Generating Smooth and Facial-Details-Enhanced Talking Head Video: A Perspective of Pre and Post Processes

Tian Lv
lt22@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Yu-Hui Wen*
wenyh1616@tsinghua.edu.cn
Tsinghua University
Beijing, China

Zhiyao Sun
sunzy21@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Zipeng Ye
yezp17@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Yong-Jin Liu*
liuyongjin@tsinghua.edu.cn
Tsinghua University
Beijing, China

ABSTRACT

Talking head video generation has received increasing attention recently. So far the quality (especially the facial details) of the videos from state-of-the-art deep learning methods is limited by either the quality of training data or the performance of generators, and needs to be further improved. In this paper, we propose a data pre- and post- processing strategy based on a key observation: generating a talking head video from a multi-modal input is a challenging problem and generating a smooth video with fine facial details makes the problem even harder. Then we propose to decompose the problem solution into a main deep model, a pre- and a post-processing. The main deep model generates a reasonably good talking face video, with the aid of a pre-process, which also contributes to a post-process for restoring smooth and fine facial details in the final video. In particular, our main deep model reconstructs a 3D face from an input reference frame, and then uses an AudioNet to generate a sequence of facial expression coefficients with an input audio clip. To ensure final facial details in the generated video, we sample the original texture from the reference frame in the pre-process with the aid of reconstructed 3D face and a predefined UV map. Accordingly, in the post-process, we smooth the expression coefficients of adjacent frames to alleviate jitters and apply a pre-trained face restoration module to recover the fine facial details. Experimental results and ablation study show the advantage of our proposed method.

CCS CONCEPTS

• **Computing methodologies** → **Appearance and texture representations**.

KEYWORDS

talking head generation, facial details, data pre- and post- process

*Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3551583

ACM Reference Format:
Tian Lv, Yu-Hui Wen*, Zhiyao Sun, Zipeng Ye, and Yong-Jin Liu*. 2022. Generating Smooth and Facial-Details-Enhanced Talking Head Video: A Perspective of Pre and Post Processes. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3503161.3551583

1 INTRODUCTION

In this paper, we study the audio-driven talking head video generation problem, which aims to generate a realistic talking video of a target person by inputting a reference frame and an audio clip. Many methods have been proposed to address this problem. E.g., some researchers directly edit the images to obtain the talking video [4, 16, 19, 28], while some others utilize 2D facial landmarks or 3D human face as intermediate representations to enhance the quality of the generated talking videos [2, 10, 20, 22, 23, 25, 26, 30]. However, the quality of fine facial details in videos generated by state-of-the-art deep learning methods is still limited by either the quality of training data or the generator performance, which needs to be further improved.

To address this problem, in this paper, we propose a simple yet effective method, based on an observation that generating a talking head video from a multi-modal input is a challenging problem and generating smooth videos with fine facial details makes the problem even harder. Our method adopts a simple data pre- and post- processing strategy to decompose this difficult problem into several simple sub-problems and solve them one by one. In more details, our method uses a simple and lightweight deep model (hereafter referred to as the main deep model) to generate a reasonably good talking face video, in which the fine facial details need to be further improved. Then with the aid of a pre-process, we apply a post-process to restore smooth and fine facial details in the final video.

We design the pre-processing to extract textures from the reference frame of the target person (instead of generating textures of the target person using a generator as did in many previous works [10, 20, 26]). For the main deep model, we use a simple audio network AudioNet[1] to generate a talking face video through manipulating the shape of 3D human face reconstructed from the input reference frame. We tolerate the imperfect results generated

[1]Our pre- and post- processing is general and can be applied to other lightweight deep model as well.

by the main deep model, i.e., the results may be jitter and miss ne facial details. Then we apply a post-process focusing on inter-frame smoothing and face restoration [21].

In the Talking Head Generation Track of ViCo Conversational Head Generation Challenge 2022 using the dataset and baseline [29], our method achieved the best results on the metric of cumulative probability of blur detection (CPBD) [14] and won the third place in the competition.

2 RELATED WORKS

2.1 Audio-Driven Talking Head Generation

Recently, many deep neural network methods have been proposed for the audio-driven talking head video generation task, which takes an audio clip and a reference face image as input and outputs a realistic talking video. Some early methods directly modify/edit the reference frame at the pixel level to generate talking videos [4, 16, 19, 24, 28]. Some other methods extract 2D facial landmarks from the reference frame and use the input audio clip to drive these landmarks [2, 30]. Recently, many researches reveal that 3D face models can better serve as a bridge to represent the variations during talking [10, 20, 22, 23, 26]. In this paper, we adopt the 3D face as our intermediate facial representation.

2.2 3D Face Reconstruction

Reconstructing the 3D shape of the human face from a 2D image or multi-view images is a challenging problem. The pioneering work [1] proposes the widely used 3DMM model, which expresses a 3D human face as a weighted sum of the principle components of template faces and mean face; accordingly the reconstruction task is converted into prediction of the weights. Another milestone is the BFM dataset [15], which is made publicly available in 2009. After that, based on the 3DMM model, many neural network methods have been proposed to regress the coe cients of identity, expression and pose, etc. [5, 7, 9]. In this paper, we use the methods in [5] and [27] to reconstruct a high accurate 3D face from a reference frame.

2.3 Face Restoration

Face restoration is a task to recover high-quality human faces with details and enhance colors from low-quality images with degradation (e.g., noise and blur). When the degradation model is unknown, such as low-quality in-the-wild images, the face restoration will be very challenging. To cope with this problem, previous researchers usually use priors to guide the restoration of facial details and textures, including geometry facial priors [3, 18], reference priors [12, 13] and generative priors [8, 21]. In this paper, we adopt GFP-GAN model [21] as our face restoration module, due to its outstanding performance in face restoration.

3 METHOD

3.1 Overview

The pipeline of our method is illustrated in Fig. 1. Firstly, the 3D reconstruction is done on the reference frame to get the BFM coe - cients [15] of the target person. Given an audio clip, our AudioNet transforms it into a sequence of expression coe cients. Meanwhile, our texture sampler generates high delity texture from the reference image based on the 3D face. Before rendering texture map and 3D face to videos, our smoothing operator applies a window to adjacent frames and smooth the BFM coe cients to avoid jitters. Since BFM does not model the details in the mouth region, the texture map lacks information in the mouth region. To compensate for this, we add a StyleGAN-based submodule [11, 17] to complete the mouth region of the rendering result to make the generated videos more natural. Finally, the intermediate video is fed into a face restoration module [21] to enhance the ne facial details, to get the output of our entire pipeline.

3.2 Texture Sampler As Pre-Processing

Compared with generating texture maps or predicting BFM texture coe cients, we consider that sampling textures from original reference image can capture high-frequency information in the texture, which is bene cial to obtain high- delity rendering result. Thanks to 3D face reconstruction, the 3D coordinates of human face's vertices in the world coordinate system can be transformed into the image space, where each vertex nds its color from the nearest pixel in the reference image. In order to generate a complete texture map, we use a prede ned UV map which records the UV coordinates of all vertices, to paint the sampled color on corresponding position in the texture map. For the unpainted pixels in the texture map, we apply barycentric interpolation with the information of triangle faces from reconstructed mesh to ll the pixels' color. An example of sampled texture is shown in the texture sampler part in Fig. 1.

3.3 Main Deep Model

3.3.1 AudioNet. In our method, the main deep model is the AudioNet, which is a neural network model that generates expression coe cients to control the variation of human face. The input of AudioNet is a 16×80 MFCC feature maps of a frame which are transformed from input audio clip and outputs the transformed expression coe cient $V \in R^{64}$ at the corresponding frame. At the inference stage, the generated expression coe cients are used to substitute the coe cients of the reconstructed result from reference image to form the talking face shape. We use two loss terms to supervise the training process, which are L1 loss l_1 and delta loss l_3:

$$l_1 = \sum_{i=1}^{64} \| V_i - \hat{V}_i \|_1 \tag{1}$$

$$l_3 = \sum_{i=1}^{63} \| (V_{i+1} - V_i) - (\hat{V}_{i+1} - \hat{V}_i) \|_1 \tag{2}$$

where V_i, $\hat{V}_i$ are the i-th element in the predicted and true expression coe cients, respectively. l_1 is designed to guide the regression, and l_3 is to ensure that AudioNet learns more accurate distribution of expression coe cient. The overall loss is expressed as:

$$l = \lambda_1 l_1 + \lambda_3 l_3 \tag{3}$$

3.3.2 Mouth Completion. After rendering texture map and 3D face, we obtain a talking video without the mouth region because the 3DMM model is not capable of modeling mouth and teeth. In order to eliminate this unnatural artifact, we need to ll the mouth

Figure 1: The pipeline of our method. The preliminary BFM coe cients are generated from 3D reconstruction of input frame, which is used to obtain texture map. The expression part of BFM coe cients is then substituted by the output of AudioNet, which is then smoothed. With texture map and BFM coe cients, the intermediate video is generated, which is followed by the mouth completion and face restoration module that signi cantly re ne the talking video.

region with reasonable content. To solve this problem, we introduce StyleGAN [11] and StyleGAN encoder pSp [17]. The training pairs for mouth completion module are obtained by cropping mouth region using mouth landmarks. We mask out the mouth region in cropped images to get source data, and regard the unmasked counterparts as target data.

3.4 Post-processing

3.4.1 Smoothing Operator. Our method does not explicitly guarantee the inter-frame continuity, which may result in jitters in generated videos. We notice that these jitters mainly take place in the facial region, therefore we introduce smoothing operator to alleviate jitters of human face by smoothing the expression coe cients (mouth completion is also applied after smoothing operation). Given a target frame $?_8$ and its corresponding expression coe cients V_8, we de ne the one-side window size ω as a hyperparameter which indicates the smoothing range, therefore the number of involved frames is $2\omega + 1$. We calculate the weights of each frame in the smoothing window as:

$$F_? = \frac{1}{2}\left(1 + \cos\left(\frac{\pi(? - (8 - \omega))}{\omega + 1}\right)\right), \quad 8 - \omega \le ? \le 8 + \omega, \tag{4}$$

where $?$ is the index of frame. Based on these weights, we normalize them and calculate the smoothed expression coe cient of frame 8 by the weighted sum of the expression coe cients in the smoothing window:

$$V'_8 = \sum_{?=8-\omega}^{8+\omega} \left(\frac{F_?}{\sum_{9=8-\omega}^{8+\omega} F_9}\right) V_? \tag{5}$$

For the boundary part of the video (i.e. the beginning and ending parts of video), we adaptively scale the size of smoothing window. For example, for the first frame, its ω is 0 because there is no frame before it.

3.4.2 Face Restoration Module. We observe that if the face regions in the training data are of low-resolution and blur, the generated talking face videos tend to suffer from vagueness and degradation. To remedy this limitation, we propose to use a face restoration module in the post process to restore face details inspired by GFP-GAN [21]. Specifically, our face restoration module uses a pretrained GAN to provide facial priors to restore diverse and rich facial details. In more details, the facial priors are incorporated on multi-resolution spatial features to produce final results of high fidelity.

4 EXPERIMENT

4.1 Implementation Details

3D Reconstruction. We adopt Deep Face Reconstruction [5] and WM3DR pretrained models [27] to reconstruct 3D faces in our method.

AudioNet. Our AudioNet is trained on the training split of ViCo dataset's talking head generation part [29]. We extract audio clips and frames from source videos, and get the BFM coe cients as well as MFCC features of each frame, which are used to train AudioNet. Our AudioNet is trained for 7,000 epochs on a single NVIDIA TITAN RTX with loss weights $\lambda_1 = \lambda_3 = 1$.

Mouth Completion Module. Our mouth completion network is trained on the processed training pairs from ViCo training set for 100,000 iterations, and other experimental settings are the same as the official implementation in StyleGAN2.

Face Restoration Module. We use the GFP-GAN pretrained model [21] for face restoration.

4.2 Qualitative Result

Based on the experiment setting described in Section 4.1, our model can generate smooth realistic talking videos with fine facial details.

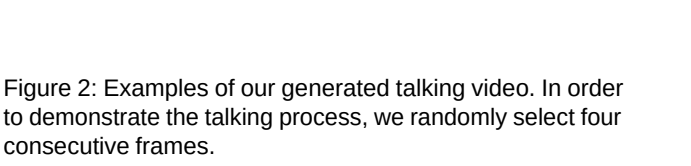

Figure 2: Examples of our generated talking video. In order to demonstrate the talking process, we randomly select four consecutive frames.

We randomly choose three identities to show some of our results in Fig 2. It is clear that the quality of our talking videos is not limited by the low-quality of the input frame, while preserving the identity and lip variations of the talking persons.

4.3 Quantitative Result

As the ground truth of test data is not available, we apply the evaluation results from leaderboard of ViCo Challenge from https://vico.solutions/ for comparison. According to the evaluations, our method is ranked No. 1 at the CPBD metric, showing that our method generates videos with best clarity. However, our method does not achieve the best results at PSNR, SSIM, and FID metrics. The reason is that we use a face restoration module that is pretrained on public high-quality face images to enhance the facial details, and thus the module increases the di erence between input frame and generated result. Furthermore, we use a simple, lightweight audio-driven module, so the performance on the lip synchronization metrics is not the best. Anyway, we regard that our post-process can help generate more realistic talking videos with ne facial details if they are appropriately combined with a more comprehensive audio-driven model as the main deep model.

4.4 Ablation Study

In order to show that our post-process helps to generate better talking head video, we remove the face restoration and smoothing modules respectively. As shown in Fig 3, our method generates low-quality talking videos after removing the face restoration module. Meanwhile, we observe that the generated videos su er from severe jitters if we remove the smoothing operator.

4.5 Limitations and Discussions

Since we design a lightweight simple network as the main deep model, our method does not achieve the best performance under some evaluation metrics, such as lip synchronization. We believe these limitations can be alleviated if a better and more comprehensive main deep model is introduced. E.g., if we use the Wav2Lip method [16], which is a talking head generation method that directly manipulate the frames on the pixel-level (we use a pretrained model from its o cial implementation), the generated results are shown in Fig 4, indicating that the lip synchronization and identity is maintained after face restoration, and the image quality is signi cantly improved. These results show that our pre- and post-processes can work with di erent main deep models and generate improved results.

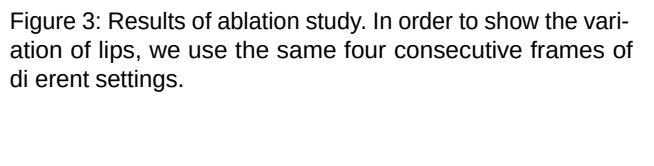

Figure 3: Results of ablation study. In order to show the variation of lips, we use the same four consecutive frames of di erent settings.

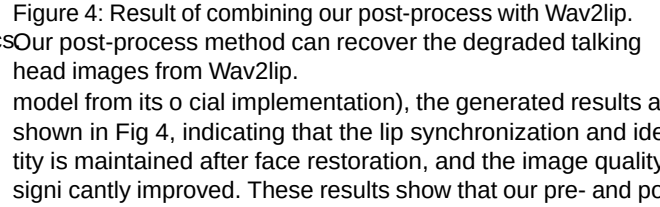

Figure 4: Result of combining our post-process with Wav2lip. Our post-process method can recover the degraded talking head images from Wav2lip.

5 CONCLUSION

In this paper, we propose to generate smooth and facial-details-enhanced talking head video with data pre- and post-process techniques. Experimental results show that these techniques can help generate more smooth talking head videos with ne facial details. Meanwhile, if we use the state-of-the-art Wav2Lip as the main deep model and combine it with our data process techniques, the good results demonstrate that our proposed method has the potential to improve the generating results of existing talking video generation methods. We believe that data processing methods are essential components in talking video generation, and by reasonably adopting these methods, the existing talking head generation models can be optimized, so as to obtain the videos of higher quality.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of China (61725204) and Tsinghua University Initiative Scienti c Research Program, China Postdoctoral Science Foundation (2021M701891).