MergeTalk: Audio-Driven Talking Head Generation From Single Image With Feature Merge

Jian Gao, Chang Shu, Ximin Zheng, Zheng Lu, and Nengsheng Bao

***Abstract*—Audio-driven talking head generation has wide real world applications but remains challenging due to the problems such as audio-lip synchronization, head poses, identity preservation, video quality, etc. We propose a novel two-stage framework that uses explicit 3D face images rendered from a 3D model based on the audio input, as intermediate features. We devise two independent 3D motion parameter generation networks to generate expression and pose parameters for the popular 3DMM model to solve the audio-lip synchronization problem and natural head poses without losing identity information. To improve the final talking head quality such as avoiding facial distortion and artifacts, we propose a novel face feature merge network to accurately extract and fuse the background, identity information, facial texture from the source image, and the lip movements and head poses from the 3D face images, and generate the final videos based on generative adversarial networks. Extensive experiments show that our framework outperforms the SOTA methods in several aspects and has good generalization ability.**

***Index Terms*—Talking head generation, 3DMM, feature merge, GAN.**

I. Introduction

TALKING head generation is a very popular but challenging research topic with many practical applications in real-world scenarios, such as digital human generation, online education, video conferencing, etc. This paper targets to use a source image of a target person and arbitrary driver audio to generate realistic talking head videos that include accurate lip movements as well as natural head motions unconstrained by identity information.

Previous works such as [1], [2], [3] focused on generating accurate lip movements associated with the audio. Despite promising results, such focus is not enough to generate realistic talking head video including both lip movements and head motions. Generating natural talking head video still has challenges, such as maintaining identity features and good image quality during motion. Although a large number of works start to focus on these challenges, the generated results are subjected to different limitations, such as identity modification [4], [5], distorting the face [4], [5], or focusing on identity-guided changes in head motion [6]. Another common method based on facial landmarks [1], [7], [8], [9] can easily cause facial distortion.

In this work we propose a novel two-stage audio-driven approach to generate a full-frame video of talking head from a single image. We use the 3D face image rendered by the popular 3DMM model as an intermediate feature. Such an intermediate feature not only retains the lighting information in the source image but also has the identity information of the person. This way greatly reduces the difficulty of extracting target features and helps the generation of the subsequent target face image. Our approach has two separate 3D parameter generation networks, expression and pose, for the first stage to generate intermediate face representations better capturing lip movements and head poses without interfering face identity information. This is based on the observation that human voice is highly correlated with lip movements but head poses have a certain degree of randomness. Our framework has a face merge network for the second stage, based on Generative Adversarial Network (GAN) [10] to keep the identity information and background content in the reference image unchanged, and merge the expression and head pose in the 3D face to output video frames.

The main contribution of this paper can be summarized as follows:

- We propose a novel two-stage framework for audio-driven talking head generation from a single image using explicit 3D face images as intermediate features.
- We devise two independent networks to learn the realistic 3D motion parameters of the 3DMM model from audio separately, ensuring lip-audio synchronization and natural head poses without identity interference.
- We design a face feature merge network, MergeBlock, which can effectively extract and fuse features from different inputs for face video generation.
- Experimental results show that our framework beats the SOTA in terms of audio-lip synchronization, identity preservation and visual quality.

II. Method

As illustrated in Fig. 1, our framework firstly generates 3D motion coefficients for expression and head poses based on audio information, and then renders the 3D face images. A face merge

Manuscript received 15 May 2024; revised 25 June 2024; accepted 25 June 2024. Date of publication 3 July 2024; date of current version 19 July 2024. This work was supported in part by the 2020 Li Ka Shing Foundation Cross-Disciplinary Research under Grant 2020LKSFG06D and in part by Ningbo Science and Technology Bureau through Major Projects Fund under Project 2021Z089 and Grant 2022Z060. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Guang Hua. *(Jian Gao and Chang Shu contributed equally to this work.) (Corresponding author: Nengsheng Bao.)*

Jian Gao and Nengsheng Bao are with the Department of Mechanical Engineering, Shantou University, Shantou 515063, China (e-mail: 21jgao@stu.edu.cn; nsbao@stu.edu.cn).

Chang Shu and Ximin Zheng are with the Ping An Technology Company, Ltd., Shenzhen 518063, China (e-mail: shuchang254@pingan.com.cn; zhengximin135@pingan.com.cn).

Zheng Lu is with the University of Nottingham Ningbo China, Ningbo 315199, China (e-mail: zheng.lu@nottingham.edu.cn).

Digital Object Identifier 10.1109/LSP.2024.3422816

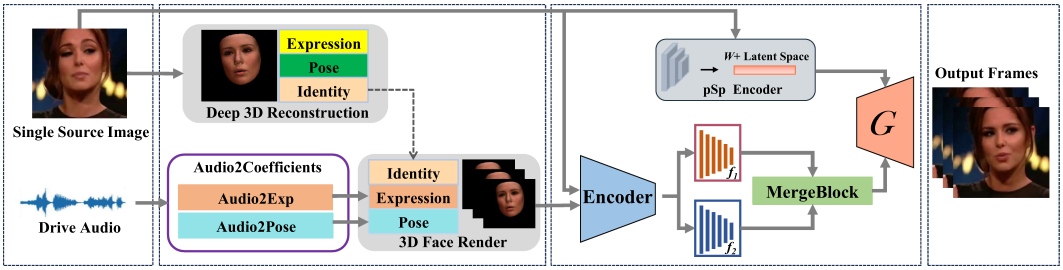

Fig. 1. Pipeline of MergeTalk. We render 3D face images based on the motion parameters generated for the 3DMM model as intermediate features, and use our MergeBlock to merge the face feature maps to generate video frames and output the video. f_1 and f_2 are feature maps of different scales obtained from the source image and 3D face image through the encoder.

network is devised to generate the final video frames. Below, we give a brief introduction to the 3D face model as preliminaries, and then 3DMM coefficients generation network and face merge network respectively.

A. 3D Face Reconstruction

Using audio solely to drive a single reference image talking head video generation is a much challenging task. For example, it is almost impossible to directly generate images with correct illumination from audio input. Instead, our framework utilizes the popular 3DMM model as our 3D face model to construct intermediate 3D face images for the final talking head video generation. 3DMM decouples the 3D face shape S as follows:

$$S = \overline{S} + \alpha B_{id} + \beta B_{exp}, \tag{1}$$

where $\overline{S}$ is the average face shape, B_{id} and B_{exp} are the Principal Component Analysis bases of identity and expression respectively, which are scaled with standard deviations. Parameters $\alpha \in \mathbb{R}^{80}$ and $\beta \in \mathbb{R}^{64}$ describe the facial shape and expression. The 3D face pose $\mathcal{P}$ is represented by rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$. To preserve identity change during parameter generation, we only construct motion parameters as the parameter set $\{\beta, R, t\} \in \mathbb{R}^{70}$. This way we can directly modify the expression and head poses of the rendered 3D face, to generate intermediate features.

B. Audio-Based Motion Parameter Generation

It is well recognized that, compared to head poses, lip movements and voice content have much stronger correlation, while head poses have a certain degree of randomness. When a person speaks the same content several times, his/her lip movements are very similar but head poses may vary. Using the same network for both expression and pose parameter generation may lead to unnatural videos for the same person. As a result, we opt to use two independent networks whose structures are very similar to each other. The structures are shown in Fig. 2. We use the popular Wav2Lip [2] to extract audio features which are used as the same input for both networks. We train an encoder consisting of MLP, GRU [11], and SAM [12] to regress expression parameters. GRU controls the flow of information through two gating units (update gate and reset gate) to capture temporal correlations and model long-term dependencies in audio feature sequences, while SAM helps the model acquire global dependencies among audio

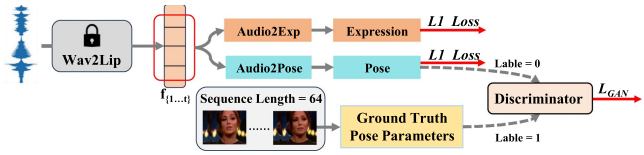

Fig. 2. Structure of the 3DMM motion parameters generation network. $f_{\{1,\ldots,t\}}$ is audio features of corresponding frames.Sampling pose parameters from 64 consecutive frames as real labels to train the discriminator.

features and focus on important features. Note that $L1$ loss is used as a constraint for training.

We design a pose parameters generator based on GAN to generate natural and identity-independent head pose parameters $P \in \mathbb{R}^6$, we add multiple one-dimensional BN layers to the expression parameter generation network to eliminate style differences between different identities. In the discriminator design, we mainly use multilayer 1D convolution with Batch Normalization and a Sigmoid activation function to determine whether the input pose sequences are naturally continuous.

For training, we fixedly sampling pose parameters from 64 consecutive frames as real labels, while $\bar{P} \in \mathbb{R}^{6\times64}$, generated using the speech features of the corresponding frames, are used as fake labels to train the discriminator. We consider that 64 frames of the image contain significant changes in head movement without causing further interference to the network. In order to prevent the amplitude of the generated head pose movements from being too large, we add the $L1$ loss with the adversarial loss L_{GAN} as a constraint for training. Inspired by SadTalker [6], we predict the difference from the initial head pose instead of predicting the head pose parameters directly.

C. Face Merge Network

Due to the limitations of 3DMM, the 3D face images do not contain details such as facial texture. Instead of attempting to generate 3D face image with texture lick Liu et al. [13], we opt to obtain the details from the source image directly. Specifically, our second stage aims to extract features for background, target character identity, and facial texture details from the source image and features for lip movements and head poses from the 3D facial images at the same time and then merge these individual features to generate a new facial image with more natural looking and actions. We use a residual network [14] based encoder to extract feature pyramid from both the source image and the 3D

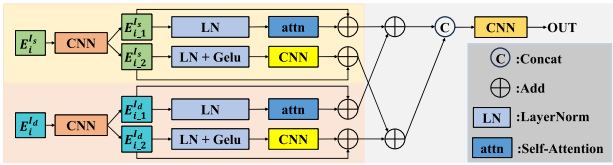

Fig. 3. Structure of MergeBlock. Feature merge by SAM and CNN and dimensionality up or down operation. i_1 and i_2 denote the two feature maps of the same dimensions after the first convolutional layer.

face image, obtaining two sets of feature maps with six different sizes, ranging from 256×256 to 8×8. Similar to LIA [15], we use a generator adapted from StyleGAN2 [16] and a pre-trained pSp Encoder [17] to convert the source image to its $\mathcal{W}+$ latent features other than directly feeding it into the generator. Such design can simultaneously encode different spatial features of the original image to achieve more controllable image editing effects.

The input to the face merge block, both feature maps hailing from distinct sources yet bearing the same dimensions, contain target features at different levels of granularity. Specifically, as shown in Fig. 3, each individual feature map increases the number of channels through a convolutional layer, leading to the derivation of two feature maps with equivalent number of dimensions. Subsequently, these feature maps are fed to a layer-normalization procedure and then either a SAM or a convolutional operation, all being summed through jump-connections. The features from their respective pathways are then added and concatenated along the channel dimension. The final feature map is obtained by applying a 1×1 convolutional operation for dimensionality reduction. The whole process is defined as follows:

$$OUT_i = \Phi_{merge}(E_i^{I_s}, E_i^{I_d}), \quad i \in \{1, \ldots, 6\}, \tag{2}$$

where I_s and I_d are the source image and 3D facial image, respectively. We obtain six levels of feature maps from 256×256 to 8×8 through the encoder.

We train our face merge network independently, using the video frames from the training data as 3D face images instead of the ones generated by the first stage, in order to minimize the impact of the generation errors in the first stage. We also use the first frame of the training videos as the source image and frames that are 20 to 80 frames away as 3D frames to ensure that the target image has a significant difference from the source image to prevent the neural network from cheating.

We use a weighted combination of several objectives to train our face merge network, including pixel-wise $L1$ distance between the output and the ground truth, and LPIPS [18] with a pre-trained VGG16 network [19] to learn perceptual similarities to preserve image quality. We also use the discriminator from PatchGAN [20] to obtain the adversarial loss L_{patch}. In summary, the total loss function of our face merge network is defined as follows:

$$\mathcal{L}_{merge} = \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{LPIPS} + \lambda_3 \mathcal{L}_{patch}, \tag{3}$$

where λ_1, λ_2, and λ_3 are constants for weight balance.

III. Experiments

A. Implementation Details and Metrics

We use the LRW dataset [21] to train the expression parameters generation network, the head pose parameters generation network and face merge network are trained using the voxceleb2 dataset [22]. We start training after loading the pre-trained models of pSp Encoder [17] and StyleGAN2 [16] in face merge network. We train all the networks separately using the Adam optimizer [23]. For both motion parameter generation networks, we adopt a learning rate of 10^{-4} for training, whereas for the face merge network we set the learning rate to 10^{-5}. we set λ_1, λ_2 and λ_1 as 10, 5, and 1 respectively. All 3DMM parameters are extracted by the pre-trained Deep3DFace reconstruction model [24].

We evaluate our framework using several metrics on both datasets. Specifically, We use structured similarity (SSIM) and peak signal-to-noise ratio (PSNR) to evaluate the low-level similarity between the generated image and the target image. To evaluate the identity retention during the generation process, we employ Arcface [25] to extract the identity embedding from the images and calculate the cosine similarity of the identity embedding between the source images and the generated images. We then normalize the results to a range from 0 to 1 for easy comparison, which are recorded in the column with the name ID in Tables I and II. We also evaluate audio-lip synchronization by applying the evaluation method in Wav2Lip and compute the distance score (LSE-D) and confidence score (LSE-C).

B. Experimental Result

1) Compare With SOTA Methods: Eight sate-of-the-art (SOTA) baselines of talking head generation are selected in this work for comparisons, i.e., MakeItTalk, Wav2Lip, Audio2head, PC-AVS, AVCT, SadTalker, IP-LAP, and AniPortrait. The authors of these methods provide pre-trained checkpoints and code for testing. The quantitative comparison is shown in Table I. Our framework obtains the best results in terms of image quality and identity preservation, and also obtains relatively high scores in terms of audio-lip synchronization. We observe that LSE metrics tend to be high when the lip movements are more exaggerated, leading to unnatural movements. In contrast, our method actually scores closer to the real videos, producing more natural output in our opinion.

Fig. 4 shows the qualitative comparison with SOTA methods. It can be seen that audio2head generates images with large areas of artifacts and distorted facial videos due to 2D distortion. Wav2Lip generates the lower half of the face based on the audio only and the synthesized video is less realistic, while PC-AVS cannot accurately guarantee the identity information. AVCT and IP-LAP, although maintaining good identity information, cannot accurately generate lip movements and the lips do not open most of the time.

2) Test With Other Style Inputs: To evaluate the generalizability of our framework, we also use multiple styles of images for testing. Fig. 5 shows the effectiveness of our framework in processing non-realistic faces. We can see that for different styles of input images, such as genders or painting styles, regardless of the facial angle and background colour, our model achieves

TABLE I
QUANTITATIVE COMPARISON WITH THE SOTA ON VOXCELEB2 DATASET [22]

Method	Voxceleb2					LRW				
	PSNR↑	SSIM↑	ID↑	LSE-C↑	LSD-D↓	PSNR↑	SSIM↑	ID↑	LSE-C↑	LSD-D↓
Ground Truth	/	1	1	7.514	7.473	/	1	1	3.713	9.846
MakeItTalk [9]	30.063	0.597	0.904	5.042	8.853	28.752	0.433	0.804	5.263	9.055
Wav2Lip* [2]	29.845	0.609	0.957	10.394	5.519	29.109	0.508	0.953	8.010	6.000
Audio2head [5]	29.160	0.499	0.499	5.862	9.550	28.566	0.460	0.785	4.827	8.792
PC-AVS [4]	28.368	0.392	0.547	8.557	6.752	27.902	0.434	0.684	6.661	6.303
AVCT [26]	28.853	0.462	0.563	4.290	9.306	28.883	0.473	0.712	6.748	7.987
SadTalker [6]	29.951	0.645	0.949	7.632	7.038	29.006	0.486	0.871	7.440	6.366
IP-LAP [27]	30.088	0.583	0.916	0.581	14.531	29.197	0.449	0.804	1.427	11.052
AniPortrait [8]	30.081	0.591	0.909	1.087	13.329	28.641	0.407	0.874	2.819	9.952
Ours	**30.147**	**0.664**	**0.954**	7.320	7.702	**29.225**	**0.492**	**0.885**	6.770	7.076

Wav2Lip* has the best identity retention because it only animates the lip area, while the rest of the area is the same as the original frame.

TABLE II
QUANTITATIVE EVALUATION OF THE RESULTS OF ABLATION STUDY

Method	PSNR↑	SSIM ↑	ID↑	LSE-C ↑	LSD-D↓
GT	/	1	1	7.514	7.473
addition	28.876	0.604	0.514	1.833	11.966
1×1 Conv	29.428	0.635	0.894	3.944	9.341
w/o attn	29.395	0.643	0.883	5.431	8.862
Ours	**30.147**	**0.664**	**0.954**	**7.320**	**7.702**

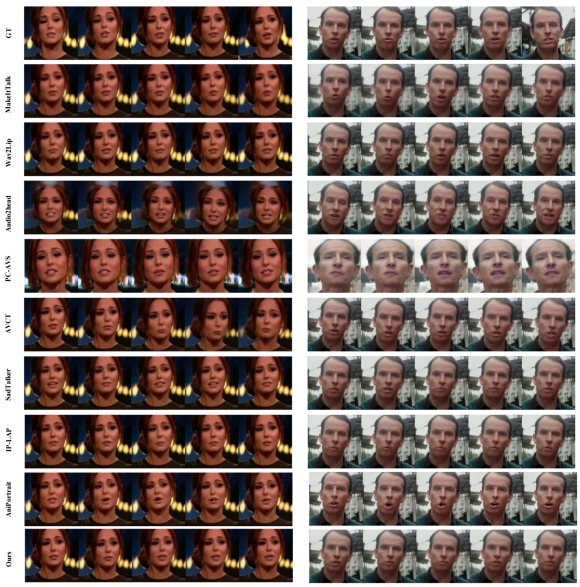

Fig. 4. Compared with the SOTA methods, our method produces better results in identity preservation and image quality.

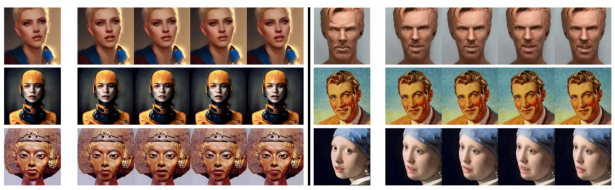

Fig. 5. Result for different styles inputs (sculpture, oil painting, etc.), demonstrating the generalization ability of our approach.

excellent performance, being close to the input real face image and preserving the subtle features in the image. Notably, the last row shows that our model works even in the case of a side face with a large angle.

Fig. 6. The ablation of MergeBlock. Comparison results of different feature merge methods with our MergeBlock.

C. Ablation Studies

To show the effectiveness of our MergeBlock, we compare three distinct feature fusion methods: direct summing of the feature maps, per-channel concatenation combined with 1×1 convolution, and replacing self-attention in the MergeBlock with a CNN, respectively. It is crucial to emphasize that in this ablation study the focus remains primarily on feature merging while keeping all other architectural components and parameters in our network intact. The 3D face image inputs remain fixed throughout both the training and inference phases. The performance differences are clearly shown in Fig. 6 and Table II. It is obvious that directly summing feature maps fails to generate realistic face images, while the other two methods have drawbacks such as loss of face details or large distortion. This empirical evidence strongly demonstrates the ability of our proposed face merge network to autonomously identify and extract key information from different input feature spaces, and subsequently perform the feature merging process.

IV. CONCLUSION

In this paper, we propose a novel method for talking face generation that relies on a single face image. We present two networks to generate accurate 3DMM expression parameters and natural and realistic head motion pose parameters which are rendered into 3D face images as intermediate representations. In addition, we devise a feature merge module, which can accurately preserve the identity, background and facial texture details and merge the lip movement and head pose from 3D face image to generate a new face image, open up a new direction of exploration for face editing or other fields. Various experiments demonstrate the advancement and generalization ability of our framework.

ACKNOWLEDGMENT

The work was done during the Jian Gao's internship at Ping An Technology.

REFERENCES

[1] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7832–7841.

[2] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 484–492, doi: 10.1145/3394171.3413532.

[3] K. Cheng et al., "Videoretalking: Audio-based lip synchronization for talking head video editing in the wild," in *Proc. SIGGRAPH Asia Conf. Papers*, 2022, pp. 1–9.

[4] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4176–4186.

[5] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," 2021, *arXiv:2107.09293*.

[6] W. Zhang et al., "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8652–8661.

[7] D. Yang, S. Huang, W. Jiang, and J. Zou, "Implicit memory-based variational motion talking face generation," *IEEE Signal Process. Lett.*, vol. 31, pp. 431–435, 2024.

[8] H. Wei, Z. Yang, and Z. Wang, "Aniportrait: Audio-driven synthesis of photorealistic portrait animation," 2024, *arXiv:2403.17694*.

[9] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: Speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, 2020.

[10] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[12] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[13] Y. Liu, Y. Fan, Z. Guo, A. Zaman, and S. Liu, "Fine-scale face fitting and texture fusion with inverse renderer," *IEEE Signal Process. Lett.*, vol. 30, pp. 26–30, 2023.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[15] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," in *Proc. Int. Conf. Learn. Representations*, 2022.

[16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[17] E. Richardson et al., "Encoding in style: A stylegan encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2287–2296.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[21] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Comput. Vis. 13th Asian Conf. Comput. Vis.*, Taipei, Taiwan, 2017, pp. 87–103.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," 2018, *arXiv:1806.05622*.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[24] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[26] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2531–2539.

[27] W. Zhong et al., "Identity-preserving talking face generation with landmark and appearance priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9729–9738.