FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning

Chenxu Zhang[1], Yifan Zhao[2], Yifei Huang[3], Ming Zeng[4], Saifeng Ni[5]
Madhukar Budagavi[5], Xiaohu Guo[1]

[1]The University of Texas at Dallas [2]Beihang University [3]East China Normal University
[4]Xiamen University [5]Samsung Research America

{chenxu.zhang, xguo}@utdallas.edu, zhaoyf@buaa.edu.cn, yifeihuang17@gmail.com
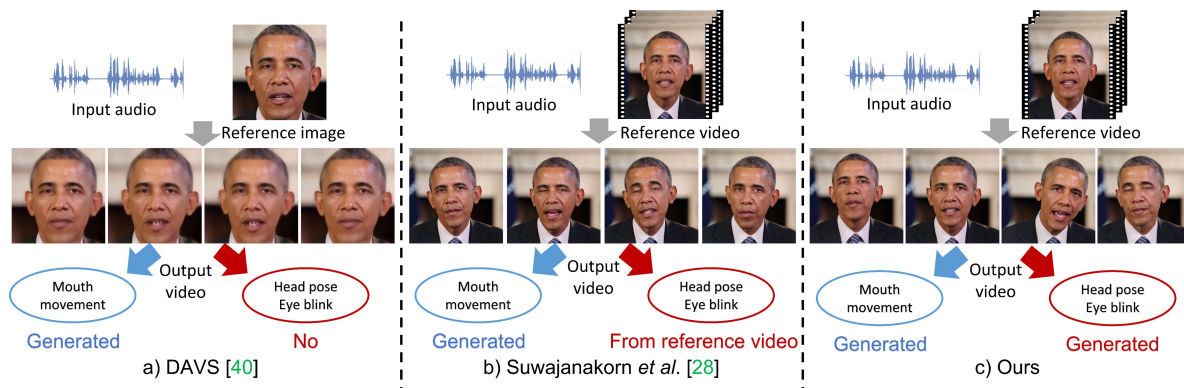zengming@xmu.edu.cn, {saifeng.ni, m.budagavi}@samsung.com

Figure 1. Illustration of three typical frameworks. a) Explicit attribute generation: only considering mouth movements of talking head. b) Explicit generation with implicit morphing: only generating explicit mouth movements and taking the implicit attributes from reference videos. c) Our implicit attribute learning framework: generating explicit and implicit attributes from input audio in one unified framework.

Abstract

In this paper, we propose a talking face generation method that takes an audio signal as input and a short target video clip as reference, and synthesizes a photo-realistic video of the target face with natural lip motions, head poses, and eye blinks that are in-sync with the input audio signal. We note that the synthetic face attributes include not only explicit ones such as lip motions that have high correlations with speech, but also implicit ones such as head poses and eye blinks that have only weak correlation with the input audio. To model such complicated relationships among different face attributes with input audio, we propose a FACe Implicit Attribute Learning Generative Adversarial Network (FACIAL-GAN), which integrates the phonetics-aware, context-aware, and identity-aware information to synthesize the 3D face animation with realistic motions of lips, head poses, and eye blinks. Then, our Rendering-to-Video network takes the rendered face images and the attention map of eye blinks as input to generate the photo-realistic output video frames. Experimental results and user studies show our method can generate realistic talking face videos with not only synchronized lip motions, but also natural head movements and eye blinks, with better qualities than the results of state-of-the-art methods.

1. Introduction

Synthesizing dynamic talking faces driven by input audio has become an important technique in computer vision, computer graphics, and virtual reality. There have been steady research progresses [4, 9, 10, 16, 26, 27, 31, 32, 40], however, it is still very challenging to generate photo-realistic talking faces that are indistinguishable from real captured videos, which not only contain synchronized lip motions, but also have personalized and natural head movements and eye blinks, etc.

The information contained in dynamic talking faces can be roughly categorized into two different levels: 1) the at-

tributes that need to be synchronized with the input audio, *e.g.*, the lip motion that has strong correlations with the signals of auditory phonetics; 2) the attributes that have only weak correlations with the phonetic signal, *e.g.*, the head motion that is related to both the context of speech and the personalized talking style and the eye blinking whose rate is mainly decided by personal health condition as well as external stimulus. Here we call the first type of attributes to be *explicit* attributes, and the second type to be *implicit* attributes.

It should be noted that the majority of existing works [9, 10, 16, 26, 27, 40] on talking face generation are focusing on *explicit* attributes only, by synchronizing the lip motions with input audio. Examples include Zhou *et al.* [40] disentangled the audio into subjected-related information and speech-related information to generate clear lip patterns, and Chen *et al.*'s Audio Transformation and Visual Generation (ATVG) networks [9] to transfer audio to facial landmarks and generate video frames conditioned on landmarks. There are only a few recent efforts [7, 36, 41] exploring the correlation between the *implicit* attributes of head pose with the input audio. For example, Chen *et al.* [7] adopted a multi-layer perceptron as the head pose learner to predict the transformation matrix of each input frame. However, it remains unclear on: (1) how the *explicit* and *implicit* attributes might potentially influence each other? (2) how to model implicit attributes, such as head poses and eye blinks, that depend not only on the *phonetic* signal, but also on the *contextual* information of speech as well as the *personalized* talking style?

To tackle these challenges, we propose a FACe Implicit Attribute Learning (FACIAL) framework for synthesizing dynamic talking faces, as shown in Fig. 2. (1) Unlike the previous work [7] predicting implicit attributes using an individual head pose learner, our FACIAL framework jointly learns the *implicit* and *explicit* attributes with the regularization of adversarial learning. We propose to embed all attributes, including Action Unit (AU) of eye blinking, head pose, expression, identity, texture and lighting, in a collaborative manner so their potential interactions for talking face generation can be modeled under the same framework. (2) We design a special FACIAL-GAN in this framework to jointly learn *phonetic*, *contextual*, and *personalized* information. It takes a sequence of frames as a grouped input and generates a contextual latent vector, which is further encoded together with the phonetic information of each frame, by individual frame-based generators. FACIAL-GAN is initially trained on our whole dataset (Sec. 4). Given a short reference video (2 ∼ 3 minutes) of the target subject, FACIAL-GAN will be fine-tuned with this short video, so it can capture the personalized information contained in it. Hence our FACIAL-GAN can well-capture all phonetic, contextual, and personalized information of the implicit at-

tributes, such as head poses. (3) Our FACIAL-GAN can also predict the AU of eye blinks, which is further embedded into an auxiliary eye-attention map for the final Rendering-to-Video module, to generate realistic eye blinking in the synthesized talking face.

With the joint learning of explicit and implicit attributes, our end-to-end FACIAL framework can generate photo-realistic dynamic talking faces as shown in Fig. 1, superior to the results produced by the state-of-the-art methods. The **contribution** of this paper is threefold: (1) We propose a joint explicit and implicit attribute learning framework to synthesize photo-realistic talking face videos with audio-synchronized lip motion, personalized and natural head motion, and realistic eye blinks. (2) We design a FACIAL-GAN module to encode the contextual information with the phonetic information of each individual frame, to model the implicit attributes needed for synthesizing natural head motions. (3) We embed the FACIAL-GAN generated AU of eye blinking into an eye-attention map of rendered faces, which achieves realistic eye blinks in the resulting video produced by the Rendering-to-Video module.

2. Related Work

Audio-driven talking face generation Most of existing talking face generation methods [4, 9, 10, 16, 26, 27, 31, 32, 40] focus on generating videos that are in sync with the input audio stream. Chung *et al.* [10] proposed an encoder-decoder CNN model using a joint embedding of face and audio to generate synthesized talking face video frames. Chen *et al.* [9] proposed a hierarchical structure to first transfer audio to landmarks and then to generate video frames conditioned on landmarks. However, in talking face videos generated by these approaches, the head pose is almost fixed during the speech. To achieve photo-realistic videos with head motion, several techniques [28, 23, 29, 34] first generate the lip area that is in sync with input audio and compose it into an original video. Suwajanakorn *et al.* [28] used an audio stream from Barack Obama to synthesize a photo-realistic video of his speech. However, this method is applicable to other characters because of the requirement of a large amount of video footage. Thies *et al.* [29] employed a latent 3D model space to generate talking face video which can be used for different people. However, those methods cannot disentangle the head motion and facial expression due to their intrinsic limitations, which means the head motion is irrelevant with the input audio. More recently, Chen *et al.* [7] and Yi *et al.* [36] focus on generating head movement directly from input audio. Yi *et al.* [36] proposed a memory-augmented GAN module to generate photo-realistic videos with personalized head poses. However, due to the limitations of network and 3D model, their generated face expression (*e.g.*, eye blinks) and head motion tend to be still. In comparison, we introduce the FACIAL-
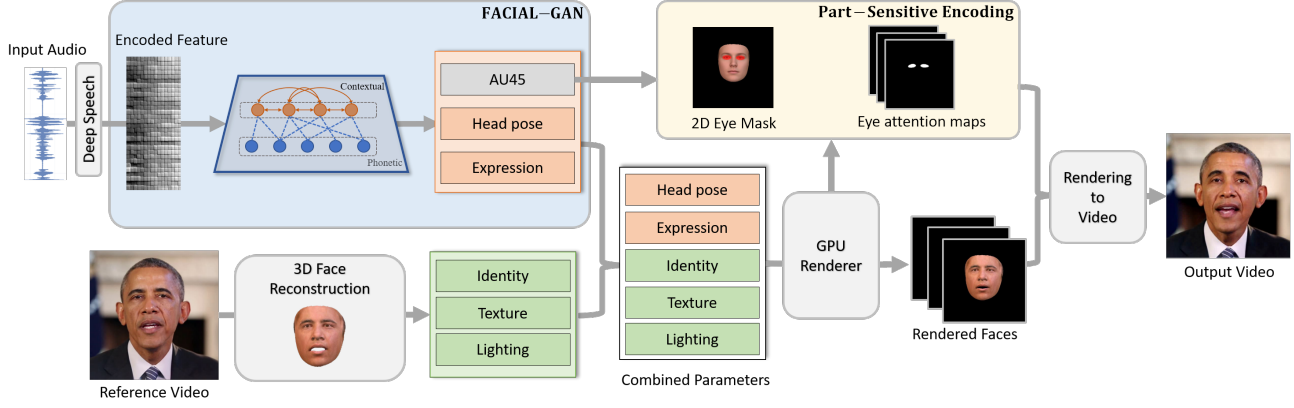
Figure 2. Overview of the proposed implicit attribute learning framework. Given input audio, the proposed FACIAL-GAN aims to generate the explicit attributes (expression) and implicit attributes (eye blinking AU45, head pose) with jointly temporal correlations and local phonetic features. The reference video is performed with a face reconstruction operation to provide a 3D model guidance for rendering operation. Besides, a part-sensitive encoding takes the eye blinking action units as input and serves as eye attention maps for rendering faces. These guidances are jointly combined to feed into the rendering-to-video network.

GAN module to integrate phonetic, contextual, and personalized information of the talking, and combine the synthesized 3D model with AU attention map to generate photorealistic videos with synchronized lip motion, personalized and natural head poses and eye blinks.

Video-driven talking face generation Video-driven talking face generation methods [20, 19, 39, 37, 24, 35, 30] transferred face expression and slight head movements from the given source video frames to target video frames. Zakharov *et al.* [37] presented a system to frame the few- and one-shot learning of neural talking head models of unseen people as adversarial training problems with high capacity generators and discriminators. Kim *et al.* [20] introduced a generative neural network to transfer head pose, face expression, eye gaze and blinks from a source actor to a portrait video based on the generated 3D model. However, since the head movements and face expressions are guided by the source video, these methods can only generate predetermined talking head movements and expressions which is consistent with the source ones.

3. Approach

3.1. Problem Formulation

Given an input audio $\mathcal{A}$ and a short ($2 \sim 3$ minutes) reference video $\mathcal{V}$ of the subject, our talking head synthesis aims to generate a speech video $\mathcal{S}$ of the subject synchronized with $\mathcal{A}$. The conventional steps to generate neural talking head can be represented as:

$$
\begin{aligned}
\mathcal{F}_{lip} &= \mathbf{G}(\mathbf{E}(\mathcal{A})), \\
\mathcal{S} &= \mathbf{R}(\mathcal{F}_{lip}, \mathcal{V}),
\end{aligned} \tag{1}
$$

where $\mathcal{F}_{lip}$ denotes the explicit features synthesized by an adversarial generator $\mathbf{G}$. $\mathbf{E}$ denotes audio feature extraction network and $\mathbf{R}$ denotes the rendering network to translate the synthesized features into the output video.

As mentioned above, this conventional synthesizing approach usually fails to capture the implicit attributes, *e.g.*, dynamic head poses $\mathcal{M}_{pose}$, and eye blinks $\mathcal{M}_{eye}$. Toward this end, we further exploit the intrinsic interrelationships among speech audio and these implicit attributes, namely FACe Implicit Attribute Learning (FACIAL). Besides, we introduce an auxiliary part attention map of eye regions $\mathcal{E}$. Our FACIAL synthesis process has the form:

$$
\begin{aligned}
\{\mathcal{F}_{lip}, \mathcal{M}_{pose}, \mathcal{M}_{eye}\} &= \mathbf{G}(\mathbf{E}(\mathcal{A})), \\
\mathcal{S} &= \mathbf{R}(\mathcal{F}_{lip}, \mathcal{M}_{pose}, \mathcal{E} \odot \mathcal{M}_{eye}, \mathcal{V}).
\end{aligned} \tag{2}
$$

The overall framework in Fig. 2 is composed of two essential parts, *i.e.*, a FACIAL Generative Adversarial Network (FACIAL-GAN) to encode the joint explicit and implicit attributes, and a Rendering-to-Video network to synthesize the output talking face video with synchronized lip motion, natural head pose, and realistic eye blinks. Furthermore, different attributes require individual encoding strategies, the explicit attributes $\mathcal{F}_{lip}$ are highly correlated with the syllables of input audio, which are decided by each audio frame. However, implicit features $\mathcal{M}_{\{eye,pose\}}$ heavily relies on the long-term information, *e.g.*, head movements of the next frames are decided by the previous states. We thus elaborate on how to embed these attributes into one unified framework in the next subsections.

3.2. FACIAL-GAN

To jointly embed the explicit and implicit attributes in one unified network, we need to: 1) generate explicit ex-
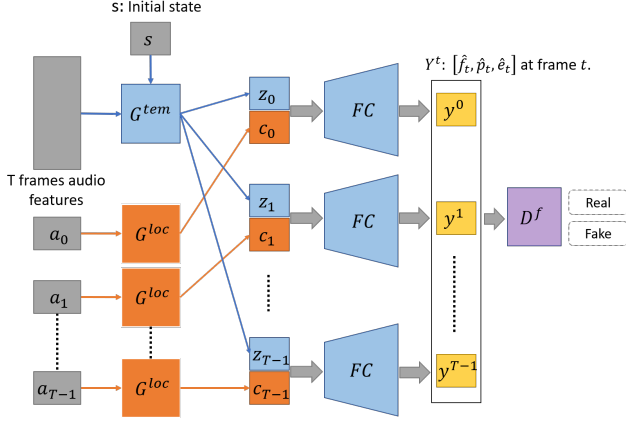
Figure 3. Framework of the proposed FACIAL-GAN. $\mathbf{G}^{tem}$ takes the whole sequence of T frames as input to generate temporal vector z, while $\mathbf{G}^{loc}$ generates the local latent vector c of each frame.

pressions corresponding to phonetic features of each frame; 2) embed the contextual information, *i.e.*, temporal correlations into the network for implicit attribute learning. We propose FACIAL-GAN as a solution to achieve these goals.

The proposed FACIAL-GAN is composed of three essential parts: temporal correlation generator $\mathbf{G}^{tem}$ to build contextual relationships and a local phonetic generator $\mathbf{G}^{loc}$ to extract the characteristics of each frame. Besides, a discriminator network $\mathbf{D}^f$ is employed to judge real or fake of the generated attributes. As shown in Fig. 3, the input audio $\mathcal{A}$ is sampled by a sliding window of T frames, and is prepossessed with DeepSpeech [17] to generate the feature $\mathbf{a} \in \mathbb{R}^{29 \times T}$. Let $\mathbf{f}$ denote the facial expression parameters, $\mathbf{p}$ denote head pose features and $\mathbf{e}$ denote eye blink AU estimation, and use f_t, p_t and e_t to represent the features of the t-th frame, respectively (See supplementary for details).

Temporal correlation generator: To extract the temporal correlations of the whole input sequence, our key idea to feed the audio sequence $\mathcal{A}$ of T frames into a contextual encoder, which generates the latent global features $\mathbf{z}$. As taking the audio sequence as a whole, each unit of latent feature $\mathbf{z}$ is able to incorporate the information from other frames. Hence the corresponding feature z_t of the t-th frame can be extracted by splitting the encoded feature $\mathbf{z}$. Given the DeepSpeech features $a[0 : T - 1]$ of input audio $\mathcal{A}$ and its initial state $\mathbf{s} = \{f_0, p_0, e_0\} \in \mathbb{R}^{71}$, we generate the predicted temporal attribute sequence z_t, $t \in [0, T - 1]$ with $\mathbf{G}^{tem}$. The initial state $\mathbf{s}$ is introduced to ensure the temporal continuity between generated sequences.

Local phonetic generator: The temporal network $\mathbf{G}^{tem}$ focuses on the whole temporal domain, where the phonetic features of each frame are not emphasized. Therefore, we employ the local phonetic network $\mathbf{G}^{loc}$ to generate local features c_t for the t-th frame. Taking the t-th frame as an example, $\mathbf{G}^{loc}$ takes audio features $a_t = a[t - 8 : t + 8]$

as input and outputs the local feature c_t. Now, we have obtained the temporal features z_t and local features c_t for time step t. One fully connected layer $\mathbf{FC}$ is used to map z_t and c_t to the predicted parameters $\hat{f}_t, \hat{p}_t, \hat{e}_t \in \mathbb{R}^{71}$. The encoding process of FACIAL-GAN can be represented as:

$$z_t = \mathcal{S}(\mathbf{G}^{tem}(\mathbf{E}(\mathcal{A})|\mathbf{s}), t),$$
$$c_t = \mathbf{G}^{loc}(\mathcal{S}(\mathbf{E}(\mathcal{A}), t)), \tag{3}$$
$$[\hat{f}_t, \hat{p}_t, \hat{e}_t] = \mathbf{FC}(z_t \oplus c_t),$$

where function $\mathcal{S}(\mathbf{X}, t)$ denotes splitting and extracting the t-th feature block of feature $\mathbf{X}$, and $\oplus$ is the feature concatenation operation. $\mathbf{E}$ denotes the audio feature extraction.

Learning objective: We supervise our generator networks $\mathbf{G}^{tem}$ and $\mathbf{G}^{loc}$ with the following loss functions:

$$\mathcal{L}_{\text{Reg}} = \omega_1 \mathcal{L}_{\exp} + \omega_2 \mathcal{L}_{\text{pose}} + \omega_3 \mathcal{L}_{\text{eye}} + \omega_4 \mathcal{L}_s, \tag{4}$$

where ω_1, ω_2, ω_3, and ω_4 are the balancing weights, and $\mathcal{L}_s$ is the L_1 norm loss for the initial state values, which guarantees the continuity between the generated sequences of sliding windows:

$$\mathcal{L}_s = \|f_0 - \hat{f}_0\|_1 + \|p_0 - \hat{p}_0\|_1 + \|e_0 - \hat{e}_0\|_1. \tag{5}$$

$\mathcal{L}_{\exp}$, $\mathcal{L}_{\text{pose}}$, and $\mathcal{L}_{\text{eye}}$ are the L_2 norm losses for facial expression, head pose, and eye blink AU respectively. We also introduce the motion loss $\mathcal{U}$ to guarantee the inter-frame continuity:

$$\mathcal{L}_{exp} = \sum_{t=0}^{T-1} \mathcal{V}(f_t, \hat{f}_t) + \omega_5 \sum_{t=1}^{T-1} \mathcal{U}(f_{t-1}, f_t, \hat{f}_{t-1}, \hat{f}_t),$$
$$\mathcal{L}_{\text{pose}} = \sum_{t=0}^{T-1} \mathcal{V}(p_t, \hat{p}_t) + \omega_5 \sum_{t=1}^{T-1} \mathcal{U}(p_{t-1}, p_t, \hat{p}_{t-1}, \hat{p}_t),$$
$$\mathcal{L}_{\text{eye}} = \sum_{t=0}^{T-1} \mathcal{V}(e_t, \hat{e}_t) + \omega_5 \sum_{t=1}^{T-1} \mathcal{U}(e_{t-1}, e_t, \hat{e}_{t-1}, \hat{e}_t),$$
$$\tag{6}$$

where $\mathcal{V}(x_t, \hat{x}_t) = \|x_t - \hat{x}_t\|_2^2$, and $\mathcal{U}(x_{t-1}, x_t, \hat{x}_{t-1}, \hat{x}_t) = \|x_t - x_{t-1} - (\hat{x}_t - \hat{x}_{t-1})\|_2^2$ is to guarantee the temporal consistency between adjacent frames. ω_5 is a weight to balance these two terms. Here we use x to represent the ground-truth values of the predicted $\hat{x}$.

The loss of facial discriminator $\mathbf{D}^f$ is defined as:

$$\mathcal{L}_{\text{F-GAN}} = \arg \min_{\mathbf{G}^f} \max_{\mathbf{D}^f} \mathbb{E}_{\mathbf{f}, \mathbf{p}, \mathbf{e}}[\log \mathbf{D}^f(\mathbf{f}, \mathbf{p}, \mathbf{e})] + \tag{7}$$
$$\mathbb{E}_{\mathbf{a}, \mathbf{s}}[\log(1 - \mathbf{D}^f(\mathbf{G}^f(\mathbf{a}, \mathbf{s}))],$$

where the generator $\mathbf{G}^f$ is composed of two sub-generators $\mathbf{G}^{tem}$ and $\mathbf{G}^{loc}$, minimizing this objective function, while discriminator $\mathbf{D}^f$ are optimized for maximization. The final loss function is then defined as:

$$\mathcal{L}_{facial} = \omega_6 \mathcal{L}_{\text{F-GAN}} + \mathcal{L}_{\text{Reg}}. \tag{8}$$

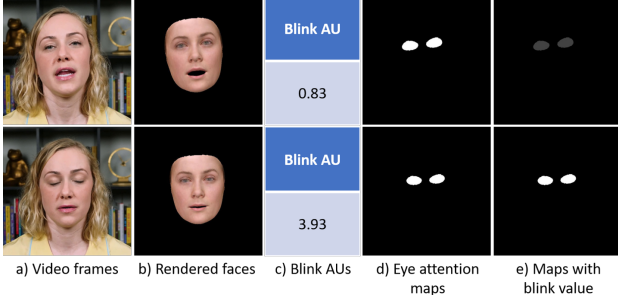a) Video frames   b) Rendered faces   c) Blink AUs   d) Eye attention maps   e) Maps with blink value

Figure 4. Illustration of the part sensitive encoding map. Our final generated encoding e) is composed of two parts: c) estimated blink AU, and d) eye attention maps.

3.3. Implicit Part-Sensitive Encoding

With the combination of geometry, texture and illumination coefficients from reference video and generated expression and head pose coefficients from input audio, we can render the 3D face with personalized head movements. 3D model describes the head pose better than 2D methods by rotating and translating the head. However, it is very difficult for 3D reconstruction methods to capture the subtle motions in the upper face region, especially for eye blinks as shown in Fig. 4. We combine the advantages of both 3D model and 2D action units to generate talking face with personalized head movements and natural eye blinks.

An intuitive solution is to directly concatenate the blink value to face image channels. However, convolutional neural networks can not recognize this channel for the eye part. We propose to use an eye attention map, which first locates the eye region, and then only changes the pixel value of this region according to the blinking AU value.

We first mark the vertices of the eye regions in 3D models. The vertices are identified from the mean face geometry of 3D Morphable Model (3DMM) by the following criteria:

$$(v_x - center_x)^2/4 + (v_y - center_y)^2 < th, \quad (9)$$

where v_x, v_y are the x, y values of vertex v, and $center_x$, $center_y$ are the x, y values of the center of each eye landmark. Threshold th is used to adjust the size of eye regions.

During the 3D face rendering, we locate pixels related with marked regions to generate the eye attention map for each face image in Fig. 4. Finally, we apply the normalized blinking value to the pixels in the eye attention map.

3.4. Rendering-to-Video Network

We employ the rendering-to-video network to translate the rendering images into the final photo-realistic images. Inspired by Kim *et al.* [20], we first combine the rendering image with the eye attention map to generate the training input data $\hat{I}$ with a size of $W \times H \times 4$ (rendering image with 3 channels, attention map with 1 channel). To ensure

temporal coherency, we use a window of size $2N_w$ with the current frame at the center of the window.

By following Chan *et al.*'s method [6], we train our rendering-to-video network consisting of a generator $\mathbf{G}^r$ and a multi-scale discriminator $\mathbf{D}^r = (\mathbf{D}_1^r, \mathbf{D}_2^r, \mathbf{D}_3^r)$ that are optimized alternatively in an adversarial manner. The generator $\mathbf{G}^r$ takes the stacked tensor $X_t = \{\hat{I}_t\}_{t-N_w}^{t+N_w}$ of size $W \times H \times 8N_w$ as input and outputs a photo-realistic image $\mathbf{G}^r(X_t)$ of the target person. The conditional discriminator $\mathbf{D}^r$ takes the stacked tensor X_t and a checking frame (either a real image I or a generated image $\mathbf{G}^r(X_t)$) as input and discriminates whether the checking frame is real or not. The loss function can be formulated as:

$$\mathcal{L}_{render} = \sum_{\mathbf{D}_i^r \in \mathbf{D}^r} (\mathcal{L}_{R-GAN}(\mathbf{G}^r, \mathbf{D}_i^r) + \lambda_1 \mathcal{L}_{FM}(\mathbf{G}^r, \mathbf{D}_i^r))$$
$$+ \lambda_2 \mathcal{L}_{VGG}(\mathbf{G}^r(X_t), I) + \lambda_3 \mathcal{L}_1(\mathbf{G}^r(X_t), I), \quad (10)$$

where $\mathcal{L}_{R-GAN}(\mathbf{G}^r, \mathbf{D}^r)$ is the GAN adversarial loss, $\mathcal{L}_{FM}(\mathbf{G}^r, \mathbf{D}^r)$ denotes the discriminator feature-matching loss proposed by [33], $\mathcal{L}_{VGG}(\mathbf{G}^r, I)$ is a VGG perceptual loss [18] for semantic level similarities, and $\mathcal{L}_1(\mathbf{G}^r, I)$ is an absolute pixel error loss.

The optimal network parameters can be obtained by solving a typical min-max optimization:

$$\mathbf{G}^{r*} = \arg \min_{\mathbf{G}^r} \max_{\mathbf{D}^r} \mathcal{L}_{render}(\mathbf{G}^r, \mathbf{D}^r). \quad (11)$$

4. Dataset Collection

As mentioned above, previous popular datasets mostly neglect the combination of explicit and implicit attributes. For example, GRID [13] provides a fixed head pose for talking head video, and some other datasets do not focus on the attributes of one specific person, *e.g.*, LRW [11] includes many short clips of different people. To jointly incorporate the explicit and implicit attributes for neural talking heads, we adopt the talking head dataset from Zhang *et al.* [38], with rich information, *i.e.*, dynamic head poses, eye motions, lip synchronization as well as the 3D face model for each frame in an automatic collection manner.

Audio preprocessing. We employ DeepSpeech [17] to extract speech features. DeepSpeech outputs the normalized log probabilities of characters in 50 Frames Per Second (FPS), which forms an array of size $50 \times D$ for each second. Here $D = 29$ is the number of speech features in each frame. We resample the output to 30 FPS using linear interpolation to match the video frames in our dataset, which generates an array of size $30 \times D$ for each second.

Head pose and eye motion field. To automatically collect the head pose as well as to detect the eye motions, we adopt OpenFace [2] for the face parameter generation of each video frame. The rigid head pose $\mathbf{p} \in \mathbb{R}^6$ is represented by Euler angles (pitch θ_x, yaw θ_y, roll θ_z) and a
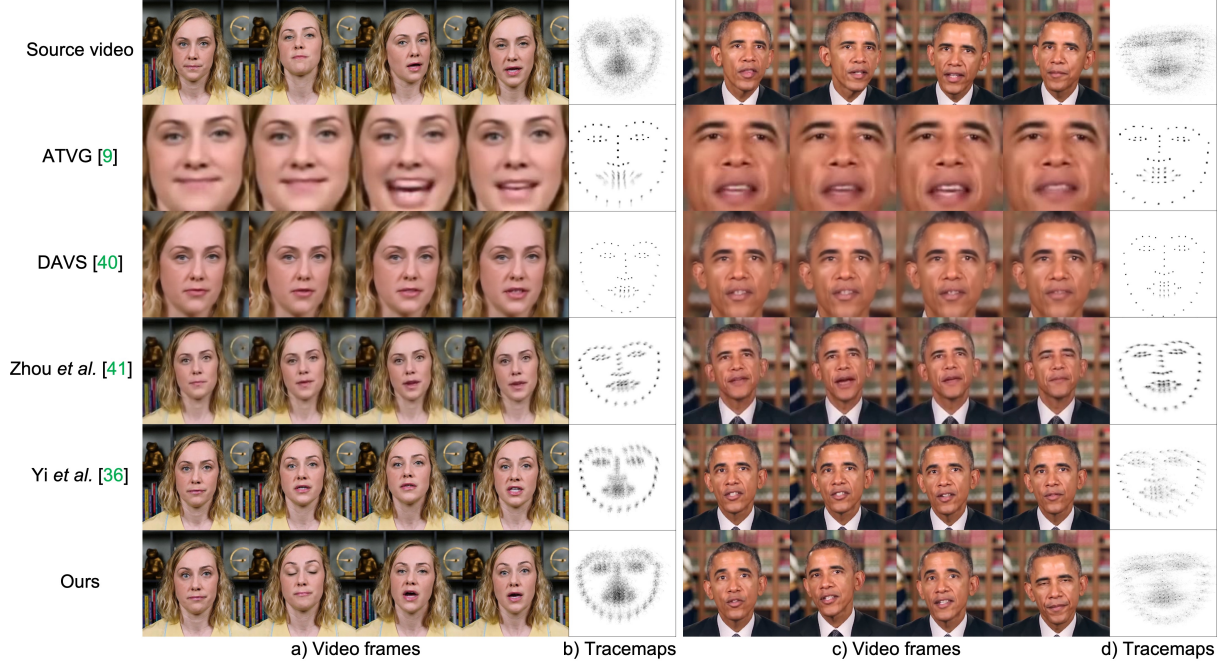
Figure 5. Comparisons with ATVG [9], DAVS [40], Zhou *et al.* [41] and Yi *et al.* [36]. The first row is corresponding video frames with input audio. a) and c) are the generated video frames. b) and d) are the corresponding tracemaps of facial landmarks in multiple frames. From the tracemaps we can see our generated head motions are highly consistent with the source videos.

3D translation vector $\mathbf{t} \in \mathbb{R}^3$. To depict the eye motions, Action Units (AUs) [15] are exploited to define the action intensities of muscle groups around the eye regions.

3D face reconstruction. To automatically generate 3D face models, we adopt Deng *et al.*'s method [14] to generate face parameters $[F_{id}, F_{exp}, F_{tex}, \gamma]$, where $F_{id} \in \mathbb{R}^{80}$, $F_{exp} \in \mathbb{R}^{64}$ and $F_{tex} \in \mathbb{R}^{80}$ are the coefficients for geometry, expression, and texture, respectively for the 3D Morphable Model (3DMM) [3]. $\gamma \in \mathbb{R}^{27}$ is the spherical harmonics (SH) [25] illumination coefficients. The parametric face model of 3DMM consists of a template triangle mesh with N vertices and an affine model which defines face geometry $S \in \mathbb{R}^{3N}$ and texture $T \in \mathbb{R}^{3N}$:

$$S = \overline{S} + B_{id}F_{id} + B_{exp}F_{exp},$$
$$T = \overline{T} + B_{tex}F_{tex}, \qquad (12)$$

where $\overline{S}$ and $\overline{T} \in \mathbb{R}^{3N}$ denote the averaged face geometry and texture, respectively. B_{id}, B_{tex} and B_{exp} are the PCA basis of geometry, texture and expression adopted from the Basel Face Model [22] and FaceWareHouse [5].

Dataset statistics. The proposed dataset contains rich samples of more than 450 video clips which are collected from the videos used by Agarwal *et al.* [1]. Each video clip lasts for around 1 minute. We re-normalize all videos to 30 FPS, forming 535,400 frames in total. We further divide our dataset using a train-val-test split of 5-1-4. Each video clip in our dataset has a stable fixed camera and appropriate lighting, with only one speaker for stable face generation.

5. Experiments

5.1. Network Learning

Training. Our training scheme consists of two steps: (1) We first optimize FACIAL-GAN loss $\mathcal{L}_{facial}$ based on our whole training dataset, which mainly considers the general mapping between the audio and the generated attributes. (2) Given a reference video $\mathcal{V}$, we first extract the audio feature a, 3D face model, head pose p and eye blinking AU e. Then we fine-tune the FACIAL-GAN loss $\mathcal{L}_{facial}$ to learn the personalized talking style. Meanwhile, we optimize the rendering loss $\mathcal{L}_{render}$ to learn the mapping from rendered faces with eye attention maps to the final video frame.

Testing. Given the input audio, we first use the fine-tuned FACIAL-GAN to map the audio features to expression $\mathbf{f}$, head pose $\mathbf{p}$ and eye blink AU $\mathbf{e}$, which have the personalized talking style of the reference video. Then, we render the corresponding face images and eye attention maps, and convert them to a photo-realistic target video with the personalized talking style.

Implementation details. All experiments are conducted on a single NVIDIA 1080-Ti GPU using Adam [21] optimizer and a learning rate of 0.0001. We use a sliding window of size $T = 128$ to extract training samples of audio

Figure 6. Comparison to 2D GAN-based Vougioukas *et al.*'s [32] and Chen *et al.*'s [7] methods.

Figure 7. Comparison to Suwajanakorn *et al.*'s Synthesizing Obama [28] and Thies *et al.*'s Neural Voice Puppetry [29].

and video, and use a sliding distance of 5 frames between neighboring samples. A total of 50 epochs are trained with a batch size of 64 for general training. For the fine-tuning step, it takes 10 epochs with a batch size of 16. For the rendering-to-video network, the training process takes 50 epochs with a batch size of 1 with a learning rate decay for the last 30 epochs. In our experiments, the parameters in Eqs. (4), (6), (8) are $\omega_1 = 2$, $\omega_2 = 1$, $\omega_3 = 5$, $\omega_4 = 10$, $\omega_5 = 10$, and $\omega_6 = 0.1$. The parameters in Eq. (10) are $\lambda_1 = 2$, $\lambda_2 = 10$, and $\lambda_3 = 50$.

5.2. Comparison with State-of-the-Arts

5.2.1 Qualitative Comparison

As shown in Fig. 5, we first compare our results with four state-of-the-art audio-driven talking face video generation methods: ATVG [9], DAVS [40], Zhou *et al.* [41] and Yi *et al.* [36]. The ATVG and DAVS are 2D-based methods that take an audio sequence and a target image as input. The head pose and eye blink in their generated videos are fully static, which is a contradiction to the human sense. Zhou *et al.* [41] uses face landmarks as an intermediate step to generate talking face videos. However, using landmark positions to represent them cannot fully capture head pose dynamics. By using 3D face model, Yi *et al.* [36] generates photo-realistic talking videos. However, its generated head pose shows subtle movements, as can be seen from its tracemaps in Fig. 5, and the eye blinks are completely static. In contrast, with collaborative learning of explicit and implicit attributes, our method generates realistic talking face video with personalized head movements and realistic eye blinking. We also compare our method with 2D GAN-based Vougioukas *et al.*'s [32] and Chen *et al.*'s [7] methods in Fig. 6. The comparisons are conducted on the

same characters, and our results are of higher visual quality than all other methods.

We further compare our method with the audio-driven facial reenactment methods [28, 29], which first generate the lip area that is in sync with the input audio, and compose it to an original video. We show qualitative results based on the same character - Barack Obama, and facial reenactment methods can generate photo-realistic talking videos in Fig. 7. However, their generated implicit attributes (*e.g.*, head pose and eye blinks) are exactly from the original video, which means that the length of the generated video is limited by the reference video, otherwise, a special video connection technology must be used.

5.2.2 Quantitative Evaluation

Landmark distance metric: We apply the Landmark Distance (LMD) proposed by Chen *et al.* [8] for accuracy evaluation of lip movement.

Sharpness metric: The frame sharpness is evaluated by the cumulative probability blur detection (CPBD).

Lip-sync metric: We evaluate the synchronization of lip motion with input audio by SyncNet [12], which calculates the Audio-Visual (AV) Offset and Confidence scores to determine the lip-sync error.

Eye blink metric: The average human eye blinking rate is 0.28-0.45 blinks/s and average inter-blink duration is 0.41s [26]. These reference values will vary for different people and speaking scenarios.

Personalization metric: One high-quality synthesizing should be able to generate personalized characteristics for different identities. To evaluate this personalization capability, we train a typical N-way pose classification network (see supplementary for more information) of N identities by matching their input head poses or eye blinks.

As shown in Tab. 1, it can be found our model can generate personalized attributes and surpasses most existing methods [9, 40, 41, 36], which verifies the effectiveness of our collaborative learning network.

5.3. Ablation Studies

In our FACIAL-GAN module, we generate temporal correlation feature $\mathbf{z}$ and local phonetic feature $\mathbf{c}$, and then use decoders to convert those two features into facial attributes including expression, head pose, and eye blinking AU. Here we evaluate the importance of these two features. As shown in Fig. 8, the generated videos result in -3 / 4.309 (AV offset / confidence) in second row (w/o $\mathbf{G}^{loc}$), and -2 / 4.051 in third row (w/o $\mathbf{G}^{tem}$), and -2 / 5.127 for our combined method. In addition, from the tracemaps we can see the head motion is more static without the $\mathbf{G}^{tem}$ network.

We also evaluate our part-sensitive encoding module. For video frames without eye attentions, the blinking fre-

Table 1. Quantitative comparisons of state-of-the-art models and our model. Better values are highlighted in bold.

Method	LMD	CPBD	AV offset	AV confidence	blinks/s	blink dur. (s)	Personalization blinks	head pose
ATVG [9]	5.31	0.119	-1	4.048	N/A	N/A	N/A	N/A
DAVS [40]	4.54	0.144	-3	2.796	N/A	N/A	N/A	N/A
Zhou [41]	4.97	0.271	-2	5.086	0.42	0.21	0.40	0.52
Yi [36]	3.82	0.291	-2	4.060	N/A	N/A	N/A	0.30
Ours	**3.57**	**0.314**	-2	**5.216**	0.47	0.26	**0.73**	**0.85**

a) Video frames   b) Tracemaps
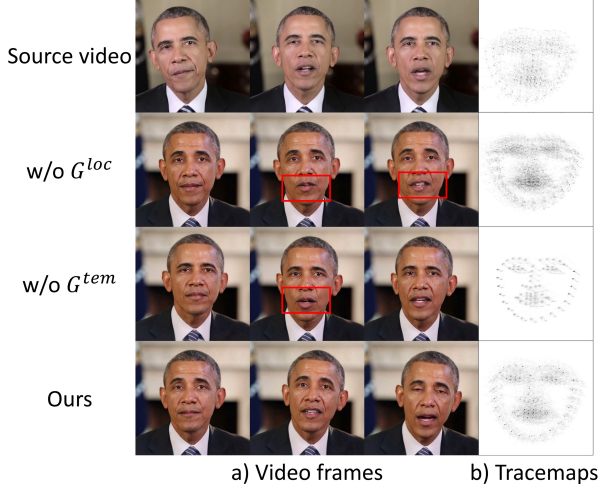
Figure 8. Ablation study for $\mathbf{G}^{tem}$ and $\mathbf{G}^{loc}$.

a) Distribution of blinks for videos   b) Distribution of blink duration
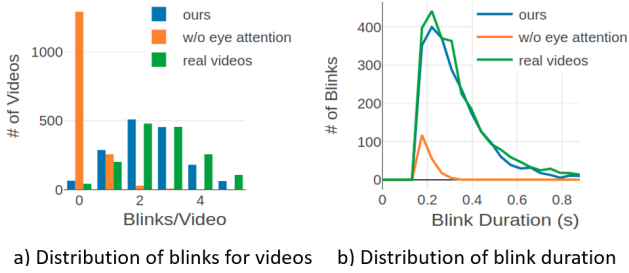
Figure 9. Ablation study for part-sensitive encoding maps.

quencies of generated videos are extremely low and unnatural. We sample 1,569 video clips from our testing dataset and each clip is about 4 seconds. Then we calculate the distribution of eye blinks for video clips and the distribution of blink duration time with and without eye attention maps in Fig. 9. The blinking frequency and duration from the results of our method are similar to that of the real videos.

5.4. User Studies

We conduct user studies to compare the generated results from the human perspective. 20 volunteers participate in the study to evaluate the video quality based on four criteria: 1) photo-realistic image quality, 2) audio-lip synchronization, 3) natural head motion, 4) realistic eye blinks. Participants

Table 2. User study analyses of our model with state of the arts.

Implicit Att.	Method	Image	Lip	Pose	Blink
✗	ATVG [9]	-1.1	0.2	-1.6	-1.3
	DAVS [40]	-1.2	-1.7	-1.7	-1.6
✓	Zhou [41]	0.8	0.9	0.6	1.0
	Yi [36]	**1.6**	0.8	0.9	0.2
	Vougioukas [32]	-0.6	1.1	-1.3	-1.4
	Chen [7]	-1.2	0.2	0.3	-0.3
	Ours	**1.6**	**1.2**	**1.7**	**1.4**
	Real video	1.9	2.0	1.9	2.0

are required to evaluate each video 4 times based on the evaluation criteria. The evaluation scores include: -2 (very bad), -1 (bad), 0 (normal), 1 (good), 2 (very good). Every participant learns 3 examples first, and then evaluates 18 videos of either real video or synthesized from face generation methods. We calculate the average value of evaluated results for each method. Participants' evaluation results are summarized in Table 2, which indicates that our method is better than state-of-the-art methods.

6. Discussion and Future Work

In this work, we focus on implicit attribute learning targeting for natural head poses and eye blinks. It should be noted that human talking videos still have other implicit attributes, *e.g.*, gaze motion, body and hand gestures, microexpressions, etc., which are guided by other dimensions of information and may require specific designs of other network components. We hope our FACIAL framework could be a stepping stone towards the future exploration of implicit attribute learning along with those directions.

Acknowledgments

References

[1] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. 6

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, 2016. 5

[3] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 6

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, pages 353–360, 1997. 1, 2

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 6

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5933–5942, 2019. 5

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision (ECCV)*, pages 35–51, 2020. 2, 7, 8

[8] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 7

[9] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7832–7841, 2019. 1, 2, 6, 7, 8

[10] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *British Machine Vision Conference (BMVC)*, 2017. 1, 2

[11] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision (ACCV)*, pages 87–103, 2016. 5

[12] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision (ACCV)*, pages 251–263, 2016. 7

[13] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 5

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[15] Paul Ekman and Wallace V Friesen. *Manual for the facial action coding system*. 1978. 6

[16] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics (TOG)*, 21(3):388–398, 2002. 1, 2

[17] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 4, 5

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 5

[19] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 3

[20] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3, 5

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[22] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 6

[23] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*, pages 484–492, 2020. 2

[24] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 3

[25] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 497–500, 2001. 6

[26] Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks, (IJCNN)*, 2020. 1, 2, 7

[27] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 919–925, 2019. 1, 2

[28] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2, 7

[29] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision (ECCV)*, pages 716–731. Springer, 2020. 2, 7

[30] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 3

[31] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–40, 2019. 1, 2

[32] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019. 1, 2, 7, 8

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 5

[34] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[35] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *European Conference on Computer Vision (ECCV)*, pages 670–686, 2018. 3

[36] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with natural head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2, 6, 7, 8

[37] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9459–9468, 2019. 3

[38] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 3d talking face with personalized pose dynamics. In *International conference on Computational Visual Media (CVM)*, 2021. 5

[39] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment. In *British Machine Vision Conference (BMVC)*, 2019. 3

[40] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 9299–9306, 2019. 1, 2, 6, 7, 8

[41] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 2, 6, 7, 8