3D Talking Face with Personalized Pose Dynamics

Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, Xiaohu Guo

Abstract—Recently, we have witnessed a boom in applications for 3D talking face generation. However, most existing 3D face generation methods can only generate 3D faces with a static head pose, which is inconsistent with how humans perceive faces. Only a few papers focus on head pose generation, but even these ignore the attribute of personality. In this paper, we propose a unified audio-driven approach to endow 3D talking faces with personalized pose dynamics. To achieve this goal, we establish an original person-specific dataset, providing corresponding head poses and face shapes for each video. Our framework is composed of two separate modules: PoseGAN and PGFace. Given an input audio, PoseGAN first produces a head pose sequence for the 3D head, and then, PGFace utilizes the audio and pose information to generate natural face models. With the combination of these two parts, a 3D talking head with dynamic head movement can be constructed. Experimental evidence indicates that our method can generate person-specific head pose sequences that are in sync with the input audio and that best match with the human experience of talking heads.

Index Terms—Audio-driven generation, 3D talking face, personalized pose, generative adversarial network.

1 Introduction

Talking face generation is an attractive research topic in computer vision and graphics. Aside from being interesting, it has a wide range of applications, for example, game animation, 3D video calls, and 3D avatars for AR/MR. Most existing works [1], [2], [3], [4], [5], [6], [7], [8], [9] have proposed generating talking faces from static images. Because of the lack of 3D face model datasets, there are only a few studies [10], [11] on generating talking faces in 3D shapes.

A synthesized talking face from state-of-the-art approaches usually has a static and fixed pose of the head model throughout the speech process. However, in any realistic talking scenario, a person's head will rotate and translate accordingly. If the 3D talking face cannot move reasonably, it will not seem authentic for the audience. In this work, we name the corresponding movement of the head as the *head pose sequence*. A convolutional neural network (CNN) has been adopted as an encoder for 3D face shape generation to achieve state-of-the-art results [11]. VisemeNet [10] adopted a long short-term memory (LSTM) network to generate a 3D talking face without any head movement. It should be noted that all these conventional methods do not take head poses into consideration when generating 3D talking faces, which severely compromises the reality of the synthesized results. The head pose sequences vary in the different video scenarios but show strong correlations with the person's identities, as illustrated in Figure 2. Therefore, generating dynamic pose animations is a crucial step for realistic 3D talking head synthesis.

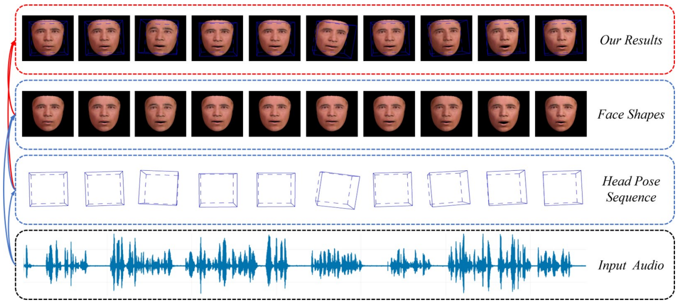


Fig. 1. Pipeline to synthesize a talking face with pose dynamics. Given an input audio, we generate the corresponding sequence of 3D head poses and face shapes.

Zhou *et al.* [12] used face landmarks as an intermediate representation to generate talking face videos with new head poses. However, because a head pose and facial expression are two very different characteristics, using landmark positions to represent them cannot fully capture the personalized head pose dynamics. We show the disadvantages of such landmark-based approaches in the experimental comparisons in Sec. 5.3.3.

In the current paper, we introduce a fully automatic generation framework for an audio-driven 3D talking face with pose dynamics (see Figure 1). To assign different persons with individual head poses, we build a person-specific head motion dataset by providing corresponding head pose sequences and face shapes for each video. During the inference phase, the input audio is first encoded with deep speech [13], and the extracted features are then fed into two proposed modules: the head pose generative adversarial network (PoseGAN) module and pose-guided face (PGFace) generation module. As shown in Figure 3, the

- *C. Zhang, H. Li and X. Guo are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75083.*
- *S. Ni and M. Budagavi are with Samsung Research America.*
- *Z. Fan is with the Tandon School of Engineering, New York University, NY 10003.*
- *M. Zeng is with the School of Informatics, Xiamen University, Xiamen 361005, China.*
- *Corresponding Author: X. Guo, Email: xguo@utdallas.edu*

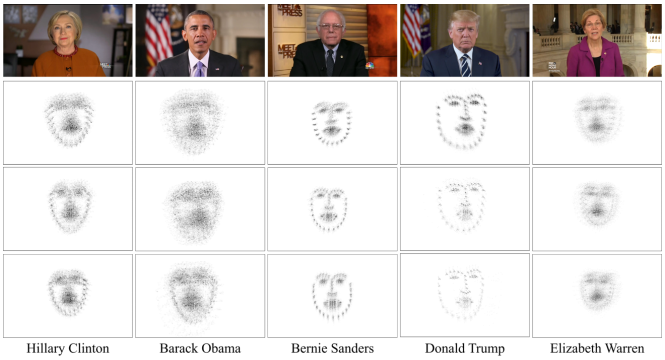


Fig. 2. Our person-specific head motion dataset. Below each person are three trace maps of face landmarks that are tracked from different videos and that depict the frequency of landmarks in different spatial locations. This visualization reveals the speaker's resting pose and their unique head movement style.

PoseGAN module is used to extract the cross-modal head pose sequence with the rotation and translation parameters. The PGFace module with head pose parameters is applied to generate face shape parameters corresponding to the audio. With the combination of the audio, head pose sequence, and face shape parameters, the final 3D talking face with pose dynamics can be synthesized.

Based on this person-specific head motion dataset, we propose an end-to-end unified approach for synthesizing a natural 3D talking head. The main contributions of our work are three-fold:

- We introduce a new method to construct a person-specific head motion dataset that includes over 535,400 frames from 450 video clips. Based on this dataset, a unified audio-driven framework is proposed to generate 3D talking faces with pose dynamics.
- Taking audio flows as the input, a new cross-modal PoseGAN module is proposed to generate the dynamic head poses. A new loss function and initial poses are introduced to ensure the consistency of long-term generations. The PGFace module is designed for pose-dependent facial shape correction, which makes the face shape rendering results more realistic.
- Extensive ablation studies and comparisons with conventional methods indicate that our method can generate a person-specific head pose sequence that is in sync with the input audio and that best matches with the human expectation of talking heads.

2 Related Work

There has been a branch of research in facial animation focusing on synthesizing facial motion from audio, and generating either 2D videos or 3D models as the outputs.

Audio-based 2D facial animation Chung *et al.* [1] proposed an encoder-decoder CNN model to generate synthesized talking face video frames. Deep bidirectional LSTM (BLSTM) was applied by Fan *et al.* [14] in their talking head system. Vougioukas *et al.* [4] used a temporal GAN with two discriminators to generate lip movements and facial expressions. Suwajanakorn *et al.* [3] proposed learning the mapping from raw audio features to mouth shapes by using a recurrent neural network. Chen *et al.* [9] devised a network to synthesize lip movements and proposed a correlation loss to synchronize lip motions and speech changes. Xie and Liu [15] used a dynamic Bayesian network to model the movements of articulators. Jalalifar *et al.* [2] produced realistic faces conditioned on landmarks using a recurrent neural network and a conditional GAN [16], [17]. The arbitrary subject talking face generation method was realized by Zhou *et al.* [5] using a disentangled audio-visual representation with GANs. Zhou *et al.* [12] used the deformation of face landmarks to generate the talking face video with the new head pose.

Our synthesized 3D talking head with personalized pose dynamics can serve as an important intermediate step for these 2D video synthesis methods, as demonstrated by the video application in Sec. 5.1.

Audio-based 3D facial animation A deep learning approach proposed by Taylor *et al.* [18] uses a sliding window predictor that learns mappings from phoneme label input sequences to mouth movements. Zhou *et al.* [10] proposed an automatic real-time lip synchronization from audio solution based on LSTM network architecture. Karras *et al.* [19] presented real-time, low-latency 3D facial animations based on speech audio input with an emotional state. Liu *et al.* [20] employed a data-driven regressor for modeling the correlation between speech data and mouth shapes with a DNN acoustic model. The dynamic facial expressions of the source subject were transferred to the target subject in [21]. Face transfer is based on a multilinear model [22] of 3D face meshes that parameterize the space of geometric variations. Most recently, Cudeiro *et al.* [11] proposed voice operated character animation (VOCA), which takes a random speech signal as the input and generates a wide range of adult faces realistically. VOCA first converts the input audio into DeepSpeech [13] features, and then, one-hot encoding with different subjects is used to train the offsets of 3D face mesh. The FLAME [23] model is applied to generate their final face shape.

For head pose generation methods, Sadoughi *et al.* [24], [25], [26] focus on synthesizing head motions for conversational agents with a synthetic speech from three aspects: 1) how to use the parallel corpus for synthetic speech [24]; 2) how to use the discourse function to generate head poses with meanings [25]; and 3) how to use conditional GAN to generate multiple realizations of head poses for the same input speech [26]. Jonell et al. [27] mainly focus on how to generate head poses for conversational agents conditioned on their interlocutors.

However, none of these works take the personalized head motions into consideration, and the results from these works highly depend on the quality of the 3D face dataset, which is hard to collect in real life. Different from these works, our method focuses on generating head poses for different personalities. The application of our method uses the 3D face model to generate realistic talking face videos for real people.

Text-based facial animation Relatively few works have worked on generating a face model directly from text input. Sako *et al.* [28] described a text-based technique to

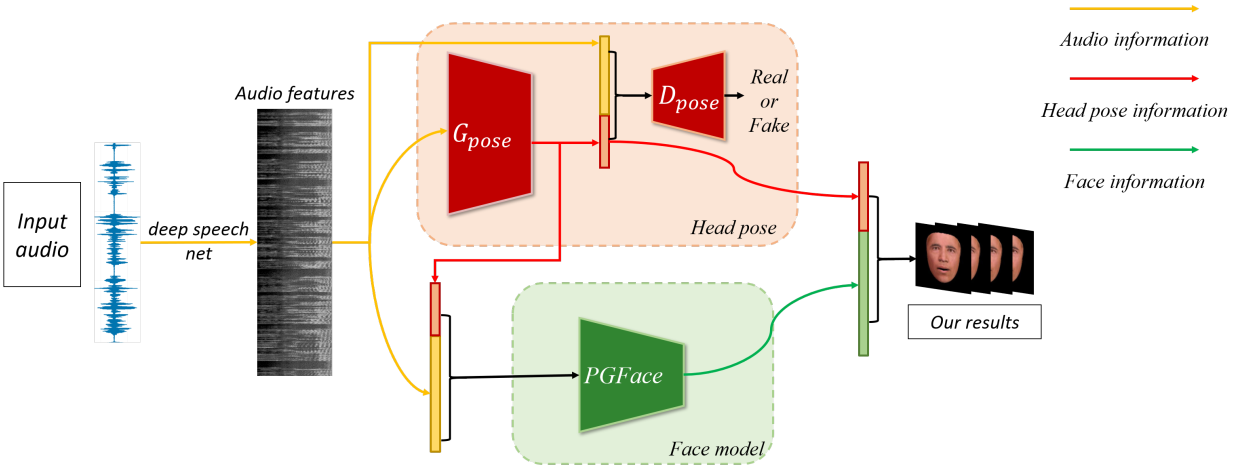


Fig. 3. An overview of our unified framework. G_{pose} denotes the generator of 3D head pose sequences, and D_{pose} is the discriminator. The face shape parameters are generated by PGFace.

generate realistic auditory speech and lip image sequences using hidden Markov models (HMMs). The system for an expressive visual text-to-speech (VTTS) was presented by Anderson *et al.* [29], in which the face is modeled using an active appearance model (AAM). Kumar *et al.* [30] presented a text-based lip-sync generation method that takes a time-delayed LSTM to generate the mouth keypoints that are synced to the audio. Hong *et al.* [31] described a visual speech synthesizer that provides a form of virtual face-to-face communication using text streams.

While in this work we focus on the generation of 3D faces from audio, it is possible to convert our framework into a text-driven model by using a text-to-speech engine (*e.g.*,Tacotron 2 [32]), which we leave to future work for further in-depth exploration.

3D face datasets There have been some datasets [33], [34], [35] used for static 3D face model analysis and some datasets [36], [37], [38], [39], [40] focusing on dynamic 3D face models and expressions. In addition, there are several datasets containing scanned face models. Cheng *et al.* [41] published the 4DFAB dataset containing 4D captures of 180 subjects, and Fanelli *et al.* [42] proposed a 3D audio-visual corpus that contains a large set of audio-4D scan pairs using a real-time 3D scanner. The VOCASET presented by Cudeiro *et al.* [11] contains 3D scans of 255 sentences with the entire head and neck. In the present paper, we constructed our dataset with a large number of face models and head pose sequences corresponding to speech.

3 Dataset

The motivation of this work is to learn and extract the pose characteristics of human talking faces from any data available in the wild. However, real-world 3D face data are labor-intensive to capture using high-speed facial scanners. Another disadvantage of such 3D capture is that these kinds of data are typically captured by a well-designed environment with tens of cameras and projectors. Hence, the participants may unintentionally suppress their natural head movements and facial expressions under such conditions. In contrast, in most videos of real-world scenarios that are available online, people usually perform more natural behaviors, which can serve our research purpose much better. To this end, we advocate for collecting dynamic 3D talking data by analyzing the videos in the wild instead of performing labor-intensive 3D facial capture.

The videos used in the current paper have a total length of approximately five hours and were collected from the videos used by Agarwal *et al.* [43] for their deepfake detection. Our dataset contains over 535,400 frames from 450 video clips along with the audios, 3D head pose parameters, and 3D face shape parameters.

Head pose parameters We adopt OpenFace [44] to generate 3D head pose parameters. Head pose $\mathbf{p} \in \mathbb{R}^6$ is represented by Euler angles (pitch θ_x, yaw θ_y, roll θ_z) in radians and a 3D translation vector $\mathbf{t}$ in millimeters. If we naively apply the head pose sequences detected in the original video by OpenFace, it will cause unstable effects in some high-frequency regions, and the head motion will appear unsatisfying. Therefore, we propose a Gaussian filtering method that filters the head pose parameters throughout the time dimension and generates convincing results. Specifically, our Gaussian filtering method removes abnormal head jittering effectively. As shown in Figure 4, the *pitch* parameter of the head pose is measured in the time dimension over the video clip. In the high-frequency region (*e.g.*, the area in the red rectangle), the curve of the pitch parameter is smoothed, as shown by the orange curve. The Gaussian density and head pose filtering functions are given as follows:

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{1}{2\delta^2}x^2}, \\ \mathbf{p}(i) &= \sum_{k=i-m}^{i+m} \mathbf{p}(k) F(k-i), \end{aligned} \tag{1}$$

where i is the frame index, $2m$ is the window size of the filter, and $\mathbf{p}(i)$ indicates the head pose of the ith frame.

The original videos are divided into small sets of video clips based on the camera parameters, the detection of the frame continuity, and the number of frames. The head pose

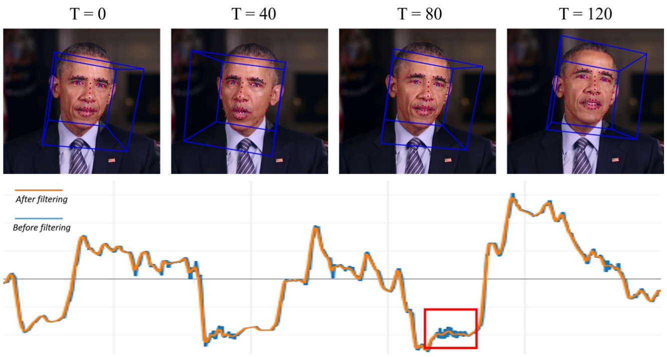


Fig. 4. Gaussian filtering. Blue curve denotes the original pitch parameter. Orange curve is for the smoothed pitch parameter.

is centralized and unified under the same coordinate system in every small video set.

3D face parameters The deep 3D face reconstruction method [45] achieves state-of-the-art performance on multiple datasets. Therefore, we apply this method to generate face parameters $[\alpha_{\text{id}}, \alpha_{\text{exp}}]$. The 3DMM [33], [46] face model is defined as follows:

$$S = \overline{S} + B_{\text{id}}\alpha_{\text{id}} + B_{\text{exp}}\alpha_{\text{exp}}, \tag{2}$$

where $\overline{S}$ is the averaged face shape; B_{id} and B_{exp} are the PCA bases of identity and expression, respectively; $\alpha_{\text{id}} \in \mathbb{R}^{80}$ and $\alpha_{\text{exp}} \in \mathbb{R}^{64}$ are the corresponding coefficients.

It is generally a non-trivial task to capture the 3D face models. We provide a unified framework to obtain precise 3D face models corresponding to video frames, along with the head pose sequence. This person-specific dataset supports our fully automatic framework for generating a 3D talking face. The proposed method for data collection and preparation can also be easily extended to the videos of other person identities available online.

4 Methodology

4.1 Head Pose Sequence Generation Network

Generating the corresponding 3D head pose sequence from an input audio source can be quite difficult. Depending on the speaking scenarios and individual speaking habits, people do not always exhibit the same head pose sequence when speaking the same words. Ginosar *et al.* [47] proposed an audio-based generation method for 2D body gestures. Specifically, they acquired the 2D landmarks of the character's arm and gestures from audio inputs, demonstrating the effectiveness of GAN for cross-modal pose generation.

The generation of the head pose sequence is also a cross-modal prediction task. Inspired by Ginosar *et al.* [47], we propose the PoseGAN to generate the corresponding head pose sequence. To ensure the correlation between the generated head pose sequence and the input audio, we introduce the conditional GAN to determine the output of the head pose sequence belonging to the specific character and a discriminator to determine the authenticity of the head pose sequence. Here, we set 256 frames as the unit sequence.

We notice that the conventional pose loss cannot guarantee consistency between the neighboring sequences and the continuity of head poses in each sequence. To address these problems, an embedding method and motion loss function are proposed. The experimental results in Sec. 5.5.3 show that with the initial pose loss constraint and motion loss function, the two discontinuity problems are solved successfully.

4.1.1 Generator

As shown in Figure 5, we develop an enhanced CNN encoder before the U-net [48] structure to build the generator G, and we embed the initial head pose $\mathbf{p}$ into the input layer and the U-net output layer to constrain the initial position and orientation of the generated head pose sequence.

The initial head pose $\mathbf{p}$ and audio $\mathbf{x}$ are simultaneously input into the generator G, as shown in Figure 5. During the training stage, the pose of the first frame is adopted as the initial pose $\mathbf{p}$ in the head pose sequence. During the inference stage, the rest pose of the same identity is adopted as $\mathbf{p}$ for the generation of the first head pose sequence. Here, we use the mean pose to approximate the rest pose. The last pose of previous sequence is adopted as $\mathbf{p}$ for subsequent head pose sequence generation. The initial pose guarantees consistency between neighboring sequences.

The output head pose sequence presents abnormal instability when directly using the L^2 norm of pose loss (defined in Equation 3) because there are no constraints for continuous motion between frames. We introduce motion loss to ensure the motion continuity of the output head pose sequence.

The L^2 norm loss functions for pose and motion are defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{pose}} &= ||\mathbf{p} - \hat{\mathbf{y}}_\mathbf{0}||_2^2 + \sum_{t=0}^{T-1} ||\mathbf{y_t} - \hat{\mathbf{y}}_\mathbf{t}||_2^2, \\ \mathcal{L}_{\text{p-motion}} &= \sum_{t=1}^{T-1} ||(\mathbf{y_t} - \mathbf{y_{t-1}}) - (\hat{\mathbf{y}}_\mathbf{t} - \hat{\mathbf{y}}_\mathbf{t-1})||_2^2, \end{aligned} \tag{3}$$

where T is the number of frames and is set to 256, $\mathbf{y}$ represents the real head pose sequence in our dataset, $\hat{\mathbf{y}}$ indicates the generated head pose sequence by the generator G, and $\mathbf{p}$ indicates the initial head pose.

The generator's loss function is defined as follows:

$$\mathcal{L}_{L^2} = \alpha\mathcal{L}_{\text{pose}} + \beta\mathcal{L}_{\text{p-motion}}, \tag{4}$$

where α and β are the weights to control the balance between the pose and motion losses.

4.1.2 Discriminator

A CNN structure is applied to discriminate the true and false head pose sequences, here by taking the generated head pose sequence $G(\mathbf{x}, \mathbf{p})$ combined with audio $\mathbf{x}$ as the input. The loss function of discriminator D is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{GAN}} = \arg\min_G \max_D \mathbb{E}_{\mathbf{x},\mathbf{y}}[\log D(\mathbf{x}, \mathbf{y})] + \\ \mathbb{E}_{\mathbf{x},\mathbf{p}}[\log(1 - D(\mathbf{x}, G(\mathbf{x}, \mathbf{p}))], \end{aligned} \tag{5}$$

where the generator G tries to minimize this objective function, while the discriminator D tries to maximize it.

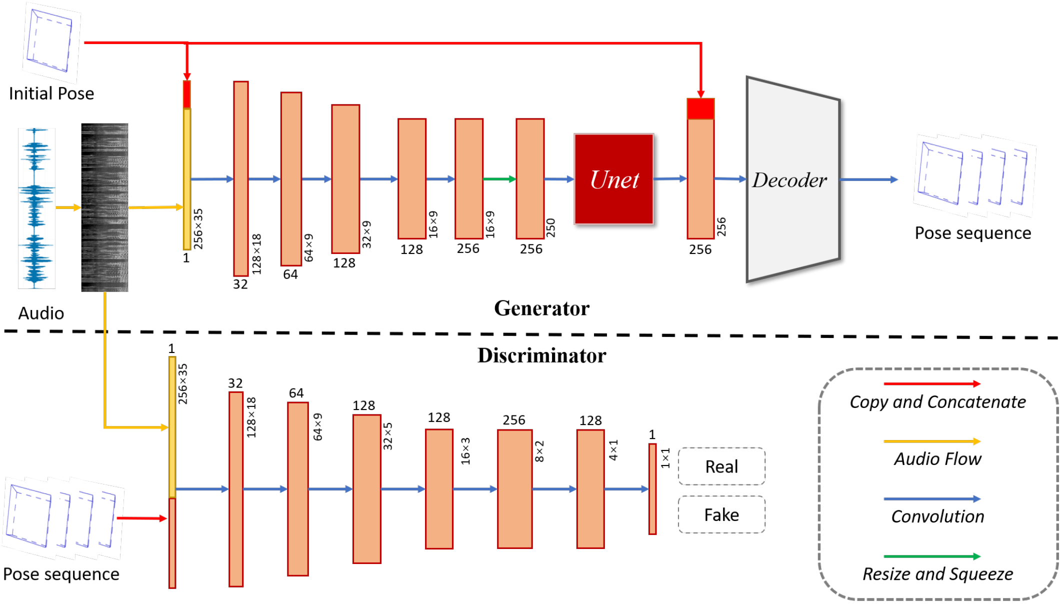


Fig. 5. The architecture of our PoseGAN for head pose estimation from the input audio.

The final PoseGAN's loss function is then defined as follows:

$$\mathcal{L}_{\text{PoseGAN}} = \lambda \mathcal{L}_{\text{GAN}} + \mathcal{L}_{L^2}, \tag{6}$$

where λ is a weight parameter, which controls the balance between the GAN loss and L^2 loss.

4.2 Pose-guided Face Generation Network

The face shape parameters are generated by the deep 3D face reconstruction method [45]. The generated identity parameters α_{id} could be different for each frame. These differences are introduced by camera parameters, speaker position, and inaccurate expression shape. The conventional methods, for example [11], have only generated expression parameters α_{exp}, which are not suitable for our case. Inspired by the VOCA network [11], we propose a pose-guided face shape generation method (PGFace) that includes the head pose parameters as the input for estimating the change of face shape to make up for the difference. We concatenate audio features $\mathbf{x} \in \mathbb{R}^{29\times16}$ and head pose parameters $\mathbf{p} \in \mathbb{R}^{6}$ for each frame as the input for the network. The network output is the corresponding face shape parameters $[\alpha_{\text{id}}, \alpha_{\text{exp}}]$.

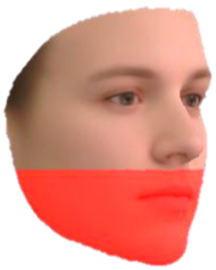


Fig. 6. The lower part of the face, which is shown in red, is used to calculate higher weights for the vertex-level loss.

Based on our experiments, the audio shows a higher correlation with the lower part of the face, as shown in Figure 6. We employ a vertex-based loss function, which attaches a 10-times weight $\mathbf{m}$ on the lower part of the face model. The loss functions can be formally represented as follows:

$$\begin{aligned} \mathcal{L}_{\text{shape}} &= \mathbb{E}_{\mathbf{v},\mathbf{f}}[\|(\mathbf{v}-\mathbf{f}) \odot \mathbf{m}\|^2], \\ \mathcal{L}_{\text{s-motion}} &= \mathbb{E}_{\mathbf{v},\mathbf{f}}[\|((\mathbf{v}_{\text{next}}-\mathbf{v}) - (\mathbf{f}_{\text{next}}-\mathbf{f})) \odot \mathbf{m}\|^2], \end{aligned} \tag{7}$$

where $\mathbf{v}$ denotes the ground-truth face vertices and $\mathbf{f}$ represents the generated face vertices; $\mathbf{v}_{\text{next}}$ and $\mathbf{f}_{\text{next}}$ indicate the values of $\mathbf{v}$ and $\mathbf{f}$, respectively, in the next frame; the mask $\mathbf{m}[i] = 10$ if the vertex i is in the lower part of the face, otherwise $\mathbf{m}[i] = 1$. The $\odot$ operation means an element-wise product. The motion loss $\mathcal{L}_{\text{s-motion}}$ represents the vertex displacement between neighboring frames in sequence.

The PGFace's loss function is then defined as follows:

$$\mathcal{L}_{\text{PGFace}} = \mu_1 \mathcal{L}_{\text{shape}} + \mu_2 \mathcal{L}_{\text{s-motion}}, \tag{8}$$

where μ_1 and μ_2 balance the shape and motion losses.

4.3 Implementation Details

The networks for head pose and face shapes are trained on an Nvidia GTX 1080 Ti using Adam [49] with a batch size of 64 and a learning rate of 10^{-4}. We divide our dataset using a train-val-test split of 7-1-2. In the PoseGAN training section, we first centralize and normalize the head poses as described in our dataset section. The frame rate of our video is 30 fps. We use a 256-frame sliding window as a training sample, and the output is a 256-frame head pose sequence. The sliding distance between neighbors is five frames. During training, α and β are set to 1 and 10, respectively. The value of λ is 0.01. A total of 50 epochs are trained. The best-performing model on the validation set is selected. In the PGFace training section, the network is learned from audio features and head pose parameters with 50 epochs. The window size used for PGFace is 16, and the output is the face shape in the 8th frame. The values of μ_1 and μ_2 are 1 and 10, respectively.

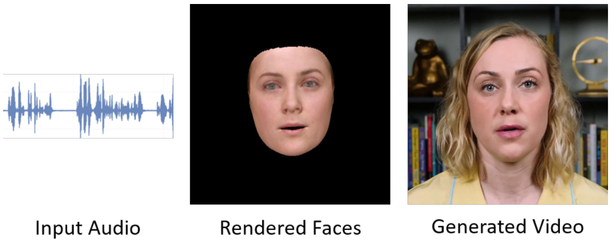


Fig. 7. Personalized training and video application. This woman is not in our dataset.

TABLE 1
One-class SVM results for verifying the correlation between the speech and head pose sequence.

Audio Feature	Corresponding Head Pose	Random Head Pose
Clinton	**0.90**	0.75
Obama	**0.88**	0.52
Sanders	**0.83**	0.71
Trump	**0.85**	0.73
Warren	**0.80**	0.59

5 Experimental Results

5.1 Personalized Training and Video Application

Personalized training: For an existing character in our dataset, we only need to use this character's training data to achieve personalized training. For any character not included in our dataset, 2–3 min videos of this new person are taken as the learning guidance, and then, we fine-tune our PoseGAN and PGFace networks to generate personalized results. Specifically, we first train our model based on the original dataset and obtain the pre-trained model parameters. Given the reference video, We use the methods from Sec. 3 to extract audio features, head poses, and face parameters for each frame. Then, we train the personalized model with the reference data by fine-tuning the pretrained model. For the fine-tuning step, it takes 20 epochs for each network with a batch size of 16. We keep the other parameters unchanged for both networks.

Video application: Our framework generates a 3D talking face with personalized pose dynamics, providing high-level guidance for video synthesis. Existing vid2vid technology can be adopted to map the rendered images into photo-realistic videos. By following Chan *et al.*'s method [50], we train a generator that takes our rendered talking face as an input and a multiscale discriminator for video synthesis. The goal of the generator is to generate a more realistic image and to fool the discriminator to regard it as real.

As shown in Figure 7, when people are not included in our dataset, we first fine-tune our networks to generate new rendered face images. Then, the vid2vid method [50] is adopted to map the rendered images to the corresponding video.

TABLE 2
L^2 distance with head pose and motion on the test set.

Method	$\mathcal{L}_{pose}$	$\mathcal{L}_{p-motion}$
Mean	0.92	0.12
Random	1.23	0.16
NN	1.20	0.14
CNN	0.84	0.11
Our PoseGAN	0.89	0.12

5.2 Evaluation of Feasibility: Correlation Verification

Because our goal is to generate the head pose sequence from speech, we first verify that there is a correlation between a person's speech and his/her head pose. DeepSpeech [13] is used to extract the speech feature for each frame, and OpenFace [44] is used to extract the corresponding head pose. Each frame corresponds to 29 speech features and 6 values of a head pose. We calculate the correlation between the speech and head pose sequence on 256 frames by Pearson's correlation function, obtaining the 29 × 6 features for each 256-frame clip:

$$F(i,j) = \frac{\sum_{k=0}^{255}(S_{ik} - \bar{S}_i)(H_{jk} - \bar{H}_j)}{\sqrt{\sum_{k=0}^{255}(S_{ik} - \bar{S}_i)^2}\sqrt{\sum_{k=0}^{255}(H_{jk} - \bar{H}_j)^2}}, \quad (9)$$

where $i \in [0,5], j \in [0,28]$. S_{ik} and H_{jk} are the ith speech feature and jth head pose value in the kth frame. $\bar{S}_i$ and $\bar{H}_j$ are their average values across 256 frames, respectively.

We then train a one-class support vector machine (SVM) [51] with 29 × 6 features on real data samples. As shown in Table 1, we replace the head pose sequence in the test dataset of each person with a random head pose sequence from the same person. The results of the one-class SVM are reduced when replacing the original head pose sequence, which indicates the existence of a correlation between the head pose sequence and the speech of a particular person. Furthermore, other works [52], [53] have also verified the direct correlation between audio and pose.

5.3 Quantitative Evaluation

We compare our PoseGAN to the following four head pose generation methods.

The mean head pose: Most 2D talking face videos [1], [4], [5], [54], [55], [56], [57], [58], [59] and 3D talking faces [10], [11], [18], [19], [20], [21], [21], [22] can only generate fixed head poses. Most of the time, the head is in a resting position and orientation during speech (see Figure 2). Thus, we use the mean pose as a comparison with these 2D and 3D methods.

Randomly chosen head pose sequence: Another simple way to quickly generate the head pose sequence is to randomly select a head pose sequence from the dataset. This choice is reasonable because they are true head poses. This random method is widely used in 2D talking face methods [3], [6], [7], [8]. Although the retiming technique is used in [3] to increase the authenticity, this method is still a random pose sequence and cannot generate new head poses based on speech. Therefore, such a randomly selected head pose sequence does not correspond to the input audio.

Nearest neighboring (NN) pose: The head pose chosen by this method is the closest to the real head pose in the audio feature space. For each test audio, the head pose sequence with the closest audio feature in the training set is selected as the final output.

TABLE 3
L^2 distance with face shape and motion on the test set.

Method	$\mathcal{L}_{shape}$	$\mathcal{L}_{s-motion}$
Fixed identity	0.96	0.43
Ours	0.81	0.41

Convolutional neural network (CNN): A conventional CNN [11] achieved state-of-the-art results with 3D face shape generation. Some 2D talking face methods [53] also use CNNs to generate head poses in videos. For example, Yi *et al.* [53] used LSTM to generate head pose sequences in their talking face video. However, the head pose estimation is a cross-modal prediction task. We find out that the head pose sequence generated without GAN tends to be close to a static head pose. It is hard to consider the results of CNN as realistic head pose sequences.

5.3.1 L^2 Distance Comparison

To compare our PoseGAN architecture to all these four baselines, we use videos in our test dataset and calculate the L^2 pose distance and motion distance of each method. In Table 2, the random method and nearest neighbor perform significantly worse when it comes to accuracy. This is because these two methods have no constraints on the head pose. The distance of the mean head pose method is low because the speaker is mostly in a static head pose while speaking. The distance with CNN is the lowest because only the pose loss and motion loss are used for training. As discussed before, the generated head pose sequence with CNN tends to be static. The L^2 distance results of our PoseGAN outperforms most of the baseline methods, except for CNN. This is expected because we add GAN loss to our generator to produce more realistic and reasonable head pose sequences.

For face shape generation, we introduce the PGFace network. Compared with the previous VOCA network [11] that adopted a fixed identity method, we generate both the identity parameters and expression parameters. In Table 3, we compare the L^2 distance of our method with the fixed identity method. It can be seen that our method is better than the fixed identity method.

5.3.2 Head Pose Classifier

A head pose classifier is optimized on our training set with five identities to evaluate the head pose results obtained by different methods. A classic CNN and dense layer structure were used to implement the head pose classifier, where the output of the last fully connected layer was set to five. The input is the head pose motion on 256 frames. We choose the best performance on the validation set, which has an accuracy rate of 92% on the test set. As shown in Table 4, the results of our method are closest to the true head pose distribution. If the confidence value is greater than 0.5, most of the data in this category are correctly classified. The results of the mean and CNN methods are close to a random distribution (0.2), which deviate from the true head pose distribution.

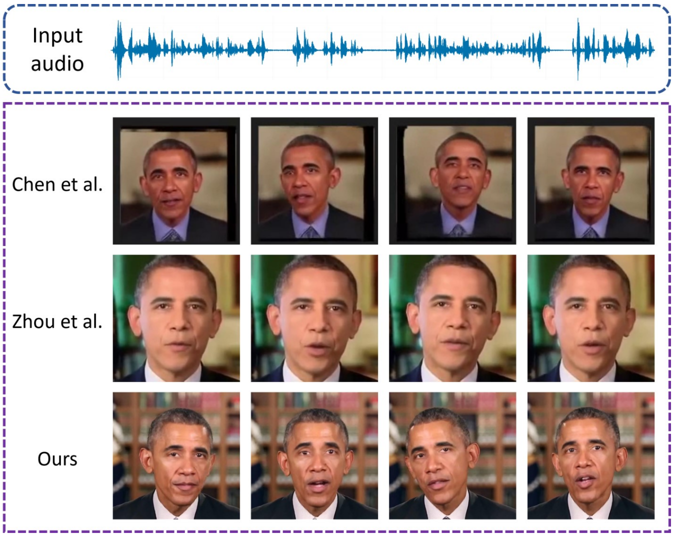


Fig. 8. Comparison with 2D face generation methods, including Chen *et al.* [60] and Zhou *et al.* [12].

5.3.3 Comparison with 2D Face Generation Methods

We compare our proposed method with the state-of-the-art methods of generating a talking video with a head pose based on facial landmarks, including Chen *et al.* [60] and Zhou *et al.* [12]. As shown in Figure 8, we conduct the experiments using the same character, and it can be found that our results are of a higher visual quality than the other methods. Please refer to the supplementary video for the detailed results.

Compared with previous methods [12], [60] that have used landmarks to represent the face shape and head pose, we use a 3D face model to generate video to guarantee more accurate generation of lip motion and personalized head poses. As shown in Table 5, SyncNet [61] is used to evaluate the synchronization of lip motion with the input audio. We calculate the audio-visual (AV) offset and confidence scores for each video to determine the lip-sync error. An offset value closer to zero with a higher confidence score means better synchronization. The head pose classifier is used to evaluate the generated head pose in videos. A higher value indicates more personalized head pose sequences.

We compared our model with Thies *et al.* [7], which also took the 3D face model as the bridge to generate talking face videos. However, they [7] only focused on mouth area generation, and the rest of the faces were all from the reference video, including head poses. As shown in Figure 9, we conducted the qualitative comparison with [7] based on the same person. Please refer to the supplementary video for the detailed results. Different from [7], which only generates expression params, we use the head pose as a guide to generate both the face shape and face expression params. We also calculate the audio-visual (AV) confidence scores to determine the lip-sync error. The AV confidence of [7] for the Obama video is 3.619, and the result of our method is 4.682. The higher confidence score means better synchronization.

TABLE 4
The results of the head pose classifier. Each value represents the confidence of correct classification.

Method	Clinton	Obama	Sanders	Trump	Warren	**Avg**
Mean	0.51	0.01	0.21	0.27	0.00	0.20
CNN	0.13	0.06	0.61	0.30	0.01	0.22
Our PoseGAN	**0.86**	**0.87**	**0.70**	**0.52**	**0.65**	**0.72**

TABLE 5
Quantitative comparisons to Chen *et al.* [60] and Zhou *et al.* [12]. Better values are highlighted in bold.

Method	AV offset	AV confidence	pose classification
Chen *et al.* [60]	-8	3.814	0.72
Zhou *et al.* [12]	**-2**	5.086	0.60
Ours	**-2**	**5.107**	**0.83**

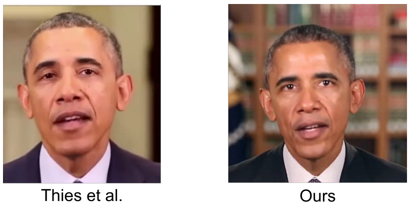


Fig. 9. Comparison to neural voice puppetry [7].

5.4 User Study

5.4.1 Head Pose

One user study is designed to compare our method to the ground truth and all baselines. We prepared 100 pairs of videos. Each includes two videos: one is the talking face with ground truth head pose sequence, and another is generated by one of the four baselines or our method. Three ground truth videos are given to participants to learn before the task. The participants are required to select the better one from each pair. Among the 100 pairs, 60 sets of videos are 4 seconds in length, 25 sets of videos are 8 seconds, and 15 sets of videos are 12 seconds. Fifty people have participated in the study to evaluate the rationality and authenticity of the synthesized 3D talking faces.

We present the results in Table 6. For each video pair (synthesized and ground truth) of different lengths, we measure the probability of selecting the face model generated by the method as the better one. Intuitively, a higher probability means better performance for that method. We find that CNN performs poorly in the user study, while the random method performs relatively better on the 4-second videos but poorly on longer videos. Our method works well on all videos of different lengths.

5.4.2 Face Shape

Our second user study shows the comparison between our pose-corrected face shape with fixed identity shape. The participants select more realistic videos among three groups of 50-second video pairs. Most of them think our results are more realistic (73%) than the fixed identity method (27%).

TABLE 6
User study results. Each value (%) represents the probability that the user selected the generated pose (the true pose is not selected). A larger value indicates that the result is more realistic.

Method	4 seconds	8 seconds	12 seconds
Mean	14.2	14.8	12.0
Random	27.3	20.4	21.3
NN	20.3	16.0	16.7
CNN	16.5	18.8	12.7
Our PoseGAN	**34.3**	**28.4**	**30.0**

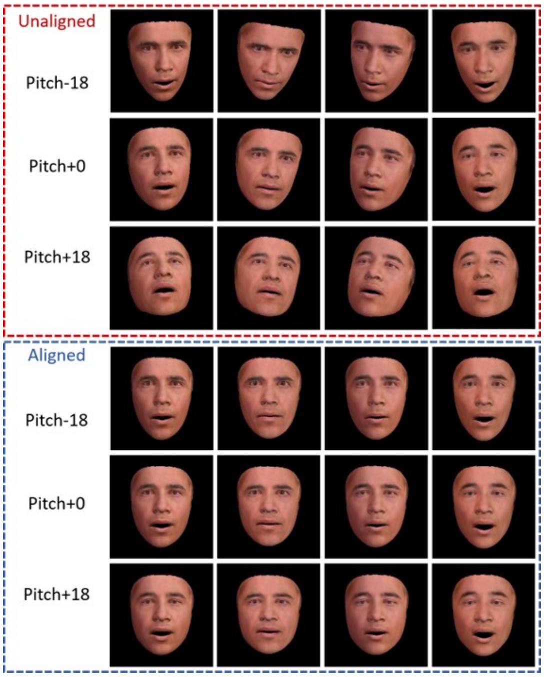


Fig. 10. The rendering results of face shape under different head poses with the same audio.

5.5 Qualitative Evaluation

5.5.1 Pose-dependent Facial Shape Correction

We propose a face shape generation method to complement the face rendering result with head pose information. To show the influence of head poses on face shapes, we conduct three experiments using different head pose parameters: i) use the normal head pose sequence ($Pitch + 0$); ii) increase the pitch angle by 18 degrees ($Pitch + 18$); and iii) control the pitch angle downward by 18 degrees ($Pitch - 18$). The results are shown in Figure 10. To visualize the results in a clear way, we also align the face shapes. In both cases, the head pose has a noticeable effect on producing a more reasonable face shape with the same input audio.

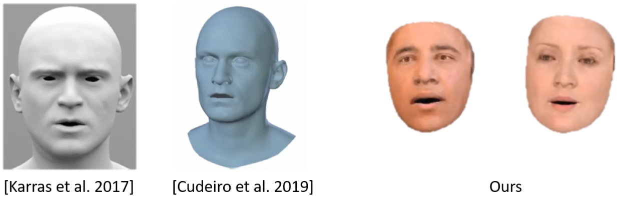


Fig. 11. Comparison with state-of-the-art 3D face generation methods, including VOCA [11] and Karras *et al.* [19].

5.5.2 Comparison with Other Methods

In the supplementary video, we compare our results with state-of-the-art 3D face generation methods, including VOCA [11] and Karras *et al.* [19]. In Figure 11, we show a representative frame of the results for generating the corresponding 3D faces based on input audio.

5.5.3 Ablation Studies

Different variants are compared for head pose generation, including no motion loss and our methods. No motion loss results in jitter problems, and no initial pose leads to discontinuities. In contrast, our proposed PoseGAN generates realistic head pose sequences. More results can be found in the supplementary video. In the supplementary video, we show that our method is still applicable under different noises. Although our training language is based on English, we also show that the method can be applied to multiple language environments.

5.5.4 More Visualization Results

Figure 12 shows the visualization results of our framework. Given an input audio, we generate a 3D talking face with personalized pose dynamics. From top to bottom, they are input audio, head pose sequence, and face shape with head pose. Here, the head pose sequence of the mean method remains the same. The head pose sequence of the CNN method tends to be close to the mean pose and changes slightly. The head poses generated by the random and NN methods change sharply. However, the head pose sequence generated by our method changes stably and reasonably.

6 Conclusion

The current study has worked to generate a 3D talking face with personalized pose dynamics based on input audios. Our 3D face database involves audio, head pose sequences, and face shape parameters. The PoseGAN is trained to generate the head pose sequence, with the initial head pose loss constraint and motion loss function, guaranteeing the continuity of the head pose sequence in the long term. The PGFace network is designed for pose-dependent facial shape correction, which makes the face shape rendering results more realistic. Our experiments verify the effectiveness of our approach, and our synthesized 3D talking head looks more realistic than other methods.

Limitations: The image-based deep 3D face reconstruction method [45] has been adopted. However, using this 3D face modeling method will cause some problems. As shown in Figure 13(a), because the texture information is fixed in a specific video, the generated 3D face cannot blink, while the person in the original video is blinking.

By following Chan *et al.*'s method [50], we extend our rendering faces to realistic video. However, for some large head poses, this vid2vid method cannot correctly generate the corresponding face images. We can see the obvious distortion in the face part shown in Figure 13(b).

Future work: In this paper, we employ image-based 3D face reconstruction, which causes problems, such as inaccurate identity shape and texture. In the future, we would like to build a video-based 3D face reconstruction method to make our training data more accurate. Further, a talking head only conveys the face part of information to audiences during the speech. In future work, we hope to build a speech-driven 3D human animation, including face, body and hands.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions. Zhang, Li, and Guo were partially supported by a grant from National Science Foundation (2007661) and research gifts from Samsung Research America. Zeng was partially supported by NSFC (No.62072382), and Fundamental Research Funds for the Central Universities, China (No.20720190003). The opinions expressed are solely those of the authors, and do not necessarily represent those of the funding agencies.

References

[1] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *British Machine Vision Conference (BMVC)*, 2017.

[2] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-driven facial reenactment using conditional generative adversarial networks," *arXiv preprint arXiv:1803.07461*, 2018.

[3] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.

[4] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven realistic facial animation with temporal gans." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 37–40.

[5] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 9299–9306.

[6] 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 International Conference on Multimedia*, 2020, pp. 484–492.

[7] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: audio-driven facial reenactment," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 716–731.

[8] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic audio-driven video portraits," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020.

[9] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.

[10] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 161, 2018.

[11] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 101–10 111.

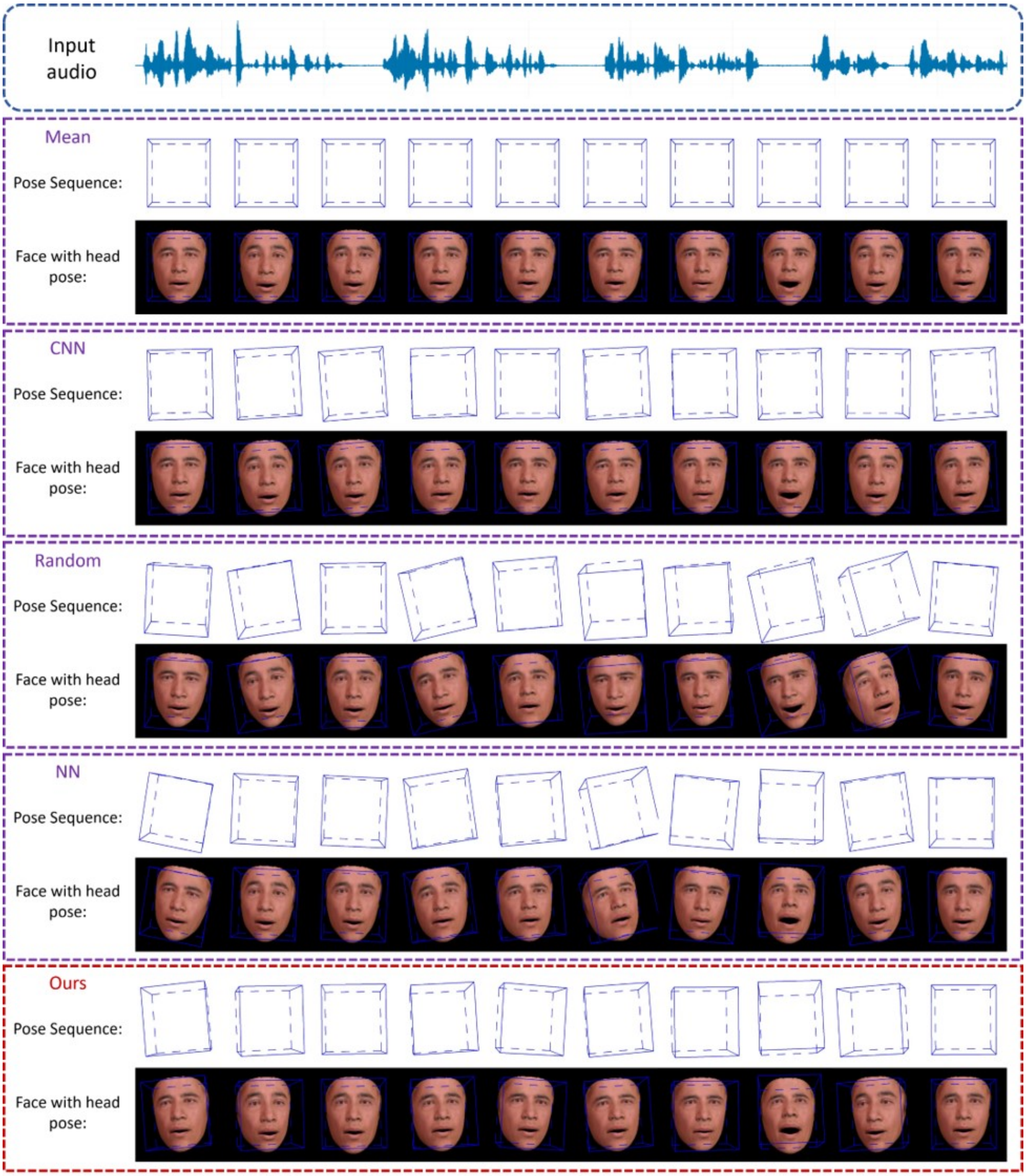
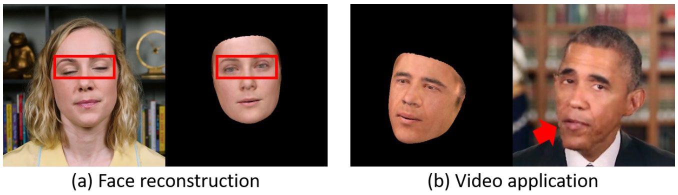


Fig. 12. Results of our framework. From the input audio, we generate a 3D talking head with personalized pose dynamics using the baseline methods and our method. The head pose and face result are sampled every 60 frames (2 seconds).


(a) Face reconstruction

(b) Video application

Fig. 13. Failure cases in face reconstruction and video application.

[12] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, “Makelttalk: speaker-aware talking-head animation,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[13] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.

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

[15] L. Xie and Z.-Q. Liu, “Realistic mouth-synching for speech-driven talking face using articulatory modelling,” *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 500–510, 2007.

[16] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[18] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 93, 2017.

[19] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 94, 2017.

[20] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo, "Video-audio driven real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 182, 2015.

[21] Y. Zhang and W. Wei, "A realistic dynamic facial expression transfer method," *Neurocomputing*, vol. 89, pp. 21–29, 2012.

[22] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," in *ACM SIGGRAPH 2006 Courses*, 2006, pp. 24–es.

[23] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 194, 2017.

[24] N. Sadoughi and C. Busso, "Head motion generation with synthetic speech: A data driven approach." in *INTERSPEECH*, 2016, pp. 52–56.

[25] N. Sadoughi, Y. Liu, and C. Busso, "Meaningful head movements driven by emotional synthetic speech," *Speech Communication*, pp. 87–99, 2017.

[26] N. Sadoughi and C. Busso, "Novel realizations of speech-driven head movements with generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6169–6173.

[27] P. Jonell, T. Kucherenko, G. E. Henter, and J. Beskow, "Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings," in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 2020, pp. 1–8.

[28] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hmm-based text-to-audio-visual speech synthesis," in *Sixth International Conference on Spoken Language Processing*, 2000.

[29] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3382–3389.

[30] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv preprint arXiv:1801.01442*, 2017.

[31] P. Hong, Z. Wen, and T. S. Huang, "iface: A 3d synthetic talking face," *International Journal of Image and Graphics*, vol. 1, no. 01, pp. 19–26, 2001.

[32] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[33] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 20, no. 3, pp. 413–425, 2013.

[34] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European Workshop on Biometrics and Identity Management*, 2008, pp. 47–56.

[35] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*, 2006, pp. 211–216.

[36] T. Alashkar, B. B. Amor, M. Daoudi, and S. Berretti, "A 3d dynamic database for unconstrained face recognition," in *Proceedings of 5th International Conference on 3D Body Scanning Technologies*, 2014, pp. 357–364.

[37] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3d facial expression analysis in videos," in *International Workshop on Analysis and Modeling of Faces and Gestures*, 2005, pp. 293–307.

[38] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[39] D. Cosker, E. Krumhuber, and A. Hilton, "A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling," in *2011 International Conference on Computer Vision*, 2011, pp. 2296–2303.

[40] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438–3446.

[41] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d database for facial expression analysis and biometric applications," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5117–5126.

[42] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[43] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.

[44] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Tech. Rep., 2016.

[45] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[46] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." in *Siggraph*, vol. 99, no. 1999, 1999, pp. 187–194.

[47] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3497–3506.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5933–5942.

[51] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[52] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

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

[54] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353–360.

[55] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7832–7841.

[56] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3, pp. 388–398, 2002.

[57] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-preserving realistic talking face generation," in *2020 International Joint Conference on Neural Networks, (IJCNN)*, 2020.

[58] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, (IJCAI)*, 2019, pp. 919–925.

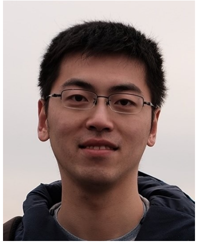
[59] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, pp. 1–16, 2019.

[60] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 35–51.

[61] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian Conference on Computer Vision (ACCV)*, 2016, pp. 251–263.

Chenxu Zhang is currently pursuing a Ph.D. degree with the department of computer science, the University of Texas at Dallas. He received his B.S. degree in software engineering in 2015 and M.S. degree in computer science in 2018, both from Beihang University, Beijing. His research interests include computer graphics, computer vision, and deep learning.

Saifeng Ni received a Ph.D. degree from the University of Texas at Dallas, 2018. She received her M.S. degree from the University of Science and Technology of China, 2012 and B.E. degree from the University of Science and Technology of China, 2009. Her research interests include several topics in computer graphics, computer vision, machine learning, and VR/AR, with an emphasis on mesh optimization, 3D human body and face modeling, reconstruction animation, and motion tracking.



Zhipeng Fan is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at New York University. He received his B.S. degree from the School of Precision Instrument and Optoelectronic Engineering, Tianjin University, China. His current research interests include human pose estimation and computer vision.

Hongbo Li is currently pursuing a master's degree at the University of Texas, Dallas. He received a bachelor's degree from East China Normal University, China, in 2019. His interests are the detection of face manipulation and fake images.

Ming Zeng is currently an associate professor at the School of Informatics, Xiamen University. He was a visiting researcher at Visual Computing Group, Microsoft Research Asia (MSRA) in 2017 and 2009–2011, respectively. He received his Ph.D. degree from State Key Laboratory of CAD&CG, Zhejiang University. His research interests include computer graphics and computer vision, especially in human-centered analysis, reconstruction, synthesis, and animation.

Madhukar Budagavi is a Senior Director R&D in the Standards and Mobility Innovation Lab at Samsung Research America. His team is working on next-gen immersive media processing, compression, and delivery for 5G and its standardization. He is a co-editor of the Springer book *High Efficiency Video Coding (HEVC): Algorithms and Architectures* published in 2014. Dr. Budagavi has coauthored 40+ book chapters and technical papers and 90+ granted patents related to multimedia compression, processing, and streaming. Dr. Budagavi received a Ph.D. degree in Electrical Engineering from Texas A&M University. He is a Senior Member of the IEEE.

Xiaohu Guo is a Full Professor of Computer Science at the University of Texas at Dallas. He received his Ph.D degree in Computer Science from Stony Brook University, and a B.S degree in Computer Science from the University of Science and Technology of China. His research interests include computer graphics, computer vision, medical imaging, and VR/AR, with an emphasis on geometric modeling and processing, as well as body and face modeling problems. He received the prestigious NSF CAREER Award in 2012. For more information, please visit https://personal.utdallas.edu/~xguo/.