Audio-Driven Talking Face Generation with Diverse yet Realistic Facial Animations

Rongliang Wu[a,b], Yingchen Yu[a], Fangneng Zhan[c], Jiahui Zhang[a], Xiaoqin Zhang[d,*], Shijian Lu[a,*]

[a]*School of Computer Science and Engineering, Nanyang Technological University, Singapore*
[b]*Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore*
[c]*Max Planck Institute for Informatics, Germany*
[d]*College of Computer Science and Artifcial Intelligence, Wenzhou University, China*

Abstract

Audio-driven talking face generation, which aims to synthesize talking faces with realistic facial animations (including accurate lip movements, vivid facial expression details and natural head poses) corresponding to the audio, has achieved rapid progress in recent years. However, most existing work focuses on generating lip movements only without handling the closely correlated facial expressions, which degrades the realism of the generated faces greatly. This paper presents DIRFA, a novel method that can generate talking faces with diverse yet realistic facial animations from the same driving audio. To accommodate fair variation of plausible facial animations for the same audio, we design a transformer-based probabilistic mapping network that can model the variational facial animation distribution conditioned upon the input audio and autoregressively convert the audio signals into a facial animation sequence. In addition, we introduce a temporally-biased mask into the mapping network, which allows to model the temporal dependency of facial animations and produce temporally smooth facial animation sequence. With the generated facial animation sequence and a source image, photo-realistic talking faces can be synthesized with a generic generation network. Extensive experiments show that

*Corresponding author
Email Address: {ronglian001, yingchen001, jiahui003}@e.ntu.edu.sg, fzhan@mpi-inf.mpg.de, zhangxiaoqinnan@gmail.com, shijian.lu@ntu.edu.sg

DIRFA can generate talking faces with realistic facial animations effectively.

Keywords: Audio-driven Talking Face Generation, Face, Face Animation, Audio-to-Visual Mapping, Image Synthesis

---

1. Introduction

Audio-driven talking face generation aims to synthesize talking faces with realistic facial animations that correspond to the input audio signals, which has attracted increasing interest from both academia and industry due to its wide applications in digital human, visual dubbing, virtual reality, etc. Facial animations are usually closely entangled with the corresponding speech, which contain strong non-verbal information that helps the audience understand the speech contents [1]. Generating naturally-looking facial animations is therefore one key factor in realistic talking face generation, which remains a very open research challenge. Specifically, modelling facial animations from audio signals is not a rigid one-to-one mapping problem, which requires to accommodate fair variations of plausible facial animations given the same audio.

Audio-driven talking face generation has been explored extensively with the prevalence of deep generative networks in recent years. One typical approach is static talking face generation [2, 3, 4] which edits lip movements only without considering other facial animations (i.e., head poses and facial expressions). Another approach focuses on dynamic talking face generation [5, 6, 7], which includes head movements for modelling full-face animations but the generated faces are still expressionless. Hence, most existing work focuses on generating lip-synced talking faces with respect to the driving audio signals (with or without head movements) but neglects facial expressions which are crucial to realistic talking face generation.

At the other end, generating realistic audio-driven facial expressions is a non-trivial task. This is largely due to the fact that facial expressions do not have very strong correlations with the corresponding audio. Specifically, there often exist fair variations of plausible facial expressions for the same audio signal, and

the variations might further expand while extending to the temporal dimension with a sequence of audio signals as input. To model such uncertainty, the desired network should be capable of accommodating fair variations of plausible facial expressions for the same input audio. Deterministic prediction of facial expressions from the input audio will drive the regression to the mean facial expression which often leads to expressionless talking face videos [5, 6, 7].

This paper presents **DIRFA**, an innovative audio-driven talking face generation method that can generate talking faces with **DI**verse yet **R**ealistic **F**acial **A**nimations (including *accurate* lip movements, *vivid* facial expression details and *natural* head poses) from the same driving audio. Specifically, we design a mapping network to model the uncertain relations between audio and visual signals. The design is inspired by the transformer architecture [8], which allows the mapping network to model the variational facial animation distribution conditioned upon the input audio and convert the audio into facial animation sequence in an autoregerssive manner. In addition, we introduce a temporally-biased mask into the mapping network, which allows to model the temporal dependency of facial animations and produce temporally coherent animation sequences effectively by assigning higher attention weights to closer facial frames while generating new animations. With the generated facial animation sequence and a source facial image, talking face videos with realistic facial animations can be synthesized with a generic generation network.

The contributions of this work can be summarized in three aspects. First, we propose DIRFA, a novel audio-driven talking face generation method that can generate talking face videos with diverse yet realistic facial animations from the same input audio. Second, we design a probabilistic mapping network that can model the uncertain relations between audio and visual signals effectively. In addition, we introduce a temporally-biased mask into the mapping network, which enables it to generate temporally coherent facial animations. Third, extensive experiments show that DIRFA can generate realistic talking face videos with naturally-looking facial animations.

2. Related Work

2.1. Talking Face Generation

Audio-driven Talking Face Generation: Audio-driven talking face generation has been studied in both computer graphics and computer vision communities for years. Early studies focus on subject-specific talking face synthesis. For example, Suwajanakorn et al. [9] generate mouth movement from the driving audio and composite it with video frames of the same speaker to synthesize talking faces. This approach requires a large amount of video footage of one speaker to train a speaker-specific model that cannot generalize to new persons. Recently, several studies [10, 11, 12, 13, 4, 14] exploit deep generative networks for subject-independent talking face generation. For example, Chen et al. [11] present a hierarchical structure that first predicts facial landmarks from the audio and then generates talking faces conditioned on the predicted landmarks. However, all aforementioned methods generate head-fixed talking faces which degrades the realism of the synthesized videos greatly. To improve the perceptual realism, some recent work [5, 6, 7, 15, 16] considers head poses while generating talking face videos.

Although modelling head movements from speech improves the realism of talking faces, most existing methods share a common constraint that they usually synthesize expressionless talking faces as they neglect to model facial expressions from the audio. This directly degrades the realism of the generated talking faces as human facial expressions usually change with the speech. Leveraging the dedicatedly collected data, some recent studies [17, 18] attempt to control the facial expressions of the generated talking faces of a specific person, but their models cannot generalize to new persons. Liang et al. [19] propose to use an additional emotional video for guiding expression synthesis. Our proposed DIRFA instead can synthesize talking faces with diverse yet realistic facial expressions from audio only and the trained model is generalizable to unseen persons.

Video-driven Talking Face Generation: The task of video-driven talking face generation aims to transfer facial animations from a reference actor to a

target person. For example, Zakharov et al. [20] propose a few-shot talking head model that generates talking faces conditioned on the facial landmarks. Ren et al. [21] leverage 3D face models to transfer facial animations. Wiles et al. [22] introduce a warping-based talking face generation network. Although these methods can generate photo-realistic talking face videos, they require videos with desired facial animations to guide the synthesis.

2.2. Modelling Facial Animations from Audio

One fundamental challenge in audio-driven talking face generation is how to accurately convert audio contents into visual signals. Existing methods tackle this challenge by either implicitly modelling facial animations from audio with latent features [10, 12] or explicitly converting audio into intermediate visual representations such as 2D facial landmarks [11, 13] and 3D face coefficients [16, 21]. Though great progress has been witnessed in generating accurate lip sync, most existing work models audio-to-visual mapping in a deterministic manner and ignores the uncertainty between audio and facial expressions. This leads to regress-to-mean problem [15] where the synthesized faces have little variation in facial expressions.

Differently, our proposed mapping network is trained to model the variational facial animation distribution conditioned upon the input audio, which allows to convert the audio signals into facial animation sequences with diverse yet realistic facial animations. Specifically, we map audio signals to facial animations by exploiting continuous Action Units (AUs) [23] for jointly modelling mouth movements and facial expressions, and a rigid 6 degrees of freedom (DoF) movement vector (pitch, yaw, roll and 3D translation) for modelling head pose.

2.3. Transformers in Vision

Transformer [8] emerges as a powerful tool for modelling long-range contextual information, which has been widely studied in natural language processing. The core of transformer is the attention mechanism that allows for interaction between sequences regardless of the relative positions. A basic transformer

building block consists of a multi-head attention-layer (Attn) followed by a feed forward layer (FF), which embeds input sequence X into an internal representation C, which is often referred to as the context vector [24]. Specifically, C is computed using the query Q, the key K and the value V via:

$$\begin{aligned} C = FF(Attn(Q,K,V,M)) = FF(softmax(\frac{QK^T+M}{\sqrt{N_k}})V), \\ Q = XW^Q, K = XW^K, V = XW^V, \end{aligned} \tag{1}$$

where N_k is the number of channels, Ws are the trainable weights and M is a mask that enables causal attention, where each token can only attend to its past inputs [25].

Dosovitskiy et al. [26] make the first attempt to apply transformer to image classification, which shows the potential of transformer in solving computer vision problems. From then on, transformer-based models have been exploited for different tasks, including object detection [27], image generation [28]), video understanding [29], etc. Some recent work explores applying transformer to model sequential data and produces impressive results on motion synthesis [30] and dance generation [24, 31]. Nevertheless, most existing methods either focus on motion generation without considering audio information, or cannot generate diverse motion sequences given audio only. In this paper, we employ transformer to convert audio signals into diverse yet realistic facial animation sequences for talking face generation.

3. Method

3.1. Overview

Fig. 1 shows the framework of our proposed DIRFA. As audio usually has strong correlations with lip shapes and relatively weak correlations with facial expressions and head poses, our goal is to generate talking faces with *accurate* lip movements, *vivid* facial expression details and *natural* head poses corresponding to the driving audio. To achieve this goal, we design a probabilistic mapping network that exploits transformer for capturing uncertain relations

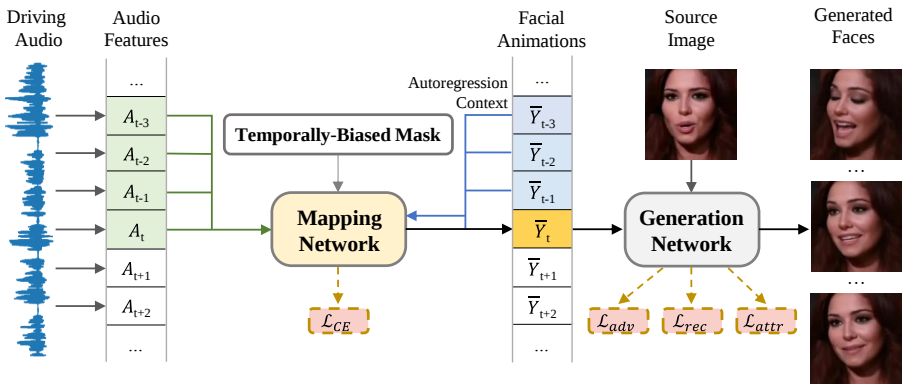


Figure 1: The framework of DIRFA: Given the driving audio, we first convert it into audio features that are temporally aligned with the video frames. The audio features are then fed to a mapping network that generates plausible facial animations in an autoregressive manner. A temporally-biased mask is introduced to the mapping network to enable it to better model the temporal dependency of facial animations and produce temporally coherent facial animation sequence. Finally, the generated facial animations and a source image are forwarded to a generation network to synthesize talking faces with *accurate* lip movements, *vivid* facial expressions and *natural* head poses.

between audio signals and facial animations. Specifically, the mapping network models the probability distribution of variational facial animations conditioned upon the driving audio signals, which can autoregressively transfer the audio into diverse yet realistic facial animation sequences for guiding talking face generation. In addition, we introduce a temporally-biased mask into the mapping network, which allows to model the temporal dependency of facial animations and produce temporally coherent facial animation sequence effectively. With the generated facial animation sequence and a source image, a generation network can directly synthesize photo-realistic talking faces. More details to be described in the ensuing subsections.

3.2. Mapping Network

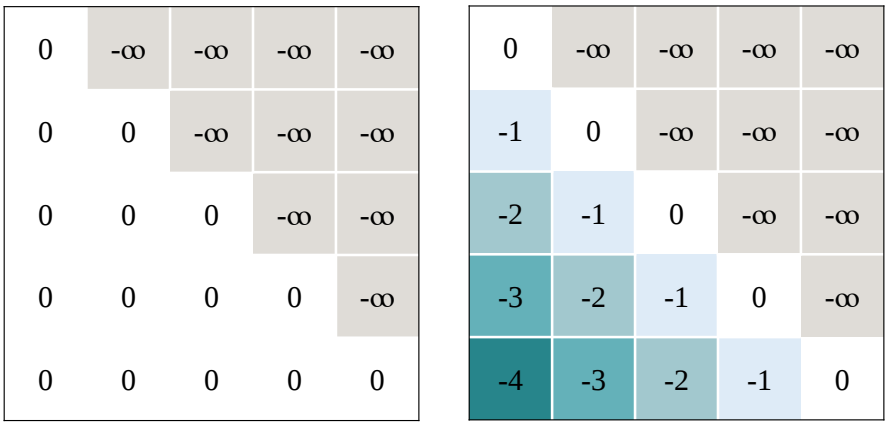
3.2.1. Facial Animations Representation

The goal of the mapping network is to convert the driving audio signals into reasonable and coherent visual signals for guiding talking face generation. Before learning the audio-to-visual mapping, it is crucial to define a compact yet informative representation for the facial animations as visual signal. In this paper, we exploit 17 continuous Action Units (AUs) [23] for jointly modelling accurate mouth movements and rich facial expression details. Meanwhile, we model the head pose as a rigid 6 degrees of freedom (DoF) movement (pitch, yaw, roll and 3D translation). In this way, we encode the facial animations of each talking face frame into a 23 dimensional vector.

Existing audio-driven talking face generation methods [11, 13] represent facial animations with continuous value (e.g., the coordinates of 2D landmarks or the 3D face coefficients) and learn the audio-to-visual mapping in a deterministic manner. This paradigm ignores the uncertainty between audio signals and facial expressions and often leads to regress-to-mean problem [15] where the synthesized faces have little variation in facial expressions. Differently, we represent facial animations as discrete categories and formulate the output of the mapping network as a probability tensor of the categorical distribution, which allows to sample diverse yet realistic facial animations at inference. Specifically, we uniformly discretize facial animations into D constant intervals and obtain 23 D-dimensional one-hot vectors (including 17 AUs and 6 DoF of head movement). Given a facial animation vector $Y_t = [y_1, y_2, ..., y_{23}]$ of frame t, its discrete representation $\overline{Y}_t = [\overline{y}_1, \overline{y}_2, ..., \overline{y}_{23}]$ can be obtained as follows:

$$\overline{y}_i = \underset{d \in D}{\operatorname{argmin}} ||y_i - y_d||_1, \tag{2}$$

where $y_i \in Y_t$, $\overline{y}_i \in \overline{Y}_t$ and y_d is the centroid value of d^{th} category. In our experiments, we empirically set D at 500.



(a) Original Mask (b) The Proposed Temporally-Biased Mask

Figure 2: Comparison of the original mask in [8] and the proposed temporally-biased mask: Our temporally-biased mask negatively biases attention weights with a linearly decreasing penalty that is proportional to the distance between the current prediction and previously predicted animations. Our mask allows the mapping network to better model the temporal dependency of facial animations and produce temporally coherent facial animation sequence effectively. The attention weights in upper triangular are set at $-\infty$ to ensure currently predicting facial animations attend to its previously generated frames only [8].

3.2.2. Audio-to-visual Mapping

Inspired by the power of transformer in modelling long-range contextual information, we introduce transformer [8] into the mapping network for modelling the uncertain relations between audio and visual signals and converting the driving audio signals into facial animations in an autoregressive manner.

Autoregressive generation has been widely explored for generating sequential data, where the probability of each new observation conditions on its previous observations and the joint distribution of the whole sequence is modelled by the product of conditional distributions. With the conditional nature of audio-driven talking face generation, we train the mapping network to autoregressively predict facial animations conditioned on both audio context and previously generated facial animations. Following Bayes' Rule, the probability of discrete facial

animation sequence $\overline{Y}_{1:T}$ under driving audio $A_{1:T}$ can be decomposed by:

$$p(\overline{Y}|A) = p(\overline{Y}_{1:\tau}|A_{1:\tau}) \cdot \prod_{t=\tau+1}^{T} p(\overline{Y}_t|\overline{Y}_{t-\tau:t-1}, A_{t-\tau:t}), \tag{3}$$

where we assume that facial animations at next time step t depends on at most τ previous facial animations (besides the audio signals). Following [8], a mask is applied to the transformer to ensure the currently predicting facial animations attend to its previously generated frames only.

Once the mapping network is trained, it allows to generate diverse and plausible facial animations from the same audio in an autoregressive manner. Specifically, given the audio signals and previously generated animations, the mapping network first predicts the likelihood of possible facial animations for current time step. The top-k sampling is then adopted to sample facial animations from the likelihood as the newly generated facial animations. This process repeats until sequence of desired length is produced. Finally, the generated discrete facial animations $\tilde{Y}_t$ are converted back to continuous facial animations $\hat{Y}_t$ by replacing the discrete categorical label with the corresponding centroid value:

$$\tilde{y}_i = k \text{ such that } \hat{y}_i = y_k, \tag{4}$$

where $\hat{y}_i \in \hat{Y}_t$, $\tilde{y}_i \in \tilde{Y}_t$, $\hat{y}_i$ is the converted continuous value of $\tilde{y}_i$ and y_k is the centroid value of the k^{th} category, respectively.

3.2.3. Improving Temporal Coherence

As facial animations change smoothly when we talk, modelling temporal dependency of facial animations is one key factor in generating realistic talking faces. Without considering the temporal dependency, the vanilla transformer [8] assigns equal attention weights to all previously generated facial animations while generating the new one, which often leads to abrupt facial animations in the generated faces.

Inspired by the intuition that the closer facial animation frames should have stronger correlations with the current frame and thus are more likely to affect its prediction, we introduce a temporally-biased mask into the mapping network, which biases the attention by assigning higher attention weights to the

closer frames. Specifically, our temporally-biased mask negatively biases the attention weights with a linearly decreasing penalty [32] that is proportional to the distance between the previously predicted frames and current prediction as illustrated in Fig. 2. After that, a softmax function is employed to obtain the final attention weights as in Eq. (1). With the proposed temporally-biased mask, the mapping network is encouraged to take into account the temporal dependency of facial animations across consecutive frames while generating facial animations at current time step, which enables it to produce temporally coherent facial animation sequence effectively. Fig. 2 shows the comparison between the original attention mask in [8] and our temporally-biased mask.

3.2.4. Loss Functions

Given the predicted facial animation probability $p(\overline{Y}|A)$ and ground-truth discrete facial animation sequence $\overline{Y}_{1:T}$, we train the mapping network by minimizing the cross-entropy loss:

$$L_{CE} = \frac{1}{T}\text{CrossEntropy}(p(\overline{Y}|A), \overline{Y}_{1:T}). \tag{5}$$

3.3. Generation Network

Given a source image and the facial animations generated by the mapping network, a generation network is designed to synthesize plausible talking faces with desired facial animations while preserving the source identity attributes.

We design the generation network based on face-vid2vid [33] and follow existing talking face generation methods [11, 7] to train it on video datasets [34, 35]. Specifically, we first randomly extract two frames from the same video track and pick one of them as the source image I_{src}. We then treat the other frame as the target image I_{tgt} and extract the facial animation attributes from the it as the driving attributes Z_{tgt}. Finally, I_{src} and Z_{tgt} are fed into the generation network G, which is trained to transform I_{src} to I_{tgt} conditioned on Z_{tgt}:

$$\hat{I}_{tgt} = G(I_{src}, Z_{tgt}), \tag{6}$$

where $\hat{I}_{tgt}$ the generated image.

With a source facial image and the facial animation sequence generated by the mapping network, the trained generation network can generate talking faces with realistic facial animations as illustrated in the right part of Fig. 1.

3.3.1. Loss Functions

The loss for training the generation network consists of three terms: 1) the adversarial loss for improving the photo-realism of the output; 2) the reconstruction loss for penalizing the reconstruction error; 3) the attribute loss that examines whether the generated image contains desired facial animations. The overall objective function is:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{attr}\mathcal{L}_{attr}. \tag{7}$$

Adversarial Loss: We adopt an adversarial loss for improving the photo-realism of the synthesized talking face images. Specifically, LSGAN [36] is employed to optimize network parameters. The adversarial loss is formulated as:

$$\mathcal{L}_{adv} = \frac{1}{2}\mathbb{E}(D_I(I_{tgt})^2) + \frac{1}{2}\mathbb{E}((1 - D_I(\hat{I}_{tgt}))^2), \tag{8}$$

where D_I is the discriminator.

Reconstruction Loss: We employ a reconstruction loss to reduce the error between the output of the generation network and the ground-truth image:

$$\mathcal{L}_{rec} = \mathbb{E}(||\phi(I_{tgt}) - \phi(\hat{I}_{tgt})||_1), \tag{9}$$

where ϕ is the pre-trained VGG network.

Attribute Loss: We adopt an attribute loss to encourage the generation network to generate a face image with similar facial animations as the ground-truth. Specifically, we include an auxiliary head on top of the discriminator (D_Z) to predict attributes, and apply L2 loss on the predicted attributes of both ground-truth and generated images:

$$\mathcal{L}_{attr} = \mathbb{E}(\|D_Z(I_{tgt}) - Z_{tgt}\|_2^2) + \mathbb{E}(\|D_Z(\hat{I}_{tgt}) - Z_{tgt}\|_2^2). \tag{10}$$

4. Experiments

4.1. Experimental Settings

Datasets: We conduct experiments over two in-the-wild audio-visual datasets Voxceleb2 [34] and LRW [35], which are widely used in existing audio-driven talking face generation studies. Specifically, Voxceleb2 [34] contains over 1 million utterances for 6,112 celebrities, extracted from videos uploaded to YouTube. LRW [35] contains 1,000 utterances of 500 different words, spoken by hundreds of different speakers. We split each dataset into training and testing sets by following the official settings.

Implementation Details: The audios are pre-processed to 16kHz, then converted to mel-spectrograms with FFT windows size 1280, hop length 160 and 20 Mel filter-banks. The facial animations (i.e., the continuous AUs and head pose parameters) of each frame are extracted by OpenFace [37]. To match facial animations of each video frame with audio features, we align the extracted mel-spectrogram segment to a short video clip that has the same time duration, where the target video frame is strategically positioned at the center. τ is set at 50. All experiments are conducted in PyTorch environment with four 11GB GeForce RTX 2080 Ti GPUs. The mapping network and the generation network are trained separately. Please refer to the supplementary material for the network training details.

4.2. Qualitative Evaluation

We first qualitatively compare DIRFA with three state-of-the-art subject-independent audio-driven talking face generation methods, including ATVG [11], MakeItTalk [13] and PC-AVS [7]. The results are shown in Fig. 3. Note all compared methods take the source image and the driving audio as inputs (unseen in training) for generating talking faces, while PC-AVS [7] takes the audio-synced videos as additional input for controlling head poses.

ATVG [11] generates talking faces conditioned on 2D landmarks predicted from the audio. It can generate faces with good lip-sync with respect to the

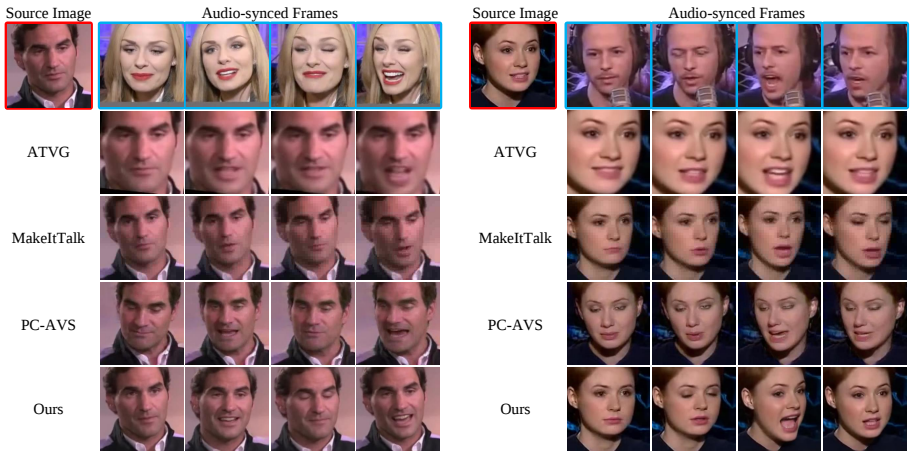


Figure 3: Qualitative comparisons of DIRFA with state-of-the-art audio-driven talking face generation approaches: The first row shows the source image and audio-synced frames that provide ground-truth lip shapes. The rest rows show the synthesized faces. DIRFA can generate *accurate* lip shapes with respect to the audio-synced frames, *vivid* facial expressions and *natural* head poses conditioned on audio only. Note PC-AVS [7] takes the audio-synced frames as additional input for head pose control.

audio, but its generated head poses are static. Leveraging speaker-aware 3D landmarks, MakeItTalk [13] can synthesize photo-realistic talking faces, but their head poses show slight movements only and the lip shapes are not well-aligned with the audio-synced frames. PC-AVS [7] can generate more diverse head motions and accurate lip movements than MakeItTalk [13], but it requires additional videos for pose guidance and struggles to preserve the identity of source image. Further, all three methods model audio-to-visual mapping in a deterministic manner without considering the uncertainty between audio and facial animations. This leads to regression-to-mean problem [15] where the synthesized faces have little variation in facial expressions. As a comparison, DIRFA generates talking faces with accurate lip shapes, vivid facial expressions and natural head poses conditioned on audio only. Its superior synthesis is largely attributed to our probabilistic mapping network that can model the variational facial animation distribution conditioned upon the input audio.

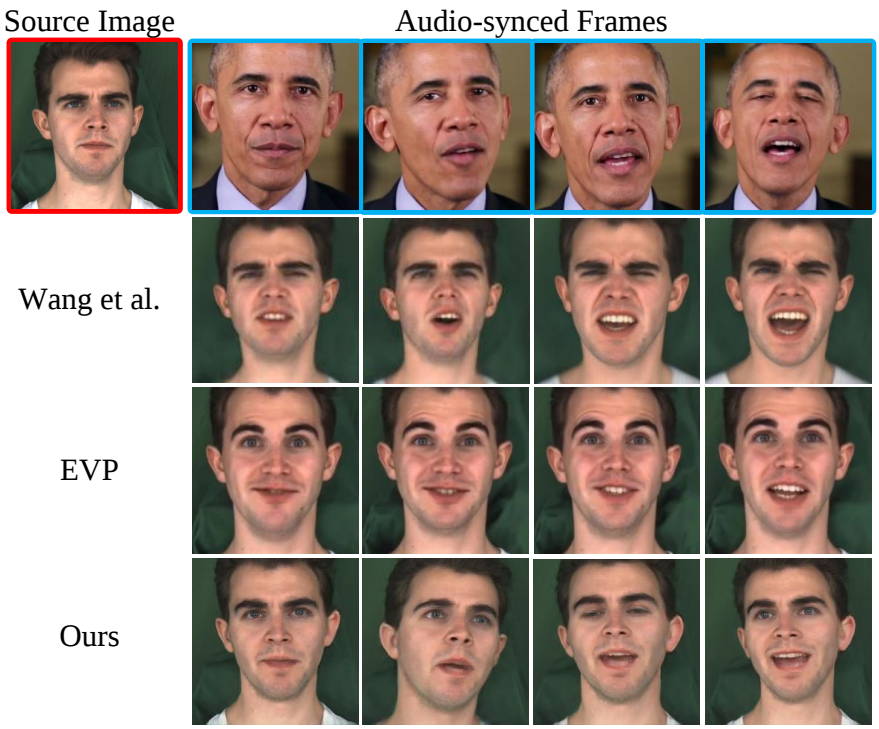


Figure 4: Qualitative comparisons of DIRFA with Wang et al.'s method [18] and EVP [17] on MEAD [18]: The audio-synced frames provide ground-truth lip shapes. The *subject-specific training* in Wang et al.'s method [18] and EVP [17] helps to generate sharper facial texture, while DIRFA can generate better lip-sync and more consistent facial expressions with respect to the audio-synced frames as well as natural head movements.

We further compare DIRFA with two state-of-the-art subject-specific audio-driven emotional talking face generation methods, including Wang et al. [18] and EVP [17], both of which are trained on the MEAD dataset [18]. It is worth mentioning that DIRFA is trained on Voxceleb2 [34] only and directly applied to MEAD images without finetuning. As Fig. 4 shows, the subject-specific training in Wang et al.'s method [18] and EVP [17] helps to generate talking faces with sharper texture, but their lip shapes and facial expressions are not consistent with the audio-synced frames and the generated head poses are static. In contrast, DIRFA generates talking faces with better lip-sync and

Table 1: Quantitative comparisons of DIRFA with state-of-the-art audio-driven talking face generation approaches on datasets Voxceleb2 [34] and LRW [35]. N/A in EBR means eye blinks are completely static in the synthesized talking faces.

Dataset	Methods	PSNR ↑	ACD ↓	LMD ↓	S_{conf} ↑	S_{off} ↓	EBR
Voxceleb2 [34]	ATVG [11]	28.79	0.376	**4.28**	5.45	0.57	N/A
	MakeItTalk [13]	29.15	**0.320**	6.13	5.11	1.08	0.39
	PC-AVS [7]	29.42	0.354	4.67	5.39	0.36	N/A
	Ours	**29.56**	0.331	4.45	**5.82**	**0.25**	0.31
LRW [35]	ATVG [11]	29.64	0.303	3.71	4.47	0.66	N/A
	MakeItTalk [13]	29.93	**0.274**	6.62	3.56	0.82	0.27
	PC-AVS [7]	30.65	0.288	3.35	6.21	**0.33**	N/A
	Ours	**30.97**	0.280	**3.16**	**6.39**	0.45	0.24

Table 2: Quantitative comparisons of DIRFA with Wang et al.'s method [18] and EVP [17] on MEAD [18] dataset. N/A in EBR means eye blinks are completely static in the synthesized talking faces.

Methods	PSNR ↑	ACD ↓	LMD ↓	S_{conf} ↑	S_{off} ↓	EBR
Wang et al. [18]	28.67	0.366	11.94	2.38	1.70	N/A
EVP [17]	**28.90**	**0.307**	8.23	4.21	0.88	N/A
Ours	28.42	0.391	**6.45**	**5.09**	**0.63**	0.35

more consistent facial expressions as well as natural head motions. In addition, DIRFA is more flexible and generalizable, which can be applied to different persons as shown in Fig. 3.

4.3. Quantitative Evaluation

Evaluation Metrics: We perform quantitative evaluations with several metrics that have been widely adopted in existing audio-driven talking face generation studies. Specifically, we use peak signal-to-noise ratio (**PSNR**) to evaluate the generation quality and average content distance (**ACD**) [38] to

Table 3: User study in mean opinion scores: Users rate videos from 1 to 5. Larger scores mean better performance. Subject-inde. and subject-spec. mean subject-independent and subject-specific, respectively.

	Methods	Video Realness	Lip Sync Quality	Expression Naturalness
Subject-inde. Methods	Ground Truth	4.83	4.75	4.64
	ATVG [11]	1.49	2.68	1.35
	MakeItTalk [13]	2.07	2.11	1.84
	PC-AVS [7]	3.16	3.53	2.80
	Ours	**3.95**	**4.02**	**3.76**
Subject-spec. Methods	Ground Truth	4.89	4.92	4.78
	Wang et al. [18]	2.57	1.33	1.16
	EVP [17]	3.25	3.64	3.12
	Ours	**3.68**	**3.91**	**3.50**

measure the facial identify preservation quality. We also adopt the landmark distance (**LMD**) around mouths [11], confidence score ($\mathbf{S}_{conf}$) and synchronization offset ($\mathbf{S}_{off}$) as described in [35] to measure the accuracy of mouth shapes and lip sync. Note the $\mathbf{S}_{off}$ here refers to the absolute difference between video-audio lags of the generated and audio-synced videos. In addition, we measure the eye blinking rate (**EBR**) [16] of the generated faces which is more realistic when it is similar to the average human eye blinking rate around 0.28-0.45 blinks per second [39].

Quantitative Results: Table 1 shows quantitative comparisons of several subjective-independent audio-driven talking face generation methods. We can see that DIRFA achieves the best PSNR on both datasets, indicating that the DIRFA synthesized faces are sharper and more realistic than those generated by other methods. In addition, DIRFA achieves the best LMD, S_{conf} and S_{off} in most cases, showing that DIRFA can generate accurate lip-sync videos from audio signals. Note DIRFA obtains slightly lower ACD than MakeItTalk [13],

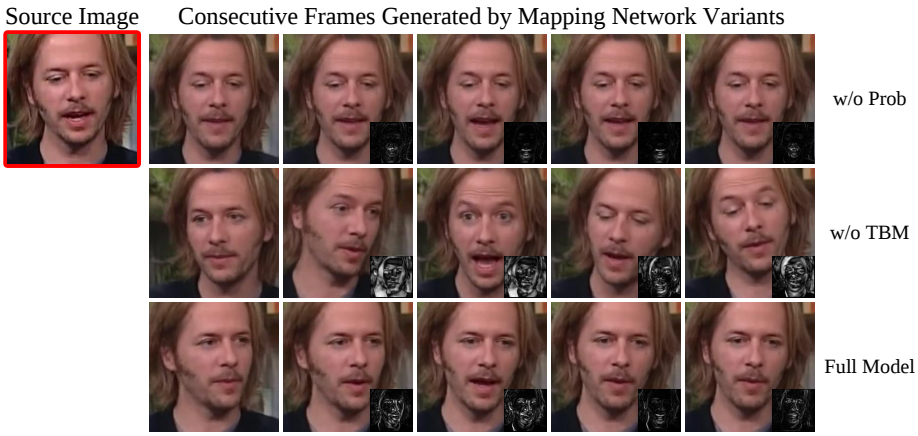


Figure 5: Qualitative results of ablation study: The probabilistic design (Prob) and the temporally-biased mask (TBM) help to model audio-to-visual uncertainty and generate temporally coherent facial animations, respectively. The L1 difference maps between the currently and the previously generated frames are illustrated in the bottom right.

largely because MakeItTalk [13] deals with a much simpler task of generating limited facial movements while DIRFA generates various face poses and facial expressions. Further, only MakeItTalk [13] and DIRFA can generate natural eye blinking but DIRFA can generate more realistic facial animations as illustrated in Fig. 3.

We also compare DIRFA with subjective-specific audio-driven talking face generation methods. As Table 2 shows, [18] and [17] can achieve better PSNR and ACD as they conduct subject-specific training. However, their generated lip shapes are not well aligned with the audio which leads to low LMD, S_{conf} and S_{off}. In addition, their generated eye blinks are static, which impairs the realism of the synthesized talking faces. As a comparison, DIRFA can synthesize more realistic facial animations with better lip-sync.

4.4. User Study

We evaluate and benchmark the DIRFA generated videos by conducting Amazon-Mechanical-Turk (AMT) user studies. The subjects are presented by

Table 4: Quantitative results of ablation study: Prob denotes the probabilistic design; TBM denotes the temporally-biased mask; N/A in EBR means eye blinks are static in the faces.

Methods	PSNR ↑	ACD ↓	LMD ↓	S_{conf} ↑	S_{off} ↓	EBR
w/o Prob	29.52	**0.318**	4.71	5.60	0.43	N/A
w/o TBM	29.38	0.347	6.22	4.39	0.74	0.59
Full Model	**29.56**	0.331	**4.45**	**5.82**	**0.25**	0.31

Table 5: Effect of different number of discrete categories D on the accuracy of generated lip shapes on Voxceleb2 [34].

D	10	100	250	500	750
LMD ↓	9.31	5.92	4.77	**4.45**	4.68

randomly-ordered videos synthesized by DIRFA and the compared methods (each method generates 3 videos). They are tasked to rate video quality (from 1 to 5) based on three criteria: 1) the realness of the video; 2) the lip sync quality; and 3) the naturalness of facial expressions. A total of 83 AMT users participated in the study and provided valid questionnaires. Table 3 shows the mean opinion scores. DIRFA consistently outperforms the competing methods under all metrics, demonstrating the effectiveness of our method. Please refer to the supplementary materials for the designed questions.

4.5. Discussion

Ablation Studies: We conduct ablation studies to analyse the effectiveness of the proposed mapping network in realistic talking face generation with three variants. The first replaces the discrete category facial animation representations with continuous representations ('w/o Prob'). The second replaces the proposed temporally-biased mask with the original mask in [8] ('w/o TBM'). The third is our full model ('Full Model'). The facial animation sequences produced by different mapping networks are fed to the same generation network to generate talking faces.

Fig. 5 shows the qualitative results, where each row shows five consecutive

generated frames and the L1 difference maps between the currently and the previously generated frames are illustrated in the bottom right. Without the probabilistic design, the mapping network tends to learn a deterministic mapping between audio and facial animations, which suffers from regression-to-mean problem and generates static and expressionless faces, resulting in minimal L1 differences between consecutive frames. Without the temporally-biased mask, the mapping network produces facial animation sequences without considering the temporal dependency of facial animations, leading to generating faces with abrupt head poses and very large L1 differences. Including the probabilistic design and temporally-biased mask, our full model generates temporally smooth talking faces with natural facial animation transition.

Table 4 shows the quantitative evaluation results, our full model achieves the best under most metrics. The mapping network without probabilistic design obtains better ACD as its generated faces have limited variations in head poses and expressions. In contrast, our full model generates natural talking faces with much more rich variations in head poses and expressions. As reported in [40], the rich variations can influence the visual perception of person identities and lead to degraded ACD.

Discrete Category Analysis: To allow sampling diverse yet realistic facial animations at inference, we represent facial animations as discrete categories and formulate the output of the mapping network as a probability tensor of categorical distribution. Specifically, we uniformly discretize facial animations into D constant intervals. We conduct an ablation study to examine the impact of the parameter D on the accuracy of lip shapes in the generated talking faces. In the experiments, we adopt the LMD metric (lower LMD indicates better lip shape accuracy) and examine D over the Voxceleb2 benchmark [34].

As Table 5 shows, the accuracy of lip shapes degrades clearly when D is very small (e.g., 10), largely because a small D restricts possible states of facial animations and leads to inaccurate and unnatural lip movements. While increasing D gradually, we first observe a drop of LMD scores and then a convergence to

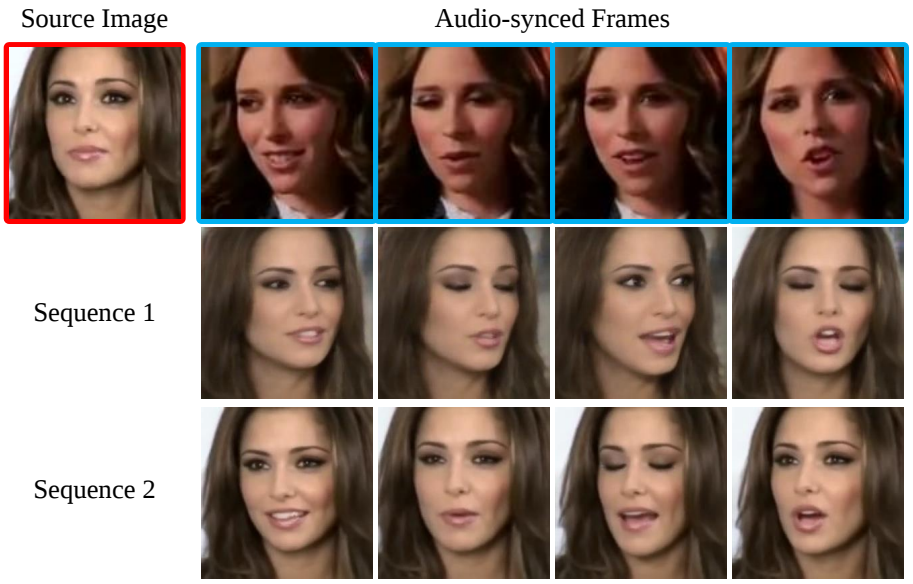


Figure 6: Diverse talking faces generation: By randomly sampling facial animation sequences from the same audio, DIRFA can synthesize talking faces with *accurate* and *consistent* lip shapes with respect to the audio-synced frames as well as *diverse* and *realistic* expressions and head poses as illustrated in Rows 2 and 3.

Table 6: Comparisons of parameter number of different methods.

Method	ATVG [11]	MakeItTalk [13]	PC-AVS [7]	Wang [18]	EVP [17]	Ours
No. of Params	88.5M	72.9M	152.3M	121M	403.3M	112.2M

a plateau while D is around 500. Further increasing D after that degrades LMD scores. One plausible reason is that too many fine-grained categories introduce ambiguity during model optimization and increase the complexity of the prediction task. We therefore use $D = 500$ as the default setting.

Comparison of Model Complexity: We compare the parameter number of different audio-driven talking face generation methods in Table 6. Our model has 112.2M parameters, including 11.2M from the mapping network and 101M from the generation network. Its parameter counts are comparable to that of ATVG [11], MakeItTalk [13] and Wang et al. [18] but clearly lower than PC-

AVS [7] and EVP [17].

Diverse Talking Faces Generation: DIRFA can generate diverse talking faces from the same driving audio. Thanks to our designed probabilistic mapping network, we can easily obtain diverse facial animation sequences by randomly sampling facial animations from the predicted likelihood (as discussed in section 3.2.2). The generation network can then generate diverse talking faces with the same source image plus diverse animation sequences as shown in Fig. 6.

Ethical Considerations: With the convenience of generating talking faces from audio and a single source image, our method may be misused by immoralists to spread misinformation. To avoid improper use, we will include a watermark to the generated videos making it clear that they are synthetic.

5. Conclusion

This paper presents DIRFA, an audio-driven talking face generation method that can generate talking faces with diverse yet realistic facial animations from the same driving audio. We design a transformer-based probabilistic mapping network to model the uncertain relations between audio and visual signals and introduce a temporally-biased mask into the mapping network to convert audio into temporally smooth facial animations in an autoregressive manner. Our generation network then takes the generated facial animations and a source image as input to synthesize talking faces. Extensive experiments show that DIRFA can generate talking faces with accurate lip movements, vivid facial expressions and natural head poses.

The proposed method effectively addresses the audio-to-visual mapping problem with a transformer-based probabilistic network, which could be applied to various tasks such as audio-driven dance synthesis or co-speech gesture generation. However, the current design generates visual signals from audio via a fully automatic pipeline. It remains an open challenge to incorporate user interaction for controlling certain desired synthesized facial animations. We will explore it in our future work.

6. Acknowledgments

This work is funded by the Ministry of Education, Singapore, under the Tier-1 Project RG94/20.

References

[1] J. Cassell, D. McNeill, K.-E. McCullough, Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information, Pragmatics & cognition 7 (1) (1999) 1–34.

[2] B. Fan, L. Wang, F. K. Soong, L. Xie, Photo-real talking head with deep bidirectional lstm, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4884–4888.

[3] L. Chen, Z. Li, R. K. Maddox, Z. Duan, C. Xu, Lip movements generation at a glance, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 520–535.

[4] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484–492.

[5] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, C. Xu, Talking-head generation with rhythmic head motion, in: European Conference on Computer Vision, Springer, 2020, pp. 35–51.

[6] R. Yi, Z. Ye, J. Zhang, H. Bao, Y.-J. Liu, Audio-driven talking face video generation with learning-based personalized head pose, arXiv preprint arXiv:2002.10137 (2020).

[7] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, Z. Liu, Pose-controllable talking face generation by implicitly modularized audio-visual representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4176–4186.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[9] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing obama: learning lip sync from audio, ACM Transactions on Graphics (ToG) 36 (4) (2017) 1–13.

[10] Y. Song, J. Zhu, D. Li, X. Wang, H. Qi, Talking face generation by conditional recurrent adversarial network, arXiv preprint arXiv:1804.04786 (2018).

[11] L. Chen, R. K. Maddox, Z. Duan, C. Xu, Hierarchical cross-modal talking face generation with dynamic pixel-wise loss, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7832–7841.

[12] H. Zhou, Y. Liu, Z. Liu, P. Luo, X. Wang, Talking face generation by adversarially disentangled audio-visual representation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 9299–9306.

[13] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, D. Li, Makelttalk: speaker-aware talking-head animation, ACM Transactions on Graphics (TOG) 39 (6) (2020) 1–15.

[14] N. Liu, T. Zhou, Y. Ji, Z. Zhao, L. Wan, Synthesizing talking faces from text and audio: an autoencoder and sequence-to-sequence convolutional neural network, Pattern Recognition 102 (2020) 107231.

[15] S. Wang, L. Li, Y. Ding, C. Fan, X. Yu, Audio2head: Audio-driven one-shot talking-head generation with natural head motion, arXiv preprint arXiv:2107.09293 (2021).

[16] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, X. Guo, Facial: Synthesizing dynamic talking face with implicit attribute learning,

in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3867–3876.

[17] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, F. Xu, Audio-driven emotional video portraits, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14080–14089.

[18] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, C. C. Loy, Mead: A large-scale audio-visual dataset for emotional talking-face generation, in: European Conference on Computer Vision, Springer, 2020, pp. 700–717.

[19] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, J. Wang, Expressive talking head generation with granular audio-visual control, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3387–3396.

[20] E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9459–9468.

[21] Y. Ren, G. Li, Y. Chen, T. H. Li, S. Liu, Pirenderer: Controllable portrait image generation via semantic neural rendering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13759–13768.

[22] O. Wiles, A. Sophia Koepke, A. Zisserman, X2face: A network for controlling face generation using images, audio, and pose codes, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 670–686.

[23] P. Ekman, W. Friesen, J. Hager, Facial action coding system (facs) a human face, Salt Lake City (2002).

[24] R. Li, S. Yang, D. A. Ross, A. Kanazawa, Learn to dance with aist++: Music conditioned 3d dance generation, arXiv e-prints (2021) arXiv–2101.

[25] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.

[28] P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12873–12883.

[29] R. Girdhar, J. Carreira, C. Doersch, A. Zisserman, Video action transformer network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 244–253.

[30] E. Aksan, M. Kaufmann, P. Cao, O. Hilliges, A spatio-temporal transformer for 3d human motion prediction, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 565–574.

[31] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, Z. Liu, Bailando: 3d dance generation by actor-critic gpt with choreographic memory, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11050–11059.

[32] O. Press, N. A. Smith, M. Lewis, Train short, test long: Attention with linear biases enables input length extrapolation, arXiv preprint arXiv:2108.12409 (2021).

[33] T.-C. Wang, A. Mallya, M.-Y. Liu, One-shot free-view neural talking-head synthesis for video conferencing, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10039–10049.

[34] J. S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, arXiv preprint arXiv:1806.05622 (2018).

[35] J. S. Chung, A. Zisserman, Lip reading in the wild, in: Asian conference on computer vision, Springer, 2016, pp. 87–103.

[36] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.

[37] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.

[38] S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, Mocogan: Decomposing motion and content for video generation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1526–1535.

[39] A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali, A. Albanese, Analysis of blink rate patterns in normal subjects, Movement disorders 12 (6) (1997) 1028–1034.

[40] F. A. Pavel, E. Iordănescu, The influence of facial expressions on recognition performance in facial identity, Procedia-Social and Behavioral Sciences 33 (2012) 548–552.