Talking Face Generation via Face Mesh - Controllability without Reference Videos

Ali Köksal, Qianli Xu and Joo-Hwee Lim
*Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore*
Email: {koksal_ali, qxu, joohwee}@i2r.a-star.edu.sg

***Abstract*—Recent development in audio-driven talking face generation strives for controlling facial features including facial expression, head pose, eye blink, *etc*. as well as accurate lip-synchronization and the ability to apply to arbitrary subjects. Existing audio-visual models that can control facial features require encoders that encode driving videos, which is both computationally expensive and limited by the availability of such driving videos. In this paper, we address this limitation and aim to control facial features without encoding driving videos. We propose a cascaded GAN-based audio-visual model, which incorporates face mesh as an intermediate representation. Different from existing cascaded methods that use facial landmarks, our method uses face mesh as a medium of informative facial feature representation. To the best of our knowledge, this is the first cascaded model that allows controllable talking face generation via face mesh. We train our audio-visual model with training samples of MEAD dataset. In the evaluation, we benchmark our model in extensive experiments on MEAD and LRW datasets. The results show our model outperforms existing ones by generating high-fidelity audio-driven talking faces on arbitrary subjects with realistic emotional expression patterns.**

***Index Terms*—talking face generation, facial animation, controllable generation**

I. Introduction

Generating talking head videos is a challenging task and it has a wide range of applications in entertainment, education, healthcare, and communication industries. The main technical issue is to generate realistic and expressive videos with high fidelity and synchronized lip motions. Meanwhile, researchers are paying more attention to the controllability of facial features, *i.e.* being able to change specific semantics/motions based on control signals. Controllable video generation can boost the flexibility of product deployment and enhance user experience by creating engaging talking heads [1], [2].

With the latest advancements in audio-driven talking face generation [1], [3], [4], the controllability of facial features (*e.g.* emotional expressions, head pose, eye blink) is achieved while ensuring precise lip-synchronization and fidelity. Many audio-driven models adopt an implicit control mechanism, *i.e.* they disentangle the driving audio to control different aspects of facial features [5]–[8]. This mechanism has shown some success in controlling lip motion owing to the high correlation between lip motion and the driving audio. However, it is less effective to control other features (*e.g.* head pose and emotion) because the correlation between these facial features and the audio is weak. Some works propose to generate talking head videos with explicit control over eye blink [9], [10]. However, eye blink is visible only within a minor facial area, so it has little effect on user perception. In essence, audio-driven models offer limited controllability over the aspects of facial features.

Another stream of works adopts driving videos to exert control over facial features, typically called face reenactment [1], [3], [4]. These audio-visual models utilize encoders to process driving videos to manipulate the facial features of the generated faces. These models are end-to-end trainable and allow for explicit control, provided that they can effectively extract disentangled information in the encoding process. However, such a practice is constrained by the content of the available driving videos, *i.e.* one needs to find suitable driving videos to achieve the desired effect of control.

To alleviate this problem, we attempt to explicitly control emotional expression in generated talking faces without encoding driving videos (Fig. 1). We propose a new cascaded model that leverages face landmarks as an intermediate representation, which serves to disentangle the face identity (appearance) and the face dynamics (motion). Facial landmarks are 2D or 3D points to localize salient regions of a face. Depending on the representation format, they range from sparse 2D points (*e.g.* 68 points) [11] that capture key features, such as face contours, eyes, eyebrows, nose, and mouth, to dense representations such as 3DMM using as many as 53,000 vertices [12]. With more condensed representation (as compared to pixel-based images), facial landmarks are conducive to controllable generation. This study adopts face mesh - a type of face landmark with intermediate representation sparsity - to represent facial features, aiming to strike a favorable trade-off of data efficiency and informativeness.

A major challenge to face generation based on face mesh is how to ensure temporal consistency with a sparse representation (as compared to pixel-based approaches). Failing to do so leads to instability in generated videos. In this research, we divide the task into two sub-tasks, namely, talking face mesh generation and face mesh-to-face translation, which are implemented in a cascaded manner (as shown in Fig. 2). In the talking face mesh generation, we develop a face mesh generator to generate face meshes according to control signals (*e.g.* happy, sad, neutral) with the given face mesh of a reference identity and the driving audio (speech). Importantly, we propose a face mesh alignment procedure to tackle the stability issue. This is achieved by training the face mesh encoder and

This project is funded by the SMU-A*STAR Joint Lab Seed Grant under the Human-AI Synergy Pillar, Project No. C232918002.

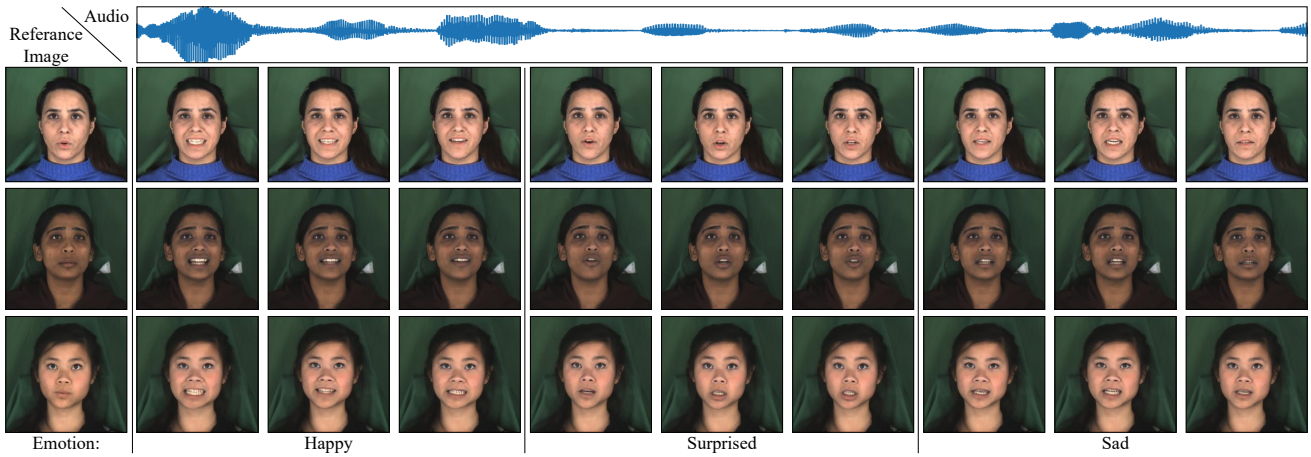

Fig. 1. Our audio-visual model generates realistic talking face videos of arbitrary faces on reference images with synchronized lips to audio and is capable of controlling emotional expression with a given emotion label. Note that the generated videos of different faces are conditioned on the same audio clip. Results with all the emotions can be found in our supplementary video.

generator to reconstruct target face meshes with an aligned reference face mesh. In the face mesh-to-face translation, our audio-visual model employs a conditional GAN to translate talking face meshes into talking faces. It includes an encoder decoder-based generator and a discriminator that are trained adversarially. The generator takes as input the concatenated embeddings of the generated face mesh and the reference image of the target identity and generates a realistic face that mimics the face mesh.

The contributions can be summarized in three-fold. **(1)** The proposed model generates audio-driven realistic talking head videos that can control emotional expression explicitly without encoding driving reference videos. **(2)** The proposed model is the first cascaded model that incorporates face meshes into two orchestrated sub-tasks. To the best of our knowledge, this is the first cascaded model that can control facial features (*i.e.* emotional expression) other than lip motion. **(3)** With the evaluations on two well-known datasets (MEAD [13] and LRW [14]), we show that our model generates realistic talking head videos with explicit control of emotional expressions.

II. Related Work

According to the source of lip motion, talking head generation can be broadly classified into two categories: (1) text-driven and (2) audio-driven. Text-driven talking head generation methods [15], [16] are usually based on phonemes, which are the unit of sound structure extracted from the given text. Audio-driven talking head generation methods usually process audio as spectrogram [1], [3], [17]–[19] or acoustic features [20]–[23]. One of the well-known audio-driven talking face generation models is Wav2Lip [24] which utilizes a powerful Lip sync expert. Some studies [1], [2], [25]–[27] utilize both the driving text and the driving audio. For example, [25] aims to build language robust talking face generation by using audio and phonemes. TalkLip [28] focuses on reading intelligibility and employs a lip-reading expert that transcribes videos to text to improve generated lip motions. Recently, DiffTalk [23] and Diffused Heads [29] have employed diffusion models for audio-driven face generation.

Regarding the controllability of facial features such as head pose [3], [30], emotion [22], eye blink, or a combination of them [1], [2], [4], [18], [31], existing methods usually use driving videos along with the driving audio. Some of the approaches utilize single-driving video to control a single feature other than lip motion. PC-AVS [3] is the first approach that uses driving video to control the pose of the generated talking head. EAMM [22] can generate realistic emotion-aware talking heads based on reference images, driving audio, and driving videos. AVFR-GAN [18] reenacts a reference image based on driving audio and video to control facial features. StyleTalk [2] generates style-controllable talking faces with a driving video that depicts the styles. Some works utilize more than one driving video and focus on the disentanglement of facial features to achieve fine control over multiple facial features. For example, GC-AVT [4] generates audio-driven talking face videos with controllable expressions and head poses with driving videos. PD-FGC [1] controls facial features with the disentanglement of lip motions, eye gaze and blink, head pose, and emotional expression.

In view of the computational complexity of controlling facial features through encoding driving videos, some studies propose controlling facial features without driving videos. However, they either control features by inferring from driving audio or exhibit limited control power on facial features [5]–[10]. In [5], head poses are inferred from the given audio to achieve personalized head poses. In addition to head pose, FACIAL [6] also generates realistic eye blinks by learning implicit facial features from driving audio along with lip motion. In [7], a cascaded method is proposed to disentangle audio as content and emotion, whereby emotional expressions are controlled by the disentangled emotion features. In [8], an end-to-end trainable model is built to control emotional expression by using driving audio.

Similar to our model, some works [2], [7], [9], [10], [16], [27], [32]–[34] incorporate facial landmarks as an intermediate medium. For example, MakeItTalk [32] is a landmark-based model that predicts facial landmarks from driven audio and then translates predicted landmarks to face images. Canonical landmarks are used as an intermediate representation, so as to generate talking head videos with spontaneous eye blinks [10]. SadTalker [9] generates eye-blink controllable talking head videos via a cascaded model based on 3DMM and a 3D-aware face renderer. Existing landmark-based models usually have two limitations: (1) they are often applied to the faces of specific persons [16], [27], and (2) they show limited power in controlling facial features [16], [27], [32]. For example, although [9], [10] can control eye blink as one of the other facial features explicitly, eye blink is a minor facial motion that is visible on only a small portion of the face. Hence, it does not affect other parts of the generated faces and is not related to lip motion. In comparison, our face-landmark-based model is capable of generating videos of an arbitrary face and exerting effective control over emotional expression.

III. Approach

Fig. 2 shows the overview of our audio-visual model that generates realistic talking face videos of an arbitrary subject with lips synchronized to the audio. Our model is capable of controlling emotional expression explicitly. By incorporating face mesh, we design a cascaded model consisting of two sub-tasks: face mesh generation and face mesh-to-face translation. In the face mesh generation, our model includes two encoders to encode audio clips and the reference face meshes and a generator to generate face meshes according to the conditioned control signals. In the face mesh-to-face translation, our model employs a GAN with a generator and a discriminator.

A. Disentanglement

As mentioned earlier, we aim to disentangle the facial features without requiring video encoders as much as possible, so as to alleviate the need of driving video at the inference time. In so doing, we leverage the facial landmarks to disentangle the face identity (appearance) and the face dynamics (motion). Motions related to the emotional expression and the lip motion are disentangled by training the face mesh generator to reconstruct face meshes conditioned on the emotion label and the latent vector of audio.

Since the space of emotional expression is well-defined, controlling the expression with the emotion label is suitable and straightforward. Our model employs the categorical model, which describes emotions with a set of emotion labels (*e.g.* happy, surprised, sad, *etc.*). For lip motion control, we employ an audio encoder to learn the space of lip motion and audio and compute latent vectors for the driven audio.

B. Talking Face Mesh Generation

Public datasets usually do not provide face mesh information. To address this issue, we resort to an off-the-shelf model, called CVZone, to extract face mesh information. The facial landmark model consists of 468 3D points. It should be noted that our method is agnostic to the face mesh extraction model.

For the talking face mesh generation, our model employs two encoders and a generator. Encoders compute latent vectors for the driven audio and the given reference face meshes. The encoded latent vectors are concatenated with the emotion label of the target expression. Then, the concatenated control signals are given to the generator as a condition to synthesize face meshes that depict the given control signals.

The audio encoder is trained to compute latent vectors for the driven audio. We leverage the audio encoder proposed in [3]. First, the raw audios are converted into spectrograms in 2D time-frequency space. Next, they are given to the audio encoder as input to build the space of the audio spectogram and lip motion of face meshes.

A major challenge to talking face generation via face mesh is how to ensure stability, which refers to generating faces with temporal consistency. Inspired by SadTalker [9] whıch tackles instability by ExpNet to generate intermediate representation based on 3DMM, we train our model to generate blocks of c consecutive frames. While the model is capable of generating temporarily consistent c frames, we observe instability between blocks of frames. To mitigate this issue, we propose an additional alignment step whereby the reference face mesh is aligned with the first target face mesh of the c consistent frames with the homography. Next, it is given to the face mesh encoder to compute a latent vector that guides the generator to generate face meshes that maintain inter-block consistency. At the inference time, the alignment is not necessary since the face mesh of the reference image is given directly and the generated face meshes are expected to be aligned to the face mesh of the reference image. In addition to the instability between blocks of consistent frames, face mesh alignment improves the controllability of lips and expression as it disentangles the head pose motion and the facial motions related to expression and lips. It is further explored in the following sub-section and Section IV-C with an ablation study.

C. Talking Face Generation

For the talking face generation, our model employs a conditional image-to-image translation GAN similar to [32] with a generator and a discriminator. The generator takes face meshes as a condition and it is concatenated with the reference frame. A GAN model is trained to generate target frames and trick the discriminator by generating realistic frames. The discriminator is trained to distinguish real and generated frames. Hence, the generator and the discriminator are trained adversarially.

At inference time, the generator translates the generated face mesh and generates a face with the appearance of the identity on the given reference frame. Thus, the generated face is a variant of the given identity that mimics the generated mesh.

D. Training

N training video clips with K-frame videos, $\{V_1 = \{I_{(1,1)}, ..., I_{(1,K)}\}, ..., V_N = \{I_{(N,1)}, ..., I_{(N,K)}\}\}$ are accompanied by audio $\{A_1 = \{a_{(1,1)}, ..., a_{(1,K)}\}, ..., A_N =$

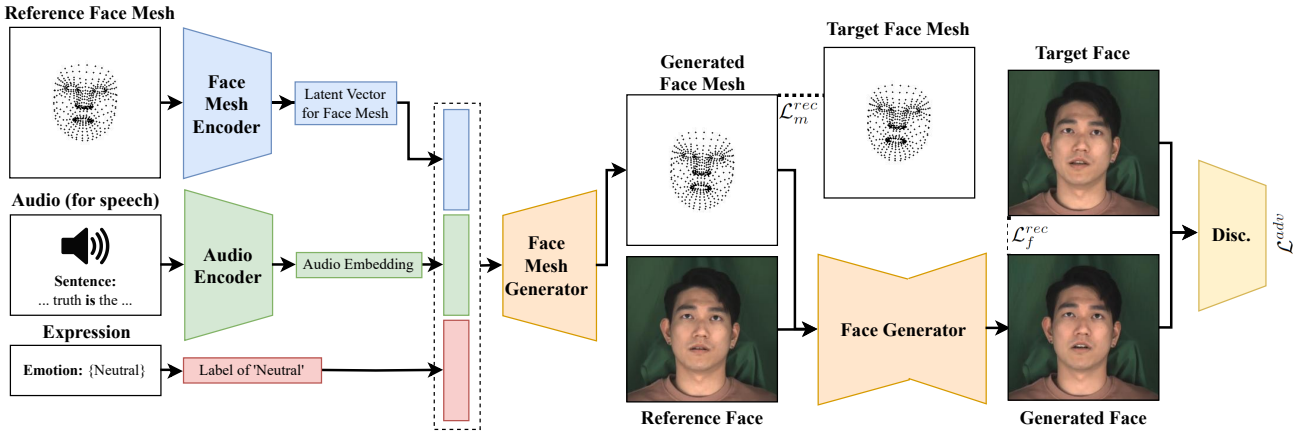

Fig. 2. Overview of the proposed model. At the inference time, the model generates the face mesh conditioned on the face mesh of the reference image, driving audio, and the target emotion label, and then the generated face mesh and the reference image are given to the face generator as conditions to generate the face that mimics the generated face mesh.

$\{a_{(N,1)}, ..., a_{(N,K)}\}\}$ that is processed into spectogram. Video clips depict the head videos of different subjects with different emotions. The annotation of training video clips is extended by extracting face mesh $m_{(n,k)}$ for each frame. The goal is to generate any target face mesh $m^t_{(n,k)}$ and any target frame $I^t_{(n,k)}$ that depicts the talking head of target identity s^t with driving target audio $a^t_{(n,k)}$ and the target emotion label e^t.

The face mesh generation model is conditioned on a randomly selected reference face mesh $m^r_{(n,k)}$ where $n \in [1, ..., N]$ and $k \in [1, ..., K]$. $m^r_{(n,k)}$ is a randomly selected face mesh of the same subject s_t with an arbitrary emotion. So, the target audio and the target emotion label can be different from the audio and emotion label of the reference face mesh as it is selected randomly from a set that consists of every video clip of the same subject. Hence, the face mesh generator is trained to generate lip motions and expressions that are not identical to those in the given reference face mesh.

Since reference face meshes are selected randomly, they are not temporally consistent with the target face mesh and there exists an arbitrary head pose motion (possibly undesired) along with the facial motion. So, giving the reference face mesh directly leads to instability in the generated talking face video and confuses the face mesh generator as it consists of head pose motion and facial motion. To solve this problem, reference face meshes are aligned by wrapping the reference face mesh with computed homography between the reference and the target face mesh.

During the training, the model is trained to reconstruct every available face mesh with aligned reference face meshes. So, the reconstruction loss (Equation 1) is the main leading loss. Audio encoder (E_a), face mesh encoder (E_m), and the face mesh generator (G_m) are trained to minimize the following Mean Squared Error (MSE) as reconstruction loss:

$$\mathcal{L}_m^{rec} = \|m^t_{(n,k)} - m^g_{(n,k)}\|_2, \quad (1)$$

where $m^t_{(n,k)}$ denotes the target face mesh. For a set of the inputs (the target face mesh $m^t_{(n,k)}$, driving audio $a^t_{(n,k)}$, emotion label e^t), $m^g_{(n,k)}$ denotes the generated face mesh and formulated as follows:

$$m^g_{(n,k)} = G_m(E_a(a^t_{(n,k)}), E_m(\omega(m^r_{(n,k)})), e^t), \quad (2)$$

where ω is a wrapping operation. The wrapping operation is omitted at test time and inference time.

The talking face generator (G_f) is conditioned on a target face mesh ($m^t_{(n,k)}$) and trained to translate reference frame ($I^r_{(n,\hat{k})}$) to target frame ($I^t_{(n,k)}$). Note that although $m^t_{(n,k)}$ is a set of 2D points in the talking face mesh generation, it denotes a binary image that represents the same face mesh in the talking face generation. $I^r_{(n,\hat{k})}$ (where $\hat{k} \in [1, ..., K]$) is a randomly selected frame from the same video clip as the target frame, considering that the outfit and the hairstyle of subjects might vary from video clip to video clip. $\hat{k}$ is selected as 1 at test time. Similar to face mesh generation, a reconstruction loss is computed based on the target frame ($I^t_{(n,k)}$) and the generated frame ($I^g_{(n,k)}$), which consists of two components: (i) MSE loss (ii) L1 norm based on VGG features [35].

$$\mathcal{L}_f^{rec} = \|I^t_{(n,k)} - I^g_{(n,k)}\|_2 + \|\phi(I^t_{(n,k)}) - \phi(I^g_{(n,k)})\|_1, \quad (3)$$

where ϕ denotes concatenated features of pretrained VGG19 network. Since the training of the talking face mesh generation and the face mesh-to-face translation are independent. The generated frame is formulated as follows:

$$I^g_{(n,k)} = G_f(I^r_{(n,\hat{k})}, m^t_{(n,k)}). \quad (4)$$

At inference time, $m^t_{(n,k)}$ is replaced by $m^g_{(n,k)}$.

In addition to reconstruction loss, the talking face generator is also trained with an adversarial loss, computed based on the discriminator's ability to distinguish real and generated frames. Least squared adversarial loss is used for adversarial training and the generator is trained to optimize the following to generate indistinguishable frames from real frames:

$$\mathcal{L}_G^{adv} = (D(I^g_{(n,k)}))^2. \quad (5)$$

Hence, the full objective function of the generator consists of the reconstruction loss and the adversarial loss and is formulated as follows:

$$\mathcal{L}^G = \lambda_{rec}\mathcal{L}_f^{rec} + \lambda_{adv}\mathcal{L}_G^{adv}, \tag{6}$$

where λ_{rec} and λ_{adv} balance the loss functions with default values 100.0 and 1.0, respectively.

The discriminator (D) is trained to optimize the following adversarial loss to distinguish real and generated frames:

$$\mathcal{L}_D^{adv} = \frac{1}{2}\left[D(I_{(n,k)}^t))^2 + (D(I_{(n,k)}^g) - 1)^2\right]. \tag{7}$$

E. Implementation Details

The talking face mesh generation (face mesh encoder, audio encoder, and face mesh generator) and the face mesh-to-face translation are trained independently, but this does not harm our model's ability to work end-to-end fashion at inference time. During training, all networks are trained from scratch with Adam optimizer for 500k iterations with batch size of 16, learning rate of 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$.

The face mesh generator is trained to generate 8 consecutive frames *i.e.* the default value of c is selected as 8 experimentally. Based on experiments, we found that 8 is an optimal number for generating stable face mesh videos. Unlike the face mesh generator, the face generator is trained to generate faces frame-by-frame.

IV. Experiments

We conduct experiments to compare our method with state-of-the-art methods including MakeItTalk [32], PC-AVS [3], PD-FGC [1], and SadTalker [9]. Their results are produced by using their publicly available pretrained models. In the evaluation, models are used to reconstruct test videos with driving audio and the first frame as the reference image. In addition, PC-AVS and PD-FGC are fed with ground truth test videos as they require driving videos, which gives them an extra advantage. In contrast, our model only requires the ground truth emotion label.

Datasets We train our model on multi-view emotional audio-visual dataset **(MEAD)** and evaluate it on both MEAD and lip reading in the wild **(LRW)** datasets. MEAD dataset [13] consists of high-quality audio-visual recordings of subjects speaking with 8 different emotions at 3 intensity levels. The dataset includes videos of more than 40 different actors, and we randomly select 6 subjects for evaluation. We use the highest intensity levels of emotions in the training of our model and the evaluation of models. LRW dataset [14] consists of 1,000 utterances of 500 different words with videos that last about 1 second. This dataset has no emotion annotation. Thus, we use its testing set for evaluation only by giving *neutral* as the emotion label.

A. Qualitative Results

Figures 3 and 4 compares our method with MakeItTalk [32], PC-AVS [3], PD-FGC [1], and SadTalker [9] on the test set of MEAD [13] and LRW [14] datasets, respectively. In addition to the generated results, the ground truth is also provided for comparing lip shape and identity. As shown in Figures 3 and 4, our method generates better talking faces than benchmark methods, with results closer to ground truth in terms of lip synchronization, identity preservation, and image quality. MakeItTalk's results have lower image quality. PC-AVS and PD-FGC cannot preserve the given identity well. SadTalker suffers from a lack of lip synchronization.

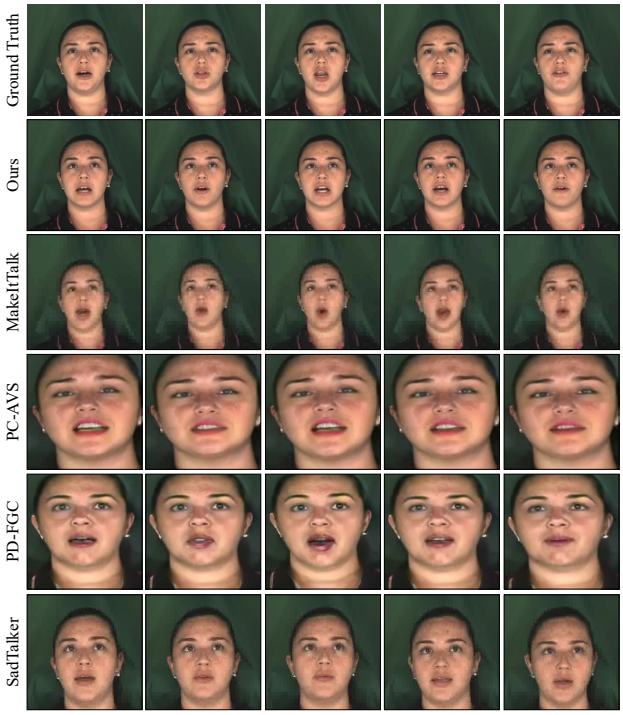

Fig. 3. Qualitative comparison on MEAD dataset [13]

B. Quantitative Results

Metrics we evaluate our method in terms of both video quality and lip synchronization based on 5 metrics that are widely used in recent works. For video quality, structural similarity index **(SSIM)** [36] and peak signal-to-noise ratio **(PSNR)** are used. Both SSIM and PSNR measure the similarity between real and generated frames. SSIM compares the outcome based on image patches whereas PSNR is a pixel-wise metric. For lip synchronization, the confidence score of SyncNet **(Conf.)** [14], facial landmark distance [37] on the mouth **(Lmd$_m$)** and the whole face **(Lmd)** are used. SyncNet measures the accuracy of mouth shape with driving audio. Facial landmark distances (Lmd$_m$, Lmd) are computed between the facial landmarks detected on the ground truth and generated frames. Note that although our method generates face mesh, the generated face meshes are not used for the computation of Lmd$_m$ and Lmd. Instead, facial landmarks are detected on the generated talking face videos separately.

Tables I and II show the results of our model and benchmark models on the test sets of MEAD and LRW datasets, respectively. Our model has the best scores in most metrics (except

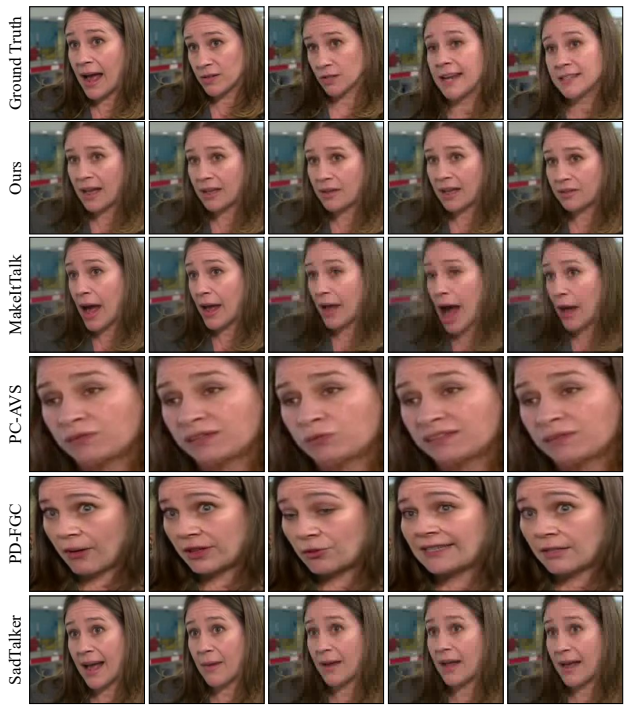

Fig. 4. Qualitative comparison on LRW [14]

Method	Video Quality		Lip Synchronization		
	SSIM ↑	PSNR ↑	Conf. ↑	Lmd_m ↓	Lmd ↓
MakeItTalk	0.537	18.061	0.469	0.803	0.549
PC-AVS	0.586	22.659	0.677	0.982	0.727
PD-FGC	0.357	16.693	**0.833**	1.176	1.314
SadTalker	0.501	17.414	0.448	0.826	1.768
Ours	**0.798**	**28.370**	0.596	**0.747**	**0.397**

TABLE I
COMPARISON OF OUR MODEL WITH MAKEITTALK [32], PC-AVS [3], PD-FGC [1], AND SADTALKER [9] ON MEAD [13] DATASET

for **Conf.**) with large margins in both datasets. Better **SSIM**, **PSNR**, **Lmd**$_m$, and **Lmd** show that our model is capable of generating talking face videos with higher quality in terms of video quality and lip synchronization. Specifically, our model achieves the best scores on Lmd$_m$ and Lmd, which means that our model can generate lip motions closer to the ground truth than benchmark models and it can generate realistic lip motions that are synchronized with the driving audio. Meanwhile, our method did not achieve the highest confidence score. This is because the confidence score of SyncNet is very sensitive to the audio, which may lead to a better score with unrealistic lip motions as discussed in [9]. In fact, it is argued that the confidence score of SyncNet need not be indicative of true lip-sync quality when a model outperforms ground truth's SyncNet score [19]. In our case, the SyncNet scores of ground truth are 0.620 and 1.503 in MEAD and LRW datasets, respectively, which are close to our model's score.

Method	Video Quality		Lip Synchronization		
	SSIM ↑	PSNR ↑	Conf. ↑	Lmd_m ↓	Lmd ↓
MakeItTalk	0.425	18.608	1.316	0.654	0.573
PC-AVS	0.409	19.571	1.551	0.619	0.676
PD-FGC	0.216	14.518	**1.698**	0.843	1.502
SadTalker	0.279	15.301	1.199	0.667	0.957
Ours	**0.615**	**20.987**	1.199	**0.450**	**0.365**

TABLE II
COMPARISON OF OUR MODEL WITH MAKEITTALK [32], PC-AVS [3], PD-FGC [1], AND SADTALKER [9] ON LRW [14] DATASET

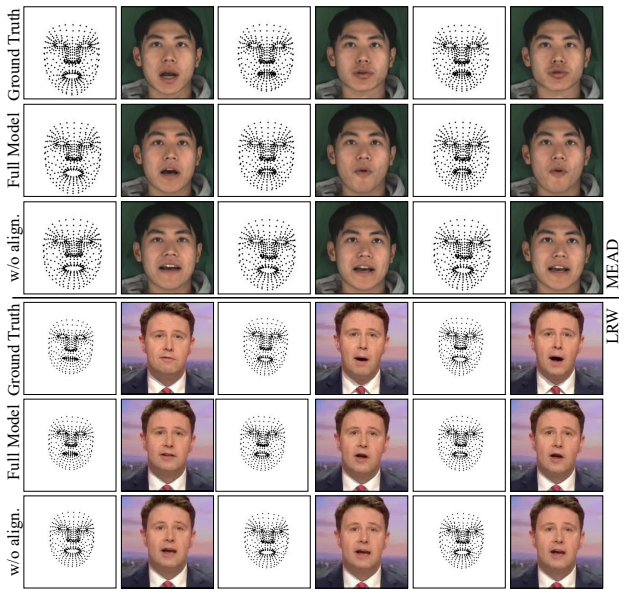

Fig. 5. Comparison of generated talking face videos of our full model with the ablation model. Note that the results are shown by zooming faces to show the details better.

C. Ablation Study

In the ablation study, we evaluate the effectiveness of the face mesh alignment with homography in training. The ablation model's face mesh generator is trained by omitting the face mesh alignment.

Method	Video Quality		Lip Synchronization			
	SSIM ↑	PSNR ↑	Conf. ↑	Lmd_m ↓	Lmd ↓	
w/o align.	0.731	26.832	0.454	0.788	0.473	MEAD
Full Model	0.798	28.370	0.596	0.747	0.397	
w/o align.	0.551	19.412	1.091	0.716	0.545	LRW
Full Model	0.615	20.987	1.199	0.450	0.365	

TABLE III
ABLATION STUDY FOR THE FACE MESH ALIGNMENT WITH HOMOGRAPHY.

Fig. 5 illustrates the generated face meshes and faces with the full model and the ablation model *(w/o align.)* with the ground truth on the test set of MEAD and LRW datasets. Although the desired lip motions (target lip motions on ground truth) are changing, the ablation model generates face meshes with the same lip motion. On the other hand, the full model is capable of generating lip motions similar to the ground truth. Moreover, Table III compares scores of the full model and the ablation model. Apparently, the full model achieves better scores than the ablation model in all metrics, indicating that face mesh alignment during training is effective. The main reason is that the face mesh alignment facilitates disentanglement of head pose motion and facial motion (motion related

to expression and lips), which in turn helps the model learn facial motions. In comparison, non-aligned input face meshes confuse the face mesh generator as it consists of the head pose motion and the facial motion together. Hence, the face mesh generator cannot relate the given inputs (audio and the emotion label) with the head pose motion.

V. Conclusion

In this paper, we propose a novel cascaded audio-visual model that can generate audio-driven talking faces with high fidelity on arbitrary subjects with realistic emotional expression patterns. Our model incorporates face mesh as an intermediate representation, which is the first to explicitly control facial features using this medium of representation. Unlike existing methods, our method allows controllable talking face generation without encoding driving videos. The results show that our model outperforms existing models in generating audio-driven talking faces. Notably, our model generates realistic video from a single reference image and driving audio and can control emotional expression without driving videos. One limitation of our model is that it does not model the head pose to control it explicitly. This limitation can be improved by modeling head pose changes. Similarly, we leave it as future work to use face mesh to control other features such as eye blink, head pose, and gaze direction.

Ethical Considerations Our model improves the flexibility of talking face generation as it enables controlling emotional expression without driving videos in the synthesis of realistic talking face videos. Although the main goal is to synthesize virtual avatars, it can be misused to synthesize harmful content that we do not condone. We also hope our method leads to progress in the forgery detection area to identify synthesized content to prevent harmful usage.

References

[1] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang, "Progressive disentangled representation learning for fine-grained controllable talking head synthesis," in *CVPR*, 2023, pp. 17 979–17 989.

[2] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in *AAAI Conference on Artificial Intelligence*, 2023.

[3] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *CVPR*, 2021, pp. 4176–4186.

[4] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *CVPR*, 2022, pp. 3387–3396.

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

[6] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *ICCV*, 2021, pp. 3867–3876.

[7] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *CVPR*, 2021, pp. 14 080–14 089.

[8] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *IJCV*, vol. 128, pp. 1398–1413, 2020.

[9] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *CVPR*, 2023, pp. 8652–8661.

[10] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded gans for learning of motion and texture," in *ECCV*. Springer, 2020, pp. 408–424.

[11] Z. Ye, Z. Jiang, Y. Ren, J. Liu, J. He, and Z. Zhao, "Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis," *ArXiv*, vol. abs/2301.13430, 2023.

[12] J. Tang, B. Zhang, B. Yang, T. Zhang, D. Chen, L. Ma, and F. Wen, "3dfaceshop: Explicitly controllable 3d-aware portrait generation." *IEEE transactions on visualization and computer graphics*, vol. PP, 2022.

[13] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, 2020.

[14] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*. Springer, 2017, pp. 87–103.

[15] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *AAAI Conference on Artificial Intelligence*, 2021.

[16] S. Zhang, J. Yuan, M. Liao, and L. Zhang, "Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary," in *ICASSP*. IEEE, 2022, pp. 2659–2663.

[17] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: real-time photorealistic talking-head animation," *ACM TOG*, vol. 40, no. 6, 2021.

[18] M. Agarwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Audio-visual face reenactment," in *WACV*, 2023, pp. 5178–5187.

[19] J. Guan, Z. Zhang, H. Zhou, T. Hu, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu *et al.*, "Stylesync: High-fidelity generalized and personalized lip sync in style-based generator," in *CVPR*, 2023.

[20] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *ECCV*, 2020.

[21] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM ToG*, 2017.

[22] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH*, 2022, pp. 1–10.

[23] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in *CVPR*, 2023.

[24] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM Multimedia*, 2020, pp. 484–492.

[25] H.-K. Song, S. H. Woo, J. Lee, S. Yang, H. Cho, Y. Lee, D. Choi, and K.-w. Kim, "Talking face generation with multilingual tts," in *CVPR*, 2022, pp. 21 425–21 430.

[26] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2531–2539.

[27] M. Liao, S. Zhang, P. Wang, H. Zhu, X. Zuo, and R. Yang, "Speech2video synthesis with 3d skeleton regularization and expressive body poses," in *ACCV*, 2020.

[28] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li, "Seeing what you said: Talking face generation guided by a lip reading expert," in *CVPR*, 2023, pp. 14 653–14 662.

[29] M. Stypułkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat gans on talking-face generation," *arXiv preprint arXiv:2301.03396*, 2023.

[30] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *ECCV*. Springer, 2020.

[31] J. Zhang, X. Zeng, C. Xu, and Y. Liu, "Real-time audio-guided multi-face reenactment," *IEEE Signal Processing Letters*, vol. 29, 2021.

[32] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM TOG*, vol. 39, no. 6, pp. 1–15, 2020.

[33] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *CVPR*, 2021, pp. 3661–3670.

[34] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," in *ICCV*, 2021.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[37] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *CVPR*, 2019.