Audio-Driven Emotion-Aware 3D Talking Face Generation from Single Image

Chun-Shuo Qiu[a,b], Feng-Lin Liu[a,b], Hongbo Fu[c], Fan Zhang[d], Yan-Pei Cao[e], Yu-Kun Lai[f], Lin Gao[a,b]✉
[a] *Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[b] *University of Chinese Academy of Sciences, Beijing, China*
[c] *Hong Kong University of Science and Technologyy, Hong Kong, China*
[d] *Shandong Technology and Business University, Shandong, China* [e] *VAST, Beijing, China*
[f] *School of Computer Science and Informatics, Cardiff University, UK*
{qiuchunshuo22s, liufenglin21s}@ict.ac.cn, fuplus@gmail.com, zhangfan51@sina.com, caoyanpei@gmail.com, LaiY4@cardiff.ac.uk, gaolin@ict.ac.cn

***Abstract*—Audio-driven talking face generation from a single source image is a popular research topic. There still exist many challenges for its practical applications, e.g., diverse motion generation, effective emotional control, and large view angle changes. In this work, we propose a novel one-shot emotion-controllable audio-driven 3D talking face generation framework, which creates free-view talking videos from one reference image. Firstly, to synchronize the motion with the input audio, we use a transformer-based motion generator to capture the context of the input audio and predict motion coefficient sequences, which are leveraged by a motion encoder to extract motion codes. Meanwhile, to reconstruct a 3D portrait from one reference image, an identity encoder is utilized to extract an identity code and generate emotion-dependent appearance with a specific emotion label. Finally, we introduce an emotion-controllable 3D portrait video generator to synthesize free-view talking videos using the disentangled motion and identity codes. Thanks to the audio-synchronized motion codes and emotion-aware identity code, we can render a talking face with realistic emotional expressions in novel views. Extensive experiments show that our method is capable of maintaining superior visual performance and motion accuracy in both front view and novel views.**

***Index Terms*—3D Face Reconstruction, Face Animation, Expression Editing, Neural Radiance Fields**

I. Introduction

Recently, talking head generation from a static portrait image has spawned abundant applications in digital human animation, visual dubbing, short video creation, etc. Creating a realistic talking face video from a single image and audio input is challenging due to the intricate relationship between audio and lip movements. Furthermore, a single image lacks sufficient information to support vivid head movements, especially when talking head animations involve substantial changes in the head pose. Moreover, the creation of a lifelike talking head requires an abundance of detailed expressions. To tackle these challenges, it is imperative to take into account a wide range of knowledge, such as vocal-visual multi-modalities, human face models, and emotions.

Previous 2D works [1], [2] mainly focused on syncing lip movements with speech. Certain studies further considered head pose [3] and emotions [4], [5]. However, these studies struggled with the video quality and the expression accuracy amid large changes in the head pose. In contrast, 3D methods handle pose changes better by rendering 3D portraits. As neural rendering has achieved remarkable progress in 3D reconstruction, several works [6], [7] utilized neural radiance fields (NeRFs) [8] to render dynamic faces. Regrettably, these methods either overfit to a single identity or lack stable movements and detailed emotional expressions.

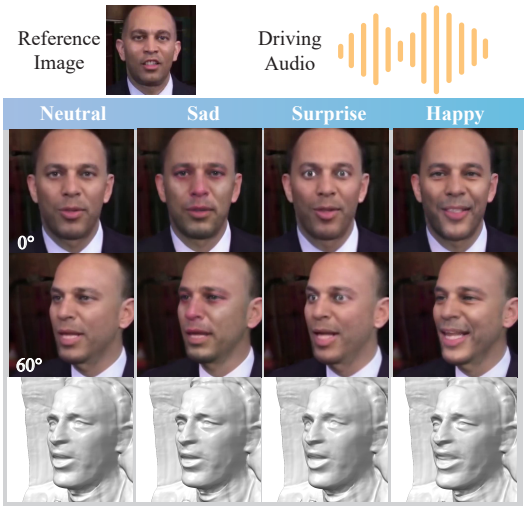

Fig. 1. Given a reference image and a driving audio, our method generates free-view talking videos with vivid emotional expressions.

In this work, we propose a novel one-shot emotion-controllable audio-driven 3D talking face generation framework, which creates an audio-synchronized talking video with diverse emotions in novel views from a single speaker image, as illustrated in Fig. 1. To achieve this goal, we design a 3D portrait video generator with a two-branch encoder, namely a motion encoder and an identity encoder to disentangle the audio-driven motion control and emotion-aware identity reconstruction. Specifically, we first leverage an audio-driven motion generator to synthesize realistic motion sequences. Then, with a reference portrait image, an identity encoder extracts a feature pyramid from the reference image and

fuses each level's features into the identity code. An emotion-aware mapper adapts the identity code using the user-specified emotion label. Finally, with the audio-synthesized motion codes and emotion-aware identity code, the 3D portrait video generator predicts tri-planes [9] as a compact 3D-aware representation and renders free-view 3D portrait videos via neural rendering. As shown by extensive experiments, thanks to the well-designed audio-driven motion generator and 3D portrait video generator with the two-branch encoder, our method is capable of generating audio-synchronized emotion-aware portrait videos with superior visual performance in novel views.

Our contributions are summarized as follows:

- We propose the first method for one-shot (i.e., single reference image-based) emotion-controllable audio-driven 3D talking face generation, which creates audio-synchronized talking videos with diverse emotions in novel views from one target speaker image;
- We present a controllable 3D portrait video generator with a two-branch encoder to disentangle motion and identity control so that the generated videos contain not only temporally consistent dynamics but also emotion-aware appearance.

II. Related Work

A. Audio-Driven Talking Face Generation

A number of investigations have significantly advanced the performance of audio-driven talking head animation. Existing methods can be divided into two principal categories: person-specific and person-agnostic approaches.

Person-specific studies [10]–[12] employed neural networks to establish a correlation between audio input and lip motion, generating higher-quality results by leveraging a target individual's video data. Despite enhanced animation quality, these methods are constrained by the demanding data requirements.

In contrast, person-agnostic strategies aim to generate high-quality talking face videos in a one-shot setting. Initial approaches [1], [2], [13] focused primarily on synchronizing lip movements with the audio content. SadTalker [3] introduces two modules that provide controllable eye blink and stylized head pose effects. And some works [4], [5], [14] embedded emotional information into generated talking faces for vivid video. While these methods are capable of generating face images with novel poses, they lack a 3D-aware representation, leading to noticeable artifacts in the synthesized results, particularly when significant head pose changes are involved.

B. Talking NeRF Face Generation

The advancement of NeRFs has laid the foundation for numerous studies [11], [16], [17] that utilize dynamic NeRFs to simulate the speaking motion of a specific individual in a 3D-aware manner, at the cost of extensive target video data. Following works [18], [19] applied lightweight or generic mechanisms to reduce the video data requirements and the convergence time cost. These studies exhibit exceptional multi-view consistency due to the use of NeRFs. However, the necessity for person-specific data and optimization remains a substantial constraint on practical applications.

Regarding the general animation of arbitrary individuals, several studies [6], [20] use facial blendshape coefficients to control a tri-plane generator explicitly. However, these methods often fail to maintain temporal stability and produce flickering videos due to the lack of consideration for temporal information. OTAvatar [7] attempted to ensure temporal stability by using a window of temporally adjacent motion coefficients as input. Unfortunately, this method struggles to generate intricate expressions, such as eye blinking or frowning. Real3D-Portrait [21] introduced an image-to-plane model with a motion adapter to achieve high-quality reconstruction and reenactment. Our method also incorporates NeRF for audio-driven 3D-aware talking face video generation. Further, it attains emotion control and temporal consistency.

III. Methodology

Now we introduce our novel framework, a 3D portrait video generator with a two-branch encoder, illustrated in Fig. 2, for generating emotion-aware photo-realistic 3D talking faces. The motion encoder branch leverages a pre-trained audio encoder and an audio-synchronized motion generator $\mathcal{G}_m$ to convert the input audio into a sequence of 3DMM expression coefficients, which is encoded as motion code by a Motion Encoder $\mathcal{E}_m$. Meanwhile, to achieve 3D facial animation using only a single image, we design an emotion-aware identity encoder $\mathcal{E}_{id}$, which encodes a reference image into an emotional identity code. Finally, with the input of the identity code and motion code, a controllable 3D portrait video generator $\mathcal{G}$ synthesizes tri-planes [9] as a 3D-aware representation, followed with a volume renderer and a super-resolution module to obtain the video clip in any novel view. Below, we introduce the motion generator and controllable 3D portrait video generator in Sec. III-A and Sec. III-B, respectively. And the training strategy of both modules is described in Sec. III-C.

A. Audio-Synchronized Motion Generation

To produce accurate facial movements based on audio input, we initially employ an audio encoder to extract contextualized audio features from the audio data. Subsequently, the audio features are converted into a motion sequence by a motion generator. Our system employs the expression coefficients of 3D Morphable Models (3DMMs) [22] as an intermediate representation, which provides crucial 3D information to improve the realism of generated 3D portrait videos and ensures identity independence owing to its well-disentangled structure.

1) Audio Encoder: Our audio encoder adopts the architecture of the state-of-the-art pre-trained speech model, wav2vec 2.0 [15], which consists of an audio feature extractor and a multi-layer transformer encoder. With audio input A, the audio feature sequence is extracted as $a_{1:T} = \mathcal{E}_a(A)$.

2) Motion Generator: With audio feature sequence $a_{1:T}$, the motion generator $\mathcal{G}_m$, a stack of several transformer encoder layers, predicts 3DMM motion sequence $\hat{m}_{1:T} = \mathcal{G}_m(a_{1:T})$ to control tri-plane synthesis for final speech video

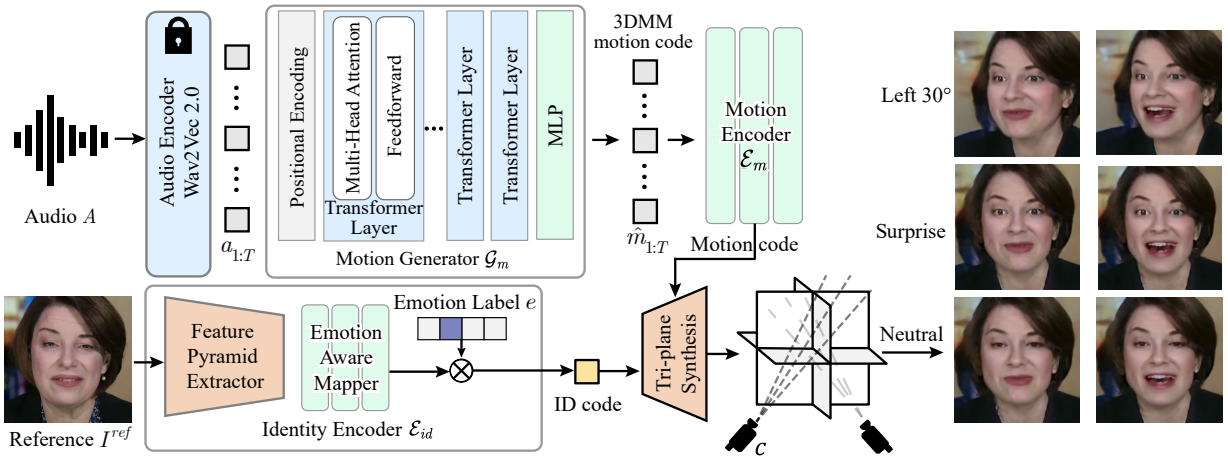

Fig. 2. **The framework of our method**. Our method first encodes audio clips A into contextualized features using Wav2Vec 2.0 [15] and then feeds the resulting features into a transformer-based motion generator $\mathcal{G}_m$ to obtain an audio-synchronized 3DMM coefficient sequence $\hat{m}_{1:T}$. Subsequently, this coefficient sequence is encoded into a motion code z_m sequence by a motion encoder. Concurrently, the feature pyramid extractor processes a reference image I^{ref} to derive an identity code w_{id}. Furthermore, utilizing an emotion label e, the emotion-aware mapper translates the identity code into an emotion-aware identity code. Finally, employing both the motion codes and the identity code, the tri-plane synthesis network generates tri-planes, which are used to obtain a free-view and emotion-manipulated video using a volume renderer and a super-resolution module.

rendering. Considering temporal consistency, the audio features are inputted as time windows, $a'_i = a_{i-\delta_a:i+\delta_a}$ and $\delta_a = 2$ in our experiments.

B. Controllable 3D Portrait Video Generation

After generating the realistic 3D motion coefficients, we animate the 3D portrait constructed from a target image through a well-designed two-branch encoder, motion encoder and identity encoder. Utilizing the tri-plane synthesis of PV3D [23] as backbone, we introduce the Motion Encoder module $\mathcal{E}_m$ to learn the relationship between the explicit 3DMM expression coefficients and the implicit motion code. Meanwhile, we utilize an additional Emotion-Aware Identity Encoder, denoted as $\mathcal{E}_{id}$, to extract the identity code from the reference image. This identity code is then adapted to generate appearance dependent on the emotion. During the inference stage, we utilize several inversion steps to enhance identity faithfulness.

1) Motion Encoder: In contrast to PV3D, which utilizes a latent code and timestamps to represent motion dynamics, we build a motion encoder $\mathcal{E}_m(\cdot)$ to map 3DMM expression coefficients m to the motion code z_m of the generator. Specifically, the motion encoder is built via several 1D convolutional layers followed by a network composed of multiple residual blocks. To ensure temporal consistency, the motion encoder uses the temporal adjacent coefficients from a window δ_m as input, where $\delta_m = 11$.

2) Emotion-Aware Identity Encoder: The emotion-aware identity encoder is composed of a basic identity encoder $E_{id}(\cdot)$ and an emotion-aware mapping network $M_e(\cdot|e)$. Given a single reference image, the basic identity encoder $E_{id}(\cdot)$ uses a feature pyramid network to produce three hierarchical layers of feature maps, from which the identity code is extracted using a simple, intermediate mapping layer. As for emotion-dependent appearance, we use a multi-level mapping network M_e to adjust each channel of the identity code w_{id}, whose outputs are summed with weights assigned corresponding to the emotion label e. We further define $\overline{w}_{id}$ to be the average identity code of the pre-trained generator. Given an input image I and emotion label e, the output of the emotion-aware identity encoder $\mathcal{E}_{id}(\cdot)$ is defined as $\mathcal{E}_{id}(I|e) = M_e(E_{id}(I) + \overline{w}_{id}|e)$.

C. Training Strategy

Due to the high memory usage of video rendering gradients, we avoid an end-to-end training strategy. In our experiments, we train the audio-synchronized motion generator $\mathcal{G}_m$ independently. And the training objectives are introduced in Sec. III-C1. Additionally, as described in Sec. III-C2, we joint train the identity encoder $\mathcal{E}_{id}$, motion encoder $\mathcal{E}_m$, and tri-plane generator $\mathcal{G}$ with super-resolution module $\mathcal{R}$ frozen. As for the emotion-aware mapping module M_e for the identity encoder, we isolate this module by freezing the remainder of the network and conduct training exclusively on it, as elaborated in Sec. III-C3.

Some detailed formulae and diagrams are presented in the supplementary material.

1) Audio-synchronized Motion Generator: We train the motion generator $\mathcal{G}_m$ utilizing a pre-trained wav2vec 2.0 to extract audio features $a_{1:T}$ from audio clip A. Drawing inspiration from SelfTalk [24], we attached a lip-reading module, which translates the lip motions of a 3DMM sequence to a sentence, behind the motion generator, thereby supervising the alignment between audio clips and motion sequences. Initially, we employ pre-trained wav2vec 2.0 to extract latent features $S^{latent}_{1:T}$ and text tokens $S^{text}_{1:T}$ as pseudo-truth values from audio clip A. Concurrently, a sequence of lip vertices is segmented from head meshes, reconstructed from the predicted motion coefficient sequence. The lip-reading module map these lip vertex sequences to latent features $\hat{S}^{latent}_{1:T}$ and extract text sequence $\hat{S}^{text}_{1:T}$. With the pseudo-truth values, a MSE loss $\mathcal{L}_{lat}$ for latent consistency and CTC loss $\mathcal{L}_{ctc}$ for temporal aligment are introduced. Moreover, we use a reconstruction

term $\mathcal{L}_{MSE}$ for motion coefficient directly. During the motion generator training phase, the total loss is

$$\mathcal{L} = \mathcal{L}_{lat} + \lambda_{ctc}\mathcal{L}_{ctc} + \lambda_{MSE}\mathcal{L}_{MSE}, \quad (1)$$

where the weights of loss terms are $\lambda_{ctc} = 0.1$ and $\lambda_{rec} = 10$.

2) Controllable 3D Portrait Video Generator: We use the pre-trained parameters of the PV3D generator for proper initialization and then train the two-branch encoder (Motion Encoder $\mathcal{E}_m$ and Identity Encoder $\mathcal{E}_{id}$) and Tri-plane Synthesis module. Noticeably, the emotion-aware mapping network M_e is trained separately.

To effectively decouple motion and identity, we employed a cross-contrastive training strategy. Given a pair of source and driving image I^s, I^d from a video, we extract their 3DMM coefficients m^s, m^d and motion code $z_m^s = \mathcal{E}_m(m^s), z_m^d = \mathcal{E}_m(m^d)$, respectively. Further, we reconstruct the source image $\hat{I}^s = \mathcal{R}(\mathcal{G}(w_{id}^s, z_m^s), c^s)$ and predict the driving image $\hat{I}^d = \mathcal{R}(\mathcal{G}(w_{id}^s, z_m^d), c^d)$, where the shared identity code $w_{id}^s = \mathcal{E}_{id}(I^s)$ is extracted from the source image I^s.

The reconstruction term $\mathcal{L}_s = \mathcal{L}(I^s, \hat{I}^s)$ and the animation term $\mathcal{L}_d = \mathcal{L}(I^d, \hat{I}^d)$ are utilized to ensure the proper appearance reconstruction and decouple the identity code and motion code, where the total loss $\mathcal{L}(I, \hat{I})$ is a weighted sum of the following loss terms: a perceptual loss $\mathcal{L}_{VGG}$ using VGG-19 network, a MAE of pixel $\mathcal{L}_{pixel}$, a regularization loss $\mathcal{L}_{reg}$ for the motion code and identity code and an identity consistency loss $\mathcal{L}_{ID}$. The total loss function is

$$\mathcal{L}(I, \hat{I}) = \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_{pixel}\mathcal{L}_{pixel} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{ID}\mathcal{L}_{ID}, \quad (2)$$

where the parameter values we use for the examples in this paper are $\lambda_{VGG} = 0.8$, $\lambda_{pixel} = 1$, $\lambda_{reg} = 0.1$, $\lambda_{ID} = 0.2$.

3) Emotion-Aware Mapping Network: For the emotion-aware mapping network M_e, we design a training strategy using CLIP loss term. With randomly sampled identity code w_{id}, we first render the original image I^{origin}. Then, we randomly choose emotion label e and obtain emotion-manipulated image I^e with the adjusted emotion-aware identity code w_{id}^e. Instead of pixel loss $\mathcal{L}_{pixel}$ used in the above training, we use CLIP loss $\mathcal{L}_{CLIP}$ to guide the mapper to minimize the cosine distance between the emotional image and text description in the CLIP latent space. Since the emotion-aware mapping network adjusts the appearance for the given identity, we use the histogram loss $\mathcal{L}_{sim}$ [25] to maintain the coarse style of emotion-manipulated image I^e, which measures the similarity between the histograms representing color distributions. The total loss function is a weighted combination of these losses:

$$\begin{aligned}\mathcal{L}(I^{origin}, I^e) =& \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_{CLIP}\mathcal{L}_{CLIP} + \lambda_{reg}\mathcal{L}_{reg} + \\ & \lambda_{id}\mathcal{L}_{id} + \lambda_{sim}\mathcal{L}_{sim},\end{aligned} \quad (3)$$

where the parameter values are $\lambda_{VGG} = 0.05$, $\lambda_{CLIP} = 1$, $\lambda_{reg} = 0.7$, $\lambda_{ID} = 1$, $\lambda_{sim} = 1$.

IV. Experiments

A. Datasets and Metrics

1) Datasets: To learn a universal audio-driven motion generator, we construct our dataset based on the widely used dataset MEAD [26]. We split the dataset into training and test parts following EAT [5], which utilizes all data of four identities as the test part and the others for training.

2) Evaluation Metrics: We demonstrate the superiority of our method on multiple metrics that have been widely used in previous studies. We adopt Peak-Signal-to-Noise Ratio (**PSNR**), Structural Similarity Index Measure (**SSIM**) , Learned Perceptual Image Patch Similarity (**LPIPS**) and Fréchet Inception Distance score (**FID**) to evaluate the visual quality. To measure the audio-visual synchronization, we use the Landmark Distance around mouths (**M-LMD**) and the confidence score of SyncNet [27] (**Sync**$_{conf}$). To assess the emotional accuracy of the generated emotions, we fine-tune the Emotion-Fan [28] using the training set of MEAD.

B. Results

1) Emotion Manipulation: We display some emotion manipulation results in Fig. 3 in different views. No existing methods can achieve such emotion control for one-shot audio-driven 3D-aware talking face generation. The first column contains the reference images, which are selected from the test set of HDTF [29]. For the remaining columns, there are three images in each cell, corresponding to three different views of each result. As shown in these images, our method is capable of synthesizing emotion details (such as upturned lips, tear-stained eyes and bulging eyes) and maintaining these expressions in novel views.

Fig. 3. **The qualitative results of one-shot emotion manipulation.**

C. Comparisons

1) Quantitative Evaluation: We compare our method with state-of-the-art 2D methods including: PC-AVS [30], EAMM [14], SadTalker [3], StyleTalk [4] and EAT [5]. Meanwhile, we also compare it with 3D animation methods, Next3D [6] and OTAvatar [7]. We conduct the experiments in the self-driven setting on the test set of MEAD, using the same test identities and audio clips as EAT [5]. We employ the first frame of each video clip as the reference image and the corresponding audio clip as the audio input for generating videos from two views (front and left 60 respectively) to verify the effectiveness of our method. Regarding emotion accuracy, we compare our method with EAT [5] and StyleTalk [4].

As presented in Tables I, II and III, our method achieves the best performance in most metrics. Our method achieves the best synchronization and visually much better results for novel views than 2D methods. Since PC-AVS and EAT are trained using SyncNet as a supervision, it is reasonable for these methods to obtain higher confidence score of SyncNet, even though the mouths and faces are blurry in their results. Compared with 3D methods, our method achieves better synchronization, along with better visual quality. As shown in Table IV, our method exhibits highest overall emotion accuracy on the entire test dataset.

TABLE I
QUANTITATIVE COMPARISON WITH 3D BASELINES.

	PSNR/SSIM ↑	LPIPS↓	FID ↓	M-LMD ↓	Sync_{conf}↑
OTAvatar	21.12/0.75	0.33	74.21	0.12	1.93
Next3D	21.18/0.78	0.31	80.57	0.13	3.82
Ours	**23.51/0.80**	**0.29**	**59.06**	**0.09**	**4.61**

TABLE II
QUANTITATIVE COMPARISONS WITH 3D METHODS(LEFT 60 VIEW).

	PSNR/SSIM ↑	LPIPS↓	FID ↓	M-LMD ↓	Sync_{conf}↑
OTAvatar	16.97/0.72	0.40	81.77	0.16	1.54
Next3D	17.03/0.70	**0.39**	73.69	0.17	3.72
Ours	**17.36/0.73**	**0.39**	**69.46**	**0.13**	**4.49**

TABLE III
QUANTITATIVE COMPARISONS WITH 2D METHODS(LEFT 60 VIEW).
*PC-AVS, EAT USE SYNCNET AS SUPERVISION FOR HIGHER SYNC_{conf}.

	PSNR/SSIM ↑	LPIPS↓	FID ↓	M-LMD ↓	Sync_{conf} ↑
PC-AVS*	16.06/0.52	0.43	137.90	0.23	**7.01**
EAMM	14.00/0.54	0.47	107.61	0.21	2.79
SadTalker	15.99/0.57	0.48	115.96	0.18	3.92
StyleTalk	15.59/0.56	0.42	119.87	**0.13**	4.39
EAT*	16.84/0.63	**0.39**	113.83	0.14	6.63
Ours	**17.36/0.73**	**0.39**	**69.46**	**0.13**	4.49

TABLE IV
EMOTION ACCURACY COMPARISONS.

	EAT	StyleTalk	Ours	GT
Avg. Acc.	0.59	0.56	**0.63**	0.87

2) Qualitative Evaluation: We first compare our method with NeRF-based animation methods. The results are displayed in Fig. 4 in three views (front, left 30 and left 60). Our method can reconstruct more vivid and detailed identities and motions, such as blinks. OTAvatar [7] hardly handles the mouth motion and lacks details. Next3D [6] has fine reconstruction per frame but cannot guarantee temporal consistency, leading to flickering results (please refer to the supplementary video).

Fig. 5 compares our method with 2D methods PC-AVS [30], EAT [5], SadTalker [3] and StyleTalk [4], which use warping fields for reenacting faces and pose control. Our method can handle extreme pose change and maintain identity consistency.

D. Ablation Study

1) Emotion-Aware Identity Encoder: Given that the standard loss functions of our dual-branch encoder are adapted from PV3D [23], we embarked on an ablation study focusing

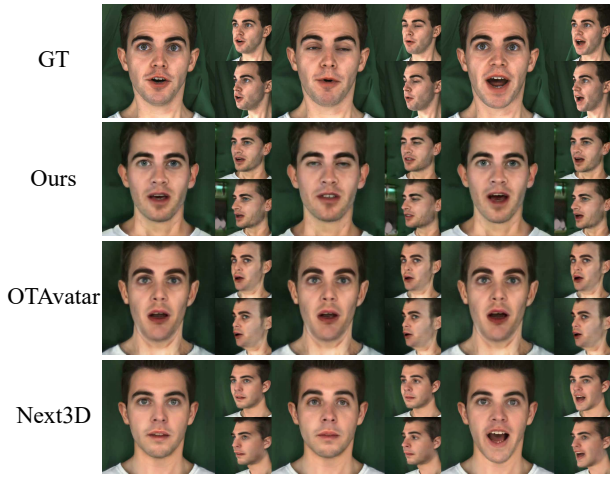

Fig. 4. **The qualitative results of the one-shot self-reenactment cases, compared with 3D methods.** Note that these methods mainly differ in motions and dynamics. For more noticeable differences, please refer to the supplementary videos.

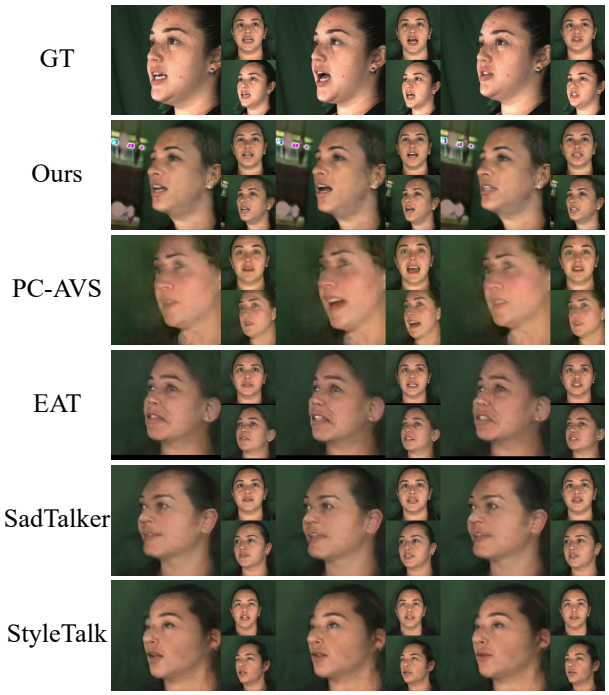

Fig. 5. **The qualitative results of the one-shot talking-head generation cases, compared with 2D methods.** As our method reconstructs the entire scene, artifacts appear in the background at the left-60 view, whereas the region of the human face remains stable.

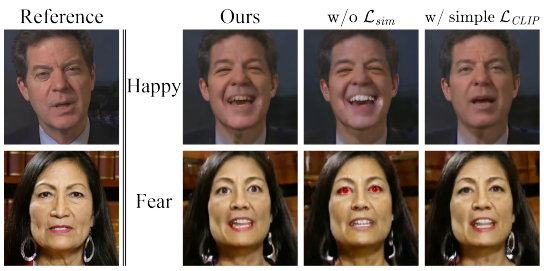

Fig. 6. **The ablation study about the loss terms of the emotion mapper.**

specifically on the contributions of the histogram loss $\mathcal{L}_{sim}$ and the CLIP loss $\mathcal{L}_{CLIP}$. The two configurations considered are: (1) w/o $\mathcal{L}_{sim}$, where the histogram loss is not utilized, and (2) w/ simple $\mathcal{L}_{CLIP}$, where CLIP loss is employed using simple text prompts such as 'realistic face with sad emotion'. Fig. 6 shows that employing the histogram loss markedly improves video quality, whereas using the CLIP loss with detailed prompts significantly enhances the generation of accurate emotional expressions.

V. Conclusion

A. Conclusion

In this paper, we propose a one-shot 3D face reconstruction with audio-driven and emotion-controllable rendering for talking head video generation. With well-designed modules, we can generate audio-synchronized talking videos with diverse emotions in novel views from one target speaker image. Experimental evaluations validate the superior performance of our proposed framework. Given the use of CLIP for emotion manipulation, our methodology presents a potential extension to other applications, such as portrait stylization.

B. Limitations and Future Work

We have demonstrated that our method can generate audio-synchronized, emotion-aware talking videos. However, our technique, in its current form, is unable to generate a complete head structure due to the absence of necessary geometric priors, such as the back of the head. Therefore, with more priors in both geometry and data, the method's usability will be further enhanced. And limited by volume rendering in NeRF, our method cannot synthesize videos in real-time. The integration of 3D Gaussian Splatting representations [31]–[33] presents a promising avenue for future research.

VI. Acknowledgement

This work was sponsored by Beijing Municipal Science and Technology Commission (No. Z231100005923031), Innovation Funding of ICT, CAS (No. E461020) and National Natural Science Foundation of China (No. 62322210).

References

[1] Yang Zhou, Xintong Han, et al., "MakeItTalk: speaker-aware talking-head animation," *ACM Transactions On Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[2] Suzhen Wang, Lincheng Li, et al., "One-shot talking face generation from single-speaker audio-visual correlation learning," in *AAAI*, 2022, pp. 2531–2539.

[3] Wenxuan Zhang, Xiaodong Cun, et al., "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *CVPR*, 2023, pp. 8652–8661.

[4] Yifeng Ma, Suzhen Wang, et al., "StyleTalk: One-shot talking head generation with controllable speaking styles," *arXiv preprint arXiv:2301.01081*, 2023.

[5] Yuan Gan, Zongxin Yang, et al., "Efficient emotional adaptation for audio-driven talking-head generation," in *ICCV*, 2023, pp. 22634–22645.

[6] Jingxiang Sun, Xuan Wang, et al., "Next3D: Generative neural texture rasterization for 3D-aware head avatars," in *CVPR*, 2023, pp. 20991–21002.

[7] Zhiyuan Ma, Xiangyu Zhu, et al., "OTAvatar: One-shot talking face avatar with controllable tri-plane rendering," in *CVPR*, 2023, pp. 16901–16910.

[8] Ben Mildenhall, Pratul P Srinivasan, et al., "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[9] Eric R. Chan, Connor Z. Lin, et al., "Efficient geometry-aware 3D generative adversarial networks," in *CVPR*. 2022, pp. 16102–16112, IEEE.

[10] Linsen Song, Wayne Wu, et al., "Everybody's talkin': Let me talk as you want," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 585–598, 2022.

[11] Yudong Guo, Keyu Chen, et al., "AD-NeRF: Audio driven neural radiance fields for talking head synthesis," in *ICCV*, 2021, pp. 5784–5794.

[12] Xin Wen, Xuening Zhu, et al., "Cad-nerf: learning nerfs from uncalibrated few-view images by cad model retrieval," *Frontiers of Computer Science*, vol. 19, no. 10, pp. 1910706, 2025.

[13] Han Bao, Xuhong Zhang, et al., "Milg: Realistic lip-sync video generation with audio-modulated image inpainting," *Visual Informatics*, vol. 8, no. 3, pp. 71–81, 2024.

[14] Xinya Ji, Hang Zhou, et al., "EAMM: One-shot emotional talking face via audio-based emotion-aware motion model," in *SIGGRAPH*, 2022, pp. 1–10.

[15] Alexei Baevski, Yuhao Zhou, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[16] Aggelina Chatziagapi, ShahRukh Athar, et al., "LipNeRF: What is the right feature space to lip-sync a NeRF?," in *FG*. IEEE, 2023, pp. 1–8.

[17] Zhenlong Yuan, Jiakai Cao, et al., "Tsar-mvs: Textureless-aware segmentation and correlative refinement guided multi-view stereo," *Pattern Recognition*, vol. 154, pp. 110565, 2024.

[18] Jiahe Li, Jiawei Zhang, et al., "Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis," in *ICCV*, 2023, pp. 7568–7578.

[19] Zhenhui Ye, Jinzheng He, et al., "Geneface++: Generalized and stable real-time audio-driven 3d talking face generation," *arXiv preprint arXiv:2305.00787*, 2023.

[20] Junshu Tang, Bo Zhang, et al., "3DFaceShop: Explicitly controllable 3D-aware portrait generation," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[21] Zhenhui Ye, Tianyun Zhong, et al., "Real3d-portrait: One-shot realistic 3d talking portrait synthesis," in *ICLR*, 2024.

[22] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999.

[23] Eric Zhongcong Xu, Jianfeng Zhang, et al., "PV3D: A 3D generative model for portrait video generation," in *ICLR*, 2023.

[24] Ziqiao Peng, Yihao Luo, et al., "Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces," in *ACM MM*, 2023, pp. 5292–5301.

[25] Kaiwen Jiang, Shu-Yu Chen, et al., "NeRFFaceEditing: Disentangled face editing in neural radiance fields," in *SIGGRAPH Asia*, 2022, pp. 1–9.

[26] Kaisiyuan Wang, Qianyi Wu, et al., "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*. Springer, 2020, pp. 700–717.

[27] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *ACCVW*, 2016.

[28] Debin Meng, Xiaojiang Peng, et al., "Frame attention networks for facial expression recognition in videos," in *ICIP*. IEEE, 2019, pp. 3866–3870.

[29] Zhimeng Zhang, Lincheng Li, et al., "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *CVPR*, 2021, pp. 3661–3670.

[30] Hang Zhou, Yasheng Sun, et al., "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *CVPR*, 2021, pp. 4176–4186.

[31] Tong Wu, Yu-Jie Yuan, et al., "Recent advances in 3d gaussian splatting," *Comput. Vis. Media*, vol. 10, no. 4, pp. 613–642, 2024.

[32] Lin Gao, Jie Yang, et al., "Real-time large-scale deformation of gaussian splatting," *TOG*, vol. 43, no. 6, pp. 1–17, 2024.

[33] Tobias Kirschstein, Simon Giebenhain, et al., "Gghead: Fast and generalizable 3d gaussian heads," in *SIGGRAPH Asia*, 2024, pp. 1–11.