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.

While many works[7, 6, 18, 13, 51, 5, 48, 29, 42, 49, 44] 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 Figure 10(e) for example), however, directly employing their model on 512×512 image will get blurry results (see Figure 10(f) for example). Several factors result in this challenge.

The first reason is that there are no appropriate datasets for high-resolution talking face videos. 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-
Table 1. Statistics of current common audio-visual datasets.

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Environment</th>
<th>Year</th>
<th>Resolution</th>
<th>Subject</th>
<th>Hours</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRW [9]</td>
<td>Wild</td>
<td>2016</td>
<td>∼1080P</td>
<td>1k+</td>
<td>173</td>
<td>1k</td>
</tr>
<tr>
<td>Voxceleb1 [26]</td>
<td>Wild</td>
<td>2017</td>
<td>∼720P</td>
<td>1251</td>
<td>352</td>
<td>100k</td>
</tr>
<tr>
<td>GRID [1]</td>
<td>Lab</td>
<td>2006</td>
<td>∼720×576</td>
<td>34</td>
<td>27.5</td>
<td>51</td>
</tr>
<tr>
<td>RAVDESS [21]</td>
<td>Lab</td>
<td>2018</td>
<td>∼1280×1024</td>
<td>24</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>MEAD [43]</td>
<td>Lab</td>
<td>2020</td>
<td>∼1920×1080</td>
<td>60</td>
<td>40</td>
<td>129k</td>
</tr>
<tr>
<td>Our HDTF</td>
<td>Wild</td>
<td>2022</td>
<td>∼720P−1080P</td>
<td>300k</td>
<td>13k</td>
<td>15k</td>
</tr>
</tbody>
</table>

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

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 $128 \times 128$ 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 leading 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 becomes too complex to handle on high-resolution videos, e.g., [49] synthesize blurry results on $512 \times 512$ resolution (see Figure 10(g) and Figure 11(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 Figure 11(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 $720P \sim 1080P$ videos. There are over 300 subjects and 10k different sentences in HDTF dataset. Our HDTF dataset has higher video resolution than previous in-the-lab 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
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 $512 \times 512$ (the second row in Figure 3). 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

$$M(c^e, c^c) = M_0 + \sum_{i=1}^{60} c^e_i \cdot V^e_i + \sum_{j=1}^{33} c^c_j \cdot V^c_j \text{ (1)}$$

where $M(c^e, c^c)$ represent the 3D facial mesh point. $M_0$ is the average facial mesh. $\{V^e_i\}_{i=1}^{60}$ and $\{V^c_j\}_{j=1}^{33}$ are the linear basis of facial shape and facial expression. $c^e$ and $c^c$ represent the coefficient of the basis. $\{V^e_i\}_{i=1}^{60}$ 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

$$\arg \min_{c^e, c^c, s, R, t} E(c^e, c^c, s, R, t)$$

$$= \arg \min_{c^e, c^c, s, R, t} \sum_{k=1}^{K} \delta_k |p_k - (s [1 0 0] R M(c^e, c^c)^{(k)} + t)|^2 \text{ (2)}$$

where $p_k$ is the $k_{th}$ landmark point and $\delta_k$ represents 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 reconstruction, in each video, we extract
4. Proposed Method

Based on our HDTF dataset, as shown in Figure 2, 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 transferred to approximate dense flow \( F^{app} \) by the 3DMM. Then, \( F^{app} \) and reference image are input into a carefully-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 Figure 4), the style-specific audio representation \( f^{audio} \) is computed from \( I^{ref} \) and \( f^{audio} \). \( f^{audio} \) encodes the speaking content of \( f^{audio} \) and the speaking style of \( I^{ref} \). In the second step (orange part in Figure 4), \( f^{audio} \) 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 \( f^{audio} \) from \( f^{audio} \). Then, considering that different identity has different speaking style[50], AdaIN[20] operation is taken to transform \( f^{audio} \) into the \( f^{ref} \) in the 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.

Furthermore, in the second step, with three branches of the decoder, mouth, eyebrow and head pose are generated simultaneously according to \( f^{audio} \). In mouth branch, a CNN-based mouth decoder is employed to decode \( f^{audio} \) 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

\[
L_{mou}^{mou} = \frac{1}{T} \sum_{t=1}^{T} ||p_{t}^{mou} - \hat{p}_{t}^{mou}||_1, \tag{3}
\]

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

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

where \( \mu_i / \hat{\mu}_i \) and \( \sigma_i / \hat{\sigma}_i \) are the mean and standard deviation of the \( i \)-th dimension of \( \hat{p}^{ebro} \), \( \mu_i \) and \( \sigma_i \) are the mean and standard deviation of the \( i \)-th dimension of \( \hat{p}^{ebro} \). The covariances, \( \delta_1 \) and \( \delta_2 \) 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.
dimension. The final objective function is written as
\[
L(G_{anc}) = L^{mou}_{GAN} + L^{ebro}_{GAN} + L^{hed}_{GAN} + \lambda_{mou} L^{mou}_{1} + \lambda_{ebro} L^{ebro}_{ssim} + \lambda_{hed} L^{hed}_{ssim},
\]
\[ (7) \]
\( \lambda_{mou}, \lambda_{ebro} \) and \( \lambda_{hed} \) represent the loss weights. All the GAN structures are conditional GAN, i.e., \( D^{mou/ebro/hed} \) takes \{audio, 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

**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
\[
F = F^{app} * M^f
\]
\[ (8) \]

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

\( f^{ref} \) represents the encoded feature map of \( I^{ref} \) by a CNN-based encoder. \( F(\cdot) \) 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
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

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 $F_{app}$, $F_{warp}$, $M^{i}$, $F$, $M^{m}$ and synthetic frame, of the animation-to-video module in Figure 9. The $M^{i}$ 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 $M^{i}$ 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,
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 $p_{\text{audio}}$, we measure MSE on mouth parameters and compute lip visual quality results. To validate our audio-to-animation module, 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 resolution. 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.

### 5.3. Evaluation of Submodules

5.3.1. Evaluation of audio-to-animation parameters

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](https://github.com/joonson/syncnet_python)
specific operation (w/o style), i.e., delete the transformation from \( f_{\text{audio}} \) to \( f_{\text{ref}} \); (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 \( F_{\text{app}} \) and generating \( F \) from dense flow in inner face with one network (w/o \( F_{\text{app}} \)). 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

\[
\hat{f}_{\text{ref}} = F(f_{\text{ref}}).
\]

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 \( F_{\text{app}} \), 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\cite{Zhou2020, Siarohin2019, Chen2016}. For fair comparison, 5 reference images are downloaded from internet to obtain 4 \times 5 = 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.

### 6. Limitations

Our work has many limitations. In the generation of \( F_{\text{app}} \), 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\cite{Siarohin2019, Zhou2020, Chen2016} 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. Figure 12 shows the experimental results. Compared with Figure 12 (a) and (c), both \cite{Zhou2020} and \cite{Siarohin2019} take facial landmark as input, \cite{Siarohin2019} utilize explicitly modeling in their network to synthesize more realistic results. Compared with Figure 12 (a) and (b), we just replace the landmark input with \( F_{\text{app}} \) in \cite{Zhou2020}, and generate more realistic frames, especially the richer texture and accurate expression in inner face. Compared with Figure 12 (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


