VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild

Kun Cheng*
Xidian University
Xi'an, China
kuncheng.xidian@gmail.com

Xiaodong Cun*
Tencent AI Lab
Shenzhen, China
vinthony@gmail.com

Yong Zhang
Tencent AI Lab
Shenzhen, China
zhangyong201303@gmail.com

Menghan Xia
Tencent AI Lab
Shenzhen, China
menghanxyz@gmail.com

Fei Yin
Tsinghua University
Shenzhen, China
yinf20@mails.tsinghua.edu.cn

Mingrui Zhu
Xidian University
Xi'an, China
mrzhu@xidian.edu.cn

Xuan Wang
Tencent AI Lab
Shenzhen, China
xwang.cv@gmail.com

Jue Wang
Tencent AI Lab
Shenzhen, China
arphid@gmail.com

Nannan Wang
Xidian University
Xi'an, China
nnwang@xidian.edu.cn

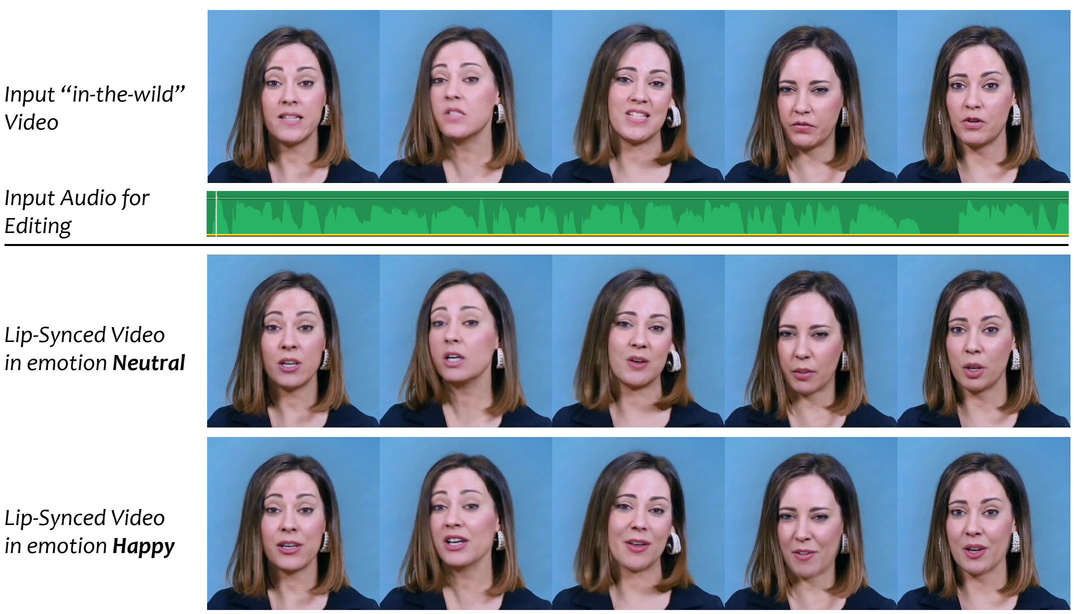

Figure 1: Given an arbitrary talking video and another audio, our method synthesizes photo-realistic talking videos with accurate lip-audio synchronization with retouched face expressions. Natural face © *European Central Bank* (CC BY).

*Both authors contributed equally to this research. Xiaodong Cun is the corresponding author. Project page: https://vinthony.github.io/video-retalking/

SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9470-3/22/12…$15.00
https://doi.org/10.1145/3550469.3555399

ABSTRACT

We present VideoReTalking, a new system to edit the faces of a real-world talking head video according to input audio, producing a high-quality and lip-syncing output video even with a different emotion. Our system disentangles this objective into three sequential tasks: (1) face video generation with a canonical expression; (2) audio-driven lip-sync; and (3) face enhancement for improving photo-realism. Given a talking-head video, we first modify the expression of each frame according to the same expression template using the expression editing network, resulting in a video with the canonical expression. This video, together with the given audio, is then fed

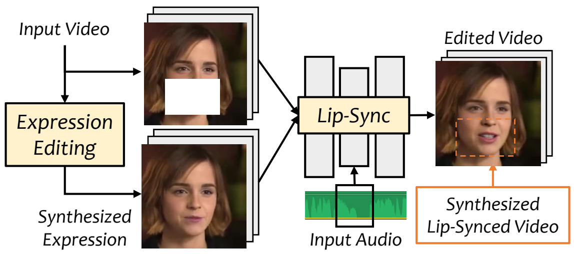

Figure 2: Our method modifies the original video and generates a lip-syncing video by an input audio through expression editing and lip-sync networks. Natural face © *ONU Brasil* (CC BY).

into the lip-sync network to generate a lip-syncing video. Finally, we improve the photo-realism of the synthesized faces through an identity-aware face enhancement network and post-processing. We use learning-based approaches for all three steps and all our modules can be tackled in a sequential pipeline without any user intervention. Furthermore, our system is a generic approach that does not need to be retrained to a specific person. Evaluations on two widely-used datasets and in-the-wild examples demonstrate the superiority of our framework over other state-of-the-art methods in terms of lip-sync accuracy and visual quality.

CCS CONCEPTS

• **Computing methodologies** → **Animation**.

KEYWORDS

Facial Animation, Video Synthesis, Audio-driven Generation

ACM Reference Format:
Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. 2022. VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3550469.3555399

1 INTRODUCTION

The task of editing a talking head video according to an input speech audio has important real-world applications, such as translating an entire video into a different language, or modifying the speech after video recording. Known as visual dubbing, this task has been studied in several prior works [Prajwal et al. 2020; Suwajanakorn et al. 2017; Thies et al. 2020; Wen et al. 2020], which edit the input talking-head video by modifying the facial animation and emotion to match the target audio, while leaving all the other motions unchanged (shown in Figure 1). Some methods [Suwajanakorn et al. 2017; Thies et al. 2020; Wen et al. 2020] can achieve satisfactory results on a specific speaker, but require training on the talking corpus of the target speaker to obtain a personalized model, which is not always available. On the other hand, the current generic methods produce blurry lower faces [Prajwal et al. 2020] or inaccurate lip synchronization [Song et al. 2022], which are visually intruding. These methods also do not support emotion editing, which is often desirable when changing the speech content.

Inspired by previous inpainting-based talking-head video editing approaches [Prajwal et al. 2020], we present a new system to edit the talking lips to match the input audio with more stable lip-sync results and better visual quality. Previous works consider the original frames in the video as the head pose references. However, we have found that lip generation is very sensitive to these references, and directly using original frames as basis for lip generation often produces out-of-sync results. To this end, as shown in Figure 2, we employ a divide-and-conquer strategy by neutralizing the facial expressions first, then use the modified frames as pose references for lip generation, which is more accurate given that all reference faces now have the same canonical expression. Finally, in contrast to previous works that often produce low-resolution and blurry results, we produce photo-realistic results via the proposed identity-aware enhancement network and the restorations [Wang et al. 2021c; Yang et al. 2021] based on StyleGAN's facial prior [Karras et al. 2019].

Specifically, given an arbitrary talking video, we first crop the face region and extract the pose and expression coefficients of the 3D Morphable Model (3DMM) by a deep neural network [Deng et al. 2019b]. We then use the parameters of the 3DMM with a standard neutral template expression and re-generate a video through the semantic-guided expression reenactment network similar to [Ren et al. 2021]. By doing so, we obtain a video with the same canonical expression across all the frames, and they will be considered as the structure references for our lip-sync network. Interestingly, we can also synthesize talking head videos with different emotions by changing the expression template. For example, by changing the lip shape of the expression template to match the "happy" emotion, this lip shape will be taken into account in the lip-sync network, causing the talking-head video exhibits the same emotion.

After expression neutralization, a lip-sync network is then applied to synthesize photo-realistic lower-half faces using the synthesized expression as the conditional structure information. Specifically, we design an hourglass-like network with the Fast Fourier Convolution block [Chi et al. 2020] as the basic learning unit, since it achieves great success in the general image inpainting task [Suvorov et al. 2021]. As for the audio injection, we use the Adaptive Instance Normalization (AdaIN) block [Huang and Belongie 2017] to modulate the audio features in global. Similar to [Prajwal et al. 2020], we use a pre-trained lip-sync discriminator to ensure the audio-visual synchronicity.

Although previous steps can synthesize talking-head videos with relatively accurate lip shapes, the visual quality is still limited by the low-resolution training datasets [Afouras et al. 2018; Nagrani et al. 2017]. To solve this problem, we design an identity-preserving face enhancement network to produce high-quality outputs by progressive training. The enhancement network is trained on an enhanced LRS2 dataset [Afouras et al. 2018] enhanced by the face restoration method [Yang et al. 2021]. We also apply the StyleGAN prior guided face restoration network [Wang et al. 2021c] to remove visual artifacts around the teeth.

All the above modules can be applied in sequential order without manual intervention or fine-tuning. We conduct extensive experiments to evaluate our framework on several existing benchmarks as well as in-the-wild videos. Results show that the proposed system can produce videos with much higher visual quality than previous methods while providing accurate lip synchronization.

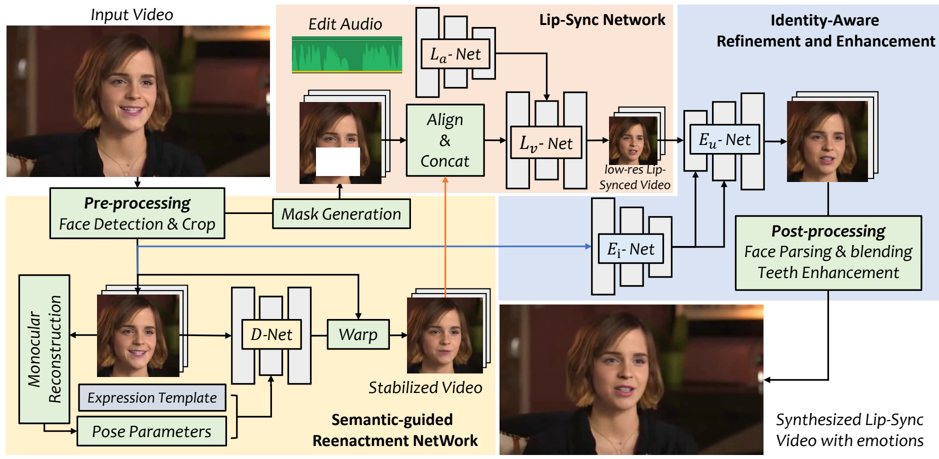

Figure 3: Our framework contains three main components for photo-realistic lip-sync video generation. Natural face © *ONU Brasil* (CC BY).

2 RELATED WORK

We review the related methods from two aspects, including the visual dubbing task which aims to edit the input video through audio, and single image animation using the audio as conditions.

2.1 Audio-based Dubbing in Video Editing

2.1.1 Arbitrary-subject methods. Arbitrary-subject methods aim at building a general model that does not need to be retrained for different identities. Speech2Vid [Chung et al. 2017] can re-dub a source video with a different segment of audio thanks to the context encoder. Reconstructing the lower-half face by inpainting is popular recently [KR et al. 2019; Park et al. 2022; Prajwal et al. 2020]. For example, LipGAN [KR et al. 2019] design a neural network to fill the lower-half face as a pose prior. Wav2Lip [Prajwal et al. 2020] extends LipGAN using a pre-trained SyncNet as the lip-sync discriminator [Chung and Zisserman 2016] to generate accurate lip synchronization. Based on Wav2Lip, SyncTalkFace [Park et al. 2022] involve the audio-lip memory to store the lip motion features implicitly and retrieve them at inference time. Another category of the methods predicts the intermediate representation first, and then, synthesizes the photo-realistic results by image-to-image translation networks, for example, the facial landmarks [Xie et al. 2021] and the facial landmarks based on 3D face reconstruction [Song et al. 2022]. However, all these methods are struggling to synthesize the high-quality results with editable emotion.

2.1.2 Personalized methods. Personalized visual dubbing is easier than the generic one, since these methods are limited to the certain person in the known environment. For example, SynthesizeObama [Suwajanakorn et al. 2017] can synthesize the mouth region of Obama by the audio-to-landmark network. Inspired by the face reenactment methods [Kim et al. 2018; Thies et al. 2019], recent visual dubbing methods focus on generating the intermediate representation from audio, and then, rendering the photo-realistic results by the image-to-image translation networks. For example, several works [Thies et al. 2020; Wen et al. 2020; Zhang et al. 2021b] focus on the expression coefficient from the audio features and render the photo-realistic results by the image generation networks [Kim et al. 2018; Thies et al. 2019; Wang et al. 2018]. Facial landmarks [Lu et al. 2021] and edges [Ji et al. 2021] are also popular choices by projecting the 3D rendered faces since it contains sparser information. Furthermore, 3D mesh-based [Lahiri et al. 2021] and NeRF [Mildenhall et al. 2020]-based methods [Guo et al. 2021] are also powerful. Although these methods can synthesize the photo-realistic results, they have relatively limited applications because they need to retrain the model on the specific person and environment.

2.2 Audio-based Single Image Facial Animation

Different from the visual dubbing, single image face animation aims to generate the animation by single driven audio, and it has also been influenced by the video-driven face animation. For example, [Song et al. 2018] generate the motion from audio using the recurrent neural network, [Zhou et al. 2019] disentangle the input to subject-related information and speech-related information by adversarial representation learning. [Vougioukas et al. 2020; Zhou et al. 2021] consider the audio as the latent code and drive the face animation by an image generator. The intermediate representation is also a popular choice in this task. ATVG [Chen et al. 2019] and MakeItTalk [Zhou et al. 2020] first generate the facial landmarks from audio, and then, render the video using a landmark-to-video network. Dense flow field is another active research direction [Siarohin et al. 2019; Yin et al. 2022]. [Zhang et al. 2021a] predict the 3DMM coefficients from audio and then transfer these parameters into a flow-based warping network. [Wang et al. 2021b,a] borrow the idea from video driven face animation [Siarohin et al. 2019].

3 FRAMEWORK

Technically, our method is a cross-modal video inpainting framework to fill the masked lower-half face under the guidance of the driven audio and the emotion-modulated reference frame. To this end, we design a lip-sync network (L-Net in Sec. 3.2), which uses the masked lower-half face frames, the given audio, and the original video frames as input to generate a lip-syncing video. However, there are two major problems if we use the L-Net only. The first

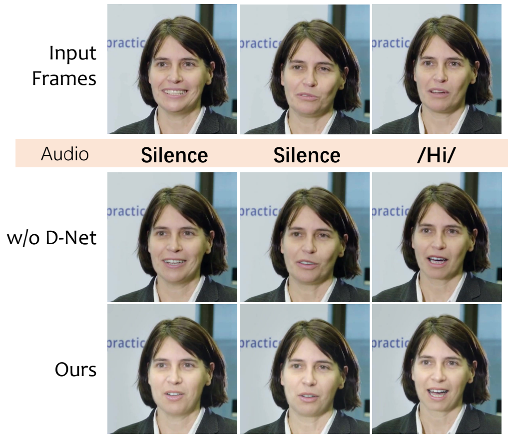

Figure 4: The proposed D-Net is used to remove the talking-related motions from the original video. W/o D-Net, the generated lip motion is heavily influenced by the source video and is still moving even when the audio is silent, indicating that information leakage affects lip synthesis. Natural face © *European Central Bank* (CC BY).

is the information leak caused by the reference frame, where the generated lip still relies on the reference heavily. The other is the low visual quality since current large-scale talking head datasets are in low resolution.

To this end, except L-Net, we propose two additional modules as shown in Figure 3. First, to solve the information leak, we generate a video with the frozen face expression by a semantic-guided expression reenactment network (D-Net in Sec. 3.1). The synthesized lips are the reference lips instead of the original ones. Then, the lower-half faces of the edited video will be used as a reference structure for our lip-synthesis network (L-Net). In L-Net, our method takes the audio as input and synthesizes the lip-sync results frame-wisely. Furthermore, we design an E-Net for the identity-aware face restoration in Sec. 3.3. Finally, we can paste the generated face back to the original video seamlessly through the post-processing in Sec. 3.4. Below, we give the details of each component.

3.1 Semantic-guided Reenactment Network

It is challenging to edit the lip-related motion in the video directly. Previous works often omit the original lip motion changes [Prajwal et al. 2020] or retiming the background [Song et al. 2022; Suwajanakorn et al. 2017] to avoid unnatural movements between the head pose and lip. Differently, we directly edit the whole lower-half face, including the facial movements with the help of a face reenactment method. Our key observation is that there is an information leak [KR et al. 2019; Prajwal et al. 2020] in conditional in-painting based methods if we use the original frame as the conditional image for lip synchronization. We give an example to show this phenomenon in Figure 4. Given the audio and the input frames, if we directly use the original frames as reference (w/o D-Net), the generated lips will be modified according to the original one. Thus, we aim at editing the expression of the whole lower-half face by the proposed semantic-guided reenactment network. Then, the frame with stable expression will serve as the reference for further lip synthesis.

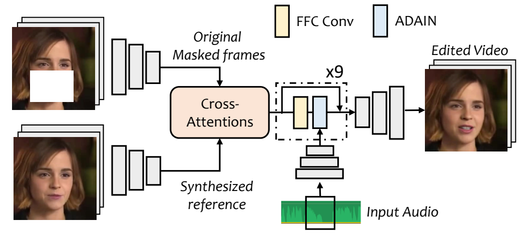

Figure 5: The detailed structure of the proposed L-Net. The skip-connections between the reference features and decoder are omitted for clarify. Natural face © *ONU Brasil* (CC BY).

As shown in Figure 3, after the face detection and crop, we extract the pose and expression coefficients from each frame using monocular face reconstruction [Deng et al. 2019b]. Then, we obtain the new driven signal by replacing the original expression coefficient with the pre-defined expression template. Thus, we can synthesize a video with the frozen expression via the produced dense warp fields of the network and the original frame. Similar to [Ren et al. 2021], the D-Net contains two encoder-decoder-like structures for coarse-to-fine training. After the expression editing, we get the stabilized expression across all the frames. Note that, since the quality of the face reenactment network is still limited, we use the edited face as the structure reference of our lip-sync network. To this end, we first detect facial landmarks, smooth them utilizing a temporal Savitzky–Golay filter, and then use the keypoints of the eye center and the nose as anchors for face alignment.

Interestingly, we can also utilize this information leak caused by the lip-sync reference frame through more expression templates (*e.g.* smile), resulting in an emotional talking-face video as shown in Figure 1. Since our expression reenactment network only edits the lower-half face of the original video, inspired by the facial action code system [Ekman and Friesen 1978], we can generate the talking faces in other emotions, *i.e.*, anger and surprise, via the image-based expression editing network [Pumarola et al. 2018] on the upper face. We consider it as a plugin and show some results in Sec.5.

3.2 Lip-Sync Network

Our lip-sync network (L-Net) is inspired by a recent conditional inpainting-based framework [Prajwal et al. 2020], which edits the original video directly through new audio. Differently, we use the pre-processed frames from D-Net as the identity and structure reference, the audio and the masked original frames as the condition, to synthesize the lip-syncing video with respect to the input audio.

In Figure 3, we give a brief overview of L-Net, which contains two sub-networks, L_a and L_v, for audio and video processing, respectively. Here, we give the detailed structure of L-Net in Figure 5. For the audio processing, we firstly extract the mel-spectrograms from the raw audio and use a ResNet-based encoder [He et al. 2016] to extract the global audio vector $F_{audio} \in \mathbb{R}^{256\times1\times1}$ of a time window. Following previous works, the time window is set to 0.2s per frame, causing the feature in the dimension of 80×16 to process. As for the image generation, we first extract the image features $F_{ref}, F_{orig} \in \mathbb{R}^{256\times H\times W}$ from the pre-processed referenced images and the original masked image by two different encoders respectively, then, these features are learned to model the relationship between pixels automatically via two cross-attention blocks [Vaswani et al. 2017]. These cross-attention blocks will calculate the pixel-wise corresponding matrix of two features and enlarge the reception fields. After that, we use nine residual Fast Fourier Convolutional blocks [Chi et al. 2020] to refine the features inspired by recent general image inpainting framework [Suvorov et al. 2021], and we inject the audio features by the AdaIN blocks [Huang and Belongie 2017] which normalize visual features channel-wise after each FFC block. Finally, a series of the convolutional up-sampling layers are used to generate the final results.

3.3 Identity-aware Enhancement Network

The result from L-Net is still unperfect since it is hard to train the model on high-resolution talking-head datasets. On the one hand, there is no public available large-scale high-resolution talking-head dataset. On the other hand, if we directly apply the GAN-prior based face restoration networks [Wang et al. 2021c; Yang et al. 2021] as the post-processing tools to improve the results, the results might not be perfect in terms of identity changes [Wang et al. 2021c] and blurry teeth and face [Yang et al. 2021] as shown in Figure 6.

To this end, we propose an identity-aware enhancement network inspired by recent image generation networks [Chan et al. 2021; Karras et al. 2020]. In detail, to acquire the high-resolution talking-head dataset and aligned domain for up-sampling, we enhance the low-resolution dataset firstly using a GAN prior-based face restoration network [Yang et al. 2021]. However, there is a domain gap between the enhanced high-resolution dataset during training and the blurry output of D-Net during testing. Then, to avoid this gap, we produce the low-resolution input of E-Net by feeding the enhanced frame and its corresponding audio to the L-Net. Ideally, L-Net should produce the same lip motions as the original frame using the conditional audio. Thus, we can use the high-resolution input as supervision directly. As for the architecture, we learn two style-based blocks [Karras et al. 2020] to up-sample the results four times and we design a ResBlock-based encoder E_i-Net to generate the identity-aware global modulation in each style block.

3.4 Post-processing

We also remove several artifacts when pasting back to the original video, including the artifacts of teeth generation and the synthesizing bounding box from the L-Net. Synthesizing the photo-realistic teeth for the face video is surprisingly hard [Suwajanakorn et al. 2017]. Unlike previous approach which uses the teeth proxy [Suwajanakorn et al. 2017], we seek help from the pre-trained face restoration network [Wang et al. 2021c] for teeth enhancement through face parsing [Yu et al. 2018]. As for the face bounding box caused by L-Net, we segment [Yu et al. 2018] the produced face and paste back to the original video using the Multi-band Laplacian Pyramids Blending [Burt and Adelson 1983].

4 TRAINING

Our framework is implemented using Pytorch [Paszke et al. 2019], and we train each module individually. After training, the whole framework can be tested in a sequence without manual intervention.

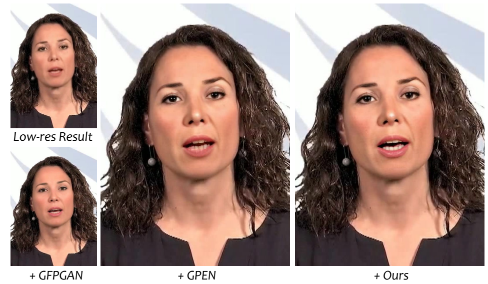

Figure 6: Comparison between different face restoration networks on the results, including GFPGAN [Wang et al. 2021c], GPEN [Yang et al. 2021], and our hybrid method. Note that, GFPGAN changes identity a lot. Natural face © *ONU Brasil* (CC BY).

Below, we give the dataset and training details of each module. More details can be found in supplementary material.

4.1 Training for each module

4.1.1 D-Net. To perform semantic-guided expression reenactment, we train our network on the VoxCeleb [Nagrani et al. 2017] dataset with the pose and expression from [Deng et al. 2019b]. This dataset contains 22496 talking head videos with diverse identities and head poses. We resize the input frames to 256×256 and train the network on the cropped faces similar to [Siarohin et al. 2019]. We train the network in 400k iterations using a progressive training setting. As for the loss function, we calculate the pixel-wise differences between the predicted image and the ground truth using perception loss [Zhang et al. 2018] and gram matrix loss [Gatys et al. 2016].

4.1.2 L-Net. We train the L-Net on the LRS2 [Afouras et al. 2018] dataset. This lip-reading dataset contains large-scale 160p videos from BBC programs. We pre-process the dataset using face detection [Bulat and Tzimiropoulos 2017] and resize the input image to 96×96 following the previous method [Prajwal et al. 2020]. We train the L-Net using perceptual loss and lip-sync discriminator for visual quality and audio-visual synchronization [Prajwal et al. 2020], respectively.

4.1.3 E-Net. The training process of E-Net is based on L-Net. We enhance the LRS2 dataset in advance to get a high-resolution dataset, and train the E-Net in 300k iterations. As for the loss function, E-Net is trained on the hybrid losses of perceptual loss [Johnson et al. 2016], pixel-wise L_1 loss, adversarial loss [Isola et al. 2017], lip-sync discriminator [Prajwal et al. 2020] and identity-loss using a pre-trained face recognition network [Deng et al. 2019a].

4.2 Evaluation

We evaluate the proposed method in terms of visual quality and lip-synchronization. As for the visual quality, since the ground-truth talking video is unavailable, we choose Fréchet inception distance (FID) [Heusel et al. 2017] and cumulative probability blur detection (CPBD) [Narvekar and Karam 2009] to evaluate the visual quality of generated videos. A lower FID score means that the generated images are closer to the dataset distribution. The CPBD reflects the sharpness of the results. Different from [Prajwal et al. 2020], we compute visual quality metrics on the full frames of the video

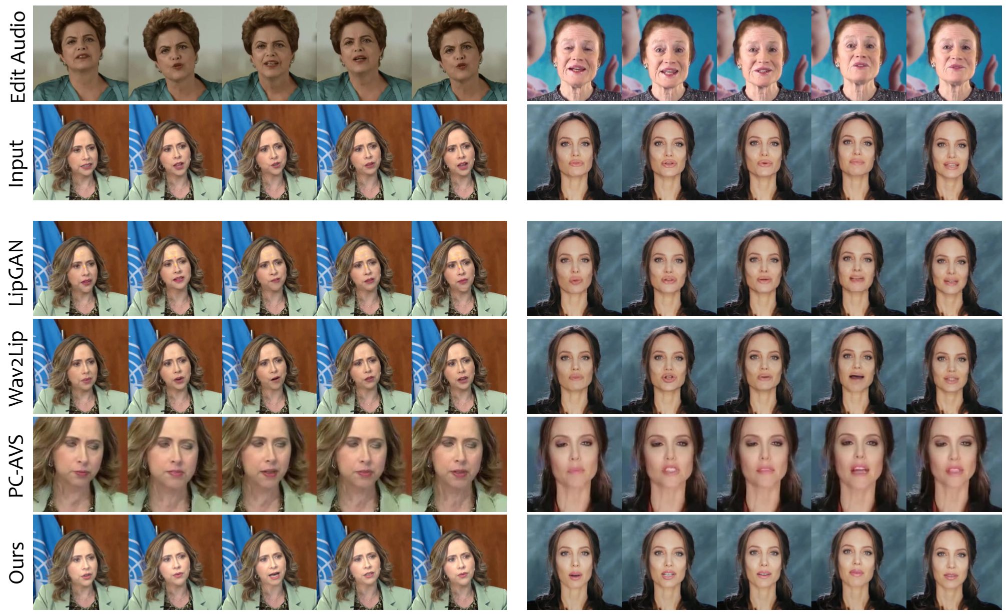

Figure 7: Qualitative comparison with LipGAN [KR et al. 2019], Wav2Lip [Prajwal et al. 2020], and PC-AVS [Zhou et al. 2021]. Above two rows show the edit audio and the input video frames, respectively. Note that, to visualize the input audio, we use the audio's corresponding face to show their mouth shapes. Natural face © *ONU Brasil* (CC BY).

Table 1: Quantitative results on LRS2 and HDTF datasets.

	LRS2 Dataset				HDTF Dataset			
	Visual Quality		Lip-Sync		Visual Quality		Lip-Sync	
	FID↓	CPBD↑	LSE-D↓	LSE-C↑	FID↓	CPBD↑	LSE-D↓	LSE-C↑
LipGAN [KR et al. 2019]	5.168	0.2615	9.609	3.062	7.684	0.2754	9.943	4.052
Wav2Lip w/o GAN [Prajwal et al. 2020]	5.069	0.2607	7.116	6.889	7.358	0.2764	**8.689**	**5.427**
Wav2Lip [Prajwal et al. 2020]	**3.911**	0.2714	7.191	6.870	5.632	0.2763	8.895	5.228
PC-AVS [Zhou et al. 2021]	12.800	0.2085	7.666	5.974	-	-	-	-
Ours	5.193	**0.2809**	**6.519**	**7.089**	**4.504**	**0.2903**	9.359	4.518

instead of cropped faces since we focus on the quality of the whole video. We choose the LSE-C and LSE-D [Prajwal et al. 2020] to evaluate the quality of lip synchronization. As for the dataset choices, we evaluate our framework on both low-resolution dataset (LRS2) and high-resolution dataset (HDTF). HDTF dataset contains 720p or 1080p videos from YouTube. Following the unpaired evaluating settings as described in [Prajwal et al. 2020], we take a video and an audio clip from the other different video to synthesize the results. We create 14*k* and 100 twenty-second audio-video pairs for LRS2 and HDTF dataset evaluation respectively.

5 RESULTS

5.1 Comparison with state-of-the-art Methods

We compare our method with three state-of-the-art methods under the same settings, including LipGAN [KR et al. 2019], Wav2Lip [Prajwal et al. 2020] and PC-AVS [Zhou et al. 2021]. LipGAN and Wav2Lip share the similar network structures. Differently, Wav2Lip uses a pre-trained lip-sync discriminator as the lip-expert, yet a better lip-sync performance. PC-AVS is originally proposed for one-shot pose-controllable talking-head generation. We use the identity code of each original video frame to replace the original single image face animation settings. We compare the proposed method with these methods using their open-sourced codes.

As shown in Table 1, the proposed method achieves much better visual qualities according to CPBD and FID. Since the LRS2 dataset is low-resolution and our method produces high-resolution results, the FID of Wav2Lip on the LRS2 dataset is better. As for the accuracy of lip-sync, our method still gets much better and comparable performance on these two datasets. We also show some examples in Figure 7 to perform the visual comparison. From this figure, our method produces high-quality results with more accurate lip-sync than previous methods. Since visual dubbing is a video editing task, we highly recommend the reader to compare our methods with others refer to the accompanying video.

For the comparison of the lip-sync quality, human evaluation is required. We perform a user study to further evaluate the performance of the proposed method. In the user study, we generate ten talking videos with different audio and video sources of our method and two state-of-the-art methods (LipGAN and Wav2Lip) on the HDTF dataset. We let the users show their opinions about each video in terms of the visual and lip-sync qualities. We set five different scores (larger is better, ranging from 1 to 5) for each option. Our form is sent to 51 people in total, getting 510 opinions. As shown in Table 2, most users prefer to give higher scores to our method with respect to the visual and lip-sync quality.

Table 2: User Study.

Method	Visual Quality↑	Lip-Sync Quality↑
LipGAN	2.867	3.058
Wav2Lip	3.173	3.398
Ours	**4.171**	**4.100**

5.2 Ablation Study

We mainly ablate three major components of our framework in Table 3. The first component is the cross-attention between two image encoders. L-Net w/o cross-attention in Table 3 means channel-wisely concatenating the features from the source and reference frames. We find cross-attention is helpful in terms of the lip-sync quality since it can capture the long-range dependencies. Besides the gains in numerical metrics, we also find it brings more vivid results (e.g, larger mouth). We then show the results of adding the E-Net in our framework. As we expected, the identity-aware face enhancement will hugely improve the visual quality. However, the additional artifacts will also influence the lip-sync quality. Finally, by using D-Net to stabilize reference frames, our framework generates better video in terms of visual and lip-sync quality.

Table 3: Major Ablation Studies on HDTF Dataset.

	Visual Quality		Lip-Sync Quality	
	FID↓	CPBD↑	LSE-D↓	LSE-C↑
L-Net w/o cross-att.	5.951	0.2743	9.788	4.164
L-Net	6.471	0.2755	9.578	4.382
L-Net + E-Net	**3.334**	0.2873	10.171	3.764
L-Net + E-Net + D-Net	4.504	**0.2903**	**9.359**	**4.518**

5.3 Extensions to Emotional Talking Video

We have already shown that the proposed method can be used for emotional talking-head video editing in Figure 1. Since our method only modifies the lower-half face, we also get inspiration from the facial action unit system [Ekman and Friesen 1978] and edit the upper face of the images using [Pumarola et al. 2018], causing different combinations as shown in Figure 8.

5.4 Limitation

Although the proposed method can work for the videos in the wild, it still contains some noticeable artifacts in some cases. As shown in Figure 9, one noticeable difference of the proposed framework will cause a slightly identity change from the original video due to the dense warping of D-Net. However, it is only one module

Figure 8: More emotional results using [Pumarola et al. 2018]. Natural face © *ONU Brasil* (CC BY).

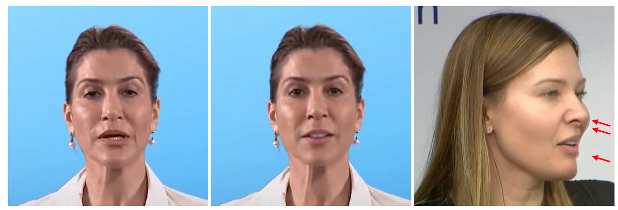

Figure 9: Failed cases on identity and extreme poses. Natural faces © *ONU Brasil* and © *European Central Bank* (CC BY).

of our method and we will replace it with another face reenactment network [Wang et al. 2021d] or 3D-based face reenactment method [Kim et al. 2018] directly. Our method also shows some artifacts in some extreme poses as shown in Figure 9. Since our method edits the video in a frame-by-frame fashion, the results may show some small temporal jittering and flashing.

6 CONCLUSION

We present a generic system for audio-based talking-head video editing by removing the lip motion first and then performing editing. As demonstrated, our framework can work on in-the-wild videos without fine-tuning and produce high-quality results using the audio as the condition. Besides, our system has the potential on the emotional talking-head generation for the lower-half face of the video. We will explore in the future to support more emotions and connect the source audio and contexts to the emotions.

Ethical Considerations. Since our system can edit the talking content of the video in the wild, we also consider the misuse of the proposed method. We will add both robust video and audio watermark to the produced video, and develop the tools to identify the trustworthiness. On the other hand, we hope our method can also help the research in the DeepFake detection.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202, in part by the National Natural Science Foundation of China under Grant 61922066, 61876142 and 62036007, in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15, in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01.

REFERENCES

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

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Peter J Burt and Edward H Adelson. 1983. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)* 2, 4 (1983), 217–236.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7832–7841.

Lu Chi, Borui Jiang, and Yadong Mu. 2020. Fast Fourier Convolution. In *NeurIPS*.

J. S. Chung, A. Jamaludin, and A. Zisserman. 2017. You said that?. In *British Machine Vision Conference*.

Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *ACCV*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019b. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *ICCV*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS* (2017).

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. 2021. Audio-Driven Emotional Video Portraits. *arXiv preprint arXiv:2104.07452* (2021).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*.

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *TOG* (2018).

Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1428–1436.

Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. 2021. LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video using Pose and Lighting Normalization. In *CVPR*.

Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation. *TOG* (2021).

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

Niranjan D Narvekar and Lina J Karam. 2009. 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, 87–91.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. 2022. SyncTalkFace: Talking Face Generation with Precise Lip-syncing via Audio-Lip Memory. In *36th AAAI Conference on Artificial Intelligence (AAAI 22)*. Association for the Advancement of Artificial Intelligence.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*. 818–833.

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *ICCV*.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *NIPS* (2019).

Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security* (2022).

Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786* (2018).

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. *arXiv preprint arXiv:2109.07161* (2021).

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *TOG* (2017).

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *TOG* (2019).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision* 128, 5 (2020), 1398–1413.

Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. 2021b. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. *IJCAI* (2021).

Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. 2021a. One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning. *arXiv preprint arXiv:2112.02749* (2021).

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2021e. High-fidelity gan inversion for image attribute editing. *arXiv preprint arXiv:2109.06590* (2021).

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601* (2018).

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021d. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021c. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*.

Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. 2020. Photorealistic Audio-driven Video Portraits. *TVCG* (2020).

Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. 2021. Towards Realistic Visual Dubbing with Heterogeneous Sources. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1739–1747.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN. *arxiv:2203.04036* (2022).

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 325–341.

Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. 2021b. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3867–3876.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021a. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *CVPR*.

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeitTalk: speaker-aware talking-head animation. *TOG* (2020).