Supplementary Material for "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

A Details of Implementation

The source code for this paper can be found at the following link: `https://github.com/DiffuseMorph/DiffuseMorph`.

A.1 Network Architecture

Diffusion Network As we described in the paper, we employed the network architecture of DDPM [8] for the diffusion network G_θ of the proposed method. Fig. 1 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 [13], the swish function [6], and convolution layers, which takes the embedded time t_e as well as the output of the preceding layer. At the last encoder block, the feature maps are attended by the self-attention module [12]. Also, each decoder block additionally takes the feature maps of the encoder outputs via 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_t are given to the diffusion network with the time step t, the network is trained to estimate the latent feature of the conditional score function $\hat{\epsilon}$ of the deformation between the moving and fixed images. Here, we configured the kernel dimension of convolution layers according to the input image dimension.

Deformation Network For the deformation network M_ψ, we implemented VoxelMorph-1 [3] that presents the network architecture for image registration. As illustrated in Fig. 2, 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 [7] layers, called CL units, which enables the network to focus on image processing and learn the complex image features. 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 latent feature of the diffusion network

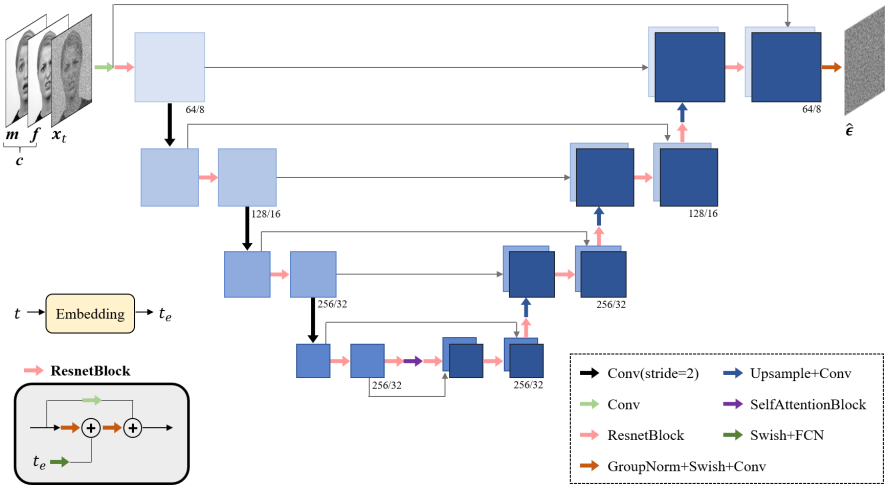

Fig. 1. The architecture of the diffusion network G_θ. The Resnet block (pink arrow) is composed of the group normalization, swish function, and convolution layers. The number of convolutional channels is shown as a/b below the square box, where a and b are for 2D and 3D networks, respectively. The downsampling 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 via skip-connection.

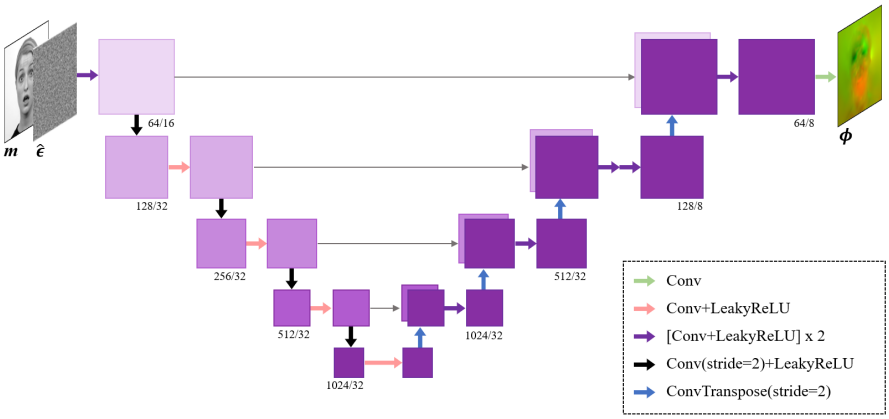

Fig. 2. The architecture of the deformation network M_ψ. It is a modified U-net structure with the convolution layer and leakyReLU activation function. The number of convolutional channels is shown as a/b below the square box, where a and b are for 2D and 3D networks, respectively. The image features 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 map of each encoder block is concatenated to that of the decoder block at the same level.

output $\hat{\epsilon}$, the deformation network estimates the registration field ϕ that warps the moving image into the fixed image. Similar to the diffusion network, the convolution layers are configured depending on the input image dimension.

A.2 Data Processing

The intensity range of grayscale facial data and medical images 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. Since the moving image is deformed by the registration field using the spatial transformation layer 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 hand, for the brain and cardiac MR data, we used random horizontal/vertical flips and random rotations with 90 degrees for data augmentation.

A.3 Baseline Methods

To implement baseline methods in our main paper, we set parameters by following several references. Specifically, for cardiac data, we implemented the baselines using the suggested parameters of [3, 5]. For the brain data, we set the parameters by following to [9], and SyN was conducted by mostly following to [4] using a step size of 0.25 at three scales with at most 100 iterations each.

B Memory and Time Costs

In this section, we analyzed the costs when training and testing the proposed model. Specifically, we computed the number of learnable parameters of our model, memory usage, and average runtime for the test. Table 1 reports the costs in the face experiment. Compared to the methods of VM and VM-diff, our method has more learnable parameters, but the image registration is performed in real-time since ours provides the registration in one step as other methods. Also, the synthetic image generation takes about 6 seconds for sampling with 80 steps that we described in the main paper.

Table 1. The memory and time costs of ours and other methods.

Method	Network	Test (Registration)		Test (Generation)	
	#Params	Memory	Time	Memory	Time
VM [3]	74.74M	0.30 GB	0.05 sec	N/A	N/A
VM-diff [5]	74.74M	0.35 GB	0.10 sec	N/A	N/A
Ours	90.67M	0.37 GB	0.07 sec	0.37 GB	6.02 sec

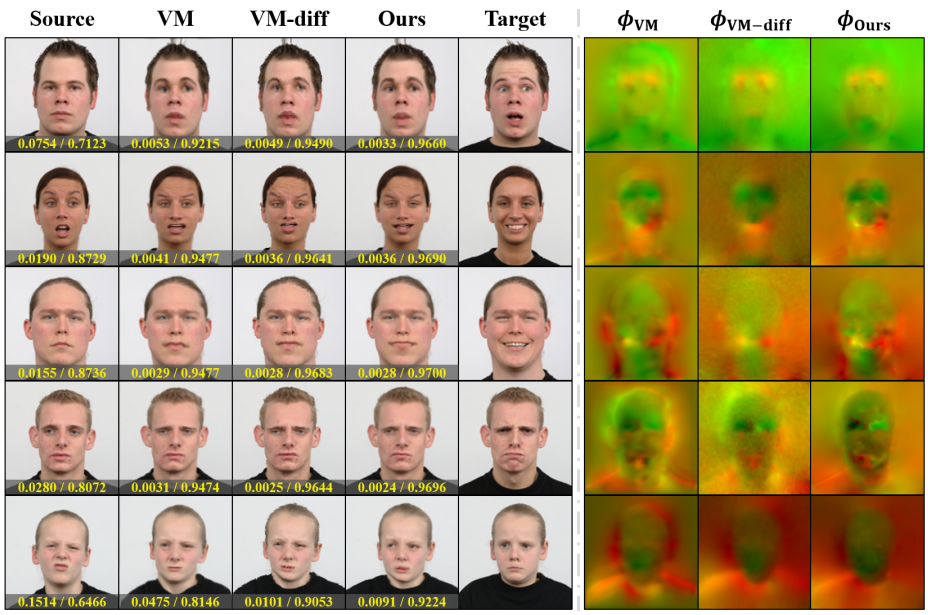

Fig. 3. Visual results of facial expression image registration (left) using the estimated registration fields (right). From top to bottom, the results are deformed from the front-gazed neutral to right-gazed surprised images, from the right-gazed surprised to front-gazed happy images, from the front-gazed neutral to right-gazed happy images, from the left-gazed neutral to front-gazed angry images, from the front-gazed disgusted to left-gazed sad images. The mean values of NMSE/SSIM are displayed on each result.

C Additional Experimental Results

C.1 Image Registration Results of the Comparisons

2D Facial Expression Image Registration For the intra-subject facial expression image registration task, we compared the proposed method with the VoxelMorph (VM) [3] and VM-diff [5]. We implemented these methods using the same architecture of the deformation network D_ψ of our model and trained the networks until the training loss converges for a fair comparison. Fig. 3 shows the results of visual comparisons on various facial expression images. Compared to the other baseline methods, we can see that our model provides more accurate deformation of the moving source image into the fixed target image by the smooth registration field. This can be also observed through the quantitative evaluation with NMSE and SSIM that are displayed on each result.

3D Cardiac MR Image Registration We also implemented our method for the intra-subject 3D cardiac MR image registration task. For the baseline methods, we compared ours with VM [3] and VM-diff [5]. As the face image

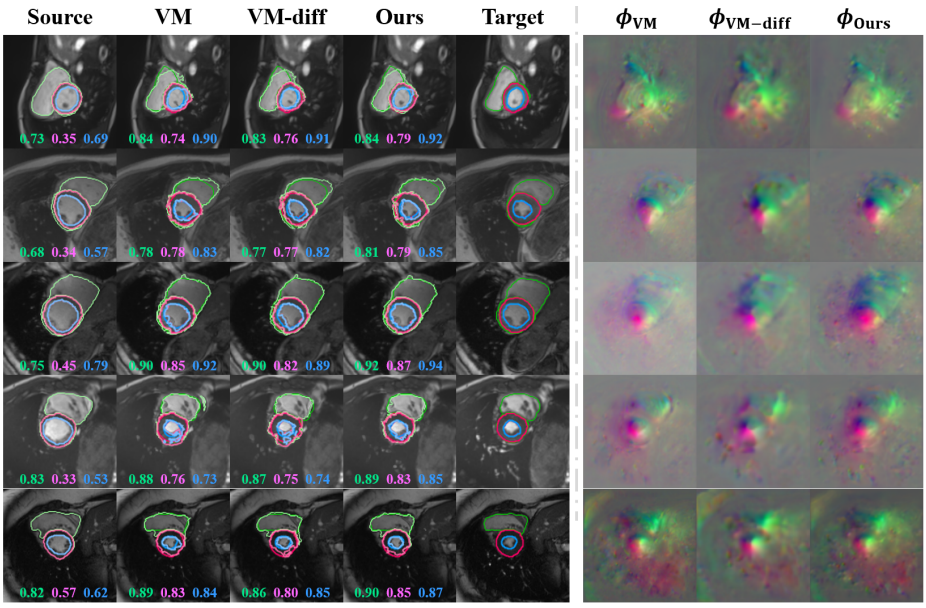

Fig. 4. Visual results of cardiac MR image registration (left) using the estimated registration fields (right). The registration results show the overlaid contours of segmentation maps (green: epicardium of the right ventricle (RV), red: myocardium of left ventricle (Myo), blue: left blood-pool (BP)). The Dice score for each structure is displayed with the corresponding color on each result.

registration, we trained these models using the deformation network architecture D_ψ of our method. Fig. 4 visualizes the registration results of the cardiac image at the end-diastolic phase to the end-systolic phase. We also display the contours of segmentation maps for several structures and their Dice scores. The results show that our proposed method achieves higher registration performance than the comparative methods in that the moving source image is more accurately aligned with the fixed target image.

3D Brain MR Image Registration To verify the performance of the atlas-based 3D brain image registration, we employed the following comparative methods: SyN [1] by Advanced Normalization Tools (ANTs) [2], VM [3], VM-diff [5], SYMNet [11], MSDIRNet [10], and CM [9]. For the learning-based methods, we used the 3D model of deformation network D_ψ as a baseline network and set the same parameters for a fair comparison. Fig. 5 shows the results of brain image registration. The Dice scores for several anatomical structures are displayed on each result, and the overall quantitative evaluation results can be found in the main paper. The visual results with the contours of segmentation maps show that the proposed DiffuseMorph deforms the moving source image more similar

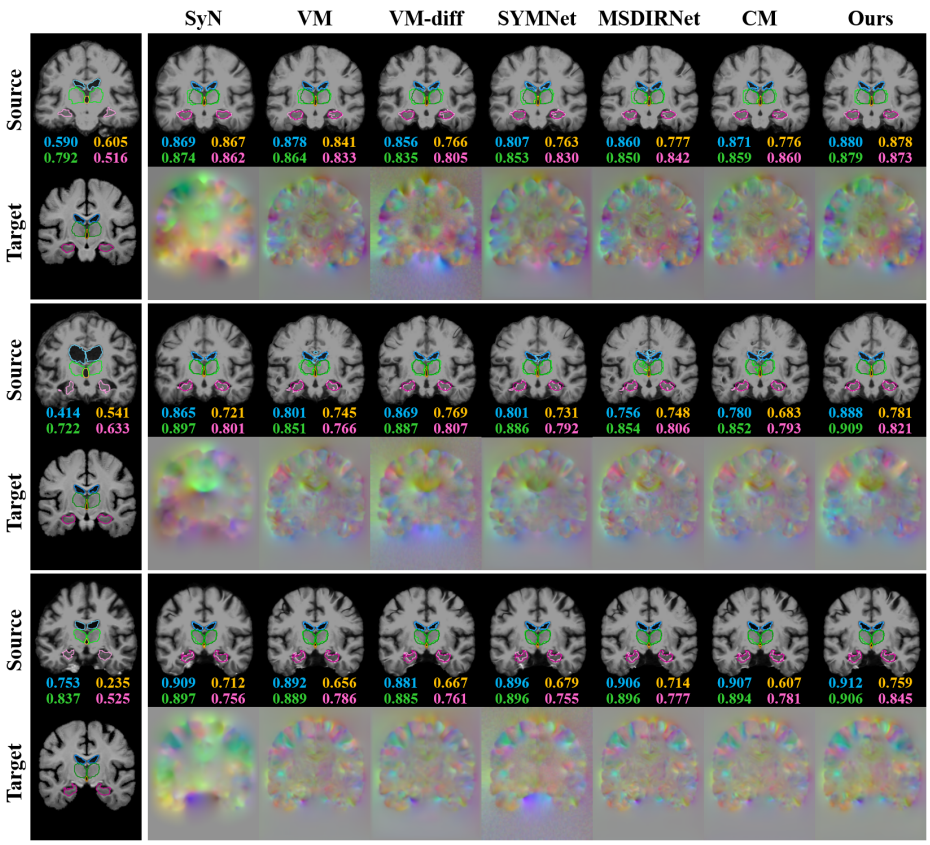

Fig. 5. Visual results of atlas-based brain MR image registration (odd rows) using the estimated registration fields (even rows). Segmentation maps of several anatomical structures are overlaid with the contours (blue: ventricles, green: thalami, orange: third ventricle, pink: hippocampi). The Dice score of each structure is displayed with the corresponding color on each result.

to the fixed target images than the others, not only in the overall shape but also in the detailed structures.

C.2 Image Registration Along Continuous Trajectory

The proposed method can provide continuous image deformation 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 of ours and comparative methods, VM and VM-diff. For the case of VM, we obtain the deformations by linearly interpolating the registration field. For the VM-diff, the continuous deformation is done by integrating the velocity field in shorter timescales. In con-

trast, our method yields the intermediate images by scaling the latent feature from the conditional score function of the deformation.

Fig. 6 shows the results on the facial expression image registration task. As can be seen from the results, the estimated registration fields of VM are only scaled so that the change of the specific facial movement such as eyes is not clearly visible. Also, VM-diff is limited in providing continuous deformation since the registration fields in early levels are near zero but deform images rapidly in late levels. On the other hand, in our proposed method, the estimated registration fields from the scaled latent feature are not just a scaled version as in the VM, but rather exhibits very dynamically changing deformation fields depending on the positions (for example, more specifically consistent movement along the eyes and mouths compared to other methods). Thus, the resulting intermediate deformed images of our method have distinct changes from the moving and fixed images. These results can be observed similarly in Fig. 7, which visualizes the comparison results for the cardiac MR image registration task and verifies the continuous deformation performance of our method.

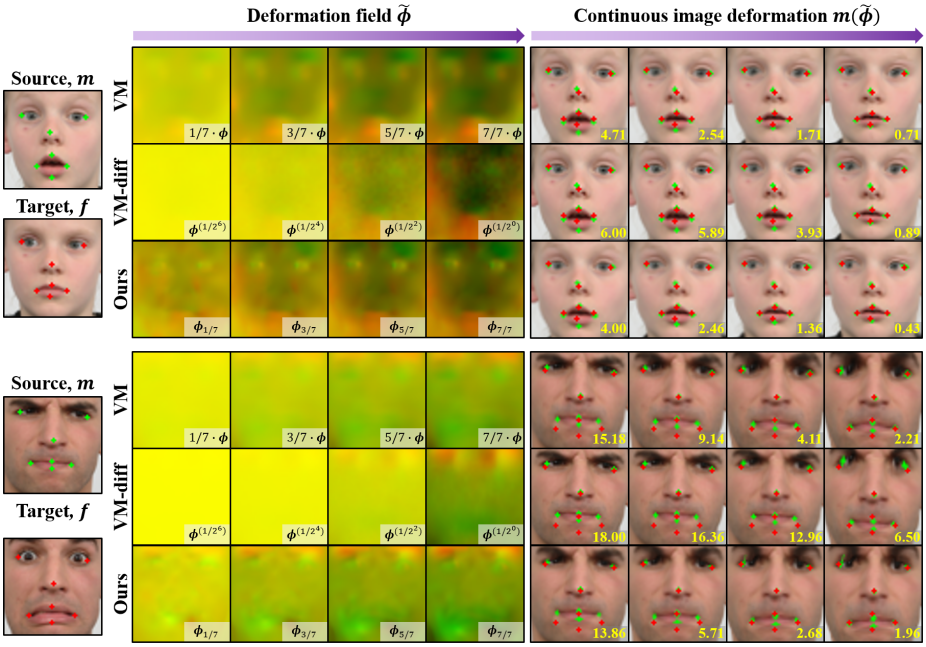

Fig. 6. Results of the continuous image deformation using the facial expression images with landmarks. Image registration is performed from the front-gazed surprised to the left-gazed neutral images (top), and from the right-gazed angry to the front-gazed fearful images (bottom). The average of MSE between the deformed and target facial landmarks is displayed on each result.

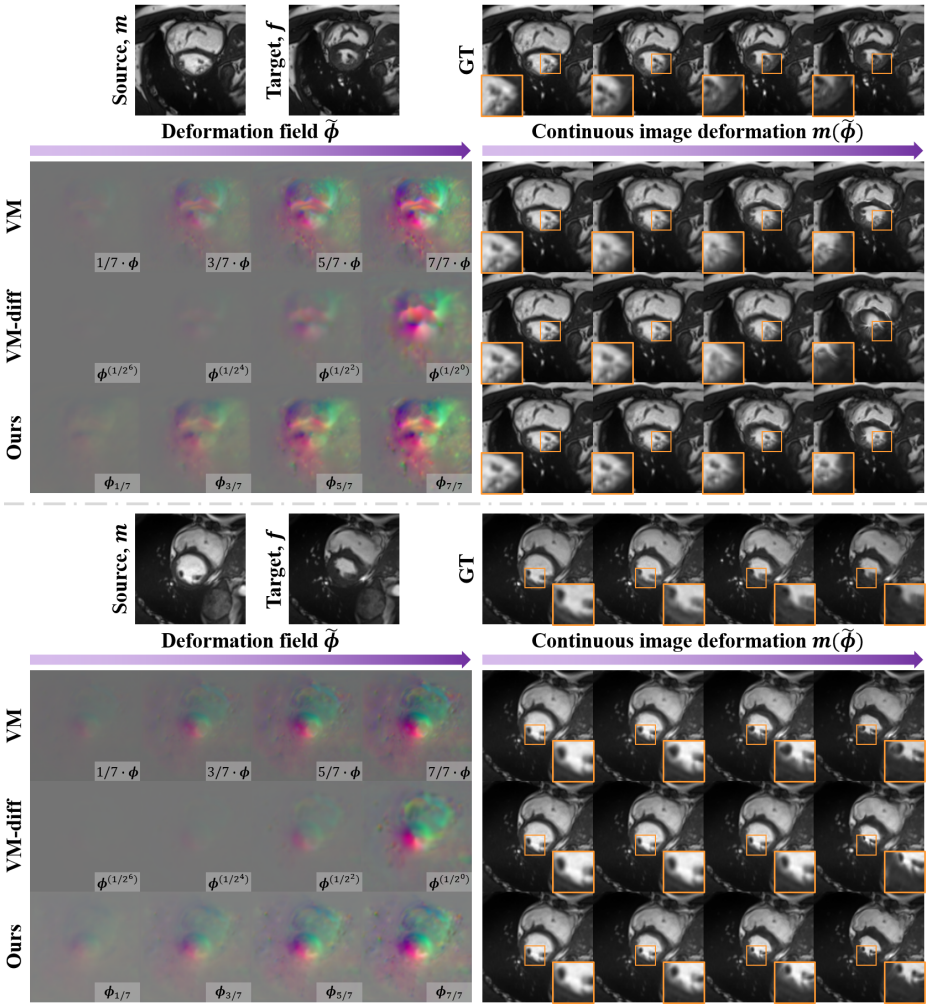

Fig. 7. Results of the continuous image deformation using the cardiac MR images. GT is the ground-truth data, and the orange box shows the remarkable regions.

C.3 Synthetic Deformed Image Generation

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 the 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 refined iteratively by the reverse diffusion steps, resulting in the synthetic deformed images aligned with the fixed image. Fig. 8 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 α_{200}. As our method learns the conditional score function of the deformation using various pairs of facial expression images, we can see that the generated samples from the moving images become similar to the fixed images. This indicates that our model has a capacity for conditional image generation as well as image registration.

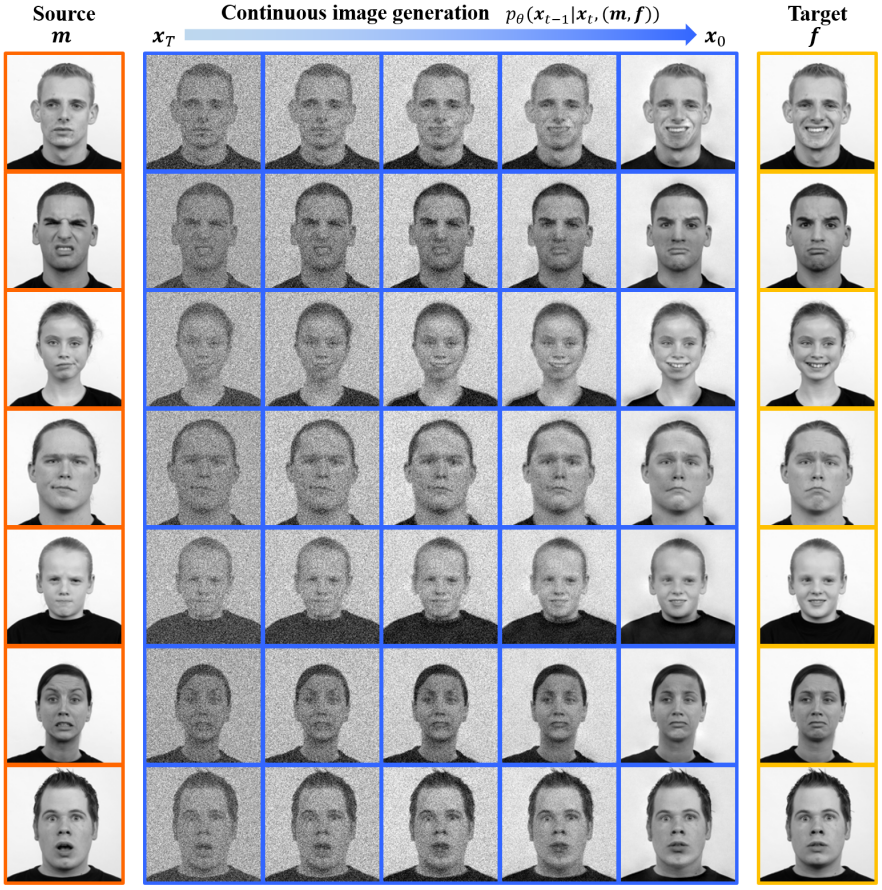

Fig. 8. Results of the synthetic deformed image generation via our generative process from $T = 80$. From top to bottom, the deformed image is generated from the left-gazed neutral to the front-gazed happy images, from the front-gazed disgusted to the front-gazed sad images, from the front-gazed contemptuous 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.

References

1. 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. Medical image analysis **12**(1), 26–41 (2008)
2. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
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. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. TMI (2019)
5. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 729–738. Springer (2018)
6. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks **107**, 3–11 (2018)
7. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
9. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: Cyclemorph: Cycle consistent unsupervised deformable image registration. Medical Image Analysis **71**, 102036 (2021)
10. Lei, Y., Fu, Y., Wang, T., Liu, Y., Patel, P., Curran, W.J., Liu, T., Yang, X.: 4d-ct deformable image registration using multiscale unsupervised deep learning. Physics in Medicine & Biology **65**(8), 085003 (2020)
11. 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)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
13. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)