DiffuseMorph: Unsupervised Deformable Image Registration Along Continuous Trajectory Using Diffusion Models

Boah Kim, Inhwa Han, Jong Chul Ye

Department of Bio and Brain Engineering
Program of Brain and Cognitive Engineering
Kim Jaechul Graduate School of AI
Department of Mathematical Sciences
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
{boahkim, inhwahan, jong.ye}@kaist.ac.kr

Abstract

Deformable image registration is one of the fundamental tasks for medical imaging and computer vision. Classical registration algorithms usually rely on iterative optimization approaches to provide accurate deformation, which requires high computational cost. Although many deep-learning-based methods have been developed to carry out fast image registration, it is still challenging to estimate the deformation field with less topological folding problem. Furthermore, these approaches only enable registration to a single fixed image, and it is not possible to obtain continuously varying registration results between the moving and fixed images. To address this, here we present a novel approach of diffusion model-based probabilistic image registration, called DiffuseMorph. Specifically, our model learns the score function of the deformation between moving and fixed images. Similar to the existing diffusion models, DiffuseMorph not only provides synthetic deformed images through a reverse diffusion process, but also enables various levels of deformation of the moving image along with the latent space. Experimental results on 2D face expression image and 3D brain image registration tasks demonstrate that our method can provide flexible and accurate deformation with a capability of topology preservation.

1. Introduction

Deformable image registration is to estimate voxel correspondences between moving and fixed image pairs. This is especially important for medical image analysis such as disease diagnosis and treatment monitoring, since the anatomical structures or shapes of medical images are different according to subjects, scanning time, imaging modality, etc. Accordingly, various image registration methods have been studied over the past decades, and these classical approaches usually attempt to align images by solving a computationally expensive optimization problem [1, 2, 24].

To address this computational issue, deep-learning-based image registration methods have been extensively studied lately [4, 7, 22, 28, 31], which train neural networks
to estimate the deformation field by taking the moving and fixed images as network inputs. These approaches provide fast deformation while maintaining registration accuracy. However, they usually require the ground-truth deformation fields for supervised training [31, 33], complicated additional diffeomorphic integration layers [10, 25], or the cycle-consistency [22] for topology preservation.

Recently, the denoising diffusion probabilistic model (DDPM) [18, 36] has shown impressive performance in many areas of computer vision [8, 15, 21, 35, 37]. DDPM learns the Markov transformation from Gaussian noise distribution to data distribution and provides diverse samples through stochastic diffusion processes. To generate images with desired semantics, conditional denoising diffusion models have been also presented [9, 34]. However, it is challenging to apply DDPM to image registration tasks, since the existing methods may generate samples with losing image identity, and a correct registration should be performed by the deformation field for the moving image rather than image generation.

In this paper, we present a new unsupervised image registration approach, dubbed DiffuseMorph, which adapts DDPM to generate deformed images as well as to allow deformation field-based registration along the continuous trajectory between moving and fixed images. Specifically, our proposed model is composed of a diffusion network and a deformation network: the former network learns the score function of the deformation between moving and fixed images, and the latter network estimates the deformation field using the score and provides deformed images. These two models are jointly trained in an end-to-end learning manner, which enables DiffuseMorph not only to measure Markov transformation in the direction that the moving image is deformed, but also to estimate the registration field for the moving image to be warped into the fixed image. Thanks to this end-to-end learning, the proposed method performs effective image registration at the inference phase.

Specifically, as shown in Fig. 1, once the networks are trained, our model can provide the intermediate deformations of the moving image to the fixed image by simply interpolating the score values that are used as an input for the deformation network. Also, the proposed DiffuseMorph can quickly generate synthetic deformed moving images similar to the fixed images. Specifically, instead of starting from a random Gaussian noise, we perform one step forward diffusion from a moving image, which is then iteratively refined using the reverse diffusion of DDPM through the diffusion network. This allows the sample to retain the original moving image identity and reduce the number of diffusion steps significantly.

To demonstrate performance of the proposed method, we apply our model to 2D face image registration and 3D brain MR image registration tasks. The experimental results verify that our model achieves high performance in deformation accuracy. Also, even though our model does not have a constraint for diffeomorphic registration, the results show topology preservation with smooth deformation fields. Furthermore, our method enables deformable registration along the continuous trajectory between the moving and fixed images, which was not possible with any existing image registration methods. Our contributions are summarized as:

- We propose DiffuseMorph, the first image registration method employing the denoising diffusion model conditioned on a pair of moving and fixed images.
- Thanks to the end-to-end learning of the diffusion and deformation networks, our model not only generates deformed images through the fast reverse diffusion, but also enables image registration along the continuous trajectory between the moving and fixed images.
- We verify that the proposed method can be applied to 2D and 3D image registration tasks and provide accurate deformation with topology preservation.

2. Background and related works

2.1. Deformable image registration

Given a moving image $m$ and a fixed image $f$, classical deformable image registration methods are performed by solving the following optimization problem:

$$
\phi^* = \arg\min_{\phi} \ L_{\text{sim}}(m(\phi), f) + L_{\text{reg}}(\phi),
$$

where $\phi^*$ is the optimal registration field to deform the moving image into the fixed image. $L_{\text{sim}}$ is the dissimilarity function to compute the similarity between the deformed and fixed images, and $L_{\text{reg}}$ is the regularization penalty of the deformation field. By minimizing the energy function, the deformed image $m(\phi)$ is estimated by warping the moving image. In particular, diffeomorphic registration can be achieved when imposing additional constraints on the fields $\phi$ such that the deformation mapping is differentiable and invertible, and provides topology preservation [2, 5, 41].

Learning-based methods As the traditional registration approaches usually require large computation and long runtime, deep learning methods have been extensively studied lately, which can estimate deformation in real time once a neural network is trained. For example, supervised learning methods train the networks with the ground-truth registration fields [6, 7, 33, 43], but they are not practical since high-quality labels of deformation fields are required for training. In contrast, unsupervised learning approaches train networks by computing similarity between the deformed image.
and the fixed reference image \([4,12,22,28,29]\). To guarantee topology preservation, diffeomorphic registration methods with probabilistic parameterization of deformation are proposed \([10,11,25]\). However, the diffeomorphic constraints are needed also for the inference, adding additional complexity.

### 2.2. Diffusion probabilistic model

Recently, denoising diffusion probabilistic model (DDPM) \([18,36]\) is presented as a generative model, which learns a Markov chain process to transit the simple Gaussian distribution to the data distribution. In the forward step, noises are gradually added to the data \(x_0\) using a Markov chain, in which each step of sampling latent variables \(x_t, t \in [0, T]\) is represented as a Gaussian transition:

\[
q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \quad (2)
\]

where \(0 < \beta_t < 1\) is a variance of the noise. The resulting distribution of \(x_t\) given \(x_0\) is then expressed as:

\[
q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1 - \alpha_t) I), \quad (3)
\]

where \(\alpha_t = \prod_{s=1}^{t}(1 - \beta_s)\). Accordingly, given \(\epsilon \sim \mathcal{N}(0, I)\), \(x_t\) can be sampled as:

\[
x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon. \quad (4)
\]

For the generative process to perform the reverse diffusion, DDPM learns parameterized Gaussian process \(p_\theta(x_{t-1}|x_t)\), which is represented as:

\[
p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta^2 I), \quad (5)
\]

where \(\sigma_t\) is a fixed variance, and \(\mu_\theta(x_t, t)\) is a denoised image defined as:

\[
\mu_\theta(x_t, t) = \frac{1}{\sqrt{1 - \beta_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t) \right), \quad (6)
\]

where a parameterized model \(\epsilon_\theta\) is estimated during the training phase. In fact, the model \(\epsilon_\theta(x_t, t)\) is just a scaled version of the score function \(s_\theta(x_t, t)\) \([38]\), which is the gradient of the \(\log p_\theta(x_t)\). Once the model \(\epsilon_\theta\) is trained, the data is sampled by the following stochastic generation step: \(x_{t-1} = \mu_\theta(x_t, t) + \sqrt{\beta_t}z\), where \(z \sim \mathcal{N}(0, I)\).

### Conditional generative methods

Since we are interested in image registration that is performed given the moving and reference images, here we review the existing conditional diffusion-based models for image generation \([9,19,30,34,37,39]\). To control the image generation by the reverse process, DDIM \([37]\) proposes a deterministic non-Markovian generative process starting from an initial condition. SR3 \([34]\) was proposed for super-resolution by training DDPM with a conditioned image. ILVR \([9]\) presents conditioning iterative generative process using an unconditional model. However, the existing methods do not deform the shape of the conditioned image or may transform the image identity, which should not happen in the image registration task. In contrast, under the condition of a pair of moving and fixed images, our diffusion model produces deformation of a moving image aligned with the fixed image without losing its identity.
3. Proposed method

3.1. Framework of DiffuseMorph

By leveraging the capability of DDPM, we aim to develop a novel diffusion model-based unsupervised deformation approach. Since the image registration task is to deform images using deformation fields, we design our model with two networks as illustrated in Fig. 2: one is a diffusion model $G_θ$ to estimate a score function of the deformation under the condition of moving and fixed images, and the other is a deformation model $M_ψ$ that actually outputs registration fields using the score function.

Specifically, for the moving source image $m$ and the fixed reference image $f$, the diffusion network $G_θ$ is trained to learn the score function of the deformation between the moving and fixed images under the condition $c = (m, f)$. For this, we sample the latent variable $x$ of the target by Eq. (4), considering the fixed image as the target, i.e. $x_0 = f$. Moreover, to make the network $G_θ$ aware of the level of noise, we directly give the adequate statistics for the variance of the noise $α$ to the network, similar to [8, 34].

On the other hand, the deformation network $M_ψ$ takes the score output of the diffusion network, $ε$, as well as the moving source image $m$, and estimates the deformation field $φ$. Then the deformed image $m(ϕ)$ is generated by warping the moving image $m$ using the spatial transformation layer [20]. To deform 2D/3D images in our experiments, we adopt the transformation function using bi-/trilinear interpolation.

**Loss function** Recall that the diffusion network $G_θ$ and the deformation network $M_ψ$ are jointly trained by end-to-end learning. Thus, for the training of our model, we design the objective function as follows:

$$\min_{θ, ψ} L_{\text{diffusion}} + λL_{\text{regist}}, \quad (7)$$

where $L_{\text{diffusion}}$ and $L_{\text{regist}}$ are the diffusion loss and the registration loss, respectively, and $λ$ is a hyper-parameter.

Specifically, the diffusion loss is given by:

$$L_{\text{diffusion}}(c, x, α) = \mathbb{E}_{ε, x, α}||G_θ(c, x, α) - ε||_2^2, \quad (8)$$

where $ε \sim \mathcal{N}(0, I)$, and $α \sim U(α_t-1, α_t)$ where $t \in [0, T_{\text{train}}]$ is a uniformly sampled time step. Also, the registration loss is to estimate the deformation field so that the deformed image has similar shape of the fixed image. Based on the traditional energy function in Eq. (1), we design the registration loss as follows:

$$L_{\text{regist}}(m, f) = -(m(φ) \otimes f) + λ_R \sum ||\nabla φ||^2, \quad (9)$$

where $φ = M_ψ(m, ˆε)$ with $ˆε$ referring to the diffusion network output, and $λ_R$ is a hyper-parameter. The first term of Eq. (9) is the local normalized cross-correlation [4] for the dissimilarity metric, and the second term is the smoothness penalty on the deformation fields. We set $λ_R = 1$.

It is remarkable that the net effect of the two loss functions is that $G_θ$ is trained to learn the score function of the deformation. Accordingly, when combined with the reverse diffusion, the score guides the reverse diffusion to generate the fixed image from the moving image initialization. Furthermore, the score helps the deformation network to generate deformation field toward the fixed image.

3.2. Image registration using DiffuseMorph

Once the networks of the proposed model are trained, at the inference phase, they provide image registration by estimating deformation field for the moving image to be warped into the fixed image. Thanks to the end-to-end training of our model, the diffusion model guides the deformation network to generate regular deformation field. Specifically, using the learned parameters of $G_θ^∗$ and $M_ψ^∗$, the deformation field $ϕ$ is estimated by:

$$ϕ = M_ψ^∗(m, G_θ^∗(c, x_0, α_0)), \quad (10)$$

where $x_0$ is set to the fixed target image $f$. The deformed image $m(ϕ)$ is computed using the estimated $ϕ$ through the spatial transformation. Therefore, our model performs image registration at a single step with smooth registration field with less folding problem.

Image registration along continuous trajectory Since the diffusion network estimates the score function of the
deformation, our model also provides the continuous deformation of the moving image along the trajectory toward the fixed image, which was not possible with existing registration methods. Specifically, if the score input to the deformation network is set to zero, the network produces the deformation field that hardly deforms the moving image. Accordingly, for a given network output $\hat{\epsilon}_f = G\theta^*(c, f, \alpha_0)$, the registration field $\phi_\eta$ that deforms the moving image along the continuous trajectory toward the fixed image can be generated by simply interpolating the score values:

$$\phi_\eta = M_{\psi^*}(m, \hat{\epsilon}_\eta), \quad \hat{\epsilon}_\eta := \eta \hat{\epsilon}_f,$$

for $0 \leq \eta \leq 1$. This interesting phenomenon occurs thanks to learning the score function of deformation, as will be confirmed later in our experiments.

**Image deformation via generative process** In addition to the continuous image registration, the learnt score function of deformation can help generate synthetic deformed images through the reverse diffusion process. Here, we describe a method of the deformed image generation only using the diffusion network $G\theta$.

Specifically, as the score function learns the deformation of the moving image toward the fixed image, our image generation starts from the moving image, in contrast to the existing conditional generative process of DDPM [9, 34] that starts from the pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$. When we set the initial state with the original moving image $m$, the one-step forward diffusion is performed by:

$$x_T = \sqrt{\alpha_T} m + \sqrt{1 - \alpha_T} \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, I)$, and $\alpha_T$ is the noise level at a time step $T \leq T_{train}$. Here, the time step $T$ is set to a value not to lose the identity of the image. This forward sampling allows the moving image distribution to be close to the fixed image distribution, as illustrated in Fig. 3, which reduces the number of reverse diffusion steps and generation time.

Then, by starting from $x_T$, the generation of deformed image $x_0$ that fits into the fixed image $f$ is performed by the following reverse diffusion process from $t = T$ to $t = 1$:

$$x_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} (x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} G\theta^*(c, x_t, \alpha_t)) + \sqrt{\beta_t} \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, I)$. Here, in choosing the total sampling steps $T$, we employ [8] that presents a more efficient inference method than DDPM. Due to the directly conditioning our denoising model on the noise level $\alpha_t$, one can flexibly set the number of sampling steps. In our experiments, we set the 200 sampling steps in maximum. The pseudocode of this generative process of DiffuseMorph is described in Algorithm 1.
4. Experiments

To demonstrate that DiffuseMorph generates deformed images from the moving to the fixed images using high-quality registration fields, we apply our method to two applications. We first conduct the experiment on the deformation of 2D facial expression images. Also, we apply the proposed method for 3D brain MR registration, in which individual brain images are deformed to a common atlas. The datasets and training details are as follows, and more details are described in Appendix.

4.1. Datasets

Facial expression image For 2D facial expression image registration, Radboud Faces Database (RaFD) [27] was used. It contains eight different facial expressions collected from 67 subjects; neutral, angry, contemptuous, disgusted, fearful, happy, sad, and surprised. For each facial expression, three different gaze directions are provided, and thus there are total 1608 images. We cropped the images to $640 \times 640$, resized them into $128 \times 128$, and converted the RGB images to gray scale.

Brain MRI We conducted the experiment of 3D brain MR registration using OASIS-3 dataset [26] which provides brain MR images and corresponding volumetric segmentation maps from FreeSurfer [16]. We used 1156 T1-weighted scans that preprocessed by image resampling to $256 \times 256 \times 256$ grid with $1mm^3$ isotropic voxels, affine spatial normalization, and brain extraction. The images were cropped by $160 \times 192 \times 224$. We used 1027 and 129 scans for training and test, respectively.

4.2. Implementation details

The proposed model was implemented by employing the network architectures of DDPM [18] and VoxelMorph [4].

| Method  | NMSE $\times 10^{-1}$ | SSIM | $|J_\phi| \leq 0$ (%) |
|---------|-----------------------|------|----------------------|
| Initial | 0.363 (0.268)         | 0.823 (0.066) | 0                    |
| VM      | 0.047 (0.057)         | 0.936 (0.024) | 0.050 (0.106)        |
| VM-diff | 0.033 (0.018)         | 0.957 (0.132) | 0.019 (0.068)        |
| Ours    | 0.032 (0.017)         | 0.964 (0.011) | 0.017 (0.056)        |

Table 1. Quantitative evaluation results of the facial expression image registration. Standard deviations are shown in parentheses.
Figure 7. Results of image registration on the brain MR images (left) using the estimated deformation fields (right). Segmentation maps of several anatomical structures are overlaid with dark target contours and light source/deformed contours (blue: ventricles, green: thalami, orange: third ventricle, pink: hippocampi). The Dice scores for each structure are displayed with the corresponding colors on each result.

Here, we configured layers of the networks according to the dimension of the image, e.g., 2D conv layer for 2D image registration. Using a single Nvidia Quadro RTX 6000 GPU, we trained our model by Adam optimization algorithm [23].

Specifically, for the face image registration, we set the hyper-parameter as $\lambda = 2$, and trained the model with the learning rate $5 \times 10^{-6}$ for 40 epochs. On the other hand, for brain MR image registration, we set the hyper-parameter as $\lambda = 10$ and trained the networks with the learning rate $1 \times 10^{-4}$ for 60 epochs. The proposed method was implemented in Python using PyTorch library.

Evaluation To evaluate the registration performance, we computed the percentage of non-positive values of Jacobian determinant on the deformation fields ($|J_\phi| \leq 0$), which indicates that one-to-one mapping of the registration has been lost. Also, for the facial images that have labels for all deformed images, we measured NMSE and SSIM between deformed and fixed target images. For the brain MR images, we applied the deformation fields to the segmentation maps of original moving images, and computed Dice score on 30 anatomical structures. When compared to the existing learning-based models, we used the same deformation network architecture for fair comparisons.

5. Results

5.1. 2D facial expression image registration

Fig. 4 shows the results of our DiffuseMorph on the facial expression image registration according to the variable $\eta$. We can see that the image is slowly moving toward the fixed image as $\eta$ increases. Moreover, the resulting deformation field is not a simple scaled version of the final deformation field. Rather, the registration field changes non-uniformly depending on the importance of variations at the intermediate deformation level. This verifies that our model provides continuous deformation between the moving source image and the fixed target image. Also, this confirms the importance of the score as the latent variable for deformation.

Furthermore, we compared our model to VoxelMorph (VM) [4] and VM-diff [10] in terms of deformation performance toward the fixed image. Here, we obtained deformed images with RGB channels by applying the same deformation fields of gray-scale images to each RGB channel. Fig. 5 shows visual comparisons of the registration results. Compared to VM, our model deforms the source image to be more accurately aligned with the target image. Also, Tab. 1 shows the quantitative evaluation results. When comparing our model to the initial source image and the other comparative methods, our model achieves lower NMSE and higher SSIM. Moreover, the metric of Jacobian determinant on deformation fields also decreases in the proposed method. These results indicate that the proposed DiffuseMorph provides high-quality image registration with less folding problem. More results of facial expression images are shown in Appendix.

Generation of synthetic deformed image To verify the capability of image generation from our DiffuseMorph, we performed the study on generative process using the facial images. Since the facial data were divided into the train-
Figure 8. Quantitative evaluation results of Dice scores on 30 brain anatomical structures in the brain MR image registration.

| Method      | Dice   | $|J_\phi| \leq 0 \%$ | Time (min) |
|-------------|--------|---------------------|------------|
| Initial     | 0.616 (0.171) | 0                   | 0          |
| SyN         | 0.752 (0.140) | 0.400 (0.100)       | 122, CPU   |
| VM          | 0.749 (0.145) | 0.553 (0.075)       | 0.01, GPU  |
| VM-diff     | 0.731 (0.139) | 0.631 (0.073)       | 0.01, GPU  |
| MSDIRNet    | 0.751 (0.142) | 0.804 (0.089)       | 2.06, GPU  |
| CM          | 0.750 (0.144) | 0.510 (0.087)       | 0.01, GPU  |
| Ours        | 0.756 (0.139) | 0.505 (0.058)       | 0.01, GPU  |

Table 2. Quantitative evaluation results of the brain MR image registration. Standard deviations are shown in parentheses.

| Method | Dice   | $|J_\phi| \leq 0 \%$ |
|--------|--------|-----------------|
| $\lambda = 2$ | 0.736 (0.152) | 0.498 (0.098) |
| $\lambda = 5$ | 0.746 (0.143) | 0.499 (0.070) |
| $\lambda = 10$ | 0.756 (0.139) | 0.505 (0.058) |

Table 3. Result of study on the effect of registration loss. Standard deviations are shown in parentheses.

5.2. 3D brain MR image registration

We evaluated the brain MRI registration results with the comparative methods: SyN [2], VM [4], VM-diff [10], MSDIRNet [28], and CM [22]. Fig. 7 shows the visual registration results. We displayed the results of all methods in Appendix. The results show that our model estimates smooth deformation fields and provides more accurate deformed moving images aligned to the fixed images, which can be also observed through contours of the segmentation maps of several brain anatomical structures. Fig. 8 and Tab. 2 also report that the proposed method achieves higher Dice scores with less non-positive Jacobian determinant over the existing learning-based methods.

Study on the effect of registration loss To analyze the effect of the registration loss in the training of our model on the performance of our model, we performed a comparison study with varying the value of $\lambda$ in Eq. (7) using brain MR data. As reported in Tab. 3, when $\lambda$ is lower, the Dice score was decreased but generating better regularity of deformation fields. This indicates that the registration loss forces the model to output more smooth deformation fields, and we need to balance between the trade-off.

6. Limitation

In this paper, we only verified the capability of Diffuse-Morph for the image generation only on 2D images, since the existing diffusion models are usually tested on 2D image generation. Future work should model the proper architectures for 3D image generation. Nevertheless, in the viewpoint of image registration, by jointly training the diffusion and deformation networks to learn the score function of deformation, our model provides very flexible tools for deformable image registration and generation.
Potential negative social impact  While the registration task is importantly used in medical image fields, the applications of our approach may be used to transform the identity of images. It could be extended for malicious use such as deepfake. Abuse of the image registration method can cause social problems such as copyright infringement, and privacy concerns. To mitigate such risks, social and privacy regulations have to be enforced.

7. Conclusion

We present a novel DiffuseMorph architecture for unsupervised image registration by employing the diffusion probabilistic model that is jointly trained with the deformation network. Thanks to the capability of learning the score function of deformation, the proposed method not only generates synthetic deformed images, but also provides high-quality registration by estimating deformation fields along the continuous trajectory toward the fixed image. We believe that DiffuseMorph can be a promising algorithm.

References

[1] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[2] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[3] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

[5] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

[6] Xiaohuan Cao, Jianhua Yang, Li Wang, Zhong Xue, Qian Wang, and Dinggang Shen. Deep learning based inter-modality image registration supervised by intramodality similarity. In *International workshop on machine learning in medical imaging*, pages 55–63. Springer, 2018.

[7] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2017.

[8] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

[9] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[10] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[11] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.

[12] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[13] P Ekman, WV Friesen, and JC Hager. Facial action coding system: The manual, 2002.

[14] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[15] Shreyas Fadnavis, Joshua Batson, and Eleftherios Garyfallidis. Patch2self: denoising diffusion mri with self-supervised learning. *arXiv preprint arXiv:2011.01355*, 2020.

[16] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.
[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. arXiv preprint arXiv:2106.15282, 2021. 3
[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. Advances in neural information processing systems, 28:2017–2025, 2015. 4
[21] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. arXiv preprint arXiv:2104.01409, 2021. 2
[22] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclesmooth: Cycle consistent unsupervised deformable image registration. Medical Image Analysis, 71:102036, 2021. 1, 2, 3, 8, 14
[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. 7
[24] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging, 29(1):196–205, 2009. 1
[25] Julian Krebs, Tommaso Mansi, Boris Mailhé, Nicholas Ayache, and Hervé Delingette. Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 101–109. Springer, 2018. 2, 3
[26] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv, 2019. 6, 12
[27] Oliver Langner, Ron Dotsch, Gjersert Bijslstra, Daniel HJ Wibboldus, Skyley T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. Cognition and emotion, 24(8):1377–1388, 2010. 6, 12
[28] Yang Lei, Yabo Fu, Tonghe Wang, Yingzi Liu, Pretesh Patel, Walter J Curran, Tian Liu, and Xiaofeng Yang. 4d-ct deformable image registration using multiscale unsupervised deep learning. Physics in Medicine & Biology, 65(8):085003, 2020. 1, 3, 8, 14
[29] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 1449–1453. IEEE, 2018. 3
[30] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. arXiv preprint arXiv:2102.09672, 2021. 3
[31] John A Onofrey, Lawrence H Staib, and Xenophon Papademetris. Semi-supervised learning of nonrigid deformations for image registration. In International MICCAI Workshop on Medical Computer Vision, pages 13–23. Springer, 2013. 1, 2
[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 12
[33] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In International conference on medical image computing and computer-assisted intervention, pages 266–274. Springer, 2017. 2
[34] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636, 2021. 2, 3, 4, 5, 12
[35] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358, 2021. 2
[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265. PMLR, 2015. 2, 3
[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020. 2, 3
[38] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. arXiv preprint arXiv:1907.05600, 2019. 3
[39] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 12

[41] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009. 2

[42] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 12

[43] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration—a deep learning approach. *NeuroImage*, 158:378–396, 2017. 2
Appendix

A. Details of datasets

For the experiments of image registration tasks in the main paper, we used the following two public datasets.

A.1. Datasets

RaFD dataset The Radboud Faces Database [27] contains photos of 67 subjects including both Caucasian adults and children. For each subject, 8 different emotional expressions with 3 gaze directions under a controlled environment are provided. All emotional expressions are based on prototypes from the Facial Action Coding System [13]. This dataset is publicly available for non-profit scientific research, according to the license agreement on its official website. Also, the dataset is the property of the Behavioural Science Institute of the Radboud University Nijmegen. Although it is unclear that the provided data is under IRB approval, the RaFD dataset is broadly used for non-commercial scientific study, and is also useful for the facial expression registration task. The dataset can be downloaded at the URL: www.rafd.nl.

OASIS-3 dataset The OASIS-3 dataset [26] is provided by the Open Access Series of Imaging Studies (OASIS) that has brain MR imaging data scanning from subjects of humans aged from 42 to 95. The dataset also contains volumetric segmentation maps on brain anatomical structures. It can be used under the OASIS Data Use Terms (DUT), which has the license of Creative Commons Attribution 4.0. The participated subjects consented to Knight ADRC-related projects that follow the procedures approved by the Institutional Review Board (IRB) of Washington University School of Medicine, thus, Protected Health Information (PHI) of all data has been excluded. The data can be downloaded at https://www.oasis-brains.org.

B. Details of implementation

B.1. Codes

We implemented our proposed DiffuseMorph based on SR3 [34] that provides the public code of conditional diffusion denoising model. Also, by employing the network architecture of VoxelMorph-1 [4], we trained and tested the proposed method for image registration, which is implemented in Python using Pytorch 1.4.0 library [32]. The source code to reproduce our model is available at https://github.com/DiffuseMorph/DiffuseMorph.

B.2. Network architecture

Diffusion network As we described in the paper, we employed the network architecture of DDPM [18] for the diffusion network $G_θ$ of the proposed method. Fig. 9 illustrates the structure of the diffusion network. Specifically, it consists of four encoder and four decoder blocks. Each block has the Resnet block composed of the group normalization [42], swish function [14], and convolution layers, which takes the embedded input of the noise level $α$ as well as the output of the preceding layer. At the last encoder block, the feature maps are attended by the self-attention module [40]. Also, each decoder block additionally takes the feature maps of encoder outputs by skip-connection that enables the decoder to use the encoding information of inputs. Accordingly, when the moving image $m$, fixed image $f$, and perturbed target image $x$ are given to the diffusion network with the noise level $α$, the network estimates the score function $ϵ$ of the deformation. The details such as the number of convolution channels of each block are described in Tab. 4. Here, we configured the kernel dimension of convolution layers according to the input image dimension. Also, we set the number of channels for 3D diffusion network as deep as possible up to the GPU memory.

Deformation network For the deformation network $M_ψ$, we implemented VoxelMorph-1 [4] that presents the network architecture for image registration. As illustrated in Fig. 10, the deformation network is a U-shape network with encoder and decoder blocks, similar to the diffusion network. However, instead of the Resnet block, each block has the convolution and leakyReLU [17] layers, called CL units, which enables the network to focus on image processing and learn feature extraction. Also, as the downsampling by the convolution is with stride 2, the upsampling is performed by the transposed convolution with stride 2. The final network output is generated by the convolution with stride 1. Thus, given the moving image $m$ and the output of the diffusion network $ϵ$, the deformation network estimates the registration field $ϕ$ that warps the moving image into the fixed image. Tab. 5 shows the detail of the deformation network architecture. Similar to the diffusion network, the convolution layers are configured depending on the input image dimension.

B.3. Data processing

The intensity range of grayscale facial data and preprocessed brain MR data used in the experiments of the main paper is [0, 1]. We scaled this intensity range of the data into [-1, 1]. Then, the noisy target is sampled using the scaled image and given to our model as an input, along with a condition of the moving and fixed images. Here, since the moving image is deformed using the deformation field through the spatial transformation with linear interpolation, we rescaled the moving image into [0, 1] just before warping the moving image. For data augmentation of the facial data, we used random horizontal flipping. On the other
Figure 9. Architecture of the diffusion network $G_\theta$. It is a U-net like architecture with the Resnet block (pink arrow) composed of the group normalization, swish function, and convolution layers. The downsampling in the encoder is performed by the convolution layer with stride 2, whereas the upsampling in the decoder is done by nearest interpolation with scale factor 2, followed by the convolution layer with stride 1. Each feature map of the encoder is given to the decoder by the skip-connection.

Figure 10. Architecture of the deformation network $M_\psi$. It is a modified U-net structure with the convolution layer and leakyReLU activation function. The image features are downsampled by the convolution with stride 2 (black arrow), while they are upsampled by the transposed convolution with stride 2 (blue arrow). The feature maps of each encoder block are concatenated to those of the corresponding decoder block.
leakyReLU (denoted as $L$). Also, $D$ is the convolution with 3D kernel and stride 1, where $d$ is the dimension of input. $R$ is the convolution with 3D kernel and stride 2, and $U$ is the nearest interpolation with scale factor 2, followed by $C$. $RB$ is the Resnet block, and SA is the self-attention block. $\text{GN}(n)$ denotes the group normalization of $n$ groups, where we set $n$ to 32 and 4 for 2D and 3D image registration tasks, respectively. $S$ is the swish function, $F$ is the fully connected layer.

### Table 5. Architecture of the deformation network.

| Unit       | Layers   | nCh (2D) | nCh (3D) |
|------------|----------|----------|----------|
| Enc#1      | C RB     | 64 8     |          |
| Enc#2      | D RB     | 128 16   |          |
| Enc#3      | D RB     | 256 32   |          |
| Enc#4      | D RB     | 256 32   |          |
| Mid        | RB SA RB | 256 32   |          |
| Dec#1      | Cat RB Cat RB U | 256 32  |          |
| Dec#2      | Cat RB Cat RB U | 256 32  |          |
| Dec#3      | Cat RB Cat RB U | 128 16  |          |
| Dec#4      | Cat RB Cat GSC | 64 8    |          |

| RB GSC     | S-F-Sum-GSC C-Sum |
| GSC        | GN(n)-S-C |

Table 4. Architecture of the diffusion network. $C$ is the convolution with 3D kernel and stride 1, where $d$ is the dimension of input. $D$ is the convolution with 3D kernel and stride 2, and $U$ is the nearest interpolation with scale factor 2, followed by $C$. $RB$ is the Resnet block, and SA is the self-attention block. $\text{GN}(n)$ denotes the group normalization of $n$ groups, where we set $n$ to 32 and 4 for 2D and 3D image registration tasks, respectively. $S$ is the swish function, $F$ is the fully connected layer.

### Table 5. Architecture of the deformation network.

| Unit       | Layers   | nCh (2D) | nCh (3D) |
|------------|----------|----------|----------|
| Enc#1      | CL×2     | 64 16    |          |
| Enc#2      | D CL     | 128 32   |          |
| Enc#3      | D CL     | 256 32   |          |
| Enc#4      | D CL     | 512 32   |          |
| Mid        | D CL U   | 1024 32  |          |
| Dec#1      | Cat CL×2 U | 1024 32  |          |
| Dec#2      | Cat CL×2 U | 512 32   |          |
| Dec#3      | Cat CL×4 U | 128 8    |          |
| Dec#4      | Cat CL×2 C | 64 8     |          |

| CL C-LeakyReLU |

C. Additional experimental results

### C.1. Registration along continuous trajectory

The proposed method can provide continuous image registration for the moving image along the trajectory toward the fixed image, which is one of the main contributions of our paper. Here, we show additional results for the continuous deformation by varying $\eta$ multiplied by the score of the diffusion network output.

Fig. 11 and Fig. 12 shows the results on the facial expression image registration task. For the comparison method of VM [4], the linear interpolation is performed in the registration field, whereas it is performed in the latent space of the score function for our method by adjusting $\eta$. As can be seen from the results, the estimated deformation field of VM only varies in their scale, but the relative spatial distribution of the registration field does not change. Accordingly, the change of the specific facial movement such as eyes is not clearly visible. On the other hand, in our proposed method, the estimated deformation field from the scaled score function values is not just a scaled version of the vector field as in VM, but rather exhibits very dynamically changing motion fields depending on the positions (for example, more specifically consistent movement along the eyes and mouths compared to other backgrounds). Thus, the resulting intermediate deformed images have distinct changes from the moving and fixed images.

### C.2. Registration results of the comparisons

#### 2D facial expression image registration

For the facial expression image registration task, we compared the proposed method to VM [4] and VM-diff [10]. We implemented these methods using the same architecture of the deformation network $D_\psi$ of our model, and trained the networks until the loss converges for fair comparison. Fig. 13 shows the results of visual comparisons on various facial expression images of men, women, and children. The quantitative evaluation results are reported in the main paper. Compared to the other methods, we can see that our model provides more accurate deformation of the moving source image into the fixed target image by the smooth deformation field. This can be also observed through the quantitative evaluation with NMSE and SSIM that are displayed on each result.

#### 3D brain MR image registration

To verify the performance of the 3D brain image registration, we employed several comparative methods: SyN [2] by Advanced Normalization Tools (ANTS) [3], VM [4], VM-diff [10], MSDIR-Net [28], and CM [22]. For the learning-based methods, we used the 3D model of deformation network $D_\psi$ as a baseline network, and set the same parameters for fair comparison. Fig. 14 shows the results of atlas-based brain image registration. The Dice scores for several structures are displayed on each result.
fixed image in not only the entire shape but also the detailed structures.

C.3. Generation of synthetic deformed image

In addition to the image registration, thanks to jointly training of the diffusion and deformation networks, our DiffuseMorph provides the image generation via the reverse of diffusion process. As described in the main paper, given the condition with a pair of moving and fixed images, the generation process starts from one step forward diffusion on the moving image. Then, the noisy moving image with a certain noise level is iteratively refined by the reverse diffusion steps. Fig. 15 shows the generative process on the facial expression images. The sampling results are obtained by 80 diffusion steps starting from the moving image with the noise level $\alpha_{200}$. As our method learns the score function for the deformation for various pairs of facial expression images, we can see that the generated samples from the moving images become similar to the fixed images. This clearly indicates that our model has a capacity of controlled image generation as well as image registration.
Figure 11. Results of facial expression image registration along the continuous trajectory. Image registration is performed from the left-gazed fearful to the front-gazed surprised images. For the case of VM, the linear interpolation is performed in the registration field domain, whereas it is performed in the latent space of the score function for our method by adjusting $\eta$. 

$\eta \phi$

VM

$m(\phi_\eta)$

Ours

$\hat{e}_\eta$

$\phi_\eta$

$m(\phi_\eta)$
Figure 12. Results of facial expression image registration along the continuous trajectory. Image registration is performed from the right-gazed disgusted to left-gazed contemptuous images. For the case of VM, the linear interpolation is performed in the registration field domain, whereas it is performed in the latent space of the score function for our method by adjusting $\eta$. 
Figure 13. Visual comparisons of facial expression image registration (left) using the estimated deformation fields (right). From top to bottom, the results are deformed from the right-gazed sad to the front-gazed contemptuous images, from the front-gazed neutral to the right-gazed surprised images, from the right-gazed surprised to the front-gazed happy images, from the left-gazed disgusted to the front-gazed fearful images, from the front-gazed neutral to the right-gazed happy images, from the left-gazed neutral to the front-gazed angry images, from the front-gazed disgusted to the left-gazed sad images. The average values of NMSE / SSIM are displayed on each result.
Figure 14. Visual comparisons of atlas-based brain MR image registration (odd rows) using the estimated deformation fields (even rows). Segmentation maps of several anatomical structures are overlaid with dark target contours and light source/deformed contours (blue: ventricles, green: thalami, orange: third ventricle, pink: hippocampi). The Dice score of each structure is displayed with the corresponding color on each result.
Figure 15. Results of the synthetic deformed image generation via our generative process from $T = 80$. From top to bottom, the deformed image is generated from the left-gazed neutral to the front-gazed happy images, from the front-gazed disgusted to the front-gazed sad images, from the front-gazed contemptuous to the right-gazed happy images, from the front-gazed contemptuous to the front-gazed sad images, from the front-gazed angry to the right-gazed happy images, from the front-gazed fearful to the left-gazed sad images, from the front-gazed surprised to the front-gazed fearful images.