ITR-Net: A Hybrid Deep Learning Architecture for Precise and Efficient Medical Image Registration

Yan Yan[1#], Liyilei Su[1#], Chengmin Zhou[1#], Yongzhi Huang[1], Jing Li[2], Rui Li[1], Haseeb Hassan[1*], BingdingHuang[1*]

1. College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China.
2. Department of Pulmonary and Critical Care Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, Guangdong, China.

# These authors contributed to this work equally

yanyan1476578701@163.com, suliyilei@sztu.edu.cn, chuengminchou@gmail.com, hhuang0823@163.com, dr.lijing@gdph.org.cn, lirui@sztu.edu.cn

Corresponding Author: Haseeb Hassan Email: haseeb@sztu.edu.cn; Bingding Huang Email: huangbingding@sztu.edu.cn

2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 979-8-3503-3366-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITAIC58329.2023.10409014

***Abstract*—This research proposes a weakly supervised-based learning registration network called Improved Transformer Registration Net (ITR-Net) to improve medical image registration accuracy. Firstly, we improved the transformer module by incorporating patch embedding and a feed-forward layer. These enhancements enable the transformer module to focus on local features in close proximity while establishing connections between distant voxels. Secondly, we embedded this module within U-Net, utilizing a CNN structure to extract image features more precisely and enhance matching accuracy. Then we evaluated the performance of our approach using three-phase CT imaging data of kidneys and lungs. The results demonstrate that our method surpasses traditional and pure CNN-based registration algorithms in terms of both registration accuracy and efficiency.**

Keywords—Medical image registration; Deep neural network; Attention mechanism; Transformer-based registration; Weakly supervised learning

I. INTRODUCTION

Image registration is a process to uncover the hidden relationships between the input and reference images, typically represented by a coordinate transformation matrix [1] [2]. Image registration is significant in numerous practical applications, such as Computer Vision, Geographic Information Systems (GIS), and medical imaging [3]. Apart from the other applications, one is to align medical images, such as MRI and CT scans, to improve the images' quality and extract useful information [4]. Further applications involve diagnosis [5], studying disease progression and monitoring [6], image-guided surgical interventions [7], or atlas construction [8]. Image registration can be viewed as a fundamental optimization challenge [9]. Image variability, noise and artifacts, illumination differences, limited overlapping regions, non-linear deformations, computational complexity, and lack of ground truth pose challenges in image registration [10]. Overcoming these requires advanced techniques, including feature-based, intensity-based, and deformable registration.

Conventional algorithms primarily rely on feature-based point matching and often encounter challenges due to human initialization, making it difficult to achieve optimal registration results [11]. Moreover, the iterative nature of these algorithms significantly extends the registration time, rendering them unsuitable for real-time operations [12]. Convolutional neural networks (CNNs) have shown promise by autonomously predicting spatial transformation parameters, leading to substantial improvements [13]. Despite the inherent limitations of CNNs, effectively capturing long-distance features, particularly in the case of lung and kidney CT images with larger displacements, poses a significant challenge. Consequently, this results in sub-optimal registration outcomes. With the introduction of the attention mechanism [14], it becomes possible to address the challenges posed by large displacements. Attention-based Transformer learning has witnessed a rapid expansion and can acquire a deep understanding of the internal relationships within data and effectively maintain connections between distantly located voxels. Therefore, this research introduces a novel registration network architecture incorporating a transformer module, showcasing superior performance compared to a standalone neural network. The proposed network is assessed using three phase lung and kidney CT datasets to validate its efficacy. The main contributions of this work are as follows:

We conduct a comprehensive investigation and analysis of the performance of both traditional registration

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.

and deep learning algorithms on 3D kidney and lung CT data. The proposed network utilizes the self-attention mechanism to establish efficient connections with voxels that are initially distant, enabling improved registration of lungs and kidneys even when large displacements are present. The proposed architecture is compared to traditional and deep learning registration algorithms, revealing that the transformer-based registration methods emphasizes global and local features. Likewise, it demonstrates superior registration performance on lung and kidney datasets.

II. Related Works

A. Traditional Algorithms

Most medical image registration techniques rely on image feature points and employ a non-learning approach. These algorithms continuously optimize through iterations to determine the optimal spatial transformation. One such algorithm is the Scale Invariant Feature Transform (SIFT) [15]. However, it has limitations in extracting image features with smooth edges. To address this, Bay et al. proposed a targeted improvement known as the Speeded Up Robust Features (SURF) [16]. Mutual Information [17] approach involves computing the mutual information between two medical images. To simplify the registration process, tools such as ANTs [18] and SimpleElastix [19] have been developed specifically for automated image registration [20]. Generally, these algorithms iteratively update parameters to find the optimal transformation. However, the time-consuming optimization process makes this approach impractical for clinical uses.

B. Supervised Learning

Supervised learning algorithms are important methods in medical image registration. Before training the neural network, such algorithms requires the actual deformation field of the image, which is often obtained by simulating the image deformation. For example, Salhei et al. utilized random transformation to convert rotation, translation, and scaling parameters into an affine transformation matrix [21]. Sokooti et al. introduced RegNet [22], which combines features from upper and lower network layers for end-to-end registration. To address irregular deformation of breathing lungs, Eppenhof et al. repeatedly superimposed small and large deformations in the grid to enhance model generalization [23]. Sloan et al. utilized CNN for weighted single-modal and multi-modal brain registration [24]. In conclusion, supervised learning methods demonstrate strong performance, however, they require extensive ground truth comparisons.

C. Unsupervised Learning

De Vos introduced the unsupervised registration network DIRNet [25] for deformable organ tissue registration and achieved better results than supervised learning on public datasets. Balakrishnan proposed a deformable framework [26] that optimizes the registration function continuously. Jiang et al. developed a multi-scale network for supervision with deformation field smoothing and image similarity losses [27]. Zhao et al. designed a shared-weight recursive cascaded network Structure, delivering promising results on brain and liver images [28]. Kori et al. utilized a pre-trained registration network for multi-modal registration of 2D T1 and T2 weighted brain MR images [29]. Mahapatra et al. employed generative adversarial learning to ensure smooth deformation fields and accurate registration [30]. Song et al. introduced the MSReg framework, a coarse-to-fine multi-level registration network, improving registration accuracy across different image data precision [31].

D. Weakly-supervised and Transformer-based

Weakly supervised methods aim to reduce reliance on labels. For example, Fan et al. applied a hierarchical approach to predict the deformation field for the brain registration network [32]. Hu et al. introduced a novel approach that involves designing a multi-scale Dice similarity measure to calculate similarity [33]. The image space transformation introduced the attention mechanism to image registration [34]. To enhance the ability of neural networks, Hu et al. proposed the SeNet network [35]. Likewise, recently the Vision Transformer (ViT) has shown that pure transformer architectures can perform better [36]. In addition to pure ViT methods, Zhang et al. combined a CNN with a transformer for deformable registration [37]. However, the CNN architecture still has inherent limitations in considering global and local feature changes. To address this, we propose improved CT image registration technique by incorporating a transformer module into the existing CNN framework.

III. Proposed Model and Methodology

Let F and M denote the fixed image and moving image. In image registration, the aim is to find a coordinate transformation, i.e., $T: F \rightarrow M$ which aligns a fixed image F and a moving image M. In conventional image registration, the similarity between these images is optimized by minimizing a dissimilarity metric **L**:

$$\widehat{\mu} = argmin\left\{L(Tu, F, M) + R(Tu)\right\} \quad (1)$$

where Tu is the spatial transformation parameter of μ and R is an optional regularization term to promote smoothness of the transformation Tu. In our approach, we use CNN to model the given registration function. In the subsequent sections, we provide details regarding our proposed improvements to the transformer module and the design of the overall registration network.

A. Transformer-improved Structure

In a deep-learning framework, the registration network aims to predict transformation parameters using inputs F and M given as:

$$\mu = f_{\theta}(F, M) \quad (2)$$

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.

where f represents the structure of our proposed network, θ represents the network's parameters. The registration based on the neural network can automatically estimate the spatial transformation of the registration image and minimize the image difference between the image pair to be registered through the loss function. The Transformer model primarily focused on processing the positional information of sentences. However, the Vision Transformer (ViT) extends the application of Transformer to image-related tasks. ViT adopts a simple approach by dividing the image into equal blocks, flattening them, and then directly mapping the original image pixels to primary features using a fully connected layer. These primary features are then input into the Transformer, where each small block is treated as a sequential token input, similar to processing text.

Although ViT model performs well in capturing long-range voxel relationships through self-attention but lacks local information incorporation and effective modeling of adjacent patches. To address this, we introduce patch embedding and feed-forward layers to the transformer module as shown in Fig 1, enhancing locality and enabling better capture of local spatial characteristics. We replace the original linear patch embedding layer with a convolutional patch embedding layer. The convolutional patch embedding layer enables the transformer module to model the local space of the registration image pair and focus on local features. On the other hand, since the feed-forward layer of ViT is composed of MLP, which is mapped on the patch embedding in a patch-wise manner, and cannot establish an effective connection between adjacent patch embeddings. Therefore, we add two 3x3x3 convolutional layers between the hidden layers.

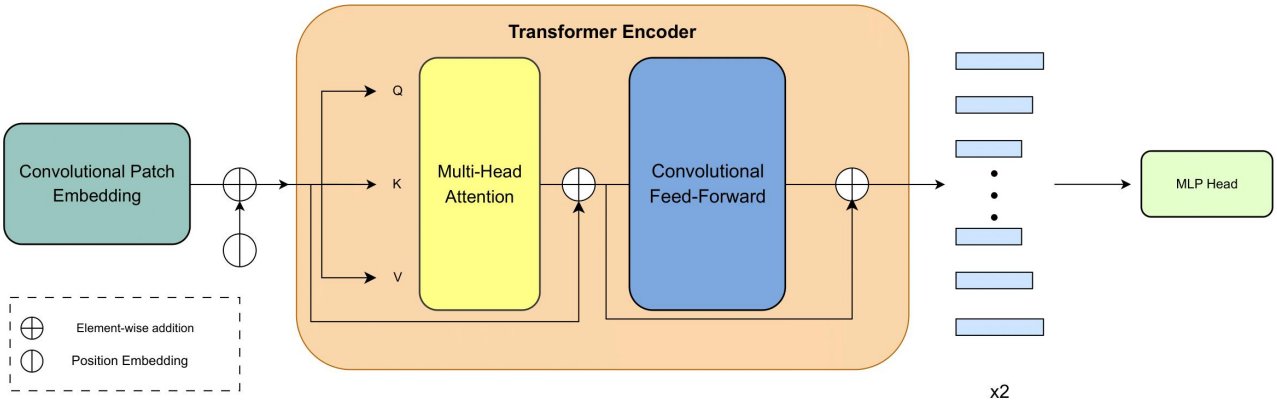

Fig. 1. Improved-transformer structure. In order to be able to take advantage of the transformer module in building long-distance voxels while focusing on adjacent local features, we add patch embedding and feed-forward layers to the transformer module. In addition, two 3x3x3 convolutional layers are also added.

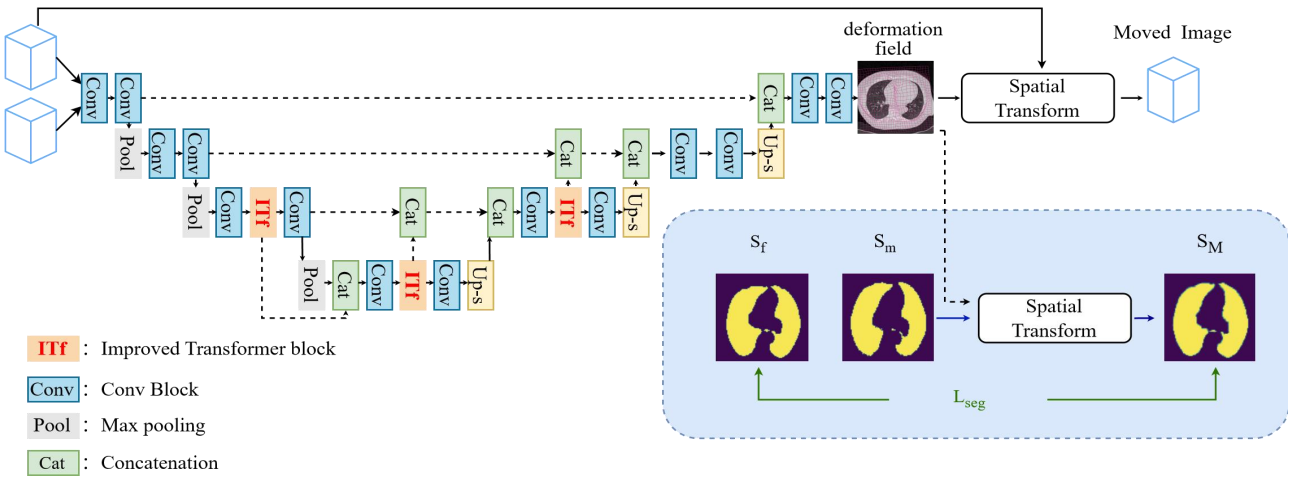

Fig. 2. The proposed registration framework is based on the Improved Transformer module. We use the improved U-Net network to learn the registration process of medical images. In the training process, we use a spatial converter to distort the registration image and obtain the final registration image. The difference is that we make use of the image labeled data at the end to evaluate the registration effect more directly by calculating the similarity between labels.

B. Improved-transformer Registration Network

We leverage the proven effectiveness of CNNs in extracting local image features and the widespread adoption of the U-Net network. We enhanced U-Net as the backbone network by incorporating an improved transformer module into the encoder-decoder framework. The network takes a pair of images to be registered as input. Within the encoder, each convolution block is

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.

followed by a ReLU activation function, maintaining a 3x3x3 kernel size and stride consistent with U-Net. Connection layers are introduced in selected encoder layers to link the improved transformer module, allowing the fusion of generated features. Concatenated layers are used in the encoder to merge features from the transformer module in the previous layer with those from the pooling layer.

The decoder combines the Conv layer, Transformer blocks, and Up-sampling layer to preserve spatial resolution across various feature map levels. At the network's end, the connected features undergo two convolutions layers to derive the final output features, representing the deformation field parameters. The deformation field generated by the network is applied to the segmented label corresponding to the moving image, resulting in the transformed label. The segmentation loss is then computed by comparing it with the fixed image label. Based on experimental evidence, we choose to incorporate the transformer module exclusively in the deep network. Adding the transformer module to shallow layers proves ineffective for capturing low-level features and does not significantly enhance registration accuracy.

C. Spatial Transformation Function

Our proposed approach aims to determine the optimal parameters by minimizing the disparity between the registered image pairs. To achieve this, we employ a space transformer network to calculate the transformed image through a differentiable operation. We specifically utilize the bi-linear interpolation definition within the 3D transformation function given as follows:

$$M \circ \phi(p) = \sum_{q \in N(\phi(p))} M(q) \prod_{d \in \{x,y,z\}} \left(1 - \left|\phi(p) - q_d\right|\right) \quad (3)$$

where p is a voxel, $M \circ \phi$ denotes the registered image generated by the moving image M through the action of the deformation field ϕ, $N(\phi(p))$ represents the neighbor of $\phi(p)$. With this function, we can back-propagate the error during optimization by computing the gradient.

D. Loss Function

We incorporate Dice loss as the loss function to regulate the registration outcomes. Additionally, the network can utilize the segmentation map as supplementary information during training. The anatomical map mask, which corresponds to the image, assigns each voxel to a specific anatomical structure. If the registered region exhibits precise anatomical correlation, with good overlapping of areas, it indicates a favorable registration effect. The Dice loss is as follows:

$$DiceLoss = 1 - \frac{2\left|L_F \cap L_M\right|}{\left|L_F\right| + \left|L_M\right|} \quad (4)$$

where L_F represents the mask label data associated with the fixed image, while L_M represents the mask label data associated with the moving image. The image registration quality is assessed by quantifying the Dice value, where a Dice score of 1 indicates a perfect anatomical match, while a score of 0 signifies no overlap between the images.

IV. EXPERIMENTS AND RESULTS

This section details the datasets, the experimental method of our proposed registration algorithm, and their comparison with other registration algorithms.

A. Dataset

We choose three-phase CT data (unenhanced CT, excretion phase CT, and cortical phase CT) of the lungs and kidneys. The unenhanced phase CT images were utilized as the standard reference, while the enhanced CT images are regarded as cortical and excretory phases. The abdominal CT data of all three phases were sourced from the Haiyin Branch of the First Affiliated Hospital of Guangzhou Medical University, resulting in 12,012 original data samples encompassing both lungs and kidneys. After discarding poor-quality images, the final data distribution across the different phases was as follows: 3,822 are unenhanced CT, 2,401 are excretion phase CT, and 2,288 are cortical phase CT. A final selection of 600 CT images (both kidneys and lungs) was made from all three phase CT images. Each CT image underwent meticulous annotation by experienced professionals, encompassing comprehensive lung and kidney contour information.

B. Experimental Details

- Data preprocessing: To optimize the initial CT image storage, preprocessing is conducted. Initially, the images are resampled to ensure consistent pixel dimensions, with all CT images being resampled to 1mm×1mm×1mm. Next, the centroid formula is applied to identify and locate the centroid positions of the lungs and kidneys. Using these centroids as reference points, the images are cropped to a standardized size of 256×256×128, comprising the entire target organ. Lastly, a uniform grayscale treatment is applied to ensure consistent image quality. These procedures unify the input registration network images by standardizing their size and eliminating image noise.
- Network training: The three-phase CT images corresponding to the kidneys and lungs were registered. In this process, the unenhanced CT images were fixed, while the moving images consisted of the excretory and cortical CT scans. The selected 600 three-phase CT images were split into training and validation sets. A total of 500 images were reserved for training, and the remaining 100 images were set aside for validation purposes. The proposed registration network is initialized with a learning rate of 0.001. Adam

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.

optimizer is used as an optimization algorithm. The registration model is saved upon reaching the best performance on the validation set. Using our trained model, we extract features from the fixed and moving images and their corresponding labeled data. We also recorded the parameters of spatial transformation and then applied the deformation field to the moving image, resulting in the registered image. We obtain a quantitative measure of the registration accuracy by calculating the Dice value between the registered image and the fixed image mask.

- Baseline methods: To effectively showcase the capabilities of our proposed deep learning registration algorithm, we compared it with various registration methods as showed in Table 1. These include two publicly available registration software packages: SimpleElastix and ANTs, as well as three traditional machine learning registration algorithms: SIFT, SURF, and MI. Additionally, we also evaluated a pure U-Net-based registration approach. SimpleElastix and ANTs are versatile registration frameworks with optimized parameters for lung and kidney datasets. SIFT and SURF are feature-based registration algorithms, with SURF outperforming others. MI utilizes grayscale values for registration. The surf algorithm has the highest registration accuracy among the traditional registration algorithms. Among deep learning-based algorithms, our proposed ITR-Net achieves the best results. Among them, ME represents the excretion phase and MC represents the cortical phase, which are used as moving images and registered with plain scan CT as fixed images respectively. The content in parentheses shows the standard deviation of the experiment.

C. Results and Discussion

The registration results for various methods using our dataset are presented in Table 1. Among these methods, SURF demonstrated the most favorable outcomes as compared to other traditional algorithms. The traditional algorithms are constrained by human factors, such as feature points selection, which limits the registration accuracy. The Dice value after their registration can only reach 0.86, which is hardly accurate. On the other hand, the registered outcomes are greatly enhanced with the incorporation of deep learning-based algorithms. Likewise, comparing our proposed ITR-Net with other deep learning models (VM, U-Net [38], and C2FVit [39]) generated better-registered outcomes. Thus, including the proposed Transformer block showed the validity and superiority of our proposed registration model.

Ablation study: The results of the ablation experiments, as depicted in Table 2, demonstrate that the registration performance of ITR-Net is enhanced through utilizing an improved transformer module and implementing weakly supervised learning techniques. For instance, it was observed that the registration based solely on U-Net was not as effective compared to the CNN structure integrated with our proposed transformer module. However, by adopting a data-driven approach and utilizing labeled data similarity for controlling the registration process, we were able to achieve the highest accuracy.

Registration Visualization results: Figure 3 visually shows several registration results. Taking the lung image as an example, the first four images are moving image, fixed image and corresponding mask image respectively. Notably, in the last row, the registration method based on ITR-Net exhibits remarkable performance, effectively aligning two images with substantial dissimilarities. This observation highlights the ability of ITR-Net to achieve favorable registration results, even in challenging scenarios where image differences are pronounced. Each row of images corresponds to different registration methods, namely: traditional method, U-Net, C2FVit, VM and the method proposed in this paper, ITR-Net. Among them, the results of traditional registration methods show the registration effect of SURF algorithm.

TABLE I. STANDARD DEVIATION PERFORMANCE EVALUATION OF PROPOSED REGISTERATION MODEL WITH TRADITIONAL REGISTERATION MODELS.

Methods	**Lung DSC**		**Kidney DSC**	
	M_E	M_C	M_E	M_C
Simple Elastix	0.834(0.015)	0.841(0.015)	0.825(0.021)	0.838(0.020)
ANTs	0.821(0.018)	0.833(0.015)	0.828(0.020)	0.834(0.017)
SIFT	0.864(0.015)	0.858(0.015)	0.861(0.018)	0.853(0.015)
SURF	0.865(0.015)	0.852(0.014)	0.865(0.015)	0.862(0.020)
MI	0.853(0.017)	0.852(0.015)	0.861(0.015)	0.843(0.020)
VM	0.895(0.015)	0.903(0.017)	0.901(0.020)	0.889(0.015)
U-Net	0.894(0.018)	0.901(0.020)	0.897(0.015)	0.902(0.016)
C2FVit	0.910(0.015)	0.922(0.015)	0.903(0.017)	0.922(0.014)
ITR-Net	**0.924(0.012)**	**0.933(0.015)**	**0.920(0.020)**	**0.929(0.018)**

TABLE II. STANDARD DEVIATION PERFORMANCE EVALUATION OF PROPOSED REGISTRATION MODEL WITH TRADITIONAL REGISTRATION MODELS FROM ABLATION EXPERIMENTS.

Methods	**Lung DSC**		**Kidney DSC**	
	Ours	***Mc***	***Ours***	***Mc***
U-Net	0.894(0.018)	0.901(0.020)	0.897(0.015)	0.902(0.016)
ITR-Net (Unsup)	0.901(0.014)	0.909(0.015)	0.910(0.017)	0.903(0.014)
ITR-Net (Weak supervision)	0.924(0.012)	0.933(0.015)	0.920(0.020)	0.929(0.018)

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.

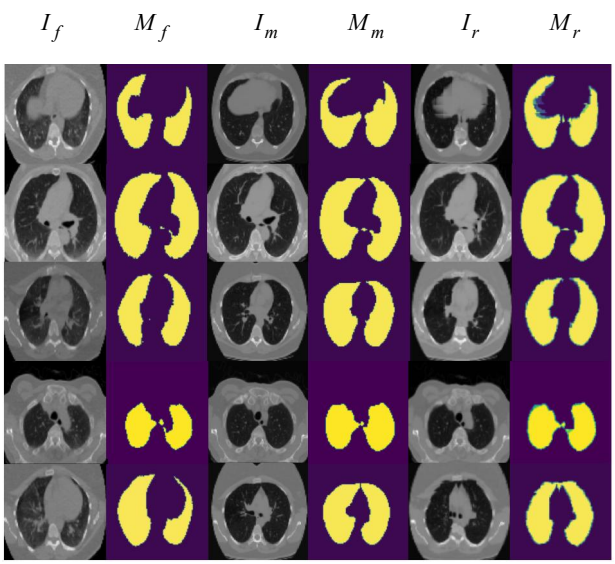

Fig. 3. Comparison of our proposed ITR-Net model with other registration techniques, where I_f represents the fixed image, and M_f represents the mask corresponding to the fixed image, I_m represents the moving image and M_m represents the mask corresponding to the moving image. I_r represents the registration image and M_r represents the mask corresponding to the registration image.

V. Conclusion

In this research, we presented an innovative registration algorithm for CT images of lungs and kidneys, utilizing attention-based mechanisms. Unlike existing approaches that rely solely on CNN-based registration, our method leverages the global connectivity of self-attention to encode the overall orientation. This enables the neural network to focus on long-term correlations of spatial transformations. Experimental results demonstrated the effectiveness of our method in handling image registration with significant displacements while exhibiting faster processing times compared to traditional registration algorithms. However, it is worth noting that transformers, as a relatively new topic in computer vision, are not as widely adopted as CNNs. Future efforts could involve expanding the dataset and incorporating specialized data augmentation techniques to enhance the performance of transformers applied to medical image registration.

Acknowledgment

This work was supported by the Project of the Educational Commission of Guangdong Province of China under Grant No. 2022ZDJS113.

References

[1] L. G. Brown, "A survey of image registration techniques," ACM computing surveys (CSUR), vol. 24, no. 4, pp. 325-376, 1992.

[2] J. A. Maintz and M. A. Viergever, "A survey of medical image registration," Medical image analysis, vol. 2, no. 1, pp. 1-36, 1998.

[3] R. Feng, Q. Du, X. Li, and H. Shen, "Robust registration for remote sensing images by combining and localizing feature-and area-based methods," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 151, pp. 15-26, 2019.

[4] T. Wang et al., "A review on medical imaging synthesis using deep learning and its clinical applications," Journal of applied clinical medical physics, vol. 22, no. 1, pp. 11-36, 2021.

[5] D. Rueckert and J. A. Schnabel, "Medical image registration," in Biomedical image processing: Springer, 2010, pp. 131-154.

[6] V. Gorbunova, P. Lo, H. Ashraf, A. Dirksen, M. Nielsen, and M. de Bruijne, "Weight preserving image registration for monitoring disease progression in lung CT," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008: 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part II 11, 2008: Springer, pp. 863-870.

[7] A. Elmi-Terander et al., "Surgical navigation technology based on augmented reality and integrated 3D intraoperative imaging: a spine cadaveric feasibility and accuracy study," Spine, vol. 41, no. 21, p. E1303, 2016.

[8] M. Sinclair et al., "Atlas-ISTN: joint segmentation, registration and atlas construction with image-and-spatial transformer networks," Medical Image Analysis, vol. 78, p. 102383, 2022.

[9] B. Balluff, R. M. Heeren, and A. M. Race, "An overview of image registration for aligning mass spectrometry imaging with clinically relevant imaging modalities," Journal of Mass Spectrometry and Advances in the Clinical lab, vol. 23, pp. 26-38, 2022.

[10] F. G. Zöllner, A. Šerifović-Trbalić, G. Kabelitz, M. Kociński, A. Materka, and P. Rogelj, "Image registration in dynamic renal MRI—current status and prospects," Magnetic Resonance Materials in Physics, Biology and Medicine, vol. 33, pp. 33-48, 2020.

[11] G. Song, J. Han, Y. Zhao, Z. Wang, and H. Du, "A review on medical image registration as an optimization problem," Current Medical Imaging, vol. 13, no. 3, pp. 274-283, 2017.

[12] E. Saiti and T. Theoharis, "An application independent review of multimodal 3D registration methods," Computers & Graphics, vol. 91, pp. 153-178, 2020.

[13] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Human brain mapping, vol. 38, no. 11, pp. 5391-5420, 2017.

[14] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[15] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," Nucleic acids research, vol. 31, no. 13, pp. 3812-3814, 2003.

[16] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," Computer vision and image understanding, vol. 110, no. 3, pp. 346-359, 2008.

[17] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," Physical review E, vol. 69, no. 6, p. 066138, 2004.

[18] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ANTS)," Insight j, vol. 2, no. 365, pp. 1-35, 2009.

[19] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," IEEE transactions on medical imaging, vol. 29, no. 1, pp. 196-205, 2009.

[20] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," Computer methods in biomechanics and biomedical engineering, vol. 17, no. 2, pp. 73-93, 2014.

[21] S. S. M. Salehi, S. Khan, D. Erdogmus, and A. Gholipour, "Real-time deep pose estimation with geodesic loss for image-to-template rigid registration," IEEE transactions on medical imaging, vol. 38, no. 2, pp. 470-481, 2018.

[22] H. Sokooti et al., "3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations," arXiv preprint arXiv:1908.10235, 2019.

[23] Z. Jin, P. Xue, Y. Zhang, X. Cao, and D. Shen, "Semantic-Aware Registration with Weakly-Supervised Learning," in MICCAI

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.

Workshop on Cancer Prevention through Early Detection, 2022: Springer, pp. 159-168.

[24] J. M. Sloan, K. A. Goatman, and J. P. Siebert, "Learning rigid image registration-utilizing convolutional neural networks for medical image registration," 2018.

[25] B. D. De Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, 2017: Springer, pp. 204-212.

[26] 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.

[27] Z. Jiang, F.-F. Yin, Y. Ge, and L. Ren, "A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration," Physics in Medicine & Biology, vol. 65, no. 1, p. 015011, 2020.

[28] S. Zhao, Y. Dong, E. I. Chang, and Y. Xu, "Recursive cascaded networks for unsupervised medical image registration," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 10600-10610.

[29] A. Kori and G. Krishnamurthi, "Zero shot learning for multi-modal real time image registration," arXiv preprint arXiv:1908.06213, 2019.

[30] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018: IEEE, pp. 1449-1453.

[31] X. Song et al., "Cross-modal attention for MRI and ultrasound volume registration," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24, 2021: Springer, pp. 66-75.

[32] J. Fan, X. Cao, P.-T. Yap, and D. Shen, "BIRNet: Brain image registration using dual-supervised fully convolutional networks," Medical image analysis, vol. 54, pp. 193-206, 2019.

[33] Y. Hu et al., "Label-driven weakly-supervised learning for multimodal deformable image registration," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018: IEEE, pp. 1070-1074.

[34] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," Advances in neural information processing systems, vol. 28, 2015.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[36] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration," arXiv preprint arXiv:2104.06468, 2021.

[37] Y. Zhang, Y. Pei, and H. Zha, "Learning dual transformer network for diffeomorphic registration," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24, 2021: Springer, pp. 129-138.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 2015: Springer, pp. 234-241.

[39] T. C. Mok and A. Chung, "Affine medical image registration with coarse-to-fine vision transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20835-20844.p. 18-23, 1996.

Authorized licensed use limited to: Shenzhen Technology University. Downloaded on March 13,2024 at 07:20:25 UTC from IEEE Xplore. Restrictions apply.