The Visual Computer (2024) 40:4913–4925
https://doi.org/10.1007/s00371-024-03490-4

RESEARCH

# Manitalk: manipulable talking head generation from single image in the wild

Hui Fang[1] · Dongdong Weng[2] · Zeyu Tian[1] · Yin Ma[3]

Accepted: 13 May 2024 / Published online: 8 June 2024

**Abstract**
Generating talking head videos through a face image and a piece of speech audio has gained widespread interest. Existing talking face synthesis methods typically lack the ability to generate manipulable facial details and pupils, which is desirable for producing stylized facial expressions. We present ManiTalk, the first manipulable audio-driven talking head generation system. Our system consists of three stages. In the first stage, the proposed Exp Generator and Pose Generator generate synchronized talking landmarks and presentation-style head poses. In the second stage, we parameterize the positions of eyebrows, eyelids, and pupils, enabling personalized and straightforward manipulation of facial details. In the last stage, we introduce SFWNet to warp facial images based on the landmark motions. Additional driving sketches are input to generate more precise expressions. Extensive quantitative and qualitative evaluations, along with user studies, demonstrate that the system can accurately manipulate facial details and achieve excellent lip synchronization. Our system achieves state-of-the-art performance in terms of identity preservation and video quality. Code is available at https://github.com/shanzhajuan/ManiTalk.

**Keywords** Facial animation · Expression manipulation · Gaze manipulation · Neural network

## 1 Introduction

Talking head generation has gained widespread interest in multimodal human–computer interaction in recent years. It is crucial for filmmaking, virtual avatars, video conferencing, and virtual education. Precise lip movements and realistic video portraits are essential for enhancing user feedback. Natural head movements, along with facial details like blinking, eye movements, and eyebrow motions, also contribute to an improved user experience. Most existing methods focus on improving lip synchronization [24] and image clarity [18] and generating stylized head motions [39, 44, 48]. Generating realistic and manipulable talking videos, which are desirable for generating diverse talking styles in complex scenarios, is typically overlooked by talking face synthesis methods.

Generating realistic talking heads contains many challenges since head motions and facial details have nearly no correlation to audio, unlike the lips [18]. Some works can generate specific facial details, such as natural head movements [33, 41] and templated [18] or controllable blinking [42, 44]. They struggle to capture all the facial details. Moreover, due to the neglect of controlling pupils, their models are constrained and need help handling source images with unnatural gaze directions or multi-target interaction situations (for generating talking videos where the subject looks at different targets). Manipulating pupils is also challenging due to the requirement for a realistic pupil rendering model and additional controllable pupil parameters. Some end-to-end methods need help to offer additional constraints for pupil rendering [24]. Our approach uses 3D sparse land-

✉ Dongdong Weng
crgj@bit.edu.cn

Hui Fang
FangHui72@163.com

Zeyu Tian
tianty97@163.com

Yin Ma
mayin98@163.com

[1] Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing, China

[2] Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Zhengzhou Research Institute, Beijing Institute of Technology, Zhengzhou, China

[3] Ningxia Baofeng Group Co. Ltd., Yinchuan, China

marks as intermediate characterization, which is a typical and straightforward approach [10] and has been used in the state-of-the-art[18, 48]. We add two additional pupil landmarks for controlling gaze direction. The pupils, eyebrows, and eyelids are constrained by corresponding landmarks and independently parameterized for ease of control.

After editing the landmarks, accurately warping the facial images corresponding to the landmark motions poses a challenge. The significant changes in gaze direction and lip movements make generating stable and realistic rendering difficult. To solve this challenge, several works concatenate a candidate image set of the source person on the facial representation [10, 18], but it is often challenging to find additional images with matching backgrounds. Neural radiance field (NeRF) has been explored in facial animation, which could preserve more details and provide better naturalness [4]. However, this method could be more effective in various scenarios and for different identities. The flow-based system considers facial deformation a conditioned action transfer and learns optical flow to represent facial changes from either supervised [10] or unsupervised [38, 47] 2D feature points. Zhao et al. [47] provide a first-order motion approximation using Thin Plate Splines (TPS) transformation [3] based on unsupervised points and efficiently model complex facial movements. However, the generated facial details could be more explicit, and mouth shapes could be more precise.

Based on Zhao et al.'s [47] system, we introduce SFWNet to deform facial images based on landmark motions. What sets our model apart from theirs is that our landmarks are strategically placed on meaningful parts of the face, making the network focus on critical facial motions. Due to task differences, their Background Motion Predictor is no longer applicable. We employ the generative adversarial training mechanism to optimize the generation of unseen pixels. To minimize expression errors, we incorporate the sketch of the driving landmarks as additional shape constraints into the Warp Module. The main contribution of this paper can be summarized as:

- We propose ManiTalk for manipulable talking head generation. The system can manipulate facial details (head pose, blink, eye gaze, and eyebrow) and generate diverse talking styles.
- The proposed SFWNet accurately deforms faces according to landmark motions. Incorporating a sketch of driving landmarks enhances visual quality and ensures more precise facial shapes in the final results.
- The system can work on source images in the wild, such as arbitrary identities, complex backgrounds, and face images with head poses and expressions.
- Experiments show that our system achieves state-of-the-art performance in identity preservation and visual quality. We also achieve state-of-the-art motion synchronization on the VoxCeleb dataset.

## 2 Related work

### 2.1 Audio-based dubbing in video editing

Audio-based dubbing in video editing can generate audio-synchronized lower face and replace the mouth region of the original video. These methods do not need to consider expressions and head poses. The main challenge is to generate a photo-realistic mouth texture matching the original video and splice it seamlessly. Personalized visual dubbing [17, 29, 36, 45] is easier since they are limited to several certain persons in the known backgrounds. Arbitrary-subject visual dubbing [24, 28] builds a general model for any identities. Most methods generate the latent embedding from audio and then render the results using the image-to-image translation network [17, 36, 45]. Several methods generate audio-related facial landmarks to produce accurate lip motions [28, 29]. 3D model-based approaches generate expression, texture, and illumination blendshapes from audio features based on existing face datasets [36, 45]. There are methods to generate mesh vertices and texture maps for more accurate expressions and realistic renderings [17].

### 2.2 Audio-based single image facial animation

Audio-based single image facial animation generates talking heads from a single facial image. With the development of deep learning, end-to-end methods [11, 33] have become a trend. For example, Sefik et al. [11] take a standard normal distribution and a categorical emotion as input and generate talking face videos synchronized with the input speech and consistent with the input emotion. Some works calculate dense motion fields between the target and source motions and take them as intermediate representations. Wang et al. [33] infer talking motions represented by keypoint-based dense motion fields from the input audio. An image generation network is then used to render videos based on the motion fields. It better governs spatial and temporal consistency in the generated videos. Because of the limited information in the input image, some works additionally take reference images or videos as input to provide more constraints for generating stable background and hair [18, 39]. The additional images have the same background as the source image, which is hard to achieve in practice.

## 2.3 Gaze redirection

Gaze redirection is manipulating an input image of a face such that the face in the output image appears to look at a given target direction [25]. Some works use graphics models to render eye gaze [37]. Freund et al. [37] correct gaze directions in video conferencing by modifying the eyes' albedo, diffuse, shading, and illuminations, which is computationally complex. Some methods mainly utilize image warping to redirect the gaze [13, 40]. They are limited when the target gaze direction is far from the source gaze [25]. Recent methods consider eye rotation as a 3D perception problem, which thinks of gaze changes as the rigid rotation of the eyeball [10, 25]. They need to model the complete eyeball in advance. Although gaze redirection has many advances, few researches have achieved gaze redirection in audio-driven face generation.

# 3 Methods

The pipeline of our system is illustrated in Fig. 1. We use 3D sparse landmarks as intermediate results to generate talking head videos synchronized with the speech.

## 3.1 Audio-related generation

This section introduces the facial landmark generation network (Exp Generator) and the pose generation network (Pose Generator). Given the input audio sequence $a_{1:T}$ and the source image $I^i$, our model generates audio embeddings $h^a_{1:T}$ and subsequently produces head poses $\hat{P}_{1:T}$ and facial landmark sequences $\hat{D}_{1:T}$. The framework is formulated as:

$$h^a_{1:T} = \Phi^a(a_{1:T}) \tag{1}$$

$$\hat{D}_{1:T} = \Phi^e(h^a_{1:T}) + I^{m'} \tag{2}$$

$$\hat{P}_{1:T} = \Phi^p(h^a_{1:T}) + I^p \tag{3}$$

where $\Phi^a$, $\Phi^e$, and $\Phi^p$ are the Audio Encoder, Exp Generator, and Pose Generator, respectively. Initial pose $I^p$ is the pose of the source image. For $I^{m'}$, we first estimate the 3D source landmarks $I^m$ from the source image using MediaPipe detector [19]. $I^{m'}$ is the de-posed source landmarks with a neutral expression.

**Audio Encoder.** For speech-related face generation tasks, the Audio Encoder should be robust to different audio sources, regardless of noise, languages, or speakers. Using a pre-trained speech model meets this requirement. The pre-trained model has already been sufficiently trained on large audio datasets and efficiently extracts relevant information from speech. Our Audio Encoder $\Phi^a$ uses the feature encoder part of the pre-trained Wav2Vec2 model [1]. It consists of several blocks containing a temporal convolution (TCN) followed by layer normalization and a GELU activation function. It can extract acoustically meaningful but contextually independent features from the raw speech signal. $\Phi^a$ is initialized using the feature encoder weights in Facebook's Wav2Vec2 model [1], which are trained using more than 50,000-h unlabeled speech. It has been applied in facial animation work and achieves surprising results [12]. The encoder weights are fixed during training.

**Exp Generator.** Exp Generator observes audio embedding $h^a_{1:T} \subseteq \mathbb{R}^{T\times 512}$ and predicts facial landmarks $\hat{D}_{1:T} \subseteq \mathbb{R}^{T\times 68\times 3}$. To avoid some audio-irrelevant facial motions influencing the lip's accuracy, Exp Generator only predicts

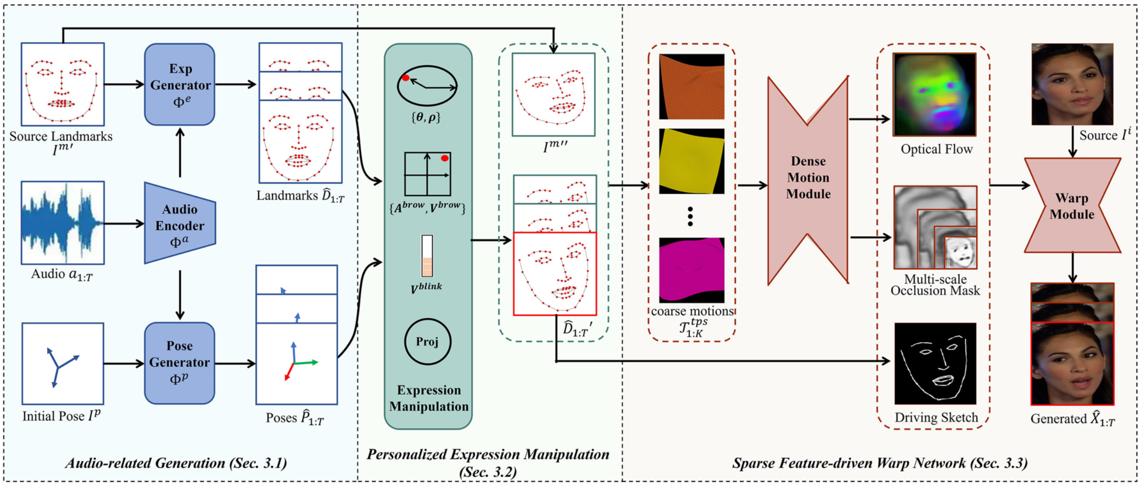

**Fig. 1** Overview of ManiTalk. It consists of three parts: Audio-related Generation (the blue part), Personalized Expression Manipulation (the green part), and Sparse Feature-driven Warp Network (the red part)

lip and jaw motions while the rest of the parts are still. The network employs a transformer-based architecture to ensure accuracy and smoothness across long sequences. It consists of a twelve-layer Transformer encoder [32], a linear projection layer, and a FaceFormer decoder [12]. The linear projection layer projects expression-related embeddings into a 64-dimensional space. The FaceFormer decoder predicts synchronized landmark offsets, which are added to unposed 3D source landmarks $I^{m\prime}$. The structure of the FaceFormer decoder is similar to the Transformer decoder but with generalization abilities for longer sequences. During training, the Transformer encoder is initialized with the Transformer weights in Facebook's Wav2Vec2 model [1] and fine-tuned.

**Pose Generator.**
Mapping from speech features to head poses is difficult due to the weak correlation between them. To simplify this task, we constrain the speaking scenario to online presentations, where the speaker's head mainly faces the camera with small motions. Training data is also readily available.

Pose Generator is trained to generate head motions $\hat{P}_{1:T} \subseteq \mathbb{R}^{T\times 6}$ (consist of rotations $\hat{P}^{\text{rot}}_{1:T}$ and translations $\hat{P}^{\text{trans}}_{1:T}$) conditioned on the audio features $h^a_{1:T}$. Similar to LSP [18] and SadTalker [44], we formulate this task as a probabilistic instead of a regression problem. The network sees the history poses and current audio features and autoregressively predicts the joint probability distribution of the current pose. The probabilistic model we use is a multi-dimensional Gaussian distribution. At time $t$, the network outputs the mean values $\mu_t$ and the standard deviations $\sigma_t$ of the estimated Gaussian conditioned on $h^a_{1:T}$ and $\{\hat{P}_{t-T'}, \hat{P}_{t-T'+1}, ..., \hat{P}_{t-1}\}$. $T'$ is the receptive field of the network. We sample the distribution and obtain the head motion offsets. The offsets are added to $I^p$ and fed to the network to predict the subsequent timesteps. The used prediction network is the conditional probabilistic generative model [18, 23]. It consists of a stack of two residual blocks with seven dilated convolution layers each. The history receptive field size $T'$ is 255 frames, equal to 4.25 seconds.

## 3.2 Personalized expression manipulation

The facial landmarks generated by Exp Generator only have lip and jaw motions, while eyebrows and eyelids remain still. We add facial detail motions to the landmarks through expression manipulation. The landmarks on different facial parts are independent. This allows us to manipulate facial parts independently.

**eyelid and brow motions.** To generate eyelid and eyebrow motions, LSP [18] samples a standard motion set from datasets for each identity. Their approach does not work for our system because we must apply movements to different individuals. To solve this, we personalize the sampled landmarks. For sampled standard motion sequences $M^b_{1:B}$ and $M^e_{1:E}$, we align them to the source landmarks $I^{m\prime}$ by aligning the neutral state frame. Then, we calculate the differences for each frame relative to the neutral frame, obtaining personalized motion offset sequences $\hat{M}^b_{1:B}{}'$ and $\hat{M}^e_{1:E}{}'$. Finally, we cyclically add the calculated offset sequences to the eyelid and eyebrow landmarks in $\hat{D}_{1:T}$ and generate personalized eyelid and eyebrow motions.

Additionally, we can control the motion patterns of eyelids and eyebrows. The blinking frequency can be controlled using the parameter $V^{\text{blink}} \in [0, 1]$, where a bigger value corresponds to a higher blinking frequency. We control eyebrow motions using parameters $\{A^{\text{brow}} \in [-1, 1],\ V^{\text{brow}} \in [0, 1]\}$. A positive value of $A^{\text{brow}}$ represents raising eyebrows, while a negative value represents frowning. Larger absolute $A^{\text{brow}}$ results in more pronounced eyebrow actions. A bigger $V^{\text{brow}}$ corresponds to a higher frequency of eyebrow motions. By default, the system is set to generate natural facial movements with a neutral expression, i.e., $V^{\text{blink}} = 0.5$ and $A^{\text{brow}} = 0$.

**Gaze Manipulation.** We introduce two pupil landmarks to control the gaze and abstract them into two parameters, $\{\theta, \rho\}$. Figure 2a shows the coordinate axes directions of landmarks $\hat{D}_{1:T}$. We project eyelid landmarks onto the $XOY$ plane. Given the landmarks $Q_1{}'$, $Q_2{}'$, $Q_3{}'$, $Q_4{}'$, $Q_5{}'$, $Q_6{}'$ on the eyelid, we roughly define the pupil motion range $Q_1Q_2Q_3Q_4Q_5Q_6$ after the observation of pupil movements. The range is shown in Fig. 2c. The motion contour can be formulated as follows:

$$Q_1 = 2Q_4{}'/7 + 5Q_1{}'/7,\ \ Q_2 = Q_6{}'/4 + 3Q_2{}'/4 \tag{4}$$

$$Q_3 = Q_5{}'/4 + 3Q_3{}'/4,\ \ Q_4 = (Q_5{}' + Q_3{}')/2 \tag{5}$$

$$Q_5 = 4Q_5{}'/5 + Q_3{}'/5,\ \ Q_6 = 4Q_6{}'/5 + Q_2{}'/5 \tag{6}$$

$$O = (Q_5{}' + Q_2{}')/2 \tag{7}$$

The motion contour of the left eye is mirror to that of the right eye (as seen in Fig. 2b). Next, we define the pupil polar coordinate system and using the ordered pair $\{\theta \in [0, 360°], \rho \in [0, 1]\}$ specifies the pupil location $S$ as shown in Fig. 2b. $\rho$ is the relative radius, defining the motion amplitude in that direction. The $\rho$ for points on the motion contour equals 1. We define the point on the contour in the $\theta$ direction as $S'$. So $\overrightarrow{OS} = \rho \cdot \overrightarrow{OS'}$. When $\rho = 0$, $S$ is at the location of pole O, indicating looking straight. When $\{\theta = 90°, \rho = 1\}$, $S$ lies on the line $Q_5Q_6$, representing looking upward. Notably, the movements of the left and right pupils are often the same. The polar axes for both eyes are consistent. We approximate the $z$-coordinate of $S$ as the average of the $z$-coordinates of $Q_1{}'$, $Q_2{}'$, $Q_3{}'$, $Q_4{}'$, $Q_5{}'$, $Q_6{}'$.

**Poses and Projections.** Because the subsequent SFWNet operates in 2D space. We project the manipulated 3D land-

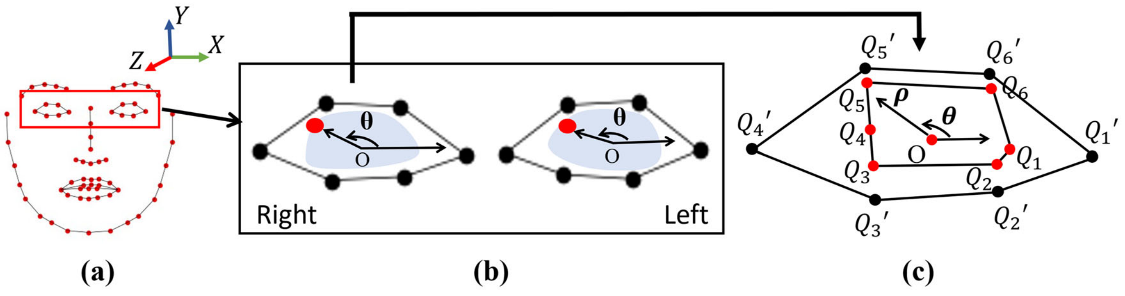

**Fig. 2** Illustrations of the pupil motion range and pupil polar coordinate system

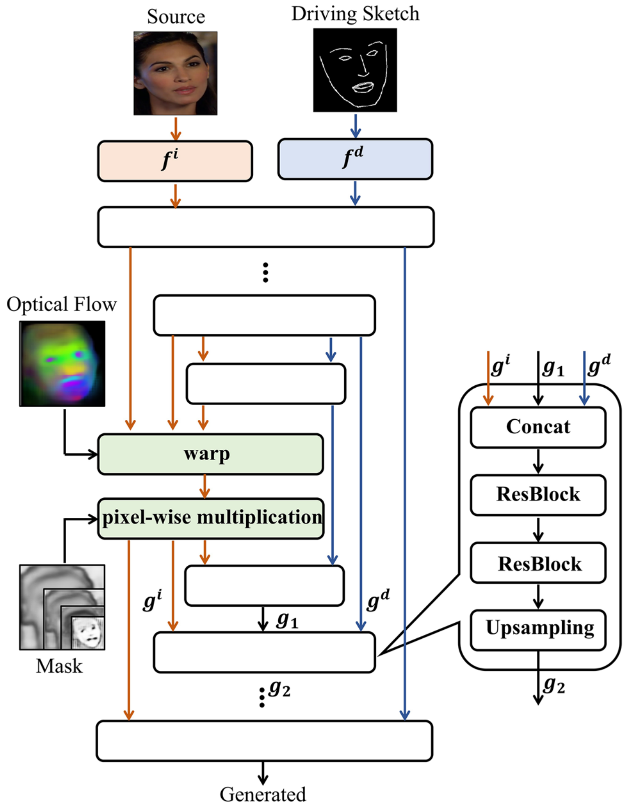

**Fig. 3** Implementation details of the Warp Module

marks to a 2D plane via an orthographic camera model $C$. The process is formulated as:

$$\hat{D}_{1:T}{}' = C \cdot (\hat{P}^{\text{rot}}_{1:T} \cdot \hat{D}_{1:T} + \hat{P}^{\text{trans}}_{1:T}) \tag{8}$$

where $\hat{D}_{1:T}{}' \subseteq \mathbb{R}^{T \times 70 \times 2}$ is the resulting 2D facial landmarks containing pupil landmarks.

### 3.3 Sparse feature-driven warp network—SFWNet

In this section, we predict the final facial animation using SFWNet. As illustrated in the red part of Fig. 1, SFWNet consists of Coarse Motion Estimation, Dense Motion Estimation, and the Warp Module. The network generates facial images frame by frame. It is robust enough to generate smooth facial animations without considering contextual dependencies.

First, the Coarse Motion Estimation module generates K sets of shape transformations from the source image to the driving. Specifically, given the orthographically projected 2D source landmarks $I^{m''}$ and the $t$th frame driving landmarks $\hat{D}_t{}'$, we select $5 \times K$ deformation points from them and divide points into $K$ groups, resulting in $I^{m}_{1:K}{}''$ and $\hat{D}_{t,1:K}{}'$. The selected landmarks are shown in Fig. 7. By minimizing distortion, we can calculate the Thin Plate Splines (TPS) [3] transformation of each group from source points to driving points. We use $\mathcal{T}_k^{tps}$ to represent the $k$th TPS transformation. $K$ is considered as a hyper-parameter. Setting K to 10 yields better results (see Sect. 4.5). Please refer to [3] for more details about TPS transformations.

Dense Motion Estimation aims to combine the $K$ coarse transformations and generate dense motion fields along with the multi-scale occlusion masks. This is achieved by using an hourglass-structured Dense Motion Module. Specifically, we individually warp the source image using $K$ TPS transformations. The $K$ warped images are concatenated and fed into the Dense Motion Module. It outputs $K$ weight maps $W_{1:K}$ corresponding to the $K$ TPS transformations, which are used to calculate dense optical flow for facial motion:

$$\widetilde{\mathcal{T}}(u, v) = \sum_{i=1}^{K} W_k(u, v)\mathcal{T}_k^{tps}(u, v) \tag{9}$$

Where the weight maps are summed to one at any pixel location. Additionally, the Dense Motion Module outputs multi-scale occlusion masks using an additional convolution layer at each decoder layer to handle missing parts in the source image.

In the Warp Module, we fuse multi-scale features to generate high-quality images. The details are shown in Fig. 3. Meaningful landmarks allow us to provide additional shape constraints by inputting the sketch of the driving landmarks as conditions to the decoder. Specifically, the feature extraction layers, denoted as $f^i$ and $f^d$, extract effective features from the source image and the driving sketch at the same resolution, respectively. The features of the source image are passed through the encoder, followed by warping the resulting feature map of each layer using the optical flow. The warped feature map is then multiplied by the occlusion masks of the corresponding scale. The sketch features are input into the same encoder, and its output is concatenated to the decoder via a skip connection along with the masked source features. The decoder finally reconstructs the driving image $\hat{X}_t$.

### 3.4 Implementation details

In this section, we describe training in relevant detail. The Exp Generator, Pose Generator, and SFWNet are trained in a decoupled way. They are trained using Adam optimizer on two NVIDIA GeForce RTX$^{\text{TM}}$ 3090Ti.

**Exp Generator.** We construct the VOCASET-sparse dataset based on VOCASET [7] as the training set. The VOCASET-sparse contains 473 sentences spoken by 12 subjects at 60 $fps$. It has 68 facial landmarks specified manually on the VOCASET facial meshes. We use the same training, validation, and testing splits as VOCA [7]. Exp Generator is trained for 100 epochs with autoregression. The learning rate is $1e-4$. The batch size is 1. The loss function is the Mean Squared Error, which computes the distance between the predicted landmarks and the ground truth.

**Pose Generator.** Pose Generator is trained to generate presentation-style head motions from audio embeddings. HDTF [46] serves as the training set since it consists of online presentation videos. The truth head poses are calculated using OpenFace [2]. We randomly split the dataset into a training, validation, and test set with an 8:1:1 ratio. At the training phase, the learning rate is $1e-4$. The batch size is 8. The negative log-likelihood of the pose distribution [18] is optimized, which forces the network to output the mean values $\mu$ and standard deviations $\sigma$ of the Gaussian distribution.

**SFWNet.** SFWNet is trained on the VoxCeleb dataset [21], which consists of interview videos of different celebrities. The videos are cropped based on face regions and resized to $256 \times 256$ pixels. The 70 facial landmarks are predicted from video frames using MediaPipe [19]. During each iteration, we extract a source image $I^i$ and a driving image $X$ from the same video. This process eliminates the influence of facial shapes, allowing the network to focus solely on expression changes. The learning rate is $2e-4$. The batch size is set to 32. The model is trained for 100 epochs. For the training loss, a pre-trained VGG-19 network [27] is used to improve rendering quality. The warp loss is used for optimizing the Warp Module similar to [26, 47]. The absence of deformation landmarks outside the facial region may result in a limited ability to generate realistic hair and background. We improve this with the adversarial training mechanism. We use the multi-scale PatchGAN [34] as the discriminator. The LSGAN loss [20] is used to optimize the discriminator. We also use a mask color loss to penalize artifacts in the eye and mouth area. The final loss is the sum of the terms:

$$L = L_{\text{GAN}} + L_{\text{VGG}} + L_{\text{warp}} + 500 * L_{\text{color}} \tag{10}$$

## 4 Results

In this section, we demonstrate the superiority of our method. We also report on user studies that evaluate the quality at a video level. Readers can refer to the supplemental videos to evaluate the results presented in this section.

### 4.1 Evaluation metrics

We evaluate our method on multiple metrics widely used in previous works. Learned Perceptual Image Patch Similarity (LPIPS) [43] is used to evaluate the perceptual similarity between images. Cumulative Probability Blur Detection (CPBD) [22] is used to evaluate the sharpness of generated frames. $L_1$ denotes the average $L_1$ distance between the pixel values (range of 0–255) of the generated and reference images. When used to evaluate landmarks and poses, $L_1$ represents the average $L_1$ distance between generated and truth values. We also use Average Keypoint Distance (AKD) to evaluate the facial alignment in generated images. We employ MediaPipe [19] to extract 68 keypoints from the generated and reference images and subsequently calculate AKD values. We choose Cosine Similarity (CSIM) to evaluate identity preservation. We use ArcFace [9] to extract identity embedding of the source image and generated frames and calculate the CSIM of embedding. Smoothness is used to assess the smoothness of generated sequences, the same as the temporal loss introduced in [31]. We choose LSE-D and LSE-C [24] to assess lip synchronization. Lower values are preferable for $L_1$, Smoothness, LPIPS, AKD, and LSE-D metrics. As for LSE-C, CSIM, and CPBD metrics, higher values are more desirable.

### 4.2 Comparison to the state-of-the-art

We compare our system with several state-of-the-art methods (MakeItTalk [48], Audio2Head [33], AVCT [30], EAMM [38], and SadTalker [44]) using their publicly available checkpoints. For EAMM, we generate facial animations with neutral emotions. The evaluation is performed on VoxCeleb test set [21] and HDTF dataset [46]. Most videos in VoxCeleb have complex backgrounds and exaggerated poses and expressions, which can be used to evaluate the model's ability to process wild images. Videos in HDTF have clean backgrounds, forward-facing faces, and high resolutions. They

**Table 1** Comparison to the state-of-the-art methods for the single image facial animation (MakeItTalk [48], Audio2Head [33], AVCT [30], EAMM [38], and SadTalker [44]) on VoxCeleb [21] and HDTF dataset [46]

| | VoxCeleb | | | | HDTF | | | |
|---|---|---|---|---|---|---|---|---|
| | LSE-D↓ | LSE-C↑ | CSIM↑ | CPBD↑ | LSE-D↓ | LSE-C↑ | CSIM↑ | CPBD↑ |
| MakeItTalk (2020) | 10.230 | 4.294 | 0.779 | 0.212 | 10.283 | 4.118 | 0.865 | 0.330 |
| Audio2Head (2021) | 8.681 | 5.891 | 0.402 | 0.149 | **7.786** | **6.684** | 0.645 | 0.265 |
| AVCT (2022) | 10.887 | 3.832 | 0.267 | 0.157 | 8.826 | 5.567 | 0.584 | 0.264 |
| EAMM (2022) | 9.443 | 4.908 | 0.420 | 0.130 | 9.921 | 3.893 | 0.554 | 0.179 |
| SadTalker (2023) | 8.321 | 6.476 | 0.643 | 0.231 | 8.325 | 6.143 | 0.780 | 0.334 |
| Ours | **8.196** | **6.935** | **0.863** | **0.294** | 8.486 | 5.771 | **0.926** | **0.348** |

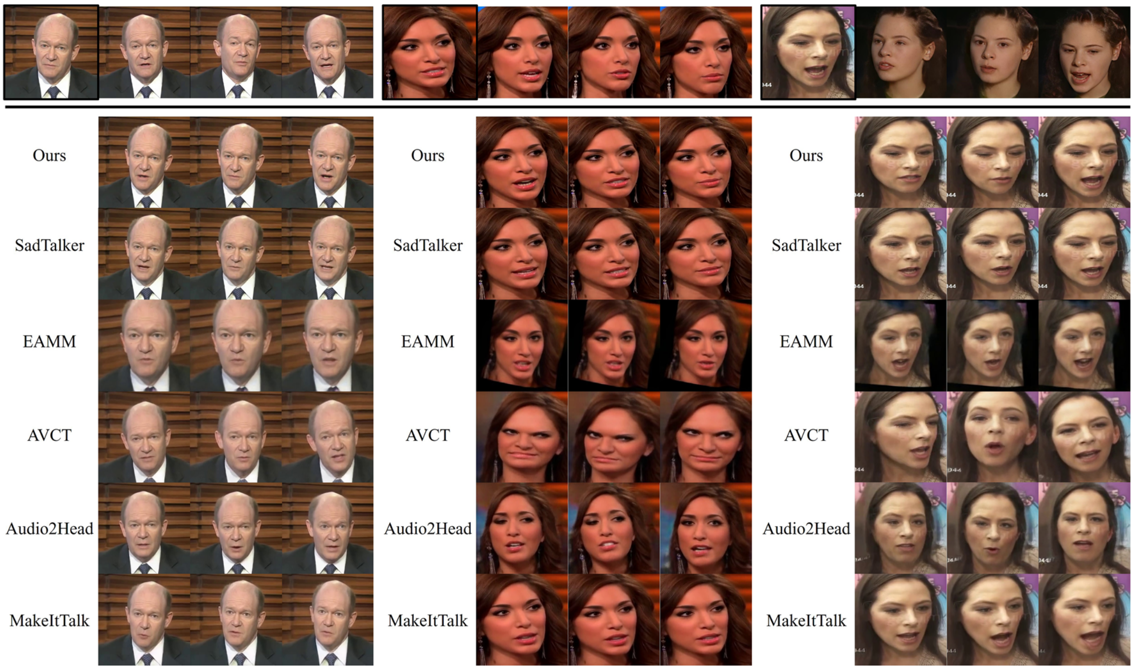

**Fig. 4** Comparison to state-of-the-art audio-based face generation methods. The top row presents source images and the actual lip shapes

are suitable for testing the clarity of generated videos. We use LSE-D, LSE-C, CSIM, and CPBD evaluation metrics.

Table 1 shows that our method achieves much better visual quality and identity consistency according to CPBD and CSIM. Despite the low resolution for VoxCeleb, our method still produces high-sharpness results. For LSE-D and LSE-C, we achieve optimal lip synchronization on the VoxCeleb but not on the HDTF. In some cases, we have observed that lip blurring can make it challenging to distinguish clearly between some consecutive and similar pronunciations. This might be a contributing factor to the relatively poorer synchronization of our method on HDTF. Nonetheless, we have still achieved comparable performance. Meanwhile, as mentioned in SadTalker [44], the lip synchronization metrics may be too sensitive to audio, where the unnatural lip movement may get a better score. Figure 4 illustrates visual results obtained by different methods. We give the truth frames to visualize the lip synchronization. Our method has visual quality and mouth shape similar to the original video. Other methods are all struggling for identity preservation. Apart from SadTalker, these methods often produce distorted faces, especially for faces from VoxCeleb. SadTalker struggles to generate precise lip movements when using a source image with an open mouth (See the third column in Fig. 4).

## 4.3 User studies

We conduct thoughtful user studies in this section. We generated a total of 10 videos for testing. These samples contain almost equal genders with different poses and expressions to evaluate the robustness. We use the binary comparison method, where two videos with the same audio clip are shown side by side [16]. Every participant is required to make 150 comparisons (10 audio clips $\times C_6^2$ comparisons). To avoid any

**Table 2** Results of the user study where we compare our method against state-of-the-art methods

| | Lip Synchronization | Video Sharpness | Overall Naturalness |
|---|---|---|---|
| MakeItTalk | 323 (32.3%) | 409 (40.9%) | 429 (42.9%) |
| Audio2Head | 487 (48.7%) | 358 (35.8%) | 454 (45.4%) |
| AVCT | 403 (40.3%) | 464 (46.4%) | 409 (40.9%) |
| EAMM | 258 (25.8%) | 156 (15.6%) | 184 (18.4%) |
| SadTalker | 731 (73.1%) | 710 (71%) | 699 (69.9%) |
| Ours | **798 (79.8%)** | **903 (90.3%)** | **825 (82.5%)** |

Each row represents the number and the weight of votes for each metric

selection bias, the order (left/right) of all methods for comparison is random for each pair. We invited 20 participants and let them choose the best method for lip synchronization, video sharpness, and overall quality. All participants are students majoring in computer science. Fourteen of them are male. The average age of them is 24.95 (Std=3.90). They were naive to the purposes of the experiment.

The results are presented in Table 2, where our method receives the most votes across all metrics. SadTalker demonstrates comparable performance in lip synchronization to our method. Audio2Head exhibits poor lip synchronization, inconsistent with Table 1. We think that the blurry and distorted faces influence participants' opinions. Our method performs quite well regarding video sharpness and overall naturalness. SadTalker occasionally produces background artifacts during head movements, resulting in a decrease in overall naturalness.

### 4.4 Expression manipulation evaluation

In addition to generating realistic facial animations, a standout feature of our approach is the ability to manipulate facial details and gaze. Figure 5 illustrates the results of manipulating faces. As shown at the top of Fig. 5, our system achieves stable and realistic rendering during blinking. In the middle of Fig. 5, Column 2 shows expressions with neutral eyebrows. Column 3 illustrates expressions with frowning, while columns 4 and 5 show expressions with different eyebrow-raising degrees. The results show that our system provides accurate eyebrow control. It still generates correct lip motions and temporally smooth renderings, even for extreme eyebrow motions.

Figure 6 illustrates the results of gaze manipulation. We also retrain the work of He et al. [14] (denoted as PRMGR) as a baseline. As illustrated in Fig. 6, the eye patches re-rendered by PRMGR are inconsistent with the source image, resulting in noticeable seams on the face. The output could be more accurate when the gaze shifts horizontally. Our results accurately simulate the gaze directions subjectively, even using the source images with head poses. It can be seen that other attributes like eyebrows, hair, and background are well preserved in the manipulated images. Our results have perceptual similarity and consistency to the source images.

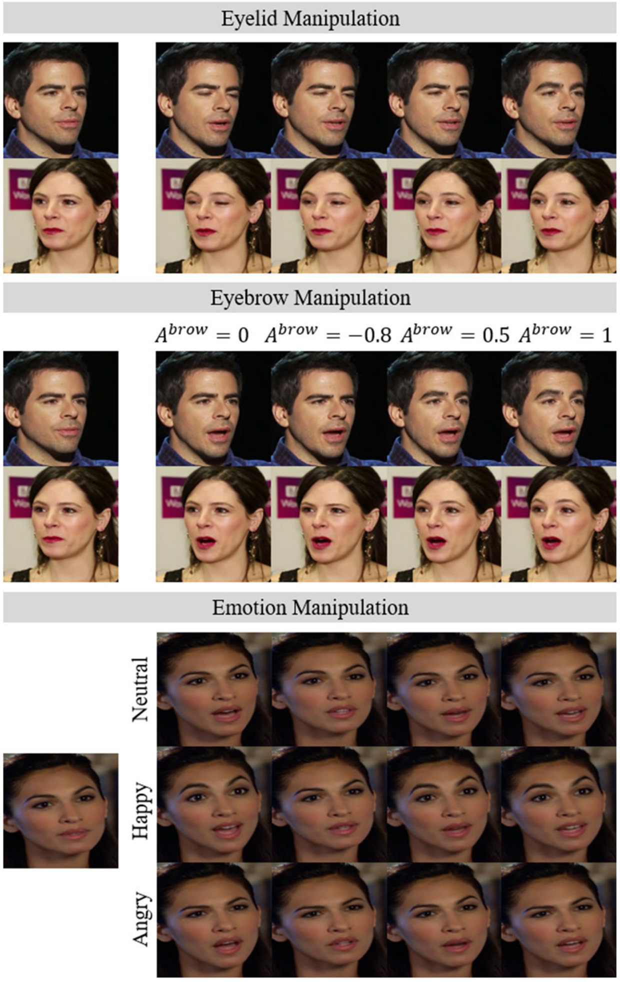

**Fig. 5** The results of manipulating eyelids, eyebrows, and emotions. The first column represents the source image. Columns 2 to 5 of the eyelid manipulation figure depict consecutive frames of blinking. Columns 2 to 5 of the eyebrow manipulation figure showcase the same expression with varying eyebrow states. Each column of the emotion manipulation figure is the same frame with the same mouth shapes but different emotions

### 4.5 Ablation

**Exp Generator.** To demonstrate the advantages of pre-trained speech features, we compare Exp Generator trained with and without pre-trained Audio Encoder weights (denoted as 'w/o pre-trained embedding'). We also train Exp Generator using fixed Transformer encoder weights without fine-tuning (denoted as 'w/o fine-tune'). We also explore whether the FaceFormer decoder has advantages over the Transformer decoder (denoted as 'w/ Transformer decoder'). The results on the test set are presented in Table 3. Without pre-trained

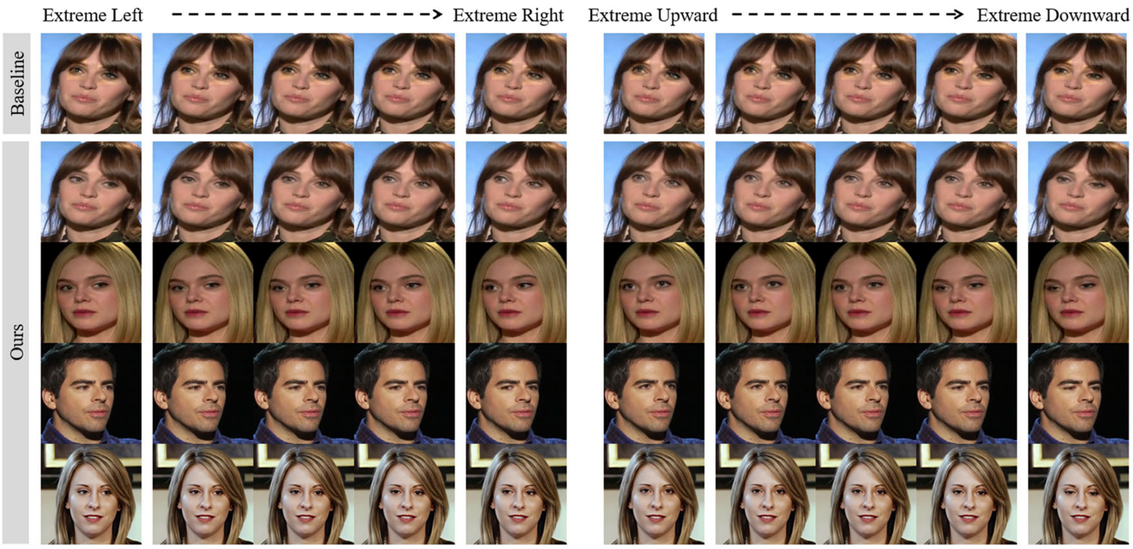

**Fig. 6** The results of gaze manipulation. Each row shows the same expression with different gaze directions. Columns 1, 5, 6, and 10 show gaze in the direction of extreme right, left, upward, and downward, respectively. The middle columns show the interpolated pupil position

**Table 3** Ablation for architecture and training modalities in Exp Generator

| | $L_1\downarrow$ ($\times10^{-3}$) | Smoothness$\downarrow$ ($\times10^{-3}$) |
|---|---|---|
| w/o fine-tune | 2.552 | 0.942 |
| w/o pre-trained embedding | 2.479 | 1.044 |
| w/ Transformer decoder | 2.679 | 0.892 |
| Ours (fine-tune+FaceFormer) | **2.453** | **0.686** |

**Table 4** Ablation for architecture and loss design in Pose Generator

| | $L_1\downarrow$ | Smoothness$\downarrow$ |
|---|---|---|
| Transformer E + Transformer D (L2) | 8.431 | 1.499 |
| Transformer E + Transformer D (P) | 7.332 | 1.356 |
| Transformer E + FaceFormer D (L2) | 11.257 | 0.521 |
| Transformer E + FaceFormer D (P) | 13.631 | 0.493 |
| Ours (L2) | 6.858 | 0.760 |
| Ours (P) | **5.634** | 0.546 |

E means encoder; D means decoder; L2 means trained using L2 loss; P means trained using probabilistic loss

weights and fine-tuning, we observe decreased accuracy and smoothness for facial movements. They fail to produce synchronized mouth movements, resulting in a noticeable temporal jitter. Due to the superiority of Transformer architecture, 'w/ Transformer decoder' provides nice smoothness. However, it tends to generate the average values, making the talking animation look mumbled.

**Pose Generator.** As shown in Table 4, we perform an ablative study for Pose Generator by training and testing five alternative variants: 'Transformer E + Transformer D (L2)', 'Transformer E + Transformer D (P)', 'Transformer E + FaceFormer D (L2)', 'Transformer E + FaceFormer D (P)', and 'Ours (L2)'. The results show that the variants produce more temporal jitters or lower accuracy than our structure, suggesting that Transformer and FaceFormer architectures may not be suitable for our task. The variant 'Ours (L2)' makes the mapping become a regression problem. It struggles to balance audio and history poses, resulting in minor motion variations. This demonstrates that the probabilistic model is better suited for generating audio-related head poses.

**SFWNet.**
As shown in Fig. 7, we compare different values of hyperparameters $K$ in our method, specifically 6, 8, and 12, which correspond to 30, 40, and 60 landmarks, respectively. As we know, SFWNet learns facial deformation based on the transformation between the source landmarks and the driving landmarks. The selected landmarks should represent key facial motions like eye, lip, and cheek motions. The test results are shown in Table 5. It shows that more or fewer landmarks negatively impact the image quality and the expression accuracy. We found that our approach exhibits superior pixel stability. In the "K=6, 8, 12" settings, temporal jitter is noticeable in the hair and background areas. This effect is particularly pronounced in the 'K=6' setting, resulting in distorted faces.

We also evaluate the design of the loss functions. Table 5 demonstrates that $L_{\text{VGG}}$ plays a more significant role in training. Other loss functions focus on enhancing the renders for facial details. Figure 8 shows the rendering samples. We can observe that the loss functions primarily affect the eye and mouth rendering. 'w/o $L_{\text{VGG}}$' causes holes on the face. The

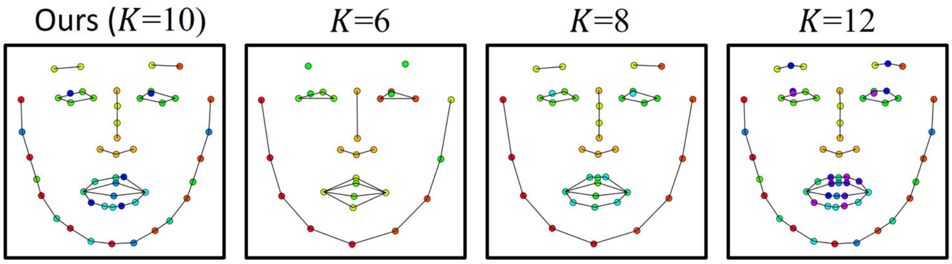

**Fig. 7** The selected landmarks for different $K$ in SFWNet ablation. The landmark colors represent their groups

**Table 5** Ablation for hyper-parameter $K$, loss design, and Warp Module structures in SFWNet

| | LPIPS ($\times 10^{-1}$)↓ | $L_1$↓ | AKD↓ |
|---|---|---|---|
| $K$=6 | 1.232 | 12.107 | 1.937 |
| $K$=8 | 1.221 | 11.769 | 1.882 |
| $K$=12 | 1.210 | 11.791 | 1.824 |
| w/o $L_{\text{color}}$ ($K$=10) | 1.210 | 11.926 | 1.914 |
| w/o $L_{\text{warp}}$ ($K$=10) | 1.249 | 12.043 | 1.906 |
| w/o $L_{\text{GAN}}$ ($K$=10) | 1.279 | 11.925 | 1.771 |
| w/o $L_{\text{VGG}}$ ($K$=10) | 2.374 | 18.408 | 3.180 |
| w/o Sketch ($K$=10) | 1.205 | 11.801 | 1.887 |
| Ours ($K$=10) | **1.201** | **11.689** | **1.709** |

rendering quality of the mouth area on ’w/o $L_{\text{GAN}}$’ is noticeably poorer.

We compare our rendering with ’w/o Sketch’, which does not incorporate the sketch of the driving landmarks in the Warp Module. As seen in Table 5, our method yields better results across all three metrics, with more pronounced changes in the $L_1$ and AKD. Figure 9 demonstrates that our approach produces accurate mouth shapes and realistic rendering. Sometimes, ’w/o Sketch’ fails to close the mouth fully or generates distorted lips.

### 4.6 Extensions

Using 3D landmarks as intermediate results allows expression manipulation, triggering several extensions. We can use facial detectors (e.g., MediaPipe [19]) to obtain facial landmarks from reference videos. The generated landmarks serve as driving landmarks to achieve video-based expression transfer. Inspired by VideoReTalking [5], our method can be used for emotional talking head generation based on manipulable facial details. Figure 5 shows the emotion editing results. Positive emotions result in more pronounced gestures and richer facial details. Users can creatively manipulate landmarks to generate multiple emotions.

### 4.7 Limitation

Although the proposed method can handle source images in the wild, it may still exhibit noticeable artifacts in some instances. Large pose changes may lead to distortion and blurriness at the hair-background junction, as shown in Fig. 10a. This could be attributed to the limited background constraints in SFWNet. One way to solve this is to separate the subject from the background using face parsing [8] and generate them independently. Although qualitative and quantitative experiments indicate that our method can generate sharp images, we still struggle to synthesize realistic teeth in some cases. This limitation can be improved via the blind face restoration networks [35], as shown in Fig. 10b. Because the training data lacks well-organized emotional talking videos, our emotion editing results are less expressive than the state-of-the-art [6]. Fine-tuning on a specialized emotional expression dataset [15] could be a solution.

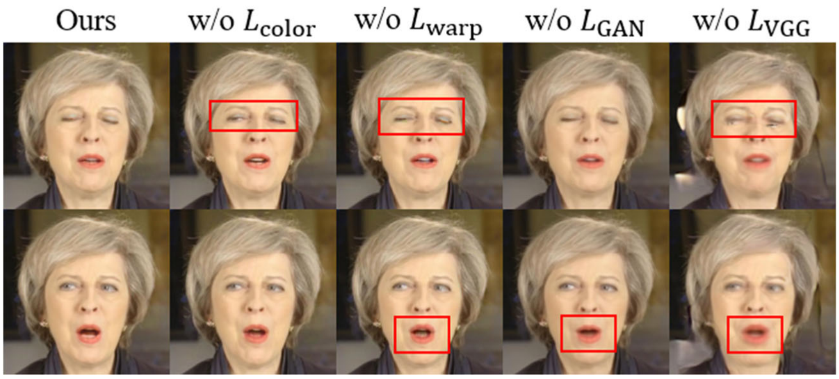

**Fig. 8** Ablation for loss design in SFWNet. The red box indicates the poor rendering parts

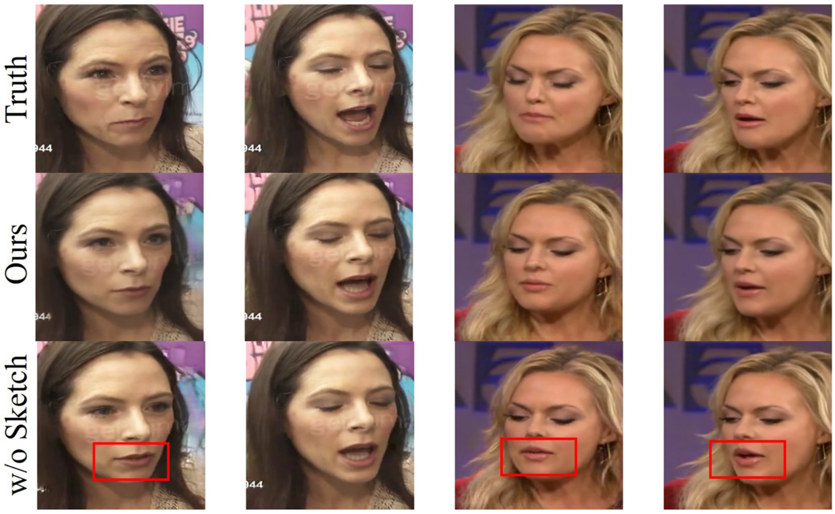

**Fig. 9** Ablation for Warp Module structures in SFWNet. 'w/o Sketch' means that the model does not incorporate the sketch of the driving landmarks in the Warp Module

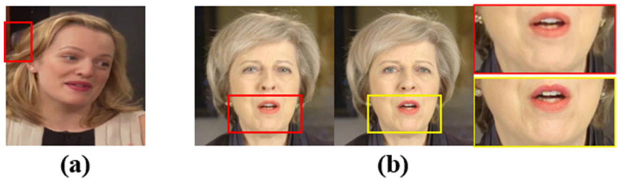

**Fig. 10** Limitations. **a** The distortion caused by large pose changes; **b** Left: The teeth rendering generated by our method. Middle: The teeth rendering generated by our method + GFPGAN [35]. Right: The zoomed mouth regions

## 5 Conclusion

This paper presents ManiTalk, the first manipulable audio-driven talking head generation system to generate personalized talking styles. We propose Exp Generator and Pose Generator to generate synchronized talking landmarks and presentation-style head poses. Personalized expression manipulation allows for manipulating facial details (eyelids and eyebrows) independently. We add two additional pupil landmarks to manipulate the gaze. We introduce SFWNet, which learns coarse and dense motion fields to model the relationships between landmarks and realistic renderings. We provide additional shape constraints by inputting the sketch of the driving landmarks to the Warp Module, enhancing the face accuracy and realism. Experimental results show that the proposed method can work on subject images in the wild. Our results not only preserve lip synchronization but also achieve state-of-the-art performance in terms of identity preservation and video quality. Furthermore, the manipulable face extends the potential application ranges. We can generate emotional talking videos and videos that talk to multiple targets with just one source image.

In the future, we will add candidate images or phoneme-related features as input to improve the quality of the generated video. Additionally, we will work on a more lightweight network for end-to-end facial generation. Generating emotional expressions that match speech is also one of our future works.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00371-024-03490-4.

**Acknowledgements** This work was supported by the National Key R&D Program of China (No. 2022YFF0902303) and the Beijing Municipal Science & Technology Commission and Administrative Commission of Zhongguancun Science Park (No. Z221100007722002) and the National Natural Science Foundation of China (No. 62072036).

## References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
3. Bookstein, F.L.: Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**(6), 567–585 (1989)
4. Chatziagapi, A., Athar, S., Jain, A., Rohith, M., Bhat, V., Samaras, D.: Lipnerf: what is the right feature space to lip-sync a nerf? In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE (2023)
5. Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers (2022)
6. Chenxu, Z., Chao, W., Jianfeng, Z., Hongyi, X., Guoxian, S., You, X., Linjie, L., Yapeng, T., Xiaohu, G., Jiashi, F.: Dream-talk: diffusion-based realistic emotional audio-driven method for single image talking face generation. arXiv preprint arXiv:2312.13578 (2023)
7. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10,101–10,111 (2019)

8. Deng, H., Han, C., Cai, H., Han, G., He, S.: Spatially-invariant style-codes controlled makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6549–6557 (2021)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
10. Doukas, M.C., Ververas, E., Sharmanska, V., Zafeiriou, S.: Free-headgan: neural talking head synthesis with explicit gaze control. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
11. Eskimez, S.E., Zhang, Y., Duan, Z.: Speech driven talking face generation from a single image and an emotion condition. IEEE Trans. Multimed. **24**, 3480–3490 (2021)
12. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18,770–18,780 (2022)
13. Ganin, Y., Kononenko, D., Sungatullina, D., Lempitsky, V.: Deepwarp: photorealistic image resynthesis for gaze manipulation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14, pp. 311–326. Springer (2016)
14. He, Z., Spurr, A., Zhang, X., Hilliges, O.: Photo-realistic monocular gaze redirection using generative adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6932–6941 (2019)
15. Houwei, C., David, G.C., Michael, K.K., Ruben, C.G., Ani, N., Ragini, V.: Crema-d: crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. **5**(4), 377–390 (2014)
16. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans. Graph. **36**(4), 1–12 (2017)
17. Lahiri, A., Kwatra, V., Frueh, C., Lewis, J., Bregler, C.: Lipsync3d: data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2755–2764 (2021)
18. Lu, Y., Chai, J., Cao, X.: Live speech portraits: real-time photorealistic talking-head animation. ACM Trans. Graph. **40**(6), 1–17 (2021)
19. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: a framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
20. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
21. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)
22. Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Trans. Image Process. **20**(9), 2678–2683 (2011)
23. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
24. 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: Proceedings of the 28th ACM International Conference on Multimedia, pp. 484–492 (2020)
25. Ruzzi, A., Shi, X., Wang, X., Li, G., De Mello, S., Chang, H.J., Zhang, X., Hilliges, O.: Gazenerf: 3d-aware gaze redirection with neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9676–9685 (2023)
26. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13,653–13,662 (2021)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Song, L., Wu, W., Qian, C., He, R., Loy, C.C.: Everybody's talkin': let me talk as you want. IEEE Trans. Inf. Forensics Secur. **17**, 585–598 (2022)
29. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. **36**(4), 1–13 (2017)
30. Suzhen, W., Lincheng, L., Yu, D., Xin, Y.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2531–2539 (2022)
31. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: audio-driven facial reenactment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, pp. 716–731. Springer (2020)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
33. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: audio-driven one-shot talking-head generation with natural head motion. arXiv preprint arXiv:2107.09293 (2021)
34. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)
35. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9168–9178 (2021)
36. Wen, X., Wang, M., Richardt, C., Chen, Z.Y., Hu, S.M.: Photorealistic audio-driven video portraits. IEEE Trans. Visual Comput. Graph. **26**(12), 3457–3466 (2020)
37. Wolf, L., Freund, Z., Avidan, S.: An eye for an eye: a single camera gaze-replacement method. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 817–824. IEEE (2010)
38. Xinya, J., Hang, Z., Kaisiyuan, W., Qianyi, W., Wayne, W., Feng, X., Xun, C.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10 (2022)
39. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.J.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137 (2020)
40. Yu, Y., Odobez, J.M.: Unsupervised representation learning for gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7314–7324 (2020)
41. Zhang, C., Ni, S., Fan, Z., Li, H., Zeng, M., Budagavi, M., Guo, X.: 3d talking face with personalized pose dynamics. IEEE Trans. Visual Comput. Graph. **29**(2), 1438–1449 (2023)
42. Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., Guo, X.: Facial: synthesizing dynamic talking face with implicit attribute learning. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3847–3856 (2021)

43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
44. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8652–8661 (2023)
45. Zhang, Y., He, W., Li, M., Tian, K., Zhang, Z., Cheng, J., Wang, Y., Liao, J.: Meta talk: learning to data-efficiently generate audio-driven lip-synchronized talking face with high definition. In: ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4848–4852 (2022)
46. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3661–3670 (2021)
47. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3657–3666 (2022)
48. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Trans. Graph. **39**(6), 1–15 (2020)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Hui Fang** received the bachelor of science degree from Beijing Jiaotong University, Beijing, in 2017. She is currently pursuing the PhD degree from the School of Optics and Photonics at Beijing Institute of Technology. Her primary research interests lie in facial animation, virtual reality, natural language processing and computer vision.

**Dongdong Weng** received the PhD degree in optical engineering from the Beijing Institute of Technology, Beijing, China, in 2006. He is currently a Professor of Optical Engineering with the School of Optics and Photonics, Beijing Institute of Technology. His research interests include virtual reality, mixed reality, human-computer interaction and digital human.

**Zeyu Tian** received the bachelor's degree from the Beijing Institute of Technology, Beijing, in 2019. He is currently pursuing the PhD degree from Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology. His main research interests include facial motion capture, 3D virtual face reconstruction and computer vision.

**Yin Ma** is currently serving as the Chief Information Officer (CIO) at Ningxia Baofeng Group Co. Ltd., Yinchuan, China. He obtained his Master's degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences in 2022. His primary research interests lie in multimodal technologies, computer vision, natural language processing, advanced process control and model predictive control.