DiffuseMorph: Unsupervised Deformable Image Registration Using Diffusion Model

Boah Kim, Inhwa Han, and Jong Chul Ye(✉)

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 in medical imaging. Classical registration algorithms usually require a high computational cost for iterative optimizations. Although deep-learning-based methods have been developed for fast image registration, it is still challenging to obtain realistic continuous deformations from a moving image to a fixed image with less topological folding problem. To address this, here we present a novel diffusion-model-based image registration method, called DiffuseMorph. DiffuseMorph not only generates synthetic deformed images through reverse diffusion but also allows image registration by deformation fields. Specifically, the deformation fields are generated by the conditional score function of the deformation between the moving and fixed images, so that the registration can be performed from continuous deformation by simply scaling the latent feature of the score. Experimental results on 2D facial and 3D medical image registration tasks demonstrate that our method provides flexible deformations with topology preservation capability.

Keywords: Image registration · Diffusion model · Image deformation · Unsupervised learning

1 Introduction

Deformable image registration is to estimate non-rigid 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.

Part of this paper is presented at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2022 [20].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19821-2_20.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13691, pp. 347–364, 2022.
https://doi.org/10.1007/978-3-031-19821-2_20

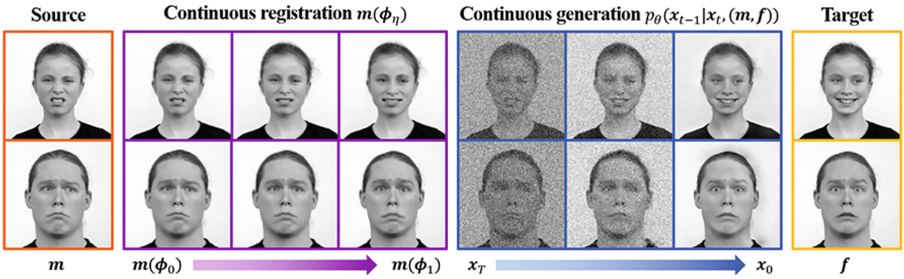

Fig. 1. DiffuseMorph provides not only deformable image registration along the continuous trajectory by simply scaling the latent features in generating deformation fields but also synthetic deformed images through continuous generation by the reverse diffusion process.

Classical image registration approaches usually attempt to align images by solving a computationally expensive optimization problem [1,2,22]. To address this computational issue, deep-learning-based image registration methods have been extensively studied [3,6,19,26,30], which train neural networks to estimate the registration field by taking the moving and fixed images as network inputs. These approaches provide fast deformation while maintaining registration accuracy. However, the supervised methods usually require the ground-truth registration fields [30,31], and some of the existing unsupervised approaches need additional diffeomorphic constraints [10,23] or the cycle-consistency [19] for topology preservation.

Recently, score-based diffusion models have shown high-quality performance in image generation [37,38]. In particular, the denoising diffusion probabilistic model (DDPM) [14,34] learns the Markov transformation from Gaussian noise to data distribution and provides diverse samples through the stochastic diffusion process by estimating the latent feature of score function, which has been applied to many areas of computer vision [8,12,18,33,35]. To generate images with desired semantics, conditional denoising diffusion models have been also presented [9,32]. However, it is challenging to apply DDPM to the image registration task, since the proper registration should be performed through the deformation field for the moving image rather than image generation.

In this paper, by leveraging the property of the diffusion model where the estimated latent feature provides spatial information to generate images, we present a novel unsupervised deformable image registration approach, dubbed DiffuseMorph, by adapting the DDPM to generate deformation fields. Specifically, our proposed model is composed of a diffusion network and a deformation network: the former network learns a conditional score function of the deformation between moving and fixed images, and the latter network estimates the deformation field using the latent feature from the score function and provides deformed images. These two networks are jointly trained in an end-to-end learning manner so that DiffuseMorph not only estimates Markov transformation in

the direction in which the moving image is deformed into the fixed image, but also produces the registration field for the moving image to be warped into the fixed image. Since the latent feature from the conditional score function of the diffusion model has spatial information of the condition, the linear scaling of the latent feature may provide deformation fields along the continuous trajectory from the moving to the fixed images.

Accordingly, as shown in Fig. 1, the proposed DiffuseMorph allows both image registration along the continuous trajectory and synthetic deformed image generation. Specifically, our trained model provide the continuous deformations from the moving image to the fixed image by simply interpolating the latent feature that is used as an input for the deformation network. In addition, the proposed model can quickly generate synthetic deformed images similar to the fixed images. Here, to further accelerate the diffusion procedure, instead of starting from random Gaussian noise, we present a generative process in which the moving image is propagated one step via forward diffusion and then iteratively refined through the reverse diffusion process of the DDPM. This reduces the number of diffusion steps significantly and makes the sample retain the original moving image content.

We demonstrate the performance of the proposed method on 2D facial expression registration and 3D medical image registration tasks. The experimental results verify that our model achieves high performance in registration accuracy. Also, thanks to the latent feature estimated from the diffusion model, our method enables real-time image registration along the continuous trajectory between the moving and fixed images, which is more realistic than the comparative learning-based registration methods. Our main 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.
- When the proposed model is trained, our model not only performs image registration along the continuous trajectory from the moving to fixed images by scaling the latent feature but also generates synthetic deformed images through the fast reverse diffusion process.
- We demonstrate that the proposed method can be applied to 2D and 3D image registration tasks and provide accurate deformation with comparable topology preservation over the existing methods.

2 Backgrounds 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:

$$\phi^* = \underset{\phi}{\operatorname{argmin}} \quad L_{sim}(m(\phi), f) + L_{reg}(\phi), \tag{1}$$

where ϕ^* is the optimal registration field to deform the moving image into the fixed image. L_{sim} is the dissimilarity function to compute the similarity between

the deformed and fixed images, and L_{reg} is the regularization penalty of the registration 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 one imposes additional constraints on the field ϕ such that the deformation mapping is differentiable and invertible, thereby preserving the topology [2,4,39].

Learning-based Registration Methods. As the traditional registration approaches usually require large computation and long runtime, deep learning methods have been extensively studied lately, which estimates the deformation field in real time once a neural network is trained. However, supervised learning methods train the networks using the ground-truth registration fields [6,7,31,42], which needs high-quality labels for training. To alleviate this, weakly-supervised registration models that use pseudo-labels such as segmentation maps have been developed [16,41]. On the other hand, unsupervised learning approaches train the networks by computing similarity between the deformed image and the fixed reference [3,19,26,27,40]. To guarantee topology preservation, learning-based diffeomorphic registration methods are also presented [10,11,23], which have the layer of scaling and squaring integration for the diffeormorphic constraint.

These existing methods may provide intermediate deformations between the moving and fixed images by scaling the registration field or integrating the velocity field in shorter timescales. However, our method produces more realistic continuous deformation by scaling the latent feature that has spatial information of the moving and fixed images, improving the performance of image registration.

2.2 Denoising Diffusion Probabilistic Model

Recently, the denoising diffusion probabilistic model (DDPM) [14,34], one of the generative models, is presented to learn the Markov transformation from the simple Gaussian distribution to the data distribution. In the forward diffusion process, noises are gradually added to the data x_0 using a Markov chain, in which each step of sampling latent variables x_t for $t \in [0, T]$ is defined as a Gaussian transition:

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

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

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

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

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon. \tag{4}$$

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

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I), \tag{5}$$

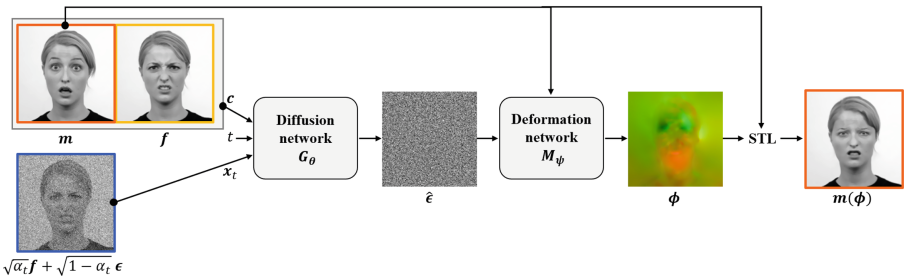

Fig. 2. The training framework of DiffuseMorph. Given a condition with a pair of a moving image m and a fixed image f, the diffusion network G_θ estimates the conditional score function of the deformation, and the deformation network M_ψ outputs the registration field ϕ. Then, using the spatial transformation layer (STL), the moving image is warped into the fixed image.

where σ_t is a fixed variance, and $\mu_\theta(x_t, t)$ is a learned mean defined as:

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

where ϵ_θ is a parameterized model. In fact, the model $\epsilon_\theta(x_t, t)$ is just a scaled version of the score function $s_\theta(x_t, t)$ [36], which is the gradient of the $\log p_\theta(x_t)$. Once the model ϵ_θ is trained, the data is sampled by the following stochastic generation step: $x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z$, where $\sigma_t^2 = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t$ and $z \sim \mathcal{N}(0, I)$.

Conditional Diffusion Models. In order to generate images with desired semantics, conditional diffusion models have been proposed recently [9,15,29,32,35,37], which gives the conditional reference image to the network or the generative process. DDIM [35] proposes a deterministic non-Markovian generative process starting from an initial condition to control the image generation of the reverse diffusion process. SR3 [32] presents a method to train the DDPM with a conditioned image for the super-resolution task. ILVR [9] proposes a conditioning iterative generative process using an unconditional model. However, these diffusion model-based generative methods are concerned about image generation, and cannot be used for image registration as they do not produce any deformation field for the registration.

3 Proposed Method

3.1 Framework of DiffuseMorph

By leveraging the capability of DDPM, we aim to develop a novel diffusion model-based unsupervised image registration approach. Since the image registration is to warp the moving image using the deformation field, we design our model with two networks as illustrated in Fig. 2: one is a diffusion network G_θ

to estimate a conditional score function, and the other is a deformation network M_ψ that actually outputs the registration field 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 conditional score function of the deformation between the moving and fixed images given the condition $c = (m, f)$. For this, we sample the latent variable x_t of the target by (4), defining 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 also give the number of time steps for the noise to the network, similar to [14].

On the other hand, the deformation network M_ψ takes the latent feature of the conditional score function $\hat{\epsilon}$ that is an output of the diffusion network, as well as the moving source image m. Then the network outputs the registration field ϕ, providing the deformed image $m(\phi)$ by warping the moving image m using the spatial transformation layer (STL) [17]. To deform 2D/3D images in our experiments, we adopt the transformation function using bi-/tri-linear interpolation.

3.2 Loss Function

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

$$\min_{G_\theta, M_\psi} L_{diffusion}(c, x_t, t) + \lambda L_{regist}(m, f), \tag{7}$$

where $L_{diffusion}$ and L_{regist} are the diffusion loss and the registration loss, respectively, and λ is a hyper-parameter. The detailed description of each loss function is as follows.

Given the condition c and the perturbed data x_t at the time step $t \in [0, T_{train}]$, the diffusion loss is to learn the conditional score function:

$$L_{diffusion}(c, x_t, t) = \mathbb{E}_{\epsilon, x_t, t} ||G_\theta(c, x_t, t) - \epsilon||_2^2, \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, I)$. Also, the registration loss is to estimate the deformation field so that the deformed source image has similar shape of the fixed image, which is designed as the traditional energy function in (1):

$$L_{regist}(m, f) = -(m(\phi) \otimes f) + \lambda_\phi \sum ||\nabla \phi||^2, \tag{9}$$

where $\phi = M_\psi(m, \hat{\epsilon})$ with $\hat{\epsilon}$ referring to the diffusion network output, and λ_ϕ is a hyper-parameter. The first term of (9) is the local normalized cross-correlation [3] between the deformed image and fixed image, and the second term is the smoothness penalty on the registration field. We set $\lambda_\phi = 1$.

It is remarkable that the net effect of the two loss functions is that G_θ is trained to estimate the latent feature for the conditional score function of the deformation, which has the spatial information of the moving and fixed images.

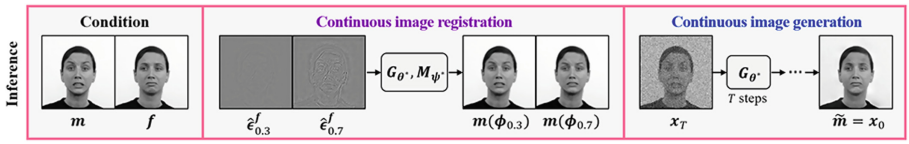

Fig. 3. In the inference phase, our model provides not only the image registration $m(\phi)$ that warps the moving image, but also generates synethetic images $\tilde{m}$.

Accordingly, the latent feature helps the proposed model to perform image registration along the continuous trajectory, which can provide topology preservation. Furthermore, when combined with the reverse diffusion process, the latent feature guides the reverse diffusion to generate the synthetic deformed image from the moving image initialization.

3.3 Image Registration Using DiffuseMorph

When the networks of the proposed model are trained, in the inference phase, they provide image registration by estimating the deformation field for the moving image to be aligned with the fixed image. Thanks to the end-to-end training of our model, the diffusion network allows the deformation network to generate the regular registration field. Specifically, using the learned parameters of G_{θ^*} and M_{ψ^*}, the registration field ϕ at $t = 0$ is estimated by:

$$\phi = M_{\psi^*}(m, G_{\theta^*}(c, x_0, t)), \tag{10}$$

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

Image Registration Along Continuous Trajectory. In the image registration that warps the moving image into the fixed image, our model provides the continuous deformation of the moving image along the trajectory toward the fixed image. This is possible since the deformation network estimates the registration field according to the latent feature. Specifically, if the latent feature from the conditional score is set to zero, the deformation network outputs the registration field that hardly warps the moving image, whereas when the latent feature is given as in (10), the deformation network estimates the registration field that deforms the moving image into the fixed image.

Accordingly, as described in Algorithm 1, for the latent feature $\hat{\epsilon}^f = G_{\theta^*}(c, f, 0)$, the registration field ϕ_η for the continuous image deformation can be generated by simply interpolating the latent feature:

$$\phi_\eta = M_{\psi^*}(m, \hat{\epsilon}^f_\eta), \qquad (11)$$

where $\hat{\epsilon}^f_\eta = \eta \cdot \hat{\epsilon}^f$ for $0 \leq \eta \leq 1$.
We believe that this interesting phenomenon occurs from learning the conditional score function of deformation, as will be observed later in our experiments.

Algorithm 1. Continuous image registration

1: **Input**: Conditional images, $c = (m, f)$
2: **Output**: Deformed moving image, $m(\phi_\eta)$
3: Set the latent feature $\hat{\epsilon}^f = G_{\theta^*}(c, f, 0)$
4: **for** $\eta \in [0, 1]$ **do**
5: $\quad \hat{\epsilon}^f_\eta \leftarrow \eta \cdot \hat{\epsilon}^f$
6: $\quad \phi_\eta \leftarrow M_{\psi^*}(m, \hat{\epsilon}^f_\eta)$
7: **end for**=
8: **return** $m(\phi_\eta)$

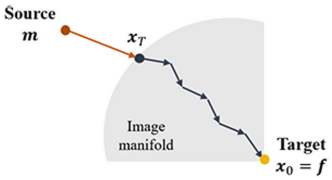

Fig. 4. Generative process toward fixed target data distribution.

Algorithm 2. Synthetic image generation process

1: **Input**: Conditional images, $c = (m, f)$
2: **Output**: Synthetic deformed image, x
3: Set $T \in (0, T_{train})$
4: Sample $x_T = \sqrt{\alpha_T} m + \sqrt{1 - \alpha_T}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$
5: **for** $t = T, T-1, ..., 1$ **do**
6: $\quad z \sim \mathcal{N}(0, I)$
7: $\quad x_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} G_{\theta^*}(c, x_t, t)) + \sigma_t z$
8: **end for**
9: **return** x_0

Synthetic Image Generation via Reverse Diffusion. In our method, the latent feature generated by the diffusion network itself also guides the generation of synthetic deformed images through the reverse diffusion process. Here, as the diffusion network learns the conditional score function for the deformation between the moving image and the fixed image, our image generation starts from the moving image, in contrast to the existing conditional generative process of DDPM [9,32] 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:

$$x_T = \sqrt{\alpha_T} m + \sqrt{1 - \alpha_T}\epsilon, \qquad (12)$$

where $\epsilon \sim \mathcal{N}(0, I)$, and α_T is the noise level at the 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. 4, which reduces the number of reverse diffusion steps and generation time.

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

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} G_{\theta^*}(c, x_t, t)\right) + \sigma_t z, \tag{13}$$

where $z \sim \mathcal{N}(0, I)$. Here, in choosing the total steps of reverse diffusion, we employ [8] that presents an efficient inference method. Thus, one can flexibly set the number of sampling steps, and in our experiments, we set the reverse steps as 200 in maximum. The pseudocode of this generative process of DiffuseMorph is described in Algorithm 2.

4 Experimental Results

To demonstrate that DiffuseMorph generates high-quality deformed images from the moving to the fixed images, we apply our method to the various image registration tasks. We conduct the experiments on the intra-subject image registration using 2D facial expression images and 3D cardiac MR scans. Also, we apply our model to 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 Supplementary Material.

Datasets. For 2D face images, Radboud Faces Database (RaFD) [25] was used. It contains 8 facial expressions collected from 67 subjects: neutral, angry, contemptuous, disgusted, fearful, happy, sad, and surprised. For each expression, 3 different gaze directions are provided. We cropped the data to 640×640, resized them into 128×128, and converted the RGB images to gray scale. We divided the data by 53, 7, and 7 subjects for training, validation, and test, respectively.

For 3D cardiac MR scans, we used ACDC dataset [5] that provides 100 4D temporal cardiac MRI data from the diastolic to systolic phases and segmentation maps at both ends of the phases. We resampled all scans with a voxel spacing of $1.5 \times 1.5 \times 3.15mm$, cropped them to $128 \times 128 \times 32$, and normalized the intensity into [-1, 1]. We used 90 and 10 scans for training and test.

Also, we used OASIS-3 dataset [24] for 3D brain MR registration. It provides brain MR images and corresponding volumetric segmentation maps from FreeSurfer [13]. 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, 93, and 129 scans for training, validation, and test, respectively.

Implementation Details. Our model was implemented using PyTorch library in Python. We used the network architecture designed in DDPM [14] for the diffusion network, and set the noise level from 10^{-6} to 10^{-2} by linearly scheduling with $T_{train} = 2000$. Also, we used the backbone of VoxelMorph-1 [3] for the

deformation network. Here, we configured layers of the networks according to the dimension of the image, e.g. 2D convolution layer for 2D image registration. For the face dataset, we set the hyper-parameter as $\lambda = 2$, and trained the model with the learning rate 5×10^{-6} for 40 epochs. For the cardiac MR data, we trained the model with $\lambda = 20$ and the learning rate 2×10^{-4} for 800 epochs. Also, we trained the model using the brain MR data for 60 epochs with $\lambda = 10$ and the learning rate 1×10^{-4}. Using a single Nvidia Quadro RTX 6000 GPU, we trained our model by Adam optimization algorithm [21].

Evaluation. To evaluate the registration performance, we computed the percentage of non-positive values of Jacobian determinant on the registration field ($|J_\phi| \leq 0$), which indicates that one-to-one mapping of the registration has been lost. Here, for the facial images, we measured NMSE and SSIM between deformed and fixed images. For the MR images, we computed Dice score between the deformed segmentation maps and the ground-truth labels for several anatomical structures. On the other hand, to evaluate the continuous deformation quality of cardiac scans, we computed PSNR and NMSE between the deformed images and the real data. For the comparative learning-based models, we used the same deformation network architecture and parameters for a fair comparison.

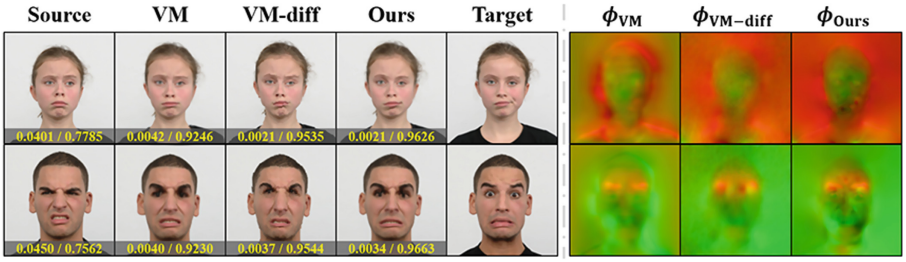

Fig. 5. Visual comparison results of image registration on the facial images (left) using the estimated registration fields (right). Results are deformed from the right-gazed sad to the front-gazed contemptuous images (top), and from the left-gazed disgusted to the front-gazed fearful images (bottom). The average values of NMSE/SSIM are displayed on each registration result.

Table 1. Quantitative evaluation results of the facial expression image registration. Standard deviations are shown in parentheses.

Method	NMSE $\times 10^{-1}$	SSIM	$\lvert J_\phi \rvert \leq 0$ (%)
Initial	0.363 (0.268)	0.823 (0.066)	0
VM [3]	0.047 (0.057)	0.936 (0.024)	0.050 (0.106)
VM-diff [10]	0.034 (0.015)	0.957 (0.013)	**0.014 (0.065)**
Ours	**0.032 (0.017)**	**0.964 (0.011)**	0.017 (0.056)

4.1 Results of Intra-subject 2D Face Image Registration

We compared DiffuseMorph against VM [3] and VM-diff [10]. We tested the image registration performance on the deformed images in RGB scale by applying the registration field to each RGB channel. Figure 5 shows visual comparisons of the registration results. Compared to VM and VM-diff, our model deforms the source image to be more accurately aligned with the target image. Also, as reported in Table 1, our model achieves lower NMSE and higher SSIM. Moreover, the metric of Jacobian determinant on registration fields of ours shows comparable values to VM-diff with the diffeomorphic constraint. These results indicate that the proposed DiffuseMorph provides high-quality image registration. More results of facial expression images can be found in Supplementary Material.

Continuous Image Deformation. We also performed continuous deformations of the facial expression from the moving source to the fixed target. Figure 6 shows the intermediate images of our model and the comparative methods. We obtained the results of VM by scaling the registration field, i.e. $\zeta \cdot \phi$ with $0 \leq \zeta \leq 1$, and those of VM-diff by integrating the velocity field along timescales, i.e. $\phi^{1/2^v}$ where v is the number of time steps. We can see that the estimated registration fields of VM only vary in their scale, but the relative spatial distribution does not change. Also, VM-diff does not provide regularly continuous deformation. On the other hand, in the proposed method, the registration field changes non-uniformly according to η, depending on the importance of variations at the intermediate deformation level. The improved performance of our method also can be also quantitatively verified using the facial landmarks extracted by *dlib* library in Python. Specifically, the visual results and MSE values in Fig. 6 show that our model provides superior performance for the continuous deformation, which indicates the importance of the conditional score for deformation.

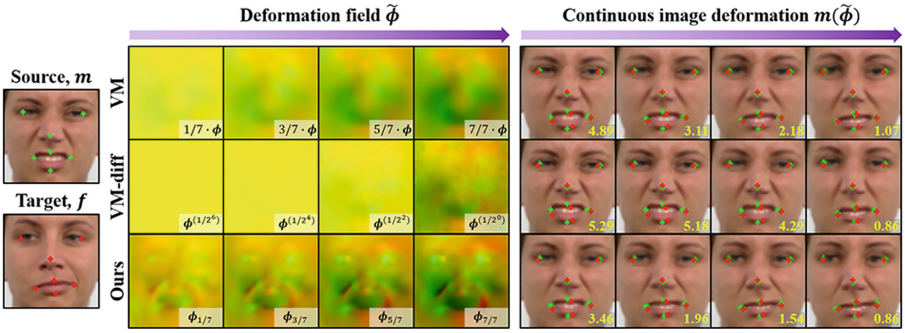

Fig. 6. Results of continuous image deformation of the facial images with facial landmarks. The deformation is performed from the right-gazed disgusted to the left-gazed contemptuous images. The average of MSE between the deformed and target landmarks is displayed on each result.

Synthetic Deformed Image Generation. To verify the capability of image generation from our DiffuseMorph, we evaluated the generation using the facial images. Here, the generative process was performed using the unseen images in the training phase. Figure 7 shows the generated samples $\tilde{m}$ given the moving source and fixed target images. The samples are obtained from the noisy moving image with the noise level α_{200} for the forward diffusion. We set the number of reverse diffusion steps to 80. As shown in the results, our model provides the synthetic deformed images similar to the target images for various pairs of facial expressions. Also, when compared to the warped image $m(\phi)$ using the registration field, we can observe that the proposed generative process is effective to provide image deformation if the moving image does not have teeth shown in the fixed image.

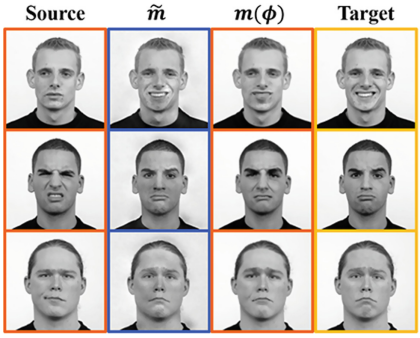

Fig. 7. Results of the synthetic deformed image generation via our generative process. The process of the samples is included in Supplementary Material.

4.2 Results of Intra-subject 3D Cardiac MR Image Registration

We tested the image registration of the end-diastolic images aligned with the end-systolic images. Table 2 reports the registration results with the average Dice score for the segmentation maps of left blood pool (BP), myocardium (Myo), left ventricle (LV), right ventricle (RV), and these total region, as well as the Jacobian metrics. Compared to the baseline methods of VM [3] and VM-diff [10], our model achieves high registration accuracy with a comparable number of folds in topology preservation.

In addition, since the cardiac dataset we used provides 4D data between the end-diastolic to the end-systolic phases, we performed quantitative evaluation for the continuous deformation. As shown in Table 2, when we measured the PSNR and NMSE between the deformed images and ground-truth reference images, our method provides continuous deformed images more similar to the ground-truth

Table 2. Quantitative comparison results of the cardiac image registration. Standard deviations are shown in parentheses.

Method	Image registration		Continuous deformation		
	Dice	$\lvert J_\phi \rvert \leq 0$ (%)	PSNR (dB)	NMSE $\times 10^{-8}$	Time (sec)
Initial	0.642 (0.188)	–	28.058 (2.205)	0.790 (0.516)	–
VM [3]	0.787 (0.113)	0.169 (0.109)	30.678 (2.652)	0.477 (0.453)	0.219
VM-diff [10]	0.794 (0.104)	0.291 (0.188)	29.481 (2.473)	0.602 (0.477)	2.902
Ours	**0.802 (0.109)**	**0.161 (0.082)**	**30.725 (2.579)**	**0.466 (0.432)**	0.456

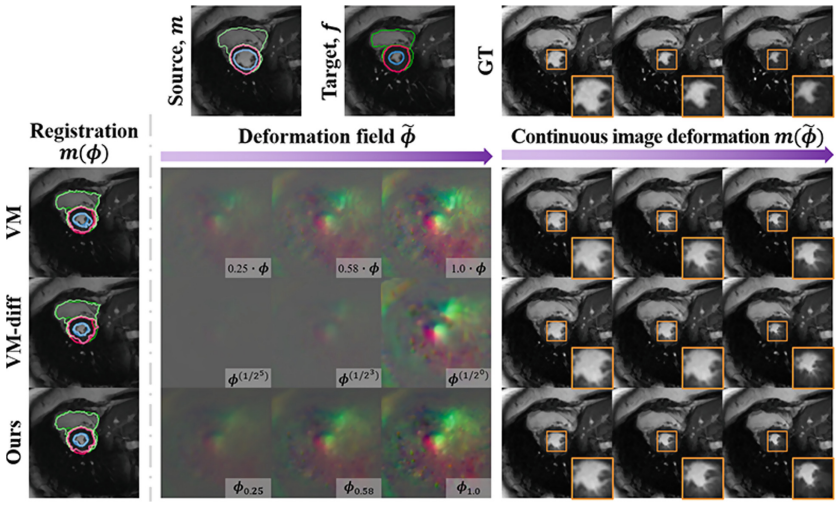

Fig. 8. Results of the image registration and continuous deformation on the cardiac MR images. The registration results show the overlaid contours of segmentation maps (green: RV, red: Myo, blue: BP). GT is the ground-truth data. (Color figure online)

images than the comparative methods. The visual comparison results in Fig. 8 also shows the superiority of our method. More visual results can be found in Supplementary Material.

4.3 Results of Atlas-based 3D Brain MR Image Registration

For the brain MR image registration, we compared our DiffuseMorph to the following comparative methods: SyN [2], VM [3], VM-diff [10], SYMNet [28], MSDIRNet [26], and CM [19]. As shown in the visual comparison results of Fig. 9, our model estimates smooth registration fields and yields deformed moving images that are more accurately aligned to the fixed images, compared to the baseline models. This also can be observed through the contours of the segmentation maps of several brain anatomical structures. We provided more visual results in Supplementary Material. Figure 10 and Table 3 report the results of quantitative evaluation. These show that the proposed method achieves higher Dice scores with less non-positive values of the Jacobian determinant over the existing learning-based methods, which empirically suggests that the proposed method can provide accurate image registration with improved topology preservation.

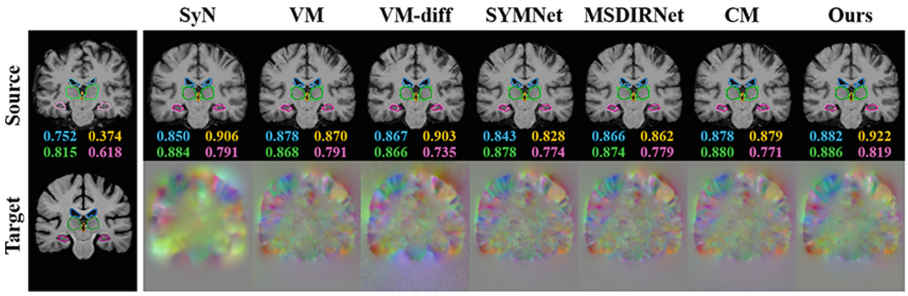

Fig. 9. Results of the image registration on the brain MR images (top) and the estimated registration fields (bottom). Segmentation maps of several anatomical structures are overlaid with the contours (blue: ventricles, green: thalami, orange: third ventricle, pink: hippocampi). The Dice scores for each structure are displayed with the corresponding colors on each result. (Color figure online)

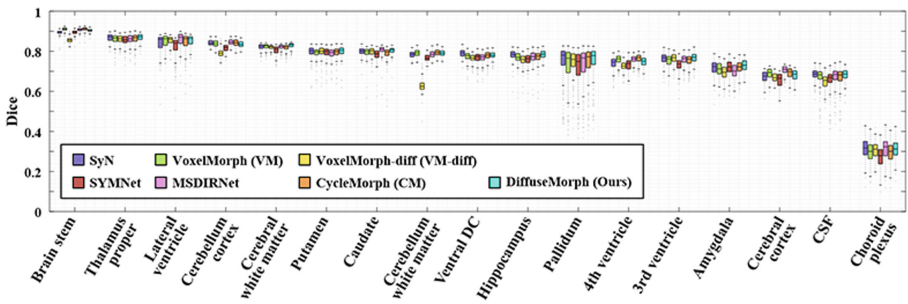

Fig. 10. Quantitative evaluation results of Dice scores on brain anatomical structures in the brain MR image registration experiment.

Table 3. Quantitative evaluation results of the brain MR image registration. Standard deviations are shown in parentheses.

Method	Dice	$\lvert J_\phi \rvert \leq 0$ (%)	Time (min)
Initial	0.616 (0.171)	0	0
SyN [2]	0.752 (0.140)	0.400 (0.100)	122, CPU
VM [3]	0.749 (0.145)	0.553 (0.075)	0.01, GPU
VM-diff [10]	0.731 (0.139)	0.631 (0.073)	0.01, GPU
SYMNet [28]	0.733 (0.148)	0.547 (0.049)	0.43, GPU
MSDIRNet [26]	0.751 (0.142)	0.804 (0.089)	2.06, GPU
CM [19]	0.750 (0.144)	0.510 (0.087)	0.01, GPU
Ours	**0.756 (0.139)**	**0.505 (0.058)**	0.01, GPU

Study on the Effect of Registration Loss. To explore the effect of the registration loss on the performance of our model, we performed a comparison study by varying the value of λ in (7) using brain MR data. As reported in Table 4, when λ is lower, the Dice score decreased but produced better regularity of registration fields. This indicates that the registration loss forces our model to provide more accurate image registration, but the trade-off needs to be balanced.

Table 4. Result of the study on the effect of registration loss. Standard deviations are shown in parentheses.

Method	Dice	$\lvert J_\phi \rvert \leq 0$ (%)
$\lambda = 2$	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)

5 Conclusion

We presented a novel DiffuseMorph model 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 conditional score function of deformation, the proposed method not only generates synthetic deformed images but also provides high-quality image registration from the continuous deformation by estimating registration fields along the trajectory for the moving image toward the fixed image. We expect that DiffuseMorph can be a promising algorithm to generate temporal data using moving and fixed images.

Acknowledgments. This work was supported in part by the National Research Foundation (NRF) of Korea under Grant NRF-2020R1A2B5B03001980, in part by Field-oriented Technology Development Project for Customs Administration through NRF of Korea funded by the Ministry of Science & ICT and Korea Customs Service(NRF-2021M3I1A1097938), and in part by the KAIST Key Research Institute (Interdisciplinary Research Group) Project.

References

1. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
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.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
4. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision **61**(2), 139–157 (2005). https://doi.org/10.1023/B:VISI.0000043755.93987.aa

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. Cao, X., et al.: Deformable image registration based on similarity-steered CNN regression. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 300–308. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_35
7. Cao, X., Yang, J., Wang, L., Xue, Z., Wang, Q., Shen, D.: Deep learning based inter-modality image registration supervised by intra-modality similarity. In: Shi, Y., Suk, H.-I., Liu, M. (eds.) MLMI 2018. LNCS, vol. 11046, pp. 55–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00919-9_7
8. Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W.: WaveGrad: estimating gradients for waveform generation. In: International Conference on Learning Representations (2020)
9. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: ILVR: conditioning method for denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14367–14376 (2021)
10. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
11. 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)
12. Fadnavis, S., Batson, J., Garyfallidis, E.: Patch2self: denoising diffusion MRI with self-supervised learning. Adv. Neural. Inf. Process. Syst. **33**, 16293–16303 (2020)
13. Fischl, B.: FreeSurfer. Neuroimage **62**(2), 774–781 (2012)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
15. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**(47), 1–33 (2022)
16. Hu, Y., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. **49**, 1–13 (2018)
17. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Adv. Neural. Inf. Process. Syst. **28**, 2017–2025 (2015)
18. Jeong, M., Kim, H., Cheon, S.J., Choi, B.J., Kim, N.S.: Diff-TTS: a denoising diffusion model for text-to-speech. arXiv preprint arXiv:2104.01409 (2021)
19. 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)
20. Kim, B., Ye, J.C.: Diffusion deformable model for 4D temporal medical image generation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention–MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13431. Springer, Cham.(2022). https://doi.org/10.1007/978-3-031-16431-6_51
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (Poster) (2015)

22. 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)
23. 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
24. LaMontagne, P.J., et al.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. MedRxiv (2019)
25. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the Radboud faces database. Cogn. Emot. **24**(8), 1377–1388 (2010)
26. Lei, Y., et al.: 4d-CT deformable image registration using multiscale unsupervised deep learning. Phys. Med. Biol. **65**(8), 085003 (2020)
27. Mahapatra, D., Antony, B., Sedai, S., Garnavi, R.: Deformable medical image registration using generative adversarial networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1449–1453. IEEE (2018)
28. 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)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
30. Onofrey, J.A., Staib, L.H., Papademetris, X.: Semi-supervised learning of nonrigid deformations for image registration. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Müller, H., Tu, Z. (eds.) MCV 2013. LNCS, vol. 8331, pp. 13–23. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05530-5_2
31. Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: SVF-Net: learning deformable image registration using shape matching. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 266–274. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_31
32. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636 (2021)
33. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
34. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
36. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Adv. Neural Inf. Processi. Syst. **32**, 1–13 (2019)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2020)
38. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. Adv. Neural. Inf. Process. Syst. **34**, 11287–11302 (2021)
39. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. Neuroimage **45**(1), S61–S72 (2009)

40. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. **52**, 128–143 (2019)
41. Xu, Z., Niethammer, M.: DeepAtlas: joint semi-supervised learning of image registration and segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 420–429. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_47
42. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration-a deep learning approach. Neuroimage **158**, 378–396 (2017)