

# Deformable Cross-Attention Transformer for Medical Image Registration

Junyu Chen<sup>1</sup>, Yihao Liu<sup>2</sup>, Yufan He<sup>3</sup>, and Yong Du<sup>1</sup>

{jchen245,yliu236,duyong}@jhmi.edu; yufanh@nvidia.com

<sup>1</sup> Russell H. Morgan Department of Radiology and Radiological Science,  
Johns Hopkins Medical Institutes, Baltimore, MD, USA

<sup>2</sup> Department of Electrical and Computer Engineering,  
Johns Hopkins University, Baltimore, MD, USA

<sup>3</sup> NVIDIA Corporation, Bethesda, MD, USA

**Abstract.** Transformers have recently shown promise for medical image applications, leading to an increasing interest in developing such models for medical image registration. Recent advancements in designing registration Transformers have focused on using cross-attention (CA) to enable a more precise understanding of spatial correspondences between moving and fixed images. Here, we propose a novel CA mechanism that computes windowed attention using deformable windows. In contrast to existing CA mechanisms that require intensive computational complexity by either computing CA globally or locally with a fixed and expanded search window, the proposed deformable CA can selectively sample a diverse set of features over a large search window while maintaining low computational complexity. The proposed model was extensively evaluated on multi-modal, mono-modal, and atlas-to-patient registration tasks, demonstrating promising performance against state-of-the-art methods and indicating its effectiveness for medical image registration. The source code for this work will be available after publication.

**Keywords:** Image Registration · Transformer · Cross-attention.

## 1 Introduction

Deep learning-based registration methods have emerged as a faster alternative to optimization-based methods, with promising registration accuracy across a range of registration tasks [1, 10]. These methods often adopt convolutional neural networks (ConvNets), particularly U-Net-like networks [21], as the backbone architecture [1, 10]. Yet, due to the locality of the convolution operations, the effective receptive fields (ERFs) of ConvNets are only a fraction of their theoretical receptive fields [12, 16]. This limits the performance of ConvNets in image registration, which often requires registration models to establish long-range spatial correspondences between images.

Transformers, which originated from natural language processing tasks [26], have shown promise in a variety of medical imaging applications [12], including

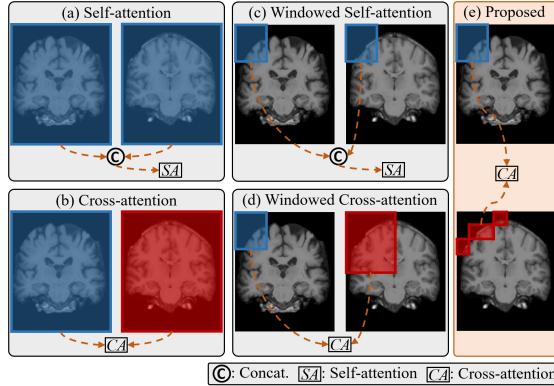


Fig. 1: Graphical illustrations of different attention mechanisms. (a) The conventional self-attention [6, 26] used in ViT-V-Net [5] and DTN [29], which computes attention for the concatenated tokens of moving and fixed images. (b) Cross-attention used in Attention-Reg [25], which computes attention between the tokens of moving and fixed images. (c) Windowed self-attention [15] used in TransMorph [4], which computes attention for the concatenated tokens of moving and fixed images within a local window. (d) Windowed cross-attention proposed in XMorpher [23], which computes attention between the tokens of fixed and moving images, specifically between two local windows of different sizes. (e) The proposed deformable cross-attention mechanism, which computes attention between tokens within a rectangular window and a deformed window with an arbitrary shape.

registration [4, 5, 29]. Transformers employ the self-attention (SA) mechanism, which can either be a global operation [6] or computed locally within large windows [15]. Consequently, Transformers have been shown to capture long-range spatial correspondences for registration more effectively than ConvNets [4].

Several recent advancements in Transformer-based registration models have focused on developing cross-attention (CA) mechanisms, such as XMorpher [23] and Attention-Reg [25]. CA improves upon SA by facilitating the efficient fusion of high-level features between images to improve the comprehension of spatial correspondences. However, the existing CA mechanisms still have drawbacks; either they compute CA globally [25], which prevents hierarchical feature extraction and applies only to low-resolution features, or they compute CA within a fixed but expanded window [23], which significantly increases computational complexity.

In this paper, we present a hybrid Transformer-ConvNet model based on a novel deformable CA mechanism for image registration. As shown in Fig. 1, the proposed deformable CA module differs from existing SA and CA modules in that it employs the windowed attention mechanism [15] with a learnable offset. This allows the sampling windows of the reference image to take on any shapes based on the offsets, offering several advantages over existing methods: 1) In contrast to the CA proposed in [23], which calculates attention between win-

dows of varying sizes, the proposed deformable CA module samples tokens from a larger search region, which can even encompass the entire image. Meanwhile, the attention computation is confined within a uniform window size, thereby keeping the computational complexity low. **2)** The deformable CA enables the proposed model to focus more on the regions where the disparity between the moving and fixed images is significant, in comparison to the baseline ConvNets and SA-based Transformers, leading to improved registration performance. Comprehensive evaluations were conducted on mono- and multi-modal registration tasks using publicly available datasets. The proposed model competed favorably against existing state-of-the-art methods, showcasing its promising potential for a wide range of image registration applications.

## 2 Background and Related Works

**Self-attention.** SA [6, 26] is typically applied to a set of tokens (*i.e.*, embeddings that represent patches of the input image). Let  $\mathbf{x} \in \mathbb{R}^{N \times D}$  be a set of  $N$  tokens with  $D$ -dimensional embeddings. The tokens are first encoded by a fully connected layer  $\mathbf{U}_{q,k,v} \in \mathbb{R}^{D \times D_{q,k,v}}$  to obtain three matrix representations, Queries  $\mathbf{Q}$ , Keys  $\mathbf{K}$ , and Values  $\mathbf{V}$ :  $[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \mathbf{x}\mathbf{U}_{q,k,v}$ . Subsequently, the scaled dot-product attention is calculated using  $SA(\mathbf{x}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_k}})\mathbf{V}$ . In general, SA computes a normalized score for each token based on the dot product of  $\mathbf{Q}$  and  $\mathbf{K}$ . This score is then used to decide which Value token to attend to.

**Cross-attention.** CA is a frequently used variant of SA for inter- and intra-modal tasks in computer vision [2, 11, 28] and has been investigated for its potential in image registration [14, 23, 25]. CA differs from SA in terms of how the matrix representations are computed. As CA is typically used between two modalities or images (*i.e.*, a base image and a reference image), the matrices  $\mathbf{Q}$ ,  $\mathbf{K}$ , and  $\mathbf{V}$  are generated using different inputs:

$$[\mathbf{K}_b, \mathbf{V}_b] = \mathbf{x}_b \mathbf{U}_{k,v}, \quad \mathbf{Q}_r = \mathbf{x}_r \mathbf{U}_q, \quad CA(\mathbf{x}) = \text{softmax}(\frac{\mathbf{Q}_r \mathbf{K}_b^\top}{\sqrt{D_k}}) \mathbf{V}_b, \quad (1)$$

where  $\mathbf{x}_b$  and  $\mathbf{x}_r$  denote, respectively, the tokens of the base and the reference. In [25], Song *et al.* introduced **Attention-Reg**, which employs Eqn. 1 to compute CA between a moving and a fixed image. To ensure low computational complexity, CA is computed globally between the downsampled features extracted by ConvNets. However, because CA is only applied to a single resolution, it does not provide hierarchical feature fusion across different resolutions, a factor that is deemed important for several successful registration models [19, 20]. More recently, Shi *et al.* introduced **XMorpher** [23], which is based on the Swin Transformer [15]. In **XMorpher**, CA is computed between the local windows of the tokens of different resolutions, enabling hierarchical feature fusion. As shown in Fig. 1 (d), the local windows are of different sizes, with a base window of size  $N_b = h \times w \times d$  and a larger search window of size  $N_s = \alpha h \times \beta w \times \gamma d$ , where  $\alpha$ ,  $\beta$ , and  $\gamma$  are set equally to 3. Using a larger search window facilitates the effective establishment of spatial correspondence, but it also increases

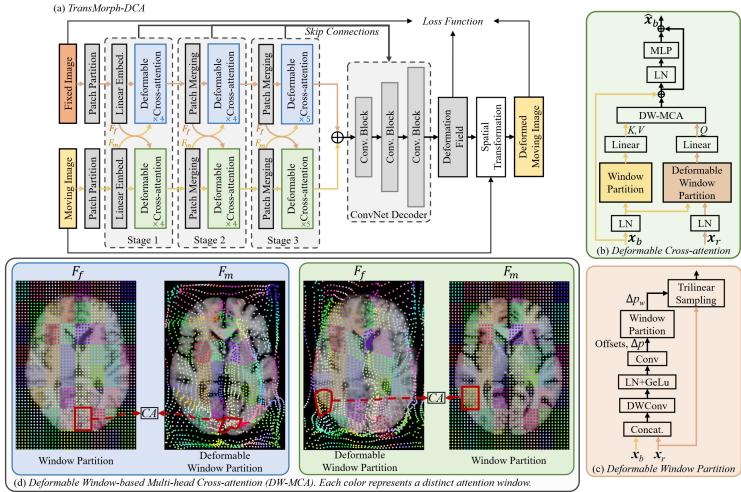


Fig. 2: The overall framework of the proposed method. (a) The proposed network architecture, which is composed of parallel Transformer encoders and a ConvNet decoder to generate a deformation field. (b) The deformable CA, which fuses features between encoders. (c) The schematic of the deformable window partitioning strategy. (d) An example of deformable CA computation in the DW-MCA.

the computational complexity of each CA module. Specifically, if the same window size of  $N_b$  is used, the complexity of CA is approximately  $O(N_b^2 D_k)$ . However, if windows of different sizes,  $N_b$  and  $N_s$ , are used, the complexity becomes  $O(N_b N_s D_k) = O(\alpha\beta\gamma N_b^2 D_k)$ , where  $\alpha\beta\gamma = 3 \times 3 \times 3 = 27$ . This means that using a larger search window would increase the computational complexity dramatically (by 27 times) and quickly become computationally infeasible.

Enlarging the search space while keeping computational costs low is challenging for 3D medical image registration. In this paper, we try to solve it with a deformable CA module that operates on equal-sized windows. This module not only provides hierarchical feature fusion, but also allows more efficient token sampling over a larger region than previously mentioned CA modules. Additionally, the proposed CA maintains a low computational complexity.

### 3 Proposed Method

The proposed model is depicted in Fig. 2 (a), which has dual Transformer encoders with deformable CA modules that enable effective communication between them. Each encoder is similar to the Swin [15] used in TransMorph [4], but the SA modules are replaced with the deformable CA modules. To integrate the features between each stage of the two encoders, we followed [25] by adding the features and passing them to the decoder via skip connections. In contrast to XMorpher [23], which uses a Transformer for the decoder, we opted for the ConvNet decoder introduced in [3,4]. This choice was motivated by the inductive bias that ConvNets bring in, which Transformers typically lack [12]. ConvNets

are also better at refining features for subsequent deformation generation, owing to the locality of convolution operations. Moreover, ConvNets have fewer parameters, making them efficient and hence speeding up the training process.

**Cross-attention Transformer.** Our model employs parallel deformable CA encoders to extract hierarchical features from the moving and fixed images in the encoding stage. At each resolution of the encoder,  $k$  successive deformable CA modules are applied to vertically fuse features between the two encoders. The deformable cross-attention module takes in a base (*i.e.*,  $\mathbf{x}_b$ ) and a reference (*i.e.*,  $\mathbf{x}_r$ ), and computes the attention between them, with the reference guiding the network on where to focus within the base. As shown in Fig. 2 (a), one encoding path uses the moving and fixed images as the base and reference, respectively, whereas the other encoding path switches the roles of the base and reference, using the moving image as the reference and the fixed image as the base.

**Deformable Cross-attention.** Fig. 2 (b) depicts the core element of the proposed model, the deformable CA. The module first applies *LayerNorm* (LN) to  $\mathbf{x}_b$  and  $\mathbf{x}_r$ , then partitions  $\mathbf{x}_b$  into non-overlapping rectangular equal-sized windows, following [15]. Next, the  $\mathbf{x}_b$  is projected into  $\mathbf{K}_b$  and  $\mathbf{V}_b$  embeddings through a linear layer. This process is expressed as  $[\mathbf{K}_b, \mathbf{V}_b] = \text{WP}(\text{LN}(\mathbf{x}_b))\mathbf{U}_{k,v}$ , where  $\text{WP}(\cdot)$  denotes the window partition operation. On the other hand, the window partitioning for  $\mathbf{x}_r$  is based on the offsets,  $\Delta p$ , learned by a lightweight offset network. As shown in Fig. 2 (c), this network comprises two consecutive convolutional layers (depth-wise and regular convolutional layers) and takes the added  $\mathbf{x}_b$  and  $\mathbf{x}_r$  as input. The offsets,  $\Delta p$ , shift the sampling positions of the rectangular windows beyond their origins, allowing tokens to be sampled outside these windows. Specifically,  $\Delta p$  are first divided into equal-sized windows,  $\Delta p_w$ , and tokens in  $\mathbf{x}_r$  are subsequently sampled based on  $\Delta p_w$  using trilinear interpolation. Note that this sampling process is analogous to first resampling the tokens based on the offsets and then partitioning them into windows. We generated a different set of  $\Delta p_w$  for each head in the multi-head attention, thereby enabling diverse sampling of the tokens across heads. The proposed deformable window-based multi-head CA (DW-MCA) is then expressed as:

$$\begin{aligned} [\mathbf{K}_b, \mathbf{V}_b] &= \text{WP}(\text{LN}(\mathbf{x}_b))\mathbf{U}_{k,v}, \\ \Delta p &= \theta_{\Delta p}(\mathbf{x}_b, \mathbf{x}_r), \quad \Delta p_w = \text{WP}(\Delta p), \quad \mathbf{Q}_r = \psi(\mathbf{x}_r; p + \Delta p_w)\mathbf{U}_k, \\ \text{DW-MCA}(\mathbf{x}) &= \text{softmax}\left(\frac{\mathbf{Q}_r \mathbf{K}_b^\top}{\sqrt{D_k}}\right)\mathbf{V}_b, \end{aligned} \quad (2)$$

where  $\theta_{\Delta p}$  denotes the offset network and  $\psi(\cdot; \cdot)$  is the interpolation function. To introduce cross-window connections, the shifted window partitioning strategy [15] was implemented in successive Transformer blocks.

The attention computation of the deformable CA is nearly identical to the conventional window-based SA employed in Swin [15], with the addition of a lightweight offset network whose complexity is approximately  $O(m^3 N_b D_k)$  ( $m$  is the convolution kernel size and  $m^3 \approx N_b$ ). As a result, the overall complexity of the proposed CA module is  $O(2N_b^2 D_k)$ , which comprises the complexity of the offset network and the CA computation. In comparison, the CA used in

**XMorpher** [23] has a complexity of  $O(27N_b^2D_k)$ , as outlined in section 2. This highlights the three main advantages of the deformable CA module: **1)** it enables token sampling beyond a pre-defined window, theoretically encompassing the entire image size, **2)** it allows sampling windows to overlap, improving communication between windows, and **3)** it maintains fixed-size windows for the CA computation, thereby retaining low computational complexity.

The deformable CA, the deformable attention (DA) [27], and the Swin DA (SDA) [9] share some similarities, but there are fundamental differences. Firstly, DA computes attention globally within a single modality or image, whereas the deformable CA utilizes windowed attention and a hierarchical architecture to fuse features of different resolutions across images or modalities. Secondly, the offset network in DA and SDA is applied solely to the Query embeddings of input tokens, and SDA generates offsets based on window-partitioned tokens, leading to square-shaped “windowing” artifacts in the sampling grid, as observed in [9]. In contrast, in the deformable CA, the offset network is applied to all tokens of both the reference and the base to take advantage of their spatial correspondences, resulting in a smoother and more meaningful sampling grid, as demonstrated in Figure 2 (d). Lastly, while DA and SDA use a limited number of reference points to interpolate tokens during sampling, deformable CA employs a dense set of reference points with the same resolution as the input tokens, allowing deformable CA to sample tokens more diversely.

## 4 Experiments

**Dataset and Pre-processing.** The proposed method was tested on three publicly available datasets to evaluate its performance on three registration tasks: **1)** inter-patient multi-modal registration, **2)** inter-patient mono-modal registration, and **3)** atlas-to-patient registration. The dataset used for the first task is the ALBERTs dataset [7], which consists of T1- and T2-weighted brain MRIs of 20 infants. Manual segmentation of the neonatal brain was provided, each consisting of 50 ROIs. The patients were randomly split into three sets with a ratio of 10:4:6. We performed inter-patient T1-to-T2 registration, which resulted in 90, 12, and 30 image pairs for training, validation, and testing, respectively. For the second and third registration tasks, we used the OASIS dataset [17] from the Learn2Reg challenge [8] and the IXI dataset<sup>4</sup> from [4], respectively. The former includes 413 T1 brain MRI images, of which 394 were assigned for training and 19 for testing. The latter consists of 576 T1 brain MRI images, which were distributed as 403 for training, 58 for validation, and 115 for testing. For the third task, we used a moving image, which was a brain atlas image obtained from [10]. All images from the three datasets were cropped to the dimensions of  $160 \times 192 \times 224$ .

**Evaluation Metrics.** To assess the registration performance, the Dice coefficient was used to measure the overlap of the anatomical label maps. Moreover,

---

<sup>4</sup> <https://brain-development.org/ixi-dataset/>

OASIS (Mono-modality)				IXI (Atlas-to-patient)			
Method	Dice↑	HdD95↓	SDlogJ↓	Method	Dice↑	% J  ≤ 0 ↓	%NDV↓
ConvexAdam [24]	0.846±0.016	1.500±0.304	0.067±0.005	VoxelMorph [1]	0.732±0.123	6.26%	1.04%
LapIRN [18]	0.861±0.015	1.514±0.337	0.072±0.007	CycleMorph [10]	0.737±0.123	6.38%	1.15%
TransMorph [4]	0.862±0.014	1.431±0.282	0.128±0.021	TM-bspl [4]	0.761±0.128	0%	0%
TM-TVF [3]	0.869±0.014	<b>1.396±0.295</b>	0.094±0.018	TM-TVF [3]	0.756±0.122	2.05%	0.36%
XMorpher* [23]	0.854±0.012	1.647±0.346	0.100±0.016	XMorpher* [23]	0.751±0.123	0%	0%
TM-DCA	<b>0.873±0.015</b>	1.400±0.368	0.105±0.028	TM-DCA	<b>0.763±0.128</b>	0%	0%

ALBERTs (Multi-modality)			
Method	Dice↑	% J  ≤ 0 ↓	%NDV↓
VoxelMorph [24]	0.651±0.159	0.04%	0.02%
TransMorph [4]	0.672±0.159	0.15%	0.04%
TM-TVF [3]	0.722±0.132	0.13%	0.03%
XMorpher* [23]	0.710±0.135	0.11%	0.03%
TM-DCA	<b>0.724±0.131</b>	0.24%	0.07%

Table 1: Quantitative results for mono-modal (OASIS) and multi-modal inter-patient (ALBERTs) registration tasks, as well as atlas-to-patient (IXI) registration tasks. Note that part of the OASIS results was obtained from Learn2Reg leaderboard [8].

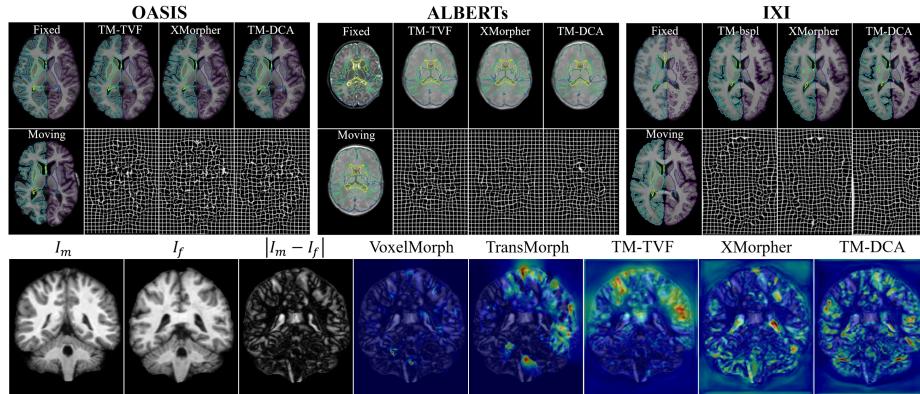


Fig. 3: Qualitative results and a visualization of Grad-CAM [22] heat maps of the comparative registration models generated for a pair of images.

for the OASIS dataset, we additionally used Hausdorff distance (HdD95) to evaluate performance and the standard deviation of the Jacobian determinant (SDlogJ) to assess deformation invertibility, in accordance with Learn2Reg. For the ALBERTs and IXI datasets, we used two metrics, the percentage of all non-positive Jacobian determinant ( $\%|J| \leq 0$ ) and the non-diffeomorphic volume (%NDV), both proposed in [13], to evaluate deformation invertibility since they are more accurate measures under the finite-difference approximation.

**Results and Discussion.** The proposed model, TM-DCA, was evaluated against several state-of-the-art models on the three registration tasks, and the corresponding quantitative outcomes are presented in Table 1. Note that our GPU was unable to accommodate the original XMorpher [23] (>48 GB) for the image size used in this study. This is likely due to the large window CA computation and the full Transformer architecture used by the model. To address this, we used the encoder of XMorpher in combination with the decoder of TM-DCA (denoted as XMorpher\*) to reduce GPU burden and facilitate a more precise comparison

between the deformable CA in TM-DCA and the CA used in **XMorpher**. On the OASIS dataset, TM-DCA achieved the highest mean Dice score of 0.873, which was significantly better than the second-best performing method, TM-TVF [3], as confirmed by a paired t-test with  $p < 0.01$ . On the IXI dataset, TM-DCA achieved the highest mean Dice score of 0.763, which was significantly better than TM-**bspl** with  $p < 0.01$ . Remarkably, TM-DCA produced diffeomorphic registration with almost no folded voxels, using the same decoder as TM-**bspl**. Finally, on the ALBERTs dataset, TM-DCA again achieved the highest mean Dice score of 0.713, which was significantly better than TM-TVF with  $p = 0.04 < 0.05$ , thus demonstrating its superior performance in multi-modal registration. It is important to note that the proposed TM-DCA model and its CA (*i.e.*, **XMorpher**) and SA (*i.e.*, TM-TVF and TM-**bspl**) counterparts differed only in their encoders, while the decoder used was identical for all models. TM-DCA consistently outperformed the baselines across the three applications, supporting the effectiveness of the proposed CA module.

Qualitative comparison results between the registration models are presented in Fig. 3. In addition, we conducted a comparison of the *Grad-CAM* [22] heat map for various learning-based registration models. The heat maps were generated by computing the NCC between the deformed moving image and the fixed image, and were then averaged across the convolutional layers at the end of the decoder, just prior to the final layer that predicts the deformation field. Notably, **VoxelMorph** exhibited inadequate focus on the differences between the image pair, which may be attributed to ConvNets’ limited ability to explicitly comprehend contextual information in the image. The SA-based models (**TransMorph** and TM-TVF) showed similar trends, wherein they focused reasonably well on regions with significant differences but relatively less attention was given to areas with minor differences. In contrast, the attention of CA-based models was more uniformly distributed, with the proposed TM-DCA method more effectively capturing differences than **XMorpher**. The presented heat maps highlight the superior performance of the proposed CA mechanism in effectively interpreting contextual information and accurately capturing spatial correspondences between images. In combination with the observed improvements in performance across various registration tasks, these results suggest that TM-DCA has significant potential as the preferred attention mechanism for image registration applications.

## 5 Conclusion

In this study, we introduced a Transformer-based network for unsupervised image registration. The proposed model incorporates a novel CA module that computes attention between the features of the moving and fixed images. Unlike the SA and CA mechanisms used in existing methods, the proposed CA module computes attention between tokens sampled from a square window and a learned window of arbitrary shape. This enables the efficient computation of attention while allowing the extraction of useful features from a large window to accurately capture spatial correspondences between images. The proposed method

was evaluated against several state-of-the-art methods on multiple registration tasks and demonstrated significant performance improvements compared to the baselines, highlighting the effectiveness of the proposed CA module.

## References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging* **38**(8), 1788–1800 (2019) [1](#), [7](#)
2. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366 (2021) [3](#)
3. Chen, J., Frey, E.C., Du, Y.: Unsupervised learning of diffeomorphic image registration via transmorph. In: Biomedical Image Registration: 10th International Workshop, WBIR 2022, Munich, Germany, July 10–12, 2022, Proceedings. pp. 96–102. Springer (2022) [4](#), [7](#), [8](#), [11](#)
4. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis* **82**, 102615 (2022) [2](#), [4](#), [6](#), [7](#), [11](#)
5. Chen, J., He, Y., Frey, E., Li, Y., Du, Y.: Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. In: Medical Imaging with Deep Learning (2021) [2](#)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020) [2](#), [3](#)
7. Gousias, I.S., Edwards, A.D., Rutherford, M.A., Counsell, S.J., Hajnal, J.V., Rueckert, D., Hammers, A.: Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *NeuroImage* **62**(3), 1499–1509 (2012) [6](#)
8. Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging* (2022) [6](#), [7](#)
9. Huang, J., Xing, X., Gao, Z., Yang, G.: Swin deformable attention u-net transformer (sdaut) for explainable fast mri. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022. pp. 538–548. Springer (2022) [6](#)
10. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical Image Analysis* **71**, 102036 (2021) [1](#), [6](#), [7](#)
11. Kim, H.H., Yu, S., Yuan, S., Tomasi, C.: Cross-attention transformer for video interpolation. In: Proceedings of the Asian Conference on Computer Vision. pp. 320–337 (2022) [3](#)
12. Li, J., Chen, J., Tang, Y., Landman, B.A., Zhou, S.K.: Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *arXiv preprint arXiv:2206.01136* (2022) [1](#), [4](#)
13. Liu, Y., Chen, J., Wei, S., Carass, A., Prince, J.: On finite difference jacobian computation in deformable image registration. *arXiv preprint arXiv:2212.06060* (2022) [7](#)

14. Liu, Y., Zuo, L., Han, S., Xue, Y., Prince, J.L., Carass, A.: Coordinate translator for learning deformable medical image registration. In: Multiscale Multimodal Medical Imaging: Third International Workshop, MMMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings. pp. 98–109. Springer (2022) [3](#)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) [2](#), [3](#), [4](#), [5](#)
16. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems **29** (2016) [1](#)
17. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: 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) [6](#)
18. Mok, T.C., Chung, A.: Conditional deformable image registration with convolutional neural network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 35–45. Springer (2021) [7](#)
19. Mok, T.C., Chung, A.: Affine medical image registration with coarse-to-fine vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20835–20844 (2022) [3](#)
20. Mok, T.C., Chung, A.C.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III **23**. pp. 211–221. Springer (2020) [3](#)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 234–241. Springer (2015) [1](#)
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017) [7](#), [8](#)
23. Shi, J., He, Y., Kong, Y., Coatrieux, J.L., Shu, H., Yang, G., Li, S.: XMorpher: Full transformer for deformable medical image registration via cross attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 217–226. Springer (2022) [2](#), [3](#), [4](#), [6](#), [7](#)
24. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 174–179. Springer (2021) [7](#)
25. Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., Wood, B.J., Sanford, T., Wang, G., Yan, P.: Cross-modal attention for multi-modal image registration. Medical Image Analysis **82**, 102612 (2022) [2](#), [3](#), [4](#)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017) [1](#), [2](#), [3](#)
27. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4794–4803 (2022) [6](#)

28. Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F., Shen, H.T.: Cross-modal attention with semantic consistence for image–text matching. *IEEE Transactions on Neural Networks and Learning Systems* **31**(12), 5412–5425 (2020) [3](#)
29. Zhang, Y., Pei, Y., Zha, H.: Learning dual transformer network for diffeomorphic registration. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 129–138. Springer (2021) [2](#)

## Appendix

### A Data Preprocessing

**ALBERTs.** We used a kernel density estimate-based method to normalize the images. Additionally, since patients exhibit considerable anatomical differences due to rapid growth, we performed affine registration to align all patients affinely with the first patient. This approach ensures that the only cause of misalignment among the volumes is nonlinear.

**IXI & OASIS.** We used FreeSurfer to perform standard procedures for brain MRI. Anatomical label maps, including over 30 ROIs, were generated for both datasets.

### B Hyperparameters Settings

	Loss	Loss Wt.	Decoder	Data Aug.	Patch Sz.	Embd. Sz.	Layer Num. ( $k$ )
OASIS	NCC+Dice+Diff.	[1, 1, 1]	TM-TVF [3]	-	4	96	{4, 4, 5}
IXI	NCC+Diff.	[1, 1]	TM-bspl [4]	Rand. Flip	4	96	{4, 4, 5}
ALBERTs	MIND+Dice+Diff.	[1, 1, 1]	TM-TVF [3]	Rand. Affine	4	96	{4, 4, 5}

Table 2: Training setups for the proposed registration model. The proposed registration model was trained using the decoder from TM-TVF and TM-bspl for different datasets. The window size was set to {5, 6, 7}, which is consistent with the window sizes used in TransMorph. The time step used in TM-TVF was set to 7. The models were trained for 500 epochs with the Adam optimizer and a learning rate of  $1e-4$ . The PyTorch framework was used for model implementation, and training was performed on an NVIDIA A6000 GPU.

### C Deformable Window Partition & Window Partition

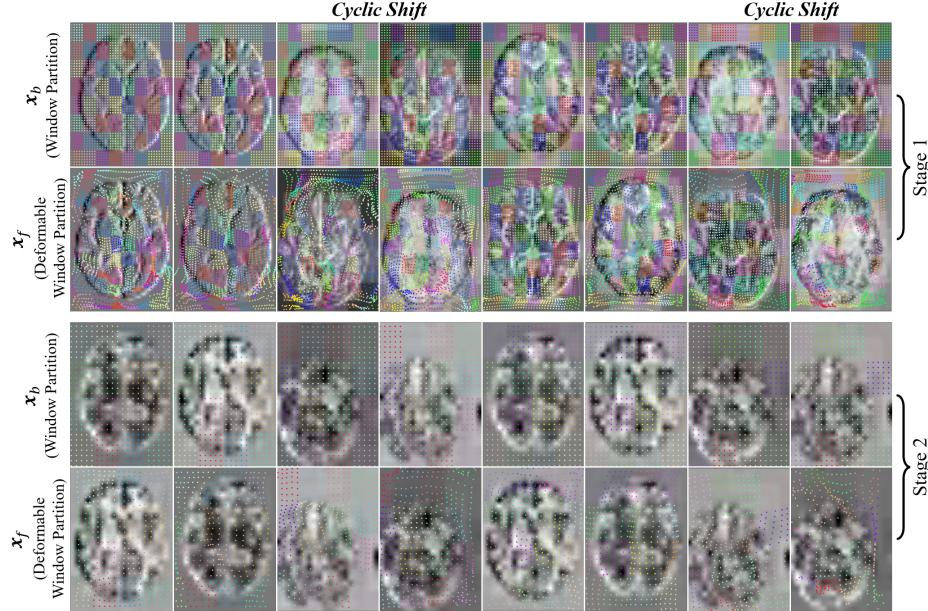


Fig. 4: The proposed window partitioning strategy in the first and second stages of the network is visualized in the figure, where each color represents a unique sampling window. Recall that the deformable window partitioning is applied to the reference image  $\mathbf{x}_r$ , while the rectangular window partitioning is applied to the base image  $\mathbf{x}_b$ . Subsequently, the cross-attention is computed between the windows in  $\mathbf{x}_b$  and  $\mathbf{x}_r$ . It is evident that the deformable window partitioning concentrates the sampling locations in information-rich regions.

## D Additional Quantitative Results

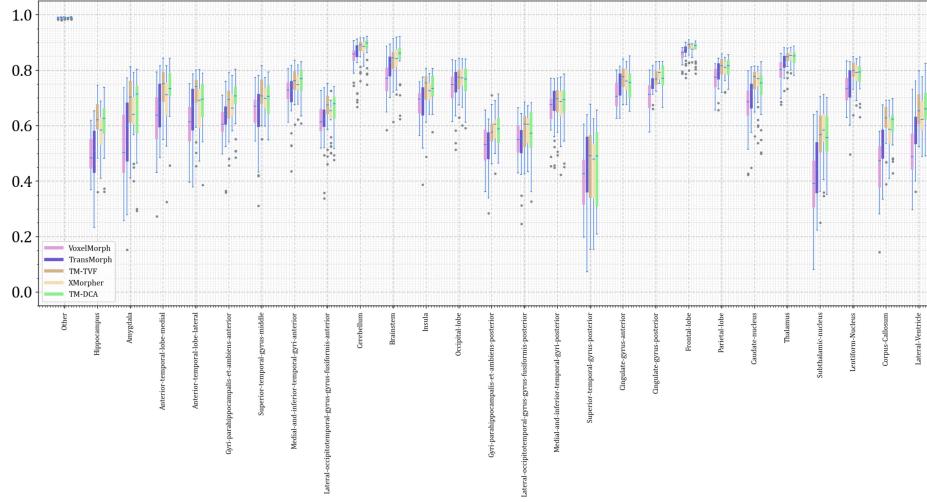


Fig. 5: Quantitative results of multi-modal registration on the ALBERTs dataset.

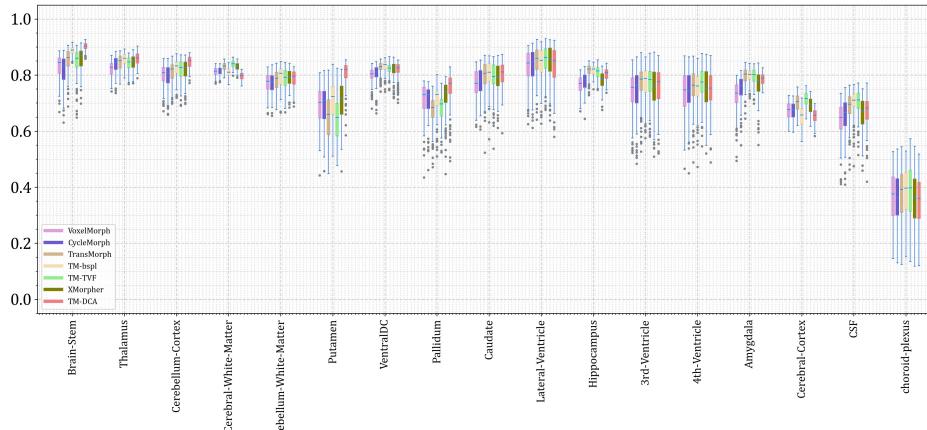


Fig. 6: Quantitative results of atlas-to-patient registration on the IXI dataset.