Normalized Total Gradient for Contrast-Robust 3D Medical Image Registration

Yimin Luo[*1] YIL4025@MED.CORNELL.EDU
[1] *Weill Cornell Medicine, New York City, 10021, USA*

Yuheng Fan[2] YUHENG.FAN@LIVERPOOL.AC.UK
[2] *Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, L69 3BX, UK*

Heejong Kim[1] HKIM@MED.CORNELL.EDU
[1] *Weill Cornell Medicine, New York City, 10021, USA*

He Zhao[2] HE.ZHAO@LIVERPOOL.AC.UK
[2] *Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, L69 3BX, UK*

Mert R. Sabuncu[*3] MSABUNCU@CORNELL.EDU
[3] *Cornell University, New York City, 10021, USA*

Qingyu Zhao[*1] QIZ4006@MED.CORNELL.EDU
[1] *Weill Cornell Medicine, New York City, 10021, USA*

Editors: Under Review for MIDL 2026

Abstract

Accurate 3D medical image affine registration remains challenging in real-world clinical settings where imaging contrast varies across scanners, acquisition protocols, and patient populations. Existing deep learning frameworks commonly rely on normalized cross-correlation (NCC), which is highly sensitive to such contrast variations and often limits generalization. We introduce a contrast-robust similarity objective based on 3D Normalized Total Gradient (NTG), which compares structural gradients between fixed and moving images rather than raw intensities. Integrated into a transformer-based registration model with a Multi-Scale Dilated Attention (MSDA) module to enlarge receptive fields, our method improves the stability of cross-dataset registration without increasing architectural complexity. Experiments across ADNI, OASIS, and LPBA demonstrate higher Dice similarity and lower deformation errors under contrast perturbations, outperforming several existing registration methods. These results highlight NTG as an effective alternative to NCC, enabling more reliable and generalizable affine registration in diverse clinical imaging environments. The code is publicly available at: https://github.com/huanlemin/MDViT-NTG.

Keywords: Affine medical image registration, Image contrast, Normalized total gradient (NTG), Vision transformer (ViT).

* Corresponding author

1. Introduction

Accurate and efficient medical image affine registration is a fundamental prerequisite for disease diagnosis and longitudinal monitoring as it defines the initial global alignment of images through translation, scaling, shearing, and rotation. Recent advances in deep learning have significantly improved registration performance and efficiency (Hu et al., 2018; Huang et al., 2021; Zhao et al., 2019; Chen et al., 2021b; De Vos et al., 2019). For example, inspired by the significant potential of ViTs in handling large-scale datasets (Wang et al., 2021; Zhang et al., 2021; Chen et al., 2021a), Mok and Chung (Mok and Chung, 2022) proposed a fast and robust learning-based framework called Coarse-to-Fine Vision Transformer (C2FViT) for 3D affine medical image registration. This registration framework leverages convolutional vision transformers with a multi-resolution strategy and significantly outperforms recent CNN-based methods (Hu et al., 2018; Miao et al., 2016) while demonstrating superior robustness and generalizability across datasets.

Despite such advancement, affine registration performance of deep learning models often degrades when faced with real clinical imaging workflows, where image contrast varies widely across hospitals due to differences in scanner hardware, scanning protocols, patient-specific factors, contrast agents, and reconstruction methods. These variations substantially alter tissue appearance and pose significant challenges for deep learning-based registration models. The core issue is that most registration frameworks (Mok and Chung, 2022; Chen et al., 2021b; De Vos et al., 2019) rely on normalized cross-correlation (NCC) as the similarity metric during model optimization. However, NCC is numerically unstable under contrast changes, causing models trained on one protocol to perform poorly on others. To make the registration robust and invariant to contrast, we propose to use 3D Normalized Total Gradient (NTG) as an alternative similarity metric for model training. 3D NTG focuses on comparing contour information between images based on gradient intensity rather than pixel correspondence, making it more resilient to changes in contrast. Appendix A displays an example of registering a fixed image with a moving image under various contrasts transformed by different Gamma scales. For each contrast, we compute the final NCC and NTG values after registration and compare them to their respective reference values based on the registration with the original moving image. Results indicate NTG is more invariant to contrast variations compared to NCC.

Based on this observation, we propose a ViT-based affine registration framework, called the Multi-scale Dilated Vision Transformer (MDViT). The framework adopts a novel multi-scale dilated attention (MSDA) module to expand the receptive field of the feature maps and is optimized by the NTG between fixed and moving images. Comprehensive experiments and analyses show that our method delivers superior robustness and generalizability across diverse neuroimaging datasets in multiple affine registration tasks.

2. Methods

Our goal is to obtain an optimal affine matrix that aligns the fixed image F and the moving image M. We parameterize the affine registration problem as a predictive network $f_\theta(F, M) = \mathcal{A}$, where θ is a set of learning parameters of the network and $\mathcal{A}$ represents the predicted affine transformation matrix. Let $\mathcal{D}$ be the training dataset of all images, our

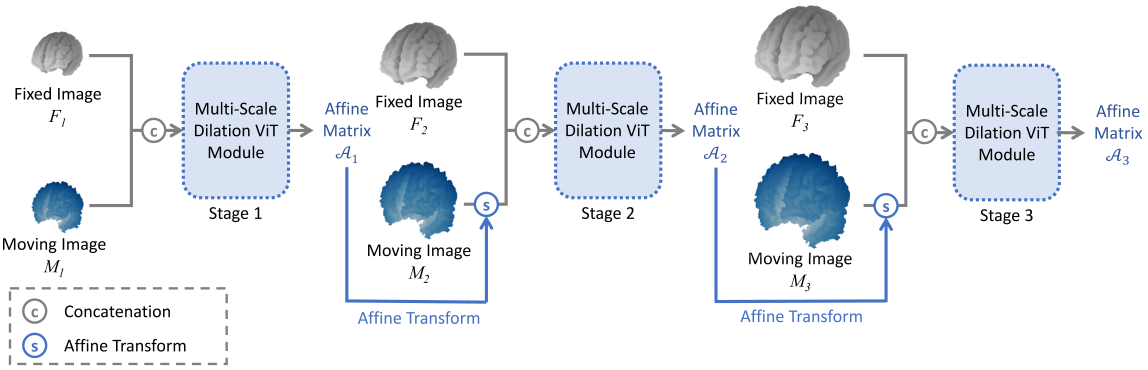

Figure 1: Overview of the proposed Multi-Scale Dilated Vision Transformer (MDViT), a 3-stage model solving the affine registration in a coarse-to-fine manner. The affine matrix estimated at each stage is used to transform the moving image for the registration at the next stage.

target is to minimize the following objective:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(F,M)\in D} \mathcal{L}_{sim}\left(F, M \circ f_\theta(F, M)\right) \tag{1}$$

where fixed and moving images (F, M) are randomly sampled from the training dataset $\mathcal{D}$, and $\circ$ is the three-dimensional affine transformation. For simplicity, let $\tilde{M}$ denote the transformed image by the affine matrix $\mathcal{A}$, i.e., $\tilde{M} = M \circ f_\theta(F, M)$. Then the loss function $\mathcal{L}_{sim}$ measures the dissimilarity between the the transformed image $\tilde{M}$ and the fixed image F. Therefore, there are two key components to be considered in the above formulation: how to design the network architecture of f_θ and which similarity measure $\mathcal{L}_{sim}$ to use. Next, we introduce our design in these two aspects.

Multi-Scale Dilated Vision Transformer. Motivated by C2FViT (Mok and Chung, 2022), our framework adopts an image pyramid of $L = 3$ stages to estimate affine registration in a coarse-to-fine manner (Fig. 1). In the first stage, the low-resolution fixed and moving images are concatenated and converted into a sequence of overlapping patch embeddings by a convolutional patch embedding layer. These embeddings are then fed into transformer encoder blocks to generate an estimated affine matrix $\mathcal{A}$ that is refined by the next two registration stages based on higher resolution images. One potential drawback of C2FViT, however, is that it cuts the input F and M into patches before sending them to the attention layer. As a result, the model tends to focus primarily on local spatial context but might overlook the global features inherent in 3D medical images.

To resolve this drawback, we design an MSDA module to enhance the receptive field and extract the feature map at different scales with dilated convolutions. Fig. 2a presents the outline of the MSDA module. Specifically, we use a 3×3 kernel size with different dilation rates. The sizes of attended receptive fields in different heads are 3×3, 5×5, and 7×7. As the receptive field increases, concatenating all different dilated features along the channel axis may lead to large and redundant feature maps, so we apply a 1D convolution to compress the feature maps, which are then fed into a multi-head self-attention module.

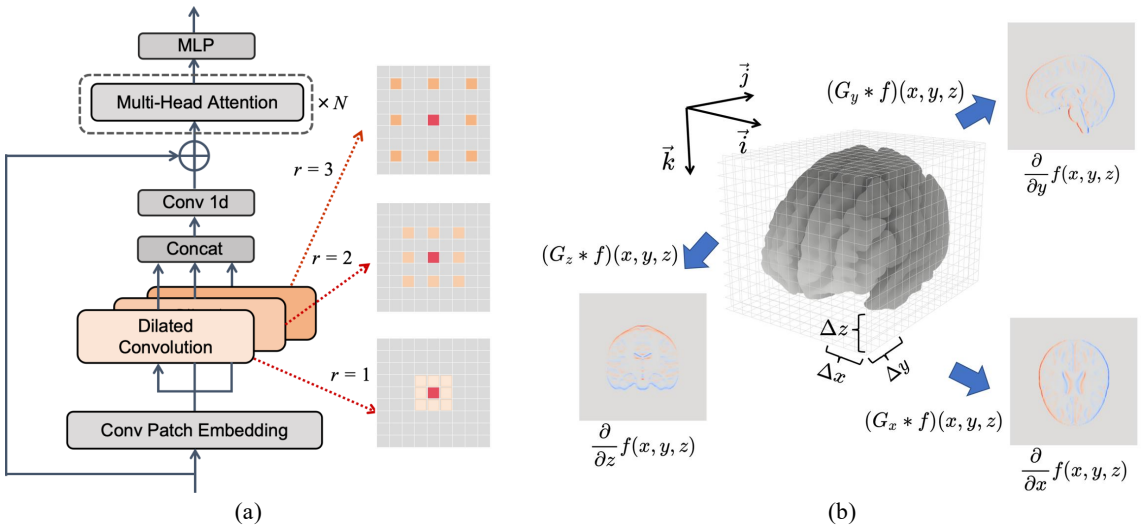

Figure 2: (a) The Multi-Scale Dilated Attention (MSDA) module; (b) Visualization of the gradient computation by the 3D Sobel filter.

This design achieves a trade-off between learning ability and model complexity, allowing for more flexible learning from 3D medical image samples.

3D Normalized Total Gradient. In terms of the objective similarity measure $\mathcal{L}_{sim}$, we use 3D NTG to train our MDViT. Specifically, the NTG measure between F and $\tilde{M}$ is defined as:

$$\mathcal{NTG}(F,\tilde{M}) = \frac{\sum_l \left\| \nabla_l \left(F - \tilde{M} \right) \right\|_1}{\sum_l \left(\|\nabla_l F\|_1 + \left\| \nabla_l \tilde{M} \right\|_1 \right)}, \tag{2}$$

where the operator $\nabla_l = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$ denotes the derivative along the direction l. By summing over all directional derivatives, the numerator is referred to as the total gradient of difference image, and the denominator represents the the total energy in F and $\tilde{M}$, defined by the L-1 norm. It can be verified that $\mathcal{NTG}(F,\tilde{M}) \in [0,1]$. Since the gradient of difference image is most sparsely distributed when two images are well aligned, NTG achieves good performance in color-channel image alignment tasks and has clear advantages over other measures. Hence, NTG is a promising alternative to NCC in defining $\mathcal{L}_{sim}$.

The key to calculating NTG is to compute the gradient of the difference image. Previous NTG (Chen et al., 2017) applied on 2D multi-spectral images is essentially a discrete operation using finite difference with stride $\Delta x = \Delta y$ for derivative approximation. To extend NTG to our 3D medical image registration tasks as a neural network training objective, we adopt the 3D Sobel filters G_x, G_y, and G_z with stride $(\Delta_x, 1, 1)$, $(1, \Delta y, 1)$, and $(1, 1, \Delta z)$ to approximate $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, and $\frac{\partial}{\partial z}$ in gradient ∇_l for neural network optimization (Fig. 2b). Appendix B further defines the formulation of the 3D Sobel filer. Finally, we compute the NTG across multiple resolution scales via the similarity pyramid in (Fig. 1). The final

objective is:

$$\mathcal{L}_{sim}(F, \tilde{M}) = \sum_{i \in [1,\cdots,L]} \frac{1}{2^{(L-i)}} \mathcal{NTG}\left(F_i, \tilde{M}_i\right), \tag{3}$$

where $(F_i, \tilde{M}_i)$ denotes the fixed and transformed moving images in the pyramid.

3. Experiments

3.1. Experimental Settings

Dataset. We adopted a pre-training and fine-tuning strategy to evaluate registration performance on small-to-medium datasets with different contrasts from the pre-training dataset. Specifically, we first pre-trained registration models on 3453 T1-weighted MRIs from the ADNI dataset (Petersen et al., 2010) and used another 433 MRIs from ADNI for validation. Based on the pre-trained models, we fine-tuned and evaluated models on 414 T1-weighted brain MRI scans from the OASIS dataset and 40 brain MRI scans from the LPBA dataset. For the OASIS dataset (Marcus et al., 2007)[1], we split the data from (Hoopes et al., 2022) into 207, 104 and 103 volumes for training, validation, and testing in a ratio of $2:1:1$. For the LPBA dataset (Shattuck et al., 2008), we split the data into 20, 10 and 10 volumes for training, validation and testing. All MRI scans were preprocessed by FreeSurfer (Fischl, 2012) including motion correction and skull stripping. We further padded all scans to the size of $256 \times 256 \times 256$ with zeros along edges, normalized to $[0, 1]$, and reduced to the size of $128 \times 128 \times 128$ by trilinear interpolation to control the space requirement of computation within an affordable range.

Registration Tasks. We evaluated our method on two registration tasks. For template-based registration, we took the MNI152 (6^{th}) brain template (Grabner et al., 2006) as the fixed image, and selected one MRI from the test dataset as the moving image. For reference-based registration, we randomly chose two random MRIs from the dataset as the fixed and moving images. During training and testing of each registration method, we followed the conventional setup of affine registration that involved adjusting each fixed-moving image pair with the center of mass initialization (CoM). The means and standard deviations of all the evaluation metrics were computed on all fixed-moving image pairs that included each sample in the test dataset.

Metrics & Comparison Methods. For evaluating the registration performance, we leveraged the Dice similarity coefficient (DSC), the 30th percentile of DSC across all cases (DSC30), the 95th percentile of the Hausdorff distance (HD95), and the Jaccard similarity coefficient (Jaccard) to measure the overlap of brain masks of the fixed image and the affinely transformed moving image. DSC scores were also computed for 12 brain regions. We compared our model with two traditional methods (ANTs and Elastix) and four deep-learning-based affine registration methods: DLIR (Miao et al., 2016), PASTA (Chen et al., 2021b), KeyMorph (Evan et al., 2022), and C2FViT (Mok and Chung, 2022).

Implementation Details. We developed all learning-based affine registration using PyTorch. All experiments were performed on a NVIDIA 3090Ti GPU with the AMD EPYC 7642 CPU and 80GB RAM. To fine-tuning models on OASIS and LPBA, we adopted the

1. https://github.com/adalca/medical-datasets/blob/master/neurite-oasis.md

Table 1: Accuracy of template-based and reference-based registration for our model and all comparison methods on LPBA and OASIS.

Dataset	Method	Template-Based				Reference-Based			
		DSC↑	DSC30↑	HD95↓	Jaccard↑	DSC↑	DSC30↑	HD95↓	Jaccard↑
LPBA	Initial	0.60±0.04	0.56±0.01	20.86±2.23	0.43±0.04	0.83±0.06	0.74±0.03	8.21±3.53	0.71±0.09
	Initial(CoM)	0.73±0.04	0.68±0.01	16.15±1.84	0.57±0.05	0.88±0.03	0.85±0.02	5.22±1.17	0.79±0.05
	ANTs	0.93±0.01	0.92±0.00	3.01±0.17	0.87±0.01	0.93±0.01	0.92±0.00	2.48±0.39	0.87±0.01
	Elastix	0.92±0.01	0.91±0.00	3.93±0.25	0.85±0.01	0.91±0.00	0.90±0.00	3.32±0.31	0.83±0.01
	DLIR	0.93±0.01	0.91±0.01	3.39±0.70	0.87±0.02	0.92±0.02	0.90±0.02	3.30±0.74	0.85±0.03
	PASTA	0.90±0.02	0.88±0.01	4.36±0.63	0.82±0.03	0.87±0.03	0.84±0.02	5.35±1.29	0.78±0.04
	KeyMorph	**0.94**±0.01	**0.93**±0.01	**2.70**±0.37	**0.89**±0.01	0.93±0.01	0.92±0.01	2.47±0.29	0.88±0.02
	C2FViT	0.93±0.01	0.91±0.01	3.20±0.50	0.87±0.02	0.92±0.01	0.92±0.00	2.96±0.38	0.86±0.01
	Ours	0.94±0.02	0.92±0.02	3.20±1.09	0.88±0.03	**0.94**±0.01	**0.93**±0.00	**2.31**±0.27	**0.88**±0.01
OASIS	Initial	0.72±0.10	0.59±0.07	17.12±5.42	0.57±0.12	0.76±0.13	0.58±0.08	11.73±5.67	0.62±0.15
	Initial(CoM)	0.84±0.04	0.80±0.02	7.10±1.65	0.73±0.06	0.90±0.03	0.86±0.03	4.87±1.62	0.82±0.05
	ANTs	0.91±0.02	0.90±0.01	4.05±1.18	0.84±0.02	0.92±0.02	0.90±0.02	3.27±1.13	0.86±0.04
	Elastix	0.93±0.02	0.90±0.01	4.57±0.71	0.86±0.04	0.92±0.02	0.89±0.01	4.73±0.92	0.85±0.04
	DLIR	0.92±0.01	0.90±0.01	4.37±0.60	0.85±0.02	0.90±0.03	0.87±0.02	5.01±1.61	0.83±0.04
	PASTA	0.91±0.02	0.89±0.01	4.82±0.89	0.83±0.03	0.91±0.03	0.87±0.02	4.70±1.57	0.83±0.04
	KeyMorph	0.92±0.01	0.91±0.01	3.87±0.62	0.86±0.02	0.94±0.01	0.92±0.01	2.97±0.56	0.88±0.02
	C2FViT	0.91±0.01	0.90±0.01	4.61±0.60	0.83±0.02	0.93±0.01	0.91±0.01	3.13±0.55	0.87±0.02
	Ours	**0.93**±0.01	**0.92**±0.00	**3.79**±0.49	**0.87**±0.01	**0.94**±0.01	**0.93**±0.01	**2.79**±0.42	**0.88**±0.02

pre-trained model at 250,000 iterations in the reference-based registration task and 150,000 iterations in the template-based registration task based on the best results validation results on ADNI. During fine-tuning, we froze the parameters of the attention layers and optimized only the convolution patch embedding layers, the multi-scale dilated layers, and the fully connected layer that predicts affine matrices. This setting allows the model to focus only on the variability of appearance across different datasets, such as scan contrast and general content, resulting in more accurate predictions. In addition, we set the learning rate of the Adam optimizer to 10^{-5} and did not apply dropout. We adjusted the weight decay to 10^{-3} and set the maximum value of the gradient clipping to 10, which kept the parameter update in a small range. We also set the maximum number of iterations to 30,000. For data augmentation on the training data, we randomly flipped both the fixed and moving volumes along one of the x, y, z axes. We then rotated both the fixed and moving volumes along the x-axis with the angle uniformly selected from the range $(-30°, 30°)$. The probability of performing such flip and rotation operations was 0.5. All methods were performed at the input resolution of $128 \times 128 \times 128$ for a fair comparison. All learning-based methods were trained in an unsupervised manner with the same data augmentation methods. The values of learning rate and weight decay were appropriately adjusted to ensure that each methods achieved optimal results on the validation set.

3.2. Results

Whole-brain Registration Accuracy. Table 1 shows the registration accuracy measured by the overlap of brain masks. In both reference-based and template-based registration, the

Table 2: Impact of loss function, model architecture, and training strategy on registration accuracy.

Dataset	Setting	Template-Based				Reference-Based			
		DSC↑	DSC30↑	HD95↓	Jaccard↑	DSC↑	DSC30↑	HD95↓	Jaccard↑
LPBA	C2FViT	$0.93_{\pm 0.01}$	$0.91_{\pm 0.01}$	$3.20_{\pm 0.50}$	$0.87_{\pm 0.02}$	$0.92_{\pm 0.01}$	$0.92_{\pm 0.00}$	$2.96_{\pm 0.38}$	$0.86_{\pm 0.01}$
	+NTG	$0.93_{\pm 0.01}$	$0.92_{\pm 0.01}$	$3.14_{\pm 0.47}$	$0.87_{\pm 0.02}$	$0.93_{\pm 0.01}$	$0.93_{\pm 0.00}$	$2.47_{\pm 0.35}$	$0.88_{\pm 0.01}$
	+MSDA	$0.93_{\pm 0.01}$	$\mathbf{0.92_{\pm 0.01}}$	$\mathbf{3.05_{\pm 0.64}}$	$\mathbf{0.88_{\pm 0.02}}$	$0.93_{\pm 0.01}$	$0.93_{\pm 0.00}$	$2.61_{\pm 0.31}$	$0.88_{\pm 0.01}$
	+fine-tuning	$\mathbf{0.94_{\pm 0.02}}$	$0.92_{\pm 0.02}$	$3.20_{\pm 1.09}$	$0.88_{\pm 0.03}$	$\mathbf{0.94_{\pm 0.01}}$	$\mathbf{0.93_{\pm 0.00}}$	$\mathbf{2.31_{\pm 0.27}}$	$\mathbf{0.88_{\pm 0.01}}$
OASIS	C2FViT	$0.91_{\pm 0.01}$	$0.90_{\pm 0.01}$	$4.61_{\pm 0.60}$	$0.83_{\pm 0.02}$	$0.93_{\pm 0.01}$	$0.91_{\pm 0.01}$	$3.13_{\pm 0.55}$	$0.87_{\pm 0.02}$
	+NTG	$0.92_{\pm 0.01}$	$0.90_{\pm 0.01}$	$8.56_{\pm 1.18}$	$0.85_{\pm 0.02}$	$0.93_{\pm 0.01}$	$0.92_{\pm 0.01}$	$3.00_{\pm 0.47}$	$0.88_{\pm 0.02}$
	+MSDA	$0.92_{\pm 0.01}$	$0.91_{\pm 0.01}$	$3.91_{\pm 0.56}$	$0.86_{\pm 0.02}$	$0.94_{\pm 0.01}$	$0.92_{\pm 0.01}$	$3.01_{\pm 0.51}$	$0.88_{\pm 0.02}$
	+fine-tuning	$\mathbf{0.93_{\pm 0.01}}$	$\mathbf{0.92_{\pm 0.00}}$	$\mathbf{3.79_{\pm 0.49}}$	$\mathbf{0.87_{\pm 0.01}}$	$\mathbf{0.94_{\pm 0.01}}$	$\mathbf{0.93_{\pm 0.01}}$	$\mathbf{2.79_{\pm 0.42}}$	$\mathbf{0.88_{\pm 0.02}}$

center of mass (CoM) initialization effectively boosted the initial Dice scores, implying that the initialization eliminated most of the misalignment due to translation. In the template-based registration task of the LPBA dataset, KeyMorph performed the best in all 4 metrics, whereas our method ranked the second among all deep learning methods. Our method did not outperform ANTs in DSC30 and HD95, indicating the paralleled strength of traditional registration in small-scale datasets. On the other hand, in the reference-based task, our method outperformed all baselines, including the two traditional methods. Notably, the performance of DLIR, PASTA, and C2FViT was significantly worse than our method, highlighting our strength in handling more diverse fixed images. Although KeyMorph was often comparable to our method on LPBA, it required manual setting of keypoint parameters, and larger parameter values led to longer training times (about 4x longer). In contrast, our method only fine-tuned the pre-trained model to achieve the same or even better registration accuracy with higher efficiency. Lastly, on the OASIS data, our method performed the best in all metrics on both registration tasks, again highlighting the strength of MSDA in handling more diverse input images in OASIS.

Ablation Study. Table 2 presents the ablation results showing the impact of MSDA module, NTG-based loss function, and pre-trained strategy on both template and reference based registration. Using the C2FViT basline, we pre-trained the model on ADNI and applied the pre-trained model on LPBA and OASIS. Changing the loss function from NCC to NTG during pre-training improved the performance in all registration tasks. Adding the MSDA module to the multi-scale registration framework mainly improved performance on the larger OASIS dataset, especially in the template-based task. This supports our assumption that larger receptive field can improve the learning of diverse samples. Lastly, the improvement of the same-task fine-tuning strategy was evident in both registration tasks on the LPBA and OASIS datasets, demonstrating the effectiveness of leveraging pre-trained strategy for 3D medical image affine registration.

Contrast Analysis. To further demonstrate the effectiveness of NTG, we used Gamma transformation to adjust the contrast of the moving image in the OASIS test set. The resulting DSC was then compared with the reference value, i.e., the DSC based on original images without contrast adjustment (Gamma=1). As shown in Fig. 3, NCC exhibited significant deviations from the reference line, indicating its high sensitivity to contrast

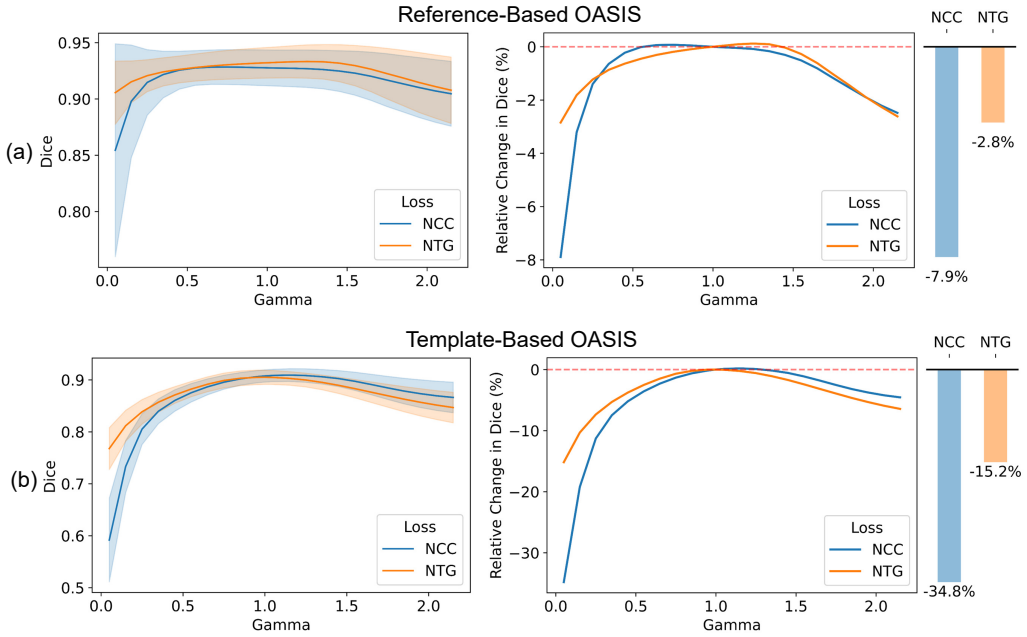

Figure 3: Performance comparison between NTG and NCC in affine registration under varying contrast conditions on the OASIS dataset. In both reference-based (a) and template-based (b) registration, NTG maintains a more stable Dice Similarity Coefficient (DSC) compared to NCC, which shows significant deviations compared to the reference value (Gamma=1.0, without contrast change).

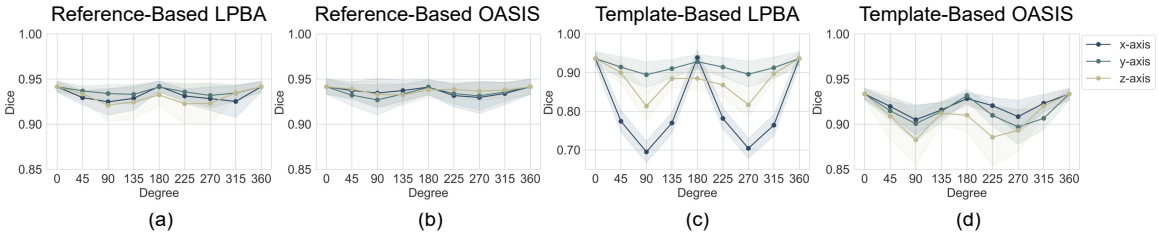

Figure 4: The DSC of our proposed method over the degree of rotation changing from 0 to 360° along x-axis, y-axis and z-axis: (a) reference-based registration on LPBA (b) reference-based registration on OASIS (c) template-based registration on LPBA (d) template-based registration on OASIS.

variations. In contrast, NTG remained much closer to the reference line, demonstrating greater robustness to such changes. Together, these results strongly indicate that NTG outperformed NCC in handling contrast variations, further validating its reliability in affine registration tasks.

Rotation Study. We also evaluated the robustness of our method for inputs with different orientation, i.e., the most challenging cases in typical affine registration tasks. Specifically, we rotated each moving image in the test set from 0 to 360° along each x, y, z axis respectively. Fig. 4 reveals that the robustness of our proposed methods was minimally impacted

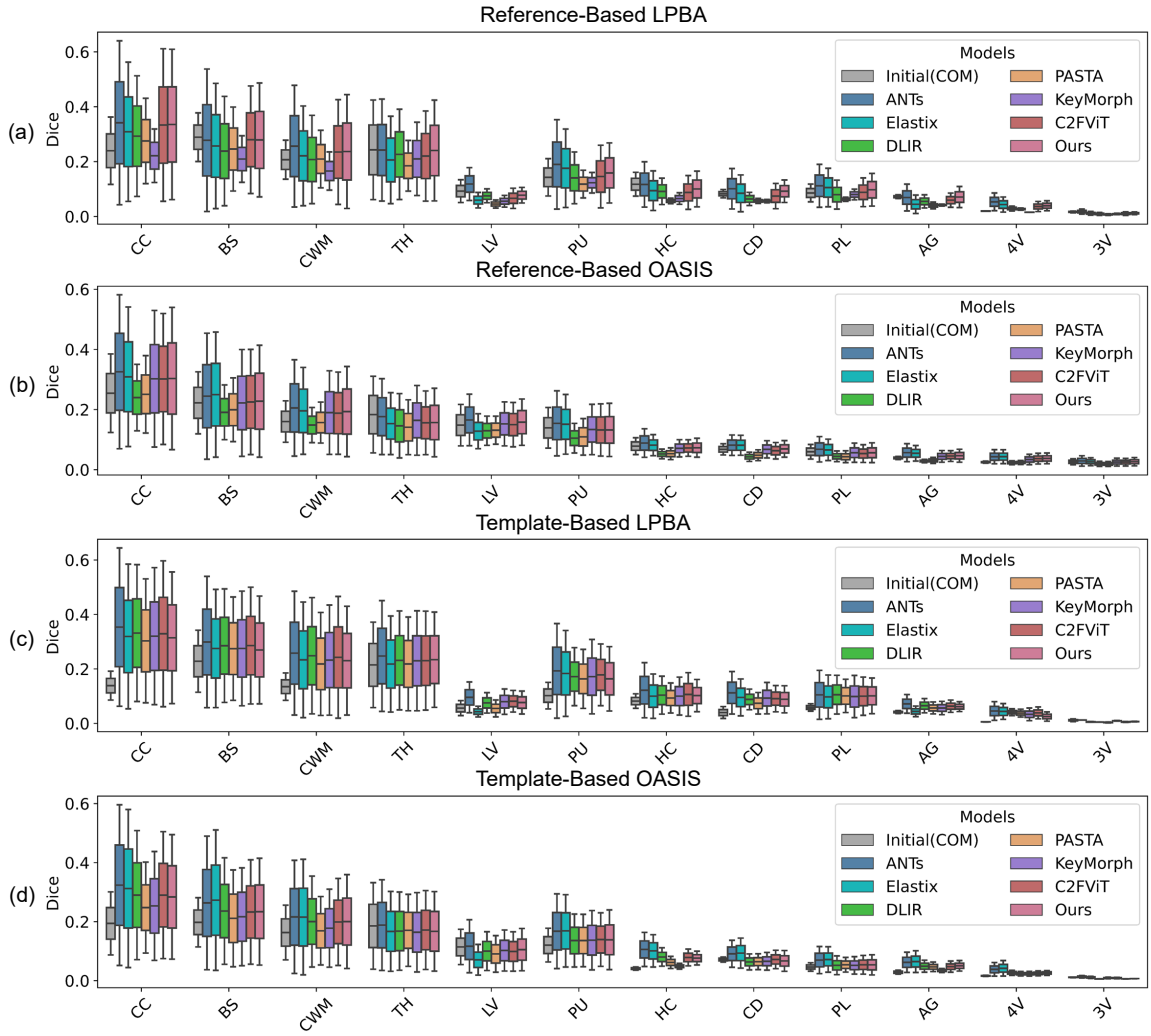

Figure 5: DSC of all registration models mesuared on 12 brain regions: (a) reference-based registration on LPBA; (b) reference-based registration on OASIS; (c) template-based registration on LPBA; (d) template-based registration on OASIS. Label information: CC: Cerebellum Cortex, BS: Brain Stem, CWM: Cerebellum White Matter, TH: represents Thalamus, PU: Putamen, LV: Lateral Ventricle, CD: Caudate, HC: Hippocampus, PL: Pallidum, AG: Amygdala, 4V: 4th Ventricle, 3V: 3rd Ventricle.

by rotation in the reference-based registration task (Fig. 4 a,b). However, for the template-based registration, the performance could be limited at a severe rotation degree of the inputs, such as 90° and 270°, especially in the small-sample scenario of the LPBA dataset. This is potentially due to the fact that the fixed template was the only registration target. Despite the appropriate data augmentation increasing the diversity of the training samples, the template-based registration model trained on smaller samples, such as datasets from a single hospital or a small clinical cohort, still faced a lack of robustness to the inputs with different orientations.

Anatomical Analysis. In addition to evaluating registration accuracy based on the whole-brain mask as shown in Table 1, we calculated the DSC for all methods separately for 12 brain regions, including the cerebellum, subcortical regions, and ventricles. The results are summarized in Fig. 5. As expected, given the small volume of the examined brain structures, the DSC were generally much lower than the whole-brain analysis. Nevertheless we made several notable observations. The performance in the template-based registration task (Fig. 5c,d) was more uniform across methods than in the reference-based task (Fig. 5a,b), and often times deep learning methods did not yield improvement over ANTs. This observation is in line with the fact that in the template-based task these brain structures in the template were smoothed out, less noisy, and identical across all registration runs. However, when the fixed images became diverse in the reference-based tasks (Fig. 5a,b), the performance gap across method emerged, with C2FViT and our method outperforming other deep learning methods. Interestingly, some methods that showed strong performance in the whole-brain analysis did not perform well in the regional analysis (e.g., KeyMorph), suggesting a trade-off between global shape matching and regional shape matching in driving the registration solution.

4. Conclusion

We propose MDViT, a novel ViT-based affine registration framework that employs dilated convolution operation to expand the receptive field of the feature maps and NTG for model optimization. In contrast to several deep learning-based registration methods, our method offers superior robustness and generalizability across two neuroimaging datasets. We envision that our method can be used in a variety of registration tasks when high-quality samples are insufficient and new samples have different contrasts than training images. While our study only explores the usage of NTG for affine registration, our model can be extended to compute non-linear deformations in the future. In this work, we only used a publicly available brain dataset for pre-training registration models. But when public datasets are insufficient for some registration tasks, synthetic samples can become a substitute. Along with the rapid development of generative models, abundant and zero-cost synthetic samples in high quality can be produced and exploited in the pre-training stage, and embedding text information to build a large registration model for multi-organ is also promising.

5. Acknolwedgement.

This work was supported in part by the National Artificial Intelligence Research Resource (NAIRR) Pilot Grant.

References

Haroon Ashraf, Wail A Mousa, and Saleh Al Dossary. Sobel filter for edge detection of hexagonally sampled 3d seismic data. *Geophysics*, 81(6):N41–N51, 2016.

Junyu Chen, Yufan He, Eric Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. In *Medical Imaging with Deep Learning*, 2021a.

Shu-Jie Chen, Hui-Liang Shen, Chunguang Li, and John H Xin. Normalized total gradient: A new measure for multispectral image registration. *IEEE Transactions on Image Processing*, 27(3):1297–1310, 2017.

Xu Chen, Yanda Meng, Yitian Zhao, Rachel Williams, Srinivasa R Vallabhaneni, and Yalin Zheng. Learning unsupervised parameter-specific affine transformation for medical images registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 24–34. Springer, 2021b.

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

M Yu Evan, Alan Q Wang, Adrian V Dalca, and Mert R Sabuncu. Keymorph: Robust multi-modal affine registration via unsupervised keypoint detection. In *Medical imaging with deep learning*, 2022.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Günther Grabner, Andrew L Janke, Marc M Budge, David Smith, Jens Pruessner, and D Louis Collins. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part II 9*, pages 58–66. Springer, 2006.

Andrew Hoopes, Malte Hoffmann, Douglas N Greve, Bruce Fischl, John Guttag, and Adrian V Dalca. Learning the effect of registration hyperparameters with hypermorph. *The journal of machine learning for biomedical imaging*, 1, 2022.

Yipeng Hu, Marc Modat, Eli Gibson, Nooshin Ghavami, Ester Bonmati, Caroline M Moore, Mark Emberton, J Alison Noble, Dean C Barratt, and Tom Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1070–1074. IEEE, 2018.

Weijian Huang, Hao Yang, Xinfeng Liu, Cheng Li, Ian Zhang, Rongpin Wang, Hairong Zheng, and Shanshan Wang. A coarse-to-fine deformable transformation framework for unsupervised multi-contrast mr image registration with dual consistency constraint. *IEEE Transactions on Medical Imaging*, 40(10):2589–2599, 2021.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

Shun Miao, Z Jane Wang, and Rui Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE transactions on medical imaging*, 35(5):1352–1363, 2016.

Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, 39(3): 1064–1080, 2008.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 129–138. Springer, 2021.

Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24(5):1394–1404, 2019.

Appendix A: Illustration of Impact of Contrast Variation on NCC and NTG

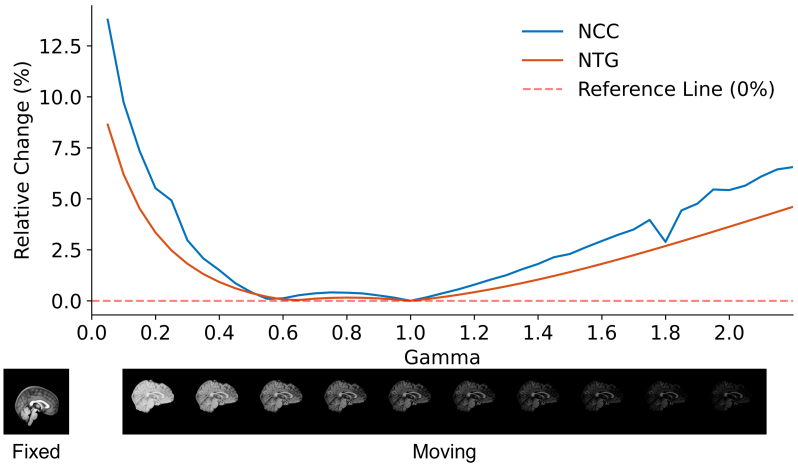

Figure 6: Relative change (%) of NTG and NCC after changing the contrast of the moving image: we apply different scales of gamma transformation to change the contrast of the moving image, register it to the fixed image, calculate NCC and NTG between the two images, and compare those metrics to the results of the registration to the original moving image (Gamma=1). Gamma = 1.0 represents the original (unmodified) image contrast. Curves closer to the Reference Line (0%) indicate higher invariance to contrast variations.

Appendix B: 3D Sobel Filter for NTG Computation

We take the axial, sagittal, and coronal axis as the x-axis, y-axis, and z-axis of the three-dimensional Cartesian coordinate respectively. Let $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ be the direction vector of x-axis, y-axis, and z-axis, we can define ∇_l in Eq. 2 as:

$$\nabla_l = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) = \boldsymbol{i}\frac{\partial}{\partial x} + \boldsymbol{j}\frac{\partial}{\partial y} + \boldsymbol{k}\frac{\partial}{\partial z}. \tag{4}$$

As an isotropic image operator, 3D Sobel operator can provide an approximation of the gradient of the image intensity function. This operator offers several advantages over other gradient operators, particularly in terms of adaptability across different volumetric data. Unlike traditional gradient operators, which may be sensitive to noise or variations in intensity distributions, 3D Sobel operator applies a weighted convolution that enhances edge detection while reducing sensitivity to local fluctuations.

Motivated by the idea of using the discrete difference with displacement along each orthogonal direction of an image (Chen et al., 2017), we adapt 3D convolution with the 3D Sobel filters G_x, G_y, and G_z and set the stride to be $(\Delta_x, 1, 1)$, $(1, \Delta y, 1)$, and $(1, 1, \Delta z)$, thus approximating the $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, and $\frac{\partial}{\partial z}$ in gradient ∇_l for neural network optimization.

These approximation can be represented as:

$$\frac{\partial}{\partial x} f(x,y,z) \approx (G_x * f)(x,y,z) \tag{5}$$

$$\frac{\partial}{\partial y} f(x,y,z) \approx (G_y * f)(x,y,z) \tag{6}$$

$$\frac{\partial}{\partial z} f(x,y,z) \approx (G_z * f)(x,y,z) \tag{7}$$

where $f(x,y,z)$ is a given 3D image with (x,y,z) coordinates, $*$ refers to the valid 3D cross-correlation operator (i.e. 3D convolution operation), and the 3D Sobel filters G_x, G_y, and G_z can be represented as (Ashraf et al., 2016):

$$G_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -2 & -4 & -2 \\ 0 & 0 & 0 \\ 2 & 4 & 2 \end{bmatrix} \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{8}$$

$$G_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 & 0 & 2 \\ -4 & 0 & 4 \\ -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{9}$$

$$G_z = \begin{bmatrix} -1 & -2 & -1 \\ -2 & -4 & -2 \\ -1 & -2 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{10}$$

With this convolutional operation, we can calculate the 3D NTG on GPU to achieve higher efficiency than the NCC-based loss. For example, under the same environment and experimental setup in training C2FViT for the template registration on LPBA dataset, it took 602.6 minutes to perform 50,000 iterations using NCC, while it took 542.9 minutes using NTG. When processing extremely noisy medical samples, like ultrasound and low-dose CT images, the effectiveness of NTG may be impacted. Overall, using NTG in the registration framework is a more robust and principled strategy. It aligns with the human visual system, which is naturally sensitive to structural changes like image contours at the beginning.