TA-MIR: TEXT AGGREGATION FOR MULTIMODAL FEATURE REPRESENTATION IN MEDICAL IMAGE REGISTRATION

*Yuhe Dai ,[1,2], Zhiyong Huang [1,2], Weimin Huang [3]**
[1]School of Computing, National University of Singapore, Singapore
[2]NUS (Chongqing) Research Institute, China
[3]A*STAR Institute for Infocomm Research (I²R), Singapore

ABSTRACT

The aggregation of multimodal features in medical image registration remains underexplored, limiting the performance of current models in capturing complex anatomical relationships. Traditional convolutional neural networks (CNNs) often overlook the rich semantic information available from text, while existing approaches lack effective methods to combine spatial and contextual cues. In this paper, we propose Text Aggregation for Medical Image Registration (TA-MIR), a novel framework that enhances encoder-decoder architecture by incorporating anatomical text embeddings throughout the registration process. By employing large kernel blocks for improved receptive fields in U-Net and fusion blocks at each level, our model effectively integrates image features with semantic text information. Extensive experiments on three brain MRI datasets—OASIS, IXI, and LPBA40—demonstrate that our approach achieves state-of-the-art performance, significantly improving registration accuracy and anatomical coherence compared to traditional CNN and Transformer-based methods.

***Index Terms*—** Registration, medical image, text embedding, deep learning

1. INTRODUCTION

Deformable image registration (DIR) plays a vital role in medical imaging, with applications in disease diagnosis, treatment planning, and intraoperative guidance. It aligns datasets, often from different modalities or time points, by transforming them into a common coordinate system based on matched anatomical structures. Using a deformation model, spatial correspondences between moving and fixed images are established and optimized [1].

Since VoxelMorph [2] demonstrated the power of skip connections and an encoder-decoder structure for medical image registration (MIR), U-Net has been widely adapted for these tasks. To manage complex deformations, methods like RCN [3] and DualPRNet++ [4] integrated cascaded and parallel architectures into U-Net to enhance feature representation learning. Further improvements were made by introducing anatomical constraints such as key point loss [5] and segmentation masks [6] to maintain the deformation boundaries between organs.

However, the limited local receptive field of the convolution operator restricts capturing long-range relationships, leading to the integration of attention-based Transformers into encoder-decoder networks, like TransMorph [7] and XMorpher [8], to capture local and global anatomical relationships. Despite these advancements, most MIR frameworks still struggle to fully utilize the diverse feature representations in medical images, prompting the exploration of multimodal models that combine both visual and textual information.

Recently, Vision-Language Models (VLMs), such as CLIP [9], have shown promise in computer vision tasks by jointly learning from images and text. In the medical domain, models like GLORIA [10], and PMC-CLIP [11] have demonstrated the potential of integrating medical reports to enhance the image analysis tasks. Models like LAVT [12] and LViT [13] have incorporated text annotations to compensate for image quality deficiencies, achieving superior segmentation results. However, VLMs' application in MIR is still limited. TextSCF [14] is one of the few methods that leverage text embeddings, improving brain MRI and abdominal CT registration. However, this model applies text embeddings only at the final decoding output, limiting the textual information's influence on the encoding and decoding of the backbone models. This leaves room for further exploration of deeper multimodal fusion strategies.

To address these limitations, we propose TA-MIR to integrate text prompts into an encoder-decoder structure at multiple scales for MIR. Our approach combines the strengths of text and image modalities, with text prompts providing semantic context while utilizing U-Net to capture spatial details. We evaluated our model on three benchmark brain MRI datasets: OASIS [15], IXI [16], and LPBA40 [17]. Our model achieved state-of-the-art results, outperforming traditional image-only registration methods. By incorporating text prompts, it enhances the registration process with added semantic and contextual relevance, making this one of the first attempts to integrate textual information at the multiscale of the encoder-decoder framework for MIR.

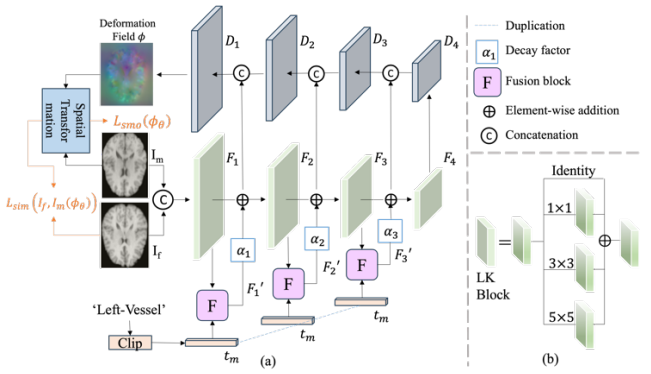

Fig. 1. (a) Illustration of the proposed TA-MIR model, which consists of an encoder-decoder structure, here we adopt the U-Net-like structure and multimodal feature aggregation blocks on each level. (b) The visualization of an LK Block [18], following the same designation in LKU-Net.

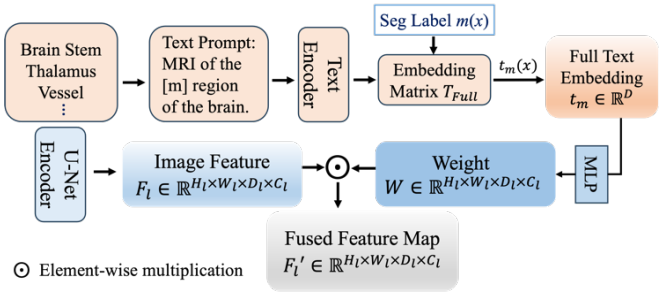

Fig. 2. The proposed fusion block in TA-MIR. Anatomical label $m(x)$ retrieves the text embedding $t_l(x)$ from matrix T.

2. METHOD

The proposed framework TA-MIR is illustrated in Fig. 1, it integrates textual prompts and U-Net architecture for multimodal feature aggregation in MIR. The key innovation lies in leveraging both image and text modalities, enhancing the traditional image encoder-decoder pipeline with semantic guidance derived from text embeddings.

2.1. Overall Framework

The overall architecture of TA-MIR builds on a U-Net backbone, enhanced by the fusion block at each level of the encoder-decoder structure and the propagation of these fused features through skip connections, as shown in Fig. 1.

The encoding path incorporates four large-kernel (LK) blocks, followed by a text-image fusion block and downsampling layers, to extract hierarchical features at different scales. The encoder progressively increases the number of kernels, starting with C kernels in the first layer and doubling them at each subsequent downsampling layer. In the expansion path, the decoder upsamples the features and combines them with the corresponding encoder features through skip connections. In TA-MIR, the skip connections not only pass spatial information but also propagate the text-infused features from the encoder to the decoder.

While the text-image fusion, as Fig. 2 shows, occurs at multiple levels of the U-Net, it allows for the integration of semantic information at three scales to achieve low-level, mid-level, and high-level fusion. Thus, our model could first capture fine-grained anatomical details, then integrate intermediate-scale features, and finally incorporate global contextual information. The fused feature at each level is modulated by an attenuation coefficient α_l:

$$F_{\#}^{*} = \alpha_l * F_{\#}', \quad (1)$$

Here, α_l is a decay factor that controls the influence of the text embedding at each level l. It ensures that the influence of the text embeddings is modulated as the feature maps propagate through the U-Net, allowing the model to balance spatial features with semantic guidance from the text. The value of α_l increases progressively from the shallower encoder levels to the deeper levels, ensuring that the text embedding's contribution is gradually amplified while guiding the entire registration process.

2.2. Backbone for Image Feature Extraction

In the main experiments of this paper, we chose the LKU- Net [18] backbone for its ability to effectively capture both detailed features and large-scale deformations. The LK encoder in our U-Net model is designed to enhance the effective receptive field by combining multiple convolutional operations in parallel while maintaining parameter efficiency. Each LK block in the encoder applies four parallel operations to the input feature map x and aggregates the outputs elementwise. Shortly, for a given x, the overall output of each LK encoder block is:

$$x_{\%\&'(\&'} = \sigma(Conv3d_{)\times)\times)}(x) + Conv3d_{+\times+\times+}(x) + Conv3d_{,\times,\times,}(x) + x), \quad (2)$$

where k represents the LK size, chosen as 5 in this paper, and $\sigma(\cdot)$ is an activation function such as ReLU.

This formulation ensures that the network captures a wide range of spatial information without dramatically increasing the number of parameters, preserving the model's training stability while enhancing its capacity to process both small-scale details and large anatomical deformations.

2.3. Textual Feature Extraction

To integrate semantic information into the registration process, we employ CLIP [9], which is pre-trained on large-scale image-text pairs, to extract text-based embeddings of regions using its original feature without fine-tuning (with a higher accuracy than One-Hot). For each anatomical region m, a descriptive prompt is generated, such as: "Magnetic Resonance Imaging (MRI) of the [m] region of the brain." CLIP's pre-trained text encoder is used to convert these prompts into corresponding text embedding vectors $t_l \in \mathbb{R}^{-}$, where D represents the embedding dimension. During this procedure, a transformer-based encoder is used to embed the textual input. The tokenized text is first converted

into a sequence of word embeddings, then passed through a series of transformer layers to extract the textual feature representation through multi-head self-attention. Finally, for a total of M anatomical regions, the embeddings are organized into a matrix $T \in \mathbb{R}^{M \times d}$, where each row is an embedding vector for a particular region:

$$T = [t_1, t_2, \ldots, t_M], \tag{3}$$

Meanwhile, the background embedding t_1 is initialized to distinguish it from the anatomical embeddings, ensuring that the network correctly differentiates background from the anatomical regions. Thus, the full embedding matrix is:

$$T_{text} = [t_1, t_2, \ldots, t_M], \tag{4}$$

2.4. Multimodal Feature Aggregation

A key aspect of TA-MIR is the fusion of text-based anatomical embeddings and image features, which occurs at multiple stages in the network. The fusion block is illustrated in Fig. 2, specifically, for each voxel x, the segmentation label $m(x)$ is used to retrieve corresponding text embedding $\boldsymbol{t}_l(x) \in \mathbb{R}^d$ from T_{text} by indexing:

$$\boldsymbol{t}_l(x) = T_{text}[m(x)] \tag{5}$$

To combine it with the image features, we gather $t_l(x)$ to construct the full embedding image $\boldsymbol{t}_l \in \mathbb{R}^{H_l \times W_l \times D_l \times d}$, and implement a feature modulation mechanism. Let $F_l \in \mathbb{R}^{H_l \times W_l \times D_l \times C_l}$ represent the image feature map at level l of the U-Net. We modulate these image features by applying the text embedding $\boldsymbol{t}_l$ to enrich the spatial information with contextual knowledge:

$$F_l^{\prime} = F_l \odot MLP(t_l), \tag{6}$$

where $\odot$ denotes element-wise multiplication, F_l' is the fused feature map, and $MLP(t_l) \in \mathbb{R}^{H_l \times W_l \times D_l \times C_l}$ is the weight generated by a multi-layer perceptron projecting $\boldsymbol{t}_l$ into the same dimensional space as the image feature map.

This fusion block is applied at multiple levels of the encoder-decoder structure, allowing the text embedding to influence low-level and high-level feature representations. The fused feature maps are used as inputs for subsequent layers of the U-Net, ensuring that image features and semantic information from the text jointly guide the process.

2.5. Loss Function

The task of medical image registration is to find the optimal transformation that aligns the moving image $\boldsymbol{I}_l$ with the fixed image I_f. With the final fused multimodal feature representation D_8 at the final layer (when L=1) of the decoder, we define the final deformation field $\phi_9(x)$ as:

$$\phi_9(x) = x + u_9(x), \tag{7}$$

where $u_9(x)$ represents the displacement field learned from the multimodal features throughout the model at location x, and θ are the parameters learned by the network.

Moreover, the loss function is composed of three terms:

$$L_{total} = L_{sim}(I_f, I_l(\phi_9)) + L_{dice}(J_f, J_l(\phi_9)) + \lambda L_{smooth}(\nabla\phi_9), \tag{8}$$

where L_{sim} is typically implemented as mean squared error (MSE) or normalized cross-correlation (NCC) depending on the data used. L_{dice} evaluates the similarity based on the deformed segmentation of the registration results $\boldsymbol{J}_l(\boldsymbol{\phi}_9)$ and the ground truth J_f using a Dice loss. $\boldsymbol{L}_{smooth}$, where the L2 norm is calculated, penalizes large gradients to enforce smoothness in the deformation field, while λ serves as the scale of smoothness modulation.

3. EXPERIMENTS AND RESULTS

3.1. Datasets

Experiments were carried out using three publicly available 3D brain MRI datasets. For all those datasets, the label and responding anatomical region names were paired and served as the text prompt of our model. Automatic segmentation processed by FreeSurfer [19] was utilized.

The OASIS dataset [15] contains 414 pre-processed 3D inter-subject brain scans with 35 structures and a resolution of 160 × 192 × 224. There were 394 scans for training, along with 19 image pairs for validation and testing.

The IXI dataset [16] includes 576 T1-weighted brain MRI scans, and label maps of 38 anatomical structures were used for Dice evaluation. All scans were cropped to 160 × 192 × 224. The training, validation, and testing set was separated to have 403, 58, and 115 images accordingly.

The LPBA40 dataset [17] consists of 40 brain MRI scans, with 56 labeled ROIs. These scans were resampled to a size of 160 × 192 × 160. A total of 30 scans were used for training, and 10 scans were reserved for testing.

3.2. Implementation Details

The model was implemented using PyTorch on a machine with four NVIDIA Quadro GV100 GPUs. It was trained for 500 epochs using the Adam optimization algorithm, with a batch size of 1, and a learning rate of 1×10^{-4} throughout the training process. The smooth factor λ and start channel C were subject to change according to the dataset used. Empirically, we used 0.2/0.4/0.6 as decay factor $\alpha_{///+}$.

3.3. Results

To demonstrate the effectiveness of the proposed TA-MIR, we compared it with multiple cutting-edge classical iterative methods, CNN-based models, Transformer-based models, and hybrid methods. Most of the quantitative results of those methods were obtained from open-sourced publications or leaderboards, while others were trained by us using the optimal parameter settings. The performance was evaluated using the Dice Similarity Coefficient (DSC) and the Jacobian determinant of the deformation field ($\%|J_T| < 0$).

Table 1. Quantitative results of different methods on OASIS, LPBA40, and IXI dataset.

Method	OASIS		LPBA40		IXI	
	Dice(%)	$\%\lvert J_{>}\rvert \leq 0$	Dice(%)	$\%\lvert J_{>}\rvert \leq 0$	Dice(%)	$\%\lvert J_{>}\rvert \leq 0$
SyN [22]	78.0	0±0.12	66.5	0±0.12	63.9	0±0.20
NiftyReg [23]	78.5	0.10±0.20	66.9	0.14±0.09	64.0	0±0.18
CycleMorph [24]	78.8	0.85±0.38	65.0	0.44±0.22	73.0	1.72±0.38
VoxelMorph [2]	84.7	1.24±0.46	64.2	0.96±0.38	72.6	1.52±0.34
TransMorph [7]	86.2	1.56±0.33	63.7	1.42±0.46	74.6	1.57±0.33
AttentionReg [25]	77.5	1.44±0.50	62.7	0.81±0.34	/	/
LapIRN [26]	76.5	0.01±0.32	**73.6**	0.01±0.30	/	/
LKU-Net [18]	88.6	0.11±0.05	68.7	0.13±0.26	75.7	0.14±0.12
TextSCF [14]	90.1	0.12±0.04	/	/	/	/
TA-MIR (Ours)	**91.4**	0.11±0.04	72.5	0.13±0.33	**78.6**	0.13±0.18

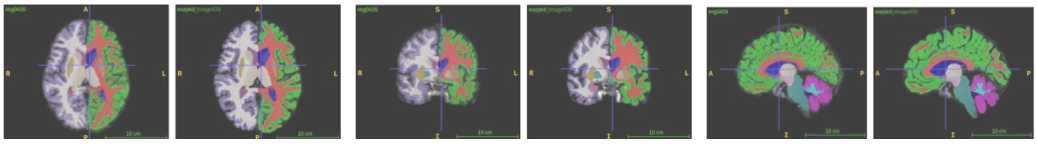

Fig. 3. One of the qualitative results on OASIS dataset in blend display, in axial, coronal, and sagittal dimension.

Table 2. Results of DMR and DMR-edited on two datasets.

Model	OASIS		LPBA40	
	Dice(%)	$\%\lvert J_{!}\rvert \leq 0$	Dice(%)	$\%\lvert J_{!}\rvert \leq 0$
Vit-V-Net [20]	78.2	2.05±0.90	61.3	1.31±0.48
DMR[21]	79.3	1.02±0.44	67.5	0.62±0.33
TA-DMR	80.8	0.98±0.43	68.9	0.56±0.33

Table 3. Results of encoder/decoder-only-fusion TA-MIR.

Model	Dice(%)	$\%\lvert \boldsymbol{J_{\phi}}\rvert \leq \mathbf{0}$
TA-MIR (Enc)	89.7	0.17±0.26
TA-MIR (Dec)	89.2	0.14±0.06

Table 1 shows the comparison of various methods across the three datasets. Our proposed TA-MIR method achieved state-of-the-art performance on the OASIS and IXI datasets and ranked second on the LPBA40 dataset, Fig. 3 shows an example. The model also generated a relatively smooth deformation field, indicated by a low percentage of voxels with non-positive Jacobian determinants.

When compared to LapIRN on the LPBA40 dataset, our model's Dice score was 1.1% lower, likely due to LapIRN's use of a similarity pyramid for multi-resolution optimization, which enhanced performance on limited data. However, compared to the LKU-Net, our model showed a 3% to 4% accuracy improvement across all datasets, highlighting the effectiveness of our multimodal feature aggregation block.

Furthermore, TA-MIR outperformed TextSCF on the OASIS dataset, emphasizing the importance of multiscale text-image fusion integrated throughout the model, rather than applied only at the final output stage. We also applied our fusion block to another U-Net-like model, Deformer (DMR) [21], named as TA-DMR. As shown in Table 2, while TA-DMR did not achieve state-of-the-art performance, it still demonstrated higher accuracy across datasets, validating the effectiveness of our multiscale fusion method.

3.4. Ablation Study

To assess the impact of image-text fusion in both the encoder and decoder of our model, we conducted an ablation study on the OASIS dataset by applying fusion only in the encoding progress or only during decoding.

In the encoder-only fusion, the text embeddings were fused during feature extraction, while the decoder used traditional skip connections. In the decoder-only fusion, the text embeddings were integrated at each decoder level, with no fusion in the encoder. Refer to Table 3, both experiments showed a decline in performance compared to the full model, with Dice scores dropping to 89.7 and 89.2, respectively. Despite this, both experiments outperformed the LKU-Net's 88.6. However, they performed worse than TextSCF, highlighting the importance of multiscale fusion across entire network.

4. CONCLUSION

In this paper, we introduced TA-MIR, a framework that integrates anatomical text embeddings with an encoder-decoder architecture for MIR. By combining spatial image features and semantic information, our model improved the registration accuracy and the robustness cross datasets. Experiments on the OASIS, IXI, and LPBA40 datasets showed state-of-the-art performance, highlighting the benefits of multimodal feature aggregation. Future work will expand the model to other data types, like abdomen CT or even cross-modal MIR, explore more sophisticated text prompts with additional anatomic details, and integrate more encoder-decoder architectures like hybrid networks, to further improve the usability and alignment precision.

Compliance with Ethical standards
This research study was conducted retrospectively using medical image data made available in open access by [15], [16], and [17]. Ethical approval was not required as confirmed by the license attached with the open-access data.

Conflicts of Interest
The authors have no conflicts to disclose.

Acknowledgements
The authors have no relevant financial or non-financial interests to disclose.
* Contact author e-mail: wmhuang@i2r.a-star.edu.sg

5. REFERENCES

[1] G. Haskins, U. Kruger, and P. Yan, "Deep Learning in Medical Image Registration: A Survey," *Machine Vision and Applications* 31, 8, 2020.

[2] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A Learning Framework for Deformable Medical Image Registration," in *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788-1800, Aug. 2019.

[3] S. Zhao, Y. Dong, E. I.-C. Chang, and Y. Xu, "Recursive Cascaded Networks for Unsupervised Medical Image Registration," *2019 IEEE Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 10599-10609.

[4] M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes, "Dual-Stream Pyramid Registration Network," Medical *Image Analysis*, *78*, 102379, May. 2022.

[5] A. Hering, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, and B. Van Ginneken. "CNN-based lung CT registration with multiple anatomical constraints," *Medical Image Analysis*, vol. 72, p. 102139, Aug. 2021.

[6] X. Chen, N. Ravikumar, Y. Xia, and A. F. Frangi, "A Deep Discontinuity-Preserving Image Registration Network," In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.

[7] 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*, *82*, 102615, Nov. 2022.

[8] J. Shi, Y. He, Y. Kong, J. Coatrieux, H. Shu, G. Yang, and S. Li. "X-Morpher: Full Transformer for Deformable Medical Image Registration via Cross Attention." In Medical Image Computing and Computer Assisted Intervention – MICCAI 2022.

[9] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. "Learning Transferable Visual Models from Natural Language Supervision," *International Conference on Machine Learning*, Feb. 2021.

[10] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931, Oct. 2021.

[11] W. Lin, Z. Zhao, X. Zhang, Y. Wang, and W. Xie. "PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents," In Medical Image Computing and Computer Assisted Intervention – MICCAI 2023.

[12] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, "LAVT: Language-Aware Vision Transformer for Referring Image Segmentation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 18134-18144.

[13] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, and L. Lu, "LViT: Language meets Vision Transformer in Medical Image Segmentation," in *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 96-107, Jan. 2024.

[14] X. Chen, M. Liu, R. Wang, R. Hu, D. Liu, and G. Li., "Spatially Covariant Image Registration with Text Prompts," in *IEEE Transactions on Neural Networks and Learning Systems*. 2024.

[15] D. Marcus, T. Wang, J. Parker, J. Morris, and R. 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*, vol. 19, no. 9, pp. 1498– 1507, 2007.

[16] IXI Brain Development Dataset. https://brain-development.org/ixi-dataset/

[17] D. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. Narr, R. Bilder, and A. Toga, "Construction of a 3D probabilistic atlas of human cortical structures," NeuroImage, vol. 39, no. 3, pp. 1064–1080, 2008.

[18] X. Jia, J. Bartlett, T. Zhang, W. Lu, Z. Qiu, and J. Duan, "U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration?" In Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022.

[19] B. Fischl, "Freesurfer," Neuroimage, vol. 62, no. 2, pp. 774–781, 2012.

[20] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration," Apr. 13, 2021, *arXiv*: 2104.06468.

[21] J. Chen, D. Lu, Y. Zhang, D. Wei, M. Ning, X. Shi, Z. Xu, Y. Zheng, "Deformer: Towards Displacement Field Learning for Unsupervised Medical Image Registration," In Medical Image Computing and Computer Assisted Intervention 2022.

[22] 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, Volume 12, Issue 1, 2008, Pages 26-41.

[23] Modat, M., Ridgway. R., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98(3), 278–284, 2010.

[24] Kim, Boah and Kim, Dong Hwan, Kim, Jieun, Lee, June- Goo and Ye, Jong Chul: CycleMorph: Cycle consistent unsupervised de- formable image registration. Medical Image Analysis 71, 102036, 2021.

[25] Song, X., Guo, H., Xu, X., Chao, H., Xu, S., Turkbey, B., Wood, B.J., Wang, G., Yan, P.: Cross-modal attention for MRI and ultrasound volume registration. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 66–75. 2021.

[26] T. C. W. Mok and A. C. S. Chung, "Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks," In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020.