TSHA: Triple-Stream Hierarchical Attention for Deformable Image Registration

Naeem Hussain [1,2], Zhiyue Yan [1,2], and Wenming Cao [1,2,*]

[1] College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
[2] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, China
Email: 2151432011@email.szu.edu.cn (N.H.); 2150432014@email.szu.edu.cn (Z.Y.); wmcao@szu.edu.cn (W.G.)
*Corresponding author

***Abstract*—Deformable Medical Image Registration (DMIR) is an important tool in medical imaging because it aligns images from several modalities or periods, resulting in a more accurate diagnosis and treatment planning. Single- and double-stream architectures frequently struggle with the challenges of patient organ movements and artifacts from numerous scanning devices. The study presents a transformer-based triple-stream network that uses hierarchical dual-attention techniques to enhance the relationships between moving-fixed and augmented moving-fixed images. We produced two different deformation fields: one with moving and fixed images and the other with augmented moving and fixed images using multi-scale deformation. To improve the overall registration accuracy, we used advanced motion-correction techniques to combine these fields to create the final deformation field. The effectiveness of the proposed model was validated using three benchmark datasets: OASIS, LPBA40, and Learn2Reg (Lung CT). The results show that our technique consistently outperforms learning-based methods, with outstanding results for both single and cross-modalities in medical image registration.**

***Keywords*—deformable image registration, transformer, multi-scale deformation, attention mechanism, unsupervised learning**

I. Introduction

Deformable Medical Image Registration (DMIR), is essential for diagnosis and treatment monitoring [1]. DMIR facilitates surgical treatment by providing spatial correspondences between image pairs, allowing the integration of several imaging modalities. Tumor examination, organ mapping, and surgical planning are essential for improving diagnostic and therapeutic outcomes [2]. Deformable medical image registration has become significant but challenging for several decades. Traditional methods, such as iterative optimization of similarity functions for each image pair, are time-consuming [3, 4]. Deep learning-based studies have improved deformable image registration [5, 6]. The absence of specific and comprehensive ground-truth data limits the use of supervised techniques [7]. These limitations have encouraged the development of unsupervised techniques [8, 9] that do not require explicit annotations, resulting in broader applicability across various medical imaging instances. Although improvements have been made to unsupervised image registration techniques, significant challenges remain. One of the most difficult tasks is to acquire helpful feature representations and establish precise spatial correspondence. Compared with traditional techniques, CNNs have significantly improved performance in this field. Transformer networks outperform CNNs in image registration by efficiently capturing these dynamics using attention processes; however, CNNs have challenges with long-range spatial relationships [10, 11]. Transformers with attention mechanisms are more significant in non-rigid medical image registration [11, 12].

To capture various motion modalities, a Motion Decomposition Transformer uses the built-in advantages of the transformer [13]. This approach increases the sensitivity of the registration process by facilitating the merging of features from different levels by integrating a competitive weighting module and decomposition transformer in the decoder stage. Furthermore, TransMorph [9], a hybrid network that combines Transformers and CNNs, improves image registration using a Swin Transformer module for encoding and convolutional blocks during the decoding stage. Transformer models use a self-attention mechanism to transform image inputs into sequences, allowing them to examine pixel relationships and build an overall understanding of the image. CNN-based image registration methods feature fundamental features in specific regions. By incorporating a Vision Transformer module into the CNN architecture, a hybrid model for image registration was developed, combining the strengths of both Transformers and CNNs. This hybrid technique offers superior registration performance on brain MR datasets compared to CNN-based models, such as VoxelMorph [12]. The attention mechanism, especially in a transformer [14], is important in medical image registration because it focuses on relevant features across image pairs, improving registration accuracy.

Our research presents a triple-stream hierarchical attention network that uses a triple-stream encoder-decoder architecture to perform feature extraction and

Manuscript received February 19, 2025; accepted April 14, 2025; published October 24, 2025.

matching. We improved the efficiency of the hierarchical attention approach in building correspondences and matching features across the three streams. The triple-stream hierarchical attention network, shown in Fig. 1, overcomes the constraints of standard registration networks, which primarily use single and dual-stream techniques.

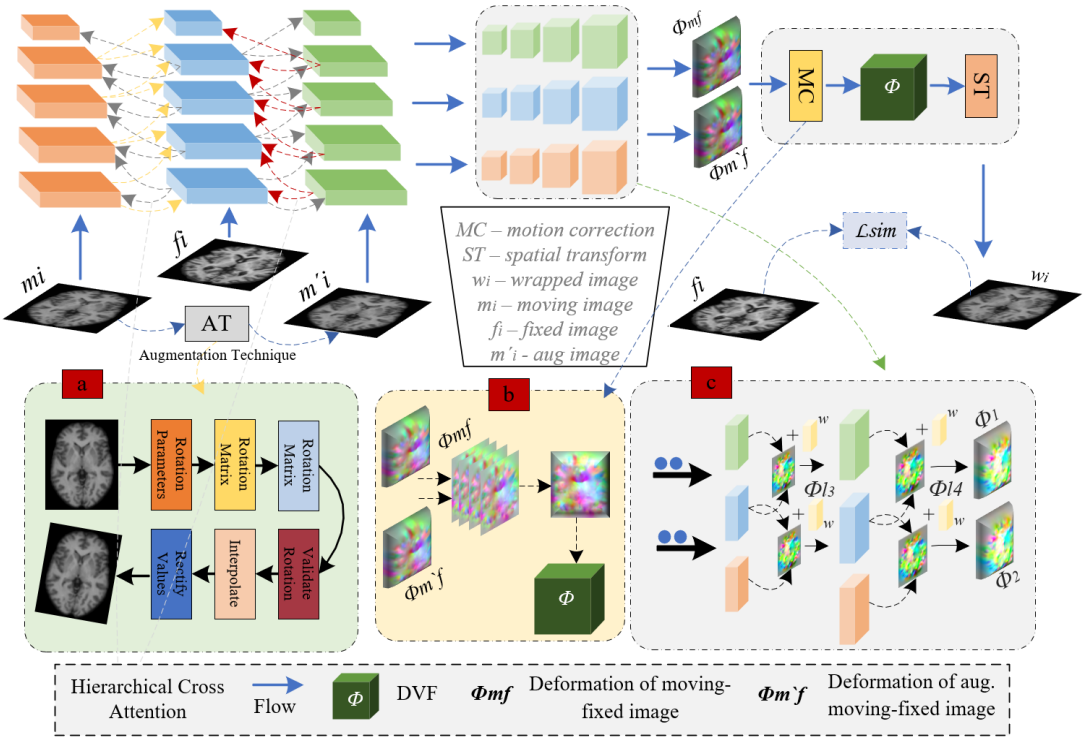

Fig. 1. The overview of our proposed triple-stream hierarchical attention method. (a) Rotation techniques. (b) Motion correction was applied to both deformation fields to get the final deformation. (c) Multi-scale deformation.

Although existing registration methods have improved, there is still substantial scope for improvement. 1) Single and dual-stream networks commonly fail because of an absence of image diversity, impair their capacity to precisely align images, and limit their clinical utility. 2) The current registration network typically uses multi-scale decoders for a single moving image despite variations like motion artifacts, which might result in inaccurate registered images.

The contributions of our work can be summarized as follows:

- We introduce a triple-stream hierarchical attention technique in our framework. This technique enables precise data sharing across parallel and hierarchical layers, considerably increasing the accuracy of our image registration method.
- We present an efficient Motion Correction (MoCo) approach for improving the deformation field by processing ϕ_{mf} and $\phi_{m'f}$ deformations. This approach significantly eliminates motion artifacts, improving the precision of deformations during movement.
- We conduct extensive experiments on three publicly available 3D medical imaging datasets – OASIS and LPBA40 (Brain MRI) and Learn2Reg (Lung CT), to validate the effectiveness of the proposed method.

II. Materials and Methods

A. Datset and Preprocessing

We evaluated our proposed method using two different 3D brain MRI datasets: OASIS and LPBA40. And one Lung CT dataset Learn2Reg. The OASIS dataset contains 405 T1W scans resampled to 80×96×112 pixels to reduce computational costs and improve registration accuracy. We used 225 and 150 scans for training and testing, respectively. We used the LPBA40 dataset for cross-validation, which consisted of 40 T1W scans with manual semi-annotations of 55 subcortical anatomical structures. These annotations offer an effective structure for cross-dataset validations. We selected five scans from the test sets of each dataset to serve as a fixed atlas. The remaining 145 scans were used as moving images for OASIS, while the remaining 35 scans were used for LPBA40. Both datasets were subjected to standard preprocessing using FreeSurfer, which included affine space normalization, motion correction, skull stripping, and subcortical structure segmentation. We also used the Learn2Reg dataset for the model's cross-evaluation, which consists of 20 training scan pairs and 10 test scan pairs, each containing a lung CT scan and the corresponding segmentation mask. These Lung CT scans offer effective structure for cross-validation.

B. Network Architecture

As illustrated in Fig. 1, our proposed method (TSHA) generates an augmented moving image using multiple augmentation techniques including noise addition and rotation. These augmented images, along with the original moving and fixed images, are processed through separate UNet. The 3D images are flattened and projected into different dimensions after being separated into non-overlapping 4×4×4 patches. The moving-fixed and augmented moving-fixed pairs use hierarchical attention techniques to capture and share complex features across the three UNet models. Each UNet decoder layers generate multi-scale outputs, resulting in two deformation fields: one derived from the moving-fixed pairs and another from the augmented moving-fixed pairs. We use extensive motion correction techniques to fine-tune the final deformation field, which significantly improves the accuracy of the registration.

C. Triple-Stream Hierarchical Attention

In this section, we describe the key importance of the triple-stream network, which is that hierarchical attention is essential for incorporating information throughout moving m_i fixed f_i and augmented moving images. Each UNet stream itself is encoded into various layers of feature maps, indicated as $\phi m(l)$, $\phi\mathrm{m(l)}$ and $\phi m'(l)$, $\phi\mathrm{m'(l)}$ for every layer. The self-attention mechanism within each layer modifies intra-stream feature representations by projecting voxel patches into query Q, key K, and value V embeddings and applying the attention operation:

$$A_{self} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_x}}\right)\mathrm{V} \quad (1)$$

Hierarchical cross-attention enhances the process by allowing for interaction between layers and streams. These layers evaluate features from distinct streams, such as $\phi m(l)$, $\phi f(l)$, and $\phi m'(l)$, as well as scales within the same stream.

$$A_{cross}(l, l+1) = softmax\left(\frac{QM(l)KF(l+1)^T}{\sqrt{d_x}}\right) VF(l+1) \quad (2)$$

$$A_{cross}(l, l+1) = softmax\left(\frac{QM'(l)KF(l+1)^T}{\sqrt{d_x}}\right) VF(l+1) \quad (3)$$

here, l represents various layers, allowing the model to incorporate spatial and contextual information across the triple-stream. This hierarchical architecture enables detailed local feature enhancement while ensuring robust global context integration, which is essential to accurately estimating the deformation fields required for perfect image registration.

D. Motion Correction in Deformation Fields

As illustrated in Fig. 1(b), we propose an efficient Motion Correction (MoCo) technique to enhance the deformation fields generated by moving-fixed ϕ_{mf} and augmented moving-fixed $\phi_{m'f}$ image pairs. Identifying the importance of effectively creating the motion artifacts, we focus more on the deformation field from the augmented moving-fixed pairs, enhancing it through rotation and other transformations to better mimic patient movements. The motion correction technique combines ϕ_{mf} and $\phi_{m'f}$ using the weighted combining technique, with the augmented pair obtaining a higher weight α, expressed as:

$$\phi = \alpha\phi_{m'f} + (1 - \alpha)\phi_{mf} \quad (4)$$

The augmented pair's increased reliability and essential function in the motion correction process are reflected in an $\alpha > 0.5$. A complex motion correction technique dynamically updated the final deformation field DVF ϕ, to improve alignment and resolve anatomical and positional errors, using vector spaces and matrix operations to assure accuracy and efficiency.

III. Experimental Results

We validated our proposed method on both single and cross-modality settings. Within the same modality, the model was first trained and tested on the OASIS Brain MRI dataset. For the cross-validation, we tested our proposed model on the LPBA40 dataset. We performed further validation of the model on the Learn2Reg Lung CT dataset to evaluate its robustness in cross-validation applications. Results demonstrate that our method is effective for image registration tasks. Comprehensive experiments were carried out to evaluate its effectiveness in various learning environments. Table I illustrates our method's memory usage across three datasets, with a focus on GPU memory for each dataset during testing.

TABLE I. GPU Memory Usage (in GB) during Testing across Different Datasets

Method	OASIS	
	Dataset	**Memory**
Ours (TSHA)	OASIS	7.54
	LPBA40	6.63
	Learn2Reg	7.01

A. Implementation Details

We used Python 3.9 and PyTorch, a deep learning framework utilizing CUDA 11.1 and an NVIDIA RTX3090 GPU in a Linux-based environment. We used the Adam optimizer to train our model, with a 10^{-4} learning rate. The hardware used for computation included an Intel Core i7 CPU @3.00 GHz with 32 GB of RAM.

B. Experimental Results on OASIS and LPBA40

Tables II and III illustrate our comparative analysis of various image registration techniques, focusing on the two datasets and multiple evaluation metrics. Our method, which utilizes the advantages of integrating CNN and Transformer structures, outperforms the others in terms of image accuracy and structural integrity. As shown in Table II, our method achieved 0.788 in OASIS and 0.735 in LPBA40 as shown in Table III, the highest DSC, optimal overlap, and fewer image alignment when compared to existing methods such as VoxelMorph, VIT-VNet, TransMorph, XMorpher, and TransMatch. The results show that our model significantly improves registration accuracy, with a 1.3% increase in DSC over TransMatch and a significant difference over other learning-based

techniques. The Jacobian matrix performance is also close to zero, with the smallest folding of $|J\phi| \leq 0$ in the deformation fields. The Dice score evaluates the entire similarity of images, whereas the Jacobian matrix examines the local consistency of structures within the images. Considering the percentage of image folding ($|J\phi| \leq 0$) in the deformation field, the positive impact of our approach is clear. The ablation study results are presented in Table IV, and provide additional support for our model performance. Qualitative evaluations demonstrate that our method performs well for both image alignment and anatomical preservation, as shown in Fig. 2. These outcomes provide a precedent for comparison in the field of medical image registration, highlighting not only our method's higher efficiency but also its increased clinical significance.

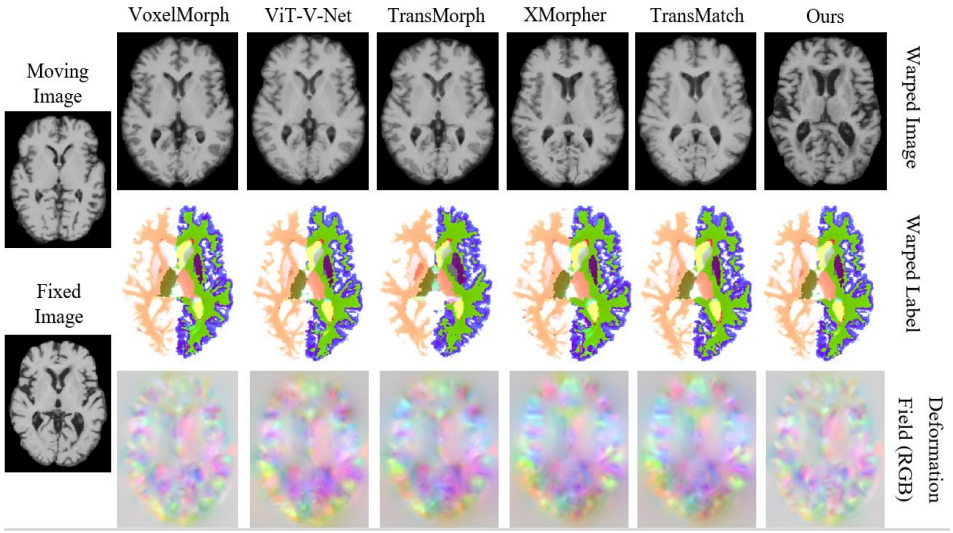

Fig. 2. Comparison of the visual results obtained from different deep learning registration methods on the OASIS dataset. (VM: VoxelMorph, ViT: ViT-V-Net, TM: TransMorph, XM: XMorpher, TransM: TransMatch).

TABLE II. QUANTITATIVE EVALUATION OF PROPOSED MODEL AGAINST OTHER METHODS ON OASIS DATASETS

Method	OASIS			
	DSC	% of $\|J\phi\| \leq 0$	STD$\|J\phi\|$	Time
VoxelMorph	0.744(0.006)	0.35(0.05)	< 0.1	0.301
ViT-V-Net	0.757(0.005)	0.35(0.08)	< 0.1	0.335
TransMorph	0.757(0.005)	0.27(0.07)	< 0.1	0.347
XMorpher	0.769(0.004)	0.26(0.05)	< 0.1	0.395
TransMatch	0.775(0.005)	0.35(0.06)	< 0.1	0.395
Ours	0.788(0.004)	0.45(0.05)	< 0.1	0.513

TABLE III. QUANTITATIVE EVALUATION OF PROPOSED MODEL AGAINST OTHER METHODS ON LPBA40 DATASETS

Method	OASIS			
	DSC	% of $\|J\phi\| \leq 0$	STD$\|J\phi\|$	Time
VoxelMorph	0.691(0.005)	0.33(0.05)	< 0.1	0.325
ViT-V-Net	0.710(0.006)	0.33(0.07)	< 0.1	0.339
TransMorph	0.719(0.004)	0.26(0.05)	< 0.1	0.350
XMorpher	0.727(0.007)	0.28(0.05)	< 0.1	0.295
TransMatch	0.729(0.006)	0.25(0.06)	< 0.1	0.365
Ours	0.735(0.005)	0.35(0.05)	< 0.1	0.552

TABLE IV. ABLATION STUDY ON THE COMPONENTS OF THE TRI-UNET AND MOCO ON THE OASIS DATASETS

Tri-UNet	Multi-Def	MoCo	Fusion	DSC
✓	✗	✗	✓	0.761
✓	✓	✗	✓	0.766
✓	✗	✓	✗	0.773
✓	✓	✓	✗	0.788

C. *Experimental Results on Learn2Reg Lung CT*

Table V illustrates the experimental results across three datasets-OASIS, LPBA40, and Learn2Reg evaluated in single and cross-modality settings. The results illustrate our proposed method's outstanding performance, particularly in the cross-modality scenario utilizing the Learn2Reg Lung CT dataset. As illustrated in Table V, qualitative evaluations demonstrate that our method performs well in both image alignment and anatomical preservation, even when applied to the cross-modality instance. The results from the Learn2Reg dataset shown in Fig. 3, which evaluates the method on Lung CT images (a different modality than the OASIS and LPBA40 brain MRI datasets), demonstrate the robustness and adaptability of our method. Considering the challenges given by modality differences, our method captured the complex deformation while effectively maintaining anatomical structures. This supports our method's ability to overcome modality restrictions, demonstrating its potential for improving medical image registration tasks across multiple imaging modalities.

TABLE V. QUANTITATIVE EVALUATION OF PROPOSED MODEL ON SINGLE AND CROSS-MODALITY. N SHOWS NOISE, AND R SHOWS ROTATION AUGMENTATION RESULTS.

Method	OASIS	
	Dataset	DSC
Ours (TSHA)	OASIS	0.779(0.005) N 0.788(0.004) R
	LPBA40	0.726(0.005) N 0.735(0.005) R
	Learn2Reg	0.695(1.02) R

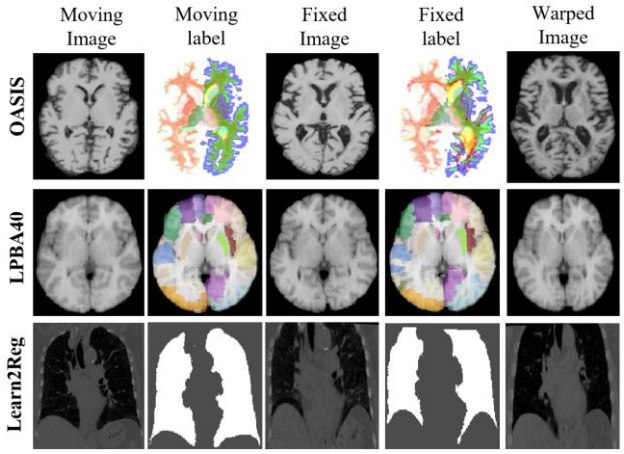

Fig. 3. Visual results of our method on different datasets.

IV. Conclusion

Our study introduces the Triple-Stream Hierarchical Attention (TSHA) approach, developed particularly for deformable medical image registration. The proposed network employs hierarchical attention processes to improve the relationship between pairs of moving-fixed and augmented moving-fixed images. Augmentation techniques create an augmented moving image from the original moving image, resulting in two separate deformation fields based on the multi-scale outputs of moving-fixed and augmented moving-fixed image pairs. Motion correction was applied to the deformation fields to obtain the final deformation. The effectiveness of the proposed model was validated using three benchmark datasets, including the OASIS, LPBA40, and Learn2Reg Lung CT dataset, to determine its performance in cross-modality validation. The comprehensive experimental findings demonstrate that our model outperforms prior methods.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Naeem Hussain conducted all aspects of the research, including conceptualization, coding, experiments, analysis, and writing the manuscript; Zhiyue Yan provided ongoing guidance and oversight throughout the project; Wenming Cao, as the corresponding author, contributed to the development of the core idea, supervised the evaluation process, and secured the necessary funding for the research, all authors had read and approved the final version of the manuscript.

Funding

The Fundamental Research Foundation of Shenzhen under Grant JCYJ20230808105705012 and the National Natural Science Foundation of China No. 61971290.

References

[1] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in Medicine & Biology*, vol. 46, no. 3, p. R1, 2001.

[2] H. Xiao, X. Xue, M. Zhu, *et al.*, "Deep learning-based lung image registration: A review," *Computers in Biology and Medicine*, 107434, 2023.

[3] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[4] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[5] X. Cao, J. Yang, J. Zhang, Q. Wang, P.-T. Yap, and D. Shen, "Deformable image registration using a cue-aware deep regression network," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1900–1911, 2018.

[6] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019.

[7] Y. Hu, M. Modat, E. Gibson, *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical Image Analysis*, vol. 49, pp. 1–13, 2018.

[8] S. Zhao, T. Lau, J. Luo, I. Eric, C. Chang, and Y. Xu, "Unsupervised 3d end-to-end medical image registration with volume tweening network," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1394–1404, 2019.

[9] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "Transmorph: Transformer for unsupervised medical image registration," *Medical Image Analysis*, vol. 82, 102615, 2022.

[10] J. Shi, Y. He, Y. Kong, J.-L. Coatrieux, H. Shu, G. Yang, and S. Li, "Xmorpher: Full transformer for deformable medical image registration via cross attention," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 217–226.

[11] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration," 2021.

[12] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[13] F. Godenschweger, U. Kägebein, D. Stucht, *et al.*, "Motion correction in MRI of the brain," *Physics in Medicine & Biology*, vol. 61, no. 5, p. R32, 2016.

[14] Z. Chen, Y. Zheng, and J. C. Gee, "Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration," *IEEE Transactions on Medical Imaging*, 2023.