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

Supplementary Material

Contents

A. Datasets and data preparation

OASIS - Open Access Series of Imaging Studies: The OASIS[7] dataset [25, 41] is a publicly accessible database of T1-weighted MRI data. In this study, this dataset was employed for inter-patient registration in alignment with the 2021 Learn2Reg challenge [25], where 394, 19, and 38 MRI scans were allocated for training, validation, and inference, respectively. FreeSurfer was utilized to pre-process the samples, providing label maps for 35 anatomical structures with the voxel size $160 \times 192 \times 224$ for subsequent evaluation.

IXI - Information eXtraction from Images: Publicly available IXI[8] contains 576 T1-weighted brain MRI samples with 30 anatomical structures that were split into 403, 58, and 115 for training, validation, and test sets, respectively. The MRI volumes were cropped into the voxel size of $160 \times 192 \times 224$. The models were trained and validated on 806 and 116 unique arbitrary pairs of MRI samples, respectively. In patient-to-atlas registration inference, three arbitrary pairs of MRI samples were selected as the atlases and 50 arbitrary samples were registered to these atlases. In total, 150 patient-to-atlas registrations were performed for assessment. For the inter-patient registration inference, 115 pairs were randomly chosen for the main evaluation of the methods.

ADNI - The Alzheimer's Disease Neuroimaging Initiative: ADNI[9] [27] is a large-scale study aimed at developing methods for the early detection and tracking of Alzheimer's disease. It encompasses longitudinal data, including MRI and PET scans from subjects with normal cognition, mild cognitive impairment, and early Alzheimer's disease. We used the T1-weighted MRI data for evaluation with 45 labels. The MRI scans are first registered to MNI152 space and then preprocessed by FreeSurfer, including skull stripping, normalization, subcortical structures segmentation, and cortical surface extraction. We utilized both the dice score of the subcortical segmentation label and the surface distance of cortical surfaces to compare the performance of the methods. The inter-patient registration inference contains 150 pairs that were selected randomly. Similar to the patient-to-atlas registration in IXI, three arbitrary pairs of MRI samples were selected as the atlases and the other 50 arbitrary samples as moving, forming a total of 150 patient-to-atlas pairs for the patient-to-atlas experiment.

LPBA - LONI Probabilistic Brain Atlas: LPBA[10] [53] comprises 40 T1-weighted 3D brain MR, each of which comes with segmentation ground truth of 56 anatomical structures. The LPBA dataset is used for inference only, so we just registered the samples to MNI152 through an affine transformation. In the inter-patient registration, a total of 120 pairs were randomly selected, and these pairs remained consistent across all employed methods. Likewise, we iterated the patient-to-atlas registration for three randomly chosen samples, each serving as an atlas in turn, and the remaining 39 MRI volumes were registered to the designated atlas. Hence, there are 117 pairs for the patient-to-atlas experiment.

Mindboggle-101: The Mindboggle dataset[11] [34] comprises 41 anatomically labeled brain surfaces derived from 101 healthy individuals. The dataset is organized into four subsets, namely HLN (containing 12 scans), MMRR (containing 23 scans), NKI (containing 42 scans), and OASIS (containing 20 scans). Since the OASIS subset was utilized in the OASIS experiment, it has been excluded from the inference in the Mindboggle experiment. The MRI volumes are registered into the MNI152 space through an affine transformation, with a resolution of $1 \times 1 \times 1\, m^3$ and a voxel grid of $160 \times 192 \times 224$ voxels. Within each subset, an arbitrary sample was selected as the atlas, and the re-

[7] https://github.com/adalca/medical-datasets/blob/master/neurite-oasis.md

[8] https://brain-development.org/ixi-dataset

[9] https://adni.loni.usc.edu/

[10] https://www.loni.usc.edu/research/atlas_downloads

[11] http://mindboggle.info

maining samples were subsequently registered to this atlas. This procedure was iterated two additional times, resulting in a total of 33, 66, and 123 to-be-registered pairs for the HLN, MMRR, and NKI subsets, respectively, in the context of patient-to-atlas registration. In the inter-patient registration, 15, 33, and 66 pairs were randomly selected from the HLN, MMRR, and NKI subsets.

B. Assessment metrics

Dice score: The Dice score measures the accuracy between the warped segmentation map of moved image A and its corresponding reference B, defined as

$$\text{Dice} = 2\frac{|A \cap B|}{|A| + |B|} \tag{9}$$

Jacobian determinant: The absolute value of Jacobian determinants provides information about the local changes in the deformation field, indicating expansion or contraction near the specified voxel location. A Jacobian determinant with a non-positive value signifies a locally non-invertible transformation. Similarly, SDlogJ calculates the standard deviation of the logarithm of the Jacobian determinant associated with the deformation field.

HD95: This metric computes 95% percentile of Hausdorff distance of segmentation results.

Surface Distance: In Sec. D.2 in Supplementary, we also reported the surface results of various registration methods for the ADNI scans. The deformed surfaces were assessed by 90-percentile Hausdorff distance (HD) and the average symmetric surface distance (ASSD), which quantifies the average boundary distance between surfaces.

C. Experimental setup

All registration models were trained for 500 epochs on an NVIDIA A100 GPU equipped with 80GB VRAM, employing the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 1. The regularization parameter $\lambda_{reg.}$ was set to 1. For the OASIS dataset, we trained H-ViT on 394 cross-sectional MRIs, and the deformation fields of the test set were assessed using the MICCAI Learn2Reg platform, accessible at https://learn2reg.grand-challenge.org/Learn2Reg2021. In the remaining datasets, we assessed the methods networks trained on IXI. All the techniques, including ours, underwent unsupervised training using normalized cross-correlation (NCC) and a 3D spatial gradient. λ_{smooth} was set to 1 during training. We used default network parameter settings recommended by authors for the competing methods, and the implementations are available at https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration.

VoxelMorph [6] adopts a CNN-based UNet, comprising encoding layers with a feature size of [16, 32, 32, 32] and corresponding decoder with [32, 32, 32, 32, 32, 16, 16] features.

MIDIR [48]: MIDIR is a diffeomorphic model that employs a B-spline FFD parameterization of the Stationary Velocity Field (SVF) for registration. This approach aims to achieve smooth diffeomorphic deformation during the registration process. The UNet network in MIDIR has encoder layers configured as [16, 32, 32, 32, 32], and the decoder layers are set to [32, 32, 32, 32].

CoTr [68]: CoTr incorporates CNN-encoder, decoder, and FFN architectures. To maintain consistency, we utilized the code available at https://github.com/YtongXie/CoTr and followed the recommended settings provided by the authors.

CycleMorph [32] employs registration UNet blocks with an encoder specified as [16, 32, 32, 32, 32] and a decoder featuring maps of [32, 32, 32, 8, 8, 3]. The parameters α, β, and λ_{regis} were configured to values of 0.1, 0.5, and 0.02, respectively, in adherence to the recommendations set forth by the authors.

PVT [63]: In the context of registration using the PVT model, we followed the recommended configuration outlined by Chen *et al.* [10]. The embedding dimension for layers is specified as [20, 40, 200, 320] with a voxel patch size of 4. The number of heads for the layers is set as [2, 4, 8, 16], and the MLP ratio follows [8, 8, 4, 4]. The depth of layers is configured as [3, 10, 60, 3]. Additionally, the spatial reduction rate for each transformer encode layer is determined as [8, 4, 2, 1].

ViT-V-Net [8] has a hybrid CNN-Transformer architecture. The CNN-based UNet component features an encoder configuration of [16, 32, 32] and a decoder of [96, 48, 32, 32, 16]. For the Transformer segment, the voxel patch size is set to 8, the embedding size is 252, and the MLP dimension is 3072, with 12 heads and 12 layers. The dropout for attention is 0, while the dropout for the Transformer is set to 0.1.

TransMorph [10] features a Swin Transformer with specific configurations, including an embedding dimension of 96, a voxel size of 4, and a window size of [5, 6, 7, 7]. The transformer has depths specified as [2, 2, 4, 2] and heads as [4, 4, 8, 8]. The MLP ratio is set to 4. In the case of TransMorph-Bspl, which involves B-spline transformation, the control point spacing is set at 2, aligning with the value utilized in MIDIR. For the TransMorph-Bayes model, the Monte-Carlo dropout probability is configured at 0.15.

nnFormer [74]: To ensure a fair comparison, we applied identical Transformer parameter values from TransMorph to nnFormer, given that nnFormer was also developed based on the Swin Transformer architecture.

	Left WM Surface		Right WM Surface		Left Pial Surface		Right Pial Surface	
	ASSD ↓	HD ↓	ASSD ↓	HD ↓	ASSD ↓	HD ↓	ASSD ↓	HD ↓
Affine	1.747± 0.114	3.852± 0.353	1.761± 0.103	3.960± 0.437	1.859± 0.104	4.047± 0.307	1.885± 0.115	4.124± 0.403
VoxelMorph [6]	1.103± 0.102	2.619± 0.241	1.118± 0.096	2.750± 0.391	1.112± 0.107	2.741± 0.273	1.146± 0.109	2.865± 0.312
MIDIR [48]	1.271± 0.075	2.921± 0.229	1.264± 0.076	2.900± 0.247	1.318± 0.089	3.019± 0.264	1.330± 0.090	3.026± 0.230
CycleMorph [32]	1.075± 0.090	2.640± 0.269	1.093± 0.109	2.716± 0.371	1.084± 0.094	2.700± 0.268	1.124± 0.112	2.812± 0.322
CoTr [68]	1.169± 0.089	2.712± 0.187	1.178± 0.093	2.762± 0.270	1.234± 0.118	2.929± 0.279	1.250± 0.113	2.977± 0.238
nnFormer [74]	1.272± 0.107	2.840± 0.231	1.246± 0.075	2.828± 0.253	1.278± 0.105	2.943± 0.262	1.272± 0.082	2.929± 0.214
PVT [63]	1.212± 0.097	2.819± 0.264	1.205± 0.104	2.848± 0.345	1.239± 0.111	2.964± 0.275	1.234± 0.115	2.976± 0.295
ViT-V-Net [8]	1.054± 0.094	2.580± 0.240	1.064± 0.096	2.666± 0.354	1.060± 0.093	2.718± 0.269	1.089± 0.106	2.820± 0.301
TransMorph-Bayes [10]	*0.935± 0.082*	*2.261± 0.192*	*0.937± 0.094*	*2.263± 0.289*	*0.908± 0.086*	*2.230± 0.216*	*0.935± 0.108*	*2.320± 0.285*
TransMorph-bspl [10]	0.975± 0.065	2.464± 0.341	0.953± 0.078	2.354± 0.276	1.038± 0.077	2.485± 0.260	1.027± 0.093	2.471± 0.255
Proposed H-ViT	**0.877± 0.075**	**2.224± 0.223**	**0.878± 0.093**	**2.233± 0.283**	**0.877± 0.069**	**2.213± 0.224**	**0.904± 0.094**	**2.311± 0.316**

Table 8. **Surface distance** results of different registration techniques for 20 arbitrary cross-sectional pairs from the ADNI dataset

C.1. H-ViT implementation details

H-ViT employs an encoder with five layers [32, 64, 128, 256, 512] and a decoder with [192, 192, 192, 192]. The H-ViT Transformer parameters were configured as follows: the embedding dimension f_e was set to 192, the number of feature maps S_h to 4, the voxel patch size to $2 \times 2 \times 2$, the model depth to 1, MLP ratio in FFN to 2, drop rate to 0, and the number of heads to 32. Due to GPU memory limitations, we utilized half resolution for the deformation field, subsequently upsampled by a factor of 2. The parameters for the small H-ViT, as used in Tab. 7 of the ablation study, were configured as follows: the number of feature maps S_h was set to 3, the embedding dimension f_e to 128, the batch size to 4, and the number of training epochs to 200.

D. Experimental results

D.1. Surface registration results on ADNI

The results of surface distance measurements, assessed by ASSD and HD metrics, for 20 surface pairs from the ADNI dataset are presented in Tab. 8.

D.2. Experiments with ADNI

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		0.531±0.082	–	0.477±0.052	–
VoxelMorph [6]	–	0.692±0.214	0.986±0.350	0.646±0.226	1.278±0.368
MIDIR [48]	–	0.666±0.220	*0.174±0.158*	0.635±0.227	*0.274±0.181*
CycleMorph [32]	–	0.687±0.217	1.062±0.407	0.655±0.225	1.414±0.409
CoTr [68]	Self-Att.	0.654±0.228	0.663±0.259	0.623±0.231	0.926±0.341
nnFormer [74]	Self-Att.+Global-Att.	0.633±0.224	1.210±0.377	0.604±0.229	1.617±0.417
PVT [63]	Self-Att.	0.610±0.230	1.938±0.421	0.579±0.234	2.297±0.420
ViT-V-Net [8]	Self-Att.	0.727±0.210	0.926±0.328	0.686±0.219	1.173±0.312
TransMorph-Bayes [10]	Self-Att.	0.726±0.209	1.207±0.350	0.696±0.215	1.562±0.429
TransMorph-bspl [10]	Self-Att.	*0.730±0.208*	<**0.001**	*0.702±0.213*	<**0.001**
Proposed H-ViT	Self-Att. + Cross-Att.	**0.760±0.203**	0.445±0.207	**0.730±0.210**	0.648±0.261

Table 9. Quantitative evaluation results for the registration methods on the *ADNI* dataset for 45 anatomical structures over 150 random pairs for inter-patient and 150 pairs for patient-to-atlas registrations.

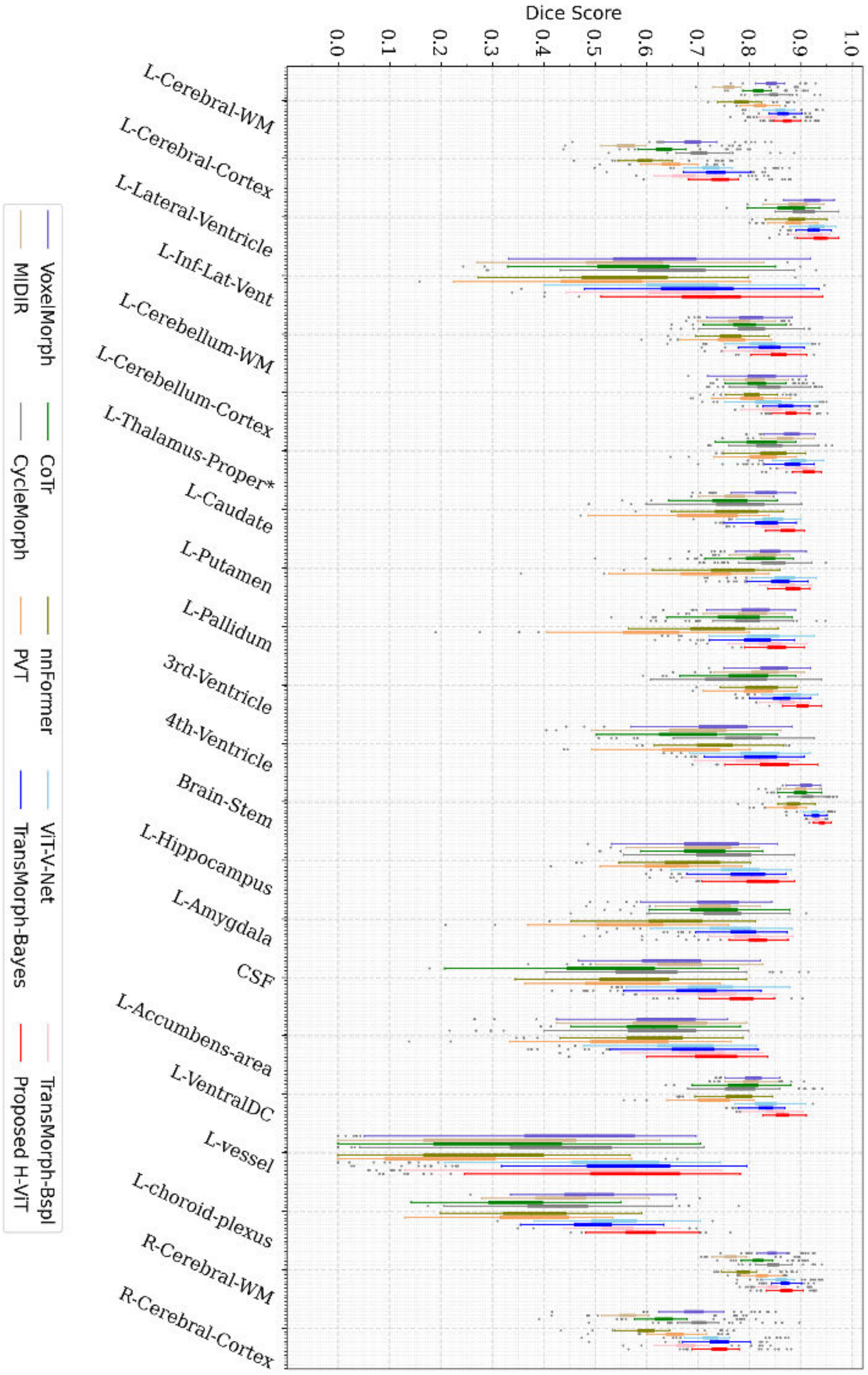


Figure 6. Dice score results for the *inter-patient* registration of various methods on the *ADNI* dataset per anatomical structure (continued)

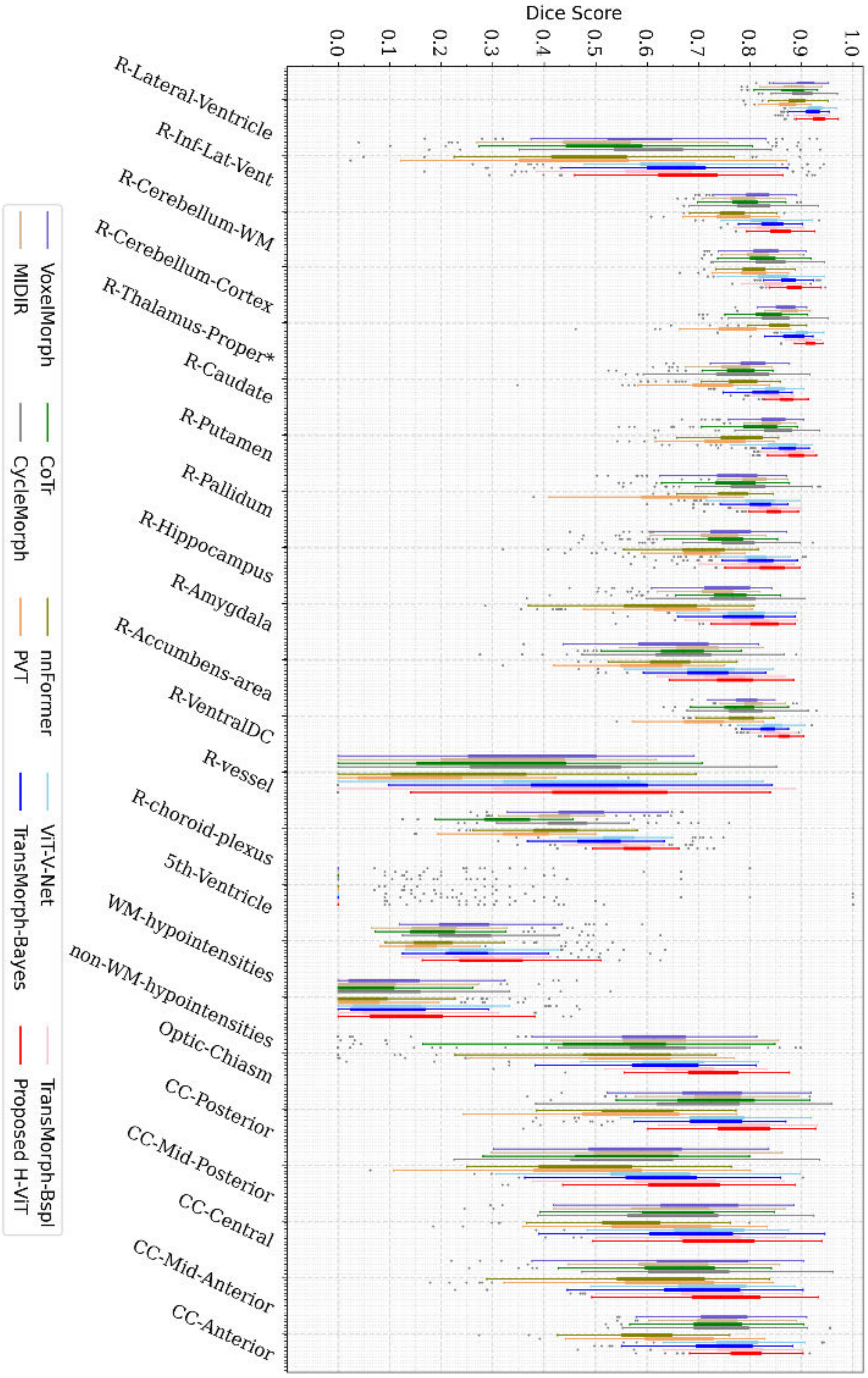


Figure 7. Dice score results for the *inter-patient* registration of various methods on the *ADNI* dataset per anatomical structure

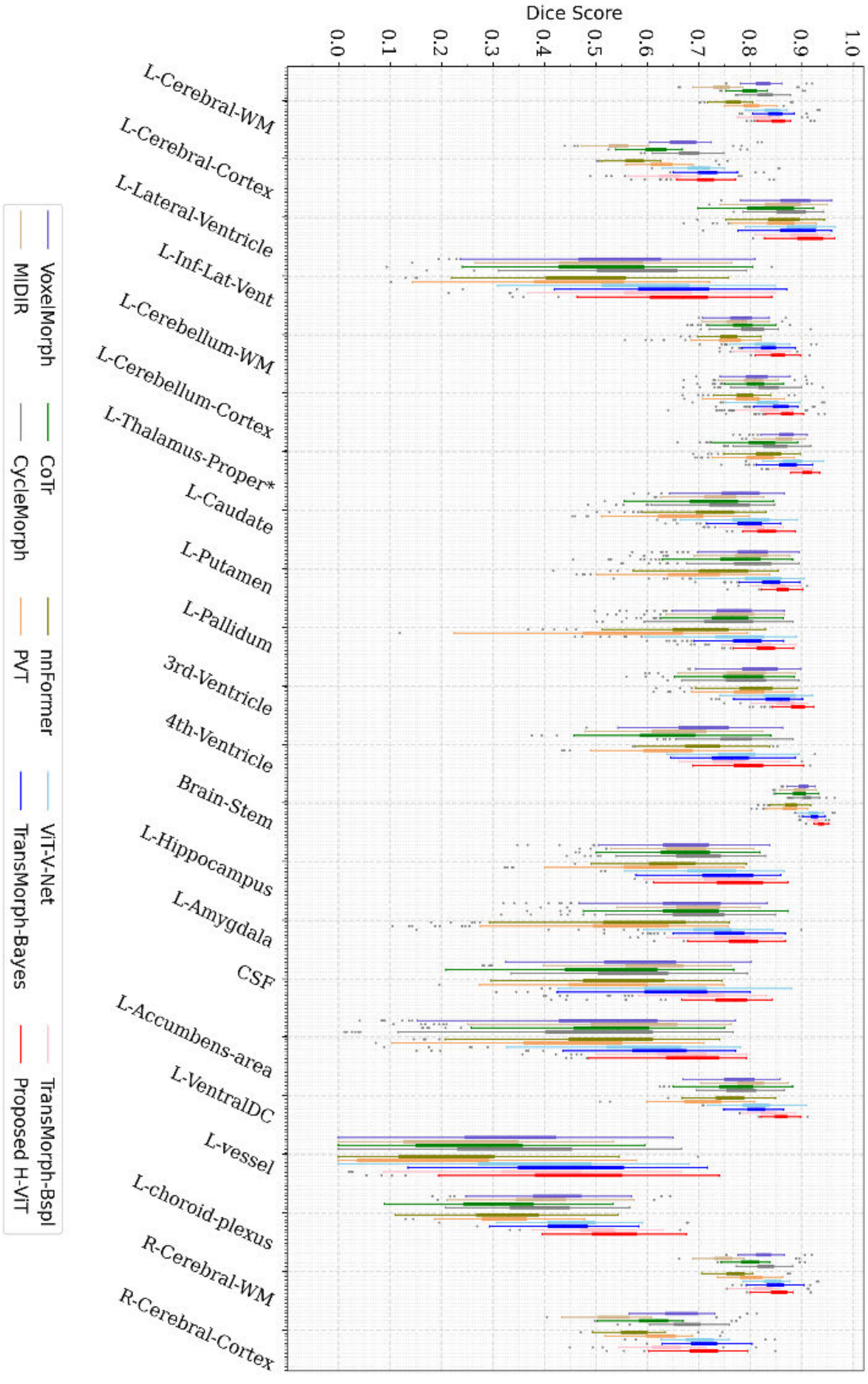


Figure 8. Dice score results for the *patient-to-atlas* registration of various methods on the *ADNI* dataset per anatomical structure (continued)

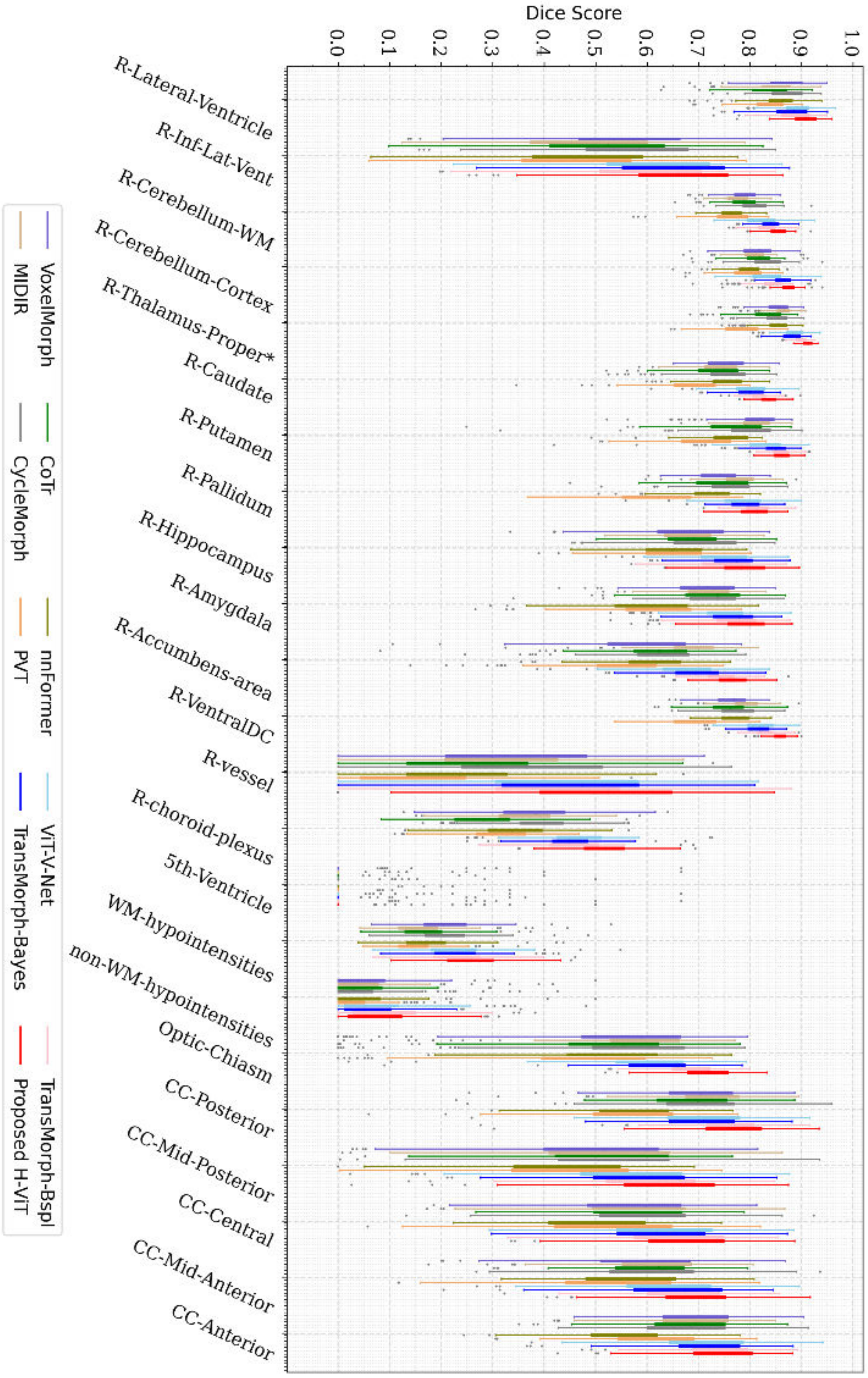


Figure 9. Dice score results for the *patient-to-atlas* registration of various methods on the *ADNI* dataset per anatomical structure

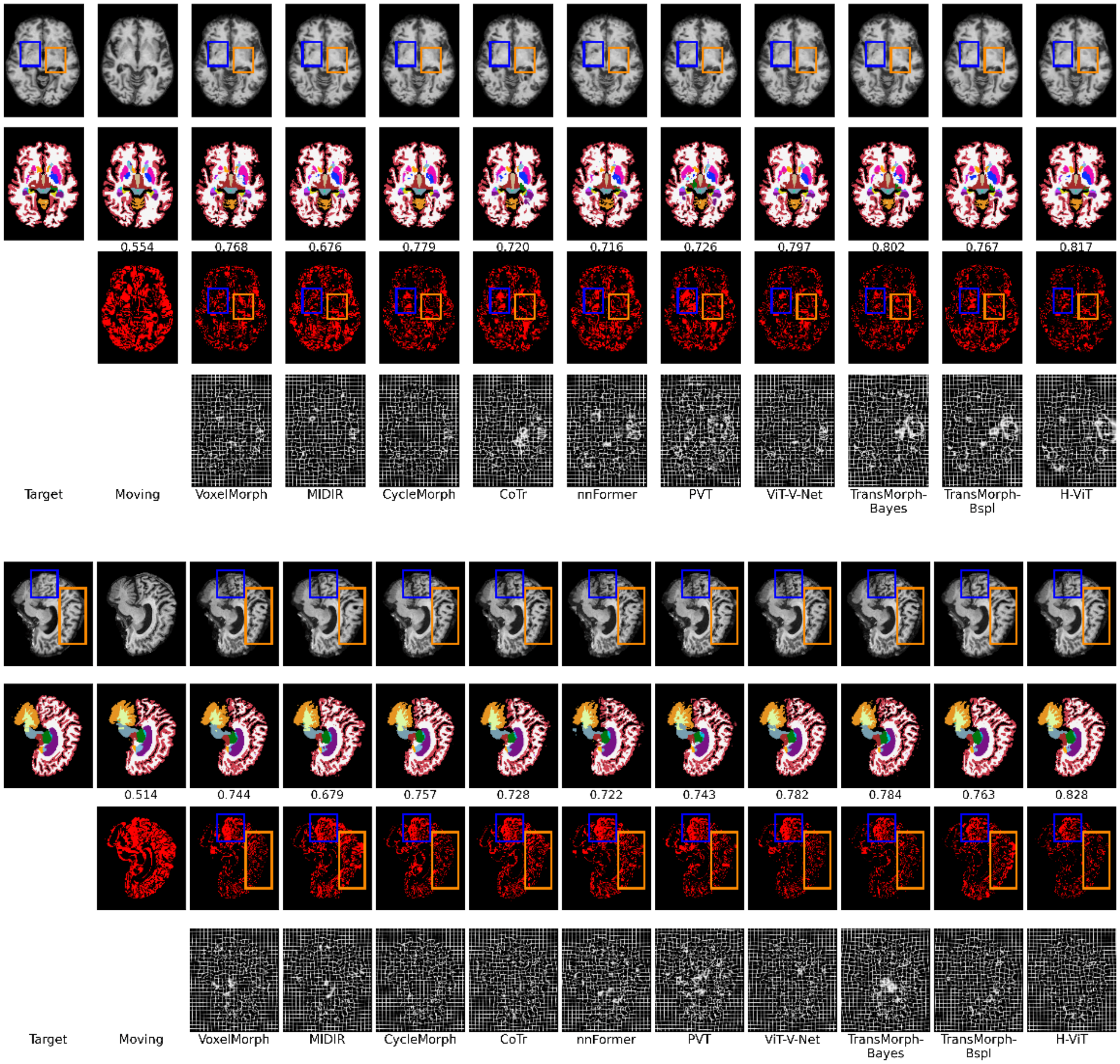


Figure 10. More examples of axial and sagittal slices from the ADNI dataset and outcomes (from top to bottom: MRI, segmentation, difference in segmentation between ground truth and segmented results, and the deformed grid) of different registration methods, with corresponding Dice scores below the segmentation 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).

D.3. Experiments with IXI

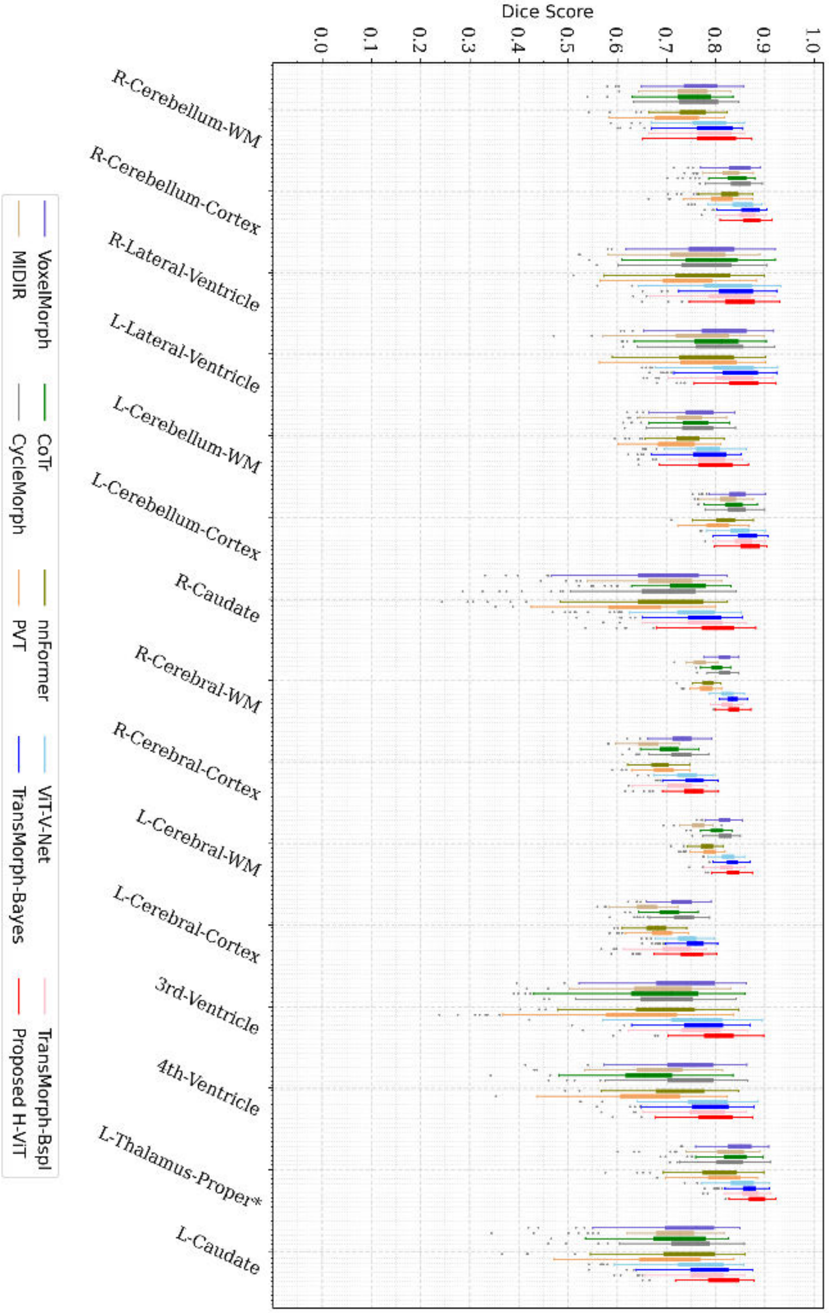


Figure 11. Dice score results for the *inter-patient* registration of various methods on the *IXI* dataset per anatomical structure (continued)

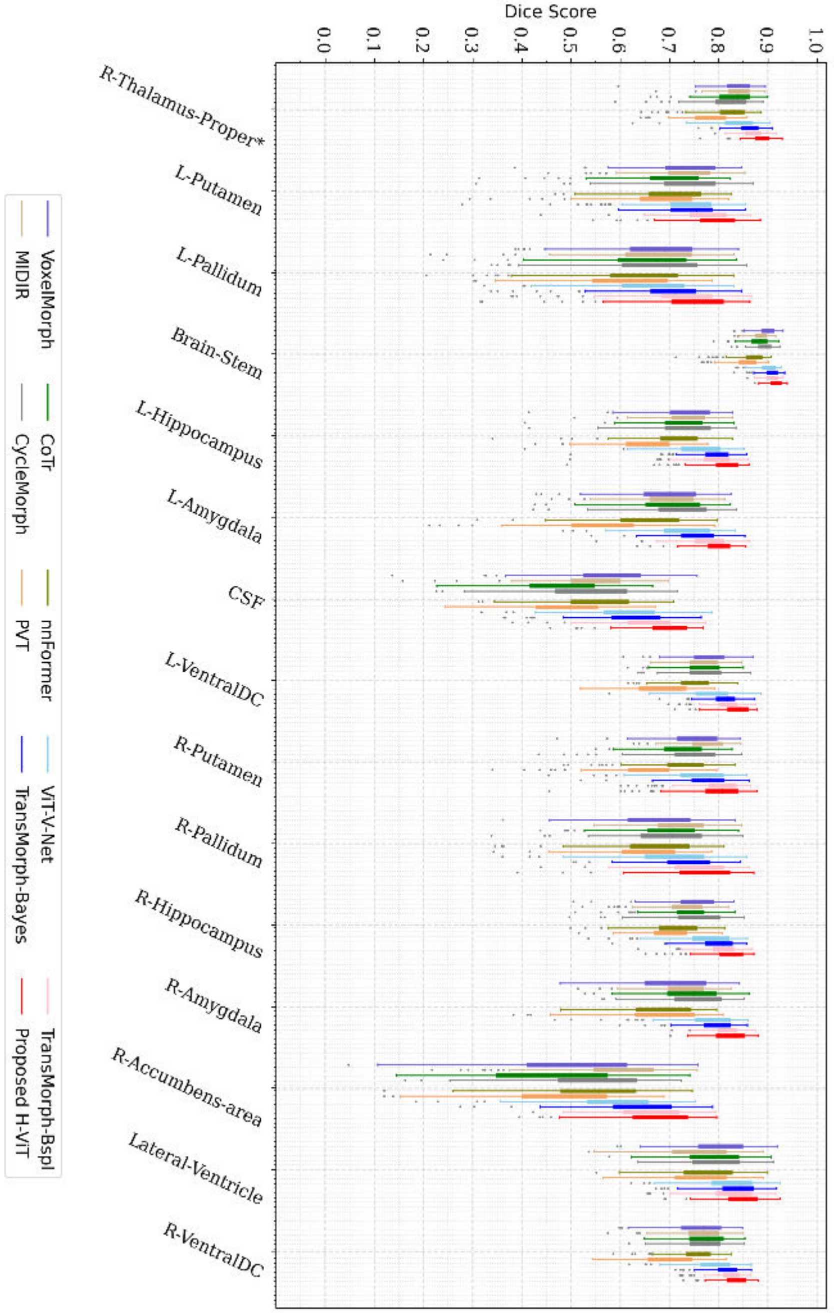


Figure 12. Dice score results for the *inter-patient* registration of various methods on the *IXI* dataset per anatomical structure

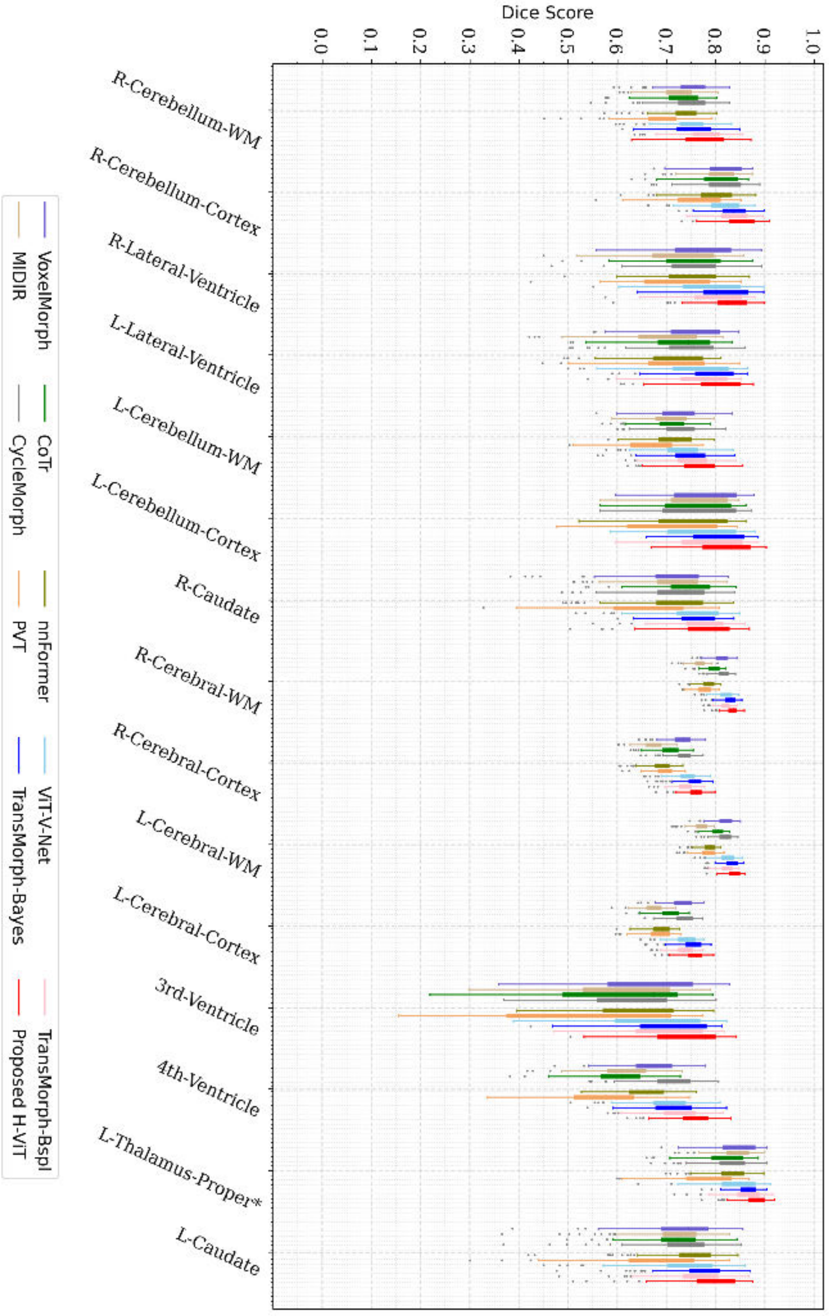


Figure 13. Dice score results for the *patient-to-atlas* registration of various methods on the *IXI* dataset per anatomical structure (continued)

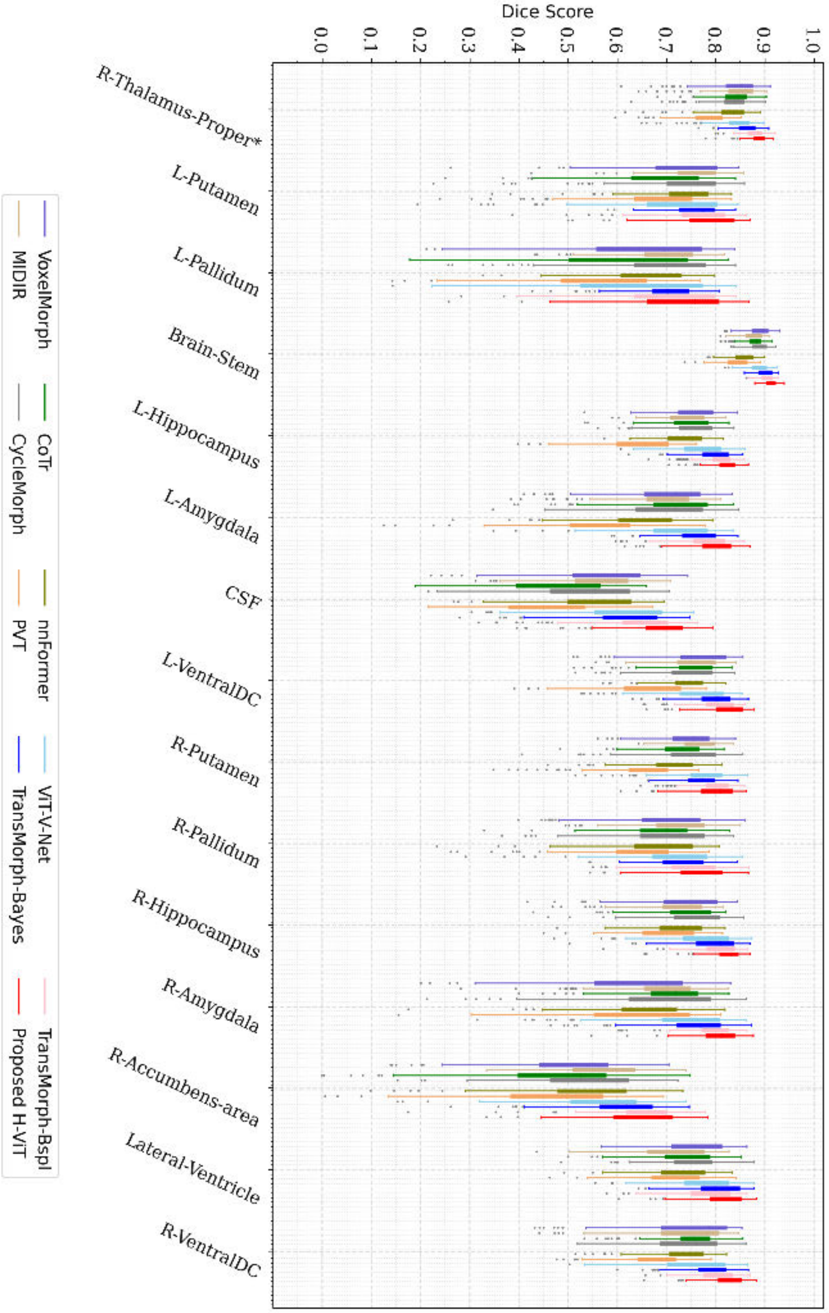


Figure 14. Dice score results for the *patient-to-atlas* registration of various methods on the *IXI* dataset per anatomical structure

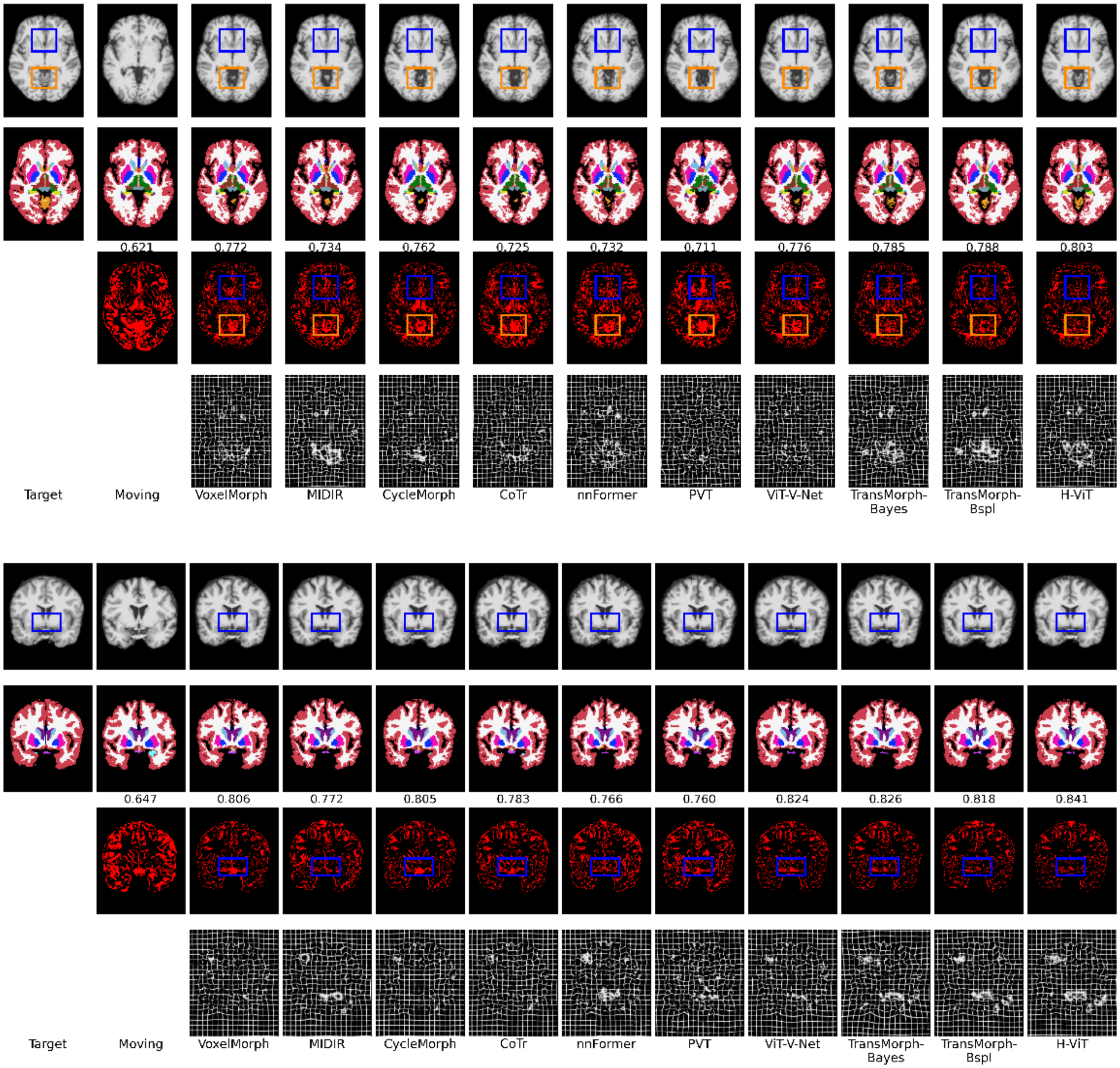


Figure 15. Examples of axial and coronal slices from the IXI dataset and outcomes (from top to bottom: MRI, segmentation, difference in segmentation between ground truth and segmented results, and the deformed grid) of different registration methods, with corresponding Dice scores below the segmentation 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).

D.4. Experiments with LPBA

We apologize for a minor error in Tab. 4 in the main draft. There is a correction regarding the total number of MRI pairs for patient-to-atlas registration, which was mistakenly stated as 108. The accurate number is 117. This will be rectified in the upcoming rebuttal.

Method	Attention	Inter-Patient Registration		Patient-to-Atlas Registration	
	Mechanism	Dice ↑	$\lvert J_\Phi \rvert \leq 0$ (%) ↓	Dice ↑	$\lvert J_\Phi \rvert \leq 0$ (%) ↓
Affine		0.561±0.018	–	0.543±0.017	–
MIDIR [48]	–	0.624±0.017	0.017±0.002	0.629±0.017	0.016±0.002
CycleMorph [32]	–	0.654±0.017	0.008±0.002	0.645±0.016	0.007±0.002
nnFormer [74]	Self-Att.+Global-Att.	0.626±0.018	0.008±0.001	0.631±0.016	0.008±0.001
PVT [63]	Self-Att.	0.637±0.016	0.013±0.001	0.642±0.016	0.013±0.001
ViT-V-Net [8]	Self-Att.	0.658±0.017	0.007±0.002	0.650±0.017	0.006±0.000
TransMorph-Bayes [10]	Self-Att.	0.658±0.017	0.005±0.000	0.655±0.015	0.005±0.000
TransMorph-bspl [10]	Self-Att.	*0.670±0.018*	<**0.001**	*0.666±0.016*	<**0.001**
Proposed H-ViT	Self-Att. + Cross-Att.	**0.704±0.016**	*0.002±<0.001*	**0.694±0.015**	*0.002±<0.001*

Table 10. Quantitative evaluation results for the registration methods on the *LPBA* dataset for 56 anatomical structures over 120 random pairs for inter-patient and 117 pairs for patient-to-atlas registrations.

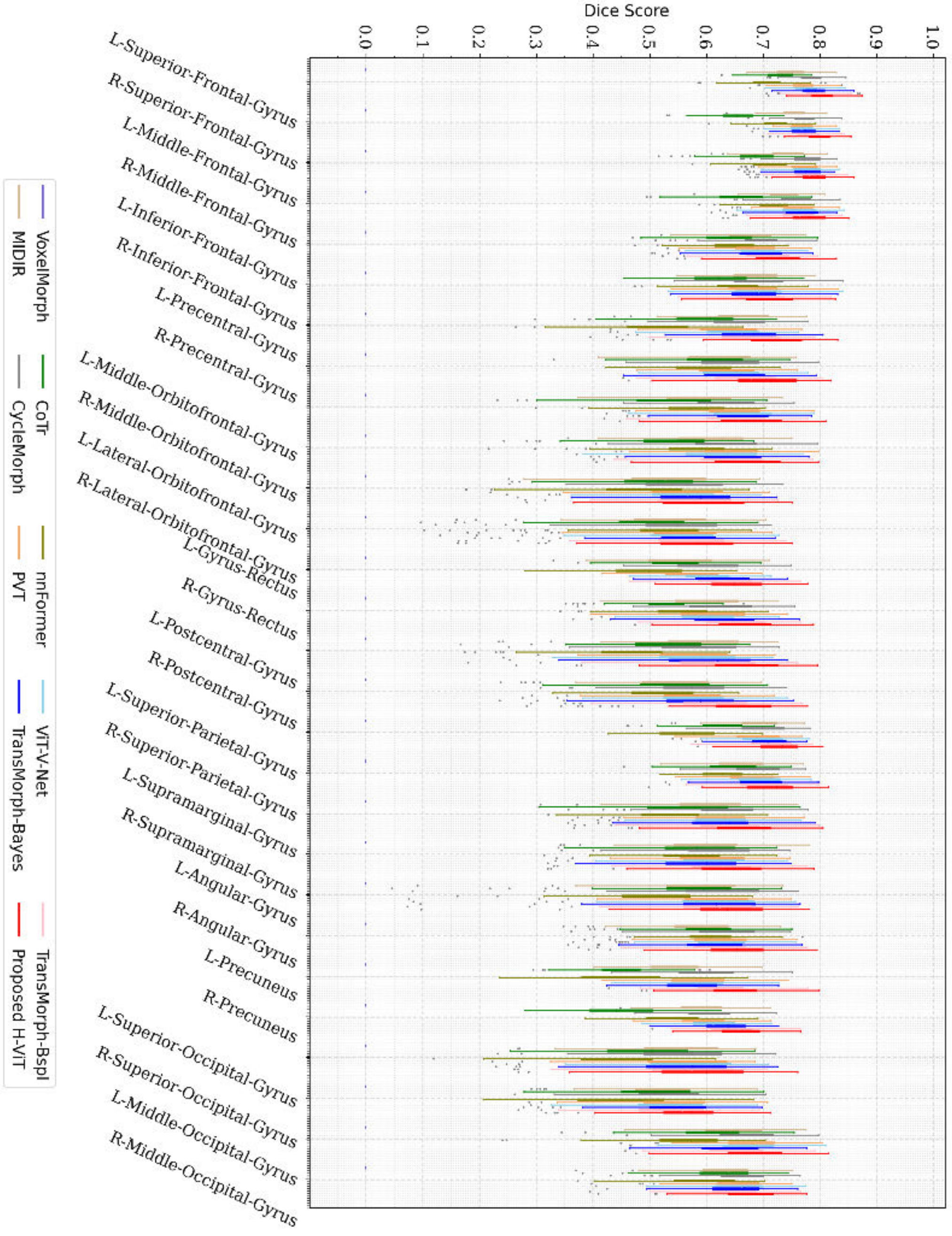


Figure 16. Dice score results for the *inter-patient* registration of various methods on the *LPBA* dataset per anatomical structure (continued)

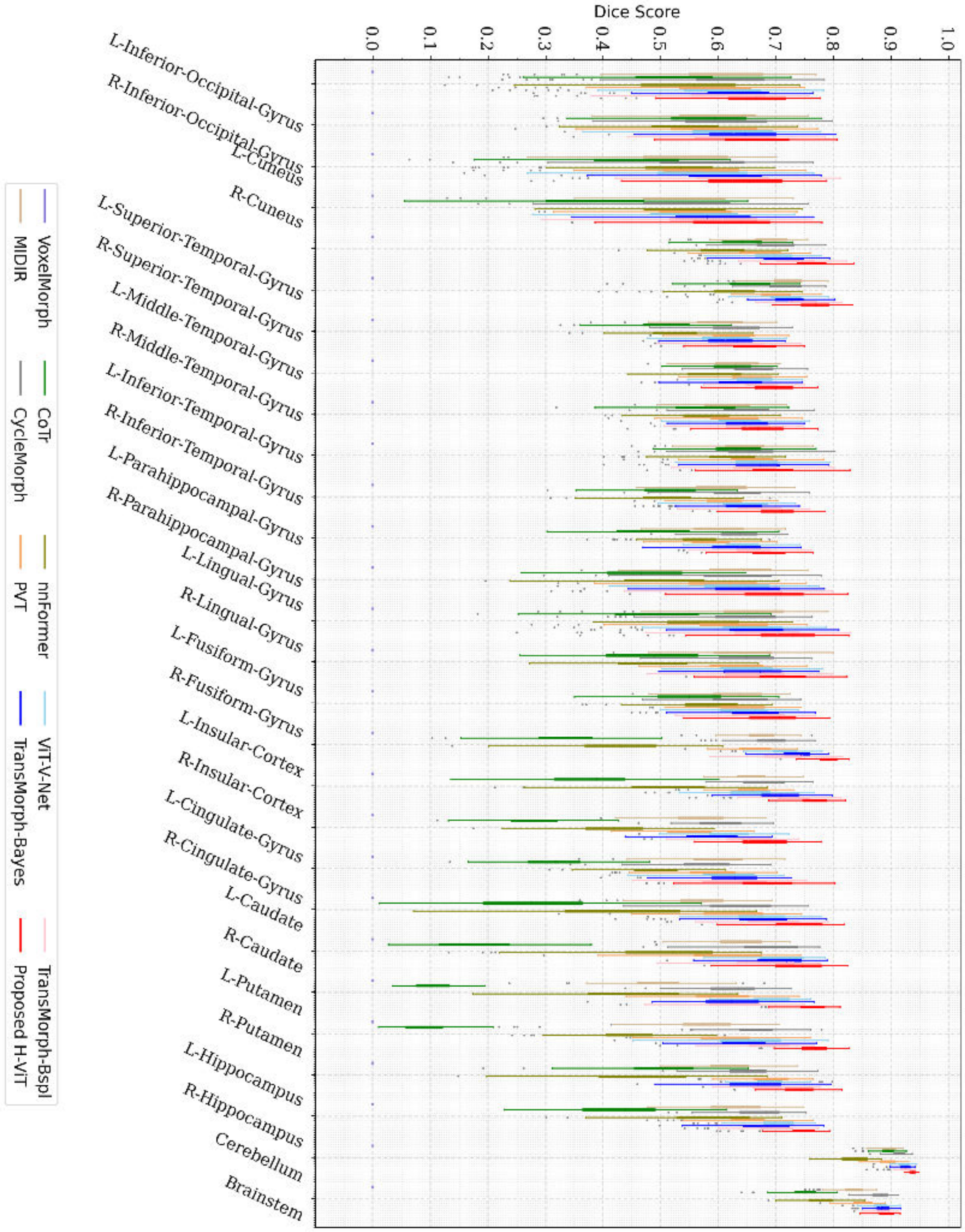


Figure 17. Dice score results for the *inter-patient* registration of various methods on the *LPBA* dataset per anatomical structure

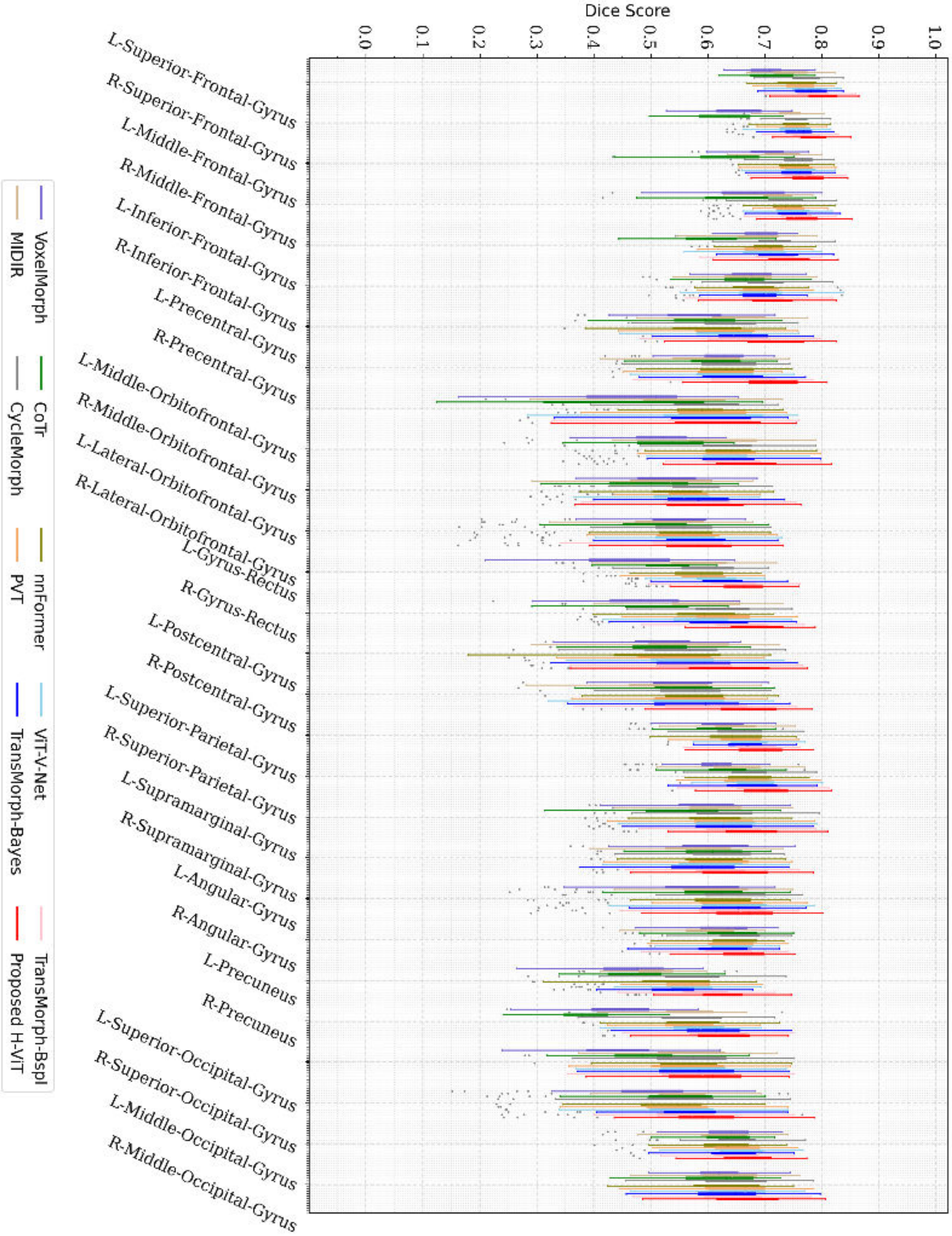


Figure 18. Dice score results for the *patient-to-atlas* registration of various methods on the *LPBA* dataset per anatomical structure (continued)

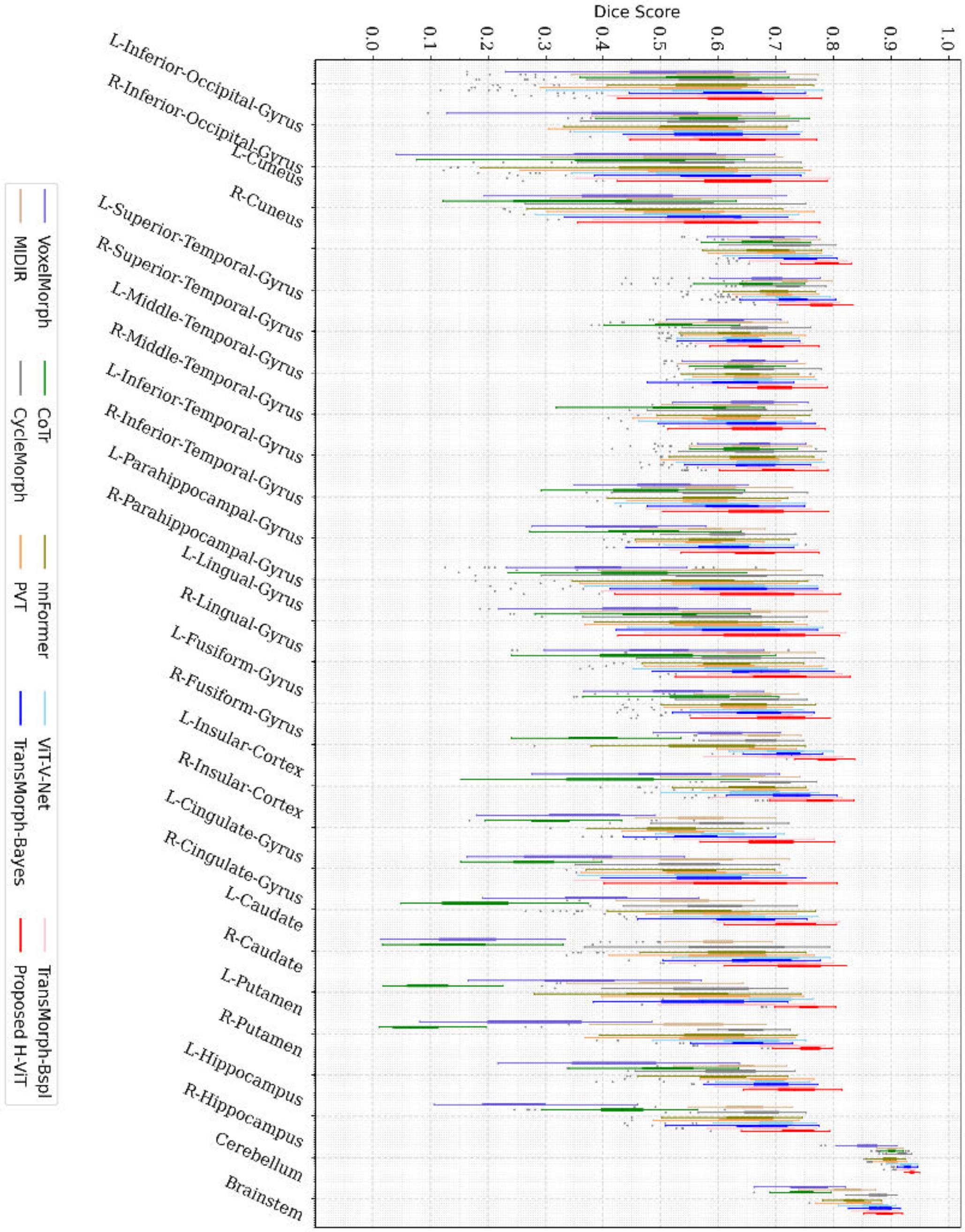


Figure 19. Dice score results for the *patient-to-atlas* registration of various methods on the *LPBA* dataset per anatomical structure

D.5. Experiments with Mindboggle

We apologize for the mistake regarding Tab. 6, where the results of ‘Inter-Patient’ and ‘Patient-to-Atlas’ were erroneously reported interchangeably. This has been rectified and reported in Tab. 11. We will modify the inaccuracy in Tab. 6 during the rebuttal.

Method	Attention	Inter-Patient Registration		Patient-to-Atlas Registration	
	Mechanism	Dice ↑	$\lvert J_{\Phi} \rvert \leq 0$ (%) ↓	Dice ↑	$\lvert J_{\Phi} \rvert \leq 0$ (%) ↓
Affine		0.537±0.041	–	0.534±0.034	–
VoxelMorph [6]	–	0.674±0.197	0.821±0.170	0.666±0.201	0.831±0.163
MIDIR [48]	–	0.637±0.197	0.403±0.215	0.539±0.292	0.347±0.205
CycleMorph [32]	–	0.679±0.194	1.044±0.211	0.671±0.199	1.064±0.189
CoTr [68]	Self-Att.	0.633±0.214	0.691±0.163	0.630±0.218	0.701±0.141
nnFormer [74]	Self-Att.+Global-Att.	0.622±0.210	1.077±0.210	0.618±0.213	1.090±0.189
PVT [63]	Self-Att.	0.588±0.214	2.006±0.254	0.583±0.216	2.034±0.217
ViT-V-Net [8]	Self-Att.	*0.700±0.186*	1.168±0.225	0.695±0.187	0.840±0.573
TransMorph-Bayes [10]	Self-Att.	0.699±0.186	0.702±0.106	0.695±0.189	0.716±0.082
TransMorph-bspl [10]	Self-Att.	0.699±0.181	<**0.001**	*0.695±0.183*	<**0.001**
Proposed H-ViT	Self-Att. + Cross-Att.	**0.731±0.170**	*0.328±0.061*	**0.726±0.173**	*0.335±0.049*

Table 11. Quantitative evaluation results for the registration methods on the *Mindboggle* dataset for 41 anatomical structures over 111 random pairs for inter-patient and 222 pairs for patient-to-atlas registrations.

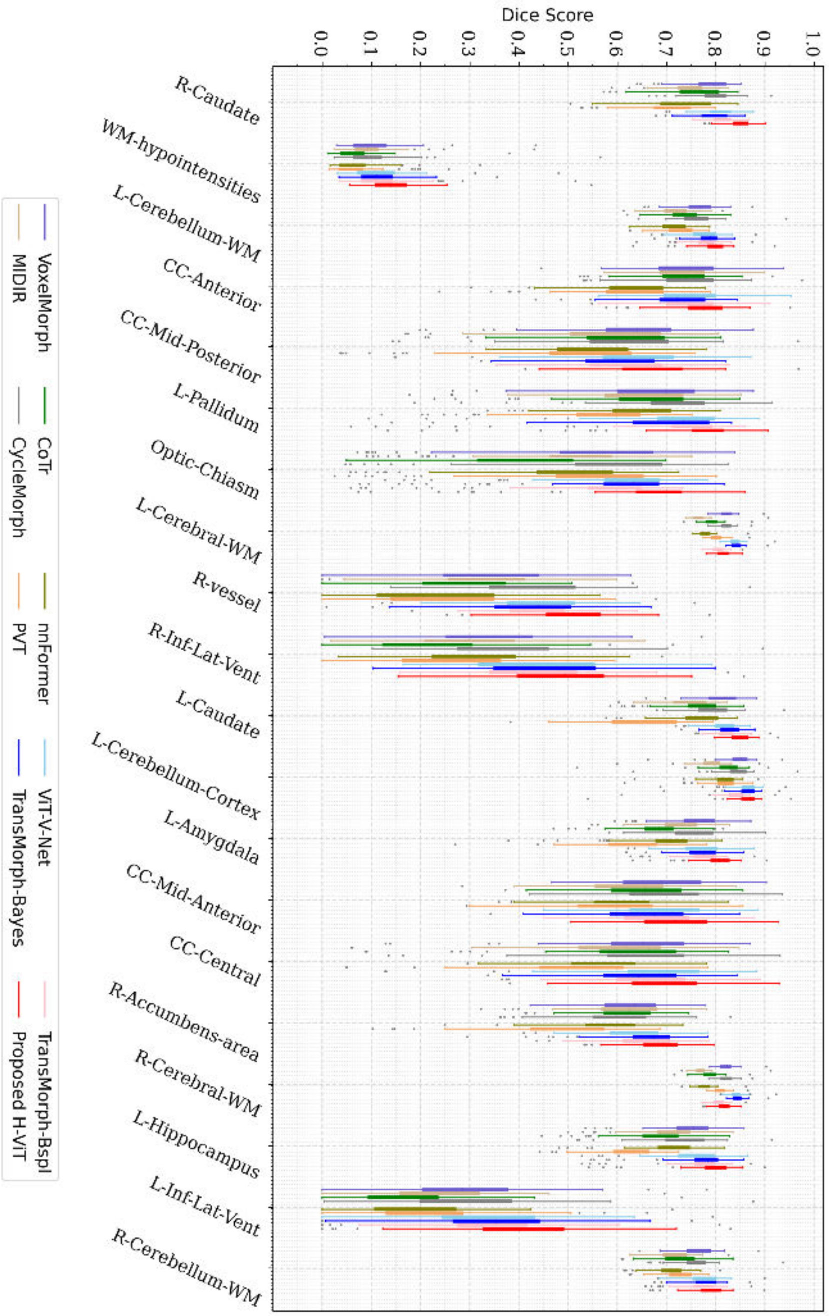


Figure 20. Dice score results for the *inter-patient* registration of various methods on the *Mindboggle* dataset per anatomical structure (continued)

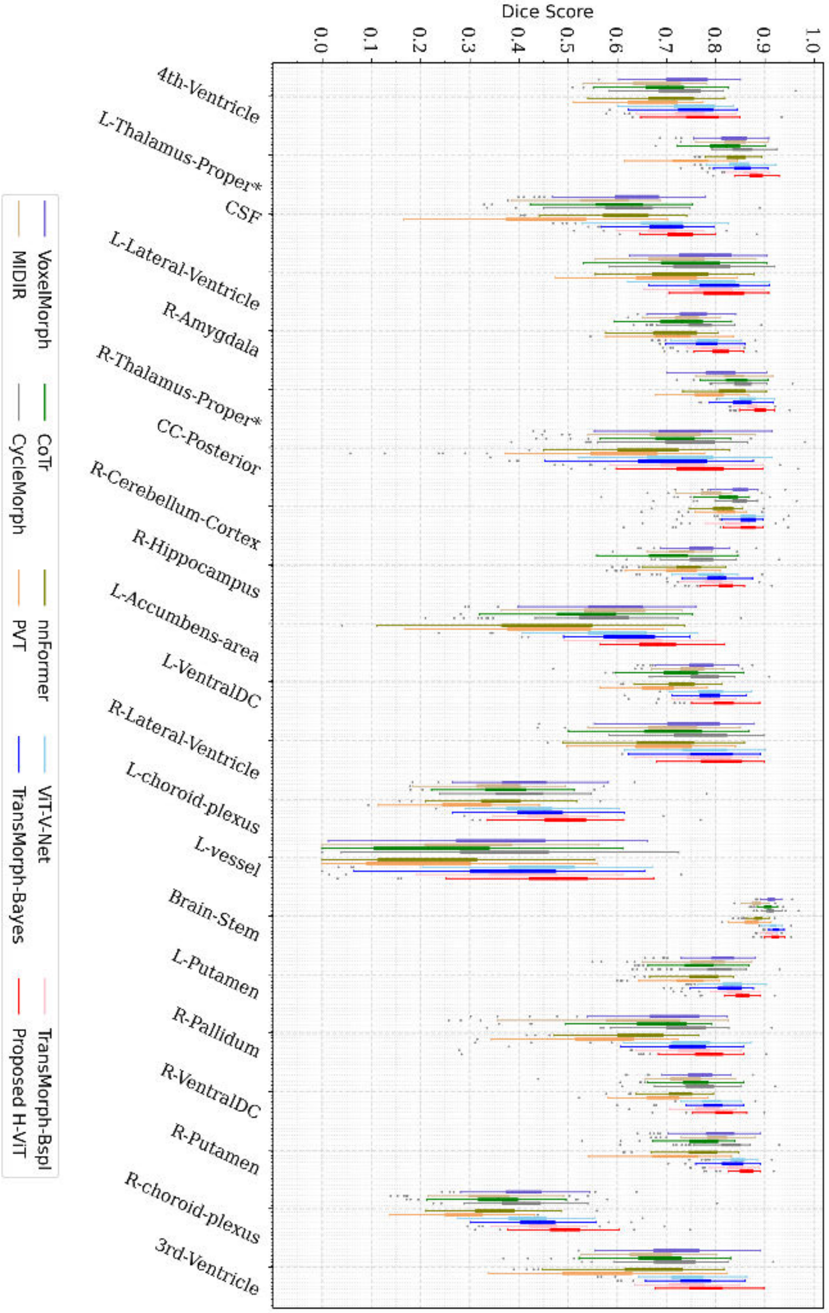


Figure 21. Dice score results for the *inter-patient* registration of various methods on the *Mindboggle* dataset per anatomical structure

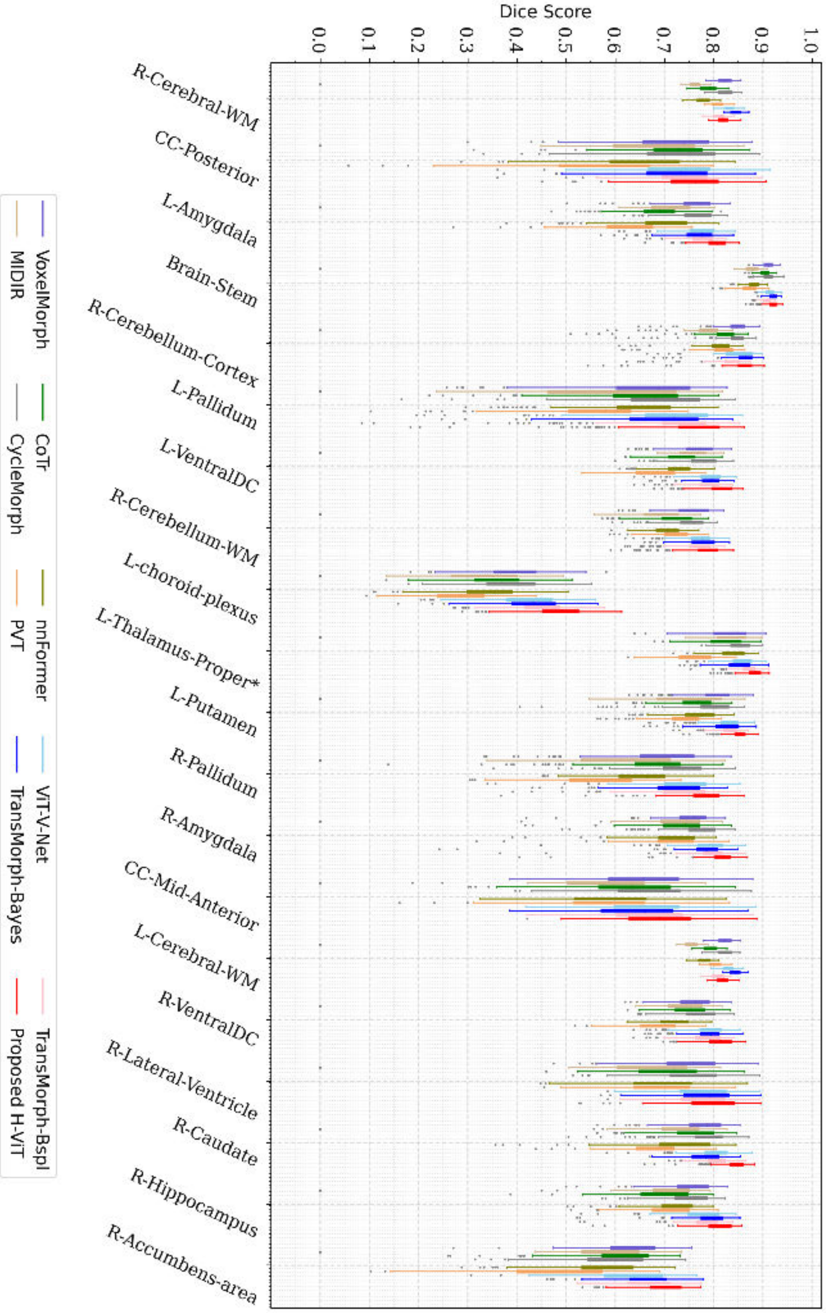


Figure 22. Dice score results for the *patient-to-atlas* registration of various methods on the *Mindboggle* dataset per anatomical structure (continued)

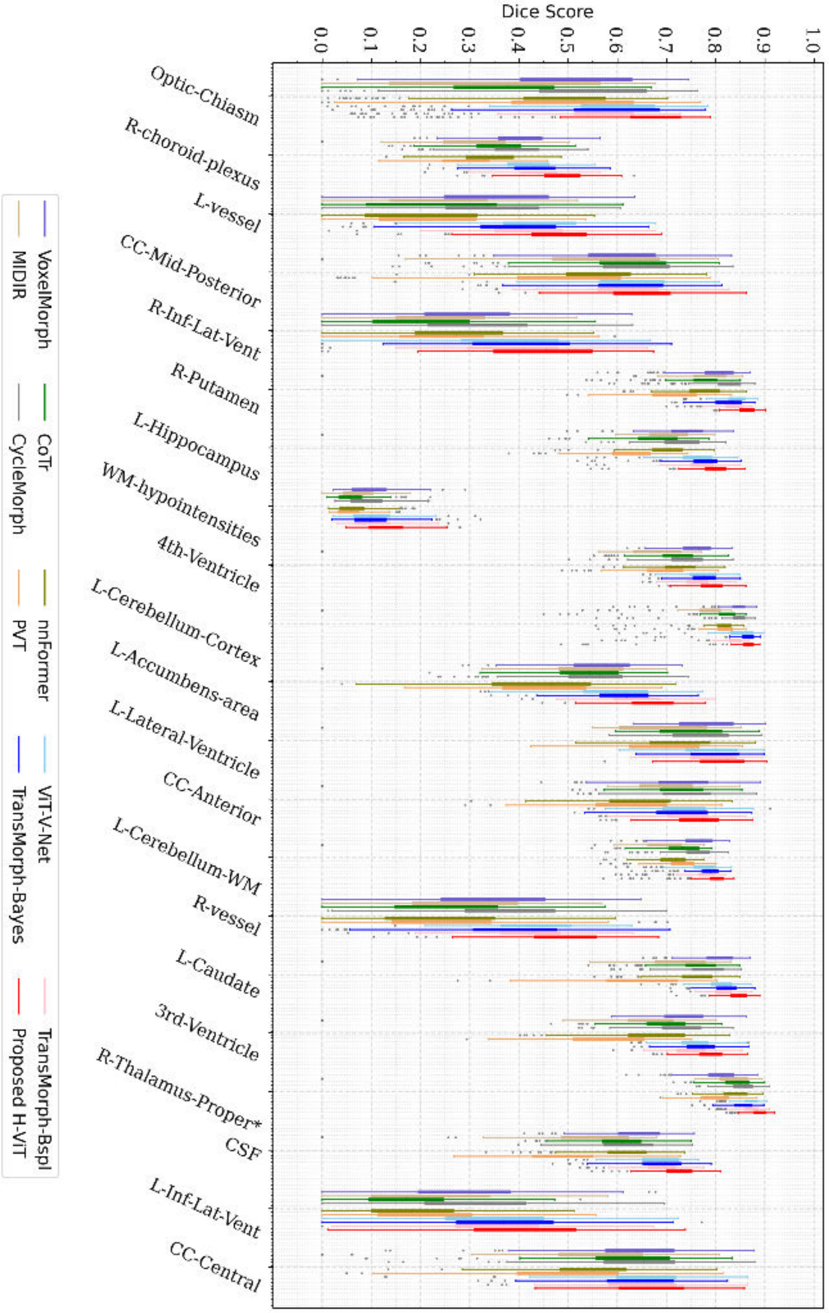


Figure 23. Dice score results for the *patient-to-atlas* registration of various methods on the *Mindboggle* dataset per anatomical structure

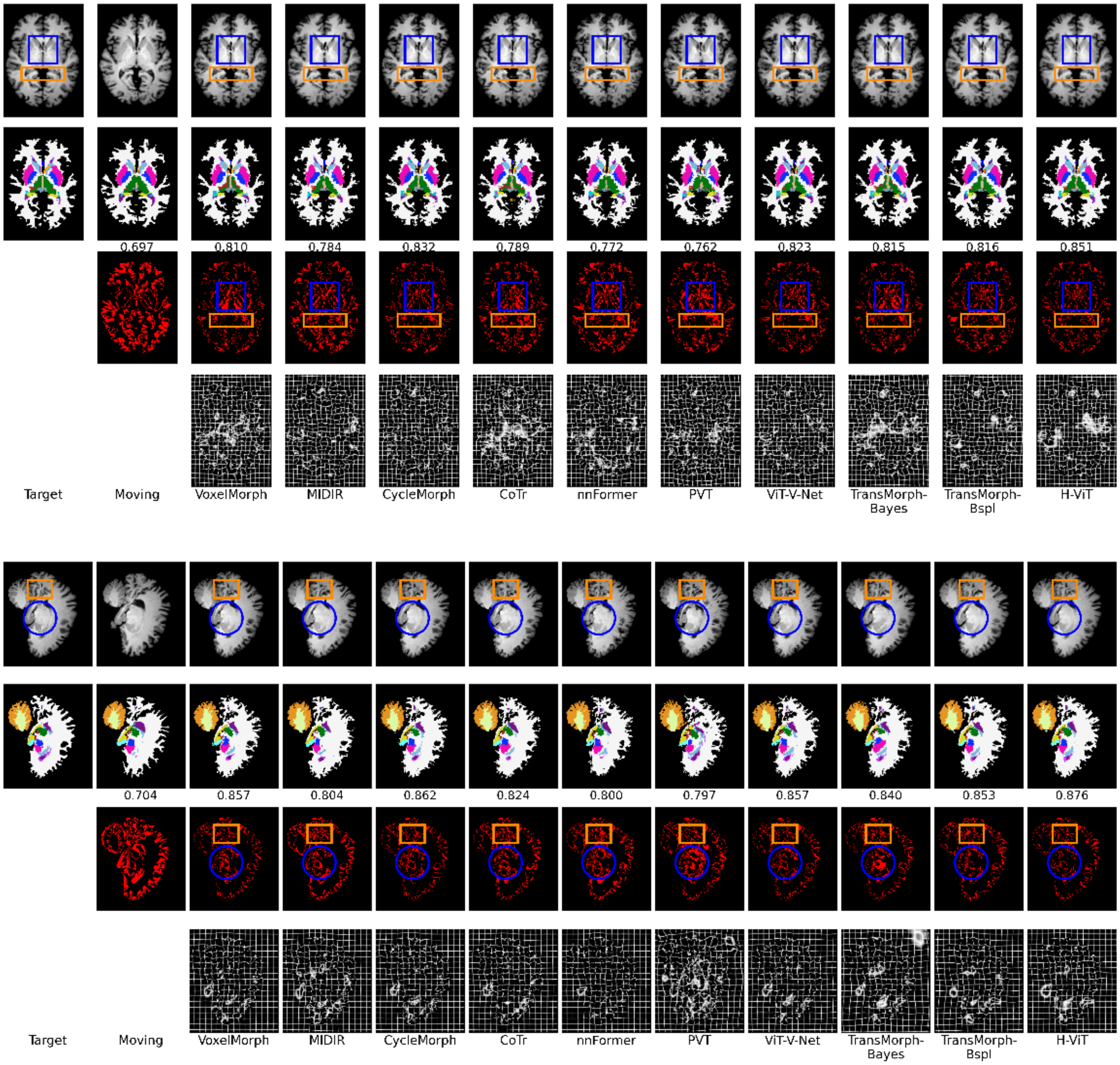


Figure 24. Examples of axial and sagittal slices from the Mindboggle-MMRR dataset (continued)

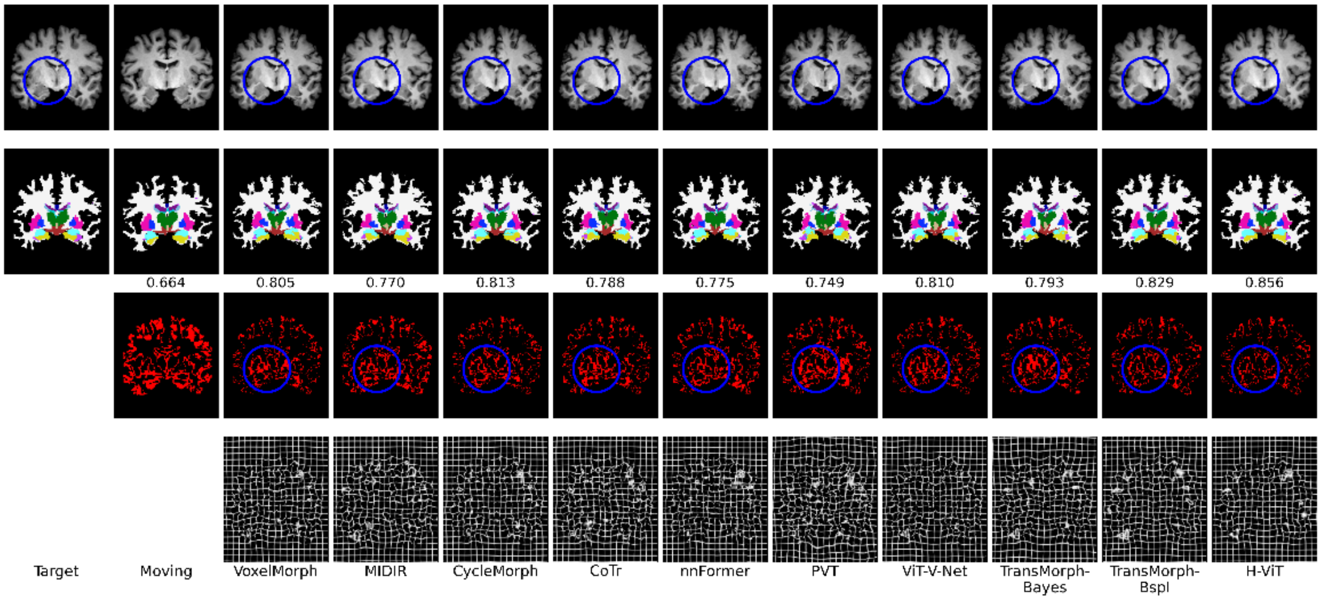


Figure 25. An example coronal slice from the Mindboggle-MMRR dataset

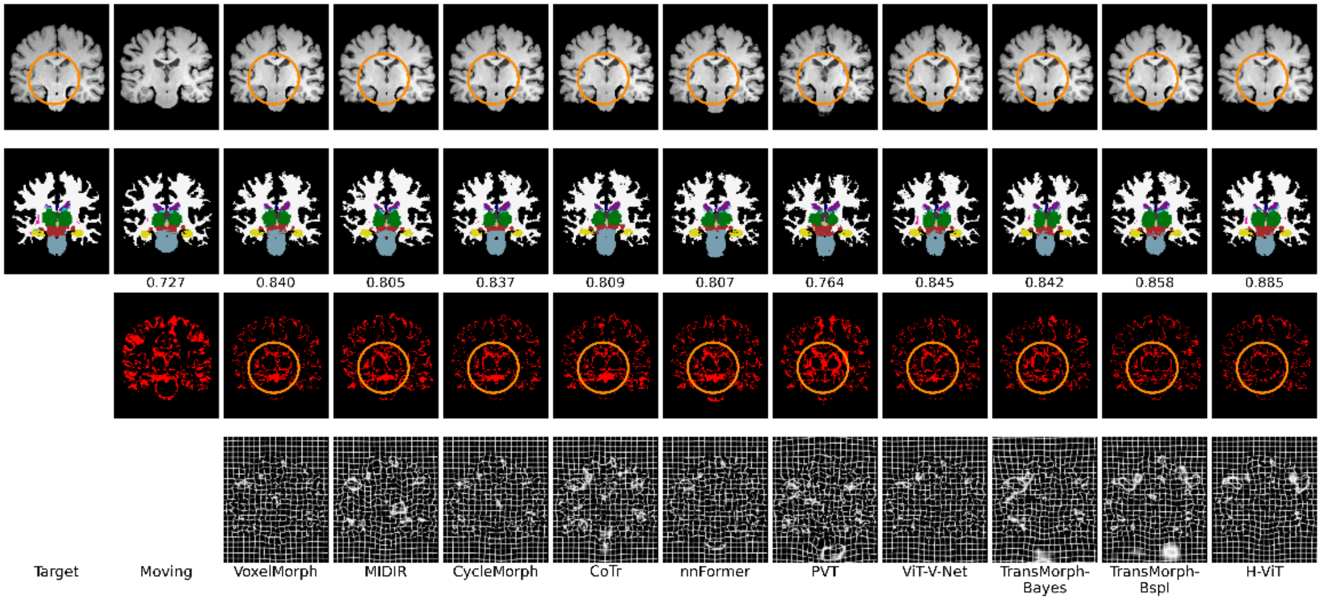


Figure 26. An example coronal slice from the Mindboggle-NKI dataset (continued)

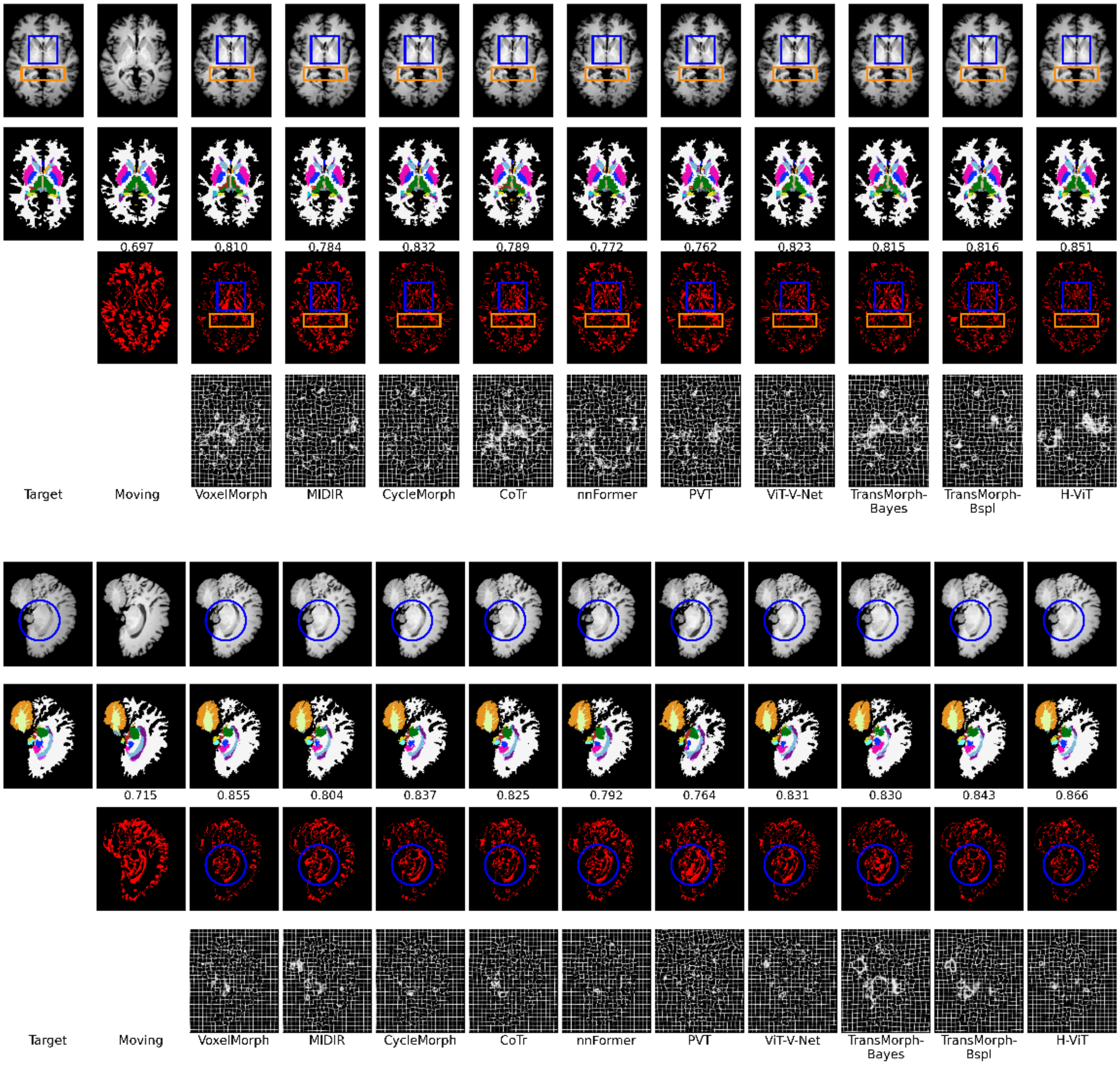


Figure 27. Examples of axial and sagittal slices from the Mindboggle-NKI dataset

D.6. Visualization of the H-ViT's dual-attention

Fig. 28 to Fig. 31 illustrate examples of the attention outputs of the dual-attention mechanism for the final layer of the deformation field, i.e. $S_h = 4$. The attention maps are averaged over the embedding channels, $\bar{\mathbf{A}} = \sum_{i=<f_e>} \mathbf{A}_i$. We also reported the energy of the averaged attention map as an indicator of activated weights: $\mathcal{E}_{\bar{\mathbf{A}}} = \frac{1}{|\Omega|} \sum_{p\in\Omega} ||\bar{\mathbf{A}}||^2$. The registration tasks were performed in 3D, while this paper presents visualizations of representative slices in 2D. As evident in the figures and highlighted by the energy values, incorporating more cross-attention units into the computation increasingly activates the feature maps. This yields an improved representation of the deformation field, shown quantitatively and qualitatively in the Experimental Sections.

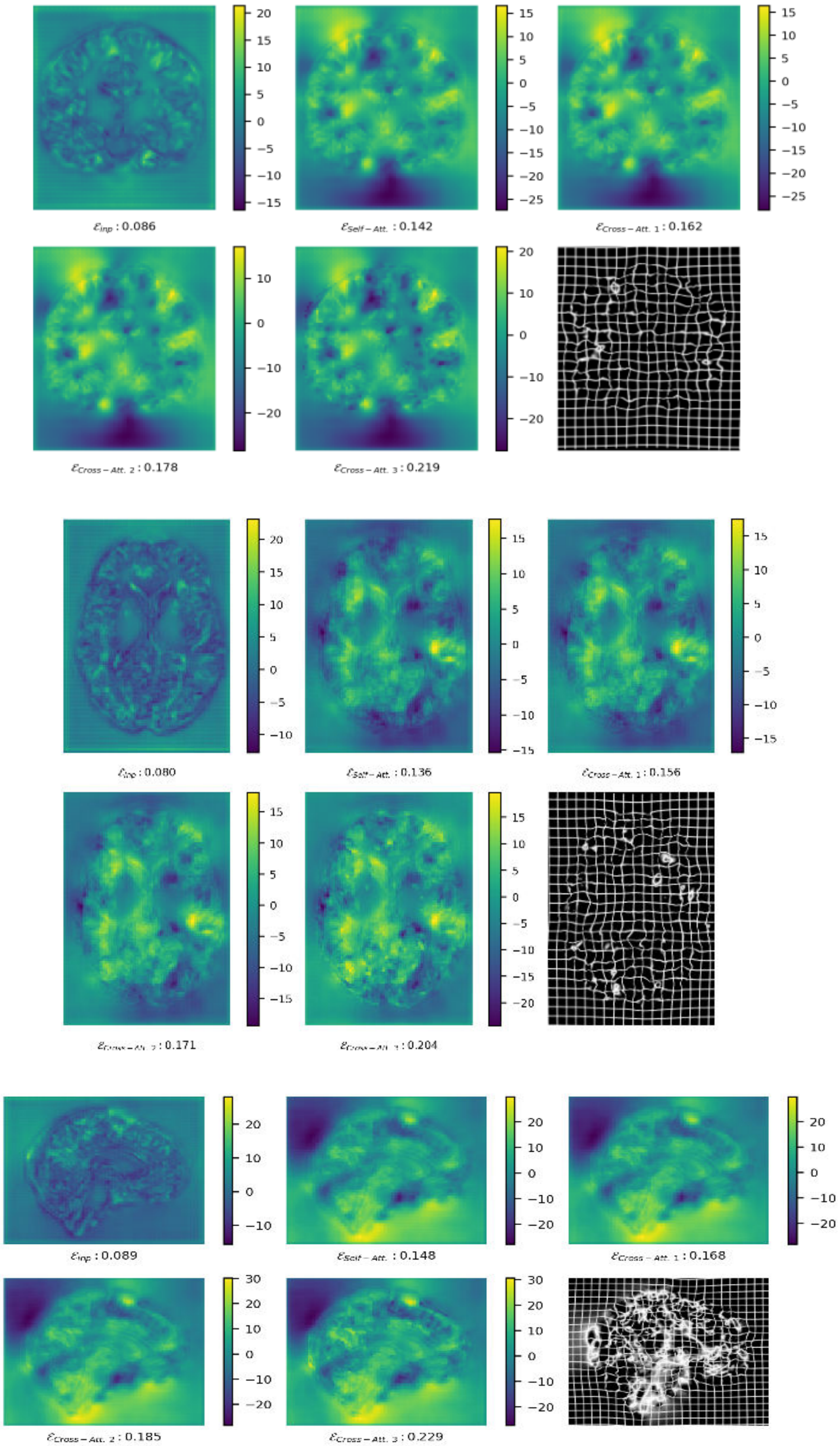
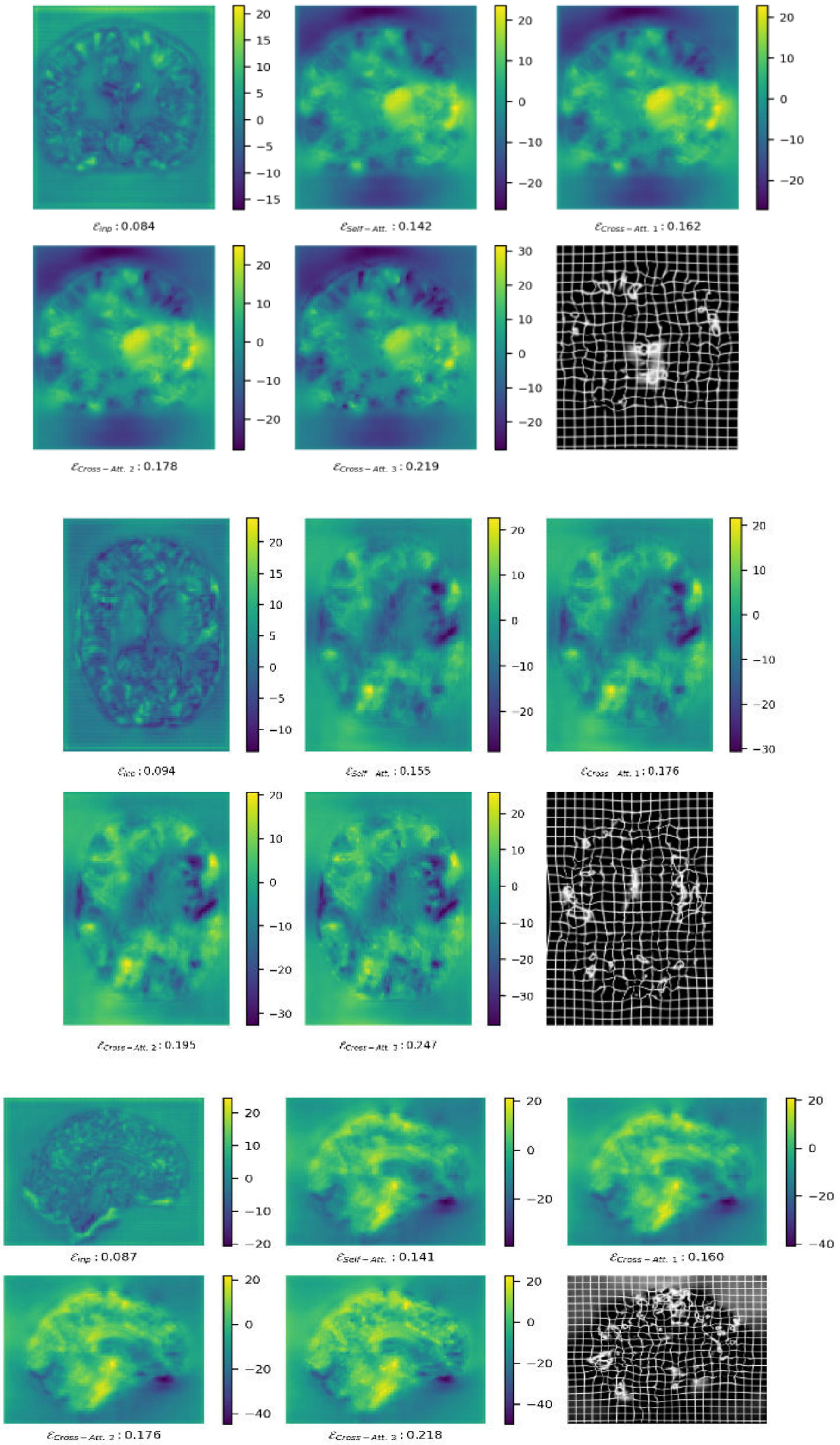

$\mathcal{E}_{inp}$: 0.086
$\mathcal{E}_{Self-Att.}$: 0.142
$\mathcal{E}_{Cross-Att.\ 1}$: 0.162
$\mathcal{E}_{Cross-Att.\ 2}$: 0.178
$\mathcal{E}_{Cross-Att.\ 3}$: 0.219
$\mathcal{E}_{inp}$: 0.080
$\mathcal{E}_{Self-Att.}$: 0.136
$\mathcal{E}_{Cross-Att.\ 1}$: 0.156
$\mathcal{E}_{Cross-Att.\ 2}$: 0.171
$\mathcal{E}_{Cross-Att.\ 3}$: 0.204
$\mathcal{E}_{inp}$: 0.089
$\mathcal{E}_{Self-Att.}$: 0.148
$\mathcal{E}_{Cross-Att.\ 1}$: 0.168
$\mathcal{E}_{Cross-Att.\ 2}$: 0.185
$\mathcal{E}_{Cross-Att.\ 3}$: 0.229

Figure 28. Example attention maps of the deformation field at the fourth layer ($S_h = 4$) in H-ViT. The attention maps are averaged over embedding channels. Numerical values beneath each map indicate respective energy levels. $\mathcal{E}_{Cross-Att:n}$ denotes the energy level for the output of cross-attention at the n-th block. 'inp' denotes the input feature map.



$\mathcal{E}_{inp}$: 0.084

$\mathcal{E}_{Self-Att.}$: 0.142

$\mathcal{E}_{Cross-Att.\ 1}$: 0.162

$\mathcal{E}_{Cross-Att.\ 2}$: 0.178

$\mathcal{E}_{Cross-Att.\ 3}$: 0.219

$\mathcal{E}_{inp}$: 0.094

$\mathcal{E}_{Self-Att.}$: 0.155

$\mathcal{E}_{Cross-Att.\ 1}$: 0.176

$\mathcal{E}_{Cross-Att.\ 2}$: 0.195

$\mathcal{E}_{Cross-Att.\ 3}$: 0.247

$\mathcal{E}_{inp}$: 0.087

$\mathcal{E}_{Self-Att.}$: 0.141

$\mathcal{E}_{Cross-Att.\ 1}$: 0.160

$\mathcal{E}_{Cross-Att.\ 2}$: 0.176

$\mathcal{E}_{Cross-Att.\ 3}$: 0.218

Figure 29. Example attention maps of the deformation field at the fourth layer ($S_h = 4$) in H-ViT

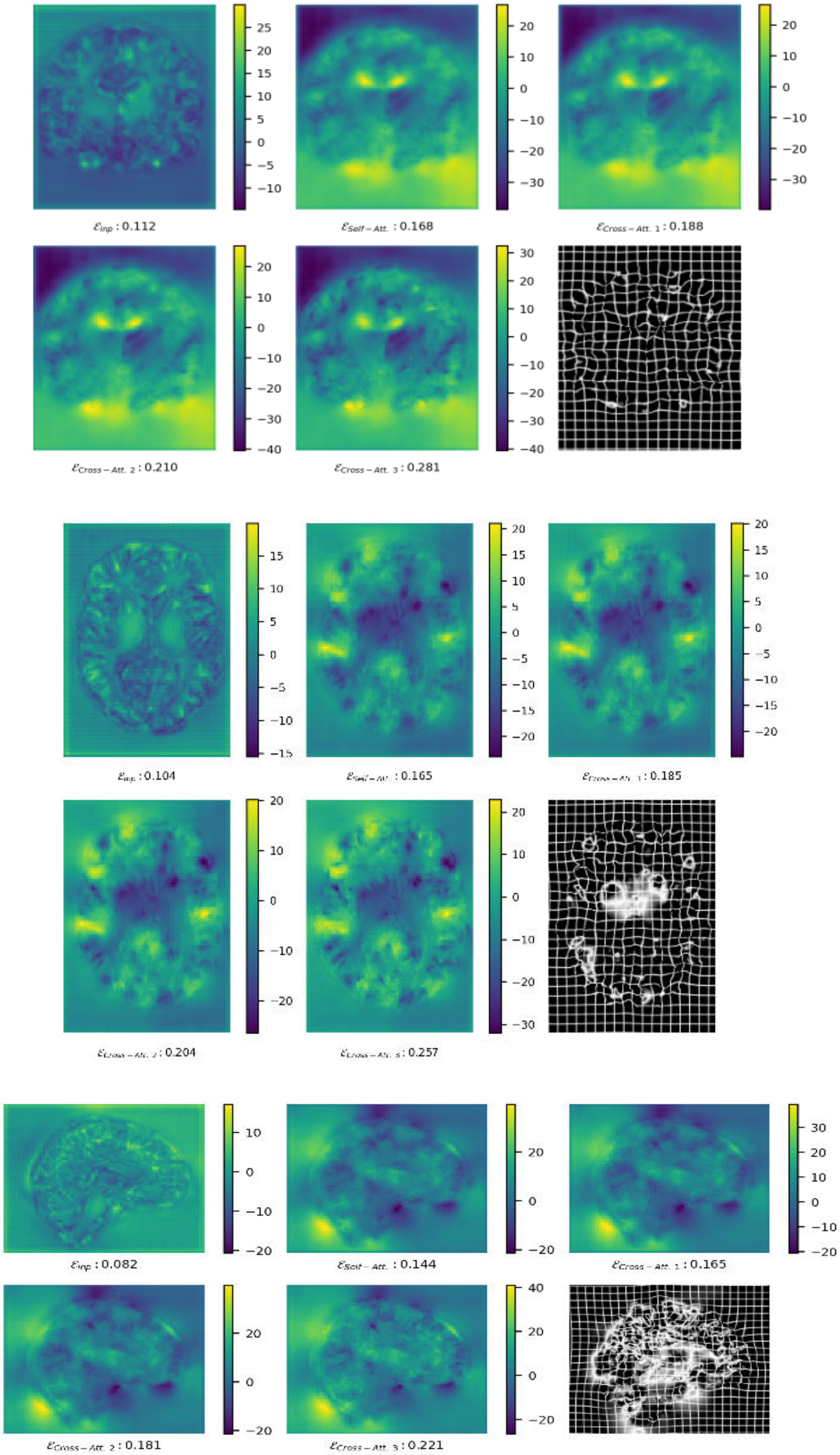


Figure 30. Example attention maps of the deformation field at the fourth layer ($S_h = 4$) in H-ViT

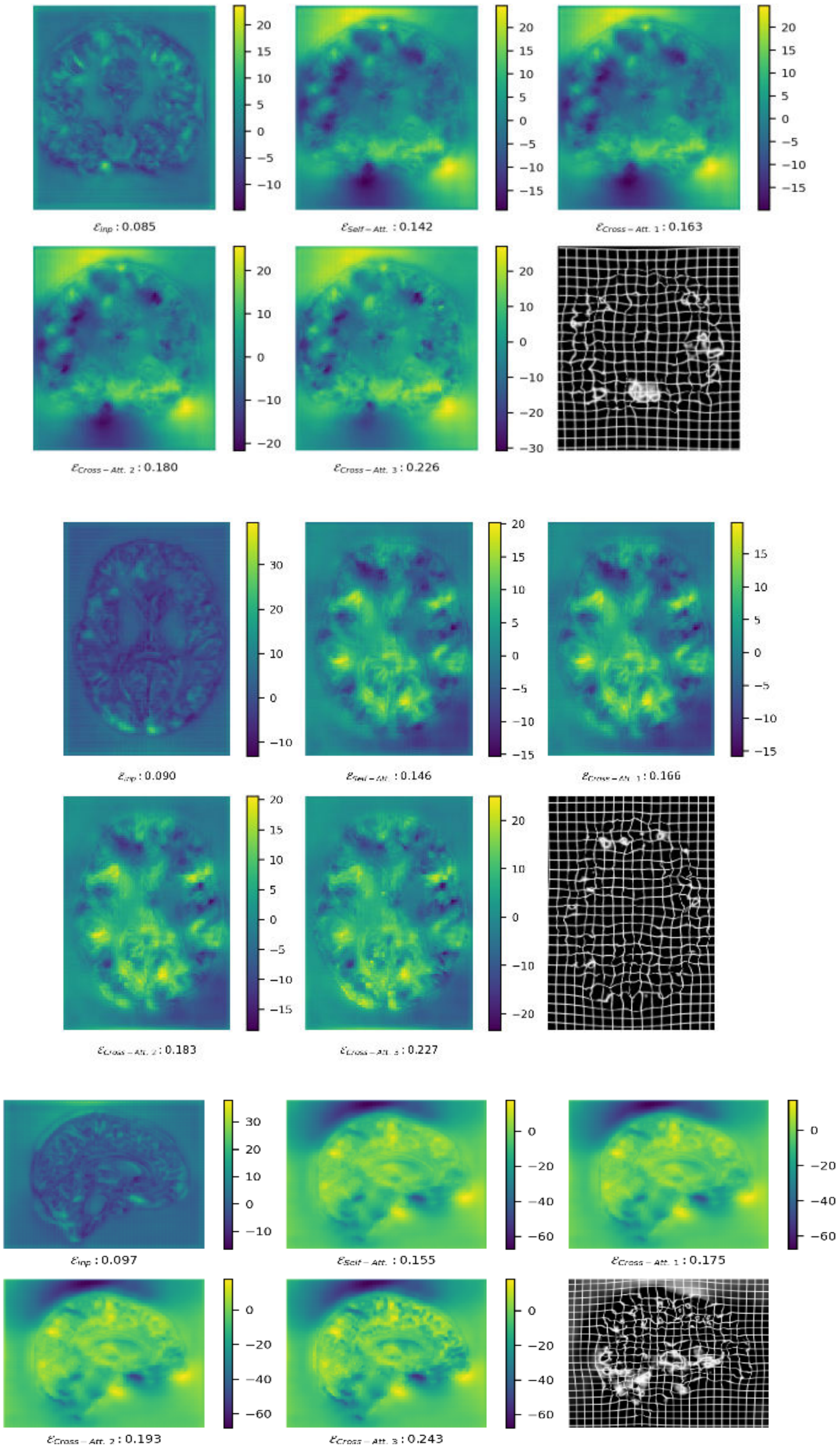

$\mathcal{E}_{inp}$: 0.085
$\mathcal{E}_{Self-Att.}$: 0.142
$\mathcal{E}_{Cross-Att.\ 1}$: 0.163
$\mathcal{E}_{Cross-Att.\ 2}$: 0.180
$\mathcal{E}_{Cross-Att.\ 3}$: 0.226
$\mathcal{E}_{inp}$: 0.090
$\mathcal{E}_{Self-Att.}$: 0.146
$\mathcal{E}_{Cross-Att.\ 1}$: 0.166
$\mathcal{E}_{Cross-Att.\ 2}$: 0.183
$\mathcal{E}_{Cross-Att.\ 3}$: 0.227
$\mathcal{E}_{inp}$: 0.097
$\mathcal{E}_{Self-Att.}$: 0.155
$\mathcal{E}_{Cross-Att.\ 1}$: 0.175
$\mathcal{E}_{Cross-Att.\ 2}$: 0.193
$\mathcal{E}_{Cross-Att.\ 3}$: 0.243

Figure 31. Example attention maps of the deformation field at the fourth layer ($S_h = 4$) in H-ViT

E. Ablation study on H-ViT's parameters

Parameter	Dice ↑	$\lvert J_\Phi \rvert \leq 0$ (%) ↓
Number of Heads		
8	0.801±0.072	0.201±0.130
16	0.803±0.072	0.201±0.130
32	0.803±0.073	0.201±0.132
64	0.805±0.072	0.211±0.132
Depth		
1	0.803±0.073	0.201±0.132
2	0.804±0.072	0.215±0.133
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
$4 \times 4 \times 4$	0.803±0.073	0.202±0.133
$6 \times 6 \times 6$	0.804±0.073	0.207±0.132
Drop rate		
0	0.803±0.073	0.201±0.132
0.1	0.802±0.073	0.186±0.131
0.2	0.801±0.073	0.179±0.128

Table 12. Ablation study on parameters of a small H-ViT for the IXI registration.