mathcal{L}_{\text{reg}}(\phi_\theta). \quad (3.4)$$

To combine the strengths of learning- and optimization-based paradigms, we propose a cascaded framework, Cas-DNVF. First, a fully convolutional network (FCN) is trained to predict an initial deformation $g_\beta(I_f, I_m) = \phi^{\text{init}}$. Its parameters β are optimized over dataset

$\mathcal{D}$ via:

$$\begin{aligned} \hat{\beta} = \arg \min_{\beta} \mathbb{E}_{(I_f, I_m) \sim \mathcal{D}} [& \mathcal{L}_{\text{sim}}(I_f, I_m \circ g_{\beta}(I_f, I_m)) \\ & + \mathcal{L}_{\text{reg}}(g_{\beta}(I_f, I_m))]. \end{aligned} \quad (3.5)$$

During inference, the pretrained g_{β} provides ϕ^{init} , which DNVF further refines by optimizing a residual deformation ϕ_{θ}^{res} . The overall deformation is computed by composition:

$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi^{init} \circ \phi_{\theta}^{res}) + \mathcal{L}_{\text{reg}}(\phi^{init} \circ \phi_{\theta}^{res}). \quad (3.6)$$

As demonstrated in Sec. 3.2.3, this cascaded design accelerates convergence, mitigates the limited generalizability of purely learning-based models, and improves dense correspondence estimation.

3.2.2.1 Neural Velocity Field Representation

Inspired by neural rendering methods [124, 101], we represent the continuous velocity field v_{θ} using an MLP. Each input coordinate $\mathbf{p}$ is mapped to a velocity vector $\mathbf{v} = v_{\theta}(\mathbf{p})$.

The complex anatomy of 3D brain MR scans requires deformation fields with high-frequency detail. Standard MLPs, however, are limited by spectral bias [130, 111, 13], which favors low-frequency representations. To address this, we employ sinusoidal activation functions with SIREN initialization [124], enabling effective modeling of high-frequency variations. The neural velocity field v_{θ} is thus defined as:

$$\begin{aligned} f_0(\mathbf{x}_0) &= \mathbf{W}_0 \mathbf{x}_0 + \mathbf{b}_0 \mapsto x_1, \\ f_i(\mathbf{x}_i) &= \mathbf{W}_i \sin(\mathbf{x}_i) + \mathbf{b}_i \mapsto x_{i+1}, \end{aligned} \quad (3.7)$$

where $\mathbf{p}$ is first mapped to a high-dimensional embedding $f_0 : \mathbb{R}^3 \mapsto \mathbb{R}^N$, and subsequent lay-

ers f_i serve as learnable Fourier mappings [16]. This enriches the neural tangent kernel [130], enabling v_θ to capture high-frequency deformations. In practice, we use a 5-layer MLP with 512 hidden units.

3.2.2.2 Diffeomorphic Deformation as Integration

Given v_θ , the deformation ϕ_θ is obtained by integration (Eq. 2). Following [35, 103, 104, 148], we assume a stationary velocity field and compute ϕ_θ using scaling and squaring [5].

Scaling and Squaring. The exponential map $\phi^{(t)} = \exp(v_\theta)$ defines a one-parameter subgroup of diffeomorphisms. To efficiently approximate $\phi^{(1)}$, we initialize with $\phi^{(1/2^T)} = \mathbf{p} + v_\theta(\mathbf{p})/2^T$, where T is the total time step, and recursively compute:

$$\phi^{(1/2^{t-1})} = \phi^{(1/2^t)} \circ \phi^{(1/2^t)}.$$

The final deformation is $\phi^{(1)} = \phi^{(1/2)} \circ \phi^{(1/2)}$. These recursive compositions, implemented with spatial transform layers and linear interpolation, are differentiable and introduce no extra learnable parameters. We set $T = 7$.

To manage memory cost, we downsample the coordinate grid by a factor of 3, compute the velocity field at reduced resolution, and upsample before integration (Fig. 3.1).

3.2.2.3 Cascaded Registration

Initial Deformation. In Cas-DNVF, the initial mapping $g_\beta(I_f, I_m) = \phi^{init}$ is parameterized by an FCN, followed by scaling-and-squaring and a spatial transform layer. The FCN adopts a U-Net-like structure (Fig. 3.2) that takes concatenated fixed and moving images as input and outputs a velocity field. Training maximizes similarity between warped and fixed images.

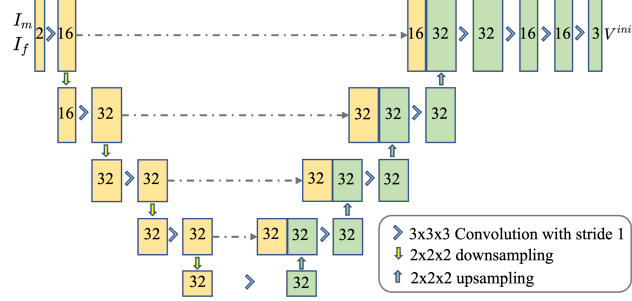


Figure 3.2: Structure of the fully convolutional network (FCN) in Cas-DNVF. Input is the concatenation of moving and fixed images; output is the velocity field.

Alternative models [35, 103, 104, 148] can also serve as g_β . Ablations in Sec. 3.2.3.5 confirm that DNVF consistently improves their performance.

Residual Deformation. After ϕ^{init} is predicted, DNVF optimizes a residual ϕ_θ^{res} (Eq. 5). The combined deformation $\phi = \phi^{init} \circ \phi_\theta^{res}$ is computed via a spatial transform layer. Since ϕ^{init} accounts for large displacements, DNVF specializes in refining local details. To stabilize optimization, we rescale the DNVF velocity vectors by 0.1 before integration.

3.2.2.4 Optimization

The loss function consists of image similarity $\mathcal{L}_{sim}$ and regularization $\mathcal{L}_{reg}$ terms.

Similarity Loss. Local normalized cross-correlation (NCC) measures similarity between warped and fixed images:

$$NCC(I_w, I_f) = \sum_{\mathbf{p} \in \Omega} \frac{\left(\sum_{\mathbf{p}_i} (I_w(\mathbf{p}_i) - \bar{I}_w(\mathbf{p})) (I_f(\mathbf{p}_i) - \bar{I}_f(\mathbf{p})) \right)^2}{\sum_{\mathbf{p}_i} (I_w(\mathbf{p}_i) - \bar{I}_w(\mathbf{p}))^2 \sum_{\mathbf{p}_i} (I_f(\mathbf{p}_i) - \bar{I}_f(\mathbf{p}))^2}. \quad (3.8)$$

We minimize negative NCC with a 9^3 window: $\mathcal{L}_{sim} = -NCC(I_m \circ \phi, I_f)$.

Jacobian Regularization. Following [103], we enforce orientation preservation by penalizing negative Jacobian determinants:

$$\mathcal{L}_{Jdet} = \frac{1}{N} \sum_{\mathbf{p} \in \Omega} \text{relu}(-|J_\phi(\mathbf{p})|), \quad (3.9)$$

where

$$J_\phi(\mathbf{p}) = \begin{bmatrix} \frac{\partial \phi_x(\mathbf{p})}{\partial x} & \frac{\partial \phi_x(\mathbf{p})}{\partial y} & \frac{\partial \phi_x(\mathbf{p})}{\partial z} \\ \frac{\partial \phi_y(\mathbf{p})}{\partial x} & \frac{\partial \phi_y(\mathbf{p})}{\partial y} & \frac{\partial \phi_y(\mathbf{p})}{\partial z} \\ \frac{\partial \phi_z(\mathbf{p})}{\partial x} & \frac{\partial \phi_z(\mathbf{p})}{\partial y} & \frac{\partial \phi_z(\mathbf{p})}{\partial z} \end{bmatrix}. \quad (3.10)$$

Smoothness Regularization. To suppress abrupt local distortions, we constrain spatial gradients:

$$\mathcal{L}_{smooth} = \sum_{\mathbf{p} \in \Omega} \|\nabla \phi(\mathbf{p})\|^2. \quad (3.11)$$

Final Objective. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{sim} + \lambda_1 \mathcal{L}_{Jdet} + \lambda_2 \mathcal{L}_{smooth}, \quad (3.12)$$

where λ_1 and λ_2 control the contribution of regularization terms.

3.2.3 Experiment

3.2.3.1 Datasets and Preprocessing

We evaluate DNVF on two public 3D brain MR datasets: **OASIS** [95] and **Mindboggle101** [73]. The OASIS dataset includes 416 T1-weighted MR scans from subjects aged 18–96, of

which 100 are diagnosed with mild to moderate Alzheimer’s disease. Subcortical segmentation maps of 35 anatomical structures are available as ground truth. Mindboggle101 contains 101 T1-weighted MR scans collected from five cohorts (e.g., HLN-12, MMRR-21, NKI-RS). Following [144], scans with erroneous labels are excluded, and performance is evaluated on 31 cortical regions.

All scans undergo standard preprocessing: skull stripping, resampling to isotropic (1 mm^3) resolution, affine alignment to the MNI-152 template, cropping to $(162 \times 192 \times 144)$, and intensity normalization by each volume’s maximum.

3.2.3.2 Experimental Setup

We benchmark DNVF against optimization-based methods (SyN [8], NiftyReg [102]), learning-based methods (VoxelMorph [11], SYMNet [103]), and recent neural-field approaches (IDIR [140], NODEO [142]). The latter two are included because they were proposed contemporaneously with this work and similarly rely on neural implicit representations.

For each dataset, 20 scans are randomly sampled as moving images and 3 scans as atlases, yielding 60 test pairs. All anatomical labels are used during evaluation. To ensure fairness, instance-specific optimization is applied to learning-based methods.

Hyperparameters follow standard practice. SyN uses gradient step size 0.25 and Gaussian smoothing parameters $[0.9, 0.2]$. Both SyN and NiftyReg employ cross-correlation as the similarity metric. Learning-based methods are trained with 86 scans (Mindboggle) and 250 scans (OASIS).

For DNVF, regularization weights are set to $\lambda_1 = 100$ (orientation consistency) and $\lambda_2 = 0.1$ (smoothness). Cas-DNVF uses $\lambda_1 = 10$ and $\lambda_2 = 4$. The FCN is optimized with Adam at a learning rate of $1e^{-4}$. All experiments are implemented in PyTorch and run on an RTX

2080 Ti GPU with an Intel i7-7700K CPU.

3.2.3.3 Evaluation Metrics

Performance is assessed using three complementary criteria:

- **Dice Similarity Coefficient (DSC)** — measures volumetric overlap between warped and fixed segmentations.
- **Negative Jacobian Determinant ratio ($|J_{<0}|$)** — quantifies topology violations, reflecting deformation invertibility.
- **Structural Similarity Index (SSIM)** [138] — evaluates local structural fidelity between warped and fixed intensities.

3.2.3.4 Results and Discussion

Accuracy. Table 3.1 summarizes the results. DNVF consistently outperforms both traditional (SyN, NiftyReg) and learning-based methods (VoxelMorph, SYMNet) in terms of DSC and SSIM, while maintaining a low $|J_{<0}|$ ratio. These results highlight the advantage of representing deformations via neural velocity fields. Cas-DNVF achieves further gains by initializing from a pretrained FCN, though with a marginally higher $|J_{<0}|$ due to residual–initial deformation composition.

Visualization. Figure 3.3 shows qualitative comparisons. While existing methods achieve coarse alignment, they often miss subtle local deformations. By contrast, DNVF and Cas-DNVF recover high-frequency structural details, particularly in cortical folds and subcortical boundaries.

Efficiency. Figure 3.4 demonstrates the accuracy–efficiency trade-off. DNVF reaches strong

Category	Dataset	Mindboggle [73]			OASIS [95]		
	Model/Metric	DSC($\uparrow$)	$ J_{<0} (\downarrow)$	SSIM($\uparrow$)	DSC($\uparrow$)	$ J_{<0} (\downarrow)$	SSIM($\uparrow$)
Traditional Methods	SyN [8]	0.548	$\approx \mathbf{0}$	0.9097	0.777	$\approx \mathbf{0}$	0.9149
	NiftyReg [102]	0.509	$\approx \mathbf{0}$	0.8650	0.773	$\approx \mathbf{0}$	0.8916
Learning-based Methods	VoxelMorph [11]	0.555	0.076%	0.9145	0.763	0.072%	0.9082
	SYMNet [103]	0.567	0.0023%	0.9191	0.777	0.0022%	0.9247
	FCN	0.548	$\approx \mathbf{0}$	0.9017	0.765	$\approx \mathbf{0}$	0.9091
Instance-specific	VoxelMorph+OPT	0.558	0.072%	0.9186	0.784	0.051%	0.9141
Optimization of	SYMNet+OPT	0.575	0.0010%	0.9271	0.791	0.0012%	0.9292
Learning-based Methods	FCN+OPT	0.556	$\approx \mathbf{0}$	0.9072	0.781	$\approx \mathbf{0}$	0.9137
Our Methods	DNVF	0.606	0.0013%	0.9496	0.794	0.0015%	0.9512
	Cas-DNVF	0.612	0.0031%	0.9531	0.815	0.0036%	0.9532

Table 3.1: The registration methods are evaluated using anatomical structures (**31** for Mindboggle and **35** for OASIS) in atlas-based fashion. FCN is the model used in C2F-DNVF which provides better topology preservation compared with other learning-based methods. Table 3.5 shows the results of replacing FCN with other SOTA methods in C2F-DNVF. For instance-specific optimization of learning-based methods, we finetune the pretrained model for each given pair of test images.

accuracy within practical runtimes: 48s (80 iterations) on Mindboggle and 72s (120 iterations) on OASIS. Cas-DNVF further accelerates convergence by narrowing the optimization search space through FCN initialization. Both methods achieve near-SOTA performance in under three minutes.

Dataset	OASIS		CANDI [71]		
	DSC ₂₈ ($\uparrow$)	$ J_{<0} (\downarrow)$	DSC ₃₂ ($\uparrow$)	DSC ₂₈ ($\uparrow$)	$ J_{<0} (\downarrow)$
IDIR	0.794	0.124%	0.774	0.811	0.113%
NODEO	0.779	0.03%	0.760	0.802	$1.8 \times 10^{-7}\%$
DNVF	0.815	0.0004%	0.786	0.823	0.0003%

Table 3.2: Following NODEO, we use the same data split and evaluation metric. The DSC for OASIS is averaged over **28** structures excluding the very small ones and the DSC for CANDI is averaged over all **32** structures and **28** large structures as stated in NODEO’s setting. We limit the optimization of DNVF to 100 iteration (60s) compared with NODEO (80s)

Comparison with Neural-Field Methods. Table 3.2 compares DNVF against IDIR

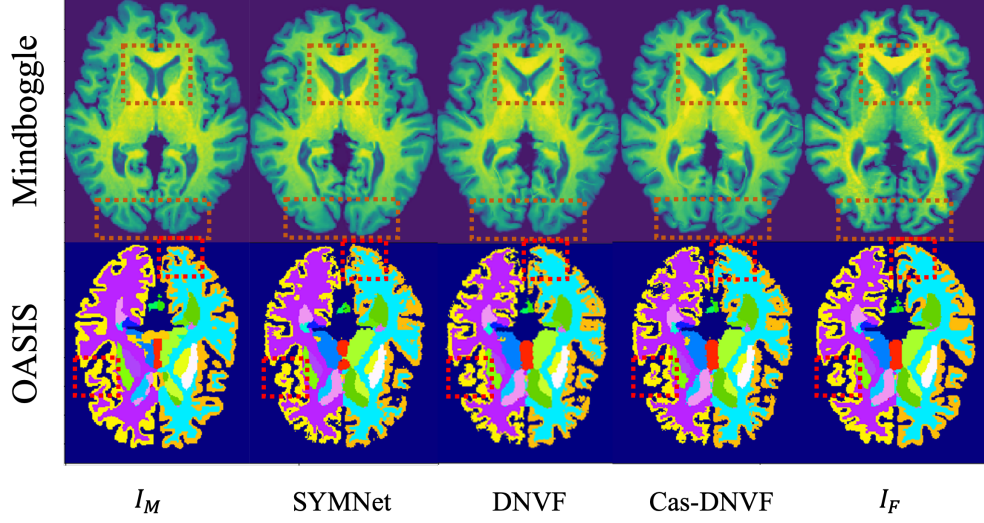


Figure 3.3: Qualitative registration results. Top: warped Mindboggle scans. Bottom: OASIS anatomical segmentations. DNVF and Cas-DNVF capture fine-scale cortical and subcortical deformations missed by other methods.

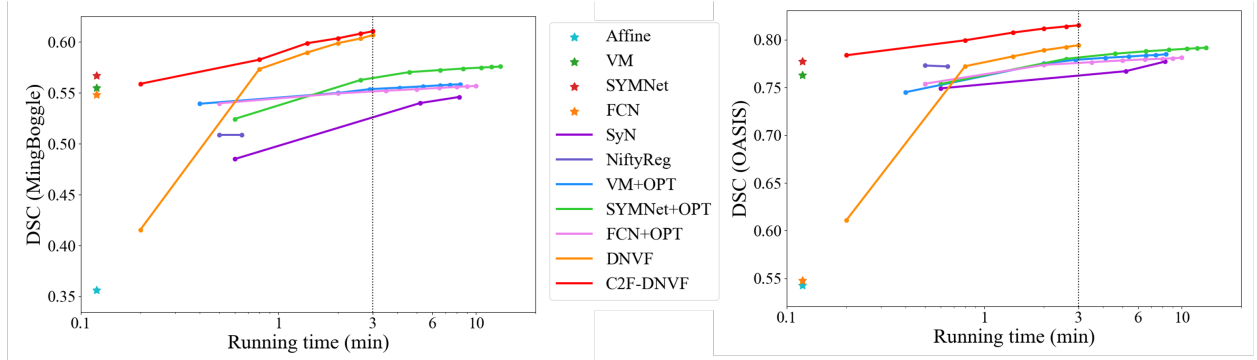


Figure 3.4: Accuracy–efficiency trade-off: DSC versus runtime across different iteration budgets.

and NODEO. IDIR directly models displacement fields and lacks diffeomorphic guarantees, leading to poorer topology preservation. NODEO incorporates convolutional layers and Neural ODE solvers, but suffers from higher runtime and reduced accuracy. DNVF achieves the best balance: highest DSC, lowest $|J_{<0}|$, and faster convergence, owing to its sinusoidal-MLP velocity field representation and efficient scaling-and-squaring integration.

3.2.3.5 Ablation Studies

Activation functions. Table 3.3 compares sinusoidal activations with ReLU (with/without positional encoding). Periodic activations provide superior capacity for capturing high-frequency deformations in unsupervised registration.

Dataset	DNVF	MLP+ReLU	MLP+ReLU+PE
Mindboggle	0.606	0.440	0.397
OASIS	0.794	0.708	0.683

Table 3.3: Results using sinusoidal and ReLU activation functions.

Velocity field representation. Table 3.4 contrasts three representations: (I) volumetric grid with TV regularization [45], (II) 3D CNN-based predictor, and (III) sinusoidal-MLP mapping (ours). The proposed neural field achieves the highest DSC, confirming its expressive yet efficient representation.

Dataset	DNVF	Grid	Conv+Sin	Conv+ReLU
Mindboggle	0.606	0.572	0.515	0.480
OASIS	0.794	0.755	0.739	0.731

Table 3.4: Results with different velocity field representations.

Choice of pretrained FCN for Cas-DNVF. Table 3.5 evaluates Cas-DNVF initialized with VoxelMorph or SYMNet. In both cases, Cas-DNVF improves over the standalone FCN, demonstrating that the cascaded framework generalizes across different initial predictors.

Dataset	FCN+DNVF	VM+DNVF	SYMNet+DNVF
Mindboggle	0.612	0.609	0.606
OASIS	0.815	0.816	0.812

Table 3.5: Results with different pretrained FCN in Cas-DNVF.

3.2.4 Conclusion

In this work, we introduced a neural field representation for continuous velocity fields, enabling diffeomorphic deformation modeling for medical image registration. The proposed DNVF and its cascaded variant Cas-DNVF establish a flexible framework that bridges optimization-based and learning-based approaches. Validation on multiple brain MR datasets demonstrated clear advantages in registration accuracy, efficiency, and topology preservation, highlighting the effectiveness of neural velocity fields combined with periodic activations.

Despite these benefits, our methods have certain limitations. First, the scaling-and-squaring integration introduces a relatively large memory footprint, suggesting that more efficient velocity-to-deformation formulations could be explored. Second, Cas-DNVF employs a two-stage design where the FCN initialization and neural velocity field optimization are decoupled. A natural extension is to integrate these components into a unified, end-to-end trainable framework.

Overall, DNVF provides a new perspective on diffeomorphic registration by leveraging neural fields, and we believe further development along these directions will continue to advance robust and efficient medical image registration.

3.3 Learning Registration with Pseudo Groundtruth Generation

3.3.1 Introduction

Medical image registration is a cornerstone of medical image analysis, supporting applications such as atlas-based segmentation, disease progression modeling, and multi-modal

fusion. The task involves estimating a transformation, represented as a deformation field, that aligns a moving image to a fixed image while preserving anatomical fidelity. Classical optimization-based methods [8, 10, 120] iteratively refine the deformation field under smoothness or topology-preserving constraints. More recently, deep learning has transformed registration research [12, 35, 104, 103, 25, 24], enabling models to directly predict deformation fields from paired images with remarkable efficiency at inference time.

Despite these advances, a persistent limitation lies in how registration networks are trained. Existing frameworks typically adopt either weakly-supervised or unsupervised strategies. Weakly-supervised approaches incorporate segmentation labels to guide alignment [12], often yielding higher accuracy but at the cost of expensive, large-scale annotations. In contrast, unsupervised methods [38, 121, 148, 25, 24] depend solely on image similarity measures such as normalized cross-correlation or mutual information, avoiding labels but struggling in cases with large or complex deformations. This trade-off raises a natural question: *Can we train registration networks without manual labels while still providing strong supervision signals for high accuracy?*

We address this challenge with a unified framework, termed **On-the-Fly Guidance (OFG)**, that dynamically generates pseudo-ground truth during training. The key idea is to couple direct prediction with lightweight instance-specific optimization in a hybrid scheme. For each moving-fixed image pair, the network first predicts a deformation field ϕ_{pre} , which is then refined by a limited-step optimizer to produce ϕ_{opt} . Crucially, ϕ_{opt} serves as a pseudo-label to supervise ϕ_{pre} via a mean squared error loss. This mechanism supplies explicit supervision without requiring precomputed ground truth. By constraining the optimizer to only a small number of iterations (e.g., 10), OFG provides incremental yet effective guidance that prevents poor local minima and steadily improves the model over training, all while keeping inference as efficient as standard prediction-only networks.

Our work builds upon two active research threads. First, learning-based methods such

as VoxelMorph [12], ViT-V-Net [25], and TransMorph [24] have shown the power of end-to-end networks in modeling complex deformations, sometimes with auxiliary segmentation to improve performance. Yet, these approaches remain constrained by the supervised-unsupervised trade-off. Second, recent *instance-specific optimization* strategies have shown promise across registration [18, 105], segmentation [151], and human pose estimation [74]. For example, Neural Instance Optimization (NIO) [105] adapts deformation fields by correcting distribution shifts, while cyclical self-training strategies [18] use pseudo labels across training phases. However, such methods typically decouple optimization from prediction, either relying on hand-crafted priors or transferring supervision only between stages, limiting adaptability. By contrast, OFG integrates prediction and optimization *within each training iteration*, establishing a continuous self-improving loop.

Contributions. Our main contributions are summarized as follows:

- We propose OFG, a unified training framework that couples direct prediction with lightweight instance-specific optimization, providing on-the-fly pseudo-supervision without requiring manual annotations.
- We show that the optimized deformation fields generated during training act as effective pseudo-ground truth, creating a closed feedback loop that steadily improves the base registration network.
- We validate OFG across multiple architectures and datasets, demonstrating consistent accuracy gains over state-of-the-art baselines while preserving the fast inference efficiency of learning-based methods.

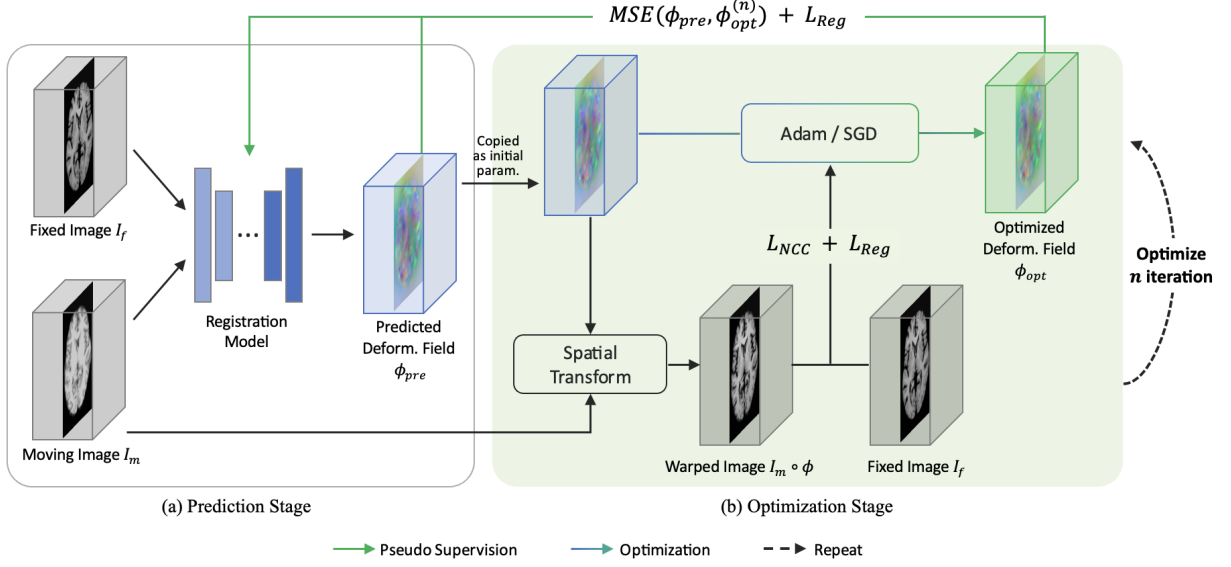


Figure 3.5: **Overview of the proposed OFG framework.** It consists of a prediction stage (any registration model predicting a deformation field) and an optimization stage. During training, the optimizer refines the predicted deformation field for a small number of iterations, and the refined field is used as pseudo ground truth to supervise the registration model.

3.3.2 Method

We propose *On-the-Fly Guidance* (OFG), a two-stage training framework that unifies direct prediction and instance-specific optimization for medical image registration. Unlike conventional learning-based methods that rely solely on image similarity losses or offline pseudo labels, OFG dynamically generates pseudo ground truth during training, providing stronger supervision without requiring manual annotations. In this section, we describe the overall architecture, the strategy for generating pseudo ground truth, the principle of on-the-fly guidance, the design of the online optimizer, and the implementation details.

3.3.2.1 Overall Structure

The OFG framework consists of two tightly coupled stages (Fig. 3.5).

Prediction stage. A learning-based registration network F_θ takes a moving image I_m and a fixed image I_f and predicts a dense deformation field ϕ_{pre} :

$$F_\theta(I_f, I_m) = \phi_{pre}, \quad (3.13)$$

where θ denotes the network parameters. The prediction stage is model-agnostic, meaning that OFG can be applied to different architectures. In our experiments, we use VoxelMorph [12], ViT-V-Net [25], and TransMorph [24].

Optimization stage. The optimizer refines ϕ_{pre} for a small number of steps (10 by default) to produce an improved deformation ϕ_{opt} . This refined field serves as pseudo ground truth to supervise the registration network, establishing a feedback loop where prediction informs optimization and optimization strengthens prediction. Importantly, the optimizer is used only during training and introduces no overhead at inference.

3.3.2.2 Training with Pseudo Ground Truth

In registration, ground truth deformations are unavailable, making supervision difficult. A natural idea is to generate pseudo labels. Two naive strategies exist: (1) using predictions from a pre-trained model, which is simple but limited by the upper bound of the teacher model; (2) fully optimizing predicted deformations until convergence, which may yield high-quality labels but often diverges from the model’s current predictions, leading to unstable training.

OFG resolves these limitations by generating pseudo labels dynamically during training. Instead of relying on fixed pseudo ground truth, we refine the current prediction ϕ_{pre} into ϕ_{opt} through lightweight optimization. This yields pseudo labels that are (i) closer to the current model state, (ii) computationally efficient, and (iii) continuously improving as the

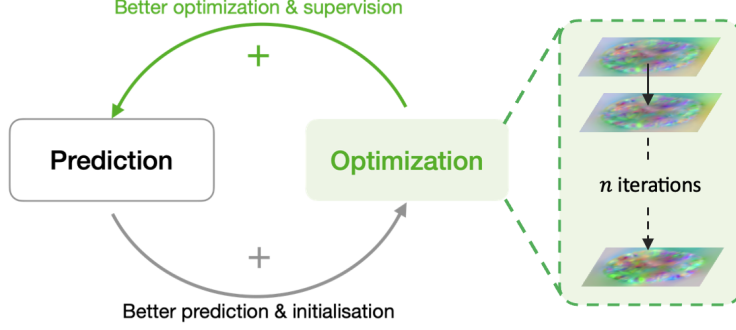


Figure 3.6: **Illustration of on-the-fly guidance.** At each iteration, the optimizer refines the predicted deformation field into a stronger pseudo label. Better predictions lead to better optimizations, which in turn provide stronger supervision, forming a positive feedback loop.

model itself improves.

3.3.2.3 On-the-Fly Guidance

The key insight of OFG is to align supervision targets with the current learning state. At each training step, ϕ_{pre} is refined by a few iterations of the optimizer to obtain ϕ_{opt} , which then supervises ϕ_{pre} via mean squared error loss. This incremental supervision provides an attainable target for the network and avoids poor local minima.

Moreover, OFG establishes a *self-improving loop* (Fig. 3.6). As the registration network improves, it produces more accurate ϕ_{pre} , which enables the optimizer to generate better ϕ_{opt} . The improved pseudo labels, in turn, guide the model toward even higher performance.

3.3.2.4 Online Optimizer

The online optimizer is designed to be efficient and lightweight. It initializes with ϕ_{pre} and treats it as optimizable parameters. Internally, it contains a Spatial Transformer Network (STN) [65], which warps the moving image using the current deformation. The optimizer

updates ϕ by minimizing an energy function through gradient descent:

$$\phi_{opt}^{(n+1)} = \phi_{opt}^{(n)} - \eta \nabla E_{opt}, \quad (3.14)$$

where n is the iteration index, η is the learning rate, and E_{opt} is defined below. Notably, no additional parameters are introduced—the optimizer directly updates ϕ . Since only a few iterations are performed per training step, the added computational cost is modest.

3.3.2.5 Implementation Details

Training loss. The registration network is trained to match its prediction ϕ_{pre} to the refined pseudo label ϕ_{opt} while encouraging smoothness. The overall training loss is:

$$L_{all} = \frac{1}{n} \sum (\phi_{pre} - \phi_{opt})^2 + \lambda L_{reg}, \quad (3.15)$$

where the first term is an MSE loss and L_{reg} is a regularization term defined as

$$L_{reg}(\phi) = \sum_{p \in \Omega} \|\nabla \phi(p)\|^2. \quad (3.16)$$

We set $\lambda = 0.02$ to balance deformation fidelity and smoothness.

Optimizer energy function. The optimizer minimizes an energy function combining image similarity and deformation smoothness:

$$E_{opt}(I_m, I_f, \phi) = LNCC(I_f, I_m \circ \phi) + L_{reg}(\phi), \quad (3.17)$$

where $\circ$ denotes warping. The similarity is computed by local normalized cross-correlation

(LNCC):

$$LNCC(I_f, I_w) = \sum_{\mathbf{p} \in \Omega} \frac{\left(\sum_{\mathbf{p}_i} (I_w(\mathbf{p}_i) - \bar{I}_w(\mathbf{p})) (I_f(\mathbf{p}_i) - \bar{I}_f(\mathbf{p})) \right)^2}{\sum_{\mathbf{p}_i} (I_w(\mathbf{p}_i) - \bar{I}_w(\mathbf{p}))^2 \sum_{\mathbf{p}_i} (I_f(\mathbf{p}_i) - \bar{I}_f(\mathbf{p}))^2} \quad (3.18)$$

where $\bar{I}_f(\mathbf{p})$ and $\bar{I}_w(\mathbf{p})$ denote local mean intensities in an n^3 window centered at voxel $\mathbf{p}$.

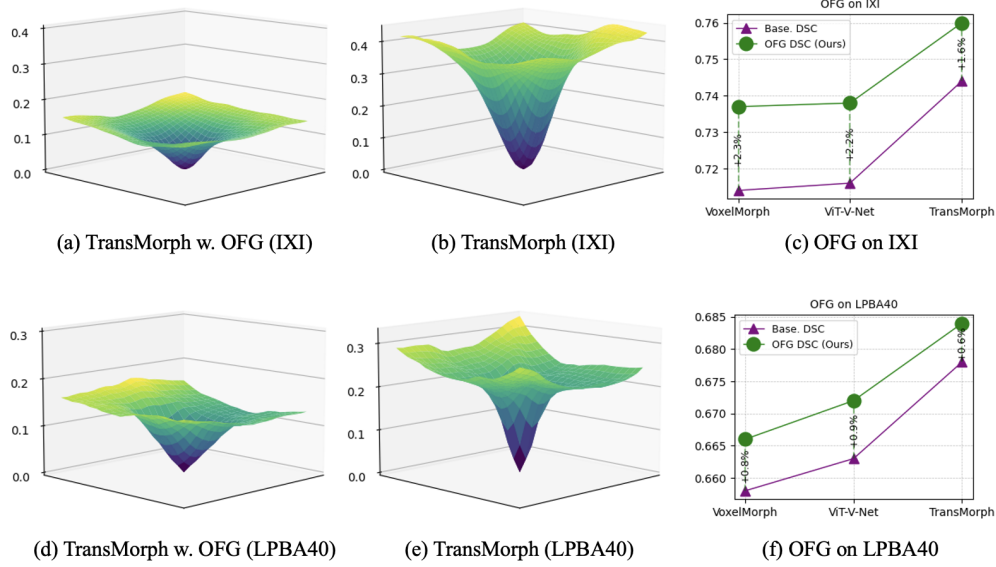


Figure 3.7: Loss landscape comparison between TransMorph with and without OFG across two datasets. OFG produces flatter and smoother optimization surfaces, improving trainability and convergence.

3.3.3 Experiments

3.3.3.1 Experimental Setup

Datasets and Preprocessing. We evaluated OFG on three public brain MRI datasets: IXI [88], OASIS [95], and LPBA40 [133]. Standard preprocessing was performed with FreeSurfer [9], including skull stripping, resampling, and affine alignment. The processed volumes have a dimension of $160 \times 192 \times 224$ for IXI and OASIS, and $160 \times 192 \times 160$ for LPBA40. Specifically, IXI was split into 200 training and 20 validation volumes; OASIS into 200 training and 19

Datasets	Methods	Base. DSC $\uparrow$	OFG DSC $\uparrow$	Base. $\% J_\phi < 0 \downarrow$	OFG $\% J_\phi < 0 \downarrow$	Infer. Time (s)
IXI [88]	SyN [8]	0.647	N/A	1.96e-6	N/A	277(CPU)
	NiftyReg [21]	0.585	N/A	0.029	N/A	22.4(CPU)
	VoxelMorph [12]	0.714	0.737 (+2.3%)	1.398	0.516 (-0.882)	0.061(GPU)
	ViT-V-Net [25]	0.716	0.738 (+2.2%)	1.543	0.545 (-0.998)	0.615(GPU)
	TransMorph [24]	0.744	0.760 (+1.6%)	1.433	0.794 (-0.639)	0.114(GPU)
OASIS [95]	SyN [8]	0.769	N/A	1.58e-4	N/A	258(CPU)
	NiftyReg [21]	0.762	N/A	0.011	N/A	25.0(CPU)
	VoxelMorph [12]	0.788	0.794 (+0.6%)	0.911	0.490 (-0.421)	0.061(GPU)
	ViT-V-Net [25]	0.794	0.809 (+1.5%)	0.887	0.487 (-0.400)	0.646(GPU)
	TransMorph [24]	0.818	0.818 (=)	0.765	0.517 (-0.248)	0.160(GPU)
LPBA40 [133]	SyN [8]	0.703	N/A	1.18e-4	N/A	172(CPU)
	NiftyReg [21]	0.691	N/A	1.13e-3	N/A	22.8(CPU)
	VoxelMorph [12]	0.658	0.666 (+0.8%)	0.288	0.023 (-0.205)	0.046(GPU)
	ViT-V-Net [25]	0.663	0.672 (+0.9%)	0.390	0.112 (-0.278)	0.446(GPU)
	TransMorph [24]	0.678	0.684 (+0.6%)	0.438	0.150 (-0.288)	0.327(GPU)

Table 3.6: Evaluation results for different methods on various datasets. The OFG architecture provides significant and substantial improvement on the unsupervised learning-based methods. These results validate OFG’s effectiveness and generalizability without compromising on inference efficiency.

validation volumes; and LPBA40 into 30 training, 9 validation, and 1 atlas volume.

Evaluation Metrics. Performance was quantified using two complementary measures. (1) *Dice similarity coefficient (DSC)* [12, 10] was computed on anatomical segmentations to assess registration accuracy. (2) *Jacobian regularity*, measured as the percentage of voxels with $|J_\phi(p)| < 0$, where $J_\phi(p) = \nabla\phi(p)$, captured violations of diffeomorphism [6]. Lower values indicate smoother and more physically plausible deformations.

Baseline Models. We compared OFG against two classical optimization-based methods and three state-of-the-art learning-based approaches:

- **SyN** [8]: trained with mean squared error (MSE) objective and three resolution levels with 160, 80, and 40 iterations.
- **NiftyReg** [21]: optimized using sum of squared differences (SSD) over three scales with 300 iterations each.

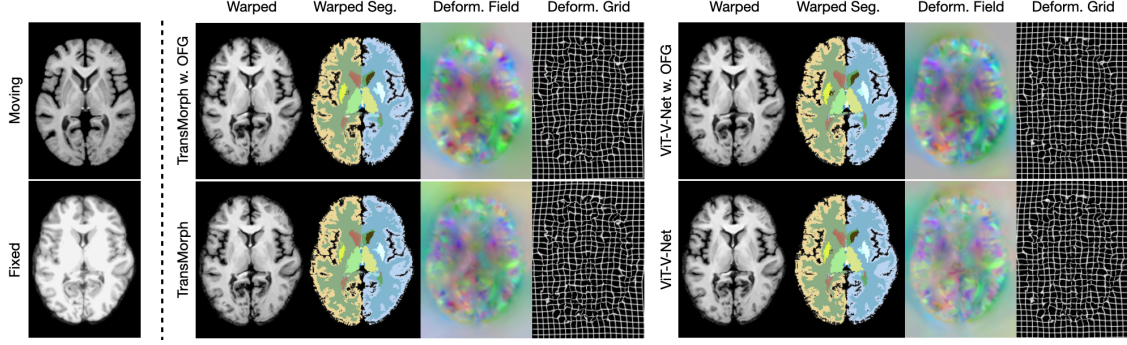


Figure 3.8: Qualitative registration results on IXI [88]. Each row shows comparison between baseline (row 2) and its OFG-enhanced counterpart (row 1) for TransMorph and ViT-V-Net. OFG produces sharper alignment while avoiding irregular deformations.

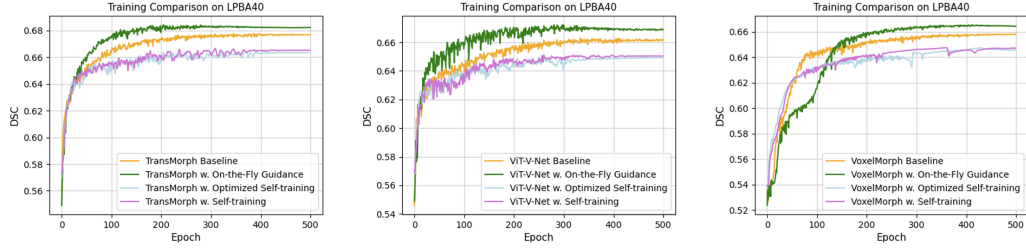


Figure 3.9: Training dynamics on LPBA40. We compare self-training (using fixed pre-trained deformations), optimized self-training (refining these fields offline), and OFG. OFG provides more stable convergence and higher validation DSC, particularly under limited data.

- **VoxelMorph** [12]: implemented following the VoxelMorph-1 configuration.
- **ViT-V-Net** [25]: applied with default hyperparameters.
- **TransMorph** [24]: applied with default hyperparameters.

Training Details. All models were trained on an NVIDIA RTX 4090 for 500 epochs using Adam [41], with a learning rate of 10^{-4} and a batch size of 1. The OFG optimizer employed an initial learning rate of 0.1 and ran for 10 iterations per training step.

Converg. Loss	IXI [88]	OASIS [95]	LPBA40 [133]
VoxelMorph [12]	-0.222	-0.259	-0.183
ViT-V-Net [25]	-0.242	-0.266	-0.215
TransMorph [24]	-0.262	-0.283	-0.232

Table 3.7: Comparison between the 3 models’ convergence losses on 3 datasets. Note that the loss function settings were the same. For TransMorph on OASIS, the convergence loss is clearly the lowest.

3.3.3.2 Image Registration Results

We extensively validated OFG on IXI, LPBA40, and OASIS. To demonstrate its plug-and-play nature, OFG was applied to three learning-based architectures: VoxelMorph, ViT-V-Net, and TransMorph.

IXI. As shown in Table 3.6, OFG significantly boosted DSC across all models. For TransMorph, OFG achieved a +1.6% DSC improvement and halved the proportion of voxels with $|J_\phi| < 0$. Similar improvements were observed on VoxelMorph (+2.3% DSC) and ViT-V-Net (+2.2% DSC). Figure 3.8 highlights qualitative improvements, where OFG suppresses over-sharpened or irregular deformations.

LPBA40. Due to its small sample size (40 volumes), LPBA40 presents a risk of overfitting. Figure 3.9 illustrates that baseline models tend to stagnate, whereas OFG’s additional pseudo supervision enables steady improvement. This highlights OFG’s utility in data-scarce regimes.

OASIS. While OFG consistently improved VoxelMorph and ViT-V-Net on OASIS, its gains on TransMorph were negligible (Table 3.6). We hypothesize that OASIS poses a relatively easier challenge for TransMorph. To validate this, we compared convergence losses across datasets (Table 3.7). Indeed, TransMorph on OASIS achieved the lowest loss, supporting our hypothesis.

	DSC	$\% J_\phi < 0$	Optimize Time (s)
VoxelMorph-1	0.691	0.456	1.174
VoxelMorph-2	0.690	0.383	2.314
VoxelMorph-5	0.693	0.385	5.856
<i>Ours</i>	0.731	0.503	4.697

Table 3.8: Comparison between instance-specific optimizer (OFG) and network-based optimizer on IXI dataset[88]. The outcomes presented are derived from the initial 200 epochs.

3.3.3.3 Training with Pseudo Ground Truth

To evaluate pseudo ground truth quality (Sec. 3.3.2.2), we analyzed the loss landscapes of TransMorph and TransMorph+OFG on IXI and LPBA40 following [80, 51, 62]. Using a pre-trained parameter set θ^* as the origin, we perturbed θ^* along two random directions and computed average validation loss. As shown in Fig. 3.7, OFG yields a smoother and flatter landscape, indicating better trainability, reduced risk of poor local minima, and improved generalization.

3.3.3.4 On-the-Fly Guidance Strategy

We next compared OFG’s incremental guidance with an alternative scheme: training from scratch using pseudo labels generated offline from a pre-trained model and refined with the OFG optimizer. As shown in Fig. 3.9, this strategy struggled with convergence, while OFG maintained stable learning. This underscores the advantage of aligning supervision with the model’s evolving state.

3.3.3.5 Instance-Specific Optimization

We further evaluated the benefit of instance-specific optimization (Sec. 3.3.2.4) against a network-based optimizer (details in Sec. 3.3.3.6). Results in Table 3.8 show that OFG’s direct

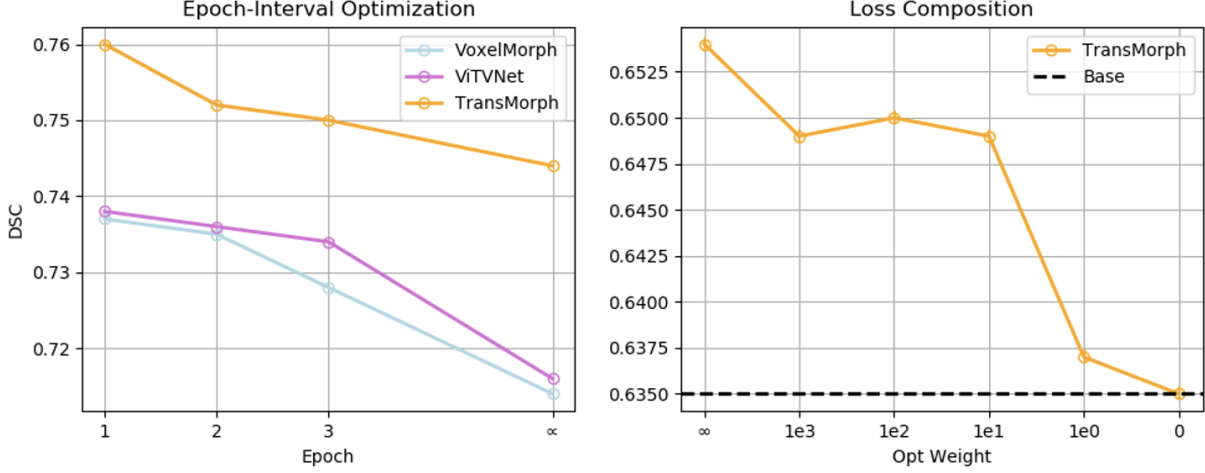


Figure 3.10: Ablation study on IXI (left) and LPBA40 (right). Increasing reliance on unsupervised losses degrades performance, highlighting the importance of supervision from OFG’s pseudo labels.

refinement of ϕ_{pre} achieves the highest DSC while maintaining low overhead, validating the design of our lightweight optimizer.

3.3.3.6 Ablation Studies

Optimizer Design. We tested several optimizer variants (Table 3.9).

- **Network-based optimizer:** Learns transformations shared across instances by iteratively updating its parameters on an optimization set before producing pseudo fields. Effective but computationally expensive and less adaptive to each case.
- **Downsample optimizer:** Optimizes a coarse deformation field [104], then upsamples to full resolution. While faster, it loses fine-grained detail, leading to degraded performance.

Our instance-specific optimizer outperforms these alternatives, striking the best balance between accuracy and efficiency.

Optimizer Design	DSC	Time	VRAM Usage
None	0.635	N/A	12.07 GB
Network-based	0.642	13.604	14.35 GB
DownSample	0.610	0.097	12.44 GB
<i>Ours</i>	0.654	3.228	13.07 GB

Table 3.9: Ablation results of the OFG Optimizer on LPBA40 Dataset. This evaluation focuses over the initial 30 epochs.

Role of Pseudo Ground Truth. We ablated OFG’s reliance on pseudo labels by introducing alternative strategies: (1) *Epoch-interval optimization*, where pseudo labels are updated every few epochs; and (2) *Threshold-based optimization*, where labels are applied stochastically. As Fig. 3.10 shows, both weaken performance relative to continuous on-the-fly supervision.

We also tested a modified loss formulation (Eq. 3.19) by adding an auxiliary similarity term L_{sim} alongside L_{reg} and L_{opt} . This diluted the supervision from ϕ_{opt} and led to degraded accuracy, reaffirming the importance of prioritizing high-quality pseudo labels.

$$L_{all} = \alpha L_{sim} + \lambda L_{reg} + \beta L_{opt}. \quad (3.19)$$

3.3.4 Conclusion

In this work, we introduced On-the-Fly Guidance (OFG), a training framework that integrates learning-based prediction with instance-specific optimization to dynamically generate pseudo ground truth during training. Experiments across multiple benchmark datasets demonstrated that OFG consistently improves registration accuracy and regularity while maintaining inference efficiency, highlighting its effectiveness and generalizability. Nonetheless, the current framework introduces additional computational cost due to the optimizer module. Future work may focus on designing more lightweight optimization strategies and

extending OFG to broader medical imaging tasks to further enhance its practicality.

3.4 Deformable Coordinate Attention for Multi-scale Registration

3.4.1 Introduction

Medical image registration is a fundamental task in medical image analysis [94, 126, 43], aiming to spatially align images from different time points, modalities, or subjects by establishing dense anatomical correspondences. Achieving accurate registration across the full spectrum of deformations—from subtle sub-voxel misalignments to large displacements caused by breathing motion, organ shifts, or inter-patient variability—remains a longstanding challenge.

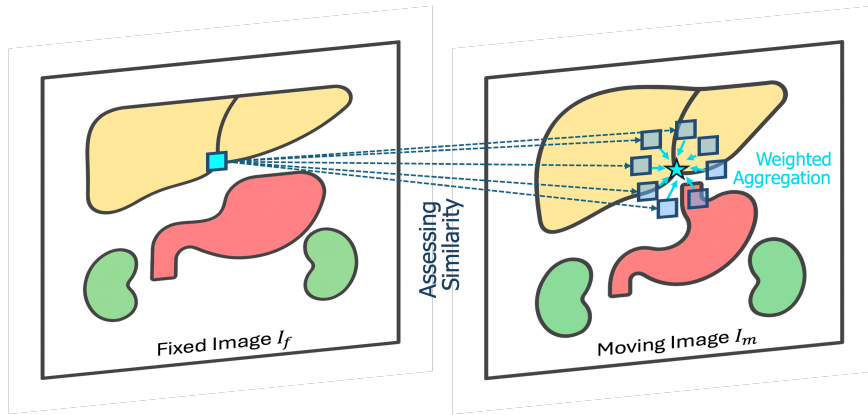


Figure 3.11: **Deformable Coordinate Attention (DCA)**. DCA computes attention between a fixed-image feature and a set of sampled moving-image features guided by predicted correspondence offsets. These offsets adaptively determine candidate locations, and the resulting attention weights produce the final correspondence for deformation propagation.

Classical registration methods approach the problem as iterative optimization of image similarity subject to regularization constraints [10, 120, 117, 15, 8, 102]. While effective in certain

settings, these methods are computationally expensive, sensitive to initialization, and struggle with complex or large deformations. With the advent of deep learning, learning-based registration frameworks [46, 27] have demonstrated remarkable efficiency and accuracy by directly learning mappings between image pairs.

Early approaches such as VoxelMorph [12] pioneered encoder–decoder architectures for end-to-end registration [36, 72, 59, 25, 24, 122, 31]. Among subsequent developments, coarse-to-fine multi-scale frameworks [104, 96, 78, 69, 92, 97, 155, 136, 99, 139] have become the dominant paradigm: coarse scales capture global alignment and large displacements, while fine scales refine local details. Despite their success, these methods share a critical limitation—**they propagate deformation blindly across scales without assessing correspondence reliability**. Unreliable predictions at coarse levels are directly inherited and amplified during refinement, leading to systematic error accumulation and degraded accuracy.

Recent efforts have attempted to mitigate this problem using correlation-based mechanisms. Methods such as Dual-PRNet++ [69], Im2Grid [87], ModeT [135], CorrMLP [98], and SACB-Net [32] leverage correlation layers to estimate voxelwise correspondences. However, these approaches rely on fixed local neighborhoods (typically $3 \times 3 \times 3$), restricting them to short-range dependencies. This design prevents discovery of distant correspondences required for large deformations, and lacks the flexibility to adapt search granularity for sub-voxel refinements.

The fundamental challenge lies in balancing computational efficiency with correspondence flexibility. Fixed neighborhoods are efficient but spatially restrictive; expanding the search region improves coverage but dramatically increases cost and ambiguity. Current frameworks lack an adaptive mechanism to focus computational resources where meaningful correspondences exist while dynamically adjusting granularity to the deformation scale.

To overcome these limitations, we propose **Deformable Coordinate Attention (DCA)**, a novel framework that introduces attention-guided correspondence discovery and reliability-aware propagation. Our key insight is that **robust multi-scale registration requires adaptive, unconstrained correspondence discovery prior to deformation propagation**. Instead of exhaustive search, DCA dynamically predicts likely correspondence locations and positions sparse attention keys at these locations, enabling efficient discovery across the entire deformation spectrum—from fine sub-voxel adjustments to large anatomical displacements. This design transforms the coarse-to-fine paradigm from blind field estimation into a reliability-aware, attention-guided process that achieves both accuracy and efficiency.

3.4.2 Methodology

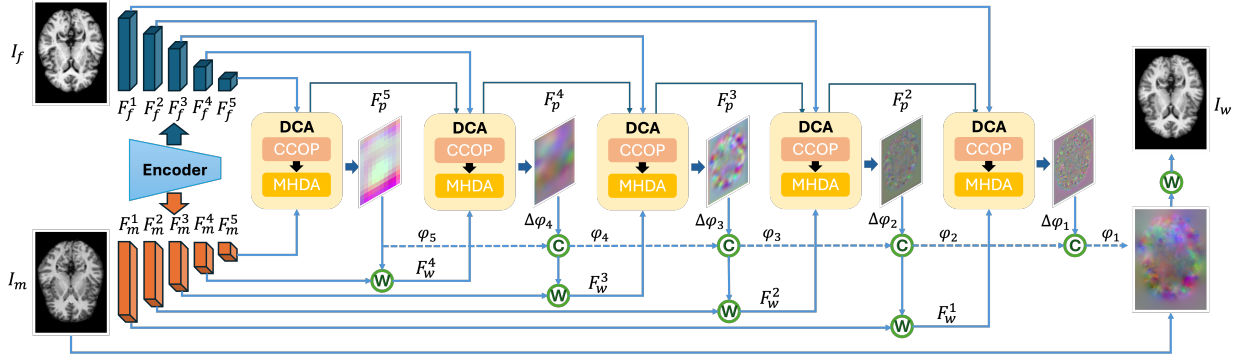


Figure 3.12: **Overview of the Proposed Method.** The shared encoder extracts multi-scale representations from moving and fixed images, which are then used in DCA (Deformable Coordinate Attention) blocks in a coarse-to-fine registration manner. At each scale, DCA combines Correlation-guided Cross-scale Offset Prediction (CCOP) with Multi-Head Deformable Aggregation (MHDA) to iteratively refine the deformation field through attention-guided correspondence discovery and propagation.

3.4.2.1 Problem Formulation

Given a moving image $I_m \in \mathbb{R}^{H \times W \times D}$ and a fixed image $I_f \in \mathbb{R}^{H \times W \times D}$, medical image registration aims to find a dense deformation field $\phi : \Omega \rightarrow \Omega$ that establishes spatial

correspondences between anatomical structures, where $\Omega \subset \mathbb{R}^3$ represents the image domain. The registered moving image is obtained as $I_m \circ \phi$, where $\circ$ denotes spatial transformation.

Multi-scale registration methods decompose this task hierarchically across L resolution levels. At each scale l , a deformation field $\phi^{(l)}$ is predicted and propagated to the next finer scale. However, this blind propagation paradigm accumulates errors when correspondence predictions are unreliable. Our work tries to address this limitation by introducing an attention-guided propagation that aggregates correspondence with reliability before cross-scale propagation.

3.4.2.2 Architecture Overview

Figure 3.12 illustrates the overall architecture, which consists of three main components: (1) a shared feature pyramid encoder that extracts multi-scale representations, (2) Deformable Coordinate Attention (DCA) modules that perform attention-guided correspondence discovery at each scale, and (3) a progressive deformation refinement scheme that propagated correspondences across scales.

3.4.2.3 Multi-Scale Feature Extraction

The shared encoder $\mathcal{E}$ extracts hierarchical representations from both the moving and fixed images across L scales:

$$F_m = \mathcal{E}(I_m), \quad F_f = \mathcal{E}(I_f), \quad (3.20)$$

where $F_m^{(l)}, F_f^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l \times D_l}$ denote the feature maps at scale l . Each level $l \in \{1, 2, 3, 4, 5\}$ corresponds to a resolution of $[\frac{H}{2^{l-1}}, \frac{W}{2^{l-1}}, \frac{D}{2^{l-1}}]$ for height, width, and depth, respectively. The number of channels increases with depth as $C_l = C_1 \times 2^{l-1}$, enabling progressively richer and

more abstract feature representations at finer scales.

At each scale, the encoder employs two convolutional layers followed by instance normalization and LeakyReLU activation, ensuring stable training and effective feature extraction. These multi-scale features provide the structural foundation for subsequent attention-guided registration, allowing the network to capture both coarse anatomical context and fine-grained spatial details in a unified manner.

3.4.2.4 Attention-Guided Deformation Propagation

The decoder performs L stages of coarse-to-fine registration using the proposed Deformable Coordinate Attention (DCA) mechanism.

At the coarsest scale L , the moving and fixed feature maps $F_m^{(L)}$ and $F_f^{(L)}$ are fed into a DCA block, which consists of a Correlation-guided Cross-scale Offset Prediction (CCOP) module and a Multi-Head Deformable Aggregation (MHDA) module. The CCOP module estimates spatial correspondences and produces an initial deformation field $\phi^{(L)}$, providing a coarse global alignment.

For each finer scale $l < L$, the propagated feature $F_p^{(l+1)}$ from the previous CCOP is combined with the warped moving feature $F_w^{(l)} = F_m^{(l)} \circ \phi^{(l+1)}$ and the fixed feature $F_f^{(l)}$. These inputs are used by the CCOP module to predict multi-head offsets $\{\Delta_h^{(l)}\}_{h=1}^{\mathbb{H}_l}$, where $\mathbb{H}_l$ is the number of attention heads at scale l .

The MHDA module then applies deformable attention at the predicted offsets. The attention weights $\text{Attention}_h^{(l)}$ indicate correspondence reliability, while the deformation patterns $A_h^{(l)}$ are regressed into a refinement field $\Delta\phi^{(l)}$ to update the deformation.

This process is repeated across all scales, with each level refining deformations from coarser levels. The final dense deformation field $\phi^{(1)}$ integrates both global alignment and fine-

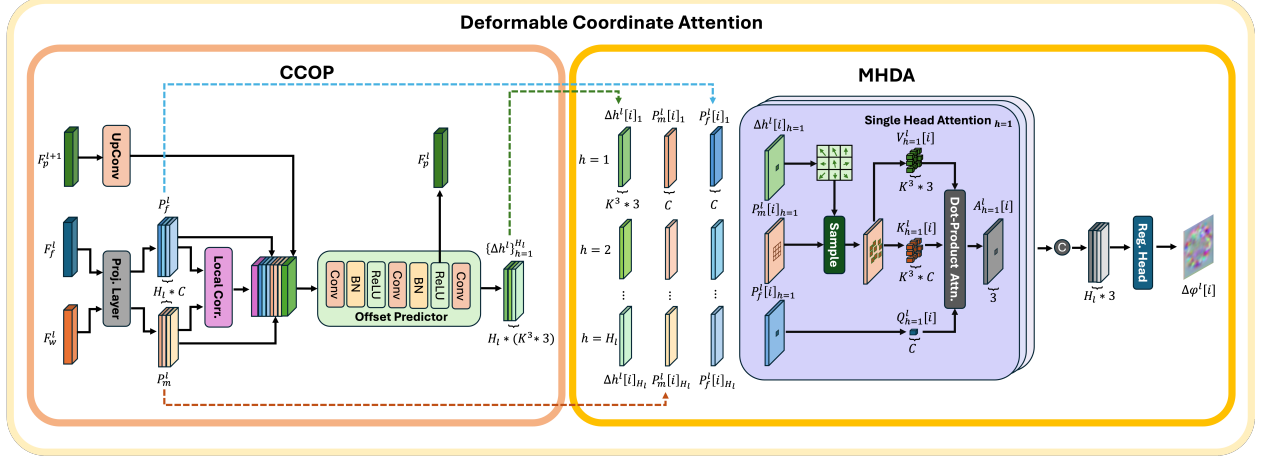


Figure 3.13: **Detailed illustration of Deformable Coordinate Attention (DCA).** DCA consists of two modules: (1) **CCOP** predicts multi-head unconstrained spatial offsets by combining projected features, local-window correlations, and propagated features from the previous scale, while also generating features for the next scale. (2) **MHDA** uses these offsets to dynamically place attention keys, where queries come from fixed features, keys from warped moving features, and values from sampled positions. Attention weights highlight correspondence reliability, and the multi-head design captures diverse deformation patterns for robust registration.

grained local adjustments.

3.4.2.5 Deformable Coordinate Attention

Figure 3.13 illustrates the core innovation of our method: Deformable Coordinate Attention (DCA) mechanism, which transforms blind deformation propagation into attention-guided correspondence discovery. DCA comprises two synergistic modules: Correlation-guided Cross-scale Offset Prediction (CCOP) and Multi-Head Deformable Aggregation (MHDA).

3.4.2.6 Correlation-guided Cross-scale Offset Prediction

The CCOP module transforms the limitation of fixed rigid neighborhood search $\mathcal{N}_K(i) = \{i + \delta : \delta \in [-\lfloor K/2 \rfloor, \lfloor K/2 \rfloor]^3\}$ where K is the local search window size, by predicting unconstrained spatial offsets that indicate where correspondences likely exist. Rather than being

restricted to predefined local windows, CCOP enables adaptive correspondence discovery across any distance.

Feature Projection and Correlation Computation: At scale l , CCOP first applies shared projection layers to decouple feature extraction from deformation prediction. The fixed image features and warped moving image features are projected into a dedicated correspondence space using a multi-layer perceptron (MLP) and layer normalization:

$$P_f^{(l)} = \text{Proj}(F_f^{(l)}), \quad P_m^{(l)} = \text{Proj}(F_w^{(l)}) \quad (3.21)$$

where $F_w^{(l)} = F_m^{(l)} \circ \phi^{(l+1)}$ and $\phi^{(l+1)}$ is the accumulated deformation from previous scale $l+1$. The number of channels for the projected feature is $\mathbb{H}_l \times C_{head}$, where $\mathbb{H}_l$ is the number of the heads for scale l and C_{head} represents the single-head dimension. The multi-head mechanism will be further introduced in Sec. 3.4.2.7 to capture various deformation pattern for robust registration.

A local correlation $C^{(l)} \in \mathbb{R}^{(H_l \times W_l \times D_l) \times (L \times L \times L)}$ is computed within an $L \times L \times L$ neighborhood window to provide the spatial prior for the following offset prediction [69, 98]:

$$C_{i,j}^{(l)} = \text{mean}(P_f^{(l)}[i] \odot P_m^{(l)}[j]) \quad (3.22)$$

where i indexes spatial positions, $j \in \mathcal{N}_L(i)$ indexes positions within the local window of size L , $\odot$ denotes element-wise multiplication, and $\text{mean}(\cdot)$ computes the average across the channel dimension.

Unconstrained Offset Prediction: To incorporate accumulated knowledge from coarser scales, the propagated feature $F_p^{(l+1)}$ from the previous scale is upsampled and concatenated

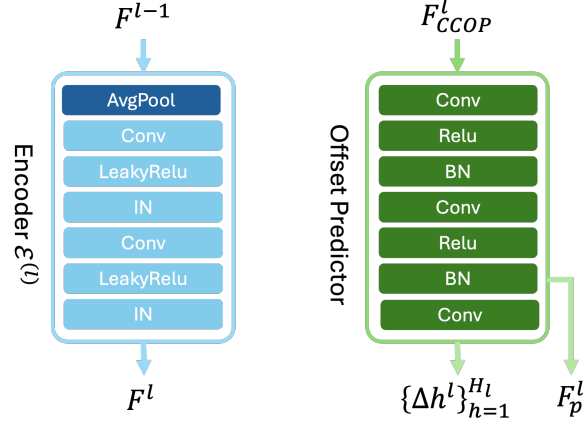


Figure 3.14: Network Structure for the Encoder layers and Offset Predictor.

with current-scale projected features and correlation volume.

$$F_{CCOP}^{(l)} = \text{Cat}([P_f^{(l)}, P_m^{(l)}, \text{Up}(F_p^{(l+1)}), C^{(l)}]) \quad (3.23)$$

The concatenated features are processed through a lightweight regression network consisting of three convolutional layers, batch normalization, and ReLU activation to generate multi-head offsets and updated propagated features:

$$\{\Delta_h^{(l)}\}_{h=1}^{\mathbb{H}_l}, F_p^{(l)} = \text{OffsetPredictor}(F_{CCOP}^{(l)}) \quad (3.24)$$

where $\mathbb{H}_l$ represents the number of attention heads, and $\Delta_h^{(l)} \in \mathbb{R}^{(H_l \times W_l \times D_l) \times (K \times K \times K) \times 3}$ represents the predicted offset $(\Delta x, \Delta y, \Delta z)$ for $K \times K \times K$ search window of each head. The enhanced cross-scale feature $F_p^{(l)}$ is then propagated to the next finer scale.

Crucially, these offsets $\Delta_h^{(l)}$ are unconstrained in magnitude, enabling the network to adaptively determine the search range based on deformation complexity. This flexibility allows CCOP to handle both large-displacement search for capturing global motion and fine-grained positioning for local refinements within a unified framework. The predicted offsets serve as dynamic positioning guides for the subsequent MHDA module.

3.4.2.7 Multi-Head Deformable Aggregation

The MHDA module performs dynamic correspondence aggregation by positioning attention keys at the updated neighbourhood location using the offset predicted by CCOP. Unlike traditional fixed-grid attention, MHDA enables flexible correspondence discovery while maintaining computational efficiency with sparse attention mechanism.

Dynamic Key Positioning: For each attention head $h \in \{1, \dots, \mathbb{H}_l\}$ of scale l and each query position i , MHDA samples $K \times K \times K$ keys from the projected warped moving features at the updated neighbourhood locations using predicted offset from the CCOP. The queries are derived from the projected fixed features, while keys are sampled at offset positions and values encode the spatial coordinates:

$$Q_h^{(l)}[i] = P_f^{(l)}[i]_h \quad (3.25)$$

$$K_h^{(l)}[i, j] = \text{Sample}(P_m^{(l)}{}_h, \mathcal{N}_K(i)[j] + \Delta_h^{(l)}[i, j]) \quad (3.26)$$

$$V_h^{(l)}[i, j] = \mathcal{N}_K(i)[j] + \Delta_h^{(l)}[i, j] \quad (3.27)$$

where $i \in \{1, \dots, H_l \times W_l \times D_l\}$ indexes the query position, and $j \in \{1, \dots, K^3\}$ indexes the $K \times K \times K$ sampled positions for query i . $\text{Sample}(\cdot, \cdot)$ performs trilinear interpolation, $V_h^{(l)}[i, j]$ represents the spatial coordinates of the sampled key positions for the head h , and $P_m^{(l)}{}_h$ and $P_f^{(l)}{}_h$ represent C_{head} -dimensional features corresponding to the head h , from projected features $P_m^{(l)}$ and $P_f^{(l)}$, respectively.

Attention-Based Correspondence Assessment: For each query position and attention head, the attention mechanism computes correspondence weights across the K^3 key positions:

$$\text{Attention}_h^{(l)}[i] = \text{Softmax} \left(\frac{Q_h^{(l)}[i] \cdot K_h^{(l)}[i, :]}{\sqrt{d_k}} \right) \quad (3.28)$$

where d_k represents the key dimension. The attended coordinate for each head is computed as:

$$A_h^{(l)}[i] = \sum_{j=1}^{K^3} \text{Attention}_h^{(l)}[i, j] \cdot V_h^{(l)}[i, j] \quad (3.29)$$

Multi-Head Coordinate Aggregation: For each query position i , we obtain $\mathbb{H}_l$ attended coordinates from all attention heads. These coordinates are concatenated and processed through a simple regression head to produce the final deformation:

$$\Delta\phi^{(l)}[i] = \text{RegHead}(\text{Cat}([A_1^{(l)}[i], \dots, A_{\mathbb{H}_l}^{(l)}[i]])) \quad (3.30)$$

$$\phi^{(l)} = \phi^{(l+1)} \circ \Delta\phi^{(l)} \quad (3.31)$$

This multi-head design from the original attention mechanism [134] helps to capture robust correspondence through complementary deformation patterns—each head can specialize in different types of anatomical motion while the concatenated representation provides comprehensive correspondence information.

3.4.2.8 Unsupervised Training Loss Function

Our method is trained using an unsupervised multi-component loss function that balances registration accuracy, deformation smoothness, and physical plausibility. All loss components are computed on the final deformation field.

Image Similarity Loss: The primary registration objective employs normalized cross-correlation (NCC) to measure alignment quality between fixed and moving images:

$$\mathcal{L}_{\text{sim}} = -NCC(I_f, I_m \circ \phi) \quad (3.32)$$

where ϕ is the final deformation field.

$$NCC(I_f, I_w) = \sum_{\mathbf{p} \in \Omega} \frac{\left(\sum_{\mathbf{p}_i} (I_w(\mathbf{p}_i) - \bar{I}_w(\mathbf{p})) (I_f(\mathbf{p}_i) - \bar{I}_f(\mathbf{p})) \right)^2}{\sum_{\mathbf{p}_i} (I_w(\mathbf{p}_i) - \bar{I}_w(\mathbf{p}))^2 \sum_{\mathbf{p}_i} (I_f(\mathbf{p}_i) - \bar{I}_f(\mathbf{p}))^2} \quad (3.33)$$

Smoothness Regularization: To ensure spatially coherent deformations, we apply smoothness constraints on the final deformation field:

$$\mathcal{L}_{\text{smooth}} = \sum_{d \in \{x, y, z\}} \|\nabla \phi_d^{(1)}\|_2^2 \quad (3.34)$$

where $\phi_d^{(1)}$ denotes the d -th component of the final deformation field, and ∇ represents the spatial gradient operator.

Negative Jacobian Determinant Loss: To preserve topology and prevent non-physical deformations such as folding or tearing, we penalize negative Jacobian determinants:

$$\mathcal{L}_{\text{jac}} = \frac{1}{N} \sum_{\mathbf{p} \in \Omega} \text{relu}(-|J_\phi(\mathbf{p})|) \quad (3.35)$$

where J_ϕ represents the Jacobian matrix of the final deformation field $\phi^{(1)}$, and $\text{relu}(\cdot)$ ensures only negative determinants contribute to the loss.

$$J_\phi(\mathbf{p}) = \begin{bmatrix} \frac{\partial \phi_x(\mathbf{p})}{\partial x} & \frac{\partial \phi_x(\mathbf{p})}{\partial y} & \frac{\partial \phi_x(\mathbf{p})}{\partial z} \\ \frac{\partial \phi_y(\mathbf{p})}{\partial x} & \frac{\partial \phi_y(\mathbf{p})}{\partial y} & \frac{\partial \phi_y(\mathbf{p})}{\partial z} \\ \frac{\partial \phi_z(\mathbf{p})}{\partial x} & \frac{\partial \phi_z(\mathbf{p})}{\partial y} & \frac{\partial \phi_z(\mathbf{p})}{\partial z} \end{bmatrix} \quad (3.36)$$

Combined Loss Function: The final loss function combines all components with appropriate weighting:

$$\mathcal{L} = \mathcal{L}_{\text{sim}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{jac}}\mathcal{L}_{\text{jac}} \quad (3.37)$$

where λ_{smooth} and λ_{jac} are hyperparameters that control relative importance of smoothness and topology preservation.

3.4.3 Experiments

3.4.3.1 Datasets

We evaluated the proposed framework on three public 3D brain MRI datasets and one abdominal CT dataset, covering diverse anatomical regions and deformation scenarios.

MindBoggle [73] contains 101 brain MRI scans with 97 manually annotated regions of interest (ROIs), including 62 cortical and 35 subcortical structures. The dataset is widely used for validating registration due to its high-quality manual segmentations of both volumes and surfaces. We used 80 scans for training, 7 for validation, and 24 for cross-site testing to assess generalizability.

LPBA40 [119] includes 40 brain MRI scans from healthy adults, each annotated with 54 anatomical labels. We split the dataset into 25 training, 5 validation, and 10 testing scans.

OASIS [95] provides 414 brain MRI scans from a cross-sectional cohort including young, middle-aged, nondemented, and demented older adults. We used 200 scans for training, 20 for validation, and the provided test pairs for evaluation.

AbdomenCT [143] consists of 30 abdominal CT scans with 13 anatomical annotations.

This dataset poses greater challenges due to large anatomical variations and deformations caused by breathing and organ shifts. We used 18 scans for training, 4 for validation, and 8 for testing.

3.4.3.2 Implementation Details

The framework was implemented in PyTorch and trained on a single NVIDIA RTX A6000 GPU. We adopted a 5-scale pyramid architecture. For offset prediction, the local correlation neighborhood size was set to $L = 3$, and the deformable attention search window was set to $K = 3$. The MHDA module employed scale-adaptive attention heads to capture diverse correspondence patterns.

Training was performed with the Adam optimizer, using a learning rate of 10^{-4} with linear annealing. The loss function hyperparameters were tuned to $\lambda_{\text{smooth}} = 1$ and $\lambda_{\text{jac}} = 50$.

3.4.3.3 Evaluation Metrics

We employed standard metrics for medical image registration. Registration accuracy was evaluated using the **Dice Similarity Coefficient (DSC)**, defined as $\text{DSC} = \frac{2|A \cap B|}{|A| + |B|}$, where A and B are the label sets. **Average Symmetric Surface Distance (ASSD)** was used to quantify boundary alignment by computing the mean distance between corresponding surfaces. To assess the smoothness and invertibility of predicted transformations, we report the percentage of **Negative Jacobian Determinants (NJD)**, which indicate topological violations.

3.4.3.4 Comparison Methods

We compared our method with state-of-the-art approaches, including **2** traditional optimization-based methods (Affine and SyN [8]) and **16** learning-based methods. The learning-based methods are grouped into three categories:

- **Single-Scale** methods, which perform registration at the finest resolution only: Vox-elMorph [12], ViT-V-Net [25], TransMorph [24], TransMatch [31], TransDCA [26], and H-ViT [48].
- **Multi-Scale** methods, which decompose registration hierarchically: PIViT [92], NICE-Net [96], NICE-Trans [97], RPNet [93], and RDPNet [136].
- **Local Correlation** methods, which rely on voxel-wise correlations in a fixed neighborhood: Dual-PR++ [69], Im2Grid [87], ModeT [135], CorrMLP [98], and SACB-Net [32].

All methods were trained and evaluated under identical conditions with the same data splits and preprocessing pipelines to ensure fairness.

3.4.3.5 Brain Registration Results

Table 3.10 reports results across three brain MRI datasets. Our method consistently outperformed all baselines in terms of DSC, ASSD, and NJD. Single-scale methods lagged behind, while multi-scale methods showed clear advantages, particularly RDPNet, which achieved strong performance with recursive refinement and diffeomorphic guarantees. Local correlation-based methods like ModeT and CorrMLP achieved competitive results but were still limited by fixed-window correspondence search.

Table 3.10: Quantitative comparisons on MindBoggle, LPBA, and OASIS datasets. **Red**, **Green** and **Blue** denote the **1st**, **2nd** and **3rd** best.

Group	Method	MindBoggle (107 ROIs)			LPBA (54 ROIs)			OASIS (35 ROIs)		
		DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$	DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$	DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$
Traditional	Affine	0.404 $\pm$ 0.021	2.137 $\pm$ 0.196	0	0.635 $\pm$ 0.020	2.120 $\pm$ 0.125	0	0.571 $\pm$ 0.052	1.618 $\pm$ 0.281	0
	SyN[8]	0.513 $\pm$ 0.031	1.692 $\pm$ 0.211	0	0.695 $\pm$ 0.017	1.780 $\pm$ 0.109	0	0.755 $\pm$ 0.029	0.837 $\pm$ 0.123	0
Single-Scale	VoxelMorph[12]	0.500 $\pm$ 0.108	2.147 $\pm$ 1.092	0.883 $\pm$ 0.289	0.638 $\pm$ 0.031	2.122 $\pm$ 0.242	0.589 $\pm$ 0.273	0.763 $\pm$ 0.031	0.823 $\pm$ 0.145	0.963 $\pm$ 0.179
	ViT-V-Net[25]	0.526 $\pm$ 0.112	2.086 $\pm$ 1.128	0.797 $\pm$ 0.332	0.648 $\pm$ 0.030	2.049 $\pm$ 0.225	0.735 $\pm$ 0.247	0.787 $\pm$ 0.031	0.733 $\pm$ 0.136	0.917 $\pm$ 0.151
	TransMorph[24]	0.551 $\pm$ 0.118	1.963 $\pm$ 1.083	0.691 $\pm$ 0.181	0.663 $\pm$ 0.030	1.982 $\pm$ 0.232	0.648 $\pm$ 0.199	0.806 $\pm$ 0.025	0.661 $\pm$ 0.103	0.827 $\pm$ 0.169
	TransMatch[31]	0.544 $\pm$ 0.112	1.950 $\pm$ 1.051	0.675 $\pm$ 0.200	0.667 $\pm$ 0.018	1.956 $\pm$ 0.218	0.515 $\pm$ 0.162	0.805 $\pm$ 0.022	0.665 $\pm$ 0.096	0.767 $\pm$ 0.138
	TransDCA[26]	0.550 $\pm$ 0.118	1.944 $\pm$ 1.071	0.604 $\pm$ 0.161	0.661 $\pm$ 0.031	1.998 $\pm$ 0.243	0.672 $\pm$ 0.240	0.803 $\pm$ 0.023	0.674 $\pm$ 0.097	0.818 $\pm$ 0.166
	H-ViT[48]	0.536 $\pm$ 0.113	2.033 $\pm$ 1.137	0.751 $\pm$ 0.342	0.667 $\pm$ 0.030	1.969 $\pm$ 0.235	0.524 $\pm$ 0.177	0.803 $\pm$ 0.021	0.676 $\pm$ 0.095	0.817 $\pm$ 0.156
Multi-Scale	PIViT[92]	0.559 $\pm$ 0.075	1.737 $\pm$ 0.694	0.041 $\pm$ 0.025	0.706 $\pm$ 0.013	1.712 $\pm$ 0.113	0.026 $\pm$ 0.012	0.797 $\pm$ 0.022	0.684 $\pm$ 0.096	0.180 $\pm$ 0.044
	NICE-Net[96]	0.589 $\pm$ 0.081	1.668 $\pm$ 0.718	0.651 $\pm$ 0.239	0.715 $\pm$ 0.157	1.657 $\pm$ 0.135	0.418 $\pm$ 0.107	0.821 $\pm$ 0.023	0.610 $\pm$ 0.103	0.820 $\pm$ 0.160
	NICE-Trans[97]	0.608 $\pm$ 0.068	1.568 $\pm$ 0.599	0.668 $\pm$ 0.231	0.726 $\pm$ 0.015	1.591 $\pm$ 0.126	0.389 $\pm$ 0.093	0.828$\pm$0.024	0.588$\pm$0.097	0.822 $\pm$ 0.158
	RPNet[93]	0.587 $\pm$ 0.076	1.659 $\pm$ 0.676	0.113 $\pm$ 0.056	0.714 $\pm$ 0.015	1.675 $\pm$ 0.136	0.060 $\pm$ 0.020	0.816 $\pm$ 0.023	0.624 $\pm$ 0.099	0.183 $\pm$ 0.057
	RDPNet[136]	0.608 $\pm$ 0.0748	1.606 $\pm$ 0.659	0.003$\pm$0.001	0.729$\pm$0.015	1.589 $\pm$ 0.124	0.001$\pm$0.001	0.830$\pm$0.023	0.581$\pm$0.096	0.008$\pm$0.003
Local-Corr.	Dual-PR++[69]	0.562 $\pm$ 0.102	1.899 $\pm$ 0.974	0.258 $\pm$ 0.111	0.695 $\pm$ 0.021	1.798 $\pm$ 0.183	0.040 $\pm$ 0.018	0.807 $\pm$ 0.021	0.657 $\pm$ 0.095	0.449 $\pm$ 0.106
	Im2Grid[87]	0.556 $\pm$ 0.098	1.768 $\pm$ 0.788	0.019 $\pm$ 0.006	0.703 $\pm$ 0.015	1.715 $\pm$ 0.129	0.008 $\pm$ 0.003	0.804 $\pm$ 0.021	0.666 $\pm$ 0.095	0.028 $\pm$ 0.012
	ModeT[135]	0.614$\pm$0.051	1.472$\pm$0.418	0.008 $\pm$ 0.006	0.729$\pm$0.013	1.573$\pm$0.110	0.002 $\pm$ 0.001	0.824$\pm$0.023	0.600 $\pm$ 0.095	0.013 $\pm$ 0.006
	CorrMLP[98]	0.617$\pm$0.057	1.533 $\pm$ 0.528	0.599 $\pm$ 0.291	0.731$\pm$0.015	1.566$\pm$0.110	0.307 $\pm$ 0.078	0.830$\pm$0.023	0.580$\pm$0.093	0.761 $\pm$ 0.143
	SACB-Net[32]	0.605 $\pm$ 0.055	1.514$\pm$0.455	0.125 $\pm$ 0.041	0.725 $\pm$ 0.013	1.581 $\pm$ 0.108	0.103 $\pm$ 0.019	0.818 $\pm$ 0.023	0.616 $\pm$ 0.095	0.208 $\pm$ 0.039
Proposed	DCA	0.624$\pm$0.049	1.468$\pm$0.447	0.005$\pm$0.004	0.736$\pm$0.015	1.549$\pm$0.123	0.002$\pm$0.001	0.830$\pm$0.024	0.580$\pm$0.099	0.006$\pm$0.003

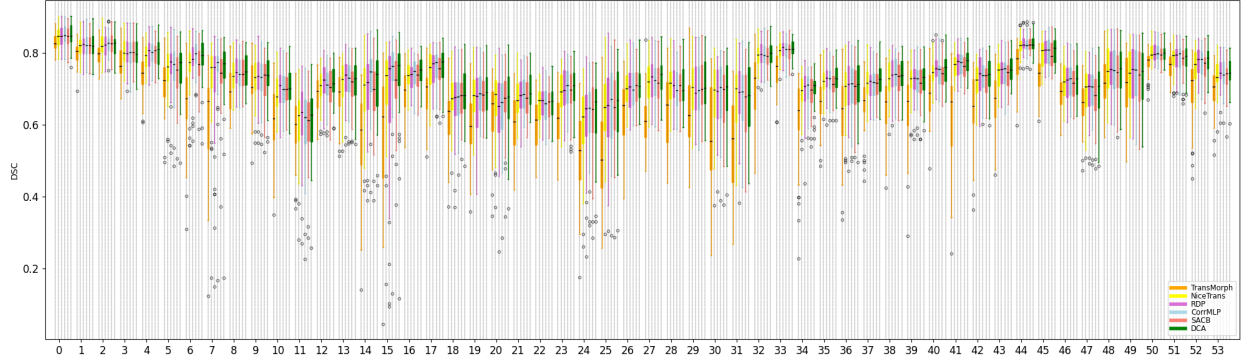


Figure 3.15: Dice score distributions on LPBA40 (54 ROIs).

Our proposed DCA achieved the highest accuracy on MindBoggle and LPBA, and matched the best performance on OASIS. These results highlight the benefits of attention-guided correspondence discovery and cross-scale propagation in capturing both global and local deformations.

3.4.3.6 AbdomenCT Registration Results

Abdominal CT registration is particularly challenging due to large non-linear deformations caused by breathing and organ shifts. Table 3.11 shows that DCA achieves state-of-the-art performance on this dataset, demonstrating robustness to large anatomical variability.

Table 3.11: Quantitative comparison on AbdomenCT dataset. **Red**, **Green** and **Blue** denote the **1st**, **2nd** and **3rd** best results.

Group	Method	AbdomenCT (13 ROIs)		
		DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$
Traditional	Affine SyN[8]	0.251 $\pm$ 0.070	6.629 $\pm$ 2.416	0
		0.319 $\pm$ 0.084	5.780 $\pm$ 2.443	0
Single-Scale	VoxelMorph[12]	0.349 $\pm$ 0.051	4.793 $\pm$ 1.047	1.746 $\pm$ 0.094
	ViT-V-Net[25]	0.385 $\pm$ 0.054	4.547 $\pm$ 1.032	4.798 $\pm$ 1.400
	TransMorph[24]	0.409 $\pm$ 0.061	4.422 $\pm$ 1.118	5.554 $\pm$ 1.368
	TransMatch[31]	0.420 $\pm$ 0.064	4.258 $\pm$ 1.097	5.188 $\pm$ 1.430
	TransDCA[26]	0.420 $\pm$ 0.058	4.317 $\pm$ 1.071	4.497 $\pm$ 1.375
	H-ViT[48]	0.429 $\pm$ 0.055	4.147 $\pm$ 0.986	5.219 $\pm$ 2.250
Multi-Scale	PIViT[92]	0.416 $\pm$ 0.044	4.005 $\pm$ 0.714	0.367 $\pm$ 0.179
	NICE-Net[96]	0.464 $\pm$ 0.058	3.666 $\pm$ 0.893	6.375 $\pm$ 2.100
	NICE-Trans[97]	0.524 $\pm$ 0.047	3.034 $\pm$ 0.689	14.938 $\pm$ 2.943
	RPNNet[93]	0.480 $\pm$ 0.055	3.543 $\pm$ 0.723	1.700 $\pm$ 0.827
	RDPNet[136]	0.554 $\pm$ 0.052	3.023 $\pm$ 0.757	0.227 $\pm$ 0.216
Local-Corr.	Dual-PR++[69]	0.466 $\pm$ 0.056	3.883 $\pm$ 0.929	2.371 $\pm$ 1.208
	Im2Grid[87]	0.455 $\pm$ 0.041	3.642 $\pm$ 0.686	0.178 $\pm$ 0.119
	ModeT[135]	0.554 $\pm$ 0.053	2.729 $\pm$ 0.646	0.061$\pm$0.029
	CorrMLP[98]	0.572$\pm$0.055	2.645$\pm$0.708	12.235 $\pm$ 2.686
	SACB-Net[32]	0.577$\pm$0.047	2.575$\pm$0.603	1.642 $\pm$ 0.651
Proposed	DCA	0.597$\pm$0.052	2.411$\pm$0.603	0.136$\pm$0.092

Among learning-based competitors, SACB-Net performed strongly by incorporating structural modeling, but its reliance on fixed neighborhoods limited its ability to capture large displacements. In contrast, our unconstrained attention-guided offsets allowed more flexible correspondence discovery, leading to superior performance while maintaining reasonable topology preservation. These results confirm that attention-guided propagation becomes especially beneficial in scenarios with complex, large-scale deformations.

3.4.3.7 Ablation Study

Component Analysis. Table 3.12 analyzes the contributions of each component. Correlation guidance (CG) significantly improved DSC and ASSD by explicitly modeling correspondences. Cross-scale propagation (CS) further improved accuracy by transferring information across resolutions. Deformable aggregation (DA) reduced topology violations, and the multi-head extension (MH) improved robustness by capturing diverse correspondence

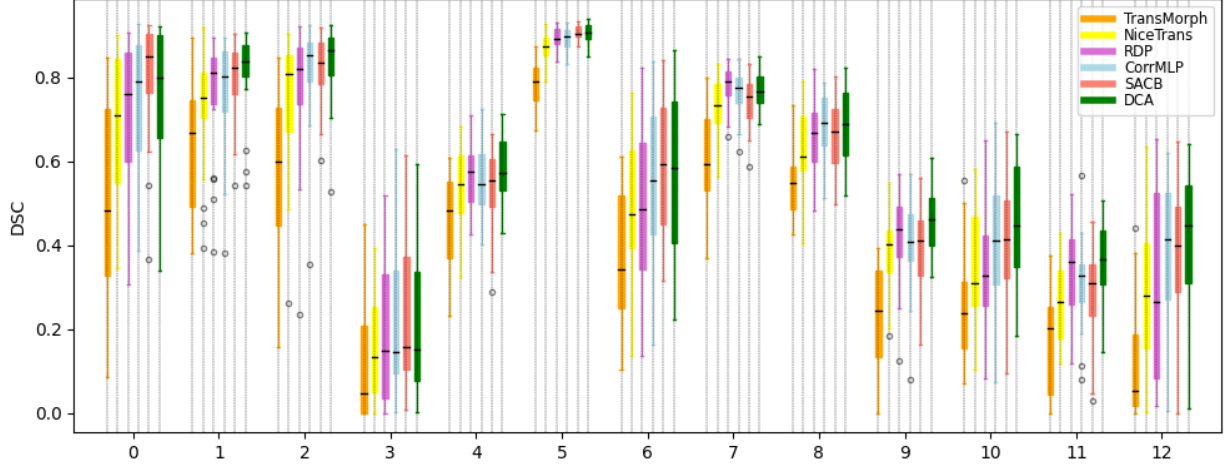


Figure 3.16: Dice score distributions on AbdomenCT (13 ROIs).

patterns. The full DCA achieved the best overall results, demonstrating the complementary benefits of all components.

Table 3.12: Ablation study on AbdomenCT dataset. Each component’s contribution is evaluated by adding modules to the baseline. **CG**: Correlation-Guidance; **CS**: Cross-Scale Feature Propagation; **DA**: Single-Head Deformation Aggregation; **MH**: Multi-Head Mechanism

Components				AbdomenCT (13 ROIs)		
CG	CS	DA	MH	DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$
				0.549 $\pm$ 0.054	2.830 $\pm$ 0.670	0.105 $\pm$ 0.080
✓				0.570 $\pm$ 0.051	2.731 $\pm$ 0.658	0.103 $\pm$ 0.066
✓	✓			0.577 $\pm$ 0.046	2.626 $\pm$ 0.646	0.151 $\pm$ 0.105
		✓		0.564 $\pm$ 0.045	2.649 $\pm$ 0.598	0.098 $\pm$ 0.064
		✓	✓	0.579 $\pm$ 0.049	2.585 $\pm$ 0.651	0.084 $\pm$ 0.048
✓	✓	✓	✓	0.597 $\pm$ 0.052	2.411 $\pm$ 0.603	0.136 $\pm$ 0.092

Effect of Jacobian Regularization. Table 3.13 evaluates the influence of the Jacobian loss L_{jac} on topology preservation. Increasing λ_{jac} consistently reduced NJD for both DCA and CorrMLP, with minimal impact on DSC, confirming the importance of topology-preserving regularization, especially for large deformations.

Attention Window Size. Table 3.14 shows the effect of the deformable attention window size. Both $K = 2$ and $K = 3$ provided strong results compared to the baseline without attention. Larger K values were not explored due to the cubic growth in computational

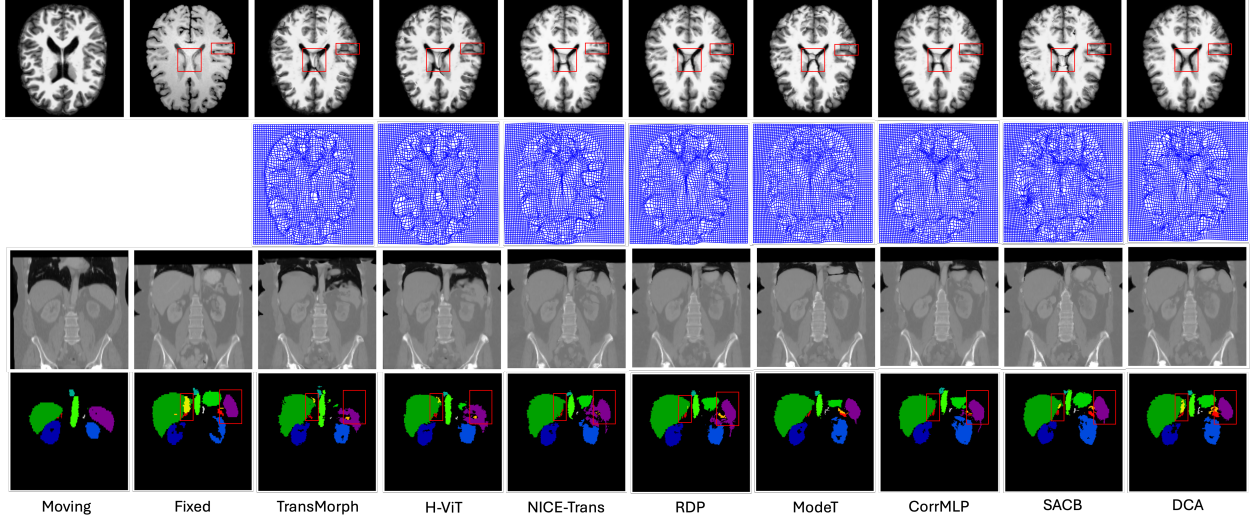


Figure 3.17: Qualitative comparisons on LPBA40 (top) and AbdomenCT (bottom).

Table 3.13: Hyperparameter Tuning for Negative Jacobian Loss

λ_{jac}	DCA		CorrMLP[98]	
	AbdomenCT	MindBoggle	AbdomenCT	MindBoggle
	DSC $\uparrow$ NJD(%) $\downarrow$	DSC $\uparrow$ NJD(%) $\downarrow$	DSC $\uparrow$ NJD(%) $\downarrow$	DSC $\uparrow$ NJD(%) $\downarrow$
0	0.590 8.381	0.625 0.191	0.572 12.235	0.617 0.599
20	0.592 0.317	0.623 0.011	0.580 0.613	0.615 0.028
50	0.597 0.136	0.624 0.005	0.575 0.126	0.615 0.014
80	0.595 0.086	0.623 0.003	0.576 0.082	0.615 0.010

cost, and $K = 3$ was selected as the optimal balance between accuracy and efficiency.

Table 3.14: Deformable Attention Size

K^3	AbdomenCT			MindBoggle		
	DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$	DSC $\uparrow$	ASSD $\downarrow$	NJD(%) $\downarrow$
$2^3 = 8$	0.595 $\pm$ 0.056	2.475 $\pm$ 0.614	0.135 $\pm$ 0.096	0.621 $\pm$ 0.054	1.512 $\pm$ 0.517	0.005 $\pm$ 0.006
$3^3 = 27$	0.597 $\pm$ 0.052	2.411 $\pm$ 0.603	0.136 $\pm$ 0.092	0.624 $\pm$ 0.049	1.468 $\pm$ 0.447	0.005 $\pm$ 0.004

3.4.3.8 Limitations

While DCA demonstrates robust performance, certain limitations remain. As illustrated in Figure 3.18, the model can introduce unnecessary deformations, such as cortical folding in brain scans, when optimizing solely for intensity similarity. In AbdomenCT, visually

aligned scans may still contain structural artifacts when examined via segmentation. These issues reflect the limitations of structure-unaware loss functions, which prioritize voxel-wise similarity over structural integrity.

Future work should explore structure-aware supervision, possibly by integrating anatomical priors directly into the attention mechanism. Such an approach could allow the model to weigh correspondences by both feature similarity and anatomical consistency, thereby improving reliability for complex structures.

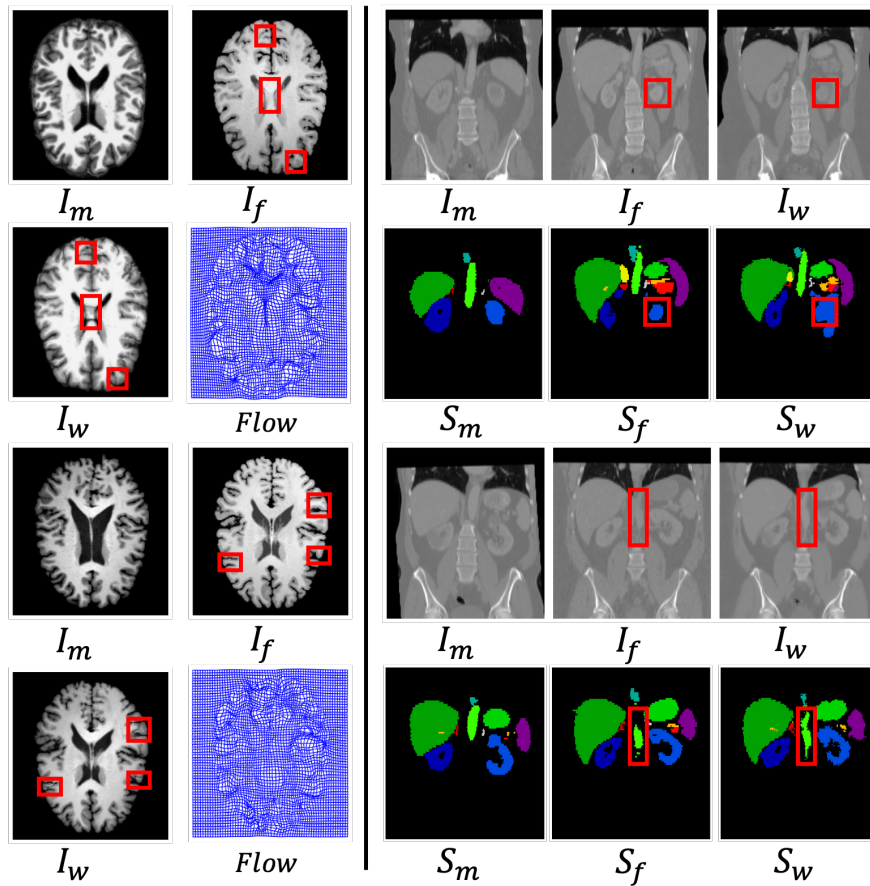


Figure 3.18: Failure cases. Left: unnecessary cortical folding (red box). Right: high image similarity but misaligned structures.

3.4.4 Conclusion

In this work, we introduced Deformable Coordinate Attention (DCA), a novel framework for multi-scale medical image registration that integrates attention-guided propagation with unconstrained correspondence discovery. By adaptively predicting correspondence candidates and applying sparse attention, DCA achieves accurate and efficient alignment across a broad spectrum of deformations. Comprehensive experiments on four benchmark datasets demonstrate its state-of-the-art performance, with particular strength in handling large and highly non-linear deformations. Looking ahead, future directions include incorporating structure-aware supervision to improve boundary-level accuracy, extending the framework to multi-modal registration, and enhancing scalability for large-scale clinical applications.

Chapter 4

Shape Representation

4.1 Introduction

Shape representation is a fundamental problem in medical image analysis, providing a structured way to capture anatomical geometry beyond raw voxel intensities. Accurate shape models enable statistical shape analysis, population studies, disease progression modeling, and atlas construction. By abstracting anatomy into geometric descriptors, they allow quantitative comparisons across individuals and cohorts.

Classical approaches—such as landmark models, meshes, and point clouds—have laid the foundation for this field. While effective in controlled scenarios, they often suffer from noise sensitivity, difficulty in establishing dense correspondences, and limited ability to represent fine-scale anatomical variability. These shortcomings are pronounced in medical imaging, where anatomical structures exhibit both global regularities and intricate local variations.

The advent of deep learning has introduced *implicit neural representations* (INRs) as a powerful alternative. INRs model shapes as continuous functions over spatial coordinates,

offering resolution independence, compact representation, and the ability to capture fine geometric detail. These properties make them highly appealing for medical imaging, where complex geometries and inter-subject variability are pervasive.

Yet, directly applying generic INR methods to medical shape analysis is insufficient. Key challenges remain: ensuring anatomical plausibility, modeling population-level variability, and enforcing consistent correspondences across subjects.

This dissertation addresses these challenges through a hierarchical neural field framework that combines the strengths of statistical shape modeling with the flexibility of implicit representations. By capturing global and local anatomical features in a unified manner, this approach provides a robust foundation for analyzing shape variability and advancing clinical shape analysis tasks.

4.2 Neural Diffeomorphic Flow in Shape Modeling

4.2.1 Introduction

Three-dimensional (3D) shape representation plays a central role in medical image analysis, providing a foundation for tasks such as reconstruction, segmentation, population studies, and disease progression modeling. Unlike intensity-based methods, shape representations enable explicit modeling of anatomical geometry, which is crucial for quantitative comparison across individuals and cohorts.

Classical shape representations, including voxel grids, point clouds, and polygon meshes, have long supported geometric modeling but suffer from inherent limitations: voxel grids require excessive memory and suffer from limited resolution, point clouds lack explicit connectivity, and meshes often depend on clean topology or predefined connectivity, which are challenging

to guarantee in medical imaging.

Recent advances in *deep implicit functions* (DIFs), or neural fields, offer a powerful alternative by representing shapes as continuous functions over spatial coordinates. DIFs provide compactness, resolution independence, and the ability to capture fine-scale geometric details. Foundational works such as DeepSDF [109], Occupancy Networks [100], and NeRF [101] established DIFs as a new paradigm, with later works improving locality and efficiency using patch-based or triplane representations [22, 66, 132, 23].

However, conventional DIFs remain limited in their ability to model semantic correspondences across shapes, a property essential for medical applications such as statistical shape analysis [90, 56], atlas construction, and cross-population studies. To address this, several methods represent shapes as deformations of a template shape, such as Deep Implicit Templates (DIT) [153], Neural Deformation Fields (NDF) [129], and ImplicitAtlas [146]. While these approaches capture shared structure, they often remain topology-agnostic, lose fine-scale details, or produce defective mappings when correspondences are weak.

In parallel, generative models have advanced shape synthesis through adversarial and diffusion-based approaches. Hybrid explicit–implicit methods such as EG3D [23] and GET3D [47] generate realistic 3D shapes, while diffusion-based models such as 3D-LDM [106] and NFD [123] have demonstrated strong diversity and fidelity. Yet, these methods typically generate shapes directly without enforcing topology preservation or dense correspondences, limiting their use in clinical contexts where such consistency is essential.

To overcome these challenges, we propose **Hybrid Neural Diffeomorphic Flow (HNDF)**, a new framework for shape representation and generation. HNDF builds on the template-deformation paradigm but introduces two key innovations. First, it encodes diffeomorphic deformations into axis-aligned triplane features, enabling both high representation capacity and local control via feature interpolation and a shared decoder. Second, it employs hybrid

supervision to enforce both global and local constraints, ensuring stable optimization and preservation of diffeomorphic properties. Furthermore, we extend HNDF to shape generation by representing per-object deformations as triplane features and synthesizing them with a 2D diffusion model. This enables deformation-based shape generation, ensuring that new shapes preserve topology and dense correspondences while achieving diversity.

The main contributions of this work are:

- We introduce HNDF, a diffeomorphic template-deformation framework that leverages triplane features to achieve high-fidelity shape representation with accurate dense correspondences.
- We propose a hybrid supervision strategy that combines global and local constraints, ensuring stable training while preserving the diffeomorphic property of deformations.
- We extend HNDF to shape generation, where novel shapes are synthesized by deforming a template through triplane features produced by a diffusion model, preserving topology and correspondences while enabling diversity.

4.2.2 Method

In this section, we present our Hybrid Neural Diffeomorphic Flow (HNDF) for shape representation and generation. Section 4.2.2.2 reviews our baseline method [129]. In Section 4.2.2.3, we introduce the utilization of triplane features, and the hybrid supervision for capturing local and global correspondences. Finally, in Section 4.2.2.4, we describe our proposed method for generating topology-preserving shapes.

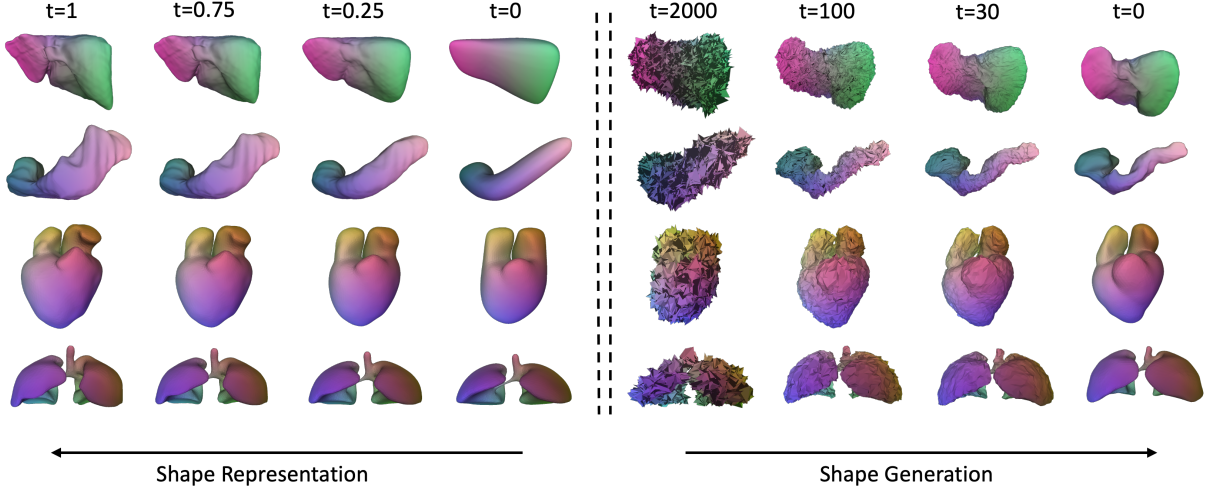


Figure 4.1: **Shape Representation** and **Shape Generation**. **Left** half presents the diffeomorphic deformation from the learned **template** ($t=0$) to instance shapes ($t=1$), with color highlighting the dense correspondence captured by triplane features. **Right** half presents the denoising process for shape generation. The shapes are generated as deformed templates and the 3D deformation is controlled by the generated triplane features from diffusion.

4.2.2.1 Preliminaries

Diffeomorphic Flow. A diffeomorphic flow is a smooth, invertible mapping that preserves topology during deformation. In 3D geometry, it is essential for establishing dense correspondences between shapes while ensuring anatomical consistency. The forward flow $\Phi(p, t) : \mathbb{R}^3 \times [0, 1] \rightarrow \mathbb{R}^3$ traces the trajectory of a point p over time under a velocity field $\mathbf{v}$:

$$\frac{\partial \Phi}{\partial t}(p, t) = \mathbf{v}(\Phi(p, t), t), \quad \Phi(p, 0) = p. \quad (4.1)$$

The inverse flow Ψ can be obtained with $-\mathbf{v}$:

$$\frac{\partial \Psi}{\partial t}(p, t) = -\mathbf{v}(\Psi(p, t), t), \quad \Psi(p, 0) = p, \quad (4.2)$$

guaranteeing that transformations between template and instance shapes remain diffeomor-

phic. Integration of these ODEs ensures continuous and topology-preserving deformations.

Diffusion Probabilistic Models (DPMs). DPMs [57] learn data distributions via a forward–reverse noise process. In the *forward process*, Gaussian noise is gradually added:

$$q(X_t | X_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t}X_{t-1}, \beta_t I), \quad (4.3)$$

with variance schedule $\{\beta_t\}$, such that $X_T \sim \mathcal{N}(0, I)$. The reverse process then learns to denoise step-by-step:

$$p_\theta(X_{t-1} | X_t) = \mathcal{N}(\mu_\theta(X_t, t), \Sigma_\theta(X_t, t)), \quad (4.4)$$

where a neural network ϵ_θ predicts the injected noise, enabling sampling of realistic shapes starting from Gaussian noise.

4.2.2.2 Review of NDF

Neural Diffeomorphic Flow (NDF) [129], inspired by DeepSDF [109] and DiT [153], represents a 3D shape S_i through a continuous signed distance field (SDF) $\mathcal{F}$. Given a 3D point p and a latent code $c_i \in \mathbb{R}^k$, $\mathcal{F}$ predicts the distance from p to the surface of S_i . Unlike DeepSDF, which directly encodes each shape, NDF introduces a diffeomorphic deformation field that maps points from an instance space to a canonical template space. Formally, the conditional SDF can be expressed as $\mathcal{F} = \mathcal{T} \circ \mathcal{D}$, where $\mathcal{D} : \mathbb{R}^3 \times \mathbb{R}^k \mapsto \mathbb{R}^3$ denotes the diffeomorphic mapping and $\mathcal{T}$ models the implicit template shape. This design enforces smooth and topology-preserving deformations across instances. However, relying on a single latent code c_i to parameterize the deformation limits its ability to accurately capture highly complex or diverse shapes.

4.2.2.3 Hybrid Shape Representation via Triplane

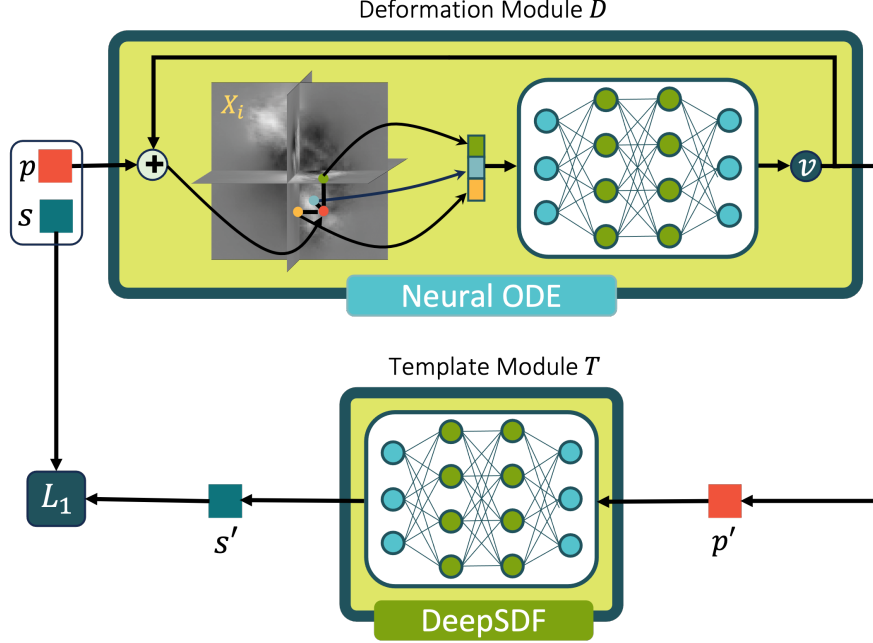


Figure 4.2: **Shape Representation.** The framework consists of a deformation module $\mathcal{D}$, a template module $\mathcal{T}$, and per-object triplane features X_i . Given a point p in the instance space, we compute its corresponding canonical point p' in the template space using Eq. 4.1. The template module then evaluates the signed distance value s' . Training optimizes the entire framework by minimizing the L_1 difference between s' and the ground-truth s , along with regularization terms.

Earlier methods that rely on a single latent vector to encode an entire shape or deformation space [109, 153, 129, 146] often fail to capture local details [22, 66, 132, 110, 30, 112]. Inspired by hybrid neural representations [23], we instead encode diffeomorphic deformations as three orthogonal, axis-aligned 2D feature planes (Fig. 4.2). This triplane representation provides fine-grained detail while remaining efficient.

For each shape S_i , we learn three feature planes

$$X_i = [F_{xy}^i, F_{xz}^i, F_{yz}^i],$$

each of resolution $L \times L \times C$. Given a query point p_i , its projections onto the planes yield

feature vectors retrieved by bilinear interpolation. Concatenating these features (instead of summing, as in [23]), we use an MLP decoder to predict a velocity vector v_i . The corresponding displacement d_i is computed by integrating v_i with a Runge–Kutta ODE solver [28], as in Eq. 4.1.

Training. An instance shape S_i is modeled as a deformed template $\mathcal{T} \circ \mathcal{D}_i$. For a query point p_i , the deformation module $\mathcal{D}$ outputs a displacement d_i , producing $p'_i = p_i + d_i$. The template module $\mathcal{T}$ then evaluates the signed distance value at p'_i . We jointly optimize $\mathcal{D}$, $\mathcal{T}$, and X_i across a dataset of S shapes. Unlike NDF [129], which requires multiple deformation modules, our approach uses a single $\mathcal{D}$, reducing both memory and computation while improving consistency.

The training loss combines reconstruction and regularization:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \quad (4.5)$$

The reconstruction loss enforces fidelity to ground-truth SDF values:

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^S \sum_{j=1}^N L_1(\mathcal{T} \circ \mathcal{D}_i(p_{i,j}), s_{i,j}), \quad (4.6)$$

where S is the number of instance shapes in the training set, N is the number of sampling points for each shape, $p_{i,j}$ is the j -th point on the i -th shape and $s_{i,j}$ is the corresponding ground truth SDF value.

The regularization term balances stability and smoothness:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{PW}} \mathcal{L}_{\text{PW}} + \lambda_{\text{L2}} \mathcal{L}_{\text{L2}} + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}. \quad (4.7)$$

Here, $\mathcal{L}_{\text{PW}}$ penalizes large point-wise deformations ($\sum_{i,j} \|\mathcal{T} \circ \mathcal{D}_i(p_{i,j}) - s_{i,j}\|_2$), $\mathcal{L}_{\text{L2}}$ regularizes feature magnitudes ($\|F_{xy}^i\|_2 + \|F_{yz}^i\|_2 + \|F_{xz}^i\|_2$), and $\mathcal{L}_{\text{TV}}$ enforces smoothness through

total variation regularization [116].

Hybrid Supervision for Inference Time Reconstruction

Unlike previous approaches that rely on a single latent vector for shape reconstruction [109, 129], our triplane-based representation introduces new challenges when reconstructing unseen shapes. During optimization, features interpolated from the triplane representation are updated locally for each query point p_i . Since the final deformation is obtained by integrating velocity vectors along entire trajectories, locally optimized features may become trapped in poor optima, leading to incorrect global correspondences. This often results in reconstruction artifacts and structural inconsistencies, as illustrated in Fig. 4.3.

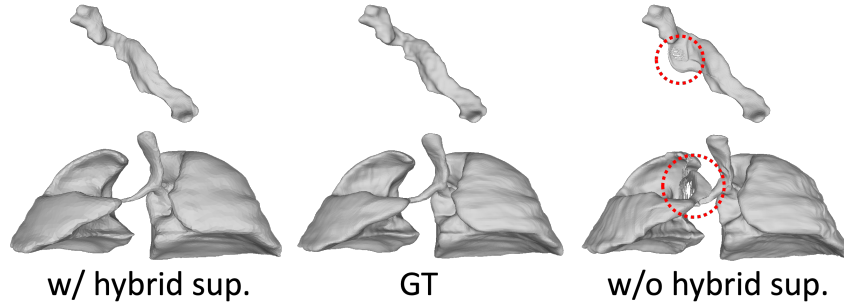


Figure 4.3: Reconstruction comparison. Left: our method with hybrid supervision. Middle: ground truth. Right: purely local supervision, which fails to recover correct global correspondence.

To address this, we propose a *hybrid supervision* strategy that enforces both local and global correspondence during optimization. In addition to randomly sampled points for local detail supervision, we downsample the entire $N \times N \times N$ coordinate grid with a predefined step size, incorporating these regularly sampled points to provide global guidance. This ensures that the deformation captures fine local structure while remaining globally consistent.

The reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{rec}}^{\text{grid}} + \lambda_{\text{random}} \mathcal{L}_{\text{rec}}^{\text{random}}, \quad (4.8)$$

where λ_{random} is initialized at 0 and gradually increased during optimization to balance global and local terms.

After obtaining the grid-structured deformation Φ , we further apply two regularization terms to preserve diffeomorphism and structural plausibility:

To enforce orientation consistency, we penalize negative Jacobian determinants:

$$\mathcal{L}_{\text{Jdet}} = \frac{1}{N} \sum_p \text{ReLU}(-|J_{\Phi}(p)|), \quad (4.9)$$

where the Jacobian of Φ at point p is

$$J_{\Phi}(p) = \begin{bmatrix} \frac{\partial \Phi_x(p)}{\partial x} & \frac{\partial \Phi_x(p)}{\partial y} & \frac{\partial \Phi_x(p)}{\partial z} \\ \frac{\partial \Phi_y(p)}{\partial x} & \frac{\partial \Phi_y(p)}{\partial y} & \frac{\partial \Phi_y(p)}{\partial z} \\ \frac{\partial \Phi_z(p)}{\partial x} & \frac{\partial \Phi_z(p)}{\partial y} & \frac{\partial \Phi_z(p)}{\partial z} \end{bmatrix}. \quad (4.10)$$

To discourage extreme distortions, we penalize sharp gradients in the deformation field:

$$\mathcal{L}_{\text{def}} = \sum_p \|\nabla \Phi(p)\|^2. \quad (4.11)$$

By combining local and global supervision with these regularization terms, our method achieves reconstructions that preserve fine anatomical details while maintaining global structural consistency and diffeomorphism.

Point Correspondence and Shape Registration

During inference, our framework leverages the learned template shape and the diffeomorphic deformations encoded by triplane features to establish dense point correspondence and shape registration. For each template point p_t , we apply the inverse diffeomorphic flow Ψ (Eq. 4.2)

to map p_t to its corresponding locations p_i and p_j on instance shapes S_i and S_j , respectively, using their triplane features X_i and X_j . This mechanism enables consistent correspondence across instances, supporting downstream tasks such as shape comparison, synthesis, and texture transfer.

4.2.2.4 Topology-preserving Shape Generation

Instead of directly generating shapes from scratch, we generate new shapes by deforming the template shape with synthesized diffeomorphic deformations. This preserves topological consistency while allowing diverse structural variations.

Training a Diffusion Model

After training the deformation module $\mathcal{D}$ and the template shape $\mathcal{T}$ (Sec. 4.2.2.3), and optimizing triplane features with hybrid supervision (Sec. 4.2.2.3), we obtain per-shape triplane features for the dataset. These features, denoted as

$$X \in \mathbb{R}^{N \times (L \times L \times 3C)},$$

serve as training data for our generative model, where N is the number of shapes, L the spatial resolution, and C the channel size per 2D plane $(F_{xy}^i, F_{xz}^i, F_{yz}^i)$.

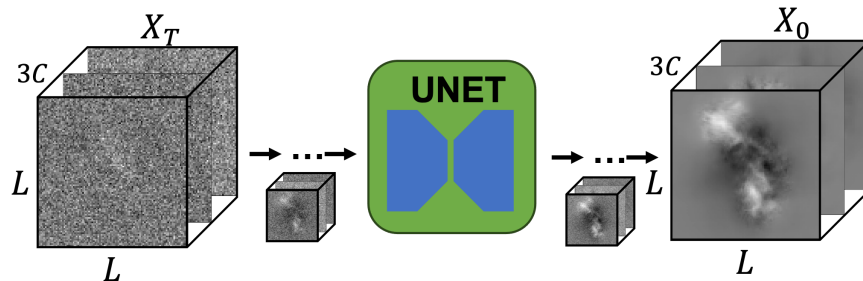


Figure 4.4: **Shape Generation.** Triplane features represented as multi-channel images. We train a 2D diffusion model to generate triplane features, which encode diffeomorphic deformations that warp the template into new shapes.

We concatenate the three 2D feature planes into a multi-channel representation and exploit the strong generative capacity of 2D diffusion models. Following Sec. 4.2.2.1, we train the diffusion model to predict added noise during the reverse process by minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{X_0 \sim q(X), \epsilon \sim \mathcal{N}(0, I), t} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right], \quad (4.12)$$

where ϵ_θ denotes the predicted noise and θ the model parameters.

New Shape Generation

At inference, new shapes are generated by sampling triplane features and applying them as diffeomorphic deformations to the template. Following [57], we begin with Gaussian noise $X_T \sim \mathcal{N}(0, I)$ and iteratively denoise over T steps:

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(X_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(X_t, t) \right) + \sigma_t \epsilon, \quad (4.13)$$

where $\epsilon \sim \mathcal{N}(0, I)$ for $t > 1$ and $\epsilon = 0$ otherwise.

The resulting triplane feature is split into three orthogonal planes $(F_{xy}^i, F_{xz}^i, F_{yz}^i)$ and interpreted as a diffeomorphic deformation field. Using the ODE-based integration in Eq. 4.2, each template point is displaced to its destination in the instance space. The generated shape thus emerges as a deformed template, preserving topology while capturing diverse geometric variations.

4.2.3 Experiments

We evaluate the proposed Hybrid Neural Diffeomorphic Flow (HNDF) on both shape **representation** and **generation** tasks.

Datasets: For shape representation, we adopt the same medical datasets used in [129]: Pan-

creas CT, In-house Liver, In-house Lung, and Multi-Modality Whole Heart Segmentation. These datasets exhibit consistent topology while containing notable inter-subject variations, making them suitable benchmarks. For shape generation, we use liver and pancreas shapes from the Abdomen1k dataset [91], and heart and lung from [129]. Additional details on data sources and preprocessing are provided in the supplementary material.

Shape Representation Evaluation: We assess shape representation with two experiments. (1) We evaluate the expressive power of triplane representation and the effectiveness of our hybrid supervision. Metrics include Chamfer Distance (CD) and Normal Consistency (NC). (2) We evaluate point correspondence and registration accuracy, introducing self-intersection (SI) as an additional metric for geometric fidelity.

Shape Generation Evaluation: For generation, we adopt an adapted version of the Fréchet Inception Distance (FID) following [123]. Specifically, FID is computed on rendered shading images of generated meshes, which better reflects perceptual quality [152]. FID is averaged across 20 views:

$$\text{FID} = \frac{1}{20} \sum_{i=1}^{20} [\|\mu_g^i - \mu_r^i\|^2 + \text{Tr}(\Sigma_g^i + \Sigma_r^i - 2(\Sigma_r^i \Sigma_g^i)^{1/2})], \quad (4.14)$$

where (μ_g^i, Σ_g^i) and (μ_r^i, Σ_r^i) are the statistics of generated and real samples, respectively. Additionally, we report precision and recall scores [118], where precision measures visual fidelity and recall reflects shape diversity.

Baseline Methods: For shape representation, we compare HNDF with DIT [153], DIF-Net [40], and NDF [129], which all represent shapes as deformations of a template. AtlasNet [52] (explicit mesh parameterization), DeepSDF [109], and NFD [123] (direct shape modeling) are also included.

For shape generation, we compare with DeepSDF [109] and NDF [129] using Gaussian sam-

pling or PCA over latent vectors, as well as more recent generative models: PVD [154] (point-cloud-based), 3D-LDM [106] (neural field), and NFD [123]. Unlike our method, these approaches do not explicitly preserve topology.

4.2.3.1 Shape Representation

Model/Data	Reconstruction								Registration											
	CD Mean(↓)				NC Mean(↑)				CD Mean(↓)				NC Mean(↑)				SI Mean(↓)			
	Pancreas	Liver	Heart	Lung	Pancreas	Liver	Heart	Lung	Pancreas	Liver	Heart	Lung	Pancreas	Liver	Heart	Lung	Pancreas	Liver	Heart	Lung
DeepSDF[109]	0.711	0.539	0.951	0.669	0.898	0.866	0.913	0.928	-	-	-	-	-	-	-	-	-	-	-	-
NFD [123]	0.080	<u>0.118</u>	<u>0.287</u>	0.255	0.982	0.898	<u>0.947</u>	0.947	-	-	-	-	-	-	-	-	-	-	-	-
AtlasNet[52]	8.08	3.46	7.55	5.01	0.703	0.823	0.808	0.824	8.08	3.46	7.55	5.01	0.703	0.823	0.808	0.824	5860	29.5	0	13.8
DIT[153]	0.63	0.509	1.05	0.712	0.903	0.87	0.919	0.934	0.677	0.528	1.07	0.736	0.893	0.868	0.918	0.931	346	11.8	0	2.06
DIF-Net[40]	4.18	1.58	2.23	1.86	0.756	0.832	0.838	0.882	10.5	2.06	2.42	1.94	0.694	0.832	0.838	0.881	2560	4.61	1090	786
NDF[129]	0.512	0.476	0.993	<u>0.643</u>	0.917	0.873	0.923	0.937	<u>0.518</u>	<u>0.49</u>	<u>1.02</u>	<u>0.67</u>	<u>0.916</u>	<u>0.873</u>	<u>0.923</u>	<u>0.936</u>	0	2	0	0
Ours	<u>0.082</u>	0.116	0.277	0.255	<u>0.961</u>	<u>0.885</u>	0.948	<u>0.945</u>	0.099	0.125	0.306	0.304	0.946	0.882	0.936	0.939	<u>15</u>	<u>8</u>	<u>6</u>	0

Table 4.1: Shape Reconstruction and Shape Registration results on Unseen Shapes. The chamfer distance results shown above are multiplied by 10^3 . **DeepSDF** and **NFD** cannot model the deformation between shapes, therefore, they are unable to conduct the shape registration task.

We follow the setting of [129] to evaluate both **representation** on training data and **reconstruction** on unseen data. For each instance point p , Eq. 4.1 maps it to the template, where the trained template module returns its signed distance. After querying all grid points, marching cubes [89] is applied to extract meshes.

In the representation comparison, per-object latent features are directly optimized, while in reconstruction, features are optimized with network weights fixed. Figure 4.5 shows reconstruction results. Detailed representation results are provided in the supplementary.

DIF-Net achieves strong training-set performance but fails to generalize to unseen shapes, suggesting overfitting. Our method and NFD both benefit from the enhanced expressiveness of triplane features, but HNDF achieves superior reconstruction even with a single deformation module, surpassing NDF with four modules. An ablation study on regularization

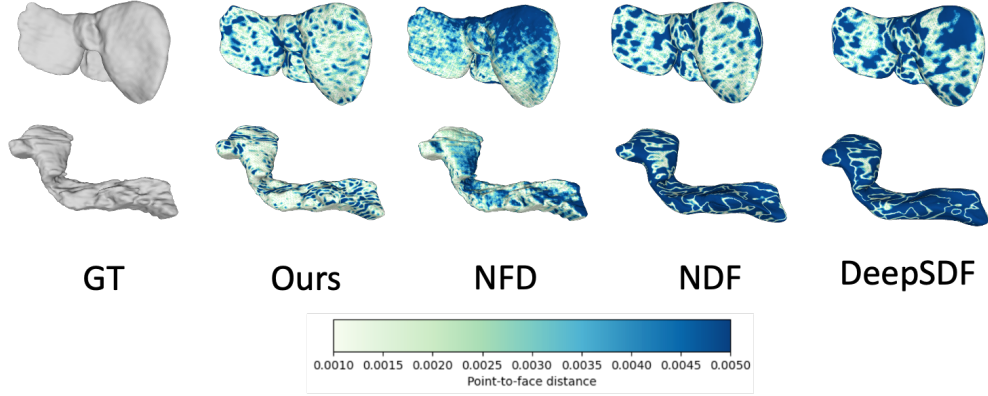


Figure 4.5: Reconstruction results on unseen data.

(Tab. 4.3) highlights the effectiveness of our hybrid supervision for accurate reconstruction.

4.2.3.2 Point Correspondence and Shape Registration

DeepSDF and NFD lack explicit correspondence modeling; thus, we compare only deformation-based methods in Tab. 4.1. Using Eq. 4.2, template points are diffeomorphically mapped to their instance correspondences. This guarantees consistent topology and connectivity. Examples of template-to-instance deformations are shown in Fig. 1 (left).

Model/Data	FID Mean($\downarrow$)				Prec. Mean($\uparrow$)				Recall Mean($\uparrow$)			
	Pancreas	Liver	Heart	Lung	Pancreas	Liver	Heart	Lung	Pancreas	Liver	Heart	Lung
DeepSDF $\triangle$	99.46	93.74	85.76	83.63	0.810	<u>0.858</u>	0.773	0.581	0.078	0.089	0.501	0.630
DeepSDF $\star$	80.03	85.64	72.22	69.46	0.729	0.810	0.771	0.670	0.430	0.534	0.735	0.664
NDF $\triangle$	69.66	60.50	65.63	60.28	0.797	0.714	0.762	0.765	0.508	<u>0.593</u>	0.615	0.695
NDF $\star$	<u>69.66</u>	<u>66.45</u>	64.21	<u>58.27</u>	<u>0.844</u>	0.821	0.804	0.813	0.505	0.571	0.842	0.794
PVD	89.26	86.32	82.44	87.12	0.760	0.821	0.764	0.732	0.420	0.466	0.671	0.712
3D-LDM	78.64	79.58	73.25	76.42	0.782	0.824	0.819	0.803	0.470	0.554	0.813	0.821
NFD	72.83	74.24	<u>61.58</u>	64.34	0.812	0.831	<u>0.832</u>	0.821	<u>0.523</u>	0.560	0.972	0.937
Ours	52.01	48.54	54.5	41.3	0.992	0.994	0.984	0.951	0.661	0.613	<u>0.893</u>	<u>0.914</u>

Table 4.2: Shape generation results. Our method achieve better performance according to the FID, precision and recall. $\triangle$ denotes sampling from Gaussian distribution while $\star$ denotes sampling from PCA.

We evaluate registration by comparing the deformed template to ground-truth shapes, and report self-intersection (SI) to assess topological fidelity. Following [129], all template meshes

are remeshed to 5000 vertices for fairness. Results (Tab. 4.1) show that HNDF achieves more accurate registration and consistent correspondence, with negligible self-intersection relative to mesh size.

4.2.3.3 Shape Generation

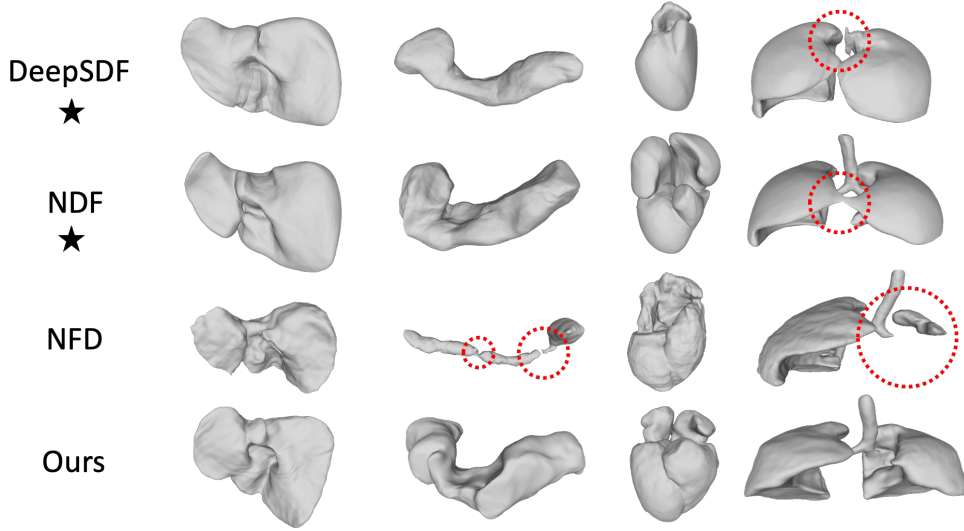


Figure 4.6: Generated 3D shapes across different methods.

Table 4.2 reports quantitative results. DeepSDF and NDF, when sampled via Gaussian or PCA, produce overly smooth shapes. PVD generates variable but coarse point clouds, while 3D-LDM inherits DeepSDF’s smoothing limitations. NFD captures shape diversity but often fails to preserve topology, producing disconnected components (Fig. 4.6). In contrast, HNDF generates diffeomorphic deformations encoded in triplane features, yielding shapes that preserve topology while achieving both high fidelity and variability.

4.2.3.4 Ablation Study

Supervision: Tab. 4.3 shows that global supervision is critical for avoiding local minima in reconstruction. Adding mesh-level supervision gives minor gains at the cost of higher

memory and computation, so we adopt global supervision only.

Model/Data	CD Mean($\downarrow$)		NC Mean($\uparrow$)	
	Pancreas	Liver	Pancreas	Liver
Ours	0.082	0.116	0.961	0.885
Ours - Global Sup.	0.264	0.368	0.932	0.877
Ours + Mesh Sup.	0.082	0.112	0.960	0.886

Table 4.3: Shape Reconstruction with various supervision.

Feature Representation: We also tested voxel-grid features as an alternative to triplanes (Tab. 4.4). While achieving comparable accuracy, voxel features incur significantly higher memory and computation. Triplane features provide a more efficient and scalable representation.

Model/Data	CD Mean($\downarrow$)		NC Mean($\uparrow$)	
	Pancreas	Liver	Pancreas	Liver
Vector	0.512	0.476	0.917	0.873
Triplane	0.082	0.116	0.961	0.885
Voxel	0.146	0.112	0.957	0.885

Table 4.4: Shape Reconstruction with various feature representations.

4.2.4 Conclusion

We introduced Hybrid Neural Diffeomorphic Flow (HNDF), a framework for topology-preserving shape representation and generation. By combining a triplane-based feature representation with diffeomorphic deformation modeling, HNDF enables accurate shape reconstruction, consistent point correspondence, and high-quality shape generation. Extensive experiments on multiple medical datasets demonstrated that HNDF achieves state-of-the-art performance in representation and registration, while its diffusion-based generative model produces diverse shapes that preserve underlying topology—overcoming limitations of prior methods that often yield oversmoothed or topologically inconsistent results.

Beyond its empirical advantages, HNDF highlights the value of incorporating topology-preserving deformations into generative modeling. In the future, we plan to extend HNDF to larger and more diverse datasets, explore conditional and controllable generation, and investigate its integration with downstream clinical tasks such as disease progression modeling and shape-based diagnostics. These directions may further broaden the applicability of HNDF in both medical imaging and general 3D shape analysis.

Chapter 5

Data Generation

5.1 Introduction

Data is the driving force of modern medical imaging research. Deep learning models achieve strong performance by leveraging large and diverse datasets, yet in medicine such resources are scarce. Unlike natural image collections that can be amassed at scale, medical imaging is constrained by **privacy regulations**, **ethical considerations**, and the **high cost of expert annotations**. Some modalities are inherently difficult to obtain, requiring specialized hardware or invasive procedures. These factors make data scarcity a central challenge in building robust and generalizable models.

Generative modeling offers a powerful path forward. By synthesizing realistic medical images, generative methods can alleviate data scarcity, expand dataset diversity, and even enable the creation of modalities that are rarely available in practice. Beyond enlarging datasets, they allow the exploration of hypothetical scenarios, the simulation of rare pathologies, and the generation of paired multi-modal data that would otherwise be infeasible. In this sense, medical data generation is not merely a convenience—it is a transformative capability that

reshapes both research pipelines and clinical applications.

This dissertation presents two complementary contributions to medical data generation. **XOCT** demonstrates the potential of **cross-modality synthesis**, translating between structural and functional imaging to enrich information available from routine scans. In contrast, **MedGen3D** advances the frontier of **volumetric synthesis**, providing anatomically faithful 3D datasets at unprecedented scale. Together, these contributions highlight how generative models can fill gaps in medical imaging, enrich datasets with greater variability, and enable more generalizable and clinically impactful analysis methods.

5.2 Enhancing OCT to OCTA Translation

5.2.1 Introduction

Optical Coherence Tomography Angiography (OCTA) has revolutionized retinal imaging by providing high-resolution, dye-free visualization of retinal and choroidal microvasculature. As illustrated in Fig. 5.1, en-face OCTA projections offer intuitive top-down assessments of vascular networks, playing a critical role in the early detection and monitoring of diseases such as diabetic retinopathy, age-related macular degeneration (AMD), and glaucoma [70]. Despite its clinical importance, the acquisition of high-quality OCTA images remains challenging due to motion artifacts and the reliance on costly OCTA-enabled devices [60, 37, 125, 3]. These limitations have spurred increasing interest in *OCT-to-OCTA translation*, where structural OCT volumes are leveraged to synthesize angiographic images, thereby broadening accessibility while reducing acquisition burdens.

Early studies addressed OCT-to-OCTA translation using 2D strategies, processing B-scans individually with encoder-decoder or adversarial frameworks [77, 149, 82, 85, 67, 86]. While

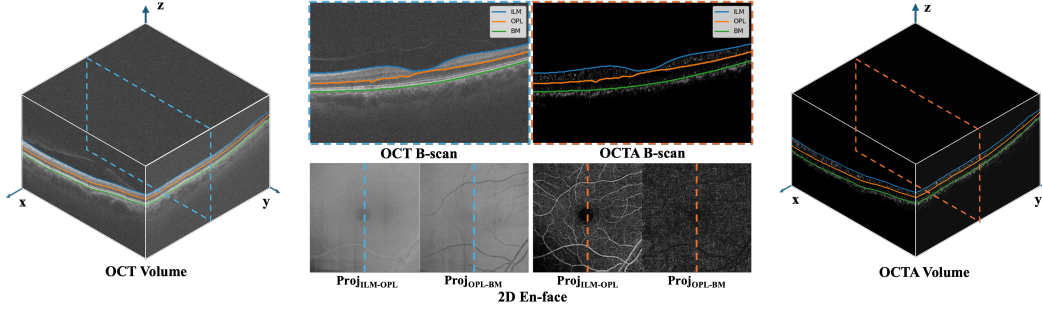


Figure 5.1: OCT/OCTA volumes with retinal layer segmentation and en-face projections. Differences between $\text{Proj}_{\text{ILM-OPL}}$ and $\text{Proj}_{\text{OPL-BM}}$ highlight the heterogeneous imaging properties arising from distinct cellular and vascular distributions across retinal layers.

these methods demonstrated feasibility, they failed to preserve three-dimensional vascular continuity and often yielded fragmented reconstructions that compromise clinical interpretability. Projection-based approaches [108] sought to alleviate this issue by mapping OCT projections into en-face OCTA angiograms. However, by collapsing volumetric data into a single 2D plane, they inevitably discarded depth information and obscured subtle microvascular structures critical for accurate analysis. More recently, volumetric translation models have been proposed to directly map 3D OCT volumes into 3D OCTA reconstructions [61, 83]. Although these methods improved vascular continuity, their reliance on standard convolutional operators limited their adaptability to the heterogeneous imaging properties of different retinal layers, resulting in incomplete recovery of the fine-scale vascular details that characterize OCTA.

In this work, we propose **XOCT**, a framework that addresses these challenges by integrating cross-dimensional constraints with multi-scale feature modeling. At its core, XOCT introduces a *Cross-Dimensional Supervision (CDS)* module, which leverages the distinct structural properties of individual retinal layers by enforcing alignment between volumetric predictions and segmentation-weighted en-face projections. This design promotes intra-layer consistency, preserves vessel coherence, and enhances recovery of subtle microvascular details often overlooked in conventional frameworks. Complementing this, the *Multi-Scale Fea-*

ture Fusion (MSFF) module refines vessel delineation by combining isotropic, anisotropic, and depth-wise large-kernel convolutions with adaptive channel reweighting, enabling the network to balance fine local details with global vascular connectivity while maintaining computational efficiency.

By combining cross-dimensional supervision with multi-scale fusion, XOCT effectively overcomes the limitations of prior approaches and provides a robust solution for OCT-to-OCTA translation. Extensive experiments on the OCTA-500 dataset [81] demonstrate that XOCT achieves superior vascular clarity, continuity, and fidelity in both volumetric OCTA reconstructions and layer-wise en-face projections, underscoring its direct clinical relevance for retinal disease assessment.

5.2.2 Method

We propose **XOCT** (Fig. 5.2), a novel deep learning framework for OCT-to-OCTA translation that extends a 3D encoder-decoder backbone with two key innovations: *Cross-Dimensional Supervision (CDS)* and a *Multi-Scale Feature Fusion (MSFF)* module. Given a volumetric OCT scan $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$, XOCT produces a reconstructed OCTA volume $\hat{\mathbf{Y}} \in \mathbb{R}^{D \times H \times W}$. The model is trained end-to-end with a composite objective that integrates volumetric and projection-based constraints, ensuring both global structural consistency and accurate vessel delineation.

5.2.2.1 Cross-Dimensional Supervision

Conventional volumetric supervision treats the OCTA volume as homogeneous, neglecting the retina’s intrinsic heterogeneity across layers. In reality, the retina is composed of multiple layers with distinct tissue compositions, cellular structures, and vascular distributions

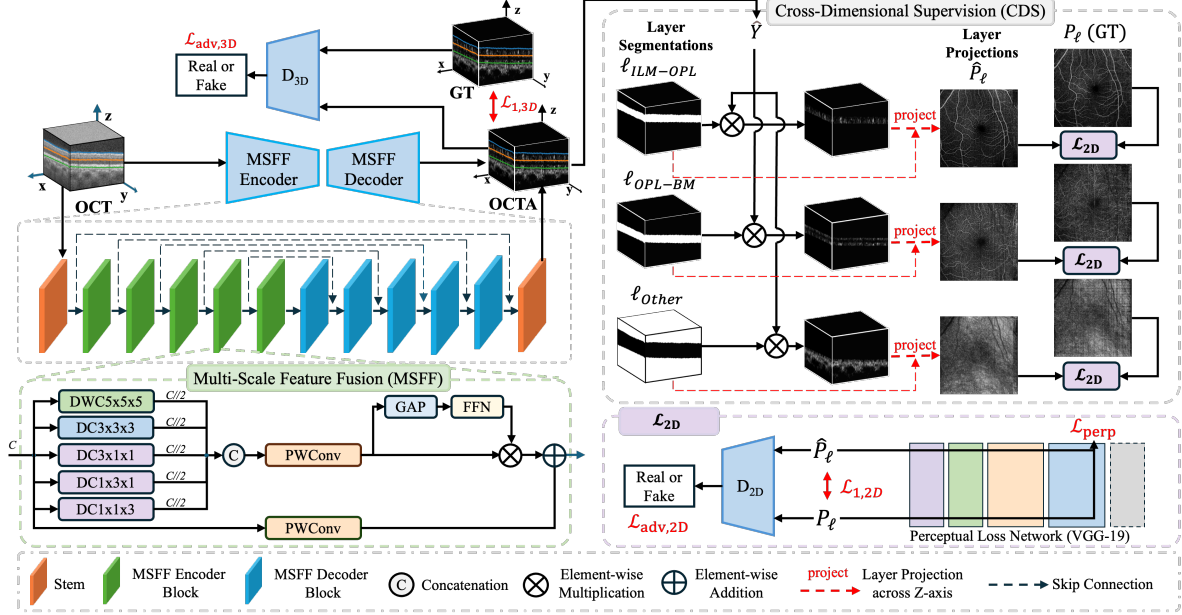


Figure 5.2: **Overview of XOCT.** The framework combines volumetric OCT-to-OCTA translation with Cross-Dimensional Supervision (CDS) and Multi-Scale Feature Fusion (MSFF).

[19]. To address this, XOCT introduces **Cross-Dimensional Supervision (CDS)**, which augments 3D supervision with layer-aware 2D projection guidance.

Given a retinal layer segmentation map $\mathbf{S} \in \mathbb{R}^{D \times H \times W}$, we compute 2D layer-specific projections $\mathbf{P}_l \in \mathbb{R}^{H \times W}$ for each layer l . For a predicted OCTA volume $\hat{\mathbf{Y}}$, the layer-specific projection is obtained as the segmentation-weighted average of voxel intensities along the depth (**z-axis**):

$$\hat{\mathbf{P}}_l = \frac{\sum_z \hat{\mathbf{Y}} \odot \mathbf{S}_l}{\sum_z \mathbf{S}_l}, \quad (5.1)$$

where $\odot$ denotes element-wise multiplication.

The predicted projections $\hat{\mathbf{P}}_l$ are compared against the ground-truth projections $\mathbf{P}_l$ using a

hybrid objective:

$$L_{2D} = \sum_l \left(\alpha_{2D} L_1(\widehat{\mathbf{P}}_l, \mathbf{P}_l) + \beta_{2D} L_{adv}(\widehat{\mathbf{P}}_l, \mathbf{P}_l) + \gamma_{2D} L_{perp}(\widehat{\mathbf{P}}_l, \mathbf{P}_l) \right), \quad (5.2)$$

where L_1 enforces pixel-level fidelity, L_{adv} promotes visual realism through adversarial learning, and L_{perp} —a perceptual loss derived from a pretrained VGG19 network—captures high-level vascular similarity.

By supervising projections at each retinal layer, CDS encourages the model to disentangle layer-specific structures, preserve intra-layer coherence, and accurately recover fine vascular details. The combined volumetric–projection objective is expressed as:

$$L = L_{3D} + L_{2D}, \quad (5.3)$$

which enforces structural fidelity both in 3D space and across layer-wise projections.

5.2.2.2 Multi-Scale Feature Fusion

Fine vascular reconstruction requires both localized detail and long-range connectivity. To achieve this, we design the **Multi-Scale Feature Fusion (MSFF)** module, which aggregates features across multiple spatial scales.

MSFF employs three types of convolutional operators: - **Isotropic 3D kernels** ($3 \times 3 \times 3$) capture balanced local spatial context. - **Anisotropic kernels** ($3 \times 1 \times 1$, $1 \times 3 \times 1$, $1 \times 1 \times 3$) emphasize elongated vascular patterns along specific axes. - **Depth-wise large-kernel convolutions** ($5 \times 5 \times 5$) extend the receptive field, enabling recovery of broader vessel connectivity.

Although larger kernels (e.g., $7 \times 7 \times 7$) were explored, they provided marginal performance

improvements while incurring cubic increases in computation. The $5 \times 5 \times 5$ configuration was therefore selected as the best trade-off.

To improve efficiency, we halve the output channels in each block, reducing parameters while retaining accuracy. Multi-scale responses are fused with point-wise convolutions and adaptive channel reweighting, which emphasizes critical vascular signals. A residual shortcut from the input preserves fine structural details and stabilizes optimization. Collectively, this design enhances vessel continuity and improves reconstruction of delicate capillaries.

5.2.2.3 Overall Framework

XOCT is trained in a generative adversarial learning paradigm [64], jointly optimizing the 3D volumetric generator G_{3D} , volumetric discriminator D_{3D} , and a set of 2D projection discriminators $\{D_{2D}^l\}$ for different retinal layers l .

The volumetric loss is defined as:

$$L_{3D} = \alpha_{3D} L_1(\hat{\mathbf{Y}}, \mathbf{Y}) + \beta_{3D} L_{adv}(\hat{\mathbf{Y}}, \mathbf{Y}), \quad (5.4)$$

where L_1 enforces voxel-wise reconstruction accuracy and L_{adv} denotes the adversarial objective:

$$L_{adv} = \mathbb{E}_{\mathbf{Y} \sim p} [\log D(\mathbf{Y})] + \mathbb{E}_{\mathbf{X} \sim p} [\log(1 - D(G(\mathbf{X})))]. \quad (5.5)$$

For 2D supervision, projections are derived directly from predicted OCTA volumes using Eq. 5.1, eliminating the need for a separate 2D generator. These projections are compared against ground truth using adversarial, pixel-wise, and perceptual losses. Importantly, retinal segmentation maps are used only during training to generate layer-specific projections; they

are not required at inference time.

Together, CDS and MSFF enable XOCT to capture layer-specific vascular properties, preserve volumetric continuity, and achieve high-fidelity OCT-to-OCTA translation.

5.2.3 Experiments

5.2.3.1 Datasets and Experimental Setup

Dataset: We conduct experiments on the publicly available **OCTA-500** dataset [81], which contains 500 paired 3D OCT–OCTA volumes with corresponding retinal layer segmentation annotations. The dataset is divided into two subsets: **OCTA-3M** and **OCTA-6M**. The **OCTA-3M** subset consists of 200 scans with a field-of-view of $3\text{ mm} \times 3\text{ mm} \times 2\text{ mm}$ and volume dimensions of $304 \times 304 \times 640$ voxels, partitioned into 140 training, 20 validation, and 40 test scans. The **OCTA-6M** subset contains 300 scans with a larger field-of-view of $6\text{ mm} \times 6\text{ mm} \times 2\text{ mm}$ and volume dimensions of $400 \times 400 \times 640$ voxels, divided into 200 training, 30 validation, and 70 test scans.

Implementation Details: XOCT is implemented as a modified 3D Pix2Pix framework and trained for 300 epochs using the Adam optimizer with a learning rate of $1\text{e-}4$ and a batch size of 1. The adversarial loss weight is fixed at 1, while a grid search determines the optimal weights for reconstruction and perceptual terms: $\lambda_{L_1} = 10$ and $\lambda_{perp} = 1$, respectively.

Evaluation Setup: We evaluate XOCT on both 2D en-face projections and 3D volumetric reconstructions. For en-face evaluation, we adopt projections not used in training, ensuring an unbiased assessment. Specifically, **Proj_{full}** spans the full retinal depth (ILM to BM), capturing the complete vasculature, while **Proj_{mean}** computes the mean projection across the entire z -axis, offering a global representation of vascular density. Quantitative evaluation employs Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR) [49], Structural

Similarity Index (SSIM) [138], and Perceptual Discrepancy (Perp.) [68].

We compare XOCT against leading baselines: 2D-based methods (BBDM [79], Pix2Pix [64], MultiGAN [108]) and 3D-based methods (Pix2Pix3D [39], TransPro [83], and BBDM3D* [33]). BBDM3D* represents a recent 2.5D diffusion strategy that mitigates GPU memory demands by stacking neighboring slices as channels, enabling a fair diffusion-based comparison [33, 84].

5.2.3.2 Experimental Results

Table 5.1: OCT to OCTA translation results on the **OCTA-3M** and **OCTA-6M** datasets. $\downarrow$ indicates lower is better, $\uparrow$ indicates higher is better.

OCTA-3M					$\text{Proj}_{\text{full}}$				$\text{Proj}_{\text{mean}}$				3D			
Method	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$
BBDM [79]	22.86	18.84	0.420	0.741	25.33	18.14	0.384	0.727	-	-	-	-	-	-	-	-
Pix2Pix [64]	24.25	18.16	0.380	0.726	27.18	17.47	0.340	0.708	-	-	-	-	-	-	-	-
MultiGAN [108]	22.54	18.87	0.417	0.722	24.73	18.32	0.382	0.708	-	-	-	-	-	-	-	-
BBDM3D* [33]	21.04	19.54	0.460	0.703	22.51	19.27	0.457	0.675	2.69	30.71	0.883	0.202	2.88	29.45	0.885	0.198
Pix2Pix3D [39]	19.87	19.91	0.556	0.656	21.61	19.47	0.541	0.620	3.30 [†]	28.46	0.866	0.203	2.78	30.08	0.893	0.184
TransPro [83]	19.54	20.14	0.580	0.608	20.93	19.83	0.573	0.582								
XOCT	19.22	20.21	0.608	0.573	20.54	19.90	0.596	0.558								

OCTA-6M					$\text{Proj}_{\text{full}}$				$\text{Proj}_{\text{mean}}$				3D			
Method	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$	MAE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	Perp. $\downarrow$
BBDM [79]	22.94	18.68	0.367	0.729	25.73	17.92	0.346	0.714	-	-	-	-	-	-	-	-
Pix2Pix [64]	22.87	18.67	0.347	0.703	26.43	17.76	0.303	0.695	-	-	-	-	-	-	-	-
MultiGAN [108]	21.22	19.32	0.374	0.713	24.69	18.43	0.332	0.714	-	-	-	-	-	-	-	-
BBDM3D* [33]	17.90	20.81	0.433	0.702	18.29	20.81	0.436	0.686	2.48	31.44	0.889	0.146	2.34	31.23	0.903	0.136
Pix2Pix3D [39]	17.39	21.06	0.494	0.641	17.21	21.33	0.510	0.609	2.72 [†]	29.92	0.887	0.155	2.34	31.26	0.905	0.131
TransPro [83]	17.40	21.07	0.524	0.591	17.42	21.24	0.539	0.570								
XOCT	16.65	21.46	0.568	0.541	16.21	21.85	0.574	0.537								

[†]The reported MAE in TransPro[83] appears higher (e.g., 0.078 vs 3.30/255=0.013) due to a `uint8` overflow error (e.g., `abs(1-2) = 255`). All reported MAE values use `float32` for a fair comparison.

Table 5.1 reports quantitative results on OCTA-3M and OCTA-6M across 2D projections (**Proj_{full}**, **Proj_{mean}**) and 3D volumes. Projection-based methods (BBDM, Pix2Pix, MultiGAN) perform poorly, as collapsing volumetric information into 2D leads to loss of fine vascular continuity and structural depth, resulting in higher MAE/Perp. and lower PSNR/SSIM. While BBDM3D* reduces volumetric error by leveraging stacked slices, it still fails to model full 3D dependencies. TransPro improves over 2D models through 3D supervision, but its reliance on pretrained 2D components limits volumetric fidelity.

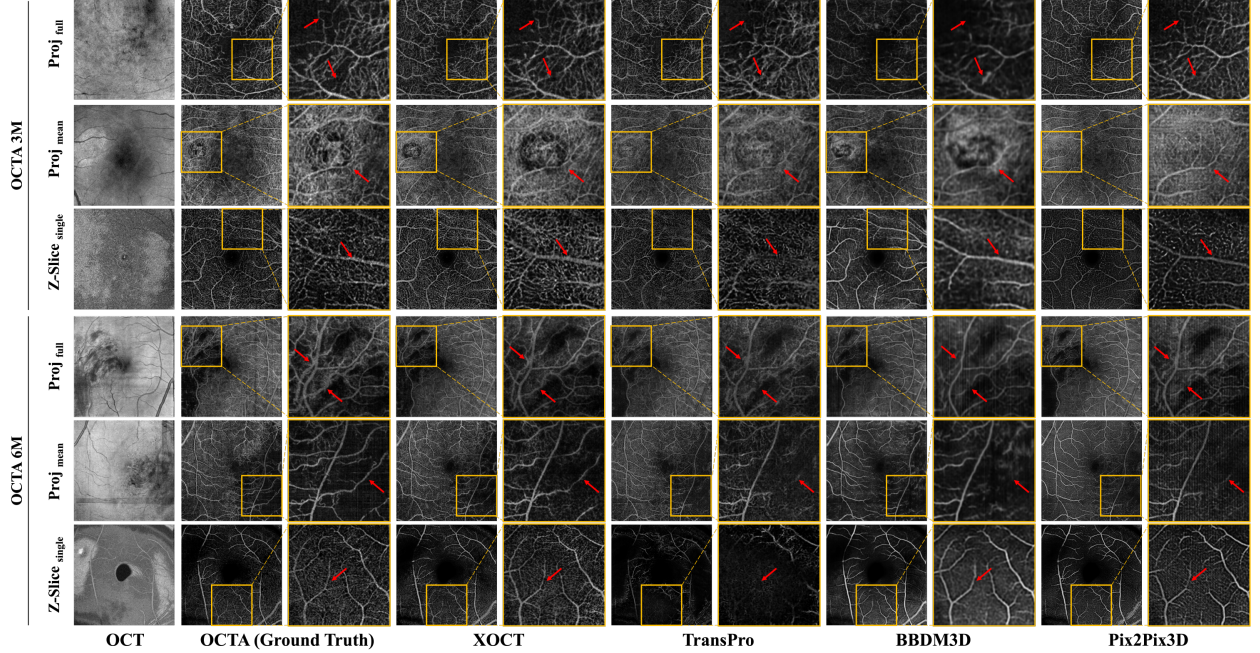


Figure 5.3: Qualitative comparison of OCT-to-OCTA translation methods across en-face projections (**Proj_{full}**, **Proj_{mean}**) and a representative 3D z -slice. Red arrows indicate regions where XOCT better preserves vascular continuity, reconstructs vessels missed by other models, and avoids over-smoothing of subtle vessel dropouts.

Although BBDM3D* achieves favorable MAE and PSNR values, qualitative inspection (Fig. 5.3) reveals substantial blurring in small vessels. This suggests that BBDM3D* prioritizes coarse structural consistency while neglecting high-frequency vascular detail. In contrast, XOCT produces sharper, anatomically faithful reconstructions, recovering microvascular structures that are often smoothed out in diffusion-based approaches. While intensity-sensitive metrics such as MAE may undervalue XOCT’s improvements, its superior vessel clarity and continuity are crucial for clinical interpretability.

XOCT consistently preserves vessel connectivity, accurately reconstructs small vessels, and captures subtle dropout patterns (Fig. 5.3). These dropout regions are clinically significant, as they indicate microvascular abnormalities linked to early diabetic retinopathy [115]. Competing models, such as Pix2Pix3D and TransPro, often “hallucinate” vessels in missing regions, potentially masking perfusion deficits. By maintaining both vessel sharpness and dropout fidelity, XOCT provides reconstructions that are not only visually superior but also

more clinically meaningful.

5.2.3.3 Ablation Study

Table 5.2: Ablation Study on Proposed Modules. ↓: lower is better, ↑: higher is better.

OCTA-3M	# Param.	Proj _{full}		Proj _{ILM-OPL}		Proj _{OPL-BM}		Proj _{Jmean}		3D	
Method		SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓
Pix2Pix3D	64.6M	0.556	0.656	0.509	0.720	0.479	0.668	0.541	0.620	0.885	0.198
+CDS	64.6M	0.600	0.581	0.563	0.606	0.522	0.616	0.590	0.563	0.889	0.195
+MSFF	52.7M	0.589	0.651	0.552	0.707	0.523	0.677	0.577	0.614	0.893	0.184
XOCT	52.7M	0.608	0.573	0.577	0.592	0.533	0.609	0.596	0.558	0.893	0.184

OCTA-6M	# Param.	Proj _{full}		Proj _{ILM-OPL}		Proj _{OPL-BM}		Proj _{Jmean}		3D	
Method		SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓	SSIM ↑	Perp. ↓
Pix2Pix3D	64.6M	0.494	0.641	0.485	0.655	0.419	0.632	0.510	0.609	0.903	0.136
+CDS	64.6M	0.560	0.552	0.532	0.556	0.494	0.562	0.567	0.542	0.902	0.135
+MSFF	52.7M	0.550	0.636	0.535	0.643	0.467	0.641	0.554	0.607	0.909	0.132
XOCT	52.7M	0.568	0.541	0.544	0.540	0.501	0.545	0.574	0.537	0.905	0.131

Table 5.2 summarizes the ablation study assessing the impact of CDS and MSFF. CDS significantly improves en-face projection metrics by introducing layer-aware constraints, which enforce intra-layer vascular coherence. MSFF enhances volumetric reconstructions by capturing fine-grained structures with reduced parameter cost. When combined, CDS and MSFF yield complementary benefits, enabling XOCT to achieve both sharp vascular delineation and volumetric continuity. This synergy results in the highest performance across all evaluation metrics, confirming the effectiveness of the proposed design.

5.2.4 Conclusion

In this work, we introduced **XOCT**, a novel framework for OCT-to-OCTA translation that integrates Cross-Dimensional Supervision (CDS) to enforce layer-specific consistency and a Multi-Scale Feature Fusion (MSFF) module to enhance fine vascular reconstruction. Experiments on the OCTA-500 dataset demonstrate that XOCT achieves substantial improvements over state-of-the-art baselines, particularly in preserving vascular continuity and microvascular details within 2D en-face projections and 3D volumetric reconstructions. These advances are critical for reliable retinal pathology assessment.

While promising, XOCT still requires validation across larger clinical datasets and heterogeneous imaging devices to ensure robustness in real-world deployment. Future directions include improving generalization to diverse patient populations, enhancing reconstruction of small capillaries, and optimizing computational efficiency for real-time applications. Extending XOCT to multi-modal retinal imaging represents another exciting avenue, with the potential to further strengthen its clinical impact.

5.3 Paired 3D Image and Mask Generation with Diffusion Model

5.3.1 Introduction

The success of deep learning in medical image analysis relies heavily on the availability of large-scale, accurately annotated datasets. High-quality 3D data is particularly critical for tasks such as segmentation and detection, where volumetric context provides essential structural information for precise diagnosis and treatment planning [114, 55, 131, 29, 145]. However, acquiring such data remains challenging due to the inherent complexity of medical images, the requirement for expert annotation, and persistent privacy concerns. These barriers severely restrict the diversity and scale of existing datasets, limiting the development of robust and generalizable models.

Synthetic data generation has emerged as a promising solution to these challenges. By reducing reliance on manual annotation and alleviating privacy risks, it enables broader data accessibility without direct dependence on patient populations. Earlier research in this domain primarily focused on 2D image synthesis [54, 14, 17, 128], often restricted to single modalities or lacking paired segmentation masks. For instance, Guibas et al. [53] employed

dual generative adversarial networks (GANs) [50] to synthesize labeled retinal fundus images, while Fernandez et al. [42] combined a label generator with an image generator for brain MRI. More recent attempts extended these ideas into three dimensions, such as Subramaniam et al. [127] who applied Wasserstein GANs (WGANs) [4] to generate small 3D patches and corresponding vessel segmentations. Yet, the task of generating complete volumetric medical images with paired multi-label masks remains largely unexplored.

Two central obstacles underlie this difficulty. First, directly training on entire 3D volumes is computationally prohibitive due to GPU memory constraints, while naive downsampling compromises anatomical fidelity. Second, the scarcity of annotated volumetric datasets makes it inefficient to treat each volume as a single training instance, resulting in limited data efficiency and poor generalization. Overcoming these limitations requires a generative framework that balances computational feasibility with the ability to capture full volumetric complexity and ensure accurate alignment between images and segmentation masks.

To this end, we propose **MedGen3D**, a diffusion-based generative framework for synthesizing realistic, complete 3D medical images together with corresponding segmentation masks. MedGen3D reformulates volumetric synthesis as an autoregressive sequence generation problem, treating 3D volumes as ordered stacks of slices. In the first stage, a Multi-Condition Diffusion Probabilistic Model (MC-DPM) generates anatomical masks by combining conditional and unconditional diffusion, enabling both free-form mask generation from noise and subsequence completion conditioned on existing slices. Slice indices are incorporated to exploit anatomical regularity across spatial positions. In the second stage, the generated masks provide structural guidance for a conditional image generator, followed by a semantic diffusion refiner to ensure anatomical consistency, spatial coherence, and accurate alignment between images and labels.

Through this two-stage design, MedGen3D addresses long-standing challenges in volumetric data generation, producing high-fidelity, diverse, and fully labeled 3D datasets. Experimental

results show that the synthesized volumes not only exhibit realistic anatomical appearance but also improve downstream segmentation performance, highlighting the utility of Med-Gen3D as a practical tool for advancing medical imaging research and applications.

5.3.2 Method

We propose a sequential framework to generate complex 3D volumetric images with paired segmentation masks, as illustrated in Figure 5.4. The pipeline consists of two stages: (1) a 3D mask generator that synthesizes multi-label segmentation volumes autoregressively, and (2) a conditional medical image generator that synthesizes realistic 3D images guided by the generated masks. Details of each component are presented below.

5.3.2.1 Preliminary

Diffusion Probabilistic Model. A diffusion probabilistic model (DPM) [57] is a parameterized Markov chain of length T , designed to learn the data distribution $p(X)$. The forward diffusion process (FDP) gradually perturbs data into Gaussian noise via

$$q(X_t | X_{t-1}) = \mathcal{N}\left(X_t; \sqrt{1 - \beta_t} X_{t-1}, \beta_t I\right), \quad (5.6)$$

where $X_0 \sim q(X_0)$ and $p(X_T) = \mathcal{N}(X_T; 0, I)$. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$. The reverse diffusion process (RDP) learns to denoise by predicting the added noise through

$$\mathcal{L}(\theta) = \mathbb{E}_{X_0, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right], \quad (5.7)$$

where ϵ_θ is the predicted noise and θ the learnable parameters.

Classifier-Free Guidance. Conditional diffusion models can be enhanced with classifier-



Figure 5.4: **Overview of MedGen3D.** The framework first autoregressively generates a 3D mask volume using MC-DPM, then synthesizes realistic images conditioned on the generated masks via a sequence-to-sequence generator and a semantic diffusion refiner.

free guidance [58]. During training, the condition c is dropped with probability p . At inference, guidance extrapolates the output toward $\epsilon_\theta(X_t | c)$ and away from $\epsilon_\theta(X_t | \emptyset)$:

$$\hat{\epsilon}_\theta(X_t | c) = \epsilon_\theta(X_t | \emptyset) + s \cdot (\epsilon_\theta(X_t | c) - \epsilon_\theta(X_t | \emptyset)), \quad (5.8)$$

where $s \geq 1$ is the guidance scale.

5.3.2.2 3D Mask Generator

Directly training on entire 3D volumes is infeasible due to memory constraints and limited annotated data. Instead, we treat 3D volumes as ordered sequences of slices and generate them autoregressively. The process begins by unconditionally generating an initial subsequence of m slices from noise, then expands both forward and backward conditioned on existing slices.

To unify unconditional, forward, and backward generation, we introduce a **Multi-Condition Diffusion Probabilistic Model (MC-DPM)**. As illustrated in Figure 5.5, MC-DPM can generate subsequences from noise or conditioned on given slices. Anatomical regularity across slices is exploited by incorporating relative slice position $\tilde{z} = z/D$, where z is the subsequence start index and D the full depth. $\tilde{z}$ is embedded and concatenated with the time embedding. Binary indicators mark which slices are provided as conditions.

Training. Given a mask volume $M \in \mathbb{R}^{D \times H \times W}$, subsequences of length m are sampled as $\{M_z, \dots, M_{z+m-1}\}$. Conditional slices X^C are chosen as either the first n slices, the last n ,

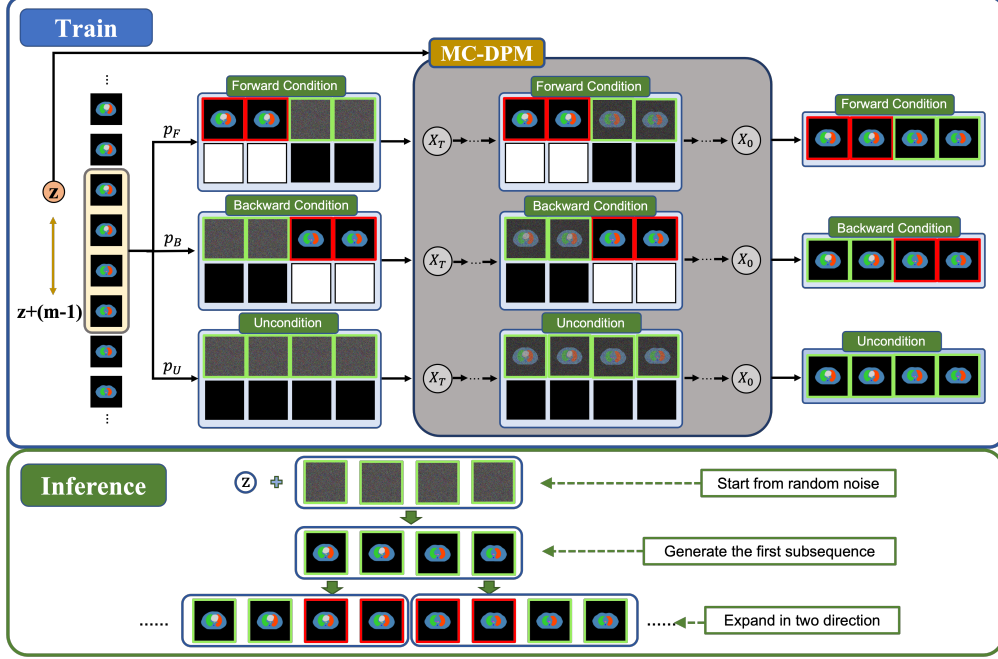


Figure 5.5: **3D mask generator with MC-DPM**. For a target position z , MC-DPM generates subsequences (length m) either unconditionally or conditioned on the first/last n slices, chosen with probabilities $p^C \in \{p_F, p_B, p_U\}$. Red boxes: conditional slices; green boxes: generated slices.

or none, based on p^C . The model predicts the remaining slices X^P . The reverse diffusion is formulated as:

$$p_\theta(X_{0:T}^P | X^C, \tilde{z}) = p(X_T^P) \prod_{t=1}^T p_\theta(X_{t-1}^P | X_t^P, X^C, \tilde{z}), \quad (5.9)$$

with $p(X_T^P) = \mathcal{N}(0, I)$. The training loss is

$$\mathcal{L}(\theta) = \mathbb{E}_{X_0, \epsilon, p^C, z, t} \left[\left\| \epsilon - \epsilon_\theta(X_t^P, X^C, \tilde{z}, t) \right\|^2 \right], \quad (5.10)$$

where $X_t^P = \sqrt{\bar{\alpha}_t} X_0^P + \sqrt{1 - \bar{\alpha}_t} \epsilon$.

Inference. At test time, MC-DPM first generates an m -slice subsequence from noise at a random position z . The volume is then autoregressively expanded forward and backward, conditioned on existing slices, to reconstruct the full mask volume.

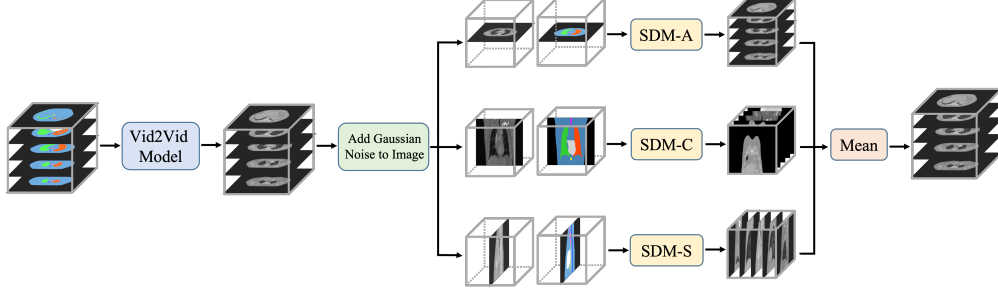


Figure 5.6: **Conditional Image Generation.** Vid2Vid synthesizes slices sequentially given masks. A semantic diffusion refiner further enhances image fidelity by adding small noise and denoising with SDMs trained on three orthogonal views. The final result averages the three refined reconstructions.

5.3.2.3 Conditional Image Generator

In the second stage, images are synthesized conditioned on the generated masks using a sequence-to-sequence model, followed by a semantic diffusion refiner (Fig. 5.6).

Image Sequence Generator. We employ Vid2Vid [137], which synthesizes each slice as a combination of the warped previous slice and newly generated texture, weighted by a learned mask. Training follows the original objective, including multi-scale adversarial losses, video discriminator loss, flow estimation loss, and feature matching loss.

Semantic Diffusion Refiner. While Vid2Vid ensures cross-slice consistency, artifacts such as blocking, blur, and suboptimal textures remain. To mitigate these, we introduce a semantic diffusion refiner (SDR). For each orthogonal view (axial, coronal, sagittal), a semantic diffusion model (SDM) is trained to refine noisy slices conditioned on masks. During inference, we add light noise (10 steps) to the generated images, then denoise with SDMs to preserve anatomical structures while enhancing realism. The final volume is obtained by averaging the outputs across three views.

This two-step design ensures both structural alignment to masks and improved texture realism in the generated images.

Algorithm 1 3D Mask Generation (Inference)

Require: n : number of conditional slices, m : length of subsequence, L : length of the sequence to generate, P_θ : the predicted probability distribution

$Z \leftarrow L - (m - 1)$

$z \sim \text{Uniform}(\{0, 1, \dots, Z\})$ $\triangleright$ Randomly pick one z as start position

Initialize an empty mask sequence $\mathcal{M}$

$X \leftarrow \{M_z, M_{z+1}, \dots, M_{z+(m-1)}\} \sim P_\theta(X_0^P \mid X^C = \emptyset, z)$

$\mathcal{M} \leftarrow \mathcal{M} \cup X$

//Forward Sampling

$z' \leftarrow z$

while $z' \leq Z$ **do**

$z' \leftarrow z' + (m - n)$

$X^C \leftarrow \mathcal{M}[-n :]$ $\triangleright$ Select the last n masks as condition

$X^P \sim P_\theta(X_0^P \mid X^C, z')$ $\triangleright$ Sample the following $(m-n)$ masks

$\mathcal{M} \leftarrow \mathcal{M} \cup X^P$ $\triangleright$ Add the generated masks to the end of sequence

end while

//Backward Sampling

$z' \leftarrow z$

while $z' \geq 0$ **do**

$z' \leftarrow z' - (m - n)$

$X^C \leftarrow \mathcal{M}[: n]$ $\triangleright$ Select the first n masks as condition

$X^P \sim P_\theta(X_0^P \mid X^C, z')$ $\triangleright$ Sample the previous $(m-n)$ masks

$\mathcal{M} \leftarrow X^P \cup \mathcal{M}$ $\triangleright$ Add the generated masks to the start of sequence

end while

return $\mathcal{M}$ $\triangleright$ Return the generated mask sequence

5.3.3 Experiments

5.3.3.1 Datasets and Setups

Datasets: We conducted experiments on two anatomical sites: thorax and brain. For both generative modeling and downstream segmentation tasks, we used the following datasets:

- SegTHOR [76]: 3D thoracic CT scans (25 training, 5 validation, 10 testing);
- OASIS [95]: 3D brain MRI T1 scans (40 training, 10 validation, 10 testing).

For downstream segmentation and transfer learning experiments, we further included 3D thoracic CT scans from StructSeg-Thorax [2] and Public-Thor [29], as well as brain MRI T1 scans from ADNI [1]. Specifically, we used 10 scans for fine-tuning, 5 for validation, and 10 for testing in each case.

Preprocessing: For thoracic datasets, CT volumes were cropped and padded to $(96 \times 320 \times 320)$. Six anatomical structures (left/right lung, spinal cord, esophagus, heart, trachea) were annotated and validated by an experienced radiation oncologist, with an additional body mask included to aid image synthesis. For brain MRI datasets, we used FreeSurfer [44] to segment four regions (cortex, subcortical gray matter, white matter, CSF), cropping each volume to $(192 \times 160 \times 160)$. Masks were assigned discrete values for each region or organ and merged into a single 3D volume. For mask synthesis, slices were resized to 128×128 , and subsequence length m was set to 6.

Implementation: All segmentation models were implemented using MONAI [20] with standard data augmentation strategies, including spatial transformations and intensity perturbations.

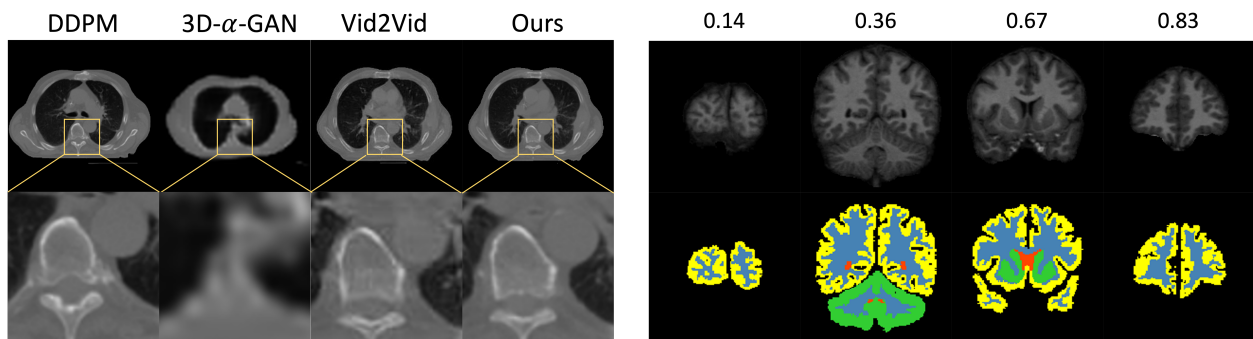


Figure 5.7: Qualitative comparison of synthetic data. Our method produces anatomically consistent images with sharper organ boundaries and more realistic textures than 3D- α -WGAN and Vid2Vid. Left: comparison across generative models. Right: visualization of synthetic MRI slices at different anatomical positions.

Comparisons: We compared synthetic image quality against DDPM [57], 3D- α -WGAN [75], and Vid2Vid [137]. For downstream segmentation, we evaluated four models (Unet2D [114], UNet3D [34], UNETR [55], and Swin-UNETR [131]) under multiple training strategies.

	Thoracic CT		Brain MRI	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
DDPM [57]	35.2	0.316	34.9	0.298
3D- α -WGAN [75]	136.2	0.286	136.4	0.289
Vid2Vid [137]	47.3	0.300	48.2	0.324
Ours	39.6	0.305	40.3	0.326

Table 5.3: Synthetic image quality comparison between baselines and ours.

5.3.3.2 Evaluation of Synthetic Image Quality

Synthetic Dataset: From limited real data (30 CT scans from SegTHOR and 50 MRIs from OASIS), we synthesized 110 additional 3D thoracic CT volumes and 110 brain MRI volumes.

We evaluated fidelity and diversity using Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) from three anatomical views.

As shown in Table 1, our method achieves slightly lower FID but comparable LPIPS compared to DDPM, which generates 2D slices without explicit anatomical constraints. In contrast, 3D- α -WGAN, despite using substantially larger training data (146 thoracic CTs and 414 brain MRIs), yields significantly worse FID and LPIPS scores. Vid2Vid also underperforms, confirming the effectiveness of our semantic diffusion refinement strategy.

5.3.3.3 Evaluation for Segmentation Tasks

We next assessed the impact of synthetic data on segmentation performance, reporting Dice similarity coefficient (DSC) across datasets. Four segmentation models—Unet2D, UNet3D, UNETR, and Swin-UNETR—were trained under five different strategies combining real and synthetic data.

Results in Tables 5.4 and 5.5 show that 3D models substantially outperform 2D models,

	SegTHOR*				StructSeg-Thorax				Public-Thor			
	Unet	Unet	UNETR	Swin	Unet	Unet	UNETR	Swin	Unet	Unet	UNETR	Swin
	2D	3D		UNETR	2D	3D		UNETR	2D	3D		UNETR
E2-1	0.817	0.873	0.867	0.878	0.722	0.793	0.789	0.810	0.822	0.837	0.836	0.847
E2-2	0.815	0.846	0.845	0.854	0.736	0.788	0.788	0.803	0.786	0.838	0.814	0.842
E2-3	0.845	0.881	0.886	0.886	0.772	0.827	0.824	0.827	0.812	0.856	0.853	0.856
E2-4	0.855	0.887	0.894	0.899	0.775	0.833	0.825	0.833	0.824	0.861	0.852	0.867
E2-5	0.847	0.891	0.890	0.897	0.783	0.833	0.823	0.835	0.818	0.864	0.858	0.867

Table 5.4: Experiment 2: DSC of different thoracic segmentation models. There are 5 training strategies, namely: **E2-1**: Training with real SegTHOR training data; **E2-2**: Training with synthetic data; **E2-3**: Training with both synthetic and real data; **E2-4**: Finetuning model from E2-2 using real training data; and **E2-5**: finetuning model from E2-3 using real training data. (* denotes the training data source.)

	OASIS*				ADNI			
	Unet	Unet	UNETR	Swin	Unet	Unet	UNETR	Swin
	2D	3D		UNETR	2D	3D		UNETR
E2-1	0.930	0.951	0.952	0.954	0.815	0.826	0.880	0.894
E2-2	0.905	0.936	0.935	0.934	0.759	0.825	0.828	0.854
E2-3	0.938	0.953	0.953	0.955	0.818	0.888	0.898	0.906
E2-4	0.940	0.955	0.954	0.956	0.819	0.891	0.903	0.903
E2-5	0.940	0.954	0.954	0.956	0.819	0.894	0.902	0.906

Table 5.5: Experiment 2: DSC of brain segmentation models. Please refer to Table 5.4 for detailed training strategies. (* denotes the training data source.)

highlighting the importance of volumetric annotations. Models trained only on synthetic data (**E2-2**) underperform compared to those trained on real data (**E2-1**). However, combining real and synthetic data (**E2-3**) improves performance in most settings, with the exception of Unet2D on Public-Thor. Fine-tuning pretrained models with real data (**E2-4**, **E2-5**) consistently achieves the best results, demonstrating the utility of synthetic pretraining.

For transfer learning (Table 5.6), pretraining on synthetic data (**E3-3**, **E3-4**) consistently improves adaptation to new datasets compared to training from scratch (**E3-1**) or using real data only. This confirms that synthetic volumes can mitigate data scarcity and accelerate domain adaptation.

	Thoracic CT		Brain MRI
	StructSeg-Thorax*	Public-Thor*	ADNI*
E3-1	0.845	0.897	0.946
E3-2	0.865	0.901	0.948
E3-3	0.878	0.913	0.949
E3-4	0.882	0.914	0.949

Table 5.6: Experiment 3: DSC of Swin-UNETR finetuned with real dataset. There are 4 training strategies: **E3-1**: Training from scratch for each dataset using limited finetuning data; **E3-2** Finetuning the model E2-1 from experiment 2; **E3-3** Finetuning the model E2-4 from experiment 2; and **E3-4** Finetuning the model E2-5 from experiment 2. (* denotes the finetuning data source.)

5.3.4 Conclusion

We presented **MedGen3D**, a diffusion-based framework for synthesizing realistic and fully labeled 3D medical volumes. By jointly generating anatomical masks and corresponding images, MedGen3D overcomes the limitations of scarce annotated datasets and demonstrates clear benefits for downstream segmentation tasks, particularly in low-data and cross-domain settings.

Looking ahead, we plan to unify the image sequence generator and semantic diffusion refiner into an end-to-end model, as well as extend MedGen3D to multi-modal synthesis across diverse imaging modalities. We believe this work provides a solid foundation for advancing large-scale, high-quality 3D medical data generation, with direct implications for model training, clinical research, and translational applications.

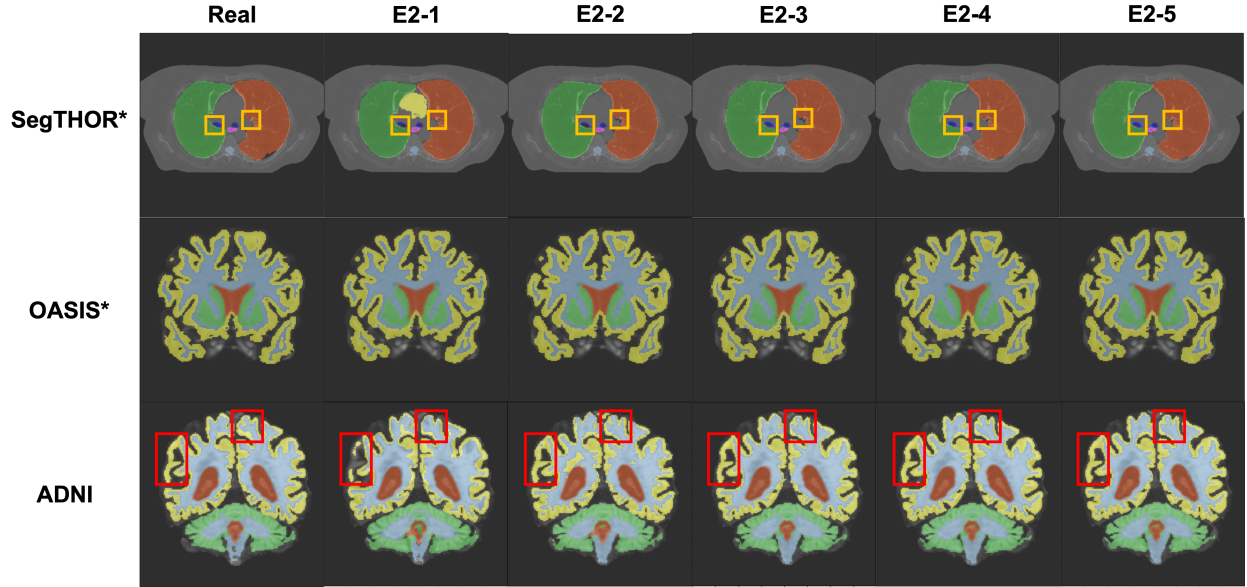


Figure 5.8: Experiment 2: Qualitative comparison of training strategies with Swin-UNETR. Improvements from synthetic data are more pronounced when training and testing are cross-dataset (e.g., train on OASIS, test on ADNI) compared to in-domain (OASIS*).

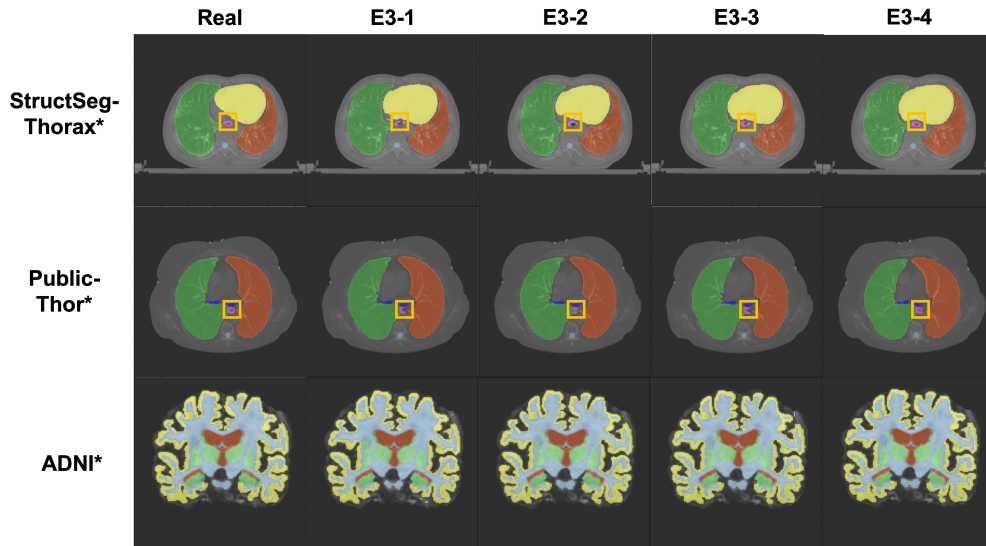


Figure 5.9: Experiment 3: Transfer learning strategies with Swin-UNETR. Synthetic pre-training enhances finetuning performance on unseen datasets (see Table 5.6).

Chapter 6

Conclusion

Medical image analysis has become indispensable in modern healthcare, advancing diagnosis, prognosis, and treatment planning. Modalities such as MRI, CT, and OCT provide unprecedented views of anatomy and physiology, yet the scale, complexity, and variability of medical data pose major challenges. High-dimensional images, subtle anatomical variations, and limited annotated datasets make it difficult to fully realize the potential of deep learning in clinical practice. Developing computational frameworks that are accurate, generalizable, and interpretable is therefore essential for the future of medical imaging.

This dissertation has pursued three interconnected directions—**image registration**, **shape representation**, and **data generation**—each addressing a fundamental bottleneck in the medical imaging pipeline. Across these domains, six novel contributions were introduced, advancing both methodological innovation and clinical applicability.

The first direction, *image registration*, focused on enabling robust alignment across individuals and timepoints. We proposed **DNVF**, a continuous neural velocity field representation for accurate and flexible diffeomorphic registration. Building on this, **OFG** introduced an optimization-for-guidance strategy, where instance-specific optimization generates reliable

pseudo-supervision, reducing dependence on manual annotations. Finally, **DCA** advanced multi-scale registration through a deformation certainty attention mechanism, improving correspondence propagation across scales. Together, these contributions demonstrate how continuous representations, self-improving supervision, and reliability correspondence can drive a new generation of registration frameworks.

The second direction, *shape representation*, explored how neural field-based formulations can model population-level anatomical variability. We introduced **HNDF**, which integrates diffeomorphic flows with implicit representations to capture both structural detail and deformation in a continuous space. This line of work highlights the promise of neural fields for compact, flexible, and generalizable shape modeling, while also pointing to future opportunities in extending these frameworks to more complex and heterogeneous anatomical structures.

The third direction, *data generation*, addressed the persistent challenge of limited annotated datasets. To alleviate data scarcity, we developed **XOCT**, a GAN-based OCT-to-OCTA translation model with cross-dimensional supervision and multi-scale feature fusion, producing high-fidelity angiograms with preserved vascular detail. Complementing this, **MedGen3D** introduced the first diffusion-based framework for large-scale volumetric synthesis with paired segmentation masks. By combining probabilistic mask generation, conditional image synthesis, and semantic refinement, MedGen3D enables anatomically consistent and diverse 3D data generation at scale. These contributions illustrate how generative modeling can enrich medical datasets, reduce reliance on costly annotations, and open new avenues for scalable clinical model development.

Taken together, these six contributions form a cohesive body of work that addresses core challenges in medical image analysis. By advancing registration, representation, and generation, they establish foundations for more accurate, efficient, and trustworthy computational methods in healthcare.

Looking forward, integrating these directions offers exciting opportunities. Registration frameworks could be coupled with generative models to synthesize anatomically consistent data for rare conditions or simulate disease progression. Shape representation could be extended to multi-modal and longitudinal imaging, capturing temporal dynamics and supporting predictive modeling of disease evolution. Generative approaches may further benefit from incorporating prior anatomical and physiological knowledge, ensuring that synthetic data are not only realistic but also clinically meaningful. Moreover, the rise of foundation models and large-scale pretraining invites exploration of domain adaptation and fine-tuning strategies to improve generalizability across modalities, populations, and institutions. Finally, interpretability and uncertainty modeling will remain essential to build systems that are not only technically powerful but also trusted in clinical practice.

In summary, this dissertation contributes a set of frameworks that advance the state of medical image registration, shape representation, and data generation. By bridging these domains and pointing toward future integration, it outlines a path toward medical image analysis that is scalable, generalizable, and clinically impactful—ultimately bringing advanced computational tools closer to everyday healthcare.

Bibliography

- [1] <https://adni.loni.usc.edu/>.
- [2] <https://structseg2019.grand-challenge.org/dataset/>.
- [3] P. Anvari, M. Ashrafkhorasani, A. Habibi, and K. G. Falavarjani. Artifacts in optical coherence tomography angiography. *Journal of ophthalmic & vision research*, 16(2):271, 2021.
- [4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv: Arxiv-1701.07875*, 2017.
- [5] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006.
- [6] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, page 95–113, Oct 2007.
- [7] J. Ashburner and K. J. Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000.
- [8] 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, 2008.
- [9] F. B. Freesurfer. *Neuroimage*, 2012. <https://surfer.nmr.mgh.harvard.edu/>.
- [10] R. Bajcsy and S. Kovačič. Multiresolution elastic matching. *Computer vision, graphics, and image processing*, 46(1):1–21, 1989.
- [11] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.
- [12] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

- [13] R. Basri, M. Galun, A. Geifman, D. Jacobs, Y. Kasten, and S. Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, pages 685–694. PMLR, 2020.
- [14] C. Baur, S. Albarqouni, and N. Navab. Melanogans: high resolution skin lesion synthesis with gans. *arXiv preprint arXiv:1804.04338*, 2018.
- [15] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.
- [16] N. Benbarka, T. Höfer, A. Zell, et al. Seeing implicit neural representations as fourier series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2041–2050, 2022.
- [17] C. Bermudez, A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick, and B. A. Landman. Learning implicit brain mri manifolds with deep learning. In *Medical Imaging: Image Processing*. SPIE, 2018.
- [18] A. Bigalke, L. Hansen, T. C. W. Mok, and M. P. Heinrich. Unsupervised 3d registration through optimization-guided cyclical self-training. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 677–687, Cham, 2023. Springer Nature Switzerland.
- [19] J. Campbell, M. Zhang, T. Hwang, S. Bailey, D. Wilson, Y. Jia, and D. Huang. Detailed vascular anatomy of the human retina by projection-resolved optical coherence tomography angiography. *Scientific reports*, 7:42201, 2 2017.
- [20] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.
- [21] U. Centre for Medical Image Computing, University College London. Niftyreg, 2023. <http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg>.
- [22] R. Chabra, J. E. Lenssen, E. Ilg, T. Schmidt, J. Straub, S. Lovegrove, and R. Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 608–625. Springer, 2020.
- [23] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [24] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022.

- [25] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021.
- [26] J. Chen, Y. Liu, Y. He, and Y. Du. Deformable cross-attention transformer for medical image registration. In *International Workshop on Machine Learning in Medical Imaging*, pages 115–125. Springer, 2023.
- [27] J. Chen, Y. Liu, S. Wei, Z. Bian, S. Subramanian, A. Carass, J. L. Prince, and Y. Du. A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Medical Image Analysis*, page 103385, 2024.
- [28] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
- [29] X. Chen, S. Sun, N. Bai, K. Han, Q. Liu, S. Yao, H. Tang, C. Zhang, Z. Lu, Q. Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.
- [30] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019.
- [31] Z. Chen, Y. Zheng, and J. C. Gee. Transmatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE transactions on medical imaging*, 43(1):15–27, 2023.
- [32] X. Cheng, T. Zhang, W. Lu, Q. Meng, A. F. Frangi, and J. Duan. Sacb-net: Spatial-awareness convolutions for medical image registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5227–5237, 2025.
- [33] K. Choo, Y. Jun, M. Yun, and S. J. Hwang. Slice-consistent 3d volumetric brain ct-to-mri translation with 2d brownian bridge diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 657–667. Springer, 2024.
- [34] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*. Springer, 2016.
- [35] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.
- [36] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.

- [37] L. R. De Pretto, E. M. Moulton, A. Y. Alibhai, O. M. Carrasco-Zevallos, S. Chen, B. Lee, A. J. Witkin, C. R. Bauman, E. Reichel, A. Z. de Freitas, et al. Controlling for artifacts in widefield optical coherence tomography angiography measurements of non-perfusion area. *Scientific reports*, 9(1):9096, 2019.
- [38] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, feb 2019.
- [39] K. Deng, G. Yang, D. Ramanan, and J.-Y. Zhu. 3d-aware conditional image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4434–4445, 2023.
- [40] Y. Deng, J. Yang, and X. Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021.
- [41] K. Diederik. Adam: A method for stochastic optimization. (*No Title*), 2014.
- [42] V. Fernandez, W. H. L. Pinaya, P. Borges, P.-D. Tudosiu, M. S. Graham, T. Vercauteren, and M. J. Cardoso. Can segmentation models be trained with fully synthetically generated data? In *MICCAI Workshop, SASHIMI*. Springer, 2022.
- [43] E. Ferrante and N. Paragios. Slice-to-volume medical image registration: A survey. *Medical image analysis*, 39:101–123, 2017.
- [44] B. Fischl. Freesurfer. *Neuroimage*, 2012.
- [45] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.
- [46] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, 2020.
- [47] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022.
- [48] M. Ghahremani, M. Khateri, B. Jian, B. Wiestler, E. Adeli, and C. Wachinger. H-vit: A hierarchical vision transformer for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11513–11523, 2024.
- [49] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 3 edition, 2008.
- [50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

- [51] I. J. Goodfellow, O. Vinyals, and A. M. Saxe. Qualitatively characterizing neural network optimization problems, 2015.
- [52] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018.
- [53] J. T. Guibas, T. S. Virdi, and P. S. Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.
- [54] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama. Gan-based synthetic brain mr image generation. In *ISBI*. IEEE, 2018.
- [55] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 2022.
- [56] T. Heimann and H.-P. Meinzer. Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009.
- [57] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [58] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv: 2207.12598*, 2022.
- [59] A. Hoopes, M. Hoffmann, B. Fischl, J. Guttag, and A. V. Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *International Conference on Information Processing in Medical Imaging*, pages 3–17. Springer, 2021.
- [60] T. T. Hormel, D. Huang, and Y. Jia. Artifacts and artifact removal in optical coherence tomographic angiography. *Quantitative imaging in medicine and surgery*, 11(3):1120–1133, 2021.
- [61] K. Huang, N. Su, Y. Tao, M. Li, X. Ma, Z. Ji, S. Yuan, and Q. Chen. Cross-device octa generation by patch-based 3d multi-scale feature adaption. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):641–653, 2 2024.
- [62] D. J. Im, M. Tao, and K. Branson. An empirical analysis of the optimization of deep network loss surfaces, 2017.
- [63] M. Incoronato, M. Aiello, T. Infante, C. Cavaliere, A. M. Grimaldi, P. Mirabelli, S. Monti, and M. Salvatore. Radiogenomic analysis of oncological data: a technical survey. *International journal of molecular sciences*, 18(4):805, 2017.
- [64] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

- [65] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks, 2016.
- [66] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, T. Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020.
- [67] Z. Jiang, Z. Huang, B. Qiu, X. Meng, Y. You, X. Liu, M. Geng, G. Liu, C. Zhou, K. Yang, et al. Weakly supervised deep learning-based optical coherence tomography angiography. *IEEE Transactions on Medical Imaging*, 40(2):688–698, 10 2021.
- [68] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016*, pages 694–711. Springer, 2016.
- [69] M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes. Dual-stream pyramid registration network. *Medical image analysis*, 78:102379, 2022.
- [70] A. H. Kashani, C.-L. Chen, J. K. Gahm, F. Zheng, G. M. Richter, P. J. Rosenfeld, Y. Shi, and R. K. Wang. Optical coherence tomography angiography: A comprehensive review of current methods and clinical applications. *Progress in retinal and eye research*, 60:66–100, 2017.
- [71] D. N. Kennedy, C. Haselgrove, S. M. Hodge, P. S. Rane, N. Makris, and J. A. Frazier. Candishare: a resource for pediatric neuroimaging data, 2012.
- [72] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis*, 71:102036, 2021.
- [73] A. Klein, S. S. Ghosh, F. S. Bao, J. Giard, Y. Häme, E. Stavsky, N. Lee, B. Rossa, M. Reuter, E. Chaibub Neto, et al. Mindboggling morphometry of human brains. *PLoS computational biology*, 13(2):e1005350, 2017.
- [74] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 2252–2261. IEEE, Oct. 2019. ISSN: 2380-7504.
- [75] G. Kwon, C. Han, and D.-s. Kim. Generation of 3d brain mri using auto-encoding generative adversarial networks. In *MICCAI*. Springer, 2019.
- [76] Z. Lambert, C. Petitjean, B. Dubray, and S. Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *IPTA*. IEEE, 2020.
- [77] C. S. Lee, A. J. Tying, Y. Wu, S. Xiao, A. S. Rokem, N. P. DeRuyter, Q. Zhang, A. Tufail, R. K. Wang, and A. Y. Lee. Generating retinal flow maps from structural optical coherence tomography with artificial intelligence. *Scientific reports*, 9:5694, 2019.

- [78] B. Li, Y. Xie, H. Zhang, and D. Shen. Recursive cascaded networks for unsupervised medical image registration. *IEEE Transactions on Medical Imaging*, 39(12):4221–4231, 2020.
- [79] B. Li, K. Xue, B. Liu, and Y.-K. Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023.
- [80] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets, 2018.
- [81] M. Li, K. Huang, Q. Xu, J. Yang, Y. Zhang, Z. Ji, K. Xie, S. Yuan, Q. Liu, and Q. Chen. Octa-500: A retinal dataset for optical coherence tomography angiography study. *Medical image analysis*, 93:103092, 2024.
- [82] P. L. Li, C. O’Neil, S. Saberi, K. Sinder, K. Wang, B. Tan, Z. Hosseinaee, K. Bizhevat, and V. Lakshminarayanan. Deep learning algorithm for generating optical coherence tomography angiography (octa) maps of the retinal vasculature. In *Proc. SPIE 11511, Applications of Machine Learning 2020*, pages 39–49, 2020.
- [83] S. Li, D. Zhang, X. Li, C. Ou, L. An, Y. Xu, W. Yang, Y. Zhang, and K.-T. Cheng. Vessel-promoted oct to octa image translation by heuristic contextual constraints. *Medical Image Analysis*, 98:103311, 2024.
- [84] Y. Li, I. Yakushev, D. M. Hedderich, and C. Wachinger. Pasta: Pathology-aware mri to pet cross-modal translation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–540. Springer, 2024.
- [85] Z. Lin, Q. Zhang, G. Lan, J. Xu, J. Qin, L. An, and Y. Huang. Deep learning for motion artifact-suppressed octa image generation from both repeated and adjacent oct scans. *Mathematics*, 12:446, 2024.
- [86] X. Liu, Z. Huang, Z. Wang, C. Wen, Z. Jiang, Z. Yu, J. Liu, G. Liu, X. Huang, A. Maier, et al. A deep learning based pipeline for optical coherence tomography angiography. *Journal of Biophotonics*, 12(10), 10 2019.
- [87] Y. Liu, L. Zuo, S. Han, Y. Xue, J. L. Prince, and A. Carass. Coordinate translator for learning deformable medical image registration. In *International workshop on multiscale multimodal medical imaging*, pages 98–109. Springer, 2022.
- [88] I. C. London. Information extraction from images, 2023. <https://brain-development.org/ixi-dataset/>.
- [89] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

- [90] C. Lorenz and N. Krahnstover. 3d statistical shape models for medical image segmentation. In *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No. PR00062)*, pages 414–423. IEEE, 1999.
- [91] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021.
- [92] T. Ma, X. Dai, S. Zhang, and Y. Wen. Pivit: Large deformation image registration with pyramid-iterative vision transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 602–612. Springer, 2023.
- [93] T. Ma, S. Zhang, J. Li, and Y. Wen. Iirp-net: iterative inference residual pyramid network for enhanced image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11546–11555, 2024.
- [94] J. A. Maintz and M. A. Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998.
- [95] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. 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, 2007.
- [96] M. Meng, L. Bi, D. Feng, and J. Kim. Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–97. Springer, 2022.
- [97] M. Meng, L. Bi, M. Fulham, D. Feng, and J. Kim. Non-iterative coarse-to-fine transformer networks for joint affine and deformable image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 750–760. Springer, 2023.
- [98] M. Meng, D. Feng, L. Bi, and J. Kim. Correlation-aware coarse-to-fine mlps for deformable medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9645–9654, 2024.
- [99] M. Meng, M. Fulham, L. Bi, and J. Kim. Advancing deformable medical image registration with multi-axis cross-covariance attention. *arXiv preprint arXiv:2412.18545*, 2024.
- [100] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.
- [101] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

- [102] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.
- [103] T. C. Mok and A. 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, 2020.
- [104] T. C. Mok and A. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–221. Springer, 2020.
- [105] T. C. W. Mok, Z. Li, Y. Xia, J. Yao, L. Zhang, J. Zhou, and L. Lu. Deformable medical image registration under distribution shifts with neural instance optimization. In X. Cao, X. Xu, I. Rekik, Z. Cui, and X. Ouyang, editors, *Machine Learning in Medical Imaging*, pages 126–136, Cham, 2024. Springer Nature Switzerland.
- [106] G. Nam, M. Khlifi, A. Rodriguez, A. Tono, L. Zhou, and P. Guerrero. 3d-ldm: Neural implicit 3d shape generation with latent diffusion models. *arXiv preprint arXiv:2212.00842*, 2022.
- [107] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019.
- [108] B. Pan, Z. Ji, and Q. Chen. Multigan: Multi-domain image translation from oct to octa. In *Pattern Recognition and Computer Vision. PRCV 2022. Lecture Notes in Computer Science*, pages 336–347. Springer, 2022. 13535.
- [109] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- [110] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020.
- [111] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.
- [112] C. Reiser, S. Peng, Y. Liao, and A. Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.

- [113] P. Risholm, A. J. Golby, and W. Wells. Multimodal image registration for preoperative planning and image-guided neurosurgical procedures. *Neurosurgery Clinics*, 22(2):197–206, 2011.
- [114] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [115] R. B. Rosen, J. S. A. Romo, B. D. Krawitz, S. Mo, A. A. Fawzi, R. E. Linderman, J. Carroll, A. Pinhas, and T. Y. Chui. Earliest evidence of preclinical diabetic retinopathy revealed using optical coherence tomography angiography perfused capillary density. *American journal of ophthalmology*, 203:103–115, 2019.
- [116] L. I. Rudin and S. Osher. Total variation based image restoration with free local constraints. In *Proceedings of 1st international conference on image processing*, volume 1, pages 31–35. IEEE, 1994.
- [117] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.
- [118] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.
- [119] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008.
- [120] D. Shen and C. Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging*, 21(11):1421–1439, 2002.
- [121] Z. Shen, X. Han, Z. Xu, and M. Niethammer. Networks for Joint Affine and Non-parametric Image Registration. *arXiv e-prints*, page arXiv:1903.08811, Mar. 2019.
- [122] J. Shi, Y. He, Y. Kong, J.-L. Coatrieux, H. Shu, G. Yang, and S. Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. *arXiv preprint arXiv:2206.07349*, 2022.
- [123] J. R. Shue, E. R. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023.
- [124] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

- [125] G. Song, K. K. Chu, S. Kim, M. Crose, B. Cox, E. T. Jelly, J. N. Ulrich, and A. Wax. First clinical application of low-cost oct. *Translational vision science & technology*, 8(3):61–61, 2019.
- [126] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.
- [127] P. Subramaniam, T. Kossen, K. Ritter, A. Hennemuth, K. Hildebrand, A. Hilbert, J. Sobesky, M. Livne, I. Galinovic, A. A. Khalil, et al. Generating 3d tof-mra volumes and segmentation labels using generative adversarial networks. *Medical Image Analysis*, 2022.
- [128] L. Sun, J. Chen, Y. Xu, M. Gong, K. Yu, and K. Batmanghelich. Hierarchical amortized gan for 3d high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 2022.
- [129] S. Sun, K. Han, D. Kong, H. Tang, X. Yan, and X. Xie. Topology-preserving shape reconstruction and registration via neural diffeomorphic flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20845–20855, 2022.
- [130] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.
- [131] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, 2022.
- [132] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Stoll, and C. Theobalt. Patch-nets: Patch-based generalizable deep implicit 3d shape representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 293–309. Springer, 2020.
- [133] L. o. N. I. University of Southern California. Loni probabilistic brain atlas (lpba40), 2023. <https://loni.usc.edu/research/atlas>.
- [134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [135] H. Wang, D. Ni, and Y. Wang. Modet: learning deformable image registration via motion decomposition transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 740–749. Springer, 2023.
- [136] H. Wang, D. Ni, and Y. Wang. Recursive deformable pyramid network for unsupervised medical image registration. *IEEE Transactions on Medical Imaging*, 43(6):2229–2240, 2024.

- [137] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.
- [138] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [139] Z. Wang, H. Wang, and Y. Wang. Pyramid attention network for medical image registration. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.
- [140] J. M. Wolterink, J. C. Zwienenberg, and C. Brune. Implicit neural representations for deformable image registration. In *Medical Imaging with Deep Learning*, 2021.
- [141] Q. Wu, Y. Li, L. Xu, R. Feng, H. Wei, Q. Yang, B. Yu, X. Liu, J. Yu, and Y. Zhang. Irem: High-resolution magnetic resonance image reconstruction via implicit neural representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 65–74. Springer, 2021.
- [142] Y. Wu, T. Z. Jiahao, J. Wang, P. A. Yushkevich, M. A. Hsieh, and J. C. Gee. Nodexo: A neural ordinary differential equation based optimization framework for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20804–20813, 2022.
- [143] Z. Xu, C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572, 2016.
- [144] Z. Xu and M. Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2019.
- [145] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3971–3981, 2022.
- [146] J. Yang, U. Wickramasinghe, B. Ni, and P. Fua. Implicitatlas: learning deformable shape templates in medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15861–15871, 2022.
- [147] M. Zhang, R. Liao, A. V. Dalca, E. A. Turk, J. Luo, P. E. Grant, and P. Golland. Frequency diffeomorphisms for efficient image registration. In *International conference on information processing in medical imaging*, pages 559–570. Springer, 2017.
- [148] Y. Zhang, Y. Pei, and H. Zha. Learning dual transformer network for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–138. Springer, 2021.

- [149] Z. Zhang, Z. Ji, Q. Chen, S. Yuan, and W. Fan. Texture-guided u-net for oct-to-octa generation. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4 Springer*, pages 42–52, 2021.
- [150] S. Zhao, Y. Dong, E. I. Chang, Y. Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10600–10610, 2019.
- [151] Z. Zhao, S. Long, J. Pi, J. Wang, and L. Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation, 2022.
- [152] X. Zheng, Y. Liu, P. Wang, and X. Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, volume 41, pages 52–63. Wiley Online Library, 2022.
- [153] Z. Zheng, T. Yu, Q. Dai, and Y. Liu. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1439, 2021.
- [154] L. Zhou, Y. Du, and J. Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021.
- [155] S. Zhou, B. Hu, Z. Xiong, and F. Wu. Self-distilled hierarchical network for unsupervised deformable image registration. *IEEE Transactions on Medical Imaging*, 42(8):2162–2175, 2023.