Identifiable Bayesian Representations for Heterogeneous Medical Imaging

Xin Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:
Linda Shapiro, Chair
Chun Yuan, Chair
Niranjan Balu

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Identifiable Bayesian Representations for Heterogeneous Medical Imaging

Xin Wang

Co-Chairs of the Supervisory Committee:
Linda Shapiro
Department of Electrical and Computer Engineering
Paul G. Allen School of Computer Science and Engineering

Chun Yuan
Department of Radiology
Department of Bioengineering

Medical images exhibit pervasive heterogeneity arising from acquisition protocols, scanner properties, reconstruction pipelines, modality and contrast mechanisms, and anatomical variability across subjects and scan coverage. While deep learning has achieved strong performance in many medical image analysis tasks, robustness under compounded heterogeneity remains fragile. This dissertation argues that such fragility reflects a representational limitation: when task-relevant generative properties and observational variability are not organized in an identifiable manner, models may rely on unstable observational cues as surrogates, leading to degraded generalization as heterogeneity intensifies.

To address this challenge, we develop a unified perspective based on Bayesian representation learning and explicit latent role specification. Using latent variable models and variational inference, we construct mechanisms that preserve task-relevant invariants while suppressing observational variability, a requirement termed *identifiable invariant preservation*. We show that strengthening the identifiability of latent organization provides a practical

pathway to both interpretability and improved predictive performance across progressively more demanding regimes of heterogeneity.

The dissertation substantiates this thesis through three projects. First, we study supervised intracranial arterial calcification segmentation from multi-contrast brain MRI under intensity-level appearance heterogeneity. Because calcification is dark and often weakly expressed in MRI, segmentation depends on fragile contextual cues that are easily perturbed by scanner- and protocol-dependent fluctuations. A variational Bayesian formulation that restricts representational complexity yields more stable internal organization and improved segmentation accuracy. Second, we address unsupervised multimodal groupwise image registration under compounded contrast/modality variability and registration-compatible geometric heterogeneity. We formulate registration as hierarchical Bayesian inference that disentangles common anatomy from image-specific geometry, enabling intrinsic multimodal similarity and stable alignment without intensity-based heuristics. Third, we study unsupervised domain adaptation for segmentation in a correspondence-free regime with unpaired source and target domains. We introduce a probabilistic anatomical manifold that provides global canonicalization through a structured latent decomposition, inducing architecture-emergent adaptation without an explicit alignment loss and yielding a unified procedure across source-accessible and source-free settings.

Together, these contributions demonstrate that interpretable, identifiable latent organization is not merely an explanatory preference, but a practical mechanism for robust medical image learning under increasing heterogeneity. By developing Bayesian, disentangled formulations that progressively strengthen latent role specification across tasks, this dissertation provides a unified methodological pathway that improves both generalization and semantic interpretability in challenging real-world imaging regimes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my Ph.D. advisors, Dr. Linda Shapiro and Dr. Chun Yuan, for their continuous guidance and support throughout my doctoral study. Working under their joint supervision has been a defining aspect of my training. From Dr. Shapiro, I learned to approach problems with technical depth and methodological clarity; from Dr. Yuan, I gained a deep appreciation for the complexity and clinical significance of medical imaging. This interdisciplinary environment has profoundly shaped how I think about research. Over time, I have come to understand that methods are not merely tools, but ways of seeing problems.

I am also grateful to the members of my supervisory committee, Dr. Yen-Chi Chen, Dr. Yiyue Luo, and Dr. Niranjan Balu, for their valuable feedback and insightful discussions, which have significantly improved this work.

My sincere thanks go to my colleagues and collaborators in the Vascular Imaging Lab (UW) and the Medical Imaging and Computational Analysis Lab (the U). I have been fortunate to work alongside an exceptional group of peers, including Dr. Yin Guo, Dr. Kaiyu Zhang, Ms. Jiamin Xia, Ms. Zhiwei Tan, Mr. Xiangjian Hou, Ms. Chang Ni, and Dr. Li Chen, as well as many fellows and postdoctoral researchers whose experience and generosity have created a stimulating research environment. I would also like to thank Ms. Yi Zhang, Mr. Zixuan Liu, and Mr. Huayu Wang for their support and companionship during my time at UW.

I am deeply grateful to my collaborators at Fudan University and Renji Hospital. In particular, I would like to thank Dr. Xiahai Zhuang and Dr. Xinzhe Luo, who played an essential role in my early research training. Looking back, that stage was less about results

than about learning how to approach uncertainty with rigor, and their guidance laid a critical foundation for my subsequent work. I also thank Dr. Huilin Zhao, Dr. Jin Zhang, Dr. Beibei Sun, and Dr. Yan Zhou for their collaboration and clinical insights.

This work was supported in part by grants from the National Institutes of Health. I gratefully acknowledge this support.

I am thankful for the opportunity to present my work at the Information Processing in Medical Imaging (IPMI) 2023 conference, where it was nominated for the Best Paper Award. The feedback and encouragement I received from the community, including from scholars such as Dr. Nassir Navab, helped me see my work as part of a broader scientific conversation. Research finds its meaning not in isolation, but in being examined and understood by others.

During my doctoral study, I also pursued a Master's degree in Statistics, which greatly enriched my perspective on data and inference. I would like to thank Dr. Yen-Chi Chen (UW) and Dr. Jiangang Ying (Fudan) for their teaching and support.

I am also grateful to my mentors and colleagues during my internships at United Imaging Intelligence America and TikTok, whose guidance broadened my perspective beyond academia.

Finally, I would like to express my deepest gratitude to my family. Throughout the years I spent abroad, their unwavering support from afar has sustained me in ways that are difficult to fully articulate. Distance, over time, has reshaped my understanding of presence and connection, and I have come to realize that no journey is ever taken alone.

I am equally thankful for my friends in Seattle, Baltimore, Minneapolis, Durham, New Haven, Austin, Boston, Chapel Hill, Shanghai, Hefei, and Beijing, whose companionship made these years not only productive, but also deeply memorable.

As this chapter of my life draws to a close, I find it increasingly difficult to separate what has been achieved from how it has been experienced. The years of doctoral study are often measured in results and milestones, yet what remains more enduring are the ways in which

they reshape one's thinking, one's patience, and one's understanding of uncertainty. Looking back, the path was neither linear nor certain, but it is precisely through this process that I have learned to navigate complexity with greater clarity.

What once felt like an individual pursuit now appears, in retrospect, as something sustained by many visible and invisible forms of support. With this awareness, I move forward not only with a sense of gratitude, but also with a renewed commitment to the work that lies ahead. There remains much to learn, much to explore, and much to contribute, and it is this open horizon that I embrace with both curiosity and resolve.

Five years ago, Mr. Wang arrived on this continent, full of anticipation and uncertainty, stepping into a world both unfamiliar and promising. Five years later, Dr. Wang moves forward with clarity, purpose, and the courage to pursue whatever lies ahead.

Chapter 1

MEDICAL IMAGE LEARNING UNDER INCREASING HETEROGENEITY

1.1 Heterogeneity in Medical Imaging

Medical imaging has become central to modern clinical decision-making, supporting diagnosis, treatment planning, disease monitoring, and longitudinal assessment across a wide spectrum of conditions. The rapid growth in imaging volume, modality diversity, and acquisition complexity has rendered manual analysis increasingly impractical, making automated and learning-based approaches indispensable components of contemporary medical workflows [189]. In particular, the emergence of deep learning has substantially reshaped the landscape of medical image analysis, enabling end-to-end learning of complex image representations and achieving state-of-the-art performance across a wide range of tasks, such as classification, registration, and segmentation [119, 35, 148].

These successes, however, often rest on an implicit assumption that observable variability does not fundamentally disrupt the coherence of the underlying representation produced by neural networks. In many medical image analysis tasks, desired outputs depend on task-relevant generative properties of the underlying biological system. Yet the observed image is not a direct encoding of these properties alone; rather, it reflects a composite realization of intrinsic generative factors together with extrinsic influences arising from acquisition and data provenance, such as imaging protocols, spatial distortions, population diversity, scanner characteristics, and institutional practices. These influences operate across multiple levels and interact in complex ways, which we refer to collectively as *heterogeneity*. To reason about this composite realization in a precise and consistent manner, we introduce a taxonomy that distinguishes major sources of heterogeneity:

- **Appearance heterogeneity**: variability in how the same underlying biological content is *observed in intensity space*, with spatial configuration treated as fixed.
  - **Intensity-level**: appearance variability within the same modality and nominal contrast setting (e.g., T1-weighted MRI), driven by scanner/vendor characteristics, acquisition parameters, noise statistics, reconstruction pipelines, and other factors that shift intensity distributions and local appearance patterns.
  - **Contrast-level**: appearance variability induced by different contrast mechanisms within a modality family (e.g., T1-weighted vs. time-of-flight MRI; non-contrast vs. angiographic CT), where intensity semantics shift because different tissue properties and physical effects are emphasized.
  - **Modality-level**: appearance variability induced by fundamentally different measurement principles (e.g., MRI vs. CT), for which intensity semantics are not directly comparable.
- **Geometric heterogeneity**: variability in *spatial configuration* across images.
  - **Registration-compatible**: spatial misalignment across images depicting the same anatomical target and admitting meaningful correspondence. Images share an underlying common organ-level shape but differ in deformation due to motion, positioning, or inter-subject variability, so alignment is well-posed in principle.
  - **Correspondence-free**: cases in which reliable cross-image correspondence is unavailable even within the same anatomical region, making direct alignment ill-posed. This includes substantial differences in anatomical extent, slice location, field-of-view, or morphology (e.g., apical vs. basal cardiac slices) such that images cannot be assumed to share a common coordinate system; observations are effectively unpaired, and registration-based coupling is not applicable.

The taxonomy above makes explicit that medical images may vary along multiple, qualitatively distinct dimensions. As these dimensions accumulate, the notion of a stable representation becomes increasingly nontrivial. The remainder of this chapter develops the conceptual foundation for analyzing this challenge. We first formalize a unifying represen-

tational requirement, then examine the limitations of widely used empirical strategies in increasingly heterogeneous settings, and finally outline how the projects of this thesis instantiate this requirement across progressively more demanding regimes.

1.2 The Representational Challenge Posed by Heterogeneity

One key question is why such variability systematically destabilizes learned representations and what must be imposed on a model so that stability can be maintained as heterogeneity intensifies. The challenges posed by medical image heterogeneity can be understood from a representation learning perspective. As illustrated in Fig. 1.1, observed medical images can be viewed as the combination of two components: the underlying anatomical structure and various sources of heterogeneity.

Robust learning therefore requires representations that preserve task-relevant anatomical structure while remaining invariant to heterogeneous observation factors. This thesis approaches the problem through structured latent representation learning. As the complexity of heterogeneity increases, from appearance variation to geometric variation, more structured latent representations are required to maintain identifiable information.

1.2.1 Identifiable Invariant Preservation

Robustness in medical image analysis requires preserving the latent properties that determine the desired output, while allowing other sources of variability to change without corrupting the internal representation. The central challenge introduced by heterogeneity is that task-relevant generative properties and observational variability are simultaneously present in the same image, and their effects are interleaved in the pixel domain.

We therefore use *identifiable invariants* to denote latent properties that should be consistently encoded across heterogeneous observations for a fixed task (e.g., anatomical boundaries for segmentation, reference shapes for registration). At the same time, we treat acquisition- and environment-dependent influences as *observational variability* that should not dominate the internal representation. The goal is to suppress such variability, or, more precisely, pre-

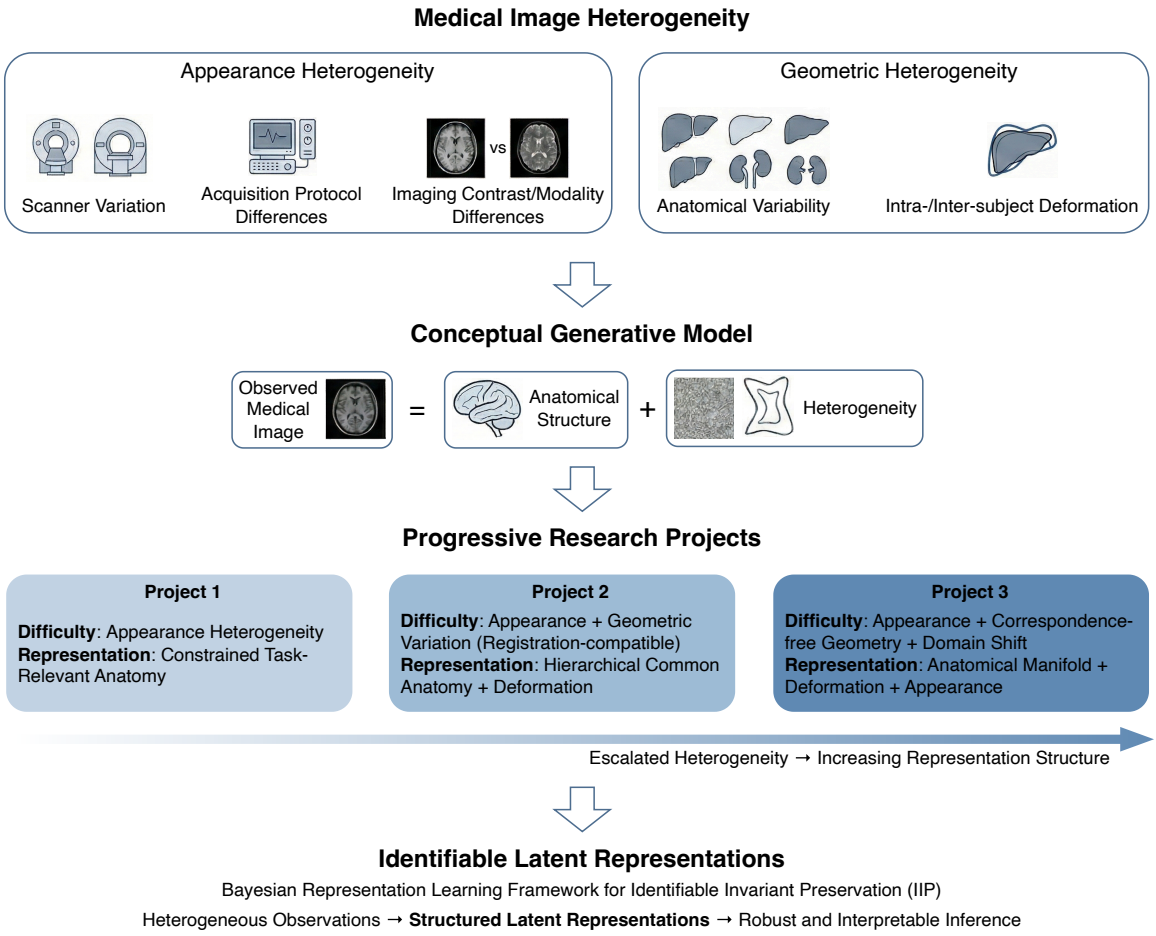


Figure 1.1: Overview of the thesis framework. Medical images can be viewed as the composition of underlying anatomical structure and heterogeneous observation factors arising from acquisition conditions and anatomical variability. As heterogeneity complexity increases, increasingly structured latent representations are required to preserve identifiable information. The three projects in this thesis progressively address appearance heterogeneity, geometric variation, and domain shifts through increasingly structured probabilistic representations.

vent it from serving as a surrogate for task-relevant content when the two are correlated in the training data, with the challenge that the underlying sources of variability are latent and unannotated. We refer to this requirement as *identifiable invariant preservation* (IIP).

1.2.2 Limitations of Empirical Compensation

A common response to IIP is to increase data scale, strengthen regularization, and refine objective functions so that empirical invariances are enhanced during training. For example, robustness is often pursued by expanding training cohorts through multi-center aggregation or large-scale pretraining to improve coverage, while introducing regularization such as weight decay [91], dropout [142], label smoothing [147], or strong augmentation policies, and refining objectives with Mixup/CutMix [184, 183], consistency-based losses (e.g., distillation [70]), and contrastive self-supervision [157]. Such strategies can be effective in moderate regimes, particularly when the relevant variability is well-covered by the training distribution and the model can interpolate.

However, these remedies do not directly constrain how latent sources of variation are internally organized. As heterogeneity increases, two failure modes become more pronounced. First, representations may exploit observational idiosyncrasies that correlate with supervision signals, yielding shortcuts that appear predictive in-distribution but are unstable under shifts in observation conditions. Second, even when additional objectives are introduced, such as generic consistency regularization, distribution-matching penalties, or alignment losses, the induced latent organization can remain underdetermined, leaving task-relevant content entangled with observational variability.

These limitations are not specific to a particular architectural family or training objective. They reflect the fact that empirical compensation alone does not ensure that identifiable invariants are encoded in a manner that remains stable when heterogeneity accumulates across multiple levels.

1.2.3 Identifiable Latent Factorization

In light of the limitations of empirical compensation in existing medical image analysis approaches, we posit that stability under increasing heterogeneity rests on a more foundational principle: *identifiable factorization* of latent sources of variation. By identifiable, we mean

that the roles of latent components are sufficiently constrained so that task-relevant generative properties can be consistently represented across heterogeneous observations, rather than being arbitrarily redistributed among latent variables or absorbed by nuisance components.

This requirement is stronger than post-hoc interpretability, because latent components are introduced with explicit semantics, and the model is constrained so that these semantics remain meaningful under changes in observation conditions. When such factorization is weakly constrained, heterogeneity can cause latent components to drift, entangle, or collapse, undermining representational coherence even if training objectives are satisfied.

1.2.4 Project Roadmap

This thesis advances medical image analysis by introducing identifiable latent factorization as a unifying representational principle. Beyond proposing isolated algorithms for individual tasks, this work argues that many challenges arising from heterogeneous medical imaging stem from a common representational problem: models must separate invariant latent structure from nuisance variability in a manner that remains identifiable. By formalizing this requirement as a design criterion, the thesis provides a general principle for constructing models that remain interpretable and robust under heterogeneous observation conditions.

As summarized in Fig. 1.1, to demonstrate the generality of this principle, the thesis instantiates it across three medical image analysis projects spanning classification (Chapter 3), registration (Chapter 4), and segmentation (Chapters 3 and 5), and covering both supervised (Chapter 3) and unsupervised (Chapters 4 and 5) learning settings. These projects are intentionally organized as a progression of increasing heterogeneity. As the variability of the data grows, the latent representation must be structured more explicitly to preserve identifiable invariants. This progression illustrates that identifiable latent factorization is not tied to a specific task or method family, but instead serves as a reusable design principle that scales with problem difficulty.

Across these settings, enforcing identifiable latent organization leads to models whose internal structure is semantically meaningful while remaining stable under changing ob-

Table 1.1: Project roadmap through the lens of identifiable latent factorization. The projects form a progression of increasing heterogeneity: each project subsumes the variability addressed by earlier ones while introducing additional, more challenging sources of heterogeneity, and correspondingly stronger mechanisms for stabilizing task-relevant latent properties under variable observation conditions.

Project	**Heterogeneity Escalation**	**Latent Factorization**
1	intensity-level appearance heterog.	task-relevant anatomy, other task-irrelevant information
2	contrast/modality-level appearance heterog., registration-compatible geometric heterog.	reference anatomy, image-specific deformation, image appearance
3	correspondence-free geometric heterog.	global anatomical prototypes, prototype selection, image-specific deformation, image appearance

servation conditions. The resulting formulations improve interpretability, robustness, and predictive performance, demonstrating that the contribution of this thesis lies not in isolated technical solutions but in establishing a coherent representational framework for heterogeneous medical image analysis.

We also summarize the projects in terms of the additional sources of heterogeneity we tackle and the latent factorization performed by our methods in Tab. 1.1. The subsequent chapters develop the methodological contributions underlying these projects. Chapter 2 formalizes the generative perspective and the constraints that promote identifiable factorization. Chapters 3 to 5 then present the three projects in detail, demonstrating how increasingly explicit latent organization supports robustness and interpretability under escalating heterogeneity.

Chapter 2

BAYESIAN REPRESENTATION LEARNING

The projects in this thesis are developed under a common methodological stance: stable learning under heterogeneity requires modeling assumptions that explicitly organize latent sources of variation and make their roles inferable from data. This chapter introduces the high-level concepts and technical vocabulary used throughout the remainder of the thesis. We first review latent variable modeling as a general framework for representing observed data through unobserved explanatory variables. We then introduce the Bayesian perspective on latent inference, and finally discuss how explicit modeling assumptions can be used to promote identifiable and disentangled representations.

This chapter is independent of particular choices of latent factors or task-specific decomposition. It provides a unified conceptual and mathematical scaffold, including probabilistic graphical representations and core inference objectives, that will be instantiated in subsequent chapters.

2.1 Latent Variable Models

Latent variable models describe observed data as the result of unobserved explanatory variables interacting through a probabilistic generative mechanism. Let $\mathbf{x} \in \mathcal{X}$ denote an observation and $\mathbf{z} \in \mathcal{Z}$ denote a latent variable that generates $\mathbf{x}$. A basic latent variable model specifies a joint distribution

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})\, p(\mathbf{x}|\mathbf{z}), \tag{2.1}$$

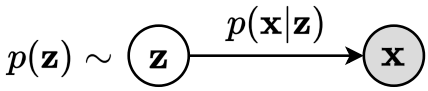


Figure 2.1: A basic latent variable model. A latent variable $\mathbf{z}$ generates an observation $\mathbf{x}$ through a likelihood $p(\mathbf{x}|\mathbf{z})$ under a prior $p(\mathbf{z})$. The observed variable $\mathbf{x}$ is shaded.

where $p(\mathbf{z})$ is a prior distribution over latent variables and $p(\mathbf{x}|\mathbf{z})$ is a likelihood that maps latent variables to the observation space. The corresponding marginal likelihood is obtained by integrating out the latent variables:

$$p(\mathbf{x}) = \int p(\mathbf{z})\, p(\mathbf{x}|\mathbf{z})\, \mathrm{d}\mathbf{z}. \tag{2.2}$$

This formulation separates *explanation* from *realization*. The latent variables $\mathbf{z}$ encode explanatory degrees of freedom, while the likelihood specifies how those degrees of freedom manifest in observations. When variability in the data arises from multiple sources, latent variable models provide a natural mechanism to represent such variability without requiring it to be directly observed or annotated. In this sense, latent modeling offers a principled alternative to treating all variation as an undifferentiated nuisance absorbed into a single feature space.

A convenient way to express the conditional independence structure implied by a latent model is through a probabilistic graphical model. The basic model above corresponds to the directed graph $\mathbf{z} \to \mathbf{x}$, indicating that $\mathbf{z}$ generates $\mathbf{x}$, as shown in Fig. 2.1. We will frequently use such diagrams to clarify modeling assumptions, especially when multiple latent variables are introduced.

Given observations, the central inferential object is the posterior distribution over latent variables,

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})}. \tag{2.3}$$

The posterior describes which latent configurations are compatible with a given observation

under the model. In many practical settings, the posterior is intractable to compute exactly because $p(\mathbf{x})$ involves a high-dimensional integral. This motivates approximate inference methods, which will be discussed in the next section.

Even at this abstract level, a key conceptual point is that the usefulness of a latent representation depends on two coupled components: the *generative specification* $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ and the *inference mechanism* used to recover $\mathbf{z}$ from $\mathbf{x}$. If the generative specification is too flexible or weakly constrained, multiple distinct latent explanations can produce similar observations, yielding an underdetermined posterior. Conversely, if the inference mechanism is poorly matched to the generative assumptions, the inferred latents may fail to reflect the intended semantics. These issues are central to the notion of identifiability developed later in this chapter, and they motivate why explicit modeling assumptions matter even before committing to any task-specific factorization.

2.2 Variational Bayesian Inference

Latent variable models become practically useful when we can infer latent explanations from data and learn model parameters from observations. From a Bayesian perspective, inference is fundamentally posterior reasoning. Given a joint model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ with parameters θ, the key quantity is the posterior

$$p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{x})}, \tag{2.4}$$

where the marginal likelihood $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z})\, \mathrm{d}\mathbf{z}$ is typically intractable in high-dimensional settings, as mentioned before. Variational Bayes provides a scalable alternative by replacing exact posterior inference with optimization over an approximating family.

2.2.1 Variational Approximation

Variational inference introduces an approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ parameterized by ϕ and seeks to bring it close to the true posterior by minimizing the Kullback–Leibler (KL)

divergence to the true posterior:

$$q_\phi^\star(\mathbf{z}|\mathbf{x}) = \arg\min_{q_\phi(\mathbf{z}|\mathbf{x})} \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x}))\,. \tag{2.5}$$

Directly optimizing this objective is inconvenient because it involves the intractable posterior normalization through $p_\theta(\mathbf{x})$. This difficulty is resolved by the standard variational decomposition obtained from Bayes' rule. For any choice of $q_\phi(\mathbf{z}|\mathbf{x})$,

$$\log p_\theta(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}))}_{\triangleq\ \mathcal{L}(\theta,\phi;\mathbf{x})} + \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x}))\,, \tag{2.6}$$

where $\mathcal{L}(\theta, \phi; \mathbf{x})$ is the evidence lower bound (ELBO). Since $\log p_\theta(\mathbf{x})$ does not depend on ϕ, maximizing the ELBO with respect to ϕ is equivalent to minimizing $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x}))$. This yields a tractable optimization objective that depends only on the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, the prior $p_\theta(\mathbf{z})$, and the variational family $q_\phi(\mathbf{z}|\mathbf{x})$.

While the argument above explains why optimizing ϕ improves posterior approximation for a fixed generative model, in representation learning the generative parameters θ are themselves unknown and must be learned from data. We therefore maximize $\mathcal{L}$ jointly with respect to θ and ϕ, which performs two coupled tasks: the first term encourages latent configurations that explain the data under the generative model, while the KL term regularizes the inferred latents toward the prior and thereby implements the assumed organization of latent variation.

2.2.2 Amortized Inference

In modern representation learning, the approximate posterior is typically amortized: rather than optimizing a separate variational distribution for each observation, one learns a shared inference function $q_\phi(\mathbf{z}|\mathbf{x})$ (often implemented as a neural network) that maps observations to posterior parameters. Under amortization, learning is driven by the dataset

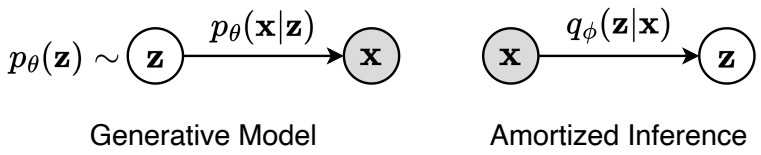


Figure 2.2: Paired graphical view of variational Bayes. Left: a generative model $p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. Right: an amortized variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ that enables scalable posterior inference.

objective

$$\max_{\theta,\phi} \sum_{n=1}^{N} \mathcal{L}(\theta, \phi; \mathbf{x}_n), \tag{2.7}$$

which couples representation learning (through $\mathbf{z}$) with Bayesian inference (through q_ϕ) under explicit generative assumptions.

Amortized inference provides a practical interface for encoding inductive bias. The prior $p_\theta(\mathbf{z})$, the likelihood family $p_\theta(\mathbf{x}|\mathbf{z})$, and the variational family $q_\phi(\mathbf{z}|\mathbf{x})$ jointly determine what latent explanations are available and how stably they can be recovered. This point is critical for identifiability: if the model admits multiple incompatible latent explanations for the same observation, amortized inference may converge to solutions that are predictive yet semantically unstable.

2.2.3 Graphical Representation

The relationship between the generative model and the variational approximation is conveniently summarized by a pair of graphical models: a directed generative graph (priors and likelihood) and an inference graph (approximate posterior), as shown in Fig. 2.2. We will use this paired view throughout the thesis to keep generative assumptions and inference mechanisms explicit.

The next section uses this variational-Bayesian perspective to motivate identifiable and disentangled representations as consequences of explicit modeling assumptions rather than post-hoc explanations.

2.3 Identifiability and Disentanglement as Modeling Assumptions

Latent variable models provide a vocabulary for explaining observations through unobserved factors, and variational Bayes provides a scalable mechanism to infer such factors. However, the existence of latent variables alone does not guarantee that the inferred representation is stable, interpretable, or reusable under heterogeneous observation conditions. These properties depend on *identifiability*: whether the roles of latent components are sufficiently specified so that their semantics can be recovered consistently from data.

2.3.1 Identifiability

At a high level, identifiability concerns whether distinct parameterizations or latent decompositions can generate the same distribution over observations. If multiple latent explanations are equally compatible with the data, then the posterior can be underdetermined: latent components may permute, drift, or redistribute explanatory responsibility without changing the likelihood. In such cases, a learned representation may be predictive in-distribution yet unstable when observation conditions change.

In representation learning, we often interpret latent variables as carrying specific semantic content. Such interpretation is meaningful only when the model restricts the space of admissible explanations. Identifiability therefore functions as a *design criterion*: modeling assumptions must constrain latent variables so that their inferred roles are stable across observations, rather than being arbitrary artifacts of optimization.

2.3.2 Disentanglement

The term *disentanglement* is often used to describe representations in which different latent components correspond to different explanatory sources. In this thesis, disentanglement is treated as an explicit modeling assumption rather than a post-hoc interpretation. Concretely, disentanglement requires specifying, through the prior, likelihood, and inference structure, which latent components are intended to vary independently and how they should

contribute to the generation of observations.

Additional constraints are often needed to prevent latent components from exchanging their explanatory responsibilities. These constraints can be implemented through architectural structure, hierarchical priors, structured variational families, or regularization terms that encode how information should be distributed across latent components.

2.3.3 Failure Modes under Weak Specification

When latent roles are weakly specified, several characteristic failure modes emerge. One common failure is *semantic drift*: latent components adopt different meanings across training runs, across subsets of data, or across changes in observation conditions, even though the overall likelihood remains high. Another is *entanglement*: multiple explanatory sources become fused within the same latent variables, so that changing one underlying factor induces changes across multiple latent dimensions. A third is *posterior collapse*, in which the variational posterior becomes insensitive to the observation (e.g., $q_\phi(\mathbf{z}|\mathbf{x}) \approx p(\mathbf{z})$), indicating that the model has learned to explain the data without meaningfully using the latent representation.

These phenomena are closely linked to heterogeneity. As observation conditions diversify, latent explanations that were sufficient under a narrow distribution may cease to be stable. Without identifiable organization, a model can satisfy training objectives by shifting explanatory burden onto observational cues that correlate with task outputs, yielding representations that do not preserve invariant content in a reusable way.

These observations suggest that identifiability is controlled by how the generative specification and the inference approximation are *coupled*.

2.4 Modeling Patterns for Identifiability

The preceding section emphasized that identifiability is governed by how generative specification and inference are coupled. This section summarizes a set of reusable modeling patterns that strengthen this coupling in Bayesian representation learning. The intent is not

to prescribe a single recipe, but to highlight design primitives that make latent roles more recoverable from data.

2.4.1 Priors

A prior $p(\mathbf{z})$ encodes assumptions about latent organization. In multi-component settings, a common starting point is a factorized prior

$$p(\mathbf{z}) = \prod_{k=1}^{K} p(\mathbf{z}_k), \tag{2.8}$$

which asserts a baseline independence across components. When this assumption is too weak to induce stable roles, one may strengthen role specification using structured prior, such as explicit allocation variables that select or gate latent substructures. Abstractly, one may write

$$p(\mathbf{z}, \mathbf{a}) = p(\mathbf{a})\, p(\mathbf{z}|\mathbf{a}), \tag{2.9}$$

where $\mathbf{a}$ parameterizes how latent explanations are composed. Such allocation variables can be discrete (mixture-style) or continuous (attention- or weighting-style), but the shared purpose is to make the space of admissible explanations more structured, thereby reducing underdetermination.

2.4.2 Likelihood

Even with a well-chosen prior, identifiability can be weak if the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ allows multiple latent explanations to produce similar observations. Likelihood structure is therefore a primary lever for role separation. A generic pattern is to introduce intermediate latent representations and define the likelihood through a structured generator,

$$\mathbf{u} = g_\theta(\mathbf{z}), \qquad \mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{u}), \tag{2.10}$$

where $\mathbf{u}$ is an explicitly organized intermediate variable (or a collection thereof). When $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_K)$ contains multiple components, g_θ can be designed to combine them through distinct pathways, thereby constraining how each component influences the observation.

A related pattern is compositional generation through conditionally independent sub-observations. Let $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$ denote multiple correlated views or measurements. One may specify

$$p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{m=1}^{M} p_\theta\big(\mathbf{x}^{(m)}|\mathbf{z}, \mathbf{c}^{(m)}\big)\,, \tag{2.11}$$

where $\mathbf{c}^{(m)}$ denotes view-specific conditions or parameters. This form does not assume that views are independent in reality; rather, it expresses a modeling choice that isolates view-dependent variability from shared latent explanations, improving recoverability of the shared content.

2.4.3 Variational Posteriors

Identifiability also depends on the variational family used to approximate the posterior. Mean-field approximations,

$$q_\phi(\mathbf{z}|\mathbf{x}) = \prod_{k=1}^{K} q_{\phi_k}(\mathbf{z}_k|\mathbf{x}), \tag{2.12}$$

often provide a useful computational baseline, but they can be mismatched when the generative model implies strong posterior dependencies. A general alternative is to adopt structured variational families that mirror conditional independences in the generative graph. Using a chain-rule factorization,

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) \prod_{k=2}^{K} q_\phi(\mathbf{z}_k|\mathbf{z}_{<k}, \mathbf{x}), \tag{2.13}$$

allows selected dependencies to be retained.

A second alignment pattern is to share or separate inference pathways in correspondence with intended latent roles. For example, one may parameterize different subsets of latent

variables using distinct inference modules, while keeping a shared trunk for common evidence extraction. Such architectural alignment is not merely an implementation choice; it defines which evidence is available to each latent component and thereby influences the stability of semantic roles.

2.4.4 Objective Shaping

Within variational Bayes, the ELBO couples data explanation and adherence to the prior through the KL term. A widely used pattern is to explicitly modulate this trade-off, e.g.,

$$\mathcal{L}_\beta(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \, \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})) \,, \tag{2.14}$$

where $\beta > 0$ controls the strength of latent regularization. Such weighting does not guarantee disentanglement, but it provides a direct mechanism for controlling information flow into latent variables and can prevent pathological solutions (e.g., overly flexible posteriors or degenerate use of latents) when combined with appropriate structural assumptions.

More generally, objective shaping can introduce additional terms that operationalize intended semantics, provided these terms are consistent with the latent organization implied by the prior and likelihood. In this thesis, such additions are used sparingly and only when they serve to reinforce a specified latent role.

2.4.5 Summary: Coupling Generation and Inference

Across these patterns, a unifying theme is explicit coupling between generation and inference. Priors specify admissible latent roles, likelihood structure constrains how roles manifest in observations, and variational families determine how roles are inferred from data. When these components are designed coherently, the resulting representation is more likely to exhibit identifiable semantics and to remain stable as observation conditions diversify. The subsequent project chapters instantiate these patterns in different regimes, progressively strengthening role specification as heterogeneity escalates.

Chapter 3

RESTRICTING FEATURE COMPLEXITY FOR ROBUST MRI SEGMENTATION

This chapter presents the first project of the thesis: automated segmentation of intracranial arterial calcification from multi-contrast brain MRI. Within the broader framework introduced in Chapter 1, this setting corresponds to the most moderate regime of heterogeneity considered in this work. The task follows the standard supervised medical image segmentation paradigm, where pixel-wise annotations are available during training and multiple co-registered MRI contrasts are provided as input channels. These contrasts serve as complementary sources of information: accurate calcification delineation typically relies on integrating structural cues from the vessel lumen, vessel wall, and surrounding anatomical context across contrasts.

Despite this seemingly conventional setup, the segmentation problem itself remains intrinsically difficult. In MRI, calcification does not exhibit a distinctive or stable signal signature. Instead, it typically appears as subtle low-intensity regions that can be easily confused with background or other anatomical structures. Reliable segmentation therefore cannot rely on simple intensity-based cues. Instead, the model must infer calcification presence indirectly from contextual relationships among neighboring tissues and vascular structures. This reliance on subtle relational patterns makes the task particularly sensitive to *intensity-level appearance heterogeneity*: scanner- and protocol-dependent variations can alter local intensity patterns and tissue contrasts in ways that distort or obscure the already weak signals used for inference. As a result, even moderate appearance variability across scans can significantly degrade the robustness of conventional segmentation pipelines.

From a representation learning perspective, the key challenge is therefore to construct

internal representations that remain stable under such appearance fluctuations while still capturing the structural information required for the segmentation task. In particular, the representation should emphasize task-relevant anatomical structure, such as vessel geometry and contextual spatial relationships, while avoiding reliance on acquisition-dependent appearance patterns that do not correspond to the underlying target.

To address this challenge, we adopt the Bayesian representation learning framework introduced in Chapter 2. The central idea is to explicitly organize the latent representation into components that account for task-relevant structural information and those that capture nuisance variability arising from appearance fluctuations or unrelated anatomical signals. Within a variational formulation, latent variables mediate the mapping from observed images to segmentation outputs, allowing the model to distribute explanatory responsibility across these components in a controlled manner.

The practical mechanism for enforcing this organization is a constraint on representational complexity. By limiting the effective capacity of the latent representation, the model is discouraged from encoding spurious appearance patterns as shortcuts for predicting the target. Instead, it is encouraged to compress the input information into a representation that preferentially preserves structural cues that are consistently predictive of calcification across heterogeneous scans. As demonstrated in the following sections, this constrained latent organization improves robustness under appearance variability while simultaneously yielding a more interpretable internal representation of the task.

The remainder of this chapter closely follows the associated journal publication:

- Xin Wang et al. Automated mri-based segmentation of intracranial arterial calcification by restricting feature complexity. Magnetic Resonance in Medicine, 93(1):384–396, 2025. doi: 10.1002/mrm.30283 [168].

3.1 Introduction

Intracranial atherosclerosis is a leading cause of ischemic stroke [5]. Atherosclerotic plaque may contain various plaque components that determine the risk of stroke. Among these plaque components, arterial calcification is known to be of pathological significance in stroke, and also associated with other diseases such as dementia and cognitive decline [26, 28, 25]. Therefore, the identification and segmentation of calcification are important in vascular image analysis for diagnosis and risk assessment.

Automated calcification segmentation on non-contrast computed tomography (CT) or contrast-enhanced CT angiography (CTA) has been extensively explored in the literature, where calcification has a high, easy-to-detect signal. For example, Lessmann et al. [92] connected two networks to label and refine calcification on chest CT. Graffy et al. [56] utilized Mask R-CNN to identify aortic calcification on abdominal CT. Weng et al. [171] and Bortsova et al. [24] used U-Nets to segment calcification around superficial femoral arteries and internal carotid arteries, respectively. In general, these studies commonly relied on existing deep networks established for general medical image segmentation tasks.

Despite its widespread use, CT exhibits inherent limitations which diminish its utility in intracranial arterial calcification studies.

- **Safety**. CT involves patient exposure to ionizing radiation, and CTA requires the administration of invasive contrast agents, especially when longitudinal scans are essential to monitor disease progression [27].
- **Comprehensive analysis with other plaque features**. Intracranial arterial calcification often coexists and interacts with other atherosclerotic plaque components, such as lipid core and fibrous tissue [178]. However, arteries cannot be visualized with non-contrast CT. Even on CTA that uses contrast, distinguishing the type of calcification (e.g., intimal or medial) and identifying other plaque components are challenging [4].

In contrast, MRI can overcome these disadvantages. Vessel wall MRI (VWI) allows for safe,

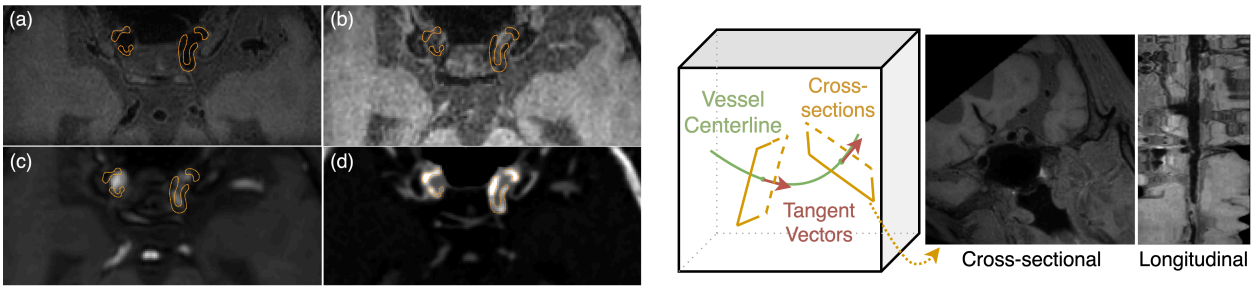


Figure 3.1: Original and preprocessed images. **Left**: Axial slices of intracranial scans from multi-sequence MRI and CT angiography (CTA). (a) T1-weighted, (b) Simultaneous Non-contrast Angiography and intraPlaque hemorrhage imaging (SNAP) [136], (c) Time-of-flight (TOF) MR angiography (MRA), (d) CTA. Calcification is delineated with orange contours. **Right**: Extraction of 2D cross-sectional slices perpendicular to the vessel centerlines of 3D scans. An example slice from a T1 image and the corresponding longitudinal view of extracted slices from the same subject are shown on the right side.

non-invasive, and radiation-free imaging of various types of atherosclerotic plaques [106, 113], preferable for serial monitoring of atherosclerosis [63]. Detection of calcification on VWI enables a clinically significant joint analysis of different plaque components. Therefore, there has been a growing demand for MRI-based assessment of arterial calcification [132], which our method can address.

However, it is difficult to detect calcification on MRI. As shown in Fig. 3.1, calcification appears dark in most MR sequences and is difficult to distinguish from noise. For example, on T1-weighted VWI, when the signal of flowing blood in the vessel lumen is suppressed, calcification adjacent to the lumen can be hard to identify; on time-of-flight MR angiography (TOF MRA), all tissue signals except flowing blood are generally suppressed. Therefore, a single MRI contrast may not suffice to segment calcification. Nonetheless, the various imaging patterns shown on different MR sequences still provide rich tissue information that may have value in calcium identification. Similar to the aforementioned CT-based segmentation, it is straightforward to train an existing network using multi-sequence MRI, by stacking sequences as a multi-channel input image. However, such a naive solution could be suboptimal, as shown in the result section.

To improve the segmentation beyond such black-box networks, we propose to *refine* the network features extracted from the MR images. Following the idea of the information bottleneck (IB) [150], we hypothesize that it is crucial for the MRI features to exhibit: 1) rich information about the target tissue (calcification); 2) constrained complexity essentially excluding irrelevant information. Such features are potentially more generalizable and better for segmentation [52]. In contrast, a naive network only satisfies the first requirement by a regular segmentation loss, but its features could be too complex, presenting a challenge in prioritizing the learning of calcification patterns, consequently impairing performance.

To reduce the complexity of MR features, we use regularization to concentrate the information in MR features on calcification. Assume we have an auxiliary image (spatially aligned to the MRI) that displays calcification but few other tissue structures. Considering features extracted separately from the MR and the auxiliary images, even though both contain information of calcification, the auxiliary feature may exhibit lower complexity due to its image consisting of simpler structures. By *aligning* the MR feature with the auxiliary one, we could *refine* the MR feature, reducing its complexity.

Such alignment of features can be achieved using variational autoencoders (VAEs) [88], which extract stochastic features (probability distributions) from input data to facilitate downstream tasks such as reconstruction, detection and segmentation. VAEs are capable of refining and learning features with compressed information [2]. In this work, we propose a novel VAE for MRI-based calcification segmentation. Particularly, during network training, we incorporate the ground-truth calcification mask as the auxiliary image input to the network. We explicitly align the MR and auxiliary features by minimizing their divergence, such that the model learns to extract MR features with a lower complexity. Once training is done, auxiliary images are no longer needed, and the network can segment calcification using only MR images during testing.

The contributions of this article are summarized as follows:

1. To the best of our knowledge, this is the first automated method for MRI-based intracranial arterial calcification detection. Notably, our model is theoretically grounded

Table 3.1: Baseline characteristics of the studied cohort. The calcification-related values are for internal carotid arteries and middle cerebral arteries, and the calcification (cal.) volume is reported as the median with interquartile range.

Characteristic	Value	Characteristic	Value	Characteristic	Value
No. of subjects	113	Smoking	23	Ischemic stroke	85
No. of women	15	Hypertension	82	Transient ischemic attack	8
Age (y)	64.8 ± 8.5	Diabetes mellitus	46	Stenosis	72
Cal. volume (mm^3)	18.9 (3.2-50.5)	Hyperlipidimia	5	Aneurysm	7
Presence of cal.	99	Coronary heart disease	8		

in the VAE framework; our strategy of feature complexity reduction provides an interpretable way to improve performance.

2. We demonstrate the superiority of our model compared to multiple widely-used state-of-the-art approaches for segmentation, and highlight the clinical significance of our work in predicting high-level measurements, including calcium volume and slice-wise calcification occurrence.
3. We quantitatively explore the effect of different MR sequences on enhancing calcification identification.

3.2 Methods

3.2.1 Data acquisition

We used a dataset of 113 subjects scanned at Renji hospital, China. Use of the data was approved by the local institutional review board. The subjects underwent multi-sequence intracranial MRI and CTA scans from 2019 to 2020 during their hospitalizations due to different clinical indications. The MRI was performed using a 3T scanner (Philips Ingenia, the Netherlands) with a dedicated 16-channel phased-array carotid artery coil (Beijing TSImaging Healthcare Technology Co., China); the CTA was performed using a 320-detector row scanner (Aquilion ONE VISION, Canon Medical System Corporation, Otawara, Japan). The demographics of the subjects are summarized in Tab. 3.1. The dataset consists of sub-

jects likely to have intracranial vascular diseases, leading to a large diversity of calcification distribution and allowing for a comprehensive assessment of model performance.

Three MR sequences were chosen for this study, *i.e.*, T1-weighted VWI, TOF MRA, and SNAP, based on their potential to visualize calcification. SNAP stands for Simultaneous Non-contrast Angiography and intraPlaque hemorrhage imaging, an MRI technique capable of acquiring a proton-density-weighted image, in which vessel wall and luminal blood are both bright [136]; T1 is typically used to image the vessel wall by suppressing blood flow within the lumen; TOF displays the lumen with high intensity while suppressing other tissues. Calcification produces no MR signal and appears dark on all three sequences. These sequences provide complementary structural information that helps distinguish calcification from other tissues or the lumen.

3.2.2 Preprocessing

Preprocessing of the scans was performed using a previously-developed tool, MOCHA, for multi-contrast vascular image analysis [62]. Specifically, we first used 3D rigid registration to spatially align T1, TOF, SNAP and CTA, with T1 as the reference image. CTA was used only because it can help human readers better visualize calcification when they reviewed the images; it was not used in model training and testing. All images were resampled to be isotropic (spacing of 0.58 mm), and normalized to a fixed intensity window (0-500 for T1, 0-2000 for TOF, and 0-220 for SNAP) for appropriate contrast.

A reviewer (G.C.) with more than ten years of experience then used the TOF sequence to track the centerlines of left and right intracranial internal carotid arteries (ICAs) and the horizontal (M1) and sylvian (M2) segments of the middle cerebral arteries (MCAs), using iCafe, a previously-developed software package for intracranial artery extraction [36].

Then, 2D cross-sectional slices through the tracked centerlines were generated, as shown in Fig. 3.1. A region-of-interest (ROI) with size 80×80 was cropped from the center of each 2D slice to obtain an appropriate field-of-view (containing the vessel but not too large). Thus, for each location on each vessel of each patient, we generated 2D slices of CTA and multiple

MR sequences, referred to as a slice group. For each group, the radiologist meticulously labeled calcification through a systematic process, encompassing the following steps: 1) careful examination of the CTA to identify the locations and shapes of calcification, 2) comprehensive review of all MR sequences to identify on MRI the calcification locations corresponding to the CTA, 3) precise delineation of calcification on MRI according to the boundaries formed by signal contrasts. Note that CTA was used to guide manual labeling, because even experienced radiologists find it challenging to minimize false positives when labeling calcification using MRI alone without concurrent reference to CTA. Besides, due to the blooming effect of CTA and imperfect registration between CTA and MRI in some instances, the calcification labels may not precisely align with the bright shapes on CTA. However, this pragmatic labeling procedure acknowledged the intricacies of the imaging modalities, and prioritized optimal precision of manual labeling on MRI. The manual labels were converted to binary masks as the ground-truth calcification segmentation for MRI.

Subjects were randomly divided into training, validation and test sets, with a ratio of 10:1:1. The vessel lengths among different subjects exhibited variations, leading to diverse numbers of slice groups for each individual. Consequently, the training, validation, and test sets comprised of 39858, 5296, and 2782 slice groups, respectively.

3.2.3 Proposed model

Overall framework

We propose a novel model to improve calcification segmentation on MRI. The overall architecture of our model is illustrated in Fig. 3.2. The inputs to the network are the 2D cross-sectional slices of multi-sequence MRI (x_1) and an auxiliary image (x_2, *i.e.*, the ground-truth segmentation mask, used for training only). The MR feature (denoted by $q_1(\boldsymbol{z}|x_1)$) is extracted via the MR branch based on all MR sequences, and utilized by a decoder to segment calcification. For training, a segmentation loss (cross-entropy) is used to measure the difference between the predicted segmentation probability map and the ground-truth calcification label.

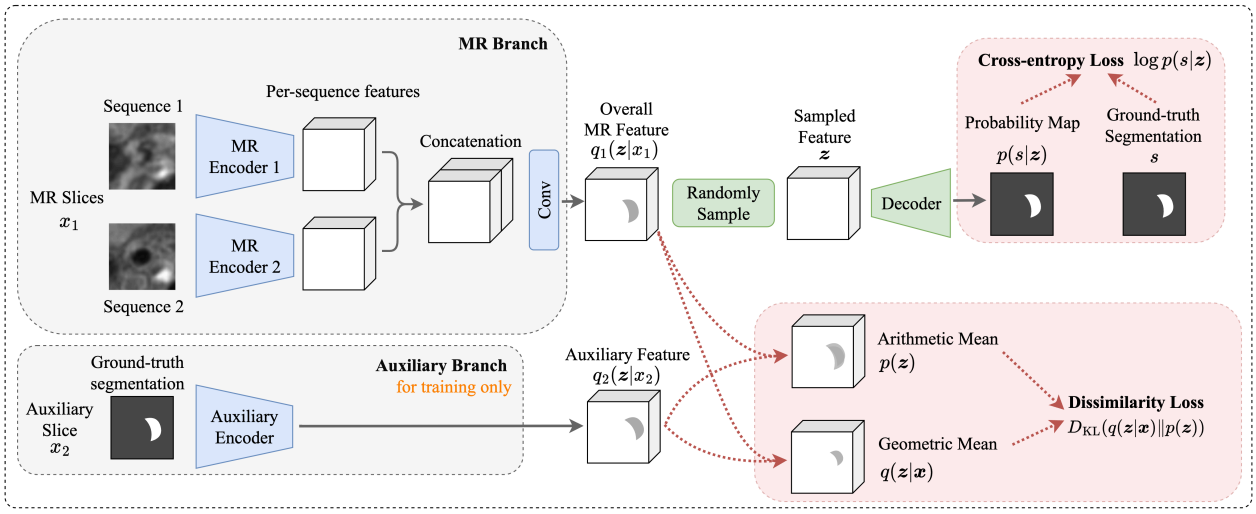


Figure 3.2: The proposed framework (with two MR sequences as an example). Each encoder or decoder is the same as in a U-Net. The cubes represent feature maps. The probability distributions (e.g., q_1, q_2) indicate the correspondence between the features and the terms in the derived objective function (the ELBO). The red arrows and boxes indicate the calculation of the two losses.

In addition to the cross-entropy loss, during training, the auxiliary branch is utilized to extract the feature (denoted by $q_2(\boldsymbol{z}|x_2)$) from the auxiliary image. We introduce a dissimilarity loss to measures the divergence between the MR and auxiliary features, which is detailed in the next section. Trained with this loss, the network can learn to extract an MR feature similar to the auxiliary one. As a binary mask, the auxiliary image has a lower complexity, resulting in simpler features. The dissimilarity loss could therefore restrict the complexity (compress the information) of the MR feature. As we described in Sec. 3.1, this could improve feature generalizability and lead to better segmentation. Since the auxiliary branch is only used for loss calculation, after the training is complete, the model does not require auxiliary images for inference and can segment calcification using only MRI scans.

Note that the model is a variational auto-encoder (VAE). Therefore, the MR and auxiliary features essentially represent two probability distribution q_1 and q_2, respectively. Random sampling from q_1 is involved before the decoder, which will be detailed in the following

sections.

Theoretical explanation

In this section, we describe the theoretical details of our framework. Building on the work of Alemi et al. [2], which demonstrated that VAEs can compress information inside the features in an unsupervised scenario, we extend these concepts to develop a new VAE model for calcification segmentation.

In general, let the input multi-sequence MRI be $x_1 \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, where H_0, W_0 are the height and width of each MR image, respectively, and C_0 is the number of MR sequences; let the corresponding auxiliary image be $x_2 \in \mathbb{R}^{H_0 \times W_0 \times 1}$. Then the input images can be written as $\boldsymbol{x} = (x_1, x_2)$. Following the variational inference framework [88], we further assume the ground-truth segmentation label $s \in \{0, 1\}^{H_0 \times W_0}$ is generated from a latent variable $\boldsymbol{z}$ by a conditional distribution $p(s|\boldsymbol{z})$. Since the true posterior distribution $p(\boldsymbol{z}|s)$ is intractable, a variational posterior $q(\boldsymbol{z}|s)$ is introduced to approximate it.

The VAE and its multimodal variants [135, 174, 146, 154] are well-established techniques to learn features in an unsupervised manner. In this work, we aim to extend this model family for calcification segmentation. Notably, we propose to reduce the information in the feature $\boldsymbol{z}$ by forcing the feature to focus on the segmentation s rather than the entire input images. Thus, we assume $q(\boldsymbol{z}|s) = q(\boldsymbol{z}|\boldsymbol{x})$. Intuitively this means s and $\boldsymbol{x}$ contain the same information of $\boldsymbol{z}$. Then a lower bound of the log-likelihood $\log p(s)$ can be derived as

$$\log p(s) \geqslant \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \log p(s|\boldsymbol{z}) - D_{\mathrm{KL}} \left(q(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}) \right) := \mathrm{ELBO}, \tag{3.1}$$

where $\mathbb{E}$ denotes the mathematical expectation over a specific distribution, $p(\boldsymbol{z})$ is the prior distribution of $\boldsymbol{z}$, and D_{KL} is the Kullback-Leibler (KL) divergence between two distributions. This is called the evidence lower bound (ELBO) [88]. Our goal is to maximize the log-likelihood $\log p(s)$, which is, however, generally intractable. Therefore, the VAE framework aims to maximize its ELBO, which is achieved by our network and will be detailed in the

next section.

We have addressed the MRI and auxiliary together as $\boldsymbol{x} = (x_1, x_2)$, and $q(\boldsymbol{z}|\boldsymbol{x})$ in the ELBO indicates that the extraction of feature $\boldsymbol{z}$ relies on both. However, we aim to build a model capable of segmenting calcification without the auxiliary image. To this end, for each m ($m = 1, 2$ for MRI or auxiliary, respectively), we propose to model an individual posterior distribution $q_m := q_m(\boldsymbol{z}|x_m)$ (*a.k.a.* "expert"). To obtain the final posterior $q(\boldsymbol{z}|\boldsymbol{x})$ and the prior $p(\boldsymbol{z})$, we propose to process the experts in two different ways (namely the geometric [71] and arithmetic [135] means), *i.e.*,

$$q(\boldsymbol{z}|\boldsymbol{x}) \propto \left[\prod_{m=1}^{M} q_m\right]^{\frac{1}{M}}, p(\boldsymbol{z}) := \frac{1}{M}\sum_{m=1}^{M} q_m. \tag{3.2}$$

The intuition is that each expert $q_m(\boldsymbol{z}|x_m)$ corresponds to extracting the feature from either the MRI or the auxiliary. To get an overall estimation of the feature by integrating information from both inputs, we merge the information of the experts by calculating their geometric mean. Our previous work has validated this strategy for multimodal image registration [167, 101], and here we extend it to image segmentation.

Based on the assumptions and derivations above, the objective to be maximized during network training, *i.e.*, the ELBO, can be expressed as

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \log p(s|\boldsymbol{z}) - D_{\text{KL}}\left(C\left[\prod_{m=1}^{M} q_m\right]^{\frac{1}{M}} \| \frac{1}{M}\sum_{m=1}^{M} q_m\right) \tag{3.3}$$

where $q_m := q_m(\boldsymbol{z}|x_m), m = 1, 2$ is the individual posterior of MRI or auxiliary, and C is a constant to normalize the geometric mean, making it a probability distribution.

Network structure

As mentioned in Sec. 3.2.3, the goal of VAE is to maximize the ELBO, and from Eq. (3.3) we see that to calculate the ELBO, we only need to estimate the probability distributions $p(s|\boldsymbol{z})$ and $q_m(\boldsymbol{z}|x_m), m = 1, 2$ (note that $q(\boldsymbol{z}|\boldsymbol{x})$ in Eq. (3.3) is the geometric mean of

q_m's). Therefore, we build a network to infer the distributions from input images. Then, by deploying the negative of the ELBO as the loss function, we can train the network to learn to produce optimal distributions given any input data. Such network is called a variational autoencoder (VAE) [88]. In our case, $q_m(\boldsymbol{z}|x_m), m = 1, 2$ requires two encoding branches to infer the distribution of $\boldsymbol{z}$ separately from MRI (x_1) and auxiliary (x_2), based on which we analytically calculate $q(\boldsymbol{z}|\boldsymbol{x})$ as the geometric mean. Besides, $p(s|\boldsymbol{z})$ requires a decoder to infer the distribution of s (*i.e.*, the segmentation probability map) given any $\boldsymbol{z}$.

Considering these requirements, the design of our VAE is shown in Fig. 3.2. In summary, the auxiliary branch extracts from the auxiliary image an $H \times W \times C$ feature map, where H, W is the height and width, and C is the number of channels. Following the convention of VAEs [155, 172], we model the distribution $q_2(\boldsymbol{z}|x_2)$ as a diagonal Gaussian distribution, with mean and covariance represented by the first and second half channels of this feature map. For multi-sequence MRI, we first use multiple encoders to extract per-sequence features, which are then concatenated and processed by a convolutional layer to obtain the MR feature (overall feature map for MRI). This feature map parameterizes the distribution $q_1(\boldsymbol{z}|x_1)$ in the same way as the auxiliary feature map. Once $q_1(\boldsymbol{z}|x_1)$ and $q_2(\boldsymbol{z}|x_2)$ are obtained through the feature maps, the final posterior and prior distributions are calculated through Eq. (3.2), and the KL divergence in the ELBO can be calculated analytically.

More importantly, the KL term in Eq. (3.3) measures the dissimilarity between the geometric and arithmetic means of q_m. As a part of maximizing the ELBO, we minimize the KL during training, encouraging the network to extract similar q_ms (features) from MRI and the auxiliary image. In other words, the MR feature q_1 will be aligned to the auxiliary feature q_2. As we discussed in Sec. 3.1, this restricts the complexity of q_1 and thus leads to better model performance. In contrast, a vanilla model may learn an MR feature intertwined with redundant information, making it difficult for the decoder to segment specific substructures. It has been demonstrated that individual distributions q_ms are helpful for learning desired features [135, 167].

To calculate the expectation term in the ELBO, we perform Monte Carlo estimation,

shown in Fig. 3.2, similar to previous works [88]. Specifically, for each training iteration, we sample a value $\boldsymbol{z}$ from $q(\boldsymbol{z}|\boldsymbol{x})$, based on which the decoder produces the segmentation probability map $p(s|\boldsymbol{z})$. The log-probability $\log p(s|\boldsymbol{z})$ can then be calculated by measuring the probability of the ground-truth segmentation s given the probability map, which is equivalent to the negative of cross-entropy [172]. This forms the estimate of $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})} \log p(s|\boldsymbol{z})$. For model evaluation, instead of sampled $\boldsymbol{z}$, the decoder directly utilizes the mode of $q(\boldsymbol{z}|x_1)$, as the input MR feature to produce segmentation probability maps.

The structures of the encoders and decoder are the same as in a U-Net. Therefore, the feature maps from the encoders are exactly the outputs of convolutional layers after downsampling (pooling) operations. Since these U-Net features are with different resolutions, we also calculate the dissimilarity loss (KL term in Eq. (3.3)) in a multi-scale manner. In other words, for each downsampling (scale) of the encoders, we obtain one $q(\boldsymbol{z}|\boldsymbol{x})$ to calculate one KL divergence, and the final KL is the sum of the multi-scale KL divergences. In addition, the encoders share all the parameters except for batch normalizations. This significantly reduces computational complexity and the number of learnable parameters, thus mitigating overfitting and helping better extract features [29].

Implementation details

We set the weights of the KL divergence and the segmentation (cross-entropy) loss to 1.5 and 500, respectively. The positive weight for the cross-entropy loss was set to 0.95. The model was implemented using PyTorch [118] and trained on an NVIDIA TITAN V GPU for 100 epochs, via the Adam optimizer [87] with a learning rate of 10^{-3} and a batch size of 400. We selected the model with the best validation performance (evaluated after each epoch) and reported the following results by applying this model to the test set.

3.2.4 Evaluation metrics

We used the Dice Similarity Coefficient to evaluate the segmentation performance. We also measured the commonly-used 95% Hausdorff Distance (HD95) and the Average Symmet-

ric Surface Distance (ASSD), where HD95 evaluates the maximum surface distance between the prediction and ground-truth, and ASSD considers the distance in an average sense. Besides, we report the area under the precision-recall curve (PR-AUC) to evaluate the overall performance by taking into account different thresholds.

3.2.5 Compared methods

We compared our model with two types of state-of-the-art deep learning methods: UNet-based and transformer-based. The first type exhibits similar U-Net-like encoder/decoder structures to ours, yet each adopts a distinctive approach to feature processing, including U-Net [130], Residual U-Net (ResU-Net) [85], Attention U-Net [115], Attention ResU-Net, U-Net++ [190], and nnU-Net [79], which are widely recognized for their performance in medical image segmentation nowadays. The second type involves more advanced transformer-based architectures, including recently proposed UNETR [65], Swin UNETR [64], and MedNeXt [131].

Since we aimed to segment calcification on MRI, all methods were trained and evaluated with the combination of the three MR sequences. For fair comparisons, the common settings (e.g., the number of scales) for all methods were also the same.

3.3 Results

3.3.1 Quantitative comparisons of different methods for segmentation

We first compared the proposed model and baseline methods for calcification segmentation. Particularly, our model involves two variants:

- **Ours (w/o mask):** For training, the auxiliary branch and dissimilarity loss are disabled.
- **Ours (w/ mask):** For training, both branches are utilized, with the ground-truth segmentation mask as the auxiliary.

The comparison between Ours (w/o mask) and Ours (w/ mask) served as an ablation study

Table 3.2: Segmentation performance of different methods on the test MR images. HD95 and ASSD were measured in millimeter (mm).

Methods	Dice	PR-AUC	HD95 (mm)	ASSD (mm)
U-Net	0.562	0.575	1.829	1.294
ResU-Net	0.539	0.545	2.431	1.080
Att U-Net	0.545	0.552	1.998	1.117
Att ResU-Net	0.539	0.564	5.087	1.347
U-Net++	0.596	0.623	1.648	0.853
nnU-Net	0.485	0.654	1.165	**0.672**
UNETR	0.606	0.652	8.114	1.535
Swin UNETR	0.587	0.633	4.201	1.260
MedNeXt	0.601	0.626	1.165	0.761
Ours (w/o mask)	0.592	0.630	6.649	1.133
Ours (w/ mask)	**0.620**	**0.660**	**0.848**	0.692

to investigate the effect of the auxiliary training input.

Comparisons with state-of-the-art segmentation models: The quantitative metrics evaluated on the test set are summarized in Tab. 3.2. Our method outperforms all baselines across Dice, PR-AUC and HD95, and our ASSD is very close to the best value achieved by nnU-Net, which, however, exhibits a very low Dice score. These results highlight the effectiveness of our approach. Besides, while UNETR achieves the best Dice among the baseline methods, its distance-based metrics, especially HD95, are significantly worse than ours. This indicates that our model produces better spatial agreement in boundary localization, effectively mitigating false positive outliers. Moreover, comparisons with the six U-Net variants highlight our model's superiority, attributed to the novel design of the theoretically grounded loss function, because our models shares similar encoder/decoder structures with those U-Net variants. Besides, our model outperforms the three networks with much more advanced transformer-based structures, reaffirming its superiority.

Comparison with inter-reader variations: To further validate that our method's results align closely with the consensus of human experts, we followed the same procedure in [127] to investigate inter-reader variations. Particularly, we invited two more radiologists

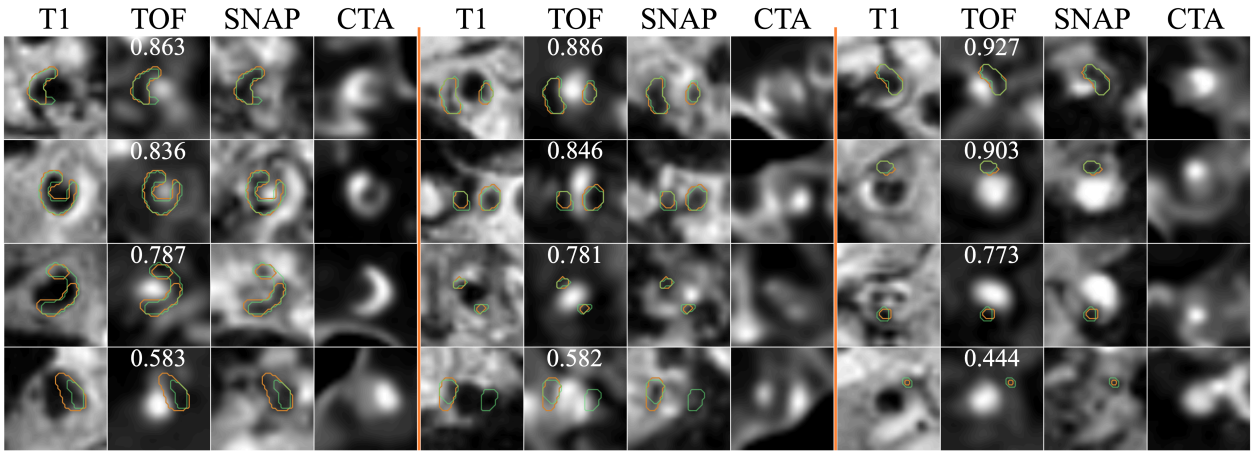


Figure 3.3: Examples of segmentation from Ours (w/ mask) compared to the ground-truth. 2D slices with various calcification shapes (ring-like, bulk and spotty) are displayed, with contours (green for ground-truth and orange for ours) overlaid on MR images, and the Dice scores for each slice group shown on the TOF images. Corresponding CTA images are not used for training or test, but are shown here only for reference.

(H.A. and E.Y.A.) to individually label calcification on 14 randomly selected subjects, resulting in an inter-reader Dice score of 0.626, which is very close to the value 0.620 achieved by our model.

Ablation study for the auxiliary encoder and dissimilarity loss: Ours (w/ mask) exhibits superior performance compared to Ours (w/o mask) trained without the auxiliary encoder. Notably, Ours (w/ mask) achieves a significantly lower HD95, demonstrating that the auxiliary encoder, coupled with the dissimilarity loss, effectively eliminated false positive outlier pixels in the predicted segmentation.

3.3.2 Qualitative results

The qualitative results from our model are visualized in Fig. 3.3, where we show 12 examples with various calcification shapes and a range of segmentation performances. Note that the CTA images are shown for reference purpose only, *i.e.*, they were not used as input for training or testing. It is evident that our method can produce relatively accurate calcification boundaries in a variety of situations. Although the dark area on SNAP is much

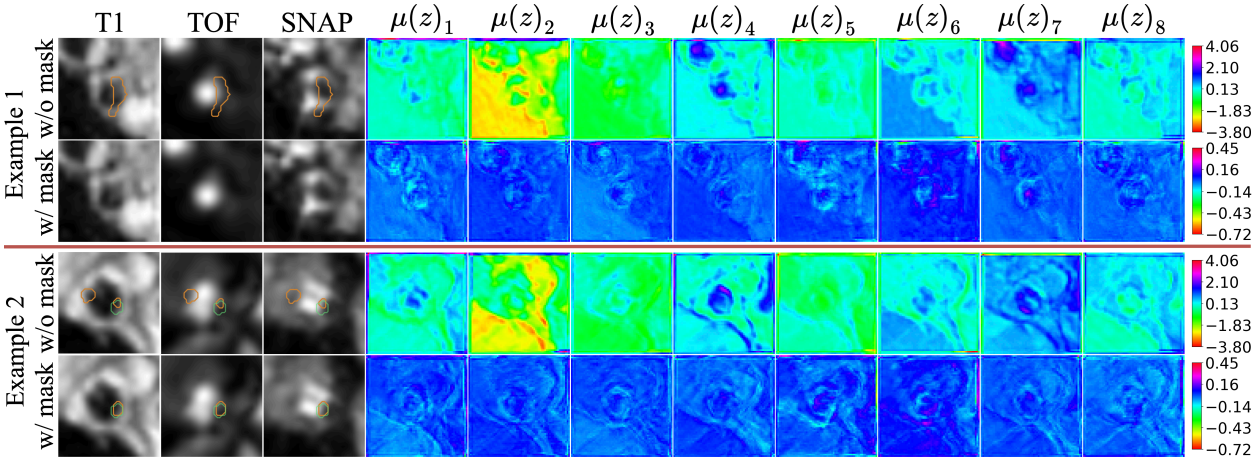


Figure 3.4: Visualization of MR features extracted by Ours (w/ mask) and Ours (w/o mask) from two example groups of multi-sequence MRI (with predicted and ground-truth segmentation delineated by orange and green contours, respectively). The features are eight channels of the mean value $\mu(z)$ of $q_1(\boldsymbol{z}|x_1)$. Ours (w/ mask) extracted features with reduced complexity, in contrast to ours (w/o mask) which extracted more complex features and predicted more false positives as seen by the segmentation contours, indicating poor generalizability of the features on the test set.

larger than the calcification region, the model still managed to combine multiple sequences to delineate the real boundaries of calcification. One can observe that even with relatively low Dice scores in some cases, the model can still localize correct calcification region, *e.g.*, the worst Dice 0.444 is just due to a very small calcification area. In the lower bottom case, the model failed to detect one of the two calcification regions, probably because a portion of the vessel wall on T1 is dark and unclear. Still, our model is powerful enough to identify calcification locations that could be hard to detect on MRI.

Feature complexity: We visualized MR features from the two variants of our model in Fig. 3.4. One can observe that Ours (w/ mask) produced visually simpler feature maps with a much smaller intensity range indicated by the color bars. Similar to [150], we also estimated an upper bound of mutual information $I(X_1; \boldsymbol{Z})$ between input MRI X_1 and its feature $\boldsymbol{Z}$, which provides a measurement for feature complexity. Ours (w/ mask) achieved a value of 6.9×10^{-8}, much smaller than the value 9.5×10^{7} from Ours (w/o mask). These

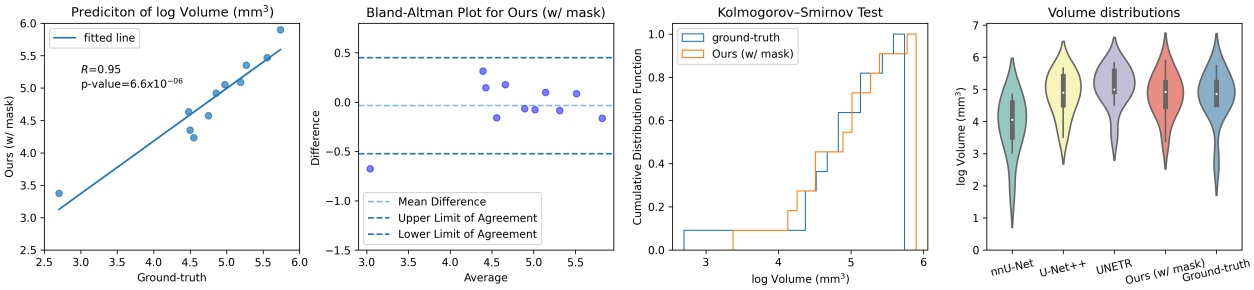


Figure 3.5: Comparisons between ground-truth and predicted calcium volumes for test subjects.

results indicates that the proposed model with the dissimilarity loss can effectively restrict feature complexity.

3.3.3 Performance in predicting clinical measurements

In clinical reviews of calcification, radiologists commonly rely on high-level measurements derived from segmentations for diagnosis and the analysis of disease progression. Frequently utilized calcification metrics include the Agatston Score and calcium volume [145, 181]. Since the former is specific to CT scans, here we investigated our model's performance in predicting calcium volume. The results are shown in Fig. 3.5.

Compared to baselines, our approach demonstrates a calcium volume distribution closer to the ground truth, and the Kolmogorov-Smirnov test for calcium volumes yields a p-value of $0.997 \gg 0.05$, indicating no significant evidence of a difference between our volume distribution and the ground-truth. The Bland-Altman plot reveals nearly all predictions fall within the range of agreement with ground-truth, and there is only one slight outlier with a small volume (typically more challenging to predict). The regression line between our predictions and the ground-truth further exhibits a substantial R value with p-value $\ll 0.001$. Collectively, these findings demonstrate robust evidence of the efficacy of our method in accurately predicting calcium volume.

Another potentially valuable evaluation involves assessing the frequency of calcium oc-

Table 3.3: Slice-wise detection performance of different methods on the test MR images, with top-2 values bolded.

Methods	F1	Recall	Precision	PR-AUC
U-Net	0.761	0.696	0.839	0.800
ResU-Net	0.776	0.740	0.814	0.805
Att U-Net	0.751	0.680	0.840	0.794
Att ResU-Net	0.764	0.720	0.813	0.797
U-Net++	**0.821**	0.772	0.876	0.848
nnU-Net	0.788	0.658	**0.982**	**0.857**
UNETR	0.804	**0.899**	0.727	0.824
Swin UNETR	0.810	**0.907**	0.731	0.829
MedNeXt	0.794	0.720	0.886	0.833
Ours (w/o mask)	0.756	0.745	0.766	0.783
Ours (w/ mask)	**0.823**	0.764	**0.892**	**0.853**

currence along the arteries on a slice-by-slice basis. Therefore, we further evaluated the performance of the compared methods in slice-wise calcification detection. In this task, each location on the vessel centerline is classified as calcified if the calcification mask of the corresponding 2D slice group contains positive pixels. The results are shown in Tab. 3.3, demonstrating the superiority of our method in detecting calcified locations along arteries, with high F1 score around 0.82 and PR-AUC around 0.85. Considering calcification is dark and nearly invisible on MRI, our approach achieves accurate automated calcification localization. This could greatly alleviate the burden of manual analysis, which requires an experienced radiologist to carefully examine suspicious dark signals on multiple MR sequences simultaneously.

3.3.4 Effect of different MR sequences

We also investigated the effect of each MR sequence on model performance. To this end, we trained and evaluated Ours (w/ mask) with different combinations of MR sequences, as shown in Tab. 3.4.

We observed that the results improved with the inclusion of more MR sequences, and

Table 3.4: Performance of Ours (w/ mask) with different combinations of MR sequences. A check mark indicates that the sequence is used for both training and test. A value in bold means it is the best among all combinations with the same number of sequences.

Combinations			Segmentation				Slice-wise detection			
T1	TOF	SNAP	Dice	PR-AUC	HD95 (mm)	ASSD (mm)	F1	Recall	Precision	PR-AUC
✓			0.215	0.140	22.950	5.588	0.533	0.632	0.460	0.585
	✓		0.246	0.133	9.712	3.212	0.579	0.605	0.554	0.622
		✓	**0.444**	**0.418**	**8.229**	**1.869**	**0.690**	**0.717**	**0.666**	**0.722**
✓	✓		0.406	0.378	15.990	2.938	0.638	0.723	0.571	0.677
✓		✓	**0.554**	**0.561**	2.959	1.184	**0.762**	**0.735**	**0.791**	**0.791**
	✓	✓	0.502	0.488	**2.171**	**1.178**	0.737	0.691	0.790	0.774
✓	✓	✓	**0.620**	**0.660**	**0.848**	**0.692**	**0.823**	**0.764**	**0.892**	**0.853**

SNAP is much more informative about calcification compared to TOF and T1, as the combination containing SNAP achieves a better performance than without SNAP, when the number of sequences is fixed. Moreover, T1 is generally better than TOF when combined with SNAP to detect calcification, but either TOF+SNAP or T1+SNAP is not sufficient to achieve a good performance. The best performance was achieved when all three sequences were included.

The results are consistent with our intuition that: 1) SNAP can provide a relatively accurate segmentation boundary, since the bright lumen area and the dark calcification area usually form an easy-to-detect contrast ratio, even though there are other non-calcified dark regions in SNAP, which may increase the false positive rate. 2) TOF can help localize the lumen and thus ameliorate the interference of dark artifacts inside the lumen in SNAP, while it cannot help exclude surrounding dark structures outside the vessel region. 3) T1 can also help localize calcification by providing an outer bound of the region-of-interest (including lumen, calcification and outer wall). An abnormal boundary could indicate the potential presence of calcification, while calcification that does not distort vessel wall contours could be missed. Thus, the combination of the three sequences produces the best result, and each of them plays an important role.

3.4 Discussion

We have developed a novel model for intracranial arterial calcification segmentation and detection on multi-sequence MRI by restricting feature complexity. The literature on automated deep learning models for MRI-based calcification assessment is very limited and unexplored. Therefore, we conducted this study and validated the feasibility of the task.

While CT/CTA is the first-line imaging in stroke, and is often considered the reference standard for calcification detection, with the increased use of VWI [113] and a healthcare push to limit imaging overutilization, patients with cerebrovascular diseases may only undergo MRI and VWI in disease assessment. By providing comprehensive feature characterization without the need for added imaging, VWI improves healthcare efficiency and reduces the burden on patients from additional radiation and iodinated contrast injection. In addition, dementia evaluation may not include CT/CTA. Considering the associations between atherosclerosis, calcification and cognitive impairment, with further evaluation and validation, comprehensive evaluation of plaque features and calcification could be beneficial in outcome assessment in the future.

The proposed method was compared with multiple cutting-edge medical image segmentation methods on a dataset of intracranial VWI. With the combination of three MR sequences (*i.e.*, T1, TOF, and SNAP) and an auxiliary image for training, our method achieved superior segmentation and slice-wise detection on the test MRI data, generally outperforming all baseline methods. This demonstrates the efficacy of our model. Moreover, the proposed method with auxiliary was better than the same model trained without auxiliary. This demonstrates that the strategy of restricting MR feature complexity by an auxiliary feature is beneficial for calcium segmentation.

We also investigated the model performance with different combinations of MR sequences. We found that T1, TOF and SNAP all contain certain structural information of calcification. In addition, with the presence of more MR sequences, the model generally achieves a better performance. In particular, SNAP seemed to be more informative than T1 or TOF, as it

presents calcified areas with darker signal intensities, which contrasts more sharply with the lumen and other surrounding tissue. Nonetheless, these three sequences support calcification assessment with complementary and unique information from their own perspectives. Note that "SNAP" in this work refers to the reference image of the SNAP VWI sequence rather than the corrected real image which is generally referred to as SNAP in the literature. The reference image of the SNAP sequence has been shown to benefit the identification of calcium [37].

Although this work utilized three specific MR sequences, our model is flexible to include more. New sequences need to be registered and then 2D cross-sectional slices can be generated. More MR encoders can be added to the network structure for the additional sequences. Then training and test can be performed as usual. The number of model parameters will not increase excessively, since we share the convolutional layers across all encoders. In addition, we only targeted the intracranial ICA and MCA. Further study is needed to determine whether our model has good generalizability and robustness for calcification segmentation for other intracranial artery segments. We plan to include the full intracranial arterial tree for comprehensive assessment in the future.

In summary, calcification is generally considered extremely challenging to segment on MRI. This work established the first automated deep learning approach to this task. This can enable comprehensive MRI review including calcification, vessel wall, plaque and hemodynamic evaluation that may improve imaging efficiency and reduce patient burden and risk.

3.5 Chapter Takeaway

This chapter addressed intracranial arterial calcification segmentation from multi-contrast brain MRI in a standard supervised setting. Although labels are available and the contrasts are co-registered, robust segmentation remains difficult because calcification is typically dark and often only weakly expressed in MRI; reliable delineation must therefore be inferred from

indirect contextual relationships rather than from direct intensity signatures. Under this condition, compounded *intensity-level appearance heterogeneity*, arising from scanner, protocol, and reconstruction variability, can readily perturb the fragile cues on which the task depends.

Methodologically, the main lesson is that stability under such appearance variability benefits from explicit control of representational capacity. By introducing a variational Bayesian formulation and enforcing constraints that restrict feature complexity, the model is encouraged to allocate explanatory responsibility to task-relevant latent content rather than to acquisition-dependent appearance fluctuations that may correlate with the labels in limited training data. This perspective aligns with the thesis-wide requirement of identifiable invariant preservation: even in the most moderate heterogeneity regime, representations can become unreliable when observational variability is implicitly absorbed, whereas explicit latent role specification can improve both robustness and predictive performance.

The next chapters extend this principle to progressively more demanding regimes. Chapter 4 moves beyond appearance variability to settings in which multiple images must be jointly explained despite substantial geometric misalignment and heterogeneous observation mechanisms. Chapter 5 further removes the availability of reliable spatial correspondence across images, requiring global canonicalization mechanisms that preserve task-relevant invariants when direct alignment is ill-posed.

Chapter 4

MULTI-MODAL GROUPWISE IMAGE REGISTRATION

This chapter presents the second project of the thesis: unsupervised multimodal groupwise image registration. Compared with the supervised segmentation setting of Chapter 3, the difficulty here arises from a fundamentally different source. Instead of predicting labels from images with fixed geometric alignment, the goal of registration is to establish anatomical correspondence across a collection of images. In this project, we consider groups of images that depict the same anatomical structure but differ simultaneously in both observation conditions and spatial configuration.

Specifically, the setting involves two interacting forms of heterogeneity. First, images may be acquired using different contrasts or modalities (e.g., different MRI contrasts or MRI versus CT), resulting in substantial differences in image appearance even when the underlying anatomy is identical. Second, each image may exhibit its own geometric configuration due to subject variability, motion, or positioning. In the terminology introduced in Chapter 1, this regime corresponds to compounded contrast/modality-level appearance heterogeneity together with *registration-compatible geometric heterogeneity*. A meaningful notion of correspondence exists in principle, since the images depict the same anatomical structure, but must be inferred without supervision and in the presence of substantial appearance variability.

This dual source of variability creates a central difficulty for registration. Establishing correspondence typically relies on comparing local image patterns across images. However, when observation mechanisms differ across modalities or contrasts, raw intensity similarity becomes unreliable as a proxy for anatomical similarity. Conventional approaches often address this problem through handcrafted similarity measures or by modeling joint intensity

relationships across the image group. While effective in limited settings, these strategies become increasingly fragile as modality differences grow or as group size increases, since the resulting similarity models must account for increasingly complex observation statistics.

From a representation learning perspective, the key challenge is therefore to construct an intrinsic notion of similarity that reflects shared anatomical structure rather than modality-dependent appearance. In other words, the representation should capture anatomical information that is consistent across modalities while separating it from factors related to observation conditions. At the same time, this representation must remain coupled with geometric transformations so that correspondence across the image group can be inferred.

To address this challenge, we adopt the Bayesian representation learning framework introduced in Chapter 2. We formulate multimodal groupwise registration under a hierarchical generative model in which two distinct latent sources jointly explain the observed images: a latent variable representing the *common anatomy* shared by the image group, and latent variables representing the *geometry* of each individual image. The observed multimodal images are then viewed as arising from the interaction of these latent factors with modality-specific observation processes.

Within this formulation, registration emerges naturally as Bayesian inference over the latent anatomy and geometry variables. We implement this idea through a hierarchical variational inference architecture that explicitly disentangles anatomical structure from geometric transformations. By organizing the latent representation in this way, multimodal similarity can be evaluated in a modality-invariant latent space while geometric alignment is inferred simultaneously across the image group. This representation-centered formulation enables stable groupwise registration even under substantial modality and appearance heterogeneity.

The remainder of this chapter closely follows the associated publications:

- (Oral, Best Paper Runner-Up) Xin Wang* and Xinzhe Luo* et al. Bingo: Bayesian intrinsic groupwise registration via explicit hierarchical disentanglement. In Alejandro Frangi, Marleen de Bruijne, Demian Wassermann, and Nassir Navab, editors, Infor-

mation Processing in Medical Imaging, pages 319–331, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34048-2. [167].

- Xinzhe Luo* and Xin Wang* et al. Bayesian Unsupervised Disentanglement of Anatomy and Geometry for Deep Group-wise Image Registration. IEEE Transactions on Pattern Analysis & Machine Intelligence, 48(01):711–729, January 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3609521. [101].

4.1 Introduction

Groupwise image registration aims to find the hidden spatial correspondence that aligns multiple observations. For medical images, when the observations reflect some common anatomy, their intrinsic structural correspondence, which can be independent of the multi-modal imaging acquisition protocol, is of particular interest. However, conventional methods on multi-modal groupwise registration usually rely on intensity-based similarity measures to iteratively optimize the spatial transformations. For instance, multivariate joint entropy and mutual information were proposed as groupwise similarity measures in [22, 143, 141, 162], followed by more computationally favourable template-based approaches [123, 100]. Nevertheless, devising proper similarity measures and choosing the correct registration hyperparameters for heterogeneous medical images can be tedious and challenging. The high computational burden and the instability in registration accuracy may also prevent real-world applications of conventional groupwise registration algorithms.

In this work, we seek to establish a new unified and interpretable learning framework for large-scale unsupervised multi-modal groupwise registration. Inspired by recent progress in disentangled representation learning [17, 69], we propose to disentangle the underlying common anatomy and geometric variations from the observed images, which can be regarded as reverting the data generating process of the observations. To this end, a probabilistic generative model is constructed, where the common anatomy and spatial transformations are disentangled as latent variables. Thus, the problem of groupwise registration (recovery

of the intrinsic structural correspondence) devolves to the estimation of the latent posterior distribution, which is then solved via variational inference. Besides, since unsupervised learning of disentangled representation is impossible without proper inductive biases [96, 86], we have designed a novel hierarchical variational auto-encoding architecture to realize the inference procedure of latent variables, where the decoder network complies with the equivariance assumption of the imaging process. Accordingly, groupwise registration is learnt in an unsupervised closed-loop self-reconstruction process: i) the encoder extracts the intrinsic structural representations from the multi-modal observations, based on which the common anatomy and the spatial correspondence are estimated; ii) the decoder emulates the equivariant image generative process from the common anatomy, and reconstructs the observations via the inverse spatial transformations.

To improve the efficiency and scalability of groupwise registration, particularly in achieving deep groupwise registration with variable (and potentially very large) group sizes for the first time, we make the following contributions:

- We propose a new learning paradigm for multi-modal groupwise registration based on Bayesian inference and disentangled representation learning. Remarkably, we can achieve registration in an unsupervised closed-loop self-reconstruction process, which spares the burden of designing complex image-based similarity measures.
- We propose a novel hierarchical variational auto-encoding architecture for the joint inference of latent variables. The network is able to reveal the underlying structural representations from the observations with visual semantics, based on which the intrinsic structural correspondence can be explicitly estimated in a mathematically interpretable fashion.
- Based on the equivariance assumption and the decomposition of the symmetry group actions on the observational space, we prove that under certain conditions, our model can identify the desired registration parameters, which constitutes the theoretical underpinnings of the established framework.
- Our registration model, while trained using small image groups, can be readily adapted

Table 4.1: Definition of the main mathematical symbols used in this paper.

Symbol	Description	Symbol	Description
L	the number of levels of latent variables	N	the number of images in the group
$\boldsymbol{Z}$	random variable of the common anatomy	$\boldsymbol{U}$	random vector of the image group $\{U_j\}_{j=1}^N$
Ω	the coordinate space of the common anatomy	Ω_j	the j-th image coordinate space
$\boldsymbol{\phi}$	the set of diffeomorphic maps $\boldsymbol{\phi} \triangleq \{\phi_j\}_{j=1}^N$	ϕ_j	the spatial transformation from Ω to Ω_j
$\boldsymbol{z}$	the latent common structural representation	$\boldsymbol{u}$	an observed sample of the image group
$\boldsymbol{z}^l$	the latent common structural representation at the l-th level	$\boldsymbol{v}^l$	the latent stationary velocity fields registering the image group at the l-th level
$q^*(\boldsymbol{z}^l \mid \boldsymbol{u}, \boldsymbol{v})$	the geometric mean variational distribution of $\boldsymbol{z}^l$ given $\boldsymbol{u} \circ \boldsymbol{\phi} \triangleq \{u_j \circ \phi_j\}_{j=1}^N$	$\boldsymbol{v}_j^l$	the stationary velocity field of the j-th diffeomorphism at the l-th level, $\boldsymbol{v}^l = \{\boldsymbol{v}_j^l\}_{j=1}^N$
$q_j^\diamond(\boldsymbol{z}^l \mid u_j, \boldsymbol{v}_j)$	the j-th single-view variational distribution of $\boldsymbol{z}^l$ given $u_j \circ \phi_j$	$\boldsymbol{\mu}_{\boldsymbol{v},j}^l, \boldsymbol{\Sigma}_{\boldsymbol{v},j}^l$	parameters of the Gaussian variational distribution $q(\boldsymbol{v}_j^l \mid \boldsymbol{u}, \boldsymbol{v}^{<l})$
$\widetilde{q}_j(\boldsymbol{z}^l \mid u_j)$	the j-th single-view variational distribution of $\boldsymbol{z}^l$ given u_j	$\boldsymbol{v}_j^+$	the total velocity field aggregating all $\{\boldsymbol{v}_j^l\}_{l=1}^L$ upsampled to the top level
$\boldsymbol{\psi}$	parameters of the variational model $\{\boldsymbol{\psi}_j\}_{j=1}^N$	$\boldsymbol{\theta}$	parameters of the generative model

to large-scale and variable-size test groups, significantly enhancing its computational efficiency and applicability.

- We validated the proposed framework on four publicly available medical image datasets, demonstrating its superiority over similarity-based methods in terms of accuracy, efficiency, scalability, and interpretability.

The remainder of the article is organised as follows. Sec. 4.2 discusses the related work to this study. Sec. 4.3 elucidates the proposed Bayesian inference framework for groupwise registration. Sec. 4.4 describes the proposed hierarchical variational auto-encoder architecture for learning and estimation of the latent variables. Sec. 4.5 presents the experimental setups and evaluation results of our method on four different public datasets. Sec. 4.6 notes implications of the proposed framework and concludes the study. For conciseness, in Tab. 4.1 we specify the main mathematical symbols used in the rest of this paper.

4.2 Related Work

In this section, we review related work to the proposed framework, including similarity-based and deep feature-based approaches to (groupwise) image registration, as well as literature on multi-modal representation learning that facilitates disentanglement of latent variables from multiple modalities. We hope to expose the connections between them that motivate our established framework.

4.2.1 Groupwise Image Registration

Let $\boldsymbol{U} = \{U_j\}_{j=1}^N$ be the random vector representing the image group and $\boldsymbol{u} = \{u_j\}_{j=1}^N$ an observed sample of $\boldsymbol{U}$, with $U_j : \mathbb{R}^d \supset \Omega_j \to \mathbb{R}$ the intensity mapping, Ω_j the image domain, and d the dimensionality. In the classical pattern matching and computational anatomy regime [57, 58, 153, 49, 111], the observed homogeneous image group is considered as samples from the orbit of a transformation group $\mathcal{G}$ acting on a deformable template U_0 (*a.k.a.* atlas), namely

$$\boldsymbol{U} \subset \mathcal{G} \cdot U_0 \triangleq \{U_0 \circ \phi_j^{-1} : \phi_j \in \mathcal{G}\}, \tag{4.1}$$

where the transformation $\phi_j^{-1} : \Omega_j \to \Omega$ maps spatial locations in the image domain to a common coordinate space $\Omega \subset \mathbb{R}^d$, and $\mathcal{G}$ is often taken as the group of diffeomorphisms. This concept has motivated the development of population averaging methods that estimate the deformations $\mathcal{G}$ and the template U_0 simultaneously, based on the observations $\boldsymbol{U}$ [60, 11, 82, 38, 19, 3, 102, 53, 186, 42, 47]. The resultant transformations encode the structural variability and the template provides a statistical representative of the images. In particular, *a priori* guess of the template as one of the observed images was proposed in [60, 11, 38], with the template updated by the average deformation in each iteration. Joshi et al. [82] and Bhatia et al. [19] proposed using the intensity mean image as a template, avoiding reference selection by estimating transformations from the common space. Allassonnière et al. [3] extended the setup to a Bayesian framework, with the template modeled as a linear combination of continuous kernel functions. Recently, Dalca et al. [42], Ding and

Niethammer [47] proposed learning-based template estimation methods, using spatial transformation networks to predict diffeomorphisms parameterized by stationary velocity fields [7, 9]. Nevertheless, groupwise registration based on the deformable template assumption could not readily accommodate the multi-modal nature of medical images.

Multi-modality can introduce additional complexity, as the observed images are not related to an intensity template directly through spatial transformations. In this case, based on the common anatomy assumption of multi-modal medical images, we can instead assume an *anatomical* or structural representation $\boldsymbol{Z}$ from which each observed image is generated through the composition of an imaging functional f_j and a spatial transformation ϕ_j^{-1}, *i.e.*,

$$U_j = f_j(\boldsymbol{Z}) \circ \phi_j^{-1}. \tag{4.2}$$

More importantly, the imaging functional is assumed to be *spatially equivariant* w.r.t. ϕ_j^{-1}, a condition motivated by the demand to learn equivariant image features [72, 89, 133, 125], which writes

$$f_j(\boldsymbol{Z}) \circ \phi_j^{-1} = f_j(\boldsymbol{Z} \circ \phi_j^{-1}), \quad \forall\, \phi_j \in \mathcal{G}. \tag{4.3}$$

Inspired by inter-modality pairwise registration based on the joint intensity distribution (JID) or particularly the mutual information (MI) [160, 105, 93, 144, 129], initial attempts to realize *multi-modal groupwise registration* focused on generalizing this information-theoretic approach directly to high-dimensional cases [22, 143, 185, 141, 162]. Later, concerning the curse of dimensionality in estimating high-dimensional JIDs, template-based groupwise registration revives through probabilistic modeling of the image generative process in Eq. (4.2) [97, 116, 21, 123, 191, 100]. These methods assumed that the JID is modeled by the marginalisation of the joint distribution between the latent template and the observed images. Thus, the spatial transformations were determined by maximum likelihood estimators (MLEs) while the template was predicted via maximum a posteriori (MAP). We direct interested readers to our previous work [100] for detailed elucidation of this maximum-likelihood perspective and its connection to information-theoretic metrics.

Over the past years, the shift from optimization to learning-based image registration [75, 45, 14, 43] has also motivated the development of learning-based multi-modal groupwise registration methods which predict the desired spatial transformations using neural network estimation [31, 99]. However, this deep-learning approach may inherent the same limitation of scalability from its optimization-based counterpart. That is, the trained network for a fixed input size can hardly be utilized to register image groups of variable sizes. This hinders the potential of learning-based frameworks on large-scale groupwise registration.

4.2.2 Deep Feature-Based Image Registration

A closely related line of research to our work is deep feature-based image registration [122, 95, 114, 138, 41, 46]. These methods usually adopt a two-stage pipeline: in the first stage a feature extraction network is pre-trained on a surrogate task, be it segmentation [138, 41], auto-encoding [41], or contrastive learning [122, 95], followed by an instance optimization procedure using conventional similarity measures with the learnt features in the second stage. They may have the advantage of being interpretable and capable of multi-modal registration based on modality-invariant features [122, 95, 138], compared to previous end-to-end counterparts [45, 14] that make predictions using a black-box network. However, the two-stage prediction pipeline and the requirement for ground-truth labels (segmentation annotation or aligned image pairs) limit their applicability. In contrast, Qin et al. [126] proposed using the multimodal unsupervised image-to-image translation framework [77] to reduce the multi-modal registration problem to a mono-modal one by disentangling shape and appearance representations. More recently, Deng et al. [46] proposed a unified framework where explicit modality-invariant feature extraction and unsupervised registration are learnt via an interpretable optimization problem.

4.2.3 Multi-Modal Representation Learning

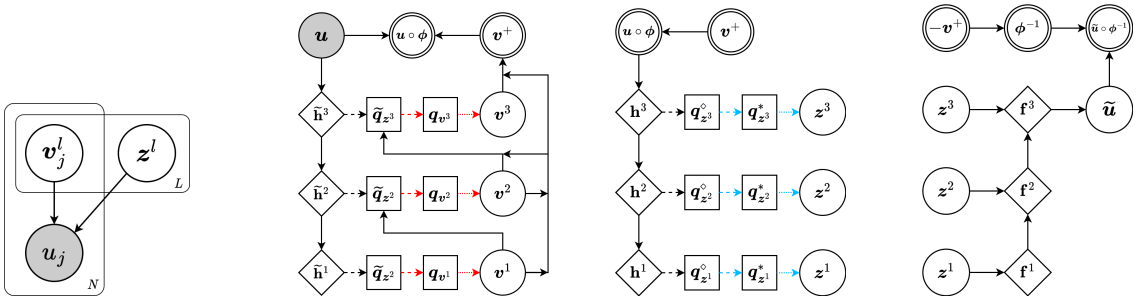
The underlying structural representations that facilitate the estimation of spatial correspondence can be regarded as some latent embedding, which encodes the information

correlating the multi-modal observations. This viewpoint, in the context of heterogeneous data, brings the need for multi-modal representation learning which exploits the commonality and complementarity of multiple modalities, and the demand for data alignment which identifies the relation and correspondence between different modalities [15].

Modern probabilistic generative models [88, 55, 156] construct a general framework for representation learning, since the data formation procedure to which they adapt is optimally suited for an unsupervised manner, and are powerful in striking a balance between fitting and generalizability taking into account the uncertainty in characterizing the data and latent distributions.

One family of methods within this scope is variational autoencoders (VAEs) [88], where the paradigm is to approximate the intractable latent distribution by a learnable variational posterior during the optimization of an evidence lower bound (ELBO) of the log likelihood of observed data. It provides an information-theoretic way to establish the foundational underpinnings of data encoding for downstream tasks [2]. The multi-modal variants of VAEs additionally imbue the capability of retrieving modality-invariant representation by factorizing it into single-view posteriors (*a.k.a.* experts) inferred from each modality. A typical factorization specifies the combination of the experts explicitly (*e.g.* mixtures [135], products [174], mixtures of products [146]) or implicitly (*e.g.* cross-modal generation [154]), compelling the summarization of the common information from any individual modality.

These approaches, nonetheless, all necessitate originally aligned inputs, such as the image and sentence that describe the same object [135], or multi-modal images that are well registered [66, 48], which limits their applicability to distorted data. Besides, they often adopt a reductionist stance by assuming an overly simplistic prior. (*e.g.*, the standard Gaussians or Laplacians), failing to respect the structure of latent manifold induced by real-world high-dimensional images, which is crucial for mitigating off-track regularisation and vanishing latent dimensions [73, 152], as well as for ensuring suitable decomposition of latent encodings [109]. The mixture-based factorization of the posterior also suffers from inevitable gap between the ELBO and the true likelihood, leading to undesirable suboptimal results [44].



(a) Generative model. (b) Inference steps #1. (c) Inference steps #2. (d) Generation steps.

Figure 4.1: The proposed hierarchical framework for Bayesian groupwise registration (3-layer example). Random variables are in circles, deterministic variables are in double circles, and observed variables are shaded. Diamonds denote network feature maps, and squares represent variational distributions. (a) Probabilistic graphical model of the generative process. (b) Inference steps #1 that predict the hierarchical velocity fields, where we denote $\widetilde{\boldsymbol{q}}_{\boldsymbol{z}^l} \triangleq \{\widetilde{q}_j(\boldsymbol{z}^l|u_j;\boldsymbol{\psi}_j)\}_{j=1}^N$ and $\boldsymbol{q}_{\boldsymbol{v}^l} \triangleq \{q(\boldsymbol{v}_j^l|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi})\}_{j=1}^N$. (c) Inference steps #2 that predict the common structural representations base on the warped images, where we denote $\boldsymbol{q}^{\diamond}_{\boldsymbol{z}^l} \triangleq \{q_j^{\diamond}(\boldsymbol{z}^l|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j)\}_{j=1}^N$ and $\boldsymbol{q}^{*}_{\boldsymbol{z}^l} \triangleq q^*(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi})$. (d) Generation steps that reconstruct the original images. Note that the inference and generation steps form a closed-loop self-reconstruction process.

In contrast, our framework embraces advancements in tackling all the limitations above: It allows learning from distorted images in a unified manner, through explicitly endowing the representations of spatial transformations and the common anatomy with a disentangled nature.

4.3 Bayesian Groupwise Registration

The fundamental principle driving the proposed approach is to disentangle the underlying common anatomy and geometric variations as latent representations *w.r.t.* the decomposition of two symmetry groups acting independently in the observational image space. To this end, we propose to formulate the estimation of these latent variables as a problem of Bayesian inference. Specifically, let $\boldsymbol{u} = (u_j)_{j=1}^N$ be a sample of the random vector $\boldsymbol{U}$ of the image group. From Eq. (4.2), each image u_j is generated from the latent variable $\boldsymbol{z}$ representing the

corresponding common anatomy, and the diffeomorphic transformation ϕ_j^{-1}. To ensure its invertibility, we assume that the transformations $\boldsymbol{\phi} \triangleq (\phi_j)_{j=1}^N$ is parameterized by stationary velocity fields $\boldsymbol{v} = (\boldsymbol{v}_j)_{j=1}^N$ [7, 9] such that $\phi_j = \exp(\boldsymbol{v}_j)$, and

$$\frac{\partial}{\partial t}\phi_j(\boldsymbol{\omega}, t) = \boldsymbol{v}_j(\phi_j(\boldsymbol{\omega}, t)), \quad \forall\, \boldsymbol{\omega} \in \Omega,\ t \in [0, 1]. \tag{4.4}$$

The diffeomorphism of the transformations is further imposed by assigning proper priors to the velocity fields.

4.3.1 Hierarchical Bayesian Inference

In practice, registration is often performed at multiple levels to facilitate convergence of the algorithm [134, 140]. Therefore, we express the latent variables with L hierarchical levels, *i.e.*, $\boldsymbol{z} = (\boldsymbol{z}^l)_{l=1}^L$ and $\boldsymbol{v}_j = (\boldsymbol{v}_j^l)_{l=1}^L$, in which higher levels indicate finer resolutions. Thus, the graphical model of the image generative process can be described as Fig. 4.1a. Note that while different levels of the latent variables are assumed to be independent, the inference procedure is performed *hierarchically*.

Since we are going to parameterize the likelihood $p(\boldsymbol{u}|\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta})$ by neural networks in the context of Bayesian deep learning, exact inference and parameter estimation become intractable [88]. Therefore, we resort to variational inference (VI) to approximate the maximum likelihood. The objective function of VI is the evidence lower bound (ELBO) of the log-likelihood function, *i.e.*,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{u}) &\triangleq \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi})}[\log p(\boldsymbol{u}|\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta})] - D_{\mathrm{KL}}\left[q(\boldsymbol{z}, \boldsymbol{v}|\boldsymbol{u}; \boldsymbol{\psi}) \parallel p(\boldsymbol{z})p(\boldsymbol{v})\right] \\ &\leq \log p(\boldsymbol{u}; \boldsymbol{\theta}) \triangleq \ell(\boldsymbol{\theta}|\boldsymbol{u}),\end{aligned} \tag{4.5}$$

where $q(\boldsymbol{z}, \boldsymbol{v}|\boldsymbol{u}; \boldsymbol{\psi})$ is the variational distribution, an approximation to the intractable true posterior $p(\boldsymbol{z}, \boldsymbol{v}|\boldsymbol{u}; \boldsymbol{\theta})$; $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are the variational and generative parameters respectively. The expectation over the likelihood can be estimated by Monte-Carlo sampling [88].

To simplify the KL divergence term, we first write

$$\begin{aligned} &D_{\mathrm{KL}}\big[q(\boldsymbol{z},\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi}) \parallel p(\boldsymbol{z})p(\boldsymbol{v})\big] \\ =&\mathbb{E}_{q(\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi})}\Big[D_{\mathrm{KL}}\big[q(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel p(\boldsymbol{z})\big]\Big] + D_{\mathrm{KL}}\big[q(\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi}) \parallel p(\boldsymbol{v})\big], \end{aligned} \tag{4.6}$$

where $q(\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi}) = \prod_{l=1}^{L} q(\boldsymbol{v}^l|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi})$, with $\boldsymbol{v}^l \triangleq (\boldsymbol{v}_j^l)_{j=1}^N$, and $\boldsymbol{v}^{<l}$ denotes the velocity fields in levels lower than l. This factorization of velocity fields can be illustrated by the inference steps #1 in Fig. 4.1b. Besides, we assume that the multi-level variational posteriors of the velocity fields factorizes as

$$q(\boldsymbol{v}^l|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi}) = \prod_{j=1}^{N} q(\boldsymbol{v}_j^l|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi}). \tag{4.7}$$

Then, the KL *w.r.t.* the velocity fields can be decomposed as

$$D_{\mathrm{KL}}\big[q(\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi}) \parallel p(\boldsymbol{v})\big] = \sum_{j=1}^{N}\sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{v}_j^{<l}|\boldsymbol{u};\boldsymbol{\psi})}\left[D_{\mathrm{KL}}\big[q(\boldsymbol{v}_j^l|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi}) \parallel p(\boldsymbol{v}_j^l)\big]\right]. \tag{4.8}$$

Likewise, the KL *w.r.t.* the common anatomy can be simplified into

$$D_{\mathrm{KL}}\big[q(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel p(\boldsymbol{z})\big] = \sum_{l=1}^{L} \mathbb{E}_{q(\boldsymbol{z}^{<l}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi})}\left[D_{\mathrm{KL}}\big[q(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel p(\boldsymbol{z}^l)\big]\right], \tag{4.9}$$

where we have assumed that

$$q(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v},\boldsymbol{z}^{<l};\boldsymbol{\psi}) = q(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}),$$

i.e., the common anatomy at level l can be inferred directly from $\boldsymbol{u}$ and $\boldsymbol{v}$ without referring to lower-level representations $\boldsymbol{z}^{<l}$, which is illustrated by Fig. 4.1c.

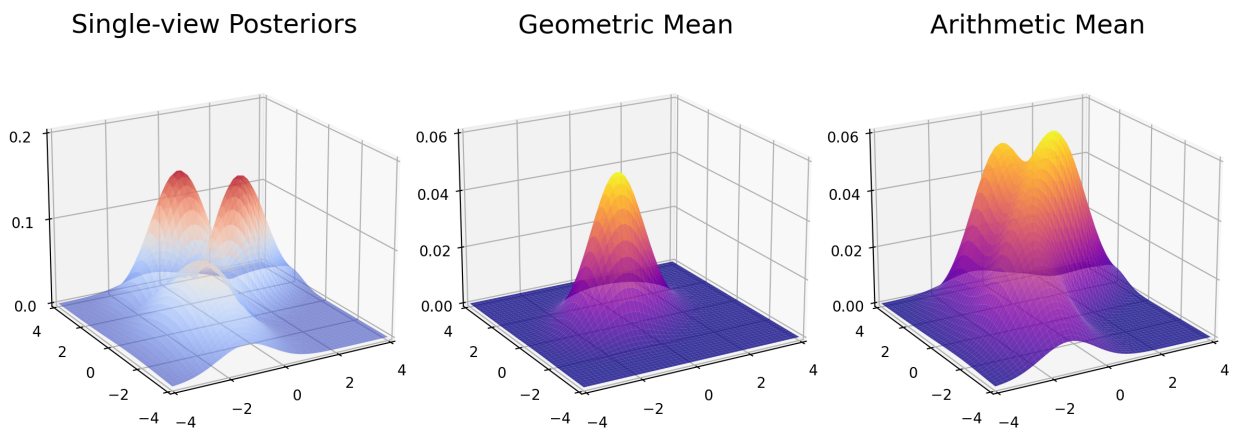


Figure 4.2: An example of the geometric and arithmetic mean for the single-view Gaussian posterior distributions.

The decomposition of KL divergence *w.r.t.* the latent variables is summarized as

$$\begin{aligned} & D_{\mathrm{KL}}\big[q(\boldsymbol{z},\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi}) \parallel p(\boldsymbol{z})p(\boldsymbol{v})\big] \\ &= \mathbb{E}_{q(\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi})}\left\{\sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{z}^{<l}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi})}\left[D_{\mathrm{KL}}\big[q(\boldsymbol{z}^{l}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel p(\boldsymbol{z}^{l})\big]\right]\right\} \qquad \text{(i)} \\ &\quad + \sum_{j=1}^{N}\sum_{l=1}^{L}\mathbb{E}_{q(\boldsymbol{v}_j^{<l}|\boldsymbol{u};\boldsymbol{\psi})}\left[D_{\mathrm{KL}}\big[q(\boldsymbol{v}_j^{l}|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi}) \parallel p(\boldsymbol{v}_j^{l})\big]\right]. \qquad \text{(ii)} \end{aligned} \tag{4.10}$$

Note that for simplicity, we have defined $q(\boldsymbol{v}_j^{<1}|\boldsymbol{u};\boldsymbol{\psi}) = q(\boldsymbol{z}^{<1}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \triangleq 1$. As will be elucidated in the following subsections, the key idea behind Eq. (4.10) is: the overall KL divergence is decomposed *w.r.t.* (i) the common structural representations $\boldsymbol{z}$, and (ii) the velocity fields $\boldsymbol{v}$. The former is used to measure the intrinsic structural dissimilarity among the observed multi-modal images, while the latter serves as smoothness regularisation to enforce diffeomorphism of the transformations.

4.3.2 Intrinsic Distance over Structural Representations

As mentioned above, the KL divergence *w.r.t.* the common structural representations is used to measure the dissimilarity of the observed images for registration. Namely, instead of

using similarity measures in the image space, we propose learning an intrinsic distance over the structural representations corresponding to the observed images.

We first extract the single-view structural representation maps corresponding to each individual image, represented by the distributions $q_j^{\diamond}(\boldsymbol{z}|u_j, \boldsymbol{v}_j; \boldsymbol{\psi}_j)$, $j = 1, \ldots, N$, which is predicted from the inference steps #2 (ref. Fig. 4.1c). Note that $\boldsymbol{\psi}_j$ denotes the modality-specific variational parameters for each image, which are the same for images of the same modality. Then we define the variational posterior of the common anatomy as the *geometric mean* of these single-view posteriors, *i.e.*,

$$q(\boldsymbol{z}^l|\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\psi}) := q^*(\boldsymbol{z}^l|\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\psi}) \propto \left[\prod_{j=1}^{N} q_j^{\diamond}(\boldsymbol{z}^l|u_j, \boldsymbol{v}_j; \boldsymbol{\psi}_j)\right]^{1/N}, \tag{4.11}$$

which captures the common information among the multi-modal images. On the other hand, the prior distribution of the common anatomy is given by the *arithmetic mean* of the single-view posteriors, *i.e.*,

$$p(\boldsymbol{z}^l; \boldsymbol{\psi}) := p^+(\boldsymbol{z}^l; \boldsymbol{\psi}) \triangleq \frac{1}{N} \sum_{j=1}^{N} q_j^{\diamond}(\boldsymbol{z}^l|u_j, \boldsymbol{v}_j; \boldsymbol{\psi}_j), \tag{4.12}$$

which expresses *a priori* knowledge of the common anatomy by the mixture of experts. This setup is closely related to the *variational mixture of posteriors prior* (VampPrior) introduced in [73, 152], where the VampPrior with pseudo-inputs is chosen to maximize the ELBO. For example, Fig. 4.2 illustrates the geometric and arithmetic mean of Gaussian distributions when $N = 3$. One can observe that the three individual modes of single-view posteriors collapse to one after computing the geometric mean. Indeed, if the experts are Gaussian, one can prove that the KL divergence from the arithmetic to the geometric mean distribution is minimized when the experts are identical.

On the other hand, applying Jensen's inequality to the KL divergence yields

$$\begin{aligned} D_S &\triangleq D_{\mathrm{KL}}\big[q(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel p^+(\boldsymbol{z}^l;\boldsymbol{\psi})\big] \\ &\leq \frac{1}{N}\sum_{j=1}^{N} D_{\mathrm{KL}}\big[q(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel q_j^{\diamond}(\boldsymbol{z}^l|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j)\big] \triangleq \widetilde{D}_S, \end{aligned} \tag{4.13}$$

where $\widetilde{D}_S$ is an upper bound of the original (possibly) intractable KL divergence D_S involving mixture distributions. Thus, in practice we use this upper bound as a surrogate to minimize the KL divergence *w.r.t.* the latent variable $\boldsymbol{z}$.

Note that when minimizing $\widetilde{D}_S$ *w.r.t.* the distribution $q(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi})$, we also obtain the geometric mean $q^*(\boldsymbol{z}^l|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi})$ in Eq. (4.11), *i.e.*,

$$q^*(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) = \arg\min_{q} \widetilde{D}_S[q],$$

where $\widetilde{D}_S[q]$ is regarded a functional of the distribution q. This also justifies the usage of the geometric mean as the variational posterior of the common anatomy. In fact, since the KL divergence is a statistical distance between probability distributions, the $\widetilde{D}_S$ in Eq. (4.13) serves as an *intrinsic distance* between the structural representations of the common anatomy and each observed image. Therefore, since the geometric mean is the distribution that minimizes this intrinsic distance, the optimization of $\widetilde{D}_S^* \triangleq \widetilde{D}_S[q^*]$ *w.r.t.* $q_j^{\diamond}$'s will then seek to force the single-view structural representations $q_j^{\diamond}(\boldsymbol{z}^l|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j)$'s (which are conditioned on the transformation variables) to be identical, thus driving the registration process.

Structural Representations for Variable Group Sizes

In our previous work [167], we have used Gaussian distributions to parameterize the structural representations. That is, we assume $q_j^{\diamond}(\boldsymbol{z}|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{\boldsymbol{z},j}^{\diamond},\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond})$ where $\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond}$ are diagonal, and *for notational conciseness we omit the superscript l indicating the level of the latent variables*. We can calculate that the geometric mean is also a Gaussian

distribution, *i.e.*, $q^*(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{\boldsymbol{z}}^*,\boldsymbol{\Sigma}_{\boldsymbol{z}}^*)$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^* = N\left[\sum_{j=1}^{N}\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond\,-1}\right]^{-1}, \quad \boldsymbol{\mu}_{\boldsymbol{z}}^* = \frac{\boldsymbol{\Sigma}_{\boldsymbol{z}}^*}{N}\sum_{j=1}^{N}\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond\,-1}\boldsymbol{\mu}_{\boldsymbol{z},j}^{\diamond}. \tag{4.14}$$

Therefore, the intrinsic distance has a closed-form expression, *i.e.*,

$$\begin{aligned}&D_{\mathrm{KL}}\big[q^*(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel q_j^{\diamond}(\boldsymbol{z}|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j)\big]\\ =\ &\frac{1}{2}\left[\log\frac{\left|\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond}\right|}{\left|\boldsymbol{\Sigma}_{\boldsymbol{z}}^*\right|} + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond\,-1}\boldsymbol{\Sigma}_{\boldsymbol{z}}^*\right) + (\boldsymbol{\mu}_{\boldsymbol{z},j}^{\diamond}-\boldsymbol{\mu}_{\boldsymbol{z}}^*)^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{z},j}^{\diamond\,-1}(\boldsymbol{\mu}_{\boldsymbol{z},j}^{\diamond}-\boldsymbol{\mu}_{\boldsymbol{z}}^*)\right] + \text{const.}\end{aligned}$$

where the quadratic term is essentially a Mahalanobis distance between the structural representations.

Here, we extend the framework using categorical distribution to improve the interpretability of the structural representations. Particularly, we propose to model them by independent categorical latent variables, *i.e.*, $q_j^{\diamond}(\boldsymbol{z}|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j) = \mathrm{Cat}(\boldsymbol{z};\boldsymbol{\pi}_{j,1}^{\diamond},\ldots,\boldsymbol{\pi}_{j,K}^{\diamond})$, where $\boldsymbol{\pi}_{j,k}^{\diamond} = (\pi_{\boldsymbol{\omega},j,k}^{\diamond})_{\boldsymbol{\omega}\in\Omega} \in [0,1]^{|\Omega|}$ and $\sum_{k=1}^{K}\boldsymbol{\pi}_{j,k}^{\diamond} = \mathbf{1} \in \mathbb{R}^{|\Omega|}$. Thus, the geometric mean becomes another categorical distribution

$$q^*(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) = \mathrm{Cat}(\boldsymbol{z};\boldsymbol{\pi}_1^*,\ldots,\boldsymbol{\pi}_K^*) = \prod_{\boldsymbol{\omega}\in\Omega}\prod_{k=1}^{K}(\pi_{\boldsymbol{\omega},j,k}^*)^{z_{\boldsymbol{\omega},k}}, \tag{4.15}$$

with

$$\boldsymbol{\pi}_k^* = \frac{\left[\prod_{j=1}^{N}\boldsymbol{\pi}_{j,k}^{\diamond}\right]^{1/N}}{\sum_{k=1}^{K}\left[\prod_{j=1}^{N}\boldsymbol{\pi}_{j,k}^{\diamond}\right]^{1/N}} \in [0,1]^{|\Omega|},$$

and the intrinsic distance takes the form

$$D_{\mathrm{KL}}\big[q^*(\boldsymbol{z}|\boldsymbol{u},\boldsymbol{v};\boldsymbol{\psi}) \parallel q_j^{\diamond}(\boldsymbol{z}|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j)\big] = \sum_{k=1}^{K}\boldsymbol{\pi}_k^*\log\frac{\boldsymbol{\pi}_k^*}{\boldsymbol{\pi}_{j,k}^{\diamond}}. \tag{4.16}$$

Note that this intrinsic distance can be applied to image groups with variable group sizes, as

the geometric mean can be computed from an arbitrary number of structural representations.

Categorical Reparameterization Using Gumbel-Rao

Unfortunately, the Gumbel-Max trick [78] for sampling discrete random variables is not differentiable *w.r.t.* its parameterization, making it unsuitable for stochastic gradient estimators. Concurrent works [80, 104] have then introduced the Gumbel-Softmax (GS) distribution as a continuous relaxation of the discrete categorical variable, which admits a biased reparameterization gradient estimator. Here, we use a variant of the GS estimator, namely the Gumbel-Rao (GR) estimator [120], to further reduce the variance in gradient estimation via Rao-Blackwellisation.

Specifically, for objective functions like ELBO, the gradient *w.r.t.* the distribution parameters of an expectation over a function $f(\boldsymbol{z})$ must be computed, namely $\nabla_{\boldsymbol{\psi}}\mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\psi})}[f(\boldsymbol{z})]$, where $f(\boldsymbol{z})$ can be the likelihood function of the observed variables. To overcome the challenge of gradient computation with discrete stochasticity, the Straight-Through Gumbel-Softmax (ST-GS) estimator [80, 120] uses a continuous relaxation, giving rise to a biased Monte-Carlo gradient estimator of the form

$$\nabla^{\text{MC}}_{\text{ST-GS}} \triangleq \frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}} \frac{\mathrm{d}\,\text{softmax}_\tau(\boldsymbol{g} + \log \boldsymbol{\pi}(\boldsymbol{\psi}))}{\mathrm{d}\,\boldsymbol{\psi}}, \tag{4.17}$$

where the forward pass in $f(\cdot)$ is computed using the non-relaxed discrete samples. Based on the ST-GS, the GR estimator takes the form as

$$\nabla^{\text{MC}}_{\text{GR}} \triangleq \mathbb{E}\left[\nabla^{\text{MC}}_{\text{ST-GS}} | \boldsymbol{z}\right] \approx \frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}} \left[\frac{1}{S}\sum_{s=1}^{S} \frac{\mathrm{d}\,\text{softmax}_\tau(\boldsymbol{G}^s(\boldsymbol{\psi}))}{\mathrm{d}\,\boldsymbol{\psi}}\right], \tag{4.18}$$

where $\boldsymbol{G}^s \overset{\text{i.i.d.}}{\sim} \boldsymbol{g} + \log \boldsymbol{\pi} | \boldsymbol{z}$ for $s = 1, \ldots, S$, and $\boldsymbol{g}$ is a random vector with its entries *i.i.d.* sampled from the Gumbel distribution.

4.3.3 Spatial Regularization for Diffeomorphisms

To impose spatial smoothness of the velocity fields, we define its prior distribution by $p(\boldsymbol{v}_j^l) = \mathcal{N}(\boldsymbol{v}; \mathbf{0}, \boldsymbol{\Lambda}_{\boldsymbol{v}}^{-1})$, where the precision matrix $\boldsymbol{\Lambda}_{\boldsymbol{v}}$ is given by the scaled Laplacian matrix of a neighborhood graph on the voxel grid [43], *i.e.*, $\boldsymbol{\Lambda}_{\boldsymbol{v}} = \lambda \boldsymbol{L}_{\boldsymbol{v}} = \lambda(\boldsymbol{D}_{\boldsymbol{v}} - \boldsymbol{A}_{\boldsymbol{v}})$, with $\boldsymbol{D}_{\boldsymbol{v}}$ and $\boldsymbol{A}_{\boldsymbol{v}}$ the degree and adjacency matrices, respectively. On the other hand, the variational posterior of the velocity fields is given by a mean-field Gaussian distribution, *i.e.*, $q(\boldsymbol{v}_j^l | \boldsymbol{u}, \boldsymbol{v}^{<l}; \boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{v}; \boldsymbol{\mu}_{\boldsymbol{v},j}^l, \boldsymbol{\Sigma}_{\boldsymbol{v},j}^l)$ where $\boldsymbol{\Sigma}_{\boldsymbol{v},j}^l$ is diagonal. Therefore, the KL divergence for the velocity fields has an explicit formula

$$\begin{aligned} & D_{\mathrm{KL}}\big[q(\boldsymbol{v}_j^l | \boldsymbol{u}, \boldsymbol{v}^{<l}; \boldsymbol{\psi}) \parallel p(\boldsymbol{v}_j^l)\big] \\ &= \frac{1}{2}\left[\operatorname{tr}(\lambda \boldsymbol{D}_{\boldsymbol{v}} \boldsymbol{\Sigma}_{\boldsymbol{v},j}^l - \log \boldsymbol{\Sigma}_{\boldsymbol{v},j}^l) + \frac{\lambda}{2} \sum_r \sum_{q \in \mathcal{N}(r)} (\boldsymbol{\mu}_{\boldsymbol{v},j}^l[r] - \boldsymbol{\mu}_{\boldsymbol{v},j}^l[q])^2\right] + \text{const.} \end{aligned} \tag{4.19}$$

where $\mathcal{N}(r)$ are the neighbors of voxel r. Thus, the quadratic term over $\boldsymbol{\mu}$ enforces the velocity fields to be spatially smooth, encouraging diffeomorphic transformations.

4.4 An Interpretable Registration Architecture via Bayesian Disentanglement Learning

Most learning-based registration methods in the recent literature are based on an end-to-end training pipeline, where a black-box neural network directly predicts the spatial transformation by optimizing some image-level similarity measures [45, 14, 43, 83, 34]. However, not only do black-box end-to-end networks lack interpretability, learning directly the complex mapping from multi-modal images to their spatial correspondence disregards the underlying structural relationship and may be prone to generalization issues [52].

Bayesian deep learning [164], however, could be a remedy. It unifies probabilistic graphical models (PGM) with deep learning, and integrates the inference and perception tasks, enabling them to benefit from each other. Particularly in our context, the *perception com-*

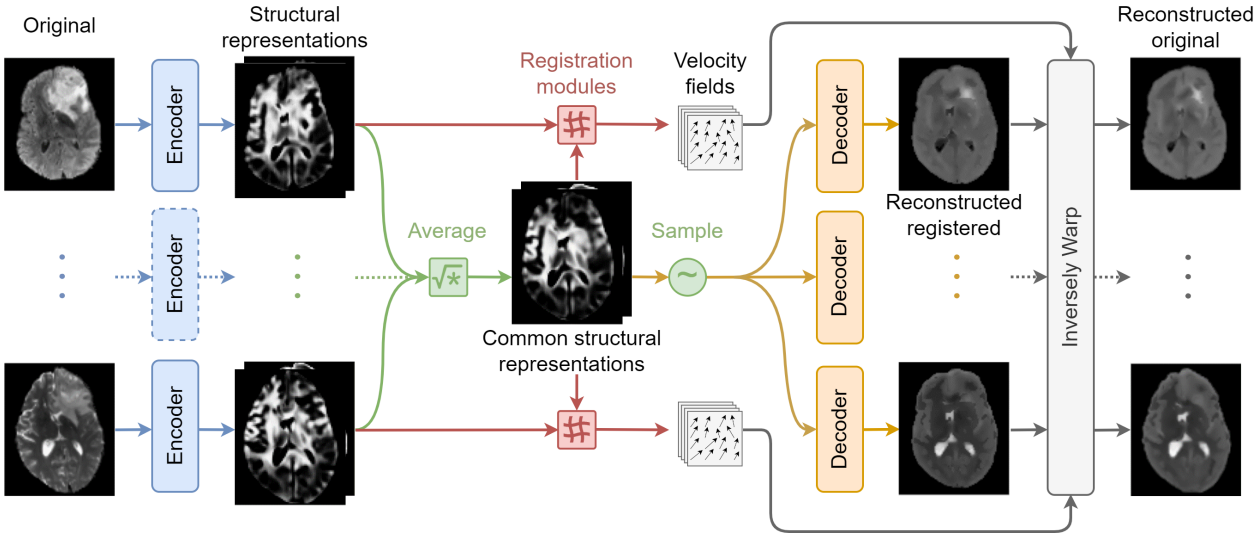


Figure 4.3: An overview of the proposed interpretable groupwise registration architecture via Bayesian disentanglement learning.

ponent includes the encoders that extract single-view posteriors from individual modalities, while the *task-specific component* comprises the registration modules and reconstruction decoders that learn jointly to infer the spatial correspondence among the input images.

Fig. 4.3 presents a schematic overview of the proposed interpretable registration architecture, specially designed based on the optimization procedure for Bayesian inference outlined in the previous sections and Fig. 4.1. The encoders first disentangle the input multi-modal images into structural representation maps with categorical distribution. Then an average representation of the common anatomy is computed within the latent space, so that spatial correspondence between the observed images and the common anatomy is predicted from these structural representations using the Diffeomorphic Demons algorithm. The disentanglement of geometric and anatomical variations is facilitated by inversely simulating the image generative process with spatially equivariant decoders, *i.e.*, reconstructing the original images from the common anatomy representation. Fig. 4.4 depicts the entire network architecture, composed of the encoders, the registration modules and the decoders. The following subsections will detail the construction of these modules.

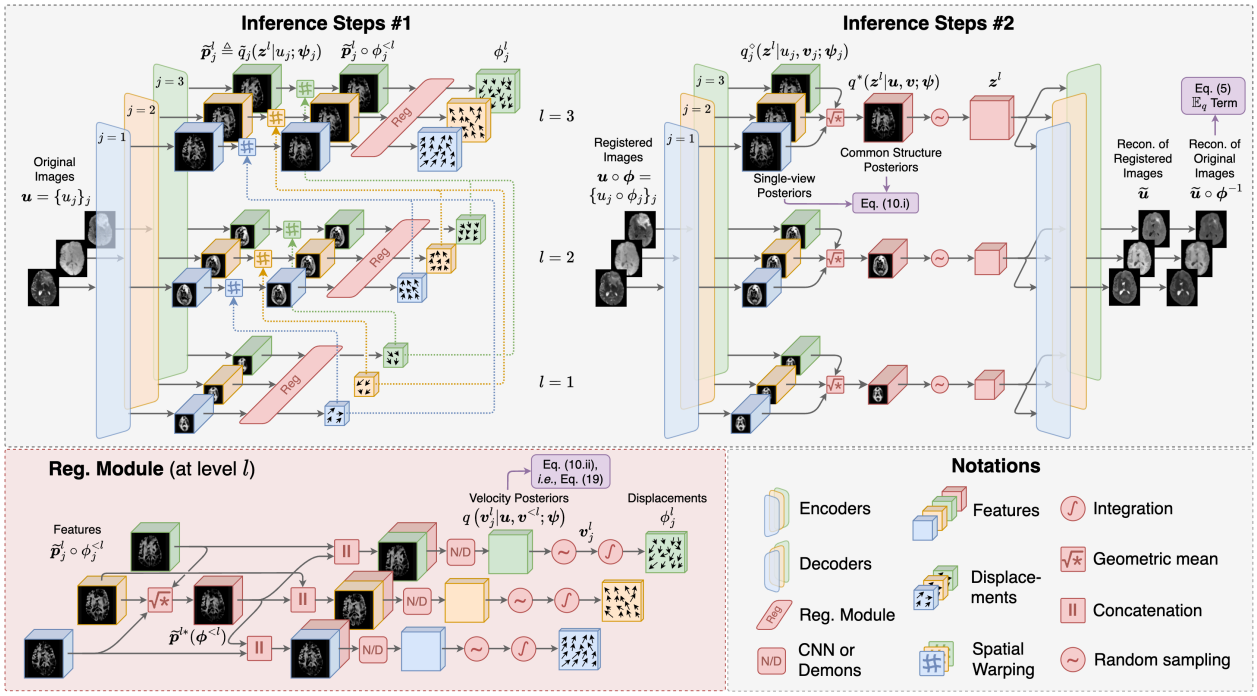


Figure 4.4: The network architecture for the proposed Bayesian groupwise registration, composed of the encoders that extract categorical structural representation maps, the registration modules that calculate multi-scale velocity fields, and the decoders that reconstruct the original images based on the common structural representations. Without loss of generality, the illustration is with $L = 3$ levels and $N = 3$ images to co-register. Note that inference steps #2 are performed only in the training stage, while in the test stage the encoder is only fed with the original image group to predict groupwise registration. The purple boxes indicate the calculation of related terms in the ELBO.

4.4.1 Inference of Structural Representations

In the inference steps #1 (ref. Fig. 4.1b), given the original misaligned image group $\boldsymbol{u}$, the encoders first estimate the multi-level categorical distributions $\widetilde{q}_j(\boldsymbol{z}^l|u_j;\boldsymbol{\psi}_j)$ for each original image u_j, where l indicates the spatial resolution of the representation. Note that $\widetilde{q}_j(\boldsymbol{z}^l|u_j;\boldsymbol{\psi}_j) \triangleq \mathrm{Cat}(\boldsymbol{z};\widetilde{\boldsymbol{p}}^l_{j,1},\ldots,\widetilde{\boldsymbol{p}}^l_{j,K})$ differs from $q^{\diamond}_j(\boldsymbol{z}^l|u_j,\boldsymbol{v}_j;\boldsymbol{\psi}_j) = \mathrm{Cat}(\boldsymbol{z};\boldsymbol{\pi}^{\diamond\,l}_{j,1},\ldots,\boldsymbol{\pi}^{\diamond\,l}_{j,K})$ regarding the network input. The former takes $\boldsymbol{u} = \{u_j\}_{j=1}^N$ as input, while the latter takes $\boldsymbol{u}\circ\boldsymbol{\phi} = \{u_j\circ\phi_j\}_{j=1}^N$. The architecture of the encoders is adapted from the Attention U-Net [115], which enhances the vanilla U-Net [130] with additive attention gates. Moreover, domain-specific batch normalization [29] is utilized in the contracting path of the encoders to cancel out modality-specific appearance variations, while all convolutional layers are shared across modalities, embedding multi-modal images into modality-invariant representations.

4.4.2 Inference of Velocity Fields

Based on the structural representation maps extracted from the original images by the encoders, the registration modules seek to estimate the transformations that spatially register the image group. This is realized in a coarse-to-fine approach, starting from prediction of the lowest-resolution velocity fields (ref. the red arrows in Fig. 4.1b).

Particularly, denoting $\widetilde{\boldsymbol{p}}^l_j \triangleq \widetilde{q}_j(\boldsymbol{z}^l|u_j;\boldsymbol{\psi}_j)$, the registration module at level l takes as input the partially registered single-view structural probability maps $\widetilde{\boldsymbol{p}}^l_j\circ\phi_j^{<l}$ and their geometric mean for the common anatomy $\widetilde{\boldsymbol{p}}^{l*}(\boldsymbol{\phi}^{<l})$, based on which the velocity field distribution $q(\boldsymbol{v}^l_j|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi})$ is predicted. Note that the spatial transformations $\boldsymbol{\phi}^{<l} \triangleq \{\phi_j^{<l}\}_{j=1}^N$ are integrated by $\phi_j^{<l} = \exp\big(\boldsymbol{v}_j^1+\cdots+\boldsymbol{v}_j^{l-1}\big)$, with $\boldsymbol{v}^l_j$ sampled from $q(\boldsymbol{v}^l_j|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{v};\boldsymbol{\mu}^l_{\boldsymbol{v},j},\boldsymbol{\Sigma}^l_{\boldsymbol{v},j})$. Finally, the total transformations that register the observed image group are given by $\phi_j = \exp\big(\boldsymbol{v}_j^+\big)$ with $\boldsymbol{v}_j^+ \triangleq \sum_{l=1}^L \boldsymbol{v}_j^l$, where $\boldsymbol{v}^l_j$ is a random sample during training while $\boldsymbol{v}^l_j = \boldsymbol{\mu}^l_{\boldsymbol{v},j}$ for testing. Note that the velocity fields are upsampled to the finest resolution before summation.

As the structural representation maps $\widetilde{\boldsymbol{p}}^l_j\circ\phi_j^{<l}$ encode the structural features of the

partially registered input images, we can in fact compute the mean $\boldsymbol{\mu}_{\boldsymbol{v},j}^{l}$ of the variational posterior $q(\boldsymbol{v}_j^l|\boldsymbol{u},\boldsymbol{v}^{<l};\boldsymbol{\psi})$ directly from them based on the Diffeomorphic Demons algorithm [149, 159]. Particularly, for each grid location $\boldsymbol{\omega}$ at level l, we define $\boldsymbol{\mu}_{\boldsymbol{v},j}^{l}$ by

$$\boldsymbol{\mu}_{\boldsymbol{v},j}^{l}(\boldsymbol{\omega}) \triangleq \alpha \cdot \widetilde{\boldsymbol{\mu}}_{\boldsymbol{v},j}^{l}(\boldsymbol{\omega}) \Big/ \max_{\boldsymbol{\omega}\in\Omega^l} \big\|\widetilde{\boldsymbol{\mu}}_{\boldsymbol{v},j}^{l}(\boldsymbol{\omega})\big\|_2,$$

where

$$\widetilde{\boldsymbol{\mu}}_{\boldsymbol{v},j}^{l}(\boldsymbol{\omega}) \triangleq -\Big[\boldsymbol{J}_{\boldsymbol{\varphi}_j^l}^{\intercal}(\boldsymbol{\omega})\boldsymbol{J}_{\boldsymbol{\varphi}_j^l}(\boldsymbol{\omega}) + \sigma_{\boldsymbol{\varphi}_j^l}^{2}(\boldsymbol{\omega})\boldsymbol{I}_d\Big]^{-1}\boldsymbol{J}_{\boldsymbol{\varphi}_j^l}^{\intercal}(\boldsymbol{\omega})\boldsymbol{\varphi}_{j,\boldsymbol{\omega}}^{l}(\boldsymbol{0})$$

is the *symmetric Demons force*, and $\alpha \in (0, \alpha_0)$ is a learnable parameter controlling the magnitude of the velocity fields. Specifically,

$$\boldsymbol{\varphi}_{j,\boldsymbol{\omega}}^{l}(\boldsymbol{u}) \triangleq \widetilde{\boldsymbol{p}}^{l*}(\boldsymbol{\omega}) - \widetilde{\boldsymbol{p}}_j^{l} \circ \phi_j^{<l} \circ (\mathrm{id} + \boldsymbol{u})(\boldsymbol{\omega})$$

is the feature difference given a displacement $\boldsymbol{u} \in \mathbb{R}^d$,

$$\boldsymbol{J}_{\boldsymbol{\varphi}_j^l}(\boldsymbol{\omega}) \triangleq -\frac{1}{2}\left[\nabla_{\boldsymbol{\omega}}^{\intercal}\,\widetilde{\boldsymbol{p}}^{l*}(\boldsymbol{\omega}) + \nabla_{\boldsymbol{\omega}}^{\intercal}\,\widetilde{\boldsymbol{p}}_j^{l} \circ \phi_j^{<l}(\boldsymbol{\omega})\right]$$

is the symmetrized Jacobian matrix at the origin, and

$$\sigma_{\boldsymbol{\varphi}_j^l}^{2} = \frac{1}{|\Omega^l| - 1}\sum_{\boldsymbol{\omega}\in\Omega^l} \big\|\boldsymbol{\varphi}_{j,\boldsymbol{\omega}}^{l}(\boldsymbol{0}) - \overline{\boldsymbol{\varphi}}_j^{l}(\boldsymbol{0})\big\|_2^2$$

is the sample variance of the feature difference norm. The variance $\boldsymbol{\Sigma}_{\boldsymbol{v},j}^{l}$ of the variational posterior is predicted by a convolutional block from the concatenation of probability maps $[\widetilde{\boldsymbol{p}}_j^{l} \circ \phi_j^{<l}; \widetilde{\boldsymbol{p}}^{l*}]$, where the network parameters are shared among different modalities.

In addition, to avoid the degeneracy that deforms the observed images to an arbitrary coordinate space, we constrain the sum of the final velocity fields to be zero by subtracting their average [11], *i.e.*, $\boldsymbol{v}_j^{+} \leftarrow \boldsymbol{v}_j^{+} - \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{v}_j^{+}$.

4.4.3 Bayesian Disentangled Representation Learning

Given the perceptual modules capturing intrinsic structural representations from multi-modal images and their spatial correspondence, we are going to discuss how such mapping functions are learnt in the first place. This is achieved by optimizing the network parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ in a closed-loop self-reconstruction process that maximizes ELBO. The single-view structural distributions for estimating the common anatomy are defined by

$$q_j^{\diamond}(\boldsymbol{z}^l|u_j, \boldsymbol{v}_j; \boldsymbol{\psi}_j) \triangleq \widetilde{q}(\boldsymbol{z}^l|u_j \circ \phi_j; \boldsymbol{\psi}_j),$$

which is obtained by feeding the transformed images $u_j \circ \phi_j$ into the encoders in the inference steps #2 (ref. Fig. 4.1c), effectively reverting the image generative process in Eq. (4.2).

Based on the common structural representations sampled from the geometric mean $q^*(\boldsymbol{z}^l|\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\psi})$, the decoders reconstruct the original images in two steps (ref. Fig. 4.1d). The images corresponding to the common anatomy $\widetilde{u}_j = f(\boldsymbol{z}; \boldsymbol{\theta}_j)$ are decoded first. Then, estimation of the original images $\widehat{\boldsymbol{u}} = \{\widehat{u}_j\}_{j=1}^N$ is obtained by inverse warping, *i.e.*, $\widehat{u}_j = \widetilde{u}_j \circ \phi_j^{-1}$ with $\phi_j^{-1} = \exp\big(-\boldsymbol{v}_j^+\big)$, simulating Eq. (4.2).

The network architecture of the decoders is given by multi-level convolutional blocks and linear upsampling. At level l, the output feature maps from upsampling are multiplied by the sampled categorical variables $\boldsymbol{z}^l$. Besides, all convolutional layers are assigned with a kernel size of 1 isotropically, to suppress spatial correlation and to form *spatial equivariance*. Thus, the decoded image $\widetilde{u}_j$ is in the same spatial orientation as the common anatomy $\boldsymbol{z}$, encouraging the registration module to identify the correct velocity fields $\boldsymbol{v}_j^+$ for reconstructing $\widehat{u}_j$. Furthermore, the batch normalization layers in the decoders are set as the inverse of the counterpart BN functions along the contracting path of the encoders, *i.e.*,

$$\boldsymbol{c}_j^{\text{out}} = \frac{\boldsymbol{c}_j^{\text{in}} - \boldsymbol{\beta}_m}{\boldsymbol{\gamma}_m + \varepsilon} \sqrt{\boldsymbol{\sigma}_m^2 + \varepsilon} + \boldsymbol{\mu}_m,$$

where for each image index j and its corresponding modality m, $\boldsymbol{c}_j^{\text{in}}, \boldsymbol{c}_j^{\text{out}} \in \mathbb{R}^{B_m \times C \times H \times W \times D}$

are the input and output feature maps, $\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m \in \mathbb{R}^C$ are the batch statistics, and $\boldsymbol{\beta}_m, \boldsymbol{\gamma}_m \in \mathbb{R}^C$ are the affine parameters. The arithmetic is computed element-wise. In fact, such preference for invertible and spatially decomposable decoders may promote learning of the true disentangled representations [86, 128].

Finally, the expectation $\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{v}|\boldsymbol{u};\boldsymbol{\psi})}[\log p(\boldsymbol{u}|\boldsymbol{z},\boldsymbol{v};\boldsymbol{\theta})]$ is approximated by Monte-Carlo ancestral sampling, where the likelihood is modeled by a Laplace distribution, *i.e.*,

$$p(\boldsymbol{u}|\boldsymbol{z},\boldsymbol{v};\boldsymbol{\theta}) \propto \prod_{j=1}^{N} \prod_{\boldsymbol{\omega}\in\Omega_j} \exp\left\{ -\frac{\left|u_j(\boldsymbol{\omega}) - f(\boldsymbol{z};\boldsymbol{\theta}_j) \circ \phi_j^{-1}(\boldsymbol{\omega})\right|}{b} \right\}, \tag{4.20}$$

where the scale parameter b is set to 1 in all experiments.

The proposed framework also encourages learning of disentangled representations formally defined by Higgins et al. [69], which formulates disentangled representations as a reflection on the decomposition of certain symmetry groups acting on the world states (or observations). This explains why we choose to decompose the latent variables of our model into the common anatomy and the corresponding spatial diffeomorphisms: because they reflect two separate symmetry structures of the observed images, *i.e.*, the ontological transformation changing the *underlying anatomy*, and the diffeomorphic transformation characterizing *geometric variation*.

4.4.4 Learning, Inference and Scalability

Learning of our model is performed in an end-to-end fashion. All model parameters, including the auto-encoders and the registration modules, are trained simultaneously by stochastic gradient ascent to maximize the ELBO. In the inference stage, the transformations that register the image group are given by $\widehat{\boldsymbol{\phi}} = \{\widehat{\phi}_j\}_{j=1}^N$, where $\widehat{\phi}_j = \exp\big(\widehat{\boldsymbol{\mu}}_j^+\big)$ and $\widehat{\boldsymbol{\mu}}_j^+$ is aggregated from the means of the velocity distributions at all levels predicted by the registration modules.

Remarkably, the proposed disentanglement of anatomy and geometry endows our model with *powerful scalability*. Since geometric and arithmetic averaging can be performed on

any number (greater than 1) of representations, our model is capable of processing image groups of arbitrary sizes during training and testing. Let the number of images in an input group be $N' = \sum_{m=1}^{M} N_m$, where N_m is the number of images of modality m. Each image of modality m is processed by the encoder for the corresponding modality to extract structural representations, and then the geometric and arithmetic means are calculated on the representations of all images to estimate the KL divergence and predict deformations for each image. For decoding, the reconstruction from the decoder of modality m is inversely warped for N_m times by the deformations for the N_m images of modality m to reconstruct original images. Thus, our model allows N_m's to be different or even 0, and only requires $N' \geq 2$ (for calculating averages) and at least two modalities are available in each group, *i.e.*, $\sum_m 1(N_m > 0) \geq 2$ (for effectively learning modality-invariant representations).

Therefore, our model can handle groups with missing modalities or variable sizes, which are common occurrences in real-world scenarios. Particularly, the scalability of our model is of great value in two aspects:

- During training, we can drop several modalities/images in each training group, thereby reducing the computational cost and/or allowing for learning from less data effectively. Typically, there are two experimental settings: 1) *complete learning*, the standard setting, where each group consists of all modalities, with one image per modality, 2) *partial learning*, where each group only contains two modalities, with one image per modality. Experiments in both settings are discussed in Sec. 4.5.
- During evaluation, our model can register groups with arbitrary sizes, enabling large-scale and variable-size multi-modal groupwise registration, as explored in Sec. 4.5.6. In contrast, conventional iterative methods optimize similarity metrics with computational complexities that escalate rapidly with group size. Deep learning baselines, on the other hand, often have fixed channel numbers to take as input the concatenation of images [67], necessitating groups to be of uniform size. These limitations significantly restrict their applicability.

4.5 Experiments and Results

We evaluated our framework on a total of four publicly available datasets, including BraTS-2021 [12, 110, 13], MS-CMRSeg [192], Learn2Reg Abdominal MR-CT [68, 176, 84], and the OASIS dataset [108]. The experiments were implemented in PyTorch [118] and conducted on NVIDIA® RTX™ 3090 GPUs. The experimental materials, baselines, evaluation metrics, and results are detailed as follows.

4.5.1 Materials

To show that our method applies to various groupwise registration tasks, especially in multi-modal scenarios, we validated it on four different datasets:

1. **MS-CMRSeg.** This dataset provides multi-sequence cardiac MR images from 45 patients, including LGE (Late Gadolinium Enhanced), bSSFP (balanced-Steady State Free Precession), and T2-weighted scans. The images were preprocessed by affine co-registration, slice selection, resampling to 1×1 mm and ROI extraction, giving 39, 15 and 44 slices for training, validation and test, respectively. Additional misalignment simulated by random FFDs using isotropic control point spacings of 5/10/20/40 mm were made to further demonstrate registration. The ROI regions were obtained by dilating the foreground mask using a circular filter of 15-pixel radius. Evaluation was conducted on the warped manual segmentations of the myocardium, left and right ventricle. *The major challenge of this dataset comes from the complex intensity patterns of the cardiac region, which can misguide the registration.*
2. **BraTS-2021.** This dataset contains multi-parametric MRI scans of glioma, including native T1, T1Gd (post-contrast T1-weighted), T2-weighted, and T2-FLAIR (Fluid Attenuated Inversion Recovery). The multi-parametric MRIs of the same patients were co-registered to the same anatomical template, interpolated to the same resolution ($1 \times 1 \times 1$ mm) and skull-stripped. We randomly selected 300, 50 and 150 patient cases for training, validation and test, respectively. The images were downsampled into

$2 \times 2 \times 2$ mm with volume size of $80 \times 96 \times 80$. To demonstrate registration, four synthetic free-form deformations (FFDs) for each image with isotropic control point spacings of 5/10/15/20 mm were generated to simulate misalignment. *The major challenge of this task stems from the complex intensity patterns around regions of tumor, yielding ambiguous anatomical correspondences among images.*

3. **Learn2Reg Abdominal MR-CT.** This dataset collects 3D T1-weighted MR and CT abdominal images. The data were resampled to $3 \times 3 \times 3$ mm and cropped to size of $112 \times 96 \times 112$. Training was performed on the unpaired 40 MR and 50 CT images [84, 176], while 8 MR-CT pairs (16 images) were used for test [39]. The masks provided by the dataset were used to confine the information used for registration. Evaluation was performed on the warped manual segmentations of the liver, spleen, left and right kidneys. *The major challenges of this dataset are the large deformation and missing correspondences between the two modalities, as well as the distribution shift between training and test datasets.*
4. **OASIS.** The OASIS-1 dataset contains 414 T1-weighted 3D MR scans from young, middle-aged, non-demented and demented older adults [108]. The images were skull-stripped, bias-corrected, and registered into an affinely-aligned, common template space with FreeSurfer [50, 74]. Evaluation was conducted on the warped manual segmentations of the cortex, subcortical grey matter, white matter, and CSF. The volumes were resampled $2 \times 2 \times 2$ mm and cropped to size of $80 \times 96 \times 96$. Besides, 287/40/87 images were randomly selected for training/validation/test. *The major challenge of this task comes from the large inter-subject variability of the brain structures.*

4.5.2 Compared Methods

We compared two types of baselines with our models in the experiments. The first type features state-of-the-art iterative methods for multi-modal groupwise registration, including accumulated pairwise estimates (APE) [162], conditional template entropy (CTE) [123], and $\mathcal{X}$-CoReg [100]. These methods do not need training, but they can be prone to image arte-

facts and can hardly register large image groups due to excessive computational burdens. The second type of baselines extends the first type, aiming to learn the groupwise spatial correspondence via a neural network using the aforementioned similarity measures. To promote a fair comparison, we used Attention U-Net [115] for the network backbone, the same architecture as the encoder in our proposed framework. *However, note that these learning-based baselines can only be trained to register images of a fixed group size, and therefore cannot be applied to data of variable group sizes in the test stage.*

For the proposed model, we implemented two variants of the registration module and adopted two training strategies. Particularly, the registration modules can be a convolutional network as in our previous work [167], or can be replaced by the Demons force proposed in this work. In addition, we compare the partial and complete learning strategies for our model. The partial learning strategy trains the network with only random image pairs, while the trained model can be used to register images of complete group sizes. Thus, the four variants of our model are:

- *Ours-PN*: with partial learning and convolutional network based registration modules.
- *Ours-CN*: with complete learning and convolutional network-based registration modules.
- *Ours-PD*: with partial learning and Demons-based registration modules.
- *Ours-CD*: with complete learning and Demons-based registration modules.

Here, "C" and "P" signify Complete and Partial learning strategies, respectively; "N" and "D" indicate Network- or Demons-based registration modules, respectively.

4.5.3 Implementation Details

During preprocessing, the intensity range of each image was linearly normalized to $[0, 1]$. For cardiac and abdominal images, they were further multiplied by the ROI mask to confine the information used by the network to predict registration. The encoder has an Attention U-Net architecture with $L = 5$ levels, and the l-th level consists of two Conv-BN-LeakyReLU blocks with kernel size 3 and $16 \times 2^{L-l}$ channels. The output of each convolutional block in

the upsampling path is passed to a Conv layer to produce logits of the categorical structural distribution, with $8 \times 2^{L-l}$ channels. Samples of the geometric mean structural distribution are passed to the decoder to reconstruct the original images. The decoder is composed of Conv-BN-LeakyReLU-Upsample blocks with kernel size 1 and $8 \times 2^{L-l}$ channels.

To facilitate learning, different weights were used for reconstruction loss, intrinsic structural distance, and registration regularisation, corresponding to a likelihood-tempered ELBO [117]. The hyperparameter for spatial regularisation was set to $\lambda = 10$ as in [43], and the maximum magnitude of the velocity calculated by the Demons algorithm was set to $\alpha_0^l = 10 \times 2^{l-L}$ at each level. The number of samples for the GR estimator was set to $S = 3$ in all experiments. Network optimization was performed by stochastic gradient descent using the Adam optimizer [87]. Besides, we used a learning rate of 1×10^{-3} and a batch size of 20 or 2 for the MS-CMRSeg or the other datasets, respectively.

4.5.4 Evaluation Metrics

For each test image group in an experiment, groupwise semantic evaluation metrics can be constructed by averaging the pairwise version over all possible pairs of the propagated segmentation masks. In other words, a metric on the n-th group $\{u_i\}_{i=1}^{N_n}$ is calculated as

$$\mathrm{Eva}_n \triangleq \frac{1}{\binom{N_n}{2}} \sum_{\substack{1 \le i,j \le N_n \\ i \neq j}} \mathrm{Eva}(y_i \circ \widehat{\phi}_i, y_j \circ \widehat{\phi}_j),$$

where Eva is the Dice similarity coefficient (DSC) or average symmetric surface distance (ASSD), y_i, y_j are distorted masks associated with their respective images u_i, u_j, and $\widehat{\phi}_i, \widehat{\phi}_j$ are the corresponding predicted transformations. The masks are used for model evaluation only.

For experiments on the BraTS-2021 dataset, since the ground-truth spatial correspondence is available, the groupwise warping index (gWI) can be implemented as an additional metric, which measures the root mean squared error for unbiased deformation recovery based

on the ground-truth and predicted deformation fields [100], *i.e.*,

$$\mathrm{gWI}_n \triangleq \frac{1}{N_n}\sum_{i=1}^{N_n}\sqrt{\frac{1}{\left|\widehat{\Omega}_i^f\right|}\sum_{\boldsymbol{\omega}\in\widehat{\Omega}_j^f}\|\bar{r}_j(\boldsymbol{\omega})\|_2^2}, \tag{4.21}$$

where $\widehat{\Omega}_j^f \triangleq \left\{\boldsymbol{\omega}\in\Omega|\phi_j^\dagger\circ\widehat{\phi}_j(\boldsymbol{\omega})\in F\right\}$ with F the foreground region of the initial phantom before distortion and $\phi_j^\dagger$ the ground-truth deformation for u_j, and

$$\bar{r}_j(\boldsymbol{\omega}) \triangleq r_j(\boldsymbol{\omega}) - \frac{1}{N_n}\sum_{j'=1}^{N_n} r_{j'}(\boldsymbol{\omega}), \quad r_j(\boldsymbol{\omega}) \triangleq \phi_j^\dagger\circ\widehat{\phi}_j(\boldsymbol{\omega}) - \boldsymbol{\omega}. \tag{4.22}$$

The gWI provides a fine-grained measurement for voxel-wise co-registration accuracies, whereas the semantic evaluation metrics attend to high-level structural concurrence of distinct tissues. Besides, since all data have isotropic physical spacing, we measured ASSD and gWI in voxel units for convenience.

4.5.5 Multi-Modal & Intersubject Groupwise Registration

Experimental Design

This experiment aims to showcase the performance of our model under various data conditions using the four datasets. These conditions include multi-modal data (all datasets except OASIS), heterogeneous data such as BraTS-2021 of brain tumors and MS-CMRSeg of myocardium infarction, and intersubject groupwise registration (OASIS). In addition, our objective was to demonstrate the efficiency of the partial learning strategy for our model. To this end, we examined four variants of our model, namely *Ours-CN*, *Ours-CD*, *Ours-PN*, and *Ours-PD*.

In this experiment, the test groups of each dataset maintained their original sizes, which was equal to the training group sizes used for both the baselines and complete learning of our model. However, in the case of partial learning for our model, the training group size was consistently set to 2, achieved by randomly selecting images before the experiment, rather

than in each training iteration. This resulted in a reduced number of training images, posing additional challenges for model generalizability. Note that our model does not have a partial learning version for the Learn2Reg dataset (as each original group is an MR-CT pair) and the OASIS dataset (as it is mono-modal).

Quantitative Results of Registration Accuracy

Tab. 4.2 summarizes the means and standard deviations of the evaluation metrics (DSC, ASSD, and gWI) on the four datasets before and after groupwise registration by different methods. The effectiveness of iterative and deep learning baselines varies across datasets. For example, all iterative approaches show inferior performance on MS-CMRSeg and Learn2Reg compared to deep learning baselines. Particularly on Learn2Reg, the evaluation reveals that APE and CTE yield marginal improvements in mean DSC, and all iterative methods even result in worse mean ASSDs after registration. In addition, deep learning baselines performed worse than APE and $\mathcal{X}$-CoReg on BraTS-2021, and all baselines (except for CTE) exhibit similar performance on OASIS. These results show the instability of both types of baselines under complex image conditions. Besides, their performance was achieved at the expense of either time-consuming iterative optimization for each test group, or a substantial increase in parameter count for deep learning baselines.

In contrast, our model with complete learning attained notable improvements in registration accuracy, consistently outperforming all other methods across all datasets. For example, the two variants of our model demonstrate an over 15% enhancement in DSC and an over 33% reduction in ASSD on Learn2Reg, compared with the best baseline. On BraTS-2021, our model also surpasses all baselines, especially in gWI. It is important to highlight that gWI assesses the registration accuracy for all spatial locations within the images, providing a comprehensive evaluation, while DSC and ASSD are based only on tumor region. Therefore, the results showcase the superiority and versatility of our model in aligning various brain structures with normal and pathological tissues and diverse imaging patterns.

Remarkably, the proposed method also exhibits a significant reduction in parameters for

Table 4.2: Evaluation metrics of the groupwise registration results on the MS-CMRSeg, BraTS-2021, Learn2Reg, and OASIS datasets. The top and second-best results for each dataset are highlighted in bold and underline, respectively. The ASSDs were measured in voxel units. The parameter counts are expressed in millions, and for our model there are test and training (in parentheses) values. The p-values were computed using the gWIs or DSCs (for the BraTS-2021 and other datasets, respectively) between the method *Ours-CD* and the others with a two-sided paired t-test. $|\det J_\phi \leq 0|$ represents the proportion (in %) of voxels with negative Jacobian determinants in the predicted displacements, where the values were first calculated for the foreground region of each registered image and then averaged over all images among all test groups.

Method	MS-CMRSeg				BraTS-2021				
	DSC ↑	ASSD ↓	#Params	p-value	gWI ↓	DSC ↑	ASSD ↓	#Params	p-value
None	$.722 \pm .104$	3.86 ± 1.43	N/A	$< 10^{-10}$	1.430 ± 0.644	$.610 \pm .150$	0.93 ± 0.42	N/A	$< 10^{-10}$
APE [162]	$.746 \pm .101$	3.48 ± 1.36	0.154	$< 10^{-10}$	0.629 ± 0.141	$\underline{.707 \pm .069}$	**0.62±0.12**	7.37	1.3×10^{-1}
CTE [123]	$.766 \pm .100$	3.15 ± 1.32	0.154	$< 10^{-10}$	1.223 ± 0.255	$.500 \pm .102$	1.27 ± 0.45	7.37	$< 10^{-10}$
$\mathcal{X}$-CoReg [100]	$.757 \pm .107$	3.31 ± 1.40	0.154	$< 10^{-10}$	0.728 ± 0.196	$.698 \pm .086$	0.65 ± 0.17	7.37	$< 10^{-10}$
APE-Att	$.799 \pm .061$	2.78 ± 0.72	8.04	$< 10^{-10}$	0.757 ± 0.153	$.690 \pm .078$	0.67 ± 0.14	22.95	$< 10^{-10}$
CTE-Att	$.820 \pm .066$	2.45 ± 0.74	8.04	$< 10^{-10}$	0.916 ± 0.210	$.661 \pm .094$	0.75 ± 0.20	22.95	$< 10^{-10}$
Ours-PN	$.803 \pm .062$	2.68 ± 0.70	5.06 (5.22)	$< 10^{-10}$	$\underline{0.608 \pm 0.115}$	$.670 \pm .071$	0.69 ± 0.13	14.91 (15.10)	$< 10^{-10}$
Ours-CN	**.842±.051**	**2.17±0.58**	5.06 (5.22)	$< 10^{-10}$	**0.538 ± 0.108**	**.709±.063**	$\underline{0.62 \pm 0.15}$	14.91 (15.10)	$< 10^{-10}$
Ours-PD	$.799 \pm .060$	2.74 ± 0.65	1.91 (2.85)	$< 10^{-10}$	0.720 ± 0.151	$.670 \pm .072$	0.70 ± 0.14	5.44 (15.06)	$< 10^{-10}$
Ours-CD	$\underline{.836 \pm .043}$	$\underline{2.21 \pm 0.47}$	1.91 (2.85)	N/A	0.663 ± 0.129	$.669 \pm .074$	0.70 ± 0.15	5.44 (15.06)	N/A

Method	Learn2Reg				OASIS				
	DSC ↑	ASSD ↓	#Params	p-value	$\lvert\det J_\phi \leq 0\rvert$ (%) ↓	DSC ↑	ASSD ↓	#Params	p-value
None	$.415 \pm .160$	5.27 ± 3.01	N/A	2.4×10^{-4}	N/A	$.614 \pm .027$	0.96 ± 0.10	N/A	$< 10^{-10}$
APE [162]	$.554 \pm .384$	6.14 ± 7.66	10.52	9.7×10^{-2}	$.5297 \pm .0814$	$.777 \pm .030$	0.59 ± 0.09	8.85	7.9×10^{-8}
CTE [123]	$.514 \pm .373$	6.85 ± 7.89	10.52	5.1×10^{-2}	$.2291 \pm .0447$	$.746 \pm .031$	0.62 ± 0.09	8.85	$< 10^{-10}$
$\mathcal{X}$-CoReg [100]	$.664 \pm .361$	6.86 ± 14.1	10.52	3.0×10^{-1}	$.1330 \pm .0404$	$.777 \pm .017$	0.54 ± 0.05	8.85	$< 10^{-10}$
APE-Att	$.687 \pm .092$	2.09 ± 0.57	22.95	2.9×10^{-3}	$.0479 \pm .0130$	$.777 \pm .018$	0.57 ± 0.05	22.95	$< 10^{-10}$
CTE-Att	$.679 \pm .069$	2.17 ± 0.56	22.95	1.4×10^{-2}	$.0552 \pm .0154$	$.773 \pm .039$	0.59 ± 0.13	22.95	4.2×10^{-6}
Ours-CN	**.803±.084**	**1.32±0.57**	14.90 (15.10)	9.3×10^{-2}	$.1746 \pm .0305$	**.803±.007**	**0.49±0.02**	13.16 (13.26)	$< 10^{-10}$
Ours-CD	$\underline{.793 \pm .086}$	$\underline{1.39 \pm 0.60}$	5.43 (15.06)	N/A	**$.0066 \pm .0029$**	$\underline{.791 \pm .008}$	$\underline{0.51 \pm 0.02}$	3.69 (13.22)	N/A

both training and test. For instance, compared with the deep learning baselines, our models with CNN or Demons based registration modules achieve reductions of 37% or 76% (for test) on MS-CMRSeg, respectively. Notably, the Demons-based models, namely *Ours-PD*

and *Ours-CD*, require even fewer parameters than iterative methods on the 3D datasets during test. This demonstrates the parameter efficiency of Demons-based variants of the proposed model.

In addition, the partial learning strategy allows our model to reduce the training image group size to 2, while maintaining performance equal to or surpassing that of the baselines. This reduction results in a 33%/50%/50% (for the MS-CMRSeg/BraTS-2021/OASIS datasets) decrease in training image resources and approximately the same reduction ratio in memory footprint during training. The results illustrate that our model, trained with partial learning and reduced computational costs on a limited number of images, achieves a level of generalizability comparable to the baselines that leverage the entire dataset. This showcases the potential of our approach in scenarios involving image groups with absent modalities and limited resources.

Moreover, examining the proportions of voxels with negative Jacobian determinants in the displacements predicted by various methods on the OASIS dataset reveals that all methods exhibit decent smoothness. Notably, *Ours-CD* achieved a significantly lower proportion than other methods, highlighting the advantage of the proposed registration modules with the Demons algorithm in predicting diffeomorphic transformations.

Robustness on Different Initial Misalignment

Fig. 4.5 presents the violin plots of the evaluation metrics for the compared methods on all test groups. For all datasets (particularly MSCMR, Learn2Reg and BraTS-2021), the initial DSCs/ASSDs/gWIs of the test image groups exhibit a broad spread, indicating considerable variability of misalignments. However, iterative methods achieved only marginal alignments for nearly all groups of the MSCMR dataset, and even led to worse distortion for certain Learn2Reg image groups, even though such methods incorporate a combination of rigid, affine, and FFD transformations to enhance registration outcomes. This implies that the limited capture range of iterative methods makes it difficult to establish reasonable spatial correspondence across long distances. This challenge may arise from the presence

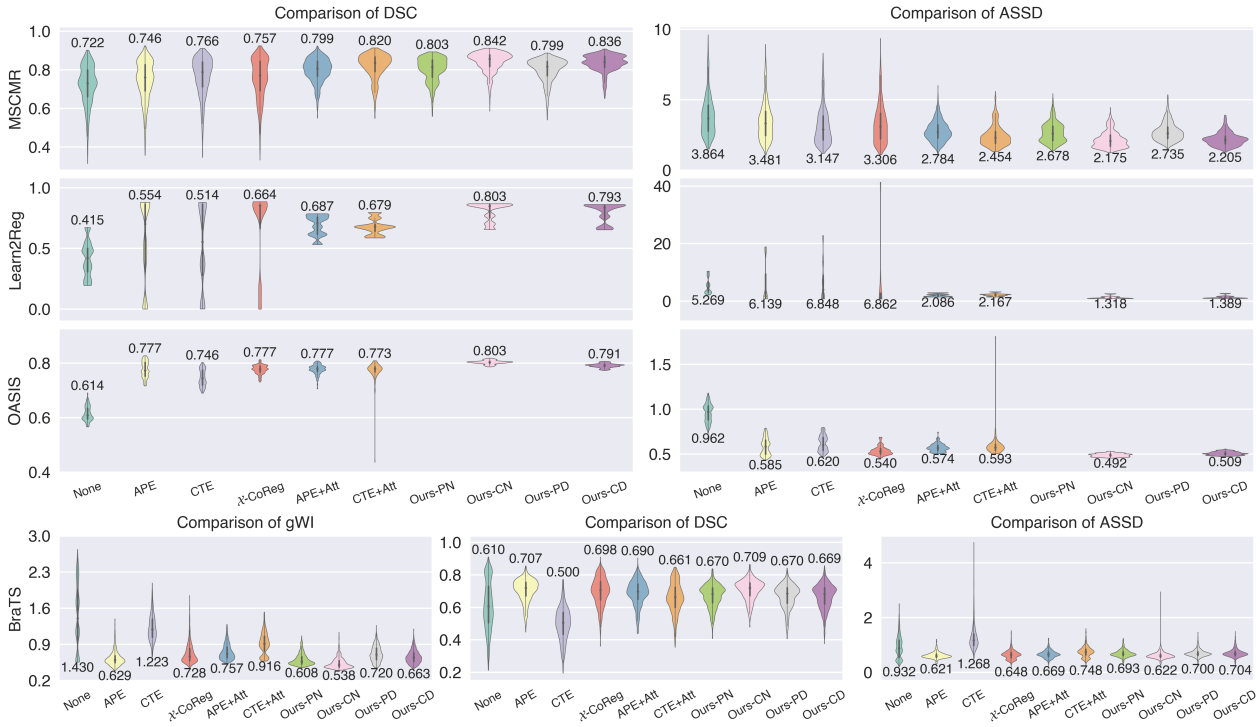


Figure 4.5: Quantitative evaluation metrics of the compared methods on the test groups of the four datasets. The mean values from each method are indicated.

of local minima in the optimization of manually crafted similarity metrics. Although deep learning baselines can mitigate the wide range of distortion in misaligned images, the overall improvement by them remains modest on certain datasets such as Learn2Reg.

In contrast, the violin plots of our models with complete learning achieve shorter tails (smaller variance in registration metrics) for all datasets, exhibiting greater stability when handling varying levels of distortion, and showing superior performance especially for image groups with significant misalignments.

Visualization of Registration Results

The results of an example test group from the MS-CMRSeg and Learn2Reg datasets are visualized in Fig. 4.6 and Fig. 4.7, respectively. For the example from MS-CMRSeg, the original images were severely distorted with initial DSCs below 0.4. Under such con-

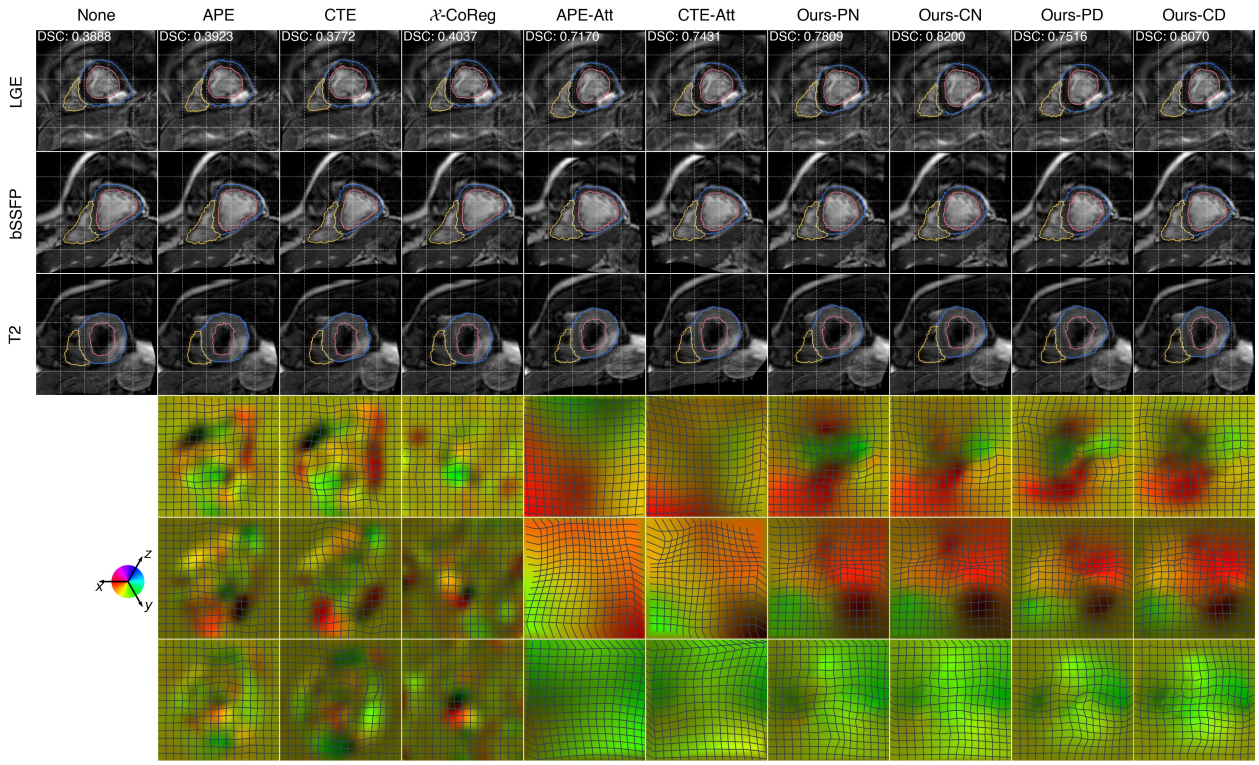


Figure 4.6: Results of an image group from the MS-CMRSeg dataset. The mean DSCs of all foreground classes on this group are shown for each method.

dition, the iterative methods were rather conservative to output small-magnitude deformations, illustrated by the deformation grids. For the group from Learn2Reg, conversely, the iterative methods estimated highly irregular deformation fields containing numerous non-diffeomorphic positions, as shown by the displacement and Jacobian determinant maps, although they achieved DSCs comparable to the proposed methods on this image group. This illustrates the instability of the deformations produced by iterative methods across different images. In contrast, while deep learning baselines avoid registration failures for all datasets, their output deformation fields imply that such an effect comes at the cost of a tendency to predict more conservative deformations with high rigidity.

The proposed model, however, demonstrates an advantage in addressing these issues, capable of generating fine-grained deformations that contain both global and localised subtle

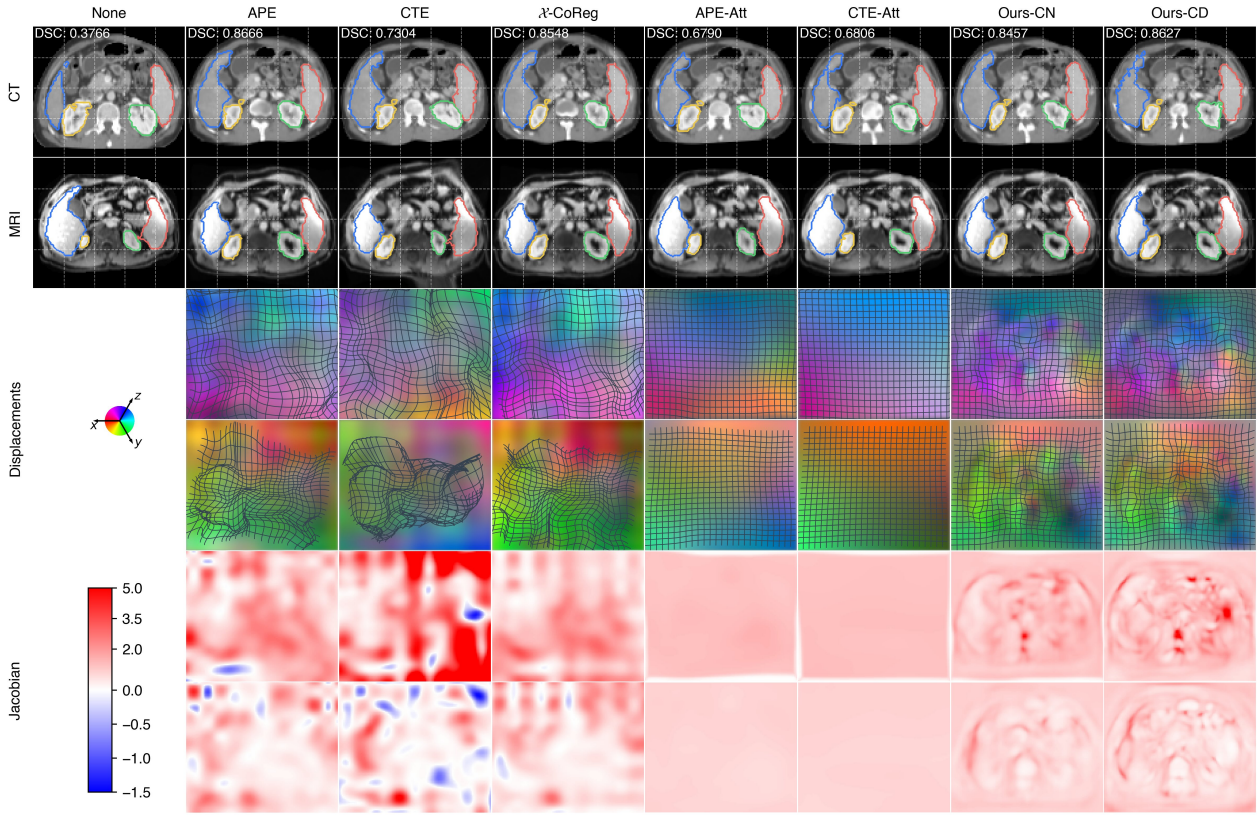


Figure 4.7: Results of an image group from the Learn2Reg dataset. The mean DSCs of all foreground classes on this group are shown for each method.

movements. To further showcase the effectiveness of the proposed hierarchical decomposition strategy for estimating the velocity fields, in Fig. 4.8 we visualize an example of the multi-level deformations produced by our model on an image group from the MS-CMRSeg dataset. The estimated spatial transformations $\{\phi_m^l\}_m$ with a lower level l tend to be smoother and consist of global displacements, while the higher-level deformations concentrate on small distortions in different local regions.

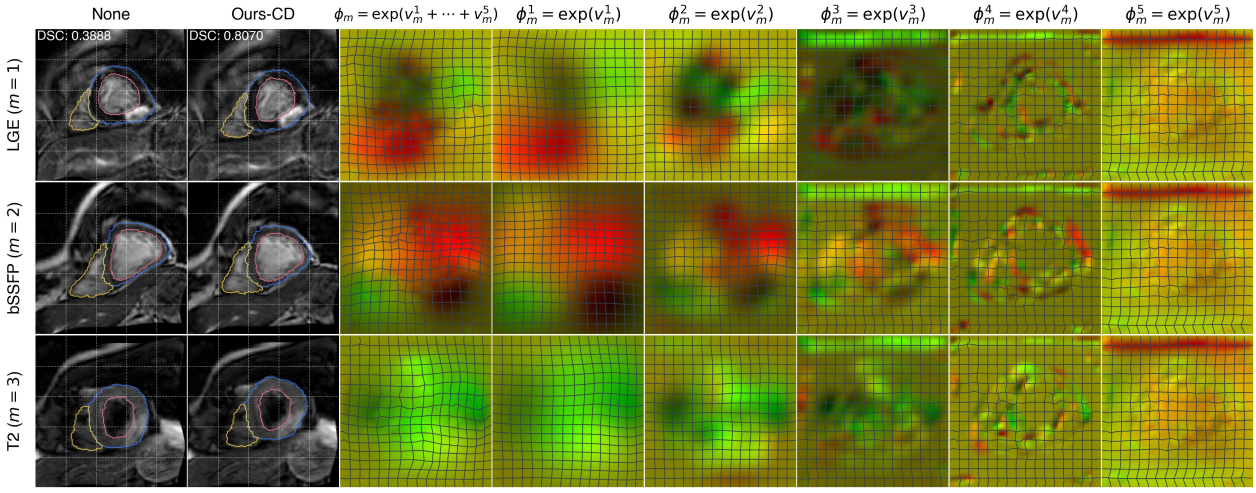


Figure 4.8: Multi-level deformations from our model *Ours-CD* on MS-CMRSeg, where the image group to register is the same as in Fig. 4.6.

4.5.6 Scalability Test on Large-scale and Variable-size Image Groups

Experimental Design

This experiment aims to evaluate the performance of our model for large-scale and variable-size groupwise registration. Co-registering large image groups poses a significant challenge for iterative methods due to their high computational complexity. On the other hand, traditional deep learning approaches typically fix the input channel number as a pre-determined group size, limiting them to handling only small groups of the same size during training and test due to GPU memory constraints. In contrast, while only requiring a small group size for training, our model can generalize effectively to much larger image groups of varying sizes during inference stage. The results in this section illustrate this scalability.

The original group sizes of the BraTS-2021, MS-CMRSeg, Learn2Reg and OASIS datasets were $N_{\text{train}} = 4, 3, 2, 4$, respectively. To test the scalability of our model, we constructed test groups of a larger size for each dataset, by merging every R original groups as one new group. Therefore, using the original test set with W groups of size N_{train}, we constructed a new test set with W/R groups of size $N_{\text{test}} = N_{\text{train}}R$, where $W \equiv 0 \pmod{R}$. The variants of our

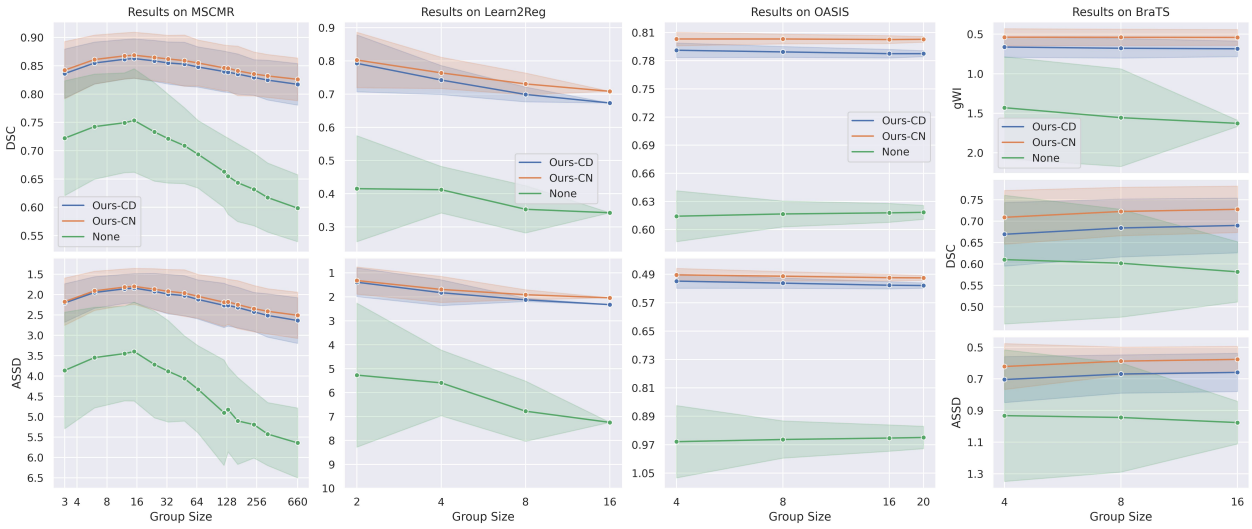


Figure 4.9: Evaluation metrics (mean values with one standard deviation bands) of registration results on image groups with different sizes.

model (*Ours-CN* and *Ours-CD*) were trained on the original training groups, and then tested on newly constructed test sets with different N_{test}.

Results

Fig. 4.9 presents the evaluation metrics versus N_{test} on different datasets. The initial DSC/ASSD/gWI metrics indicate that as N_{test} increases, the initial misalignment becomes significantly severe, posing greater challenges for co-registration. However, our model sustains good performance even when group size is very large (e.g., more than 600 2D images from MS-CMRSeg and 20 3D images from OASIS). Besides, while the pre-registration metrics on BraTS-2021 worsen with increasing group size, our models achieved even better registration accuracy when co-registering larger image groups. These demonstrate robustness of the proposed framework on large-scale multi-modal groupwise registration. In addition, the two variants of our model exhibit similar scalability, showing the effectiveness of both types of registration modules.

Note that the maximum test group size for Learn2Reg and BraTS-2021 is 16, because Learn2Reg contains only 16 test images, and for BraTS, it is infeasible to register intersubject

images due to difference in tumor structures, and there are only 16 distorted images for each subject.

4.5.7 Integration with Other State-of-the-Art Registration Methods

Experimental Design

This experiment aims to integrate state-of-the-art (SoTA) methods from other tasks (e.g., pairwise or mono-modal registration) into the proposed framework, and therefore evaluates the compatibility and versatility of our method. To this end, we selected the following recent works:

- TransMorph [34]. We replaced the Attention U-Net encoder of our model with TransMorph to extract features from different image modalities.
- PIViT [103]. We replaced the registration modules of our model by the registration networks proposed by PIViT. Particularly, as both our and the PIViT models have five levels of scales (resolutions) for spatial transformation inference, we follow the same settings in that paper, i.e., we use the LCD modules (Long-range Correlation Decoder) proposed by the PIViT paper for the three coarsest levels, and use CNNs for the other two levels.
- ModeT [166]. We replaced the registration modules of our model by the registration networks proposed in the ModeT paper. Particularly, as both our and the ModeT models have five levels, we follow the same settings in that paper, i.e., for each of the three coarest levels, we first use a ModeT module (Motion Decomposition Transformer) to infer multiple transformations, and then use a CWM (Competitive Weighting Module) to generate a single transformation; for each of the other two levels, we use a ModeT module to infer a single transformation directly.

Note that the model Ours-CN uses the networks in the registration modules to produce both mean and log variance of the velocity field distributions, while Ours-CD calculated the mean of the velocity field distributions using the Demons algorithm and uses the networks in the registration modules to produce logarithmic variance only. Therefore, with Transmorph or

Table 4.3: The results of our models integrated with SoTA methods on the MS-CMRSeg dataset.

Model	Encoder	Reg. Module	DSC↑	ASSD↓
Ours-CN	Attention U-Net	Convs	0.842±0.051	2.17±0.58
Ours-CD	Attention U-Net	Convs	0.836±0.043	2.21±0.47
Ours-CN	TransMorph	Convs	0.838±0.051	2.21±0.54
Ours-CD	TransMorph	Convs	0.848±0.048	2.09±0.52
Ours-CN	Attention U-Net	PIViT	0.841±0.043	2.18±0.47
Ours-CD	Attention U-Net	PIViT	0.838±0.045	2.26±0.49
Ours-CN	Attention U-Net	ModeT	0.848±0.040	2.12±0.42

PIViT networks integrated, our model still has these two types. As for ModeT, since it was designed to calculate spatial transformations based on a set of basis vectors, it is not suitable to be used as a predictor of logarithmic variance. As a result, we integrate ModeT only into Ours-CN.

Results

The results on the MS-CMRSeg dataset are shown in Tab. 4.3. Our models integrated with other SoTA methods achieved a similar level of performance compared to the results we obtained before (first two rows in the table). In particular, Ours-CD with TransMorph encoders and Ours-CN with ModeT registration modules even surpass the previous best performance. Therefore, it is demonstrated that our method has good versatility, compatible with different SoTA methods. Besides, this indicates that we can effectively apply methods that dominate various registration tasks to multi-modal groupwise registration, just through integration with our framework.

4.5.8 Model Interpretability

We further demonstrate the interpretability of the proposed framework by 1) visualizing the structural representation maps learnt by our model, and 2) validating the symmetry structures respected by our model.

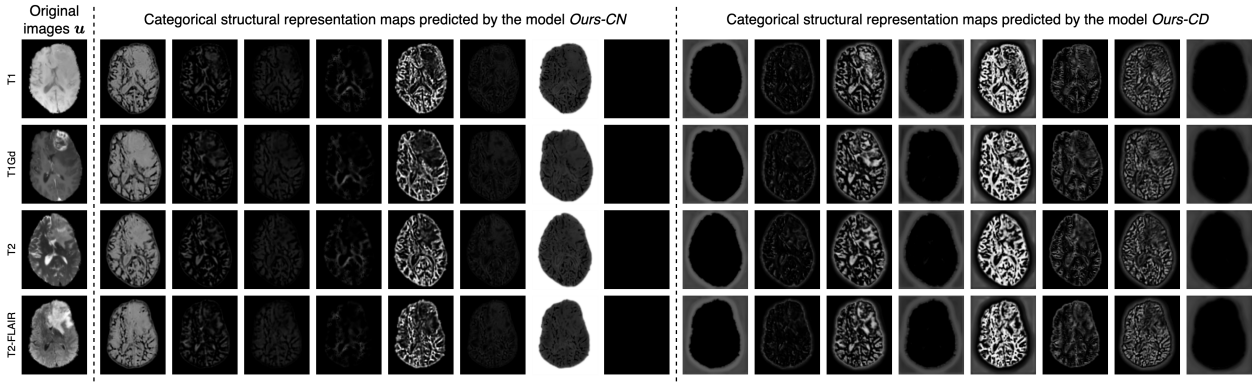


Figure 4.10: Categorical structural representations from the proposed models. One can see that complementary brain structures are revealed from these representations. Particularly, the representations from the model *Ours-CD* look more fine-grained than those from *Ours-CN*.

Structural Representation Maps

We visualize the structural representation maps as the categorical latent distribution learnt by the proposed model. For instance, Fig. 4.10 presents the finest-scale structural representation maps estimated by the proposed models of an image group from the BraTS-2021 dataset. One can observe that meaningful semantic features are extracted from the input images. Particularly, the model *Ours-CN* tends to learn high-level information, e.g. structures and boundaries, while *Ours-CD* prefers to extract more detailed features such as skeletons and textures. This difference is reasonable because the registration is computed directly from these representations for the Demons algorithm, which needs more fine-grained information.

These structural representations also demonstrate the clear advantage of our proposed models over conventional learning-based counterparts in terms of model interpretability and registration accuracy. This suggests that the trade-off between model interpretability and performance is not inevitable, while one may believe that more complex models are more accurate. In fact, since medical images are scarce and highly correlated [100], the observations are well structured and only occupy a low-dimensional landscape (or manifold) within

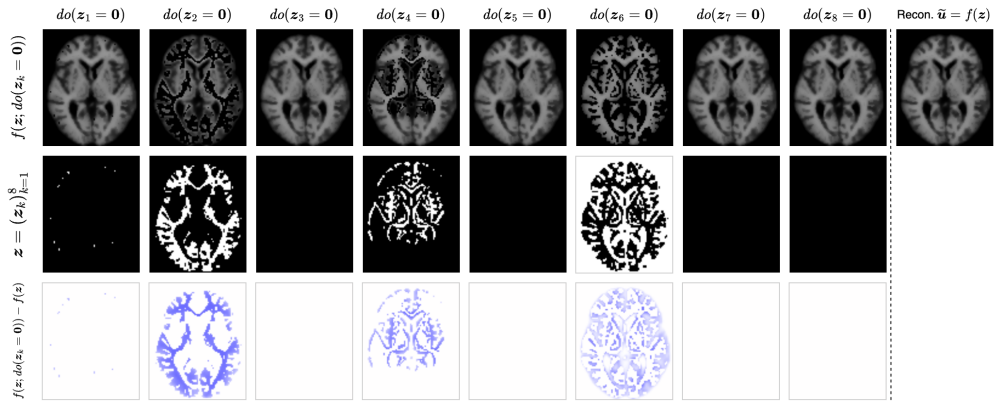
the entire data space [6]. Therefore, by respecting the underlying generating process and the inherent symmetry structures, the proposed models achieved better interpretability and higher accuracy, while using fewer parameters.

Counterfactual Validation of Latent Symmetries

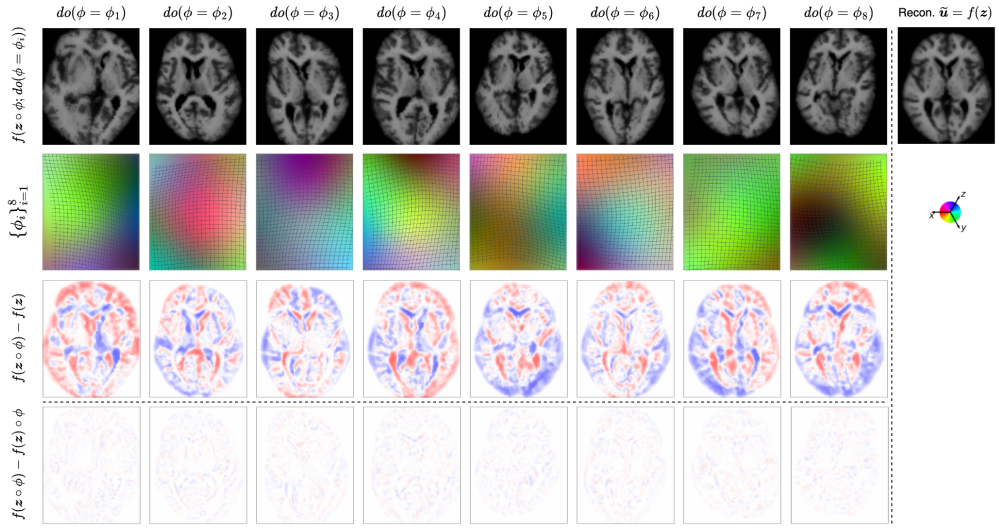
A deeper level of interpretability can be assessed by examining if the model respects the underlying symmetries of the data-generating process, namely the disentanglement of anatomy and geometry. We validate this by performing counterfactual reconstructions [121, 18], where we actively intervene on the learnt latent variables and observe the outcome. This allows us to ask "what-if" questions and verify that the model behaves in a predictable and semantically meaningful way.

We first investigate the disentangled anatomy representation $\boldsymbol{z}$. For a given test image group, we compute the latent anatomy $\boldsymbol{z}$ and then perform an ontological intervention by setting a specific channel $\boldsymbol{z}_k$ to zero (i.e., a do($\boldsymbol{z}_k = \mathbf{0}$) operation). This corresponds to asking the model: "What would the registered images look like if this specific anatomical feature k did not exist". The results of this experiment on the OASIS dataset are shown in Fig. 4.11a. The figure demonstrates that zeroing out a single channel cleanly removes a specific anatomical structure (e.g., the white matter) from the reconstruction. The difference map between the original and counterfactual reconstructions–the direct effect–clearly isolates this structure. To quantify this observation, we measure the Normalized Cross-Correlation (NCC) between the map of the latent channel $\boldsymbol{z}_k$ that was set to zero and the corresponding direct effect map. For the example in Fig. 4.11a, the four non-zero channels $\boldsymbol{z}_1$, $\boldsymbol{z}_2$. $\boldsymbol{z}_4$ and $\boldsymbol{z}_6$ yielded high NCC values of 0.99, 0.98, 0.97 and 0.62, respectively. This high correlation confirms that the model has learnt a highly disentangled representation, where individual latent channels have a direct and exclusive correspondence to specific anatomical structures.

This level of granular control and semantic correspondence is a key advantage over conventional end-to-end registration networks. In such "black-box" models, the latent features are typically entangled, and manipulating a single feature is unlikely to produce a coherent



(a) Counterfactual reconstruction by ontological transformations.



(b) Counterfactual reconstruction by diffeomorphic transformations.

Figure 4.11: Counterfactual reconstruction on an image from the OASIS dataset using the underlying symmetry transformations. (a) For ontological transformations acting in the anatomy domain, one can see that the difference calculated by $f(\boldsymbol{z}; do(\boldsymbol{z}_k = \mathbf{0})) - f(\boldsymbol{z})$ indeed corresponds to the k-th latent structure $\boldsymbol{z}_k$, which means that the learnt decoder respects the ontological symmetry. (b) For diffeomorphic transformations acting in the spatial domain, one can observe that the equivariance difference $f(\boldsymbol{z} \circ \phi_i) - f(\boldsymbol{z}) \circ \phi_i$, where $f(\boldsymbol{z} \circ \phi_i) = f(\boldsymbol{z} \circ \phi; do(\phi = \phi_i))$, are almost zero except for interpolation errors, which indicates that the learnt decoder is indeed transformation-equivariant.

or predictable semantic change. Our framework, by explicitly modeling anatomy and geometry as disentangled latent variables, provides interpretable latent that can be used to understand and validate the model's internal logic, a crucial step towards trustworthy medical AI. We also validate the geometric equivariance in Fig. 4.11b, showing that the model correctly disentangles spatial transformations.

4.6 Conclusion and Discussion

In this work, we have developed a Bayesian framework for unsupervised multi-modal groupwise image registration. In marked contrast to similarity-based registration approaches, the proposed method builds on a principled generative modeling of the imaging process. Moreover, a specially designed network architecture realizes the explicit disentanglement of anatomy and geometry from the observed images, so that registration is learnt from their underlying structural representations in a unified closed-loop self-reconstruction process. The experiments on four different datasets of cardiac, brain, and abdominal medical images have demonstrated the advantage of the proposed modeling for image registration.

There have also been works in the literature trying to learn disentanglement of appearance and geometry [137, 139, 23, 175], or inference of semantic or geometric variations [180, 107]. However, they usually involve modeling of an intensity template, and do not consider a more general multi-modality setup. The generated template trained on one dataset can be suboptimal for groupwise registration on a new dataset. This highlights the advantage of disentangling each separate image groups into their corresponding anatomical representations in our framework as a subroutine to groupwise registration, rather than optimizing for intensity variations.

The intrinsic distance proposed in Sec. 4.3.2 also has further implications. Developed based on the variational inference framework, the intrinsic distance is primarily intended to estimate the difference between structural representations of the common anatomy and the input images. Nevertheless, there could be alternative methods to calculate the average

representation and the intrinsic distance. For instance, conventional label fusion methods for multi-atlas segmentation [165, 1, 10] could be used to estimate the average probability maps of the common anatomical representations, while the intrinsic metrics between probabilistic shapes could be improved by respecting the underlying manifold structure [51, 8].

Furthermore, recent years have witnessed a growing academic interest in the field of interpretable or explainable artificial intelligence (XAI) [61, 6, 151, 188]. Our proposed model exhibits the local and active interpretability characterized by [188] in that the structural representations encoded by the network are deliberately devised to have visual semantics in a human-understandable fashion. The proposed registration module and the network architecture also demonstrate certain levels of algorithmic transparency [6]. Indeed, the Demons algorithm for calculating the velocity fields is mathematically interpretable, as it depends linearly on the gradient and difference of structural representations. On the other hand, the decoder architecture is spatially decomposable, as the reconstruction of image intensities is determined locally by the anatomy at each independent location (owing to the use of convolutions with kernel size 1 isotropically), thus respecting the spatial equivariance in the imaging process *w.r.t.* diffeomorphic transformations. One of the potential trade-offs or limitations of the proposed architecture is that the usage of convolutions with kernel size 1 in the decoders may provoke a trade-off between the reconstruction fidelity and the equivariance constraint. However, by sacrificing a small degree of reconstruction fidelity, we gain a significant improvement in the identifiability and accuracy of the geometric transformations. This is an example of how imposing a strong, domain-appropriate inductive bias can guide a learning system to a more robust and accurate solution for its primary task. Future studies could be addressed to represent the generative process of multi-modal images from the underlying common anatomy in a more compact and unified manner using techniques from, for example, normalizing flows [90] that construct an invertible mapping between the observation and the latent variable.

4.7 Chapter Takeaway

This chapter addressed unsupervised multimodal groupwise registration under a substantially stronger heterogeneity regime than that of Chapter 3. In addition to contrast- or modality-dependent appearance variability, the inputs are geometrically misaligned, so correspondence must be inferred rather than assumed. The central difficulty is that cross-image similarity cannot be reliably measured in observation space when modalities differ, yet registration fundamentally requires a correspondence-consistent notion of similarity.

The main methodological takeaway is that robust groupwise registration benefits from an explicit separation of latent roles that would otherwise be confounded. By formulating registration as Bayesian inference in a hierarchical latent variable model, the proposed approach disentangles common anatomy from image-specific geometry while treating appearance as an observation-dependent realization. This yields an intrinsic representation in which multimodal similarity becomes meaningful and stable, enabling alignment to be driven by task-relevant anatomical content rather than by modality-specific intensity statistics or heuristic similarity metrics. In the terminology of this thesis, the chapter instantiates identifiable invariant preservation in a regime where invariants must be preserved *across both observation mechanisms and geometric configurations*.

This project also clarifies an important escalation in representational requirements. In Chapter 3, appearance heterogeneity perturbs otherwise aligned inputs, and stability can be improved by constraining representational complexity. Here, heterogeneity additionally includes registration-compatible geometric variation, requiring a generative organization in which anatomy provides the invariant reference and geometry accounts for deformable misalignment. The next chapter removes the availability of reliable cross-image structural correspondence altogether, demanding mechanisms that can preserve invariants without relying on direct alignment between samples.

Chapter 5

UNIFIED DOMAIN ADAPTIVE MEDICAL IMAGE SEGMENTATION

This chapter presents the third project of the thesis: unsupervised domain adaptation (UDA) for medical image segmentation under the most challenging heterogeneity regime considered in this dissertation. In contrast to the multimodal registration setting of Chapter 4, where images can be organized into groups that share an underlying anatomical reference, domain adaptation operates in a fundamentally *correspondence-free* setting. In UDA we are given a labeled *source domain* and an unlabeled *target domain* that share the same downstream task but differ in acquisition environment and image appearance. Samples from the two domains are unpaired and cannot be aligned directly.

This setting introduces a stronger form of heterogeneity than those considered in the previous projects. In addition to appearance variability arising from differences in imaging contrast, modality, or acquisition protocol, the source and target datasets may also differ in anatomical configuration due to inter-subject variability, field-of-view differences, or scan coverage. Crucially, because images from the two domains are not paired and do not admit registration-based correspondence, there is no shared reference shape at the sample level that can be used to couple the domains. In the taxonomy of Chapter 1, this regime therefore combines appearance heterogeneity with *correspondence-free geometric heterogeneity*, representing the most difficult point along the heterogeneity progression studied in this thesis.

Most existing UDA approaches address this challenge by introducing auxiliary alignment signals in feature space. These methods typically attempt to match aggregated source and target representations through explicit feature alignment, or, when source samples are unavailable during adaptation, rely on heuristic signals such as pseudo-labels, entropy min-

imization, or knowledge distillation. While effective in certain scenarios, these strategies share a fundamental limitation: they do not explicitly model anatomical structure as a stable source of knowledge. As a result, the learned representations may remain sensitive to domain-specific appearance patterns, making adaptation fragile when the target domain deviates significantly from the source.

From a representation learning perspective, the central question is therefore whether anatomical knowledge can be encoded in a way that remains stable across domains even when individual images differ in both appearance and geometry. To address this problem, we adopt the Bayesian representation learning framework introduced in Chapter 2 and construct a generative representation in which adaptability arises directly from the latent organization of the model.

Specifically, we introduce a domain-agnostic probabilistic manifold that represents the global space of anatomical regularities relevant to the segmentation task. Each observed image is interpreted through two complementary latent components: a *canonical anatomical prototype* retrieved from this manifold, and an *image-specific geometric transformation* that accounts for individual anatomical variation. Within this formulation, segmentation predictions are generated through the interaction of these latent factors rather than through direct reliance on domain-specific appearance statistics.

Because anatomical structure is represented explicitly in the manifold while geometric variation is modeled separately, the resulting representation suppresses domain-dependent appearance fluctuations while preserving task-relevant anatomical invariants. Importantly, this architecture enables adaptation without requiring explicit cross-domain alignment. The same generative framework naturally supports both source-accessible and source-free adaptation scenarios, differing only in what information is available during the adaptation stage.

The remainder of this chapter closely follows the associated publication and submission:

- Xin Wang et al. Remind: Remembering anatomical variations for interpretable domain adaptive medical image segmentation. In Ipek Oguz, Shaoting Zhang, and Dimitris N. Metaxas, editors, Information Processing in Medical Imaging, pages 327–341, Cham,

2026. Springer Nature Switzerland. ISBN 978-3-031-96628-6 [170].

- Xin Wang et al. Unified and semantically grounded domain adaptation for medical image segmentation. IEEE Transactions on Medical Imaging, 2026, doi: 10.1109/TMI.2026.3672802 [169].

5.1 Introduction

Learning-based methods have dominated medical image segmentation by enabling automatic and accurate delineation of anatomical structures, which facilitates a variety of clinical applications [94]. This success, however, relies on large, well-annotated datasets that match the image characteristics of the intended domain. In practice, such data are often difficult to obtain due to variations in hardware, imaging protocols, patient populations, and disease manifestations. These domain shifts can severely degrade model performance, which has motivated unsupervised domain adaptation (UDA) to transfer knowledge from a labeled source domain to an unlabeled target domain with differing image characteristics, thereby alleviating the need for costly manual annotation of target data [59].

Existing UDA methods primarily assume that source-domain data remain accessible during adaptation. In this *source-accessible* setting, models can jointly utilize source and target data to learn domain-invariant representations. Numerous strategies have been explored, such as adversarial training, semi-supervised learning, and statistical alignment [161, 187, 172]. These methods often operate in high-dimensional feature spaces where alignment is computationally expensive and difficult to interpret. On the other hand, retaining source data during adaptation is not feasible in many real-world applications, due to privacy regulations, institutional policies, and data sharing restrictions. This gives rise to the *source-free* setting, where only a pre-trained source model (but no source data) is available for target-domain adaptation, presenting greater challenges. Prior works for source-free adaptation frequently rely on complex self-training or entropy minimization techniques that are prone to instability, overfitting, and loss of anatomical fidelity [33, 163, 16]. Crucially, existing

source-accessible and source-free methods share common limitations: they lack explicit and explainable mechanisms to ensure that adapted features capture valid anatomical structures, which often results in implausible or fragmented segmentations.

This absence of explicit anatomical reasoning represents a fundamental inconsistency: surprisingly, despite the trivial difference between the two settings, *i.e.*, source-free simply requires completing source-domain learning before target-domain adaptation, existing literature has produced markedly different methodological designs. While source-accessible methods typically build upon domain alignment, source-free methods introduce entirely distinct pipelines based on self-supervision, pseudo-labeling, or distillation. This divergence in methodological design appears disproportionate to the underlying problem difference. From a human perspective, adapting to a new domain, whether or not previous learning examples remain accessible, relies on the same conceptual understanding of anatomy [54]. Therefore, we argue that the separation of current methods for the two settings reflects their inherent limitations, highlighting the need for a unified, interpretable framework that generalizes naturally across both scenarios.

In this work, we set out to close this gap. Motivated by how humans adapt to unfamiliar imaging conditions, we propose a unified and semantically grounded Bayesian framework that applies seamlessly to both source-accessible and source-free adaptation. Humans typically form a conceptual understanding of anatomy by memorizing typical shape patterns from labeled examples. When confronted with an unseen image, intuitively they 1) recall a representative shape and 2) deform it moderately to account for individual-specific structural variations [124]. To emulate this process, we construct a low-dimensional, domain-agnostic latent probabilistic manifold, which encodes the full spectrum of representative structural patterns as weighted compositions of a few prototypical anatomical representations. Fine-grained geometric variations are then captured through additional spatial deformations. This anatomy-aware manifold, shared across different domains and individual images, enables a clear disentanglement between domain-invariant canonical shape templates and individual-specific geometric details, which ensures structurally consistent and anatomically plausi-

ble predictions. This design confers two key advantages. First, it enables adaptability to emerge naturally from the model architecture itself, without requiring explicit cross-domain alignment objectives. Second, the manifold serves as a compact memory that encapsulates anatomical priors, capable of adapting to target images whether or not source data are persistently available.

Our contributions can be summarized as follows:

1. We propose a unified framework that seamlessly supports both source-accessible and source-free adaptation, achieving source-free performance closely matching its source-accessible counterpart.
2. We introduce semantically grounded anatomical modeling, which emulates human visual understanding by explicitly disentangling canonical anatomy from individual geometry. This formulation leads to structurally consistent, robust, and interpretable predictions.
3. Our method's adaptability emerges naturally as an intrinsic property of the framework design. To the best of our knowledge, this is the first work to realize adaptation without explicit cross-domain alignment strategies.

5.2 Related Work

5.2.1 Unsupervised Domain Adaptation

Previous source-accessible UDA works fundamentally relied on domain alignment through various strategies, such as adversarial networks, semi-supervised learning and statistical divergence. For example, ADVENT [161] and DARUNet [179] aligned the output or feature space by discriminators. MAPSeg [187] leveraged masked auto-encoding to improve feature integrity. Generative approaches using variational inference, including VarDA [172] and VAMCEI [40], derived statistical divergences between feature distributions. While these approaches captured domain-invariance, they often suffered from semantic ambiguity due to the absence of anatomical constraints, and required costly high-dimensional calculations.

Recent works have also investigated the more challenging source-free setting, relying on conventional strategies such as pseudo-labeling, entropy minimization, and distillation. For example, while Tent [163] and AdaMI [16] minimized prediction entropy, the latter also enforced class ratio priors. UPL-SFDA [173] selected pseudo-labels via multiple prediction headers and refined results via dual-pass supervision. ProtoContra [182] aligned target features to source model parameters and applied contrastive learning on unreliable samples. Despite their promising mitigation of prediction errors, existing source-free methods universally relied on refinement strategies that are vulnerable to noise in the outputs of source-trained models.

Few prior works have explored unsupervised adaptation from an anatomy-aware perspective. In contrast, our framework encapsulates structural priors from source data via a domain-agnostic manifold shared across all images. This formulation offers an interpretable means of preserving anatomical coherence, enabling adaptation to naturally emerge in both source-accessible and source-free settings.

5.2.2 Variational Autoencoders in Medical Imaging

Variational autoencoders (VAEs) [88] provided a principled generative framework for effectively learning latent representations. For example, in image segmentation and synthesis, conditional VAEs were developed to disentangle anatomy from appearance, allowing representations to generalize across modalities [30]. Hierarchical VAEs were able to improve feature expressiveness through multiscale latent feature factorization in a theoretically-grounded probabilistic manner [155].

In image registration, VAE-based models could effectively capture inter-modality anatomical consistency and learn diffeomorphic deformations. For example, BInGo explicitly disentangled anatomy and geometry for scalable and interpretable groupwise image registration [167, 101] through unsupervised self-reconstruction. While powerful, learning domain-invariant structural features in this method relied on paired multi-modal images with a common anatomy, which is infeasible in domain adaptation since input images come from

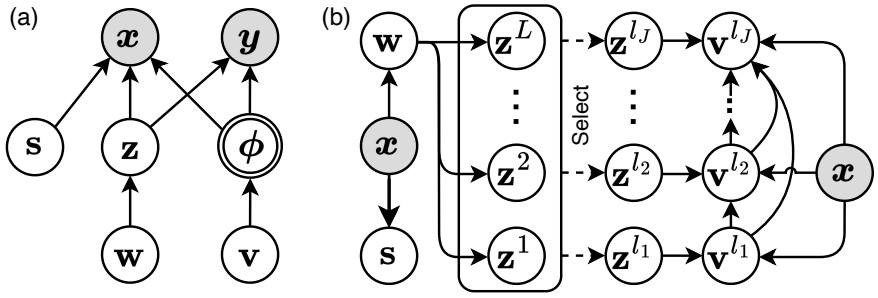


Figure 5.1: Graphical models of the proposed framework. (a) Generative model. (b) Inference model with hierarchical decomposition. Deterministic variables are in double circles, and observed variables are shaded. Dashed arrows denote selecting the subset $\{\mathbf{z}^{l_j}\}_{l_j \in \Lambda = \{l_1, \ldots, l_J\}}$.

different subjects and spatial locations.

The proposed framework builds upon hierarchical and disentangled VAE formulations and introduces two key innovations: 1) a global latent space that is shared by all images and can generalize naturally across different domains and settings, 2) improving anatomical plausibility of segmentation predictions by encoding images as weighted combinations of anatomical representatives. This structured design greatly enhances interpretability and adaptability, while maintaining the flexibility and expressiveness of variational models.

5.3 Methodology

Let $(\mathcal{X}, \mathcal{Y})$ denote the observation space, where $\mathcal{X} \subseteq \mathbb{R}^D$ is the image space, with D the number of pixels, and $\mathcal{Y} \subseteq \{0, 1, \ldots, K\}^D$ is the segmentation label space, with K the number of foreground classes. The available data for learning include a labeled source dataset $\mathcal{D}^s = \{(\mathbf{x}_s^i, \mathbf{y}_s^i)\}_{i=1}^{N_s}$ sampled from a joint distribution $\mathbb{P}_s$ over $\mathcal{X} \times \mathcal{Y}$, and an unlabeled target dataset $\mathcal{D}^t = \{\mathbf{x}_t^i\}_{i=1}^{N_t}$ sampled from a marginal distribution $\mathbb{P}_t$ over $\mathcal{X}$. We begin by introducing our framework from both theoretical and network implementation perspectives, and then describe how it is applied to both source-accessible and source-free settings.

5.3.1 Disentangled Probabilistic Modeling

The generative structure of our model is presented in Fig. 5.1(a). Unlike prior methods that encoded all anatomical information into a single latent variable, we draw inspiration from how humans approach segmentation: intuitively, one first recalls a representative anatomical shape from prior knowledge, and then adapts it to the image through moderate spatial warping to account for individual-specific geometric details [124]. Motivated by this perspective, we propose to explicitly disentangle the structural content of an image $\mathbf{x}$ into two distinct components: a canonical anatomical template $\mathbf{z}$ and a spatial deformation $\boldsymbol{\phi}$ that is parameterized by a stationary velocity field (SVF) $\mathbf{v}$ through $\boldsymbol{\phi} = \exp(\mathbf{v})$ [9], such that $\mathbf{x} \circ \boldsymbol{\phi}$ is spatially aligned to $\mathbf{z}$. This formulation leads to interpretable and geometry-aware representations of anatomical variations. In addition, we introduce a variable $\mathbf{s}$ to encode image style.

While the template $\mathbf{z}$ encodes canonical anatomy, a single fixed $\mathbf{z}$ lacks expressiveness to account for topological diversity across images; on the other hand, extracting $\mathbf{z}$ from $\mathbf{x}$ without constraints may cause $\mathbf{z}$ to capture all anatomical variability and the deformation $\boldsymbol{\phi}$ to degenerate. To ensure both flexibility and disentanglement, we propose to condition $\mathbf{z}$ on a low-dimensional vector $\mathbf{w} \in \mathbb{R}^M$. This induces a structured latent manifold that supports interpretable and controllable shape retrieval, as detailed in Secs. 5.3.2 and 5.4.4.

We leverage the variational Bayesian framework [88] to effectively learn to infer the latent variables. Since $\mathbf{x}$ contains all information about the variables, we assume $\mathbf{s}$ is conditionally independent of all structure-related variables given $\mathbf{x}$. Thus, given a labeled source sample $(\mathbf{x}, \mathbf{y})$, the joint and variational posterior distributions can respectively be written as

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}, \mathbf{v}, \mathbf{s}) &= p(\mathbf{w})p(\mathbf{s})p(\mathbf{v})p(\mathbf{z}|\mathbf{w})p(\mathbf{x}|\mathbf{z}, \mathbf{v}, \mathbf{s})p(\mathbf{y}|\mathbf{z}, \mathbf{v}), \\ q(\mathbf{w}, \mathbf{z}, \mathbf{v}, \mathbf{s}|\mathbf{x}, \mathbf{y}) &= q(\mathbf{s}|\mathbf{x})q(\mathbf{w}|\mathbf{x})q(\mathbf{z}|\mathbf{w})q(\mathbf{v}|\mathbf{x}, \mathbf{z}). \end{aligned} \tag{5.1}$$

Consequently, the evidence lower bound (ELBO) of the log-likelihood $\log p(\mathbf{x}, \mathbf{y})$ is derived

as

$$\begin{aligned}
\text{ELBO} \coloneqq & \mathbb{E}_{q(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{s}|\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{s})}{q(\mathbf{w},\mathbf{z},\mathbf{v},\mathbf{s}|\mathbf{x},\mathbf{y})}\right] \\
= & \mathbb{E}_{q(\mathbf{s}|\mathbf{x})q(\mathbf{w}|\mathbf{x})q(\mathbf{z}|\mathbf{w})q(\mathbf{v}|\mathbf{x},\mathbf{z})} \log p(\mathbf{x}|\mathbf{z},\mathbf{v},\mathbf{s}) && \cdots \mathcal{L}_{\text{recon}} \\
& + \mathbb{E}_{q(\mathbf{w}|\mathbf{x})q(\mathbf{z}|\mathbf{w})q(\mathbf{v}|\mathbf{x},\mathbf{z})} \log p(\mathbf{y}|\mathbf{z},\mathbf{v}) && \cdots \mathcal{L}_{\text{seg}} \\
& - \mathbb{E}_{q(\mathbf{w}|\mathbf{x})} D_{\text{KL}}\left[q(\mathbf{z}|\mathbf{w}) \parallel p(\mathbf{z}|\mathbf{w})\right] && \cdots \mathcal{L}_{\text{tem}} \\
& - \mathbb{E}_{q(\mathbf{w}|\mathbf{x})q(\mathbf{z}|\mathbf{w})} D_{\text{KL}}\left[q(\mathbf{v}|\mathbf{x},\mathbf{z}) \parallel p(\mathbf{v})\right] && \cdots \mathcal{L}_{\text{vel}} \\
& - D_{\text{KL}}\left[q(\mathbf{s}|\mathbf{x}) \parallel p(\mathbf{s})\right] - D_{\text{KL}}\left[q(\mathbf{w}|\mathbf{x}) \parallel p(\mathbf{w})\right],
\end{aligned} \tag{5.2}$$

where D_{KL} is the Kullback-Leibler (KL) divergence. A similar decomposition applies to unlabeled target samples, with the omission of $\mathcal{L}_{\text{seg}}$. The reconstruction probability $p(\mathbf{x}|\mathbf{z},\mathbf{v},\mathbf{s})$ is modeled using pixel-wise Laplacian distributions, with their parameters predicted by a neural network. The segmentation term $\mathcal{L}_{\text{seg}}$ is computed as the negative sum of the cross-entropy and Dice losses, following prior works [172, 40]. Moreover, we assume deterministic posteriors for both $\mathbf{w}$ and $\mathbf{s}$, *i.e.*, $q(\mathbf{w}|\mathbf{x}) \coloneqq \delta(\mathbf{w} - \widetilde{\mathbf{w}}(\mathbf{x}))$, where δ is the Dirac delta and $\widetilde{\mathbf{w}}$ is predicted from $\mathbf{x}$. As a result, inference yields $\mathbf{w} = \widetilde{\mathbf{w}}$. The same formulation applies to $\mathbf{s}$. Therefore, the last two KL terms in the ELBO can be omitted during training.

We further factorize $\mathbf{z}, \mathbf{v}$ as $\mathbf{z} = (\mathbf{z}^l)_{l=1}^L$ and $\mathbf{v} = (\mathbf{v}^{l_j})_{l_j \in \Lambda}$ [155, 167, 101, 168], where $\Lambda = \{l_1, \ldots, l_J\}$ is a subsequence of $\{1, \ldots, L\}$, and a smaller l indicates a coarser resolution. This hierarchical decomposition facilitates effective learning by capturing complex anatomical variability through progressively refined components. Thus, the spatial transformation is computed as $\boldsymbol{\phi} = \boldsymbol{\phi}^{l_1} \circ \cdots \circ \boldsymbol{\phi}^{l_J}$, with $\boldsymbol{\phi}^{l_j}$ parameterized by the velocity $\mathbf{v}^{l_j}$. We assume that different scales of $\mathbf{z}^l$ are independently conditioned on $\mathbf{w}$, and that $\mathbf{v}^{l_j}$ can be inferred given $\mathbf{x}$ and the template $\mathbf{z}^{l_j}$ at the same scale. In addition, we set $p(\mathbf{v}^{l_j}|\mathbf{v}^{<l_j}) = p(\mathbf{v}^{l_j})$ by design,

where $< l_j$ denotes scales below l_j. Thus, $\mathcal{L}_{\text{tem}}$ and $\mathcal{L}_{\text{vel}}$ are simplified as

$$\begin{aligned}\mathcal{L}_{\text{tem}} &= \mathbb{E}_{q(\mathbf{w}|\mathbf{x})} \sum_{l=1}^{L} D_{\text{KL}} \left[q(\mathbf{z}^l|\mathbf{w}) \parallel p(\mathbf{z}^l|\mathbf{w}) \right], \\ \mathcal{L}_{\text{vel}} &= \mathbb{E}_{q(\mathbf{w}|\mathbf{x})q(\mathbf{z}|\mathbf{w})} \sum_{l_j \in \Lambda} \mathbb{E}_{q(\mathbf{v}^{<l_j}|\mathbf{x},\mathbf{z}^{<l_j})} \left(D_{\text{KL}} \left[q(\mathbf{v}^{l_j}|\mathbf{x}, \mathbf{z}^{l_j}, \mathbf{v}^{<l_j}) \parallel p(\mathbf{v}^{l_j}) \right] \right).\end{aligned} \tag{5.3}$$

where $q(\mathbf{v}^{<l_1}|\mathbf{x}, \mathbf{z}^{<l_1}) \coloneqq 1$. The inference structure of our model is thereby presented in Fig. 5.1(b). While each scale of $\mathbf{z}^l$ is computed independently, the inference of $\mathbf{v}^{l_j}$ proceeds in a coarse-to-fine manner: each scale depends on the template $\mathbf{z}^{l_j}$ and velocities $\mathbf{v}^{<l_j}$ inferred at coarser resolutions. This enables incrementally refining spatial transformations across scales. The KL terms in $\mathcal{L}_{\text{vel}}$ are computed via the probabilistic SVF formulation [43] to regularize $\boldsymbol{\phi}^{l_j}$ to be diffeomorphic.

5.3.2 Semantically Grounded Encoding with Shared Bases

Prior UDA works typically extract structural encodings directly from the input image. This conventional strategy often yields opaque and entangled representations, making interpretation and cross-domain adaptation difficult. In sharp contrast, we propose a structured extraction of the latent template $\mathbf{z}$, where its formation is not freely learned but explicitly modulated by the vector $\mathbf{w}$. To instill semantic organization and interpretability in the latent space, we introduce a small set of *learnable* basis distributions $\{q_m(\mathbf{z}^l)\}_{m=1}^{M}$ for each scale l, *shared across all images*, where M is the length of $\mathbf{w}$. The posterior distribution of $\mathbf{z}^l$ is then modeled as a log-linear mixture [97] of the bases weighted by $\mathbf{w}$, *i.e.*,

$$q(\mathbf{z}^l|\mathbf{w}) \propto \prod_{m=1}^{M} \left[q_m(\mathbf{z}^l) \right]^{w_m}, \tag{5.4}$$

where we constrain $\mathbf{w}$ on the probability simplex $\Delta \coloneqq \{\mathbf{w} \in \mathbb{R}^M | \mathbf{w} \succeq \mathbf{0}, \mathbf{1}^\top \mathbf{w} = 1\}$. This formulation restricts $q(\mathbf{z}^l|\mathbf{w})$ to lie within the convex geometry spanned by the bases, introducing a strong inductive bias with several advantages:

1. **Domain-agnostic regularization**: The domain-agnostic bases $q_m(\mathbf{z}^l)$ capture global anatomical regularities, enhancing robustness and generalizability.
2. **Expressive shape composition**: The weight $\mathbf{w}$ blends prototypical structures encoded by bases, allowing the template $\mathbf{z}^l$ to express morphologically rich shapes via compositions of structural primitives.
3. **Interpretable memory-like manifold**: The latent space is organized into a semantically structured manifold through the simplex-constrained $\mathbf{w}$, enabling interpretable traversal and feature extraction. This mechanism mimics how humans retrieve learned anatomical patterns from memory [20].

Similar to the posterior, we define the prior as a uniform mixture over the bases, *i.e.*, $p(\mathbf{z}^l|\mathbf{w}) \coloneqq p(\mathbf{z}^l) \propto \prod_{m=1}^{M} \left[q_m(\mathbf{z}^l)\right]^{1/M}$, without dependence on $\mathbf{w}$. To allow for analytical tractability, we model each basis as a multivariate Gaussian, *i.e.*, $q_m(\mathbf{z}^l) = \mathcal{N}(\boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)$, with diagonal covariance. Thus, we have $q(\mathbf{z}^l|\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}^l, \boldsymbol{\Sigma}^l)$, where

$$\boldsymbol{\Sigma}^l = \left[\sum_{m=1}^{M} w_m \left(\boldsymbol{\Sigma}_m^l\right)^{-1}\right]^{-1}, \quad \boldsymbol{\mu}^l = \boldsymbol{\Sigma}^l \sum_{m=1}^{M} \left[w_m \left(\boldsymbol{\Sigma}_m^l\right)^{-1} \boldsymbol{\mu}_m^l\right]. \tag{5.5}$$

The same form applies to $p(\mathbf{z}^l|\mathbf{w})$ with $\mathbf{w} = \frac{1}{M}\mathbf{1}$.

Since $\mathcal{L}_{\text{tem}}$ requires computing KL divergences for each image $\mathbf{x}$, we propose to replace it by the average of KLs over basis distributions, *i.e.*,

$$\widetilde{\mathcal{L}}_{\text{tem}} = \sum_{l=1}^{L} \frac{1}{M} \sum_{m=1}^{M} D_{\text{KL}}\left[q_m(\mathbf{z}^l) \parallel p(\mathbf{z}^l)\right]. \tag{5.6}$$

This reduces the number of KL terms per batch from LB to LM, where the batch size B is typically much larger than M, substantially reducing computational cost. Despite the simplification, $\widetilde{\mathcal{L}}_{\text{tem}}$ serves as an effective surrogate for $\mathcal{L}_{\text{tem}}$: minimizing the former encourages each basis to stay close to the prior $p(\mathbf{z}^l|\mathbf{w})$, which regularizes the mixture $q(\mathbf{z}^l|\mathbf{w})$ to also remain close to the prior, thus effectively minimizing the original term $\mathcal{L}_{\text{tem}}$. The regularization $\widetilde{\mathcal{L}}_{\text{tem}}$ can be interpreted as encouraging the bases to remain close to their

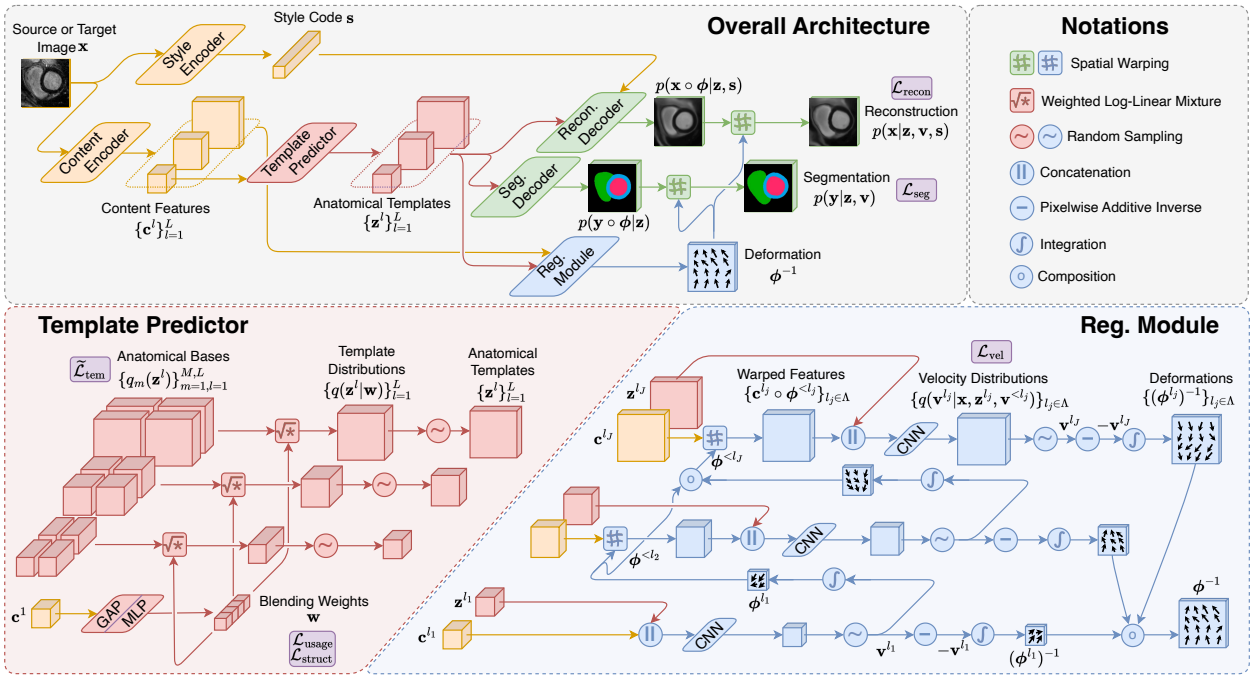


Figure 5.2: Network architecture for the proposed framework. Without loss of generality, the illustration utilizes $L = 3$, $\Lambda = \{1, 2, 3\}$, and $M = 4$. The Gaussian (*resp.* Laplacian) distributions are represented by feature maps whose two halves of channels correspond to the mean and variance (*resp.* scale), with the latter obtained via a Softplus function. Random samplings are performed during training, and replaced by taking the mathematical expectations during evaluation. The purple boxes correspond to the calculation of loss terms using related outputs.

average, and thus to each other. This constraint reduces the sensitivity of $q(\mathbf{z}^l|\mathbf{w})$ to small perturbations in $\mathbf{w}$, facilitating more stable and reliable control of the latent template $\mathbf{z}^l$ via $\mathbf{w}$.

5.3.3 Manifold Structuring for Emergent Adaptation

A central premise of our framework is that canonical anatomical representations of all images are embedded into a shared, domain-agnostic manifold. In principle, this design allows domain adaptation to emerge naturally: the model only needs to interpret each image through this unified space, regardless of its domain origin. To fully realize this potential,

we introduce two principled constraints that guide the internal organization of the manifold. These constraints ensure that the manifold is richly populated with structurally diverse and well-supervised representations, allowing target-domain inputs to be projected into regions with coherent anatomical semantics.

First, we promote balanced activation of anatomical bases within each source or target batch $\{\mathbf{x}_i\}_{i=1}^{B}$ through

$$\mathcal{L}_{\text{usage}} \coloneqq \sum_{m=1}^{M} \max\left(0, \tau - \frac{1}{B}\sum_{i=1}^{B} w_m\left(\mathbf{x}_i\right)\right), \tag{5.7}$$

where $\tau = 0.05$ defines a minimum usage threshold for each basis distribution. This encourages the model to utilize the full representational capacity of the manifold, which not only prevents basis underuse but also enhances the expressiveness and flexibility of the anatomical encoding.

Second, to ensure that the diversity in composition weights $\mathbf{w}$ reflect meaningful structural differences, we introduce a semantic dispersion constraint based on each labeled source batch $\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{B_s}$:

$$\mathcal{L}_{\text{struct}} \coloneqq \sum_{\substack{i,j=1 \\ i<j}}^{B_s} \left[\text{Sim}\left(\mathbf{y}_i^s \circ \boldsymbol{\phi}_i^s, \mathbf{y}_j^s \circ \boldsymbol{\phi}_j^s\right) - C\left(\mathbf{w}_i^s, \mathbf{w}_j^s\right)\right]^2 \tag{5.8}$$

where Sim denotes the Dice similarity between ground-truth segmentations warped by the inferred spatial deformations, and $C(\cdot,\cdot)$ is a similarity between their corresponding composition weights $\mathbf{w}$. To respect the non-Euclidean geometry of the probability simplex Δ, we impose the Fisher-Rao metric D_{FR} on the simplex to measure the distances among $\mathbf{w}$, *i.e.*,

$$D_{\text{FR}}\left[\mathbf{w} \parallel \mathbf{w}'\right] \coloneqq 2\arccos\left(\sum_{m=1}^{M}\sqrt{w_m w'_m}\right), \ \forall \mathbf{w}, \mathbf{w}' \in \Delta, \tag{5.9}$$

which corresponds to the geodesic distance on a Riemannian statistical manifold [112]. There-

fore, $C(\cdot,\cdot)$ is calculated by transforming D_{FR} into a bounded similarity score in [0,1], *i.e.*, $C(\mathbf{w},\mathbf{w}') \coloneqq 1 - D_{\mathrm{FR}}\left[\mathbf{w} \parallel \mathbf{w}'\right]/\pi$. The loss $\mathcal{L}_{\mathrm{struct}}$ explicitly associates latent diversity with anatomical diversity, thereby encouraging the manifold to be semantically organized and broadly covered by source-domain knowledge.

Together, these two constraints shape the manifold into an expressive, semantically coherent and fully utilized space, allowing adaptation to arise naturally through latent encoding. Importantly, these regularizers do not introduce additional degrees of freedom. Instead, they restrict the solution space of an otherwise underconstrained objective, thereby reducing the risk of degenerate solutions rather than increasing overfitting. For this reason, such regularization is particularly important in limited-data regimes, where unsupervised adaptation is most prone to instability.

5.3.4 Network Architecture

We design a specialized network architecture to support disentanglement of anatomy and geometry within the established probabilistic framework, as shown in Fig. 5.2.

Inference of Anatomical Template

A content encoder first extracts multiscale content features $\{\mathbf{c}^l\}_{l=1}^{L}$ from the input image. The coarsest feature map $\mathbf{c}^1$ is aggregated via global average pooling (GAP) and passed through a multilayer perceptron (MLP) followed by a Softmax to infer the composition weight $\mathbf{w}$. The anatomical template distributions $q(\mathbf{z}^l|\mathbf{w})$ are then computed according to Eq. (5.5), leveraging a set of learnable feature maps that parameterize the basis distributions $q_m(\mathbf{z}^l)$. The templates $\mathbf{z}^l$ are then inferred based on $q(\mathbf{z}^l|\mathbf{w})$.

Inference of Velocity Fields

Velocities and deformations are inferred via a dedicated registration module, conditioned on the content features $\mathbf{c}^{l_j}$ and anatomical templates $\mathbf{z}^{l_j}$. As indicated by the hierarchical factorization in Eq. (5.3), the inference of each velocity $\mathbf{v}^{l_j}$ is conditioned on coarser-scale velocities $\mathbf{v}^{<l_j}$. Specifically, we first estimate the distribution $q(\mathbf{v}^{l_1}|\mathbf{x},\mathbf{z}^{l_1})$ using a convolu-

tional neural network (CNN) that takes $\mathbf{c}^{l_1}$ and $\mathbf{z}^{l_1}$ as input, yielding the initial velocity field $\mathbf{v}^{l_1}$ and corresponding deformation $\boldsymbol{\phi}^{l_1}$. For each scale $l_j > l_1$, we warp $\mathbf{c}^{l_j}$ by the composed deformation $\boldsymbol{\phi}^{<l_j} := \boldsymbol{\phi}^{l_1} \circ \cdots \circ \boldsymbol{\phi}^{l_{j-1}}$, resulting in a content representation partially aligned to the anatomical template. The warped feature $\mathbf{c}^{l_j} \circ \boldsymbol{\phi}^{<l_j}$ and $\mathbf{z}^{l_j}$ are then fed into a CNN to estimate $\mathbf{v}^{l_j}$. After inferring all velocities $\{\mathbf{v}^{l_j}\}_{l_j \in \Lambda}$, the overall deformation $\boldsymbol{\phi}$ and its inverse $\boldsymbol{\phi}^{-1}$ are computed deterministically via integration [9] and composition.

Segmentation and Reconstruction

The anatomical template $\mathbf{z} = \{\mathbf{z}^l\}_{l=1}^L$ are fed into a segmentation decoder to generate a categorical probability map $p(\mathbf{y} \circ \boldsymbol{\phi} | \mathbf{z})$. This map is subsequently warped by the inverse transformation $\boldsymbol{\phi}^{-1}$ to yield the final segmentation prediction $p(\mathbf{y}|\mathbf{z}, \mathbf{v})$. For image reconstruction, a style encoder first extracts the style code $\mathbf{s}$ from the input image. Conditioned on both $z = \{\mathbf{z}^l\}_{l=1}^L$ and $\mathbf{s}$, a reconstruction decoder produces a Laplacian distribution $p(\mathbf{x} \circ \boldsymbol{\phi} | \mathbf{z}, \mathbf{s})$, which is then warped by $\boldsymbol{\phi}^{-1}$ to obtain the final reconstruction $p(\mathbf{x}|\mathbf{z}, \mathbf{v}, \mathbf{s})$ of the input image.

We note that although the blending vector $\mathbf{w}$ is low-dimensional, it parameterizes an anatomical manifold spanned by dense, pixel-wise bases, from which full-resolution canonical segmentations are constructed. The deformation $\boldsymbol{\phi}$ then specifies the pixel-to-pixel mapping from canonical space to image space. As a result, pixel-wise spatial detail is carried by dense bases and deformation, rather than by $\mathbf{w}$ itself, and the proposed canonical-deformation factorization does not constitute a spatial-bandwidth bottleneck for pixel-wise segmentation prediction. This representational design is further analyzed in the ablation study in Sec. 5.4.5.

5.3.5 A Unified Paradigm for Source-Accessible and Source-Free Domain Adaptation

We have established a theoretically and semantically grounded UDA framework through latent bases that encode and memorize global anatomical information. We can now describe how our unified paradigm applies to the source-accessible and source-free settings, as a substantial extension of our conference paper [170].

Formally, the ELBO has been derived as

$$\begin{aligned} \text{ELBO}^s(\mathbf{x}, \mathbf{y}) =& \lambda_1 \mathcal{L}_{\text{seg}}(\mathbf{x}, \mathbf{y}) + \lambda_2 \mathcal{L}_{\text{recon}}(\mathbf{x}) - \lambda_3 \mathcal{L}_{\text{vel}}(\mathbf{x}), \\ \text{ELBO}^t(\mathbf{x}) =& \lambda_2 \mathcal{L}_{\text{recon}}(\mathbf{x}) - \lambda_3 \mathcal{L}_{\text{vel}}(\mathbf{x}), \end{aligned} \tag{5.10}$$

for a source or target observation, respectively, where λ's are term weights, and $\widetilde{\mathcal{L}}_{\text{tem}}$ has been separated out due to its independence from the data. We denote source and target batches as $\mathcal{B}^s \coloneqq \{\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^{B_s}$ and $\mathcal{B}^t \coloneqq \{\mathbf{x}_i^t\}_{i=1}^{B_t}$, respectively, and the mean ELBO over a batch $\mathcal{B}$ as $\mathcal{L}_{\text{LB}}(\mathcal{B})$. In addition, we have introduced $\mathcal{L}_{\text{usage}}, \mathcal{L}_{\text{struct}}$ for manifold structuring.

Source-Accessible

This conventional UDA setting involves a single-stage training using the source and target datasets $\mathcal{D}^s$ and $\mathcal{D}^t$ simultaneously. Thus, the overall loss is

$$\mathcal{L} = -\frac{1}{2}\left[\mathcal{L}_{\text{LB}}(\mathcal{B}^s) + \mathcal{L}_{\text{LB}}(\mathcal{B}^t)\right] + \lambda_4 \widetilde{\mathcal{L}}_{\text{tem}} + \lambda_5 \mathcal{L}_{\text{struct}}(\mathcal{B}^s) + \frac{1}{2}\left[\mathcal{L}_{\text{usage}}(\mathcal{B}^s) + \mathcal{L}_{\text{usage}}(\mathcal{B}^t)\right]. \tag{5.11}$$

Source-Free

In this setting, training is divided into two stages: the first has access only to the source data, while the second operates solely on the target domain. To this end, the loss for the first stage is

$$\mathcal{L}_1 = -\mathcal{L}_{\text{LB}}(\mathcal{B}^s) + \lambda_4 \widetilde{\mathcal{L}}_{\text{tem}} + \lambda_5 \mathcal{L}_{\text{struct}}(\mathcal{B}^s) + \mathcal{L}_{\text{usage}}(\mathcal{B}^s). \tag{5.12}$$

After the first-stage training, the model has memorized structural patterns through the latent bases. Therefore in the second stage, we fix the bases $q_m(\mathbf{z}^l)$ and the segmentation decoder, and only optimize other parts of the network through

$$\mathcal{L}_2 = -\mathcal{L}_{\text{LB}}(\mathcal{B}^t) + \mathcal{L}_{\text{usage}}(\mathcal{B}^t). \tag{5.13}$$

Note that $\mathcal{L}_1$ and $\mathcal{L}_2$ together constitute an exact decomposition of loss terms in $\mathcal{L}$. Thus,

our formulation provides a unified and principled paradigm that seamlessly supports both source-accessible and source-free settings.

5.4 Experiments and Results

5.4.1 Datasets

We conducted comprehensive experiments on two public datasets across the source-accessible and source-free settings to evaluate the accuracy, robustness, and interpretability of the proposed framework. The datasets encompass a wide spectrum of imaging characteristics, including various organs, modalities, protocols, pathologies, and populations.

MS-CMRSeg

The MS-CMRSeg 2019 challenge [192] provides cardiac MRI scans in three sequences, bSSFP, LGE and T2, acquired from 45 patients. Ground-truth segmentations are available for the left ventricle (LV), right ventricle (RV), and myocardium (Myo). We designated 35 bSSFP subjects as the source domain and 45 LGE subjects as the target domain, where 5 LGE subjects are used for validation and model selection, and the remaining 40 LGE subjects are used for final testing. During training, images from all 45 LGE subjects are used without accessing any target-domain labels. To simulate an unpaired scenario, the 2D slices were randomly shuffled after the subject-wise split has been fixed. The protocol above is identical to that used in prior studies [172, 40]. All images were resampled to an in-plane resolution of 0.76 mm, cropped to 192×192 pixels to standardize the field of view, and min-max normalized. *The major challenges of this dataset include 1) limited number of training slices, and 2) considerable domain shifts due to complex intensity patterns, imaging noise, artifacts, weak contrasts across substructures, and shape deformation induced by pathology (*e.g.*, scarring).*

Table 5.1: Batch sizes and loss weights for training our model.

Dataset	Setting (Stage)	Batch Size	λ_1–λ_5
MS-CMRSeg	Source Access.	55	1,15,65,0.5,1
	Source Free (1)	100	1,15,65,2,1
	Source Free (2)	70	0,15,65,0,0
AMOS22	Source Access.	45	20,15,25,1e-4,10
	Source Free (1)	80	20,15,25,1e-4,10
	Source Free (2)	60	0,15,25,0,0

AMOS22

The AMOS 2022 challenge [81] provides a multi-center collection of unpaired abdominal CT and MRI scans, encompassing multiple diseases and imaging protocols. In this study, we focused on segmenting four key organs: liver, spleen, left kidney (LK), and right kidney (RK). Following previous works [32], we selected 25 MRI scans as the source domain and 35 CT scans as the target domain. The CT data were further split into 25 scans for training, 5 for validation, and 5 for testing, with all splits being subject-wise disjoint. To ensure consistency across samples, axial slices were extracted from the 3D volumes, resampled to a uniform spacing of 1.5 mm, and cropped to a consistent field of view centered on the organs of interest. Pixel values were clipped to $[0, 250]$ for CT and the 0–99.5 percentile range for MRI, and then min-max normalized. *The major challenges of this dataset include large structural differences across images and domain shifts due to multi-modality, noise and artifacts.*

5.4.2 Experimental Setups

Implementation Details

We set the number of hierarchical levels to $L = 5$ with $\Lambda = \{1, 3, 5\}$, and the number of basis distributions M to be 6 and 10 for the MS-CMRSeg and AMOS22 datasets, respectively. The content encoder used an attention U-Net [115], with the outputs from attention layers

serving as multi-scale content features $\mathbf{c}^l$ of the input image. The style encoder is a Conv-LeakyReLU-AvgPool-Linear sequence, producing a 128-dimensional style code $\mathbf{s}$. While the reconstruction and segmentation decoders do not share parameters, they both adopted the U-Net decoding structure that takes multi-scale feature maps as input. The reconstruction decoder also incorporated adaptive instance normalization [76] to modulate the output using the style code. Each CNN in the registration module comprised four Conv-Norm-LeakyReLU sequences followed by a 1×1 Conv. The model was trained using the AdamW optimizer [98] (learning rate: 10^{-3}, weight decay: 10^{-4}). The batch size and loss weights are presented in Tab. 5.1. Experiments were conducted using PyTorch [118] on an NVIDIA RTX 4090 GPU.

Compared Methods and Evaluation Metrics

To demonstrate the superiority of our unified framework in both source-accessible and source-free adaptation, we compared it against state-of-the-art methods specifically designed for each respective setting. The baselines cover a variety of strategies, such as adversarial learning, variational inference, pseudo labels, image translation, distillation, contrastive learning, etc. Additionally, a source-trained attention U-Net (Att-UNet) was directly tested on the target domain, serving as a baseline without adaptation (w/o Adapt.). All methods were trained using the same data preprocessing pipeline to ensure a fair comparison. Evaluation metrics include Dice Similarity Coefficient (DSC) and Average Symmetric Surface Distance (ASSD).

5.4.3 Comparison with State-of-the-Art Methods

Tabs. 5.2 and 5.3 report quantitative results compared with state-of-the-art methods developed specifically for each respective setting. Notably, our approach adopts a unified architecture under both settings, with the only difference being that, in the source-free case, the loss terms are split into two groups and applied in separate stages. Moreover, our method's adaptability emerges purely from the framework design, while all baseline methods heavily rely on explicit alignment objectives, such as feature matching or pseudo-labeling.

Table 5.2: Comparison on the MS-CMRSeg dataset with state-of-the-art methods. #Adapt denotes the number of adaptation strategies. In each setting, best results are marked in bold, and * indicates $p < 0.05$ (paired t-test) compared with Ours.

Setting	Method	#Adapt	DSC (%) ↑				ASSD (mm) ↓			
			Average	Myo	LV	RV	Average	Myo	LV	RV
w/o Adapt.	Att-UNet [115]	0	54.2±10.2	48.3±10.3	70.7±12.6	43.5±13.2	9.52±2.32	7.16±2.26	8.07±3.43	13.3±3.50
Source Access.	ADVENT [161]	3	69.7±17.1*	58.1±17.2	77.8±17.2	73.3±19.7	3.75±3.38*	4.05±7.12	4.01±3.56	3.19±2.06
	VarDA [172]	1	79.8±9.35*	73.0±8.32	88.1±4.83	78.5±14.9	2.60±1.33*	1.73±0.56	2.55±1.18	3.51±2.24
	DARUNet [179]	7	82.0±6.78*	75.0±9.47	88.4±5.41	82.7±9.17	2.26±1.05*	1.64±0.71	2.15±1.17	2.99±1.97
	MAPSeg [187]	3	64.3±17.2*	51.4±15.0	78.0±17.2	63.6±22.0	5.17±4.76*	3.88±3.93	5.56±6.28	6.08±4.79
	VAMCEI [40]	3	82.5±5.39*	75.8±6.64	88.2±5.27	83.6±8.45	1.93±0.89*	1.54±0.70	2.01±1.03	**2.25±1.50**
Source Free	Tent [163]	1	59.9±15.6*	56.0±12.7	67.2±17.7	56.4±21.3	9.58±3.20*	6.62±2.48	12.7±5.50	9.43±3.86
	FSM [177]	5	67.8±15.0*	61.0±13.8	77.4±14.1	65.2±22.5	4.49±2.12*	3.35±1.87	4.54±2.75	5.57±3.54
	AdaMI [16]	2	71.6±8.06*	68.7±9.90	84.9±6.38	61.0±11.3	5.56±1.79*	2.63±1.20	4.71±2.82	9.33±3.31
	UPL [173]	2	67.8±10.4*	58.4±14.0	82.3±8.68	62.6±17.7	6.13±2.68*	4.39±2.02	4.27±2.69	9.75±6.58
	ProtoContra [182]	3	72.5±12.1*	64.8±11.3	82.5±11.5	70.3±17.6	4.07±1.99*	2.69±1.38	3.70±2.35	5.81±3.76
Source Access.	**Ours**	0	**84.1±5.35**	**78.5±5.16**	**90.0±4.28**	**83.9±8.64**	**1.72±0.80**	**1.27±0.44**	**1.64±0.79**	2.26±1.53
Source Free	**Ours**	0	**83.1±5.55**	**77.0±5.80**	**89.5±4.60**	**82.7±8.56**	**1.88±0.77**	**1.38±0.48**	**1.78±0.87**	**2.49±1.42**

Across both datasets and settings, our method achieves the best performance in average DSC and ASSD and consistently outperforms prior approaches. The superiority is particularly pronounced under the source-free setting, where our method not only outperforms existing approaches by substantial margins across all evaluation metrics, but also narrows the long-standing performance gap with source-accessible models. Particularly on the MS-CMRSeg dataset, our model in the source-free setting even outperforms the best source-accessible baseline. On the AMOS22 dataset, the Att-UNet without adaptation produces a very poor performance, indicating severe domain shift. Many source-free baselines fail because their adaptation strategies heavily rely on the initial predictions of the source-trained models, which are severely degraded. In contrast, the source-free variant of our method maintains a strong performance approaching that of the source-accessible counterpart. We attribute this improvement to our framework's ability to retain a semantically structured shape memory through learnable latent bases, leading to stronger generalizability and adaptability, even without persistent source supervision.

Table 5.3: Comparison on the AMOS22 dataset with state-of-the-art methods. In each setting, best results are marked in bold, and * indicates $p < 0.05$ (paired t-test) compared with Ours.

Setting	Method	DSC (%) ↑					ASSD (mm) ↓				
		Average	Liver	LK	RK	Spleen	Average	Liver	LK	RK	Spleen
w/o Adapt.	Att-UNet	9.38±8.67	29.3±25.3	0.17±0.35	4.85±6.06	3.17±5.91	50.3±11.4	38.2±7.58	49.1±10.5	45.3±16.8	68.7±19.7
Source Access.	ADVENT	65.0±6.00*	71.0±6.61	60.0±11.5	54.2±13.2	74.9±31.2	6.70±1.78*	6.11±3.11	5.69±1.75	9.01±2.18	5.98±5.99
	VarDA	81.4±4.73*	85.2±6.70	79.8±8.86	78.4±6.76	82.0±7.47	4.49±1.35*	5.72±3.43	5.08±1.19	3.15±0.65	4.01±2.87
	DARUNet	85.8±5.20	87.1±5.56	82.1±9.46	82.4±12.8	**91.6±3.85**	4.61±2.28	5.24±4.29	5.89±3.15	3.61±2.16	3.70±2.50
	MAPSeg	88.5±3.02	**92.2±2.45**	84.0±5.62	87.9±4.38	90.0±8.10	4.21±2.34	4.95±6.24	5.15±2.52	3.46±2.61	3.26±3.24
	VAMCEI	87.1±3.15	88.6±5.01	85.7±6.72	84.5±3.72	89.8±3.09	3.53±1.92	**3.64±1.90**	2.33±0.88	5.07±3.72	**3.10±2.34**
Source Free	Tent	25.7±25.0*	17.2±26.8	12.2±24.3	18.3±35.4	54.9±28.9	35.3±12.6*	33.0±16.7	49.2±12.7	44.6±20.5	14.4±5.53
	FSM	2.57±5.09*	0.00±0.00	0.30±0.59	0.71±1.42	9.28±18.4	70.1±21.4*	88.4±23.8	42.7±13.4	78.4±32.3	71.2±29.1
	AdaMI	70.6±2.92*	85.6±4.35	71.5±7.56	37.2±4.37	88.1±6.70	14.4±3.97*	5.42±2.13	18.2±8.69	29.7±4.49	4.11±2.75
	UPL	44.4±18.6*	75.9±7.51	17.5±35.0	12.2±22.1	71.9±36.0	31.2±7.71*	17.4±9.01	40.6±11.5	50.3±15.7	16.4±23.1
	ProtoContra	76.4±8.26*	81.9±21.9	57.1±20.0	78.3±5.37	88.2±4.30	11.0±2.92*	5.20±5.76	18.1±8.26	16.0±5.16	4.64±4.04
Source Access.	**Ours**	**89.7±1.30**	89.4±5.22	**90.5±1.85**	**89.5±1.55**	89.3±3.37	**3.03±1.12**	4.33±2.10	**1.33±0.17**	**1.36±0.27**	5.11±2.74
Source Free	**Ours**	**87.0±3.27**	**87.7±6.39**	**85.9±4.98**	**86.1±5.50**	**88.2±3.96**	**3.28±1.30**	**4.36±2.04**	**2.64±1.06**	**2.64±1.43**	**3.50±2.43**

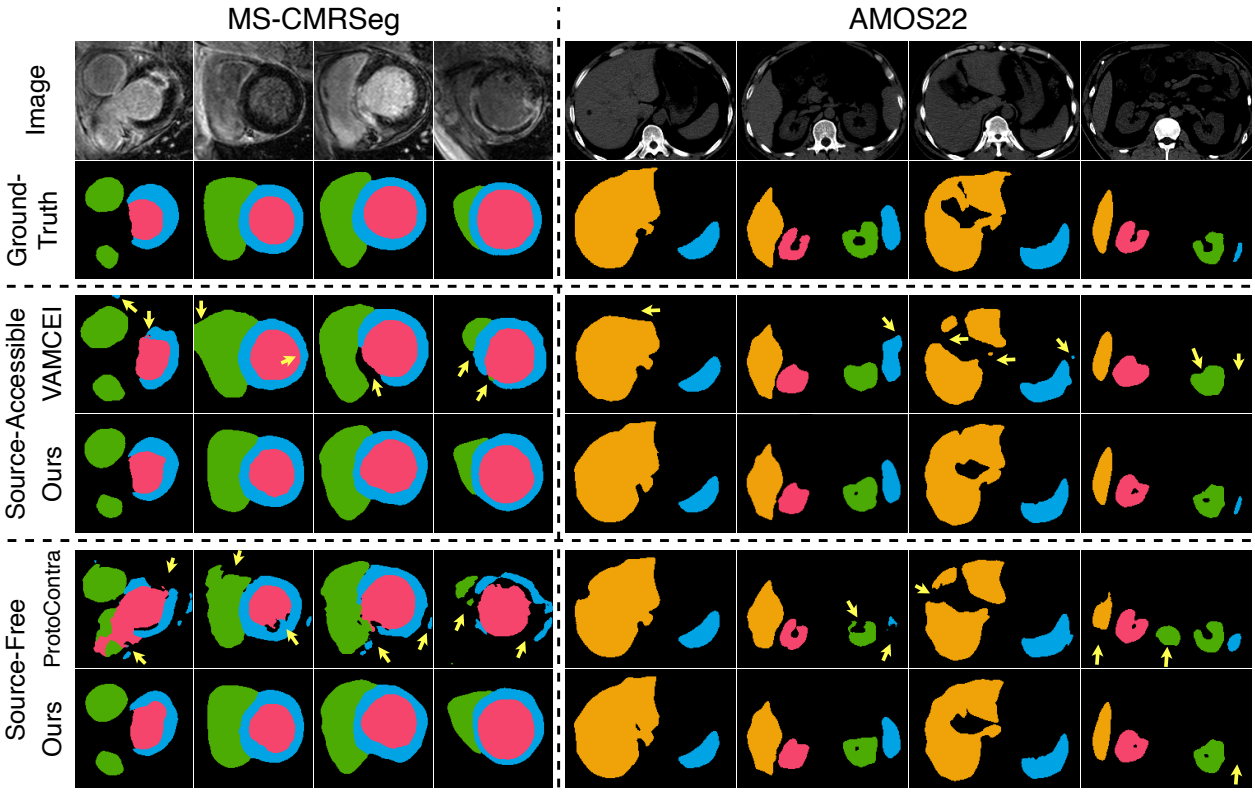


Figure 5.3: Qualitative comparison of our method and the baselines that achieve best overall performance (VAMCEI/ProtoContra for the source-accessible/source-free settings). Yellow arrows indicate inferior results.

Qualitative comparisons in Fig. 5.3 further highlight the strength of our approach. For challenging cases with poor image quality, low contrast, or imaging artifacts, even the best baselines produce physiologically invalid segmentations with fragmented shapes, particularly in the source-free setting. Our method effectively mitigates this issue in both settings, with predictions that preserve coherent topological structure. This robustness stems from the clear disentanglement of canonical anatomy and spatial deformation, which together ensure geometrically smooth and anatomically plausible predictions, even when pixel intensities are unreliable.

In summary, despite using a single model design, our method consistently outperforms all competing approaches, demonstrating superior generalizability, stability, and adaptability

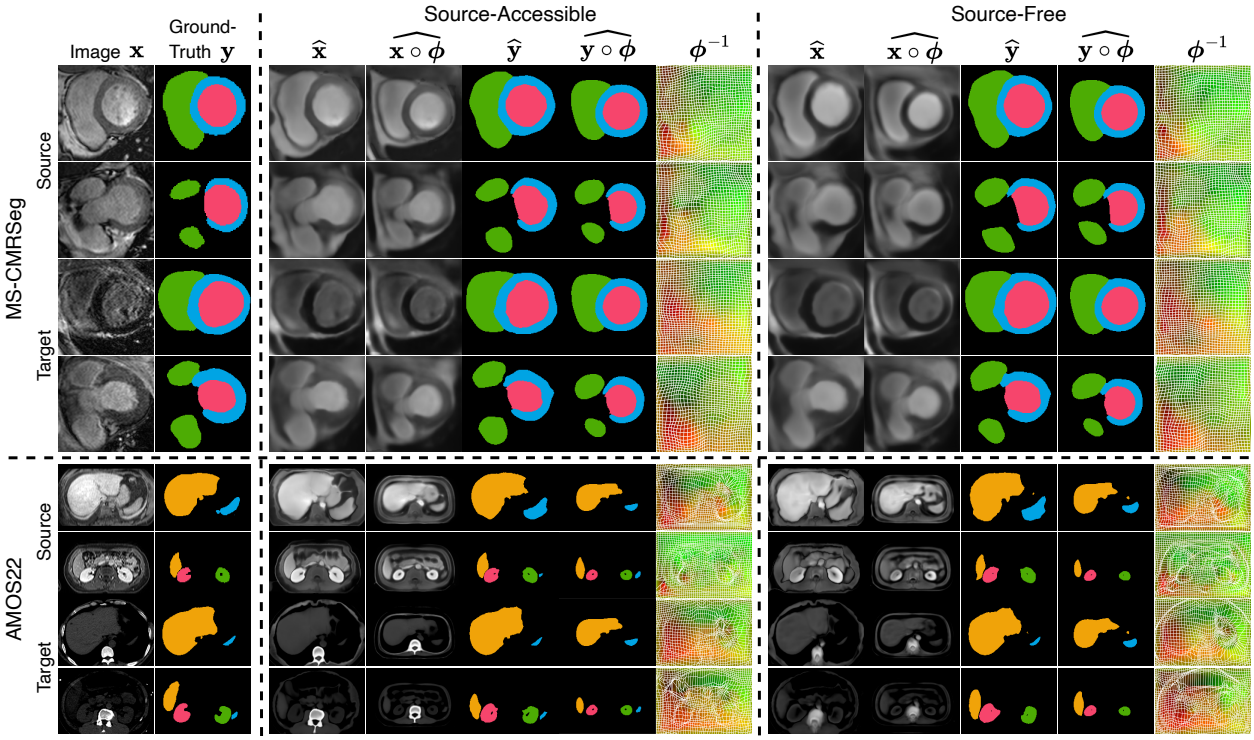


Figure 5.4: Disentanglement of canonical anatomy and geometry by our model. We visualize the templates $\mathbf{z}$ by decoding them into intermediate segmentations $\widehat{\mathbf{y} \circ \boldsymbol{\phi}}$ and reconstructions $\widehat{\mathbf{x} \circ \boldsymbol{\phi}}$ using the segmentation and reconstruction decoders. We also show the corresponding deformations $\boldsymbol{\phi}^{-1}$, as well as the final segmentations $\widehat{\mathbf{y}}$ and reconstructions $\widehat{\mathbf{x}}$ obtained after warping by $\boldsymbol{\phi}^{-1}$.

across multiple datasets and settings.

5.4.4 Interpretability of Latent Manifold

While latent representations often contain entangled structures in conventional UDA studies, our model explicitly disentangles anatomical structure and individual-specific geometry through a low-dimensional and semantically organized latent manifold. In this section, we present a series of visualizations to illustrate the interpretability, generalizability, robustness and domain consistency of the proposed framework.

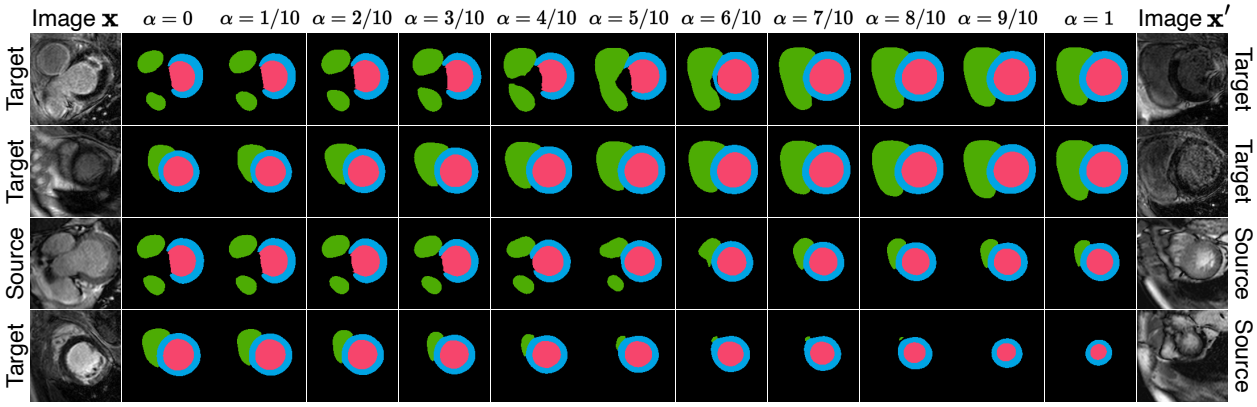


Figure 5.5: Inter-image traversal on the MS-CMRSeg dataset. Each row denotes the decoded segmentations corresponding to an interpolation $\mathcal{T}_\alpha(\mathbf{w}, \mathbf{w}')$ between the composition weights $\mathbf{w}, \mathbf{w}'$ extracted from two images $\mathbf{x}, \mathbf{x}'$. "Target" and "Source" indicates the image domains.

Disentanglement of Canonical Anatomy and Geometry

Fig. 5.4 visualizes the disentanglement of canonical anatomy and individual-specific geometry for representative images. The results indicate that the templates $\mathbf{z}$ accurately capture the underlying anatomy topology and are semantically coherent across domains, illustrating that the learned latent manifold encodes domain-invariant structural priors. The deformations further adapt these templates to capture geometric variations, such as thickening, asymmetry, or tissue movement. This decoupling leads to both strong generalizability and anatomically plausible predictions.

Traversal of the Latent Composition Space

To investigate the semantic structure of the latent manifold, we perform two types of traversals over the composition weights $\mathbf{w}$ via an interpolation operator $\mathcal{T}$, which manipulates the template $\mathbf{z}$. To respect the Riemannian geometry of the simplex Δ endowed by the Fisher-Rao metric, $\mathcal{T}$ proceeds along the geodesics on the positive orthant $\mathbb{S}_+^{M-1}$ of a unit sphere, *i.e.*, $\mathcal{T}_\alpha(\mathbf{w}, \mathbf{w}') := ([\frac{\sin((1-\alpha)\theta)}{\sin\theta}\sqrt{w_i} + \frac{\sin(\alpha\theta)}{\sin\theta}\sqrt{w_i'}]^2)_{i=1}^M$, where $\alpha \in [0,1], \theta = \arccos\left(\sum_{i=1}^M \sqrt{w_i w_i'}\right)$.

Inter-Image Traversal: Given two images with distinct anatomical characteristics, we interpolate between their corresponding composition weights to generate $\mathbf{z} = \{\boldsymbol{\mu}^l\}_{l=1}^L$ through

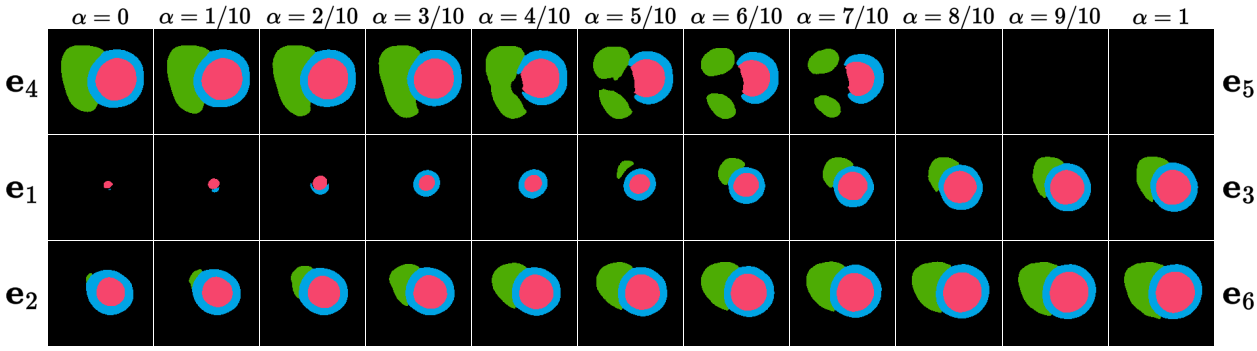


Figure 5.6: Inter-basis traversal on the MS-CMRSeg dataset. Each row denotes the decoded segmentations corresponding to an interpolation $\mathcal{T}_\alpha(\mathbf{e}_i, \mathbf{e}_j)$ between a pair of one-hot composition weights $\mathbf{e}_i, \mathbf{e}_j$. All topological patterns observed in the displayed segmentations are anatomically valid, as some ground-truth labels in the dataset exhibit the same structures.

Eq. (5.5) and decode them into segmentations. As shown in Fig. 5.5, the resulting templates transition smoothly and plausibly, reflecting a continuous shape morphing. This confirms that the latent space captures semantically meaningful variations. Notably, interpolation between source and target-domain samples yields similarly coherent transitions, underscoring domain-invariance of the learned manifold.

Inter-Basis Traversal: We further explore the semantic content encoded in the basis distributions. Specifically, we construct two one-hot vectors $\mathbf{e}_i, \mathbf{e}_j \in \mathbb{R}^M$, *i.e.*, the composition weights that select only the base distributions $\{q_i(\mathbf{z}^l)\}_l$, $\{q_j(\mathbf{z}^l)\}_l$, respectively, and interpolate between them as $\mathcal{T}_\alpha(\mathbf{e}_i, \mathbf{e}_j)$, which are decoded into segmentations (Fig. 5.6), similar to inter-image traversal. The results reveal smooth transitions, indicating that the learned bases form a diverse set of morphological primitives, which can be meaningfully blended via simplex-weighted mixture. In addition to interpretability, this traversal offers a diagnostic tool: if interpolations involving certain bases yield incoherent shapes or lack variability, those bases may be underutilized or redundant. We observe that the loss $\mathcal{L}_{\text{usage}}$ for basis usage plays a key role in maintaining basis diversity and preventing mode collapse.

Together, these two forms of traversal demonstrate that our model learns an anatomical plausible and geometrically smooth latent space. The ability to maintain structural

Figure 5.7: t-SNE results by our method on the MS-CMRSeg dataset in the source-accessible and source-free settings. The projection maps 6D composition weights to a 2D space.

coherence when moving continuously through this space enables both emergent adaptability and interpretation of the latent anatomy, as well as potential applications such as data augmentation, anomaly characterization, and interactive editing.

Cross-Domain Alignment

To visualize how our approach harmonizes the latent spaces of source and target domains, we project the predicted composition vectors $\mathbf{w}$ from the two domains onto a 2D space using t-SNE [158]. Notably, since $\mathbf{w}$ are low-dimensional (6D for the MS-CMRSeg dataset), t-SNE introduces minimal information loss. Fig. 5.7 shows the results in both settings, where the source and target samples are well aligned and form overlapping clusters. This demonstrates that our approach effectively achieves domain alignment. To the best of our knowledge, this is the first work that harmonizes source and target representations without explicit cross-domain alignment objectives in both source-accessible and source-free settings.

5.4.5 Ablation Studies

We conduct comprehensive ablation studies to assess the effectiveness and necessity of each component in our framework.

Table 5.4: Quantitative comparison of segmentation and reconstruction performance between our disentanglment architecture and a direct-decoding baseline on the source domain of the two datasets. ASSD is reported in millimeters (mm), and PSNR is reported in decibels (dB).

Dataset	Model	Segmentation		Reconstruction	
		DSC (%) ↑	ASSD ↓	PSNR ↑	SSIM ↑
MS-CMRSeg	Direct-Dec	92.4±1.08	0.41±0.13	35.4±4.08	0.982±0.007
	Ours-Sup	92.7±0.99	0.38±0.17	20.6±1.13	0.611±0.026
AMOS22	Direct-Dec	94.7±2.13	1.17±0.39	36.8±0.65	0.966±0.019
	Ours-Sup	94.5±1.21	1.44±0.14	21.1±0.85	0.737±0.034

Pixel-Wise Expressiveness for Dense Prediction

This ablation evaluates whether the proposed canonical-deformation factorization limits pixel-wise spatial prediction. To this end, we compare two models under a fully supervised setting: (i) Ours-Supervised, which uses our full architecture to model image structure; and (ii) Direct-Decoding, where segmentation and reconstruction are predicted directly from dense image features using two decoders attached to the same encoder. Both models use the same encoder and decoder backbones and are trained with the same segmentation and reconstruction losses. As shown in Tab. 5.4, our model achieves segmentation accuracy comparable to direct decoding on both datasets, including the more complex multi-organ scenario. This result indicates that relying on a low-dimensional blending vector and deformation-based warping does not impose a spatial-bandwidth bottleneck for pixel-wise segmentation, nor increase learning difficulty in practice. Moreover, maintaining segmentation accuracy while sharing the bases across reconstruction, segmentation, and deformation prediction suggests that the global bases are not overburdened by multi-task usage, but instead capture an anatomical substrate that is jointly useful across tasks. While direct decoding attains higher reconstruction fidelity, this difference is expected due to its unconstrained use of dense features and does not contradict the strong segmentation accuracy achieved by our method.

Table 5.5: Ablation studies on MS-CMRSeg by setting corresponding loss weights to 0 (for w/o $\widetilde{\mathcal{L}}_{\text{tem}}$) or a large value (for w/o $\boldsymbol{\phi}$).

Method	DSC (%) ↑	ASSD (mm) ↓	Epochs to Converge
w/o $\boldsymbol{\phi}$	58.4±7.59	5.88±1.07	24
w/o $\widetilde{\mathcal{L}}_{\text{tem}}$	82.1±6.02	1.96±0.78	2018
Proposed	84.1±5.35	1.72±0.80	2067

Anatomical Disentanglement

This ablation evaluates the effectiveness of disentangling anatomy and geometry by evaluating segmentation performance without spatial transformation. Specifically, by setting the loss weight λ_3 to a large constant, the inferred deformation field collapses to identity. As a result, segmentation is directly decoded from the canonical anatomical template without warping. As shown in Tab. 5.5 row 2, removing deformation leads to a noticeable drop in segmentation accuracy. This confirms that modeling image-specific geometric variations is critical for adapting canonical shape priors to individual anatomy, and the disentanglement of anatomy and geometry leads to greatly improved accuracy, which aligns with human visual recognition.

Effect of Structural and Usage Regularizers

This ablation analyzes the effects of $\mathcal{L}_{\text{struct}}$ and $\mathcal{L}_{\text{usage}}$. As these regularizers are employed in both source-accessible and source-free settings, we conduct the analysis under the source-free setting, where the two-stage training procedure enables a more comprehensive, stage-wise examination of their roles.

Experimental Design In Stage-1 (source-supervised training), we consider four model variants, denoted as S1–S4, corresponding to all combinations of enabling or disabling $\mathcal{L}_{\text{struct}}$ and $\mathcal{L}_{\text{usage}}$. Specifically, S1 uses neither regularizer, S2 uses $\mathcal{L}_{\text{struct}}$ only, S3 uses $\mathcal{L}_{\text{usage}}$ only, and S4 uses both regularizers. This 2×2 design allows us to isolate the individual and

combined effects of the two regularizers on the learned latent representation. In Stage-2 (target-only adaptation), $\mathcal{L}_{\text{struct}}$ is disabled by design due to the absence of target-domain labels, and thus we compare two variants (both initialized from the full Stage-1 configuration S4): T1, where $\mathcal{L}_{\text{usage}}$ is removed, and T2 (ours), where $\mathcal{L}_{\text{usage}}$ is retained. This comparison isolates the marginal effect of $\mathcal{L}_{\text{usage}}$ during unsupervised adaptation under a fixed latent manifold.

Evaluation Metrics To quantitatively characterize how $\mathcal{L}_{\text{struct}}$ and $\mathcal{L}_{\text{usage}}$ affect the learned latent variables, we report both segmentation accuracy and a set of diagnostic metrics:

- **Basis utilization.** Let $\{\mathbf{w}_i \in \Delta^{M-1}\}_{i=1}^{N}$ denote the blending vectors predicted for a dataset of N images. We define the dataset-level basis usage rate as $\overline{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i$, where the m-th element $\overline{w}_m$ represents the average usage of the m-th basis. To measure how evenly the bases are utilized, we compute the usage entropy $H(\overline{\mathbf{w}}) = -\sum_{m=1}^{M}\overline{w}_m \log \overline{w}_m$, and the number of effective bases $N_{\text{eff}} = \exp\big(H(\overline{\mathbf{w}})\big)$, which reaches its maximum value M when all bases are used uniformly.
- **Simplex geometry.** To quantify how broadly the blending vectors occupy the simplex Δ^{M-1}, we define a dispersion metric based on the generalized variance. Since Δ^{M-1} lies in a $(M-1)$-dimensional affine subspace of $\mathbb{R}^M$, we first center the samples by $\widetilde{\mathbf{w}}_i = \mathbf{w}_i - \overline{\mathbf{w}}$ and project them onto the intrinsic subspace $\{\boldsymbol{u} \in \mathbb{R}^M : \mathbf{1}^\top \boldsymbol{u} = 0\}$ using a fixed orthonormal basis $U \in \mathbb{R}^{M\times(M-1)}$. The empirical covariance is computed as $\Sigma = \frac{1}{N-1}\sum_{i=1}^{N}(U^\top\widetilde{\mathbf{w}}_i)(U^\top\widetilde{\mathbf{w}}_i)^\top$, and the dispersion is then measured by $Q = \log\det(\Sigma + \epsilon I)$, with $\epsilon = 10^{-6}$ for numerical stability. Larger values of Q indicate a more dispersed distribution of blending vectors on the simplex.
- **Structural consistency.** We also compute the Spearman rank correlation r_s between blending vector distances $D_{\text{FR}}\left[\mathbf{w}_i \parallel \mathbf{w}_j\right]$ and canonical segmentation dissimilarities $1 - \text{DSC}\big(\mathbf{y}_i \circ \boldsymbol{\phi}_i,\ \mathbf{y}_j \circ \boldsymbol{\phi}_j\big)$ over all sample pairs (i, j). A high value of r_s indicates that variations in $\mathbf{w}$ consistently induce corresponding changes in the segmentation outcomes.

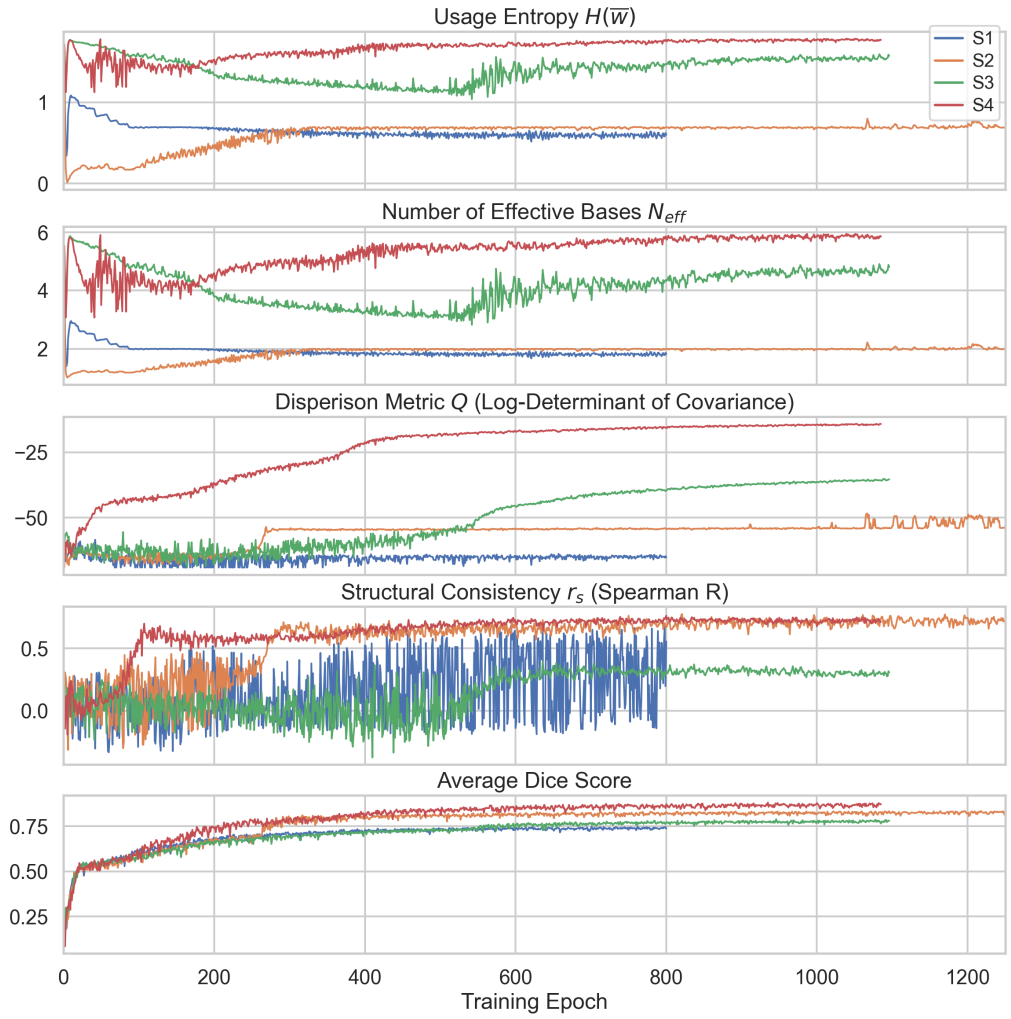


Figure 5.8: Stage-1 training dynamics and latent-space quality on the source domain under different regularizer configurations, including evolution of (a) usage entropy $H(\overline{\mathbf{w}})$, (b) number of effective bases N_{eff}, (c) dispersion metric Q, (d) structural consistency r_s, and (e) segmentation Dice during source-domain supervised training. All curves share the same training timeline.

Results Fig. 5.8 summarizes the Stage-1 training dynamics under different regularizer configurations (S1–S4). Without $\mathcal{L}_{\text{usage}}$, basis utilization remains highly uneven throughout training, with persistently low usage entropy and N_{eff} (S1, S2). Enabling $\mathcal{L}_{\text{usage}}$ substantially increases both metrics (S3, S4), indicating broader and more stable basis participation. In contrast, $\mathcal{L}_{\text{struct}}$ primarily affects the geometric organization of the latent space: settings

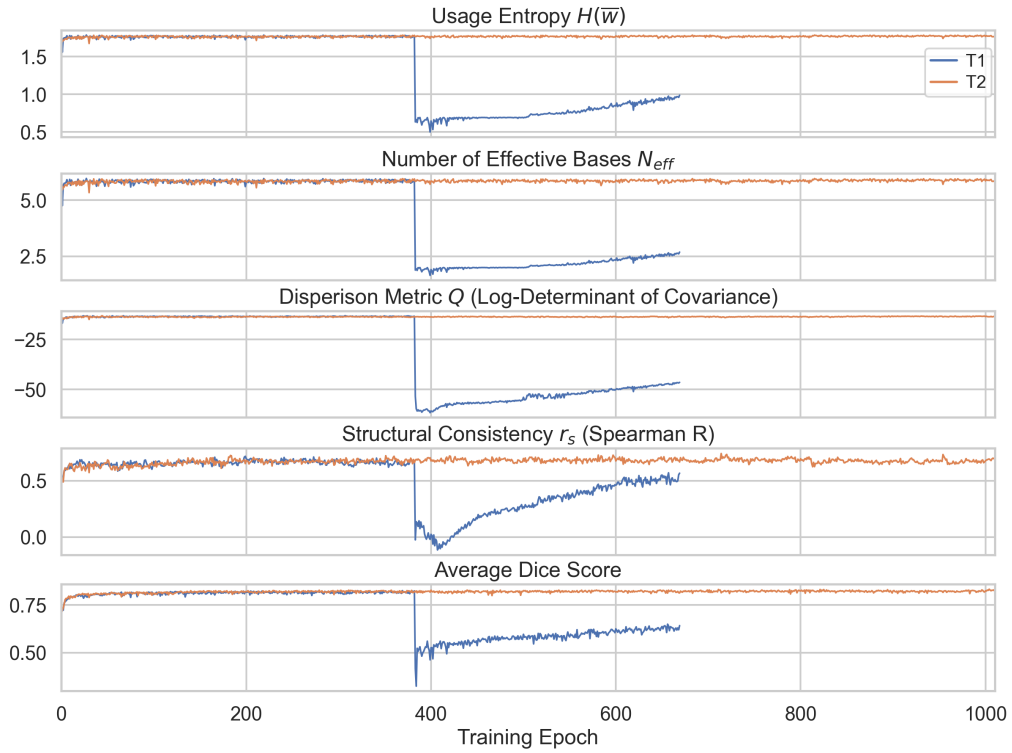


Figure 5.9: Stage-2 adaptation dynamics on the target domain, including evolution of basis-utilization metrics, simplex-geometry metrics, and target-domain segmentation Dice during unsupervised adaptation, comparing models without ($T1$) and with ($T2$) the usage regularizer. The latent simplex learned in stage 1 is kept fixed.

with $\mathcal{L}_{\text{struct}}$ (S2, S4) exhibit markedly higher dispersion Q, when compared with variants under the same usage-regularizer setting, as well as stronger structural consistency r_s. The full model (S4) consistently achieves the highest values across all metrics, corresponding to faster convergence and higher segmentation accuracy. Fig. 5.9 shows the Stage-2 adaptation dynamics on the target domain. When $\mathcal{L}_{\text{usage}}$ is removed during adaptation (T1), basis utilization exhibits an abrupt drop, accompanied by a collapse in dispersion Q and a decrease in r_s. This degradation coincides with unstable target-domain Dice. In contrast, retaining $\mathcal{L}_{\text{usage}}$ (T2) preserves balanced basis utilization and stable simplex geometry, leading to smooth and monotonic improvement in segmentation performance. These results indicate that while the latent manifold is fixed during Stage-2, the usage regularizer plays a critical

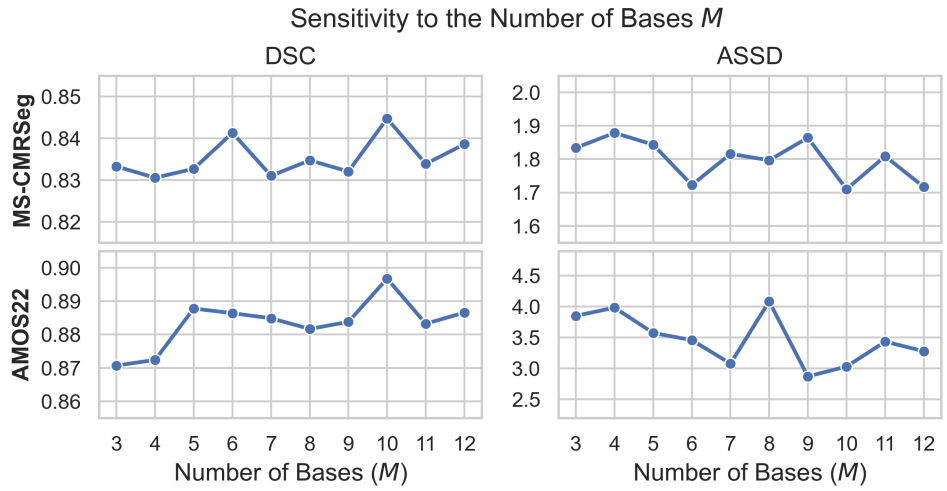


Figure 5.10: Sensitivity analysis with respect to the number of bases M. Segmentation performance is evaluated using DSC (left column) and ASSD (right column, in mm) on MS-CMRSeg (top row) and AMOS22 (bottom row).

role in stabilizing how target-domain samples populate and exploit the learned simplex.

Robustness to the Number of Bases

This ablation examines the effect of the number of bases M by sweeping M from 3 to 12 on the two datasets, while keeping all other training and evaluation settings fixed. As shown in Fig. 5.10, performance remains consistently strong over a wide range of M values on both datasets, forming a broad performance plateau. Importantly, neither dataset exhibits a preference for a narrowly tuned M. For the more anatomically complex AMOS22 dataset, slightly increased sensitivity is observed only at very small M, while performance quickly stabilizes as M increases. Overall, these results demonstrate that our method is robust to the choice of M. A moderately sized basis set is sufficient in practice, without requiring detailed tuning.

Effect of Stage-2 in the Source-Free Setting

This ablation investigates the contribution of Stage-2 target-only adaptation beyond Stage-1 source-supervised training in the source-free setting. While Stage-1 learns seman-

Table 5.6: Target-domain segmentation and reconstruction performance of our model on the MS-CMRSeg dataset after Stage-1 training and after Stage-2 target-only adaptation. Stage-1 refers to supervised training on the source domain, and Stage-2 refers to subsequent adaptation using unlabeled target data. Δ denotes the performance change from Stage-1 to Stage-2.

Training Stage	Segmentation		Reconstruction	
	DSC (%) ↑	ASSD (mm) ↓	PSNR (dB) ↑	SSIM ↑
Stage 1	74.0±10.3	3.16±1.43	17.9±0.83	0.420±0.045
Stage 2	83.1±5.55	1.88±0.77	21.8±1.21	0.514±0.063
Δ	+9.1	−1.28	+3.9	+0.094

Table 5.7: Target-domain segmentation and reconstruction performance of our model on the AMOS22 dataset after Stage-1 training and Stage-2 target-only adaptation.

Training Stage	Segmentation		Reconstruction	
	DSC (%) ↑	ASSD (mm) ↓	PSNR (dB) ↑	SSIM ↑
Stage 1	48.9±17.7	20.8±7.94	12.5±1.13	0.303±0.050
Stage 2	87.0±3.27	3.28±1.30	17.5±0.27	0.570±0.044
Δ	+38.1	−17.52	+5.0	+0.267

tic anatomical representations under source supervision, Stage-2 recalibrates appearance-sensitive mappings to align unlabeled target images with the learned semantic manifold. As shown in Tabs. 5.6 and 5.7, directly applying the Stage-1 model to the target domain leads to substantial performance degradation, while Stage-2 yields consistent and non-trivial improvements across both datasets. The paired visualizations in Fig. 5.11 further show that Stage-2 systematically corrects appearance-induced reconstruction biases and stabilizes anatomical structures, which in turn leads to improved segmentation accuracy. These results demonstrate that Stage-2 is not a minor fine-tuning step, but plays a critical role in recalibrating target-domain image–manifold mappings under unlabeled data.

Sensitivity to Source Pretraining Quality in the Source-Free Setting

This ablation analyzes the sensitivity of the proposed source-free adaptation framework

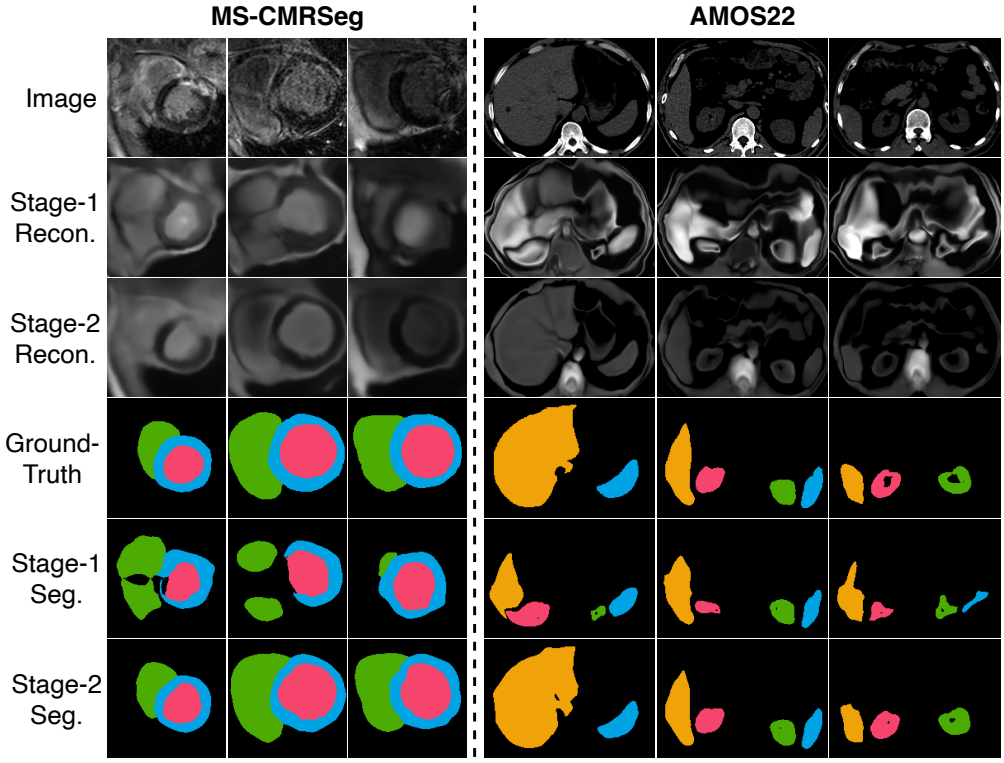


Figure 5.11: Qualitative comparison of model outputs on representative target-domain cases after Stage-1 and Stage-2 training.

to the quality of source pretraining by performing a checkpoint sweep on the source-domain training stage. Because the latent bases and segmentation decoder are fixed during Stage-2, source pretraining quality defines the semantic prior available for adaptation. Specifically, we early-stop source training at different validation Dice levels, obtaining a series of pretrained models spanning from under-trained to near-optimal initializations. Each checkpoint is then used to initialize Stage-2 target-only adaptation under identical settings.

Fig. 5.12 plots the target-domain Dice after adaptation as a function of the source-domain Dice achieved at the end of Stage-1. As expected, target performance depends on the reliability of source pretraining, since it provides the only semantic prior in the source-free setting. However, the empirical trend exhibits a much weaker dependence than the theoretical upper bound $y = x$, indicating that target performance degrades gracefully as source pretraining

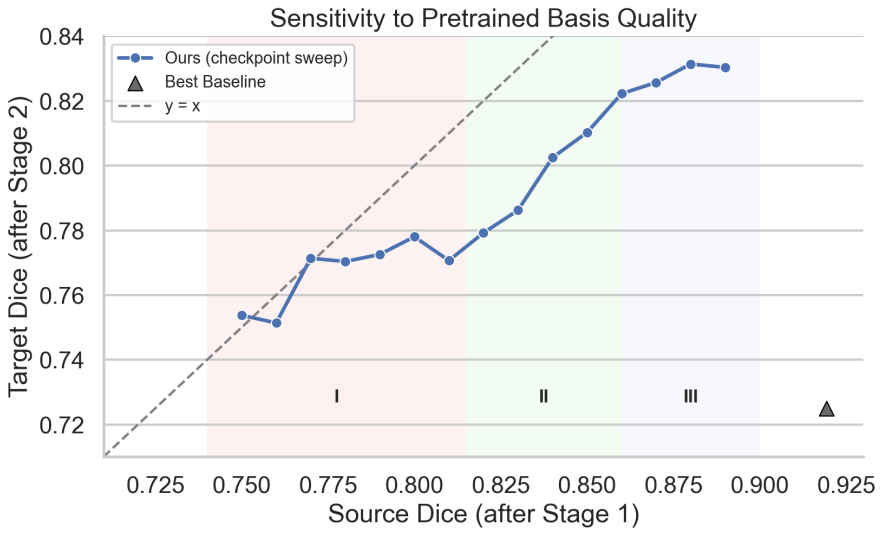


Figure 5.12: Sensitivity of source-free adaptation performance to the quality of source pretraining on MS-CMRSeg. Target-domain Dice after Stage-2 adaptation is shown as a function of the source-domain Dice achieved at the end of Stage-1. Each point corresponds to a pretrained checkpoint selected via early stopping. The dashed line $y = x$ indicates the theoretical upper bound.

quality decreases. We also observe three qualitative regimes: (i) under-trained source models lead to limited adaptation gains, (ii) once a moderate source performance threshold is reached, target adaptation becomes effective and stable, and (iii) further improvements in source Dice yield diminishing returns on the target domain. These results suggest that while source pretraining quality is necessary, the proposed framework operates robustly across a broad range of source initialization quality.

For reference, we also report the performance of the strongest baseline (ProtoContra) evaluated using its best-performing source-domain checkpoint. Despite this favorable initialization, the baseline remains clearly below the proposed method across the sensitivity range, including cases where our model is initialized from substantially weaker source checkpoints. This observation indicates that the superior target-domain performance of the proposed framework cannot be attributed solely to stronger source pretraining, but instead arises from its architectural design for source-free adaptation.

Table 5.8: Quantitative effect of removing the reconstruction loss $\mathcal{L}_{\text{recon}}$ in the source-accessible setting.

Dataset	$\mathcal{L}_{\text{recon}}$	Dice (%) ↑	ASSD (mm) ↓
MS-CMRSeg	✓	84.1±5.35	1.72±0.80
	–	63.8±8.43	4.92±1.25
AMOS22	✓	89.7±1.30	3.03±1.12
	–	42.0±11.0	19.9±6.13

Table 5.9: Quantitative effect of hierarchical warping scales on MS-CMRSeg. Different configurations correspond to using a subset Λ of multi-scale warping levels.

Warping Scales Λ	DSC (%) ↑	ASSD (mm) ↓
Single-scale ($\{5\}$)	77.9±6.39	2.59±0.87
Reduced-scale ($\{3, 5\}$)	82.8±4.53	1.90±0.69
Ours ($\{1, 3, 5\}$)	84.1±5.35	1.72±0.80

Effect of the Reconstruction Loss

This ablation evaluates the contribution of the reconstruction loss $\mathcal{L}_{\text{recon}}$ by removing it from the training objective under the source-accessible setting, while keeping all other components and hyperparameters unchanged. As shown in Tab. 5.8, removing $\mathcal{L}_{\text{recon}}$ leads to a substantial degradation in segmentation accuracy on both datasets. This behavior is consistent across datasets with different anatomical scope and imaging characteristics. These results indicate that the reconstruction pathway is essential for effective adaptation. Indeed, it provides the only image-based supervision that anchors the latent manifold to the observed target images, and is also a necessary component implied by the generative formulation as the observational likelihood.

Effect of Hierarchical Decomposition of Anatomical Structure

This ablation studies the effect of hierarchical anatomical priors by varying the set of canonical-space warping scales used in the model. In our formulation, anatomical bases are

defined across a fixed levels $\{1, \ldots, L\}$ of network resolutions, while hierarchical composition is controlled by selecting a subset Λ of these resolutions at which deformation is applied, as detailed in Eq. (5.3). This subset governs how anatomical variation is distributed across spatial scales.

We compare three variants: $\Lambda = \{5\}$ (single full-resolution scale), $\{3, 5\}$ (reduced-scale), and $\{1, 3, 5\}$ (ours). All variants share the same backbone, latent bases, and training protocol. As summarized in Tab. 5.9, segmentation performance improves consistently as additional warping scales are incorporated. These results indicate that hierarchical anatomical priors, realized through multi-scale warping, play an important role in capturing structural variability. In principle, using a larger set of scales (e.g., $\Lambda = \{1, \ldots, 5\}$) could further increase modeling capacity. Accordingly, we adopt $\Lambda = \{1, 3, 5\}$ as a practical trade-off between representational expressiveness and computational cost.

Effect of the Template Loss

This ablation evaluates the effect of $\widetilde{\mathcal{L}}_{\text{tem}}$. As shown in Tab. 5.5 row 3, removing the KL $\widetilde{\mathcal{L}}_{\text{tem}}$ over basis distributions causes the bases to drift independently, weakening segmentation performance. This highlights the value of basis regularization in preserving a coherent manipulation of the template $\mathbf{z}$ through $\mathbf{w}$ for effective adaptation.

5.5 Limitations and Future Directions

Despite the strong empirical performance of the proposed framework, several limitations warrant discussion.

2.5D/3D Extensions

Our method is in 2D, as the considered datasets exhibit substantial through-plane anisotropy. Under these conditions, enforcing a 3D or 2.5D architecture may introduce inappropriate inductive bias. Moreover, extending the method to 3D dramatically increases requirements on memory, training time, and annotated data to ensure stable optimization. Nevertheless,

the method remains straightforward to extend to 2.5D or 3D variants in implementation, such as slice-stacked encoders or volumetric canonical spaces with 3D deformations, when sufficiently isotropic data are available.

Cross-slice Consistency

The current framework processes slices independently, which may lead to through-plane inconsistencies in volumetric reconstructions. This limitation is partially mitigated by the large inter-slice spacing of the datasets, where anatomical correspondence between adjacent slices is inherently weak. Improving cross-slice coherence remains an important direction for future work. Given the explicit modeling of anatomical structure, natural extensions include enforcing smooth trajectories of latent anatomical codes along the slice direction or encouraging consistency in the canonical shape space across adjacent slices.

Computational Considerations

Our framework introduces additional computational and memory costs beyond standard encoder–decoder segmentation networks, primarily due to the anatomical bases, registration module, and multi-scale warping. This overhead is configurable through the number of bases and warping scales, allowing a flexible trade-off between accuracy and efficiency. At the same time, our method avoids computationally intensive objectives such as adversarial training or explicit cross-domain feature alignment, and remains practical under typical GPU constraints. Further efficiency gains may be achieved through basis pruning or distillation of the learned anatomical representations.

Modeling Assumptions of the Semantic Manifold

Our work models the bases using Gaussians to ensure stable optimization and tractable inference, which may limit expressiveness for complex anatomical variability. Exploring richer formulations, such as mixture-based priors or normalizing flows, is a natural direction for future work, albeit with increased challenges in identifiability, regularization, and optimization. Moreover, we adopt a simple weighted log-linear aggregation to compose bases

conditioned on the blending vector $\mathbf{w}$, which may limit expressiveness in highly heterogeneous anatomical settings. More expressive aggregation mechanisms, such as cross-attention, can be integrated easily, at the cost of additional computational and optimization complexity.

Dataset Bias

As with most publicly available benchmark datasets, the evaluated datasets may exhibit inherent biases related to cohort composition, acquisition protocols, distribution of anatomical variability, and annotation conventions, which implicitly introduce data and label noise. Our evaluation follows standard benchmark protocols without introducing additional selection criteria or noise-specific assumptions, and thus reflects model behavior under naturally occurring, non-ideal conditions, enabling fair comparison with prior work. Nevertheless, the disease spectrum and demographic diversity of these datasets may not fully reflect real-world clinical populations, and validation on larger and more diverse cohorts remains an important direction for future work.

Extensions Beyond the Evaluated Setting

Although we focus on unsupervised adaptation, our method is not tied to a specific supervision pattern, as it performs generative modeling at the level of individual images: each image contributes to learning through reconstruction, while segmentation supervision, when available, is incorporated as an additional objective. As a result, the framework naturally generalizes to more general settings, including multi-source and partially annotated datasets, where labeled and unlabeled images can be jointly incorporated into training. A systematic evaluation of these extensions is left as future work. Moreover, our framework is evaluated on pixel-wise segmentation of structural medical images, where anatomical shape can be meaningfully modeled. While the current instantiation focuses on MRI and CT segmentation, the underlying principle of decomposing images into shared semantic knowledge and subject-specific variations is not tied to a specific modality or task. Extending the framework to other modalities (e.g., PET or ultrasound) or to tasks like detection or classification would require specific observation models and supervision schemes, and is left as a promising

direction for future work.

5.6 Conclusion

We have proposed a unified and semantically grounded framework for unsupervised domain adaptation in medical image segmentation, which seamlessly supports both source-accessible and source-free scenarios. Our method explicitly disentangles canonical anatomy and individual-specific geometry through a shared latent manifold within a theoretically grounded Bayesian framework. By leveraging a structured composition of learnable anatomical bases, our method enables explainable and emergent domain adaptation indirectly via a structured, shared semantic space, without relying on explicit cross-domain alignment strategies. Extensive experiments on public multi-organ and multi-modality benchmarks demonstrate state-of-the-art performance of our model, particularly in the highly challenging source-free setting, with strong generalization and robustness under various domain shifts. Beyond segmentation accuracy, we illustrate the strong interpretability of our framework by visualizing the disentanglement, manifold traversal, and domain alignment results.

5.7 Chapter Takeaway

This chapter addressed unsupervised domain adaptation for medical image segmentation in a correspondence-free regime. The setting is defined by a labeled source domain and an unlabeled target domain that share a task but differ in acquisition environment and appearance, with no paired samples and no reliable notion of cross-image alignment. In this regime, heterogeneity is compounded: appearance variability is coupled with substantial sample-to-sample anatomical diversity and coverage differences, so adaptation cannot be grounded in registration-based correspondence.

The main methodological takeaway is that adaptation can be induced by *explicit anatomical organization*. By learning a probabilistic anatomical manifold as a global repository of canonical structures and by decomposing each observation into a selected canonical prototype

together with image-specific variation, the model ties prediction to task-relevant generative content and suppresses acquisition-dependent shortcuts. This yields architecture-emergent adaptation: transfer arises from the model's latent organization and inference procedure, without requiring an explicit alignment loss. The same latent formulation further supports a unified procedure across different source-data access assumptions, enabling both source-accessible and source-free adaptation within a single model family.

Relative to the previous projects, this chapter completes the escalation of representational requirements. Chapter 3 stabilized supervised segmentation under intensity-level appearance variability by constraining representational complexity. Chapter 4 enforced identifiable separation of common anatomy and geometry when correspondence exists but must be inferred. Here, correspondence itself is unavailable, so invariants must be preserved through global canonicalization: the model must maintain task-relevant structure across domains without relying on direct sample-level coupling. This establishes the thesis conclusion that interpretable, identifiable latent organization is a practical mechanism for robustness and performance under increasing heterogeneity.

Chapter 6

CONCLUSION

6.1 Summary of thesis

This dissertation studied medical image learning under progressively increasing heterogeneity through the lens of *identifiable latent organization*. The central claim developed in Chapter 1 is that performance degradation under compounded heterogeneity reflects a representational limitation: when task-relevant generative properties and observational variability are not explicitly and identifiably separated, models can exploit unstable observational cues as surrogates, leading to fragile generalization. Chapter 2 established Bayesian representation learning as a unifying technical framework for addressing this issue, emphasizing that semantic stability is governed by the coupling between generative specification and inference, and summarizing reusable design primitives, including priors, likelihood structure, variational assumptions, and objective shaping, that promote identifiability without committing to task-specific latent semantics.

The three projects then instantiated this thesis-wide requirement in increasingly demanding regimes. Chapter 3 addressed supervised MRI-based intracranial arterial calcification segmentation under intensity-level appearance heterogeneity. Although the setup follows a conventional supervised paradigm, calcification in MRI is often dark and weakly expressed, so segmentation depends on fragile contextual cues that are easily perturbed by scanner- and protocol-dependent appearance variation. By introducing a variational Bayesian formulation and restricting representational complexity, the method encouraged stable allocation of explanatory responsibility to task-relevant content rather than to acquisition-dependent appearance fluctuations, improving both robustness and performance.

Chapter 4 escalated the heterogeneity regime to multimodal and geometrically misaligned inputs in an *unsupervised* groupwise registration setting. Here, correspondence must be inferred, yet intensity similarity is unreliable across modalities. The proposed hierarchical Bayesian formulation disentangled common anatomy from image-specific geometry, enabling similarity to be evaluated in an intrinsic representation space while transformations were inferred jointly. This explicit separation promoted identifiable roles for latent components and yielded stable registration behavior under compounded modality and geometric variability.

Chapter 5 further removed the availability of reliable sample-level correspondence by studying unsupervised domain adaptation for segmentation. With a labeled source domain and an unlabeled target domain, adaptation must proceed from unpaired data under appearance shifts and substantial anatomical diversity. The proposed probabilistic manifold framework introduced global canonicalization via a structured latent organization in which each observation is explained through a selected canonical prototype and image-specific variation. This induced architecture-emergent adaptation: transfer arose from the model's latent organization rather than from an explicit alignment loss, and a unified procedure applied across different assumptions about source-data accessibility.

Taken together, these projects support the broader conclusion of the dissertation: interpretable robustness under heterogeneity is achieved by making latent roles identifiable. As heterogeneity escalates from within-contrast appearance variation, to multimodal observation differences coupled with registration-compatible geometric variation, and finally to correspondence-free domain shifts, the representational requirements become progressively stronger. Empirical compensation through data scale, generic regularization, or increasingly elaborate objective engineering is often insufficient to guarantee stability. Instead, robust performance emerges when models are designed with explicit latent organization and inference mechanisms that preserve task-relevant invariants while suppressing observational variability.

6.2 Limitations and Future Directions

Several limitations and future directions follow naturally from this perspective. First, identifiability in deep latent variable models remains sensitive to model mismatch and optimization, suggesting the need for stronger theoretical characterizations of when specific priors and likelihood structures yield stable latent semantics in practice. Second, the canonical structures learned by probabilistic manifolds depend on the diversity and coverage of training data; extending these models to more complex anatomies, broader populations, and longitudinal changes remains an open challenge. Third, while the methods in this thesis emphasize interpretability through latent organization, rigorous clinical validation requires connecting these representations to clinically meaningful uncertainty, decision-making workflows, and downstream outcomes. Fourth, the framework does not explicitly model or compensate for *imaging artifacts*. Artifacts (e.g., bias field in MRI, beam-hardening or metal artifacts in CT) are systematic, physics-linked intensity distortions that are central to image quality and differ from random noise. Addressing them typically requires modality-specific or physics-based forward models of the imaging pipeline. This thesis focuses on a general post-processing generative framework for structural images and does not incorporate such artifact models; robustness is pursued through identifiable latent organization rather than explicit artifact correction. When artifacts are severe or structured, preprocessing or dedicated correction steps may remain necessary.

Nevertheless, the results of this dissertation indicate that identifiable latent organization is not merely an interpretability preference, but a practical mechanism for robust learning in medical imaging. By treating disentanglement as an explicit modeling assumption and by using Bayesian inference as a disciplined way to allocate explanatory responsibility, one can obtain models that generalize more reliably as heterogeneity increases, while preserving semantic structure that supports scientific understanding and clinical trust.

6.3 Toward Representation-Centric Clinical Imaging Systems

While this dissertation focused on three specific methodological problems, the broader implication of the work lies in how medical imaging systems may be designed to cope with increasingly heterogeneous data environments.

Modern clinical imaging workflows are evolving toward large-scale, multi-center, and multi-modal settings in which acquisition conditions, patient populations, and imaging protocols vary substantially. In such environments, traditional pipelines that rely primarily on direct intensity-based representations often encounter instability, since observational variability can dominate the signals relevant to downstream clinical tasks. The results of this dissertation suggest an alternative perspective: robustness in medical image analysis can be achieved by explicitly organizing the representation space so that task-relevant generative structure is separated from observational variability.

Across the three projects, this idea manifested through progressively stronger forms of latent organization. In the segmentation setting of Chapter 3, constraining representational complexity encouraged the model to allocate explanatory responsibility to structural cues rather than to acquisition-dependent appearance fluctuations. In the multimodal registration framework of Chapter 4, anatomical correspondence emerged from an explicit hierarchical decomposition of common anatomy and image-specific geometry. In the domain adaptation setting of Chapter 5, this principle was extended further through a probabilistic anatomical manifold, allowing segmentation to be performed in a canonical representation space even when source and target domains lacked direct correspondence.

Viewed collectively, these developments point toward a representation-centric design paradigm for future clinical imaging systems. Instead of treating segmentation, registration, and cross-domain transfer as isolated tasks with separate heuristics, one may construct unified analysis frameworks in which anatomical structure, geometric variability, and observational conditions are modeled as distinct yet interacting latent factors. In such systems, downstream tasks become different forms of inference within a shared generative repre-

sentation, rather than independent prediction problems operating directly on raw image intensities.

This perspective also suggests a pathway for integrating methodological advances into larger clinical analysis platforms. Contemporary pipelines for neurovascular imaging, such as those used for intracranial arterial analysis, typically combine acquisition, preprocessing, segmentation, and quantitative analysis modules. Embedding representation-centered models within such systems may enable more stable operation across heterogeneous imaging environments, facilitating reliable deployment in multi-institutional and longitudinal clinical studies. In this sense, the contributions of this dissertation should be understood not only as individual algorithms, but as conceptual building blocks for future imaging systems capable of handling complex heterogeneity while preserving interpretable anatomical structure.

More broadly, these results illustrate how principled latent organization can transform the role of machine learning in medical imaging. By designing models in which anatomical knowledge is encoded directly in the representation space and inferred through structured probabilistic reasoning, it becomes possible to reconcile robustness, interpretability, and adaptability—three properties that are often in tension in purely data-driven approaches. Developing such representation-centered frameworks for increasingly complex imaging scenarios remains an important direction for future research.

BIBLIOGRAPHY

[1] P. Agrawal, R. T. Whitaker, and S. Y. Elhabian. An optimal, generative model for estimating multi-label probabilistic maps. *IEEE Transactions on Medical Imaging*, 39 (7):2316–2326, 2020.

[2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[3] S. Allassonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 2007.

[4] K. Amann. Media calcification and intima calcification are distinct entities in chronic kidney disease. *Clinical Journal of the American Society of Nephrology*, 3(6), 2008. ISSN 1555-9041. URL https://journals.lww.com/cjasn/fulltext/2008/11000/media_calcification_and_intima_calcification_are.3.aspx.

[5] J. F. Arenillas. Intracranial atherosclerosis: Current concepts. *Stroke*, 42(1_-suppl_1):S20–S23, 2011. doi: 10.1161/STROKEAHA.110.597278. URL https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.110.597278.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[7] 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.

[8] G. Arvanitidis, M. Gonz'alez-Duque, A. Pouplin, D. Kalatzis, and S. Hauberg. Pulling

back information geometry. In *International Conference on Artificial Intelligence and Statistics*, 2021.

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

[10] B. Audelan, D. Hamzaoui, S. Montagne, R. Renard-Penna, and H. Delingette. Robust bayesian fusion of continuous segmentation maps. *Medical Image Analysis*, 78:102398, 2022.

[11] B. Avants and J. C. Gee. Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage*, 23:S139–S150, 2004.

[12] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[13] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1): 1–13, 2017.

[14] 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.

[15] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2019.

[16] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82:102617, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102617.

[17] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):

1798–1828, 2013.

[18] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations*, 2020.

[19] K. K. Bhatia, J. V. Hajnal, A. Hammers, and D. Rueckert. Similarity metrics for groupwise non-rigid registration. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 10 Pt 2:544–52, 2007.

[20] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[21] C. Blaiotta, P. Freund, M. J. Cardoso, and J. Ashburner. Generative diffeomorphic modelling of large mri data sets for probabilistic template construction. *NeuroImage*, 166:117–134, 2018.

[22] J. L. Boes and C. R. Meyer. Multi-variate mutual information for registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–612. Springer, 1999.

[23] A. Bône, P. Vernhet, O. Colliot, and S. Durrleman. Learning joint shape and appearance representations with metamorphic auto-encoders. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 202–211. Springer, 2020.

[24] G. Bortsova, D. Bos, F. Dubost, M. W. Vernooij, M. K. Ikram, G. van Tulder, and M. de Bruijne. Automated segmentation and volume measurement of intracranial internal carotid artery calcification at noncontrast ct. *Radiology: Artificial Intelligence*, 3(5):e200226, 2021. doi: 10.1148/ryai.2021200226. URL https://doi.org/10.1148/ryai.2021200226.

[25] D. Bos, M. W. Vernooij, S. E. Elias-Smale, B. F. Verhaaren, H. A. Vrooman, A. Hofman, W. J. Niessen, J. C. Witteman, A. van der Lugt, and M. A. Ikram. Atherosclerotic calcification relates to cognitive function and to brain changes on magnetic resonance imaging. *Alzheimer's & Dementia*, 8(5S):S104–S111, 2012. doi: https:

//doi.org/10.1016/j.jalz.2012.01.008. URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2012.01.008.

[26] D. Bos, M. L. P. Portegies, A. van der Lugt, M. J. Bos, P. J. Koudstaal, A. Hofman, G. P. Krestin, O. H. Franco, M. W. Vernooij, and M. A. Ikram. Intracranial Carotid Artery Atherosclerosis and the Risk of Stroke in Whites: The Rotterdam Study. *JAMA Neurology*, 71(4):405–411, 04 2014. ISSN 2168-6149. doi: 10.1001/jamaneurol.2013.6223. URL https://doi.org/10.1001/jamaneurol.2013.6223.

[27] D. J. Brenner and E. J. Hall. Computed tomography — an increasing source of radiation exposure. *New England Journal of Medicine*, 357(22):2277–2284, 2007. doi: 10.1056/NEJMra072149. URL https://doi.org/10.1056/NEJMra072149. PMID: 18046031.

[28] J.-M. Bugnicourt, C. Leclercq, J.-M. Chillon, M. Diouf, H. Deramond, S. Canaple, C. Lamy, Z. A. Massy, and O. Godefroy. Presence of intracranial artery calcification is associated with mortality and vascular events in patients with ischemic stroke after hospital discharge. *Stroke*, 42(12):3447–3453, 2011. doi: 10.1161/STROKEAHA.111.618652. URL https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.111.618652.

[29] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han. Domain-specific batch normalization for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi: 10.1109/CVPR.2019.00753.

[30] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101535.

[31] T. Che, Y. Zheng, X. Sui, Y. Jiang, J. Cong, W. Jiao, and B. Zhao. Dgr-net: Deep groupwise registration of multispectral images. In *International Conference on Information Processing in Medical Imaging*, pages 706–717. Springer, 2019.

[32] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng. Unsupervised bidirectional cross-

modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(7):2494–2505, 2020. doi: 10.1109/TMI.2020.2972701.

[33] C. Chen, Q. Liu, Y. Jin, Q. Dou, and P.-A. Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 225–235, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3.

[34] 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. ISSN 1361-8415.

[35] 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*, 100:103385, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103385. URL https://www.sciencedirect.com/science/article/pii/S1361841524003104.

[36] L. Chen, M. Mossa-Basha, N. Balu, G. Canton, J. Sun, K. Pimentel, T. S. Hatsukami, J.-N. Hwang, and C. Yuan. Development of a quantitative intracranial vascular features extraction tool on 3d mra using semiautomated open-curve active contour vessel tracing. *Magnetic Resonance in Medicine*, 79(6):3229–3238, 2018. doi: https://doi.org/10.1002/mrm.26961. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.26961.

[37] S. Chen, H. Zhao, J. Li, Z. Zhou, R. Li, N. Balu, C. Yuan, H. Chen, and X. Zhao. Evaluation of carotid atherosclerotic plaque surface characteristics utilizing simultaneous noncontrast angiography and intraplaque hemorrhage (snap) technique. *Journal of Magnetic Resonance Imaging*, 47(3):634–639, 2018. doi: https://doi.org/10.1002/jmri.25815. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.25815.

[38] G. E. Christensen, H. J. Johnson, and M. W. Vannier. Synthesizing average 3d anatom-

ical shapes. *NeuroImage*, 32(1):146–158, 2006.

[39] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26:1045–1057, 2013.

[40] H. Cui, Y. Li, Y. Wang, D. Xu, L.-M. Wu, and Y. Xia. Toward accurate cardiac mri segmentation with variational autoencoder-based unsupervised domain adaptation. *IEEE Transactions on Medical Imaging*, 43(8):2924–2936, 2024. doi: 10.1109/TMI.2024.3382624.

[41] S. Czolbe, P. Pegios, O. Krause, and A. Feragen. Semantic similarity metrics for image registration. *Medical Image Analysis*, 87:102830, 2023.

[42] A. Dalca, M. Rakic, J. Guttag, and M. Sabuncu. Learning conditional deformable templates with convolutional networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[43] 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. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.07.006.

[44] I. Daunhawer, T. M. Sutter, K. Chin-Cheong, E. Palumbo, and J. E. Vogt. On the limitations of multimodal VAEs. In *International Conference on Learning Representations*, 2022.

[45] B. D. De Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Isgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, 2019.

[46] X. Deng, E. Liu, S. Li, Y. Duan, and M. Xu. Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. *IEEE Transactions on Image Processing*, 32:1078–1091, 2023.

[47] Z. Ding and M. Niethammer. Aladdin: Joint atlas building and diffeomorphic regis-

tration learning with pairwise alignment. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20752–20761, 2022.

[48] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 74–82. Springer, 2019.

[49] P. Dupuis, U. Grenander, and M. I. Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of Applied Mathematics*, pages 587–600, 1998.

[50] B. Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012.

[51] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1): S143–S152, 2009.

[52] S. Gao, H. Zhou, Y. Gao, and X. Zhuang. Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. *Medical image analysis*, 89: 102889, 2023.

[53] X. Geng, G. E. Christensen, H. Gu, T. J. Ross, and Y. Yang. Implicit reference-based group-wise image registration and its application to structural and functional mri. *NeuroImage*, 47(4):1341–1351, 2009.

[54] D. Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(83)80009-3.

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[56] P. M. Graffy, J. Liu, S. D. O'Connor, R. M. Summers, and P. J. Pickhardt. Automated segmentation and quantification of aortic calcification at abdominal ct: application of a deep learning-based algorithm to a longitudinal screening cohort. *Abdominal Radiology*,

44:2921–2928, 2019.

[57] U. Grenander. *General pattern theory: A mathematical study of regular structures*. Oxford University Press on Demand, 1993.

[58] U. Grenander and M. I. Miller. Computational anatomy: an emerging discipline. *Quarterly of Applied Mathematics*, 56:617–694, 1998.

[59] H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022. doi: 10.1109/TBME.2021.3117407.

[60] A. Guimond, J. Meunier, and J.-P. Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192–210, 2000.

[61] D. Gunning and D. Aha. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2):44–58, 2019.

[62] Y. Guo, G. Canton, L. Chen, J. Sun, D. B. Geleri, N. Balu, D. Xu, M. Mossa-Basha, T. S. Hatsukami, and C. Yuan. Multi-planar, multi-contrast and multi-time point analysis tool (mocha) for intracranial vessel wall characterization. *Journal of Magnetic Resonance Imaging*, 56(3):944–955, 2022. doi: https://doi.org/10.1002/jmri.28087. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.28087.

[63] Y. Guo, G. Canton, D. Baylam Geleri, N. Balu, J. Sun, M. Kharaji, N. Zanaty, X. Wang, K. Zhang, D. L. Tirschwell, T. S. Hatsukami, C. Yuan, and M. Mossa-Basha. Plaque evolution and vessel wall remodeling of intracranial arteries: A prospective, longitudinal vessel wall mri study. *Journal of Magnetic Resonance Imaging*, n/a(n/a), 2023. doi: https://doi.org/10.1002/jmri.29185.

[64] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In A. Crimi and S. Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08999-2.

[65] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. A. Landman,

H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 1748–1758. IEEE, 2022. ISBN 978-1-6654-0915-5. URL http://dblp.uni-trier.de/db/conf/wacv/wacv2022.html#HatamizadehTNOM22.

[66] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016.

[67] Z. He and A. C. S. Chung. Unsupervised end-to-end groupwise registration framework without generating templates. In *IEEE International Conference on Image Processing*, pages 375–379, 2020. doi: 10.1109/ICIP40778.2020.9191141.

[68] A. Hering, L. Hansen, T. C. Mok, A. C. Chung, H. Siebert, S. Häger, A. Lange, S. Kuckertz, S. Heldmann, W. Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 2022.

[69] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[70] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

[71] G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 08 2002. ISSN 0899-7667. doi: 10.1162/089976602760128018. URL https://doi.org/10.1162/089976602760128018.

[72] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011.

[73] M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NeurIPS*, volume 1, 2016.

[74] 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.

[75] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical Image Analysis*, 49:1–13, 2018.

[76] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. doi: 10.1109/ICCV.2017.167.

[77] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, pages 172–189, 2018.

[78] I. A. Huijben, W. Kool, M. B. Paulus, and R. J. Van Sloun. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1353–1371, 2022.

[79] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203—211, February 2021. ISSN 1548-7091. doi: 10.1038/s41592-020-01008-z. URL https://doi.org/10.1038/s41592-020-01008-z.

[80] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

[81] Y. Ji, H. Bai, J. Yang, C. Ge, Y. Zhu, R. Zhang, Z. Li, L. Zhang, W. Ma, X. Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022.

[82] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[83] M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes. Dual-stream pyramid registration network. *Medical Image Analysis*, 78:102379, 2022.

[84] A. E. Kavur, N. S. Gezer, M. Baris, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Ozkan, et al. Chaos challenge-combined (ct-mr) healthy

abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[85] A. Khanna, N. D. Londhe, S. Gupta, and A. Semwal. A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images. *Biocybernetics and Biomedical Engineering*, 40(3):1314–1327, 2020. ISSN 0208-5216. doi: https://doi.org/10.1016/j.bbe.2020.07.007. URL https://www.sciencedirect.com/science/article/pii/S0208521620300887.

[86] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020.

[87] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[88] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[89] J. J. Kivinen and C. K. Williams. Transformation equivariant boltzmann machines. In *International Conference on Artificial Neural Networks*, pages 1–9. Springer, 2011.

[90] I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

[91] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Proceedings of the 5th International Conference on Neural Information Processing Systems*, NIPS'91, page 950–957, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558602224.

[92] N. Lessmann, B. van Ginneken, M. Zreik, P. A. de Jong, B. D. de Vos, M. A. Viergever, and I. Isgum. Automatic calcium scoring in low-dose chest ct using deep neural networks with dilated convolutions. *IEEE Transactions on Medical Imaging*, 37(2):615–625, 2018. doi: 10.1109/TMI.2017.2769839.

[93] M. E. Leventon and W. E. L. Grimson. Multi-modal volume registration using joint

intensity distributions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1057–1066. Springer, 1998.

[94] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.07.005.

[95] F. Liu, K. Yan, A. P. Harrison, D. Guo, L. Lu, A. L. Yuille, L. Huang, G. Xie, J. Xiao, X. Ye, et al. Same: Deformable image registration based on self-supervised anatomical embeddings. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 87–97. Springer, 2021.

[96] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

[97] P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi. Multi-modal image set registration and atlas formation. *Medical image analysis*, 10(3):440–451, 2006.

[98] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[99] X. Luo and X. Zhuang. Mvmm-regnet: A new image registration framework based on multivariate mixture model and neural network estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 149–159. Springer, 2020.

[100] X. Luo and X. Zhuang. X-metric: An n-dimensional information-theoretic framework for groupwise registration and deep combined computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9206–9224, 2023.

[101] X. Luo, X. Wang, L. Shapiro, C. Yuan, J. Feng, and X. Zhuang. Bayesian Unsupervised Disentanglement of Anatomy and Geometry for Deep Groupwise Image Registration . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(01):711–729, Jan.

2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3609521.

[102] J. Ma, M. I. Miller, A. Trouvé, and L. Younes. Bayesian template estimation in computational anatomy. *NeuroImage*, 42(1):252–261, 2008.

[103] T. Ma, X. Dai, S. Zhang, and Y. Wen. Pivit: Large deformation image registration with pyramid-iterative vision transformer. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 602–612. Springer, 2023. ISBN 978-3-031-43999-5.

[104] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[105] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[106] D. Mandell, M. Mossa-Basha, Y. Qiao, C. Hess, F. Hui, C. Matouk, M. Johnson, M. Daemen, A. Vossough, M. Edjlali, D. Saloner, S. Ansari, B. Wasserman, and D. Mikulis. Intracranial vessel wall mri: Principles and expert consensus recommendations of the american society of neuroradiology. *American Journal of Neuroradiology*, 38(2):218–229, 2017. ISSN 0195-6108. doi: 10.3174/ajnr.A4893. URL https://www.ajnr.org/content/38/2/218.

[107] W. Mao, M. Zhu, Z. Sun, S. Shen, L. Y. Wu, H. Chen, and C. Shen. De novo protein design using geometric vector field networks. In *International Conference on Learning Representations*, 2024.

[108] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010.

[109] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages

4402–4412, 2019.

[110] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[111] M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41(1):61–84, 2001.

[112] H. K. Miyamoto, F. C. C. Meneghetti, J. Pinele, and S. I. R. Costa. On closed-form expressions for the fisher–rao distance. *Information Geometry*, Sep 2024. ISSN 2511-249X. doi: 10.1007/s41884-024-00143-2.

[113] M. Mossa-Basha, C. Yuan, B. A. Wasserman, D. J. Mikulis, T. S. Hatsukami, N. Balu, A. Gupta, C. Zhu, L. Saba, D. Li, J. K. DeMarco, V. T. Lehman, Y. Qiao, H. Jager, M. Wintermark, W. Brinjikji, C. P. Hess, and D. Saloner. Survey of the american society of neuroradiology membership on the use and value of extracranial carotid vessel wall mri. *American Journal of Neuroradiology*, 43:1756 – 1761, 2022.

[114] D. Moyer, E. Abaci Turk, P. E. Grant, W. M. Wells, and P. Golland. Equivariant filters for efficient tracking in 3d imaging. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 193–202. Springer, 2021.

[115] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018.

[116] J. Orchard and R. Mann. Registering a multisensor ensemble of images. *IEEE Transactions on Image Processing*, 19(5):1236–1247, 2009.

[117] K. Osawa, S. Swaroop, M. E. E. Khan, A. Jain, R. Eschenhagen, R. E. Turner, and R. Yokota. Practical deep learning with bayesian principles. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[118] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin,

N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[119] C. Patricio, J. C. Neves, and L. F. Teixeira. Explainable deep learning methods in medical image classification: A survey. *ACM Comput. Surv.*, 56(4), Oct. 2023. ISSN 0360-0300. doi: 10.1145/3625287. URL https://doi.org/10.1145/3625287.

[120] M. B. Paulus, C. J. Maddison, and A. Krause. Rao-blackwellizing the straight-through gumbel-softmax gradient estimator. In *International Conference on Learning Representations*, 2021.

[121] J. Pearl. *Causality*. Cambridge University Press, 2009.

[122] N. Pielawski, E. Wetzer, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, and N. Sladoje. Comir: Contrastive multimodal image representation for registration. In *Advances in Neural Information Processing Systems*, volume 33, pages 18433–18444, 2020.

[123] M. Polfliet, S. Klein, W. Huizinga, M. M. Paulides, W. J. Niessen, and J. Vandemeulebroucke. Intrasubject multimodal groupwise registration with the conditional template entropy. *Medical Image Analysis*, 46:15–25, 2018.

[124] M. I. Posner. *The Foundations of Cognitive Science*. The MIT Press, 11 1989. ISBN 9780262281805. doi: 10.7551/mitpress/3072.001.0001.

[125] G.-J. Qi, L. Zhang, F. Lin, and X. Wang. Learning generalized transformation equivariant representations via autoencoding transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[126] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. In *International Conference on Information Processing in Medical Imaging*, pages 249–261. Springer, 2019.

[127] J. Qiu, L. Li, S. Wang, K. Zhang, Y. Chen, S. Yang, and X. Zhuang. Myops-net: Myocardial pathology segmentation with flexible combination of multi-sequence cmr

images. *Medical Image Analysis*, 84:102694, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102694. URL https://www.sciencedirect.com/science/article/pii/S136184152200322X.

[128] P. Reizinger, L. Gresele, J. Brady, J. Von Kügelgen, D. Zietlow, B. Schölkopf, G. Martius, W. Brendel, and M. Besserve. Embrace the gap: Vaes perform independent mechanism analysis. In *Advances in Neural Information Processing Systems*, volume 35, pages 12040–12057, 2022.

[129] A. Roche, G. Malandain, and N. Ayache. Unifying maximum likelihood approaches in medical image registration. *International Journal of Imaging Systems and Technology*, 11(1):71–80, 2000.

[130] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[131] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jäger, and K. H. Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation. 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 405–415, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43901-8.

[132] L. Saba, C. Yuan, T. Hatsukami, N. Balu, Y. Qiao, J. DeMarco, T. Saam, A. Moody, D. Li, C. Matouk, M. Johnson, H. Jäger, M. Mossa-Basha, M. Kooi, Z. Fan, D. Saloner, M. Wintermark, D. Mikulis, and B. Wasserman. Carotid artery wall imaging: Perspective and guidelines from the asnr vessel wall imaging study group and expert consensus recommendations of the american society of neuroradiology. *American Journal of Neuroradiology*, 39(2):E9–E31, 2018. ISSN 0195-6108. doi: 10.3174/ajnr.A5488. URL https://www.ajnr.org/content/39/2/E9.

[133] U. Schmidt and S. Roth. Learning rotation-aware features: From invariant priors to

equivariant descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2050–2057. IEEE, 2012.

[134] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, et al. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 573–581. Springer, 2001.

[135] Y. Shi, S. N, B. Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems*, pages 15718–15729, 2019.

[136] H. Shu, J. Sun, T. S. Hatsukami, N. Balu, D. S. Hippe, H. Liu, T. R. Kohler, W. Zhu, and C. Yuan. Simultaneous noncontrast angiography and intraplaque hemorrhage (snap) imaging: Comparison with contrast-enhanced mr angiography for measuring carotid stenosis. *Journal of Magnetic Resonance Imaging*, 46(4):1045–1052, 2017. doi: https://doi.org/10.1002/jmri.25653. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.25653.

[137] Z. Shu, M. Sahasrabudhe, R. A. Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *European Conference on Computer Vision*, pages 650–665, 2018.

[138] H. Siebert, L. Hansen, and M. P. Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 174–179. Springer, 2021.

[139] N. Skafte and S. Hauberg. Explicit disentanglement of appearance and perspective in generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[140] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013.

[141] Z. Spiclin, B. Likar, and F. Pernus. Groupwise registration of multimodal images by an

efficient joint entropy minimization scheme. *IEEE Transactions on Image Processing*, 21(5):2546–2558, 2012.

[142] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[143] C. Studholme and V. Cardenas. A template free approach to volumetric spatial normalization of brain anatomy. *Pattern Recognition Letters*, 25(10):1191–1202, 2004.

[144] C. Studholme, D. L. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[145] D. Subedi, U. S. Zishan, F. Chappell, M.-L. Gregoriades, C. Sudlow, R. Sellar, and J. Wardlaw. Intracranial carotid calcification on cranial computed tomography. *Stroke*, 46(9):2504–2509, 2015. doi: 10.1161/STROKEAHA.115.009716. URL https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.115.009716.

[146] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal ELBO, 2021.

[147] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

[148] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101693. URL https://www.sciencedirect.com/science/article/pii/S136184152030058X.

[149] J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[150] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc.*

of the 37-th Annual Allerton Conference on Communication, Control and Computing, pages 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

[151] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11): 4793–4813, 2020.

[152] J. Tomczak and M. Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223, 2018.

[153] A. Trouvé. Diffeomorphisms groups and pattern matching in image analysis. *International Journal of Computer Vision*, 28(3):213–221, 1998.

[154] X. Tu, Z.-J. Cao, x. chenrui, S. Mostafavi, and G. Gao. Cross-linked unified embedding for cross-modality representation learning, 2022.

[155] A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc., 2020.

[156] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756, 2016.

[157] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

[158] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[159] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[160] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[161] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *2019 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 2512–2521, 2019. doi: 10.1109/CVPR.2019.00262.

[162] C. Wachinger and N. Navab. Simultaneous registration of multiple images: similarity metrics and efficient optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1221–1233, 2012.

[163] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[164] H. Wang and D.-Y. Yeung. A survey on bayesian deep learning. *ACM Computing Surveys*, 53(5):1–37, 2020.

[165] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623, 2012.

[166] H. Wang, D. Ni, and Y. Wang. Modet: Learning deformable image registration via motion decomposition transformer. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 740–749. Springer, 2023. ISBN 978-3-031-43999-5.

[167] X. Wang, X. Luo, and X. Zhuang. Bingo: Bayesian intrinsic groupwise registration via explicit hierarchical disentanglement. In A. Frangi, M. de Bruijne, D. Wassermann, and N. Navab, editors, *Information Processing in Medical Imaging*, pages 319–331, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34048-2.

[168] X. Wang, G. Canton, Y. Guo, K. Zhang, H. Akcicek, E. Yaman Akcicek, T. Hatsukami, J. Zhang, B. Sun, H. Zhao, Y. Zhou, L. Shapiro, M. Mossa-Basha, C. Yuan, and N. Balu. Automated mri-based segmentation of intracranial arterial calcification by restricting feature complexity. *Magnetic Resonance in Medicine*, 93(1):384–396, 2025. doi: https://doi.org/10.1002/mrm.30283.

[169] X. Wang, Y. Guo, J. Xia, K. Zhang, N. Balu, M. Mossa-Basha, L. Shapiro, and

C. Yuan. Unified and semantically grounded domain adaptation for medical image segmentation, 2025. URL https://arxiv.org/abs/2508.08660.

[170] X. Wang, Y. Guo, K. Zhang, N. Balu, M. Mossa-Basha, L. Shapiro, and C. Yuan. Remind: Remembering anatomical variations for interpretable domain adaptive medical image segmentation. In I. Oguz, S. Zhang, and D. N. Metaxas, editors, *Information Processing in Medical Imaging*, pages 327–341, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-031-96628-6.

[171] W. Weng, Y. Ku, Z. Chen, H. Zheng, C. Xu, H. Ding, L. Li, and G. Wang. Superficial femoral artery calcification segmentation and detection in ct angiography using convolutional neural network. *Computers in Biology and Medicine*, 148:105951, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105951. URL https://www.sciencedirect.com/science/article/pii/S0010482522006862.

[172] F. Wu and X. Zhuang. Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Transactions on Medical Imaging*, 40(12): 3555–3567, 2021. doi: 10.1109/TMI.2021.3090412.

[173] J. Wu, G. Wang, R. Gu, T. Lu, Y. Chen, W. Zhu, T. Vercauteren, S. Ourselin, and S. Zhang. Upl-sfda: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42 (12):3932–3943, 2023. doi: 10.1109/TMI.2023.3318364.

[174] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[175] X. Xing, R. Gao, T. Han, S.-C. Zhu, and Y. N. Wu. Deformable generator networks: unsupervised disentanglement of appearance and geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1162–1179, 2020.

[176] 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.

[177] C. Yang, X. Guo, Z. Chen, and Y. Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79:102457, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102457.

[178] C. X.-y. Yang Wen-jie, Wong Ka-sing. Intracranial atherosclerosis: From microscopy to high-resolution magnetic resonance imaging. *J Stroke*, 19(3):249–260, 2017. doi: 10.5853/jos.2016.01956. URL http://www.j-stroke.org/journal/view.php?number=192.

[179] K. Yao, Z. Su, K. Huang, X. Yang, J. Sun, A. Hussain, and F. Coenen. A novel 3d unsupervised domain adaptation framework for cross-modality medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(10):4976–4986, 2022. doi: 10.1109/JBHI.2022.3162118.

[180] K. Ying, Q. Zhong, W. Mao, Z. Wang, H. Chen, L. Y. Wu, Y. Liu, C. Fan, Y. Zhuge, and C. Shen. Ctvis: Consistent training for online video instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023.

[181] W. J. Yoon, P. Crisostomo, P. Halandras, C. F. Bechara, and B. Aulivola. The use of the agatston calcium score in predicting carotid plaque vulnerability. *Annals of Vascular Surgery*, 54:22–26, 2019. ISSN 0890-5096. doi: https://doi.org/10.1016/j.avsg.2018.08.070. URL https://www.sciencedirect.com/science/article/pii/S0890509618307568.

[182] Q. Yu, N. Xi, J. Yuan, Z. Zhou, K. Dang, and X. Ding. Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and contrastive learning. 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 3–12, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43990-2.

[183] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. doi:

10.1109/ICCV.2019.00612.

[184] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

[185] J. Zhang and A. Rangarajan. Multimodality image registration using an extensible information metric and high dimensional histogramming. In *International Conference on Information Processing in Medical Imaging*, pages 725–737. Springer, 2005.

[186] M. Zhang and P. T. Fletcher. Bayesian principal geodesic analysis for estimating intrinsic diffeomorphic image variability. *Medical Image Analysis*, 25(1):37–44, 2015.

[187] X. Zhang, Y. Wu, E. Angelini, A. Li, J. Guo, J. M. Rasmussen, T. G. O'Connor, P. D. Wadhwa, A. P. Jackowski, H. Li, J. Posner, A. F. Laine, and Y. Wang. Mapseg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3d masked autoencoding and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5851–5862, June 2024.

[188] Y. Zhang, P. Tino, A. Leonardis, and K. Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5 (5):726–742, 2021.

[189] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021. doi: 10.1109/JPROC.2021.3054390.

[190] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, editors, *Deep Learning in Medical Image Analy-*

sis and Multimodal Learning for Clinical Decision Support, pages 3–11, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00889-5.

[191] X. Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (12):2933–2946, 2019.

[192] X. Zhuang, J. Xu, X. Luo, C. Chen, C. Ouyang, D. Rueckert, V. M. Campello, K. Lekadir, S. Vesal, N. RaviKumar, et al. Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge. *Medical Image Analysis*, 81:102528, 2022.

VITA

Xin Wang was born in Anqing, Anhui, China. He received the Bachelor of Engineering degree from the Department of Electronic Engineering, School of Information Science and Technology, Fudan University, where he was enrolled in the Excellence Engineer Program and advised by Professor Jinhua Yu. After completing his undergraduate studies, he worked as a visiting scholar for one year in the School of Data Science, Fudan University, under the supervision of Professor Xiahai Zhuang.

He subsequently pursued doctoral studies in the Department of Electrical and Computer Engineering at the University of Washington, advised by Professor Linda Shapiro and Professor Chun Yuan. During his PhD program, he also earned the Master of Science degree in Statistics at the University of Washington. He worked as a Graduate Research Assistant in the Vascular Imaging Lab in the Department of Radiology, focusing on Bayesian and disentangled representation learning methods for robust and interpretable medical image analysis under increasing heterogeneity. In addition to methodological research, Xin contributed to the development of software tools that strengthened the lab's cardiovascular imaging analysis workflow, including MOCHA Viewer and Vessel Voyager.

Beyond academic research, Xin gained industry experience through internships at United Imaging Intelligence and TikTok. At United Imaging Intelligence, he worked on medical AI methods for imaging analysis, while at TikTok he contributed to multimodal large language models for video understanding. He has served as a reviewer for leading venues in medical imaging and machine learning, including MICCAI and IEEE Transactions on Medical Imaging.