

# LKCA-Net: Medical Image Registration Network Enhanced by Large Kernel Convolution Attention

Peiwen Yan. Zhuang Wang\*

Qingdao University, Qingdao 266071, China

Email: 445416349@qq.com

**How to cite this paper:** Yan, P., & Wang, Z. (2026). LKCA-Net: Medical Image Registration Network Enhanced by Large Kernel Convolution Attention. Journal of Computer Science and Frontier Technologies, 4(2), 43–54. ISSN Print: 3104-4204, ISSN Online: 3104-4212.

<https://doi.org/10.63313/JCSFT.9102>

Published: 2026-09-28

Copyright © 2026 by author(s) and Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



## Abstract

In medical image registration, current models cannot simultaneously achieve favorable registration performance and maintain a small number of computational parameters. To solve this problem, this paper proposes a method based on U-shaped structure. By introducing Convolutional Block Attention Module (CBAM) and multi-scale large kernel convolution (LK) into this structure (LKCA-Net), it enhances feature expression ability, reduces information loss and increases the receptive field in the registration process. This method adopts the attention mechanism at the early stage of the encoder to simulate the way human eyes preferentially focus on key target areas when initially observing images. Meanwhile, at the  $1/4$ ,  $1/8$  and  $1/16$  resolution scales of the encoding path, large kernel convolution modules with parallel  $1 \times 1 \times 1$ ,  $3 \times 3 \times 3$  and  $5 \times 5 \times 5$  branches are deployed, and more contextual information can be effectively captured under controllable computational overhead by means of grouped convolution. The experimental results show that compared with indicators including Dice and parameter size, the LKCA-Net model using attention mechanism and multi-scale large kernel is obviously superior to the currently popular TransMorph and VoxelMorph models in terms of computational parameters and registration accuracy.

## Keywords

Medical Image; Attention Mechanism; Large Kernel Convolution; Unsupervised Deformable Registration

## 1. Introduction

In the field of medical image analysis, unsupervised medical image registration[1] is a basic but important technology. Its goal is to find the optimal spatial transformation[2] for images obtained at different times, with different equipment or from different modalities, and then align one medical image to the coordinate system of another image. This technology plays a decisive role in practical clinical procedures such as surgical planning and therapeutic effect evaluation. However, traditional methods suffer from a large number of parameters, high computational

complexity and long running time, which fail to meet the requirements of real-time clinical applications. However, such methods have extremely high computational costs[3] and cannot meet clinical demands in practical situations.

In recent years, unsupervised registration methods based on deep learning have achieved breakthrough progress. As one of the representative works in this field, VoxelMorph firstly models the registration task as a function approximation problem and uses convolutional neural network (CNN) to directly predict deformation fields from image pairs so as to realize fast end-to-end registration[4]. This method adopts U-Net as the basic network structure[5]. Multi-layer encoders extract features of different resolutions, then decoders generate deformation fields with the same resolution as input images after up sampling, and the final step realizes image deformation through the spatial transformation layer[6]. Compared with traditional methods, VoxelMorph improves efficiency by orders of magnitude, yet its standard convolution brings insufficient receptive field and fails to effectively capture long-range spatial dependencies; Furthermore, they cannot focus on the anatomical information that is truly beneficial to the registration task.

To address the above limitations, this paper proposes an enhanced deformable registration network[7]. First, we introduce the Convolutional Block Attention Module (CBAM)[8] into the shallow layers of the encoder to simulate the behavior that human eyes preferentially focus on key target areas when initially observing images. Next, inspired by the outstanding performance of large kernel convolution in dense prediction tasks, we deploy multi-scale large kernel convolution module (LK-Block)[9] at the  $1/4$ ,  $1/8$  and  $1/16$  resolution paths of the encoder. The module is constructed as a multi-branch structure with three convolution sizes of  $1 \times 1 \times 1$ ,  $3 \times 3 \times 3$  and  $5 \times 5 \times 5$  simultaneously. Residual connections are introduced to facilitate gradient propagation, which can effectively expand receptive field while keeping computational efficiency[10]. Such multi-scale enhancement enables the network to integrate local details and global context from images of different resolutions, which is greatly helpful for processing anatomical structures with obvious scale changes in medical images.

In summary, the main contributions of this paper are listed below:

- Introduce attention mechanism into the VoxelMorph framework to realize adaptive weighting on feature channel and spatial dimensions.
- A multi-scale large-kernel convolution module is constructed to effectively expand the receptive field of the network.
- Significantly improve registration accuracy for complex anatomical deformation while retaining end-to-end training and registration efficiency. Subsequent chapters will introduce the network architecture, loss function and experimental results of the proposed method in detail sequentially.

## 2. Methods

## 2.1. Unsupervised Registration Structure

This registration framework adopts the popular unsupervised end-to-end learning paradigm. Given a pair of three-dimensional medical images to be registered, namely moving image and fixed image, they are imported into dual channels and then fed into Large Kernel Convolution Attention Enhanced Registration Network (LKCA-Net). LKCA-Net can learn the rules of nonlinear transformation, which are used to determine the registration positions of voxels in the moving image relative to the fixed image. Then, the differentiable spatial transformation layer performs resampling on the moving image by combining trilinear interpolation according to the information of the deformation field and finally obtains the registered image. During the model training phase, on the one hand, the similarity loss function is adopted to constrain the grayscale similarity of images and drive the network to continuously optimize the deformation field. On the other hand, gradient regularization is introduced to ensure the spatial smoothness of the deformation field and prevent the deformation field from distortion. The entire framework realizes end-to-end training via backpropagation without requiring pre-provided real deformation fields as supervision labels.

## 2.2. Multi-scale Large Kernel Convolution Attention Mechanism Framework

Aiming at the problems of insufficient feature extraction and huge parameter consumption in practical clinical medical registration, this paper proposes a novel unsupervised registration network named LKCA-Net. The VoxelMorph model adopts single-scale convolution, which leads to inadequate feature extraction and insufficient learning capability caused by a small receptive field. To tackle the above issues, cascaded attention mechanism is added during down sampling of VoxelMorph backbone network. First, the CBAM module is embedded in the shallow layers of the encoder to highlight the features of lesion areas and improve the feature representation of these regions. Then the LK-Block is arranged at the 1/4, 1/8 and 1/16 resolution levels of encoder, where parallel  $5 \times 5 \times 5$  large kernel convolution,  $3 \times 3 \times 3$  standard convolution and  $1 \times 1 \times 1$  point convolution are adopted to capture contextual information of images[11]. This method can not only extract discriminative features better but also obtain more contextual information through multi-scale convolution. The detailed network framework is shown in Figure 1.

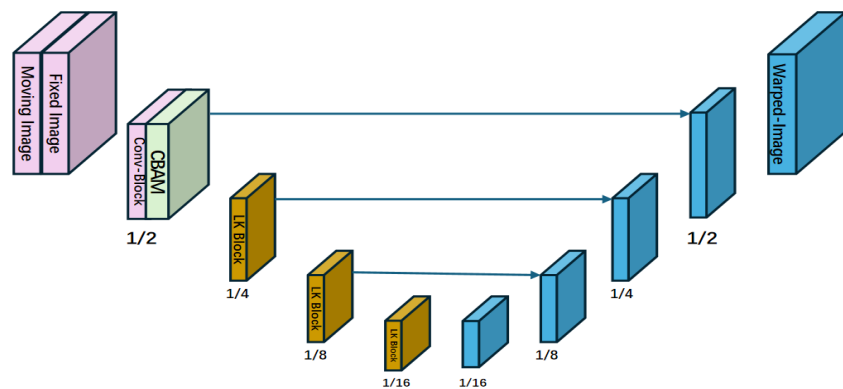


**Figure 1.** Infrastructure and composition of the unsupervised registration network LKCA-Net framework

### 2.2.1. Multi-scale Large Kernel Convolution

The LK-Block adopts a multi-branch parallel mode to realize effective extraction of multi-scale features. This module is constructed with three convolution kernels of different sizes in parallel: the large-size convolution branch with a default  $5 \times 5 \times 5$  kernel is used to effectively expand the receptive field to obtain more contextual information and adapt to various 3D medical image registration tasks, and the large convolution kernel can be adjusted according to the demands of different tasks; the medium-size ordinary convolution branch adopts a  $3 \times 3 \times 3$  kernel to capture local detailed features; the point-size convolution branch uses a  $1 \times 1 \times 1$  kernel to complete feature transformation on the channel dimension. Next, the LK-Block also applies residual connection, and the outputs of three branches are fused with input features via direct addition across different scales, which can avoid gradient attenuation and feature degradation caused by the increase of network depth. The final output is obtained through the PReLU activation function. This mode of "parallel extraction with different sizes - residual fusion - nonlinear activation" can not only effectively improve feature representation capability, but also greatly cut down computational parameters compared with the TransMorph model, and its structural framework is shown in Figure 2.

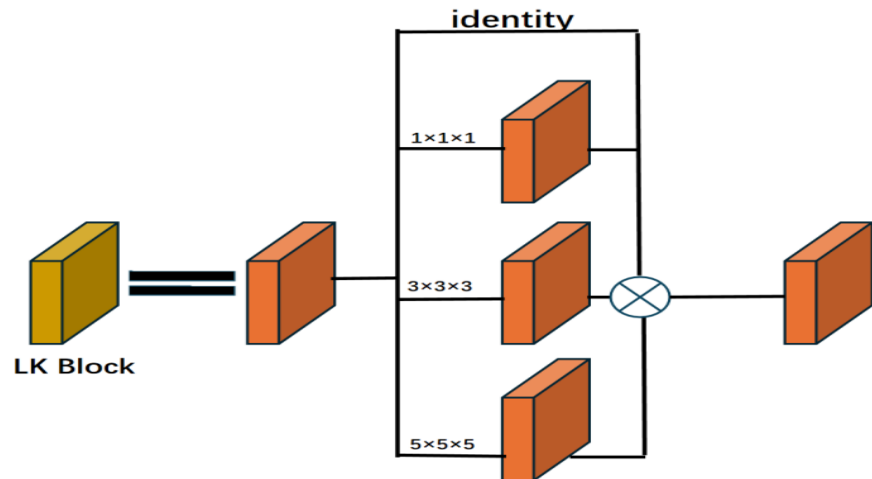


**Figure 2.** Structural framework of multi-scale large kernel convolution

### 2.2.2. CBAM Module

The CBAM module is composed of two parts: channel attention and spatial attention. When we observe an image, we will be attracted to certain regions, and areas outside these regions will not be noticed, which further increases the visual weight of such attractive regions. Specifically, this attention mechanism[12] can raise the weights of key channels or spatial feature maps, enhance the feature representation of key features, and thus improve registration performance. The structural framework is shown in Figure 3.

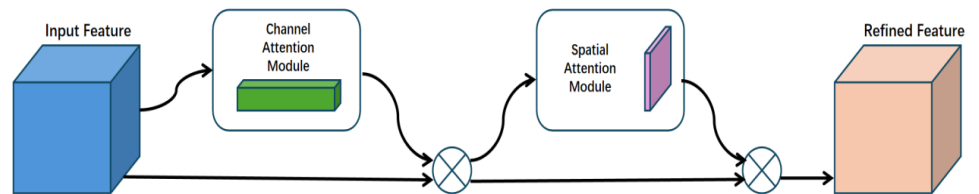


**Figure 3.** Structural block diagram of CBAM attention mechanism

The channel attention module first performs global average pooling and global max pooling operations on the input feature map respectively, thus obtaining two one-dimensional vectors for describing features. Afterwards, these two vectors for describing features are processed through a multi-layer perceptron with shared weights. The perceptron is composed of two  $1 \times 1 \times 1$  convolutional layers, and ReLU activation function is adopted between the two layers to introduce nonlinear features, so as to complete the dimensionality reduction and dimensionality boosting operations on the channel dimension. Finally, the outputs of the two branches are

summed, and Sigmoid function is used for normalization to generate channel attention weights, which are multiplied with the original input feature channel by channel to realize the key calibration of channel features. The structural framework is shown in Figure 4.

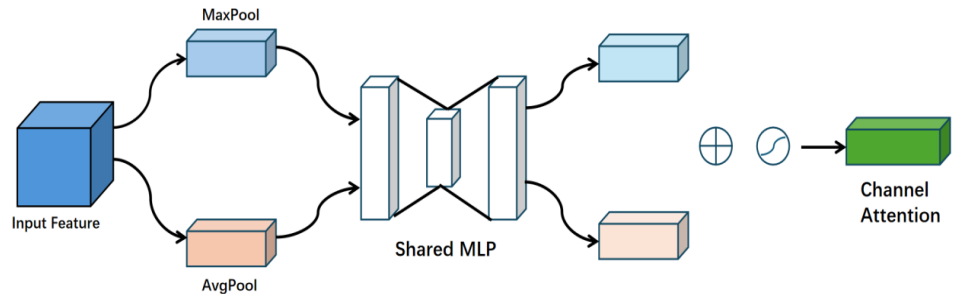


**Figure 4.** Structural block diagram of channel attention module

The input of the spatial attention module is the output of the channel attention module. It first calculates the feature average value and features maximum value along the channel dimension respectively to obtain two single-channel feature description maps, which separately represent the general feature intensity of each spatial position and the intensity of salient target regions. Then these two feature description maps are concatenated along the channel dimension, a  $7 \times 7 \times 7$  convolutional layer is used to complete feature fusion, and a single-channel spatial attention map is subsequently obtained. Finally, an activation function is applied to realize weighting on the spatial dimension. A convolution kernel with a relatively large size is adopted to obtain a sufficient receptive field, which can effectively capture spatial context information between adjacent voxels. In LKCA-Net, the CBAM module is placed after the first-layer encoder, which can enhance shallow features and provide a more effective input representation for subsequent deep feature extraction. Meanwhile, it can avoid the gradient propagation problem caused by introducing attention in the deep layers of the network. The structural framework is shown in Figure 5.

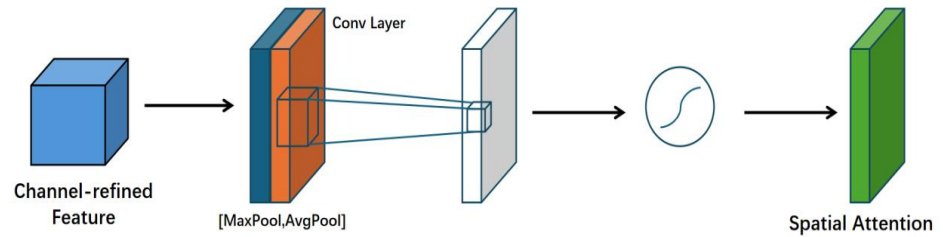


**Figure 5.** Structural block diagram of spatial attention module

### 3. Experiment

#### 3.1. Dataset

The OASIS dataset is used for model training and validation in this study. This dataset contains T1-weighted brain MRI scan data of 414 subjects, covering young, middle-aged, non-dementia elderly and dementia elderly groups, and it is highly representative in terms of age span and cases[13]. All images have undergone preprocessing, the three-dimensional images cropped to appropriate sizes are adopted, and pixel values are normalized. In addition, this dataset also includes structural label maps for evaluating the registration performance of the model.

#### 3.2. Loss Function

The rationality of loss function directly determines the performance of unsupervised medical image registration. Without annotated anatomical labels, the core of loss function construction lies in balancing the accurate intensity alignment between fixed and moving images as well as the rationality of deformation field. In this paper, the total loss consists of image intensity similarity loss and deformation field smooth regularization loss. The weighted combination effectively balances the impact of the two loss terms, ensuring the registration result achieves favorable registration accuracy as well as reasonable and smooth deformation field. The overall loss function is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sim}(I_{fixed}, I_{warp}) + \lambda \cdot \mathcal{L}_{smooth}(\phi) \quad (1)$$

Among them,  $\mathcal{L}_{sim}$  is used to quantify the global grayscale deviation between the fixed image and the registered image,  $\mathcal{L}_{smooth}(\phi)$  is used to solve the smoothness and continuity problems of the deformation field, and  $\lambda$  denotes the regularization weight hyperparameter[14].

##### 3.2.1. Intensity Similarity Loss

The grayscale distribution of medical images serves as the core basis for the

optimization of grayscale similarity loss. Mean Squared Error (MSE) is adopted as the measurement of image intensity similarity in this paper. Benefiting from high computational efficiency and stable gradient, MSE can effectively quantify the global intensity error between fixed image and registered image[15]. Minimizing this loss enables the network to learn the deformation field for accurate intensity alignment.

Let the voxel space of three-dimensional medical images be  $\Omega \in \mathbb{R}^3$ ,  $p = (x, y, z) \in \Omega$  be the spatial coordinate of any voxel,  $I_{warp}(p) = I_{moving}(\phi(p))$  represent the image obtained by transforming the moving image  $I_{moving}$  via the deformation field  $\phi$ , then MSE is defined as:

$$\mathcal{L}_{sim} = \frac{1}{|\Omega|} \sum_{p \in \Omega} (I_{fixed}(p) - I_{warp}(p))^2 \quad (2)$$

where  $|\Omega|$  denotes the total number of voxels within voxel space  $\Omega$ . This loss calculates the average of squared intensity differences voxel by voxel, which can effectively penalize global intensity discrepancies, make the intensity distribution of the warped image consistent with that of the fixed image, and effectively constrain the accurate alignment of anatomical structures.

### 3.2.2. Deformation Field Gradient Smooth Regularization Loss

In unsupervised medical image registration tasks, if only the intensity similarity loss is adopted, the predicted deformation field may suffer from excessive local deformation and topological folding inconsistent with physiological characteristics. Such registration results deviate from real human anatomical structures and reduce the clinical application value of registration outcomes. Therefore, introducing the deformation field smooth regularization loss is of great significance. This loss term can constrain the spatial continuity characteristic of the deformation field, meanwhile ensure that the deformation transformation conforms to human physiological characteristics, and significantly improve the overall smoothness and physiological rationality of registration results.

Let  $u = \phi$  denote the three-dimensional displacement deformation field output by the model network,  $\nabla u(p)$  represent the three-dimensional spatial gradient vector of this deformation field at any voxel coordinate  $p$  within the domain  $\Omega$ , then the gradient smooth regularization loss is defined as:

$$\mathcal{L}_{smooth}(\phi) = \sum_{p \in \Omega} \|\nabla u(p)\|^2 \quad (3)$$

### 3.3. Baseline and Ablation Experiments

This paper compares the performance of LKCA-Net with that of current mainstream registration models. In addition, we conduct ablation experiments on the CBAM attention module, the multi-scale large-kernel convolution module as well as the cascaded module combining CBAM attention and multi-scale large-kernel convolution, so as to verify that the proposed LKCA model can boost the performance of unsupervised registration. The environmental configuration of all

network models remains unchanged throughout the experiments.

### 3.4. Implementation

The LKCA-Net is trained on an NVIDIA GeForce RTX 3080 Ti graphics card with 12 GB video memory under PyTorch 2.8.0. The batch size is fixed at 1, and the initial learning rate is set to  $10^{-4}$ . The loss function is formulated as the weighted sum of mean squared error similarity loss and deformation field gradient regularization loss, with the regularization coefficient  $\lambda$  empirically set to 0.01. To avoid computational loss caused by unexpected training interruption, model checkpoint is saved every 500 iterations.

## 4. Results

### 4.1. Generalization Experiment Results

To verify the generalization ability and registration accuracy of the proposed LKCA-Net, classical mainstream models in unsupervised medical image registration are selected as baselines, including VoxelMorph[16], CSAIN and TransMorph[17].

**Table 1.** Comparison of Generalization Performance Among Different Models

| Model      | Dice (%)         | Parameters |
|------------|------------------|------------|
| VoxelMorph | $74.06 \pm 2.76$ | 301,411    |
| CSAIN      | $75.40 \pm 1.29$ | 503,744    |
| TransMorph | $79.13 \pm 1.47$ | 46,771,251 |
| Ours       | $80.45 \pm 1.20$ | 867,188    |

As shown in the experimental data in Table 1, the proposed LKCA-Net outperforms the VoxelMorph model and the CSAIN model[18] by a notable margin in registration accuracy. Its accuracy is only slightly higher than that of the TransMorph model, while the total parameter quantity of LKCA-Net is far smaller than that of TransMorph, which effectively reduces the computational overhead of the model. The above results demonstrate that LKCA-Net achieves lightweight design while improving registration precision.

### 4.2. Results of Ablation Experiments

To further verify the respective contributions of the CBAM attention module and the multi-scale large-kernel (LK) convolution module to registration performance, a set of ablation experiments are carried out as illustrated in Figure 6 and Figure 7. The experimental groups consist of the network equipped only with the CBAM attention module, the network equipped only with the multi-scale large-kernel (LK) convolution module, and the complete LKCA-Net integrating both CBAM and LK modules. All other training configurations are kept identical across all groups.

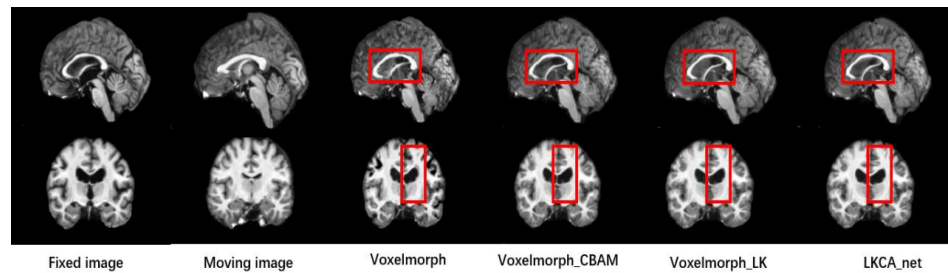


**Figure 6.** Visualization results of ablation experiments (sagittal and transverse planes)

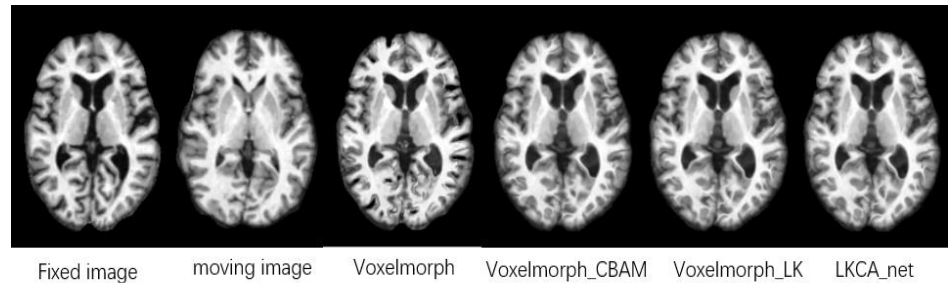


**Figure 7.** Visualization results of ablation experiments (coronal plane)

**Table 2.** Results of Ablation Experiments

| Model           | Dice (%)         | Parameters |
|-----------------|------------------|------------|
| Voxelmorph      | $74.06 \pm 2.76$ | 301,411    |
| Voxelmorph_CBAM | $79.93 \pm 1.35$ | 397,169    |
| Voxelmorph_LK   | $80.33 \pm 1.28$ | 866,470    |
| Ours            | $80.45 \pm 1.20$ | 867,188    |

According to the quantitative experimental results in Table 2, registration accuracy is greatly improved when only the CBAM attention module is adopted, which verifies that this module enables the model to adaptively strengthen valid features and suppress interference from irrelevant background contents. The standalone multi-scale large-kernel convolution module also brings prominent performance gains, proving that it can expand the receptive field to capture abundant contextual information. The complete LKCA-Net integrating both CBAM and LK modules achieves optimal registration performance, indicating that the two modules complement and cooperate with each other to cut computational cost and raise registration precision simultaneously. In terms of visualized experimental outcomes, compared with the original VoxelMorph, the variant with only CBAM and the variant with only multi-scale large-kernel convolution, LKCA-Net realizes the most precise alignment on cerebral tissue boundaries and deep brain nuclei, generates smoother deformation fields and better conforms to the actual deformation rules of anatomical structures.

## 5. Conclusion

To solve the problem that favorable registration performance and a small number of computational parameters cannot be guaranteed simultaneously, this paper designs

a registration network LKCA-Net combining multi-scale large convolutions and attention mechanisms. This method adds cascaded attention mechanisms during the down sampling process of the VoxelMorph backbone network, adopts the series connection of channel attention and spatial attention to adaptively enhance the feature representation of lesion areas and reduce the interference of irrelevant backgrounds; then large-scale convolutions are applied to the low-resolution layers of the encoder to capture image context information. Experimental results verify that LKCA-Net can effectively reduce the computational cost of the model while maintaining favorable registration performance, and it can fully exert its value in practical clinical procedures such as surgical planning and therapeutic effect evaluation.

## References

- [1] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2018). An unsupervised learning model for deformable medical image registration. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9252-9260).
- [2] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. *Advances in neural information processing systems*, 28.
- [3] vs Transformer, U. N., & Is, U. (2022, December). Outdated in Medical Image Registration?. In *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings* (p. 151). Springer Nature.
- [4] Mok, T. C., & Chung, A. (2020). Fast symmetric diffeomorphic image registration with convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4644-4653).
- [5] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.
- [6] Boveiri, H. R., Khayami, R., Javidan, R., & Mehdizadeh, A. (2020). Medical image registration using deep neural networks: a comprehensive review. *Computers & Electrical Engineering*, 87, 106767.
- [7] Miao, S., Wang, Z. J., & Liao, R. (2016). A CNN regression approach for real-time 2D/3D registration. *IEEE transactions on medical imaging*, 35(5), 1352-1363.
- [8] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). Cbam: Convolutional block attention module. In *European conference on computer vision* (pp. 3-19). Cham: Springer International Publishing.
- [9] Ding, X., Zhang, X., Han, J., & Ding, G. (2022, June). Scaling up your kernels to  $31 \times 31$ : Revisiting large kernel design in CNNs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11953-11965). IEEE.
- [10] Araujo, A., Norris, W., & Sim, J. (2019). Computing receptive fields of convolutional neural networks. *Distill*, 4(11), e21.
- [11] Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B. P., Išgum, I., & Staring, M. (2017, September). Nonrigid image registration using multi-scale 3D convolutional neural networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 232-239). Cham: Springer International Publishing.
- [12] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3), 331-368.
- [13] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L.

- (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9), 1498-1507.
- [14] Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., & Dalca, A. V. (2021, June). Hypermorph: Amortized hyperparameter learning for image registration. In *International conference on information processing in medical imaging* (pp. 3-17). Cham: Springer International Publishing.
- [15] Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, 7(3), 8-18.
- [16] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8), 1788-1800.
- [17] Chen, J., Frey, E. C., He, Y., Segars, W. P., Li, Y., & Du, Y. (2022). Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82, 102615.
- [18] Wang, Y., Qiu, H., & Qin, C. (2023, April). Conditional deformable image registration with spatially-variant and adaptive regularization. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.