

Everybody’s Talkin’: Let Me Talk as You Want

Linsen Song^{1*} Wayne Wu^{2,3} Chen Qian² Ran He¹ Chen Change Loy³

¹NLPR, CASIA ²SenseTime Research ³Nanyang Technological University

songlinsen2018@ia.ac.cn {wuwenyan, qianchen}@sensetime.com
rhe@nlpr.ia.ac.cn ccloy@ntu.edu.sg



Figure 1: **Audio-based video editing.** Speech audio of an arbitrary speaker, extracted from any video, can be used to drive any videos featuring a random speaker.

Abstract

We present a method to edit a target portrait footage by taking a sequence of audio as input to synthesize a photo-realistic video. This method is unique because it is highly dynamic. It does not assume a person-specific rendering network yet capable of translating arbitrary source audio into arbitrary video output. Instead of learning a highly heterogeneous and nonlinear mapping from audio to the video directly, we first factorize each target video frame into orthogonal parameter spaces, i.e., expression, geometry, and pose, via monocular 3D face reconstruction. Next, a recurrent network is introduced to translate source audio into expression parameters that are primarily related to the audio content. The audio-translated expression parameters are then used to synthesize a photo-realistic human subject in each video frame, with the movement of the mouth regions precisely mapped to the source audio. The geometry and pose parameters of the target human portrait are re-

tained, therefore preserving the context of the original video footage. Finally, we introduce a novel video rendering network and a dynamic programming method to construct a temporally coherent and photo-realistic video. Extensive experiments demonstrate the superiority of our method over existing approaches. Our method is end-to-end learnable and robust to voice variations in the source audio. Some results are shown in Fig. 1. Video results are shown on our project page¹.

1. Introduction

I’m going where the sun keeps shining. Through the pouring rain. Going where the weather suits my clothes.

—————
Fred Neil, *Everybody’s Talkin’*

¹Project Page: <https://wywu.github.io/projects/EBT/EBT.html>

*This work was done during an internship at SenseTime Research.

Video portrait editing is a highly sought-after technique in view of its wide applications, such as filmmaking, video production, and telepresence. Commercial video editing applications, such as Adobe Premiere and Apple iMovie, are resource-intensive tools. Indeed, editing audio-visual content would require one or more footages to be reshot. Moreover, the quality of the edited video is highly dependent on the prowess of editors.

Audio-based approach is an attractive technique for automatic video portrait editing. Several methods [6, 62] are proposed to animate the mouth region of a still image to follow an audio speech. The result is an animated static image rather than a video, hence sacrificing realism. Audio-driven 3D head animation [46] is an audio-based approach but aiming at a different goal, namely to drive stylized 3D computer graphic avatars, rather than to generate a photo-realistic video. Suwanakorn *et al.* [45] attempted to synthesize photo-realistic videos driven by audio. While impressive performance was achieved, the method assumes the source audio and target video to come from the same identity. The method is only demonstrated on the audio tracks and videos of Barack Obama. Besides, it requires long hours of single-identity data (up to 17 hours [45]) for training using relatively controlled and high-quality shots.

In this paper, we investigate a learning-based framework that can perform many-to-many audio-to-video translation, *i.e.*, without assuming a single identity of source audio and the target video. We further assume a scarce number of target video available for training, *e.g.*, at most a 15-minute footage of a person is needed. Such assumptions make our problem *non-trivial*: 1) Without sufficient data, especially in the absence of source video, it is challenging to learn direct mapping from audio to video. 2) To apply the framework on arbitrary source audios and target videos, our method needs to cope with large audio-video variations between different subjects. 3) Without explicitly specifying scene geometry, materials, lighting, and dynamics, as in the case of a standard graphics rendering engine, it is hard for a learning-based framework to generate photo-realistic yet temporally coherent videos.

To overcome the aforementioned challenges, we propose to use the expression parameter space, rather than the full pixels, as the target space for audio-to-video mapping. This facilitates the learning of more effective mapping, since the expression is semantically more relevant to the audio source, compared to other orthogonal spaces, such as geometry and pose. In particular, we manipulate the expression of a target face by generating a new set of parameters through a novel LSTM-based Audio-to-Expression Translation Network. The newly generated expression parameters, combined with geometry and pose parameters of the target human portrait, allow us to reconstruct a 3D face mesh with the same identity and head pose of the target but with new expression (*i.e.*, lip movements) that matches the phonemes

of the source audio.

We further propose an Audio ID-Removing Network that keeps audio-to-expression translation agnostic to the identity of the source audio. Thus, the translation is robust to variations in the voices of different people in different source audio. Finally, we solve the difficult face generation problem as a face completion problem conditioned on facial landmarks. Specifically, after reconstructing a 3D face mesh with new expression parameters, we extract the associated 2D landmarks from the mouth region and represent them as heatmaps. These heatmaps are combined with target frames where the mouth region is masked. Taking the landmark heatmaps and the masked target frames as inputs, a video rendering network is then used to complete the mouth region of each frame guided by dynamics of the landmarks.

We summarize our contributions as follows: 1) We make the first attempt at formulating an end-to-end learnable framework that supports audio-based video portrait editing. We demonstrate coherent and photo-realistic results by focusing specifically on expression parameter space as the target space, from which source audios can be effectively translated into target videos. 2) We present an Audio ID-Removing Network that encourages an identity-agnostic audio-to-expression translation. This network allows our framework to cope with large variations in voices that are present in arbitrary audio sources. 3) We propose a Neural Video Rendering Network based on the notion of face completion with a masked face as input and mesh landmarks as conditions. This approach facilitates the generation of photo-realistic video for arbitrary people within one single network.

Ethical Considerations. Our method could contribute greatly towards advancement in video editing. We envisage relevant industries, such as filmmaking, video production, and telepresence to benefit immensely from this technique. We do acknowledge the potential of such forward-looking technology being misused or abused for various malevolent purposes, *e.g.*, aspersions, media manipulation, or dissemination of malicious propaganda. Therefore, we strongly advocate and support all safeguarding measures against such exploitative practices. We welcome enactment and enforcement of legislation to obligate all edited videos to be distinctly labeled as such, to mandate informed consent be collected from all subjects involved in the edited video, as well as to impose hefty levy on all law defaulters. Being at the forefront of developing creative and innovative technologies, we strive to develop methodologies to detect edited video as a countermeasure. We also encourage the public to serve as sentinels in reporting any suspicious-looking videos to the authority. Working in concert, we shall be able to promote cutting-edge and innovative technologies without compromising the personal interest of the general public.

2. Related Work

Audio-based Facial Animation. Driving a facial animation of a target 3D head model by input source audio learns to associate phonemes or speech features of source audio with visemes. Taylor *et al.* [47] propose to directly map phonemes of source audio to face rig. Afterward, many speech-driven methods have been shown superior to phoneme-driven methods under different 3D models, *e.g.* face rig [61, 12], face mesh [25], and expression blendshapes [37].

Compared to driving a 3D head model, driving a photo-realistic portrait video is much harder since speaker-specific appearance and head pose are crucial for the quality of the generated portrait video. Taylor *et al.* [46] present a sliding window neural network that maps speech feature window to visual feature window encoded by active appearance model (AAM) [10] parameters. To improve the visual quality, in several methods [23, 52, 60], a still face image is taken as a reference for video generation. However, the result is an animation of a still image rather than a natural video.

Recently, Suwajanakorn *et al.* [45] obtain the state-of-the-art result in synthesizing the Obama video portrait. However, it assumes the source and target to have the same identity and requires long hour of training data (up to 17 hours). Thus, it is not applicable in audio-based video editing that need to cope with different sources of voice and target actors. In addition, the target video data is relatively scarce. Fried *et al.* [13] proposed a method to edit a talking-head video based on its transcript to produce a realistic video. While it produce compelling results, a person-specific face rendering network need to be trained for each target person. Besides, it takes a long time for viseme search (up to 2 hours for 1-hour recording) and relies on phoneme, thus it cannot be scaled to different languages.

Video-based Facial Reenactment. It is inherently difficult to synthesize mouth movements based solely on speech audio. Therefore, many methods turn to learning mouth movements from videos comprising the same/intended speech content. [17, 55, 48, 27, 32]. From source portrait video, facial landmarks [17, 40] or expression parameters [55] are estimated to drive the target face image. In all of these methods [17, 40, 55], the generated portrait videos are frame-wise realistic but they suffer from poor temporal continuity. ReenactGAN [58] is the first end-to-end learnable video-based facial reenactment method. It introduces a notion of “boundary latent space” to perform many-to-one face reenactment. However, ReenactGAN needs person-specific transformers and decoders, which makes the model size increase linearly with person identities raising.

Many model-based methods [48, 27, 32] leverage a 3D head model to disentangle facial geometry, expression, and pose. Face2Face [48] transfers expressions in parameter space from the source to the target actor. To synthesize a

realistic target mouth region, the best mouth image of the target actor is retrieved and warped [48]. Kim *et al.* [27] present a method that transfers expression, pose parameters, and eye movement from the source to the target actor. These methods are person-specific [27] therefore rigid in practice and suffer audio-visual misalignment [48, 27, 32] therefore creating artifacts leading to unrealistic results.

Deep Generative Models. Inspired by the successful application of GAN [18] in image generation [41, 31, 63, 20, 57], many methods [6, 60, 62, 38, 27] leverage GAN to generate photo-realistic talking face images conditioned on coarse rendering image [27], fused audio, and image features [6, 62, 60]. Generative inpainting networks [34, 21, 29, 59] are capable of modifying image content by imposing guided object edges or semantic maps [59, 44]. We convert talking face generation into an inpainting problem of mouth region, since mouth movement is primarily induced by input speeches.

Monocular 3D Face Reconstruction. Reconstructing 3D face shape and texture from a single face image has extensive applications in face image manipulation and animation [14, 16, 48, 43, 45]. In general, monocular 3D face reconstruction produces facial shape, expression, texture, and pose parameters by solving a non-linear optimization problem constrained by a statistical linear model of facial shape and texture, such as Basel face model [35], FaceWarehouse model [5], and Flame [28]. Recently, different 3D head models have been increasingly applied in talking portrait video synthesis [48, 45, 27, 32, 13, 26].

3. Methodology

The architecture of the proposed method is shown in Fig. 2. First, we register a parametric 3D face model [5] in the target video, for every portrait video frame to extract face geometry, pose, and expression parameters. Then, the Audio-to-Expression Translation Network learns the mapping from the source audio feature to face expression parameters. We design an Audio ID-Removing Network to alleviate the issues on large variations caused by multiple speakers. Lastly, we formulate the talking face generation problem as a face completion problem guided by mouth region landmarks, in which the landmarks are projected from the restructured 3D facial mesh. We propose a Neural Video Rendering network to complete the mouth region of each frame, guided by the dynamics of the landmarks to generate a photo-realistic portrait video.

3.1. 3D Face Modeling

We leverage a parametric 3D face model [5] on portrait video frame to recover low dimensional geometry, expression, and pose parameters. To reduce parameter dimension, geometry and expression bases are computed based on high-quality head scans [2] and facial blendshapes [5, 1]

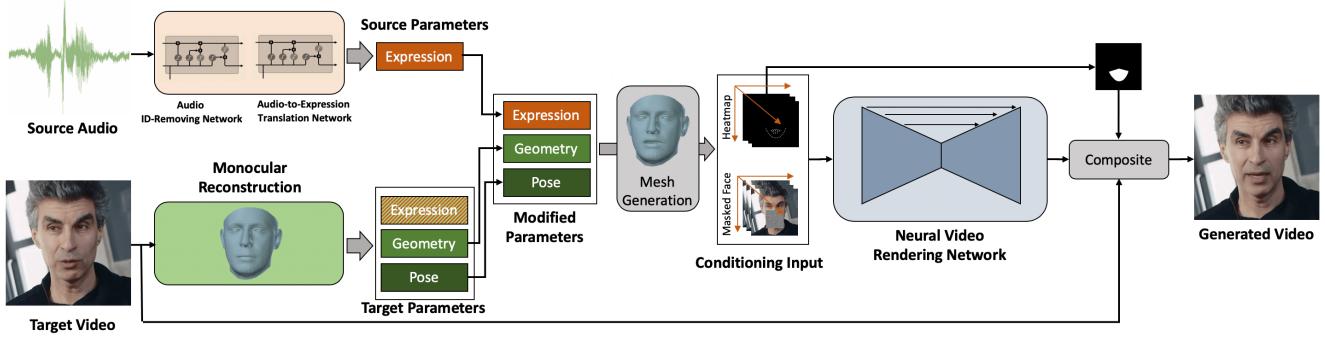


Figure 2: **Architecture.** Our network contains an Audio-to-Expression Translation Network that learns facial expression parameters from speech audio and a Neural Video Rendering Network that generates mouth region guided by projected mouth landmarks.

via principal component analysis (PCA). The geometry parameters $s \in \mathbb{R}^{199}$ and the expression parameters $e \in \mathbb{R}^{29}$ are the coefficients of geometry and expression principle components in the PCA, respectively. The pose of the head $p \in \mathbb{R}^6$ which contains 3 head rotation coefficients, 2 translation coefficients (x and y directions on the screen surface), and 1 scaling coefficient. All the parameters are computed by solving a non-linear optimization problem, constrained by the statistical linear 3D face model [2]. By optimizing the geometry, expression, and pose parameters of a given monocular face image based on its detected facial landmarks, portrait video frames will be automatically annotated with low dimensional vectors [2]. The recovered expression parameters are used as the learning target in the Audio-to-Expression Translation Network. Then, the recovered geometry and pose parameters, together with the expression parameters inferred by the Audio-to-Expression Translation Network, are employed for reconstructing the 3D facial mesh.

3.2. Audio-to-Expression Translation

3.2.1 Audio ID-Removing Network

We empirically find that identity information embedded in the speech feature degrades the performance of mapping speech to mouth movement. Inspired by recent advances of the speaker adaptation method in the literature of speech recognition [50, 39], we transfer the speech feature lies in different speaker domains onto a “global speaker” domain by applying a linear transformation, in the form [50]:

$$x' = W_i x + b_i = \bar{W}_i \bar{x}, \\ \text{where } \bar{W}_i = (W_i, b_i), \bar{x} = (x; 1), \\ \bar{W}_i = I + \sum_{j=1}^k \lambda_j \bar{W}^j \quad (1)$$

Here, x and x' represent the raw and transferred speech feature, respectively, while $\bar{W}_i = I + \sum_{j=1}^k \lambda_j \bar{W}^j$ represents the speaker-specific adaptation parameter that is factorized into an identity matrix I plus weighted sum of k components \bar{W}^j [39]. In speech recognition, these parameters are iteratively optimized by fMLLR [11, 15] and EM

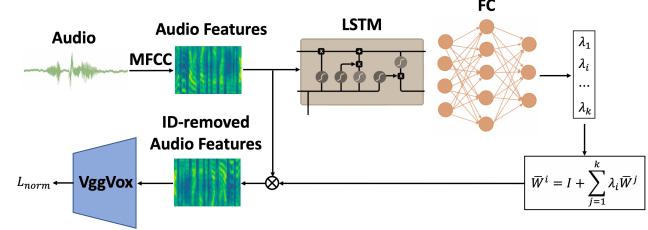


Figure 3: **Audio ID-Removing Network.** We formulate the speaker adaptation method from speech recognition [50, 39] as a neural network. The network removes identity in speech MFCC spectrum by transferring it to the “global speaker” domain.

algorithms. We formulate the above method into a neural network to be integrated with our end-to-end deep learning network.

From Eq.(1), the parameters λ_j need to be learned from the input speech feature, while the matrix components \bar{W}^j is general speech features of different speakers. Thus, we design an LSTM+FC network to infer λ_j from the input and set the matrix components \bar{W}^j as the optimizing parameter of the Audio ID-Removing Network. The matrix components \bar{W}^j of the Audio ID-Removing Network are updated by the gradient descent-based algorithm. The details of the network is depicted in Fig. 3. The output of the Audio ID-Removing Network is a new MFCC (Mel-frequency cepstral coefficients) spectrum. We apply a pre-trained speaker identity network VGGVox [33, 8] on the new MFCC spectrum and constrain the Audio ID-Removing Network by the following cross-entropy loss function:

$$\mathcal{L}_{norm} = - \sum_{c=1}^N \frac{1}{N} \log p(c|x'), \quad (2)$$

where N is the number of speakers, c is the speaker class label. The $p(c|x')$ is the probability of assigning MFCC x' to speaker c , which is inferred from the pre-trained VGGVox. Eq. (2) enforces the Audio ID-Removing Network to produce an MFCC spectrum that is not distinguishable by the pre-trained VGGVox.

3.2.2 Audio-to-Expression Translation Network

We formulate a simple but effective Audio-to-Expression Translation Network that learns the mapping from the ID-removed MFCC feature to the corresponding facial expression parameters. To infer the expression parameters at time t , the translation network observes a sliding window speech clip of 1 second, which contains 0.8 seconds before time t and 0.2 seconds after time t .

We empirically find it challenging to train a network to solely regress the expression parameters. The underlying reason could be that the expression parameters are defined and related to the 3DMM model that is hard to model by the network. To facilitate the learning, we introduce a shape constraint. In particular, with the predicted expression parameters from audio and the ground truth geometry/pose parameters of the video portrait, we can obtain a predicted reconstructed 3D facial mesh. Then, we project 3D points of mouth area to the 2D space to obtain the predicted 2D mouth landmarks. Using a similar method, we can obtain a set of ground-truth 2D mouth landmarks from the ground-truth expression parameters. The shape constraint can be introduced between the predicted 2D mouth landmarks and ground-truth 2D mouth landmarks. The whole process of generating mouth landmarks from expression parameters only involves linear operations and thus is differential. The loss function is written as follows:

$$\mathcal{L}_{trans} = \mathcal{L}_{exp} + \mathcal{L}_{shape} = \|\hat{e} - e\|_2 + \|\hat{l} - l\|_2, \quad (3)$$

where e and l are the ground truth expression and landmark, respectively, and \hat{e} and \hat{l} are the output expression and landmark of the translation network, respectively. The Audio ID-Removing and Audio-to-Expression Translation Networks are trained jointly, whose objective function is weighted sum of L_{norm} (Eq. (2)) and L_{trans} (Eq. (3)).

3.3. Neural Video Rendering Network

3.3.1 Network Architecture

Our final step is to generate photo-realistic talking face video that is conditioned on dynamic background portrait video and is guided by the mouth region landmark heatmap sequence. We design a completion-based generation network that completes the mouth region guided by mouth landmarks. First, to obtain the masked face images, a tailored dynamic programming based on retiming algorithm inspired by [45] is introduced to select frame sequence whose head shaking and blink of eyes look compatible with the source speech. Then, the mouth area that contains lip, jaw, and nasolabial folds are manually occluded by a square mask filled with random noise. To make the conversion from the landmark coordinates to heatmap differentiable, we follow [22, 56] to generate heatmaps with Gaussian-like functions centered at landmark locations. We mod-

ify a Unet [42, 40]-based network as our generation network. The employed skip-connection enables our network to transfer fine-scale structure information. In this way, the landmark heatmap at the input can directly guide the mouth region generation at the output, and the structure of the generated mouth obeys the heatmaps [40, 53].

We composite the generated mouth region over the target face frame according to the input mouth region mask. To obtain the mouth region mask, we connect the outermost mouth landmarks as a polygon and fill it with white color, then we erode the binary mask and smooth its boundaries with a Gaussian filter [26]. With the soft mask, we leverage Poisson blending [36] to achieve seamless blending. To improve the temporal continuity of generated video, we apply a sliding window on the input masked video frames and heatmaps [27, 26]. The input of the Neural Video Rendering Network is a tensor stacked by 7 RGB frames and 7 heatmap gray images [26]. It works well in most cases while a little lip motion jitters and appearance flicker might emerge in the final video. Then, a video temporal flicker removal algorithm improved from [3] is applied to eliminate these artifacts. Please refer to *appendix* for more details of the flicker removal algorithm.

3.3.2 Loss Functions

The loss function for training the Neural Video Rendering Network is written as follows:

$$\mathcal{L}_{render} = \mathcal{L}_{recon} + \mathcal{L}_{adv} + \mathcal{L}_{vgg} + \mathcal{L}_{tv} + \mathcal{L}_{gp}. \quad (4)$$

The reconstruction loss L_{recon} is the pixel-wise L1 loss between the ground truth and generated images. To improve the realism of the generated video, we apply the LSGAN [30] adversarial loss L_{adv} and add the gradient penalty term L_{gp} [19] for faster and more stable training. We also apply the perception loss L_{vgg} [24] to improve the quality of generated images by constraining the image features at different scales. The total variation regularization term L_{tv} is used to reduce spike artifact that usually occurs when L_{vgg} is applied [24]. The network is trained end-to-end with $L_{total} = L_{norm} + L_{trans} + L_{render}$ (Equations (2), (3), and (4)) with different coefficients. Due to the limited space, we report the details of the loss function, network architecture, and experimental settings in our *appendix*.

4. Results

We show qualitative and quantitative results on a variety of videos to demonstrate the superiority of our method over existing techniques and the effectiveness of proposed components.

Datasets. We evaluate our method on a talking face benchmark dataset GRID [9] and a speech video dataset we newly

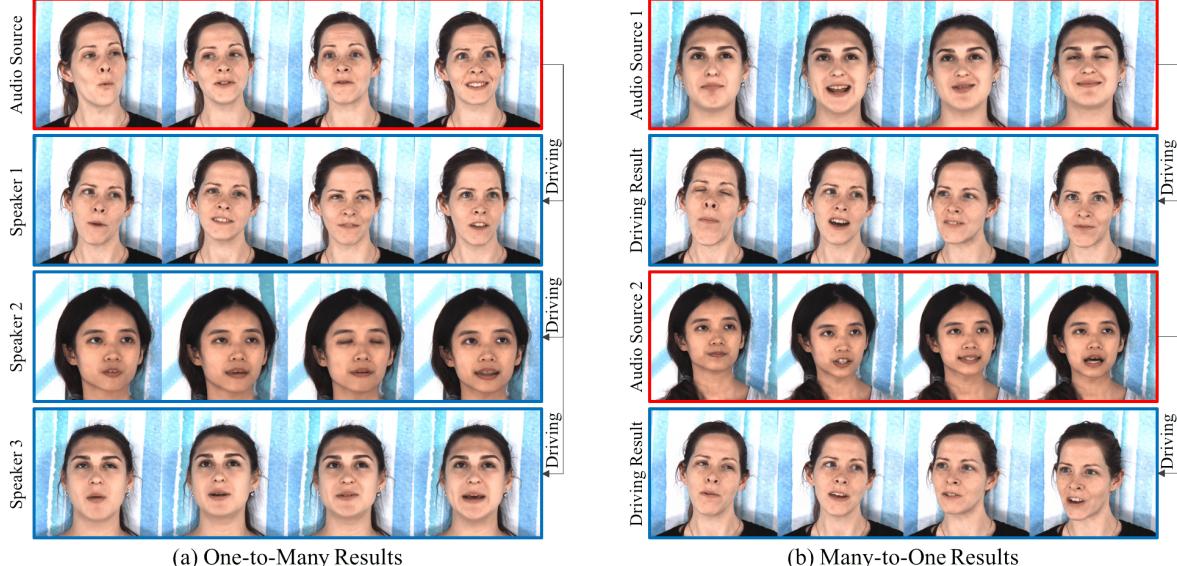


Figure 4: **Many-to-many results.** (a) One-to-many results: we use speech audio of 1 speaker to drive face of 3 different speakers. (b) Many-to-one results: we use speech audio of 2 different speakers to drive 1 same speaker. For video results we refer to the video on our project page.

collected. The former contains 1,000 sentences spoken by 18 males and 16 females. We follow Chen *et al.* [6] to split training and testing sets on the GRID dataset. Since GRID only provides frontal face videos of minor head movement, we record a video dataset that contains multiple head poses and time-varying head motion. The collected dataset contains speech videos of 4 speakers. Each speaker contributes 15 minutes video for training and 2 minutes video for testing, all videos are captured from 7 viewpoints to provide 7 head poses. Resolution of each video is 1920×1080 . We also take several videos downloaded from YouTube with the same percentage of the training and testing split of recorded data to evaluate our approach.

Evaluation Metrics. To evaluate the accuracy of the expression parameters and the projected landmarks under various head poses and motions, we apply the following distance metric:

$$E_{exp} = \frac{1}{N_{exp}} \sum_{i=1}^{N_{exp}} \|\hat{e}(i) - e(i)\|_2, \quad (5)$$

$$E_{ldmk} = \frac{1}{N_{ldmk}} \sum_{i=1}^{N_{ldmk}} \|\hat{l}(i) - l(i)\|_2,$$

where N_{ldmk} and N_{exp} are the number of landmarks and expression parameters respectively and i is the index of landmarks or expression parameters. To quantitatively evaluate the generated quality of portrait videos, we apply common image quality metrics like PSNR [54] and SSIM [54]. To qualitatively evaluate the generated quality of portrait videos, we conduct a user study in Section 4.4 and demonstrate some generated video results on our project page.

4.1. Audio-to-Video Translation

Many-to-Many Results. To prove that the audio-to-expression network is capable of handling various speakers and the face completion network is generalized on multiple speakers, we present one-to-many results and many-to-one

results in Fig. 4 and on our project page. In the one-to-many results, we use the speech audio of one speaker to drive different speakers. Note that different speakers share a single generator instead of multiple person-specific generators. In the many-to-one results, we use the speech audio of different speakers to drive the same speaker. This is in contrast to recent methods, where the whole pipeline [45] or part of components [26, 13] is designed for a specific person, which disables these methods in handling different voice timbres and facial appearances.

Large Pose Results. The main purpose of leveraging 3D face model is to handle head pose variations in generating talking face videos. As far as we know, majority of the recent audio-driving methods focus on generating frontal face video no matter whether a 3D head model is applied [45, 26] or not [51, 61, 7]. Our method, however, can generate portrait videos under various large poses driven by audio input. Thanks to the decomposition of audio-driving facial animation problem in our framework, which makes the audio only relate to expression parameters of face rather than shape or pose parameters. Results are shown in Fig. 5 and the video on our project page. Note that in previous methods [45, 61, 7], they directly learn a mapping from audio to landmarks, which involves the shape and pose information that is actually independent to the input audio.

Audio Editing & Singing Results. Our method can also be used to edit the speech contents of a pre-recorded video by splitting and recombining the words or sentences taken from any source audio. We show our audio editing results in Fig. 6 and video on our project page. In addition, we also ask a person to record singing and the audio is fed into our network. The driving result can be viewed in Fig. 7 and video on our project page. This demonstrates the general-



Figure 5: **Large pose results.** We demonstrate 4 head poses including up, down, right and left. Video results of all of 7 head poses can be viewed the video on our project page.

ization capability of our method and its potential in more complex audio-to-video tasks.



Figure 6: **Audio editing.** We select “knowledge is” and “virtue” from “Knowledge is one thing, virtue is another” in the source audio, then recombine them as “Knowledge is virtue” as input.



Figure 7: **Singing.** We evaluate our network on the singing audio clips. Video result can be viewed in the video on our project page.

4.2. Comparison with State-of-the-Art

We compare our method with the recent state-of-the-art portrait video generation methods, *e.g.*, Audio2Obama [45], Face2Face [48], DVP (Deep Video Portrait) [27] and Text-based Editing (TBE) [13]. The comparative results are demonstrated in Fig. 8 and video on our project page.

First, the Audio2Obama [45] combines a weighted median texture for synthesizing lower face texture and a teeth proxy for capturing teeth sharp details. Our GAN-based rendering network generates better texture details compare to the weighted median texture synthesis [45], *e.g.*, nasolabial folds (Fig. 8 (a)). Then, we compare our method to Face2Face [48] that supports talking face generation driving by source video in Fig. 8 (b). Face2Face [48] directly

transfers facial expression of source video in the parameter space while our method infers facial expression from source audio. The similar lip movement of Face2Face and our method in Fig. 8 (b) suggests the effectiveness of our Audio-to-Expression Translation Network in learning accurate lip movement from speech audio. Moreover, our GAN-based rendering network generates better texture details, such as mouth corners and nasolabial folds. We also compare to another video-driving method DVP [27] that supports talking face generation (Fig. 8 (c)). In DVP, a rendering-to-video translation network is designed to synthesize the whole frame other than the face region. It avoids the blending of face region and background that might be easily detectable. The DVP might fail in a complex and dynamic background as shown in Fig. 8 (c). In contrast, our method uses the original background and achieves seamless blending that is hard to distinguish. Finally, we compare our method with the contemporary text-based talking face editing method TBE [13] in Fig. 8 (d). In TBE, the mouth region is searched by phoneme and a semi-parametric inpainting network is proposed to inpaint the seam between the retrieved mouth and the original face background. This method requires training of a person-specific network per input video while our method can generalize on multiple speakers and head poses. Besides, our generation network produces competitive mouth details as shown in Fig. 8 (d).

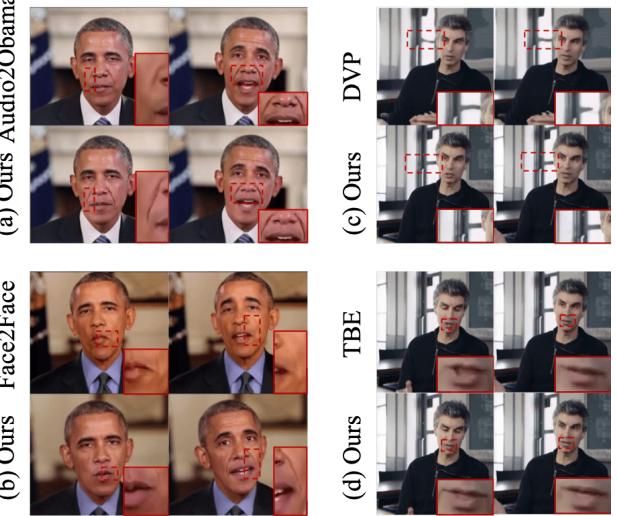


Figure 8: **Comparison to state-of-the-art methods.** Comparison between our method with Audio2Obama [45], Face2Face [48], DVP [27], and TBE [13].

4.3. Ablation Study

Evaluation of Parameter Regression. To prove the superiority of incorporating the 3D face model, we compare our network with the one that replaces Audio-to-Expression Translation Network with an Audio-to-Landmark Trans-

lation Network as performed in [45]. The Audio-to-Landmark Translation Network modifies the last fully connected layer of the Audio-to-Expression Translation Network so that its output dimension is the coordinate number of mouth region landmarks. The visualized comparison can be viewed in the video on our project page and Fig. 9 (a). We also compare the quantitative metric on GRID and collected dataset as shown in Tab. 1. In the collected dataset that contains more head motion and poses, our method achieves better lip synchronization results as the mouth generated by the one that applies Audio-to-Landmark Translation Network does not even open.

Table 1: **2D vs 3D quantitative comparison.** E_{exp} , E_{ldmk} , PSNR, and SSIM comparison of 2D and 3D parameter regression.

Parameter Regression	E_{exp}	E_{ldmk}	PSNR	SSIM
GRID	2D	-	3.99	28.06 0.89
	3D	0.65 2.24	31.19 0.95	
Collected	2D	-	3.13	26.76 0.93
	3D	0.595 1.82	29.16 0.95	

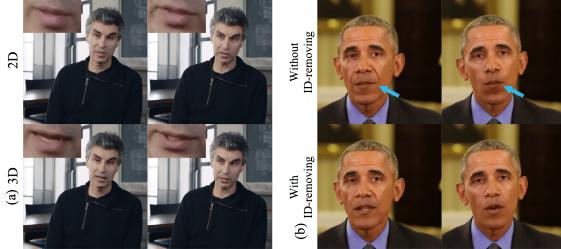


Figure 9: (a) **2D vs 3D qualitative comparison.** 3D parameter regression outperforms 2D parameter regression for head motion and poses. (b) **ID-removing qualitative comparison.** Improvement can be observed on a failure case caused by not applying Audio ID-Removing Network.

Table 2: **ID-removing quantitative comparison.** “BL” is the 3D parameter regression baseline; “IR” is “id-removing”.

Method	E_{exp}	E_{ldmk}	PSNR	SSIM
GRID	BL	0.84	3.09	30.88 0.94
	+IR	0.65 2.24	32.23 0.97	
Collected	BL	0.68	1.92	26.86 0.95
	+IR	0.59 1.83	31.21 0.96	

Evaluation of ID-removing. Our Audio ID-Removing Network transfers the speech feature of different speakers to a “global speaker”, which can be directly proven by tSNE [49] maps in *supplementary materials*. We also demonstrate the lip synchronization improvement in Fig. 9 (b) and the video on our project page. Quantitative metrics on GRID and collected dataset also validate its effectiveness as shown in Tab. 6.

Evaluation of Completion-Based Generation. We evaluate the effects of the proposed completion-based generation that benefits from jointly training on data of different people. As shown in Tab. 3, jointly training completion-based generators outperform separately training person-specific generators with much fewer network parameters when the number of speakers increases, regardless of the time length of the training data.

Table 3: **Training data size and training style.** PSNR/SSIM of different amount of training data and training style. The “Joint” means jointly training one generator for all speakers and “Split” means separately training multiple person-specific generators. One generation network contains 75 million parameters and $N(N = 4$ in the table) is the speaker number.

Time	2 mins	5 mins	10 mins	15 mins
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Split ($N \times 75M$)	29.621/0.868	29.338/0.875	29.487/0.849	29.650/0.876
Joint (75M)	30.421/0.886	30.664/0.888	30.787/0.892	31.072/0.897

4.4. User Study

To quantitatively evaluate the visual quality of generated portrait videos, following [13], we conduct a web-based user study involving 100 participants on the collected dataset. The study includes 3 generated video clips for each of the 7 cameras and for each of the 4 speakers, hence a total of 84 video clips. Similarly, we also collect 84 ground truth video clips and mix them up with the generated video clips to perform the user study. We separately calculate the study results of the generated and ground truth video clips.

In the user study, all the $84 \times 2 = 168$ video clips are randomly shown to the participants and they are asked to evaluate its realism by evaluating if the clips are real on a likert scale of 1-5 (5-absolutely real, 4-real, 3-hard to judge, 2-fake, 1-absolutely fake) [13]. As shown in Tab. 4, the generated and the ground truth video clips are rated as “real”(score 4 and 5) in 55.0% and 70.1% cases, respectively. Since humans are highly tuned to the slight audio-video misalignment and generation flaws, the user study results demonstrate that our method can generate deceptive audio-video content for large poses in most cases.

Table 4: **User study.** User study results on generated and ground truth video clips for videos of 7 poses.

score	Generated Videos					Ground Truth Videos						
	1	2	3	4	5	“real”	1	2	3	4	5	“real”
front	5.2	8.5	20.6	42.6	23.2	65.8%	0.6	12.1	10.1	29.7	47.5	77.2%
up-30°	4.6	25.0	14.2	36.9	19.3	56.3%	0.8	13.5	13.2	29.1	43.4	72.5%
down-30°	4.8	22.0	15.2	39.7	18.3	58.0%	0.9	13.6	14.1	30.2	41.3	71.5%
right-30°	3.9	22.9	15.8	42.1	15.3	57.4%	1.3	15.6	14.4	29.2	39.6	68.8%
right-60°	7.3	33.8	11.4	36.9	10.6	47.5%	0.8	17.8	15.1	29.1	37.2	66.3%
left-30°	3.2	20.9	21.9	40.1	13.9	54.0%	1.1	12.8	16.1	31.5	38.6	70.1%
left-60°	7.9	33.7	12.1	35.1	11.3	46.3%	0.7	17.5	14.1	27.2	40.5	67.7%
all poses	5.3	23.8	15.9	39.0	16.0	55.0%	0.9	14.7	13.9	29.4	41.2	70.6%

5. Conclusion

In this work, we present the first end-to-end learnable audio-based video editing method. At the core of our approach is the learning from audio to expression space bypassing the highly nonlinearity of directly mapping audio source to target video. Audio ID-Removing Network and Neural Video Rendering Network are introduced to enable generation of photo-realistic videos given arbitrary targets and audio sources. Extensive experiments demonstrate the robustness of our method and the effectiveness of each pivotal component. We believe our approach is a step forward towards solving the important problem of audio-based video editing and we hope it will inspire more researches in this direction.

References

- [1] Oleg Alexander, M. Rogers, W. Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul E. Debevec. The digital emily project: Achieving a photorealistic digital actor. *CG&A*, 30:20–31, 2010. [3](#)
- [2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. [3](#), [4](#)
- [3] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM TOG*, 34:196, 2015. [5](#)
- [4] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *TOG*, 34:196, 2015. [11](#)
- [5] Chen Cao, Yanlin Weng, Shun Zhou, Yiyi Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20:413–425, 2014. [3](#)
- [6] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *ECCV*, 2018. [2](#), [3](#), [6](#), [11](#), [12](#)
- [7] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. [6](#)
- [8] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018. [4](#)
- [9] Martin Cooke, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *JASA*, 2006. [5](#)
- [10] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *ECCV*, 1998. [3](#)
- [11] Vassilios V. Digalakis, Dimitry Ristichev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3:357–366, 1995. [4](#)
- [12] Peter Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM TOG*, 35:127, 2016. [3](#)
- [13] Ohad Fried, Maneesh Agrawala, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, and Christian Theobalt. Text-based editing of talking-head video. *ACM TOG*, 38:68, 2019. [3](#), [6](#), [7](#), [8](#)
- [14] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul E. Debevec. Driving high-resolution facial scans with video performance capture. *ACM TOG*, 34:8, 2014. [3](#)
- [15] Mark JF Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech & Language*, 12(2):75–98, 1998. [4](#)
- [16] Pablo Garrido, Levi Valgaerts, H. Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *CGF*, 34:193–204, 2015. [3](#)
- [17] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM TOG*, 37:231, 2018. [3](#)
- [18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. [3](#)
- [19] Ishaaq Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. [5](#)
- [20] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017. [3](#)
- [21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM TOG*, 36:107, 2017. [3](#)
- [22] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. In *NeurIPS*, 2018. [5](#)
- [23] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *IJCV*, pages 1–13, 2019. [3](#), [11](#), [12](#)
- [24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. [5](#)
- [25] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM TOG*, 36:94, 2017. [3](#)
- [26] Hyeyoungwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM TOG*, 38:178, 2019. [3](#), [5](#), [6](#)
- [27] Hyeyoungwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37:163, 2018. [3](#), [5](#), [7](#)
- [28] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 36:194, 2017. [3](#)
- [29] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. [3](#)
- [30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. [5](#)

- [31] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2016. 3
- [32] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM TOG*, 37:258, 2018. 3
- [33] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Interspeech*, 2017. 4
- [34] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CVPR*, pages 2536–2544, 2016. 3
- [35] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. *AVSS*, pages 296–301, 2009. 3
- [36] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM TOG*, 22:313–318, 2003. 5
- [37] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *CVPRW*, 2017. 3
- [38] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. *CoRR*, abs/1803.07716, 2018. 3
- [39] Daniel Povey and Kaisheng Yao. A basis representation of constrained mllr transforms for robust adaptation. *Computer Speech & Language*, 26:35–51, 2012. 4
- [40] Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. Make a face: Towards arbitrary high fidelity face manipulation. In *ICCV*, 2019. 3, 5
- [41] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 3
- [42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5
- [43] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, 2016. 3
- [44] Linsen Song, Jie Cao, Linxiao Song, Yibo Hu, and Ran He. Geometry-aware face completion and editing. In *AAAI*, 2018. 3
- [45] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM TOG*, 36:95, 2017. 2, 3, 5, 6, 7, 8, 13
- [46] Sarah Taylor, Akihiro Kato, Iain A. Matthews, and Ben P. Milner. Audio-to-visual speech conversion using deep neural networks. In *Interspeech*, 2016. 2, 3
- [47] Sarah L. Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica K. Hodgins, and Iain A. Matthews. A deep learning approach for generalized speech animation. *ACM TOG*, 36:93, 2017. 3
- [48] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 3, 7
- [49] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008. 8, 12
- [50] Karthik Visweswarah, Vaibhava Goel, and Ramesh Gopinath. Structuring linear transforms for adaptation using training time information. In *ICASSP*, 2002. 4
- [51] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016. 6, 11, 12
- [52] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. In *BMVC*, 2018. 3
- [53] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter. M. Hall, and Shi-Min Hu. Example-guided style consistent image synthesis from semantic labeling. In *CVPR*, 2019. 5
- [54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 6
- [55] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. 3
- [56] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Disentangling content and style via unsupervised geometry distillation. In *ICLRW*, 2019. 5
- [57] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019. 3
- [58] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018. 3
- [59] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. *CoRR*, abs/1806.03589, 2018. 3
- [60] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2018. 3
- [61] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemnet: Audio-driven animator-centric speech animation. *ACM TOG*, 37:161, 2018. 3, 6
- [62] Hao Zhu, Aihua Zheng, Huaibo Huang, and Ran He. High-resolution talking face generation via mutual information approximation. *CoRR*, abs/1812.06589, 2018. 2, 3, 11
- [63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3

Appendix

A. Details of Audio-to-Expression Translation Network

The network architecture of our Audio-to-Expression Translation Network can be viewed in Figure 10. In the training phase, we use paired audio and video frames from the training footage as network input. Ground truth facial shape, expression and pose parameters are calculated from video frame by monocular reconstruction. From the input audio, our Audio-to-Expression Translation Network infers predicted expression parameters that are supervised by the ground truth expression parameters. The loss function is $L_{exp} = \|\hat{e} - e\|_2$ in Eq. 6. We reconstruct facial 3D mesh by predicted expression parameters and ground truth shape parameters, then we use the ground truth pose parameters to project 2D mouth landmarks that are supervised by the ground truth 2D mouth landmarks. The loss function is $L_{shape} = \|\hat{l} - l\|_2$ in Eq. 6. The ground truth 2D mouth landmarks are projected in a similar way where ground truth expression parameters are used. In the testing phase, the predicted expression parameters from the source audio together with ground truth shape and pose parameters from target video are used to estimate 2D mouth landmarks. The embedded mouth identity and head pose of estimated 2D mouth landmarks are the same as those of the target video while the mouth movement in accord with the source audio.

$$L_{trans} = L_{exp} + L_{shape} = \|\hat{e} - e\|_2 + \|\hat{l} - l\|_2 \quad (6)$$

B. Temporal Flicker Removal Algorithm

In our approach, talking face video is generated frame by frame and temporal information in video frames is only concerned in landmark estimation. During testing, we find the generated talking face videos demonstrate acceptable frame continuity even in the circumstance that the video temporal flicker removal algorithm is not applied. It is due to that audio series clips used to generate time-adjacent frames contain vast overlap in time. The remaining temporal flicker in the video can be attributed to two reasons: 1) The inferred mouth and jawline landmarks contain slight jitter. 2) Appearance flicker, especially color flicker exists in the video frames generated by the inpainting network. Based on the above analysis, our flicker removal algorithm contains two parts: mouth landmark motion smoothing and face appearance deflicker. Algorithm 1 demonstrates the mouth landmark motion smooth algorithm.

Algorithm 1 Mouth Landmark Smoothing Algorithm

```

Require:  $l_{t-1}$ : mouth and jawline landmarks at time  $t - 1$ ;  

 $l_t$ : mouth and jawline landmarks at time  $t$ ;  

 $d_{th}$ : mouth movement distance threshold,  $s$ : mouth  

movement smooth strength;  

Ensure:  $l'_t$ : smoothed mouth and jawline landmarks at time  

 $t$ ;  

get mouth center position  $c_t$  at time  $t$  from  $l_t$   

get mouth center position  $c_{t-1}$  at time  $t - 1$  from  $l_{t-1}$   

if  $\|c_t - c_{t-1}\|_2 > d_{th}$  then  

     $l'_t = l_t$   

else  

     $\alpha = \exp(-s\|c_t - c_{t-1}\|_2)$   

     $l'_t = \alpha l_{t-1} + (1 - \alpha)l_t$   

end if  

return  $l'_t$ 

```

The appearance deflicker algorithm is modified from [4]. We take mouth movement into consideration. If the mouth does not move, then the color flicker is more obvious, and then we increase the deflicker strength. We denote the generated frame and processed frame at time t as P_t and O_t , respectively. The mouth center moving distance between time $t - 1$ and t is denoted as d_t . The processed frame at t is written as:

$$\mathcal{F}(P_t) = \frac{4\pi^2 f^2 \mathcal{F}(P_t) + \lambda_t \mathcal{F}(\text{warp}(O_{t-1}))}{4\pi^2 f^2 + \lambda_t} \quad (7)$$

where $\lambda_t = \exp(-d_t)$. Here, \mathcal{F} is the Fourier transform and f means frequency. Function $\text{warp}(O_{t-1})$ uses optical flow from P_{t-1} to P_t to warp input frame O_{t-1} . Compared with [4], the weight of previous frame λ_t is measured by the strength of mouth motion instead of global frame consistency.

C. Other Experiments

C.1. Quantitative Comparison on GRID dataset

Our method mainly focuses on talking face video editing, which is different from the recent methods that generate full face from input audio and reference still face image [51, 23, 6, 62]. Here we quantitatively compare our method with these methods [51, 23, 6, 62] on image generation metrics. For a fair comparison, in our method, we do not apply any post-process and we also modify the input of the inpainting network to generate the full face other than the mouth region. Specifically, the original network input tensor is stacked by 7 RGB frames and 7 heatmap gray images (from time $t - 6$ to time t), we remove the RGB frame and heatmap gray image at time t and require the network to generate the complete frame image at time t . Table 5 demonstrates that our approach outperforms these methods in PSNR and achieves comparable performance in SSIM.

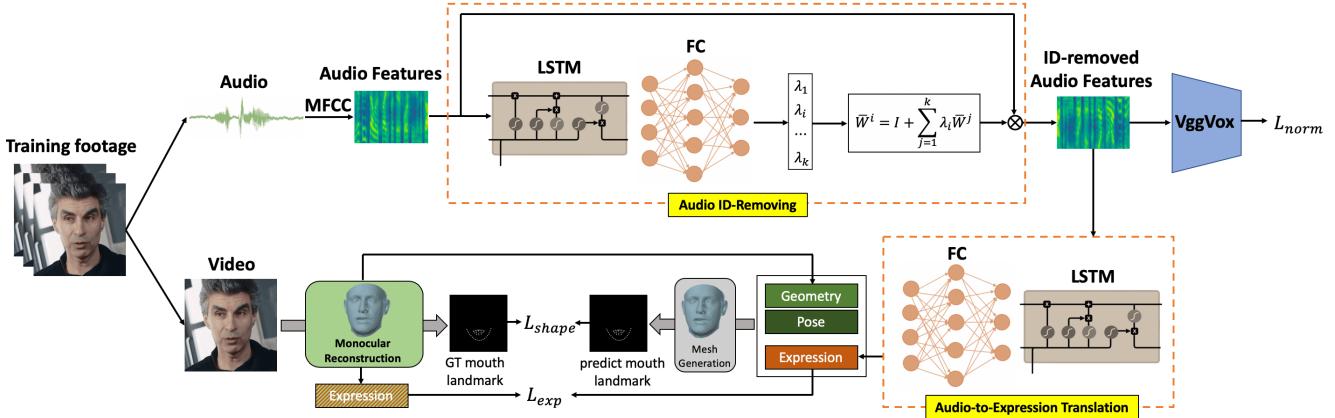


Figure 10: **Architecture of Audio-to-Expression Network.** The Audio ID-Removing Network eliminates identity information in speech audio. The Audio-to-Expression Translation Network estimate expression parameters from input audio. We constrain the predicted expression parameters and the projected 2D mouth landmark from the reconstructed facial mesh.

Table 5: SSIM, PSNR, IS and FID score comparison of our method and recent methods on GRID dataset. for fair comparison, we generate the full face and do not apply any post process.

Method	PSNR	SSIM	IS	FID
Vondrick <i>et al.</i> [51]	28.45	0.60	-	-
Jamaludin <i>et al.</i> [23]	29.36	0.74	-	-
Chen <i>et al.</i> [6]	29.89	0.73	-	-
Ours	30.01	0.94	23.53	9.01

C.2. Ablation Study on Temporal Flicker Removal Algorithm

The temporal flicker removal algorithm tries to smooth the output landmark coordinates and eliminate the appearance flicker. The quantitative improvement is slight as shown in Table 6 but the temporal continuity improvement is obvious as shown in the video on our project page, especially when the mouth does not open. We demonstrate the quantitative results of our 3D parameter regression baseline, Audio ID-Removing Network and Temporal Flicker Removal Algorithm in Table 6.

C.3. Audio ID-removing Effects in tSNE map

The tSNE [49] maps in Figure 11 demonstrate the 2D visualized distribution of the input MFCC spectrum and the identity removed MFCC spectrum produced by our Audio ID-Removing Network. We can see that the speaker identity can not be distinguished after removing identity in the MFCC spectrum.

D. Runtime Performance

We conduct the inference phrase on a commodity desktop computer with an NVIDIA GTX 1060 and an Intel Core

Table 6: **ID-removing & Deflicker quantitative comparison.** “BL” is the 3D parameter regression baseline; “IR” is “id-removing”; “DF” is “deflicker”. The metrics validate the effectiveness of the proposed components except for the deflicker algorithm. The deflicker algorithm mainly focus on removing temporal discontinuity that can be viewed in the video on our project page.

Method	E_{exp}	E_{ldmk}	PSNR	SSIM
GRID	BL	0.84	3.09	30.88 0.94
	+IR	0.65	2.24	32.23 0.97
	+DF	0.84	3.07	27.71 0.92
	+IR+DF	0.65	2.24	31.19 0.95
Collected	BL	0.68	1.92	26.86 0.95
	+IR	0.59	1.83	31.21 0.96
	+DF	0.68	1.92	27.46 0.93
	+IR+DF	0.59	1.82	29.16 0.95

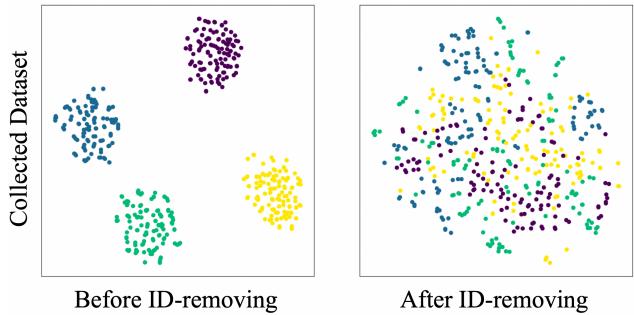


Figure 11: **tSNE before and after id-removing.** 2D visualized distributions of input MFCC and normalized MFCC. Different color represents different speaker. Our audio normalization erases the identity information embedded in the MFCC spectrum.

i7-8700. The audio to expression network takes 17 ms per frame and the inpainting network takes 77 ms per frame. The post processes including deflicker and teeth proxy take 1.3s and 300 ms per frame respectively. The deflicker al-

gorithm involves the calculation of optical flow that dominates the inference time. Thus, it takes about 1.7s/100ms to generate one video frame with/without the post-process on average.

E. Limitations

Emotion: Our method does not explicitly model facial emotion or estimate the sentiment from the input speech audio. Thus, the generated video looks unnatural if the emotion of the driving audio is different from that of the source video. This problem also appears in [45] and we leave this to future improvement.

Tongue: In our method, our Neural Video Rendering Network produces lip fiducials and the teeth proxy adds the teeth high-frequency details. Our method ignores the tongue movement when some phonemes (*e.g.* "Z" in the word "result") are pronounced. The tongue texture can not be well generated according to lip fiducials and teeth proxy as shown in Figure 12 (a).

Accent: Our method performs poorly when the driving speech audio contains an accent. For example, the generated results driven by a speech with a strong Russian accent do not achieve visually satisfactory lip-sync accuracy as shown in Figure 12 (b). We owe it to the fact that English speech with a strong accent is an outlier to our Audio-to-Expression Translation Network and we leave it to future research.

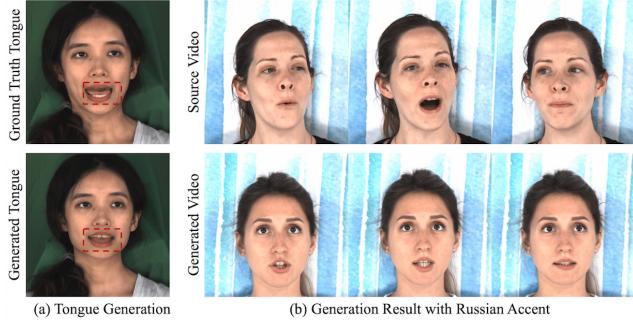


Figure 12: **Failure Cases.** (a) Poor tongue generation result on phoneme "Z" that require the use of tongue. (b) Poor lip-sync accuracy when we use normal speech audio to drive a speaker with strong Russian accent.