Medical Image Registration Based on Uncoupled Learning and Accumulative Enhancement

Yucheng Shu[1,2], Hao Wang[1,2], Bin Xiao[1,2(✉)], Xiuli Bi[1,2], and Weisheng Li[1,2]

[1] Chongqing University of Posts and Telecommunications, Chongqing 400065, China
{shuyc,xiaobin,bixl,liws}@cqupt.edu.cn, s190201027@stu.cqupt.edu.cn
[2] Chongqing Key Laboratory of Image Cognition, Chongqing 400065, China

Abstract. As a basic building block in medical image analysis, image registration has been greatly developed since the emergence of modern deep neural networks. Compared to non-learning-based methods, the latest approaches can learn task-specific features spontaneously, thus generate the registration results with one round of inference. However, when large inter-image distortion occurs, the stability of existing methods can be strongly affected. To alleviate this problem, the iterative framework based on coarse-to-fine strategies has been introduced in recent works. However, their networks at each iteration step are relatively independent, which is not an optimal solution for the reinforcement of image features. What is more, the moving and the fixed images are often concatenated or fed to identical network layers. Consequently, the iterative learning and warping on the moving image can be entangled with the fixed image. In order to address these issues, we present a novel medical image registration framework, namely ULAE-net, to continuously enhance the spatial transformation and establish more profound contextual dependencies under a compact network layout. Extensive experiments on 3D brain MRI data sets demonstrate that our method has greatly improved the registration performance, thereby outperforms state-of-the-art methods under large-scale deformations (https://github.com/wanghaostu/ULAE-net).

Keywords: Medical image registration · Uncoupled learning · Accumulative enhancement

1 Introduction

Deformable image registration plays an essential role in the field of medical image analysis and diagnosis. By aligning the anatomical structures in the moving and the fixed image, it can lay a solid foundation for many subsequent tasks such as

This research was funded in part by the National Natural Science Foundation of China 61906024, 61976031, and 61801068, National Major Scientific Research Instrument Development Project of China 62027827, National Key R&D Program of China 2019YFE0110800, 2016YFC1000307-3.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12904, pp. 3–13, 2021.
https://doi.org/10.1007/978-3-030-87202-1_1

medical image fusion [5,9], medical image reconstruction [7,13], medical image segmentation [16,20], etc.

In the traditional non-learning-based approaches [22–24], deformable registration is often performed in an iterative manner. By gradually minimizing a predefined spatial-structural energy function, they have achieved great accuracy in most of the medical image registration tasks. However, when the large-scale spatial displacement exists between the input images (e.g. 3D brain MRI), the optimization algorithms may become time-consuming, and more likely, be trapped into local minimum.

Recently, with the theoretical development of machine learning techniques, deep-learning-based methods have shown promising quality and speed in a variety of medical image registration tasks [12,21]. For instance, by employing a U-shape network to calculate the pixel-level deformation field, Voxelmorph [2,4] is able to generate 3D registration result through one round of network inference, which has tremendously sped up the medical image registration process. Nevertheless, even with the help of automatic feature learning, it is also a great challenge when two scans have large differences in appearance.

Although large data transformations can be dealt with by building long-range dependencies via pooling operation, the spatial resolution of the image features will be affected at the network bottleneck. According to our experiments, it may reduce the network's ability to generate local subtle deformations, thus resulting in the decrease of registration accuracy. To circumvent this issue, some existing methods attempt to apply the coarse-to-fine strategy with multiple registration steps. For instance, Dual-PRNet [10] was proposed to calculate multi-scale deformation fields based on two separated network branches. Hering et al. [8] proposed a hierarchical layout to align the inhale-to-exhale lung scans. Zhao et al. [25] designed the Recursive Cascaded Networks to register images iteratively.

However, the methods mentioned above still have certain limitations. Firstly, the net branches at each iteration are often relatively independent, which may directly increase the network complexity. More importantly, such discrete optimizing strategy cannot make full use of the cumulative ability of the data-driven learning, thus the spatial transformations may not be continuously enhanced. What is more, the moving and the fixed images are often concatenated as the input of a single net routine, or independently fed to similar feature encoders, which is relatively rigid and may not be suitable for the iterative learning and warping of the moving images. In order to address these issues, in this paper, we propose a novel medical image registration framework, namely ULAE-net, based on uncoupled feature learning and spatial accumulative enhancement.

The main contributions of this work can be summarized as follows:

–We present a new accumulative medical image registration framework to effectively enhance the spatial transformations from continuous optimizations under a more compact network layout, thereby construct a balanced combination based on the coarse-to-fine strategy and informative feature learning.

–By introducing a novel uncoupled feature learning algorithm, we are able to capture richer long-range dependencies and build better visual correlations between the moving and fixed images for the registration task under large deformations.

–Finally, a simple yet effective multi-window loss is designed to cooperate with our proposed framework, to further expand the network's learning potentials while avoiding the local minimum. Extensive experiments on two large scale MR brain data sets, Mindboggle101 and IXI, show that our proposed algorithm greatly improved the registration accuracy and robustness, thereby outperforms the SOTAs among traditional and deep-learning-based methods.

2 Method

2.1 Overview

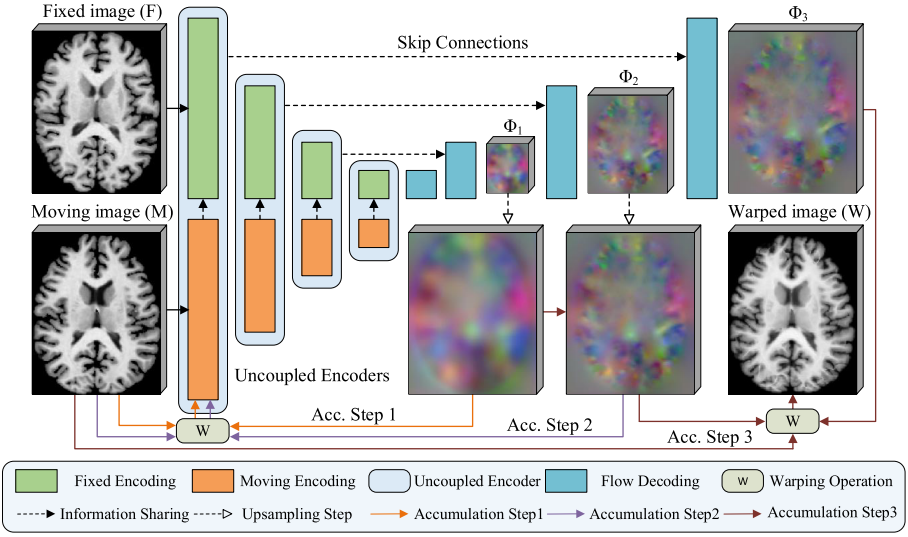

Fig. 1. Overall architecture of our proposed registration framework.

As shown in Fig. 1, given a fixed 3D image F and a moving 3D image M, the output of our registration model is a 3-channel deformation field ϕ which containing the voxel-to-voxel correspondence:

$$\phi = U_\theta(M, F) \tag{1}$$

where U corresponds to our proposed network with parameter θ. Specifically, we employ the U-shape [19] architecture with encoding routine to learn contextual features, and decoding routine to generate spatial displacements. Skip connection is also applied for gradient transduction and information sharing. Typically, the $2 \times 2 \times 2$ stride convolution is used in the encoder, and $\times 2$ trilinear up-sampling and $3 \times 3 \times 3$ convolution is applied in the decoder. As introduced in the previous section, our basic rationale is to devise a compact network layout under an

unsupervised framework, thus address the large-deformation problem by learning and accumulating long-range information. Therefore, a novel uncoupled learning scheme is proposed to feed the moving and fixed image into different learning paths, and then form an integrated encoder with feature fusion. At the end of the decoder, an accumulative enhancement mechanism is proposed to cooperate the moving encoding path to continually enhance the network's potential transformation ability. Detailed introductions will be given in the following sections.

2.2 Uncoupled Spatial Encoder

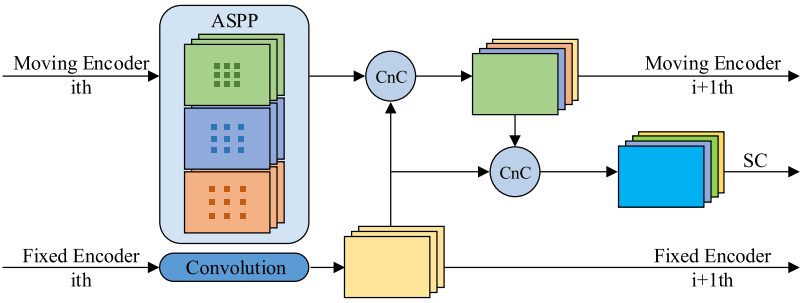

Fig. 2. The Uncoupled Spatial Encoder. CnC denotes Concatenation and Convolution, SC corresponds to Skip Connection.

During the voxel-wised registration process, the moving image will undergo spatial transformations based on the displacement field calculated by the network. Therefore, compared to the fixed image, there should be specified treatments imposed on the moving image to discover and learn more long-range dependencies. However, in the standard U-shape registration framework, the moving and fixed images are usually concatenated as the input of a same encoding routine, or independently feed to similar feature encoders, which is not an optimal solution, especially under large image deformations.

Therefore, in order to capture both long-range and short-range corresponding information between the moving and fixed image, a novel encoder structure, namely Uncoupled Spatial Encoder, is proposed for the deformable image registration. As shown in Fig. 2, the learning path of moving and fixed images are separated into two different data routines. Specifically, we apply the typical stride convolution on fixed image encoding path, and utilized the atrous spatial pyramid pooling [3] (ASPP) as the basic building block in the moving image encoding module. ASPP layer use dilated convolutions with several different sampling rates to expand the receptive field of moving routine, which is far more flexible, and can capture the context information at multi-scales. Moreover, as shown in Fig. 2, in order to learn a better contextual relationship between two images, we also apply an alternating fusion module based on concatenation and

convolution, to append the fixed encoding information to the moving encoding routine. Under this feature learning mechanism, the network is able to capture much more longer-range information at each voxel and acquire a more flexible transformation field for the moving image.

2.3 Accumulative Warping Enhancement

As analyzed in the previous section, although the coarse-to-fine strategy has shown its effectiveness in large-scale deformable registration [8,10,17,25], the iterations are often performed with independent encoders and decoders, which has limitations in terms of feature learning and computational complexity. In order to continuously enhance the network's transformation ability, we propose a novel spatial accumulative enhancement mechanism. Specifically, three accumulative steps (Acc_1, Acc_2, Acc_3) are employed within the proposed network. Note that the accumulative iterations are performed compactly without the need for new network branches or extra weight sharing.

In the beginning, F and M are passed through 4 uncoupled encoder blocks to capture semantic information, and the first two decoder blocks will transmit the encoded feature to the coarsest deformation field ϕ_1. Then, ϕ_1 is up-sampled to the original shape, and the moving image M is warped accordingly to obtain roughly aligned image M_1. M_1 and fixed image F is re-input to the network as the next accumulative step. The calculation of Acc_2 and Acc_3 is similar to Acc_1, but they will utilize one and two more decoder layers, respectively. At last, the finest flow ϕ_3 will be generated at the final step, and the transformation field ϕ from M to F can be obtained by accumulating the ϕ_1, ϕ_2, and ϕ_3 as:

$$\phi = S^{2,3}(S^{1,2}(\phi_1) \circ \phi_2 + \phi_2) \circ \phi_3 + \phi_3 \tag{2}$$

Where $\circ$ is the warping operation based on trilinear interpolation [11], S is the up-sampling function, and $+$ corresponds to the voxel-wised addition operation on the transformation tensor. The final warped image W can be acquired by $M \circ \phi$ as the final registration result.

Note that the final deformation ϕ is not simply computed by adding ϕ_1, ϕ_2, and ϕ_3, as the input at each iteration step will change accordingly. On the contrary, we calculate the final flow recursively. For example, at Acc_2, the flow ϕ_1 from Acc_1 has to be warped firstly by ϕ_2 as coordinate alignment and then added back to ϕ_2. This step will guarantee that we can finally acquire the correct flow. By closely cooperate with the uncoupled encoder, the corresponding information at different scales can be effectively recovered, and the potential abilities of the network can be continuously reinforced.

2.4 Multi-window Loss

In many deformable image registration methods [2,15,17,25], negative local normalized cross-correlation (NCC) is successfully utilized to be a similarity metric for gradient-descent optimization. However, in our framework, we propose to

learn and accumulate richer contextual information with various types of spatial correlations. Thus different parts of the network should receive more specified guidance, and the direct use of NCC would be not suitable enough. Therefore, in order to prevent our model from being trapped into local optimal, we propose a new multi-window loss, which is simple yet effective, by calculate the weighted sum of NCC with different window size as the similarity metric:

$$L_{sim}(F, M \circ \phi) = -\sum_{i=1}^{K} \gamma^{i-1} NCC_{w_i}(F, M \circ \phi) \tag{3}$$

where K is the number of different windows, γ is a hyperparameter, and w_i is the window size of ith NCC part. In our experiment, we set $\gamma = 0.5$, and $K = 3$ according to the number of accumulation steps in our method, and $(w_1, w_2, w_3) = (11, 9, 7)$. We also constrain the smoothness of the deformation field ϕ with a $L2$ regularizer:

$$L_{smooth}(\phi) = \sum_{p \in \Omega} \|\nabla \phi(p)\|^2 \tag{4}$$

In order to provide more spatial transformation flexibility and transit the gradient through hierarchical aggregation, we only constrain the smoothness at the final computed deformation field ϕ in Eq. 2. Finally, with the regularization parameter λ, the loss of our method is defined as:

$$L_{total}(F, M \circ \phi) = L_{sim}(F, M \circ \phi) + \lambda L_{smooth}(\phi) \tag{5}$$

3 Experiments and Results

Data Set. In order to demonstrate the effectiveness of our proposed method, extensive experiments were conducted on two 3D brain MRI data sets: the Mindboggle101 [14] data set and IXI data set[1]. Mindboggle101 contains 101 T1 weighted MR scans, which were annotated with 25 cortical segmentation parts. It can be used to evaluate registration accuracy for large-scale deformation. We follow [15] to combine all 25 categories into 5 areas for a better review. IXI data set contains nearly 600 MR images from normal and healthy subjects. We randomly selected 100 scans for the experiment and use FreeSurfer [6] for preprocessing, including skull stripping, affine transformation, and segmentation for evaluation.

Experimental Settings. For Mindboggle101 dataset, we followed the protocols in [15] to prepare the training and testing set. Our model was trained on NKI-RS-22 and NKI-TRT-20 subset, (42×41, 1722 pairs in total), and tested on OASIS-TRT-20 subset (20×19, 380 pairs in total). For IXI data set, 80 scans (80×79, 6320 pairs in total) were used for training and 20 scans (20×19, 380 pairs

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

in total) were used for testing. Each scan is cropped to 160×192×160, affinely warped to MNI152 space, and normalized by the maximum voxel intensity of each brain volume. Our model and other comparative deep learning models are all implemented with Pytorch [18] and trained on 1 T V100 GPU. We set learning rate = 10^{-4}, $\lambda = 1$ and batch-size = 1. Instance Normalization and Leaky-ReLU activation was used after each convolution except the last output.

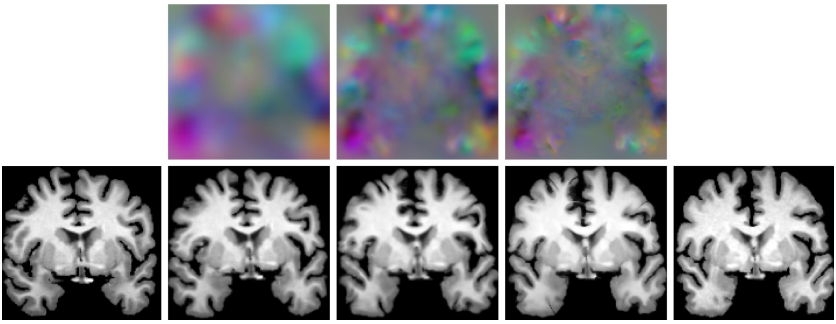

Fig. 3. Top: The generated displacement fields. Bottom (left to right): the moving image, the images warped accumulatively by three deformation flow, and the fixed image.

Visualization, Comparison, and Discussion. Firstly, the generated deformation fields and the warped images at each accumulation step are visualized in Fig. 3. It can be found that the first generated flow has the coarsest resolution and mainly focus on the alignment of the large image areas, while the last estimated flow has the finest resolution and is able to provide rich textural guidance for the registration.

We further compared our method with the state-of-the-art approaches, including SyN [1], VoxelMorph (VM) [2], RCN [25], and LapIRN [17]. VM is one of the most popular registration methods in the past three years, RCN and LapIRN are the latest methods to perform iterative large-scale deformable registration. We implemented SyN by using ANTs [1], and used the original codes of VM and LapIRN. Moreover, as the amount of parameters of default VM is smaller than other methods, we also doubled the channels of VM (VMx2), to make a relatively fair comparison. As shown in Table 1, our proposed method has demonstrated its superior registration accuracy by achieving best Dice score over all 9 categories on both Mindboggle101 and IXI data set.

As mentioned above, one merit of our method is to effectively perform registration under the accumulative layout with end-to-end training. Unlike other iterative methods, we do not have to divide the training into multiple stages, and the loss is only computed once. Therefore, the gradients at the frontend of the net can be computed by multiple times. Specifically, in our network, if

a tensor is used more than one time, the gradient associated to it is the sum of the gradients corresponding to each iteration. We also evaluated the network complexity in terms of the Flops and the number of parameters (the last two rows in Table 1).

Table 1. In Mindboggle101, FL: Frontal region, PL: Parietal region, OL: Occipital region, TL: Temporal region, CL: Cingulate region; In IXI: LWm: Left white matter, RWm: Right white matter, LGm: Left gray matter, RGm: Right gray matter.

Class	Initial	SyN	VM	VMx2	RCN	LapIRN	Ours
FL	0.327	0.521	0.576	0.605	0.623	0.626	**0.664**
PL	0.312	0.452	0.526	0.555	0.564	0.566	**0.607**
OL	0.269	0.402	0.447	0.479	0.493	0.507	**0.549**
TL	0.352	0.543	0.583	0.609	0.624	0.616	**0.657**
CL	0.456	0.630	0.674	0.698	0.695	0.696	**0.722**
Avg.	0.343	0.510	0.561	0.589	0.600	0.602	**0.640**
LWm	0.664	0.788	0.839	0.841	0.843	0.847	**0.860**
RWm	0.663	0.788	0.839	0.842	0.844	0.847	**0.862**
LGm	0.495	0.661	0.792	0.806	0.808	0.813	**0.817**
RGm	0.493	0.660	0.792	0.806	0.807	0.812	**0.820**
Avg.	0.579	0.724	0.815	0.824	0.826	0.830	**0.840**
Flops	–	–	164.74G	643.67G	494.76G	588.57G	543.18G
Params.	–	–	396.45K	1.58M	1.19M	923.75K	789.56K

Moreover, to better illustrate the characteristic of our framework, we compared our accumulative enhancement process to the similar iteration steps in LapIRN, and both of them have 3 levels for large-scale deformation registration. As shown in Fig. 4, we visualize the intensity difference of the moving images between different warping stages as: $|M_1 - M|$, $|M_2 - M_1|$, and |final warping$-M_2|$. It shows that both LapIRN and our method can acquire finer flow of smaller structures at later iteration stage, which demonstrates that the iterative-based method is able to model long-range dependencies with lager receptive field, while obtain more detailed local transformations. Further more, compared to LapIRN's results (Fig. 4(b)), it can be observed that even at the finest warping process, our method (Fig. 4(c)) is still able to discover more detailed correspondences and perform the continuous transformation, thanks to uncoupled feature learning and spatially accumulated information.

Ablation Study. We conducted an additive ablation study on Mindboggle101 to illustrate the effectiveness of our proposed modules. The result is shown in Table 2. We choose the basic U-shaped transformation network with a single window NCC loss as the baseline. As shown by the statistics, after progressively

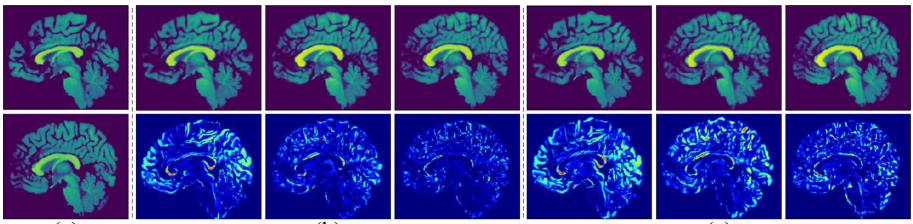

Fig. 4. (a) The moving (up) and fixed (down) image, (b) three levels of visualization of LapIRN, (c) three levels of visualization of our method. Brighter pixel corresponds to higher intensity difference.

adding Uncoupled Spatial Encoder, Accumulative Warping Enhancement, and Multi-window Loss module to the baseline, the performance of our proposed framework also improved accordingly, which implies that by actively capture richer spatial relationships within the images, all the network modules in our method are very effective for the large-scale deformable registration.

Table 2. The ablation study results on: 1) Accumulative Warping Enhancement (AWE), 2) Uncoupled Spatial Encoder (USE), and 3) Multi-window Loss (ML).

Model name	1)	2)	3)	FL	PL	OL	TL	CL	Avg.
Baseline	×	×	×	0.576	0.526	0.447	0.583	0.674	0.561
AWE	√	×	×	0.631	0.569	0.506	0.624	0.695	0.605
AWE+US	√	√	×	0.654	0.597	0.541	0.650	0.718	0.632
AWE+US+ML	√	√	√	**0.664**	**0.607**	**0.549**	**0.657**	**0.722**	**0.640**

4 Conclusion

In this paper, we presented a novel unsupervised medical image registration method, ULAE-net, for large-scale deformations. By introducing an uncoupled spatial encoder, we are able to effectively build long-range visual correlations between the moving and fixed image. During the spatial transformation stage, the proposed accumulative warping enhancement mechanism is applied to perform additive iterations integrally within the network. To provide stronger training guidance, we also proposed a simple yet effective multi-window setting for the local similarity measurement. Experiments on two 3D brain MRI data sets show that our model is able to obtain superior results compared with other state-of-the-art methods, especially in the case of large-scale deformations.

References

1. Avants, B.B., Tustison, N., Song, G.: Advanced normalization tools (ants). Or Insight **1–35** (2008)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
4. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 729–738. Springer (2018). https://doi.org/10.1007/978-3-030-00928-1_82
5. Du, J., Li, W., Lu, K., Xiao, B.: An overview of multi-modal medical image fusion. Neurocomputing **215**, 3–20 (2016)
6. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012)
7. García, H.F., Torres, C.A., Cardona, H.D.V., Álvarez, M.A., Orozco, Á.Á., Padilla, J.B., Arango, R.: 3d brain atlas reconstruction using deformable medical image registration: Application to deep brain stimulation surgery. In: 2014 XIX Symposium on Image, Signal Processing and Artificial Vision, pp. 1–5. IEEE (2014)
8. Hering, A., van Ginneken, B., Heldmann, S.: mlvirnet: Multilevel variational image registration network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 257–265. Springer (2019). https://doi.org/10.1007/978-3-030-32226-7_29
9. Hou, R., Zhou, D., Nie, R., Liu, D., Ruan, X.: Brain CT and MRI medical image fusion using convolutional neural networks and a dual-channel spiking cortical model. Medical and Biological Engineering and Computing (2019)
10. Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M.: Dual-stream pyramid registration network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–390. Springer (2019). https://doi.org/10.1007/978-3-030-32245-8_43
11. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. p. 2017–2025. NIPS'15, MIT Press, Cambridge, MA, USA (2015)
12. Jingfan, F., Xiaohuan, C., Pew-Thian, Y., Dinggang, S.: Birnet: Brain image registration using dual-supervised fully convolutional networks. Medical Image Analysis (2019)
13. Kaur, H., Kumar, S.: A review on decomposition/reconstruction methods for fusion of medical images (2020)
14. Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. **6,** 171 (2012)
15. Kuang, D., Schmah, T.: Faim-a convnet method for unsupervised 3d medical image registration. In: International Workshop on Machine Learning in Medical Imaging. pp. 646–654. Springer (2019). https://doi.org/10.1007/978-3-030-32692-0_74
16. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV) (2016)

17. Mok, T.C., Chung, A.C.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 211–221. Springer (2020). https://doi.org/10.1007/978-3-030-59716-0_21
18. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. arXiv:1912.01703 (2019)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Shu, Y., Wu, X., Li, W.: Lvc-net: Medical image segmentation with noisy label based on local visual cues. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 558–566. Springer (2019). https://doi.org/10.1007/978-3-030-32226-7_62
21. Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B.P., Išgum, I., Staring, M.: Nonrigid image registration using multi-scale 3d convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 232–239. Springer (2017). https://doi.org/10.1007/978-3-319-66182-7_27
22. Sommer, S., Nielsen, M., Lauze, F., Pennec, X.: A multi-scale kernel bundle for lddmm: Towards sparse deformation description across space and scales. In: Székely, G., Hahn, H.K. (eds.) Information Processing in Medical Imaging. pp. 624–635. Springer, Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22092-0_51
23. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. NeuroImage **45**(1), S61–S72 (2009)
24. Wu, G., Kim, M., Wang, Q., Shen, D.: Hierarchical attribute-guided symmetric diffeomorphic registration for MR brain images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 90–97. Springer (2012). https://doi.org/10.1007/978-3-642-33418-4_12
25. Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al.: Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10600–10610 (2019)