# Emotion-Aware Talking Head Synthesis with Landmark Translation Networks

Anonymous CVPR 2021 submission

Paper ID 2925

## Abstract

*We present a new approach for speech-driven talking head generation. Our goal is to synthesize a talking character that matches the given speech audio and conveys a similar emotion. Our formulation explicitly models the emotion of a speaker and generates plausible talking heads with high visual quality, temporal smoothness, and strong emotional expressiveness. We first extract the speaker's emotion features from the speech audio clip using an encoder network and generate 3D facial landmarks with lip motions synchronized with the speech audio. Next we modify these 3D landmarks to express corresponding face emotions through edge deformations. Finally, we translate the deformed landmarks sequence into a realistic talking head video by transferring the facial components. We generate the face video by integrating these these components and ensuring temporal consistency. We conduct extensive quantitative and qualitative experiments along with user studies, which demonstrate that our method generates talking heads with high visual quality and conveys the perceived emotion.*

## 1. Introduction

> "Let's not forget that the little emotions are the great captains of our lives and we obey them without realizing it."
>
> —Vincent van Gogh

As Vincent van Gogh said, emotion is subtle, but it helps people express themselves and understand each other. It serves as an important social signal that conveys people's physiological states, intentions, and relationships. Humans are sensitive to various kinds of expressions, including voice tone, facial expressions, body gestures, etc. Therefore, animating a realistic emotional talking head has a wide range of applications in both interpersonal and human computer interactions. For example, when having a teleconference, people can drive their own avatar to express both content and emotions through only their voice, without using any cameras. This is extremely helpful when the participants have limited bandwidth. In addition, emotional talking head generation will enable virtual agents to help people in a more amicable way. For instance, avatars can have calm expressions when they are dealing with humans who are angry or agitated. Similarly, avatars can convey a happy facial expressions while expressing birthday or festival wishes.

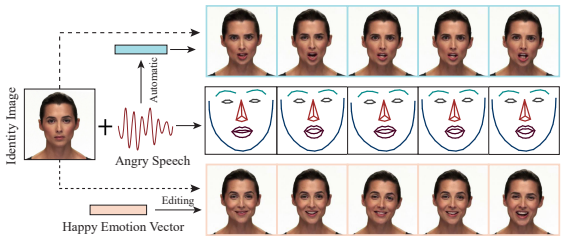


Figure 1. Problem Formulation. Given a neutral face image and an emotional speech clip as input, our model automatically synthesizes a talking head video with expressive emotions of the audio. Our model further enables conditional emotional talking face generation, provided a given emotion feature vector.

Many works have been proposed in the context of speech-driven talking head generation, i.e., modeling the dynamics or movements of the talking character in synchronization with the speech signal. Some works enable facial expression generation by randomization [57] or by analyzing the audio signal [6]. However, their generated facial expressions and head motion may not be expressive enough to convey the emotional state corresponding to the input speech audio. Other methods [40, 26] can generate talking heads that convey a given emotion, but they cannot automatically detect and generate emotions from audio. Some techniques [6, 66] use facial landmarks as an intermediate representation for expression and motion manipulation, as it is easier to control these landmarks as opposed to the pixels. However, such methods may not be able to account for geometric consistency when the facial landmarks are deformed. [8] uses an attention-based mechanism to improve the visual quality. Some methods [48, 60, 65, 8, 57, 6, 66] use discriminators on the integrated face to generate realistic facial images. Overall, there are still many challenges in

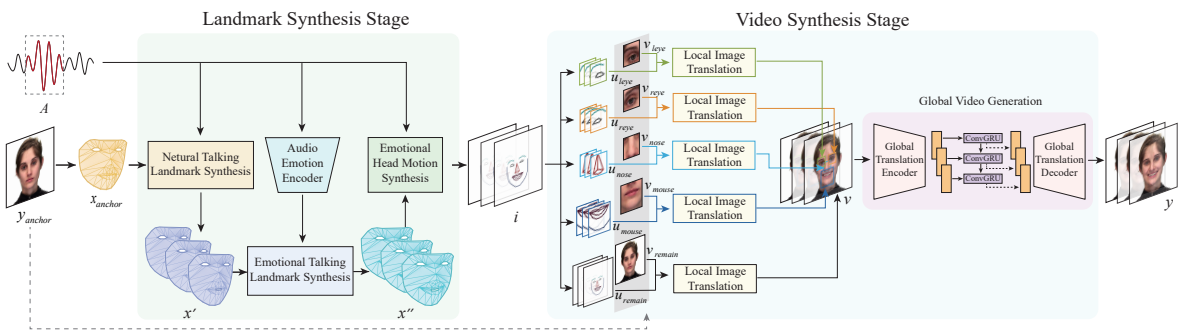


Figure 2. Illustration of our two-stage pipeline for audio-driven emotional talking head synthesis. The inputs are a speech clip $A$ and a facial image $y_{anchor}$. The *neutral talking landmark synthesis module* takes extracted facial landmarks $x_{anchor}$ and audio feature as input, and generates a talking landmark sequence $x'$. The *audio emotion encoder* and *emotion talking landmark synthesis* modules extract emotion features from the speech clip, and generate emotional talking landmark sequence $x''$. The *emotional head motion synthesis* module applies transformations and generates landmark sequence $x$, which are projected to obtain facial landmark images $i$. The *video synthesis stage* decomposes $i$ and $y_{anchor}$ into facial component patches. *Local image translation* module first translates landmark patches to images patches, which are composed together to concatenate the facial images $v$. These are encoded by the *global translation encoder* and sent into a 3-layer *ConvGRU* to be further encoded with temporal dependencies. Finally, they are decoded into emotional talking head frames and used to generate the talking face video $y$.

terms of generating talking heads with high visual quality and temporal consistency.

**Main Results:** We present a new approach for speech-driven emotional talking head generation. Given a speech audio clip and a facial image, our method is able to produce a talking head in synchronization with the audio and expresses the emotion corresponding to the audio, e.g., happy, sad, angry, etc. We propose a two-stage pipeline consisting of a *landmark synthesis* stage and a *video synthesis* stage. The landmark synthesis stage computes the facial landmarks of the input character, which convey audio information, including word utterance and encompassed emotions. These landmarks are used to guide video synthesis in the second stage. Our formulation is inspired by the studies in anatomy related to muscle relaxation and contraction for facial movements and expressions [54], we connect landmark points with edges and represent landmarks via edge features, including edge lengths and dihedral angles. The video synthesis stage generates a high-quality talking head video from facial landmarks using local and global computations. Specifically, we first translate component-level features, which helps preserve local facial details. Next, we use a global video translator with a 3-layer *ConvGRU* [2], which improves global face quality and enhances the temporal coherence. The novel components of our work include:

- We propose a general and automatic method that generates talking heads conveying the emotion of the speech audio.
- We use geometry-aware features that accurately control facial landmarks to convey expressive emotions.
- We transform landmark sequences to realistic talking heads in using local and global methods, that result in higher visual quality and temporal coherence. We also propose a new loss function to measure facial component similarity.

We validate our results on two emotional speech video datasets: RAVDESS [35] and CREMA-D [5]. In practice, our method out-performs state-of-the-art methods based on qualitative and quantitative metrics. We also highlight the perceptual benefits based on a user study.

## 2. Related Work

We briefly summarize recent works from three research areas: audio-based talking head generation, facial expression synthesis, and conditional face generation.

### 2.1. Audio-based Talking Head Generation

Previous research has proposed methods for generating talking heads from audio. [18] synthesizes talking heads through motion transfer from a reference motion model and then aligns lip motions with the audio clip. Without referencing a motion model, some methods directly use audio to drive 3D talking head. [26] maps audio features to 3D vertex positions and disentangles an emotion vector from the input audio. At run time, the emotion vector is chosen from the database to drive an emotional talking face. [52] first maps the input audio to a phonemic transcript and then translates the phoneme to corresponding face model parameters. [41, 55] use 3D blender-shapes, which map audio features to blender-shape parameters directly. Similarly, [13] uses the DeepSpeech network [21] as an intermediate representation of audio signal and then regresses parameters of

a FLAME head model [33]. [16, 51] propose methods that use audio to control the lower half of the face or lip motion. After that, the textured lower half of the face is reattached to a talking head.

While 3D model-based methods require post-processing for texture mapping, image-based methods can directly generate textured talking heads. [34, 12, 7] focus on generating talking lips and reattaching them to facial images. [4, 48, 65, 8, 57] directly extract visual features of the entire face from audio and combine them with the identity features. For example, [57] develops a generative adversarial network (GAN) based solution. It contains a generator that encodes identity, lip motion, and expression features to the latent space to generate a plausible talking face. It also has three discriminators to enhance video quality, identity consistency, and spatial consistency. In addition to generating a plausible talking face, [19, 60, 6, 66] propose methods that can control the motion of the generated talking head. [6] infers the head motion pattern from a clip of input video, while [66] learns speaker-aware head motion patterns from input audio features. Although current methods can generate talking heads with plausible expressions and head motions, they cannot express the emotion state.

### 2.2. Facial Expression Synthesis

Facial expression synthesis has been actively studied in computer graphics and computer vision. Traditional data-driven methods synthesize facial expressions by reusing similar image patches [38] or by retrieving similar expression patterns [28, 32] from a large database. Other methods synthesize facial expressions by image warping [62, 61, 17, 63]. For example, [61] adopts an expression flow map to perform image warping. [63] proposes a Flow Variational Autoencoder that projects image differences to a flow map.

3D deformation-based methods [3, 53, 30, 56] are widely used for facial expression synthesis. [53] achieves real-time expression transfer by mapping facial deformation of expressions to a low-dimensional expression space. [30] synthesizes high-quality facial expressions by deforming a 3D facial model fitted by a 3D Morphable Model (3DMM). Their methods not only enable the synthesis of six basic expressions, but also support emotion detailed VA (valence and arousal) control. [56] constructs SliderGAN which can manipulate facial expressions through a continuous deformation space to generate more diverse facial expressions than previous methods.

Many works use a generative model to synthesize facial expressions. [45, 10] disentangle the latent space, enabling facial expression synthesis by manipulating latent vectors. Other methods adopt conditional generative networks but differ in terms of the underlying conditions and features. [15, 64, 31] use one-hot labels as conditions to generate corresponding emotions. [50, 42] use action units (AU) as condition labels, which can generate more diverse expressions. However, the AU labels are hard to acquire. [47, 44, 49, 59] use landmarks as condition labels. Inspired by these works, we decide to use facial landmarks as condition labels for facial expression synthesis, since landmarks not only contain information about the facial geometry, but also enable the generation of more continuous and diverse facial expressions.

### 2.3. Conditional Face Generation

Conditional GANs [37] generate images that are conditioned by given attributes. Based on this architecture, many methods [24, 58, 27] are able to generate high-quality facial images with some conditional attributes. To address the lack of supervised data of some instances that have corresponding attributes , [67, 23] propose self-consistent constraints that enable domain transfer of different attributes. Further, [27] introduces StyleGAN, which generates high-quality facial images by separating high-level facial attributes in the semantic space. Some works [20, 9] introduced component-level face generation. These methods first generate facial components with separate conditional labels and then fuse the component-level features to generate the integrated face. Inspired by these methods, we first generate local facial features and then combine them to generate an integrated face, enabling local facial component control and high-quality facial image generation.

## 3. Our Talking Head Synthesis Algorithm

| Symbol | Description |
|---|---|
| $A$ | input speech audio |
| $y_{anchor}$ | input neutral face image |
| $x_{anchor}$ | 3D landmarks extracted from $y_{anchor}$ |
| $x'$ | neutral talking landmark sequence |
| $x''$ | emotional talking landmark sequence |
| $x$ | emotional talking landmark sequence with head motion |
| $i$ | facial landmark images rasterized from $x$ |
| $u_{t_x}$ | landmark patch of component at time $t$ |
| | $x \in \mathcal{C} := \{leye, reye, mouth, nose, remaining\}$ |
| $v_{a_x}$ | original neutral image patch of component $x \in \mathcal{C}$ |
| $v_{t_x}$ | translated local image patch of component $x \in \mathcal{C}$ |
| $a$ | audio feature sequence encoded by AutoVC [43] |
| $h$ | LSTM hidden vector sequence |
| $\Delta x_t$ | predicted landmark relative to $x_{anchor}$ at time $t$ |
| $f$ | LogMel Spectrogram feature |
| $\hat{x}'$ | landmark sequence $x'$ in its edge-based representation |
| $s$ | emotion feature vector from *audio emotion encoder* |
| $\tilde{f}$ | the concatenated feature sequence of $\hat{x}'$ and $s$ |

Table 1. Symbols and notation used in the paper

In this section, we present our overall algorithm. Our method (Fig. 2) takes a single neutral face image $y_{anchor}$ (e.g., $256 \times 256$ resolution) and a speech audio clip $A$ as

inputs. We generate a talking head sequence $y$ that is a 2D image sequence conveying basic emotions, such as happy, sad, angry, disgusted, fear, etc. We ensure that $y$ conveys an emotion corresponding to audio $A$ and preserves the identity of the input image $y_{anchor}$. To achieve this goal, we propose a two-stage pipeline that utilizes a facial landmark sequence as an intermediate representation of a face. These landmarks are concise and capable of expressing lip dynamics, facial expressions and head motions simultaneously.

The *landmark synthesis* stage generates landmark sequences corresponding to lip motion, facial expression, and head motion. We first synthesize neutral talking landmark sequences $x'$ by predicting the deformation of landmarks from the input landmarks $x_{anchor}$ driven by the audio $A$ (Section 3.1). Next,an emotional talking landmark synthesis module deforms $x'$ to a landmark sequence $x''$ that exhibits facial expressions conditioned on the emotion features extracted from the input audio $A$ (Section 3.2). At the end of this stage, an emotional head motion synthesis module translates and rotates the expressive landmark sequence $x''$ into a talking landmark sequence $x$ that moves in accordance with the input audio $A$ (Section 3.3).

For the *video synthesis* stage, we observe that the positions of the generated talking head landmarks $x$ tend to change frequently. Thus, facial details will be compromised if we directly translate the position and facial expression of landmarks to facial images. Therefore, to enhance the translation quality in terms of facial local details and temporal consistency, we adopt a hierarchical network to first translate the facial details at a component level, and then translate global information including temporal information(Section 3.4).

### 3.1. Neutral Talking Landmark Synthesis

This module is responsible for driving the static neutral facial landmark $x_{anchor} \in \mathbb{R}^{68\times3}$ , extracted by Openface [1], to a dynamic landmark sequence $x' \in \mathbb{R}^{T\times68\times3}$ , which synchronizes with the audio input $A$. Following [66], we use an LSTM network $lstm_{neu}$ to model the temporal dependency between audio features and landmark movements. Specifically, we first extract the audio content feature sequence $a$ from $A$ using the content encoder of AutoVC [43]. Then, $lstm_{neu}$ takes $a$ as input and outputs a hidden feature sequence $h$. Finally, an MLP (Multi-Layer Perceptron) transforms $h$ and $x_{anchor}$ into landmark displacements $\Delta x_t$ during each frame. For each frame, the displaced landmarks are calculated by $x'_t = x_{anchor} + \Delta x_t, t = 1, ..., T$. The output of this module is a landmark sequence $x'$ that utters sentence $A$.

### 3.2. Emotional Talking Landmark Synthesis

Using a landmark sequence with accurate lip motions and a neutral facial expression as input, we want to modify these landmarks so that they can convey the emotions such as happy, sad, angry, fearful, disgusted, etc. from input audio. To achieve this goal, we propose an end-to-end network comprised of an *emotion encoder* $E$ and an *emotional landmark generator* $G$ to analyze the emotion of speech audio and stylize the neutral landmark, respectively, as illustrated in Fig. 3.

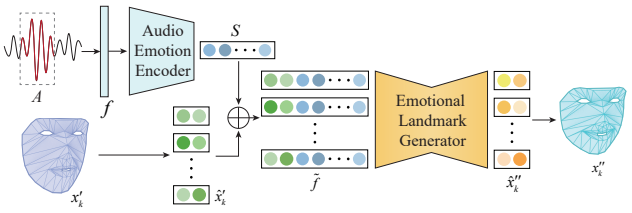


Figure 3. Our end-to-end network used to generate emotional talking landmarks. The network takes speech audio $A$ and neutral talking landmarks $x_k$ as input. First, the audio emotion encoder extracts emotion vector $s$ from speech audio feature $f$, which is then concatenated edge-wise with landmark feature $\hat{x}$. The concatenated feature $\hat{f}$ is then fed into the *emotional landmark generator*, to generate edge feature $\hat{x}'$ of emotional talking landmarks. Finally, the emotional talking landmarks $x'$ are recovered from their edge representation $\hat{x}'$.

Here we assume that the speech audio $A$ only contains a single emotion. To encode the emotion feature from $A$, we first extract its LogMel Spectrogram feature with Librosa [36]. We set the input with windows of size 2048 and a hop length of 512. Then the frequency domain is evenly separated into 128 Mel bands. Therefore, for each input audio, we obtain its LogMel Spectrogram feature $f \in \mathbb{R}^{128\times n_{mel}}$, where $n_{mel}$ is the number of LogMel Spectrogram features. With this base feature, we use a 6-layer 2D CNN encoder $s = E(f)$, where $s \in [0, 1]^c$ is the emotion feature vector and $c$ is the number of discrete emotion classes. Our method further enables emotional talking face editing. Results will be shown in the supplementary material together with the network architecture.

The extracted emotion feature vector stylizes the neutral facial landmark to express the same emotion as the speech audio. Conditioned on it, we are generating a emotional landmark sequence with a conditional generator $G$.

However, representing the facial landmarks during stylization remains a challenge. Turning the landmarks into an image results in rasterization artifacts, while representing the landmarks through point position produces unreasonable, non-smooth results due to lack of constraints. To address this, we propose connecting adjacent landmarks to form edges, which provide strong spatial constraints. Specifically, we use Delaunay triangulation to triangulate a template landmark configuration. Then the obtained connectivity is also applied to other landmark configurations to ensure topological consistency. Landmarks $x'_k$ can be rep-

resented as a set of edge lengths and their corresponding dihedral angles:

$$\hat{x}'_k = \left\{ L_k,\, \Theta_k |\, L_k \in \mathbb{R}_+^e,\, \Theta_k \in [-\pi, \pi)^e \right\}, \quad (1)$$

where $L_k \in \mathbb{R}_+^e$ contains all $e$ edge lengths, and $\Theta_k \in [-\pi, \pi)^e$ includes the dihedral angles of all edges.

Inspired by [11], we concatenate the landmark feature $\hat{x}'_k$ and emotion feature vector $s$ edge-wise and feed the combined feature $\tilde{f}_k \in \mathbb{R}^{e\times(2+c)}$ to the generator network $G$. The generator outputs $x''_k \in \mathbb{R}^{e\times 2}$, corresponding to the edge length and dihedral angles of the stylized landmarks. $G$ is trained in an adversarial manner, where two discriminators $D_{cls}$ and $D_{src}$ discriminate between the classes of emotion conveyed by the landmark and whether the generated landmarks are real or fake, respectively. Network architectures corresponding to $G$, $D_{src}$, and $D_{cls}$ are detailed in the supplementary material.

To train the end-to-end network, we adopt losses as follows. First, to have the emotion encoder $E$ classify the emotion correctly, we use an emotion classification loss:

$$L_{emo} = -\mathbb{E}_f[E(f)\log\hat{s} + (1 - E(f))(1 - \hat{s})], \quad (2)$$

where $\hat{s}$ is the ground truth emotion label of the input speech audio $A$. To have the generator network $G$ generate plausible emotional landmarks, we adopt the adversarial loss:

$$L_{adv} = \mathbb{E}_{x'_k}[\log D_{src}(x'_k)] + \mathbb{E}_{x'_k,s}[\log((1 - D_{src}(G(x'_k, s))]. \quad (3)$$

To ensure generated landmarks convey the target emotion, we introduce a loss term to constrain the style of the generated landmarks:

$$L_{cls} = \mathbb{E}_{x'_k,s}[\|\hat{s} - D_{cls}(G(x'_k, s))\|_2^2]. \quad (4)$$

In addition, to preserve the content, an extra reconstruction constraint is needed. Due to lack of paired stylized data, we adopt cycle consistency loss, as in [67]:

$$L_{rec} = \mathbb{E}_{x'_k,s,s'}[\|x'_k - G(G(x'_k, s), s')\|_2^2], \quad (5)$$

where $s'$ is the emotion label of landmarks $x'_k$. The overall loss is given as:

$$L = L_{emo} + \lambda_{adv}L_{adv} + \lambda_{cls}L_{cls} + \lambda_{rec}L_{rec}. \quad (6)$$

In our implementation, $\lambda_{adv}$, $\lambda_{cls}$ and $\lambda_{rec}$ are set to 1, 10, and 10, respectively.

### 3.3. Emotional Head Motion Synthesis

After generating the talking landmark sequence $x''$, we now synthesize head motions to match the landmarks to improve the naturalness of the generated videos. Assuming that each speech audio is a single sentence clip, we first extract the onset strengths $o$ of the speech audio $A$ over time

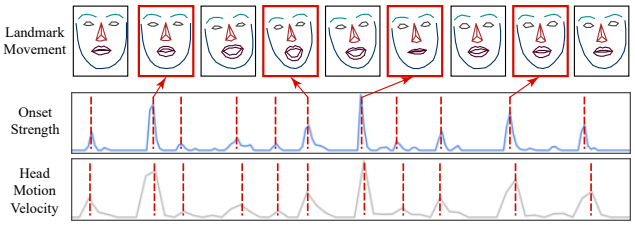


Figure 4. Illustration of the synthesized head motion pattern. The top row is the generated landmark sequence with head motion. Red boxes highlight the talking landmarks with higher velocity. The second row is the onset strength extracted by Librosa [36]. The third row is the head motion velocity computed as the sum of the absolute of row, yaw, pitch angle velocity.

using the Librosa library [36] and only retain those higher than a threshold $o_{thres}$. Then we randomize a moving direction of the talking head and set the head motion velocity synchronized with the onset strengths. As shown in Fig. 4, higher onset strengths of speech audio correspond to larger motion velocities. After head motion synthesis, the generated talking landmarks $x$ are projected to the image plane using the camera matrix, to obtain a talking landmark image sequence $i$ with head motion.

### 3.4. Hierarchical Landmark Sequence Translation

The second stage of the pipeline (Fig.2) is to translate the talking head landmark sequence. The problem becomes more challenging with the added head motion. We propose a hierarchical translation network with ConvGRU (Gated Recurrent Unit) [2] cells. It translates landmark sequences in a local-to-global manner and ensures the time consistency of the generated image sequences.

As shown in Fig. 2, the translation network comprises two parts. In the first part, we conduct local image translation. We decompose the facial landmark image $i_t$ and input neutral face image $y_{anchor}$ into each of five components: left-eye, right-eye, nose, mouth and the remaining parts of the face. For each component, we train an individual *component translation network* to translate landmark patches to images patches. For example, for the left-eye, the procedure of translating landmarks $u_{t_{leye}}$ to a left eye image patch is formulated as:

$$v_{t_{leye}} = T_{leye}(u_{t_{leye}}, v_{a_{leye}}), \quad (7)$$

where $T_{leye}$ is a GAN-based translation network of the left-eye component and $v_{t_{leye}}$ is the translated facial patch image with the same identity feature of $v_{a_{leye}}$ and the same pose as $u_{t_{leye}}$.

In the second part, we conduct global image translation. We place the translated nose, mouth, left-eye, and right-eye patches on the remaining patch of their respective positions of the original landmark. Although the combined facial image may have an inconsistent optical flow and artifacts on the patch borders, the translated image sequence should


Figure 5. Results of our algorithm on RAVDESS dataset corresponding to four different emotions (the bottom four rows), compared with neutral faces (the top row). Each row displays selected frames of the generated video under a specific emotion (labeled on the left). Our results demonstrate vivid expressions and plausible head motions.

| Method | RAVDESS | | | | CREMA-D | | | |
|---|---|---|---|---|---|---|---|---|
| | CPBD↑ | FID↓ | CSIM↑ | LMD↓ | CPBD↑ | FID↓ | CSIM↑ | LMD↓ |
| Vougioukas et al.[57] | 0.243 | 265.86 | 0.165 | 0.901 | 0.176 | 135.20 | 0.214 | 0.778 |
| Chen et al.[8] | 0.042 | 66.78 | 0.265 | 0.989 | 0.073 | 103.06 | 0.458 | 0.674 |
| Zhou et al.[66] | 0.325 | 77.05 | 0.457 | 0.713 | 0.395 | 89.025 | 0.761 | 0.658 |
| Ours | **0.336** | **14.62** | **0.516** | **0.705** | **0.442** | **22.25** | **0.773** | **0.326** |

Table 2. Quantitative evaluation results of different methods on RAVDESS and CREMA-D datasets (best results in bold). In all items, we achieve the best results.

be free of artifacts between patches and temporally coherent. To achieve these goals, as shown in Fig. 2, we use a global translation network to generate an integrated face with time consistency. Specifically, a 3-layer encoder network *Enc* extracts image features from the combined facial image, and a 3-layer *ConvGRU* generates sequentially consistent image features. These image features are decoded by a 3-layer decoder network *Dec* to generate a realistic and time-consistent facial image sequence $y$. A discriminator $D_{global}$ is applied to train the global generator in an adversarial manner. The network architecture is detailed in the supplementary material.

We adopt several loss terms to train the network to generate realistic talking faces. Both local and global translation networks are trained in a supervised manner with ground truth landmark image pairs, so we use mean squared error (MSE) loss to constrain the image reconstruction process. Also, all translation networks are trained in an adversarial manner, so adversarial loss is employed. In addition, we use perceptual loss [25] to constrain the high-level image features. To preserve both local identity consistency between patches in the input neutral face image $y_{anchor}$ and the generated facial images $y$, we propose a local identity loss inspired by [46]. Specifically, we place the generated component image patch onto its averaged position of this component on a $224 \times 224$ all-zero tensor and resize the tensor to $112 \times 112$ with a max-pooling operation. The max-pooled tensor is fed into Arcface [14] to get a feature vector that represents image identity. The identity loss is defined as the MSE between the feature vector of $y_{anchor}$ and that of $y$. These operations ensure that the component is recognized by the Arcface and the loss is capable of back-propagating gradients through the network. With these loss terms, the translation network is able to synthesize high-quality time-consistent identity-preserved talking head image sequences.

## 4. Implementation and Results

**Datasets** We quantitatively evaluate our results on the RAVDESS [35] dataset and CREMA-D [5] dataset. The RAVDESS dataset contains 24 professional actors speaking and singing with eight general emotions: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. We only use the speaking videos and discard the singing ones. The CREMA-D dataset contains 91 speakers speaking 12

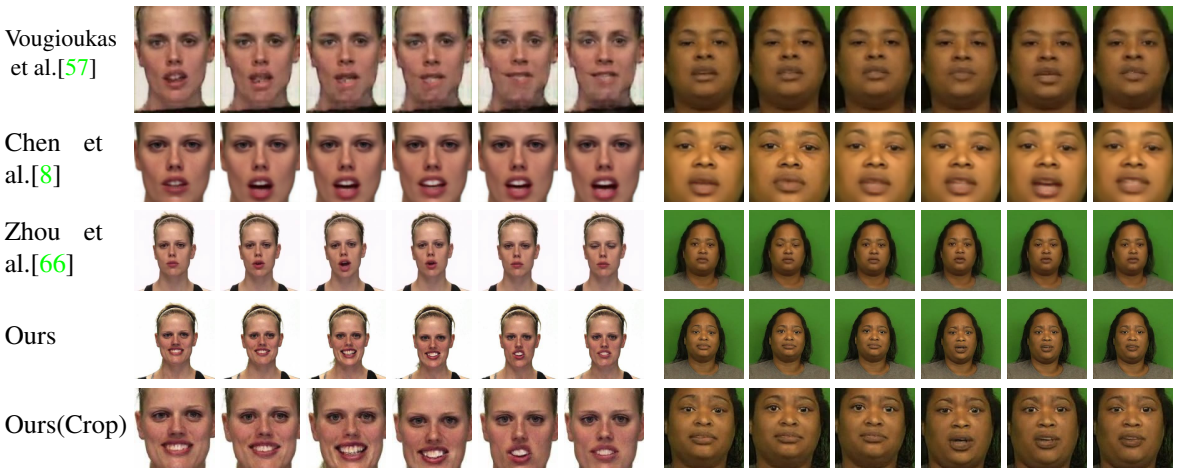


Figure 6. Comparison between our method and sother methods on talking head generation on RAVDESS (left) and CREMA-D (right) datasets. Our results are displayed in the bottom two lines where the last line is the crop of the second-last line.

| Method | RAVDESS | | | | CREMA-D | | | |
|---|---|---|---|---|---|---|---|---|
| | CPBD↑ | FID↓ | CSIM↑ | LMD↓ | CPBD↑ | FID↓ | CSIM↑ | LMD↓ |
| Full Model | **0.336** | **14.621** | **0.516** | **0.705** | **0.442** | **22.252** | **0.773** | **0.326** |
| w/o local translation | 0.312 | 39.059 | 0.463 | 0.736 | 0.437 | 58.764 | 0.567 | 0.387 |
| w/o identity loss | 0.319 | 54.052 | 0.478 | 0.775 | 0.427 | 60.631 | 0.496 | 0.689 |
| w/o convGRU | 0.274 | 52.527 | 0.491 | 0.720 | 0.392 | 39.083 | 0.555 | 0.369 |

Table 3. Ablation study results. We remove one component from the model in each set of experiment and compare the performance with the rest modules. As is shown in the table, performance is downgraded if any of the modules is removed, thus we prove that each module plays an essential role in our pipeline.


Figure 7. Ablation Studies: We show results with and w/o local identity loss. The identity information is preserved with local identity loss. The last two rows demonstrate the importance of local translation network, which preserves details as eyebrow shape.

sentences in six emotions: angry, disgusted, fearful, happy, neutral, and sad. For data preparation, we sample each video at 30FPS, crop the video by human face center, and resize the cropped image to $256 \times 256$. After sampling, the RAVDESS and CREMA-D datasets contain 159,222 and 570,204 video frames, respectively. We split each dataset into training and test sets. The subject IDs of the test set are illustrated in Table 4.

**Implementation Details.** The experiments were carried out on a computer with an Intel i7-6850 CPU, 128GB RAM, and a GTX 1080Ti GPU. The networks in our method are implemented in the PyTorch framework. All networks are

| Dataset | Test Subjects |
|---|---|
| RAVDESS | 3, 4, 5, 6 |
| CREMA-D | 15, 20, 21, 30, 33, 52, 62, 81, 82, 89 |

Table 4. The subject IDs of the test sets of the RAVDESS and CREMA-D datasets.

optimized by Adam optimizer [29]. For $lstm_{neu}$ used in *neutral talking landmark synthesis* module, the hidden size of the LSTM cell is 256. The learning rate of *neutral talking landmark synthesis*, *Audio Emotion Encoder*, *emotional landmark generator*, *local image translation* and *global video generation* module are set to $1 \times 10^{-4}$, $1 \times 10^{-4}$, $1 \times 10^{-4}$, $2 \times 10^{-4}$, $2 \times 10^{-4}$ respectively. Other network details will be illustrate in supplementary material.

### 4.1. Results and Comparisons

Figure 5 demonstrates the speech-driven talking head generation results of our method applied to the RAVDESS and CREMA-D datasets. More in-the-wild results are shown in the supplementary material. We demonstrate a character uttering the same sentence with five different emotions: neutral, happy, sad, angry, and disgusted. The talking heads generated by our method clearly display expressive emotions and emotion-aware head motions. The neutral talking head barely has any head motion, happy

emotion and sad emotion have slight head motion, and angry emotion and disgusted emotion have obvious head motion. The results show that our method is able to generate realistic talking heads conveying the same emotion as the input audio. Further, our model is able to realize emotional talking face editing. Results of this functionality will be also provided in the supplementary material.

Figure 6 compares our method with other state-of-the-art methods [57, 8, 66]. We use an expressive speech audio and a neutral face image as inputs. The generated talking heads of all methods are placed in rows. Vougioukas et al. [57] generate images of size $128 \times 96$, which suffer from poor identity preservation. Chen et al. [8] generate images of size $128 \times 128$. Their method preserves identity information, but the generated mouth area suffers from poor visual quality. Zhou et al. [66] produce talking head images with head motion. All these methods do not generate expressive emotions. We compare our methods with other state-of-the-art talking head generation methods quantitatively in three aspects: video quality, identity consistency, and word pronouncing accuracy. To evaluate the generated video frame quality, we use cumulative probability blur detection (CPBD) [39] to evaluate the frame sharpness and Fréchet Inception Distance (FID) [22] to measure the quality and diversity of the generated images. We calculate the FID value between the generated image and the ground truth images of the same identity and the same emotion. For identity preservation, we follow [6] and use a CSIM criterion, which calculates the cosine similarity between the feature vectors extracted from two candidate facial images. To evaluate word pronouncing accuracy, we align the generated lip landmarks with ground truth lip landmarks and then compute the landmark distance (LMD) as proposed in [7]. The evaluation metrics on the RAVDESS and CREMA-D datasets are shown in Table 2.

### 4.2. Ablation Study

We perform the following ablation studies. First, to validate the effectiveness of the local and global structures in image translation, we design a comparison group that replaces the component-level image translation network with a global translation network, which is denote as *'w/o local'*. Second, to demonstrate the benefits of the identity loss in the image translation networks, we design a comparison group that removes this loss in both local translation networks, which is denote as *'w/o identity'*. Figure 7 provides some visual examples of images generated when these two modules are removed individually. Third, to evaluate the performance of ConvGRU in global video synthesize, we design a comparison group that removes the ConvGRU network, which is denote as *'w/o local convGRU'* (see Table 3).

### 4.3. User Study

We conduct an online user study for perceptual evaluation. Users were provided with 8 video sets, 4 for RAVDESS dataset and 4 for CREMA-D dataset. In each video set, users were provided with a single neutral face image and four generated talking head videos with audio generated by Vougioukas et al. [57], Chen et al. [8], Zhou et al. [66] and our method, respectively. Users were asked to score each generated video from 0 to 10 in two aspects: visual quality and emotional expressiveness. We present the results in Table 8 that demonstrate the benefits of our approach.

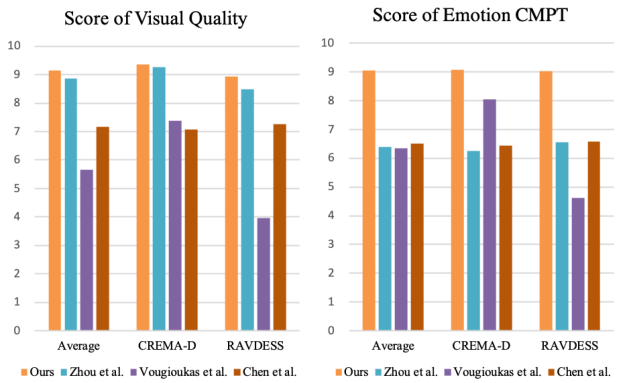


Figure 8. User study results (higher is better) on the visual quality and CMPT, which measures compatibility between audio and visual emotions. Overall, our approach is more promising.

## 5. Conclusion and Limitations

We present a new approach to generate geometrically-aware, emotionally-expressive, high-quality talking heads driven by speech audio. Our approach computes facial landmarks for the character and use them to control the video synthesis. We perform local and global computations to improve the visual quality and maintain temporal coherence. We have evaluated the performance on RAVDESS and CREMA-D datasets and highlight the benefits over prior methods. Our approach has some limitations. The emotion expression patterns of different subjects are similar and it does not take into account personalized characteristics and all arousals. Ultimately, we would like to generate continuous pattern of varying emotions based on the valence-arousal dimensions. It would be useful to incorporate more ideas from anatomy to generate a larger set of face expressions.

## References

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2), 2016. 4

[2] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015. 2, 5

[3] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library, 2003. 3

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997. 3

[5] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377—390, 2014. 2, 6

[6] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *arXiv preprint arXiv:2007.08547*, 2020. 1, 3, 8

[7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 3, 8

[8] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 1, 3, 6, 7, 8

[9] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: deep generation of face images from sketches. *ACM Transactions on Graphics (TOG)*, 39(4):72–1, 2020. 3

[10] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014. 3

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 5

[12] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 3

[13] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 2

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6

[15] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. *arXiv preprint arXiv:1709.03842*, 2017. 3

[16] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015. 3

[17] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4217–4224, 2014. 3

[18] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer graphics forum*, volume 34, pages 193–204. Wiley Online Library, 2015. 2

[19] David Greenwood, Iain Matthews, and Stephen Laycock. Joint learning of facial expression and head pose from speech. Interspeech, 2018. 3

[20] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019. 3

[21] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 2

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 8

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 3

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[26] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1, 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[28] Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M Seitz. Being john malkovich. In *European Conference on Computer Vision*, pages 341–353. Springer, 2010. 3

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer ence*, 2014. 7

[30] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, pages 1–30, 2020. 3

[31] Ying-Hsiu Lai and Shang-Hong Lai. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 263–270. IEEE, 2018. 3

[32] Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. A data-driven approach for facial expression synthesis in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 57–64. IEEE, 2012. 3

[33] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3

[34] Kang Liu and Joern Ostermann. Realistic facial expression synthesis for an image-based talking head. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011. 3

[35] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Apr. 2018. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. 2, 6

[36] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015. 4, 5

[37] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[38] Umar Mohammed, Simon JD Prince, and Jan Kautz. Visiolization: generating novel facial images. *ACM Transactions on Graphics (TOG)*, 28(3):1–8, 2009. 3

[39] Niranjan D Narvekar and Lina 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*, pages 87–91. IEEE, 2009. 8

[40] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 80–88, 2017. 1

[41] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. End-to-end learning for 3d facial animation from speech. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 361–365, 2018. 2

[42] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 3

[43] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. AutoVC: Zero-shot voice style transfer with only autoencoder loss. volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 3, 4

[44] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018. 3

[45] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439, 2014. 3

[46] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020. 6

[47] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635, 2018. 3

[48] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 1, 3

[49] Kritaphat Songsri-in and Stefanos Zafeiriou. Face video generation from a single image and landmarks. *arXiv preprint arXiv:1904.11521*, 2019. 3

[50] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pages 421–440, 2008. 3

[51] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 3

[52] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017. 2

[53] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[54] Ying-Li Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression analysis. In *Handbook of face recognition*, pages 247–275. Springer, 2005. 2

[55] Panagiotis Tzirakis, Athanasios Papaioannou, Alexander Lattas, Michail Tarasiou, Björn Schuller, and Stefanos Zafeiriou. Synthesising 3d facial motion from" in-the-wild" speech. *arXiv preprint arXiv:1904.07002*, 2019. 2

[56] Evangelos Ververas and Stefanos Zafeiriou. Slidergan: Synthesizing expressive face images by sliding 3d blendshape parameters. *International Journal of Computer Vision*, pages 1–22, 2020. 3

[57] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019. 1, 3, 6, 7, 8

[58] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[59] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7083–7092, 2018. 3

[60] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 1, 3

[61] Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas. Facial expression editing in video using a temporally-smooth factorization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 861–868. IEEE, 2012. 3

[62] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3d-aware face component transfer. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011. 3

[63] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016. 3

[64] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3359–3368, 2018. 3

[65] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019. 1, 3

[66] Yang Zhou, DIngzeyu Li, Xintong Han, Evangelos Kalogerakis, Eli Shechtman, and Jose Echevarria. Makeittalk: Speaker-aware talking head animation. *arXiv preprint arXiv:2004.12992*, 2020. 1, 3, 4, 6, 7, 8

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3, 5