Real-Time Audio-Guided Multi-Face Reenactment

Jiangning Zhang, Xianfang Zeng, Chao Xu, and Yong Liu

Abstract—**Audio-guided face reenactment aims to generate authentic target faces that have matched facial expression of the input audio, and many learning-based methods have successfully achieved this. However, most methods can only reenact a particular person once trained or suffer from the low-quality generation of the target images. Also, nearly none of the current reenactment works consider the model size and running speed that are important for practical use. To solve the above challenges, we propose an efficient *A*udio-guided *M*ulti-face reenactment model named *AMNet*, which can reenact target faces among multiple persons with corresponding source faces and drive signals as inputs. Concretely, we design a *Geometric Controller* (GC) module to inject the drive signals so that the model can be optimized in an end-to-end manner and generate more authentic images. Also, we adopt a lightweight network for our face reenactor so that the model can run in real-time on both CPU and GPU devices. Abundant experiments prove our approach's superiority over existing methods, *e.g.*, averagely decreasing FID by 0.12↓ and increasing SSIM by 0.031↑ than APB2Face, while owning fewer parameters (×4 ↓) and faster CPU speed (×4 ↑).**

Index Terms—**Real-time face reenactment, audio-guided face reenactment, multi-face reenactment, generative adversarial nets.**

I. Introduction

AUDIO-GUIDED face reenactment aims to generate authentic target faces under the condition of audio information along with auxiliary pose and eye blink signals, which has promising applications such as animation production, virtual human, and game. However, most current methods can only reenact a particular person once finishing the training procedure or suffer from the low-quality problem of the generated target images. Also, nearly none of the current reenactment works take the model size and running speed into account that is important for practical use. This work focuses on solving the above problems, and we improve previous APB2Face [1] to an efficient end-to-end model to handle audio-guided multi-face reenactment, where different target faces among multiple persons can be reenacted by only one unified model.

Benefitted from advances in computing devices and deep learning, many methods have achieved good results in the audio-to-face task [2]–[4]. Duarte *et al.* [5] propose the Wav2Pix to generate the target face by an embedded audio vector in an adversarial manner, while Zhang *et al.* [1] design an APB2Face model that employs a two-stage generation procedure and dramatically improves the quality of the generated image. However, these methods can only model one specific person once trained, meaning that we need to store and transfer an equal number of models as actual persons. Recently, Zhou *et al.* [6] propose to generate expressive talking-head videos by disentangling the content and speaker information in the input audio signal that can handle the multi-face reenactment task. However, it suffers from low-quality generated images and a redundant structure that is hard to follow. In this paper, we perform audio-guided face reenactment among multiple persons built on the previous APB2Face and design a novel *Geometric Controller* (GC) module to inject the drive signals more efficiently, enabling the end-to-end training and high-quality generation procedure. To further alleviate the problem of slow running speed for different scenarios, *e.g*., GPU for a sever and CPU for a mobile device, we adopt a lightweight neural structure [7] for our Audio-guided Multi-face Reenactor (AMNet) so that the model can run in real-time on different devices. Specifically, we make the following three contributions:

i) We propose an efficient *AMNet* to reenact different target faces among multiple persons by one unified model.

ii) A novel GC module is proposed to enable the end-to-end training and authentic generation procedure.

iii) A lightweight backbone is adopted so that our approach can run in real-time on both CPU and GPU.

II. Related Works

Generative Adversarial Networks. Since generative adversarial network (GAN) is first proposed [8], many excellent works are came up to generate authentic images in succession. These methods mainly fall into two categories: the vector-based method [9]–[14] that uses noise or embedded vector as input to generate the target image; and the pixel-based method [15]–[19] that uses the image as input. Theoretically, each GAN-based method contains a generator G with parameter θ_g to capture the data distribution for generating authentic images, as well as a discriminator D to authenticate generated images for enhancing the capability of G in an adversarial manner. To learn the distribution of G over data x from a prior distribution $p_z(z)$ $(G(z;\theta_g) \in p_{data}(x))$, D plays a two-player minimax game with G in the following value function $V(D,G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log(D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))]. \quad (1)$$

Our method belongs to the pixel-based category, and we employ the lightweight network [7] as the primary backbone.

Manuscript received September 2, 2021; accepted September 26, 2021. Date of publication September 29, 2021; date of current version January 20, 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61836015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Corresponding author: Yong Liu.)*

The authors are with the APRIL Lab, College of Control Science and Engineering, Zhejiang University, Hangzhou 310000, China (e-mail: 186368@zju.edu.cn; zzlongjuanfeng@zju.edu.cn; 21832066@zju.edu.cn; yongliu@iipc.zju.edu.cn).

(the PDF file is attached to the Supporting Documents). Code is available at https://github.com/zhangzjn/APB2FaceV2.

Digital Object Identifier 10.1109/LSP.2021.3116506

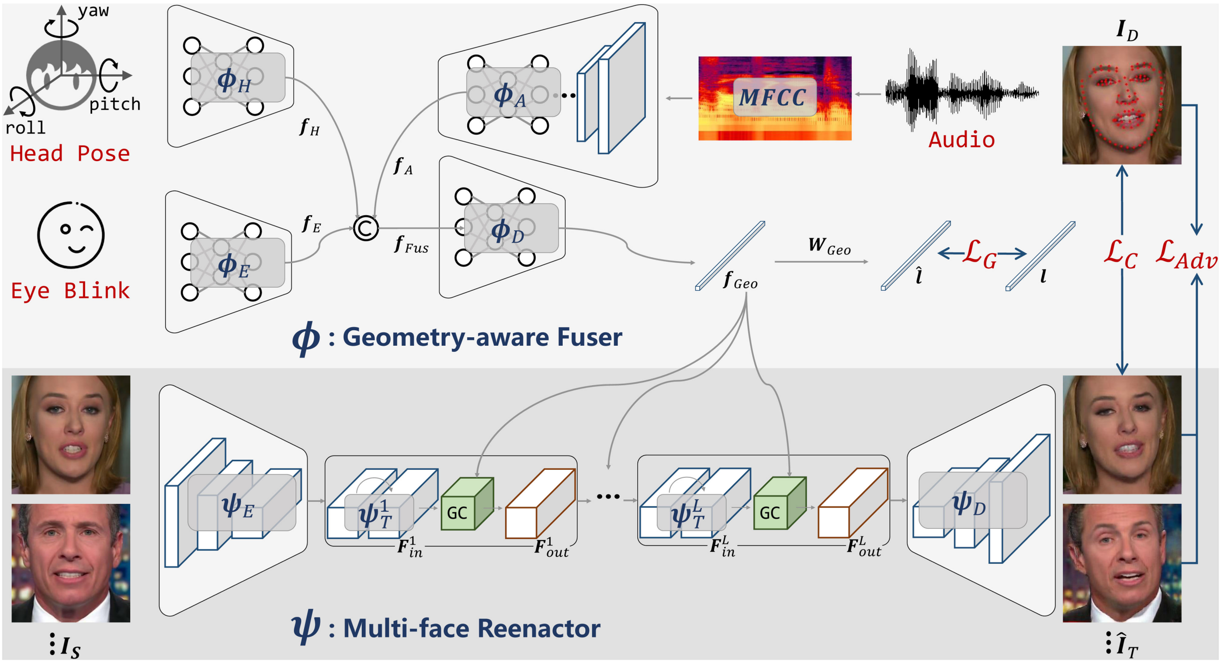

Fig. 1. **Overview of the proposed AMNet that consists of a Geometry-aware Fuser (ϕ) and a Multi-face Reenactor (ψ).** ϕ inputs audio, head pose, and eye blink signals from the drive frame $\boldsymbol{I}_D$, which are extracted by ϕ_A, ϕ_H, and ϕ_E to obtain embedded features $\boldsymbol{f}_A$, $\boldsymbol{f}_H$, and $\boldsymbol{f}_E$, respectively. These features are concatenated and go through ϕ_D to obtain the facial geometric representation $\boldsymbol{f}_{Geo}$. Subsequently, ψ inputs the source face $\boldsymbol{I}_S$ and $\boldsymbol{f}_{Geo}$ to reenact the target face $\hat{\boldsymbol{I}}_T$ that has matched facial expression with the input drive signals. Specifically, the GC module is used to inject facial geometric information into the reenactor flexibly and efficiently. Red points in the $\boldsymbol{I}_D$ represent the detected landmark.

Face Reenactment via Audio. Many works have yielded promising results in the audio-to-face task that uses the audio signal to control orofacial movements. Works [2]–[4], [20] use the audio signal to predict parameters of the predefined 3D model, while Suwajanakorn *et al.* [21] and Prajwal *et al.* [22] propose to predict the lip rather than the whole face. These methods need extra post-operations such as 3D rendering or face fusion that is cumbersome and unsuitable for practical applications. Works [23]–[27] propose to design an end-to-end model to generate the target face. X2Face [23] is a self-supervised network architecture that achieves the face puppeteering of a source face given a driving vector, while Bai *et al.* [26] design an SF2F model to improve the generation quality and address the poor connection between vocal feature domain and modern image generation models. Choi *et al.* [27] exploits the development of GANs and proposes to generate target face directly from the speech waveform. Recently, Zhou *et al.* [6] propose to generate expressive talking-head videos by disentangling the content and speaker information in the input audio signal that can handle multi-face reenactment task, but it suffers from low-quality generated images and the redundant structure. This paper proposes an efficient AMNet to reach audio-guided multi-face reenactment by one end-to-end model, which can generate authentic target faces in real-time.

III. Method

Fig. 1 shows our proposed AMNet that consists of an *Geometry-aware Fuser* (ϕ) and a *Multi-face Reenactor* (ψ), which is capable of completing the more challengeable audio-guided multi-face reenactment task efficiently.

Geometry-aware Fuser (ϕ). Audio, head pose, and eye blink signals from the drive frames are fed to ϕ, which are further extracted by ϕ_A, ϕ_H, and ϕ_E to obtain $\boldsymbol{f}_A$, $\boldsymbol{f}_H$, and $\boldsymbol{f}_E$, respectively. Specifically, ϕ_A contains five convolutional layers for extracting the feature of each time node and additional five convolutional layers for fusing them, while both ϕ_H and ϕ_E are composed of linear layers. Subsequently, the concatenated fusion feature $\boldsymbol{f}_{Fus}$ goes through ϕ_D to obtain the embedded representation $\boldsymbol{f}_{Geo}$, which represents the geometric features of the face. Specifically, the facial landmark is used as the geometric supervisory signal in the training stage.

$$\boldsymbol{f}_{Geo} = \phi_D(\boldsymbol{f}_{Fus}) = \phi_D([\boldsymbol{f}_A, \boldsymbol{f}_H, \boldsymbol{f}_E]). \tag{2}$$

Multi-face Reenactor (ψ). Given a source face image $\boldsymbol{I}_S$ and the referential drive feature $\boldsymbol{f}_{Geo}$, ψ reenacts the target face $\hat{\boldsymbol{I}}_T$ that has matched facial expression with the input signals. Specifically, ψ consists of a chain of sub-modules: an image encoder ψ_E, a feature transformer $\psi_T = \{\psi_T^1, \psi_T^2, \ldots, \psi_T^L\}$ (L is the repetition number and is set to 9 in the paper), and an image decoder ψ_D. The chain process can be described as:

$$\hat{\boldsymbol{I}}_T = \psi_D(\psi_T(\psi_E(\boldsymbol{I}_S))). \tag{3}$$

Note that ψ_T is designed in a decoupling idea that simultaneously learns appearance information from $\boldsymbol{I}_S$ as well as the geometric information from $\boldsymbol{f}_{Geo}$. The proposed *GC* module is used to inject facial geometric information into each block.

Geometry Controller. Different from two-stage APB2Face that injects facial movement by first plotting the landmark image and then concatenating it with deep features, we propose an elegant *Geometric Controller* module to directly inject geometric information to control the face movement in an end-to-end manner.

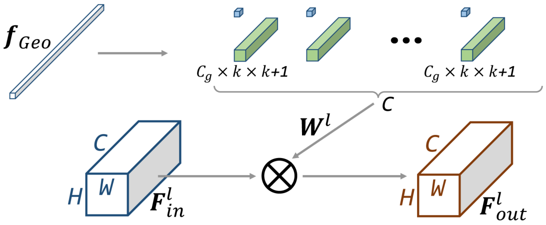

Fig. 2. **Overview of the proposed GC module.** Green and blue cuboids are predicted weight and bias parameters. ⊗ represents the convolution.

As depicted in Fig. 2, a geometric feature $\boldsymbol{f}_{Geo}$ and a deep feature map $\boldsymbol{F}^l_{in} \in \mathbb{R}^{C\times H\times W}$ are fed into GC module, and it outputs a modified feature $\boldsymbol{F}^l_{out} \in \mathbb{R}^{C\times H\times W}$. In detail, $\boldsymbol{f}_{Geo}$ goes through two linear layers to generate a set of parameters $\boldsymbol{W}^l \in \mathbb{R}^{C\times(C_g\times k\times k+1)}$ that are viewed as weight and bias parameters of the convolution layer:

$$\begin{aligned}\boldsymbol{F}^l_{out} &= Conv(\boldsymbol{F}^l_{in}; \boldsymbol{W}^l)\\ &= Conv(\boldsymbol{F}^l_{in}; Linear^l(\boldsymbol{f}_{Geo})).\end{aligned} \tag{4}$$

Specifically, $\boldsymbol{W}^l$ contains $C \times C_g \times k \times k$ weight parameters and C bias parameters, where k, C, and g are kernel size, channel number, and group number, and $C_g = C//g$. Thus we can control the intensity of injected information by controlling these parameters. Specifically, GC reduces to AdaIN [28] when $\{k = 1, g = C\}$, and we set $\{k = 3, g = C\}$ for better fusing local information, *i.e.*, 3×3 regions.

Objective Function. We adopt geometry and content losses to supervise geometric information and generated images. Adversarial loss is used to improve the authenticity of the reenacted image. The overall loss function is:

$$\mathcal{L}_{All} = \lambda_G \mathcal{L}_G + \lambda_C \mathcal{L}_C + \lambda_{Adv} \mathcal{L}_{Adv}, \tag{5}$$

where λ_G, λ_C, and λ_{Adv} represent weight parameters to balance different terms and are set 1, 100, and 1, respectively.

i) Geometry loss $\mathcal{L}_G$ calculates ℓ_1 error between the predicted facial landmark $\hat{\boldsymbol{l}}$ and the real facial landmark $\boldsymbol{l}$:

$$\mathcal{L}_G = ||\hat{\boldsymbol{l}} - \boldsymbol{l}||_1 = ||\boldsymbol{W}_{Geo}\boldsymbol{f}_{Geo} - \boldsymbol{l}||_1, \tag{6}$$

where the geometric feature $\boldsymbol{f}_{Geo}$ goes through a linear layer to regress the facial landmark $\hat{\boldsymbol{l}}$.

ii) Content loss $\mathcal{L}_C$ calculates ℓ_1 error between the reenacted target face $\hat{\boldsymbol{I}}_T$ and corresponding real face $\boldsymbol{I}_T$.

$$\mathcal{L}_C = ||\hat{\boldsymbol{I}}_T - \boldsymbol{I}_T||_1. \tag{7}$$

iii) Adversarial loss $\mathcal{L}_{Adv}$ adopts an extra discriminator D to form an adversarial training against the reenactor G. This loss significantly improves the quality of the generated image.

$$\mathcal{L}_{Adv} = \mathbb{E}_{\hat{\boldsymbol{I}}_T\sim p_f}[D(\hat{\boldsymbol{I}}_T)] - \mathbb{E}_{\boldsymbol{I}_T\sim p_r}[D(\boldsymbol{I}_T)], \tag{8}$$

where p_f and p_r stand distributions for generated and real images, and D contains five convolution layers in the paper.

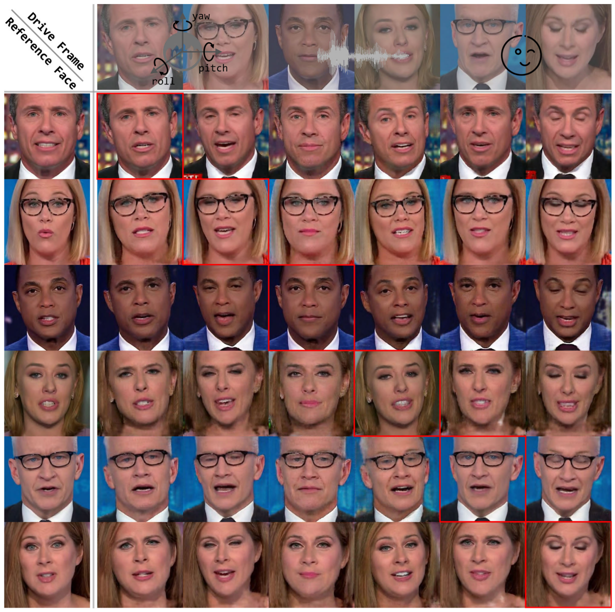

Fig. 3. **Qualitative results among multiple persons on the AnnVI dataset.** The first column contains six randomly selected source faces, while the first row shows drive frames from different persons that supply audio, head pose, and eye blink signals. The generated faces marked with red rectangles are driven by their own audio, while other faces are driven by different persons.

IV. Experiments

Dataset. Most experiments are conducted on the AnnVI dataset that contains six announcers (three men and three women) and 23,790 frames totally (sorted by name, the first 20,000 frames for training and the rest 3,790 for testing) with a synchronous audio clip, head pose, eye blink, and landmark [29] information.

Evaluation Metrics. We use SSIM [30] and LPIPS [31] to measure the quality of the generated images at pixel level, while FID [32] at semantic level.

Implementation Details. We use Adam optimizer [33] with $\{\beta_1 = 0.5,\ \beta_2 = 0.999\}$ and train the model for 110 epochs. The learning rate is set to $2e^{-4}$, and the batch size is 16. PatchGAN [15] is used as the discriminator. **Qualitative Results**. Fig. 3 shows qualitative generated images on the AnnVI dataset to visually demonstrate the superiority of the proposed approach. Specifically, we randomly select one face in all identities (6 faces totally) as the source face and one drive frame of each identity (supplying audio, head pose, and eye blink signals) to reenact the target face. Results indicate that our proposed method successfully achieves the multi-face reenactment task with one unified network and can generate authentic target images. Thanks to the decoupling design, AMNet has strong generalization ability capable of generating identity-consistent target images by non-self geometric information.

Quantitative Results. Table I shows quantitative evaluation with the SOTA APB2Face on the AnnVI dataset, and each metric result is averaged score for six persons in the dataset. Note that *Detection Rate* (DR, using the dlib to detect whether the generated image contains a face), *Average Landmark Error* (ALE), *Average Pose Error* (APE), and *Average Blink Error* (ABE) metrics are used for fair comparison [1]. Experimental

TABLE I
QUANTITATIVE ASSESSMENT WITH THE SOTA APB2FACE

Method	FID ↓	SSIM ↑	LPIPS ↓	DR ↑	ALE ↓	APE ↓	ABE ↓
APB2Face [1]	11.862	0.799	0.0195	98.8	1.429	0.0195	0.0413
Ours	**11.206**	**0.805**	**0.0187**	**99.1**	**1.382**	**0.0187**	**0.0405**

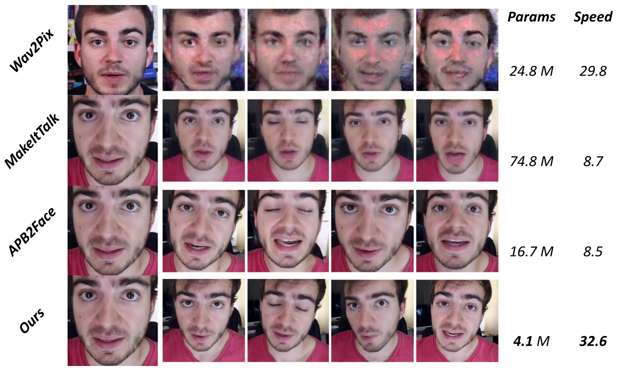

Fig. 4. Qualitative comparison with SOTA face reenactment methods.

TABLE II
QUALITATIVE COMPARISON WITH SOTAS ON THE YOUTUBERS DATASET

Method	FID ↓	SSIM ↑	LPIPS ↓	Params ↓ (M)	FPS ↑ (CPU)	FPS ↑ (GPU)
Wav2Pix [5]	214.280	0.537	0.0826	24.771	29.8	**275.7**
MakeItTalk [6]	33.640	0.657	0.0387	74.828	8.7	65.6
APB2Face [1]	13.375	0.734	0.0225	16.696	8.5	200.4
Ours	**13.256**	**0.765**	**0.0211**	**4.085**	**33.8**	158.9

TABLE III
QUANTITATIVE ABLATION STUDY FOR DIFFERENT LOSS TERMS

$\mathcal{L}_C$	$\mathcal{L}_G$	$\mathcal{L}_{Adv}$	FID ↓	SSIM ↑	LPIPS ↓
✓	✗	✗	15.556	0.735	0.0242
✓	✓	✗	14.172	0.786	0.0203
✓	✗	✓	12.528	0.749	0.0215
✓	✓	✓	**11.206**	**0.805**	**0.0187**

TABLE IV
QUANTITATIVE ABLATION STUDY FOR THE GC MODULE

Method	FID ↓	SSIM ↑	LPIPS ↓
AdaIN [28]	11.515	0.792	0.0196
GC	**11.206**	**0.805**	**0.0187**

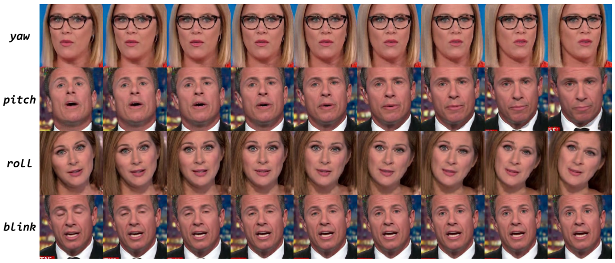

Fig. 5. Decoupling experiments of pose and eye blink signals.

results indicate that the proposed method obtains better metric scores than the basic APB2Face, demonstrating the superiority of AMNet for generating more authentic faces. Note that we use only one unified model that is more practical, while APB2Face needs to train six models totally for six persons.

Comparison with SOTAs. We further conduct a comparison experiment with most related SOTA methods on the Youtubers dataset [5]. As shown in Fig. 4, our approach can generate more authentic and clearer face images compared with other methods. We also conduct quantitative experiments in Table II, and results indicate that our approach obtains better scores on several image-quality assessment metrics, *i.e.*, FID=13.256, SSIM=0.765, and LPIPS=0.0211, as well as is more efficient than other approaches. Concretely, our method owns the smallest parameters (*i.e.*, 4.1 M, ×6 ↓ than Wav2Pix and ×4 ↓ than APB2Face) and can run in real-time on both CPU (**33.8** FPS on i9-10900 K) and GPU (**158.9** FPS on 2080 Ti). In general, our method obtains better performance than other approaches and takes the model efficiency into account, *e.g.*, parameters and running speed, which is of great practical value.

Ablation Study. We conduct ablation studies on the AnnVI dataset to assess the contribution of each loss function and the effectiveness of the proposed GC module. a) We quantitatively evaluate the effectiveness of each loss function in Table III. Results show that each loss function helps to improve the model performance, and the model obtains the best scores when all loss functions are applied. b) We also conduct an ablation study by replacing GC with AdaIN [28], and Table IV illustrates the superiority of the proposed GC than AdaIN for obtaining higher metric scores. **Decoupling Experiment**. Considering that the input of the three drive signals themselves (*i.e.*, audio, head pose, and eye blink) are decoupled, we evaluate whether the model has learned the decoupling generation from the above signals. As shown in Fig. 5, the first three rows are generated images that only change one component of the head pose signal, *i.e.*, yaw, pitch, or roll, while the last row shows the results only changing the eye blink signal. Results indicate that our method learns how to control the facial properties of the generated faces, where continuous interpolation generation and single-dimensional manipulation become possible.

V. CONCLUSION

This paper proposes a real-time AMNet to address the challenging audio-guided multi-face reenactment task. Specifically, a Geometry-aware Fuser is used to predict the geometric representation from drive signals, and the Multi-face Reenactor controls the face reenactment procedure by fusing geometric information with the source face that supplies appearance information. Besides, a novel GC module is proposed to inject geometric information so that the model can be optimized in an end-to-end manner and generate more authentic images. Extensive experiments demonstrate the effectiveness and efficiency of our approach.

We will further combine Neural Architecture Search with our approach to search for a more accurate and faster model for better practical applications, and we hope our work will help users achieve better works.

References

[1] J. Zhang, L. Liu, Z. Xue, and Y. Liu, "APB2Face: Audio-guided face reenactment with auxiliary pose and blink signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4402–4406.

[2] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 10101–10111.

[3] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with learning-based personalized natural head pose," 2020, *arXiv:2002.10137*.

[4] G. Tian, Y. Yuan, and Y. Liu, "Audio2Face: Generating speech/face animation from single audio with attention-based bidirectional LSTM networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2019, pp. 366–371.

[5] A. Duarte *et al.*, "WAV2PIX: Speech-conditioned face generation using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8633–8637.

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

[7] Y. Fu, W. Chen, H. Wang, H. Li, Y. Lin, and Z. Wang, "AutoGAN-distiller: Searching to compress generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3292–3303.

[8] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–26.

[11] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–35.

[12] A. Karnewar and O. Wang, "MSG-GAN: Multi-scale gradients for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 7799–7808.

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4401–4410.

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

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

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2017, pp. 2223–2232.

[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8798–8807.

[18] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 8188–8197.

[19] T. Wei *et al.*, "A simple baseline for styleGAN inversion," 2021, *arXiv:2104.07661*.

[20] N. Sadoughi and C. Busso, "Speech-driven expressive talking lips with conditional sequential generative adversarial networks," *IEEE Trans. Affect. Comput.*, early access, 2019. doi: 10.1109/TAFFC.2019.2916031.

[21] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, pp. 1–13, 2017.

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

[23] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. IEEE Eur. Conf. Comput. Vision*, 2018, pp. 670–686.

[24] T.-H. Oh *et al.*, "Speech2Face: Learning the face behind a voice," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7539–7548.

[25] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio" *Int. J. Comput. Vision*, vol. 127, no. 11, pp. 1767–1779, 2019.

[26] Y. Bai, T. Ma, L. Wang, and Z. Zhang, "Speech fusion to face: Bridging the gap between human's vocal characteristics and facial imaging," 2020, *arXiv:2006.05888*.

[27] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–18.

[28] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2017, pp. 1501–1510.

[29] "Face Attributes - Face++ Cognitive Services.", Accessed: Sep. 16, 2017. [Online]. Available: https://www.faceplusplus.com/attributes/

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[31] 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. Vision Pattern Recognit.*, 2018, pp. 586–595.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.