H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration

Morteza Ghahremani[1,2] Mohammad Khateri[3] Bailiang Jian[1,2] Benedikt Wiestler[1]
Ehsan Adeli[4] Christian Wachinger[1,2]
[1]Technical University of Munich [2]Munich Center for Machine Learning
[3]University of Eastern Finland [4]Stanford University
{morteza.ghahremani, bailiang.jian, b.wiestler, christian.wachinger}@tum.de
mohammad.khateri@uef.fi eadeli@stanford.edu

Abstract

This paper introduces a novel top-down representation approach for deformable image registration, which estimates the deformation field by capturing various short- and long-range flow features at different scale levels. As a Hierarchical Vision Transformer (H-ViT), we propose a dual self-attention and cross-attention mechanism that uses high-level features in the deformation field to represent low-level ones, enabling information streams in the deformation field across all voxel patch embeddings irrespective of their spatial proximity. Since high-level features contain abstract flow patterns, such patterns are expected to effectively contribute to the representation of the deformation field in lower scales. When the self-attention module utilizes within-scale short-range patterns for representation, the cross-attention modules dynamically look for the key tokens across different scales to further interact with the local query voxel patches. Our method shows superior accuracy and visual quality over the state-of-the-art registration methods in five publicly available datasets, highlighting a substantial enhancement in the performance of medical imaging registration. The project link is available at https://mogvision.github.io/hvit.

1. Introduction

Image registration facilitates the comparison or integration of mono- or multi-modal visual data in the same field of view. The image registration techniques are generally split into rigid/affine and non-rigid/deformable categories. Deformable image registration aims to find the underlying non-linear mapping between a pair of images. The displacement between a moving image and a target, commonly regarded as a continuous deformation field, can be modeled in various ways that have introduced many methods. The progressive optimization for gradually estimating a deformation field demands substantial computational time. Moreover, modeling the deformation field often requires a higher degree of freedom, which is difficult to fix and tune by conventional algorithms with limited parameters. Either affine or deformable registration finds numerous applications in computer vision and medical image analysis, such as image diagnosis [9, 57, 63], image-guided surgical navigation [43, 53], and has been under active research for several decades.

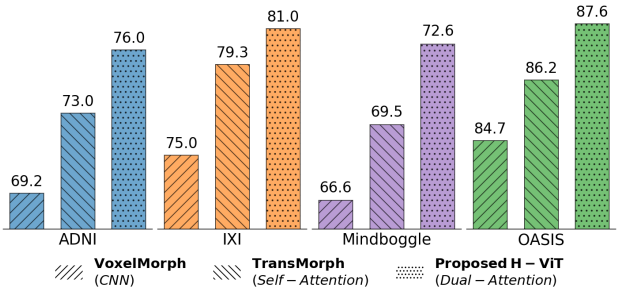

Figure 1. H-ViT has achieved state-of-the-art performance on five publicly available databases (four reported above), thanks to the integration of a dual attention mechanism. Under identical training conditions, our H-ViT consistently outperforms compared to the state-of-the-art registration methods, yielding an average Dice score ∼2.5% higher.

Deformable registration techniques incorporating convolutional neural networks (CNN) [5, 6, 17, 18, 20, 21, 24, 29, 30, 32, 50, 59, 62, 66, 73, 74] have demonstrated remarkable advancements in terms of both inference time and accuracy when compared to conventional methods [41], and there has been a rapid growth of deep learning-based approaches in recent years. More recently, deformable image registration methods have incorporated Vision Transformers (ViTs) into their architecture to address the limitations associated with the constrained receptive fields often encountered in CNN-based approaches [8, 10–12, 72]. Despite the advances in Transformer-based registration tech-

niques, they encounter challenges in the accurate representation of deformation fields.

In this paper, we introduce H-ViT as a novel hierarchical top-down representation method for effectively capturing the deformation field. Deformation fields exhibit flow patterns at various scale levels in the deformation domain. These patterns can offer valuable insights for reconstructing and estimating the deformation field, especially those laid in the high-level flow feature maps, which contain more abstract flow information (Fig. 2). Consequently, the inclusion of these patterns contributes to a more comprehensive and accurate representation of the deformation field. In our proposed top-down representation paradigm, we incorporate both short- and long-range encoding mechanisms. The former pertains to employing self-attentions akin to those in conventional Transformers, which excel at capturing short-range patterns within a specific scale level. The latter is a hierarchical cross-attention mechanism tailored to capture long-range flow patterns spanning across the deformation pyramid. To our knowledge, this is the first study that introduces a hierarchical dual-attention system for predicting the deformation field. H-ViT achieved state-of-the-art performance in deformable image registration across five publicly available MRI databases (Fig. 1). The key contributions of this study are summarized as follows:

- We propose an innovative top-down representation approach using a hybrid Transformer-CNN architecture. Our H-ViT method introduces a dual attention mechanism that strategically captures a wide spectrum of long- and short-range flow feature patterns within and between various layers.
- We overcome the problem of less accurate estimation of a deformation field in medical imaging by facilitating the stream of deformation information between layers, which provides richer flow patterns for an accurate representation of the deformation field.
- We extensively compared our unsupervised method in an identical training and inference setting with both CNN-based and Transformer-based methods on five publicly available datasets to demonstrate the superiority of the proposed H-ViT.

2. Related Work

Deformable image registration techniques are typically classified into two categories: supervised [19, 31, 47, 51, 70] and unsupervised methods [5, 8, 10, 11, 20, 32, 35, 73, 74, 77]. In supervised learning, the advantage lies in leveraging extrinsic information, such as label maps, during the training process. In contrast, unsupervised methods primarily focus on registration via uncovering intrinsic data properties. Given the high expenses associated with label collection, there is a growing interest in unsupervised registration methods. In the seminal work by Balakrishnan

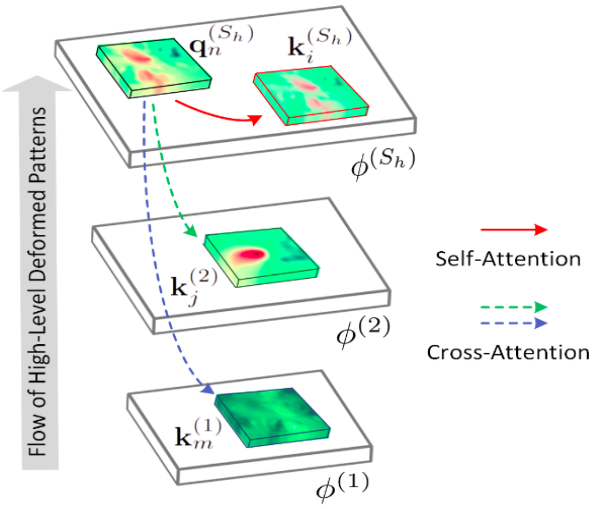

Figure 2. H-ViT with a dual attention mechanism called *self-attention* and *cross-attention* that operates in the deformation field. Given current feature map S_h, the queries $\mathbf{q}_n^{(S_h)}$ interact with local key and value tokens $\mathbf{k}_i^{(S_h)}$ in the self-attention stage, allowing the capture of short-range information within a given scale level. The same queries are further interacted with key and value tokens, $\mathbf{k}_j^{(2)}$ & $\mathbf{k}_m^{(1)}$, in higher levels of the flow feature maps, hence allowing a wide spectrum of long-range flow information between the given scale level and its higher feature maps in a hierarchical way. The hierarchical cross-attention enhances the flow stream, handles variable-sized input visual data, and improves translation invariance.

et al. [5], a formal framework for learning-based registration methods is introduced that takes the moving and target images as input, stacks them, and feeds them into a trainable neural network to extract the deformation field. Subsequently, a spatial transformer is employed to apply the obtained displacement to the moving image, resulting in the transformed or wrapped image. Thus far, many techniques have been developed for each component mentioned above. Studies [13, 47, 48, 55, 67] enhance deformable regularizers, while works [15, 16, 21, 45, 52, 60] explore novel loss functions in deformable image registration.

The majority of studies are focused on deep neural network architecture, including CNN-based methods [6, 16, 32, 61, 62, 74], Transformer-based methods [8, 10–12, 28, 38, 46, 56, 72], and diffusion methods [33, 44]. Despite the massive success of CNN-based networks in image deformable registration, these methods exhibit inadequate field-of-view, limited generalization capabilities, and insufficient valuable information in the representation of the deformation field [11, 28, 38, 56]. Consequently, recent studies have shifted towards incorporating Transformers within their architectures. The self-attention module in Transformers addresses the limitations of the CNN paradigm, particularly in capturing long-range information within the defor-

mation field. However, this study uncovers that the self-attention mechanism alone is not adequate for accurately capturing the deformation field. Thus far, a range of ViTs featuring multiscale cross-attention architectures have been proposed for tasks such as classification, segmentation, and object detection [7, 14, 22, 26, 65, 76]. Zhou *et al*. [75] developed global attention in the bottleneck for the segmentation task. FasterViT [23] provides global information propagation through the implementation of a windowing approach and hierarchical attention. HiViT [71] removes the local inter-unit operations and keeps only the global attention between tokens through several spatial merge operations and MLP layers. In deformable image registration, Chen *et al*. [12] computed the attention between the tokens of moving and target images. As mentioned earlier, this study demonstrates that incorporating self-attention at a particular scale level reveals deficiencies in accurately representing deformation fields. This inclination highlights the necessity for a sufficient patch representation to effectively capture long-range information within Transformer-based architectures. Moreover, a considerable portion of information regarding flow patterns resides in high-level features, frequently overlooked when reconstructing the flow feature maps at lower levels.

3. Methodology

3.1. Architecture

Let I_M and I_T denote a moving and a target image, respectively, defined over n spatial dimensions $\Omega \subset \mathcal{R}^n$ (in this study, $n = 3$). The goal is to establish a spatial transformation that maps the grids of the moving image into the target ones, i.e., $\phi_{M\to T} : \Omega_M \to \Omega_T$. Following the stationary velocity field approach [3, 35, 36], deformation field ϕ (also called flow field) is parameterized through the ordinary differential equation: $\frac{\partial \phi^{(t)}}{\partial t} = \nu\big(\phi^{(t)}\big)$, where $t \in [0, 1]$ and $\phi^{(0)}$ is the identity transformation ($\phi^{(0)} = Id$). The desired deformation field $\phi^{(1)}$ is obtained by integrating the stationary velocity field ν. With considering a spatially smooth velocity field ν, the recent equation is a diffeomorphic deformation, computed via scaling and squaring [2]. Diffeomorphic registration provides a differentiable and invertible solution that preserves topology, so it is widely used in deformable image registration [5, 21, 47]. $\phi^{(1)}$ is fed into a spatial transformation function[1] to deform the moving image space to the target image space and vice versa, i.e.,

$$I_{M\to T} = I_M \circ \phi_{M\to T}(p), \qquad \forall p \in \Omega \tag{1}$$

and

$$I_{T\to M} = I_T \circ \phi_{T\to M}(p), \qquad \forall p \in \Omega \tag{2}$$

where $I \circ \phi$ denotes 'I warped by ϕ'. We introduce H-ViT to represent the deformation field as $\phi = \Phi_\theta(M, T)$ with learnable parameters θ.

The framework of our approach is illustrated in Fig. 3. The moving and target images are stacked and fed into a CNN-based backbone like FPN [37, 68] to generate a set of S feature maps in the deformation field. The CNN-based backbone is comprised of $S = \min\big(\log_2(\frac{H}{h}), \log_2(\frac{W}{w}), \log_2(\frac{D}{d})\big)$ convolutional layers[2] with f_s feature number at stage $s, s \in \{1, 2, \cdots, S\}$. f_s is formed by $32 \times 2^{\lceil s/2 \rceil}, s \in \{1, 2, \cdots, S\}$, where $\lceil . \rceil$ denotes a floor function. Within this set, S_h high-level feature maps ($S_h \leq S$) are considered as input for the H-ViT's dual-attention unit. Each feature map is then mapped into f_e features by a convolutional layer, forming a feature representation $\{\phi^{(s)}\}_{s=1}^{S_h}$, where $\phi^{(1)}$ denotes the highest-level feature map. The embedding features are then fed into a dual-attention mechanism (detailed in the following section) for encoding short- and long-range flow information, providing a comprehensive representation of the deformation field for the input CNN-based layers. Utilizing the estimated deformation field and a grid sampler, we warp the to-be-registered input MRIs in both directions via Eq. (1) and Eq. (2), enabling the model to compute similarity loss of the corresponding mono-modal images:

$$\bar{\mathcal{L}}_{\text{sim}} = \frac{1}{2}\big(\mathcal{L}_{\text{sim}}(L_M, L_T, \phi_{M\to T}) + \mathcal{L}_{\text{sim}}(L_T, L_M, \phi_{T\to M})\big), \tag{3}$$

where $\mathcal{L}_{\text{sim}}$ measures the similarity score between its two inputs. To prevent producing a discontinuous deformation field, the spatial gradient $\mathbf{u}$[3] is often smoothed by a regularization term: $\mathcal{L}_{\text{smooth}} = \frac{1}{2}||\nabla \mathbf{u}||^2$. The parameters of the H-ViT network θ are optimized via minimization of the functions defined in Eq. (3) and backpropagation:

$$\mathcal{L}_{\text{tot.}} = \bar{\mathcal{L}}_{\text{sim}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}, \tag{4}$$

where λ_{smooth} is a pre-determined regularization hyperparameter, which is set to 1.

3.2. Deformation field representation by a dual-attention approach

The S_h feature maps $\{\phi^{(s)}\}_{s=1}^{S_h}$, extracted by the CNN network, are input into the H-ViT block (Fig. 3). We divide the feature maps into $h \times w \times d$ non-overlapping voxel patches, i.e. $\mathbf{X}^{(s)} = \{\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{N_s}^{(s)} | \mathbf{x}^{(s)} \in \mathbb{R}^{h\times w\times d\times f_e}\}$, $s \in \{1, \ldots, S_h\}$; N_s denotes the number of voxel patches at the ℓ-th scale level, and $\mathbf{x}_i^{(s)}$ is the vector representation of the i-th voxel patch at the s-th layer. Fig. 4 explains how

[1] Grid sampler in PyTorch: `nn.functional.grid_sample`

[2] Let the dimension of an input MRI be represented by $H \times W \times D$, where H, W, and D represent the height, width, and depth of MRI, respectively. Likewise, h, w, and d represent the height, width, and depth of voxel patches, respectively.

[3] $\phi = Id + \mathbf{u}$

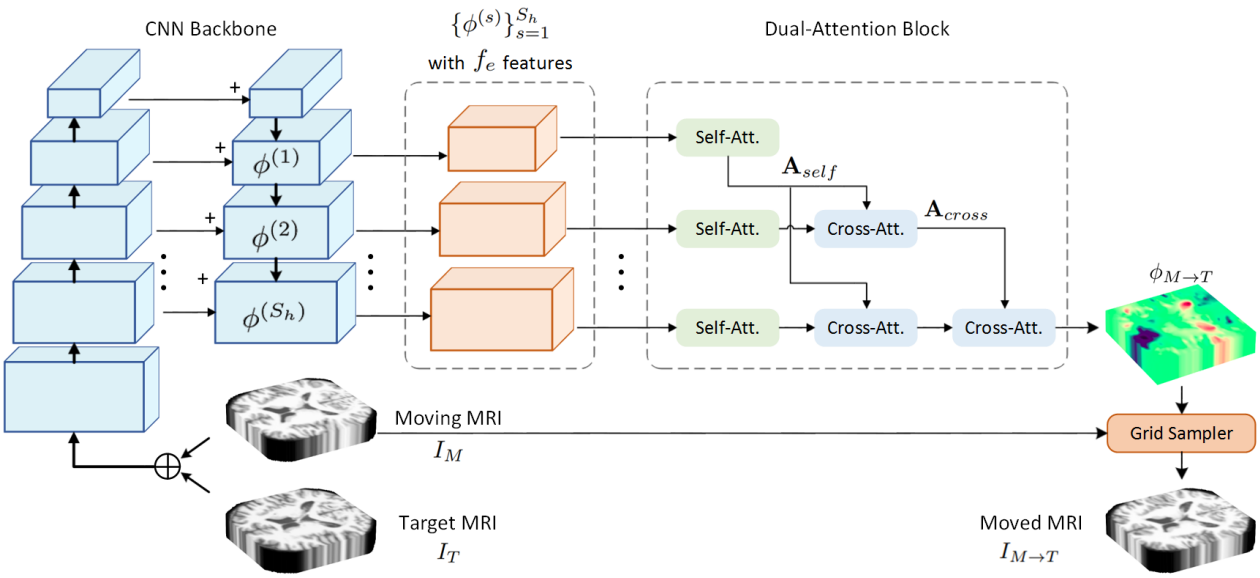

Figure 3. Overview of the proposed H-ViT architecture. Input moving and target MRIs are stacked and processed through a CNN backbone to form a translation-invariant pyramid-like representation of the deformation field. The top S_h feature maps from the CNN backbone are selected and mapped into f_e channels before being input into the dual-attention mechanism. The dual-attention block utilizes a sequence of self-attention and cross-attention blocks to encode short- and long-range flow information. The resultant deformation field is employed as input for a grid sampler, which then warps the moving MRI to produce the warped MRI.

the attention modules of H-ViT are hierarchically applied to input feature maps from the CNN-based backbone. H-ViT incorporates two attention mechanisms: i) *self-attention*, which focuses on capturing an attention map within the current feature map (short-range), and ii) *cross-attention*, which explores attention maps between interlayer patch embeddings, particularly those situated in higher-level feature layers. Abstract patterns in high-level features tend to be repeated within a specific scale/layer or across various scale levels. The hierarchy-based cross-attention module can recognize those patterns that span a larger region of the visual data. Incorporating distant patches from various layers in attention mechanisms also enables the propagation of flow patterns across all patch embeddings, regardless of their spatial proximity. This strategy facilitates the efficient dissemination of essential flow information across the entire visual data, which is particularly beneficial when dealing with large-scale 3D visual data like MRI.

Self-attention module: Self-attention in H-ViT resembles the traditional Transformers' self-attention [39, 64]. It computes attention scores among all voxel patch embeddings within the designated scale level, capturing short-range dependencies[4]. H-ViT's self-attention projects $\mathbf{X}^{(s)} \in \mathbb{R}^{N_s \times h \times w \times d \times f_e}$ into query, key and value via three matrices $\mathbf{W}_Q^{(s)} \in \mathbb{R}^{f_e \times f_q}$, $\mathbf{W}_K^{(s)} \in \mathbb{R}^{f_e \times f_k}$, and $\mathbf{W}_V^{(s)} \in \mathbb{R}^{f_e \times f_v}$, respectively:

$$\mathbf{Q}^{(s)} = \mathbf{X}^{(s)}\mathbf{W}_Q^{(s)}, \mathbf{K}^{(s)} = \mathbf{X}^{(s)}\mathbf{W}_K^{(s)}, \mathbf{V}^{(s)} = \mathbf{X}^{(s)}\mathbf{W}_V^{(s)}, \quad s = 1, \ldots, S_h. \quad (5)$$

Then self-attention $\mathbf{A}_{\text{self}}$ is computed via:

$$\mathbf{A}_{\text{self}}\left(\mathbf{X}^{(s)}\right) = \text{attention}\left(\mathbf{Q}^{(s)}, \mathbf{K}^{(s)}, \mathbf{V}^{(s)}\right) = \text{softmax}\left(\frac{\mathbf{Q}^{(s)}\mathbf{K}^{(s),\top}}{\sqrt{f_q}}\right)\mathbf{V}^{(s)}, \quad s = 1, \ldots, S_h, \quad (6)$$

where $\top$ denotes the transpose operation. As shown in Fig. 4, the output of the self-attention block serves as the input to the hierarchical cross-attention blocks. These blocks investigate the attention map between the voxel patch embeddings at layer s and those on *higher* feature scales.

Cross-attention module: The cross-attention module is a long-range-based attention mechanism designed to investigate the extended relationships between voxel patch embeddings across different scales. Since deformed patterns tend to repeat their features across several feature levels, it is tempting to capture such features to provide a comprehensive representation of the given scale level's patches. Hence, we extract key and value tokens at various scales to enable further interactions with local query tokens. While

[4]Throughout this study, the term 'long-range' refers to the relationship between voxel patch embeddings across different scale levels.

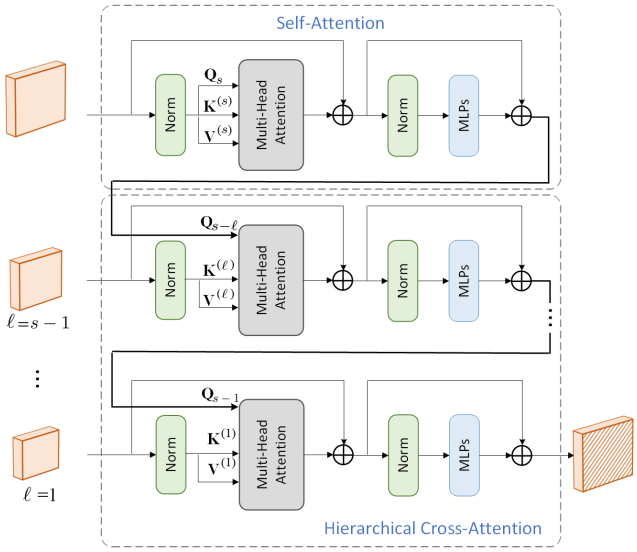

Figure 4. H-ViT's attention mechanism at scale s consists of two key components: 1) A self-attention layer responsible for calculating attention maps within the input resolution. 2) '$s-1$' hierarchical cross-attention layers for the computation of attention maps between s feature maps.

the self-attention unit focuses solely on local voxel patches within the input dimension, the cross-attention examines the connections between the same local queries and key voxel patches at different scales. If $\mathbf{X}_1$ represents the output of the self-attention layer (that serves as the input for the computation of the cross-attention), the cross-attention layer $\mathbf{A}_{cross}$ for the ℓ-th layer is computed through:

$$\begin{aligned}\mathbf{A}_{\text{cross}}\left(\mathbf{X}_1, \mathbf{X}^{(\ell)}\right) &= \text{attention}\left(\mathbf{Q}_1, \mathbf{K}^{(\ell)}, \mathbf{V}^{(\ell)}\right) \\ &= \text{softmax}\left(\frac{\mathbf{Q}_1 \mathbf{K}^{(\ell),\top}}{\sqrt{f_q}}\right)\mathbf{V}^{(\ell)}, \\ &\qquad \ell \in \{s-1, \ldots, 1\}. \quad (7)\end{aligned}$$

Note that the highest-level feature map, i.e., ϕ_1, contains no long-range cross-attention. Eq. (7) is applied recursively across all possible values of l, yielding the long-range attention map for the respective layer:

$$\mathbf{X}_s = \prod_{\ell=<s-1>} \mathbf{A}_{\text{cross}}\left(\mathbf{X}_{s-\ell}, \mathbf{X}^{(\ell)}\right), \quad (8)$$

where $\mathbf{X}_0$ is the input feature map into the cross-attention module at the given layer s, i.e., $\mathbf{A}_{\text{self}}\left(\mathbf{X}^{(s)}\right)$. After each multi-head self- and cross-attention layer, an MLP block, functioning as a feed-forward network (FFN), is applied to the voxel patch embeddings, as shown in Fig. 4. The MLP block consists of two linear transformations [4] combined with dropout layers, followed by ReLU non-linear activation functions [1].

4. Experiments

Datasets and metrics: The H-ViT method is employed in the analysis of five popular T1 MRI databases, including OASIS [25, 42], IXI[5], ADNI [27], LPBA [54], and Mindboggle [34]. Detailed information regarding the datasets and their preparation procedures are presented in Sec. A of the Supplementary Material. For quantitative comparisons on the OASIS dataset, Dice scores, the 95th percentile Hausdorff distance (HD95), and the standard deviation of the logarithm of the Jacobian determinant (SDlogJ) of the displacement field were computed via 2021 Learn2Reg[6]. For the other datasets, we computed the Dice scores of cortical and subcortical brain structures and the percentage of voxels with a non-positive Jacobian determinant (Supplementary Sec. B).

State-of-the-art deformable registration techniques: We conducted an extensive performance comparison of our method against a wide range of CNN- and Transformer-based models, including VoxelMorph [6], MIDIR [49], CycleMorph [32], ViT-V-Net [8], and TransMorph [10]. It is noteworthy that this study includes both versions of the TransMorph technique, one utilizing a cubic B-spline transformation model (denoted by TransMorph-Bspl) and the other employing Bayesian learning (TransMorph-Bayes). Additionally, we compared H-ViT with other state-of-the-art Transformer-based networks that are designed for various applications. These networks included PVT [64] and CoTr [69], which employ a hybrid Transformer-CNN architecture, as well as nnFormer [75], which relies on a pure Transformer-based architecture. The architectures of these methods were adapted to handle 3D deformation fields, replaced with the CNN backbone of VoxelMorph[7]. We preserved the fundamental elements of VoxelMorph, including the spatial transformation function, loss function, and network training procedures. We also kept all methods' configurations identical as they recommended, reported in Sec. C in Supplementary Material. The configuration details of H-ViT are also reported in Sec. C.1 in Supplementary.

Training details: We ensured the uniformity of the training settings, which included maintaining identical configurations for the optimizer and its hyperparameters, batch size, loss functions, and the number of epochs. Training was conducted on two datasets IXI and OASIS, followed by testing across five diverse datasets. We employed NVIDIA A100 GPUs with 80GB VRAM for running the experiments. Sec. C in Supplementary provides more details about the experiment settings. Due to the page limit, we present a condensed version of the results below, with the complete set of results and accompanying visualizations

[5] https://brain-development.org/ixi-dataset/
[6] https://learn2reg.grand-challenge.org/
[7] https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration

Method	Attention Mechanism	Dice ↑	HD95 ↓	SDlogJ ↓
VoxelMorph [6]	–	0.847±0.014	1.546±0.306	0.133±0.021
DFI-NFF [40]	–	0.827±0.013	1.722±0.318	0.121±0.015
LapIRN [47]	–	0.861±0.015	1.514±0.337	*0.072±0.007*
ConvexAdam [58]	–	0.846±0.016	1.500±0.304	**0.067±0.005**
TransMorph [10]	Self-Att.	*0.862±0.014*	*1.431±0.282*	0.128±0.021
Proposed H-ViT	Self-Att. + Cross-Att.	**0.876±0.014**	**1.301±0.264**	0.539±0.069

Table 1. Inference evaluation results for various registration methods, including H-ViT, on the OASIS dataset with 35 anatomical structures in inter-patient registration. Bold numbers represent the highest scores, while italicized numbers indicate the second-highest scores.

Method	Attention Mechanism	Inter-Patient Registration		Patient-to-Atlas Registration	
		Dice ↑	$\lvert J_\Phi\rvert \leq 0$ (%) ↓	Dice ↑	$\lvert J_\Phi\rvert \leq 0$ (%) ↓
Affine		0.494±0.050	–	0.445±0.055	–
VoxelMorph [6]	–	0.750±0.106	1.013±0.285	0.734±0.111	0.997±0.197
MIDIR [49]	–	0.735±0.093	*0.295±0.188*	0.722±0.096	*0.247±0.107*
CycleMorph [32]	–	0.750±0.101	1.022±0.293	0.736±0.105	0.992±0.215
CoTr [69]	Self-Att.	0.736±0.112	0.702±0.290	0.717±0.116	0.678±0.205
nnFormer [75]	Self-Att.+Global-Att.	0.727±0.101	1.284±0.349	0.718±0.097	1.282±0.256
PVT [64]	Self-Att.	0.696±0.116	1.868±0.398	0.690±0.124	1.736±0.248
ViT-V-Net [8]	Self-Att.	0.772±0.093	1.022±0.289	0.749±0.102	1.033±0.208
TransMorph-Bayes [10]	Self-Att.	0.790±0.081	1.136±0.377	0.772±0.082	1.078±0.236
TransMorph-Bspl [10]	Self-Att.	*0.793±0.075*	<**0.001**	*0.778±0.080*	<**0.001**
Proposed H-ViT	Self-Att. + Cross-Att.	**0.810±0.073**	0.525±0.209	**0.797±0.075**	0.565±0.161

Table 2. Quantitative evaluation results for the registration methods on the IXI dataset for 30 anatomical structures over 115 random pairs for inter-patient and 150 pairs for patient-to-atlas registrations.

available in supplement Sec. C.

4.1. Main results

In accordance with [25], the OASIS experiment was utilized for inter-patient registration, incorporating a total of 451 brain T1 MRI scans, wherein 394, 19, and 38 scans were allocated for training, validation, and testing, respectively. Tab. 1 reports the numerical results, where H-ViT has the highest Dice score. For the other four datasets, we employed pre-trained models from methods trained on the IXI training set under identical conditions. Tab. 2 details the results of the methods on the IXI dataset, where our method yields the highest score compared to others with a margin of +0.017 in patient-to-patient registration and +0.019 in patient-to-atlas registration. While TransMorph-Bspl achieved the best non-zero Jacobian determinant score, H-ViT also attained a score of around 0.5%, which is within an acceptable range.

The superior performance of our method is also evident in the other datasets. For ADNI (Tab. 3), H-ViT demonstrates a Dice score of approximately +0.03 higher than the second-best performing techniques in both patient-to-patient and patient-to-atlas registration scenarios. Example results of the methods are depicted in Fig. 5, with H-ViT demonstrating more accurate warping of the moving MRI compared to other techniques, particularly in frontal gyrus and in temporal, indicated by the blue and the orange rectangles, respectively. H-ViT replicated the results observed in the ADNI dataset on LPBA (Tab. 4) and Mindboggle (Tab. 6), achieving higher Dice scores of +0.034 and +0.031 in inter-patient registration compared to the second-best technique, respectively. Similarly, H-ViT demonstrated a Dice performance increase of +0.028 and +0.032 in the patient-to-atlas registration. Detailed results for all methods, including Dice scores per anatomical structure and additional visualization results, are provided in the Supplementary Material in Sec. D.

4.2. Ablation study on the H-ViT model

Dual-Attention: Tab. 5 reports the performance of the H-ViT model under various scenarios, both with and without self-attention blocks, and for different numbers of cross-attention units. While the self-attention unit enhances the performance of the CNN backbone, the cross-attention blocks further leverage this improvement, resulting in enhanced warped outcomes. The importance of

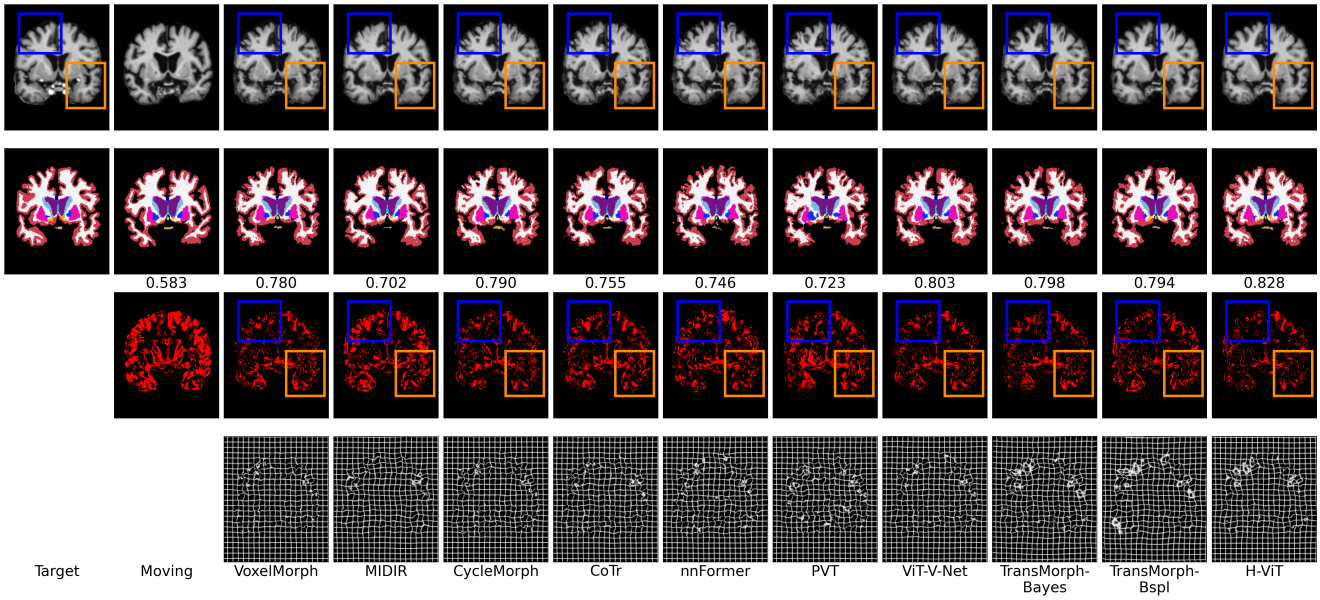

Figure 5. Example coronal slice from the ADNI dataset and outcomes (from top to bottom: MRI, segmentation, difference in segmentation between ground truth and segmented results, and the deformed grid) of different registration methods, with corresponding Dice scores below the segmentation results. In the third row, red highlights signify segmentation disparities between ground truth and segmented results, while the black ones represent accurate segmentation (optimal with fewer red pixels).

Method	Inter-Patient Dice ↑	Patient-to-Atlas Dice ↑
Affine	0.531±0.082	0.477±0.052
VoxelMorph [6]	0.692±0.214	0.646±0.226
MIDIR [49]	0.666±0.220	0.635±0.227
CycleMorph [32]	0.687±0.217	0.655±0.225
ViT-V-Net [8]	0.727±0.210	0.686±0.219
TransMorph-Bspl [10]	*0.730±0.208*	*0.702±0.213*
Proposed H-ViT	**0.760±0.203**	**0.730±0.210**

Table 3. The averaged Dice score results for ADNI registration for 45 anatomical structures over 150 random pairs for inter-patient and 150 pairs for patient-to-atlas registrations. Presented are the outcomes of the six techniques exhibiting the highest Dice scores. The percentage of folded voxels for the reported techniques is below 1.5%. A detailed table containing all methods is reported in Supplementary Material in Sec. D.2.

Method	Inter-Patient Dice ↑	Patient-to-Atlas Dice ↑
Affine	0.561±0.018	0.543±0.017
CycleMorph [32]	0.654±0.017	0.645±0.016
nnFormer [75]	0.626±0.018	0.631±0.016
PVT [64]	0.637±0.016	0.642±0.016
ViT-V-Net [8]	0.658±0.017	0.650±0.017
TransMorph-Bspl [10]	*0.670±0.018*	*0.666±0.016*
Proposed H-ViT	**0.704±0.016**	**0.694±0.015**

Table 4. The averaged Dice score results for LPBA registration for 56 anatomical structures over 120 random pairs for inter-patient and 117 pairs for patient-to-atlas registrations. The percentage of folded voxels for the reported techniques is below 0.2%. The detailed results are reported in Supplementary Material in Sec. D.4.

cross-attention units is comparable to that of self-attention units, with both contributing to a Dice score of approximately 0.80. This can be observed by comparing H-ViT with self-attention only (without cross-attention) to H-ViT with three cross-attention units but without self-attention. An increase in the number of cross-attention units leads to an improvement in the Dice score, signifying that the deformation field benefits from the flow of higher-level features into its representation. From a computational standpoint, the cross-attention units do not significantly impact model loading. The number of FLOPs increases from 1.7X in H-ViT without cross-attention to 2.2X for H-ViT with three cross-attention blocks. This trend is similarly reflected in the number of trainable parameters, with H-ViT without cross-attention containing 17.87M parameters compared to 21.23M for H-ViT with three cross-attention blocks.

H-ViT parameters: Tab. 7 presents the impact of various parameters of H-ViT on the IXI registration. In this experiment, we employed a small version of H-ViT with a reduced number of training steps, as described in Supplementary Sec. C.1. Tab. 7 indicates that an increased number of heads contributes to improved performance. Simi-

Method	CNN Backbone	Self-Attention	Cross-Attention (#)	Dice ↑	$\lvert J_\Phi\rvert \leq 0$ (%) ↓	Params. (#M)	Max. Mem. (GB)	FLOPs
H-ViT	✓	✗	✗	0.785±0.081	0.442±0.210	16.14	5.96	1.0X
H-ViT	✓	✓	✗	0.801±0.072	0.205±0.130	17.87	15.50	1.7X
H-ViT	✓	✓	1	0.803±0.072	0.216±0.133	19.55	22.60	1.9X
H-ViT	✓	✓	2	0.806±0.073	0.437±0.188	20.68	22.60	2.1X
H-ViT	✓	✓	3	**0.810±0.073**	0.525±0.209	21.23	22.60	2.2X
H-ViT	✓	✗	3	0.802±0.075	0.531±0.238	19.35	22.46	2.0X

Table 5. Ablation study on the dual-attention mechanism of H-ViT on the IXI dataset. The symbol 'X' denotes the number of FLOPs for H-ViT with only the CNN backbone that is 803.5G for an MRI scan with a size of $160 \times 192 \times 224$.

Method	Inter-Patient Dice ↑	Patient-to-Atlas Dice ↑
Affine	0.537±0.041	0.534±0.034
VoxelMorph [6]	0.674±0.197	0.666±0.201
CycleMorph [32]	0.679±0.194	0.671±0.199
CoTr [69]	0.633±0.214	0.630±0.218
ViT-V-Net [8]	*0.700±0.186*	0.695±0.187
TransMorph-Bspl [10]	0.699±0.181	*0.695±0.183*
Proposed H-ViT	**0.731±0.170**	**0.726±0.173**

Table 6. The averaged Dice score results for Mindboggle registration with 41 anatomical structures over 111 random pairs for inter-patient and 222 pairs for patient-to-atlas registrations. The percentage of folded voxels for the reported techniques is below 1.0%. The detailed results are reported in Supplementary in Sec. D.5.

Parameter	Dice ↑	$\lvert J_\Phi\rvert \leq 0$ (%) ↓
Number of Heads		
8	0.801±0.072	0.201±0.130
64	0.805±0.072	0.211±0.132
Depth		
1	0.803±0.073	0.201±0.132
4	0.805±0.071	0.222±0.135
Voxel Patch Size		
$2 \times 2 \times 2$	0.803±0.073	0.201±0.132
$6 \times 6 \times 6$	0.804±0.073	0.207±0.132
Drop rate		
0	0.803±0.073	0.201±0.132
0.2	0.801±0.073	0.179±0.128

Table 7. Ablation study on parameters of a small H-ViT for the IXI registration. The detailed results are reported in Supplementary Material in Tab. 9.

larly, elevating the depth enhances the performance of the smaller H-ViT model. Considering the scores for the voxel size experiment, opting for voxel patches of size 2 is advisable while using a drop rate is not advisable.

4.3. Discussion and Conclusion

This paper introduced H-ViT as a novel approach for registering medical imaging data. H-ViT benefits from a dual-attention mechanism, consisting of self-attention and cross-attention. Self-attention operates by interacting with local voxel patches, facilitating the capture of short-range flow information at a specific scale level. Cross-attentions draw voxel patches from higher-level feature levels, enabling the utilization of a broad spectrum of long-range flow information across scale levels in a hierarchical manner. The cross-attention mechanism enhances various facets of the representation of the deformable field, including:

- *Abstract patterns and information flow*: The hierarchical cross-attention module enables the recognition of complex flow patterns, facilitating the flow information stream across all scale levels in the deformation field, irrespective of their spatial proximity (Tab. 5).
- *Translation invariance*: In contrast to traditional Transformers that primarily depend on local patch embeddings, utilizing long-range flow patches in the cross-attention mechanism enables a greater level of translation invariance. This approach emphasizes significant features and patterns within distinct layers of flow without being limited by their specific layer positions.
- *Dealing with MRI scans of varying sizes*: The cross-attention mechanism in H-ViT guarantees the capture of relevant deformed patterns across diverse flow feature maps, irrespective of the deformation field dimension.

Broader impacts: The proposed H-ViT can be used as an effective network to enhance image representation in any conventional CNN and ViTs.

Acknowledgments

This work was supported by the Munich Center for Machine Learning (MCML), DFG, and BMBF. We gratefully acknowledge the computational resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 5

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *MICCAI*, pages 924–931. Springer, 2006. 3

[3] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007. 3

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[5] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *CVPR*, pages 9252–9260, 2018. 1, 2, 3

[6] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE TMI*, 38(8): 1788–1800, 2019. 1, 2, 5, 6, 7, 8, 3

[7] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, pages 357–366, 2021. 3

[8] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021. 1, 2, 5, 6, 7, 8, 3

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1

[10] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022. 1, 2, 5, 6, 7, 8, 3

[11] Jiashun Chen, Donghuan Lu, Yu Zhang, Dong Wei, Munan Ning, Xinyu Shi, Zhe Xu, and Yefeng Zheng. Deformer: Towards displacement field learning for unsupervised medical image registration. In *MICCAI*, pages 141–151. Springer, 2022. 2

[12] Junyu Chen, Yihao Liu, Yufan He, and Yong Du. Deformable cross-attention transformer for medical image registration. *arXiv preprint arXiv:2303.06179*, 2023. 1, 2, 3

[13] Junyu Chen, Yihao Liu, Yufan He, and Yong Du. Spatially-varying regularization with conditional transformer for unsupervised image registration. *arXiv preprint arXiv:2303.06168*, 2023. 2

[14] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, pages 16144–16155, 2022. 3

[15] Steffen Czolbe, Oswin Krause, and Aasa Feragen. Semantic similarity metrics for learned image registration. In *Medical Imaging with Deep Learning*, pages 105–118. PMLR, 2021. 2

[16] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019. 2

[17] Neel Dey, Mengwei Ren, Adrian V Dalca, and Guido Gerig. Generative adversarial registration for improved conditional deformable templates. In *ICCV*, pages 3929–3941, 2021. 1

[18] Zhipeng Ding and Marc Niethammer. Aladdin: Joint atlas building and diffeomorphic registration learning with pairwise alignment. In *CVPR*, pages 20784–20793, 2022. 1

[19] Jingfan Fan, Xiaohuan Cao, Pew-Thian Yap, and Dinggang Shen. Birnet: Brain image registration using dual-supervised fully convolutional networks. *Medical image analysis*, 54: 193–206, 2019. 2

[20] Xuan Gong, Luckyson Khaidem, Wentao Zhu, Baochang Zhang, and David Doermann. Uncertainty learning towards unsupervised deformable medical image registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2484–2493, 2022. 1, 2

[21] Daniel Grzech, Mohammad Farid Azampour, Ben Glocker, Julia Schnabel, Nassir Navab, Bernhard Kainz, and Loïc Le Folgoc. A variational bayesian method for similarity learning in non-rigid image registration. In *CVPR*, pages 119–128, 2022. 1, 2, 3

[22] Bo Guo, Liwei Deng, Ruisheng Wang, Wenchao Guo, Alex Hay-Man Ng, and Wenfeng Bai. Mctnet: Multiscale cross-attention based transformer network for semantic segmentation of large-scale point cloud. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

[23] Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention. *arXiv preprint arXiv:2306.06189*, 2023. 3

[24] Malte Hoffmann, Andrew Hoopes, Douglas N Greve, Bruce Fischl, and Adrian V Dalca. Anatomy-aware and acquisition-agnostic joint registration with synthmorph. *arXiv preprint arXiv:2301.11329*, 2023. 1

[25] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pages 3–17. Springer, 2021. 5, 6, 1

[26] Yanyan Huang, Weiqin Zhao, Shujun Wang, Yu Fu, Yuming Jiang, and Lequan Yu. Conslide: Asynchronous hierarchical interaction transformer with breakup-reorganize rehearsal for continual whole slide image analysis. In *ICCV*, pages 21349–21360, 2023. 3

[27] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008. 5, 1

[28] Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? In *International Work-*

shop on Machine Learning in Medical Imaging, pages 151–160. Springer, 2022. 2

[29] Bailiang Jian, Mohammad Farid Azampour, Francesca De Benetti, Johannes Oberreuter, Christina Bukas, Alexandra S Gersing, Sarah C Foreman, Anna-Sophia Dietrich, Jon Rischewski, Jan S Kirschke, et al. Weakly-supervised biomechanically-constrained ct/mri registration of the spine. In *MICCAI*, pages 227–236. Springer, 2022. 1

[30] Ankita Joshi and Yi Hong. R2net: Efficient and flexible diffeomorphic image registration using lipschitz continuous residual networks. *Medical Image Analysis*, 89:102917, 2023. 1

[31] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, and Salah A Baker. Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers. In *ICCV*, pages 3235–3245, 2021. 2

[32] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical Image Analysis*, 71:102036, 2021. 1, 2, 5, 6, 7, 8, 3

[33] Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: unsupervised deformable image registration using diffusion model. In *ECCV*, pages 347–364. Springer, 2022. 2

[34] Arno Klein, Satrajit S Ghosh, Forrest S Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, Brian Rossa, Martin Reuter, Elias Chaibub Neto, et al. Mindboggling morphometry of human brains. *PLoS computational biology*, 13(2): e1005350, 2017. 5, 1

[35] Julian Krebs, Tommaso Mansi, Boris Mailhé, Nicholas Ayache, and Hervé Delingette. Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 101–109. Springer, 2018. 2, 3

[36] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE TMI*, 38(9):2165–2176, 2019. 3

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[38] Yihao Liu, Lianrui Zuo, Shuo Han, Yuan Xue, Jerry L Prince, and Aaron Carass. Coordinate translator for learning deformable medical image registration. In *MMMI*, pages 98–109. Springer, 2022. 2

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 4

[40] Jinxin Lv, Zhiwei Wang, Hongkuan Shi, Haobo Zhang, Sheng Wang, Yilang Wang, and Qiang Li. Joint progressive and coarse-to-fine registration of brain mri via deformation field integration and non-rigid feature fusion. *IEEE TMI*, 41 (10):2788–2802, 2022. 6

[41] Bahram Marami, Benoit Scherrer, Onur Afacan, Simon K Warfield, and Ali Gholipour. Motion-robust reconstruction based on simultaneous multi-slice registration for diffusion-weighted mri of moving subjects. In *MICCAI*, pages 544–552. Springer, 2016. 1

[42] 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. 5, 1

[43] Sasan Matinfar, Mehrdad Salehi, Daniel Suter, Matthias Seibold, Shervin Dehghani, Navid Navab, Florian Wanivenhaus, Philipp Fürnstahl, Mazda Farshad, and Nassir Navab. Sonification as a reliable alternative to conventional visual surgical navigation. *Scientific Reports*, 13(1):5930, 2023. 1

[44] Mingyuan Meng, Michael Fulham, Dagan Feng, Lei Bi, and Jinman Kim. Autofuse: Automatic fusion networks for deformable medical image registration. *arXiv preprint arXiv:2309.05271*, 2023. 2

[45] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *CVPR*, pages 4644–4653, 2020. 2

[46] Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *CVPR*, pages 20835–20844, 2022. 2

[47] Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural network. In *MICCAI*, pages 35–45. Springer, 2021. 2, 3, 6

[48] Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. Metric learning for image registration. In *CVPR*, pages 8463–8472, 2019. 2

[49] Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. In *Medical Imaging with Deep Learning*, 2021. 5, 6, 7, 2, 3

[50] Mengwei Ren, Neel Dey, Martin Styner, Kelly Botteron, and Guido Gerig. Local spatiotemporal representation learning for longitudinally-consistent neuroimage analysis. *NeurIPS*, 35:13541–13556, 2022. 1

[51] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: learning deformable image registration using shape matching. In *MICCAI*, pages 266–274. Springer, 2017. 2

[52] Robin Sandkühler, Christoph Jud, Simon Andermatt, and Philippe C Cattin. Airlab: autograd image registration laboratory. *arXiv preprint arXiv:1806.09907*, 2018. 2

[53] F. Sauer. Image registration: Enabling technology for image guided surgery and therapy. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 7242–7245, 2005. 1

[54] 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. 5, 1

[55] Zhengyang Shen, François-Xavier Vialard, and Marc Niethammer. Region-specific diffeomorphic metric mapping. *NeurIPS*, 32, 2019. 2

[56] Jiacheng Shi, Yuting He, Youyong Kong, Jean-Louis Coatrieux, Huazhong Shu, Guanyu Yang, and Shuo Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. In *MICCAI*, pages 217–226. Springer, 2022. 2

[57] Afshin Shoeibi, Marjane Khodatars, Mahboobeh Jafari, Navid Ghassemi, Parisa Moridian, Roohallah Alizadesani, Sai Ho Ling, Abbas Khosravi, Hamid Alinejad-Rokny, HK Lam, et al. Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review. *Information Fusion*, 2022. 1

[58] Hanna Siebert, Lasse Hansen, and Mattias P Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In *MICCAI*, pages 174–179. Springer, 2021. 6

[59] Shanlin Sun, Kun Han, Deying Kong, Hao Tang, Xiangyi Yan, and Xiaohui Xie. Topology-preserving shape reconstruction and registration via neural diffeomorphic flow. In *CVPR*, pages 20845–20855, 2022. 1

[60] Maarten L Terpstra, Matteo Maspero, Alessandro Sbrizzi, and Cornelis AT van den Berg. A symmetric loss function for magnetic resonance imaging reconstruction and image registration with deep learning. *Medical Image Analysis*, 80: 102509, 2022. 2

[61] Minh Q Tran, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Light-weight deformable registration using adversarial learning with distilling knowledge. *IEEE TMI*, 41(6):1443–1453, 2022. 2

[62] Jian Wang and Miaomiao Zhang. Deepflash: An efficient network for learning-based medical image registration. In *CVPR*, pages 4444–4452, 2020. 1, 2

[63] Monan Wang and Pengcheng Li. A review of deformation models in medical image registration. *Journal of Medical and Biological Engineering*, 39(1):1–17, 2019. 1

[64] 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 *ICCV*, pages 568–578, 2021. 4, 5, 6, 7, 2, 3

[65] Wenxiao Wang, Wei Chen, Qibo Qiu, Long Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wei Liu. Crossformer++: A versatile vision transformer hinging on cross-scale attention. *IEEE TPAMI*, 2023. 3

[66] Yibo Wang, Wen Qian, Mengqi Li, and Xuming Zhang. A transformer-based network for deformable medical image registration. In *CAAI International Conference on Artificial Intelligence*, pages 502–513. Springer, 2022. 1

[67] Yinsong Wang, Huaqi Qiu, and Chen Qin. Conditional deformable image registration with spatially-variant and adaptive regularization. *arXiv preprint arXiv:2303.10700*, 2023. 2

[68] Bastian Wittmann, Fernando Navarro, Suprosanna Shit, and Bjoern Menze. Focused decoding enables 3d anatomical detection by transformers. *Machine Learning for Biomedical Imaging*, 2:72–95, 2023. 3

[69] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI*, pages 171–180. Springer, 2021. 5, 6, 8, 2, 3

[70] Xuzhe Zhang, Xinzi He, Jia Guo, Nabil Ettehadi, Natalie Aw, David Semanek, Jonathan Posner, Andrew Laine, and Yun Wang. Ptnet: A high-resolution infant mri synthesizer based on transformer. *arXiv preprint arXiv:2105.13993*, 2021. 2

[71] Xiaosong Zhang, Yunjie Tian, Wei Huang, Qixiang Ye, Qi Dai, Lingxi Xie, and Qi Tian. Hivit: Hierarchical vision transformer meets masked image modeling. *arXiv preprint arXiv:2205.14949*, 2022. 3

[72] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *MICCAI*, pages 129–138. Springer, 2021. 1, 2

[73] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *ICCV*, pages 10600–10610, 2019. 1, 2

[74] 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. 1, 2

[75] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 3, 5, 6, 7, 2

[76] S. Zhou et al. Self-distilled hierarchical network for unsupervised deformable image registration. *IEEE TMI*, 2023. 3

[77] Yongpei Zhu and Shi Lu. Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer. In *MICCAI*, pages 78–87. Springer, 2022. 2