Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation

Lincheng Li,[1*] **Suzhen Wang,**[1*] **Zhimeng Zhang,**[1*] **Yu Ding,**[1*]
Yixing Zheng,[1] **Xin Yu,**[2] **Changjie Fan**[1]
[1] Netease Fuxi AI Lab
[2] University of Technology Sydney
{lilincheng, wangsuzhen, zhangzhimeng, dingyu01, zhengyixing01, fanchangjie}@corp.netease.com
xin.yu@uts.edu.au

Abstract

In this paper, we propose a novel text-based talking-head video generation framework that synthesizes high-fidelity facial expressions and head motions in accordance with contextual sentiments as well as speech rhythm and pauses. To be specific, our framework consists of a speaker-independent stage and a speaker-specific stage. In the speaker-independent stage, we design three parallel networks to generate animation parameters of the mouth, upper face, and head from texts, separately. In the speaker-specific stage, we present a 3D face model guided attention network to synthesize videos tailored for different individuals. It takes the animation parameters as input and exploits an attention mask to manipulate facial expression changes for the input individuals. Furthermore, to better establish authentic correspondences between visual motions (i.e., facial expression changes and head movements) and audios, we leverage a high-accuracy motion capture dataset instead of relying on long videos of specific individuals. After attaining the visual and audio correspondences, we can effectively train our network in an end-to-end fashion. Extensive experiments on qualitative and quantitative results demonstrate that our algorithm achieves high-quality photo-realistic talking-head videos including various facial expressions and head motions according to speech rhythms and outperforms the state-of-the-art.

Introduction

Talking-head synthesis technology aims to generate a talking video of a specific speaker with authentic facial animations from an input speech. The output talking-head video has been employed in many applications, such as intelligent assistance, human-computer interaction, virtual reality, and computer games. Due to its wide applications, talking-head synthesis has attracted a great amount of attention.

Many previous works that take audios as input mainly focus on synchronizing lower facial parts (*e.g.*, mouths), but often neglect animations of the head and upper facial parts (*e.g.*, eyes and eyebrows). However, holistic facial expressions and head motions are also viewed as critical channels to deliver communicative information (Ekman 1997). For example, humans unconsciously use facial expressions and head movements to express their emotions (Mignault and Chaudhuri 2003). Thus, generating holistic facial expressions and head motions will lead to more convincing person-talking videos.

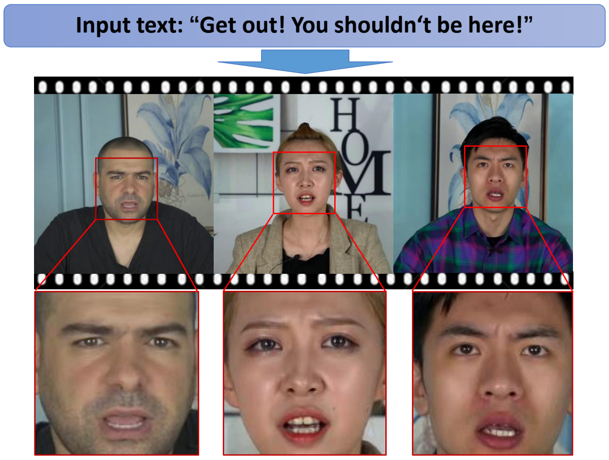

Figure 1: Our method produces emotional, rhythmic and photo-realistic talking-head videos from input texts.

Furthermore, since the timbre gap between different individuals may lead the acoustic features in the testing utterances to lying outside the distribution of the training acoustic features, prior arts built upon the direct association between audio and visual modalities may also fail to generalize to new speakers' audios (Chou et al. 2018). Consequently, the acoustic feature-based frameworks do not work well on input speeches from different people with distinct timbres or synthetic speeches (Sadoughi and Busso 2016).

Unlike previous works, we employ time-aligned texts (*i.e.*, text with aligned phoneme timestamps) as input features instead of acoustics features to alleviate the timbre gap issue. In general, time-aligned texts can be extracted from audios by speech recognition tools or generated by text-to-speech tools. Since the spoken scripts are invariant to different individuals, our text-based framework is able to achieve robust performance against different speakers.

This paper presents a novel framework to generate holistic facial expressions and corresponding head animations according to spoken scripts. Our framework is composed of two stages, *i.e.*, a speaker-independent stage and a speaker-

[*] Equal contribution. Yu Ding is the corresponding author.

specific stage. In the speaker-independent stage, our networks are designed to capture generic relationships between texts and visual appearances. Unlike previous methods (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Taylor et al. 2017; Fried et al. 2019) that only synthesize and blend mouth region pixels, our method intends to generate holistic facial expression changes and head motions. Hence, we design three networks to map input texts into animation parameters of the mouth, upper face and head pose respectively. Furthermore, we employ a motion capture system to construct the correspondences between high-quality facial expressions as well as head motions and audios as our training data. Thus, our collected data can be used for training our speaker-independent networks effectively without requiring long-time talking videos of specified persons.

Since the animation parameters output by our speaker-independent networks are generic, we need to tailor the animation parameters to the specific input speaker to achieve convincing generated videos. In the speaker-specific stage, we take the animation parameters as input and then exploit them to rig a given speaker's facial landmarks. In addition, we also develop an adaptive-attention network to adapt the rigged landmarks to the speaking characteristics of the specified person. In doing so, we only require a much shorter reference video (around 5 minutes) of the new speaker, instead of more than one hour speaker-specific videos often requested by previous methods (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Fried et al. 2019).

Overall, our method produces photo-realistic talking-head videos from a short reference video of a target performer. The generated videos also present rich details of the performer, such as realistic clothing, hair, and facial expressions.

Related work

Facial Animation Synthesis

Facial animation synthesis pre-defines a 3D face model and generates the animation parameters to control the facial variation. LSTM (Hochreiter and Schmidhuber 1997) is widely used in facial animation synthesis for sequential modeling. Several works take BiLSTM (Pham, Cheung, and Pavlovic 2017), CNN-LSTM (Pham, Wang, and Pavlovic 2017) or carefully-designed LSTM (Zhou et al. 2018) with regression loss, GAN loss (Sadoughi and Busso 2019) or multi-task training strategy (Sadoughi and Busso 2017) to synthesize full facial/mouth animation. However, LSTM tends to work slower due to the sequential computation. CNN is proven to have comparable ability to deal with sequential data (Bai, Kolter, and Koltun 2018). Some works employ CNN to animate mouth or full face from acoustic features (Karras et al. 2017; Cudeiro et al. 2019) or time-aligned phonemes (Taylor et al. 2017). Head animation synthesis focuses on synthesizing head pose from input speech. Some works direct regress head pose with BiLSTM (Ding, Zhu, and Xie 2015; Greenwood, Matthews, and Laycock 2018) or the encoder of transformer (Vaswani et al. 2017). More precisely, head pose generation from speech is a one-to-many mapping, Sadoughi and Busso (2018) employ GAN (Goodfellow et al. 2014; Mirza and Osindero 2014; Yu et al. 2019b,a) to retain the diversity.

Face Video Synthesis

Audio-driven. Audio-driven face video synthesis directly generates 2D talking video from input audio. Previous works (Vougioukas, Petridis, and Pantic 2019; Chen et al. 2018; Zhou et al. 2019; Wiles, Sophia, and Zisserman 2018; Prajwal et al. 2020) utilize two sub-modules to compute face embedding feature and audio embedding feature for the target speaker, then fuse them as input to a talking-face generator. Another group of works decouple geometry generation and appearance generation into two stages. The geometry generation stage infers appropriate facial landmarks, which is taken as input by the appearance generation stage. Landmarks are inferred with speaker-specific model (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Das et al. 2020; Zhou et al. 2020) or linear principal components (Chen et al. 2019, 2020). Thies et al. (2020) generate expression coefficients of a 3D Morphable Model (3DMM), then employ a neural renderer to generate photo-realistic images. Fried et al. (2019) infer expression parameters by searching and blending existing expressions of the reference video, then employ a recurrent neural network to generate the modified video. Although also taking text as input, their method generates novel sentences inefficiently (10min-2h) due to the viseme search. Besides, both works fail to control the upper face and head pose to match the speech rhythm and emotion.

Video-driven. Video-driven methods transfer expressions of one person to another. Several works (Ha et al. 2020; Zeng et al. 2020; Song et al. 2019; Siarohin et al. 2019) take a single image as the identity input. Other works take videos (Thies et al. 2015, 2018) as identity input to improve visual quality. Thies et al. (2016) reconstruct and renders a mesh model and fill in the inner mouth as output, the reconstructed face texture stays constant while talking. Some works directly generate 2D images with GAN instead of 3D rendering (Nirkin, Keller, and Hassner 2019; Zakharov et al. 2019; Wu et al. 2018; Thies, Zollhöfer, and Nießner 2019). Kim et al. (2019) preserve the mouth motion style based on sequential learning on the unpaired data of the two speakers. Alternatively, our work generates paired mouth expression data to make the style learning easier. Kim et al. (2018) also employs a 3DMM to render geometry information. Instead of transferring existing expressions, our method generates new expressions from text. Furthermore, our method preserves the speaker's mouth motion style and designs an adaptive-attention network to obtain higher image resolution and better visual quality.

Text-based Talking-head Generation

Our framework takes the time-aligned text as input and outputs the photo-realistic talking-head video. It can be generalized to a specific speaker with about 5 minutes of his/her talking video (reference video). Figure 2 illustrates the pipeline of our framework. Taking time-aligned text as input, G^{mou}, G^{upp} and G^{hed} separately generate speaker-independent animation parameters of mouth, upper face and

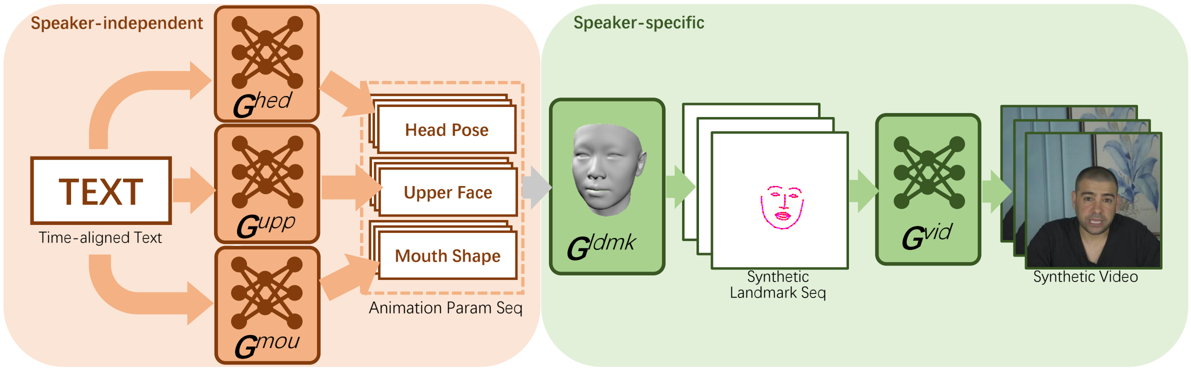

Figure 2: Pipeline of our method. The speaker-independent stage takes the time-aligned text as input and generates head pose, upper face, and mouth shape animation parameters. The speaker-specific stage then produces synthetic talking-head videos from the animation parameters.

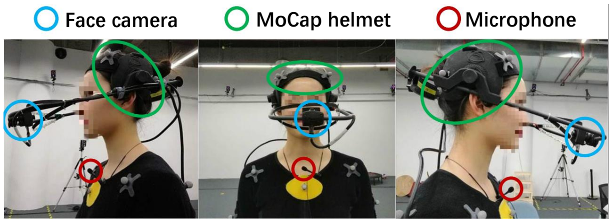

Figure 3: The collection of Mocap dataset. The recording is carried out by a professional actress wearing a helmet. Markers on the helmet offer information of head pose. The infrared camera attached to the helmet records accurate facial expressions.

head pose. Instead of learning from the reference video, they take advantage of a Mocap dataset for higher accuracy. Since a small error in geometry inference may lead to obvious artifacts in appearance inference, we introduce a 3D face module G^{ldmk} to incorporate the head and facial expression parameters and convert them to speaker-specific facial landmark sequence. Finally, G^{vid} synthesizes the speaker-specific talking-head video according to the facial landmark sequence by rendering the texture of hair, face, upper torso and background.

Mocap Dataset

To obtain high-fidelity full facial expressions and head pose, we record an audiovisual dataset relying on a motion capture (Mocap) system[1] shown in Figure 3. The collected data includes the mouth parameter sequence $m^{mou} = \{m_t^{mou}\}_{t=1}^T$ where $m_t^{mou} \in \mathbb{R}^{28}$, the upper face parameter sequence $m^{upp} = \{m_t^{upp}\}_{t=1}^T$ where $m_t^{upp} \in \mathbb{R}^{23}$ and the head pose parameter sequence $m^{hed} = \{m_t^{hed}\}_{t=1}^T$ where $m_t^{hed} \in \mathbb{R}^6$. T is the length of frames in an utterance. m^{mou} and m^{upp} are defined as blendshape weights following the definition of Faceshift. Each blendshape stands for some part of the face movement, *e.g.*eye-open, mouth-left. We record 865 emotional utterances of a professional actress in English (203 surprise, 273 anger, 255 neutral and 134 happiness), each of which lasts from 3 to 6 seconds. A time alignment analyzer [2] is employed to compute the duration of each phoneme and each word from audio. According to the alignment result, we represent the word sequence and phoneme sequence as $w = \{w_t\}_{t=1}^T$ and $ph = \{ph_t\}_{t=1}^T$ separately, where w_t and ph_t are the word and phoneme uttered at the t-th frame. In this way, we build a high-fidelity Mocap dataset including m^{mou}, m^{upp}, m^{hed}, w and ph, which is then used to train the speaker-independent generators. Another Chinese dataset (925 utterances from 3 to 6 seconds) is similarly built. Both datasets are released for research purposes[3].

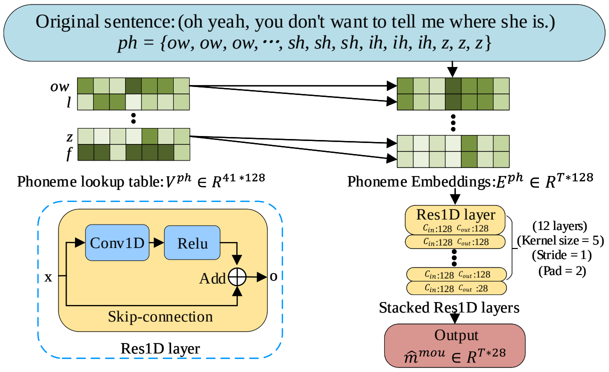
Original sentence: (oh yeah, you don't want to tell me where she is.)
$ph = \{ow, ow, ow, \cdots, sh, sh, sh, ih, ih, ih, z, z, z\}$
ow
l
z
f
Phoneme lookup table: $V^{ph} \in R^{41*128}$
Phoneme Embeddings: $E^{ph} \in R^{T*128}$
Conv1D
Relu
x
Add
o
Skip-connection
Res1D layer
Res1D layer
C_{in}:128 C_{out}:128
C_{in}:128 C_{out}:128
C_{in}:128 C_{out}:128
C_{in}:128 C_{out}:28
(12 layers)
(Kernel size = 5)
(Stride = 1)
(Pad = 2)
Stacked Res1D layers
Output
$\hat{m}^{mou} \in R^{T*28}$

Figure 4: Mouth animation generator.

Mouth Animation Generator

Since the mouth animation mainly contributes to uttering phonemes instead of semantic structures, G^{mou} learns a mapping from ph to m^{mou} ignoring w, as shown in Figure 4. The first step is to convert ph from phoneme space into the embedding vectors E^{ph} in a more flexible space. We construct a trainable lookup table (Tang et al. 2014) V^{ph} to meet the goal, which is randomly initialized and updated

[1] Dynamixyz, http://www.dynamixyz.com

[2] qdreamer.com

[3] https://github.com/FuxiVirtualHuman/Write-a-Speaker

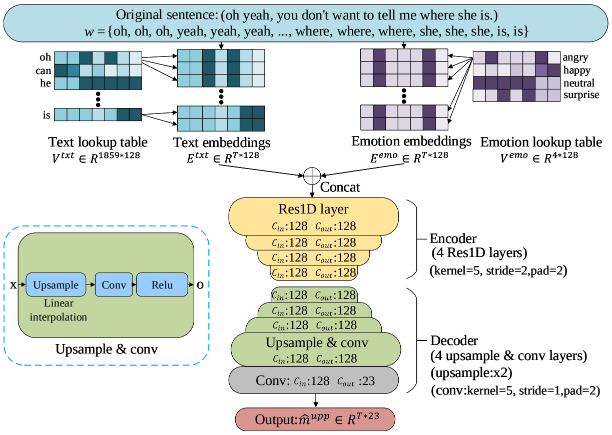
Original sentence: (oh yeah, you don't want to tell me where she is.)
w = {oh, oh, oh, yeah, yeah, yeah, ..., where, where, where, she, she, she, is, is}
oh
can
he
is
Text lookup table
$V^{txt} \in R^{1859*128}$
Text embeddings
$E^{txt} \in R^{T*128}$
Emotion embeddings
$E^{emo} \in R^{T*128}$
Emotion lookup table
$V^{emo} \in R^{4*128}$
angry
happy
neutral
surprise
Concat
Res1D layer
C_{in}:128 C_{out}:128
Encoder
(4 Res1D layers)
(kernel=5, stride=2,pad=2)
Upsample & conv
Decoder
(4 upsample & conv layers)
(upsample:x2)
(conv:kernel=5, stride=1,pad=2)
Conv: C_{in}:128 C_{out}:23
Output: $\hat{m}^{upp} \in R^{T*23}$
x
Upsample
Linear interpolation
Conv
Relu
o
Upsample & conv

Figure 5: Upper facial expression generator.

in the training stage. Afterwards, The stacked Res1D layers take E^{ph} as input and output synthetic mouth parameter sequence $\hat{m}^{mou}$ according to co-articulation effects. We design the structure based on CNN instead of LSTM for the benefits of parallel computation.

We apply L_1 loss and LSGAN loss (Mao et al. 2017) for training G^{mou}. The L_1 loss is written as

$$L_1^{mou} = \frac{1}{T}\sum_{i=1}^{T}(\|m_i^{mou} - \hat{m}_i^{mou}\|_1), \tag{1}$$

where m_i^{mou} and $\hat{m}_i^{mou}$ are the real and generated vector of the ith frame separately. The adversarial loss is denoted as

$$L_{adv}^{mou} = arg \min_{G^{mou}} \max_{D^{mou}} L_{GAN}(G^{mou}, D^{mou}). \tag{2}$$

Inspired by the idea of patch discriminator (Isola et al. 2017), D^{mou} is applied on temporal trunks of blendshape which also consists of stacked Res1D layers. The objective function is written as

$$L(G^{mou}) = L_{adv}^{mou} + \lambda_{mou} L_1^{mou}. \tag{3}$$

Upper Face/Head Pose Generators

While mouth motions contribute to speech co-articulation, upper facial expressions and head motions tend to convey emotion, intention, and speech rhythm. Therefore, G^{upp} and G^{hed} are designed to capture longer-time dependencies from w instead of ph. They share the same network and differ from that of G^{mou}, as illustrated in Figure 5. Similar to V^{ph}, a trainable lookup table V^{txt} maps w to embedding vectors E^{txt}. In order to generate m^{upp} with consistent emotion, an emotion label (surprise, anger, neutral, happiness) is either detected by a text sentiment classifier (Yang et al. 2019), or explicitly assigned for the specific emotion type. Another trainable lookup table V^{emo} projects the emotion label to embedding vectors E^{emo}. E^{txt} and E^{emo} are fed to an encoder-decoder network to synthesize m^{upp}. Benefits from the large receptive field, the encoder-decoder structure captures long-time dependencies between words.

Since synthesizing m^{upp} from text is a one-to-many mapping, the L_1 loss is replaced with SSIM loss (Wang et al. 2004). SSIM simulates the human visual perception and has benefit of extracting structural information. We extend SSIM to perform on each parameter respectively, formulated as

$$L_S^{upp} = 1 - \frac{1}{23}\sum_{i=1}^{23}\frac{(2\mu_i\hat{\mu}_i + \delta_1)(2cov_i + \delta_2))}{(\mu_i^2 + \hat{\mu}_i^2 + \delta_1)(\sigma_i^2 + \hat{\sigma}_i^2 + \delta_2))}. \tag{4}$$

$\mu_i/\hat{\mu}_i$ and $\sigma_i/\hat{\sigma}_i$ represent the mean and standard deviation of the i dimension of real/synthetic m^{upp}, and cov_i is the covariance. δ_1 and δ_2 are two small constants. The GAN loss is denoted as

$$L_{adv}^{upp} = arg \min_{G^{upp}} \max_{D^{upp}} L_{GAN}(G^{upp}, D^{upp}). \tag{5}$$

where D^{upp} shares the same structure with D^{mou}. The objective function is written as

$$L(G^{upp}) = L_{adv}^{upp} + \lambda_{upp} L_S^{upp}. \tag{6}$$

G^{hed} shares the same network and loss but ignores V^{emo} to generate m^{hed}, as the variation of head poses in different emotions is less significant than that of facial expressions.

Style-Preserving Landmark Generator

G^{ldmk} reconstructs the 3D face from the reference video, then drive it to obtain speaker-specific landmark images. A multi-linear 3DMM $U(s, e)$ is constructed with shape parameters $s \in \mathbb{R}^{60}$ and expression parameters $e \in \mathbb{R}^{51}$. The linear shape basis are taken from LSFM (Booth et al. 2018) and scaled by the singular values. We sculpture 51 facial blendshapes on LSFM as the expression basis following the definition of Mocap dataset, so that e is consistent with (m_t^{upp}, m_t^{mou}). A 3DMM fitting method is employed to estimate s of the reference video. Afterwards, we drive the speaker-specific 3D face with generated $\hat{m}^{hed}$, $\hat{m}^{mou}$ and $\hat{m}^{upp}$ to get the landmark image sequence. Our earlier experiments show that videos generated from the landmark images and rendered dense mesh are visually indifferent, we therefore choose landmark images to cut down a renderer.

Furthermore, speakers may use different mouth shapes to pronounce the same word, e.g. some people tend to open their mouths larger than others, and people are sensitive to the mismatched styles. Meanwhile, the generic $\hat{m}^{upp}$ and $\hat{m}^{hed}$ work fine among different people in practice. Hence, we retarget $\hat{m}^{mou}$ to preserve the speaker's style while leaving $\hat{m}^{upp}$ and $\hat{m}^{hed}$ unchanged. On one hand, we extract time-aligned text from the reference video and generate $\hat{m}^{mou}$ using G^{mou}. On the other hand, we estimate personalized $\breve{m}^{mou}$ from the reference video using 3DMM. In this way, we obtain paired mouth shapes pronouncing the same phonemes. With the paired data, the style-preserving mapping from $\hat{m}^{mou}$ to $\breve{m}^{mou}$ is easily learnt. A two-layer fully-connected network with MSE loss works well in our experiments. We use the mapped $\breve{m}^{mou}$ to produce the landmark images.

Photo-realistic Video Generator

G^{vid} produces the talking-head video $\{\hat{I}_t\}_{t=1}^{T}$ frame by frame from the landmark images. $\hat{I}_t$ depicts the speaker's

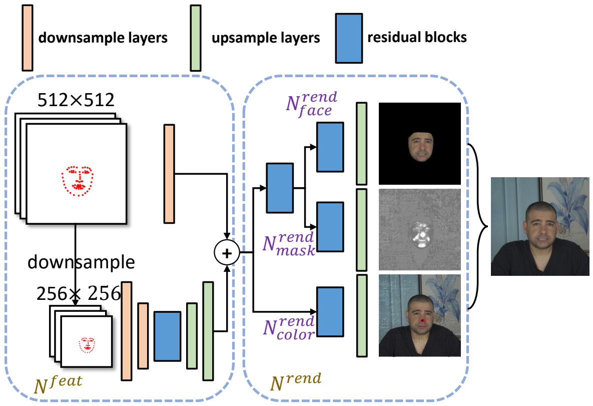

Figure 6: Photo-realistic video generator.

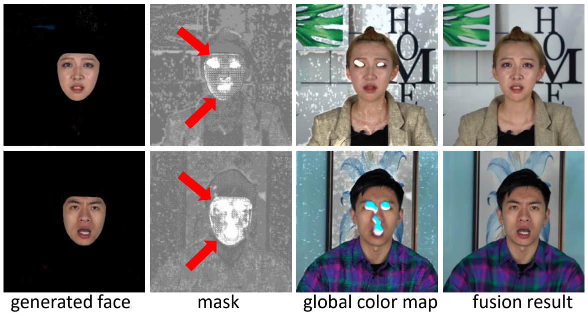

Figure 7: Sample outputs of our photo-realistic video generator. It shows that the adaptive-attention mask is able to distinguish the region of mouth and eyes from other regions.

full facial expression, hair, head and upper torso poses, and the background at the t-th frame. Considering the high temporal coherence, we construct the conditional space-time volume V as input of G^{vid} by stacking the landmark images in a temporal sliding window of length 15.

Although typical image synthesis networks (Isola et al. 2017; Wang et al. 2018; Yu and Porikli 2016, 2017a,b) are able to produce reasonable head images, their outputs tend to be blurry on areas with high-frequency movements, especially the eye and mouth regions. The possible explanation is that the movements of eye and mouth are highly correlated with landmarks while the torso pose and background are less, so it is not the best solution to treat all parts as a whole. Motivated by the observation, we design an adaptive-attention structure. As shown in Figure 6, G^{vid} is composed by a feature extraction network N^{feat} and self-attention rendering network N^{rend}. To extract features from high resolution landmark images, N^{feat} consists of two pathways of different input scales. The extracted features of the two pathways are element-wise summed. N^{rend} renders talking-head images from the latent features. To model the different correlations of body parts, we design a composite of three parallel sub-networks. N^{rend}_{face} produces the target face $\hat{I}^{face}$. N^{rend}_{clr} is expected to compute the global color map $\hat{I}^{color}$, with hair, upper body, background and so on. N^{rend}_{mask} produces the adaptive-attention fusion mask M that focus on the high-frequency-motion regions. The final generated image $\hat{I}_t$ is given by

$$\hat{I}_t = M * \hat{I}^{face} + (1 - M) * \hat{I}^{color}. \qquad (7)$$

Figure 7 shows the details of our attention mask.

We follow the discriminators of pix2pixHD (Wang et al. 2018), consisting of 3 multi-scale discriminators D_1^{vid}, D_2^{vid} and D_3^{vid}. The inputs of them are $\hat{I}_t/I_t$ and V, where I_t is the real frame. The adversarial loss is defined as:

$$L_{adv}^{vid} = \min_{G^{vid}} \max_{D_1^{vid}, D_2^{vid}, D_3^{vid}} \sum_{i=1}^{3} L_{GAN}(G^{vid}, D_i^{vid}), \qquad (8)$$

To capture the fine facial details we adopt the perceptual loss (Johnson, Alahi, and Fei-Fei 2016), following Yu et al. (2018)

$$L_{perc} = \sum_{i=1}^{n} \frac{1}{W_i H_i C_i} \|F_i(I_t) - F_i(\hat{I}_t)\|_1, \qquad (9)$$

where $F_i \in \mathbb{R}^{W_i \times H_i \times C_i}$ is the feature map of the i-th layer of VGG-19 (Simonyan and Zisserman 2014). Matching both lower-layer and higher-layer features guides the generation network to learn both fine-grained details and a global part arrangement. Besides, we use L_1 loss to supervise the generated $\hat{I}_{face}$ and $\hat{I}_t$:

$$L_1^{img} = \|I_t - \hat{I}_t\|_1, L_1^{face} = \|I_t^{face} - \hat{I}_t^{face}\|_1. \qquad (10)$$

I_{face} is cropped from I_t according to the detected landmarks (Baltrusaitis et al. 2018).

The overall loss is defined as:

$$L(G^{vid}) = \alpha L_{perc} + \beta L_1^{img} + \gamma L_1^{face} + L_{adv}^{vid}. \qquad (11)$$

Experiments and Results

We implement the system using PyTorch on a single GTX 2080Ti. The training of the speaker-independent stage takes 3 hours on the Mocap dataset. The training of the speaker-specific stage takes one day on a 5 mins' reference video. Our method produces videos of 512 × 512 resolution at 5 frames per second. More implementation details are introduced in the supplementary material. We compare the proposed method with state-of-the-art audio/video driven methods, and evaluate the effectiveness of the submodules. Video comparisons are shown in the supplementary video.

Comparison to Audio-driven Methods

We first compare our method with Neural Voice Puppetry (NVP) (Thies et al. 2020) and Text-based Editing (TE) (Fried et al. 2019), which achieve state-of-the-art visual quality by replacing and blending mouth region pixels of the reference video. As shown in Figure 8, while achieving similar visual quality on non-emotional speech, our method additionally controls the upper face and head motion to match the sentiment and rhythm of emotional audios. In contrast, NVP and TE do not have mechanisms to model sentiments of audio.

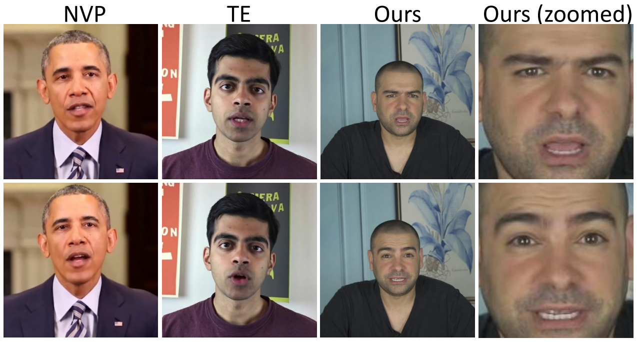

Figure 8: Comparison with NVP and TE. Our approach matches the sentiment and rhythm of emotional audios.

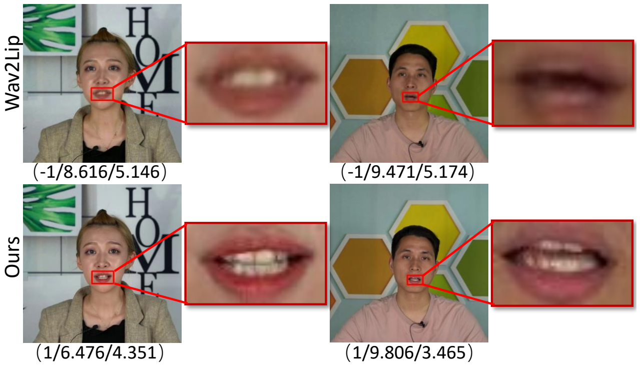

Figure 9: Comparison with Wav2Lip. Metrics of SyncNet are listed below (offset/distance/confidence).

We then compare our method with Wav2Lip (Prajwal et al. 2020) in Figure 9, which only requires a reference video of a few seconds. Metrics of SyncNet (Chung and Zisserman 2016) are listed below each image. Although their method produces accurate lip shapes from audio, we can observe the obvious artifacts in the inner mouth. Our method is compared to ATVGNet (Chen et al. 2019) in Figure 10, which produces talking head videos from a single image. Their method focuses on low resolution cropped front faces while our method generates high-quality full head videos. Considering their method learns identity information from one image instead of a video, the visual quality gap is as expected.

Comparison to Video-driven Methods

We also compare our method with Deep Video Portrait (DVP) (Kim et al. 2018), whose original intention is expression transfer. We reproduce DVP and replace their detected animation parameters with our generated animation parameters for fair comparison. Results are shown in Figure 11. Although our method uses sparse landmarks instead of rendered dense mesh, we synthesize better details on mouth and eye regions.

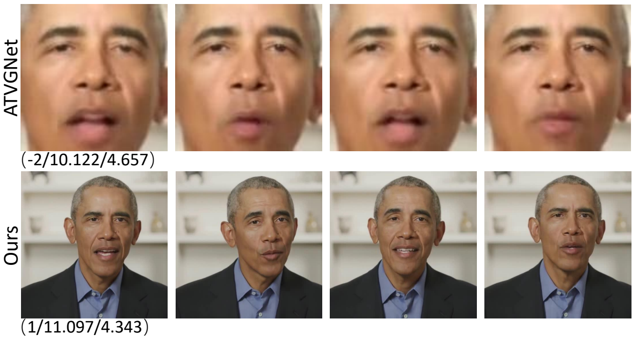

Figure 10: Comparison with ATVGNet. Metrics of SyncNet are listed below (offset/distance/confidence).

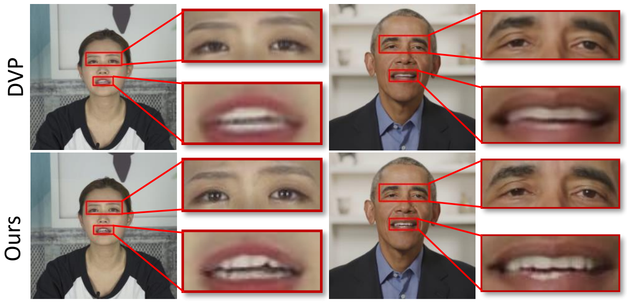

Figure 11: Comparison with DVP.

Evaluation of Submodules

In order to evaluate G^{mou} and G^{upp}, we reproduce the state-of-the-art facial animation synthesis works (Karras et al. 2017; Pham, Wang, and Pavlovic 2017; Sadoughi and Busso 2017; Taylor et al. 2017; Cudeiro et al. 2019; Sadoughi and Busso 2019). For fair comparison, their input features and network structures are retained and the output is replaced with facial expression parameters. To further evaluate the loss terms, we additionally conduct an experiment by removing the GAN loss in Equation 3 and 6 (Ours w/o GAN). The groundtruth test data is selected from the high accuracy Mocap dataset. For mouth parameters, we measure MSE of m^{mou} and lips landmark distance (LMD) on 3D face mesh. LMD is measured on 3D face mesh instead of 2D images to avoid the effect of head pose variation. For upper face parameters, we measure SSIM of m^{upp}. Results are shown in Table 1. Both G^{mou} and G^{upp} perform better than the above methods.

To prove the superiority of G^{vid}, we compare G^{vid} with pix2pix (Isola et al. 2017), pix2pixHD (Wang et al. 2018) and photo-realistic rendering network of DVP (denoted as DVPR). To evaluate the results, we apply multiple metrics including SSIM, Fréchet Inception Distance (FID) (Heusel et al. 2017), Video Multimethod Assessment Fusion (VMAF) and Cumulative Probability of Blur Detection (CPBD) (Narvekar and Karam 2011). For fair comparison, we take the same space-time volume as the input of all networks and train them on the same datasets. Table 2 shows the quantitative results, and Figure 12 shows the qualitative

Table 1: Quantitative evaluattion G^{mou} and G^{upp}.

	MSE (mouth)↓	LMD (mouth)↓	SSIM (upper face)↑
(Karras et al. 2017)	88.7470	0.0690	0.0931
(Pham, Wang, and Pavlovic 2017)	109.0163	0.0742	0.0889
(Sadoughi and Busso 2017)	103.3375	0.0721	0.0793
(Taylor et al. 2017)	89.5907	0.0699	–
(Cudeiro et al. 2019)	91.2150	0.0713	–
(Sadoughi and Busso 2019)	89.2853	0.0694	–
Ours w/o GAN	89.2143	0.0693	0.1879
Ours	**87.2378**	**0.0684**	**0.2655**

Table 2: Quantitative evaluation of G^{vid}.

		SSIM↑	FID↓	VMAF↑	CPBD↑
speaker1	pix2pix	0.9466	0.1279	62.75	0.1233
	pix2pixHD	0.9455	0.02711	65.42	0.2517
	DVPR	0.9371	0.02508	57.75	0.2607
	Ours	**0.9490**	**0.01452**	**66.68**	**0.2682**
speaker2	pix2pix	0.9026	0.04360	60.32	0.1083
	pix2pixHD	0.8998	0.01883	60.37	0.2572
	DVPR	0.9031	0.009456	62.27	0.2859
	Ours	**0.9042**	**0.003252**	**63.76**	**0.2860**
speaker3	pix2pix	0.9509	0.04631	72.15	0.2467
	pix2pixHD	0.9499	0.005940	74.64	0.3615
	DVPR	0.9513	0.005232	71.12	0.3642
	Ours	**0.9514**	**0.003262**	**74.76**	**0.3661**

comparison. Our approach is able to produce higher quality of images, especially on teeth and eyes regions.

Ablation Study

We perform an ablation study to evaluate other components of our framework, results are shown in Figure 13. We remove or replace several submodules to construct the input of G^{vid}. The first condition removes G^{ldmk} and directly input animation parameters to G^{vid} (w/o LDMK). Due to the lack of explicit geometry constraint, the output contains some twisted and jittered face regions. The second condition uses G^{ldmk} but removes the mouth style mapping (w/o MM). The speaker in the output video opens his mouth smaller than in the reference video for pronunciation, preserving the mismatched style of the actress of the Mocap dataset. The third condition additionally replaces the sparse landmarks with dense 3D face mesh (dense). The visual quality of the output is visually indifferent with that of our method, indicating that the sparse geometry constraint is good enough for G^{vid}. Figure 14 shows another ablation study to evaluate the effectiveness of each loss terms in G^{vid}. All loss terms contribute to the visual quality.

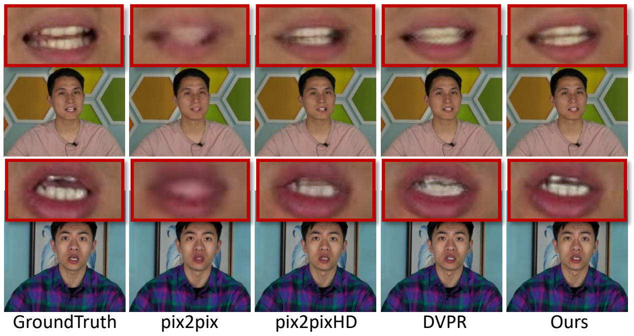

Figure 12: Comparison of G^{vid} and the state-of-the-arts.

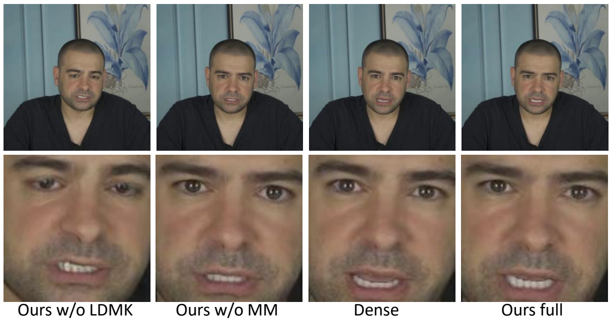

Figure 13: Results of different conditions.

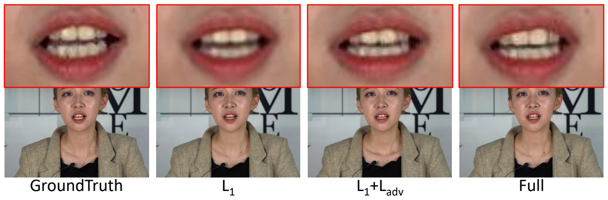

Figure 14: Results from different loss terms of G^{vid}.

User Study

We further conduct an online user study to evaluate the quality of the output videos. We compare our method with groundtruth videos (GT), ours with extracted m^{upp} and m^{hed} from reference videos instead of generated (Ours w/o E&H), DVP, Wav2Lip. We generate 5 sentences of the same speaker in the same resolution for each method, to obtain $5 \times 5 = 25$ video clips. The audios are extracted from the reference video. 60 participants are asked to rate the realism of each video clip. Results are listed in Table 3 ($60 \times 5 = 300$ ratings for each method). Only 91% of GT are judged as real, indicating that participants are overcritical when trying to detect synthesized videos. Even with the comparison of real videos, our results are judged as real in 52% of the cases. our method outperforms all compared methods significantly ($p < 0.001$) in both mean score and 'judged as real' proportion. Results of 'Ours w/o E&H' contain expression and head motion that do not match the speech sentiment and rhythm. The difference between 'Ours' and 'Ours w/o E&H' validates the effectiveness of our generated emotional upper face expressions and rhythmic head motions. The main reason of lower scores of DVP and Wav2Lip may be the artifacts in the inner mouth.

Limitations

Our work has several limitations. The proposed method takes advantage of a high-quality Mocap dataset. Our approach is restricted to produce speakers uttering in English or Chinese, because we have only captured Mocap datasets of the two languages. The amount of Mocap data is also insufficient to capture more detailed correspondences of motions and semantic and syntactic structures of text input. In

Table 3: Results of the user study. Participants are asked to rate the videos by 1-completely fake, 2-fake, 3-uncertain, 4-real, 5-completely real. Percentage numbers are rounded.

	1	2	3	4	5	Mean	'real'(4+5)
GT	0%	1%	8%	37%	54%	4.45	91.3%
Wav2Lip	11%	29%	28%	30%	3%	2.85	32.3%
DVP	11%	29%	42%	17%	1%	2.67	18.0%
Ours w/o E&H	3%	22%	43%	26%	7%	3.13	33.0%
Ours	1%	9%	38%	37%	15%	**3.56**	**51.7%**

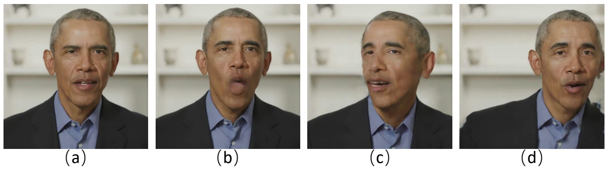

Figure 15: Failure cases from extreme parameters, including (a) upper facial expression; (b) mouth expression; (c) head rotation; (d) head translation.

the near future, we will record Mocap data of more languages and release them for the research purpose. Our rendering network cannot tackle with dynamic background and complex upper torso movements, such as shrugging, swinging arms, hunching back, extreme head poses an so on. The generated videos will degenerate if the expected expression or head motion is beyond the scope of the reference video. The effect of emotion is ignored on the generated lip and head animations. Figure 15 shows some failure cases. In the future, we will be devoted to addressing the above problems.

Conclusion

This paper presents a text-based talking-head video generation framework. The synthesized video displays the emotional full facial expressions, rhythmic head motions, the upper torso movements, and the background. The generation framework can be adapted to a new speaker with 5 minutes of his/her reference video. Our method is evaluated through a series of experiments, including qualitative evaluation and quantitative evaluation. The evaluation results show that our method can generate high-quality photo-realistic talking-head videos and outperforms the state-of-the-art. To the best of our knowledge, our work is the first to produce full talking-head videos with emotional facial expressions and rhythmic head movements from the time-aligned text representation.

Appendix

3DMM Fitting

We select $N_k = 30$ keyframes and aim to find the optimal variable set $\boldsymbol{X} = (s, m_1^{hed}, e_1, ..., m_{N_k}^{hed}, e_{N_k})$, where m_k^{hed} and e_k are the pose and expression parameters of the k-th keyframe. We focus on the $N_l = 68$ facial landmark consistency by minimizing the following energy function:

$$F(X) = \sum_{k=1}^{N_k}(\sum_{i=1}^{N_l} Dis(p_{k,i}, P(U(s,e_k)^{(i)}, m_k^{hed})) + \lambda_e \left\|e_k\right\|_2^2) + \lambda_s \left\|s\right\|_2^2, \qquad (12)$$

where $p_{k,i}$ is the coordinate of the i-th landmark detected from the k-th keyframe (Baltrusaitis et al. 2018), and $U^{(i)}$ is the i-th 3D landmark on mesh U. $P(U^{(i)}, m_k^{hed})$ projects $U^{(i)}$ with pose m_k^{hed} into image coordinates. $Dis(\cdot,\cdot)$ measures the distance of the projected mesh landmark and the detected image landmark. The regularization weights are set to $\lambda_e = 10^{-4}$ and $\lambda_s = 10^{-4}$. We employ the Levenberg-Marquard algorithm for the optimization.

Network Structure and Training

The size of V^{ph} is 41×128, where 41 is the number of phonemes and 128 is the phoneme embedding size. The row vectors of $E^{ph} \in \mathbb{R}^{T\times 128}$ are picked up from V^{ph} according to the phoneme indexes. The size of V^{txt} is 1859×128, where 1859 means 1858 words and one 'unknown' flag for all other words, and 128 is the word embedding size. The size of V^{emo} is 4×128. Each row of V^{emo} represents an emotion embedding. N_{face}^{rend} and N_{mask}^{rend} share the first 3 residual blocks. The top layer of N_{face}^{rend}/N_{clr}^{rend} is activated by tanh and that of N_{mask}^{rend} is done by sigmoid. The loss weights are set to $\lambda_{mou} = 50$, $\lambda_{upp} = 100$, $\alpha = 10$, $\beta = 100$, and $\gamma = 100$. We use the Adam (Kingma and Ba 2014) optimizer for all networks. For training G^{mou}, G^{upp} and G^{hed}, we set $\beta_1 = 0.5, \beta_2 = 0.99, \epsilon = 10^{-8}$, batch size of 32, and set the initial learning rate as 0.0005 for the generators and 0.00001 for the discriminators. The learning rates of G^{mou} stay fixed in the first 400 epoches and linearly decay to zero within another 400 epoches. The learning rates of G^{upp} and G^{hed} keep unchanged in the first 50 epoches and linearly decay to zero within another 50 epoches. We randomly select $1 \sim 3$ words as 'unknown' in each sentence to improve the performance from limited training data. For training G^{vid}, we set $\beta_1 = 0.5, \beta_2 = 0.999, \epsilon = 10^{-8}$, batch size of 3, and initial learning rate of 0.0002 with linear decay to 0.0001 within 50 epochs.

Ethical Consideration

To ensure proper use, we firmly require that any result created using our algorithm must be marked as synthetic with watermarks. As part of our responsibility, for the positive applications, we intend to share our dataset and source code so that it can not only encourage efforts in detecting manipulated video content but also prevent the abuse. Our text-based talking head generation work can contribute to many positive applications, and we encourage further discussions and researches regarding the fair use of synthetic content.

References

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* .

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *FG 2018*, 59–66. IEEE.

Booth, J.; Roussos, A.; Ponniah, A.; Dunaway, D.; and Zafeiriou, S. 2018. Large scale 3D morphable models. *IJCV* 126(2-4): 233–254.

Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; and Xu, C. 2020. Talking-head Generation with Rhythmic Head Motion. *arXiv preprint arXiv:2007.08547* .

Chen, L.; Li, Z.; K Maddox, R.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *ECCV*, 520–535.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 7832–7841.

Chou, J.-c.; Yeh, C.-c.; Lee, H.-y.; and Lee, L.-s. 2018. Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations. *Interspeech* 501–505.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *ACCVW*.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, learning, and synthesis of 3D speaking styles. In *CVPR*, 10101–10111.

Das, D.; Biswas, S.; Sinha, S.; and Bhowmick, B. 2020. Speech-driven Facial Animation using Cascaded GANs for Learning of Motion and Texture. In *ECCV*.

Ding, C.; Zhu, P.; and Xie, L. 2015. BLSTM neural networks for speech driven head motion synthesis. In *INTERSPEECH*, 3345–3349.

Ekman, R. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D. B.; Genova, K.; Jin, Z.; Theobalt, C.; and Agrawala, M. 2019. Text-based editing of talking-head video. *TOG* 38(4): 1–14.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, 2672–2680.

Greenwood, D.; Matthews, I.; and Laycock, S. D. 2018. Joint Learning of Facial Expression and Head Pose from Speech. In *Interspeech*, 2484–2488.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. MarioNETte: Few-shot Face Reenactment Preserving Identity of Unseen Targets. In *AAAI*, volume 34, 10893–10900.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 6626–6637.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 1125–1134.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711. Springer.

Karras, T.; Aila, T.; Laine, S.; Herva, A.; and Lehtinen, J. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *TOG* 36(4): 94.

Kim, H.; Elgharib, M.; Zollhöfer, M.; Seidel, H.-P.; Beeler, T.; Richardt, C.; and Theobalt, C. 2019. Neural style-preserving visual dubbing. *TOG* 38(6): 1–13.

Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Nießner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; and Theobalt, C. 2018. Deep video portraits. *TOG* 37(4): 1–14.

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

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *ICCV*, 2794–2802.

Mignault, A.; and Chaudhuri, A. 2003. The Many Faces of a Neutral Face: Head Tilt and Perception of Dominance and Emotion. *J. Nonverbal Behav.* 27: 111–132.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Narvekar, N. D.; and Karam, L. J. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *TIP* 20(9): 2678–2683.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 7184–7193.

Pham, H. X.; Cheung, S.; and Pavlovic, V. 2017. Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach. In *CVPRW*, 80–88.

Pham, H. X.; Wang, Y.; and Pavlovic, V. 2017. End-to-end learning for 3d facial animation from raw waveforms of speech. *arXiv preprint arXiv:1710.00920* .

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In *MM*, 484–492.

Sadoughi, N.; and Busso, C. 2016. Head Motion Generation with Synthetic Speech: A Data Driven Approach. In *INTERSPEECH*, 52–56.

Sadoughi, N.; and Busso, C. 2017. Joint learning of speech-driven facial motion with bidirectional long-short term memory. In *IVA*, 389–402. Springer.

Sadoughi, N.; and Busso, C. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *ICASSP*, 6169–6173. IEEE.

Sadoughi, N.; and Busso, C. 2019. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing* .

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. In *NeurIPS*, 7137–7147.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Song, L.; Cao, J.; Song, L.; Hu, Y.; and He, R. 2019. Geometry-aware face completion and editing. In *AAAI*, volume 33, 2506–2513.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *TOG* 36(4): 1–13.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*, 1555–1565.

Taylor, S.; Kim, T.; Yue, Y.; Mahler, M.; Krahe, J.; Rodriguez, A. G.; Hodgins, J.; and Matthews, I. 2017. A deep learning approach for generalized speech animation. *TOG* 36(4): 93.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural Voice Puppetry: Audio-driven Facial Reenactment. *ECCV* .

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *TOG* 38(4): 1–12.

Thies, J.; Zollhöfer, M.; Nießner, M.; Valgaerts, L.; Stamminger, M.; and Theobalt, C. 2015. Real-time expression transfer for facial reenactment. *TOG* 34(6): 183–1.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2387–2395.

Thies, J.; Zollhöfer, M.; Theobalt, C.; Stamminger, M.; and Nießner, M. 2018. Headon: Real-time reenactment of human portrait videos. *TOG* 37(4): 1–13.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2019. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In *CVPRW*, 37–40.

Wang, T.; Liu, M.; Zhu, J.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 8798–8807.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P.; et al. 2004. Image quality assessment: from error visibility to structural similarity. *TIP* 13(4): 600–612.

Wiles, O.; Sophia, K.; and Zisserman, A. 2018. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 670–686.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Change Loy, C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 603–619.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 5753–5763.

Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face super-resolution guided by facial component heatmaps. In *ECCV*, 217–233.

Yu, X.; Fernando, B.; Hartley, R.; and Porikli, F. 2019a. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *TPAMI* .

Yu, X.; and Porikli, F. 2016. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 318–333.

Yu, X.; and Porikli, F. 2017a. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI*.

Yu, X.; and Porikli, F. 2017b. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 3760–3768.

Yu, X.; Shiri, F.; Ghanem, B.; and Porikli, F. 2019b. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *TPAMI* .

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 9459–9468.

Zeng, X.; Pan, Y.; Wang, M.; Zhang, J.; and Liu, Y. 2020. Realistic Face Reenactment via Self-Supervised Disentangling of Identity and Pose. In *AAAI*, volume 34, 12757–12764.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, volume 33, 9299–9306.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. MakeltTalk: speaker-aware talking-head animation. *TOG* 39(6): 1–15.

Zhou, Y.; Xu, Z.; Landreth, C.; Kalogerakis, E.; Maji, S.; and Singh, K. 2018. Visemenet: Audio-driven animator-centric speech animation. *TOG* 37(4): 161.