ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094807

WAVSYNCSWAP: END-TO-END PORTRAIT-CUSTOMIZED AUDIO-DRIVEN TALKING FACE GENERATION

Weihong Bao[1,*], *Liyang Chen*[1], *Chaoyong Zhou*[2], *Sicheng Yang*[1], *Zhiyong Wu*[1,3,†]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Ping An Technology, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong SAR, China
{bwh21,cly21,yangsc21}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

ABSTRACT

Audio-driven talking face with portrait customization enhances the flexibility of avatar applications for different scenarios, such as online meetings, mixed reality, and data generation. Among the existing methods, audio-driven talking face and face swapping are typically viewed as separate tasks that are cascaded to achieve the objective. Using state-of-the-art methods Wav2Lip and SimSwap for this purpose, we meet some issues: affected mouth synchronization, lost texture information, and slow inference speed. To resolve these issues, we propose an end-to-end model that combines the advantages of both approaches. Our approach generates highly-synchronized mouth with the aid of a pre-trained lip-sync discriminator. And identity information is provided by ArcFace and the ID injection module in the model because of its strong correlation with facial texture. Experimental results demonstrate that our method achieves lip-sync accuracy comparable to real synced videos, preserves more texture details than cascade methods, and alleviates the blurring of Wav2Lip. Also, our approach improves the inference speed.[1]

Index Terms— Talking face generation, Face swapping, Audio driven animation

1. INTRODUCTION

Audio-driven talking face generation has been applied in various scenes, such as online conferencing, visual dubbing, and virtual customer service. The ability to customize portraits can significantly enhance the flexibility and practicality of this technology in real-world applications, such as using it to voice a stunt double in a movie. Combining the recently flourishing face-swapping technique with talking face generation can provide a viable solution to achieve this goal. Face swapping can migrate the identity of the source face to the target face while keeping the expression, pose, lighting, and other attributes of the target face. The combination of these two technologies allows us to customize a personalized virtual portrait in the talking face application, for example, by dynamically adapting the virtual customer service portrait based on user characteristics. Also, the combination of these two techniques has been used to generate realistic multimodal datasets for deep forgery video detection [1].

Talking face generation methods. With the development of deep learning techniques, talking face generation methods have made great progress and have been able to generate realistic-enough results [2–7]. Wav2lip [2] employs a pre-trained lip-synchronized expert discriminator to supervise the generation of speech-synchronized mouth shapes without the aid of a structured representation. The method [3] assumes a latent space consisting of orthogonal motion directions of the static portrait, and the portrait is driven by linear navigation in the latent space corresponding to desired image manipulation. Some works [4, 5] predict the landmarks of the face from speech as an intermediate representation and use it to drive cartoon images or real people. Not only landmarks but also some structured information is used as intermediate representations to be applied in GAN-based methods, such as the 3D morphable model (3DMM) coefficients [7, 8] and the 3D blendshape basis [6].

Face swapping methods. Face swapping methods have been extensively investigated by the academic community. The open-source algorithm [9] consists of a common encoder and two identity-specific decoders allows face swapping between two specific identities. HifiFace [10] controls the face shape with the help of prior knowledge of 3DMM. SimSwap [11] proposes an ID injection module using Adaptive Instance Normalization (AdaIN) [12] to inject the identity feature extracted from the source face into the target face and can generalize to arbitrary identity. Smooth-Swap [13] proposes a method to construct a smoother identity feature space, which can provide more stable gradients during training.

Combination of talking face generation and face swapping methods in a two-step cascade is a straightforward way to achieve audio-driven talking face generation with portrait customization. First, the talking face generation module is fed audio and video to generate lip-synchronized video. Afterwards, the face swapping module modifies the identity of the generated video. Alternately, the two steps may be performed in reverse, as demonstrated by the audio-video multimodal deepfake dataset FakeAVCeleb [1]. This cascading method permits the generation of mouth shapes synchronized with arbitrary speech, as well as the ability to freely switch identities.

Wav2Lip has received considerable attention due to the accuracy of its generated mouth shapes among the existing methods for talking face generation. SimSwap has a simple training procedure and can generalize to arbitrary identity while preserving the attributes of the target face. However, there are some issues with cascading these two advanced techniques to achieve the desired outcomes. While generating videos with excellent lip-sync mouth shapes, Wav2Lip is limited by the output resolution of 96x96 and the blurring of results. For face swapping module, the modification of the face also has an impact on the accuracy of the mouth shape. In the meantime, the two-step cascade approach requires training for the talking face generation module and the face swapping module respectively. This process not only brings about a lengthy training process, but also results in the accumulation of errors and loss of information during

[*] Work performed during an internship at Ping An Technology.

[†] Corresponding author.

[1] Demo videos url: https://tontboy.github.io/projects/wav_sync_swap

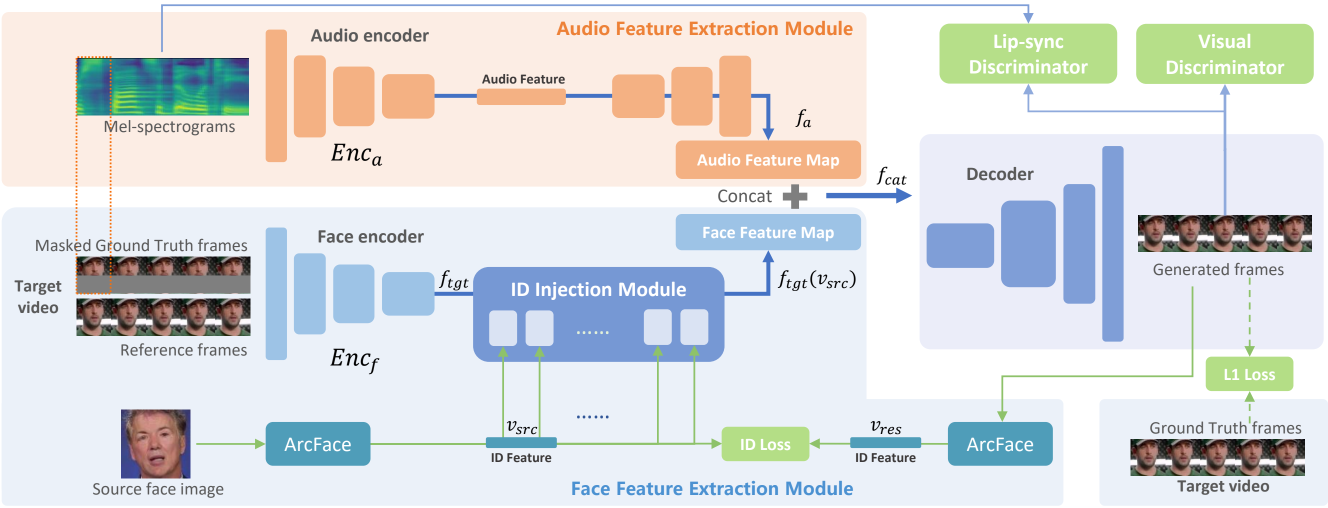

Fig. 1. The structure of our proposed framework as a whole.

data transfer. In addition, the stacking of modules in the cascade approach inevitably results in a slower inference speed.

Therefore, we propose an end-to-end model that can accommodate both talking face generation as well as personalized portrait customization. We summarize the contributions as follows:

1. We propose an end-to-end audio-driven talking face generation model that supports face swapping and is able to generate highly accurate mouth shapes.

2. The proposed approach allows the generated results to retain more facial texture information compared to the cascade approach and also alleviates the blurring of Wav2Lip.

3. The effectiveness of the proposed end-to-end model is tested on the HDTF dataset [14], and the result shows considerable performance improvement of inference speed.

2. METHODOLOGY

Given a target video, a source face image, and an arbitrary audio, our method can generate a talking face video in which the identity information comes from the source face image, the mouth shape is synchronized with the given audio, and the other parts come from the given target video.

2.1. Model architecture

To implement an end-to-end talking face generation method that supports face swapping, we combine the benefits of advanced talking face generation and face swapping techniques. We propose a model that combines concepts from Wav2Lip and SimSwap, as shown in Figure 1.

The model has three modules: the audio feature extraction module, the face feature extraction module, and the decoder. The first two parts of the model take audio, the target video frames and the source face image as inputs, and generate talking face video frames synchronized with the input audio and with the identity of source face.

The Mel-spectrograms of audios are sent to the audio feature extraction module as input. Firstly, an audio encoder Enc_a consisting of 14 residual convolutional blocks with batch normalization layer and ReLU activation encodes the Mel-spectrograms to 512-D audio feature. Then the module maps it to an intermediate feature map $f_a \in \mathbb{R}^{256\times12\times12}$ through several transposed convolutional blocks so that the audio feature is consistent with the face feature at dimension.

The face feature extraction module is composed of a face encoder Enc_f, an identity extractor ArcFace [15], and an ID injection module borrowed from SimSwap. The face encoder takes the target video frames corresponding to the audio and converts them into feature maps $f_{tgt} \in \mathbb{R}^{512\times12\times12}$. The identity extractor ArcFace extracts the 512-dimensional ID feature v_{src} from the source face. Then the ID injection module takes v_{src} and f_{tgt} as input, and transfer the identity feature of source face v_{src} into the feature map of target face f_{tgt} through a stack of 9 residual ID-Blocks consists of CNN layers and AdaIN layers. After that, the ID injection module changes f_{tgt} to $f_{tgt}(v_{src})$.

After getting the extracted audio feature f_a and face feature $f_{tgt}(v_{src})$, these two features are concatenated together. Finally, the concatenated feature $f_{cat} \in \mathbb{R}^{768\times12\times12}$ is sent to the decoder consisting of 5 transposed convolutional blocks and 3 convolutional blocks. The decoder outputs the video frame with the identity of the source face and lips synchronized with the audio frame.

2.2. Lip-synchronized

Whether the mouth shape generated by the talking face generation method is synchronized with audio is an important factor affecting performance. Inspired by Wav2Lip, we use the pre-trained lip-sync discriminator [2, 16] to supervise the model to generate accurate mouth shapes. This lip-sync discriminator takes the generated video frame and paired audio frame as input and outputs the probability of lip-sync P_{sync} . The model is optimized by the loss of lip-sync L_{sync} :

$$\mathcal{L}_{\text{sync}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(P^{i}_{\text{sync}}\right) \tag{1}$$

2.3. Identity information injection

In some existing face swapping methods [11,17,18], the model only takes the identity feature extracted from an arbitrary source face as the channel offers identity information to generate a target image with the identity feature of the source face. Identity feature makes a substantial impact on facial texture. Consequently, we introduce the ID injection module from SimSwap and employ identity infomation to improve the texture details of final results. The identity injection module allows our proposed end-to-end model to both support face swapping and preserve more texture details while generating highly accurate mouth shapes. As demonstrated in Figure 1, our model employs ArcFace to extract the identity feature of the source face I_{src}. And the ID injection module transfers identity infomation to the target video frame from I_{src}.

The ID injection module consists of 9 ID-Blocks. Each ID-Block consists of two convolutional blocks with AdaIN, and an ReLU activation function is inserted between these them. The formulation of AdaIN in our task can be written as:

$$\mathrm{AdaIN}\left(f_{tgt}, v_{src}\right) = \sigma_{src}\frac{f_{tgt} - \mu(f_{tgt})}{\sigma(f_{tgt})} + \mu_{src} \tag{2}$$

Here, $\mu(f_{tgt})$ and $\sigma(f_{tgt})$ is the channel-wise mean and standard deviation of the input feature f_{tgt}. σ_{src} and μ_{src} are predicted by full connected layers from v_{src}. Since there's no Ground Truth in face swapping task, we use the cosine similarity loss between the identity features of the source face v_{src} and the result video frame v_{res} to guide network optimization and convergence. The loss function is formulated as follows:

$$\mathcal{L}_{Id} = 1 - \frac{v_{res} \cdot v_{src}}{\left\|v_{res}\right\|_2 \left\|v_{src}\right\|_2} \tag{3}$$

where v_{res} and v_{src} denote the identity feature extracted by ArcFace from the generated result and source face, respectively.

2.4. Training

We follow the framework of adversarial generative learning for model training. Architecture of generator is described as Section 2.1 and Figure 1. And we adopt the multi-scale PatchGAN architecture [19] as the backbone of the visual discriminator. We apply LSGAN [20] adversarial loss $\mathcal{L}_{Adv}$ because it performs more stably during training. Besides, we use a Weak Feature Matching Loss [11] as shown below:

$$\mathcal{L}_{wFM}(D) = \sum_{i=m}^{M} \frac{1}{N_i} \left\| D^{(i)}\left(I_{gen}\right) - D^{(i)}\left(I_{tgt}\right) \right\|_1, \tag{4}$$

which can reduce the introduction of target texture information compared to the Feature Matching Loss [21]. And this helps the modification of identity information in target frames. Here m indicates the layer of discriminator where we start to calculate the loss.

During training, we select 5 random consecutive video frames from the dataset cropped according to faces as Ground Truth frames, and then select 5 random consecutive video frames from the same video as reference frames. After masking the lower half of Ground Truth frames, masked Ground Truth frames and reference frames are concatenated by color channel and fed into face encoder Enc_f. And the reference frames keep the same as the Ground Truth frames while inference. When batching the input frames in training, we train one batch that the source face maintains the same identity as the target video frame and another batch the source face with different identity from the target video frame. For the batches with source faces of the same identity, we use the following loss function:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{L1} + \lambda_i\mathcal{L}_{Id} + \lambda_s\mathcal{L}_{Sync} + \lambda_f\mathcal{L}_{wFm} + \mathcal{L}_{Adv} \tag{5}$$

where the λ_* are balancing coefficients. For the batches with source faces of the different identities, we remove the $\mathcal{L}_{L1}$ loss:

$$\mathcal{L}_{L1} = \frac{1}{N}\sum_{i=1}^{N} \left\|L_{gen} - L_{GT}\right\|_1 \tag{6}$$

where the $\mathcal{L}_{L1}$ loss means L1 loss between Ground Truth frames and generated frames at a pixel level.

3. EXPERIMENTS

3.1. Dataset

An issue in training a model capable of generating both lip-synchronized faces as well as identity feature swapping in an end-to-end manner is how to accommodate the different types of datasets required for the two tasks. The training data for talking face generation is generally from video datasets with continuous pictures, such as VoxCeleb2 [22] and LRS2 [23]. However, face swapping methods usually use face image databases with a lot of identities, such as VGGFace2-HQ [24, 25]. Thousands of identities are included in VGGFace2-HQ, and each identity has tens of photographs, but each photograph was taken in distinct settings.

We utilize the HDTF video dataset [14] for training due to its high-quality audio-video synchronization and its resolution similarity to VGGFace2-HQ after cropping face. The dataset contains about 300 identities, and each video has only one identity. We treat the frames of each video as different photos of the same person.

We divided the HDTF video dataset into training and test sets in the ratio of 9:1. Half of the videos in the test set have IDs from the training set but are not involved in training. The IDs of the other half are not in the training set. We shuffle the whole test set into pairs of target video and source face and use the original audio of the target video as model input. We select the first frame of the video as the source face. For fair comparison, Wav2Lip, SimSwap, and the proposed end-to-end model are in the 96-resolution version of the training set. And the test set keeps the same settings. We evaluated the effectiveness of the proposed method on the test set and compared it with the cascade method that SimSwap follows Wav2Lip.

The audios are pre-processed to 16kHz, then converted to Mel-spectrograms with FFT window size 800, hop length 200, and 80 Mel filter-banks. The target videos with cropped faces are in 25 frames-per-seconds and each video frame is corresponding to the audio with a 200ms window length.

Table 1. Quantitative evaluation on the test sets of videos.

Method	LSE-D ↓	LSE-C ↑	LMD ↓	LMD-m ↓
GT	7.19	7.74	–	–
Wav2Lip	6.02	8.65	**2.648**	**2.721**
Cascade	8.14	6.02	3.361	3.114
Proposed	**5.88**	**9.09**	3.161	2.984

3.2. Results

The results of the quantitative evaluation regarding lip-sync are shown in Table 1. LSE-D and LSE-C [2] are adopted for quantitative evaluation of lip-sync performance. Then we use LMD and LMD-m [26] to account for the accuracy of mouth shapes and lip sync. Our method performs better than the cascade method on all of these metrics. Notably, LSE-D and LSE-C of Wav2Lip and our proposed method outperform Ground Truth. This proves that their lip-sync results are comparable to the Ground Truth. The LMD and LMD-m of the cascade method and our method demonstrate a slight drop compared to Wav2Lip. This is in line with our assumption because the shape of the mouth and face are somewhat altered after identity modification as Figure 2 and Figure 3 show.

Table 2. Visual quality evaluation on the test sets of videos.

Method	FID ↓	CPBD ↑	ID-Retrieve ↑
Ground Truth	–	0.427	–
Wav2Lip	53.9	0.353	–
Cascade	74.8	0.340	**100.0**
Proposed	**49.9**	**0.470**	85.7

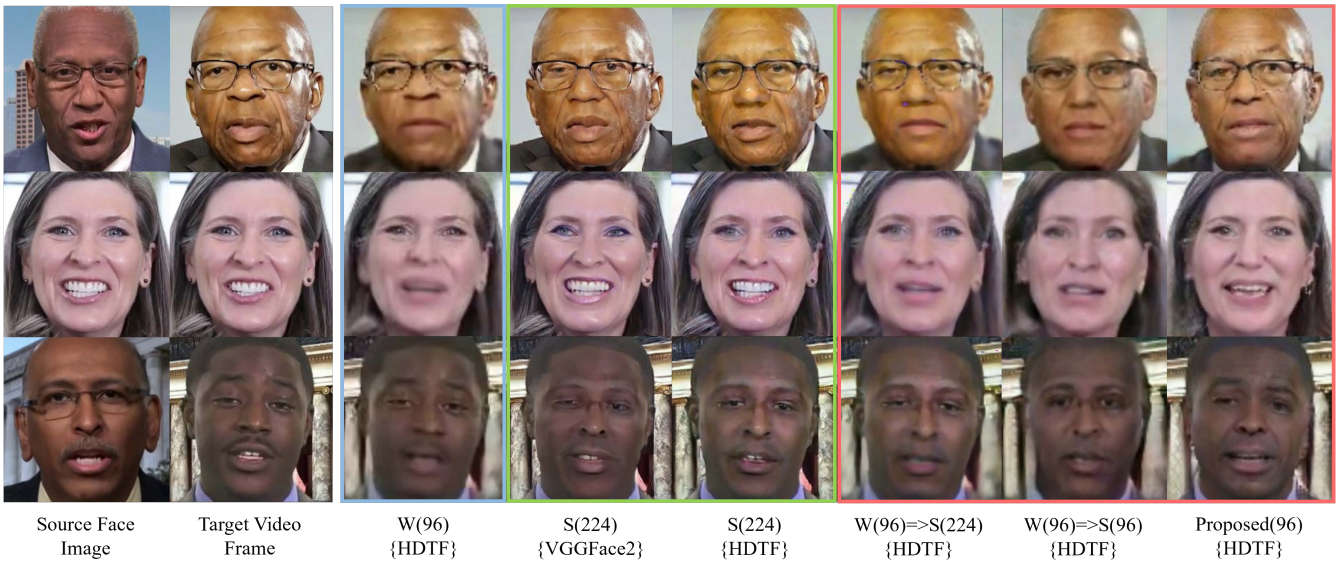

Fig. 2. The visual results of different methods according to single video frame. The identity information comes from the source face image, the mouth shape is synchronized with the original audio, and the other parts come from the given target video frame. The 'W' means **Wav2Lip**, and the 'S' means **SimSwap.** The numbers in the pointed brackets indicate the resolution of the training data. Words in parentheses designate the dataset used. The arrows indicate the direction of data transfer in the cascade method. **Observing the results in the red box**, we discover that our proposed method generates images with more facial texture information and mouth details.

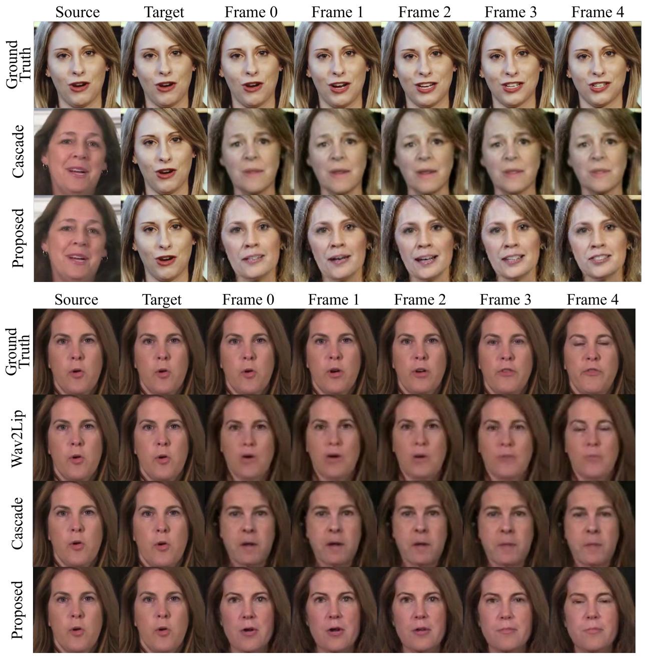

Fig. 3. Comparison between the cascade method and our proposed method. The mouth shapes generated by our method are more similar to Ground Truth than the cascade method. Moreover, compared to the blurred results generated by the cascade method and Wav2Lip, our approach preserves richer texture information.

We evaluated the visual quality of the generated frames. The cumulative probability blur detection (CPBD) [27] measure is adopted to evaluate the sharpness of the results. To measure the quality of the generated faces, we also report the Fréchet Inception Distance (FID) [28]. The results of Table 2 and Figure 2 provide supporting evidence that our proposed method can better preserve texture information. The cascade method performs worse than Wav2Lip on the metrics FID and CPBD, indicating that more image details are lost after face swapping. Observing the results in the colorful boxes of Figure 2, we discover that SimSwap alone can successfully switch identities while maintaining the facial texture. However, the cascade method results in blurry images because of quality loss introduced by Wav2Lip. Then ID-Retrieve is used to evaluate the performance of face swapping.

Table 3. The comparison of inference RTF. The evaluation is conducted on a single NVIDIA V100 GPU. RTF denotes the average time to generate one-second videos.

Method	RTF	Speedup
Cascade	0.2005 ± 0.0239	1.00×
Proposed	0.1024 ± 0.0168	**1.96**×

Additionally, we compare the average inference real time factor (RTF) of the cascade method and our proposed model as shown in Table 3. Each video for testing is around 10 seconds and we repeat the test procedure for 20 times. According to the results, our proposed end-to-end method is 1.96 times faster compared to the cascade method.

4. CONCLUSION

In this work, we propose an end-to-end model that combines the advantages of Wav2Lip and SimSwap and enables audio-driven face generation with customizable portraits. Our method achieves lip-sync accuracy comparable to real synced videos, preserves more texture details than the cascade method, and also alleviates the blurring of Wav2Lip. In addition, our approach streamlines process pipeline and improves the inference speed. Experiments demonstrate the effectiveness of the proposed method in both quality and efficiency.

5. ACKNOWLEDGE

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004) and Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030).

6. REFERENCES

[1] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo, "Fakeavceleb: a novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[2] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, p. 484–492, Association for Computing Machinery.

[3] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," in *International Conference on Learning Representations*, 2022.

[4] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[5] Yuanxun Lu, Jinxiang Chai, and Xun Cao, "Live speech portraits: real-time photorealistic talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–17, 2021.

[6] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European conference on computer vision*. Springer, 2020, pp. 716–731.

[7] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3867–3876.

[8] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[9] DeepFakes, "faceswap," https://github.com/deepfakes/faceswap, 2000.

[10] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji, "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *CoRR*, vol. abs/2106.09965, 2021.

[11] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge, "Simswap: An efficient framework for high fidelity face swapping," in *MM '20: The 28th ACM International Conference on Multimedia*, 2020.

[12] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[13] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang, "Smoothswap: A simple enhancement for face-swapping with smoothness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10779–10788.

[14] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[16] Joon Son Chung and Andrew Zisserman, "Out of time: Automated lip sync in the wild," *asian conference on computer vision*, 2016.

[17] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5074–5083.

[18] Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He, "Faceinpainter: High fidelity face adaptation to heterogeneous domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5089–5098.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[20] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[21] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[22] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[23] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[24] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[25] NNNNAI and neuralchen, "Vggface2-hq," https://github.com/NNNNAI/VGGFace2-HQ, 2 2021.

[26] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7832–7841.

[27] Niranjan D Narvekar and Lina J Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience*. IEEE, 2009, pp. 87–91.

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.