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

Abstract. One-shot talking face generation aims at synthesizing a high-quality talking face video from an arbitrary portrait image, driven by a video or an audio segment. In this work, we provide a solution from a novel perspective that differs from existing frameworks. We first investigate the latent feature space of a pre-trained StyleGAN and discover some excellent spatial transformation properties. Upon the observation, we propose a novel unified framework based on a pre-trained StyleGAN that enables a set of powerful functionalities, *i.e.*, high-resolution video generation, disentangled control by driving video or audio, and flexible face editing. Our framework elevates the resolution of the synthesized talking face to 1024×1024 for the first time, even though the training dataset has a lower resolution. Moreover, our framework allows two types of facial editing, *i.e.*, global editing via GAN inversion and intuitive editing via 3D morphable models. Comprehensive experiments show superior video quality and flexible controllability over state-of-the-art methods. Code is available at https://github.com/FeiiYin/StyleHEAT.

1 Introduction

One-shot talking face generation refers to the task of synthesizing a high-quality talking face video from a given portrait image, guided by a driving video or audio segment. The synthesized face inherits the identity information from the portrait image, while its pose and expression are transferred from the driving video or generated based on the driving audio. Talking face generation has a variety of important applications such as digital human animation, film production, *etc.*.

Recent one-shot talking face generation methods [47,38,28] have made notable progress in driving expression and pose. However, they fail to generate high-resolution video frames. The video resolution of the ordinary methods still remains at 256×256. Few methods such as [38] and [47] have achieved the resolution of 512×512 by exploiting newly collected high-resolution datasets, but

†Corresponding author: Yong Zhang (zhangyong201303@gmail.com) and Yujiu Yang (yang.yujiu@sz.tsinghua.edu.cn). Work done during an internship at Tencent AI Lab.

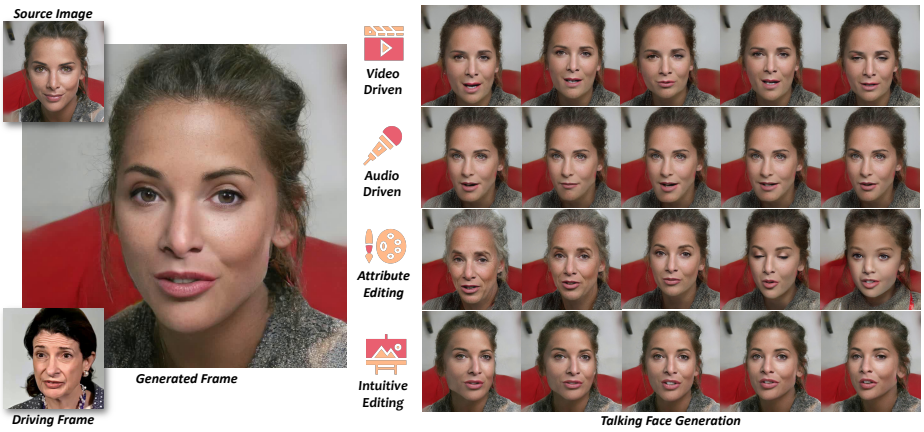

Fig. 1. Our unified framework enables high-resolution talking face generation, disentangled control by a driving video or audio, and flexible face editing.

they are still bounded by the resolution of the training data. More importantly, improving the resolution requires properly designed network architectures and training strategies. Adding upsampling layers in a straightforward way into the network usually does not work well. Warping-based methods can be directly applied to higher-resolution images but will introduce inevitable artifacts. [28] and [14] utilize a post refining network to eliminate the artifacts, but limit the resolution of the finally synthesised results at the same time.

Face editing is an useful technique to enhance talking face videos, *e.g.,* the modification of facial appearance, pose, and expression. It has two categories, *i.e.*, intuitive editing (*e.g.,* pose and expression) and semantic facial attribute editing (*e.g.,* makeup, beard, and age). Only few talking face generation methods [28,14] enable intuitive editing via 3D morphable models (3DMMs). But there is no existing work that incorporates semantic attribute editing into the talking face generation framework. Besides, almost all previous methods provide frameworks for either the video-driven case or the audio-driven case, but few consider both except for [28]. Therefore, it is another challenge to integrate the driving and editing modules of different modalities into a unified framework.

We raise two ambitious questions: can we further improve the resolution of one-shot talking face to 1024×1024 even though the existing datasets have a lower resolution? Can we build a unified framework that enables different types of driving modalities as well as semantic and intuitive face editing? To achieve these goals, we resort to a powerful pre-trained generative model: StyleGAN [23]. StyleGAN has shown impressive results in various applications, *e.g.,* facial attribute editing [11], blind image restoration [39], portrait stylization [34], *etc.*. These methods utilize the learned image prior of StyleGAN to facilitate downstream tasks, removing the need of training a large model from scratch. The resolution is retained at 1024×1024 and visual details are also reserved. Despite

these successes, to the best of our knowledge, there is no existing work that uses a pre-trained StyleGAN for one-shot talking face generation.

In this work, we first investigate the latent style space and the feature space of a pre-trained StyleGAN. The style space is also called $\mathcal{W}$ space. The style space is extensively explored by GAN inversion methods for face editing. The feature space is also called $\mathcal{F}$ space. In a talking face video, different facial expressions are achieved by deforming different facial regions in different ways. Given that style codes do not contain accurate spatial information, the style space might not be an appropriate choice for injecting facial motion. We then systematically study the feature space by applying a set of spatial transformations on the feature map of StyleGAN. Interestingly, we discover that the pre-trained model is robust to some geometric transformations as it can steadily generate high-quality images accordingly, indicating that the feature space has satisfying spatial properties.

Upon the above observation, we propose a novel unified framework for high-quality one-shot talking face generation based on a pre-trained StyleGAN. Specifically, we directly deform the StyleGAN spatial features using flow fields predicted by the video-based or audio-based motion generator, and then a calibration network is proposed to modulate the warped features. Such a design preserves the facial prior of the StyleGAN, enabling our model to generate high-resolution results while eliminating warping-induced flaws. Thanks to the pre-trained StyleGAN, our framework also allows two types of face editing, *i.e.,* global editing via GAN inversion and intuitive editing pose and expression based on 3DMM. Fig. 1 illustrates the functionalities of the proposed framework.

Our main contributions are as follows:

- We propose a unified framework based on a pre-trained StyleGAN for one-shot talking face generation. It enables high-resolution video generation, disentangled control by driving video and audio, and flexible face editing.
- We conduct comprehensive experiments to illustrate the various capabilities of our framework and compare it with many state-of-the-art methods.

2 Related Work

3D structure-based talking-face generation. Traditionally, 3D faces model priors (such as 3DMM [7]) provide a powerful tool for rendering and editing the portrait images by the parameters modulation. For example, DVP [25] modifies the parameters from source and target, then, a network is used to render the shading to video. Recent 3D model-based methods [25,16,14,28,10] can also do a good job for subject-agnostic face synthesis. HeadGAN [14] pre-processes the 3d mesh as input. PIRenderer [28] predicts a flow field for feature warping.

2D-based talking-face generation. Instead of controlling the model parameters, mimicking the motions of another individual by the neural network is also a popular direction. Subject-agnostic approaches [9,4,30,32,38,31], which only need a single image of the target person are the most popular type. For the representative methods, Monkey-Net [31] propose a network to transfer the deformation

Fig. 2. Latent feature space investigation of a pre-trained StyleGAN. Different geometric transformations are applied to modify the feature maps.

from sparse to dense motion flow. FOMM [30] extends Monkey-Net via the first-order local affine transformations. Then, Face-vid2vid [38] improves FOMM via a learned 3D unsupervised key-points for free-view talking head generation. A concurrent work [8] also tries to explore the style space for face animation, but they need to fine-tune the pre-trained StyleGAN on the specific domain.

3 Investigating Feature Space of StyleGAN

StyleGAN2 [24] draw attention from the community since it can generate high-quality face images and the feature space is highly disentangled. To allow a pre-trained StyleGAN [24] for high-resolution talking-head video generation, one possible direction is StyleGAN based video generation [35,15], where they learn to generate videos via discovering an ideal trajectory in $\mathcal{W}^+$ latent space. However, the motion is randomly sampled without any control and the content is corrupted when the current pose differs from the initial one. This is because $\mathcal{W}^+$ is a highly semantic-condensed space and lacks explicit spatial prior [37]. Moreover, editing in $\mathcal{W}^+$ space [29,41,3,2,1,36] only allows changing high-level facial attributes, which cannot generate out-of-alignment images [20] since the StyleGAN is trained on aligned faces.

Thus, image editing in $\mathcal{F}$ feature space draws our close attention. Specifically, the latent code $\boldsymbol{f}$ in $\mathcal{F}$ feature space represents a spatial feature map in the generator. For StyleGAN [24], we define $\boldsymbol{f}$ as the feature map after a pair of upsampling and convolution layers at a certain scale. There are only a few previous methods [37,48,20,39,5] that edit the spatial features for GAN inversion [37,20,5], image composition [48], and blind face enhancement [39]. These approaches harvest the potential of spatial feature space editing and apply the spatial modulation (*e.g.* spatial feature transformation [40]) to the features. However, it has not been fully investigated whether the feature space of a pre-trained

StyleGAN can still be used to generate realistic images after various geometric transformations.

We therefore conduct a detailed experiment to verify the spatial property of StyleGAN features and fully excavate its potential capability. 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. At the same time, various spatial features $[\boldsymbol{f}_{4\times4}, \boldsymbol{f}_{8\times8}, \ldots, \boldsymbol{f}_{1024\times1024}]$ in $\mathcal{F}$ space can be obtained. We choose the feature resolution of 64×64 for a balanced trade-off between the inversion quality and editing capacity. To test the spatial property of the pre-trained StyleGAN features, several geometric transformations, including translation, rotation, zoom, shear, occlusion and Thin Plate Spline (TPS [42]), are used to manipulate $\boldsymbol{f}_{64\times64}$ directly. Finally, the transformed image can be generated by the forward pass with the edited feature map as input. Our experimental results are shown in Fig. 2. Either with simple affine transformations or complicated TPS deformations, we observe that the generated images maintain the same geometric changes as the deformations applied in the feature space. The generated images also share the same identity and appearance with a minor difference. This phenomenon demonstrates that the learned convolution kernels in the pre-trained generator perform in a transformation-invariant manner.

Based on the above observation, we can conclude that the features in a pre-trained StyleGAN preserve strong spatial prior and can be directly modified with geometric transformations. This spatial property makes it a promising direction to edit the feature space for talking face generation.

4 Methodology

We are interested in the task of controllable talking-head generation. Let I be the source image and $\{d_1, d_2, \cdots, d_N\}$ be a talking-head video, where d_i is the i-th video frame and N is the total number of frames. An ideal framework is supposed to generate video $\{y_1, y_2, \cdots, y_N\}$ with the same identity as I and the consistent motions derived from $\{d_1, d_2, \cdots, d_N\}$.

Inspired by our observation in Sec. 3, we propose a unified framework based on the $\mathcal{F}$ space excavation of the pre-trained StyleGAN G. As shown in Fig. 3, our approach contains several steps to achieve this goal. Given a single source image, we first use the GAN inversion method [37] to get the latent style code $\boldsymbol{w}$ and feature maps $\boldsymbol{f}$ of the source image. Then, to inject the accurate motion guidance, we predict a dense flow field by the motion generator Φ_{warp} from video (Sec. 4.1). Since the warping operation may introduce artifacts due to the occlusions and error mapping, a calibration network Φ_{cali} is introduced to renovate the edited spatial feature map (Sec. 4.2). Our framework can be extended to audio-driven via similar flow prediction module (Sec. 4.3). The whole framework can be summarized as:

$$\hat{I}_i = G\left(\Phi_{cali}(\Phi_{warp}(I, d_i) \circ \boldsymbol{f}), \boldsymbol{w}\right), \tag{1}$$

where $\circ$ denotes the warping transformation.

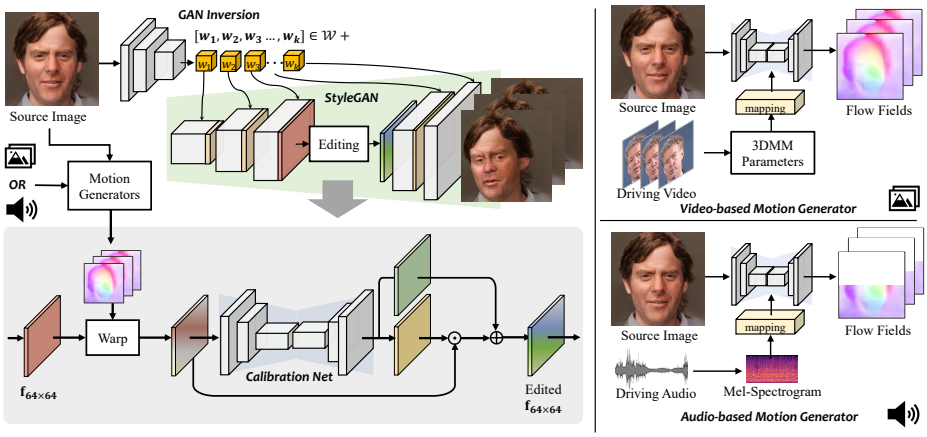

Fig. 3. The pipeline of our unified framework. The framework consists of four components, *i.e.,* a pre-trained StyleGAN, a video-driven motion generator, an audio-driven motion generator, and a calibration network. Given a source image, we can obtain the style codes and feature maps by the encoder of GAN inversion. The driven video or audio along with the source image are used to predict motion fields by the corresponding motion generator. The selected feature map is warped by the motion fields, followed by the calibration network for rectifying feature distortions. The refined feature map is then fed into the StyleGAN for the final face generation.

4.1 Video-Driven Motion Generator

The goal of the video-driven motion generator is to generate dense flows with the driving video and the source image as inputs. Then, these flow fields will manipulate the feature map of the pre-trained StyleGAN for talking face generation. In this part, we first demonstrate the intermediate motion representation in our settings. Then, we give the details of the network structure and the training process for the dense motion field generation.

Motion Representation. To achieve accurate and intuitive motion control, semantic medium plays an important role in the generation process. Following previous works [28,14], we take advantage of the 3DMM [6] parameters for motion modeling. In 3DMM, the 3D shape $\boldsymbol{S}$ of a face can be decoupled as:

$$\boldsymbol{S} = \overline{\boldsymbol{S}} + \boldsymbol{\alpha}\boldsymbol{U}_{id} + \boldsymbol{\beta}\boldsymbol{U}_{exp}, \tag{2}$$

where $\overline{\boldsymbol{S}}$ is the average shape, $\boldsymbol{U}_{id}$ and $\boldsymbol{U}_{exp}$ are the orthonormal basis of identity and expression of LSFM morphable model [7]. Coeffcients $\boldsymbol{\alpha} \in \mathbb{R}^{80}$ and $\boldsymbol{\beta} \in \mathbb{R}^{64}$ describe the person identity and expression, respectively. To preserve pose variance, coefficients $\boldsymbol{r} \in SO(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$ denote the head rotation and translation. Then, we can model the motion of the driving face with a parameter set $\boldsymbol{p} = \{\boldsymbol{\beta}, \boldsymbol{r}, \boldsymbol{t}\}$ extracted by an existing 3D face reconstruction model [13].

Due to the inevitable prediction errors between consecutive frames in the same video, the parameters from a single input frame will cause jitter and instability in the finally generated video. Hence, we adopt a windowing strategy for better temporal consistency, where the parameters of the neighboring frames are also taken as the descriptor of the center frame to smooth the motion trajectory. Thus, the motion coefficient of the i-th driving frame is defined as:

$$\boldsymbol{p}_i \equiv \boldsymbol{p}_{i-k:i+k} \equiv \{\boldsymbol{\beta}_{i-k}, \boldsymbol{r}_{i-k}, \boldsymbol{t}_{i-k}, \ldots, \boldsymbol{\beta}_i, \boldsymbol{r}_i, \boldsymbol{t}_i, \ldots, \boldsymbol{\beta}_{i+k}, \boldsymbol{r}_{i+k}, \boldsymbol{t}_{i+k}\}, \tag{3}$$

where k is the radius of the window.

Network Structure. Our network is built on a U-Net structure that requires the source image and the driving video as inputs, and the outputs are the desired flow fields for feature warping. It contains a 5-layer convolutional encoder and a 3-layer convolutional decoder for multi-scale feature extraction. We use the 3DMM parameters $\boldsymbol{p}_t$ from the driving frame d_t as the motion representation. Specifically, these parameters are first mapped to a latent vector via a 3-layer MLP to aggregate the temporal information. Then, the motion parameters are injected into each convolutional layer via the adaptive instance normalization (AdaIN [18]). Next, the network can be trained by the source image I and the motion condition $\boldsymbol{p}_t$ as inputs. Finally, the loss functions will be calculated between the target image d_t and the generated image by the backward warping.

Pre-training Strategy. As we have investigated in Sec. 3, the feature warping shares the same geometry deformation with the final image. Thus, to simplify the learning of the whole framework, before joint training we first pre-train the video-based motion generator on the widely-used talking-face datasets (VoxCeleb [26]) to generate trustful flow fields in a self-supervised manner. Then, the low-resolution flow field are used to drive the spatial feature map of the pre-trained StyleGAN. Specifically, as the ground truth flow fields are not available, we predict the flow fields $\boldsymbol{n}$ using the network, and then the source frame $\boldsymbol{I}$ will be used to calculate the warped frame by $\hat{\boldsymbol{I}}_n = \boldsymbol{I} \circ \boldsymbol{n}$. Then, given the target frame $\boldsymbol{I}_t$, we use the perceptual loss [19] to calculate the $\mathcal{L}_1$ distance between the activation maps of the pre-trained VGG-19 network [33]:

$$\mathcal{L}^v = \sum_i \|\phi_i(\hat{\boldsymbol{I}}_n) - \phi_i(\boldsymbol{I}_t)\|_1, \tag{4}$$

where ϕ_i denotes the activation map of the i-th layer of the VGG-19 network. Similar to [30], we calculate the perceptual loss on a number of resolutions by applying pyramid down-sampling on $\boldsymbol{I}_t$ and $\hat{\boldsymbol{I}}_n$. After training, the generated flow field can be used to edit the feature map of StyleGAN.

4.2 Feature Calibration and Joint Training

The video-driven motion generators are pre-trained without considering any information about the pre-trained StyleGAN. Though the predicted motion fields can be used to warp the feature map of StyleGAN, it will inevitably introduce artifacts. For example, making a closed mouth open through 2D warping cannot

fill correct teeth within the mouth. To alleviate the feature map distortion, we introduce a calibration network to rectify artifacts in the feature space.

Calibration Network. A calibration network is needed since the warped features still suffer from artifacts. As shown in Fig. 3, we adopt a U-Net architecture to extract multi-resolution spatial features. It consists of a 4-layer encoder and a 4-layer decoder. We feed the warped feature map $\boldsymbol{f}_w$ as the network's input. Then, the multi-scale conventional layers are used to refine the warped features. However, due to the high complexity of the intermediate features, instead of directly predicting the features, our calibration network performs the spatial feature transformation (SFT [38]) to the warped features, which is defined as:

$$\hat{\boldsymbol{f}}_c = SFT(\boldsymbol{f}_w|\boldsymbol{r}, \boldsymbol{t}) = \boldsymbol{r} \odot \boldsymbol{f}_w + \boldsymbol{t}, \tag{5}$$

where $\odot$ denotes element-wise multiplication. Then, the final high-quality and high-resolution result can be achieved as $\hat{I} = G(\hat{\boldsymbol{f}}_c, \boldsymbol{w})$.

Overall End-to-end Training. Directly applying the introduced calibration network is easy to encounter blur results (as shown in Fig. 12) since the quality of the frames in the video dataset is much lower than the high-resolution face dataset for training StyleGAN. Furthermore, inevitable detail lost of identity, attribute, texture, and background raised by the GAN inversion method will enlarge the gap between the generated images and the real images, which will further mislead the direction of the optimization.

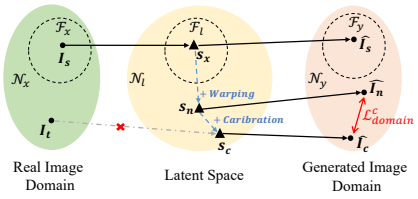

Fig. 4. Illustration of the domain loss.

Thus, we joint train the whole network except the pre-trained StyleGAN and design loss functions to solve the above problem. We first design a domain loss to restrict the differences between the reconstructed image of the warped feature map and that of the calibrated feature map in the generated image domain. As shown in Fig. 4, given a natural source image $\boldsymbol{I}_s$ in the aligned StyleGAN space $\mathcal{F}_x$, the GAN inversion method can invert and reconstruct the image in the latent space and the generated image domain, respectively. Differently, for the target image $\boldsymbol{I}_t$ which is out of the aligned domain, GAN inversion is hard to be applied. Thus, to obtain the desired latent space $\boldsymbol{s}_c$, the proposed method utilizes the flow fields to edit the images in the latent space. After editing, the warped feature $\boldsymbol{s}_n$ may not be in the aligned StyleGAN latent space anymore but it can still generate a high-quality image $\hat{\boldsymbol{I}}_n$ by forwarding pass as we have discussed in Sec. 3. Unfortunately, the warping artifacts may occur because of the low quality of the flow fields. Thus, we propose the calibration network to further edit the feature map as introduced previously. However, the results $\hat{\boldsymbol{I}}_c$ become blurry due to the feature shift. To preserve both advantages of $\hat{\boldsymbol{I}}_n$ and $\hat{\boldsymbol{I}}_c$, the domain loss is defined to measure their difference. Further, we take a masking strategy to enhance the weight of different areas. The calibration mask M is

comprised of the bounding boxes of the eyes and mouth because the artifacts often occur around them. Thus, the domain loss is:

$$\mathcal{L}^c_{domain} = \sum_i \|(1-M)\cdot\phi_i(\hat{\boldsymbol{I}}_n) - (1-M)\cdot\phi_i(\hat{\boldsymbol{I}}_c)\|_1. \tag{6}$$

Besides, for eliminating the artifacts of local facial features, the driving image $\boldsymbol{I}_t$ provides the most accurate high-frequency information. Hence, we calculate the $\mathcal{L}_1$ loss and the perceptual loss with the ground truth, which is weighted on the masked region:

$$\mathcal{L}^c_t = \sum_i \|M\cdot\phi_i(\boldsymbol{I}_t) - M\cdot\phi_i(\hat{\boldsymbol{I}}_c)\|_1 + \lambda^c_1\cdot\|M\cdot\boldsymbol{I}_t - M\cdot\hat{\boldsymbol{I}}_c\|_1, \tag{7}$$

where λ^c_1 is the weight of the $\mathcal{L}_1$ loss.

Finally, to maintain the high fidelity of face generation, we also impose adversarial loss. Note that we freeze the parameters of the discriminator since the low-quality video frames may decline its performance. The adversarial loss is:

$$\mathcal{L}^c_{adv} = -\mathbb{E}[\log(D(\hat{\boldsymbol{I}}_c))], \tag{8}$$

where D is a well-trained discriminator of StyleGAN2.

The framework is trained in an end-to-end manner together with the loss of the corresponding motion generators. Here, we calculate the perceptual loss between the intermediate results from the motion generator and the ground truth, which is the same as Eq. 4. The weight of other components (the StyleGAN generator and the inversion encoder) are frozen.

In summary, the overall loss is a weighted summation as follows:

$$\mathcal{L}^c = \mathcal{L}^c_t + \lambda^c_d\cdot\mathcal{L}^c_{domain} + \lambda^c_{adv}\cdot\mathcal{L}^c_{adv} + \beta^v\cdot\mathcal{L}^v, \tag{9}$$

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

4.3 Extension on Audio-Driven Reenactment

We can further extend our framework to tackle the audio-driven facial reenactment task by extracting sequential motions from audio input. Audio-driven motion transfer is similar to video-driven motion transfer, but requires modeling the relationships between audio and face motions. Directly predicting the visual semantic parameters from audio information only is a difficult task and the two-stage converting procedure may accumulate errors. Consequently, we directly predict the motion from audio features as shown in the audio-based motion generator of Fig. 3.

Fig. 5. Paired training data generation for audio-driven motion generator training.

In detail, we train the generator to predict the flow fields in the lower half face, since audio is closely related to lip movements. However, a major challenge, generating a video from audio lacks a paired dataset because the videos with the same pose but different lip shapes are hard to obtain. To address this issue, we construct the paired data with the same pose but different expressions under different audio conditions by utilizing the pre-trained video-driven motion generator in Sec. 4.1. Specifically, we generate the proxy input by mixing the 3DMM parameters extracted from the source and driving frame, *i.e.,* the proxy input has the same pose as the driving frame and the same expression as the source frame. We illustrate the main process in Fig. 5, where the head pose of the proxy input is high-aligned with the driving frame. By training on the paired dataset, our audio-driven generator will focus on the flow generation of expression.

After training the audio-driven motion generator, it can be added to the framework as a plugin to control the lip movement independently. The details of the network structure and the training procedure will be discussed in the supplementary materials.

5 Experiments

Datasets. We train the two motion generators on the VoxCeleb dataset [26] which consists of over 100K videos of $1,251$ subjects. We joint train the whole framework on the HDTF dataset [47] which consists of 362 videos of over 300 subjects. HDTF is split into non-overlapping training and test sets. The test set contains 20 videos with around 10K frames. For cross-identity motion transfer evaluation, we select 1K high-resolution images from the CelebA-HQ dataset [21].
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. Then, we pre-train the audio-based generator with synthesized audio-motion pairs. As the motion from the pre-trained generators cannot be seamlessly applied to feature maps of StyleGAN, we need to finetune them along with the calibration network in the second stage. During inference, the two motion generators can be used individually or jointly.

The GAN inversion [44] is used to get the spatial feature maps in our framework. We exploit a learning-based inversion method [37] during training and an optimization-based inversion method [48] to optimize latent feature maps for more accurate reconstruction during inference.
Evaluation Metrics. We use the following metrics for evaluation: Learned Perceptual Image Patch Similarity (LPIPS) [46], Peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM), Frechet Inception Distance (FID) [17], the cosine similarity (CSIM) of identity embeddings extracted from ArcFace [12], Average Expression Distance (AED) [28], and Average Pose Distance (APD) [28].

5.1 Video-Driven Face Reenactment

To evaluate the performance of video-driven motion transfer, we conduct two facial reenactment tasks, *i.e.,* same-identity reenactment and cross-identity reen-

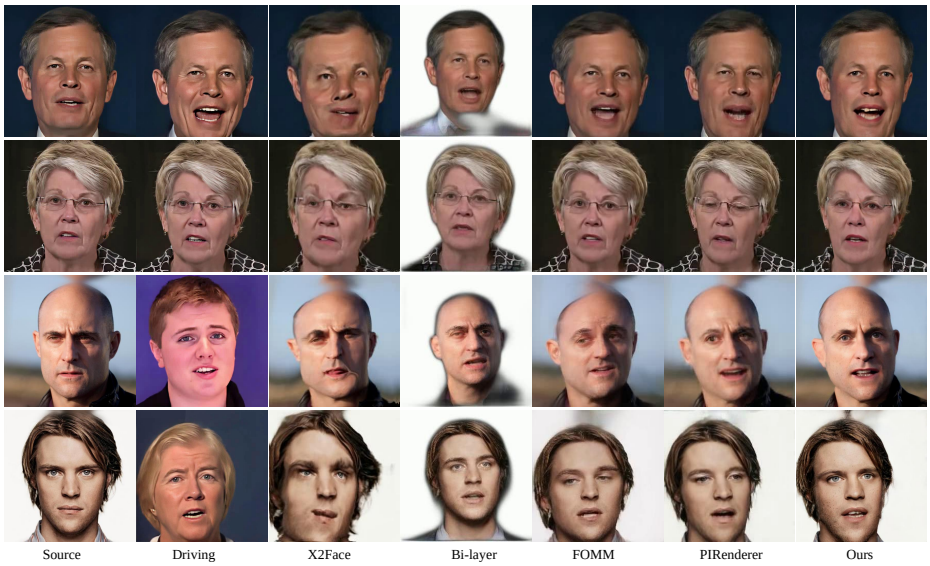

Fig. 6. Qualitative comparisons with state-of-the-art methods on the task of same-identity reenactment and cross-identity reenactment.

	Same-Identity Reenactment							Cross-Identity Reenactment			
	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	CSIM ↑	AED ↓	APD ↓	FID ↓	CSIM ↑	AED ↓	APD ↓
X2Face [43]	44.32	0.2687	31.09	0.5926	0.6965	0.1680	0.03719	128.19	0.4449	0.3415	0.05156
Bi-layer [45]	118.46	0.5758	28.23	0.2906	0.3033	0.1219	0.01322	189.64	0.2252	0.2654	**0.02054**
FOMM [30]	29.17	0.2036	31.12	**0.6353**	**0.8121**	**0.0946**	**0.00914**	108.93	0.4517	0.2692	0.02576
PIRenderer [28]	27.14	0.2252	30.96	0.6028	0.7797	0.1073	0.01459	108.56	0.4812	**0.2554**	0.02962
Ours	**18.02**	**0.1729**	**31.21**	0.6019	0.7475	0.1151	0.01664	**91.28**	**0.4890**	0.2630	0.03484

Table 1. Quantitative comparisons on talking face motion transfer.

actment. For the same-identity reenactment, the identity of the source portrait is the same as that of the driving video. For cross-identity reenactment, the identity of the source portrait differs from that of the driving video.

Qualitative Evaluation. The visual results of the same-identity and cross-identity are shown in Fig. 6. Our method can achieve superior image resolution and quality over other methods. Here we focus on other aspects. In the same-identity case, all methods perform well in transfer pose except X2Face. For expression, our method outperforms other methods when there is a large expression difference between the source and driving images, especially when the mouth of the source is closed while that of the driving image is opened by a large margin. In the cross-identity case, more issues occur for other methods while our method can work stably.

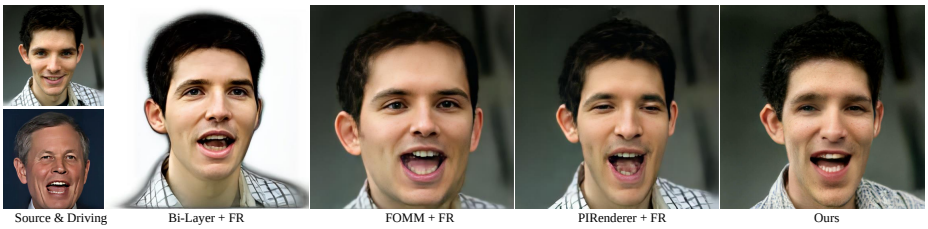

Fig. 7. Comparisons with enhanced state-of-the-art methods. We use face restoration (FR) method GFP-GAN [39] to enhance the visual quality of these competing methods.

Fig. 8. Global attribute editing via GAN inversion. The attribute is gradually modified in each generated talking video.

Quantitative Evaluation. Quantitative results of the two reenactment tasks are shown in Table 1. Our FID is the best in the cases, which indicates our synthesized faces are more realistic than those of other methods. Our better LPIPS and PSNR mean that we have better reconstruction performance. As our method uses a GAN inversion method to get the feature maps, it inevitably loses some identity information in the reconstruction. This might cause our lower CSIM in the same-identity case. On the contrary, our best CSIM in the cross-identity case indicates that our method can work stably in this more challenging setting and suffer from less distortion. Meanwhile, our AED and APD show comparable results to other methods although the StyleGAN inversion is imperfect and the extracted facial parameters will be further distorted.

Comparisons with Enhanced Methods. To eliminate the effect of resolution on our comparisons, we combine competing methods with a state-of-the-art blind face restoration method, *i.e.,* GFP-GAN [39], to improve the resolution and image quality. GFP-GAN improves the resolution of these methods to 1024 × 1024, which is shown in Fig. 7. We can observe that GFP-GAN greatly improves the visual quality for these low-resolution methods. However, it brings some side effects, including skin over-smoothing, details missing and color tone changing. Moreover, the face restoration cannot remedy the generated artifacts. The results

Fig. 9. Comparison with wav2lip[27].

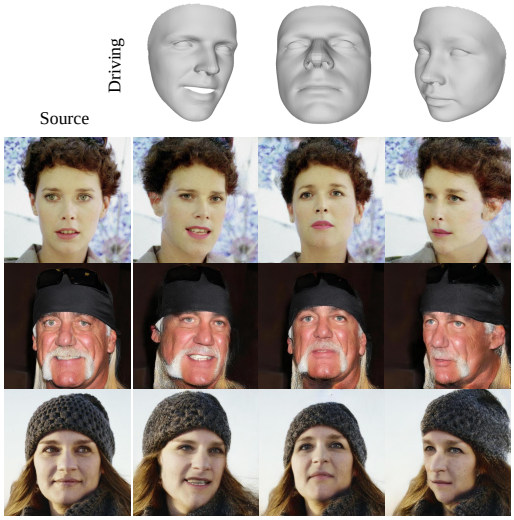

Fig. 10. The results of intuitive face editing.

demonstrate that our method outperforms the competing methods in terms of image quality even though they are enhanced by the face restoration method.

5.2 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. We compare with the state-of-the-art audio-driven method, wav2lip [27]. As shown in Fig. 9, our method have much better visual quality. The mouth of wav2lip is blurred and no teeth are synthesized. More results are shown in the supplementary material.

5.3 Talking Face Video Editing

Global Attribute Editing. As our model is built upon a pre-trained StyleGAN, it inherits a powerful property of StyleGAN, *i.e.,* facial attribute editing in the latent style space via existing GAN inversion methods. Our framework is convenient to edit attributes globally. We apply the GAN inversion method [37] to obtain the latent style codes for the first frame. Then, we can freely apply pre-defined style directions to change the style codes with a controllable extent in the video generation process at any timestamp. The results are shown in Fig. 8.
Intuitive Editing. Our video-based motion generator uses 3DMM parameters of the driving image to guide the motion generation for the source image. As 3DMM based talking face generation methods [28,14] always enable the intuitive editing on pose and expression, this also enables us to control the motion generation by directly modifying the 3DMM parameters, resulting in the intuitive editing on the final synthesis. The results are shown in Fig. 10.

Fig. 11. Ablation study of calibration net

Fig. 12. Ablation study of the domain loss

5.4 Ablation Study

Calibration Network. Directly applying the flow fields to the feature map will lead to apparent artifacts around eyes and mouth, *e.g.,* 2D warping is unable to generate teeth for a closed mouth. Hence, we design the calibration network to rectify the artifacts. We compare the performance with or without the calibration network. The results are shown in Fig. 11. The calibration network greatly improves the shape and content around the eyes and mouth.
Domain Loss. The calibration network modifies the feature maps. To prevent the edited feature maps from going far away from the original feature maps, we design the domain loss. We compare the performance with or without it. As shown in Fig. 12, we can observe that dropping the loss makes the synthetic images blurry and lose facial details such as wrinkles and hair texture.

6 Conclusion

We propose a novel framework for one-shot talking face generation based on a pre-trained StyleGAN by exploring the properties of the latent feature space. Our framework supports video-driven and audio-driven reenactment. Besides, our framework allows two types of face editing, *i.e.*, global attribute editing via GAN inversion and intuitive editing based on 3DMM. We conduct comprehensive experiments to illustrate various capabilities of our unified framework.
Limitation and Discussion. As proposed in [22], there exist texture-sticking artefacts of images generated by StyleGAN2, which means the hair and face in synthesised videos typically do not move in unison. Alias-Free GAN [22] designs a specific architecture to overcome the problem. Our framework can be migrated to the new generator when high-quality GAN inversion methods are studied.
Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under grant No. 61991450 , the Shenzhen Key Laboratory of Marine IntelliSense and Computation under grant NO.ZDSYS20200811 142605016. Baoyuan Wu is supported by Shenzhen Science and Technology Program under grant No.ZDSYS20211021111415025.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: CVPR (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: CVPR (2020)
3. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: ICCV (2021)
4. Anonymous: Latent image animator: Learning to animate image via latent space navigation. In: ICLR (2022)
5. Bai, Q., Xu, Y., Zhu, J., Xia, W., Yang, Y., Shen, Y.: High-fidelity gan inversion with padding space. arXiv preprint arXiv:2203.11105 (2022)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
7. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: CVPR (2016)
8. Bounareli, S., Argyriou, V., Tzimiropoulos, G.: Finding directions in gan's latent space for neural face reenactment. arXiv preprint arXiv:2202.00046 (2022)
9. Burkov, E., Pasechnik, I., Grigorev, A., Lempitsky, V.: Neural head reenactment with latent pose descriptors. In: CVPR (2020)
10. Cao, M., Huang, H., Wang, H., Wang, X., Shen, L., Wang, S., Bao, L., Li, Z., Luo, J.: Unifacegan: A unified framework for temporally consistent facial video editing. IEEE TIP (2021)
11. Chen, A., Liu, R., Xie, L., Chen, Z., Su, H., Yu, J.: Sofgan: A portrait image generator with dynamic styling. arXiv preprint arXiv:2007.03780 (2020)
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
13. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: CVPR Workshops (2019)
14. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: ICCV (2021)
15. Fox, G., Tewari, A., Elgharib, M., Theobalt, C.: Stylevideogan: A temporal generative model using a pretrained stylegan. arXiv preprint arXiv:2107.07224 (2021)
16. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. TOG (2019)
17. 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)
18. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
20. Kang, K., Kim, S., Cho, S.: Gan inversion for out-of-range images with geometric transformations. In: CVPR (2021)
21. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
22. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NIPS (2021)

23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
25. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. TOG (2018)
26. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
27. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: ACM Multimedia (2020)
28. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: ICCV (2021)
29. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: CVPR (2021)
30. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NIPS (2019)
31. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: CVPR (2019)
32. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: CVPR (2021)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
34. Song, G., Luo, L., Liu, J., Ma, W.C., Lai, C., Zheng, C., Cham, T.J.: Agilegan: stylizing portraits by inversion-consistent transfer learning. TOG (2021)
35. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. In: ICLR (2021)
36. Tzaban, R., Mokady, R., Gal, R., Bermano, A.H., Cohen-Or, D.: Stitch it in time: Gan-based facial editing of real videos. arXiv preprint arXiv:2201.08361 (2022)
37. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. arXiv preprint arXiv:2109.06590 (2021)
38. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: CVPR (2021)
39. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: CVPR (2021)
40. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018)
41. Wei, T., Chen, D., Zhou, W., Liao, J., Zhang, W., Yuan, L., Hua, G., Yu, N.: A simple baseline for stylegan inversion. arXiv preprint arXiv:2104.07661 (2021)
42. Wikipedia contributors: Thin plate spline — Wikipedia, the free encyclopedia (2020), https://en.wikipedia.org/wiki/Thin_plate_spline
43. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV (2018)
44. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
45. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: ECCV (2020)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

47. 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)
48. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Barbershop: Gan-based image compositing using segmentation masks. arXiv preprint arXiv:2106.01505 (2021)