A Plug-In Curriculum Scheduler for Improved Deformable Medical Image Registration

Anonymous authors
Paper under double-blind review

Abstract

Deformable image registration is a crucial task in medical image analysis, and its complexity has spurred significant research and ongoing progress. Much of the work in this area has concentrated on achieving incremental performance gains by adjusting network architectures or introducing new loss functions. However, these modifications are often tailored to specific tasks or datasets, which limits their general applicability. To address this limitation, we propose an innovative solution: a plug-in curriculum scheduler that can be seamlessly integrated into existing methods without changing their core architecture. Our scheduler, inspired by curriculum learning, progressively increases task difficulty to enhance performance, incorporating sample difficulty and matching accuracy as key criteria. Sample difficulty is assessed at voxel and volume levels, using Variance of Gradients for voxel complexity and Gaussian blurring for volume evaluation, while matching accuracy involves gradually increasing supervision for improved alignment and accuracy. We empirically demonstrate that this scheduler achieves superior accuracy and visual quality in various tasks and datasets.

1 Introduction

Deformable medical image registration is vital for medical image analysis because it allows precise alignment of images from various times or modalities. This accuracy is key for identifying changes, planning treatments, and combining data from different imaging sources. The field's complexity and its significant role in diagnosis, treatment, and personalized care have led to extensive research and ongoing advancements over the years. However, because of the difficulties in accurately representing deformation fields, research in deformable medical image registration has primarily focused on incremental performance improvements. These improvements often involve minor changes to network architectures, the integration of hierarchical or iterative processes, or the addition of loss functions. In most cases, these methods follow a common structure: a network inputs two images and generates a deformation field to align them. The loss function generally consists of an image similarity measure combined with a regularization term for the deformation field.

Recent studies continue to follow this similar flow. For example, H-ViT Ghahremani et al. (2024) introduces a top-down approach for estimating deformation fields by capturing multi-scale short- and long-range flow features, utilizing dual self- and cross-attention to enhance low-level features with high-level representations. CorrMLP Meng et al. (2024) presents the first correlation-aware MLP-based network for deformable medical image registration, improving efficiency and capturing long-range dependencies at full resolution without using self-attention. IIRP-Net Ma et al. (2024) develops a pyramid registration network that integrates a feature extractor with residual flow estimators to improve the generalization of feature extraction and registration. Despite the difference in architectures used to generate the registration field, all three methods share a common framework; a network for estimating deformation fields, with losses that include an image similarity measure and a regularization term. In addition to these recent methods, a summary of the taxonomy of representative approaches is provided in Table 1.

While modifying model architectures can lead to performance improvements, these approaches are often tailored to specific tasks or datasets, limiting their broader applicability. To overcome this

Method	CNN	U-Net	Encoder-Decoder	Transformer	MLP
plain		VoxelMorph, DALT, SVF-Net, AVSM, LKU-Net, AUM-Net, MIFR, MIDIR	ADRIR	TransMorph, ViT-V-Net	
dual / inverse	CycleMorph	ICNet, SYMNet		DTN	
iterative / recursive	IIRP-Net	VTN, RCN, VR-Net			
hierarchical / pyramid / coarse-to-fine	SDHNet, DLIR	Dual-PRNet	LapIRN, PMDR, Im2grid, NICE-Net	H-ViT, PIViT, Deformer, C2FViT, ModeT	CorrMLP
patch-wise	DIRNet		Quicksilver	TM-DCA, XMorpher	

Table 1: Taxonomy of design choices for deep learning based medical image registration methods.

limitation, we propose a novel approach that maintains the integrity of existing model structures. Rather than altering the architecture, we introduce a plug-in curriculum scheduler that integrates seamlessly with current methods. Our scheduler is inspired by curriculum learning, a training strategy that involves progressively increasing the difficulty of tasks. This approach has proven effective in various fields, which demonstrated how gradually evolving tasks and network architectures can benefit language processing models. We extend this concept to medical image registration, showing that it can significantly enhance performance on complex tasks.

Our approach incorporates two key criteria to guide the network's learning: sample difficulty and matching accuracy. Sample difficulty involves gradually selecting more challenging training examples, while matching accuracy focuses on ensuring precise alignment by strengthening supervision over time. Sample difficulty is assessed at two levels: voxel and volume. For voxel difficulty within an MRI volume, we use the Variance of Gradients (VoG) to prioritize more complex voxels for better alignment as training progresses. Evaluating difficulty across MRI volumes is more challenging, so we simplify the samples using Gaussian blurring, starting with highly blurred images and gradually shifting to sharper ones. In terms of matching accuracy, we progressively increase the strictness of supervision. This approach allows for more flexibility in the early stages, focusing on broader patterns, and then demands closer adherence to ground truths as training advances, ultimately improving accuracy.

Unlike most curriculum learning methods in medical image analysis, our approach offers an automated curriculum learning solution for medical image registration that does not require expert knowledge or prior experience. This scheduler enhances adaptability and flexibility, enabling improved performance across various tasks and datasets. By focusing on a plug-in curriculum scheduler rather than architectural changes, we aim to develop more robust and widely applicable registration techniques. In summary, our main contributions are as follows.

- We propose a novel approach that introduces a plug-in curriculum scheduler, allowing for seamless integration with existing model structures without modifying their architecture.
- In contrast to most curriculum learning methods in medical image analysis, our approach provides an automated curriculum learning solution for medical image registration that operates without the need for expert knowledge or prior experience.
- This scheduler improves adaptability and flexibility for better performance across diverse tasks and datasets, focusing on robust registration techniques without requiring architectural changes.

2 Related Works

2.1 Deformable Medical Image Registration

Traditional Algorithms. Traditional algorithms for image registration include popular models such as elastic registration Kybic (2001); Shen & Davatzikos (2002); Zhang et al. (2013); Heinrich et al. (2015), b-spline registration Jiang & Shackleford (2015); Sorzano et al. (2005); Zufeng et al. (2016); Modat et al. (2010), and viscous fluid-flow registration Bro-Nielsen & Gramkow (1996); D'agostino et al. (2003). These methods typically involve numerical optimization steps that iteratively adjust a predefined transformation space to find the optimal solution based on explainable metrics. Optical flow models Brox & Malik (2010); Chen et al. (2013); Ranjan & Black (2017) treat the moving and fixed images as continuous time samples of a sequence. To address challenges with large deformations and reverse fields, enhancements have been made through algorithms such as the Demons algorithm Thirion (1998) and diffeomorphism techniques Ashburner (2007); Avants et al. (2008); Dalca et al. (2018); Janssens et al. (2011); Krebs et al. (2019); Zhang (2018); Beg et al. (2005). These approaches provide smooth, continuous, and invertible velocity fields while preserving the topological structure of images.

Deep Learning Image Registration (DLIR) Algorithms. As shown in Tab. 1, DLIR algorithms can be categorized into five types based on their network architectures. Over the past decade, CNNs have been a primary focus of research in medical image registration, including approaches like CycleMorph Kim et al. (2021), IIRP-Net Ma et al. (2024), SDHNet Zhou et al. (2023a), DLIR de Vos et al. (2019), and DIRNet de Vos et al. (2017). U-Net architectures, along with encoder-decoder models, have become popular choices in DLIR due to their efficiency in capturing hierarchical features at multiple resolutions, which is crucial for accurately modeling complex image transformations. Notable examples include VoxelMorph Balakrishnan et al. (2019; 2018); Dalca et al. (2018), DALT Zhao et al. (2019a), SVF-Net Rohé et al. (2017), AVSM Shen et al. (2019), LKU-Net Jia et al. (2022), AUM-Net Xu et al. (2020), MIFR Shin & Lee (2023), MIDIR Qiu et al. (2021), IC-Net Zhang (2018), SYMNet Mok & Chung (2020b), VTN Zhao et al. (2020), RCN Zhao et al. (2019b), VR-Net Jia et al. (2021), and Dual-PRNet Kang et al. (2022) for U-Net, as well as ADRIR Hu et al. (2018), LapIRN Mok & Chung (2020a), PMDR Krebs et al. (2019), Im2grid Liu et al. (2022), NICE-Net Meng et al. (2022), and Quicksilver Yang et al. (2017) for encoder-decoder structures. With the advancement of Vision Transformers (ViT), transformers have also been applied in DLIR, with examples including TransMorph Chen et al. (2022a), ViT-V-Net Chen et al. (2021), DTN Zhang et al. (2021), H-ViT Ghahremani et al. (2024), PIViT Ma et al. (2023), Deformer Chen et al. (2022b), C2FViT Mok & Chung (2022), ModeT Wang et al. (2023), TM-DCA Chen et al. (2023), and XMorpher Shi et al. (2022). Their significantly larger receptive fields enable a more accurate understanding of spatial relationships between images. CorrMLP Meng et al. (2024) introduces the first MLP-based architecture for deformable medical image registration, overcoming transformers' limitations in capturing fine-grained long-range dependencies at full resolution due to high computational costs.

For each architecture, additional procedures can be integrated. Dual and inverse mechanisms are employed to ensure that two images deform symmetrically towards each other. Iterative and recursive methods are used to align images with significant displacements. Hierarchical, pyramid, and coarse-to-fine structures enhance the deformation field by leveraging high-resolution feature maps. Lastly, patch-wise techniques selectively sample a diverse range of features across a large search area while keeping computational overhead low.

2.2 Curriculum Learning (CL)

Imposing curriculum in neural networks can be traced back to Elman (1993). Inspired by the manners of how humans learn languages, Elman (1993) points out the importance of starting easily and gradually hardening the learning process when training networks. Bengio et al. (2009) further extends the idea to various vision and language tasks - where multi-stage curriculum strategies give rise to improved generalization and faster convergence. Named "curriculum learning" for these strategies, they are employed to different algorithms such as image classification Wang et al. (2019); Wei et al. (2021), object detection Zhang et al. (2019; 2017a), semantic segmentation Zhang et al. (2017b), self or semi supervised learning Murali et al. (2018), multi-task learning Sarafianos et al. (2017; 2018); Dong et al. (2017); Wang et al. (2018), multi-modal learning Gong et al. (2016); Gong (2017), etc Jiang et al. (2015); Matiisen et al. (2019); Weinshall et al. (2018). In the medical field, there has been limited prior research on curriculum learning (CL), with some examples in classification Jiménez-Sánchez et al. (2019); Luo et al. (2021), semantic segmentation Kervadec et al. (2019), self-supervised or semi-supervised learning Tang et al. (2018), and multi-modal learning Lotter et al. (2017). However, only a few studies leverage external knowledge from human experts, and even fewer integrate CL with image registration Burduja & Ionescu (2021); Zhou et al. (2023b). Our work introduces an automated CL approach for medical image registration that operates without the need of expert knowledge or prior experience.

3 PRELIMINARY: MEDICAL IMAGE REGISTRATION

Medical image registration, also known as image alignment, involves aligning two or more anatomically related images according to their spatial features. This process establishes non-linear dense correspondences between n-D medical images that are acquired from different patients, scanners, or at different times. It has been extensively studied due to its importance in clinical applications, such as monitoring tumor growth or conducting group analysis.

Deformable image registration is commonly framed as an optimization problem, aiming to minimize an energy function. This function generally consists of two main components: a penalty function that evaluates the similarity between the aligned and reference images, and a regularization term that enforces constraints on the registration field, such as promoting smoothness through a gradient loss penalty. In this process, the *fixed image* (or target image) serves as the baseline or template to which the *moving image* (or source image) is aligned. The fixed image provides the spatial coordinates used for alignment, while the moving image is adjusted to match the fixed image as closely as possible. The goal is to determine the optimal transformation that aligns the moving image with the fixed image. Below is a comprehensive explanation and formulation of a widely used framework for medical image registration.

The moving image $\hat{I_m}$ and the fixed image I_f are initially transformed into a common coordinate system using an affine transformation. The affine-aligned moving image is referred to as I_m. Next, I_m is deformed to align with I_f using a deformation field ϕ, which is generated by a specific network f_θ as

$$f_\theta(I_f, I_m) = \phi \tag{1}$$

where θ is the parameters of the network. The overall loss function used for training the network is based on the energy function from traditional image registration techniques. This loss function comprises two components: one evaluates the similarity between the deformed image and the fixed image, while the other regularizes the deformation field to ensure smoothness. It is expressed as:

$$\mathcal{L}(I_f, I_m, \phi) = \mathcal{L}_{sim}(I_f, I_m, \phi) + \lambda \mathcal{R}(\phi), \tag{2}$$

where $\mathcal{L}_{sim}$ denotes the image fidelity measure, and $\mathcal{R}$ is the regularization term for the deformation field. A common metric for evaluating the similarity between I_f and I_m is the local normalized cross-correlation:

$$LNCC(I_f, I_m, \phi) = \sum_{\mathbf{p}\in\Omega} \frac{\left(\sum_{\mathbf{p}_i}(I_f(\mathbf{p}_i) - \overline{I}_f(\mathbf{p}))([I_m \circ \phi](\mathbf{p}_i) - [\overline{I}_m \circ \phi](\mathbf{p}))\right)^2}{\left(\sum_{\mathbf{p}_i}(I_f(\mathbf{p}_i) - \overline{I}_f(\mathbf{p}))^2\right)\left(\sum_{\mathbf{p}_i}([I_m \circ \phi](\mathbf{p}_i) - [\overline{I}_m \circ \phi](\mathbf{p}))^2\right)}, \tag{3}$$

where $\overline{I}_f(\mathbf{p})$ and $\overline{I}_m(\mathbf{p})$ denote the mean value within the local window of size n^3 centered at voxel $\mathbf{p}$. A higher $LNCC$ indicates a better alignment, yielding the loss function: $\mathcal{L}_{sim}(I_f, I_m, \phi) = -LNCC(I_f, I_m, \phi)$. The regularizer $\mathcal{R}$ promotes similarity in displacement values between a given location and its neighboring locations. A commonly used diffusion regularizer can be formulated as:

$$\mathcal{R}_{diffusion}(\phi) = \sum_{\mathbf{p}\in\Omega} \parallel \nabla u(\mathbf{p}) \parallel^2, \tag{4}$$

where ∇u is the spatial gradients of the displacement field u. The spatial gradients are approximated using forward differences, that is, $\frac{\partial u(\mathbf{p})}{\partial\{x,y,z\}} \approx u(\mathbf{p}_{\{x,y,z\}} + 1) - u(\mathbf{p}_{\{x,y,z\}})$.

4 METHOD

Research in deformable medical image registration has primarily focused on incremental performance improvements through minor changes to model architectures or the addition of new loss functions. While these adjustments can improve performance metrics, they often result in methods that are task- or dataset-specific, limiting their broader applicability. To overcome these limitations, we propose a novel approach that avoids modifying existing model architectures. Instead, we introduce a plug-in curriculum scheduler that integrates seamlessly with current methods. This scheduler, inspired by curriculum learning, organizes the learning process by increasing task difficulty based on two main criteria: sample difficulty and accuracy tolerance. For sample difficulty, we employ two strategies: weighting challenging voxel-level samples within the brain and starting with blurred images before transitioning to sharper images as training progresses. The second criterion focuses on the rigor of supervision by utilizing ground truth data for more comprehensive training. This approach aims to improve the generalizability and flexibility of registration methods, making them more robust and widely applicable.

4.1 SAMPLE DIFFICULTY

Curriculum learning involves training models in a structured sequence, beginning with simpler examples and progressively advancing to more complex ones. A key challenge in this approach is developing automatic and objective metrics to assess the difficulty of each sample. However, in the medical domain, determining the difficulty of individual samples is not straightforward. For instance, identifying which specific MRI volume would be easier for learning the deformation field in registration tasks is not intuitive. Previous research has utilized domain knowledge from human experts to qualitatively assess the classification difficulty of medical images to guide curriculum learning. This approach, however, requires additional annotation efforts, depends on subjective human experience, and introduces potential bias. Instead of relying on this computationally intensive approach, we propose an automated method for assessing sample difficulties at two levels: voxel-level and volume-level.

4.1.1 VOXEL-LEVEL SAMPLE DIFFICULTY

In medical image registration tasks, each voxel in a 3D volume contributes to the overall alignment between the moving and fixed images. In standard registration methods, all voxels within an MRI volume are assigned identical weights during training, based on the assumption that each voxel has equal significance. However, not all voxels are equally crucial for registration; for instance, areas like edges or regions with high contrast tend to be more difficult to align, while homogeneous areas are often easier. To address this, we propose a method to evaluate the varying difficulty in learning the displacement field between voxels. Based on this measure, we design a training schedule that gradually prioritizes voxels of different difficulty levels as training progresses.

Difficulty Measure. For a given MRI image I, it can be decomposed into a set of voxels x_i, where $i = \{1, ..., N\}$ and N represents the total number of voxels in the image. The difficulty of these voxels is then evaluated using the Variance of Gradients (VoG) Agarwal et al. (2022) method. The VoG method is applied to the deformation fields in the following order.

The gradient calculation for each voxel involves determining how sensitive the deformation field is to changes in the input voxel values. Specifically, for each voxel in the input image, the gradient of the deformation field (displacement vector) is computed with respect to the voxel itself. This is expressed as $S_i = \frac{\partial D(x_i)}{\partial x_i}$, where $D(x_i)$ is the deformation vector at voxel x_i and S_i denotes the sensitivity of the deformation field to variations in that voxel. The gradient matrices for the deformation field are computed at different epochs or iterations during training, forming a series of checkpoints, denoted as $\{S_{i,1}, S_{i,2}, ..., S_{i,K}\}$, where $S_{i,t}$ represents the gradient matrix at checkpoint t. This process enables the monitoring of how the deformation field's sensitivity to voxel changes evolves over time. The average of the gradient matrices across all checkpoints for each voxel x_i is then calculated as

$$\mu_i = \frac{1}{K} \sum_{t=1}^{K} S_{i,t}. \tag{5}$$

This mean represents the average sensitivity of the deformation field with respect to each voxel over time. Finally, the Variance of Gradients (VoG) is calculated across the checkpoints as:

$$VoG_i = \sqrt{\frac{1}{K} \sum_{t=1}^{K} (S_{i,t} - \mu)^2}. \tag{6}$$

This variance provides a measure of how much the deformation field's sensitivity to each voxel varies over time. A higher variance indicates that the voxel is more difficult to register, while a lower variance suggests that it is easier to register.

Training Scheduler. Once the difficulty of each voxel is determined through VoG, the loss function can be modified by multiplying it with calculated weights for each voxel: $L = \sum_i w_i \cdot \mathrm{Loss}(x_i)$, adjusting the loss contribution during backpropagation.

To align with the goal of curriculum learning, which trains models by starting with simpler examples and gradually moving to more complex ones, the weights w_i are scheduled to update at each epoch. This allows the model to adjust to new weights progressively as training proceeds. Each voxel x_i is associated with an initial weight $w_{i,(1)}$ given by:

$$w_{i,(1)} = \frac{s_i}{\sum_{j=1}^{N} s_j} \tag{7}$$

where s_i represents the rank of the voxel x_i in the sorted VoG scores from highest to lowest, and N is the total number of voxels in the MRI image. In this initial probability assignment, easier samples, which have lower VoG scores, are given higher probabilities. To reflect the changing importance of each voxel over training epochs, the weight for voxel x_i at epoch e is updated as:

$$w_{i,(e)} = w_{i,(e-1)} \times \lambda_i, \qquad \lambda_i = \sqrt[L]{\frac{1/N}{w_{i,(1)}}}. \tag{8}$$

Here, L is the number of epochs over which the scheduling will last. By the final epoch, the probability for every voxel x_i is set to:

$$w_{i,(final)} = 1/N. \tag{9}$$

This training scheduler ensures that the probability distribution over voxels smooths out over time, with each voxel eventually being assigned an equal probability by the end of the scheduling.

4.1.2 VOLUME-LEVEL SAMPLE DIFFICULTY

In the previous analysis, we identified variations in difficulty within a single MRI volume, leading to different contributions for individual voxels. On a broader scale, we aim to assess the difficulty across different volume samples.

When working with a set of MRI volumes, determining which volume poses a greater challenge in learning the deformation field is difficult. In situations where directly assessing the difficulty of data samples is not possible, a curriculum learning approach can be employed by deliberately simplifying the inputs. This is achieved by blurring the images with a Gaussian filter, reducing their complexity and information content Burduja & Ionescu (2021). Training begins with highly blurred images, gradually transitioning to sharper images as the process advances.

Specifically, at a given training epoch e, the degree of blur σ is adjusted according to the following rule:

$$\sigma = \begin{cases} \sigma_{max} \times (1 - e/e_{sch}) & \text{if } e < e_{sch} \\ 0 & \text{if } e > e_{sch} \end{cases} \tag{10}$$

where σ_{max} represents the initial level of blur at the start of training, and e_{sch} is the final epoch during which the blur is applied.

4.2 ACCURACY TOLERANCE

In the context of model accuracy, curriculum learning generally involves presenting training data in a structured manner, starting from easier examples and gradually moving to more difficult ones to aid the model's learning process. An additional dimension to this approach is adjusting the level of strictness in supervision using ground truth data. This method is particularly effective in tasks that require mastering intricate details, such as medical imaging, where accurate alignment or segmentation is critical and develops over time.

In this approach, the level of strictness in evaluating the model against the ground truths is progressively increased. Early in training, the model is allowed more flexibility in its predictions (i.e., accepting a wider range of outputs as correct), focusing on general patterns. As the model progresses, the supervision becomes more stringent, requiring closer alignment with the exact ground truth (e.g., penalizing even slight deviations from the correct output). This strategy helps the model avoid being overwhelmed by complex or highly detailed data in the early stages, enabling it to build a foundational understanding before being challenged to achieve higher accuracy levels.

At a given training epoch e with a tolerance parameter ϵ_e, the image similarity loss is defined as:

$$\mathcal{L}_{\epsilon}(\mathcal{L}_{sim}(I_f, I_m, \phi)) = \max\{\mathcal{L}_{sim}(I_f, I_m, \phi) + \epsilon, 0\} \tag{11}$$

when a higher similarity value corresponds to better alignment (e.g., for metrics like $LNCC$ or $SSIM$). The value of ϵ_e increases progressively as training continues. In contrast, when a lower similarity measure indicates better alignment (e.g., for metrics like MSE), the image similarity loss is given by:

$$\mathcal{L}_{\epsilon}(\mathcal{L}_{sim}(I_f, I_m, \phi)) = \max\{\mathcal{L}_{sim}(I_f, I_m, \phi) - \epsilon, 0\} \tag{12}$$

with the value of ϵ_e gradually decreasing as training progresses.

5 Experimental Setup

5.1 Datasets and Evaluation Metrics

Our method is employed in the analysis of two popular MRI databases, including IXI [1] and Mind-Boggle [2]. Detailed information regarding the datasets and their preparation procedures are presented in Section A.1 of the supplementary material. For quantitative comparisons on the datasets, Dice similarity coefficients (DSC) Dice (1945) and non-positive values in the Jacobian determinant (NJD) are computed. Specifically, DSC are measured between the segmentation labels of the fixed and moved images. The smoothness and invertibility of the predicted displacement fields are evaluated by determining the percentage of NJD of the deformation fields (i.e., $\%$ of $|J_\phi| \leq 0$). Adjusting the regularization parameter often involves a trade-off between registration accuracy and transformation smoothness Mok & Chung (2020b); Meng et al. (2022; 2024). Hence, registration methods should be evaluated using both DSC and NJD metrics. Further details and formulations can be found in Section A.2 of the supplementary material.

5.2 Comparison Methods

Our method is extensively evaluated against state-of-the-art deformable image registration techniques, encompassing four traditional optimization-based methods (SyN Avants et al. (2008), NiftyReg Modat et al. (2010), LDDMM Beg et al. (2005), deedsBCV Heinrich et al. (2015)) and nine deep learning-based registration methods (VoxelMorph Balakrishnan et al. (2019), CycleMorph Kim et al. (2021), MIDIR Qiu et al. (2021), ViT-V-Net Chen et al. (2021), PVT Wang et al. (2021), CoTr Xie et al. (2021), nnFormer Zhou et al. (2021), TransMorph Chen et al. (2022a), H-ViT Ghahremani et al. (2024)). The methods and their hyperparameter settings are described in Section A.3 and A.4 of the supplementary material.

5.3 Implementation Details

Our method builds upon two state-of-the-art deformable image registration models: TransMorph and H-ViT. These designs are referred to as TM+ and HV+, respectively. Both models were trained for 500 epochs on an NVIDIA A6000 GPU, using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 1. Hyperparameter selections are specified in the ablation study results. Section A.4 provides more details about the experiment settings.

6 Results and Discussion

6.1 Comparison with Baseline Methods

We conducted inter-patient and atlas-to-patient registration experiments on both the IXI and Mind-Boggle datasets. Table 2 presents the registration metrics for IXI, comparing our method against several baselines using DSC and NJD. Our approach achieves the highest scores, including a DSC

[1] `https://brain-development.org/ixi-dataset/`
[2] `https://mindboggle.info/`

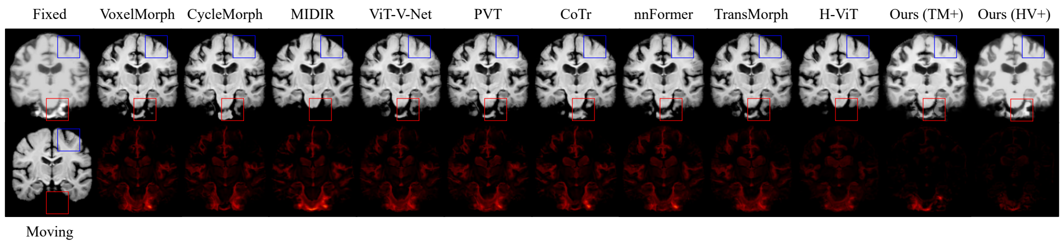

Figure 1: Example coronal slices from the IXI dataset and outcomes (moved MRI images and difference between fixed image and moved results) of baseline registration methods, compared with our methods. Red highlights indicate misalignments between the ground truth and the aligned results, while black highlights signify successful registration, with fewer red pixels indicating better performance.

Method	Inter-patient Registration		Atlas-to-patient Registration	
	DSC ↑	NJD (%) ↓	DSC ↑	NJD (%) ↓
SyN	0.639 ± 0.197	< 0.001	0.590 ± 0.210	< 0.001
NiftyReg	0.626 ± 0.183	0.068 ± 0.085	0.587 ± 0.223	0.020 ± 0.046
LDDMM	0.730 ± 0.134	< 0.001	0.638 ± 0.186	< 0.001
deedsBCV	0.717 ± 0.180	0.188 ± 0.059	0.706 ± 0.151	0.147 ± 0.050
VoxelMorph	0.720 ± 0.139	0.799 ± 0.103	0.695 ± 0.162	1.586 ± 0.339
CycleMorph	0.704 ± 0.167	0.651 ± 0.197	0.706 ± 0.155	1.719 ± 0.382
MIDIR	0.721 ± 0.156	0.151 ± 0.069	0.711 ± 0.158	< 0.001
ViT-V-Net	0.736 ± 0.128	0.999 ± 0.201	0.702 ± 0.155	1.609 ± 0.319
PVT	0.733 ± 0.117	1.314 ± 0.600	0.695 ± 0.159	1.858 ± 0.314
CoTr	0.741 ± 0.132	0.719 ± 0.269	0.706 ± 0.164	1.298 ± 0.343
nnFormer	0.744 ± 0.130	0.800 ± 0.283	0.719 ± 0.157	1.595 ± 0.358
TransMorph	0.763 ± 0.119	0.617 ± 0.210	0.724 ± 0.150	1.502 ± 0.342
H-ViT	0.779 ± 0.078	0.589 ± 0.182	0.740 ± 0.139	0.707 ± 0.185
Ours (TM+)	0.808 ± 0.063	0.493 ± 0.101	0.755 ± 0.107	0.401 ± 0.090
Ours (HV+)	0.812 ± 0.060	0.285 ± 0.071	0.772 ± 0.091	< 0.001

Table 2: Quantitative evaluation results for the registration methods on the IXI dataset for 34 anatomical structures over 115 random pairs for inter-patient and 115 pairs for atlas-to-patient registrations.

performance improvement of $+0.045$ and $+0.033$ in inter-patient registration compared to TransMorph and H-ViT, respectively. While the differences between the baselines are relatively small, our method shows a significantly larger margin, yielding meaningful results. Figure 1 presents the visualized registration results, highlighting the differences between the fixed image and the transformed outputs as well. Our method shows more precise warping of the moving MRI compared to other approaches, especially in the areas marked by the red and blue rectangles. The registration results for the MindBoggle dataset are shown in Table 3, where our method achieves an increase in DSC values by $+0.056$ and $+0.070$ compared to TransMorph and H-ViT, respectively, in atlas-to-patient registration experiments, representing a substantial performance improvement. Detailed results, including DSC values for each anatomical structure, are provided in the Supplementary material in Section B.

6.2 Visualization results by Epochs

In Figure 2, the progression of the moved image at key training epochs is illustrated. Our curriculum learning-based approach initially focuses on learning coarse spatial transformations during the early stages of training, deliberately omitting intricate details to prioritize broader structural changes. By starting with simpler, larger-scale adjustments, the model establishes a solid foundation for subsequent refinement. As training progresses, the model gradually hones in on finer, localized regions, refining these areas based on the broader shapes learned in the earlier stages. This phased approach allows the network to incrementally build upon its initial understanding, ultimately leading to more accurate and precise registration results. By progressively tackling the complexity of the transformation, this training strategy significantly enhances the overall performance of the registration process, resulting in improved alignment and greater detail capture in the final output.

Method	Inter-patient Registration		Atlas-to-patient Registration	
	DSC ↑	NJD (%) ↓	DSC ↑	NJD (%) ↓
VoxelMorph	0.674 ± 0.197	0.821 ± 0.170	0.666 ± 0.201	0.831 ± 0.163
CycleMorph	0.679 ± 0.194	1.044 ± 0.211	0.671 ± 0.199	1.064 ± 0.189
MIDIR	0.637 ± 0.197	0.403 ± 0.215	0.539 ± 0.292	0.347 ± 0.205
ViT-V-Net	0.700 ± 0.186	1.168 ± 0.225	0.695 ± 0.187	0.840 ± 0.573
PVT	0.588 ± 0.214	2.006 ± 0.254	0.583 ± 0.216	2.034 ± 0.217
CoTr	0.633 ± 0.214	0.691 ± 0.163	0.630 ± 0.218	0.701 ± 0.141
nnFormer	0.622 ± 0.210	1.077 ± 0.210	0.618 ± 0.213	1.090 ± 0.189
TransMorph	0.699 ± 0.186	0.702 ± 0.106	0.695 ± 0.189	0.716 ± 0.082
H-ViT	0.731 ± 0.170	0.328 ± 0.061	0.726 ± 0.173	0.335 ± 0.049
Ours (TM+)	0.749 ± 0.126	0.209 ± 0.039	0.751 ± 0.086	0.199 ± 0.020
Ours (HV+)	0.781 ± 0.099	< 0.001	0.796 ± 0.062	0.128 ± 0.011

Table 3: Quantitative evaluation results for the registration methods on the MindBoggle dataset for 41 anatomical structures over 114 random pairs for inter-patient and 222 pairs for atlas-to-patient registrations.

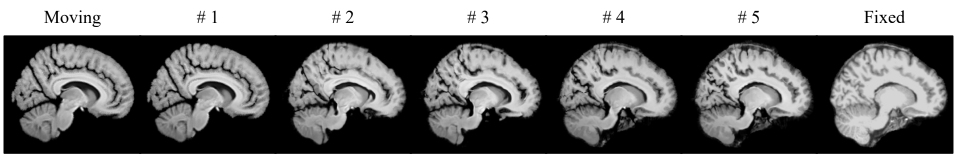

Figure 2: The visualization of moved MRI images at the end of each training stage.

6.3 Ablation Study

6.3.1 Components of Curriculum Learning

We perform an ablation study on the components of curriculum learning: voxel-level sample difficulty (VLSD), volume-level sample difficulty (VMSD), and accuracy tolerance (AT), with the resulting metrics shown in Table 4. Incorporating all three components in the learning schedule yields the best performance.

Components	VLSD	VMSD	AT	VLSD, VMSD	VLSD, AT	VMSD, AT	Full
DSC ↑	0.754 ±0.110	0.745 ±0.132	0.749 ±0.124	0.761 ±0.105	0.767 ±0.099	0.758 ±0.127	**0.772 ±0.091**

Table 4: Ablation study on the curriculum learning components for the IXI registration.

6.3.2 Parameters of voxel-level sample difficulty

Sample difficulties. In the curriculum for voxel-level sample difficulty, we update the weights of voxels at each epoch to encourage the network to increasingly focus on challenging samples. We contrast this with a simpler approach that maintains a fixed difficulty level for each voxel, utilizing only the difficulty measure and omitting the training scheduler. Table 5 presents the DSC measures for various fixed difficulty values. Our method demonstrates superior performance by progressively increasing the probabilities for difficult samples, allowing the network to concentrate more on challenging details.

Scheduling epochs. An ablation study is conducted on different scheduling epochs for voxel-level sample difficulty. Here, e_{start} indicates the epoch at which scheduling begins; thus, the VoG values for individual voxels are computed and stacked for the difficulty measure during the epochs prior to e_{start}. Meanwhile, e_{end} signifies the epoch when scheduling concludes. Consequently, the difficulty weights of voxels are updated between e_{start} and e_{end}, after which equal probabilities are assigned to all voxels. Note that $e_{start} = K$ and $e_{end} = K + L$, with K and L detailed in Section 4.1.1. Table 6 shows the DSC score of methods trained with varying scheduling epochs, showing that setting e_{start} to 100 epochs and e_{end} to 300 epochs yields the best DSC measures.

w_i	$w_{i,(1)}$	$w_{i,(final/4)}$	$w_{i,(final/2)}$	$w_{i,(final)}$	Ours
DSC ↑	0.750 ± 0.135	0.767 ± 0.105	0.764 ± 0.119	0.758 ± 0.127	$\mathbf{0.772 \pm 0.091}$

Table 5: Ablation study on methods with fixed difficulties of voxel-level sample difficulty for the IXI registration.

e_{start}, e_{end}	100, 200	100, 300	200, 300	200, 400
DSC ↑	0.765 ± 0.097	$\mathbf{0.772 \pm 0.091}$	0.767 ± 0.102	0.760 ± 0.099

Table 6: Ablation study on the scheduling epochs of voxel-level sample difficulty for the IXI registration.

6.3.3 Parameters of volume-level sample difficulty

The effects of varying degrees of initial blur and scheduling epochs are presented in Table 7. The ablation study on initial blur was conducted using a fixed scheduling epoch of $e_{sch} = 300$, while the ablation study on scheduling epochs was carried out with a fixed initial blur value of $\sigma_{max} = 1.0$. Our optimal method is configured with hyperparameters of $e_{sch} = 300$ and $\sigma_{max} = 1.0$.

σ_{max}	0.5	0.75	1.0	1.5	2.0
DSC ↑	0.760 ± 0.119	0.769 ± 0.098	$\mathbf{0.772 \pm 0.091}$	0.767 ± 0.101	0.757 ± 0.125
e_{sch}	200	250	300	350	400
DSC ↑	0.765 ± 0.120	0.768 ± 0.115	$\mathbf{0.772 \pm 0.091}$	0.763 ± 0.116	0.759 ± 0.120

Table 7: Ablation study on the initial blur and the scheduling epoch of volume-level sample difficulty for the IXI registration.

6.3.4 Parameters of accuracy tolerance

The DSC measurements for various methods trained with different ϵ values are presented in Table 8. The chosen ϵ values were determined by analyzing the model's training curve without any tolerances. Our optimal method is configured with the following ϵ values: 0.24, 0.26, 0.28, 0.3, 0.31, 0.32, 0.34, and 0.38 for epochs under 3, 5, 30, 100, 200, 300, 400, and 500, respectively. It is important to note that these values increase as training progresses, as we utilize the image similarity metric $LNCC$.

	ϵ (epochs)	DSC ↑
# 1	$0.24(<3), 0.26(<5), 0.28(<30), 0.3(<100),$ $0.31(<200), 0.32(<300), 0.34(<400), 0.38(<500)$	$\mathbf{0.772 \pm 0.091}$
# 2	$0.24(<5), 0.26(<30), 0.29(<100), 0.3(<150),$ $0.305(<200), 0.31(<300), 0.45(<500)$	0.754 ± 0.141
# 3	$0.24(<5), 0.26(<30), 0.29(<100), 0.3(<150),$ $0.305(<200), 0.31(<270), 0.32(<400), 0.5(<500)$	0.748 ± 0.160

Table 8: Ablation study on the accuracy tolerance scheduling for the IXI registration.

7 Conclusion

We introduced an innovative approach that marks a notable advancement in deformable medical image registration through the development of a plug-in scheduler inspired by curriculum learning. This technique improves the adaptability and flexibility of existing network architectures without the need for substantial alterations, enabling broader applicability across diverse tasks and datasets. By concentrating on the dual aspects of sample difficulty and matching accuracy, we effectively steer the network's learning process towards achieving accurate image alignment. Our distinctive application of the Variance of Gradients (VoG) for assessing voxel difficulty and the gradual shift from blurred to clearer images for volume difficulty sets our work apart from conventional methods. In contrast to earlier curriculum learning techniques in medical imaging, our automated solution removes the necessity for expert knowledge, making it accessible to a wider audience.

REFERENCES

C. Agarwal, D. D'souza, and S. Hooker. Estimating example difficulty using variance of gradients. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10368–10378, 2022.

J. Ashburner. A fast diffeomorphic image registration algorithm. NeuroImage, 38(1):95–113, 2007.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis, 12(1):26–41, 2008.

G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca. An unsupervised learning model for deformable medical image registration. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 9252–9260, 2018.

G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. IEEE Trans. on Medical Imaging (TMI), 38(8):1788–1800, 2019.

M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis., 61(2):139–157, 2005.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In Proc. Int. Conf. on Machine Learning (ICML), pp. 41–48, 2009.

M. Bro-Nielsen and C. Gramkow. Fast fluid registration of medical images. In Proc. Int. Conf. on Visualization in Biomedical Computing, pp. 265–276, 1996.

T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(3):500–513, 2010.

M. Burduja and R. Ionescu. Unsupervised medical image alignment with curriculum learning. In IEEE Int. Conf. on Image Processing (ICIP), pp. 3787–3791, 2021.

J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. CoRR, abs/2104.06468, 2021.

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, 2022a.

J. Chen, D. Lu, Y. Zhang, D. Wei, M. Ning, X. Shi, Z. Xu, and Y. Zheng. Deformer: Towards displacement field learning for unsupervised medical image registration. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 141–151, 2022b.

J. Chen, Y. Liu, Y. He, and Y. Du. Deformable cross-attention transformer for medical image registration. In International Workshop on Machine Learning in Medical Imaging, pp. 115–125, 2023.

Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu. Large displacement optical flow from nearest neighbor fields. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2443–2450, 2013.

E. D'agostino, F. Maes, D. Vandermeulen, and P. Suetens. A viscous fluid model for multimodal non-rigid image registration using mutual information. Medical Image Analysis, 7(4):565–575, 2003.

A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 729–738, 2018.

B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 204–212. 2017.

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.

L. R. Dice. Measures of the amount of ecologic association between species. Ecology, 26(3): 297–302, 1945.

Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In Proc. Winter Conf. on Applications of Computer Vision (WACV), pp. 520–529, 2017.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proc. Int. Conf. on Learning Representations (ICLR), 2021.

J. L. Elman. Learning and development in neural networks: The importance of starting small. Cognition, pp. 71–99, 1993.

M. Ghahremani, M. Khateri, B. Jian, B. Wiestler, E. Adeli, and C. Wachinger. H-vit: A hierarchical vision transformer for deformable image registration. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 11513–11523, 2024.

C. Gong. Exploring commonality and individuality for multi-modal curriculum learning. In Proc. AAAI Conf. on Artificial Intelligence (AAAI), volume 31, 2017.

C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang. Multi-modal curriculum learning for semi-supervised image classification. IEEE Trans. on Image Processing (TIP), 25(7):3249–3260, 2016.

M. P. Heinrich, O. Maier, and H. Handels. Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. In Proc. of the VISCERAL Anatomy3 Organ Segmentation Challenge co-located with IEEE Int. Symposium on Biomedical Imaging (ISBI), volume 1390, pp. 27–30, 2015.

M. Hoffmann, B. Billot, J. E. Iglesias, B. R. Fischl, and A. V. Dalca. Learning image registration without images. arXiv: Computer Vision and Pattern Recognition, 2020.

Y. Hu, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, T. Vercauteren, J. A. Noble, and D. C. Barratt. Adversarial deformation regularization for training image registration neural networks. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 774–782, 2018.

G. Janssens, L. Jacques, J. O. de Xivry, X. Geets, and B. Macq. Diffeomorphic registration of images with variable contrast enhancement. Int. J. of Biomedical Imaging, 2011:891585:1–891585:16, 2011.

X. Jia, A. Thorley, W. Chen, H. Qiu, L. Shen, I. B. Styles, H. Chang, A. Leonardis, A. De Marvao, D. P. O'Regan, et al. Learning a model-driven variational network for deformable image registration. IEEE Trans. on Medical Imaging (TMI), 41(1):199–212, 2021.

X. Jia, J. Bartlett, T. Zhang, W. Lu, Z. Qiu, and J. Duan. U-net vs transformer: Is u-net outdated in medical image registration? In International Workshop on Machine Learning in Medical Imaging, pp. 151–160, 2022.

L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann. Self-paced curriculum learning. In Proc. AAAI Conf. on Artificial Intelligence (AAAI), volume 29, 2015.

P. Jiang and J. A. Shackleford. B-spline registration of neuroimaging modalites with map-reduce framework. In Proc. Int. Conf. on Brain Informatics and Health (BIH), pp. 285–294, 2015.

A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. A. González Ballester, and G. Piella. Medical-based deep curriculum learning for improved fracture classification. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 694–702, 2019.

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

H. Kervadec, J. Dolz, É. Granger, and I. Ben Ayed. Curriculum semi-supervised segmentation. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 568–576, 2019.

B. Kim, D. Kim, S. Park, J. Kim, J. Lee, and J. Ye. Cyclemorph: Cycle consistent unsupervised deformable image registration. Medical Image Analysis, 71:102036, 2021.

J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. Learning a probabilistic model for diffeomorphic registration. IEEE Trans. on Medical Imaging (TMI), 38(9):2165–2176, 2019.

J. Kybic. Elastic image registration using parametric deformation models. Technical report, EPFL, 2001.

Y. Liu, L. Zuo, S. Han, Y. Xue, J. L. Prince, and A. Carass. Coordinate translator for learning deformable medical image registration. In International workshop on multiscale multimodal medical imaging, pp. 98–109, 2022.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proc. Int. Conf. on Computer Vision (ICCV), pp. 9992–10002, 2021.

W. Lotter, G. Sorensen, and D. Cox. A multi-scale cnn and curriculum learning strategy for mammogram classification. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA, pp. 169–177, 2017.

J. Luo, G. Kitamura, E. Doganay, D. Arefan, and S. Wu. Medical knowledge-guided deep curriculum learning for elbow fracture diagnosis from x-ray images. In Medical Imaging 2021: Computer-Aided Diagnosis, volume 11597, pp. 247–252, 2021.

T. Ma, X. Dai, S. Zhang, and Y. Wen. Pivit: Large deformation image registration with pyramid-iterative vision transformer. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 602–612, 2023.

T. Ma, S. Zhang, J. Li, and Y. Wen. Iirp-net: Iterative inference residual pyramid network for enhanced image registration. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 11546–11555, 2024.

T. Matiisen, A. Oliver, T. Cohen, and J. Schulman. Teacher–student curriculum learning. IEEE transactions on neural networks and learning systems, 31(9):3732–3740, 2019.

M. Meng, L. Bi, D. Feng, and J. Kim. Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 88–97, 2022.

M. Meng, D. Feng, L. Bi, and J. Kim. Correlation-aware coarse-to-fine mlps for deformable medical image registration. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 9645–9654, 2024.

M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed., 98(3):278–284, 2010.

T. C. W. Mok and A. C. S. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 12263 of Lecture Notes in Computer Science, pp. 211–221, 2020a.

T. C. W. Mok and A. C. S. Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4643–4652, 2020b.

T. CW Mok and A. Chung. Affine medical image registration with coarse-to-fine vision transformer. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 20835–20844, 2022.

A. Murali, L. Pinto, D. Gandhi, and A. Gupta. Cassl: Curriculum accelerated self-supervised learning. In Int. Conf. on Robotics and Automation (ICRA), pp. 6453–6460, 2018.

H. Qiu, C. Qin, A. Schuh, K. Hammernik, and D. Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. In Medical Imaging with Deep Learning (MIDL), volume 143, pp. 645–664, 2021.

A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2720–2729, 2017.

M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec. Svf-net: learning deformable image registration using shape matching. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 266–274, 2017.

N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris. Curriculum learning for multi-task classification of visual attributes. In Proc. of the IEEE Int. Conf. on Computer Vision Workshops (ICCVW), pp. 2608–2615, 2017.

N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris. Curriculum learning of visual attribute clusters for multi-task classification. Pattern Recognition, 80:94–108, 2018.

D. Shen and C. Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. on Medical Imaging (TMI), 21(11):1421–1439, 2002.

Z. Shen, X. Han, Z. Xu, and M. Niethammer. Networks for joint affine and non-parametric image registration. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4224–4233, 2019.

J. Shi, Y. He, Y. Kong, J. Coatrieux, H. Shu, G. Yang, and S. Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 217–226, 2022.

J. Shin and J. Lee. Mri imputation based on fused index-and intensity-registration. In Proc. Winter Conf. on Applications of Computer Vision (WACV), pp. 1949–1958, 2023.

C. O. S. Sorzano, P. Thévenaz, and M. Unser. Elastic registration of biological images using vector-spline regularization. IEEE Trans. on Biomedical Engineering (TBME), 52(4):652–663, 2005.

Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In Machine Learning in Medical Imaging: 9th International Workshop, MLMI, pp. 249–258, 2018.

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

C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proc. European Conf. on Computer Vision (ECCV), pp. 365–381, 2018.

H. Wang, D. Ni, and Y. Wang. Modet: Learning deformable image registration via motion decomposition transformer. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 740–749, 2023.

W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proc. Int. Conf. on Computer Vision (ICCV), pp. 548–558, 2021.

Y. Wang, W. Gan, J. Yang, W. Wu, and J. Yan. Dynamic curriculum learning for imbalanced data classification. In Proc. Int. Conf. on Computer Vision (ICCV), pp. 5017–5026, 2019.

J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, M. Nasir-Moin, N. Tomita, et al. Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification. In Proc. Winter Conf. on Applications of Computer Vision (WACV), pp. 2473–2483, 2021.

D. Weinshall, G. Cohen, and D. Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In Proc. Int. Conf. on Machine Learning (ICML), pp. 5238–5246, 2018.

Y. Xie, J. Zhang, C. Shen, and Y. Xia. Cotr: Efficiently bridging CNN and transformer for 3d medical image segmentation. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 12903, pp. 171–180, 2021.

Z. Xu, J. Luo, J. Yan, R. Pulya, X. Li, W. Wells, and J. Jagadeesan. Adversarial uni- and multi-modal stream networks for multimodal image registration. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 12263, pp. 222–232, 2020.

Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage, 158:378–396, 2017.

D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. arXiv preprint arXiv:1703.01290, 2017a.

D. Zhang, J. Han, L. Zhao, and D. Meng. Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. Int. Journal of Computer Vision (IJCV), 127:363–380, 2019.

J. Zhang. Inverse-consistent deep networks for unsupervised deformable image registration. CoRR, abs/1809.03443, 2018.

J. Zhang, J. Wang, X. Wang, and D. Feng. The adaptive fem elastic model for medical image registration. Physics in Medicine & Biology, 59(1):97, 2013.

Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In Proc. Int. Conf. on Computer Vision (ICCV), pp. 2020–2030, 2017b.

Y. Zhang, Y. Pei, and H. Zha. Learning dual transformer network for diffeomorphic registration. In Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 129–138, 2021.

A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 8543–8553, 2019a.

S. Zhao, Y. Dong, E. I. Chang, and Y. Xu. Recursive cascaded networks for unsupervised medical image registration. In Proc. Int. Conf. on Computer Vision (ICCV), pp. 10599–10609, 2019b.

S. Zhao, T. Fung Lau, J. Luo, E. I. Chang, and Y. Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE J. of Biomedical and Health Informatics (JBHI), 24(5):1394–1404, 2020.

H. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu. nnformer: Interleaved transformer for volumetric segmentation. CoRR, abs/2109.03201, 2021.

S. Zhou, B. Hu, Z. Xiong, and F. Wu. Self-distilled hierarchical network for unsupervised deformable image registration. IEEE Trans. on Medical Imaging (TMI), 42(8):2162–2175, 2023a.

Z. Zhou, J. Luo, D. Arefan, G. Kitamura, and S. Wu. Human not in the loop: objective sample difficulty measures for curriculum learning. In Proc. IEEE Int. Symposium on Biomedical Imaging (ISBI), pp. 1–5, 2023b.

W. Zufeng, L. Tian, W. Jiang, D. Yi, and Q. Zhiguang. Medical image registration using b-spline transform. Int. J. of Simulation Systems, Science & Technology, 17(48), 2016.

SUPPLEMENTARY MATERIAL

A EXPERIMENTAL SETUP

A.1 DATASETS AND PREPROCESSING

IXI (Information eXtraction from Images). The publicly available IXI [3] dataset contains 576 T1-weighted MRI scans. These were split into 403 for training, 58 for validation, and 115 for testing. The MRI volumes were cropped to dimensions of $160 \times 192 \times 224$ and underwent preprocessing using FreeSurfer. Registration performance was assessed using label maps corresponding to 34 anatomical structures. For inter-patient registration inference, 115 pairs were randomly selected for the primary evaluation of the methods. In the atlas-to-patient registration tasks, the IXI images served as the fixed images, while the moving image was an atlas brain MRI from Kim et al. (2021).

MindBoggle-101. The MindBoggle dataset [4] contains 41 anatomically labeled brain surfaces from 101 healthy individuals, divided into four subsets: HLN (12 scans), MMRR (23 scans), NKI (42 scans), and OASIS (20 scans). MRI volumes from the HLN, MMRR, and NKI subsets were registered to MNI152 space [5] using affine transformations, with a resolution of $1 \times 1 \times 1mm^3$ and a voxel grid size of $160 \times 192 \times 224$. Registration performance was evaluated based on label maps of 41 anatomical structures. For inter-patient registration, 15, 33, and 66 pairs were randomly selected from the HLN, MMRR, and NKI subsets, respectively. In the patient-to-atlas registration task, one random sample from each subset was chosen as the atlas, and the remaining samples were registered to it, with the process repeated twice to produce 33, 66, and 123 registration pairs for the HLN, MMRR, and NKI subsets, respectively.

A.2 EVALUATION METRICS

Dice Score. The Dice score Dice (1945), also known as the Dice coefficient, is a statistical measure used to gauge the similarity between two sets, commonly applied in image segmentation. It quantifies the overlap between the predicted segmentation and the ground truth by calculating the ratio of twice the area of overlap to the total number of pixels in both sets. The formula is:

$$\text{Dice} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{13}$$

where $|A|$ and $|B|$ are the sizes of the two sets (predicted and ground truth), and $|A \cap B|$ is the size of their intersection. The Dice score ranges from 0 to 1, with 1 indicating perfect agreement and 0 indicating no overlap.

Jacobian determinant. The Jacobian determinant measures local volume changes induced by a deformation field. It is computed from the Jacobian matrix, which contains partial derivatives of the deformation field with respect to spatial coordinates. A positive Jacobian determinant indicates a local volume expansion or contraction, while a non-positive Jacobian determinant indicates a problematic deformation, such as folding or inversion of the image. The percentage of non-positive values in the Jacobian determinant provides an indication of the quality of the deformation; ideally, this percentage should be low to ensure topological preservation and avoid unrealistic transformations.

A.3 COMPARISON METHODS

Four traditional optimization-based methods and nine deep learning-based registration methods are introduced as baselines for the IXI atlas-to-patient registration task, as outlined below.

SyN Avants et al. (2008) introduces a novel symmetric image normalization method to maximize cross-correlation within the space of diffeomorphic maps, along with the necessary Euler-Lagrange equations for optimization.

[3] `https://brain-development.org/ixi-dataset/`
[4] `https://mindboggle.info/`
[5] `https://www.lead-dbs.org/about-the-mni-spaces/`

NiftyReg Modat et al. (2010) presents a GPU-optimized, parallel-friendly algorithm that performs MR image registration in less than one minute, achieving the same accuracy as conventional serial methods for segmentation propagation.

LDDMM Beg et al. (2005) explores the Euler-Lagrange equations for large deformation diffeomorphic metric mapping, deriving the minimizing vector fields and implementing a semi-Lagrangian method to compute particle flows and metric distances on anatomical structures.

deedsBCV Heinrich et al. (2015) presents a automated discrete medical image registration framework for multi-organ segmentation across various modalities, using local self-similarity context (SSC) for similarity assessment and a Markov random field (MRF) to ensure smoothness efficiently.

VoxelMorph Balakrishnan et al. (2019) formulates registration as a function mapping an input image pair to a deformation field aligned by a CNN, using two training strategies: an unsupervised method maximizing image intensity matching and a second method leveraging auxiliary segmentations from the training data.

CycleMorph Kim et al. (2021) proposes a cycle-consistent deformable image registration method that enhances performance by preserving topology during deformation, applicable to both 2D and 3D problems and easily extendable to multi-scale implementations for large volume registration.

MIDIR Qiu et al. (2021) introduces a deep learning registration framework for fast mono-modal and multi-modal image registration using differentiable mutual information and B-spline free-form deformation to achieve smooth, efficient diffeomorphic deformation.

ViT-V-Net Chen et al. (2021) integrates ViT and ConvNets to improve volumetric medical image registration, drawing inspiration from ViT-based image segmentation methods that combine ConvNets for better localization.

PVT Wang et al. (2021) addresses dense prediction tasks by providing high output resolution and lower computational costs than ViT, while combining the strengths of both CNNs and transformers as a versatile backbone for various vision applications.

CoTr Xie et al. (2021) proposes a framework that combines CNNs with a deformable Transformer (DeTrans) for accurate 3D medical image segmentation, efficiently addressing long-range dependencies while reducing computational complexities by focusing on key positions through deformable self-attention.

nnFormer Zhou et al. (2021) is a 3D transformer for volumetric medical image segmentation that integrates interleaved convolution with self-attention, employs local and global volume-based self-attention mechanisms, and enhances the U-Net architecture by replacing skip connections with skip attentions.

TransMorph Chen et al. (2022a) is a hybrid model combining Transformer and ConvNet architectures for volumetric medical image registration, featuring diffeomorphic variants that ensure topology preservation and a Bayesian variant for assessing registration uncertainty.

H-ViT Ghahremani et al. (2024) introduces a deformable image registration method that uses dual self-attention and cross-attention mechanisms to capture multi-scale flow features, enabling high-level features to inform the representation of low-level ones across spatially distant voxel patches.

A.4 Implementation Details

All registration models, including the baselines and our proposed method, were trained for 500 epochs on an NVIDIA A6000 GPU, using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 1. For the competing methods, the default network parameter settings recommended by their respective authors were applied.

SyN Avants et al. (2008): For both inter-patient and atlas-to-patient brain MRI registration tasks, the mean squared difference (MSQ) is employed as the objective function, applying a default Gaussian smoothing of 3 and utilizing three scales with 180, 80, and 40 iterations, respectively.

NiftyReg Modat et al. (2010): The sum of squared differences (SSD) is employed as the objective function, while bending energy acts as a regularizer for all registration tasks. In inter-patient brain MRI registration, the regularization weight is empirically set to 0.0002, utilizing three scales with

300 iterations each. For atlas-to-patient brain MRI registration, the regularization weight is modified to 0.0006, with three scales and 500 iterations applied for each scale.

LDDMM Beg et al. (2005): The mean squared error (MSE) serves as the default objective function. For both inter-patient and atlas-to-patient brain MRI registration, a smoothing kernel size of 5, a smoothing kernel power of 2, a matching term coefficient of 4, a regularization term coefficient of 10, and an iteration count of 500 is applied.

deedsBCV Heinrich et al. (2015): The default objective function is self-similarity context (SSC). For both inter-patient and atlas-to-patient brain MRI registration, the hyperparameter values recommended by Hoffmann et al. (2020) is utilized for neuroimaging, setting the grid spacing, search radius, and quantization step to $6 \times 5 \times 4 \times 3 \times 2$, $6 \times 5 \times 4 \times 3 \times 2$, and $5 \times 4 \times 3 \times 2 \times 1$, respectively.

VoxelMorph Balakrishnan et al. (2019): For inter-patient and atlas-to-patient brain MRI registration, the regularization hyperparameter λ is set to 0.02 and 1, respectively, as these values are identified as optimal by the authors.

CycleMorph Kim et al. (2021): In CycleMorph, the hyperparameters α, β, and λ denote the weights for cycle loss, identity loss, and deformation field regularization, respectively. For inter-patient brain MRI registration, the hyperparameters are set to $\alpha = 0.1$, $\beta = 0.5$, and $\lambda = 0.02$, whereas for atlas-to-patient brain MRI registration, they are set to $\alpha = 0.1$, $\beta = 0.5$, and $\lambda = 1$. The authors suggest these values as optimal for neuroimaging.

MIDIR Qiu et al. (2021): The same loss function and λ value as those used in VoxelMorph are applied. Additionally, the control point spacing δ for the B-spline transformation was set to 2 for all tasks, which was identified as the optimal value by the authors.

ViT-V-Net Chen et al. (2021): This registration network was built on the Vision Transformer (ViT) framework Dosovitskiy et al. (2021). The default network hyperparameter settings recommended by the authors are utilized.

PVT Wang et al. (2021): In the context of registration with the PVT model, we adhered to the configuration recommended by TransMorph. Specifically, the default settings are implemented, with the exception that the embedding dimensions are adjusted to $\{20, 40, 200, 320\}$, the number of heads is set to $\{2, 4, 8, 16\}$, and the depth is increased to $\{3, 10, 60, 3\}$ to ensure a comparable number of parameters to those in TransMorph.

CoTr Xie et al. (2021): Default network settings by the authors are used for all registration tasks.

nnFormer Zhou et al. (2021): To ensure a fair comparison, the same Transformer parameter values from TransMorph are used for nnFormer, as nnFormer is also built on the Swin Transformer Liu et al. (2021) architecture.

TransMorph Chen et al. (2022a): The same loss function parameters as those used in VoxelMorph are applied to all tasks of TransMorph.

H-ViT Ghahremani et al. (2024): The default network settings provided by the authors are applied to all registration tasks. Specifically, the encoder consists of five layers with sizes $[32, 64, 128, 256, 512]$ and the decoder with sizes $[192, 192, 192, 192]$. The parameters for H-ViT are configured as follows: the embedding dimension f_e is set to 192, the number of feature maps S_h to 4, the voxel patch size to $2 \times 2 \times 2$, the model depth to 1, the MLP ratio in the feedforward network to 2, the drop rate to 0, and the number of heads to 32.

B Results and Discussion

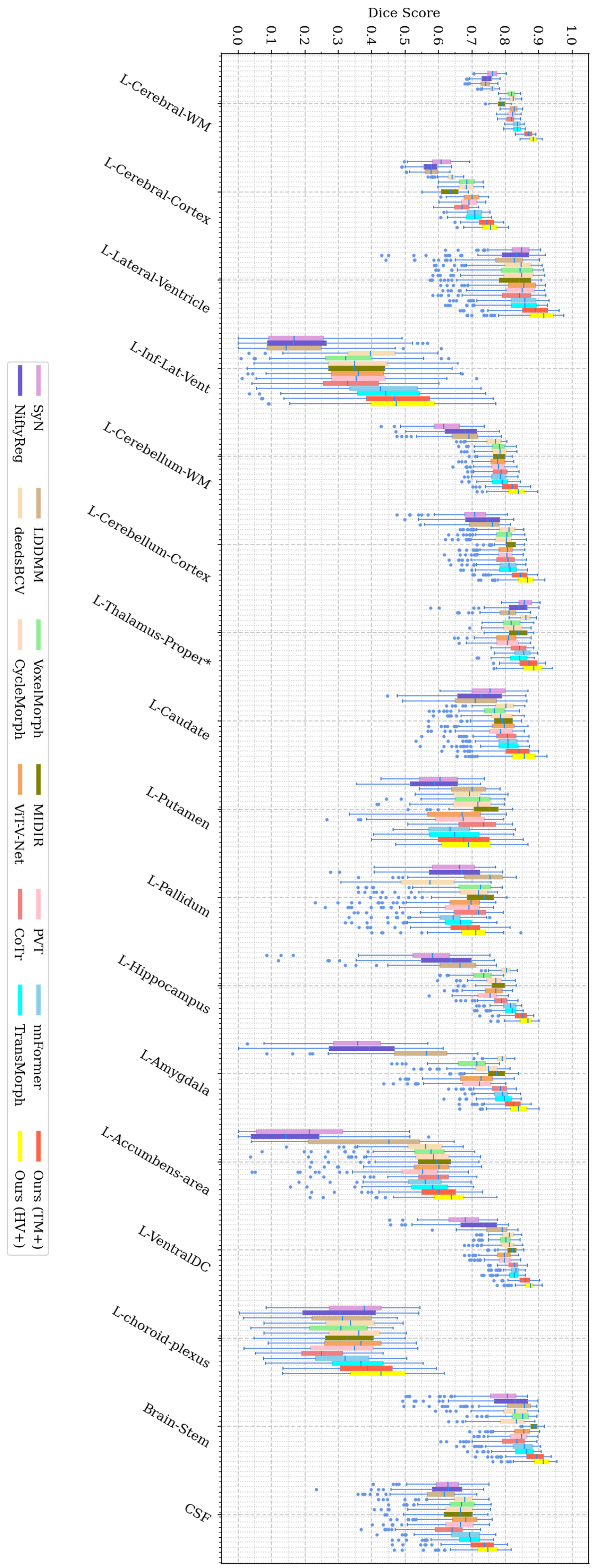

Figure 3: Dice score results for the atlas-to-patient registration of various methods on the IXI dataset (continued).

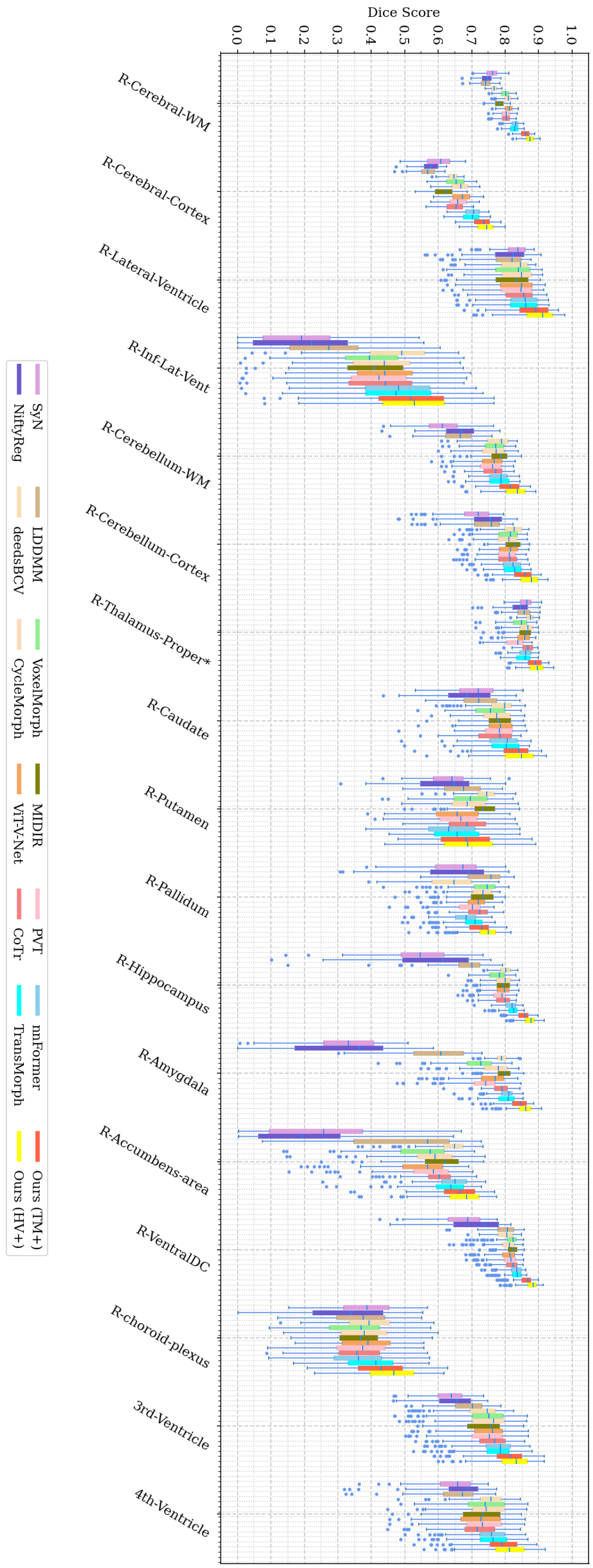

Figure 4: Dice score results for the atlas-to-patient registration of various methods on the IXI dataset.