This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3207606

Predicting Personalized Head Movement from Short Video and Speech Signal

Ran Yi, Zipeng Ye, Zhiyao Sun, Juyong Zhang, *Member, IEEE,* Guoxin Zhang, Pengfei Wan, Hujun Bao, *Member, IEEE,* and Yong-Jin Liu, *Senior Member, IEEE*

Abstract—**Audio-driven talking face video generation has attracted much attention recently. However, few existing works pay attention to machine learning of talking head movement, especially based on the phonetic study. Observing that real-world talking faces often accompany natural head movement, in this paper, we model the relation between speech signal and talking head movement, which is a typical one-to-many mapping problem. To solve this problem, we propose a novel two-step mapping strategy: (1) in the first step, we train an encoder that predicts a head motion behavior pattern (modeled as a feature vector) from the head motion sequence of a short video of 10-15 seconds, and (2) in the second step, we train a decoder that predict a unique head motion sequence from both the motion behavior pattern and the auditory features of an arbitrary speech signal. Based on the proposed mapping strategy, we build a deep neural network model that takes a speech signal of a source person and a short video of a target person as input, and outputs a synthesized high-fidelity talking face video with personalized head pose. Extensive experiments and a user study show that our method can generate high-quality personalized head movement in synthesized talking face videos, and meanwhile, has comparable facial animation quality (e.g., lip synchronization and expression) with the state-of-the-art methods.**

Index Terms—**Generative models, Head motion behavior pattern, Talking face video synthesis, Speech-driven animation.**

I. Introduction

Visual and auditory modalities are two important sensory channels in human-to-human interaction. The information in these two modalities are strongly correlated [1]. Recently, cross-modality learning has attracted more and more attention in interdisciplinary research, including computer vision, multimedia and computer graphics, e.g., [2]–[5].

In this paper, we focus on talking face video generation that transfers a speech signal of a source person into the visual information of a target person. This kind of audio-driven vision models have a wide range of applications, such as bandwidth-limited video transformation, virtual newsreader and role-playing game generation, etc. Recently, an increasing number of researches have been proposed for this purpose, e.g., [2], [3], [12]–[16]. Most of these works either only consider facial animation with *fixed* head pose (e.g., [2], [3], [12], [13]), or simply *copy* the head pose from the given video of the target person [14]–[16] or an additional pose source video [17]. Some recent works [6]–[10] learn to predict the head motion from audio. However, they did not address the fundamental *one-to-many* problem for speech-to-head-pose mapping as we point out below. A recent person-specific work [11] predicts the head motion by assuming a one-to-many mapping (conditioned on history poses). But it only models the head motion of one target person, and fails to generate head motion for multiple subjects (ref. Table I).

TABLE I
Comparison with methods that PREDICT head poses from audio (rather than fixed or copy from another video). Option A: no need to retrain pose module for new character; Option B: one-to-many mapping.

Methods	speech-to-pose network (pose module)		
	inputs of pose module	option A	option B
TGRHM [6]	speech+reference video	✓	✗
MakeItTalk [7]	speech	✓	✗
HDTF [8]	speech + face photo	✓	✗
3D-Talking-Face [9]	speech + initial pose	✗	✗
FACIAL-GAN [10]	speech	✗	✗
LiveSpeechPortrait [11]	speech + history poses	✗	✓(history poses)
Ours	speech + short video	✓	✓(motion behavior)

In real-world scenarios, natural head movement plays an important role in high-quality communication and psychological studies had shown the evidence that head movement indeed improves auditory speech perception [18]. In fact, humans can easily feel uncomfortable in communication by talking with a fixed head pose. In this paper, we propose a deep neural network to model the relation between speech signal and talking head movement in a short video.

Inferring head movement from speech has been studied in the speech community, e.g., [19]. Although some measurable correlations have been observed between speech and head movement [4], it is well recognized that speech-to-head-pose is a typical one-to-many mapping, and there is a lack of theoretical support why a unique head motion sequence can be predicted from a speech signal. In our study, we draw a key observation from the phonetic study in [20]: there is a correlation between some acoustic features of speech and some components of head motion, but their coupling varies from utterance to utterance; e.g., the fundamental frequency F_0 may correlate to the x-axis component of head rotation in a very short speech, but in another short speech, F_0 may

R. Yi is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

Z. Ye, Z. Sun, Y.-J. Liu are with BNRist, MOE-Key Laboratory of Pervasive Computing, the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Y.-J. Liu is the corresponding author. E-mail: liuyongjin@tsinghua.edu.cn.

J. Zhang is with the School of Mathematical Sciences, University of Science and Technology of China, Hefei, China.

G. Zhang and P. Wan are with Kuaishou Technology, China.

H. Bao is with the college of Computer Science and Technology, Zhejiang University, Hangzhou, China.

R. Yi and Z. Ye are joint first authors.

correlate to the y-axis component of head rotation, even for the same person.

In this paper, we characterize the basic unit of utterance by a short video V of 10-15 seconds, and propose a novel two-step mapping strategy: (1) in the first step, we train an encoder that predicts a head motion behavior pattern (modeled as a feature vector x) from the head motion sequence of V, and (2) in the second step, we train a decoder that predict a unique head motion sequence from both x and the acoustic features of an arbitrary speech signal. Our strategy elegantly solves the one-to-many mapping problem: for the same speech, given different short reference videos, our strategy can predict different personalized head poses. It is worth noting that we set the length of a short video to be 10-15 seconds, because for a relatively long video (e.g., 2 minutes) it may mix multiple inconsistent head motion behavior patterns and the decoder cannot predict a unique one. Meanwhile, short videos of 10-15 seconds have been widely used in many social media applications (such as video sharing in the popular TikTok and WeChat APPs).

Based on the proposed mapping strategy, we build a system that can transfer the speech signal of an *arbitrary* source person into the talking face video of an *arbitrary* target person with learning-based personalized head pose. We model the speech-to-head-pose mapping as a one-to-many problem, and generate head motion for multiple subjects. The contributions of this paper are three-fold:

- We propose a novel two-step mapping strategy that efficiently solves the one-to-many problem for speech-to-head-pose mapping and generates personalized head motion for multiple subjects.
- We propose a novel talking face video generation system that can transfer a speech signal of an arbitrary source person into a high-quality talking face video of an arbitrary target person, with *personalized* talking head movement learned from a short video of 10-15 seconds.
- We propose a rendering-to-realistic GAN module which can generate photo-realistic video frames for *various* face identities, i.e., corresponding to different target persons in different reference short videos.

II. Related Work

A. Talking face video generation

Existing talking face video generation methods can be broadly categorised into two classes according to the driven signal: video-driven and audio-driven.

Video-driven methods (a.k.a face reenactment) transfer expression and sometimes head pose from a driving frame to a face image of target actor. Traditional optimization methods transferred expression using 3DMM parameters [21] or image warping [22]. Learning-based methods were also proposed by training with videos of target actor or general audio-visual data, using GAN model conditioned on different characteristics, e.g., image [23], additional landmarks [24], or motion representation [25].

Audio-driven methods take both auditory and visual information as input, where the auditory data (from source person) provides the driving signal for facial movements and the visual data provides the information of target person. Audio-driven methods can be further categorized into following sub-classes, based on the time span of input visual information:

(1) *A single face image.* Chung et al. [3] learned a joint embedding of input face image and audio signal, and used an encoder-decoder CNN model to generate a talking face video. Similarly Zhou et al. [13] also learned a joint audio-visual representation. Chen et al. [2] first transferred the audio signal to facial landmarks and then generated video frames conditioned on the landmarks. Song et al. [12] proposed a conditional recurrent adversarial network that integrated audio and image features in recurrent units. Recently, Zhou et al. [17] proposed a pose-controllable audio-visual system which takes head poses from another pose source video rather than learning from audio. The above methods either focused on the facial animation and fixed the head pose during animation, or copy head motion from a reference video.

Some recent works explored learning head motions. Zhou et al. [7] proposed MakeItTalk, which generates speaker-aware talking-head animation from a face photo and a speech signal. It proposed to disentangle speaker-agnostic content and speaker identity information from audio, and designed two branches to generate facial landmark offsets from audio, for general and speaker-aware motion respectively. However, it only extracted identity from audio, and ignored visual information and head motion behavior in the target person's video. Zhang et al. [8] proposed a flow-guided talking face generation system (HDTF) which learns a mapping from audio and reference face image to head poses. It regarded the reference face identity as the style for generating head poses, but a single reference face is unable to provide information about head motions. The above methods failed to address the one-to-many mapping between audio and head motion caused by different head motion behaviors.

(2) *A short video of less than 15 seconds.* Chen et al. [6] proposes a head motion learner for talking head generation, which learns the relation between audio $x_{1:\tau}$ and its paired head motion $h_{1:\tau}$, and predicts the head motion of subsequent audio $x_{\tau+1:T}$. But it did not address the one-to-many mapping problem and could not generate personalized head motion of the target person given another person's audio. Different from that, we extract personalized head motion behavior pattern from head motion sequence only, and predict a new head motion sequence from the extracted pattern and acoustic features. Observing that short videos of 10-15 seconds have been widely used in popular free messaging APPs such as TikTok and WeChat, we provide a method by making use of this kind of video data.

(3) *A short video of 2-3 minutes.* Given a speech signal of a source person, Thies et al. [14] proposed NeuralVoicePuppetry, which outputs high-quality talking video of a target person by using a latent 3D face model. This method contains a general audio-to-expression network and a specific neural rendering network, where the specific network is trained on a target person's video of 2-3 minutes. Wen et al. [15] proposed AudioDVP to first predict expression from audio and re-render the face, and then use a neural renderer to transform the

lower face regions of rendering images into realistic ones. The network models in this method were also trained on a target person's video of 3 minutes. Both methods simply copied the pose sequence from the reference video and did not consider personalized head pose.

Recently, some person-specific methods [9]–[11] have been proposed to generate head poses, i.e., a new model need to be trained for each new character. Zhang et al. [9] proposed 3D-Talking-Face, which generates 3D talking face with personalized head pose. This method adopts personalized training: for a new character, the pose generator needs to be trained using a 2-3min video of the new person. Zhang et al. [10] proposed FACIAL-GAN to jointly learn explicit (facial expression) and implicit (head poses, eye blinks) attributes from audio features. It utilizes a temporal correlation generator, a local phonetic generator and a discriminator to learn mapping from audio features to explicit and implicit attributes, where the speech-to-head-pose is modeled as a one-to-one mapping. The model is finetuned on a 2-3min reference video of the target person. Lu et al. [11] proposed LiveSpeechPortraits, a person-specific talking dynamic estimation method by learning from a 3-5min video of the subject. It models the speech-to-head-pose as a one-to-many mapping, but the model is trained for only one target person. Different from [11], we learn an audio-to-head-pose mapping for multiple subjects.

(4) *Long video data of more than 10 hours.* A representative work in this sub-class is [5] in which a talking face video generation method with personalized head pose was proposed for a specified person, e.g., it used Obama's weekly address videos of 17 hours to train the model. The main obstacle in this method is to collect high quality video training data of many hours for a specified person.

(5) *Video data without specified time constraint.* Prajwal et al. [16] proposed a speech-to-lip generation method, which designed a novel lip-sync discriminator to achieve accurate lip synchronization. This method can deal with the target person's video of any time lengths and thus cover the above sub-classes (2-4). However, this method simply copied the head pose from reference video to output synthesized video.

As a summary, none of the above methods solve the following problems in a simple yet effective framework like ours: 1) predicting head pose from audio-visual information using an one-to-many mapping model; 2) modeling head motion of multiple characters; and 3) generate talking face video with personalized head movement of an arbitrary target person given an arbitrary audio (ref. Table I).

B. 3D face reconstruction

Another challenge in our work is that natural head movement often causes out-of-plane head rotations, and it is difficult to synthesize high-quality facial animation with smooth background transition. To address this challenge, we reconstruct 3D face animation and re-render it into video frames. Below we summarize the related work.

3D face reconstruction aims to reconstruct 3D shape and appearance of human face from 2D images. Many methods have been proposed [26] and most of them were based on 3D Morphable Model, which learned a PCA basis from scanned 3D face dataset to represent general face shapes. Traditional methods fit 3DMM by an analysis-by-synthesis approach, which optimized 3DMM parameters by minimizing difference between rendered reconstruction and the given image (e.g., [27]). Learning-based methods used CNN to learn a mapping from face images to 3DMM parameters. To deal with the lack of sufficient training data, some methods used synthetic data, e.g., [28], [29], while others use unsupervised or weakly-supervised learning, e.g., [30], [31]. In this paper, we adopt the method [31] for 3D face reconstruction.

C. GANs

The video frames by re-rendering reconstructed 3D face animation are often not very realistic. In the system proposed in this paper, we propose a rendering-to-realistic GAN module to fine tune these rendered frames into realistic ones with smooth transition. Below, we summarize the related work.

Generative Adversarial Networks (GANs) have been successfully applied to many computer vision problems (e.g., [23], [32]–[38]. The Pix2Pix proposed by [37] has shown great power in image-to-image translation between two different domains. Later it was extended to video-to-video synthesis, e.g., [39]. It has also been applied to the field of facial animation and texture synthesis. Wang et al. [23] applied a GAN conditioned on rendered face images to generate realistic video frames. Although this method achieved good results, it was only suitable for a specific target person, and it needed thousands of training samples related to this specific person. In addition to generate image, [38] proposed a GAN model to generate realistic dynamic facial textures.

III. Predicting Head Movement from Multi-Modal Input

In this paper, we aim to generate high-quality talking face video, given a speech signal of a source person and a short video (about 10-15 sec) of a target person. In addition to learning the transformation from the speech signal to lip motion and facial expression, our system specially considers the generation of personalized head movement of the target person. Below we propose a novel two-step mapping strategy to predict head movement from input speech signal and short video. The whole system is presented in Section IV.

People naturally move their heads when speaking. It is well recognized [18] that this accompanying head movement conveys linguistic information and there is a *weak* correlation between speech and head movement. Yehia et al. [20] explain the meaning of weak correlation by that speech-to-head-motion is a one-to-many mapping and the manner in which speech and head motion are coupled changes from utterance to utterance. For a simple example, they found that in a short speech of a few seconds, the fundamental frequency F_0 is coupled with head rotation in the axial axis; while in another short speech of a few seconds, F_0 is coupled with head rotation in the sagittal plane and translation in the vertical and protrusion axes. Our analysis in the appendix A also supports these findings in [20].

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3207606

In our study, we set the time length of input video to be 10-15 seconds, due to the following reasons: (1) based on the study in [20] as well as our experiment in the appendix A, we assume that a *basic utterance unit* (in which the mapping from speech signal to head movement is one-to-one) ranged from a few seconds to a dozen seconds; that is, for a video of more than one minute, it may contain multiple inconsistent styles of utterances, (2) a user study in the appendix B shows that using a video of 4 or 8 seconds leads to relative bad fine-tuning results from our GAN module, and using a video of 12 or 20 seconds has the same quality of fine-tuning results, and (3) videos of 10-15 seconds are widely used in social media such as TikToK and WeChat.

We characterize the head motion in each video of 10-15 seconds by a *motion behavior pattern*. Our modeling strategy has the following characteristics. First, different people have different motion behavior patterns. Second, since the coupling between speech and head motion changes from utterance to utterance, one person can have multiple motion behavior patterns in multiple short videos, but most of the time one short video of a person (which is 10-15 seconds and only contains one basic utterance unit) only corresponds to one motion behavior pattern. Finally, given one motion behavior pattern of a target person and a speech signal of a source person, our method can output a high-quality talking face video with personalized head movement.

A. Motion behavior pattern

We represent the motion behavior pattern as a vector in the feature space R^k, where k is empirically set to be 16 in our experiment. Given a short video of a target person, we extract the head pose sequence using 3D face reconstruction, and map the head pose sequence into a vector in R^k (to represent motion behavior pattern). We model the mapping as a function $g(\mathbf{y}) : P^* \rightarrow R^k$, where P^* is the domain of all head pose sequences. The input to the function $g(\mathbf{y})$ is the head pose sequence $\mathbf{y} = \{y^{(1)}, \ldots, y^{(T)}\}$ extracted from the short video, and the output is a motion behavior pattern $x \in R^k$. We construct a new dataset which contains different types of head motions of different people to train the mapping. The details of this new dataset are presented in Section V-A.

B. Speech to head pose prediction conditioned on motion behavior pattern

We model the one-to-many mapping between speech and head motion as a function $f(\mathbf{a}, x) : A^* \times R^k \rightarrow P^*$, where A^* is the domain of auditory features. The inputs to the function $f(\mathbf{a}, x)$ include (1) an auditory feature sequence $\mathbf{a} = \{a^{(1)}, \ldots, a^{(T)}\} \in A^*$, and (2) a motion behavior pattern $x \in R^k$. And the output of the function $f(\mathbf{a}, x)$ is a predicted pose sequence $\mathbf{y_f} = \{y_f^{(1)}, \ldots, y_f^{(T)}\} \in P^*$.

Given a fixed speech signal with audio feature $\mathbf{a}$, if we change the motion behavior pattern x, then using the function f we can get different predicted head pose sequences. In this way one speech signal can correspond to different head motions. However, if we fix the motion behavior pattern x to a specific pattern x_0, only one predicted pose sequence will be generated for each speech signal.

C. Head motion encoder and head motion decoder

We design a head motion encoder E_m for $g(\mathbf{y})$ and a head motion decoder D_m for $f(\mathbf{a}, x)$.

Network design. The head motion encoder E_m takes a head pose sequence as input, and outputs the motion behavior pattern (a k-dimension vector). It consists of (1) an LSTM network to extract the head pose feature[1], and (2) a multi-layer perceptron (MLP) network to map the pose feature into a motion behavior pattern in R^k.

The head motion decoder D_m takes the MFCC audio feature $\mathbf{a}$ and a motion behavior pattern x as inputs. And the output is a predicted head pose sequence. It consists of (1) an LSTM to extract deep auditory features from MFCC features, (2) a 1D convolution layer to collect deep auditory features in each window of 280ms, (3) a concatenation of the pattern x with the collected audio features, and (4) a MLP network to predict the pose sequence from the concatenated feature.

Loss function. We first pre-train the head motion encoder E_m using the contrastive loss: given a triplet of head pose sequences $(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)$, where $\mathbf{y}_1$ and $\mathbf{y}_2$ are head pose sequences extracted from the same video, and $\mathbf{y}_3$ is head pose sequence extracted from a different video, we minimize the distance between $E_m(\mathbf{y}_1)$ and $E_m(\mathbf{y}_2)$ and maximize the distance between $E_m(\mathbf{y}_1)$ and $E_m(\mathbf{y}_3)$. The loss function is

$$\begin{aligned}\mathcal{L}_{contra}(E_m) = \mathbb{E}_{(\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3)}\{&||E_m(\mathbf{y}_1) - E_m(\mathbf{y}_2)||^2 \\ &+ [max(\tau - ||E_m(\mathbf{y}_1) - E_m(\mathbf{y}_3)||, 0)]^2\},\end{aligned} \tag{1}$$

where τ is the margin, and we set $\tau = 2$ in the experiment.

We then train the head motion encoder and decoder jointly, on a newly constructed head motion dataset. The pretrained model provides the initial network state for the head motion encoder. For each video in the dataset, we extract the MFCC feature $\mathbf{a}$ from the audio track and the pose sequence $\mathbf{y}_r$ from the frames, which forms a pair $(\mathbf{a}, \mathbf{y}_r)$. We train the model to recover the ground-truth pose sequence. The loss function is

$$\begin{aligned}\mathcal{L}(E_m, D_m) = \mathcal{L}_{recon}(E_m, D_m) + \lambda_{p1}\mathcal{L}_{smooth}(E_m, D_m) \\ + \lambda_{p2}\mathcal{L}_{contra}(E_m) + \lambda_{p3}\mathcal{L}_{contra2}(E_m, D_m)\end{aligned} \tag{2}$$

which consists of a reconstruction term[2],

$$\mathcal{L}_{recon}(E_m, D_m) = \mathbb{E}_{(\mathbf{a},\mathbf{y}_r)}[(\mathbf{y}_r - D_m(\mathbf{a}, E_m(\mathbf{y}_r)))^2], \tag{3}$$

a smooth term,

$$\begin{aligned}\mathcal{L}_{smooth}(E_m, D_m) = \mathbb{E}_{(\mathbf{a},\mathbf{y}_r)}[&\mathbb{E}_t(\mathbf{y}_f^{(t+1)} - \mathbf{y}_f^{(t)})^2 \\ &+\mathbb{E}_t(2\ \mathbf{y}_f^{(t)} - \mathbf{y}_f^{(t-1)} - \mathbf{y}_f^{(t+1)})^2],\end{aligned} \tag{4}$$

where $\mathbf{y}_f$ is the predicted sequence $D_m(\mathbf{a}, E_m(\mathbf{y}_r))$, a contrastive term for pattern, i.e., $\mathcal{L}_{contra}(E_m)$ in Eq.(1), and a contrastive term for predicted pose,

$$\begin{aligned}\mathcal{L}_{contra2}(E_m, D_m) = \mathbb{E}_{(x_1,x_2,x_3)}\{&||D_m(\mathbf{a}, x_1) - D_m(\mathbf{a}, x_2)||^2 \\ &+[max(\tau_2 - ||D_m(\mathbf{a}, x_1) - D_m(\mathbf{a}, x_3)||, 0)]^2\}.\end{aligned} \tag{5}$$

[1] We take the hidden and cell states of the LSTM network as the pose feature.

[2] The distance between two sets of rotation angles are calculated as mean square error between elements in two rotation matrices.

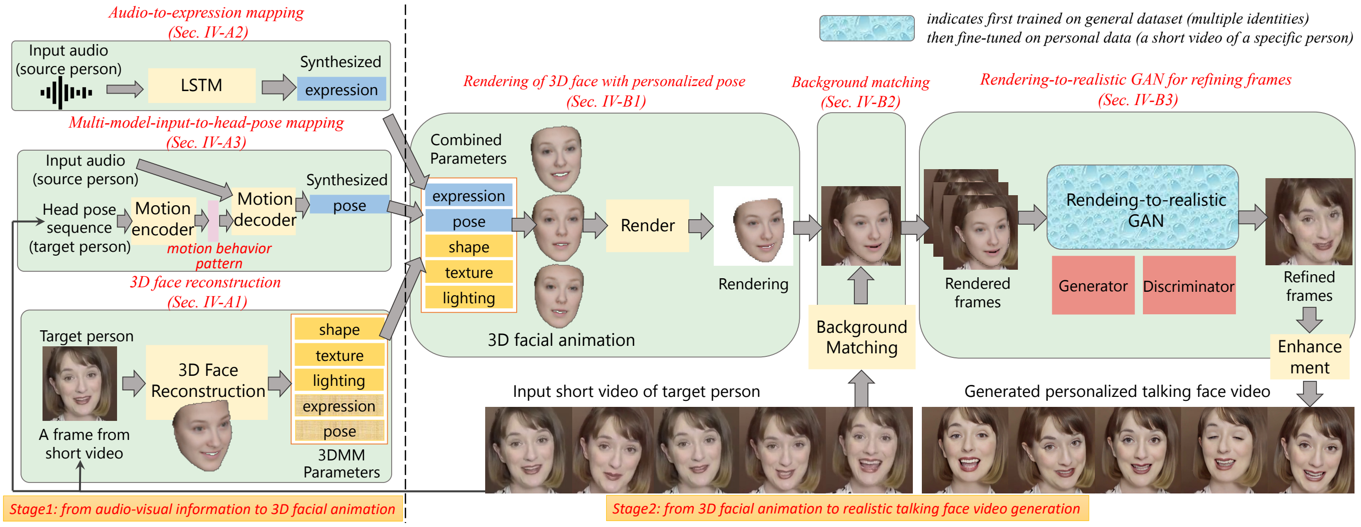

Audio-to-expression mapping (Sec. IV-A2)
Input audio (source person)
LSTM
Synthesized expression
Multi-model-input-to-head-pose mapping (Sec. IV-A3)
Input audio (source person)
Head pose sequence (target person)
Motion encoder
Motion decoder
motion behavior pattern
Synthesized pose
3D face reconstruction (Sec. IV-A1)
Target person
A frame from short video
3D Face Reconstruction
shape
texture
lighting
expression
pose
3DMM Parameters
Rendering of 3D face with personalized pose (Sec. IV-B1)
Combined Parameters
expression
pose
shape
texture
lighting
Render
Rendering
3D facial animation
Background matching (Sec. IV-B2)
Background Matching
indicates first trained on general dataset (multiple identities) then fine-tuned on personal data (a short video of a specific person)
Rendering-to-realistic GAN for refining frames (Sec. IV-B3)
Rendered frames
Rendeing-to-realistic GAN
Generator
Discriminator
Refined frames
Enhance ment
Input short video of target person
Generated personalized talking face video
Stage1: from audio-visual information to 3D facial animation
Stage2: from 3D facial animation to realistic talking face video generation

Fig. 1. Flowchart of our method. Stage 1: from audio-visual information to 3D facial animation, including (1) reconstructing 3D face of the target person (Sec. IV-A1), (2) training a general mapping from speech to the facial expression (Sec. IV-A2), and (3) training an encoder and decoder for personalized head movement for multiple subjects (Sec. IV-A3). Stage 2: from 3D facial animation to realistic talking face video generation, including (1) rendering 3D facial animation into video frames using a lightweight graphic engine (Sec. IV-B1), (2) background matching (Sec. IV-B2), (3) fine tuning rendered frames into realistic ones using a rendering-to-realistic GAN module (Sec. IV-B3), and (4) an enhancement module to obtain high quality results.

where $\tau_2 = 2$ in the experiment, x_1, x_2, x_3 are the pattern encoding of $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ respectively (note that x_1, x_2 are the encoding of the head pose from the same short video, x_3 is the encoding of a different video). In the term (5), for the same auditory feature sequence $\mathbf{a}$, we minimize the predicted pose from $(\mathbf{a}, x_1)$ and $(\mathbf{a}, x_2)$, and maximize the predicted pose from $(\mathbf{a}, x_1)$ and $(\mathbf{a}, x_3)$.

IV. Talking Face Video Generation System

Based on the two-step mapping strategy for head pose prediction, in this section, we propose a talking face video generation system with personalized head movement, when given a speech signal of a source person and a short video (about 10-15 sec) of a target person as input.

To achieve this goal, our idea is to use *3D facial animation with personalized head pose* as the kernel to bridge the gap between audio-visual-driven head pose learning and realistic talking face video generation. The flowchart of our method is illustrated in Fig. 1, which consists of two stages:

Stage 1: from audio-visual information to 3D facial animation. First, we reconstruct 3D face of the target person. Then we train a general mapping from the speech signal to the facial expression on the LRW video dataset [40]. To model personalized head motion behavior, we construct a head motion dataset to train the head motion encoder (for extracting a behavior pattern x) and the head motion decoder (for establishing a mapping from speech to head movements, conditioned on x and auditory features).

Stage 2: from 3D facial animation to realistic talking face video generation. We render the 3D facial animation into video frames using the texture and lighting information obtained from input video. With these limited information, the graphic engine can only provide a rough rendering effect that is usually not sufficient to obtain realistic frames. We then use a rendering-to-realistic GAN module that can refine these rendered frames into realistic ones. This GAN module is trained in two steps: (1) first trained on the publicly available LRW video dataset (for various identities) and then (2) fine tuned using the input short video (for the specific target person). Therefore, our GAN module can deal with various identities and generate high-quality talking face video frames that match the face identity of the target person in input video. Finally we apply an enhancement module [41] to obtain high quality results.

Our system contains six main components in two stages: (1) stage 1 contains the 3D face reconstruction module, the audio-to-expression mapping module, and the multi-modal-input-to-head-pose mapping module, and (2) stage 2 contains the rendering of 3D face module, the background matching module, and the rendering-to-realistic GAN module. In particular, our novel GAN module can generate photo-realistic video frames for various face identities. The details of each module are introduced below.

A. Stage 1: from audio-visual information to 3D facial animation

1) 3D face reconstruction: We adopt a state-of-the-art deep learning method [31] for 3D face reconstruction. It uses a CNN to fit a parametric model (for 3D face geometry, texture and illumination) to an input face photo $\mathbf{I}$. This method reconstructs the 3DMM coefficients $\chi(\mathbf{I}) = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p}\} \in \mathbb{R}^{257}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}$ is the coefficient vector for face identity, $\boldsymbol{\beta} \in \mathbb{R}^{64}$ is for expression, $\delta \in \mathbb{R}^{80}$ is for texture, $\boldsymbol{\gamma} \in \mathbb{R}^{27}$ is the coefficient vector for illumination, and $\mathbf{p} \in \mathbb{R}^{6}$ is the pose vector including rotation and translation. Then the face shape S and face texture T can be represented as $\mathbf{S} = \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}$, $\mathbf{T} = \bar{\mathbf{T}} + \mathbf{B}_{tex}\boldsymbol{\delta}$, where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are average shape and texture, $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$ and $\mathbf{B}_{tex}$ are PCA basis for shape, expression and texture separately. Basel Face Model [42] is used for $\mathbf{B}_{id}$ and $\mathbf{B}_{tex}$, and FaceWareHouse [43] is used for $\mathbf{B}_{exp}$.

The illumination is computed using the Lambertian surface assumption and approximated with spherical harmonics (SH) basis functions [44]. The irradiance of vertex v_i with normal vector $\mathbf{n}_i$ and texture $\mathbf{t}_i$ is $C(\mathbf{n}_i, \mathbf{t}_i, \boldsymbol{\gamma}) = \mathbf{t}_i \sum_{b=1}^{B^2} \gamma_b \Phi_b(\mathbf{n}_i)$, where $\Phi_b : \mathbb{R}^3 \to \mathbb{R}$ are SH basis functions, γ_b are SH coefficients and $B = 3$ is the number of SH bands. The pose is represented by rotation angles and translation vectors. A perspective camera model is used to project the 3D face model onto the image plane.

2) Audio-to-expression mapping: It is well recognized that the audio signal has strong correlation with lip and lower-half face movements. However, talking faces with only lower-half face movements are stiff and far from natural. In other words, upper-half face (including eyes and brows) movements and head pose are also essential for a natural talking face. Below we establish the mapping from the speech to the facial expression.

We extract the Mel-frequency cepstral coefficients (MFCC) feature of the input speech (using sample rate 16,000 and window step 10 ms), and model the facial expression using 3DMM coefficients. To establish the mapping inbetween, we design an LSTM network as follows. Given the MFCC features of an audio sequence $\mathbf{s} = \{s^{(1)}, \ldots, s^{(T)}\}$, and a ground-truth expression coefficient sequence $\boldsymbol{\beta} = \{\beta^{(1)}, \ldots, \beta^{(T)}\}$, we generate predicted expression coefficient sequence $\widetilde{\boldsymbol{\beta}} = \{\widetilde{\beta}^{(1)}, \ldots, \widetilde{\beta}^{(T)}\}$. Denoting the LSTM network as R, the mapping can be formulated as

$$[\widetilde{\beta}^{(t)}, h^{(t)}, c^{(t)}] = R(E(s^{(t)}), h^{(t-1)}, c^{(t-1)}), \tag{6}$$

where E is an additional audio encoder (which is applied to the MFCC feature of audio sequences $s^{(t)}$), and $h^{(t)}$ and $c^{(t)}$ are hidden and cell states of LSTM unit at time t respectively.

We design a loss function containing two loss terms to optimize the network: a mean squared error (MSE) loss for expression coefficients, and an inter-frame continuity loss for dynamic expression. Denoting the shorthand notation of Eq. (6) as $\widetilde{\boldsymbol{\beta}} = \phi_1(\mathbf{s})$, the loss function is formulated as:

$$\begin{aligned}\mathcal{L}(R, E) &= \mathbb{E}_{\mathbf{s},\boldsymbol{\beta}}[(\boldsymbol{\beta} - \phi_1(\mathbf{s}))^2] \\ &\quad + \lambda_{e1}\mathbb{E}_{\mathbf{s}}[\sum_{t=0}^{T-1}(\phi_1(\mathbf{s})^{(t+1)} - \phi_1(\mathbf{s})^{(t)})^2],\end{aligned} \tag{7}$$

where inter-frame continuity loss is computed by the squared L_2 norm of the gradient of expression coefficients.

3) Multi-modal-input-to-head-pose mapping: In this module, we establish the mapping from the speech to the head pose. In Section III, we propose a two-step mapping strategy for head movement prediction: 1) characterize the head motion in each short video by a motion behavior pattern (modelled by mapping $g(\mathbf{y})$ in Sec. III-A), and 2) predict head pose from speech conditioned on the motion behavior pattern (modelled by mapping $f(\mathbf{a}, x)$ in Sec. III-B). Then we propose a head motion encoder E_m for $g(\mathbf{y})$ and a head motion decoder D_m for $f(\mathbf{a}, x)$ (Sec. III-C).

We use the head motion encoder and the head motion decoder to predict head pose. As shown in Figure 1, (1) the head motion encoder first extracts a motion behavior pattern from the head pose sequence of the input short video; (2) then the head motion decoder predicts a head pose sequence from both the audio features of the input speech and the motion behavior pattern.

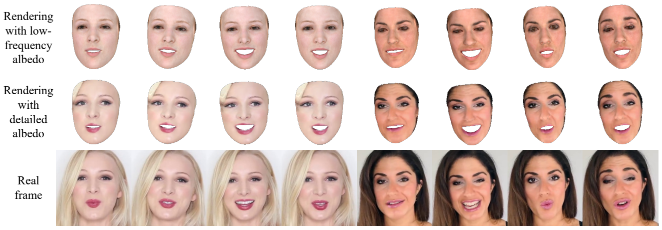

Rendering with low-frequency albedo
Rendering with detailed albedo
Real frame

Fig. 2. Comparison of the rendering effects using the low-frequency albedo and the detailed albedo.

B. Stage 2: from 3D facial animation to realistic talking face video generation

1) Rendering of 3D face with personalized pose: After reconstructing the 3D face of the target person and generating the expression and pose sequences, we obtain a sequence of 3DMM coefficients synchronized with the speech signal, in which (1) expression coefficients are from the speech of source person, (2) the identity, texture and illumination coefficients are from the input video of target person, and (3) head pose coefficients are determined by both the speech signal (of source person) and the input video (of target person). Given this sequence of 3DMM coefficients, we can render a video frame sequence of a talking face with personalized head poses, using the rendering engine [30].

If we compute the albedos from reconstructed 3DMM coefficients, these albedos are of low-frequency and too smooth, resulting in the rendered face images that do not appear visually similar to the input face images. An alternative is to compute a *detailed* albedo from input video; that is, we first project the reconstructed 3D shape (i.e., a face mesh) onto the image plane, and then we assign the pixel color to each mesh vertex. In this way, the albedo is computed by dividing illumination. Finally, the albedo from the frame with the most neutral expression and the smallest rotation angles is set as the albedo of the video.

The rendering effect using the low-frequency albedo is a bit far away realistic, while the rendering with the detailed albedo is much better; see Figure 2 for two examples; see demo video for visual comparison. In our system, we use both types of albedos. We propose a GAN module which refines rendered frames into realistic ones. Since the differences between the rendered frames of low-frequency albedo and real frames (in the facial region) are much larger than the differences between the rendered frames of detailed albedo and real frames (Figure 2), i.e., the domain gap is larger for low-frequency albedo, training the GAN module from rendered frames of low-frequency albedo to real frames needs more training data than detailed albedo. To train this GAN module, two steps are involved: (1) training on the common LRW dataset that contains multiple identities, and (2) using the input video of the target person to fine tune the GAN module. In the first step, we use the detailed albedo for rendering, because the common dataset contains few data (i.e., videos of only 1-2

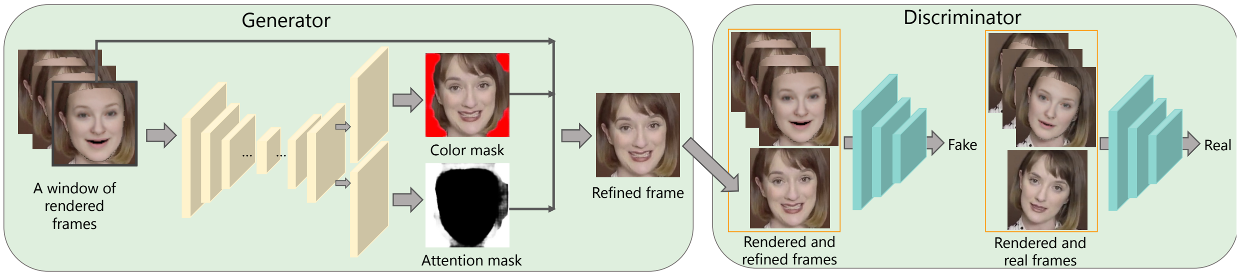


Fig. 3. Our rendering-to-realistic GAN for refining rendered frames into realistic ones. The generator takes a window of rendering frames as input, and generate a refined frame based on attention mechanism. The discriminator judges whether a frame is real or not.

seconds) for each person, which is not enough for refining the frames rendered with low-frequency albedo. In the second step, we use the low-frequency albedo, since the input video (about 10-15 seconds) can provide more training data of the target person to fine tune the synthesized frames (rendered with a low-frequency albedo) into realistic ones.

2) Background matching: So far the rendered frames only contain the facial region, without the hair and background that are also essential for a realistic talking face video. An intuitive solution is to reuse a background from the input video by matching the head pose. However, for a short video of 10-15 seconds, we only have about 300 frames to select a suitable background, which is very few and can be regarded as very sparse points in the possible high-dimensional pose space. Our experiment also shows that this intuitive solution cannot produce good video frames.

In our system, we propose to extract some keyframes from the synthesized pose sequence, where the keyframes correspond to critical head movements in the synthesized pose sequence. We choose the key frames to be the frames with largest head orientation in one axis in a short period of time, e.g., the frame with leftmost or rightmost head pose. Then we only match backgrounds for these keyframes. We call these matched backgrounds as key backgrounds. For those frames between two neighboring keyframes, we use linear interpolation to determine their backgrounds (similar to the background retiming technique in [45]). The pose in each frame is also modified to fit the background. Finally the whole rendered frames are assembled by including the matched backgrounds.

3) Rendering-to-realistic GAN for refining frames: The synthesized frames rendered by the light-weight graphic engine [30] are usually a bit far from realistic. To refine these frames into realistic ones, we propose to use a rendering-to-realistic GAN. The differences between our method and the previous GAN-based face reenactment (FR) methods (e.g., [23]) are:

- FR only refines the frames for a single, specified face identity, while our method can deal with various face identities; i.e., our GAN module can generate realistic frames given rendering results of different identities.
- FR uses thousands of frames to train a network for a single, specified face identity, while we only use a few frames for each identity when pretraining on a common LRW dataset containing over a thousand identities. After pretraining, we fine tune the network using the frames in the input short video for the target person.

The difference between our GAN module and the previous few-shot (FS) talking head method [24] is that FS deals with head pose and expression transfer in realistic frames, while we deal with rendering frame (less realistic) to realistic frame transfer.

We model the frame refinement process as a function Φ that maps from the rendered frame (i.e., synthesized frame rendered by the graphic engine) domain $\mathcal{R}$ to the real frame domain $\mathcal{T}$ using paired training data $\{(r_i, g_i)\}$, $r_i \in \mathcal{R}$ and $g_i \in \mathcal{T}$. Given a real frame g_i, the corresponding frame r_i in the training data is synthesized by rendering the 3D face reconstructed from g_i. To handle multiple-identity refinement, we build a GAN network that consists of a conditional generator G, and a conditional discriminator D (Figure 3). The conditional generator takes a window of rendered frames (i.e., a set of 3 adjacent frames r_{t-2}, r_{t-1}, r_t) as input, and synthesize a refined frame o_t using the U-Net [46] with attention mechanism. The conditional discriminator takes a window of rendered frames and either a refined frame or a real frame as input, and decides whether the frame is real or synthesized.

Attention-based generator G: We use an attention-based generator to refine rendered frames into realistic ones. Given a window of rendered frames (r_{t-2}, r_{t-1}, r_t) , the generator synthesizes both a color mask C_t and an attention mask A_t, and outputs a refined frame o_t that is the weighted average of the rendered frame and color mask:

$$o_t = A_t \cdot r_t + (1 - A_t) \cdot C_t \tag{8}$$

The attention mask specifies how much each pixel in the generated color mask contributes to the final refinement. The attention mask should attend to the regions that need to be changed during the translation, and the output color for these regions are mainly taken from the generated color mask. Since the main differences between the rendered frames and realistic frames are in the facial region, the attention mask A_t should attend to the facial region, and the color mask C_t should learn the realistic texture for facial region.

Our generator architecture is based on a U-Net structure and modify the last convolution block to two parallel convolution blocks in order to generate two outputs (i.e., color and attention masks), in which each one generates one mask. Experimental results show that our network can generate delicate target-person-dependent texture for various identities.

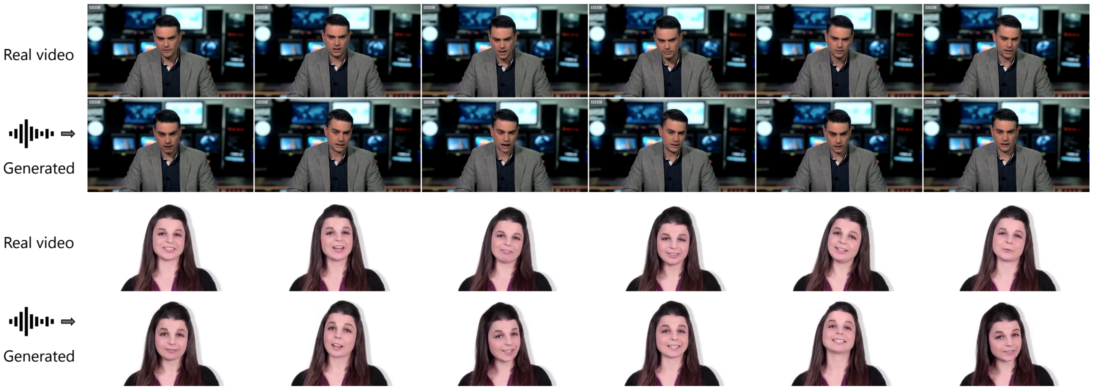


Fig. 4. Comparison of real talking face videos and our generated videos.

Discriminator D: The conditional discriminator takes a window of rendered frames and a checking frame (either a refined frame or a real frame) as input, and determines whether the checking frame is real or not. We adopt PatchGAN [37] architecture as our discriminator.

Loss function: The loss function of our GAN model contains three terms: an adversarial loss, an L_1 loss, and an attention loss [47] to prevent saturation of the attention mask A, which also enforces the smoothness of the attention mask. Denoting the input rendered frames as r, and the ground truth real frames as g, the loss function is formulated as:

$$\begin{aligned}\mathcal{L}(G,D) = &(\mathbb{E}_{r,g}[\log D(r,g)] + \mathbb{E}_r[\log(1-D(r,G(r)))]) \\ &+ \lambda_1\mathbb{E}_{r,g}[||g-G(r)||_1] + \lambda_2\mathbb{E}_r[||A||_2] \\ &+ \lambda_3\mathbb{E}_r[\sum_{i,j}^{H,W}(A_{i+1,j}-A_{i,j})^2 + (A_{i,j+1}-A_{i,j})^2]\end{aligned} \tag{9}$$

We train the GAN model to optimize the loss function:

$$G^* = \underset{G}{\operatorname{argmin}}\,\max_D \mathcal{L}(G,D) \tag{10}$$

V. Experiments

We implemented our method in PyTorch. All experiments are performed on a PC with a Titan Xp GPU. The comparison to a video-driven method are presented in the appendix C. Dynamic results can be found in the supplementary demo video.

A. Datasets

Head motion dataset. Talking face videos in most public lip reading datasets either are very short (e.g., about 1 second in LRW [40] and 3 seconds in GRID [48]), or contain rich camera movements (e.g., VoxCeleb [49] and LRS3-TED [50]). They are not suitable for learning head motion, because too short video does not contain sufficient head movement and for videos with rich camera movements, it is hard to decouple camera movement and head movement. Therefore, in our study we build a new head motion dataset by collecting high-resolution talking face videos from the Internet. The new dataset consists of 751 single-person talking videos and each video contains only one shot[3] without camera movements. The time length of these videos is in the range of 37-116 seconds (the average is 56 seconds). Various types of head movements (e.g. gentle, moderate and fierce) are contained in this dataset. We use this dataset to train the multi-modal-input-to-head-pose network.

LRW dataset. The LRW video dataset [40] is a publicly available lip reading dataset including five hundreds of distinct words, thousands of lip-reading instances for each word, and over a thousand speakers. Each instance is of 29 frames (about 1.16 seconds). The training set contains about 489k instances, and the test set contains 25k instances. We use this dataset to train the audio-to-expression network and the GAN module.

B. Training and test settings

The audio-to-expression network is trained on the LRW training set using the loss fuction in Eq.(7) for 20 epochs (learning rate 0.0002). The parameter λ_{e1} in Eq.(7) is 0.0001.

The head motion encoder (i.e., the mapping $g(\mathbf{y})$) is first pretrained on the head motion dataset using the contrastive loss in Eq.(1) for 10 epochs (learning rate 0.0001). Then the head motion encoder $g(\mathbf{y})$ and head motion decoder (i.e., the mapping $f(\mathbf{a},x)$) are trained jointly on the head motion dataset using the Eq.(2) in main paper for 10 epochs (learning rate 0.0001); the parameters in the equation are $\lambda_{p1}=5$, $\lambda_{p2}=1$, $\lambda_{p3}=1$.

The rendering-to-realistic GAN is trained in two steps. First, it is trained on the LRW training set[4] using the loss function in Eq.(9) for 1 epoch with learning rate 0.0001. The parameters in Eq.(9) are $\lambda_1=100$, $\lambda_2=2$, $\lambda_3=1e^{-5}$. Secondly, we fine-tune the GAN module using the input short video of 10-15 seconds (about 300 frames) using the same loss function for 60 epochs (learning rate 0.0001).

C. Comparison with state of the arts

Our method is proposed with a focus on personalized head pose predicted from an input short video (Fig.4), which meanwhile has comparable talking facial quality (e.g., lip synchronization and expression) as good as state-of-the-art methods. Since our method takes a speech signal (of a source

[3] A shot is a sequence of consecutive frames that was continuously captured by the same camera.

[4] The training set includes 489k videos and each video contains 29 frames, which are very large for GAN training. Then we randomly select a subset of 402k frames for training.

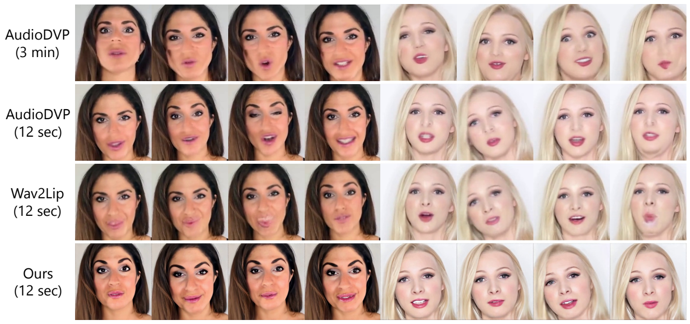


Fig. 5. Qualitative comparison of our method and state-of-the-art open-source methods using S+V inputs. AudioDVP (12 sec), Wav2Lip (12 sec) and ours use the same 12-second-long video as input.

person) and a short video (of a target person) as input, we compare our method with state-of-the-art methods that use similar input (we denote this kind of input as S+V input, where S for speech and V for short video).

It is worth noting that by removing (1) head motion encoder and decoder and (2) the fine tuning step of the GAN module from our system, our method (denoted as Ours-p) can take a speech signal (of a source person) and a face photo (of a target person) as input (which we denote as S+P input, where S for speech and P for photo), and output a synthesized talking face video without personalized head pose. We then compare Ours-p with state-of-the-art methods that use the same S+P input.

1) **Comparison with methods using S+V input***:* The state-of-the-art methods that use similar S+V (12s short video) input consists of: A) Wav2Lip [16]: can use video of arbitrary time length as input. So we can test it by inputting video of 12 seconds. B) AudioDVP [15]: is originally designed with input of 3-minute-long video. We adapt the source code provided by the authors to use input of 12-second-long video. So in this section, we compare two versions, AudioDVP (3 min) and AudioDVP (12 sec). C) NeuralVoicePuppetry [14]: is also designed with input of 3-minute-long video. D) TGRHM [6]: takes audio and K reference frames from a short video as inputs during test (we evaluate on its generated results of $K = 32$). E) FACIAL-GAN [10]: trains a person-specific model from a 2-3 minute video of a target person. F) LiveSpeechPortraits [11]: trains a person-specific model from a 3-5 minute video of a target person.

Wav2Lip, AudioDVP and NeuralVoicePuppetry methods copy the head poses from the input video to synthesize the output video. TGRHM learns the relation between audio and its paired head motion at time t, and predicts the head motion of subsequent audio at time $t + 1$. It cannot generate personalized head motion of the target person given another person's audio. LiveSpeechPortraits and FACIAL-GAN learn person-specific models and learn head motion for each subject separately. While our method propose a speech-to-head-pose mapping that generates personalized head poses for various identities, and can generate personalized head pose with any person's audio and generate high quality talking face video.

Comparison details:

- For A) and B), we retrain the models using the official training code and the same input short video data of different subjects as ours, and generate comparison groups to compare with ours. We compare with A) and B) using a user study and quantitative evaluation.
- For others, either code is partial or only generated examples are available: for C) and D), we only have access to the generated examples; and for E) and F), training code is not released, only a few pretrained models (each for one subject) are available.[5] So we use the results of given subjects for quantitative evaluation.

TABLE II
SUBJECTIVE SCORE EVALUATION OF OUR METHOD AND STATE-OF-THE-ART METHODS USING S+V INPUT. THE 2ND AND 3RD COLUMNS SHOW THE PERCENTAGE OF PARTICIPANTS WHO CHOSE EACH METHOD AS THE BEST FOR HEAD MOTION (HM) CONSISTENCY (WITH THE SPEECH) AND OVERALL QUALITY, RESPECTIVELY. THE LAST COLUMN SHOWS THE POSE VARIETY SCORE.

Methods	HM consistency ↑	Overall ↑	Pose variety score ↑
AudioDVP (3 min)	23.4%	26.8%	1.15
AudioDVP (12 sec)	5.1%	3.3%	1.16
Wav2Lip (12 sec)	15.9%	29.2%	1.31
Ours-P (12 sec)	**55.6%**	**40.8%**	**2.96**

User study. It is challenging to evaluate the visual quality and naturalness of synthesized videos, in particular regarding the human face and head pose. We designed a user study to compare these methods based on subjective score evaluation. Randomly selected thirteen short videos of different identities were used as input to generate thirteen comparison groups (two of which are illustrated in Figure 5) for evaluation. 33 participants attended the user study. Each of them compared talking face videos generated by different methods and answered three questions for each group. For a fair comparison, each group was presented by a randomly shuffled order of four videos, produced by AudioDVP (3 min), AudioDVP (12 sec), Wav2Lip and our method.

Three questions used in the study: 1) The first question was to ask participants to select the best video in term of **concordant head movement and speech**. 2) The second question was ask participants to select the best video in term of **overall quality** (including head movement, lip synchronization and video quality). 3) Finally, to evaluate the personalized head movements, given a short video of the target person, we generate two synthesized videos using two different speeches. Then the third question was to ask the participants to watch two generated videos side by side and rate the head motion difference between them: 1 (exactly the same), 2 (slightly different), 3 (quite different), and 4 (totally different). We calculate the **pose variety** score as the average rating. High pose variety score indicates that the method can generate diverse head poses for different speeches.

The statistics data of 33 participants are summarized in Table II. The results show that our method achieves the best consistency between head motion and speech, the best overall quality, and the best pose variety.

[5] For C), source code is not available and synthesis request is disabled in online demo, we can only use the generated examples provided in the demo. For D), the source code for audio-driven scenario is partial, lacking audio2expression module, we only have access to its generated examples. For E) and F), the models are person-specific, since the training codes are not released, only pretrained models for a few target subjects are available, we can only test on the provided target subjects. (accessed 2021-Nov-26.)

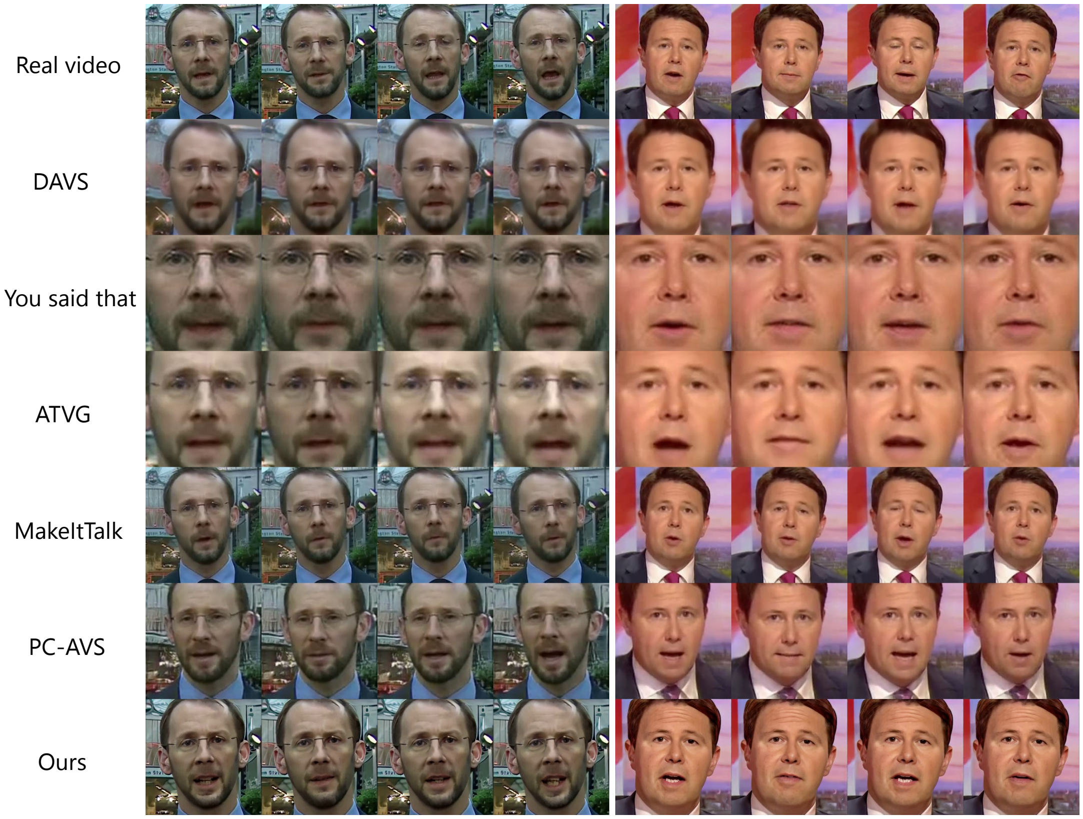


Fig. 6. Qualitative comparison of Ours-p and state-of-the-art methods using S+P input.

TABLE III
QUANTITATIVE EVALUATION OF DIFFERENT METHODS USING S+V INPUT.

Methods	V length	AV offset ↓	CPBD ↑
	3-min video		
LiveSpeechPortraits	3-5 min	1.93	0.20
FACIAL-GAN	2-3 min	3.18	0.17
NeuralVoicePuppetry	3 min	2.04	0.29
AudioDVP	3 min	1.65	**0.31**
	12-sec video		
TGRHM	32 ref frames	**1.49**	0.16
AudioDVP	12 sec	2.92	0.30
Wav2Lip	12 sec	2.00	0.26
Ours	12 sec	1.65	**0.31**

TABLE IV
QUANTITATIVE RESULTS OF OURS-P AND STATE-OF-THE-ART METHODS USING S+P INPUT.

Metric	PSNR ↑	SSIM ↑	LMD ↓
Chen [53]	29.65	0.73	1.73
Wiles [54]	29.82	0.75	1.60
You said that [3]	29.91	0.77	1.63
SDA [55]	29.44	0.68	2.32
DAVS [13]	29.81	0.73	1.73
ATVG [2]	30.91	**0.81**	**1.37**
MakeItTalk [7]	30.21	0.72	2.11
PC-AVS [17]	28.77	0.57	2.59
Ours-p	**31.19**	0.76	1.56

Quantitative evaluation. In addition to the user study, we further evaluate different methods using the following two quantitative metrics. Since the input speech and short video are from different subjects, the ground truth are not available, we adopt two widely used metrics to evaluate the lip synchronization and visual quality: (1) **AV offset**: This metric measures the quality of audio-to-video synchronization. We use SyncNet [51], which is an audio-to-video synchronisation network, to evaluate the audio and visual streams offset (AV offset) of a video. Lower AV offset value indicates better audio-to-video synchronization. 2) **Cumulative probability blur detection (CPBD) measure [52]**: is a no-reference objective sharpness metric. We use CPBD to evaluate the sharpness of the results. Higher value indicates shaper results, less blur and better visual quality.

For Wav2Lip, AudioDVP and Our method, we evaluate the metrics on all comparison groups. For NeuralVoicePuppetry and TGRHM, since we only have access to the generated examples in their online demo or shared zipped files, we only evaluate the metrics on provided examples. For FACIAL-GAN and LiveSpeechPortraits (person-specific and training not available), we test on the provided subjects.

The results are summarized in Table III, showing that (1) TGRHM, Ours and AudioDVP (3 min) are better than other methods on lip synchronization, (2) Ours and AudioDVP (3 min) are better than others on sharpness and visual quality.

2) **Comparison with methods using S+P input***:* By a simple adaption of our method, the model Ours-p can generate a talking face video based on S+P input. In this section, we compare Ours-p with state-of-the-art methods using S+P input, i.e., Chen [53], Wiles [54], You said that [3], DAVS [13], ATVG [2], SDA [55], MakeItTalk [7], PC-AVS [17] using their pretrained models.

We evaluate these methods on the test set of LRW dataset [40], which contains 25,000 videos. We use the first frame and the audio of each video as input to generate results for each method[6]. In this setting, the real video provides the ground-truth. We compare the generated results of different methods with the ground-truth videos, and follow ATVG [2] to use three widely-used metrics for the talking face generation evaluation: the classic PSNR and SSIM metrics for image quality and identity preservation evaluation; and the landmark distance (LMD) for lip synchronization. The results are summarized in Table IV, showing that Ours-p has the best PSNR, comparable quality with ATVG [2] on SSIM and LMD, and outperforms the other seven methods on all three metrics. Some qualitative comparisons are illustrated in Fig. 6.

In addition to the comparison with methods in the audio-driven class, we also compare our method with a state-of-the-art method [25] in the video-driven class. Details are in the appendix C.

D. Head Pose Behavior Quantitative Analysis

To objectively evaluate the quality of personalized head pose, we propose a new metric HS to measure the similarity of head poses between the generated video and real video. We follow [56] to use the three Euler angles to model head movements, i.e., pitch, yaw, and roll corresponding to the movement of head nod, head shake/turn, and head tilt, respectively. We compute a histogram P_{real} of pose angles in the input real video, and a histogram P_{gen} of pose angles in the generated video. Then we compute the normalized Wasserstein distance W_1 [57] between P_{real} and P_{gen}. The lower the distance, the more similar the two head pose distribution. Our new metric HS is formulated as

$$HS = 1 - W_1(P_{real}, P_{gen}) \qquad (11)$$

where HS is in the range $[0, 1]$ and larger HS value indicates higher similarity of head pose behavior. We evaluate the HS metric on real videos and corresponding videos generated by our method. The average HS score of our method is 0.871, and the maximum and minimum score are 0.958 and 0.703, respectively. While the average HS score of MakeItTalk [7] is 0.636, and the maximum and minimum score are 0.847 and 0.239, respectively. This shows that our generated videos have a high similarity to real videos in term of head pose behavior.

E. Ablation study on head pose prediction

We predict personalized head pose from multi-modal input. If we remove the head pose estimation, as shown in the first row of Figure 7, the generated results are good in lip synchronization, but look rigid due to the fixed head position, which is not natural. In comparison, with the head pose estimation network, our method generates natural head motion.

F. Ablation study on the rendering-to-realistic GAN module

We use the rendering-to-realistic GAN module to refine rendered frames into realistic frames. If we remove the GAN module, i.e., taking the rendered frames and applying the enhancement module, as shown in the first row of Figure 8, the generated results lack many facial details, and look unreal. In comparison, with the rendering-to-realistic GAN module, our method generates much more realistic results.

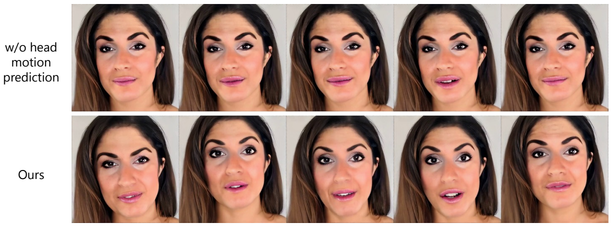


Fig. 7. Ablation study on head pose prediction. Row 1 shows the generated results without pose estimation in the first stage. Row 2 shows the results of our method. See demo video for animation.

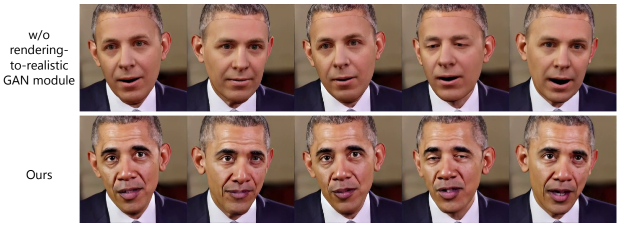


Fig. 8. Ablation study on the rendering-to-realistic GAN module. Row 1 shows the generated results without the GAN module in the second stage. Row 2 shows the results of our method.

G. Discussion and limitations

In our system, since we use 3DMM 3D face reconstruction, which represents texture using low-dimensional coefficients and makes the albedo too smooth and lack some facial details, there is some gap between the rendered faces and the target person. During experiments we found the GAN module can bridge this gap and refine the less realistic renderings to realistic frames for different people. However, if the 3D face reconstruction fails, and the rendered faces are far away from realistic frames, the system may not generate good quality results.

VI. Conclusion

In this paper, to learn and predict personalized head pose from the multi-modal input, we propose a novel two-step mapping: 1) we extract a motion behavior pattern from the head motion sequence of the input video; 2) conditioned on the motion behavior pattern, we map the auditory feature of the input speech signal to the personalized head pose sequence. Based on the two-step mapping, we propose a talking face video generation system. Since talking head movements often contain in-plane and out-of-plane head rotations, to overcome the difficulty, we reconstruct the 3D face and use the 3D facial animation to bridge the gap between audio-visual-driven head pose learning and realistic talking face video generation. Experiments results and user studies show that our method generates high-quality talking head video with personalized head pose given an arbitrary audio. We generate personalized

[6] For PC-AVS, since it requires another pose source video besides S+P input, we randomly take one video from the LRW test set as the pose source for each S+P input.

head motion using a *one-to-many mapping* based on head motion behavior, which can apply to multiple subjects.

ACKNOWLEDGEMENT

This work was partially supported by the National Key Research and Development Program of China (2019YFC1521104), the Natural Science Foundation of China (61725204, 72192821, 61972157), Tsinghua University Initiative Scientific Research Program (20211080093), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300), and Art major project of National Social Science Fund (I8ZD22).

REFERENCES

[1] J. R. Nazzaro and J. N. Nazzaro, "Auditory versus visual learning of temporal patterns," *Journal of Experimental Psychology*, vol. 84, no. 3, pp. 477–8, 1970.

[2] 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.

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

[4] T. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7539–7548.

[5] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 95:1–95:13, 2017.

[6] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *16th European Conference (ECCV)*, vol. 12354, 2020, pp. 35–51.

[7] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 221:1–221:15, 2020.

[8] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3661–3670.

[9] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo, "3D talking face with personalized pose dynamics," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[10] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3867–3876.

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

[12] 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)*, 2019, pp. 919–925.

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

[14] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *16th European Conference (ECCV)*, 2020, pp. 716–731.

[15] X. Wen, M. Wang, C. Richardt, Z. Chen, and S. Hu, "Photorealistic audio-driven video portraits," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 12, pp. 3457–3466, 2020.

[16] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *The 28th ACM International Conference on Multimedia (MM)*, 2020, pp. 484–492.

[17] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4176–4186.

[18] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.

[19] D. Greenwood, I. Matthews, and S. D. Laycock, "Joint learning of facial expression and head pose from speech," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 2484–2488.

[20] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, no. 3, pp. 555–568, 2002.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.

[22] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, "Bringing portraits to life," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 196:1–196:13, 2017.

[23] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 163:1–163:14, 2018.

[24] E. Zakharov, A. Shysheya, E. Burkov, and V. S. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9459–9468.

[25] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 7135–7145.

[26] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt, "State of the art on monocular 3D face reconstruction, tracking, and applications," *Computer Graphics Forum*, vol. 37, no. 2, pp. 523–550, 2018.

[27] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3D face rigs from monocular video," *ACM Trans. Graph.*, vol. 35, no. 3, pp. 28:1–28:15, 2016.

[28] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1585–1594.

[29] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1294–1307, 2019.

[30] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3D morphable model regression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8377–8386.

[31] J. Zhang, L. Lin, J. Zhu, and S. C. H. Hoi, "Weakly-supervised multi-face 3d reconstruction," *CoRR*, vol. abs/2101.02000, 2021.

[32] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.

[33] M. Zhang and Q. Ling, "Supervised pixel-wise GAN for face super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1938–1950, 2021.

[34] D. Wei, X. Xu, H. Shen, and K. Huang, "GAC-GAN: A general method for appearance-controllable human video motion transfer," *IEEE Trans. Multimedia*, vol. 23, pp. 2457–2470, 2021.

[35] F. Peng, L. Yin, L. Zhang, and M. Long, "CGR-GAN: CG facial image regeneration for antiforensics based on generative adversarial network," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2511–2525, 2020.

[36] L. Zhu, S. Kwong, Y. Zhang, S. Wang, and X. Wang, "Generative adversarial network-based intra prediction for video coding," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 45–58, 2020.

[37] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[38] K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, and H. Li, "Realistic dynamic facial textures from a single image using GANs," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5439–5448.

[39] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in *13th Asian Conference on Computer Vision (ACCV 2016)*, 2016, pp. 87–103.

[41] T. Yang, P. Ren, X. Xie, and L. Zhang, “GAN prior embedded network for blind face restoration in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 672–681.

[42] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3D face model for pose and illumination invariant face recognition,” in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)*, 2009, pp. 296–301.

[43] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, “Facewarehouse: A 3D facial expression database for visual computing,” *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, 2014.

[44] R. Ramamoorthi and P. Hanrahan, “An efficient representation for irradiance environment maps,” in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2001)*, 2001, pp. 497–500.

[45] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, “Text-based editing of talking-head video,” *ACM Trans. Graph.*, vol. 38, no. 4, pp. 68:1–68:14, 2019.

[46] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, 2015, pp. 234–241.

[47] A. Pumarola, A. Agudo, A. M. Martínez, A. Sanfeliu, and F. Moreno-Noguer, “GANimation: Anatomically-aware facial animation from a single image,” in *15th European Conference (ECCV)*, 2018, pp. 835–851.

[48] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5 Pt 1, pp. 2421–4, 2006.

[49] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2616–2620.

[50] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: a large-scale dataset for visual speech recognition,” *CoRR*, vol. abs/1809.00496, 2018.

[51] J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in *ACCV 2016 Workshops*, ser. Lecture Notes in Computer Science, vol. 10117, 2016, pp. 251–263.

[52] N. D. Narvekar and L. J. Karam, “A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection,” in *2009 International Workshop on Quality of Multimedia Experience*. IEEE, 2009, pp. 87–91.

[53] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, “Lip movements generation at a glance,” in *15th European Conference (ECCV)*, 2018, pp. 538–553.

[54] O. Wiles, A. S. Koepke, and A. Zisserman, “X2face: A network for controlling face generation using images, audio, and pose codes,” in *15th European Conference (ECCV)*, 2018, pp. 690–706.

[55] K. Vougioukas, S. Petridis, and M. Pantic, “End-to-end speech-driven facial animation with temporal GANs,” in *British Machine Vision Conference (BMVC)*, 2018, p. 133.

[56] R. E. Kaliouby and P. Robinson, “Generalization of a vision-based computational model of mind-reading,” in *Affective Computing and Intelligent Interaction, First International Conference (ACII 2005)*, 2005, pp. 582–589.

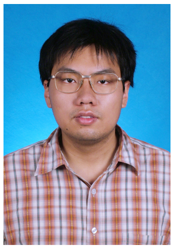
[57] C. Villani, *Topics in optimal transportation*. American Mathematical Society, 2003, no. 58.

Ran Yi is an Assistant Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. She received the BEng degree and the PhD degree from Tsinghua University, China, in 2016 and 2021. Her research interests include computer vision, computer graphics and computational geometry.

Zipeng Ye is a Ph.D student with Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tsinghua University, China, in 2017. His research interests include computational geometry and computer vision.

Zhiyao Sun is a Ph.D student with Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tsinghua University, China, in 2021. His research interests include computer vision and computer graphics.

Juyong Zhang is an Associate Professor in the School of Mathematical Sciences at University of Science and Technology of China. He received the BS degree from the University of Science and Technology of China in 2006, and the PhD degree from Nanyang Technological University, Singapore. His research interests include computer graphics, computer vision, and numerical optimization. He is now an associate editor of IEEE Trans. Multimedia.



Guoxin Zhang is currently with Kuaishou Technology. He received the B.E degree and the Ph.D degree from Tsinghua University, China, in 2007 and 2012. His research interests include computer graphics and computer vision.

Pengfei Wan received the B.E. degree in electronic engineering and information science from the University of Science and Technology of China, and the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology. He is currently with Kuaishou Technology. His research interests include image/video processing, computer vision, machine learning, and AR/VR/MR.

Hujun Bao is a Cheung Kong Professor in the school of computer science and technology in Zhejiang University, and the director of State Key Laboratory of CAD & CG. He received the BS and PhD degrees in applied mathematics from Zhejiang University in 1987 and 1993, respectively. His research interests include geometry computing, vision computing, real-time rendering and virtual reality.

Yong-Jin Liu is a Professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include computational geometry, computer graphics and computer vision. He is a senior member of the IEEE and ACM.