Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset

Zhimeng Zhang Lincheng Li Yu Ding * Changjie Fan
Virtual Human Group, Netease Fuxi AI Lab
{zhangzhimeng, lilincheng, dingyu01, fanchangjie }@corp.netease.com

Abstract

One-shot talking face generation should synthesize high visual quality facial videos with reasonable animations of expression and head pose, and just utilize arbitrary driving audio and arbitrary single face image as the source. Current works fail to generate over 256×256 resolution realistic-looking videos due to the lack of an appropriate high-resolution audio-visual dataset, and the limitation of the sparse facial landmarks in providing poor expression details. To synthesize high-definition videos, we build a large in-the-wild high-resolution audio-visual dataset and propose a novel flow-guided talking face generation framework. The new dataset is collected from youtube and consists of about 16 hours 720P or 1080P videos. We leverage the facial 3D morphable model (3DMM) to split the framework into two cascaded modules instead of learning a direct mapping from audio to video. In the first module, we propose a novel animation generator to produce the movements of mouth, eyebrow and head pose simultaneously. In the second module, we transform animation into dense flow to provide more expression details and carefully design a novel flow-guided video generator to synthesize videos. Our method is able to produce high-definition videos and outperforms state-of-the-art works in objective and subjective comparisons[*].

1. Introduction

Given one reference facial image and one driving audio, one-shot talking face generation aims at synthesizing a talking avatar video with reasonable facial animations corresponding to the driving audio. Talking face generation is of importance for many applications, including virtual assistants, mixed realities, animation movies, and so forth. Due to its wide applications, talking face generation draws considerable attention for a long time.

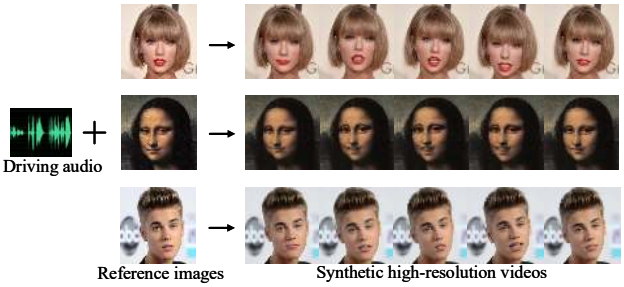


Figure 1. Our method synthesizes high-resolution talking face videos with one driving audio and one reference facial image.

While many works[7, 6, 18, 13, 51, 5, 48, 29, 42, 49, 4] make great efforts to synthesize realistic-looking videos, the generation of high-resolution videos is still a challenge. Current best work[49] just generate videos with 256×256 resolution(see Figure10(e) for example), however, directly employing their model on 512×512 image will get blurry results (see Figure10(f) for example). Several factors result in this challenge.

The first reason is that there are no appropriate datasets for high-resolution talking face generation. Table 1 illustrates some common audio-visual datasets (all available datasets are listed in [3]). As shown in Table 1, current audio-visual datasets consist of in-the-wild datasets and in-the-lab datasets. In-the-wild datasets contain larger scale and more subjects, but they all lack video resolution. There are two main reasons: On one hand, their videos are collected from the internet published in the past 2∼5 years, and at that time the internet videos generally have low resolution. On the other hand, most in-the-wild datasets do not focus on the task of talking face generation, e.g., Voxceleb[26, 8] is built for speaker identification and LRW[9] is built for word recognition, so they do not pay attention to the video resolution. For in-the-lab datasets, while they record high resolution face videos, the number of subjects and sentences is limited because of the expensive labor costs. The largest MEAD[43] only records 159 sentences with 60 actors.

The second reason is that previous works are not de-

[*]Yu Ding is the corresponding author.
[*]The HDTF dataset etc. for research purpose are at https://github.com/MRzzm/HDTF

Table 1. Statistics of current common audio-visual datasets.

Dataset name	Environment	Year	Resolution	Subject	Hours	sentence
LRW [9]	Wild	2016	360P∼480P	1k+	173	1k
Voxceleb1[26]	Wild	2017	360P∼720P	1251	352	100k
Voxceleb2[8]	Wild	2018	360P∼720P	6112	2442	1128k
GRID[11]	Lab	2006	720×576	34	27.5	51
RAVDESS[24]	Lab	2018	1280×1024	24	7	8
MEAD[43]	Lab	2020	1920×1080	60	40	159
Our HDTF	Wild	2020	**720P∼1080P**	**300+**	15.8	10k+

signed reasonably to handle high-resolution videos and are limited by the input of sparse facial landmarks. Initial works[7, 5, 42] directly utilize an end-to-end framework to synthesize the video from audio. Their synthetic results even have a low definition on 128×128 videos. Other recent advances[6, 49, 13, 4] leverage facial landmarks to split the pipeline into two cascaded modules. They produce sparse facial landmarks in the first module, and further generate videos from synthetic landmarks in the second module. Two modules are trained separately to alleviate the pressure of the network, thus lead to high visual quality results. However, in the second module, their methods are still hard to generate high resolution videos. We carefully discuss the reasons in Section 7. On one hand, some works directly utilize the network to learn the sophisticated mapping from landmark to image. This mapping become too complex to handle on high-resolution videos, e.g., [49] synthesize blurry results on 512×512 resolution(see Figure10(g) and Figure11(a) for example). On the other hand, although some works carefully design their network to explicitly model the process of image synthesis, the sparse landmark is too coarse and lose many facial expression details, e.g., [34] synthesize facial image with inaccuracy mouth shape and poor wrinkles(see Figure11(c) for example).

In order to achieve above challenge and promote the development of high-resolution talking face generation, we first build a large in-the-wild high-resolution audio-visual dataset, named High-definition Talking Face Dataset (HDTF). The HDTF dataset is collected from youtube website published in recent two years and consists of about 16 hours 720P∼1080P videos. There are over 300 subjects and 10k different sentences in HDTF dataset. Our HDTF dataset has higher video resolution than previous in-the-wild datasets and more subjects/sentences than in-the-lab datasets.

Next, we propose a novel flow-guided framework to synthesize high visual quality videos. Figure 2 illustrates the pipeline of our method. Our work first leverages 3DMM[1] to split the framework into two cascaded modules, named audio-to-animation module and animation-to-video module. Compared with the facial landmarks, 3DMM is insensitive to noise due to the prior knowledge of the face. In audio-to-animation module, 3DMM is used to decouple the face into facial animation parameters (mouth, eyebrow and head pose) and we propose a novel style-specific animation generator to produce the full animation parameters with multi-task learning strategy. Our generator considers the difference of speaking style between different identity[50], and has capacity to synthesize subject-dependent animations. In animation-to-video module, we propose a flow-guided framework to synthesize high visual quality videos. Our method utilizes 3DMM to transform animation parameters to dense flow. Dense flow has benefits of providing richer facial details than sparse landmarks. Then, a novel video generator is proposed to synthesize talking face videos from dense flow. Our generator is carefully designed to explicitly control the process of frame generation, so it is easy to generate more realistic results.

Our contributions are summarized as follows:

- We build a large in-the-wild audio-visual dataset, with higher video resolution than previous in-the-wild datasets and more subjects/sentences than in-the-lab datasets.
- We propose a novel style-specific animation generator to produce specific style animation parameters depending on the reference identity.
- To the best of our knowledge, we are the first to utilize one animation generator with multi-task learning to produce the animation parameters of mouth, eyebrow and head pose simultaneously in one-shot talking face generation.
- We propose a novel carefully-designed flow-guided framework to synthesize higher visual quality videos than previous landmark-based approaches.

2. Related Work

2.1. Talking Face Generation

One-shot talking face generation. One-shot talking face generation is identity-independent. In the inference stage, the reference identity and driving audio are not restricted to appear in training data. Early works[7, 51, 5, 29, 42] always take two sub-encoders to extract identity features and spoken features from the reference image and driving audio. Then, they fuse two features as input into a decoder to synthesize talking face videos in an end-to-end fashion. For more accurate lip-sync results, some works use the audio-mouth mutual information loss[51], audio-mouth correlation loss[5] and audio-visual disentangle learning[48]. In order to improve the visual quality, some works add an extra deblurring module[7] or just repair the mouth region[51, 5, 29].

Recent advances[6, 4, 13, 49] utilize facial landmarks to split the framework into two cascaded modules. In the first module, PCA component[6, 4] or spatial displacement[13, 49] of the landmark are used to represent facial animation parameters. [4] take two networks to synthesize the facial expression and head motion. [49] utilize two branches to

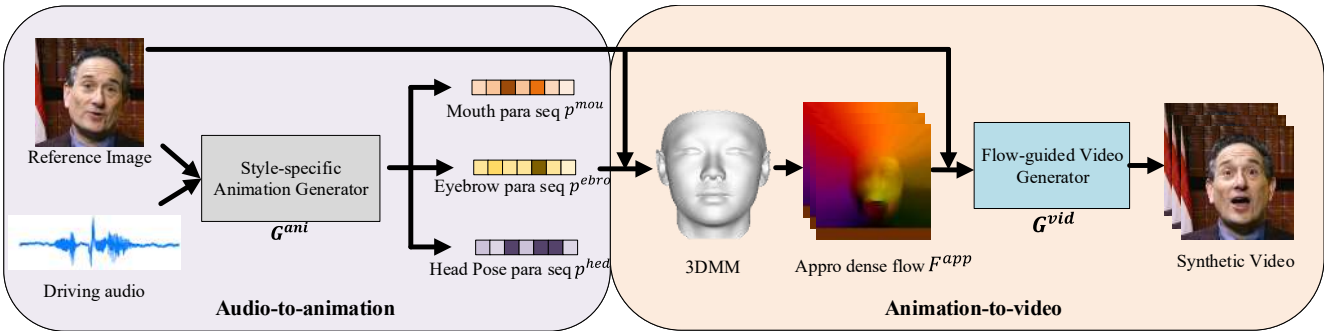


Figure 2. The pipeline of our method. Our method has two cascaded modules: audio-to-animation module (purple part) and animation-to-video module (orange part).

synthesize mouth displacement and head pose/eyebrow displacement. However, both above two works[4, 49] separately train the two animation generators. In our work, we use one animation generator to synthesize mouth, eyebrow and head pose simultaneously with multi-task learning. In the second module, they employ various landmark-to-video generators to synthesize talking head videos. Different from them, our method takes dense flow as input to generate more realistic videos.

Person-specific talking face generation. Person-specific talking face generation has benefits of synthesizing high-resolution talking face videos because the identity is in training data. [36] carefully design a framework to synthesize Obama videos with about 17 hours footage. [16] utilize dynamic programming algorithm to reduce the training data to 1 hour. [35] train a shared generator for all identities and they only require 15 minutes footage. [39] leverage a pre-trained audio-to-mouth model to reduce the required footage to 2∼3 minutes. [23] use a motion capture dataset to synthesize videos with emotion and rhythmic head pose. In our work, we synthesize videos with competitive resolution and only need one reference image for a new subject.

2.2. Animation Synthesis

Animation synthesis aims at generating animation trajectories to drive a pre-defined 3D talking avatar. In mouth animation generation, the mouth shape is related to spoken co-articulation[38]. Several works use CNN-based[22, 37, 12] or LSTM-based[32, 30] framework to capture co-articulation effects. Some works[22, 32, 30] focus on expressive animation generation. Other works[37, 12] focus on improving the generalization of input speech. In head pose/eyebrow animation generation[15, 14], there is a one-to-many mapping between speech and head pose/eyebrow[31], so [31] utilize Generative Adversarial Network(GAN)[17] to retain the diversity of head pose. Besides, head pose/eyebrow animation is related to speech prosody and syntactic structure[15, 14], so [49] take self-attention module[41] to capture this long-time dependencies.

3. Dataset

A large in-the-wild high resolution audio-visual dataset, named High-definition Talking Face Dataset (HDTF), is built for talking face generation. Some snapshots of HDTF are shown in Figure 3. In order to collect high quality videos, we only collect online videos published in recent two years. HDTF dataset consists of about 362 different videos for 15.8 hours. The resolution of origin video is 720P or 1080P. In our work, a landmark detector is first leveraged to crop the face region. The crop window is fixed during each video. Then, each cropped video is resized into 512×512 (the second row in Figure3). Due to the high quality of origin videos, our final cropped videos also have high visual quality.

Then, the 3DMM[1] is employed to decouple the cropped face into facial shape parameters and facial animation parameters(mouth, eyebrow and head pose). The 3DMM is a bilinear morphable model. It is represented as

$$M(c^s, c^e) = M_0 + \sum_{i=1}^{60} c_i^s \cdot V_i^s + \sum_{j=1}^{33} c_j^e \cdot V_j^e \qquad (1)$$

where $M(c^s, c^e)$ represent the 3D facial mesh point. M_0 is the average facial mesh. $\{V_i^s\}_{i=1}^{60}$ and $\{V_j^e\}_{j=1}^{33}$ are the linear basis of facial shape and facial expression. c^s and c^e represent the coefficient of the basis. $\{V_j^e\}_{j=1}^{33}$ is combined with 28 mouth basis and 5 eyebrow basis.

We take scaled orthogonal projection[19] to reconstruct 3D face according to facial landmark points, e.g., dlib. The objective is

$$\begin{aligned} &\arg\min_{c^s, c^e, s, R, t} E(c^s, c^e, s, R, t) \\ &= \arg\min_{c^s, c^e, s, R, t} \sum_{k=1}^{K} \delta_k [p_k - (s \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} RM(c^s, c^e)^{(k)} + t)]^2 \end{aligned} \qquad (2)$$

where p_k is the k_{th} landmark point and δ_k represent its weight. K is the number of landmark points. $R \in SO(3)$ is rotation matrix and $t \in R^2$ represents translation vector. s is the scale value. In our paper, we solve above objective with weighted least squares method.

After the 3D face restruction, in each video, we extract

Figure 3. The snapshots of HDTF dataset.

the face shape parameter $p^s \in R^{60}$, mouth parameter sequence $p^{mou} = \{p_t^{mou} \in R^{28}\}_{t=1}^T$, eyebrow parameter sequence $p^{ebro} = \{p_t^{ebro} \in R^5\}_{t=1}^T$ and head pose parameter difference sequence $p^{hed} = \{p_t^{hed} \in R^5\}_{t=1}^T$. In our method, we do not directly synthesize head pose but synthesize the difference. The main reason is that the initial head pose in different videos are different. We also extract audio feature sequence $f^{audio} = \{f_t^{audio} \in R^{15}\}_{t=1}^T$. The audio feature consists of 13-dim MFCC feature and 2-dim pitch feature. The video frames is denoted as $I = \{I_t\}_{t=1}^T$. T is the length of frames in the video. Finally, our training data is represented as $\{I, p^{mou}, p^{ebro}, p^{hed}, f^{audio}, p^s\}$ in each video.

4. Proposed Method

Based on our HDTF dataset, as shown in Figure2, we propose a novel high-quality one-shot talking face generation framework. The framework consists of one audio-to-animation module and one animation-to-video module. In the first module, a novel style-specific audio-to-animation generator G^{ani} is designed to translate reference image and driving audio to full animation parameters. In the second module, animation parameters are first transformed to approximate dense flow F^{app} by the 3DMM. Then, F^{app} and reference image are input into a careful-designed flow-guided video generator G^{vid} to synthesize the talking face videos.

4.1. Audio-to-animation

Style-specific audio-to-animation generator G^{ani}. The structure of G^{ani} is illustrated in Figure 4. G^{ani} aims at translating reference image I^{ref} and driving audio f^{audio} into the style-specific animation parameters corresponding to reference face. The parameters consist of mouth parameter $\hat{p}^{mou}$, eyebrow parameter $\hat{p}^{ebro}$ and head pose parameter $\hat{p}^{hed}$. G^{ani} utilizes two steps to realize this purpose. In the first step (the purple part in Figure4), the style-specific audio representation $\hat{f}_{ref}^{audio}$ is computed from I^{ref} and f^{audio}. $\hat{f}_{ref}^{audio}$ encodes the speaking content of f^{audio} and the speaking style of I^{ref}. In the second step (orange part in Figure 4), $\hat{f}_{ref}^{audio}$ is used to synthesize animation parameters with specific speaking style.

In the first step, a CNN-based audio feature extractor is first employed to extract audio representation $\hat{f}^{audio}$ from f^{audio}. Then, considering that different identity has different speaking style[50], AdaIN[20] operation is taken to transform $\hat{f}^{audio}$ into the $\hat{f}_{ref}^{audio}$. In specifically, a pre-trained VGG-face model[2] is first used to extract identity embedding vector from the reference image I^{ref}. Then, the identity embedding is input into fully-connected layers to generate the scale and shift parameters of the AdaIN.

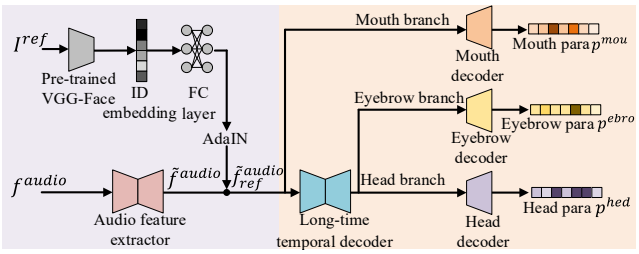


Figure 4. Structure of audio-to-animation generator G^{ani}.

Furthermore, in the second step, with three branches of the decoder, mouth, eyebrow and head pose are generated simultaneously according to $\hat{f}_{ref}^{audio}$. In mouth branch, a CNN-based mouth decoder is employed to decode $\hat{f}_{ref}^{audio}$ to $\hat{p}^{mou}$. In eyebrow and head pose branch, a long-time temporal decoder is first employed to capture the long-time dependencies. Different from [49], our long-time decoder is based on an encoder-decoder network, which has benefits of faster forward speed. Then, a CNN-based eyebrow decoder and a CNN-based head pose decoder are taken to synthesize $\hat{p}^{ebro}$ and $\hat{p}^{hed}$.

Loss function. In training stage, G^{ani} is trained with multi-task learning strategy. In mouth synthesis, we use L1 loss and LSGAN loss[25].L1 loss is written as

$$L_1^{mou} = \frac{1}{T}\sum_{t=1}^{T} \|p_t^{mou} - \hat{p}_t^{mou}\|_1, \tag{3}$$

where p_t^{mou} and $\hat{p}_t^{mou}$ are the real and synthetic mouth parameters. LSGAN loss is denoted as

$$L_{GAN}^{mou} = \min_{G^{ani}} \max_{D^{mou}} L_{GAN}(G^{ani}, D^{mou}). \tag{4}$$

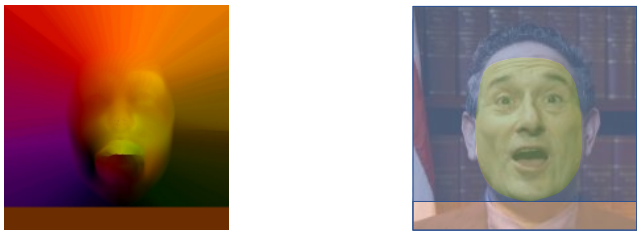
In eyebrow and head pose generation, we utilize Structural Similarity (SSIM) loss[47] and LSGAN loss. SSIM simulates the human visual perception and has benefits of extracting structural information. In our work, SSIM extends to evaluate the eyebrow and head pose on each parameter. SSIM loss in eyebrow generation is written as

$$L_{ssim}^{ebro} = 1 - \frac{1}{5}\sum_{i=1}^{5} \frac{(2\mu_i\hat{\mu}_i + \delta_1)(2cov_i + \delta_2)}{(\mu_i^2 + \hat{\mu}_i^2 + \delta_1)(\sigma_i^2 + \hat{\sigma}_i^2 + \delta_2))}, \tag{5}$$

where $\mu_i/\hat{\mu}_i$ and $\sigma_i/\hat{\sigma}_i$ are the mean and standard deviation of the i_{th} dimension of $p^{ebro}/\hat{p}^{ebro}$. cov_i is the covariance. δ_1 and δ_2 are two small constants. LSGAN loss in eyebrow generation is denoted as

$$L_{GAN}^{ebro} = \min_{G^{ani}} \max_{D^{ebro}} L_{GAN}(G^{ani}, D^{ebro}). \tag{6}$$

The loss in head pose generation has the same form (SSIM & GAN) as in eyebrow generation except for the parameter



(a) (b)

Figure 5. (a) Visualization of approximate dense motion flow F^{app} in pseudo color. (b) Different cropped parts in F^{app}, including inner face part (green), head-related part(blue) and upper torso part (orange).

dimension. The final objective function is written as

$$L(G^{ani}) = L^{mou}_{GAN} + L^{ebro}_{GAN} + L^{hed}_{GAN} + \lambda_{mou}L^{mou}_{1} + \lambda_{ebro}L^{ebro}_{ssim} + \lambda_{hed}L^{hed}_{ssim}, \quad (7)$$

λ_{mou}, λ_{ebro} and λ_{hed} represent the loss weights. All the GAN structures are conditional GAN, i.e., $D^{mou/ebro/hed}$ takes $\{f^{audio}, \hat{p}^{mou/ebro/hed}\}$ as input. The structure details are in supplementary materials.

4.2. Animation-to-video

In animation-to-video module, the animation parameters are first transformed to approximate dense motion flow F^{app} by 3DMM. However, limited by the ability of 3DMM, F^{app} is not accurate enough. Then, to solve above problem, a novel flow-guided video generator G^{vid} is proposed. G^{vid} is carefully designed to revise F^{app} and synthesize high visual quality videos.

Approximate dense motion flow F^{app}. F^{app} describes the approximate motion direction of each pixel between two frames. Figure 5 (a) visualizes the F^{app} in pseudo color. In the generation of F^{app}, given a pair of facial animation parameters, 3DMM is able to generate accurate dense motion flow in the inner face (the green part in Figure 5 (b)). However, 3DMM is incapable of describing the motion out of the face region(the blue and orange part in Figure 5 (b)). In order to solve this problem, we estimate the approximate motion flow out the facial region.

As shown in Figure 5 (b), we crop the facial image into three parts: the inner face part (green), the upper torso part (orange) and the head-related part (blue). In the inner face part, the dense motion flow is computed from 3DMM. In the upper torso part, we assume the upper torso moves with the head, so we take the average movements of inner face as the motion value in upper torso part. In the head-related part, we focus on the hair, ear and other ornaments, and assume they move rigidly follow the nearest facial edge. The flow of each pixel in head-related part is as same as its nearest facial edge pixel. Combining the flow of three parts, we obtain the final F^{app}. However, the background is ignored in the construction of F^{app}, so the flow value in background is absolutely incorrect. This incorrectness will be revised in G^{vid}.

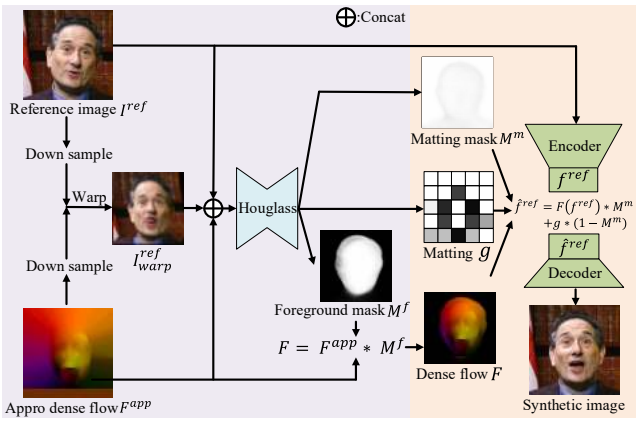


Figure 6. Structure of flow-guided video generator G^{vid}.

Flow-guided video generator G^{vid}. The structure of G^{vid} is shown in Figure 6. G^{vid} is designed to revise F^{app} and further synthesize high-resolution talking face videos. In order to realize above purpose, G^{vid} also contains two steps inside the network. In the first step (purple part in Figure 6), the network revises the F^{app}, and produces an accurate dense motion flow F, an intermediate matting image g and a matting mask M^m. In the second step (orange part in Figure 6), F, g and M^m are used to synthesize high quality videos.

In the first step(purple part in Figure 6), to revise the incorrect flow in background of F^{app}, we assume that the background is static, which is also used in many recent works[33, 34]. Upon this assumption, a foreground mask M^f is generated to transform F^{app} to accurate dense motion flow F. M^f is a soft mask with a range $0 \sim 1$. The transformation is written as

$$F = F^{app} * M^f \quad (8)$$

M^f revises the background to static. In order to generate M^f, inspired from [34], we first warp reference image I^{ref} with F^{app} to get warped image I^{ref}_{warp}. Then, I^{ref}, F^{app} and I^{ref}_{warp} are concatenated into a Hourglass network[28] to generate M^f. Besides, Hourglass network also outputs g and M^m for the second step.

In the second step(orange part in Figure 6), inspired from [45, 44], we synthesize the image by combining the warped version of I^{ref} and g. The combination is balanced by a matting mask M^m. To reduce the parameters of the network, inspired from [34], above combination is done in feature map space, and is written as

$$\hat{f}^{ref} = F(f^{ref}) * M^m + g * (1 - M^m) \quad (9)$$

f^{ref} represents the encoded feature map of I^{ref} by a CNN-based encoder. $F(.)$ is the warp operation with F. $\hat{f}^{ref}$ is the combined result. Finally, $\hat{f}^{ref}$ is input into a CNN-based decoder to synthesize facial image.

Loss function. In training stage, G^{vid} is trained with

Figure 7. Our synthetic results driving by the same audio.

LSGAN loss, perceptual loss[21] and feature matching loss[46]. The GAN loss is represented as

$$L_{GAN}^{vid} = \min_{G^{vid}} \max_{D^{vid}} L_{GAN}(G^{vid}, D^{vid}). \quad (10)$$

The perceptual loss is written as

$$L_{perc}^{vid} = \sum_{i=1}^{n} \frac{1}{W_i H_i C_i} \|N_i(I_t) - N_i(\hat{I}_t)\|_1, \quad (11)$$

where $N_i(.)$ denotes the i_{th} layer with $W_i * H_i * C_i$ elements of a specific VGG-19 network. The feature matching loss is written as

$$L_{FM}^{vid} = \sum_{j=1}^{m} \frac{1}{W_j H_j C_j} \|D_j^{vid}(I_t) - D_j^{vid}(\hat{I}_t)\|_1, \quad (12)$$

where $D_j^{vid}(.)$ is the j_{th} layer in D^{vid}. The final loss function of G^{vid} is written as

$$L(G^{vid}) = L_{GAN}^{vid} + \lambda_{perc} L_{perc}^{vid} + \lambda_{FM} L_{FM}^{vid}. \quad (13)$$

λ_{perc} and λ_{FM} are the weights of loss. The structure details of G^{vid} and D^{vid} are in supplementary materials.

5. Experiments and Results

In this section, we first display some synthetic results of our method. Then, we compare our method with state-of-the-art talking face generation works. Next, to validate the effectiveness of each sub-module, we also do quantitative and qualitative comparisons with other related works. Next, we do ablation study to evaluate the components in two sub-modules. Finally, an online user study is conducted to validate our proposed method.

5.1. Synthetic results

Figure 7 shows some high-resolution synthetic frames driven by the same audio. Our method synthesizes high visual quality results. We further draw the curve of animation parameters of three different identities driven by the same audio in Figure 8. Figure 8(a) draws the sequence of mouth parameter controlling the opening of mouth. While

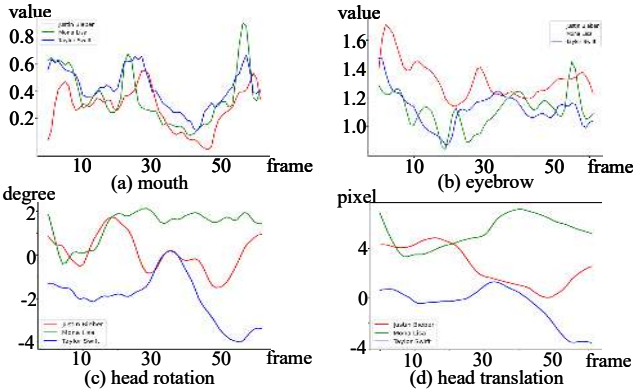


Figure 8. Animation parameters of three different subjects driven by the same audio. Different color represent different identity. The ordinate represents the value of parameter in (a)(b), the degree of head rotation in (c) and the pixel of head translation in (d).

there exist slight temporal shift and slight scale variance on mouth parameters, the tendency of the sequence is still similar on different subjects. It implies that the mouth shape mainly depends on the speech content. Figure 8 (b-d) draw the eyebrow parameter (eyebrow down), head rotation (roll) and head translation (horizontal) respectively. Obviously, there has more variance in these parameters. It demonstrates that our G^{ani} has ability to synthesize identity-dependent speaking styles for different reference subjects.

We also visualize the intermediate results, including F^{app}, I_{warp}^{ref}, M^f, F, M^m and synthetic frame, of the animation-to-video module in Figure9. The M^f focuses on separating the moving foreground and static background. The M^m leads to the matting operation impacts on the foreground. With the joint action of M^f and M^m, our G^{vid} synthesizes high visual quality videos.

5.2. Comparison with State-of-the-art

We compare our method with state-of-the-art one-shot talking face generation works[42, 4, 6, 49, 29] in Figure 10. Vougioukas et al.[42](Figure 10(a)) and Chen et al.[4,

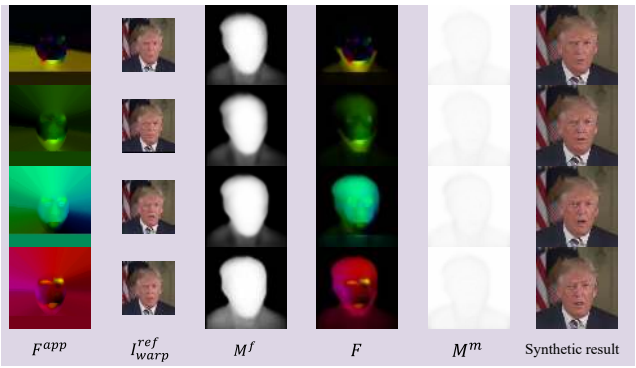


Figure 9. Visualization of intermediate results in animation-to-video module.

Table 2. Quantitative comparison of lip synchronization. Lower AV offset and higher AV confidence represents better lip synchronization.

Method	Real video	Chen et al.[6]	Prajwal et al.[29]	Zhou et al.[49]	Ours
AVOff↓	-1	-2	-2	-2	-2
AVConf↑	9.627	4.122	**5.227**	2.770	5.166

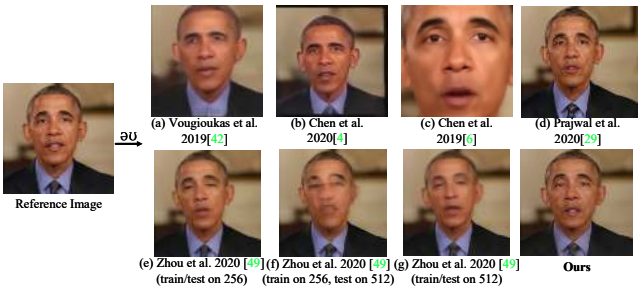


Figure 10. comparison with state-of-the-art works.

6] (Figure 10(b),(c)) synthesize low-resolution(128×128) talking face videos. The visual quality gap is obvious. Prajwal et al.[29](Figure 10(d)) has ability to synthesize videos with 512×512 resolution, but they just focus on repairing the mouth region. The eyebrow and head pose keep static when just given one reference image. Zhou et al.[49] is able to synthesize 256×256 resolution videos (in Figure10(e)). However, they fail to generate 512×512 videos. We try to directly test their model on 512×512 reference image (shown in Figure10(f)) or reproduce the model train/test on HDTF dataset (shown in Figure10(g)), but still synthesize blurry results. The reason is carefully discussed in section 7. Compared with previous works, our method synthesizes higher visual quality results.

We also carry out quantitative comparisons with state-of-the-art works[6, 49, 29] to evaluate the accuracy of lip synchronization. The experiments are conducted on HDTF dataset with the metric of audio-visual synchronization[10][†]. Table 2 illustrate the experimental results. Our method synthesizes competitive synchronous lip compared with previous works.

[†]https://github.com/joonson/syncnet_python

Table 3. Quantitative evaluation of audio-to-animation module.

	MSE(mouth)↓	LMD^{3D}(mouth)↓	SSIM(eyebrow)↑	CCA(head pose)↑
Taylor et al.[37]	0.1237	0.2355	-	-
Cudeiro et al.[12]	0.1235	0.2350	-	-
Karras et al.[22]	0.1251	0.2365	0.0801	-
Sadoughi et al.[30]	0.1347	0.2470	0.0372	-
Sadoughi et al.[31]	-	-	-	0.7615
Ours (w/o style)	0.1153	0.2308	0.0747	0.7609
Ours (w/o multi-task)	0.0912	0.1922	0.0978	0.7779
Ours	**0.0875**	**0.1899**	**0.1023**	**0.7860**

Table 4. Quantitative evaluation of animation-to-video module.

	PSNR↑	SSIM↑	CPBD↑
Zhou et al.[49]	23.2454	0.8020	0.1226
Zhou et al.[49](interpolate to 512)	23.3482	0.8128	0.0936
Zhou et al.[49](add layer)	22.8777	0.7995	0.1112
Zhou et al.[49](dense)	24.1604	0.8102	0.1273
Zhou et al.[49](dense & add layer)	23.7314	0.8045	0.1209
Siarohin et al.[34, 33]	23.4079	0.8167	0.1345
Siarohin et al.[34, 33] (add layer)	23.1355	0.8062	0.1204
Ours w/o F^{app}	23.9650	0.8220	0.1399
Ours w/o matting	24.3691	0.8384	0.1500
Ours	**24.4174**	**0.8400**	**0.1530**

5.3. Evaluation of Submodules

To validate our audio-to-animation module, we reproduce state-of-the-art animation generation works[22, 37, 30, 12, 31]. For fair comparison, we keep the input and structure setting of their model unchanged and synthesize our animation parameters. To evaluate p^{mou}, we measure MSE on mouth parameters and compute lips landmark distance (LMD^{3D}) on 3D facial mesh. LMD^{3D} has benefits of handling the variance of head posture. In the evaluation of p^{ebro} and p^{hed}, we employ SSIM and Canonical Correlation Analysis(CCA)[40] as metrics respectively. Quantitative results are shown in Table 3. Our method performs better than the above works.

To validate the superiority of our animation-to-video module on high-resolution one-shot talking video generation. We reproduce previous landmark based works[33, 34, 49] on our HDTF dataset, and do quantitative and qualitative comparisons with them. In [49], all setting is as same as original paper. In [34, 33] we replace key points with facial landmarks and ignore the affine transformation in their work. However, considering that above frameworks are designed for 256×256 resolution videos, to make the experiments more convincing, we also conduct extra experiments that make their framework easy to handle 512×512 videos. We add one extra convolutional layer with stride=2 before their network to downsample the input image to 256×256. Figure 11(a-d,g) illustrate the qualitative results. Our approach synthesizes frames with higher visual quality. Table 4 also show the quantitative compared results. PSNR, SSIM and CPBD[27] are utilized as metrics to measure the visual quality. Our approach also acquires the best results.

5.4. Ablation Experiments

Ablation Experiments are conducted to evaluate each component in two sub-modules. In audio-to-animation module, we set two conditions: (1) removing the style-

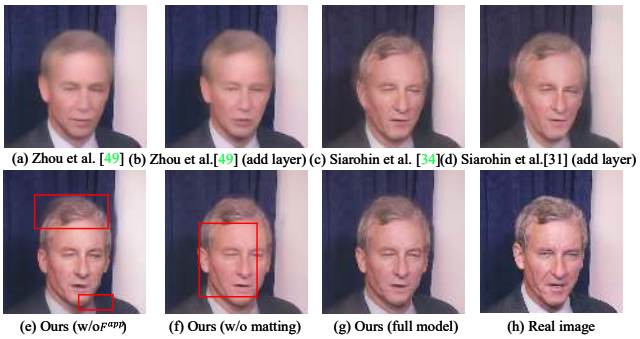


Figure 11. Qualitative results of animation-to-video module.

Table 5. The results of user study.

Method	Chen et al.[6]	Prajwal et al.[29]	Zhou et al.[49]	**Ours**
Mean	2.96	2.88	3.12	3.60
Std	0.95	1.03	0.95	0.74

specific operation (w/o style), i.e., delete the transformation from f^{audio} to f^{audio}_{ref}; (2) synthesizing animation parameters separately (w/o multi-task). Table 3 illustrates the results of two conditions. Both style-specific operation and multi-task training strategy are beneficial to animation generation and our full model synthesizes the best animation results. The style-specific operation significantly improves the synthetic animation. It implies that it is important to consider the speaking style of different identities in animation generation.

In animation-to-video module, we also set two conditions: (1) removing F^{app} and generating F from dense flow in inner face with one network (w/o F^{app}). This condition discard the assumption of motion flow out of the face; (2) removing the matting operation (w/o matting), thus lead to equation 9 as

$$\hat{f}^{ref} = F(f^{ref}). \tag{14}$$

Table 4 shows the quantitative results of two conditions. Our full model presents the best results. Figure 11 (e)(f) also illustrate the synthetic results of two conditions. Without F^{app}, the network is possible to generate inaccurate dense motion flow out of facial region, thus leads to blurry results, e.g., the hair region in figure 11(e). Without matting operation, as shown in figure 11(f), the facial region lose some texture details. This indicates that the matting operation is capability to refine the foreground.

5.5. User Study

An online user study is also conducted to validate our proposed approach. We compare our method with previous state-of-the-art one-shot talking face generation works[49, 6, 29]. For fair comparison, 5 reference images are download from internet to obtain $4\times5 = 20$ videos with 5 different driving audio. 25 volunteers are invited to rate the realism of each video between 1(pretty fake)-5(pretty real). Table 5 illustrates the results of user study. Our method achieves the highest scores and lowest standard deviation.

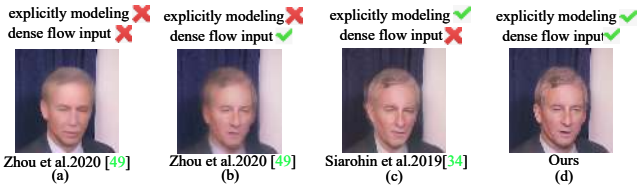


Figure 12. Results of control variate experiments.

6. Limitations

Our work has many limitations. In the generation of F^{app}, the cropped region is very coarse. The synthetic videos largely depend on the M^f. Inaccurate M^f causes the failure results. Our method does not consider the temporally coherent, so if given one reference image with mouth close, the generated face may has flicker tooth. The style-specific operation in animation generator is still hard to synthesize the speaking style as same as real value. We only utilize rule-based method to generate the eye blink movements. The head pose is not extreme enough.

7. Discussion and Conclusion

Discussion. We utilize control variate method to exploring the reason that our flow-guided animation-to-video module performs better than previous landmark-to-video module[33, 34, 49] on high-resolution video generation. We set two conditions: (1)whether to carefully design the network to explicitly model the process of image synthesis; (2) whether to take dense flow as the network input. Figure12 shows the experimental results. Compared with Figure12 (a) and (c), both [49] and [34] take facial landmark as input, [34] utilize explicitly modeling in their network to synthesize more realistic results. Compared with Figure12 (a) and (b), we just replace the landmark input with F^{app} in [49], and generate more realistic frames, especially the richer texture and accurate expression in inner face. Compared with Figure12 (c) and (d), fixing the explicitly modeling in network, our method takes dense flow as input and also generate facial image with richer wrinkles. The experiments indicate that both two conditions are beneficial to improve the visual quality of synthetic videos. Table 3 also illustrate the quantitative results with consistent conclusion. Our framework consists of above two conditions, so the results are more realistic.

Conclusion. In this paper, we build a large in-the-wild high-resolution audio-visual dataset, named HDTF dataset, with higher resolution than previous in-the-wild datasets and more subjects/sentences than in-the-lab datasets. We also propose a novel flow-guided framework, including one style-specific animation generator and one careful-designed flow-guided video generator, to synthesize high visual quality videos. Our method outperforms the state-of-the-art works in high-resolution talking face generation. In the future, we will make great efforts to solve above limitations.

References

[1] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254, 2018. 2, 3

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 4

[3] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020. 1

[4] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *arXiv preprint arXiv:2007.08547*, 2020. 1, 2, 3, 6, 7

[5] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 1, 2

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

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 1, 2

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1, 2

[9] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 1, 2

[10] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 7

[11] 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. 2

[12] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 3, 7

[13] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, pages 408–424. Springer, 2020. 1, 2

[14] Yu Ding, Catherine Pelachaud, and Thierry Artières. Modeling multimodal behaviors from speech prosody. In *Intelligent Virtual Agents*, pages 217–228. Springer Berlin Heidelberg, 2013. 3

[15] Yu Ding, Mathieu Radenen, Thierry Artières, and Catherine Pelachaud. Speech-driven eyebrow motion synthesis with contextual markovian models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3756–3760, 2013. 3

[16] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[18] Kuangxiao Gu, Yuqian Zhou, and Thomas S Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *AAAI*, pages 10861–10868, 2020. 1

[19] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 4

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[22] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017. 3, 7

[23] Lincheng Li, suzhen Wang, zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 3

[24] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 2

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 4

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 1, 2

[27] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. 7

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 5

[29] 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. 1, 2, 6, 7, 8

[30] Najmeh Sadoughi and Carlos Busso. Expressive speech-driven lip movements with multitask learning. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 409–415. IEEE, 2018. 3, 7

[31] Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6169–6173. IEEE, 2018. 3, 7

[32] Najmeh Sadoughi and Carlos Busso. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing*, 2019. 3

[33] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 5, 7, 8

[34] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, pages 7137–7147, 2019. 2, 5, 7, 8

[35] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 3

[36] 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. 3

[37] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93, 2017. 3, 7

[38] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284, 2012. 3

[39] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020. 3

[40] Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005. 7

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[42] 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, 6

[43] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 1, 2

[44] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. *Advances in Neural Information Processing Systems*, 32:5013–5024, 2019. 5

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 5

[46] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 6

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

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

[49] Yang Zhou, DIngzeyu Li, Xintong Han, Evangelos Kalogerakis, Eli Shechtman, and Jose Echevarria. Makeittalk: Speaker-aware talking head animation. *arXiv preprint arXiv:2004.12992*, 2020. 1, 2, 3, 4, 6, 7, 8

[50] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 2, 4

[51] Hao Zhu, Aihua Zheng, Huaibo Huang, and Ran He. High-resolution talking face generation via mutual information approximation. *arXiv preprint arXiv:1812.06589*, 2018. 1, 2