Supplementary Materials: StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN

Fei Yin[1], Yong Zhang[2†], Xiaodong Cun[2], Mingdeng Cao[1], Yanbo Fan[2], Xuan Wang[2], Qingyan Bai[1], Baoyuan Wu[3], Jue Wang[2], and Yujiu Yang[1†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Tencent AI Lab
[3] School of Data Science, Secure Computing Lab of Big Data, The Chinese University of Hong Kong, Shenzhen

1 Details of Audio-Driven Reenactment

We introduce the extension of our framework to tackle the audio-driven facial reenactment task via directly predicting the motion from audio features in the manuscript. Here, we complete the remaining details of network structure and training objectives.

Network Structure. The network structure of the audio-driven motion generator is similar to the proposed video-driven motion generator. Differently, the driving signal comes from audio. Thus, we transform the original audio to Mel-Spectrogram first. Then we use an MLP to squeeze the temporal dimension. Finally, these features are injected into the network via AdaIN.

Training Objectives. As for the loss function, similar to our video-driven motion generator, we calculate the loss between the driving image $\boldsymbol{I}_t$ and the warped image $\hat{\boldsymbol{I}}_{audio}$ using the perceptual loss and the $\mathcal{L}_1$ loss. Differently, we use a mask strategy to increase the weight of the mouth area. The mask is obtained by calculating the bounding-box of the landmark points around mouth. The loss is defined as:

$$\mathcal{L}_t^a = \sum_i \|M \cdot \phi_i(\boldsymbol{I}_t) - M \cdot \phi_i(\hat{\boldsymbol{I}}_{audio})\|_1 + \lambda_1^a \cdot \|M \cdot \boldsymbol{I}_t - M \cdot \hat{\boldsymbol{I}}_{audio})\|_1, \quad (1)$$

where λ_1^a is the hyper-parameter. And in practice, the mask M is in the soft form.

Besides, since the artifacts always happen in the masked region, we design a regularization loss to make sure the consistency of the non-masked region between the proxy input $\hat{\boldsymbol{I}}_{visual}$ and the warped image $\hat{\boldsymbol{I}}_{audio}$:

$$\mathcal{L}_{reg}^a = \sum_i \|(1-M) \cdot \phi_i(\hat{\boldsymbol{I}}_{visual}) - (1-M) \cdot \phi_i(\hat{\boldsymbol{I}}_{audio})\|_1. \quad (2)$$

† Corresponding author: Yong Zhang (zhangyong201303@gmail.com) and Yujiu Yang (yang.yujiu@sz.tsinghua.edu.cn).

Fig. 1. Study for which layer to operate on.

Finally, to make the lip movement be more consistent with audio, we employ a lip-sync discriminator D_{sync} that is trained for the synchronization between audio and video by SyncNet [1]. The synchronization objective can be defined as:

$$\mathcal{L}^a_{sync} = -\mathbb{E}[\sum_{t=i-2}^{i+2} \log(D(\hat{\boldsymbol{I}}_{audio}, \boldsymbol{a}))], \tag{3}$$

where D_{sync} requires 5 consecutive frames as input.

The total loss can be defined as:

$$\mathcal{L}^a = \mathcal{L}^a_t + \lambda^a_r \cdot \mathcal{L}^a_{reg} + \lambda^a_s \cdot \mathcal{L}^a_{sync}, \tag{4}$$

where λ^a_r and λ^a_s are the corresponding weights.

Joint Training with Calibration Network. To alleviate the artifacts introduced by warping, we joint train the whole framework after pre-training the motion generator. The objective is a weighted summation of the loss of calibration network and the audio-driven motion generator as follows:

$$\mathcal{L}^c = \mathcal{L}^c_t + \lambda^c_d \cdot \mathcal{L}^c_{domain} + \lambda^c_{adv} \cdot \mathcal{L}^c_{adv} + \beta^a \cdot \mathcal{L}^a, \tag{5}$$

where λ^c_d, λ^c_{adv}, and β^a are the corresponding weights.

2 Study on Warping Layer Selection

We investigate the strong spatial prior preserved in the feature space of pre-trained StyleGAN in Sec. 3 in the manuscript. Here, we provide a study to determine the proper layer for performing the spatial transformation. We first randomly sample the style latent code $\boldsymbol{w}$ in $\mathcal{W}^+$ space to generate a random face image with the pre-trained StyleGAN and obtain spatial features

$[\boldsymbol{f}_{4\times4}, \boldsymbol{f}_{8\times8}, \ldots, \boldsymbol{f}_{1024\times1024}]$ in $\mathcal{F}$ space. Then we perform a warping operation on each feature map individually and feed it with the rest fixed style code into StyleGAN. The results are shown in Fig. 1. We can observe that warping lower layers cannot accurately control pose and expression while warping higher layers yields ghost shadows on the synthetic image. In order to reduce the semantic information lost by subsequent operations as much as possible and reduce the number of parameters for the calibration network, we choose the layer $\boldsymbol{f}_{64\times64}$ as a balanced choice.

3 Additional Experiment Details

3.1 Settings

Dataset Preprocessing. We train the two motion generators on the VoxCeleb dataset [5]. Following [7], we preprocess the data by cropping faces from the videos and then resizing them to 256×256. Faces are not aligned and can move freely within a fixed bounding box. To be consistent with videos at 25 fps, we extract Mel Spectrogram as acoustic features $a \in \mathbb{R}^{80}$ per second by Fast Fourier Transform (FFT) in advance. We take the continuous temporal window coefficients as the current time feature. We joint train the whole framework on the HDTF dataset [10]. The resolution of original videos is $720P$ or $1080P$, which is higher than that of VoxCeleb. The videos are cropped in the same manner as processing VoxCeleb and then resized to 512×512.

Implementation Details. We train the two motion generators and the calibration network in two stages. In the first stage, we pre-train the video-based motion generator on VoxCeleb for 200K iterations. Then, we formulate training pairs for the audio-based motion generator by using the predicted motion as the pseudo label. We pre-train the audio-based generator with synthesized audio-motion pairs for 200K iterations. The trade-off hyper-parameters are set to $\lambda_1^a = 10$, $\lambda_r^a = 0.1$ and $\lambda_s^a = 1$. The optimizer for both pretraining processes is ADAM [4] with an initial learning rate of 10^{-4}. The batch size is set to 20 for all experiments.

In the second stage, we first jointly optimize the calibration network and the video-based motion generator in an end-to-end manner on HDTF for 20K iterations. The hyper-parameters are set to $\lambda_1^c = 10$, $\lambda_d^c = 0.01$, $\lambda_{adv}^c = 0.1$, $\beta^v = 0.01$, and $\beta^a = 0$. The learning rates are set to 10^{-4} and 2×10^{-5} for them, respectively. Then, we fix the video-based motion generator and jointly optimize the calibration network and the audio-based motion generator for 20K iterations. The hyper-parameters are set to $\lambda_1^c = 10$, $\lambda_d^c = 0.01$, $\lambda_{adv}^c = 0.1$, $\beta^v = 0$, and $\beta^a = 0.01$. The learning rates are set the same as the above optimization.

During inference, the two motion generators can be used individually or jointly. When using both of them, the video-based motion generator controls the head pose while the audio-based motion generator controls the lip movement.

The GAN inversion is used to get the spatial feature maps in our framework. Optimization techniques can achieve more accurate reconstruction results, but

Fig. 2. Qualitative results of audio-driven talking face generation.

they are not efficient. While learning-based techniques are much faster, they encounter lower reconstruction quality. Considering the efficiency, we exploit a state-of-the-art learning-based inversion method [8] during training. During inference, we first use [8] to obtain the style codes and feature maps. For motion transfer tasks, we further exploit an optimization-based inversion method [11] to optimize latent feature maps for 250 iterations with fixed style codes. For editing tasks, we directly use the style codes and feature maps from [8] to preserve edited semantic information.

Evaluation Metrics. We exploit a set of metrics to evaluate image quality and motion transfer quality. For image quality, Learned Perceptual Image Patch Similarity (LPIPS) [9], Peak signal-to-noise ratio (PSNR) are utilized as metrics to measure the reconstruction quality. Structural Similarity (SSIM) is utilized to measure the structural similarity between patches of the input images. Frechet Inception Distance (FID) [3] is utilized to measure the realism of the synthesized results. To measure identity preservation, we compute the cosine similarity (CSIM) of identity embeddings between the source images and the generated videos extracted from ArcFace [2]. For motion transfer quality, following [6], Average Expression Distance (AED), and Average Pose Distance (APD) are used to compute the differences between generated images and target images in terms of 3DMM expression and pose, respectively.

3.2 Additional Results

Results of Audio-Driven Talking Face Generation. In our framework, the audio-based motion generator can work either individually or jointly with the video-based motion generator. The visual results of both cases are illustrated in Fig. 2. The first row represents the videos that provide the audios. The first column represents the source portraits to be animated. Synthesized faces from the $2nd$ to $4th$ column are generated purely by the driving audio. While synthesized faces in the last three columns are generated according to both the driving video and audio. The driving video controls the head pose while the audio controls the lip movement.

For the audio-driven case, it can be observed that the generated lip movements are consistent with those of the ground-truth video for different source portraits. For the audio-and-video-driven case, the results show that the pose is accurately controlled by the video and the lip movements are still consistent with those of the video. Both the visual and acoustic control can generalize to different identities.

Attribute Editing Results. We provide additional global attribute editing results via modifying the style codes gradually in the video generation process. The edited attributes include decreasing age, increasing age, adding makeup and adding beard. The results are shown in Fig. 3 and Fig. 4.

3.3 Video Results.

To better demonstrate the temporal consistency and flexible editability of our synthesised results, we provide a demo video in the supplementary.

References

1. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV (2016)
2. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
6. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: ICCV (2021)
7. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NIPS (2019)
8. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. arXiv preprint arXiv:2109.06590 (2021)

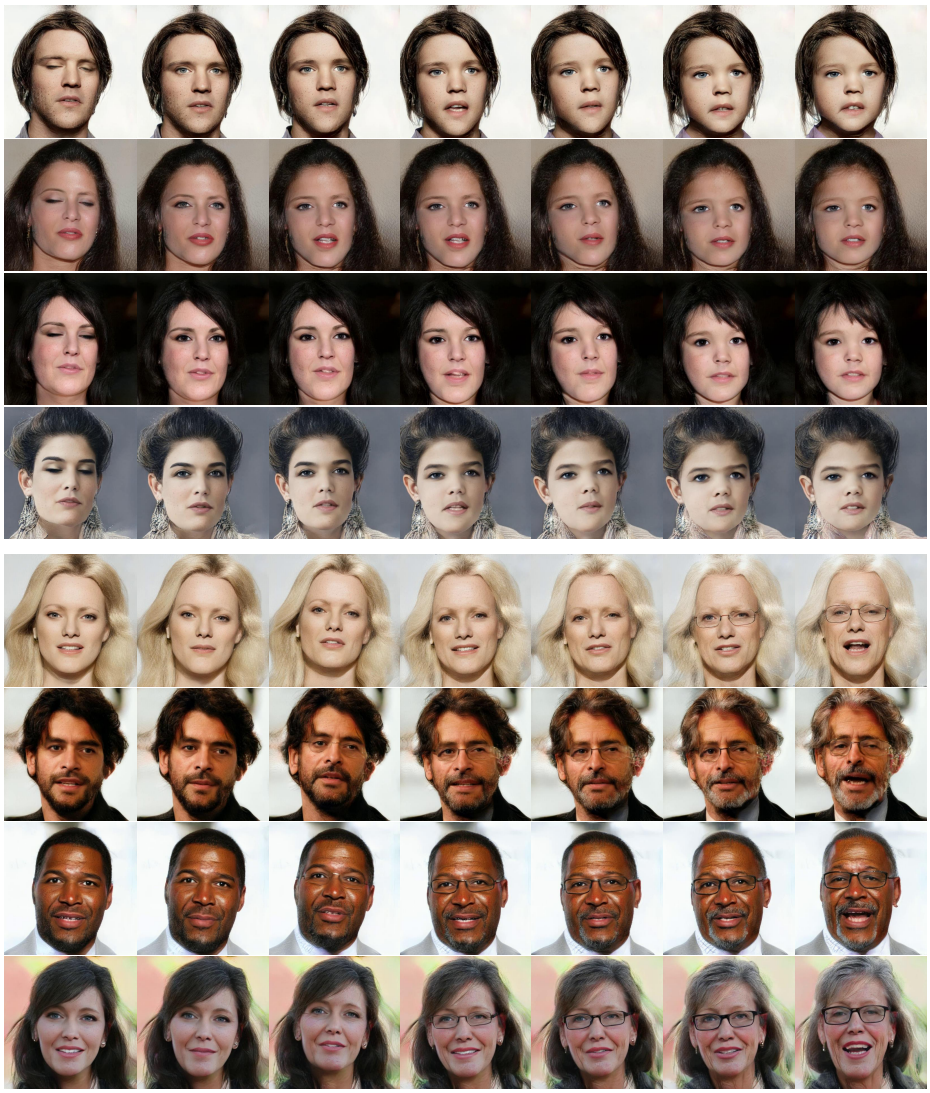

Fig. 3. Global attribute editing via GAN inversion. The attribute is gradually modified in each generated talking video.

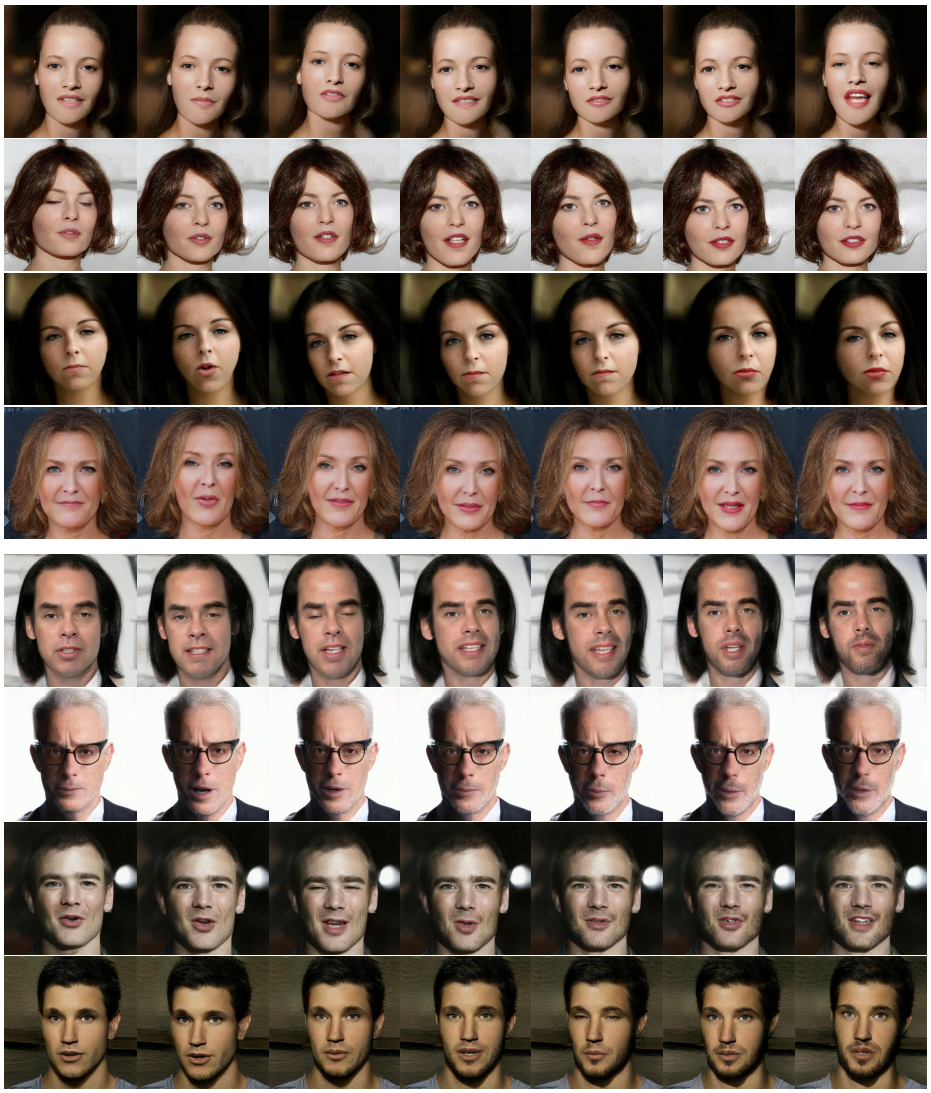

Fig. 4. Global attribute editing via GAN inversion. The attribute is gradually modified in each generated talking video.

9. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
10. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: CVPR (2021)
11. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Barbershop: Gan-based image compositing using segmentation masks. arXiv preprint arXiv:2106.01505 (2021)