FSDiffReg: Feature-Wise and Score-Wise Diffusion-Guided Unsupervised Deformable Image Registration for Cardiac Images

Yi Qin and Xiaomeng Li[(✉)]

The Hong Kong University of Science and Technology, Kowloon, Hong Kong SAR, China
eexmli@ust.hk

Abstract. Unsupervised deformable image registration is one of the challenging tasks in medical imaging. Obtaining a high-quality deformation field while preserving deformation topology remains demanding amid a series of deep-learning-based solutions. Meanwhile, the diffusion model's latent feature space shows potential in modeling the deformation semantics. To fully exploit the diffusion model's ability to guide the registration task, we present two modules: Feature-wise Diffusion-Guided Module (FDG) and Score-wise Diffusion-Guided Module (SDG). Specifically, FDG uses the diffusion model's multi-scale semantic features to guide the generation of the deformation field. SDG uses the diffusion score to guide the optimization process for preserving deformation topology with barely any additional computation. Experiment results on the 3D medical cardiac image registration task validate our model's ability to provide refined deformation fields with preserved topology effectively. Code is available at: https://github.com/xmed-lab/FSDiffReg.git.

Keywords: Deformable Image Registration · Score-based Generative Model

1 Introduction

Deformable image registration is the process of accurately estimating non-rigid voxel correspondences, such as the deformation field, between the same anatomical structure of a moving and fixed image pair. Fast, accurate, and realistic image registration algorithms are essential to improving the efficiency and accuracy of clinical practices. By observing dynamic changes, such as lesions, physicians can more comprehensively design treatment plans for patients [8,11]. When images during surgery align with preoperative ones, surgeons can locate instruments better and improve surgical prognosis [1]. As reported in [12], cardiac image registration is especially vital in improving heart chamber analysis accuracy,

correcting cardiac imaging errors, and guiding cardiac surgeries. Thus, several studies have explored classical [2,16] and deep-learning-based [3,10,14,20] registration methods over the years.

Classical registration methods [16] used hand-crafted features to align images by solving computational-expensive optimization problems. Recently, researchers explored the deep-learning-based unsupervised deformable image registration [3,14,19,20] to address the computational burden while reducing the need for accurate ground truth in the registration task. VoxelMorph [3], as the baseline, took moving and fixed image pairs as the input and maximized image pair similarity to train a registration network. To achieve higher accuracy, most unsupervised methods adopted a cascaded network with several sub-networks or an iterative refinement strategy [6,10,14,20]. These strategies made the training procedure complicated and computational resources demanding. Meanwhile, to obtain smoother and more realistic deformation fields, i.e., topology preservation, many existing works introduced explicit diffeomorphic constraints [7,17,19] or additional calculations on cycle consistency [14]. For example, CycleMorph [14] utilized the bidirectional registration consistency to preserve the topology during training. VoxelMorph-Diff [7] adopted velocity field-based deformation field and new diffeomorphic estimation. SYMNet [19] used symmetric deformation field estimation to achieve the goal. However, these schemes did not fully exploit the inherent network features, thereby overlooking these features' ability for better topology preservation.
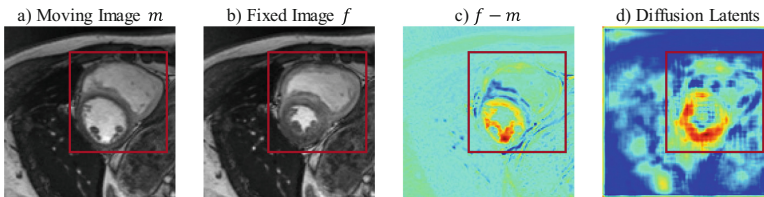

Fig. 1. $f - m$ intuitively indicates where significant deformation occurs, as the red-boxed area shows. We calculated the per voxel energy score in the diffusion model's latent obtained in [13] as [10] suggests to identify the areas where complex deformation is likely to happen. The result indicates the same area, which was not explicitly utilized in prior work [13]. (Color figure online)

Recently, Kim *et al.* [13] first proposed a diffusion model [9], which is simpler to train than other generative models yet rich in semantics, for the registration task. They used the latent feature from the diffusion model's score function, i.e., the gradient field of a distribution's log-likelihood function [22], as one of the registration network's inputs for a better registration result. However, this method only used the final diffusion score as an image level guidance, which *ignored diffusion model's rich task-specific semantics in the feature levels*, as proven in [4,18,23]. This resulted in the latent semantics of the diffusion model

not being able to directly guide the features learned at the hidden layers of the registration network. As a result, the informativeness of these features for image registration was reduced. Moreover, this method only preserved deformation topology by simply using the diffusion score as the input, thereby *ignoring the informative details about areas where significant deformations occur*; see Fig. 1d for unexploited informative semantics. Therefore, the registration network was unable to explicitly prioritize hard-to-register areas, thereby limiting its effectiveness in preserving the deformation topology.

To address these issues, we present two novel modules, namely **F**eature-wise **D**iffusion-**G**uided Module (**FDG**) and **S**core-wise **D**iffusion-**G**uided Module (**SDG**) in the registration network. FDG introduces a direct feature-wise diffusion guidance technique for generating deformation fields by utilizing cross-attention to integrate the intermediate features of the diffusion model into the hidden layer of the registration network's decoder. Furthermore, we embed the feature-wise guidance into multiple layers of the registration network and produce the feature-level deformation fields in multiple scales. Finally, after obtaining deformation fields at multiple scales, we upsample and average them to generate the full-resolution deformation field for registration. Our SDG introduces explicit score-wise diffusion guidance for deformation topology preservation by reweighing the similarity-based unsupervised registration loss based on the diffusion score. Through this reweighing scheme, direct attention is given during the optimization process to ensure the preservation of the deformation topology. Our main contribution can be summarized as follows:

– We propose a novel feature-wise diffusion-guided module (FDG), which utilizes multi-scale intermediate features from the diffusion model to effectively guide the registration network in generating deformation fields.
– We also propose a score-wise diffusion-guided module (SDG), which leverages the diffusion model's score function to guide deformation topology preservation during the optimization process without incurring any additional computational burden.
– Experimental results on the cardiac dataset validated the effectiveness of our proposed method.

2   Method

2.1   Baseline Registration Model

Figure 2a shows the overview of our proposed method. We first sample a perturbed noisy image x_t from the fixed target image f following the same scheme in [9], which can be formulated as Eq. 1:

$$x_t = \sqrt{\alpha_t} f + \sqrt{1 - \alpha_t}\epsilon,$$
$$\text{where } \alpha_t = \prod_{s=1}^{t}(1 - \beta_s), \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where $0 < \beta_s < 1$ is the variance of the noise, t is the noise level. Then we perform the registration training task. Given an input x_{in} consisting of a fixed reference image f, a moving unaligned image m, and the perturbed noisy image x_t, we feed this input $x_{in} = \{f, m, x_t\}$ into the registration network's shared encoder E_β, followed by the registration decoder R_θ. Then, the registration decoder R_θ outputs a deformation field ϕ, guided by our *Feature-wise Diffusion-Guided* module G_σ. Afterward, we feed m and ϕ into the spatial transformation layer (STL) to generate the warped image $m(\phi)$. Finally, by optimizing the similarity-based loss function $L_{scoreNCC}$ guided by our *Score-wise Diffusion-Guided* module, we can obtain the final registration model.
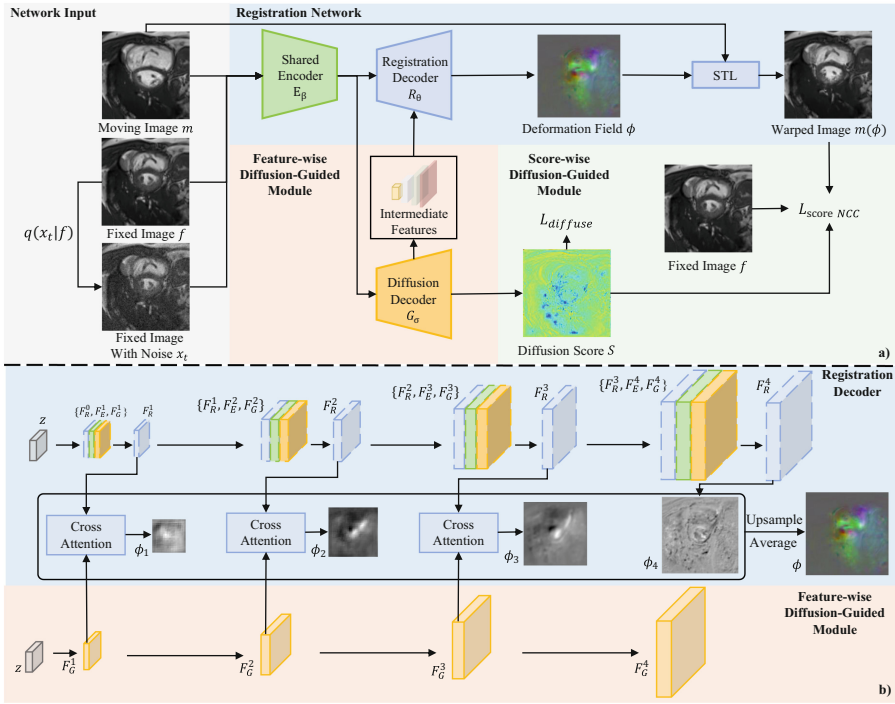

Fig. 2. a) The workflow of FSDiffReg. b) The illustration of the guidance process of the Feature-wise Diffusion-Guided Module.

2.2   Feature-Wise Diffusion-Guided Module

The main component of the Feature-wise Diffusion-Guided module (FDG) is an auxiliary denoising diffusion decoder G_σ. The workflow of FDG is shown in Fig. 2b. Given the input $x_{in} = \{f, m, x_t\}$, the UNet shared encoder E_β extracts

the representation z. z is then fed into the diffusion decoder G and the registration decoder R to get intermediate feature map pairs $F_i = \{(F_G^i, F_R^i)\}, i = 1, ..., N$ from the i-th layer of the decoder. Of note, we generate the registration decoder's feature map by incorporating the guidance from the diffusion decoder, which can be formulated as Eq. 2:

$$F_R^i = r_i(\text{concat}(F_R^{i-1}, F_E^i, F_G^i)), \text{ where } i = 1, ..., N \tag{2}$$

where r_i is the i-th layer of the registration decoder, and F_E^i is the skip connection of features from the shared encoder layer at the same depth.

After obtaining the feature map pairs, our FDG module estimates the i-th feature level deformation field ϕ_i from the feature map pair (F_G^i, F_R^i) using linear cross attention [21], which can be defined as Eq. 3:

$$\phi_i = \text{Conv}(\text{softmax}(F_R^i(\text{GroupNorm}(F_G^i)^T \cdot F_G^i)) + F_R^i) \tag{3}$$

After obtaining all feature-level deformation fields from the shallowest layer to the deepest layer, we generate the final deformation field ϕ by enlarging and averaging all feature-level deformation fields. This is a commonly adopted method for multi-scale deformation field merging so as to merge features which attend different scales and granularity. The final ϕ is then fed into the spatial transformation layer with the moving image m to generate the registered image $m(\phi)$.

2.3   Score-Wise Diffusion-Guided Module

Given the representation z encoded by the shared encoder $z = E_\beta(x_{in})$, the diffusion decoder G_σ outputs a diffusion score estimation $S = G_\sigma(z)$. Then, the Score-wise Diffusion-Guided Module (SDG) uses this score to reweigh the similarity-based normalized cross-correlation loss function, formulated as Eq. 4:

$$L_{scoreNCC}(m, f, S) = (\frac{1}{1 + e^{-S}})^\gamma \odot -(m(\phi) \otimes f) \tag{4}$$

where $m(\phi)$ is the warped moving image, $\odot$ defines the Hadamard product, and $\otimes$ defines the local normalized cross-correlation function. γ is a hyperparameter to amplify the reweighing effect.

By this means, SDG utilizes the diffusion score to explicitly indicate the hard-to-register areas, i.e., areas where deformation topology is hard to preserve, then assigning higher weights in the loss function for greater attention, and vice versa for easier-to-register areas. Therefore, the information on deformation topology is effectively incorporated into the optimization process without additional constraints by the SDG module.

2.4   Overall Training and Inference

Loss Function. Our network predicts the deformable fields at the feature level and then outputs the registered image. The total loss function of our method is defined as Eq. 5:

$$L_{total} = L_{diffusion}(x_{in}, t) + \lambda L_{scoreNCC}(m, f, S) + \lambda_\phi \sum ||\nabla_\phi||^2 \qquad (5)$$

$$L_{diffusion}(x_{in}, t) = \mathbb{E}_z ||G_\sigma(E_\beta(x_{in}, t)) - \epsilon||_2^2 \text{ ,where } \epsilon \sim \mathcal{N}(0, I) \qquad (6)$$

where $L_{diffusion}$ is the auxiliary loss function for training the diffusion decoder G_σ (Eq. 6), and t is the noise level of x_t, following the method in [9]. Our proposed $L_{scoreNCC}$ encourages maximizing the similarity between the registered and reference images while preserving the deformation topology. $\sum ||\nabla_\phi||^2$ is the conventional smoothness penalty on the deformation field. λ and λ_ϕ are hyperparameters, and we empirically set them to 20 in our experiments.

Inference. In the inference stage, we perform image registration in the same style as [13]. Instead of the perturbed image x_t, we input the original reference image f into the network, and the total network input becomes $x_{in} = \{f, m, f\}$. Given this network input x_{in}, our network first generates the deformation field ϕ between the moving image m and the reference image f and produces the registered moving image $m(\phi)$ by feeding the moving image m and the deformation field ϕ into the spatial transformation layer (STL). The registered moving image is the final output of our network.

3   Experiments and Results

Dataset and Preprocessing. Following the previous work [13], we used the publicly available 3D cardiac MR dataset ACDC [5] for experiments. The dataset includes 100 4D temporal cardiac MRI data with corresponding segmentation maps. We selected the 3D image at the end of the diastolic stage as the fixed image and the image at the end of the systolic stage as the moving image. We resampled all scans to the voxel spacing of $1.5 \times 1.5 \times 3.15$ mm, then cropped them to the voxel size of $128 \times 128 \times 32$. We normalized the intensity of all images to $[-1, 1]$. The training set contains 90 image pairs, while the remaining 10 pairs form the test set. The abovementioned preprocessing steps were performed in accordance with the approach described in prior work [13] to ensure a fair comparison.
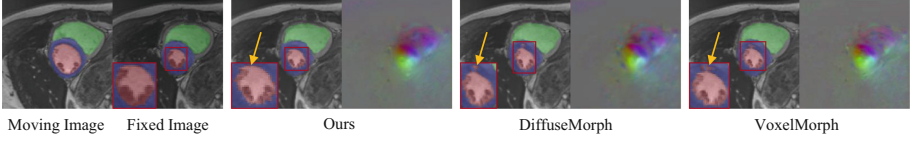
Moving Image   Fixed Image   Ours   DiffuseMorph   VoxelMorph

Fig. 3. The visualization of registration results. As the red-boxed area shows, our deformation field and the corresponding image are more refined in the area where larger deformation happens. (Color figure online)

Implementation Details. The proposed framework was implemented using the PyTorch library, version 1.12.0. Following [13], we used DDPM UNet's 3D encoder as our shared encoder and DDPM UNet's 3D decoder as our diffusion decoder. For the registration part, instead of a complete 3D UNet in [13], we only used DDPM UNet's 3D decoder as our registration decoder to generate the deformation field. During the diffusion task, we gradually increased the noise schedule from 10^{-6} to 10^{-2} over 2000 timesteps. We utilized an Nvidia RTX3090 GPU and the Adam optimization algorithm [15] to train the model with $\lambda = 20$, $\lambda_\phi = 20$, $\gamma = 1$, batch size $\mathbb{B} = 1$, a learning rate of 2×10^{-4}, and a maximum of 700 epochs.

Evaluation Metrics. We employed three evaluation metrics, i.e., DICE, $|J| \leq 0(\%)$, and $SD(|J|)$ to measure the image registration performance, following existing registration methods [3,13,14]. DICE measures the spatial overlap of anatomical segmentation maps between the warped moving image and the fixed reference image. A higher Dice score indicates better alignment between the warped moving image and the fixed reference image, thus reflecting an improved registration quality. $|J| \leq 0(\%)$ indicates the percentage of non-positive values in the Jacobian determinant of the registration field. This metric indicates the percentage of voxels that lacks a one-to-one registration mapping relation, causing unrealistic deformations and roughness. $SD(|J|)$ refers to the standard deviation of the Jacobian determinant of the registration field. A lower standard deviation indicates that the registration field is relatively smooth and consistent across the image.

Compare with the State-of-the-Art Methods. Table 1 shows the comparison of our method with existing state-of-the-art methods including Voxel-Morph [3], VoxelMorph-Diff [7], and DiffuseMorph [13] on the same training and testing dataset. We produced baseline results using the recommended hyperparameters in their paper. The result shows that our proposed method outperforms existing baseline methods by a substantial margin (Wilcoxon signed-rank test, $p < 0.005$) (Also see Fig. 3). Furthermore, our method aligned better in areas where larger deformation happened, such as myocardium (myo).

Table 1. Image registration results with standard deviation in parenthesis on the 3D cardiac dataset. "LV", "Myo", "RV" refers to Left Ventricle, Myocardium, and Right Ventricle, respectively. "Overall" refers to the averaged registration result of the left blood pool, myocardium, left ventricle, right ventricle, and these total region, following [13]. ↑: the higher, the better results. ↓: the lower, the better results.

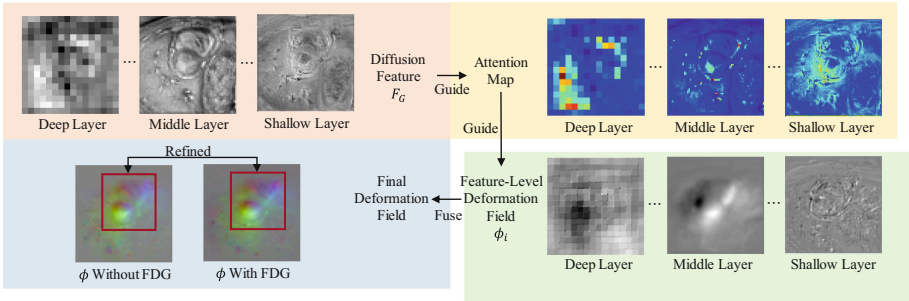| Method	DICE ↑				$|J| \leq 0(\%) \downarrow$	$SD(|J|) \downarrow$
	LV	Myo	RV	Overall		
Initial	0.585(0.074)	0.357(0.120)	0.741(0.069)	0.655(0.188)	–	–
VM [3]	0.770(0.086)	0.679(0.129)	0.816(0.065)	0.799(0.110)	0.079(0.058)	0.183
VM-Diff [7]	0.755(0.092)	0.659(0.137)	0.815(0.066)	0.789(0.117)	0.083(0.063)	0.182
DiffuseMorph [13]	0.783(0.086)	0.678(0.148)	0.821(0.067)	0.805(0.114)	0.061(0.038)	0.178
Ours	**0.809(0.077)**	**0.724(0.119)**	**0.827(0.061)**	**0.823(0.096)**	**0.054(0.026)**	**0.176**

Fig. 4. The visualization example of the effectiveness of FDG. The deformation field shows more details and is more distinct when guided by the FDG, thus improving registration accuracy, as the area in the red box shows. (Color figure online)

Ablation Study. To validate the effectiveness of our proposed learning strategies, including the Feature-wise Diffusion-Guided module(FDG) and Score-wise Diffusion-Guided module(SDG), we conducted ablative experiments, as shown in Table 2. The network without FDG also uses the denoising diffusion decoder but generates the deformation field from the encoded feature directly, and without SDG means that we optimize the network using the vanilla NCC loss. By integrating multi-scale intermediate latent diffusion features into generating deformation fields, we can see that the network's performance increased by 1%. By deploying the reweighing loss, the Jacobian metric decreased by 60.5%. The result achieved a balance when all components were deployed. These results demonstrated that our proposed components could effectively guide the deformation field generation by using multi-scale diffusion features (Also see Fig. 4). Optimization guided by diffusion score led to better preservation of deformation topology. It is worth noticing that the results without FDG or SDG showed only marginal improvement over baseline results, indicating the importance of feature-level deformation field generation and the reweighing scheme. The ablative study of hyperparameter λ is illustrated in Supp. Fig. 1.

Table 2. Ablation study on FDG and SDG.

FDG	SDG	DICE↑	$|J| \leq 0(\%) \downarrow$
		0.811(0.098)	0.114(0.076)
✓		0.818(0.102)	0.062(0.038)
	✓	0.810(0.188)	0.045(0.025)
✓	✓	0.823(0.096)	0.054(0.026)

Table 3. Ablation Study on hyperparameter γ in SDG.

γ	0.5	1	2
DICE↑	0.817(0.100)	0.823(0.096)	0.816(0.104)
$|J| \leq 0(\%) \downarrow$	0.069(0.029)	0.054(0.026)	0.042(0.023)

Analysis of γ. Furthermore, to validate SDG's effectiveness on topology preservation, we conducted another ablative study on SDG's hyperparameter γ, as Table 3 shows. Increased γ indicates a more substantial reweighing effect. The results showed that by adding stronger reweighing influence, we could obtain deformation fields with better topology preservation almost without compromising accuracy.

4   Conclusion

This work proposes two novel modules for unsupervised deformable image registration: the Feature-wise Diffusion-Guided module (FDG) and the Score-wise Diffusion-Guided module (SDG). Among these modules, FDG can effectively guide the deformation field generation by utilizing the multi-scale intermediate diffusion features. SDG demonstrates its ability to guide the optimization process for better deformation topology preservation using the diffusion score. Extensive experiments show that the proposed framework brings impressive improvements over all baselines. The proposed work models the non-linear deformation semantics using the diffusion model. Therefore, it is sound to generalize to other registration tasks and images, which may be one of the future research directions.

References

1. Alam, F., Rahman, S.U., Ullah, S., Gulati, K.: Medical image registration in image guided surgery: issues, challenges and research opportunities. Biocybern. Biomed. Eng. **38**(1), 71–89 (2018)
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
3. 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)

4. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=SlxSY2UZQT

5. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)

6. Che, T., et al.: AMNet: adaptive multi-level network for deformable registration of 3D brain MR images. Med. Image Anal. **85**, 102740 (2023)

7. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. **57**, 226–236 (2019)

8. Giger, M.L., Karssemeijer, N., Schnabel, J.A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu. Rev. Biomed. Eng. **15**, 327–357 (2013)

9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)

10. Huang, Y., Ahmad, S., Fan, J., Shen, D., Yap, P.T.: Difficulty-aware hierarchical convolutional neural networks for deformable registration of brain MR images. Med. Image Anal. **67**, 101817 (2021)

11. Jain, M., Rai, C., Jain, J., Gambhir, D.: Amalgamation of machine learning and slice-by-slice registration of MRI for early prognosis of cognitive decline. Int. J. Adv. Comput. Sci. Appl. **12**(1) (2021)

12. Khalil, A., Ng, S.C., Liew, Y.M., Lai, K.W.: An overview on image registration techniques for cardiac diagnosis and treatment. Cardiol. Res. Pract. **2018** (2018)

13. Kim, B., Han, I., Ye, J.C.: DiffuseMorph: unsupervised deformable image registration using diffusion model. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13691, pp. 347–364. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19821-2_20

14. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)

15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

16. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)

17. Krebs, J., Mansi, T., Mailhé, B., Ayache, N., Delingette, H.: Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS 2018. LNCS, vol. 11045, pp. 101–109. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_12

18. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=pd1P2eUBVfq

19. Mok, T.C., Chung, A.: 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 (2020)

20. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with Laplacian pyramid networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21

21. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: attention with linear complexities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3531–3539 (2021)
22. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
23. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. arXiv preprint arXiv:2211.12572 (2022)