Speech-driven Facial Animation using Cascaded GANs for Learning of Motion and Texture

Dipanjan Das*, Sandika Biswas*, Sanjana Sinha, and Brojeshwar Bhowmick

Embedded Systems and Robotics,
TCS Research, India
{dipanjan.da,biswas.sandika,sanjana.sinha,b.bhowmick}@tcs.com

Abstract. Speech-driven facial animation methods should produce accurate and realistic lip motions with natural expressions and realistic texture portraying target-specific facial characteristics. Moreover,the methods should also be adaptable to any unknown faces and speech quickly during inference. Current state-of-the-art methods fail to generate realistic animation from any speech on unknown faces due to their poor generalization over different facial characteristics, languages, and accents. Some of these failures can be attributed to the end-to-end learning of the complex relationship between the multiple modalities of speech and the video. In this paper, we propose a novel strategy where we partition the problem and learn the motion and texture separately. Firstly, we train a GAN network to learn the lip motion in a canonical landmark using DeepSpeech features and induce eye-blinks before transferring the motion to the person-specific face. Next, we use another GAN based texture generator network to generate high fidelity face corresponding to the motion on person-specific landmark. We use meta-learning to make the texture generator GAN more flexible to adapt to the unknown subject's traits of the face during inference. Our method gives significantly improved facial animation than the state-of-the-art methods and generalizes well across the different datasets, different languages, and accents, and also works reliably well in presence of noises in the speech.

Keywords: Realistic facial animation, meta-learning, Cascaded GAN

1 Introduction

Speech-driven facial animation can be used for many applications such as video games, virtual assistants, animation movies, etc. and has thus garnered broad interest. The problem of generating such facial animation is multifaceted, requiring accurate lip-sync, a natural expression like eye blinks, head orientations, capturing subject-specific traits like identity, lip deformations, etc. Also, the generation of such animation should not be overly dependent on the training set, and the method, therefore, should be adaptable to unknown faces and speeches quickly. Existing end-to-end learning methods [36,32] show poor adaptability given an

[*] equal contribution

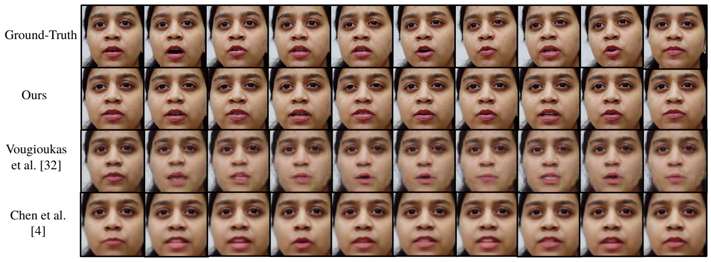


Fig. 1: Recent state-of-the-art methods [4,32][1]for speech-driven facial animation fail to accurately capture the mouth shapes and detailed facial texture on an unknown test subject whose facial characteristics differ from the training data. In these methods, the generated face can appear to be very different from the given target identity [32], or there can be a significant blur in the mouth region [4], leading to unrealistic face animation. On the other hand, our generated facial texture and mouth shapes can accurately resemble the ground-truth animation sequence.

unknown speech or face resulting in implausible animation. In order to overcome the problems of generating images directly from speech, Chen *et al.* [4] learn an intermediate high-level representation of motion from audio followed by texturing. Although this method preserves the identity but fails to produce accurate and realistic lip synchronization, as shown in Fig. 1 (the last row). On the other hand, [32] produces plausible lip motion but renders incorrect identity, as shown in Fig. 1 (third row). Therefore, the key challenges existing in the talking face problem are i) accurate lip synchronization along with identity preservation, ii) presence of natural expression like eye blinks, iii) fast adaptation to unknown subjects, and speeches for all practical purposes. Fig. 1 shows that none of the most recent state-of-the-art methods produce animations which solves all the above challenges.

In this paper, we propose a novel strategy to solve the above-mentioned challenges. In essence, our method partitions the problem into four stages. First, we design a GAN network for learning motion on canonical (person-independent) landmark from DeepSpeech features obtained from audio. GAN is powerful in learning the subtle deformations in lips due to speech, and learning motion in a canonical face makes the method invariant to the person-specific face geometry. Along with this, DeepSpeech features alleviates the problems due to different accents and noises. With all these together, our method is able to learn motion from speech robustly and also adaptable to the unknown speech. Next, we impose eye blinks predicted from a separate network and transfer this learned canonical facial landmark motion to person-specific landmark motion using Procrustes alignment [29]. Subsequently, we train another GAN network for texture generation conditioning with the person-specific landmark. For better adaptation to the unknown subject and unknown head orientation, we meta-learn this

[1] evaluated using their publicly available pre-trained models trained on LRW and TCD-TIMIT datasets respectively.

GAN network using Model-Agnostic-Meta-Learning (MAML) algorithm [12]. At test time, we fine-tune the meta-learned model with few samples (20 images) to adapt quickly (approx. 100 secs fine-tuning) to the unseen subject. Our method produces significantly better results (Fig. 1, second row) with more accurate lip synchronization, better identity preservation, and easy adaptation to the unseen subjects over the state-of-the-art techniques. Fig. 2 shows a conceptual diagram of our approach. The contributions of our work can be summarized as follows:

1. We design a GAN network for learning canonical facial landmark motion from a speech by using DeepSpeech features. The use of GAN helps to learn subtle deformations in lips accurately. DeepSpeech and motion learning in canonical face alleviates the problems in learning due to the variety of person-specific faces and speeches. Therefore the method is more robust to noises, accent, and different face geometry.
2. We use model-agnostic-meta-learning to train another GAN for texture generation conditioned on the person-specific texture. GAN produces high fidelity face images from given landmarks and because the network is meta-learned it provides quick as well as a better adaptation to the unseen subject using a few examples at the fine-tuning stage.

2 Related Work

Speech-driven face animation: In recent years many researchers have focused on the synthesis of 2D talking face video from audio input [7,3,4,32,30,36,28]. The methods which are most relevant to us are [7,4,31,32,37,36,28] which animate an entire face from speech. Earlier methods that learn subject-specific $2D$ facial animation [30,13,11] require a large amount of training data of the target subject. The first subject-independent learning method [7] achieves good lip synchronization, but images generated require additional de-blurring. Hence GAN-based methods [5,4,32,31,36,28] were proposed for generating sharp facial texture in speech-driven $2D$ facial animation. Although these methods animate the entire face, they mainly target lip synchronization with audio [5,4,36,28], by learning disentangled audio representations [22] for robustness to noise and emotional content in audio, and disentangled audio-visual representations [36] to segregate identity information from speech [36,4]. However, these methods have not addressed the other aspects for the realism of synthesized face video, such as natural expressions, identity preservation of target, etc.

Beyond lip synchronization - Realistic facial animation: The absence of spontaneous movements such as eye blinks in synthesized face videos can be easily perceived as fake [21]. Recent works [31,32] have tried to address the problem of video realism by using adversarial learning of spontaneous facial gestures such as blinks. However, the generated videos with natural expressions may still imperfectly resemble the target identity, which can also be perceived as being fake. To retain facial identity information from the given identity image of target, image attention has been learnt with the help of facial landmarks in a hierarchical approach [4]. In this approach [4] the audio is used to generate motion

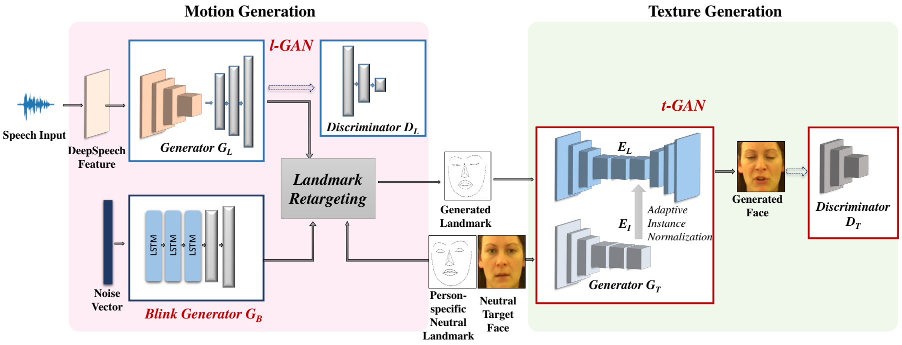


Fig. 2: Block diagram of our proposed method for speech-driven facial animation.

on 2D facial landmarks, and the image texture is generated by conditioning on the landmarks. Although the generated texture in static facial regions can retain the texture from the identity image, the generated texture in regions of motion, especially the eyes and mouth, can differ from the target identity. Hence identity-specific texture generation is needed for realistic rendering of a target's talking face.

3 Proposed Methodology

Given an arbitrary speech and a set of images of a target face, our objective is to synthesize speech synchronized realistic animation of the target face. Inspired by [4], we capture facial motion in a lower dimension space represented by 68 facial landmark points and synthesize texture conditioned on the motion of predicted landmarks. To this end, we use a GAN based cascaded learning approach consisting of the following: (1) Learning speech-driven motion on $2D$ facial landmarks independent of identity, (2) Learning eye blink motion on landmarks, (3) Landmark retargeting to generate target-specific facial shape along with motion, (4) Generating facial texture from the motion of landmarks. Fig. 2 shows our overall approach.

3.1 Speech-driven Motion Generation on Facial Landmarks

Let, A be an audio signal represented by a series of overlapping audio windows $\{W_t | t \in [0, T]\}$ with corresponding feature representations $\{F_t\}$. Our goal is to generate a sequence of facial landmarks $\{\ell_t \in \mathbb{R}^{68\times 2}\}$ corresponding to the motion driven by speech. We learn a mapping $\mathcal{M}_L : F_t \rightarrow \delta\ell_t^m$ to generate speech-induced displacement $\{\delta\ell_t^m \in \mathbb{R}^{68\times 2}\}$ on a canonical landmark (person-independent) in neutral pose ℓ_p^m. Learning the speech-related motion on a canonical landmark ℓ_p^m, which represents the average shape of a face, is effective due to the invariance of any specific facial structure. In order to generalize well over different voices, accent etc. we use a pre-trained DeepSpeech [15] model to extract the feature $F_t \in \mathbb{R}^{6\times 29}$.

Adversarial Learning of Landmark Motion We use an adversarial network *l-GAN* to learn the speech-induced landmark displacement $\mathcal{M}_L$. The generator network G_L generates displacements $\{\delta\ell_t^m\}$ of a canonical landmark from a neutral pose ℓ_p^m. Our discriminator D_L takes the resultant canonical landmarks $\{\ell_t^m = \ell_p^m + \delta\ell_t^m\}$ and the ground-truth canonical landmarks as inputs to learn the real against fake. The loss functions used for training *l-GAN* are as follows:

Distance loss: This is L_2 loss between generated canonical landmarks $\{\ell_t^m\}$ and ground-truth landmarks $\{\ell_t^{m*}\}$ for each frame t.

$$\mathcal{L}_{dist} = ||\ell_t^m - \ell_t^{m*}||_2^2 \tag{1}$$

Regularization loss: We use L_2 loss between consecutive frames for ensuring temporal smoothness in predicted landmarks.

$$\mathcal{L}_{reg} = ||\ell_t^m - \ell_{t-1}^m||_2^2 \tag{2}$$

Direction loss: We also impose a consistency in the motion vectors $(\overrightarrow{\delta\ell_t^m})$ by:

$$\mathcal{L}_{dir} = ||\overrightarrow{\delta\ell_t^m} - \overrightarrow{\delta\ell_t^{m*}}||_2^2 \qquad \text{where,} \quad \overrightarrow{\delta\ell_t^m} = \begin{cases} 1, & \textit{if } \ell_{t+1}^m > \ell_t^m \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

GAN loss: We use an adversarial loss for capturing detailed mouth deformations.

$$\mathcal{L}_{gan} = \mathbb{E}_{\ell_t^{m*}}[log(D_L(\ell_t^{m*})] + \mathbb{E}_{F_t}[log(1 - D_L(G_L(\ell_p^m, F_t))] \tag{4}$$

The final objective function which is to be minimized is as follows:

$$\mathcal{L}_{motion} = \lambda_{dist}\mathcal{L}_{dist} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{dir}\mathcal{L}_{dir} + \lambda_{gan}\mathcal{L}_{gan} \tag{5}$$

where, λ_{dist}, λ_{reg}, λ_{dir}, λ_{gan} are experimentally set to 1, 0.5, 0.5 and 1, as presented in the ablation study (Section 4.3).

3.2 Spontaneous Eye Blink Generation on Facial Landmarks

Eye blinks are essential for realism of synthesized face animation, but not dependent on speech. Therefore, we propose an unsupervised method for generation of realistic eye blinks through learning a mapping $\mathcal{M}_B : Z_t \to \delta\ell_t^e$ from a random noise $Z_t \sim \mathcal{N}(\mu, \sigma^2)|t \in (0, T)$ to eye landmark displacements $\{\delta\ell_t^e \in \mathbb{R}^{22\times 2}\}$. Our blink generator network G_B learns the blink pattern and duration through the mapping $\mathcal{M}_B$ and generates a sequence of eye landmark displacements $\{\delta\ell_t^e\}$ on the canonical face by minimizing the MMD (Maximum Mean Discrepancy) [14] loss defined as follows:

$$L_{MMD} = \mathbb{E}_{X,X'\sim p}\mathcal{K}(X, X') + \mathbb{E}_{Y,Y'\sim q}\mathcal{K}(Y, Y') - 2\mathbb{E}_{X\sim p,Y\sim q}\mathcal{K}(X, Y) \tag{6}$$

where, $\mathcal{K}(x, y)$ is defined as $exp(-\frac{|x-y|^2}{2\sigma})$, p and q represents samples from distributions X and Y of GT $\{\delta\ell_t^{e*}\}$ and generated eye landmark motion $\{\delta\ell_t^e\}$

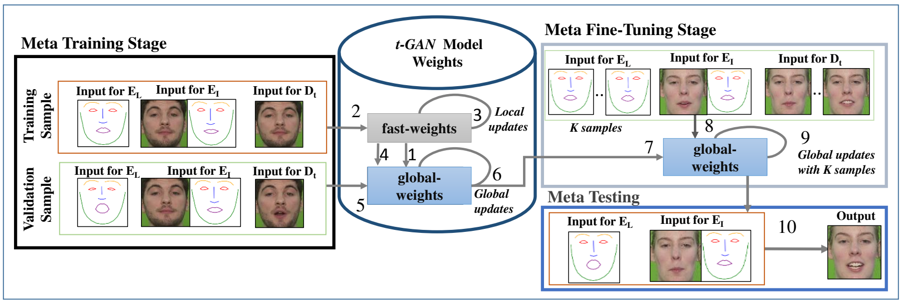


Fig. 3: State transitions of fast-weights (FW) and global-weights (GW) of *t-GAN* during meta-training. The sequence of training schedule: (1) copying FW to GW to keep global state unchanged during the training, (2)-(3) update the FW in iterations, (4)-(5) compute validation loss using FW, (6)update GW using total validation loss, (7) copy GW for the fine-tuning, (8)-(9) updating the GW using K sample images, (10) using the updated GW to produce target subject's face.

respectively. We also use a Min-max regularization to ensure that the range of the generated landmarks matches with the average range of average displacements present in the training data. We augment the eye blink with the speech-driven canonical landmark motion (Section 3.1) and retarget (Section 3.3) the combined landmarks $\ell_t^M = \{\ell_t^m \bigcup \ell_t^e\}$, where $\{\ell_t^e = \ell_p^e + \delta\ell_t^e\}$, to generate the person-specific landmarks $\{\ell_t\}$ for subsequent use for texture generation.

3.3 Landmark Retargeting

We retarget the canonical landmarks $\{\ell_t^M\}$ generated by G_L and G_B, to person-specific landmarks $\{\ell_t\}$ (used for texture generation) as follows:

$$\ell_t = \ell_p + \delta\ell_t \text{ where, } \delta\ell_t = \delta\ell_t^{'} * S(\ell_t)/S(\mathcal{T}(\ell_t^M)) \; ; \; \delta\ell_t^{'} = \mathcal{T}(\ell_t^M) - \mathcal{T}(\ell_p^m) \tag{7}$$

where, ℓ_p is the person-specific landmark in neutral pose (extracted from the target image), $S(\ell) \in \mathbb{R}^2$ is the scale (height × width) of ℓ and $\mathcal{T} : \ell \rightarrow \ell^{'}$ represents a Procrustes (rigid) alignment of ℓ with ℓ_p.

3.4 Image Generation from Landmarks

We use the person-specific landmarks $\{\ell_t\}$ containing motion due to the speech and the eye blink to synthesize animated face images $\{I_t\}$ by learning a mapping $\mathcal{M}_T : (\ell_t, \{\mathcal{I}^n\}) \rightarrow I_t$ using given target images $\{\mathcal{I}^n | n \in [0, N]\}$.

Adversarial Generation of Image Texture We use an adversarial network *t-GAN* to learn the mapping $\mathcal{M}_T$. Our generator network G_T consists of a texture encoder E_I and landmark encoder-decoder E_L influenced by E_I. E_I encodes the texture representation as $e = E_I(\mathcal{I}^n)$ for the input N images. We use Adaptive Instance Normalization [17] to modulate the bottleneck of E_L using e. Finally

we use a discriminator network D_T to discriminate the real images from the fake. The losses for training the *t-GAN* are as follows:
Reconstruction loss: L_2 distance between synthesized $\{I_t\}$ and GT images $\{I_t^*\}$,

$$\mathcal{L}_{pix} = ||I_t - I_t^*||_2^2 \tag{8}$$

Adversarial loss: For sharpness of the texture an adversarial loss is minimized.

$$\mathcal{L}_{adv} = \mathbb{E}_{I_t^*}[log(D_T(\mathcal{I}^n, I_t^*)] + \mathbb{E}_{\ell_t}[log(1 - D_T(\mathcal{I}^n, G_T(\ell_t, \mathcal{I}^n))] \tag{9}$$

Perceptual loss: We use a perceptual loss [18] which is the difference in feature representations vgg_1 and vgg_2 of generated and ground truth images obtained using pre-trained $VGG19$ and $VGGFace$ [26] respectively.

$$\mathcal{L}_{feat} = \alpha_1||vgg_1(I_t) - vgg_1(I_t^*)||_2^2 + \alpha_2||vgg_2(I_t) - vgg_2(I_t^*)||_2^2 \tag{10}$$

The total loss minimized for training G_T network is defined as,

$$\mathcal{L}_{texture} = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{feat}\mathcal{L}_{feat} \tag{11}$$

Meta-learning We use model-agnostic meta-learning (MAML) [12] to train our *t-GAN* for quick adaptation to the unknown face at inference time using few images. MAML trains on a set of tasks T called episodes. For each task, the number of samples for training and validation is d_{trn} and d_{qry}, respectively. For our problem, we define subject specific task as $T^s = (I_i^s, l_j^s) \cdots (I^s_{i_{d_{trn}+d_{qry}}}, l^s_{j_{d_{trn}+d_{qry}}})$ of task set $\{T^s\}$, where s is the subject index, I_i^s is the i^{th} face image for subject s, l_j^s is the j^{th} landmark for the same subject s. During meta-training, MAML store the current weights of the *t-GAN* into global-weights and train the *t-GAN* with d_{trn} samples for m iteration using a constant step size. During each iteration, it measures the loss L^i with the validations samples d_{qry}. Then the total loss $L = L^1 + L^2 \cdots + L^m$ is used to update global-weights as shown in Fig. 3. The resultant direction of the global-weights encodes a global information of the *t-GAN* network for all the tasks, which is used as an initialization for fine-tuning during inference.

During fine-tuning, we initialize the *t-GAN* from the global-weights and update the weights by minimizing the loss as described in Equation 11. We use a few ($K = 20$) example images of the target face for the fine-tuning.

4 Experimental Results

In this section, we present the experimental results of our proposed method on different datasets along with the network ablation study. We also show that the accuracy of our cascaded GAN based approach is quite higher than an alternate regression-based motion and texture generation. Our meta-learning based texture generation strategy makes our method to be more adaptable to any unknown faces. The combined result is a significantly better facial animation from speech than the state-of-the-art methods in terms of both quantitative and qualitative results. In what follows, we present detailed experiments for each of the building blocks of our pipeline.

4.1 Datasets

We use TCD-TIMIT[16], GRID [9], and Voxceleb [24] datasets for our experiments. We train our model only on TCD-TIMIT and test the model on GRID as well as our own recorded data for showing the efficacy of our method on cross datasets with completely unknown faces. Our training split contains 3378 videos from 49 subjects with around 6913 sentences uttered in a limited variety of accents. Test split (same as [32]) of TCD-TIMIT and GRID datasets contains 1631 and 9957 videos respectively.

4.2 Motion Generation on Landmarks

Network Architecture of *l-GAN*: The architecture of the generator network G_L of *l-GAN* is built upon the encoder-decoder architecture used in [10] for generating mesh vertices. LeakyReLU [33] activation is used after each layer of the encoder network. The input DeepSpeech features are encoded to a 33 dimensional vector (PCA coefficients), which is decoded to obtain the canonical landmark displacements from the neutral pose. The discriminator network D_L consists of 2 linear layers, which re-encodes the predicted or ground-truth landmarks into PCA coefficients to discriminate between real and fake. We initialize weights of the last layer of the decoder in G_L and the first layer of D_L with 33 PCA components computed over the landmark displacements in training data.
Network Architecture of Blink Generator G_B: We use RNN to predict a sequence of displacements $\mathbb{R}^{n\times 75\times 44}$, i.e x, y coordinates of eye landmarks $\{\ell_t^e \in \mathbb{R}^{22\times 2}\}$ over 75 timestamps from given noise vector $z \sim \mathcal{N}(\mu, \sigma^2)$ with $z \in \mathbb{R}^{n\times 75\times 10}$. Similar to the G_L of our *l-GAN* network, the last linear layer weights are initialized with PCA components (with 99% variants) computed over ground-truth eye landmark displacements.
Training Details: We extract audio features from the second last layer (before softmax) of the DeepSpeech [15] network. We consider sliding windows of Δt features for providing a temporal context to each video frame. To compute accurate ground-truth facial landmark required for our training, we experiment with different existing state-of-the-art methods [1,34,19] and find that the combination of OpenFace [1] and face segmentation [34] to be most effective for our purpose. Our speech-driven motion generation network is trained on the TCD-TIMIT dataset. The canonical landmarks used for training *l-GAN* are generated by an inverse process of the landmark retargeting method, as described in Section 3.3. We train our *l-GAN* network with a batch size of 6. Losses saturate after 40 epochs, which takes around 3 hours on a single GPU of Quadro P5000 system. We use Adam [20] optimization with a learning rate of $2e-4$ for training both of our *l-GAN* and blink generator network.

Quantitative Results: We present our quantitative results in Table 1 and 2. For comparative analysis we use publicly availabe pre-trained models of state-of-the-art methods [4,32,36]. Our model is trained on TCD-TIMIT [16], while models of [4] and [36] are pre-trained on LRW [8] dataset. [32] is trained on both TCD-TIMIT and GRID separately.

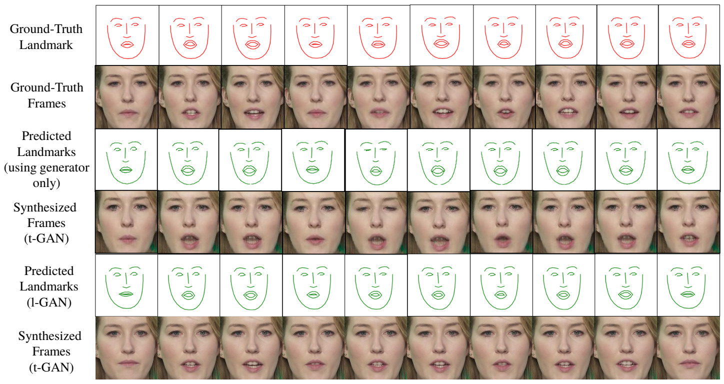

Ground-Truth Landmark
Ground-Truth Frames
Predicted Landmarks (using generator only)
Synthesized Frames (t-GAN)
Predicted Landmarks (l-GAN)
Synthesized Frames (t-GAN)

Fig. 4: Performance comparison of *l-GAN*using only generator (third row) and the complete GAN (fifth row). The regression-based approach cannot capture the finer details like "a" and "o" of lip motion without the help of the discriminator.

For evaluating and comparing the accuracy of lip synchronization produced by our method, we use a) LMD, Landmark Distance (as used in [4],[3]) and b) Audio-Visual synchronization metrics (AV Offset and AV confidence produced by Syncnet [6]). For all methods, LMD is computed using lip landmarks extracted from the final generated frames. A lower value of LMD and AV offset with higher AV confidence indicates better lip synchronization. Our method shows better accuracy compared to state-of-the-art methods. Our models trained on TCD-TIMIT also shows good generalization capability in cross-dataset evaluation on GRID dataset (Table 2). Although [4] also generates facial landmarks from audio features(MFCC), unlike their regression-based approach, our use of DeepSpeech features, landmark retargeting, and adversarial learning results in improved accuracy of landmark generation.

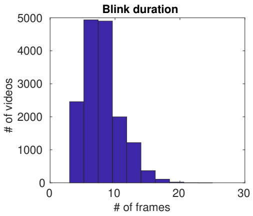


Fig. 5: Statistics of average blink duration.

Moreover, our facial landmarks contain natural eye blink motion for added realism. We detect eye blinks using a sharp drop in EAR (Eye Aspect Ratio) signal [32] calculated using landmarks of eye corners and eyelids. Blink duration is calculated as the number of consecutive frames between the start and end of the sharp drop in the EAR. The average blink duration and blink frequencies generated from our method is similar to that of natural human blinks. Our method produces a blink rate of 0.3 blinks/s and 0.38 blinks/s (Table 1 and 2) for TCD-TIMIT and GRID datasets respectively which is similar to the average human blink rate of 0.28 - 0.4 blinks/s. Also, we achieve an average blink duration of 0.33s and 0.4s, which is similar to as reported in ground-truth (Table 1 and 2). In Fig. 5 we present distribution of blink durations (in no. of frames) in synthesized videos of GRID and TCD-TIMIT datasets. So, our method can produce realis-

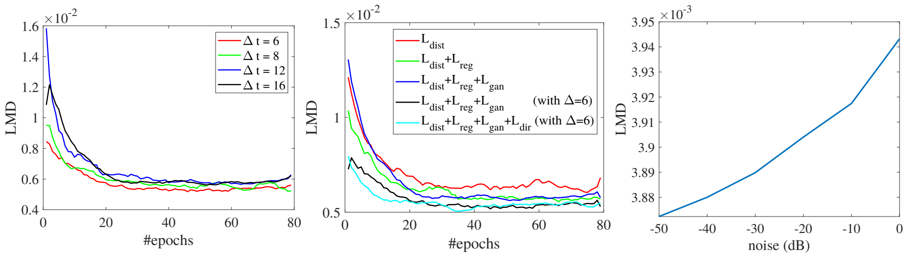


Fig. 6: **Left:** Landmark Distance (LMD) with varying context window (Δt) of deep-speech features. **Middle:** LMD with different losses used for training speech-driven motion generation network. **Right:** Error in lip synchronization (LMD) with different noise levels.

tic eye blinks similar to [32], but with better identity-preserved texture, due to our decoupled learning of eye blinks on landmarks.

Ablation Study: An ablation study of window size Δt (Fig. 6) has indicated a value of $\Delta t = 6$ frames (duration of around $198ms$) results in the lowest LMD. In Fig. 6 we also present an ablation study for different losses used for training our motion prediction network. It is seen that the proposed loss L_{motion} achieves the best accuracy. Use of L_2 regularization loss helps to achieve temporal smoothness and consistency on predicted landmarks over consecutive frames. We use direction loss (Equation 3) to capture the relative movements of landmarks over consecutive frames. Using direction loss helps to achieve faster convergence of our landmark prediction network. Use of DeepSpeech features helps us to achieve robustness in lip synchronization even for audios with noise, different accents, and different languages (Please refer to supplementary video). We experiment to evaluate the robustness of our *l-GAN* with different levels of noise by adding synthetic noise in the audio input. Fig. 6 shows upto $-30dB$, the lip motion does not get affected by the noise and starts degrading afterward. In Fig. 4 we present a qualitative result of the landmark generation network on the TCD-TIMIT dataset. It shows the effectiveness of using discriminator in *l-GAN*.

4.3 Texture Generation from Landmark Motion:

Network Architecture of *t-GAN*: We adapt a similar approach of an image-to-image translation method proposed by [18] for implementation of our texture generator G_T. Our landmark encoder-decoder network E_L takes generated person-specific landmarks represented as images of size $\mathbb{R}^{3\times 256\times 256}$ and E_I takes channel-wise concatenated face images with corresponding landmark images of the target subject. We use six downsampling layers for both E_I and the encoder of E_L and six upsampling layers for the decoder of the E_L. To generate high fidelity images, we use residual block for our downsampling and upsampling layers similar to [2]. We use instance normalization for the residual blocks and adaptive instance normalization on the bottle-neck layer of the E_L using the activation

Table 1: Comparative results on TCD-TIMIT [16].

Methods	Trained on	PSNR	SSIM	CPBD	LMD	ACD (10^{-4})	FaceNet	AVOff.	AVConf.	Blink/s	Blink Dur.(s)
[32]	TCD-TIMIT	24.2	0.73	0.30	2.1	1.76	0.578	1	5.5	0.19	0.33
[4]	LRW	20.31	0.59	0.16	1.71	1.1	0.409	1	3.91	NA	NA
[36]	LRW	23.82	0.63	0.14	1.9	1.24	0.472	1	1.94	NA	NA
Ours	TCD-TIMIT	**30.7**	**0.74**	**0.61**	**1.4**	**0.98**	**0.377**	**1**	**5.91**	**0.3**	**0.33**

Table 2: Comparative results on GRID [9] (our cross-dataset evaluation).

Methods	Trained on	PSNR	SSIM	CPBD	LMD	ACD (10^{-4})	FaceNet	WER(%)	AVOff.	AVConf.	Blink/s	Blink Dur.(s)
[32]	GRID	27.1	0.81	0.26	1.56	1.47	0.802	23.1	1	7.4	0.45	0.36
[4]	LRW	23.98	0.76	0.06	1.29	1.57	0.563	31.1	1	5.85	NA	NA
[36]	LRW	22.79	0.76	0.04	1.49	1.78	0.628	36.72	2	4.29	NA	NA
Ours	TCD-TIMIT	**29.9**	**0.83**	**0.29**	**1.22**	**1.12**	**0.466**	**19.33**	**1**	**7.72**	**0.38**	**0.4**

produced by the last layer of E_I. Moreover, to generate sharper images, we use a similar self-attention method as [35] at the 32×32 layer activation of downsampling and upsampling layers. Our discriminator network D_T consists of 6 residual blocks similar to E_I, followed by a max pooling and a fully connected layer. To stabilize our GAN training, we use spectral normalization [23] for both generator and discriminator network.

Training and Testing Details: We meta-train our *t-GAN* network using ground-truth landmarks following the teacher forcing strategy. We use fixed step size [12] $1e-3$ and Adam as the meta-optimizer [12] with learning rate $1e-4$. The values of α_1, α_2, λ_{pix}, λ_{adv} λ_{feat} and are experimentally set to $1e-1$, $2e-3$, 0.5, 1.0 and 0.3 respectively. At test time, we use 5 images of the target identity, and the person-specific landmark generated by the *l-GAN* to produce the output images. Before testing, we perform a fine-tuning of the meta-trained network using 20 images of the target person and the corresponding ground-truth landmarks. We use a clustered GPU of NVIDIA Tesla V100 for meta-training and Quadro P5000 for fine-tuning our network.

Quantitative Results: Here, we present the comparative performance of our GAN-based texture generation network with the most recent state-of-the-art methods [4], [36] and [32]. Similar to *l-GAN*, the *t-GAN* is trained on TCD-TIMIT and evaluated on the test split of GRID, TCD-TIMIT and the unknown subjects. We compute the performance metrics PSNR, SSIM (Structural Similarity), CPBD (Cumulative Probability Blur Detection) [25], ACD (Average Content Distance) [32] and similarity between FaceNet [27] features for reference identity image (1st frame of ground truth video) and the predicted frames. Our method outperforms (Table 1 and 2) the state-of-the-art methods for all the datasets indicating better image quality. Due to inaccessibility of LRW [8] dataset we have evaluated our texture generation method on Voxceleb [24] dataset which gives average PSNR, SSIM and CPBD of 25.2, 0.63, 0.11 respectively. Our method does not produce head motion and synthesizes texture with frontal face. Hence, for Voxceleb, our method gives poor performance than that of TCD-TIMIT and GRID.

Qualitative Results: Fig. 10 shows qualitative comparison against [4], [36]

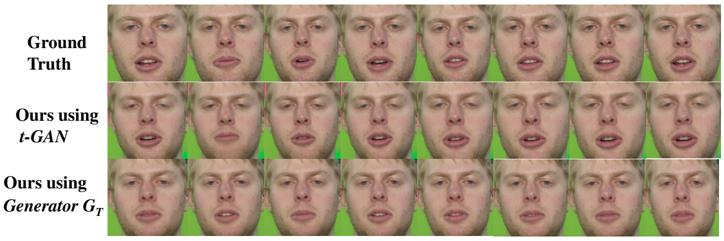


Fig. 7: Qualitative comparison between our *t-GAN* based method (Row 2) against the regression based generator G_T (Row 3) method. Use of GAN results in more accurate mouth shape.

Table 3: Ablation Study of our model. CC = channel wise concatenation.

Methods	PSNR	SSIM	CPBD	LMD
Model+CC+L_{pix}	27.2	0.62	0.51	1.65
Model+ADIN+L_{pix}	28.3	0.66	0.56	1.57
Model+ADIN+L_{pix} +L_{feat}	28.9	0.70	0.58	1.5
Model+ADIN+L_{pix} +L_{feat}+L_{adv}	**30.7**	**0.74**	**0.61**	**1.4**

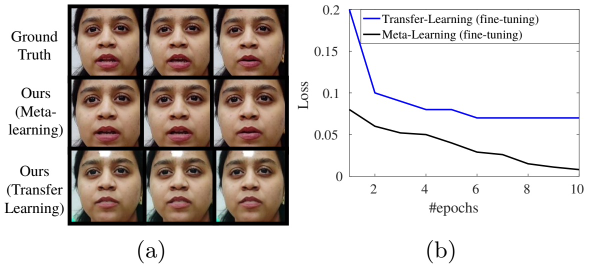
Table 4: Epoch-wise quantitative analysis in fine-tuning.

DataSet	Epoch	PSNR	SSIM	CPBD	LMD
GRID	1	21.5	0.58	0.04	6.70
	5	27.3	0.76	0.08	1.47
	10	29.8	0.83	0.29	1.22
TCD-TIMIT	1	20.6	0.59	0.38	7.80
	5	28.1	0.70	0.58	1.64
	10	30.7	0.74	0.61	1.4

and [32]. It can be seen that [32] and [36] fail to preserve the identity of the test subject over frames in the synthesized video. Although [4] can preserve the identity, there is a significant blur, especially around the mouth region. Also, it lacks any natural movements over face except lip or jaw motion yielding an unrealistic face animation. On the other hand, our method can synthesize high fidelity images (256 × 256) with preserved identity and natural eye motions. Fig. 7 shows the qualitative comparison of our GAN based texture generation against a regression-based (without discriminator) network output where it is evident that our GAN based network gives more accurate lip deformation with similar motion as ground-truth.

Ablation Study: We show a detailed ablation study on the TCD-TIMIT dataset to find out the effect of different losses (Table 3). Among channel-wise concatenation and adaptive instance normalization, which are the two different approaches in neural style transfer, adaptive instance normalization works better for our problem. Fig. 7 and quantitative result (Table 3) show that GAN based method produces more accurate lip deformation than the regression-based method, which always produces an overly smooth outcome. Fig. 9 shows the ablation study for the number of images required for fine-tuning. Using single image for fine-tune yields average PSNR, SSIM and CPBD values of 27.95, 0.82, and 0.27 respectively for GRID dataset. Our method can produce accurate motion and texture after 10 epochs (Table 4) of fine-tuning with $K = 20$ sample images.

Meta-learning vs. Transfer-learning: We compare the performance of MAML [12] and transfer-learning for our problem. To this end, we train a model with the same model architecture until it converges to similar loss values as meta-learning.



(a) (b)

Fig. 8: Comparison between the fine-tuning stage of meta-learning and transfer-learning. Meta-learning (black) provides better initialization than the transfer-learning (blue).

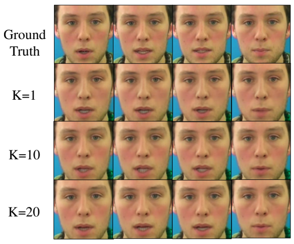


Fig. 9: Ablation study for no. of images during fine-tuning on GRID dataset.

After 10 epochs of fine-tuning with 20 images, the loss of meta-learning is much lower (Fig. 8b) than the transfer-learning (fine-tuning) and produces significantly better visual results (Fig. 8a) than transfer-learning. Moreover, fine-tuning of meta-learned network takes nearly 10 epochs with 20 images, which is much smaller than the transfer-learning based fine-tuning.

User Study: We assess the realism of our animation through a user study, where 25 users are asked to rate between 1(fake)-10(real) for 30 (10 videos from each of the methods) synthesized videos randomly selected from TCD-TIMIT and GRID. Our method achieves better realism scores with an average of 72.76% compared to state-of-the-art methods [4] and [32] with average scores 58.48% and 61.29% respectively.

5 Conclusion

In this paper, we present a novel strategy for speech driven facial animation. Our method produces realistic facial animation for unknown subjects with different languages and accents in speech showing generalization capability. We attribute this advantage due to our separate learning of motion and texture generator GANs with meta-learning capability. As a combined result, our method outperforms state-of-the-art methods significantly. In the future, we would like to study the effect of meta-learning for learning landmark motion from speech to mimic personalized speaking styles.

Acknowledgement

We would like to thank Dr. Angshul Majumdar, Professor at IIIT Delhi, India for helping us to get access of TCD-TIMIT dataset required for this research. We would also like to thank Govind Gopal from our infrastructure team for his immense support for creating a clustered GPU setup for training.

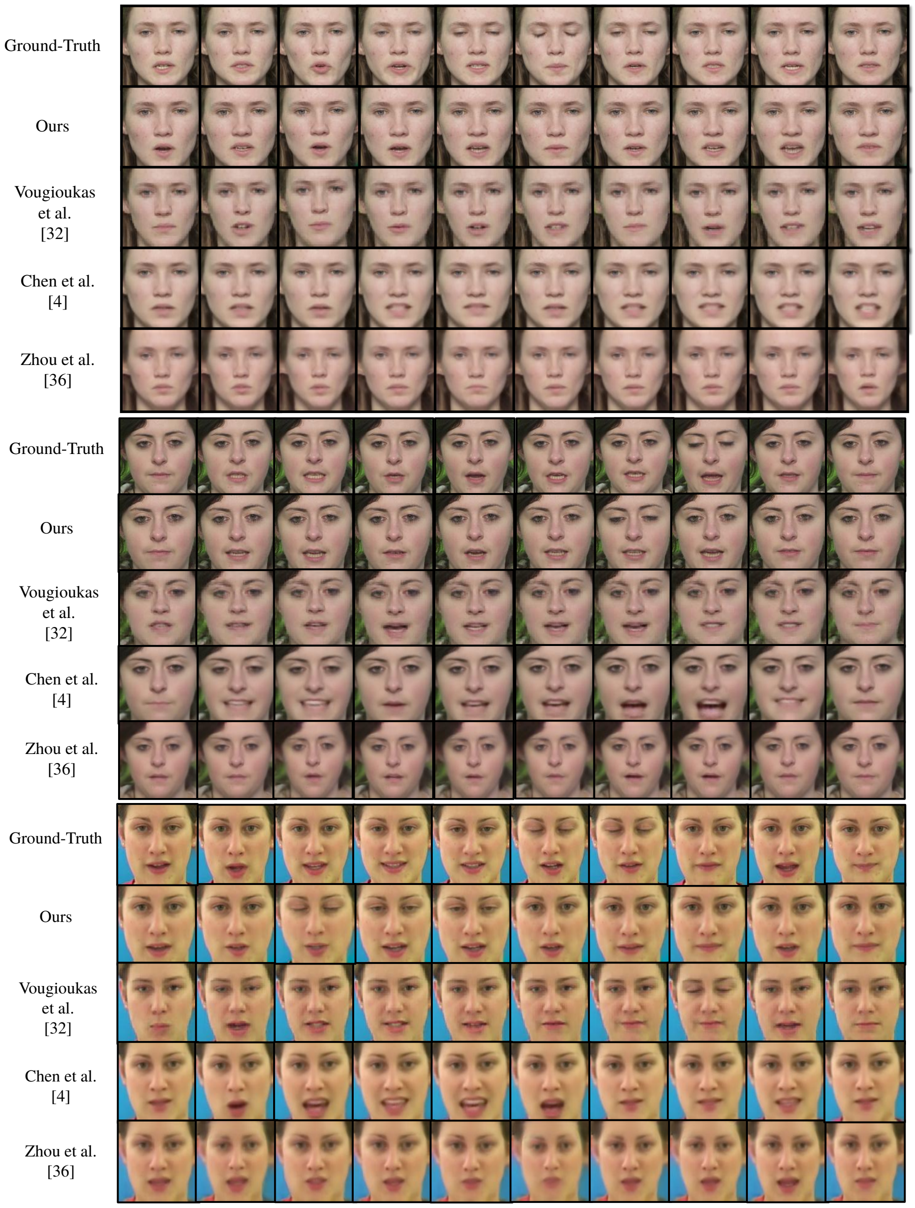


Fig. 10: Qualitative comparison with the latest SoA methods on TCD-TIMIT dataset (Upper 10 rows) and GRID dataset (Lower 5 rows). Our results indicate improved identity preservation of the subject, good lip synchronization, detailed texture (such as teeth), lesser blur, and presence of randomly introduced eye blinks.

References

1. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 59–66. IEEE (2018)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
3. Chen, L., Li, Z., K Maddox, R., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 520–535 (2018)
4. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7832–7841 (2019)
5. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. pp. 349–357. ACM (2017)
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)
7. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? arXiv preprint arXiv:1705.02966 (2017)
8. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision. pp. 87–103. Springer (2016)
9. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America **120**(5), 2421–2424 (2006)
10. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019)
11. Fan, B., Wang, L., Soong, F.K., Xie, L.: Photo-real talking head with deep bidirectional lstm. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4884–4888. IEEE (2015)
12. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017)
13. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer graphics forum. vol. 34, pp. 193–204. Wiley Online Library (2015)
14. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In: Advances in neural information processing systems. pp. 513–520 (2007)
15. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
16. Harte, N., Gillen, E.: Tcd-timit: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia **17**(5), 603–615 (2015)
17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)

18. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
19. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1867–1874 (2014)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)
22. Mittal, G., Wang, B.: Animating face using disentangled audio representations. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 3290–3298 (2020)
23. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
24. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)
25. Narvekar, N.D., Karam, L.J.: A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In: 2009 International Workshop on Quality of Multimedia Experience. pp. 87–91. IEEE (2009)
26. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: bmvc. vol. 1, p. 6 (2015)
27. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
28. Song, Y., Zhu, J., Wang, X., Qi, H.: Talking face generation by conditional recurrent adversarial network. arXiv preprint arXiv:1804.04786 (2018)
29. Srivastava, A., Joshi, S.H., Mio, W., Liu, X.: Statistical shape analysis: Clustering, learning, and testing. IEEE Transactions on pattern analysis and machine intelligence **27**(4), 590–602 (2005)
30. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) **36**(4), 95 (2017)
31. Vougioukas, K., Center, S.A., Petridis, S., Pantic, M.: End-to-end speech-driven realistic facial animation with temporal gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 37–40 (2019)
32. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision pp. 1–16 (2019)
33. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
34. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 325–341 (2018)
35. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
36. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9299–9306 (2019)
37. Zhu, H., Zheng, A., Huang, H., He, R.: High-resolution talking face generation via mutual information approximation. arXiv preprint arXiv:1812.06589 (2018)