Supplementary Material
Face Animation with an Attribute-Guided Diffusion Model

Bohan Zeng[1]*, Xuhui Liu[1]*, Sicheng Gao[1]*, Boyu Liu[1], Hong Li[1]
Jianzhuang Liu[2], Baochang Zhang[1,3]†
[1]Beihang University
[2]Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences
[3]Zhongguancun Laboratory, Beijing, China

In this supplementary material, we first provide examples of using FADM to improve existing videos in Section A. Then, we give more details of evaluation metrics in Section B and more experimental results in Section C. Finally, we state the ethical impact in Section D. The source code and more generated videos will be released soon.

A. Improving Existing Videos

Given an existing face animation video without real source and driving images, FADM is also effective to improve its overall visual quality, as stated in Sec. 3.2.2 in the main paper. In this case, we assume that the first frame and the current frame of the video are the source image and driving frame, respectively, and then extract their poses and expressions with the 3D face reconstruction module, so as to estimate the motion weight for establishing the motion condition. One example is shown in Fig. 1. Despite the lack of explicit source and driving images, FADM is still effective to synthesize impressive fine-grained facial details, demonstrating its great practical value for improving existing animation videos.

It is worth noting that FADM is a one-shot video generation method, which can be applied to other video generation tasks in addition to face animation. For example, we can provide FADM with a suitable scaling factor or required frames per second (fps) to generate super-resolution videos.

B. Metrics

We provide more detailed description about the metrics used to evaluate the effectiveness of FADM in the main paper below:

$\mathcal{L}_1$ represents the average $\mathcal{L}_1$ distance between the generated and ground-truth images.

Learned Perceptual Image Patch Similarity (LPIPS) [14] measures the perceived distance between the generated image and the ground-truth image. It is obtained by computing the cosine distances between their features of each layer in the VGG network [10] and averaging them.

Peak Signal-to-Noise Ratio (PSNR) is estimated using the ratio of the maximum possible power of the ground-truth image to the mean square error between the reconstructed image and the ground-truth to reflect the image reconstruction quality.

Structure Similarity Index Measure (SSIM). From the perspective of image distortion modeling discussed in [13], the structural similarity of two images is measured by considering three different factors, brightness, contrast, and structure, where the mean is used as the estimate of brightness, the standard deviation as the estimate of contrast, and the covariance as the measure of structural similarity.

Average Keypoint Distance (AKD) evaluates the average key-point difference between the generated face and the ground-truth. We extract the facial landmarks from the reconstructed image and the ground-truth image with the face alignment approach [2] and calculate the average distance between the two groups of landmarks.

*These authors contributed equally.
†Corresponding Author: bczhang@buaa.edu.cn.

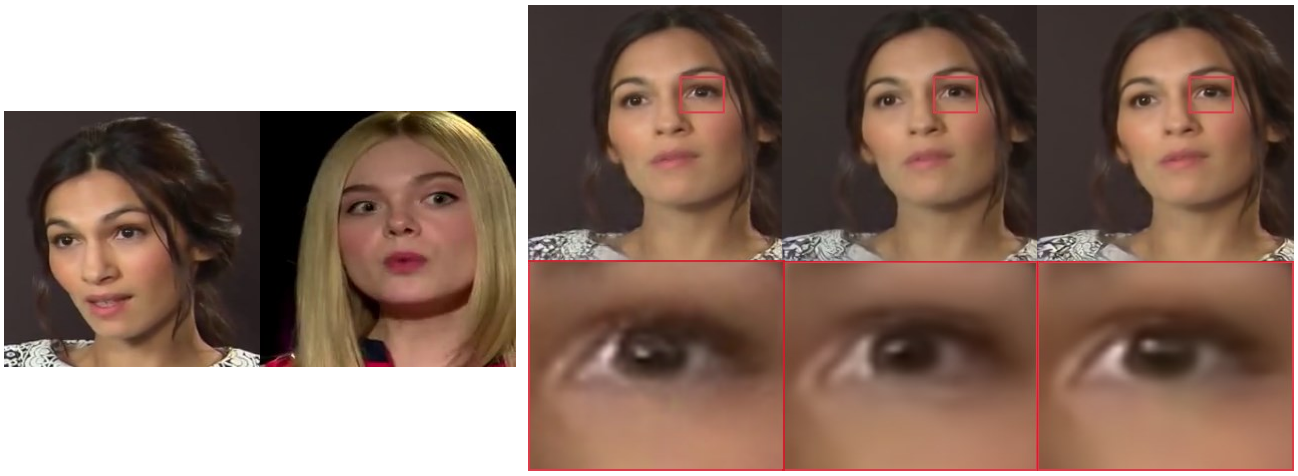

Figure 1. Using FADM to improve existing videos. **(a)** One frame of an existing video. **(b)** The improved results by FADM, **(c)** The result by the full generation process of FADM based on the source and driving frames. Note that **(b)** only takes **(a)** as input without the source and driving images. We can observe that FADM can enrich the fine facial details, such as hair, mouth, and eyes.

Average Euclidean Distance (AED) reflects the performance of preserving identity in the generated video. We use the face recognition network [1] to calculate the feature representations of the ground-truth image and the generated video frame and their averaged Euclidean distance.

Frechet Inception Distance score (FID) [4] evaluates the image quality by imitating human perception of image similarity. We leverage a pre-trained Inception V3 [11] to compare the distributions of the generated images with those of the ground-truth images.

Cosine Similarity (CSIM) measures the difference between two individuals by using the cosine value of the angle between two vectors in a vector space. We leverage the pre-trained CuricularFace [6] to compute the cosine similarity to assess the quality of identity preservation.

C. More Results

In this section, we provide more results of our FADM against the state-of-the-art (SOTA) methods.

Reenactment. In Fig. 2, we visualize more results covering different genders and multi-age groups with SOTA FOMM [9], DaGAN [5], and Face vid2vid [12] on the VoxCeleb [8], Voxceleb2 [3], and CelebA [7] datasets for the reenactment task. We see that these SOTA models suffer from the unnatural artifact problem and may encounter head distortions, while our FADM, using Face vid2vid [12] as the coarse generative module, performs better than other methods in rectifying the distortions and synthesizing fine-grained facial details on the three datasets.

Reconstruction. To further demonstrate the effectiveness of FADM, we select several difficult examples from the VoxCeleb [8] dataset with poses and expressions changing dramatically between the source images and driving frames, and show the comparison results with the SOTA on the reconstruction task in Fig. 3. These results show that FADM obtains more realistic details compared to other methods. We highlight some parts of the faces, and it is evident that FADM can alleviate unnatural artifacts and distortion effectively.

D. Ethic Impact

This work does not have a direct negative social impact. However, since it can synthesize realistic talking-head videos, which should be prevented from being abused for malicious purposes.

Figure 2. Visual comparisons of different methods on cross-identity face reenactment over three datasets.

Figure 3. Visual comparisons of same-identity face reconstruction on VoxCeleb [8]. We zoom in some parts of the faces to demonstrate the generation capability of FADM.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016. 2

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 1

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 2

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[5] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022. 2

[6] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *CVPR*, 2020. 2

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 2018. 2

[8] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 2, 4

[9] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 2

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2

[12] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 2

[13] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 1

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1