2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3315-1083-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/WACV61041.2025.00137

EFFICIENTMORPH: Parameter-Efficient Transformer-Based Architecture for 3D Image Registration

Abu Zahid Bin Aziz * Mokshagna Sai Teja Karanam* Tushar Kataria
Shireen Y. Elhabian**
{zahid.aziz,mkaranam,tushar.kataria,shireen}@sci.utah.edu
Scientific Computing and Imaging Institute & Kahlert School of Computing
University of Utah, Salt Lake City, UT, USA
***Equal Contribution**, ****Corresponding Author**

Abstract

Transformers have emerged as the state-of-the-art architecture in medical image registration, outperforming convolutional neural networks (CNNs) by addressing their limited receptive fields and overcoming gradient instability in deeper models. Despite their success, transformer-based models require substantial resources for training, including data, memory, and computational power, which may restrict their applicability for end users with limited resources. In particular, existing transformer-based 3D image registration architectures face two critical gaps that challenge their efficiency and effectiveness. Firstly, although window-based attention mechanisms reduce the quadratic complexity of full attention by focusing on local regions, they often struggle to effectively integrate both local and global information. Secondly, the granularity of tokenization, a crucial factor in registration accuracy, presents a performance trade-off: smaller voxel-size tokens enhance detail capture but come with increased computational complexity, higher memory usage, and a greater risk of overfitting. We present EFFICIENTMORPH, *a transformer-based architecture for unsupervised 3D image registration that balances local and global attention in 3D volumes through a plane-based attention mechanism and employs a Hi-Res tokenization strategy with merging operations, thus capturing finer details without compromising computational efficiency. Notably,* EFFICIENTMORPH *sets a new benchmark for performance on the OASIS dataset with ∼16-27× fewer parameters. https://github.com/MedVIC-Lab/Efficient_Morph_Registration*

1. Introduction

3D image registration [23, 60] is a critical task for various medical imaging applications in fields such as image-guided surgery [2], radiation therapy planning [49], image fusion for multimodality imaging [27], and quality enhancement [6]. Registration involves determining the spatial alignment between two volumes, typically referred to as the *fixed* and *moving* images, by identifying correspondences between similar structures or features and their relative spatial positions. Conventional approaches such as ANTs [5], Elastix [36], and NiftiReg [45] employ optimization-based frameworks [4, 20, 22, 67]. This iterative search for the optimal transformation makes these methods inherently slow, especially when dealing with large datasets or high-resolution images [33, 37, 47]. To address these challenges, there is increasing interest in adopting learning-based approaches. In particular, deep learning methods offer significantly faster inference times and currently achieve state-of-the-art performance in 3D image registration [7, 9, 28].

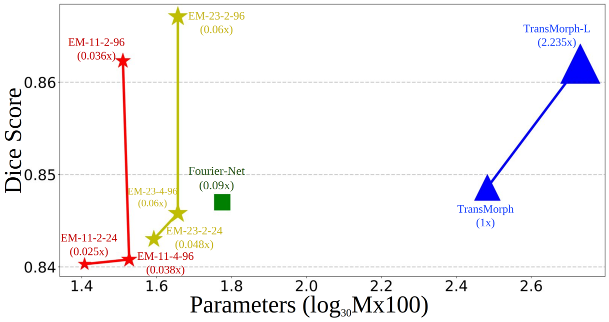

Figure 1. **Parameter Count Comparisons with performance on OASIS Dataset.** The proposed variants are formatted as EfficientMorph-11-stride-C and EfficientMorph-23-stride-C. Comparison of parameter count in millions(M) and Dice scores between the proposed variants and baselines.

Learning-based registration methods predominantly rely on convolutional architectures (e.g., [3, 7, 28, 35, 46]), using U-Net-based architectures to generate the deformation fields. However, the effectiveness of convolutional lay-

2642-9381/25/$31.00 ©2025 IEEE
DOI 10.1109/WACV61041.2025.00137

ers for registration tasks can be compromised due to their limited receptive fields that hinder capturing global context [9] and their increased susceptibility to vanishing gradients as network depth grows to enhance learning capacity [21]. Since the advent of Vision Transformers [59], transformer-based architectures have shown superior performance across various tasks, such as classification, segmentation, and registration [9, 15, 59, 69], thanks to their long-range modeling capabilities. In particular, they offer promising mitigations to CNN limitations. Specifically, transformers leverage global contextual information through self-attention mechanisms and provide more stable gradient flow across layers via techniques such as layer normalization and skip connections that are integral to transformers' design [9,15,59]. Despite their success, transformers' advantages come at the expense of a significant increase in parameter count, requiring approximately 10 to 20 times more parameters than convolutional counterparts [9, 19], making them impractical for deployment to end-user applications.

Specifically, existing transformer-based registration methods, including TransMorph [9], the current state-of-the-art transformer-based model for medical image registration, encounter *two* main significant limitations that compromise its efficiency and overall performance. *Firstly*, windowed attention approaches (e.g., the Swin transformer [40] backbone used in TransMorph [9]) optimize computational efficiency through local attention and shifted windows, enhancing interactions between adjacent windows. However, this limits global context capture, particularly in shallow layers, due to within-window constraints(masks for calculating attention) compared to methods that interact globally. *Secondly*, the pixel granularity of tokenization plays a crucial role in registration accuracy. To fit within available GPU memory, tokenization is applied to downsampled volumes, with each dimension reduced by 4 to 8. Increasing the token resolution within the same volume can capture finer details, but it also escalates computational and memory demands due to the corresponding rise in the number of tokens. [9, 34, 59]. Existing multi-resolution architectures, such as GradICON [58], HRNet [63], and HRFormer [73], leverage features at various resolutions to enhance performance. However, these approaches necessitate the simultaneous training of multiple models to optimize performance at each resolution level. These training methods significantly increase complexity and computational cost, leading to a substantial rise in the number of parameters. As a result, the models become more resource-intensive and challenging to deploy in environments with limited resources.

In this paper, we propose EFFICIENTMORPH, a novel transformer-based framework for unsupervised 3D image registration that addresses the aforementioned challenges. We introduce a *plane attention* mechanism inspired by 3D anatomical views (coronal, sagittal, and axial) to enhance the balance between local and global feature recognition [24]. We propose Hi-Resolution tokenization to capture finer image details. To further reduce model complexity within the encoded representation, we introduce a method for merging neighboring tokens in a high-resolution feature space, thereby decreasing the computational load of self-attention calculations. By integrating Hi-Res tokenization, EFFICIENTMORPH becomes a highly parameter-efficient registration architecture (see Figure 1A). Additionally, we introduce a multi-resolution EFFICIENTMORPH, which concatenates latent features from different resolutions to produce more precise deformation fields. This approach leverages multi-resolution data without needing to train separate models.

The main contributions of this paper are:

- A novel attention module for 3D registration that focuses attention across the coronal (xy), sagittal (yz), or Axial (zx) planes within a single transformer block.
- A Hi-Resolution tokenization mechanism to encode high-resolution voxel features without increasing computation complexity.
- Proposed Multi-Resolution EFFICIENTMORPH leverages the concatenation of multi-resolution latent space features to enhance model performance.
- A new parameter-efficient architecture achieves performance within ±0.05 Dice score of existing methods, surpassing state-of-the-art on 2 out of 3 datasets (single and multi-modal Registration), with 16-27x fewer parameters (Figure 1) and 5x faster convergence. Comprehensive ablation studies on regularization losses, attention mechanisms, and key design choices are also presented.

2. Related Works

3D Volume Registration. Learning-based approaches for 3D image registration can generally be divided into two main categories: supervised and unsupervised. *Supervised methods* [53, 54, 71] require estimates of deformation fields derived from traditional optimization-based approaches, the acquisition of which can be prohibitively costly for large datasets. Moreover, the efficacy of supervised approaches is contingent upon the availability of high-quality deformation fields for supervised training, with their performance capped by the accuracy of the method used to obtain these fields. In contrast, *unsupervised methods* do not require deformation fields and use image similarity as a self-supervised signal to train a registration network. Most unsupervised 3D registration methods [7,9,28,43,46] are trained to produce a 3D deformation field that is then used to transform (or warp) the moving image. Loss (L1 or L2) between the warped moving image and the fixed image is used to train the network. With sufficient data and training time, the model learns to produce realistic deformation fields that outperform optimization-based methods in accuracy and inference speed [9, 28]. Additionally, unsupervised methods incorporate regularization losses to promote spatial smoothness in the deformation field, often employing techniques such as bending energy [30, 51], total variation minimization [61], and consistency penalties [57, 58, 74], among others. To enhance the accuracy of registration, segmentations of the underlying anatomies are incorporated as regularization losses [9, 28]. However, this approach makes the registration problem fully or semi-supervised due to the requirement for manual segmentation. Ideally, unsupervised registration methods should perform effectively without the need for additional supervision. We present a parameter-efficient registration architecture that outperforms state-of-the-art models on three public datasets while significantly reducing the number of parameters. Our proposed model not only performs well in unsupervised 3D volume registration but can also leverage available segmentation data to outperform state-of-the-art models, all while maintaining a lower parameter count.

Efficient Transformer Attention Architectures. As deep learning models continue to grow in size each year [1, 16, 62, 68], deploying them on end-user devices such as mobile phones or desktops becomes increasingly impractical. Users often need access to a server API for model inference or a local desktop with substantial computing power to run these models. These requirements restrict the applicability of many deep learning models, particularly in medical applications where data privacy is paramount [29, 44, 72]. In many cases, patient data cannot be transferred to a server, requiring computations to be performed locally to comply with HIPAA guidelines. As a result, developing efficient architectures that preserve the accuracy of large models while being deployable on end-user devices is both essential and highly relevant. Examples of such architectures and methods proposed for efficient deep learning models include EfficientNet [55], MobileNet [26], LLM-pruner [41], GPTQ [18], and Mobillama [56] [8, 32, 64, 66].

Applying transformer self-attention to high-resolution medical images presents substantial computational challenges due to its quadratic complexity with respect to input size [34, 59, 75], making it difficult and resource-intensive to scale to large datasets and model sizes. As a result, these models often cannot be deployed in end-user applications in hospitals and clinics. Various strategies have been developed to address the computational challenges of applying transformer self-attention to high-resolution images. An effective strategy involves optimizing the attention matrix through techniques such as approximations—like Linformer [65], Memory efficient attention [48], and sparse attention [12, 50]—or by limiting exact attention to localized windows, as seen in models like the SWIN Transformer [40]. Moreover, efficiency can be significantly improved through GPU optimizations, as demonstrated by Flash Attention [13, 14]. An alternative strategy is to stack multiple sparse attention layers with restricted contexts, which allows overlapping layers to achieve full-context modeling. For example, the Strided Sparse Transformer [17] employs custom GPU kernels to implement block-sparse matrix multiplications, enhancing computational efficiency. Similarly, the Axial Transformer [24] maintains full conditioning contexts by processing both masked and unmasked tokens during each decomposition stage. In contrast, our proposed module is specifically designed for 3D medical volumes, introducing a novel 3D *plane-based* attention mechanism that selectively operates on a subset of planes (*axial, sagittal, coronal*) within each decomposition block. This approach allows for the creation of models with significantly fewer parameters. Furthermore, by utilizing high-resolution voxel tokens, our model matches and surpasses the performance of state-of-the-art models.

3. Methods

Given a 3D volume represented as $\mathbf{A} \in \mathbb{R}^{H\times W\times D}$, where H, W, and D denote the height, width, and depth dimensions, respectively. If each voxel is mapped into the latent space, the number of tokens for the volume would be *WHD*, which may exceeds the memory capacity of a GPU for a single sample. Therefore, strided convolutions are used in the patch embedding layer (with stride s) to project the voxels in $\mathbf{A}$ into a high-dimensional feature space, resulting in $\mathbf{A'} \in \mathbb{R}^{H'\times W'\times D'\times C}$, where C is the embedding dimension, $(H', W', D') = (\frac{H}{s}, \frac{W}{s}, \frac{D}{s})$. The resulting feature space is tokenized to train the downstream transformer layers. In the following sections, we detail the Hi-Res tokenization process, and the plane attention mechanism of the proposed EFFICIENTMORPH, as illustrated in Figure 2A. We have provided a comprehensive explanation of the overall training process in Sections 6.2.2 and 6.2.3 of the supplementary materials.

3.1. Hi-Res Tokenization

For a fixed embedding dimension C, using each voxel of a 3D volume of size 1x1x1 for tokenization would create N tokens, where $N = H \times W \times D$. Voxel-level tokenization results in attention matrices of more than a trillion parameters with a complexity of $\mathcal{O}(N^2)$. Transformer architectures for images often rely on s-strided convolutions (e.g., $s = 4$ [9]) for volume tokenization and patch embedding, trading off computational complexity, which is now $\mathcal{O}\left(\left(\frac{N}{s^3}\right)^2\right)$, at the cost of detailed features. However, fine-grained spatial information is critical for medical segmentation and registration tasks, which may be lost due to strided

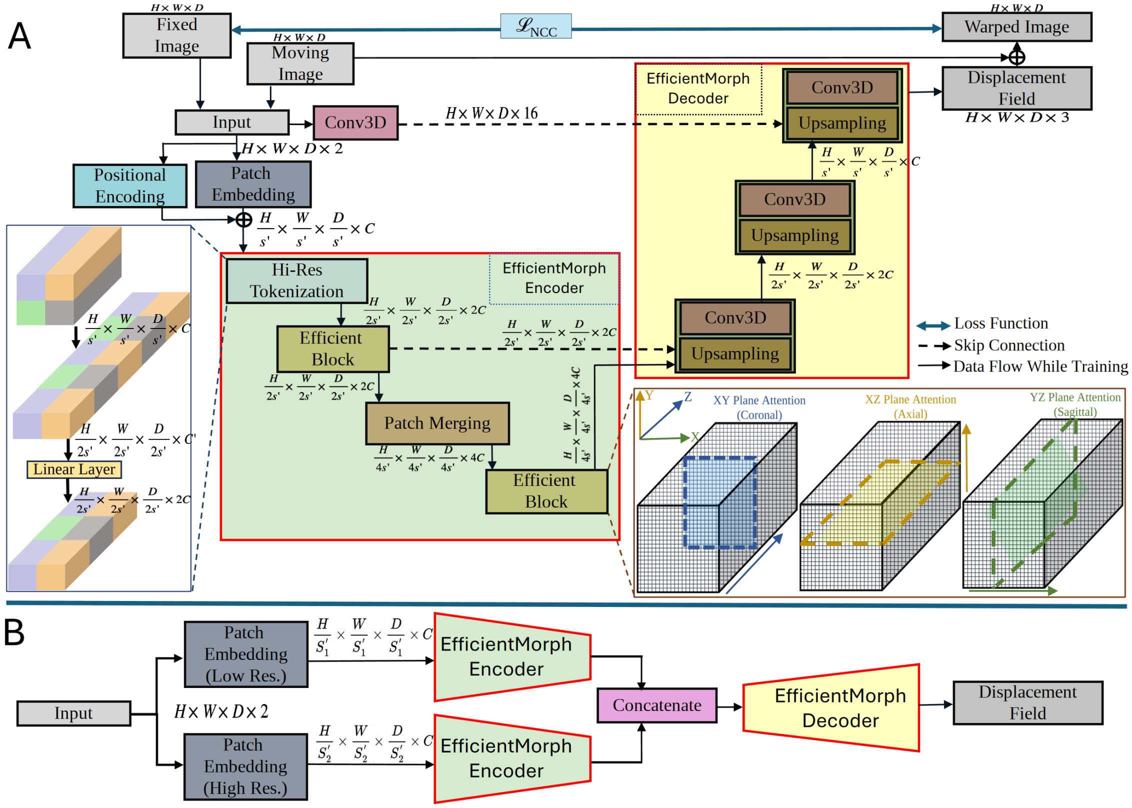
A
H × W × D
Fixed Image
$\mathscr{L}_{NCC}$
H × W × D
Warped Image
H × W × D
Moving Image
Input
Conv3D
H × W × D × 16
H × W × D × 2
EfficientMorph Decoder
Conv3D
Upsampling
Displacement Field
H × W × D × 3
Positional Encoding
Patch Embedding
$\frac{H}{s'} \times \frac{W}{s'} \times \frac{D}{s'} \times C$
Hi-Res Tokenization
EfficientMorph Encoder
Efficient Block
Patch Merging
Efficient Block
$\frac{H}{2s'} \times \frac{W}{2s'} \times \frac{D}{2s'} \times 2C$
$\frac{H}{4s'} \times \frac{W}{4s'} \times \frac{D}{4s'} \times 4C$
$\frac{H}{2s'} \times \frac{W}{2s'} \times \frac{D}{2s'} \times C'$
Linear Layer
Loss Function
Skip Connection
Data Flow While Training
Y
Z
X
XY Plane Attention (Coronal)
XZ Plane Attention (Axial)
YZ Plane Attention (Sagittal)
B
Input
H × W × D × 2
Patch Embedding (Low Res.)
Patch Embedding (High Res.)
$\frac{H}{S_1'} \times \frac{W}{S_1'} \times \frac{D}{S_1'} \times C$
$\frac{H}{S_2'} \times \frac{W}{S_2'} \times \frac{D}{S_2'} \times C$
EfficientMorph Encoder
EfficientMorph Encoder
Concatenate
EfficientMorph Decoder
Displacement Field

Figure 2. EFFICIENTMORPH **Architecture.** (A) EFFICIENTMORPH utilizes utilizes plane attention mechanism on the whole volume as shown in *Efficient Transformer Block*. We use different numbers and types of plane attentions (xy, yz, or zx planes) for each block in the transformer backbone (Table 1). Hi-Res Tokenization is shown in the left end of the figure. (B) Shows the architectural modification for multi-resolution variant where $S_1^{'} > S_2^{'}$.

convolutions.

We propose *Hi-Resolution tokenization* strategy that uses a smaller stride ($s' < 4$) within the embedding layer, thereby utilizing better spatial information available in the volume. However, this increases the number of tokens increasing the computational complexity. To reduce the computation complexity, we propose a token merging operation. These high-resolution tokens are positionally encoded and merged by grouping and concatenating the features of adjacent non-overlapping $d \times d \times d$ voxel token blocks (along the embedding dimension), resulting in $N' = \frac{H}{d} \times \frac{W}{d} \times \frac{D}{d}$ tokens with an embedding dimension of $C' = C \times d^3$. Then, C' is projected into a linear layer to attain a reduced embedding dimension of $C \times d$, as shown in Hi-Res tokenization block in Figure 2A. This approach to tokenization enables us to use high-resolution features and reduces the overall complexity of the model. Refer to Figure 2A for pictorial depiction of the Hi-Res Tokenization process.

3.2. Plane Attention Mechanism

Despite using Hi-Res tokenization, the number of tokens generated from each volume remains high. Running full attention on these tokens, while feasible, demands considerable computational resources. We introduce a novel attention framework called *plane attention* to address this challenge. Instead of performing full 3D attention on all tokens, this method utilizes attention along coronal (xy), sagittal (yz), or axial (zx) planes, as shown in Figure 2A. Although attention confines focus to a specific plane, EFFICIENTMORPH achieves volume attention by sequentially employing different attention combinations xy followed by yz or zx, thus covering all plane directions.

$$\text{Attn}(\mathbf{A}'_{dim}) = \text{softmax}\left(\frac{\mathbf{Q}_{dim}\mathbf{K}^T_{dim}}{\sqrt{d_k}}\right)\mathbf{V}_{dim} \quad (1)$$

Here, $dim \in \{xy, yz, zx\}$, $\mathbf{A}'_{dim}$ can be represented as $\mathbf{A}'_{xy} \in \mathbf{R}^{H' \times W' \times C}$, $\mathbf{A}'_{yz} \in \mathbf{R}^{W' \times D' \times C}$ and $\mathbf{A}'_{zx} \in \mathbf{R}^{D' \times H' \times C}$ for xy, yz, and zx planes, respectively. By decomposing the 3D attention into 2D plane attention, the proposed attention mechanism significantly reduces the parameter count while preserving the ability to capture essential volumetric features necessary for registration. Figure 2A shows different plane attention blocks across the efficient transformer block.

3.3. Multi-Resolution EfficientMorph

HiRes Tokenization effectively harnesses high-resolution spatial information early in the network but limits all tokens to a single resolution. However, prior work [57, 58, 63, 73] has demonstrated that multi-resolution features significantly improve registration accuracy. Therefore, we propose a multi-resolution variant for EfficientMorph.

Multi-resolution EfficientMorph processes the input image along two distinct paths, each with its patch embedding block tailored to tokenize patches of different sizes ($S_1^{'}$ and $S_2^{'}$, where $S_1^{'} > S_2^{'}$). This approach captures patches at multiple resolutions, enhancing the richness of the feature representation. Instead of training the entire architecture simultaneously, a phased training strategy is adopted. In this process, patches pass through two stages of Efficient Transformer blocks, as illustrated in Figure 2B, where their latent dimensions are merged. The merged representation is then fed into the decoder, ensuring that the output integrates comprehensive information from both resolution levels. This methodology not only improves the model's ability to capture fine details but also enhances computational efficiency during training.

4. Results and Discussion

4.1. Datasets and Implementation Details

OASIS Brain MRI. We evaluated EfficientMorph on the publicly available dataset OASIS [42], obtained from the Learn2Reg challenge [23] for inter-patient registration and pre-processed from [25]. It has a total of 451 brain T2 MRI images. Among these, 394, 19, and 38 scans are used for training, validation, and testing, respectively.

ReMIND2Reg. This dataset aims to register 3D iUS images with either ceT1 or T2 MRI images to account for brain shift during tumor resection, requiring models to handle large deformations and missing data scenarios. The dataset is divided into 155 image pairs for training, 10 image pairs for validation, and 40 for testing.

Atlas-to-Patient Brain MRI (IXI). We additionally evaluated the proposed model on IXI dataset that contains 600 MRI scans. Among these, 576 T1-weighted brain MRI images were employed as moving images, while the fixed image for this task was an atlas brain MRI [35]. The dataset was partitioned into training, validation, and test sets, comprising 403, 58, and 115 volumes, respectively. For more details on datasets, refer to supplementary section 6.2.1

Implementation Details. EfficientMorph was trained on NVIDIA A100 GPUs with 40GB RAM and a batch size of 1. We used the same splits for both datasets as the existing works [9, 28]. We limited training epochs to 100 to prioritize parameter efficiency and quick convergence within resource limits. We used the Adam optimizer with a learning rate of $5e^{-4}$ for OASIS & Remind2Reg and $3e^{-4}$ for IXI. We used a cosine annealing scheduler for OASIS and stepLR for IXI. We evaluated different variants: EfficientMorph-11, which includes one plane attention transformer (*xy*, *yz*, or *zx*) per efficient block as shown in Figure 2, and EfficientMorph-23, which features two plane attention transformers in the first efficient block and three in the second. The specific plane attentions used in these variants are detailed in Table 1. Note that no data augmentation was applied during training.

Table 1. **EfficientMorph Variants.** EfficientMorph-AB denotes a configuration with A plane attention transformers in the first efficient block and B plane attention transformers in the second efficient block.

Variants	Planes
EfficientMorph-11	(xy, yz)
EfficientMorph-23	(xy-yz, xy-yz-zx)

Loss function. In the unsupervised registration setting, we utilized normalized cross-correlation with bending energy regularization, consistent with other registration frameworks in the literature [9, 11, 28]. Let I_F and I_M be the fixed and moving image volumes and S_F and S_M be the associated anatomy segmentations (if available).

$$\mathcal{L}_{\text{UnSupReg}} = L_{NCC}(I_F, \text{Warp}(I_M)) + \text{BendingEnergy}$$

To ensure a thorough comparison on the OASIS dataset, we also incorporated an additional segmentation loss (Dice coefficient) during training, aligning with the approaches used in other methods [7, 9, 28].

$$\begin{aligned}\mathcal{L}_{\text{OASIS}} =& L_{NCC}(I_F, \text{Warp}(I_M)) + \text{BendingEnergy} \\ &+ \text{DiceLoss}(S_F, \text{Warp}(S_M))\end{aligned}$$

Comparisons Methods. We compare EfficientMorph with convolutional-based methods, including VoxelMorph [7] and Fourier-Net [28], as well as different Tranformer based methods such as TransMorph [9], including TransMorph-Tiny, TransMorph, and TransMorph-L, TransMatch [11] and Vit-V-Net [10]. All methods were trained on the same GPU as previously mentioned, using their original implementations.

Evaluation Metrics. To evaluate the results on the OASIS and IXI datasets, we utilized the Dice score for anatomical segmentation (35 regions for OASIS and 29 for IXI) and computed the percentage of negative values of Jacobian determinant. For the ReMIND2Reg, we used the Learn2Reg [23] leaderboard evaluation system, where the output deformation fields were submitted to obtain Target Registration Error (TRE) and percentage of negative values of Jacobian determinant of deformation.

4.2. Experimental Results

OASIS Results. The results on the OASIS dataset are shown in Table 2. Among the variants, EfficientMorph-23 achieves the highest Dice score with just 2.8M parameters—16 times fewer than TransMorph and 8 times fewer than TransMatch—outperforming all compared baselines, including TransMorph-L, which has over 100M parameters. Despite having fewer parameters, EfficientMorph-11 delivers comparable performance to the other baselines. Both variants maintain consistently low percentage of negative values in Jacobian determinant of deformation, demonstrating that even with fewer parameters, EfficientMorph learns a more robust representation of the underlying data, leading to superior registration performance. Table 2 further demonstrates the results when segmentation loss is added as an additional training loss for the registration network. While all models show improved accuracy with added segmentation supervision, EfficientMorph still outperforms all others, achieving the highest average Dice score and the lowest Jacobian determinant score.

Supplementary Figure 6 presents a comparison of dice scores between the EfficientMorph variants and the baseline across different brain MR substructures, highlighting significant improvements with our proposed models. Our models consistently achieve the highest test dice scores across all brain segments. Supplementary Figure 5 also provides qualitative results of the segmentations obtained after registration of three anatomies, along with their corresponding dice scores. The figure includes the best, median, and worst-performing cases for analysis. Notably, the worst-performing case, characterized by a fixed image with non-smooth boundaries, challenges all models in registration accuracy; however, our proposed model still outperforms others, achieving the highest Dice score for the case.

EfficientMorph has a significantly low parameter count and coverges faster. Figure 1 (Page 1) shows that compared to TransMorph [9], EfficientMorph proposed architectures have between 2.5-6% of the total parameters, however with comparable and even better dice scores for EfficientMorph-23 variant. Similarly, when compared with Fourier-Net [28], Efficient morph EM-11-2-96 has $\frac{1}{3}$rd parameter count and with higher dice score. These results clearly show that EfficientMorph achieves better registration accuracy than other models and a very low parameter count. Models trained with segmentation loss were used for this analysis, as using an extra loss doesn't have an impact on number of parameters.

The convergence curves in Figure 3 clearly show that TransMorph learns quickly in a few initial epochs but then slowly saturates to the final performance, whereas all EfficientMorph variants slowly and steadily converge to higher dice scores. EfficientMorph starts to outperform TransMorph by a significant margin as early as 10 epochs. This result clearly shows that efficient morph is not only parameter efficient but requires less compute for converging to a better solution.

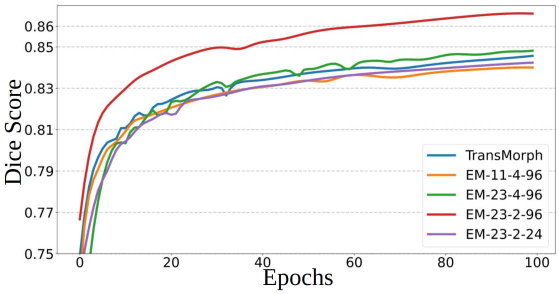

Figure 3. **Convergence Curves.** The proposed variants are formatted as EfficientMorph-11-stride-C and EfficientMorph-23-stride-C. Dice score curves of EfficientMorph variants as a function of epochs.

Multiresolution EfficientMorph is better for Unsupervised Registration. We employed a multi-resolution architecture to enhance unsupervised registration results on the OASIS dataset. The outcomes, detailed in Table 3, demonstrate that incorporating multi-resolution features improves registration accuracy across all cases. Notably, the best performance is achieved with EfficientMorph when using patch embedding blocks with strides of 2 and 4 (refer to Figure 2B).

ReMIND2Reg Results. The *ReMIND2Reg* dataset presents two key challenges: (a) It comprises multi-modal data, which introduces complexity in processing and analysis, and (b) It has a significantly smaller number of training samples—approximately half of those available in the OASIS dataset—further increasing the difficulty of achieving accurate results. Table 4 presents the target registration error (TRE) and Jacobian determinant results on the ReMIND2Reg dataset. EfficientMorph achieves the lowest TRE and the smallest percentage of negative values in Jacobian determinant of deformation among all methods. This indicates that EfficientMorph not only excels in efficiency, with a much smaller model size, but also learns superior representations for multi-modal registration. Additionally, EfficientMorph demonstrates robustness to limited dataset sizes, further widening the performance gap on smaller datasets.

IXI Results. Results of the IXI dataset are presented in Supplementary Table 9. EfficientMorph outperforms traditional optimization-based methods such as SyN, NiftiReg, and various convolutional-based approaches such as VoxelMorph-H [7] and CycleMorph [35] by a significant margin. EfficientMorph variants EM-11 and EM-23 with 4x4x4 strides achieve comparable performance (within ±0.003) with less than 3 million parameters compared to TransMorph's 46 million parameters and 5× fewer

Table 2. **OASIS Registration Results Using Single Resolution.** Average Dice Score Evaluated over 35 anatomies and percentage of negative values in Jacobian determinant of deformation are obtained for all test samples. *w/o Seg Loss* is the full unsupervised registration setting where only similarity and regularization loss between fixed and moving images are used for training. For *Seg Loss* setting, segmentation loss between segmentation of fixed and moving image anatomies are also used for training. *Param* are listed in Millions of parameters used for training the model. We can clearly see that EFFICIENTMORPH performs on par or better than other models with fewer parameters.* indicates the performance numbers taken from TransMorph [9] and Fourier-Net [28].

				w/o Seg Loss		**with Seg Loss**	
Methods	**stride**	**C**	**Param**	**Dice** ↑	$\mathbf{\|J\|} < 0\%$ ↓	**Dice** ↑	$\mathbf{\|J\|} < 0\%$ ↓
VoxelMorph [7]	-	-	0.063	0.6783 ± 0.039	2.981 ± 0.105	0.78 ± 0.024	0.1304±0.011
Fourier-Net [28]	-	-	4.19	0.770 ± 0.021	0.031 ± 0.003	0.847±0.013*	-
ViT-V-Net [10]	8x8x8	252	9.8	0.3632±0.0072	0.0149 ± 0.0001	0.4659±0.0052	0.1272 ± 0.0145
TransMatch [11]	4x4x4	96	26.39	0.4037 ± 0.055	0.1167 ± 0.0082	0.4612±0.0582	0.0546 ± 0.0011
TransMorph-Tiny [9]	4x4x4	6	0.24	0.441 ± 0.021	0.013 ± 0.001	0.801±0.056	0.081 ± 0.010
TransMorph [9]	4x4x4	96	46.5	0.801 ± 0.003	0.03 ± 0.002	0.8458±0.0137	0.119 ± 0.019
TransMorph-L [9]	4x4x4	128	108.11	0.804 ± 0.024	**0.009 ± 0.001**	0.862±0.014*	0.128 ± 0.021*
EfficientMorph-23	4x4x4	96	2.8	0.796 ± 0.035	0.091 ± 0.0006	0.846 ± 0.013	0.125 ± 0.020
EfficientMorph-11	2x2x2	96	**1.8**	0.803 ± 0.070	**0.011 ± 0.002**	**0.8623±0.0133**	**0.010 ± 0.001**
EfficientMorph-23	2x2x2	96	**2.8**	**0.810 ± 0.062**	**0.010 ± 0.001**	**0.870 ± 0.016**	**0.017 ± 0.001**

Table 3. **Multiresolution Unsupervised Registration Results on OASIS.**

	Methods	**stride**	**Param**	**Dice** ↑
Single Res	EM-11	$(2 \times 2 \times 2)$	1.8	0.803 ± 0.070
	EM-11	$(4 \times 4 \times 4)$	1.8	0.795 ± 0.071
	EM-11	$(8 \times 8 \times 8)$	1.8	0.765 ± 0.021
	EM-23	$(2 \times 2 \times 2)$	2.8	0.810 ± 0.062
	EM-23	$(4 \times 4 \times 4)$	2.8	0.796 ± 0.067
	EM-23	$(8 \times 8 \times 8)$	2.8	0.768 ± 0.026
Multi Res	EM-11	$(2 \times 2 \times 2)$,$(4 \times 4 \times 4)$	6.8	**0.820 ± 0.041**
	EM-11	$(2 \times 2 \times 2)$,$(8 \times 8 \times 8)$	6.8	**0.821 ± 0.015**
	EM-11	$(4 \times 4 \times 4)$,$(8 \times 8 \times 8)$	6.8	0.812 ± 0.037
	EM-23	$(2 \times 2 \times 2)$,$(4 \times 4 \times 4)$	9.0	0.817 ± 0.023
	EM-23	$(2 \times 2 \times 2)$,$(8 \times 8 \times 8)$	9.0	**0.818 ± 0.019**
	EM-23	$(4 \times 4 \times 4)$,$(8 \times 8 \times 8)$	9.0	0.811 ± 0.021

Table 4. **ReMIND2Reg Unsupervised Registration Results.** Average Target Registration Error and Jacobian Determinant are obtained from Learn2Reg 2024 Challenge Page. *Param* are listed in Millions of parameters used for training the model.

Methods	**C**	**Param**	**TRE** ↓	$\mathbf{\|J\|} < 0\%$ ↓
Siebert et al. [52]	-	-	3.87 ± 1.05	0.18 ± 0.009
Fourier-Net [28]	-	1.1	4.128 ± 0.890	7.047 ± 1.113
TransMorph-Tiny [9]	6	0.22	3.944 ± 0.693	0.013 ± 0.004
TransMorph [9]	96	46.5	3.916 ± 0.77	0.024 ± 0.007
TransMorph-L [9]	128	108	3.902 ± 0.763	0.018 ± 0.003
EfficientMorph-11	96	1.8	3.734 ± 0.798	0.011 ± 0.002
EfficientMorph-23	96	2.8	**3.599 ± 0.620**	**0.010 ± 0.001**

epochs. Variants employing the Hi-Res tokenization technique with a stride 2 do not perform well for IXI. However, the ablations experiment with fewer embedding dimensions (C=24) improved the performance of 0.7317 to TransMorph's 0.7293 at 100 epochs, achieving similar accuracy as Fourier-Net-s. If trained for a longer period (> 100 epochs), EFFICIENTMORPH may probably be as accurate as TransMorph (maybe even higher), however this is left for future experiments. Accuracy vs epochs curves shown in supplementary Figure 7 indicate that most EFFICIENTMORPH variants outperform TransMorph in initial epochs, but then performance tends to saturate. Qualitative segmentations for the IXI dataset, shown in supplementary Figure 9, show that EFFICIENTMORPH produces results of similar quality to TransMorph. For different substructures, EfficientMorph performs on par with the baseline, as shown in supplementary Figure 8.

4.3. Ablation Studies

Most ablation studies were conducted using the OASIS dataset. Additionally, specific studies, such as those on stride and embedding dimensions, were also carried out on the IXI dataset.

Percentage of Segmentation Data. Segmentation annotations are often unavailable for registration datasets, particularly in the medical field, where obtaining them is both time-consuming and labor-intensive. This challenge arises because multiple radiologists are typically required to mitigate human bias, which significantly increases the effort and time needed to generate accurate annotations. In this ablation study, we trained registration models with varying levels of segmentation data availability using the OASIS dataset. The results, shown in Figure 4, indicate that the performance curve is skewed. A substantial improvement in registration accuracy is observed when the initial 20%-40% of the dataset includes segmentations, but beyond this point, the relative performance improvements diminish with

Table 5. **Stride and Embedding Dimension Ablations.** Mean average dice score and standard deviation are evaluated on 35 segmented anatomies in OASIS. ‘stride’ and ‘C’ are the strides and embedding dimensions.

				w/o Seg Loss		with Seg Loss	
Methods	**stride**	**C**	**Param(M)**	**Dice Score** ↑	$\|\mathbf{J}\| < 0\%$ ↓	**Dice Score** ↑	$\|\mathbf{J}\| < 0\%$ ↓
EfficientMorph-11	4x4x4	96	1.8	0.795 ± 0.071	0.109 ± 0.012	0.841 ± 0.013	0.121 ± 0.017
EfficientMorph-23	4x4x4	96	2.8	0.796 ± 0.035	0.091 ± 0.0006	0.846 ± 0.013	0.125 ± 0.020
EfficientMorph-11	2x2x2	96	1.8	0.803 ± 0.070	0.011 ± 0.002	**0.8623±0.0133**	**0.010 ± 0.001**
EfficientMorph-23	2x2x2	96	2.8	**0.810 ± 0.062**	**0.010 ± 0.001**	**0.870 ± 0.016**	0.017 ± 0.001
EfficientMorph-11	2x2x2	24	1.2	0.796 ± 0.067	0.108 ± 0.008	0.840 ± 0.011	0.125 ± 0.016
EfficientMorph-23	2x2x2	24	2.25	0.799 ± 0.024	0.110 ± 0.014	0.8426 ± 0.013	0.126 ± 0.019
EfficientMorph-11	2x2x2	16	0.5	0.765 ± 0.004	0.164 ± 0.001	0.8311 ± 0.071	0.118 ± 0.011
EfficientMorph-23	2x2x2	16	1.3	0.796 ± 0.003	0.149 ± 0.065	0.8345 ± 0.102	0.130 ± 0.102

further increases in annotated data.

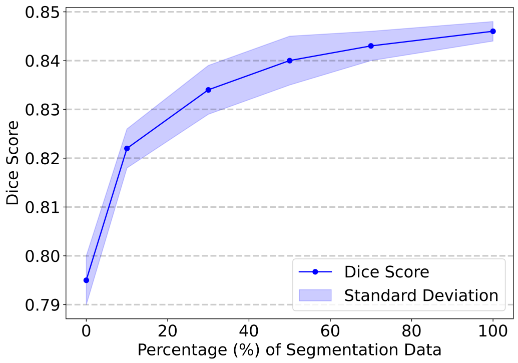

Figure 4. **Impact of Annotated Segmentation Available for Training.** These models were trained for EM-23 variant with stride 4x4x4 and embedding dimension 96.

Stride and Embedding Dimension Ablations. We fully evaluate the impact of different hyper-parameters such as the stride of the voxel used for tokenization and the embedding dimension used in the patch embedding block. Results of these ablation studies are shown in Table 5. From the results, we see that increasing the embedding dimension with the same stride always performs better. Also, models trained with a smaller stride are always performing better, this proves that utilizing high-resolution spatial information for unsupervised registration results in better accuracy.

Plane Order Ablations. We also investigated the effect of varying the plane order (*xy* vs *yx*) in the EM-11 and EM-23 variants. The results of these experiments are shown in supplementary Table 6 and Table 7. The findings suggest that the order of plane attention has minimal impact on performance, as all variants cover all three volume axes, making the plane order unimportant.

Attention Type Ablation. We also explored the impact of various attention optimizations mentioned in related works, including Sparse [12], Linformer [65], Memory Efficient [48], Nystrom [70], and Flash [14]. The results, shown in supplementary Table 8, indicate that for models using a stride of 4, different attention mechanisms have minimal effects on both performance and parameter count. This may be because these methods are optimized for processing billions of tokens, whereas 3D volumes typically involve only a few thousands tokens per sample. When experimenting with a stride of 2, we found that Flash attention reduced the parameter count by approximately 150k while maintaining similar performance to our best-performing EM-23 variant on OASIS dataset.

5. Conclusion and Future Work

We propose EFFICIENTMORPH, a parameter-efficient transformer-based architecture for unsupervised 3D deformable image registration. EFFICIENTMORPH uses a novel plane attention mechanism, which attends to 3D volumetric features by sequentially placing different plane attention blocks xy followed by yz or zx, thus attending to features along all three axes. Additionally, we propose a Hi-Res tokenization strategy to capture higher spatial resolution information while maintaining computational complexity. Evaluations of three datasets demonstrate that EFFICIENTMORPH can achieve state-of-the-art results with a considerably lower parameter count (∼16-27×). EFFICIENTMORPH with higher resolution token consumes larger memory while training, therefore in future work, we plan to explore memory-efficient model architectures using multi-resolution for 3D registration, segmentation, and synthesis applications. Additionally, incorporating frameworks such as Fourier-Net [28] or SegFormer [69] to reduce decoder complexity can further enhance the efficiency and effectiveness of our proposed model.

Acknowledgements

The National Science Foundation supported this work under grant number NSF2217154. The content is solely the authors’ responsibility and does not necessarily represent the official views of the National Science Foundation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Fakhre Alam, Sami Ur Rahman, Sehat Ullah, and Kamal Gulati. Medical image registration in image guided surgery: Issues, challenges and research opportunities. *Biocybernetics and Biomedical Engineering*, 38(1):71–89, 2018. 1

[3] Keyvan Ansarino and Emad Fatemizadeh. Two convolutional neural networks for the rigid and affine registration of two-dimensional ct-mri images of the human brain. In *2022 29th National and 7th International Iranian Conference on Biomedical Engineering (ICBME)*, pages 287–292, 2022. 1

[4] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008. 1

[5] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009. 1

[6] M Adeel Azam, K Bahadar Khan, Muhammad Ahmad, and Manuel Mazzara. Multimodal medical image registration and fusion for quality enhancement. *Computers, Materials & Continua*, 68(1):821–840, 2021. 1

[7] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 1, 2, 5, 6, 7, 13, 16

[8] Brian R Bartoldson, Bhavya Kailkhura, and Davis Blalock. Compute-efficient deep learning: Algorithmic trends and opportunities. *Journal of Machine Learning Research*, 24(122):1–77, 2023. 3

[9] 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, 3, 5, 6, 7, 13, 14, 16

[10] 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. 5, 7

[11] Zeyuan Chen, Yuanjie Zheng, and James C Gee. Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE transactions on medical imaging*, 43(1):15–27, 2023. 5, 7

[12] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3, 8, 14

[13] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 3

[14] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 3, 8, 14

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[17] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8458–8468, 2022. 3

[18] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022. 3

[19] Morteza Ghahremani, Mohammad Khateri, Bailiang Jian, Benedikt Wiestler, Ehsan Adeli, and Christian Wachinger. H-vit: A hierarchical vision transformer for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11513–11523, 2024. 2, 13

[20] Ben Glocker, Aristeidis Sotiras, Nikos Komodakis, and Nikos Paragios. Deformable medical image registration: setting the state of the art with discrete methods. *Annual review of biomedical engineering*, 13(1):219–244, 2011. 1

[21] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31, 2018. 2

[22] Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012. 1

[23] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022. 1, 5

[24] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 2, 3

[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, 13

[26] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[27] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision*, pages 539–555. Springer, 2022. 1

[28] Xi Jia, Joseph Bartlett, Wei Chen, Siyang Song, Tianyang Zhang, Xinxing Cheng, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. Fourier-net: Fast image registration with band-limited deformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1015–1023, 2023. 1, 2, 3, 5, 6, 7, 8, 13, 16

[29] Hao Jin, Yan Luo, Peilong Li, and Jomol Mathew. A review of secure and privacy-preserving medical data sharing. *IEEE access*, 7:61656–61669, 2019. 3

[30] Hans J Johnson and Gary E Christensen. Consistent landmark and intensity-based image registration. *IEEE transactions on medical imaging*, 21(5):450–461, 2002. 3

[31] Parikshit Juvekar, Reuben Dorent, Fryderyk Kögl, Erickson Torio, Colton Barr, Laura Rigolo, Colin Galvin, Nick Jowkar, Anees Kazi, Nazim Haouchine, et al. Remind: The brain resection multimodal imaging database. *Scientific Data*, 11(1):494, 2024. 13

[32] Jean Kaddour, Oscar Key, Piotr Nawrot, Pasquale Minervini, and Matt J Kusner. No train no gain: Revisiting efficient training algorithms for transformer-based language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[33] Tushar Kataria, Saradha Rajamani, Abdul Bari Ayubi, Mary Bronner, Jolanta Jedrzkiewicz, Beatrice S Knudsen, and Shireen Y Elhabian. Automating ground truth annotations for gland segmentation through immunohistochemistry. *Modern Pathology*, 36(12):100331, 2023. 1

[34] Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *International Conference on Algorithmic Learning Theory*, pages 597–619. PMLR, 2023. 2, 3

[35] 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, 5, 6, 16

[36] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009. 1

[37] Guang Li, Huchen Xie, Holly Ning, Deborah Citrin, Jacek Capala, Roberto Maass-Moreno, Peter Guion, Barbara Arora, Norman Coleman, Kevin Camphausen, et al. Accuracy of 3d volumetric image registration based on ct, mr and pet/ct phantom experiments. *Journal of Applied Clinical Medical Physics*, 9(4):17–36, 2008. 1

[38] Zi Li, Lin Tian, Tony CW Mok, Xiaoyu Bai, Puyang Wang, Jia Ge, Jingren Zhou, Le Lu, Xianghua Ye, Ke Yan, et al. Samconvex: Fast discrete optimization for ct registration using self-supervised anatomical embedding and correlation pyramid. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–569. Springer, 2023. 13

[39] Fengze Liu, Ke Yan, Adam P Harrison, Dazhou Guo, Le Lu, Alan L Yuille, Lingyun Huang, Guotong Xie, Jing Xiao, Xianghua Ye, et al. Same: Deformable image registration based on self-supervised anatomical embeddings. 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 87–97. Springer, 2021. 13

[40] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3

[41] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023. 3

[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, 09 2007. 5

[43] Mingyuan Meng, Lei Bi, Dagan Feng, and Jinman Kim. Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–97. Springer, 2022. 2

[44] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*, 2020. 3

[45] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010. 1

[46] Tony CW Mok and Albert CS Chung. 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*, pages 211–221. Springer, 2020. 1, 2

[47] Mircea Mujat, James D Akula, Anne B Fulton, R Daniel Ferguson, and Nicusor Iftimia. Non-rigid registration for high-resolution retinal imaging. *Diagnostics*, 13(13):2285, 2023. 1

[48] Markus N Rabe and Charles Staats. Self-attention does not need o(n2) memory. *arXiv preprint arXiv:2112.05682*, 2021. 3, 8, 14

[49] Bastien Rigaud, Antoine Simon, Joël Castelli, Caroline Lafond, Oscar Acosta, Pascal Haigron, Guillaume Cazoulat, and Renaud de Crevoisier. Deformable image registration for radiation therapy: principle, methods, applications and evaluation. *Acta Oncologica*, 58(9):1225–1237, 2019. 1

[50] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. 3

[51] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999. 3

[52] Hanna Siebert, Lasse Hansen, and Mattias P Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–179. Springer, 2021. 7

[53] Hessam Sokooti, Bob de Vos, Floris Berendsen, Mohsen Ghafoorian, Sahar Yousefi, Boudewijn PF Lelieveldt, Ivana Isgum, and Marius Staring. 3d convolutional neural networks image registration based on efficient supervised learning from artificial deformations. *arXiv preprint arXiv:1908.10235*, 2019. 2

[54] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 232–239. Springer, 2017. 2

[55] Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 3

[56] Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*, 2024. 3

[57] Lin Tian, Hastings Greer, Roland Kwitt, Francois-Xavier Vialard, Raul San Jose Estepar, Sylvain Bouix, Richard Rushmore, and Marc Niethammer. unigradicon: A foundation model for medical image registration. *arXiv preprint arXiv:2403.05780*, 2024. 3, 5

[58] Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, Richard Jarrett Rushmore, Nikolaos Makris, Sylvain Bouix, and Marc Niethammer. Gradicon: Approximate diffeomorphisms via gradient inverse consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18084–18094, 2023. 2, 3, 5

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[60] Max A Viergever, JB Antoine Maintz, Stefan Klein, Keelin Murphy, Marius Staring, and Josien PW Pluim. A survey of medical image registration–under review, 2016. 1

[61] Valery Vishnevskiy, Tobias Gass, Gabor Szekely, Christine Tanner, and Orcun Goksel. Isotropic total variation regularization of displacements in parametric image registration. *IEEE transactions on medical imaging*, 36(2):385–395, 2016. 3

[62] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, et al. Virchow: a million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023. 3

[63] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2, 5

[64] Shaoru Wang, Jin Gao, Zeming Li, Xiaoqin Zhang, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. In *International Conference on Machine Learning*, pages 35624–35641. PMLR, 2023. 3

[65] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 3, 8, 14

[66] Yulin Wang, Yizeng Han, Chaofei Wang, Shiji Song, Qi Tian, and Gao Huang. Computation-efficient deep learning for computer vision: A survey. *Cybernetics and Intelligence*, 2024. 3

[67] William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51, 1996. 1

[68] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 3

[69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2, 8

[70] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148, 2021. 8, 14

[71] Xiao Yang, Roland Kwitt, and Marc Niethammer. Fast predictive image registration. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 48–57. Springer, 2016. 2

[72] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pages 12208–12218. PMLR, 2021. 3

[73] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021. 2, 5

[74] Jun Zhang. Inverse-consistent deep networks for unsupervised deformable image registration. *arXiv preprint arXiv:1809.03443*, 2018. 3

[75] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020. 3