6. Supplementary

6.1. Additional Related Works

The field of medical image registration is undergoing a paradigm shift with the rise of transformer-based methods, which are increasingly surpassing traditional convolutional neural networks (CNNs). While CNN-based approaches like SAME [39] and SAMConvex [38] relied on pre-trained models and carefully designed pipelines to capture global context, transformer architectures such as H-ViT [19] leverage multi-resolution high-level features to represent low-level voxel patches, overcoming the localized feature extraction limitations of CNNs.

Our proposed method advances medical image registration by introducing Hi-Res tokenization for efficient high-resolution feature extraction and a plane-based attention mechanism to balance localized focus with global context. Additionally, the multi-resolution variant captures spatial information early in the network, enabling the model to dynamically learn features across resolution scales and effectively capture intricate structural variations in medical images.

6.2. Additional Dataset and Training Details

6.2.1 Datasets Details

OASIS. Preprocessing involves bias correction, skull stripping, alignment and cropping to dimensions of 160 x 192 x 224. Registration accuracy is reported by performing evaluation of corresponding segmentation masks for 35 anatomical structures. Additionally, FreeSurfer [25] was used for pre-processing the brain MRI images, and it has label maps for 35 anatomical structures. Automated segmentation masks are used for evaluation of registration accuracy.

Remind2Reg. ReMIND2Reg dataset is a pre-processed subset of the ReMIND dataset [31], containing multi-modal pre- and intra-operative data from patients who underwent brain tumor resection at Brigham and Women's Hospital between 2018 and 2024. This dataset is part of the Learn2Reg 2024 challenge, which 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. The images are preprocessed into NIfTI format, cropped to a size of 256 $\times$ 256 $\times$ 256 with 0.5mm isotropic spacing, and co-registered where necessary.

IXI. The IXI dataset was augmented by flipping in random directions while training, as done by baselines. Evaluation was performed on corresponding segmentation masks for 29 anatomical structures with preprocessed size of 160 $\times$ 192 $\times$ 224.

6.2.2 Further Training Pipeline Details of EFFICIENTMORPH: Data Flow

Figure 2A illustrates the proposed end-to-end registration network. The encoder processes two input volumes—a fixed and a moving image—by dividing them into non-overlapping 3D patches with dimensions $2 \times \frac{H}{S} \times \frac{W}{S} \times \frac{D}{S} \times C$, where $S = 2, 4, 8$. As S increases, high-resolution features are progressively lost. For $S = 2$, a Hi-Res Tokenization stage is employed to add depth-wise patch features and use a linear projection layer, enhancing fine-grained features while maintaining complexity similar to $S = 4$ (see Section 3.1). This stage increases the channel dimension by d and incorporates positional encoding to preserve high-resolution feature tracking. For $S = 4$ and $S = 8$, Hi-Res Tokenization is omitted to see if Hi-Res block actually utilizing the features.

Each patch, now termed as tokens, is processed through two efficient transformer blocks separated by a patch merging layer for downsampling. These transformer blocks may feature different plane attention modules (xy or yz or zx) depending on the variant. The bottleneck features from the encoder are then passed through convolutional layers (decoder) to generate a nonlinear 3D deformation field. This field is applied to the moving image using a PyTorch spatial transformer, similar to VoxelMorph [7] and TransMorph [9]. The loss function integrates image similarity (local normalized cross-correlation) and regularization (bending energy) losses, following the approach of TransMorph [9] and Fourier-Net [28]. The architecture utilizes default hyperparameters weighted similarly to TransMorph's [9].

6.2.3 Further Training Pipeline Details of Multi Resolution EFFICIENTMORPH: Data Flow

Figure 2B illustrates the end-to-end data flow for the Multi-Resolution version of EFFICIENTMORPH. The input is processed through two distinct patch embedding layers, each with different strides, S_1' and S_2'. When a stride of 2 is used, the Hi-Resolution tokenization strategy described in Section 3.1 is applied, allowing for efficient handling of high-resolution patches. Otherwise, the input flows through two parallel encoders with similar configurations, each capturing features at different resolutions.

The latent dimensions from these encoders are concatenated, enabling the model to leverage features across varying levels of detail. These bottleneck features are then passed into decoder to produce a nonlinear 3D deformation field, which is applied to the moving image via a PyTorch spatial transformer. The loss functions and hyperparameters

remain same across both versions of the architecture, ensuring smooth integration of multi-resolution features while maintaining performance stability.

6.3. Additional Ablation Experiments And Qualitative Results

6.4. Qualitative Results

Supplementary Figure 5 presents the best, median, and worst cases for both the baseline TransMorph [9] and the proposed model variants. It can be observed that the proposed variants achieve higher Dice scores for most anatomical structures, though performance slightly declines when segmenting smaller anatomies. Supplementary Figure 6 further compares the Dice scores of the proposed models with the baseline across 19 anatomical substructures, where the proposed variants consistently outperform the baseline in the majority of cases.

6.5. Additional Ablation Results

Table 6. **Plane Attention Order Ablation.** Mean average dice score and standard deviation are evaluated on 35 segmented anatomies in OASIS with stride = 4 and C=96 for the EfficientMorph-11 variant.

	w/o Seg Loss	with Seg Loss
Planes	**Dice Score ↑**	**Dice Score ↑**
yz-xy	0.795 ± 0.002	0.843 ± 0.0029
xy-zx	0.795 ± 0.005	0.844 ± 0.0033
yz-zx	0.795 ± 0.003	0.844 ± 0.0032
zx-xy	0.795 ± 0.004	0.844 ± 0.0041
zx-yz	0.795 ± 0.003	0.843 ± 0.0036

Table 7. **Plane Attention - Pattern Ablations - EM-23.** Mean average dice score and standard deviation are evaluated on 35 segmented anatomies in OASIS with stride = 4 and C=96 for EM-23 variant with segmentation loss.

Planes	Dice Score ↑
xy-zx — zx-xy-yz	0.8378 ± 0.0040
yz-xy — zx-xy-yz	0.8436 ± 0.0044
yz-zx — zx-xy-yz	**0.8446 ± 0.0038**
zx-xy — zx-xy-yz	0.846 ± 0.0042
zx-yz — zx-xy-yz	0.843 ± 0.0037
zx-xy — yz-xy-zx	0.844 ± 0.0039
zx-xy — xy-yz-zx	0.844 ± 0.0041

Supplementary Table 7 presents the results of the ablation study on plane order variants for the EfficientMorph-23 model. The table demonstrates that the proposed model achieves comparable accuracy regardless of the plane order.

Supplementary Table 8 highlights the evaluation of various memory-efficient attention mechanisms applied to the proposed plane attention. The results indicate that these approaches yield performance comparable to the proposed method, primarily because the tested mechanisms perform optimally with larger token sizes. Notably, Plane + Flash attention achieved the best performance when trained with a stride of 2, outperforming the stride 4 configuration supporting the claim.

Table 8. **Attention Type Ablation.** Discuss about the different attention types that are added on top of the proposed plane attention, dice scores are evaluated on 35 segmented anatomies in OASIS with stride = 4 and C=96 for the EfficientMorph-23 variant.*Param* as Parameter of model in millions of parameters.

Attention	Param	Dice Score ↑
Plane	2.8	0.8458 ± 0.0137
Plane + Sparse [12]	2.8	0.843 ± 0.0040
Plane + Linformer [65]	4.69	0.848 ± 0.0035
Plane + Memory Efficient [48]	2.82	0.848 ± 0.005
Plane + Nystrom [70]	2.82	0.845 ± 0.0034
Plane + Flash [14]	2.82	0.845 ± 0.0042
Plane + Flash [14] (Stride= 2)	**2.80**	**0.87 ± 0.0042**

6.5.1 IXI dataset Results

Table 9 presents the performance results on the IXI dataset. Notably, ablation experiments reducing the embedding dimensions (C=24) showed an improvement in performance from 0.7317, surpassing TransMorph's 0.7293 at 100 epochs. This also brought the model's accuracy in line with Fourier-Net-s while offering superior inference speed compared to all baselines. While extended training beyond 100 epochs could potentially result in even higher accuracy, this is left for future work.

Supplementary Figure 7 illustrates that EfficientMorph variants generally outperform TransMorph during the initial training stages, though performance plateaus as training progresses. Qualitative results in supplementary Figure 9 indicate that EfficientMorph produces segmentation results comparable to TransMorph, with EfficientMorph performing similarly to the baseline for various substructures, as depicted in supplementary Figure 8.

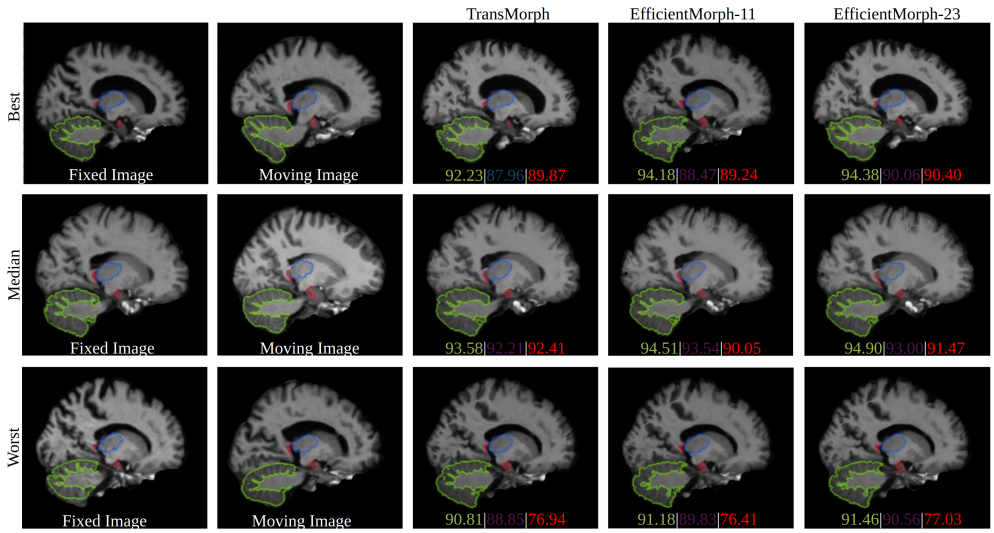

Figure 5. **OASIS qualitative results.** Comparison among the best, median, and worst output of TransMorph with the variants of the proposed method. Here, EfficientMorph-23 and EfficientMorph-11 are the different variants with 2x2x2 stride size and 96 embedded dimension; CGA means variants with cascaded group attention.

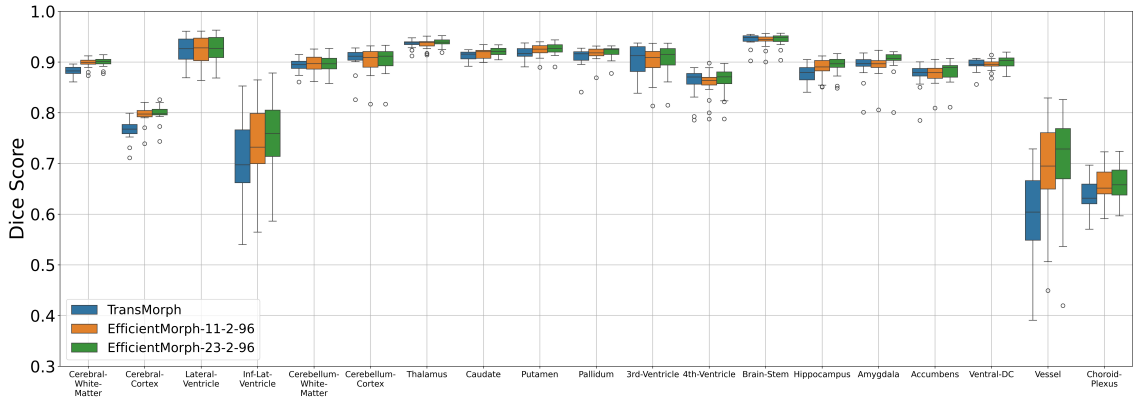

Figure 6. **OASIS boxplot.** Quantitative comparison of the proposed models with TransMorph showing dice scores for 19 anatomical substructures.

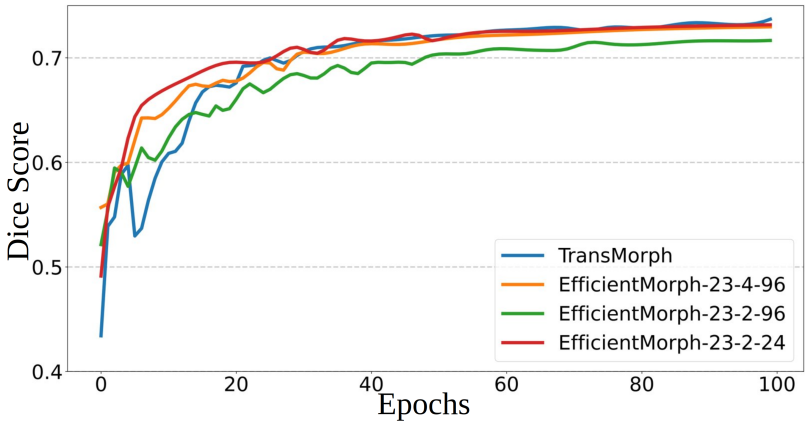

Figure 7. Dice scores as a function of number of epochs(IXI).

Table 9. **IXI Results.** Mean average dice score and standard deviation are evaluated on 29 segmented anatomies in IXI. * indicates the performance numbers taken from TransMorph and Fourier-Net; for all others, we ran these baselines on our system for fair comparison. 'stride' and 'C' are the strides and channel layer for initial embedding layer. 'Multi-Add' is the number of Multiply add operations needed for a forward pass.

Methods	stride	C	Epochs	Param(M)	Dice Score ↑ Val	Test
SyN*	-	-	-	-	-	0.645±0.152
NiftiReg*	-	-	-	-	-	0.645±0.167
voxelMorph-1* [7]	-	-	-	0.3	-	0.548±0.317
cycleMorph* [35]	-	-	-	-	-	0.528±0.321
Fourier-Net-s [28]	-	-	200	1.05	0.729±0.024	0.730±0.025
Fourier-Net-s [28]	-	-	1000	1.05	0.735±0.026	0.736±0.027
Fourier-Net [28]*	-	-	1000	4.19		**0.760±0.132**
TransMorph-Tiny* [9]	4x4x4	6	500	0.24	0.545±0.180	0.543±0.180
TransMorph [9]	4x4x4	96	100	46.7	0.7293±0.029	0.7324±0.0314
TransMorph [9]	4x4x4	96	500	46.7	0.7405±0.0283	0.7408±0.0299
TransMorph-L [9]*	4x4x4	128	500	108.34	0.753 ±0.130	**0.754±0.128**
EfficientMorph-11	4x4x4	96	100	2.01	0.7233±0.0305	0.7224±0.0324
EfficientMorph-23	4x4x4	96	100	3.04	0.7291±0.0303	0.7298±0.0322
EfficientMorph-11	2x2x2	96	100	1.7	0.6739±0.0322	0.6749±0.0323
EfficientMorph-23	2x2x2	96	100	2.8	0.7159±0.0307	0.7174±0.0330
EfficientMorph-23	2x2x2	24	100	3.0	**0.7312±0.0298**	**0.7317±0.0320**

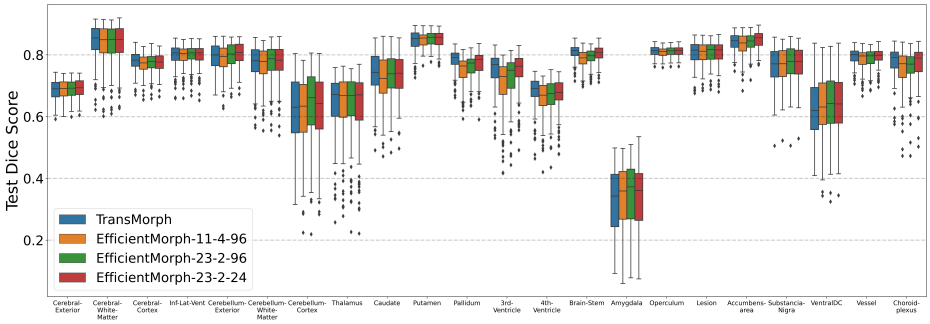

Figure 8. **IXI boxplot.** Quantitative comparison of the proposed models with TransMorph showing dice scores for 22 anatomical substructures.

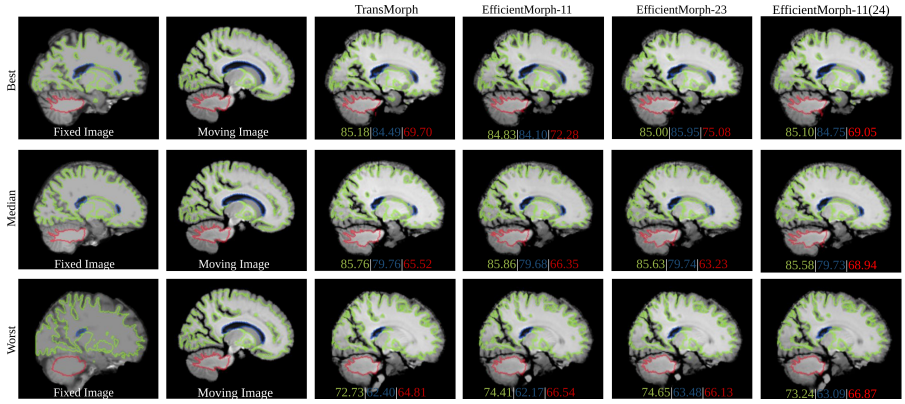

Figure 9. **IXI qualitative results.** Comparison among the best, median, and worst output of TransMorph with the variants of the proposed method. EfficientMorph-23 and EfficientMorph-11 are the different variants with 4x4x4 stride size and 96 embedded dimensions; EfficientMorph-11(24) has 24 embedding dimensions.