One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning

Suzhen Wang[1], **Lincheng Li**[1], **Yu Ding**[1*], **Xin Yu**[2]

[1] Virtual Human Group, Netease Fuxi AI Lab
[2] University of Technology Sydney
{wangsuzhen, lilincheng, dingyu01}@corp.netease.com
xin.yu@uts.edu.au

Abstract

Audio-driven one-shot talking face generation methods are usually trained on video resources of various persons. However, their created videos often suffer unnatural mouth shapes and asynchronous lips because those methods struggle to learn a consistent speech style from different speakers. We observe that it would be much easier to learn a consistent speech style from a specific speaker, which leads to authentic mouth movements. Hence, we propose a novel one-shot talking face generation framework by exploring consistent correlations between audio and visual motions from a specific speaker and then transferring audio-driven motion fields to a reference image. Specifically, we develop an Audio-Visual Correlation Transformer (AVCT) that aims to infer talking motions represented by keypoint based dense motion fields from an input audio. In particular, considering audio may come from different identities in deployment, we incorporate phonemes to represent audio signals. In this manner, our AVCT can inherently generalize to audio spoken by other identities. Moreover, as face keypoints are used to represent speakers, AVCT is agnostic against appearances of the training speaker, and thus allows us to manipulate face images of different identities readily. Considering different face shapes lead to different motions, a motion field transfer module is exploited to reduce the audio-driven dense motion field gap between the training identity and the one-shot reference. Once we obtained the dense motion field of the reference image, we employ an image renderer to generate its talking face videos from an audio clip. Thanks to our learned consistent speaking style, our method generates authentic mouth shapes and vivid movements. Extensive experiments demonstrate that our synthesized videos outperform the state-of-the-art in terms of visual quality and lip-sync.

Introduction

[1]Synthesizing audio-driven photo-realistic portraits is of great importance to various applications, such as digital human animation (Ji et al. 2021; Zhu et al. 2021), visual dubbing in movies (Prajwal et al. 2020; Ha et al. 2020) and fast short video creation (Zhou et al. 2021; Zeng et al. 2020).

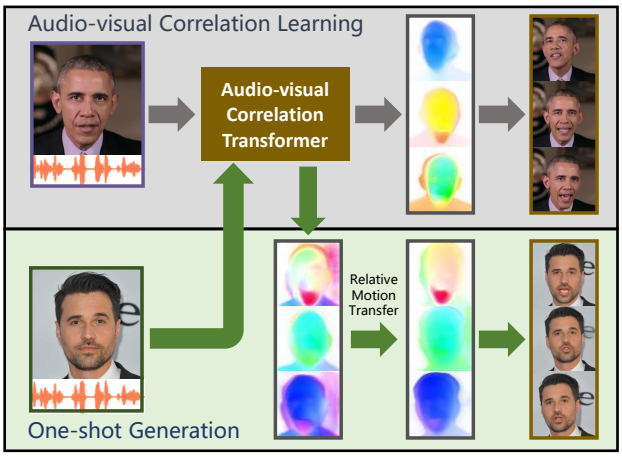

Figure 1: Illustration of our proposed talking head generation framework. Our approach takes one-shot reference image as input and generates audio-driven talking faces with rhythmic head motions, natural mouth shapes and accurate lip synchronization. Although our audio-visual correlation model is trained on a specific speaker, our framework supports arbitrary one-shot reference image and voice as input and renders the photo-realistic talking face videos.

One-shot talking face generation methods are designed to animate video portraits for unseen speakers and voice. When watching a synthetic talking head video, humans are mainly affected by three aspects: visual quality (clear and jitter-free), natural head motions, and synced lip movements. Existing one-shot methods (Chen et al. 2019; Prajwal et al. 2020; Zhou et al. 2020, 2021; Lahiri et al. 2021) are usually trained on video resources of various persons, and their results suffer unnatural lip shapes and bad lip-sync. This is mainly because their networks trained on multiple speech styles try to fit the common style among different identities while treating personalized variations as noise. Therefore, it is very challenging to synthesize natural and synchronous lip movements for one-shot speakers.

We observe that it is much easier to learn a consistent speech style from a specific speaker. Motivated by this, we propose a new one-shot talking face generation framework, which first learns a consistent speaking style from a single

*Corresponding author.

[1]https://github.com/FuxiVirtualHuman/AAAI22-one-shot-talking-face

speaker and then animates vivid videos of arbitrary speakers with new speech audio from the learned style. The key insight of the proposed method is to utilize the advantage of authentic lip movements and rhythmic head motions learned from an individual, and then to seek migration from an individual to multiple persons. Put differently, our method explores a consistent speech style between audio and visual motions from a specific speaker and then transfers audio-driven keypoint based motion fields to a reference image for talking face generation.

Towards this goal, we firstly develop a speaker-independent Audio-Visual Correlation Transformer (AVCT) to obtain keypoint-base dense motion fields (Siarohin et al. 2019) from audio signals. To eliminate the timbre effect among different identities, we adopt phonemes to represent audio signals. With the input phonemes and head poses, the encoder is expected to establish the latent pose-entangled audio-visual mapping. Considering vivid mouth movements are closely related to audio signals (e.g., the mouth amplitudes are affected by the fierce tones), we use the embedded acoustic features as the query of the decoder to modulate mouth shapes for more vivid lip movements. Moreover, due to the keypoint representation of a reference image, AVCT is agnostic against appearances of the training speaker, allowing us to manipulate face images regardless of different identities. Furthermore, considering different faces have diverse shapes, these variations would lead to different facial motions. Thus, a relative motion transfer module (Siarohin et al. 2019) is employed to reduce the motion gap between the training identity and the one-shot reference. Once obtaining the dense motion field, we generate talking head videos by an image renderer.

Thanks to our learned consistent speaking style, our method is able to not only produce talking face videos of the training speaker on par with speaker-specific methods, but also animate vivid portrait videos for unseen speakers with more accurate lip synchronization and more natural mouth shapes than previous one-shot talking head approaches. Remarkably, our method can also address talking faces with translational and rotational head movements whereas prior arts usually handle rotational head motions. Extensive experimental results on widely-used VoxCeleb2 and HDTF demonstrate the superiority of our proposed method.

In summary, our contributions are three-fold:

- We propose a new audio-driven one-shot talking face generation framework, which establishes the consistent audio-visual correlations from a specific speaker instead of learning from various speakers as in prior arts.
- We design an audio-visual correlation transformer that takes phonemes and facial keypoint based motion field representations as input, thus allowing it to be easily extended to any other audio and identities.
- Although the audio-visual correlations are only learned from a specific speaker, our method is able to generate photo-realistic talking face videos with accurate lip synchronization, natural lip shapes and rhythmic head motions from a reference image and a new audio clip.

Related Work

Animating talking faces from audio or text has received more attention in the field of artificial intelligence. As there exists a considerable audio-visual gap, early works (Edwards et al. 2016; Taylor et al. 2017; Pham, Cheung, and Pavlovic 2017; Karras et al. 2017; Zhou et al. 2018; Cudeiro et al. 2019) focus on driving animations of 3D face models. With the development of image generation (Yu and Porikli 2016; Yu et al. 2019b; Li, Yu, and Yang 2021; Yu et al. 2019a), an increasing number of works have been proposed for 2D photo-realistic talking face generation. These methods can mainly be divided into two categories, speaker-specific methods and speaker-arbitrary methods.

Speaker-specific Talking Face Generation

For a given new speaker, speaker-specific methods retrain part or all of their models on the videos of that speaker. Most works (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Song et al. 2020; Yi et al. 2020; Fried et al. 2019; Thies et al. 2020; Li et al. 2021; Lahiri et al. 2021; Ji et al. 2021; Zhang et al. 2021a,b; Lahiri et al. 2021) synthesize photo-realistic talking head videos guided by 3D face models. Suwajanakorn, Seitz, and Kemelmacher-Shlizerman (2017) synthesize videos from audio in the region around the mouth. Several methods (Thies et al. 2020; Li et al. 2021) consist of speaker-independent components relying on 3D face models and speak-specific rendering modules. Fried et al. (2019) present a framework for text based video editing. Guo et al. (2021) propose the audio-driven neural radiance fields for talking head generation.

Speaker-arbitrary Talking Face Generation

Speaker-arbitrary methods aim to build a single universal model for various subjects. Some works (Chung, Jamaludin, and Zisserman 2017; Chen et al. 2018; Song et al. 2018; Zhou et al. 2019; Chen et al. 2019; Vougioukas, Petridis, and Pantic 2019; Das et al. 2020) focus on learning a mapping from audio to the cropped faces, but their fixed poses and cropped faces in the videos are unnatural for human observations. Other works (Wiles, Koepke, and Zisserman 2018; Chen et al. 2020; Prajwal et al. 2020; Zhou et al. 2020; Zhang et al. 2021c; Wang et al. 2021; Zhou et al. 2021) try to create vivid results with natural head poses. Zhou et al. (2020) predict 2D landmarks with the head pose, and then generate talking faces. Chen et al. (2020) and Zhang et al. (2021c) use 3D face models to acquire landmarks and the dense flow respectively as their intermediate representations. Zhou et al. (2021) propose a pose-controllable talking face generation method by implicitly modulating audio-visual representations. Prajwal et al. (2020) only edit the lip-synced mouth regions of a reference image from audio. Although Prajwal et al. (2020) and Zhou et al. (2021) are able to generate videos with poses, their reference head poses are obtained from another videos rather than audio. Wang et al. (2021) employ keypoint-based dense motion fields as intermediate representations and achieve rhythmic head motions in generated videos, but their lip-sync exhibits artifacts. Among previous speaker-arbitrary works, only Zhang et al.

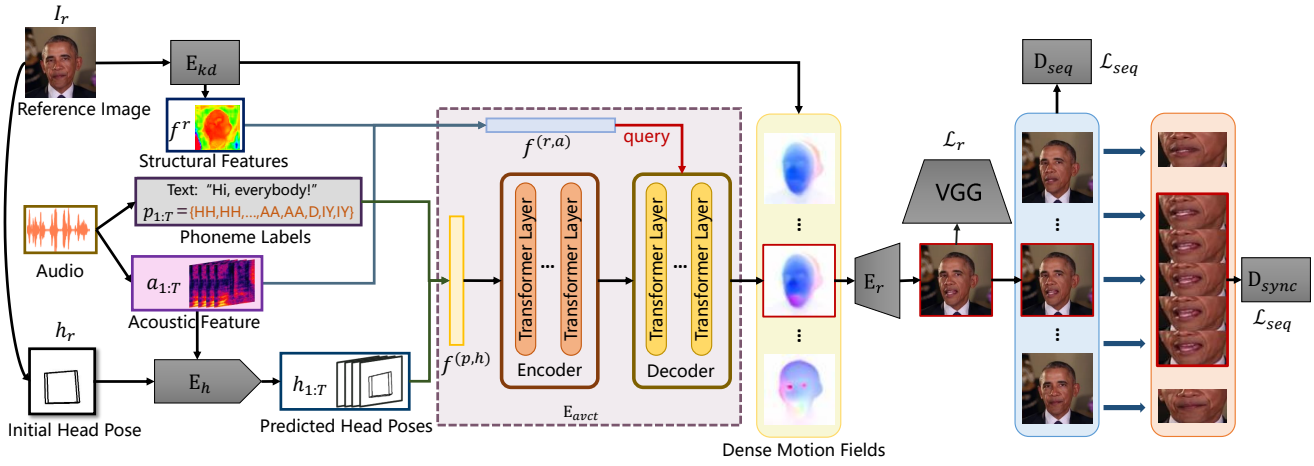

Figure 2: Illustration of our pipeline. We first extract an initial pose h_r from a reference image, and extract acoustic features $a_{1:T}$ and phoneme labels $p_{1:T}$ from the audio. The latent representation of keypoints of the reference image f^r is extracted by the keypoint detector $\mathbf{E}_{kd}$. The head motion predictor $\mathbf{E}_h$ predicts the head motion sequence $h_{1:T}$ from the input $\{a_{1:T}, h_r\}$. Then $\{a_{1:T}, p_{1:T}, h_{1:T}, f^r\}$ are sent to the audio-visual correlation transformer $\mathbf{E}_{avct}$ to generate pose-aware keypoint-based dense motion fields. Finally, the image renderer $\mathbf{E}_r$ renders the output videos. We also use a temporal discriminator $\mathbf{D}_{seq}$ and a lip-sync discriminator $\mathbf{D}_{sync}$ to improve the temporal stability and lip synchronization respectively in training. Particularly, in the inference stage, the generated dense motions are refined with a relative motion transfer module.

(2021c) and Wang et al. (2021) create videos with translational and rotational head movements while keeping background still in generated videos. However, since all these methods are trained on the corpus of multiple speakers, they often struggle to learn a consistent speaking style, and their results suffer unnatural mouth shapes.

Proposed Method

We propose a new talking face generation framework to make audio-driven portrait videos for arbitrary speakers by learning audio-visual correlations on a specific speaker. Giving a reference image I_r and an audio clip A, our method creates talking face images $\boldsymbol{y}$=$I_{1:T}$. The whole pipeline is shown in Figure 2. Our pipeline consists of four modules: (1) a head motion predictor $\mathbf{E}_h$ estimates the head motion sequence $h_{1:T}$ ($h_i \in \mathbb{R}^6$ includes the 3D rotation and the 3D translation) from audio; (2) a keypoint detector $\mathbf{E}_{kd}$ extracts initial keypoints from the reference image; (3) an Audio-visual Correlation Transformer (AVCT) $\mathbf{E}_{avct}$ maps audio signals to keypoint-based dense motion fields; and (4) an image renderer $\mathbf{E}_r$ produces output images from the dense motion fields. we adopt the architectures of $\mathbf{E}_{kd}$ and $\mathbf{E}_r$ as in FOMM (Siarohin et al. 2019).

We extract the audio channels from the training videos and transform them into audio features and phonemes as pre-processing. To be consistent with videos at 25 fps, we extract acoustic features $a_i \in \mathbb{R}^{4\times41}$ and one phoneme label $p_i \in \mathbb{R}$ per 40ms. The acoustic features include Mel Frequency Cepstrum Coefficients (MFCC), Mel-filterbank energy features (FBANK), fundamental frequency and voice flag. The phoneme is extracted by a speech recognition tool[2].

[2]https://cmusphinx.github.io/wiki/phonemerecognition/

Audio-visual Correlation Transformer (AVCT)

The core of the proposed method is to build accurate audio-visual correlations which can be extended to any other audio and identities. Such correlations are learned via a speaker-independent audio-visual correlation transformer. Considering the high temporal coherence, $\mathbf{E}_{avct}$ takes the assembled features in a sliding window as input. Specifically, for the i-th frame, $\mathbf{E}_{avct}$ takes the paired conditioning input $c_i = \{f_r, a_{i-n:i+n}, h_{i-n:i+n}, p_{i-n:i+n}\}$ and outputs the keypoints $k_i \in \mathbb{R}^{N\times2}$ and their corresponding Jacobian $j_i \in \mathbb{R}^{N\times2\times2}$. f^r is the latent representation of keypoints from the reference image I_r through the keypoint detector $\mathbf{E}_{kd}$. n indicates the window length and is set to 5 in our experiments. N is the number of keypoints and is set to 10. The paired (k_i, j_i) represents the dense motion field (Siarohin et al. 2019). The head motions $h_{1:T}$ are extracted by OpenFace (Baltrusaitis et al. 2018).

The proposed AVCT is able to aggregate dynamic audio-visual information within the temporal window, thus creating more accurate lip movements. To model the correlations among different modalities, we employ Transformer (Vaswani et al. 2017) as the backbone of AVCT due to its powerful attention mechanism. For a better extension to any other audio and identities, we carefully design the Encoder and the Decoder of E_{avct} as follows.

Encoder. We employ the phoneme labels as input instead of acoustics features to bridge the timbre gap between the specific speaker and arbitrary ones. We establish a latent pose-entangled audio-visual mapping by encoding the input phonemes $p_{i-n:i+n}$ and poses $h_{i-n:i+n}$ in the encoder. The attentions between sequential frames $2n$+1 allow for obtaining the refined latent mouth motion representation at frame

i. Specifically, we employ a 256-dimension word embedding (Levy and Goldberg 2014) to represent the phoneme label, then reshape and upsample it as $f_i^p \in \mathbb{R}^{1\times64\times64}$. h_i is converted to the projected binary image $f_i^h \in \mathbb{R}^{1\times64\times64}$ as in (Wang et al. 2021). Then, the concatenated features $\{f_i^p, f_i^h\}$ are fed into a residual convolution network consisting of 5 2× downsampling ResNet blocks (He et al. 2016) in order to obtain the assembled feature $f_i^{(p,h)} \in \mathbb{R}^{1\times512}$.

The input to the encoder is the sequential features $f^{(p,h)} \in \mathbb{R}^{(2n+1)\times512}$ by concatenating all the frame features along the temporal dimension. Since the architecture of the transformer is permutation-invariant, we supplement $f^{(p,h)}$ with fixed positional encoding (Vaswani et al. 2017).

Decoder. Practically, the mouth amplitude is affected by the loudness and energy of the audio in addition to the phoneme. To create more subtle mouth movements, we employ the acoustics features in the decoding phase for capturing energy changes. We extract audio features $f_i^a \in \mathbb{R}^{32\times64\times64}$ from a_i using an upsampling convolution network. To reduce the dependency on the identities, we cannot directly take the reference image as input but employ the latent representation, $f^r \in \mathbb{R}^{32\times64\times64}$, of the keypoints of the reference image I_r. f^r is extracted from the pretrained keypoint detector $\mathbb{E}_{kd}$. It mainly retains the pose-based structural information of body, face and background, weakening identity-related information. Such initial structural information dominates the low-frequency holistic layout in the generated dense motion fields.

f^r is repeated by $2n$+1 times. Then the concatenation of f_i^r and f_i^a is fed into another residual convolutional network to obtain the embedding $f_i^{(r,a)}$. Similarly, we acquire the features $f^{(r,a)} \in \mathbb{R}^{(2n+1)\times512}$ by concatenation, and supplement it with positional encodings. $f^{(r,a)}$ is used as the initial query of the decoder to modulate the layout of the body, head and background, and to refine the subtle mouth shape. Following the standard transformer, the decoder creates $2n$+1 embeddings. Only the i-th embedding is taken and projected to keypoints k_i and Jacobians j_i with two different linear projections.

Batched Sequential Training

Since AVCT generates the dense motion fields of each frame individually, we develop a batched sequential training strategy to improve the temporal consistency. The training samples on each batch consist of the T successive conditional inputs $c_{i:T}$ of T successive images from the same video. Then, we generate image sequence $\hat{I}_{1:T}$ in parallel in each batch. This design allows us to apply constraints on the image sequence within each batch rather than on single images. We call the above strategy as Batched Sequential Training (BST). In addition to the pixel loss on each frame image, the sequential constraint is imposed by a temporal discriminator $\mathbf{D}_{seq}$. Besides, as the common pixel reconstruction loss is insufficient to supervise the lip-sync, we employ another lip-sync discriminator $\mathbf{D}_{sync}$ to improve the lip-sync.

Temporal Discriminator. $\mathbf{D}_{seq}$ follows the structure of PatchGAN (Goodfellow et al. 2014; Isola et al. 2017; Yu and Porikli 2017a,b; Yu et al. 2018). We stack the T successive image frames along the channel dimension as the input of $\mathbf{D}_{seq}$. $\mathbf{D}_{seq}$ tries to distinguish whether the input is natural or synthetic. $\mathbf{D}_{seq}$ and $\mathbf{E}_{avct}$ are learned jointly with the generative-adversarial learning.

Lip-sync Discriminator. $\mathbf{D}_{sync}$ employs the structure of SyncNet (Chung and Zisserman 2016) in Wav2Lip (Prajwal et al. 2020). $\mathbf{D}_{sync}$ is trained to discriminate the synchronization between audio and video by randomly sampling an audio window that is either synchronous or asynchronous with a video window. The discriminated frame lies in the middle of the window, and the window size is set to 5. $\mathbf{D}_{sync}$ computes the visual embedding e_v from an image encoder and the audio embedding e_a from an audio encoder. We adopt the cosine-similarity to indicate the probability whether e_v and e_a are synchronous:

$$P_{sync} = \frac{e_v \cdot e_a}{max(\|e_v\|_2 \cdot \|e_a\|_2, \epsilon)}. \tag{1}$$

Loss Function. Based on the batched sequential training, the loss function for each batch image sequence is defined as:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{seq}(\hat{I}_{1:T}) + \frac{\lambda_{sync}}{T-4}\sum_{i=3}^{T-2}\mathcal{L}_{sync}(\hat{I}^{crop}_{i-2:i+2}) + \\ + \frac{1}{T}\sum_{i=1}^{T}(\lambda_v\mathcal{L}^{mul}_{vgg}(\hat{I}_i, I_i) + \lambda^P_{eq}\mathcal{L}^K_{eq}(\hat{k}_i) + \lambda^J_{eq}\mathcal{L}^J_{eq}(\hat{j}_i))\end{aligned} \tag{2}$$

where $\mathcal{L}_{seq}$ is the GAN loss of $\mathbf{D}_{seq}$. $\mathcal{L}_{sync}$ is the lip-sync loss from the pretrained $\mathbf{D}_{sync}$ and is defined as $-log(P_{sync})$. Note that $\hat{I}^{crop}$ means the cropped mouth area and that we ignore the boundary frames to fit the temporal input of $\mathbf{D}_{sync}$. We crop the mouth region in each training iteration dynamically according to the detected bounding boxes of real videos. $\mathcal{L}^{mul}_{vgg}$ is the multi-layer perceptual loss that relies on a pretrained VGG network. $\mathcal{L}^K_{eq}$ and $\mathcal{L}^J_{eq}$ are the equivariance constraint loss (Siarohin et al. 2019) to ensure the consistency of estimated keypoints and jacobians. In our experiments, T is set to 24 (on RTX 3090), λ_{sync},λ_v,λ^P_{eq} and λ^J_{eq} are set to 10,1,10,10 respectively.

Head Motion Predictor

The head motion predictor $\mathbf{E}_h$, developed to generates $h_{1:T}$ in the inference stage, is also trained on the specific speaker. $\mathbf{E}_h$ adopts the network structure of the head motion predictor of Audio2Head (Wang et al. 2021) but has two differences. First, instead of being trained on a large number of identities, $\mathbf{E}_h$ is trained on a specific speaker. Therefore, in order to avoid overfitting to the appearance of the specific speaker, we replace the input reference image of Audio2Head with the projected binary pose image. Secondly, for the convenience of the relative motion transfer (see below), the starting point of the predicted head pose sequence should be the same as the head pose of the reference image. Hence, we add an $\mathbf{L}1$ loss term between the head poses

Method	HDTF					VoxCeleb2				
	FID↓	CPBD↑	LMD↓	AVOff(→0)	AVConf↑	FID↓	CPBD↑	LMD↓	AVOff(→0)	AVConf↑
Wav2Lip	0.180	**0.790**	0.289	-2.92	6.97	0.203	0.541	0.273	-2.92	6.65
MakeitTalk	0.210	0.694	0.546	-2.93	4.87	0.230	0.550	0.482	-2.83	4.38
Audio2Head	0.176	0.732	0.483	**0.33**	3.90	0.224	0.532	0.314	0.50	2.47
FGTF	0.187	0.738	0.387	1.05	4.24	0.212	0.559	0.283	0.491	4.54
PC-AVS	0.238	0.725	0.318	-3.00	**7.18**	0.276	0.514	**0.251**	-3.00	6.83
Ground Truth	0	0.827	0	0.15	8.58	0	0.612	0	-2.33	7.16
Ours	**0.172**	0.751	**0.271**	**-0.33**	7.09	**0.194**	**0.564**	0.252	**-0.08**	**6.98**

Table 1: The quantitative results on HDTF and VoxCeleb2.

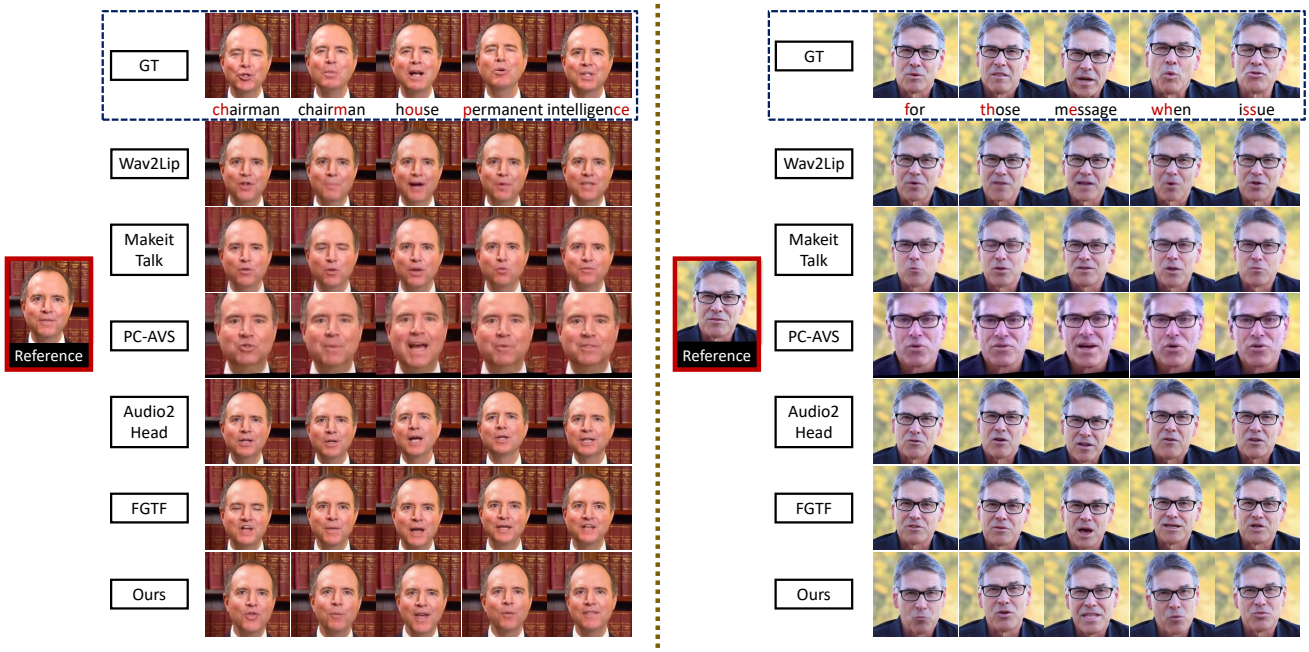

Figure 3: Comparisons with the speaker-arbitrary methods. We select the frames that pronounce the same phonemes marked in red. Please see our demo videos for more details.

of the first predicted frame and the reference image to the primitive loss function when training $\mathbf{E}_h$.

Relative Motion Transfer

As the generated motion fields are inevitably entangled with the specific speaker, we adopt the relative motion transfer (Siarohin et al. 2019) in the inference stage to reduce the motion gap between the training identity and the one-shot reference image. We transfer the relative motion between $(\hat{k}_1, \hat{j}_1)$ and $(\hat{k}_{1:T}, \hat{j}_{1:T})$ to (k_r, j_r). (k_r, j_r) are detected from the reference image. This operation is defined by:

$$\hat{k}'_i = \hat{k}_i - \hat{k}_1 + k_r, \quad \hat{j}'_i = \hat{j}_i \hat{j}_1^{-1} j_r. \tag{3}$$

Then, we use $(\hat{k}'_{1:T}, \hat{j}'_{1:T})$ to render the output videos.

Experiments

Dataset. We collect Obama Weekly Address videos from Youtube. Since we aim to generate talking face videos with translational and rotational head movements, we crop and resize the original videos to 256×256 pixels as in FOMM (Siarohin et al. 2019) without aligning speakers' noses across frames. After processing all the crowdsourcing videos, we obtain 20 hours of videos of Obama. Although our audio-visual correlation transformer is only trained on Obama, we employ two in-the-wild audio-visual datasets to evaluate our method, HDTF (Zhang et al. 2021c) and VoxCeleb2 (Chung, Nagrani, and Zisserman 2018).

Evaluation Metrics. We conduct quantitative evaluations on several metrics that have been wildly used in previous methods. As the generated videos have different head motions from ground truth, we use the Fréchet Inception Distance (FID) (Heusel et al. 2017) and the Cumulative Probability of Blur Detection (CPBD) (Narvekar and Karam 2009) to evaluate the image quality. To evaluate the mouth shape and lip synchronization, we adopt the Landmark Distance (LMD) (Chen et al. 2019) around mouths and audio-

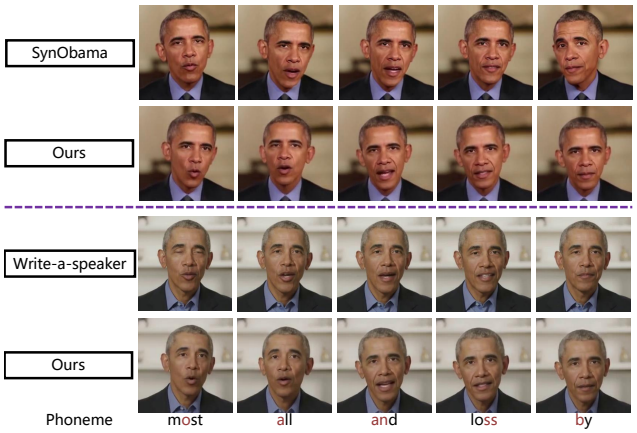

Figure 4: Comparisons with the speaker-specific methods. We select the frames that pronounce the same phonemes.

Method	FID↓	CPBD↑	LMD↓	AVOff(→0)	AVConf↑
w/o Pho	0.155	0.740	0.376	-0.503	6.13
w/o Aud	0.151	**0.747**	0.293	-0.538	6.59
w/o Dec_{kp}	0.184	0.707	0.418	-0.308	5.24
w/o Kp	0.210	0.704	0.461	**-0.170**	4.92
w/o BST	0.151	**0.747**	0.388	-0.743	4.55
w/o $\mathbf{D}_{sync}$	0.150	0.741	0.385	-0.385	5.53
Full	**0.148**	0.740	**0.274**	-0.417	**7.17**

Table 2: Results of ablation study on HDTF.

visual metrics (AVOff and AVConf) proposed in SyncNet (Chung and Zisserman 2016). Note that we calculate the normalized relative landmark distance instead of the absolute landmark distance to avoid the influence of both head poses and image resolution.

Implementation Details. All models are implemented by PyTorch, and we adopt Adam (Kingma and Ba 2014) optimizer for all experiments. Before training our AVCT $\mathbf{E}_{avct}$, we train the keypoint detector and image renderer on the combination of VoxCeleb (Nagrani, Chung, and Zisserman 2017) and Obama datasets to obtain the pretrained $\mathbf{E}_{kd}$ and $\mathbf{E}_r$. $\mathbf{D}_{sync}$ is trained with a fixed learning rate 1e-4 on videos of Obama. $\mathbf{E}_{kd}$, $\mathbf{E}_r$ and $\mathbf{D}_{sync}$ are frozen when training $\mathbf{E}_{avct}$ on the videos of Obama. $\mathbf{E}_{avct}$ is trained on 4 GPU for about 5 days using the batched sequential training mechanism, with an initial learning rate of 2e-5 and a weight decay of 2e-7. $\mathbf{E}_h$ is trained on a single GPU for about 12 hours with a learning rate of 1e-4.

Quantitative Evaluation

We compare our method with recent state-of-the-art methods, including Wav2Lip (Prajwal et al. 2020), MakeitTalk (Zhou et al. 2020), Audio2Head (Wang et al. 2021), FGTF (Zhang et al. 2021c), and PC-AVS (Zhou et al. 2021). The samples of each method are generated using their released codes with the same reference image and audio. The reference images are specially cropped to fit the input of PC-AVS. Since Wav2Lip and PC-AVS cannot obtain head poses from audio, the head poses are fixed in their generated videos. For other methods, the head poses are controlled separately by each method. The quantitative results are reported in Table 1. Our method achieves the best performance under most of evaluation metrics on HDTF and VoxCeleb2. As Wav2Lip only edits the mouth regions and keeps most parts of the reference image unchanged, it reaches the highest CPBD score on HDTF, but their synthetic mouth areas are noticeably blurry, as visible in Figure 3. These results validate that our method achieves high-quality videos, even though our audio-visual correlation transformer is learned from a specific speaker.

Figure 5: Samples are generated from different representations of the same audio clip, *i.e.*, phoneme features and the combination of phoneme and audio features. Here, the mouth shapes are further controlled by the intensity of the pronunciation in our method.

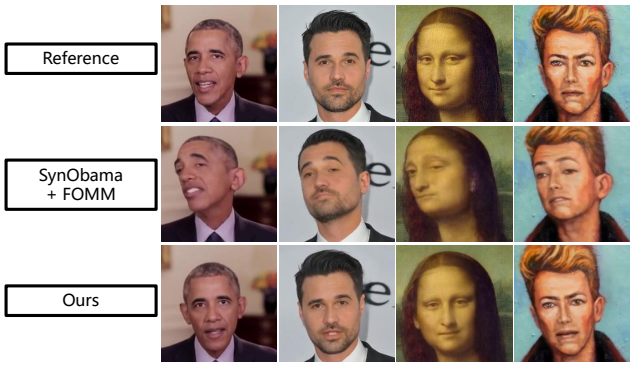

Figure 6: Compare our results with videos created by the combination of SynObama and FOMM. The combinated method is sensitive to the initial pose, while our method preserves the authentic head motions.

Qualitative Evaluation

Comparisons with Speaker-arbitrary Methods. We first compare our method with speaker-arbitrary (one-shot) methods qualitatively. The results are shown in Figure 3. Among all methods, our method creates the most accurate mouth shape and preserves the best identity (see our demo video for more clearly comparisons). Only Wav2Lip and PC-AVS achieve similar lip-sync to ours, but their mouth shapes look mechanical and unnatural because these methods struggle to

Method	Wav2Lip	MakeitTalk	Audio2Head	FGTF	PC-AVS	Ground Truth	Ours
Lip Sync Quality	3.80	2.65	2.07	2.43	3.74	4.96	**4.32**
Head Movement Naturalness	1.21	2.04	3.79	3.74	1.79	4.89	**4.11**
Video Realness	1.68	1.70	3.21	3.08	2.14	4.89	**3.86**

Table 3: Results of user study. Participants rate each video from 1 to 5. Large scores indicate better visual quality. Here, the average scores across 21 videos are reported.

produce consistent talking styles. Moreover, both of them cannot obtain head poses from audio and PC-AVS can only deal with aligned faces. Notably, PC-AVS alters the identity information of the reference image in Figure 3. While MakeitTalk creates subtle head motions, its results are obviously out of sync. Audio2Head, FGTF and our method are able to produce talking face videos with moving head poses. However, Audio2Head still suffer inferior lip-sync while our method produce authentic lip-sync. In addition, FGTF produces distorted mouth shapes in its results while our generated mouths look very natural.

Comparisons with Speaker-specific Methods. We compare our method with two speaker-specific methods, *i.e.*, SynObama (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017) and Write-a-speaker (Li et al. 2021). We first extract the reference image and audio from their demo videos and then generate our results. The comparisons are shown in Figure 4. As shown in Figure 4, our method synthesizes comparable videos of Obama with the methods that are customized for Obama. SynObama synthesizes the region around the mouth from audio, and uses compositing techniques to borrow the rest regions from real videos. This composition sometimes results in visible artifacts around the mouth. Our method is an end-to-end approach without requiring additional editing, and achieves accurate mouth shapes than Write-a-speaker. Please see our supplementary video for more details.

Ablation Study

To evaluate the effectiveness of each component in our framework, we conduct an ablation study with 7 variants: (1) remove phonemes from the encoder (**w/o Pho**), (2) remove audio features from the decoder (**w/o Aud**), (3) remove the keypoint features in the decoder (**w/o $\mathbf{Dec_{kp}}$**), (4) replace the extracted keypoint features from $\mathbf{E}_{kd}$ with the reference image (**w/o Kp**), (5) remove the temporal discriminator $\mathbf{D}_{seq}$ and lip-sync discriminator $\mathbf{D}_{sync}$ (**w/o BST**), (6) only remove $\mathbf{D}_{sync}$ (**w/o $\mathbf{D}_{sync}$**), and (7) our full model (**Full**). For evaluation, we replace generated head motions with poses extracted from real videos to create the samples. The numerical results on HDTF are shown in Table 2. As all the variants employ the same pretrained image renderer, most of them achieve similar FID and CPBD scores. However, it can be seen that the image quality drops dramatically when removing the reference image or replacing the keypoint features with the reference image. Without the supervision of $\mathbf{D}_{seq}$ and $\mathbf{D}_{sync}$, the results show poor temporal stability and bad lip synchronization. The model **w/o Pho** fails to extend to the unseen timbre by removing the input of phonemes, producing out-of-sync videos. In Table 2, the model **w/o Aud** obtains good quantitative results without using audio features. However, as seen in Figure 5, the audio features indeed affect the vivid mouth movements.

We also conduct another ablation study to evaluate the superiority of the proposed method. We compare our method with the combination of a speaker-specific work (*i.e.*, SynObama) and a expression transfer work (*i.e.*, FOMM). Specifically, we transfer the expressions in videos created by SynObama to the reference image using FOMM. The results are shown in Figure 6. Since FOMM requires the initial poses of the reference image and one-shot to be similar while the pose of the first frame from SynObama is possibly different from that of the reference one, the combination of SynObama and FOMM would lead to inferior results. In contrast, our model is able to preserve the initial pose and thus generates satisfactory facial motions.

User Study

We conduct a user study of 19 volunteers base on their subjective perception of talking head videos. We create 3 videos for each method with the same input and obtain 21 videos in total. We adopt the questions used in the user study of PC-AVS (Zhou et al. 2021), and participants are asked to give their ratings (1-5) of each video on three questions: (1) the Lip sync quality, (2) the naturalness of head movements, and (3) the realness of results. The mean scores are listed in Table 3. Note that we do not offer reference poses for Wav2Lip and PC-AVS, so their scores on head movements and video realness are reasonably low. Our method outperforms competing methods in all the aspects, demonstrating the effectiveness of our method.

Conclusion

In this paper, we propose a novel framework for one-shot talking face generation from audio. Particularly, our method learns consistent audio-visual correlations from a single speaker, and then transfers the talking styles to different subjects. Differs from prior one-shot talking face works, our method provides a new perspective to address this task and achieves vivid videos of arbitrary speakers. The extensive quantitative and qualitative evaluations illustrate that our method is able to generate photo-realistic talking-face videos with accurate lip synchronization, natural lip shapes and rhythmic head motions from a reference image and a new audio clip. Besides face photography, we can animate talking head videos for non-photorealistic paintings, demonstrating a promising generalization ability of our method.

References

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 59–66. IEEE.

Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; and Xu, C. 2020. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, 35–51. Springer.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 520–535.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7832–7841.

Chung, J. S.; Jamaludin, A.; and Zisserman, A. 2017. You said that? *arXiv preprint arXiv:1705.02966*.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, 251–263. Springer.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10101–10111.

Das, D.; Biswas, S.; Sinha, S.; and Bhowmick, B. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, 408–424. Springer.

Edwards, P.; Landreth, C.; Fiume, E.; and Singh, K. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4): 1–11.

Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D. B.; Genova, K.; Jin, Z.; Theobalt, C.; and Agrawala, M. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.; Bao, H.; and Zhang, J. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. *arXiv preprint arXiv:2103.11078*.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. MarioNETte: Few-shot Face Reenactment Preserving Identity of Unseen Targets. In *AAAI*, volume 34, 10893–10900.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C. C.; Cao, X.; and Xu, F. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14080–14089.

Karras, T.; Aila, T.; Laine, S.; Herva, A.; and Lehtinen, J. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4): 1–12.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; and Bregler, C. 2021. 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*, 2755–2764.

Levy, O.; and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27: 2177–2185.

Li, L.; Wang, S.; Zhang, Z.; Ding, Y.; Zheng, Y.; Yu, X.; and Fan, C. 2021. Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1911–1920.

Li, P.; Yu, X.; and Yang, Y. 2021. Super-Resolving Cross-Domain Face Miniatures by Peeking at One-Shot Exemplar. *arXiv preprint arXiv:2103.08863*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Narvekar, N. D.; and Karam, L. J. 2009. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, 87–91. IEEE.

Pham, H. X.; Cheung, S.; and Pavlovic, V. 2017. Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 80–88.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32: 7137–7147.

Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2020. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*.

Song, Y.; Zhu, J.; Li, D.; Wang, X.; and Qi, H. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Taylor, S.; Kim, T.; Yue, Y.; Mahler, M.; Krahe, J.; Rodriguez, A. G.; Hodgins, J.; and Matthews, I. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4): 1–11.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, 716–731. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2019. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 1–16.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. *IJCAI*.

Wiles, O.; Koepke, A.; and Zisserman, A. 2018. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, 670–686.

Yi, R.; Ye, Z.; Zhang, J.; Bao, H.; and Liu, Y.-J. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*.

Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face super-resolution guided by facial component heatmaps. In *ECCV*, 217–233.

Yu, X.; Fernando, B.; Hartley, R.; and Porikli, F. 2019a. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *TPAMI*.

Yu, X.; and Porikli, F. 2016. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 318–333.

Yu, X.; and Porikli, F. 2017a. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI*.

Yu, X.; and Porikli, F. 2017b. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 3760–3768.

Yu, X.; Shiri, F.; Ghanem, B.; and Porikli, F. 2019b. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *TPAMI*.

Zeng, X.; Pan, Y.; Wang, M.; Zhang, J.; and Liu, Y. 2020. Realistic Face Reenactment via Self-Supervised Disentangling of Identity and Pose. In *AAAI*, volume 34, 12757–12764.

Zhang, C.; Ni, S.; Fan, Z.; Li, H.; Zeng, M.; Budagavi, M.; and Guo, X. 2021a. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, C.; Zhao, Y.; Huang, Y.; Zeng, M.; Ni, S.; Budagavi, M.; and Guo, X. 2021b. FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3867–3876.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021c. 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*, 3661–3670.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9299–9306.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. MakeltTalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6): 1–15.

Zhou, Y.; Xu, Z.; Landreth, C.; Kalogerakis, E.; Maji, S.; and Singh, K. 2018. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4): 1–10.

Zhu, H.; Luo, M.-D.; Wang, R.; Zheng, A.-H.; and He, R. 2021. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 1–26.