# Emotion-Controllable Generalized Talking Face Generation

**Sanjana Sinha**[1*], **Sandika Biswas**[1*], **Ravindra Yadav**[2†] and **Brojeshwar Bhowmick**[1]
[1]TCS Research, India
[2]IIT Kanpur, India
{sanjana.sinha, biswas.sandika, b.bhowmick}@tcs.com, ravin@iitk.ac.in

## Abstract

Despite the significant progress in recent years, very few of the AI-based talking face generation methods attempt to render natural emotions. Moreover, the scope of the methods is majorly limited to the characteristics of the training dataset, hence they fail to generalize to arbitrary unseen faces. In this paper, we propose a one-shot facial geometry-aware emotional talking face generation method that can generalize to arbitrary faces. We propose a graph convolutional neural network that uses speech content feature, along with an independent emotion input to generate emotion and speech-induced motion on facial geometry-aware landmark representation. This representation is further used in our optical flow-guided texture generation network for producing the texture. We propose a two-branch texture generation network, with motion and texture branches designed to consider the motion and texture content independently. Compared to the previous emotion talking face methods, our method can adapt to arbitrary faces captured in-the-wild by fine-tuning with only a single image of the target identity in neutral emotion.

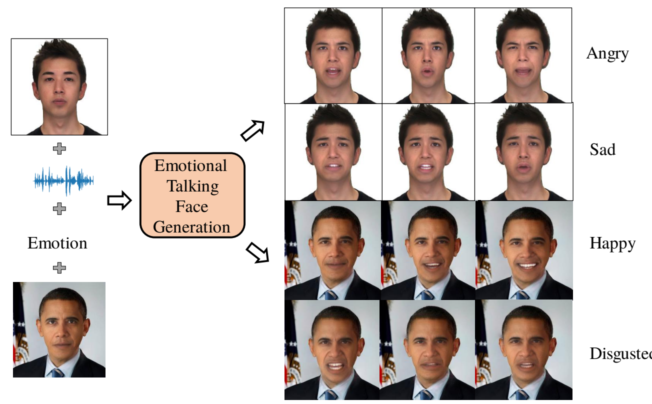


Figure 1: Results of our proposed emotional talking face generation method on arbitrary faces.

## 1 Introduction

Audio-driven realistic talking face generation is a widely studied research problem, with diverse applications in animation, virtual assistant, telepresence, gaming etc. Most of the existing methods [Chung *et al.*, 2017; Suwajanakorn *et al.*, 2017; Chen *et al.*, 2019; Das *et al.*, 2020; Zhou *et al.*, 2019; Sinha *et al.*, 2020; Chen *et al.*, 2020; Zhou *et al.*, 2020; Zhou *et al.*, 2021; Zhang *et al.*, 2021] mainly focus on generating realistic lip synchronization, identity preservation, eye blinks or head motion in the synthesized talking face video. Very few of these methods can render realistic facial emotions (Table 1), due to the limited availability of annotated emotional audio-visual datasets. Some earlier methods [Vougioukas *et al.*, 2019; Chen *et al.*, 2020] have tried to learn the facial emotions implicitly from the audio. However, these methods fail to control the facial emotion and often fail to produce realistic animation.

Recently, MEAD [Wang *et al.*, 2020] has proposed a method for emotional talking face generation with explicit emotion control and released the MEAD dataset [Wang *et al.*, 2020] containing well-defined emotions at varying intensities, and a wide variety of sentences. This method [Wang *et al.*, 2020] generates emotion only in the upper face (from external emotion control using one-hot emotion vector) and the lower part of the face is animated from audio independently, which results in inconsistent emotions over the face. A recent video editing method EVP [Ji *et al.*, 2021] focuses on generating consistent emotions over the entire face using a disentangled emotion latent feature learned from the audio. However, all these methods rely on intermediate global landmarks (or edge maps) to generate the texture directly with emotions. To generalize the texture deformation for any unknown face for a given emotion, it is important to learn the relationship between the facial geometry and the emotion-induced local deformations within the face. None of these methods consider learning this relationship, hence show a limited scope of generalization to an arbitrary unknown target face (Fig. 3, Row 3 & 4, refer to the caption for evaluation details). Moreover, MEAD[1] and EVP[2] train target-specific texture models.

*Equal contribution
†Former intern at TCS Research

[1] https://github.com/uniBruce/Mead
[2] https://github.com/jixinya/EVP

In this work, we propose a generalized one-shot learning-based emotional talking face generation method. Unlike the previous video-based method EVP (Table 1), for emotion rendering, we need only a single image of the target person, along with speech and an emotion vector as input. We want to achieve speech-independent emotion control so that the same audio can be animated using different emotions. We use features from a pre-trained automatic speech recognition model DeepSpeech [Hannun *et al.*, 2014] for disentangling emotion from speech content of audio. We first propose a graph neural network that encodes the desired emotion and speech content to render emotion and speech-induced motion on a geometry-aware graph representation of the facial landmarks. Unlike previous landmark-based talking face methods [Chen *et al.*, 2019; Zhou *et al.*, 2020; Chen *et al.*, 2020; Ji *et al.*, 2021; Zhang *et al.*, 2021], we construct a graph representation of facial landmarks using [Delaunay *et al.*, 1934] for capturing the spatial configuration of facial landmarks and their inter-dependencies during emotional speech. In the texture generation stage, we learn an emotion-guided optical flow map from the intermediate predicted landmarks to consider the facial structure and emotion-induced local deformations around the landmarks. Despite having high-quality, well-defined emotional speech videos, MEAD dataset has low variety in illumination, background, etc. We carefully design a two-branch texture generation network to disentangle the speech and emotion-induced motion from identity-related texture content. At inference time, we propose one-shot learning for adapting the texture generation model to the identity of the input target face. This helps in generalization while generating emotions for any arbitrary target face.

We demonstrate the generalization ability of our method by evaluating on different faces outside our training dataset MEAD (Fig.s 1, 4 and 5). To the best of our knowledge, this is the first work on emotional talking face generation that is generalized for any arbitrary face. Our contributions are summarized below:

- We propose a pipeline for facial geometry-aware one-shot emotional talking face generation from audio with independent emotion control.
- We propose a graph convolutional network for inducing speech and emotion on graph-representation of facial landmarks to preserve facial structure and geometry for emotion rendering.
- We propose an optical flow-guided texture generation network that renders emotional talking face animation from a single image of any arbitrary target face in neutral emotion.

## 2 Related Work

**Emotional Talking Face Generation.** Recent methods in audio-driven talking face generation are listed in (Table 1). Video-based methods that generate only the mouth in a driving video of target [Thies *et al.*, 2019; Song *et al.*, 2020; Prajwal *et al.*, 2020; Wen *et al.*, 2020] are capable of generating photo-realistic facial animation. However, since the facial texture (except the mouth) is copied from the input video frames, facial expressions and emotions in the upper part of the face cannot be manipulated using these methods. Our method uses a single image of the target for generating emotional talking faces without the need for a driving video.

| **Audio-driven Talking Face Methods** | **Input Image/Video** | **Arbitrary face** | **Emotion generation** |
|---|---|---|---|
| [Das *et al.*, 2020] | Image | ✓ | ✗ |
| MakeItTalk [Zhou *et al.*, 2020] | Image | ✓ | ✗ |
| [Zhang *et al.*, 2021] | Image | ✓ | ✗ |
| [Wang *et al.*, 2021] | Image | ✓ | ✗ |
| [Zhou *et al.*, 2021] | Image | ✓ | ✗ |
| [Thies *et al.*, 2019] | Video | ✓ | ✗ |
| [Song *et al.*, 2020] | Video | ✓ | ✗ |
| Wav2Lip [Prajwal *et al.*, 2020] | Video | ✓ | ✗ |
| [Wen *et al.*, 2020] | Video | ✓ | ✗ |
| [Vougioukas *et al.*, 2019]* | Image | ✗ | ✓ |
| [Chen *et al.*, 2020]* | Image | ✗ | ✓ |
| [Eskimez *et al.*, 2020] | Image | ✗ | ✓ |
| MEAD, [Wang *et al.*, 2020] | Image | ✗ | ✓ |
| EVP, [Ji *et al.*, 2021] | Video | ✗ | ✓ |
| **Ours** | **Image** | **✓** | **✓** |

Table 1: Recent talking face generation methods. The emotional talking face methods cannot generalize to arbitrary faces. (*) Emotion is not learned explicitly in these methods, derived implicitly from audio.

Some earlier methods [Vougioukas *et al.*, 2019; Chen *et al.*, 2020] render emotional talking face videos that learn the emotion implicitly from the audio. In contrast, we aim for an explicit control for generating consistent emotions in the talking face. Some recent methods MEAD, EVP, [Eskimez *et al.*, 2020] have proposed methods with external control on emotion in the talking face. EVP learns a disentangled emotion latent feature representation from speech input and tries to generate varying emotions by interpolating the emotion latent space. However, the latent emotion representation in EVP depends on the accuracy of the audio-emotion disentanglement; hence it is difficult to achieve completely independent control of emotion from speech. In contrast to the previous methods MEAD, EVP, our method manipulates emotions in the entire face using an emotion control input that is fully independent of the audio.

**Generalized Arbitrary-Subject Talking Face.** Talking face generation methods (Table 1) that can generalize to arbitrary faces are trained on large-scale audio-visual datasets such as Voxceleb [Chung *et al.*, 2018] having a wide diversity of faces, illumination and background. However these methods cannot render animation in different emotions. Existing emotional talking face generation methods trained on emotional audio-visual datasets CREMA-D [Cao *et al.*, 2014] and MEAD [Wang *et al.*, 2020] have limited scope of generalization owing to lower diversity of these datasets. Previous methods [Vougioukas *et al.*, 2019; Chen *et al.*, 2020; Eskimez *et al.*, 2020] which are trained on CREMA-D lack generalization to faces outside CREMA-D. Recently, MEAD and EVP have used a high quality emotional audio-visual dataset MEAD for training. However, they have trained target subject-specific texture generation models [1 2]; hence they cannot generalize to arbitrary identities. On the other hand, our method is capable of generalization to any unknown target subject.

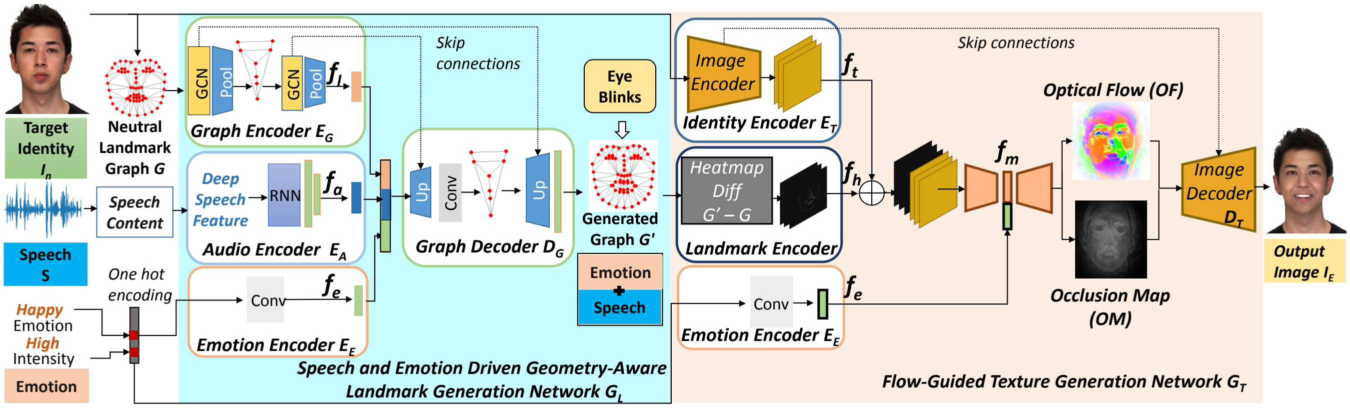


Figure 2: Our proposed method for arbitrary-face emotional talking face generation. The Geometry-Aware Landmark Generation Network $G_L$, encodes speech content of input speech $S$, neutral face landmark graph $\mathcal{G}$, target emotion $e$ (along with emotion intensity), and reconstructs landmark graph $\mathcal{G}'$ containing speech and emotion. For realism spontaneous, eye blinks are added to the landmarks in $\mathcal{G}'$. In the Texture Generation stage, the heatmap difference of the target identity's facial landmarks, encoded identity face, and encoded target emotion are used to generate emotion-induced optical flow and occlusion map, which are subsequently decoded to generate the speech and emotion-induced facial texture image of the target identity.

## 3 Methodology

Fig. 2 shows the detailed architecture of our network for generating emotion-controllable talking faces. For a given speech ($S$), an emotion input, and a single image of the target subject in neutral emotion ($I_n$), our method generates an animated face delivering the speech with desired emotion and intensity.

### 3.1 Speech and Emotion Driven Landmark Generation

We propose facial geometry-aware speech and emotion generation ($G_L$, Fig. 2) on facial landmarks using a graph neural network.

**Audio Encoder.** $E_A$ is a recurrent neural network which creates an emotion-invariant speech embedding feature $\mathbf{f_a} \in \mathbb{R}^d$ ($d = 128$) from speech audio input $S$. For each audio window of size $W$ corresponding to a video frame, features $\mathcal{A} = \{a_t \in \mathbb{R}^{W\times 29}\}$ are extracted from the output layer of a pre-trained DeepSpeech network (before applying Softmax). The output layer of Deepspeech represents log probabilities of 29 characters; hence the features are emotion-independent.

**Emotion Encoder.** $E_E$ encodes an emotion vector ($e$,$i$). $e$ denotes six types of emotions i.e. happy, angry, sad, surprise, fear and disgust, at two types of intensity levels $i$ (high or low) into a fixed feature representation $\mathbf{f_e} \in R^d$ ($d = 128$).

**Graph Encoder.** $E_G$ is a graph convolutional network that encodes the geometry of an ordered graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V} = \{v_i\}$ denotes the set of $L = 68$ facial landmark vertices, $\mathcal{E} = \{e_{ij}\}$ is the set of edges, computed using delaunay triangulation [Delaunay *et al.*, 1934] on facial landmarks, $A$ is the adjacency matrix of $\mathcal{G}$. $\mathbf{X} = [X_{ij}](X_{ij} \in \mathbb{R}^2)$ is a matrix of vertex feature vectors, i.e, coordinates of the $L = 68$ facial landmarks of a neutral image (face in neutral emotion and with closed lips). We apply spectral graph convolution [Kipf and Welling, 2016] with a modified propagation rule including learnable edge weights [Yan *et al.*, 2018] :

$$f_{k+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\boldsymbol{\omega}(A + I)\tilde{D}^{-\frac{1}{2}} f_k W_k), \tag{1}$$

where $I$ represents the identity matrix, $\tilde{D}^{ii} = \sum_j (A^{ij} + I^{ij})$, $\boldsymbol{\omega} = \{\omega^{ij}\}$ are learnable edge weights for determining the contribution of each edge in $\mathcal{G}$, $f_k$ is the output of the $k$th layer, ($f_0 = \mathbf{X}$), $W_k$ is a trainable weight matrix of the $k$th layer, $\sigma(\cdot)$ is the activation function. Since edges between landmark vertices of semantically connected regions of the face are more significant than the edges connecting two different facial regions, the learnable edge weight $\boldsymbol{\omega}$ signifies the contribution of the vertex's feature to its neighboring vertices. Unlike lip movements, emotion has an effect over the entire face and not only a specific region. Inspired by [Cai *et al.*, 2019] we apply a hierarchical "local-to-global" scheme for graph convolution to capture facial deformations. Graph pooling operation helps to aggregate feature level information in different facial regions, which helps local deformations caused by facial expressions. The face landmark graph structure is first divided into $K$ subsets of vertices, each representing a facial region, e.g., eye, nose, etc. Hierarchical graph convolution (GCN) and pooling is done (as shown in Fig. 2) to generate feature $\mathbf{f_l} \in R^d$ ($d = 128$) representing the entire graph.

**Graph Decoder.** $D_G$ reconstructs the output landmark graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}, A)$ from the concatenation of the feature vectors $\mathbf{f_a}, \mathbf{f_l}, \mathbf{f_e}$. It learns the mapping $f : (\mathbf{f_a}, \mathbf{f_l}, \mathbf{f_e}) \rightarrow \mathbf{X}'$, where $\mathbf{X}' = \mathbf{X} + \boldsymbol{\delta}$ represents the vertex positions of the reconstructed facial landmarks with generated displacements $\boldsymbol{\delta}$ induced by speech and emotion. $\hat{\mathbf{X}}$ are the ground landmarks. The losses for training $G_L$ are as follows:
*Landmark vertex distance loss*:

$$\mathcal{L}_{ver} = ||\hat{\mathbf{X}} - (\mathbf{X} + \delta)||_2^2. \tag{2}$$

*Adversarial loss*: A graph discriminator $D_L$ evaluates the realism of the facial expression in a generated graph $\mathcal{G}'$. $G_L$

and $D_L$ are trained using the LSGAN loss function [Mao *et al.*, 2017]:

$$\mathcal{L}_{gan}(D_L) = \big(\mathbb{E}[(D_L(\hat{\mathcal{G}}, e) - 1)^2] + \mathbb{E}[D_L(\mathcal{G}', e)^2]\big)/2$$
$$\mathcal{L}_{gan}(G_L) = \mathbb{E}[(D_L(\mathcal{G}', e) - 1)^2]/2, \tag{3}$$

where $\mathcal{G}'$ is the generated graph and $\hat{\mathcal{G}}$ is the ground truth graph. The combined loss function for training the landmark generation networks are:

$$\mathcal{L}_{lm} = \lambda_{ver} L_{ver} + \lambda_{gan} L_{gan}, \tag{4}$$

where the loss hyperparameters $\lambda_{ver} = 1$ and $\lambda_{gan} = 0.5$ are experimentally set using validation data.

### 3.2 Texture Generation

Fig. 2 shows our proposed Texture Generation network $G_T$ that generates an emotional talking face from a single image $I_n$ of the target identity subject in neutral expression and predicted landmarks $\mathcal{G}'$ from $G_L$. For realism, spontaneous eye blink displacements [Das *et al.*, 2020] are added to the landmark vertices of $\mathcal{G}'$ before texture generation.

**Image Encoder.** $E_T$ encodes the target identity image $I_n$ into identity feature $f_t$, that is used for predicting the optical flow and occlusion map in the subsequent stage. The emotion feature $f_e$ is generated in a similar manner as presented in the landmark generation network $G_L$.

**Heatmap Difference.** A heatmap is generated by creating a Gaussian distribution centered at each of the vertices of the landmark graph. The heatmap representation captures the structural information of the face in the image space and the local deformations around the landmark vertices. The difference $f_h$ between heatmaps of input graph $\mathcal{G}$ and generated graph $\mathcal{G}'$ models the motion of facial landmarks.

**Optical Flow and Occlusion Map Prediction.** Optical flow ($OF$) captures the local deformations over different regions of the face due to speech and emotion induced motions. Whereas, occlusion map ($OM$) denotes the regions which need to be newly generated (e.g., inside the mouth region for happy emotion) in the final texture. $OF$ and $OM$ are learned in an unsupervised manner (Eqn. 5) and no ground-truth optical flow or occlusion map are used for supervision. At an intermediate stage the network generates $OF$ and $OM$ from heatmap difference, target identity image conditioned on emotion condition. The heatmap difference ($f_h$) and the encoded target identity image feature ($f_t$) are concatenated channel-wise and passed through an encoder network to produce $f_m$. Further, to influence the facial motion by the necessary emotion, the encoded emotion feature $f_e$ is concatenated channel-wise with $f_m$ and decoded to produce the dense flow map ($OF$) and occlusion map ($OM$). Flow-guided texture generation from heatmap differences of facial landmarks helps to learn the relationship between the face geometry and emotion-related deformations within the face.

**Final Animation Generation.** The concatenated occlusion map and optical flow maps are given as input to the image decoder $D_T$, which produces the final output image ($I_E$) containing speech and emotion.

$$I_E = D_T(OF \oplus OM, f_t). \tag{5}$$

Skip connections are added between the layers of target identity encoder ($E_T$) and the decoder $D_T$. The losses used for training the network are as follows:
*Reconstruction loss* between predicted $I_E$ and GT image $\hat{I}$:

$$\mathcal{L}_{rec} = |I_E - \hat{I}|. \tag{6}$$

*Perceptual loss* between VGG16 features of $I_E$ and $\hat{I}$:

$$\mathcal{L}_{per} = |VGG16(I_e) - VGG16(\hat{I})|. \tag{7}$$

*Adversarial loss* with a frame discriminator $D$:

$$\mathcal{L}_{adv} = \min_{G} \max_{D} \mathbb{E}_{\hat{I}}[log(D(\hat{I}))] + \mathbb{E}_{I_E}[log(1 - D(I_E))]. \tag{8}$$

The total loss function for training $G_T$:

$$\mathcal{L}_{img} = \lambda_{rec} L_{rec} + \lambda_{per} L_{per} + \lambda_{adv} L_{adv}, \tag{9}$$

where the loss hyperparameters $\lambda_{rec}$, $\lambda_{per}$, $\lambda_{adv}$ are experimentally set to 1, 10, and 1 respectively.

## 4 Experiments and Training Details

### 4.1 Datasets

We use 3 emotional audio-visual datasets MEAD [Wang *et al.*, 2020], CREMA-D [Cao *et al.*, 2014], and RAVDESS [Livingstone and Russo, 2018] for our experiments. We have selected 24 subjects of diverse ethnicity from MEAD for the training of our proposed pipeline, and our method is evaluated on test splits of MEAD, CREMA-D, RAVDESS and also arbitrary unknown faces and speech.

### 4.2 Implementation Details

The Landmark Generation Network $G_L$ and Texture Generation Network $G_T$ are trained independently. The architectures of $G_L$ and $G_T$ are shown in Fig. 2. For training $G_L$ and $G_T$, the ground-truth landmarks are extracted (at $30fps$) using a combination of 3D landmarks from [Guo *et al.*, 2020] and face parsing [Yu *et al.*, 2018] for accurate mouth shapes. $G_T$ uses ground-truth landmarks during training, and predicted landmarks from $G_L$ during inference. We train both $G_L$ and $G_T$ using Pytorch on NVIDIA Quadro P5000 GPUs (16 GB) using Adam Optimizer, with a learning rate of $2e-4$. Training of $G_L$ takes around a day with batch size 256 (2GB GPU usage), and the training of $G_T$ takes around 7 days (batch size 4 on 16GB GPU).

**One-shot learning.** MEAD dataset contains a limited variety in illumination and background, which limits generalization to arbitrary target faces. By fine-tuning our texture generation network $G_T$ on a single image of any unseen target face *in neutral emotion*, we can generate emotional talking face generation for the target *in different emotions*. In order to adapt to the identity of the unknown target neutral, we only update the image encoder ($E_T$) and decoder ($D_T$) layer weights using the single image, while keeping the network weights for the rest of $G_T$ unchanged. One-shot learning helps bridge the color and illumination gap between the training and testing samples and adapt the generated texture to the identity of the target face while keeping the speech and emotion-induced motion intact.

| Dataset | Method | Texture Quality | | | | Landmark quality | | | | Emotion accuracy | Identity | Lip Sync |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR ↑ | SSIM ↑ | CPBD ↑ | FID ↓ | M-LD ↓ | M-LVD ↓ | F-LD ↓ | F-LVD ↓ | $Emo_{Acc}$ ↑ | CSIM ↑ | $Sync_{conf}$ ↑ |
| MEAD | MEAD [Wang *et al.*, 2020] | 28.61 | 0.68 | 0.29 | 22.52 | 2.52 | 2.28 | 3.16 | 2.01 | 76.00 | **0.86** | 1.83 |
| | EVP [Ji *et al.*, 2021] | 29.53 | 0.71 | 0.35 | **7.99** | 2.45 | 1.78 | 3.01 | 1.56 | 83.58 | 0.67 | 1.21 |
| | Ours | **30.06** | **0.77** | **0.37** | 35.41 | **2.18** | **0.77** | **1.24** | **0.50** | **85.48** | 0.79 | **3.05** |
| CREMA-D | [Vougioukas *et al.*, 2019] | 23.57 | 0.70 | 0.22 | 71.12 | 2.90 | **0.42** | 2.80 | **0.34** | 55.26 | 0.51 | 1.12 |
| | [Eskimez *et al.*, 2020] | 30.91 | 0.85 | 0.39 | 218.59 | 6.14 | 0.49 | 5.89 | 0.40 | 65.67 | **0.75** | **4.38** |
| | Ours | **31.07** | **0.90** | **0.46** | **68.45** | **2.41** | 0.69 | **1.35** | 0.46 | **75.02** | **0.75** | 3.53 |

Table 2: Quantitative comparison of our method with SOTA emotional talking face generation methods. [Eskimez *et al.*, 2020; Vougioukas *et al.*, 2019] have trained their models on CREMA-D dataset, while MEAD, EVP have trained on MEAD dataset. Our model is trained only on MEAD and evaluated on both MEAD and CREMA-D.

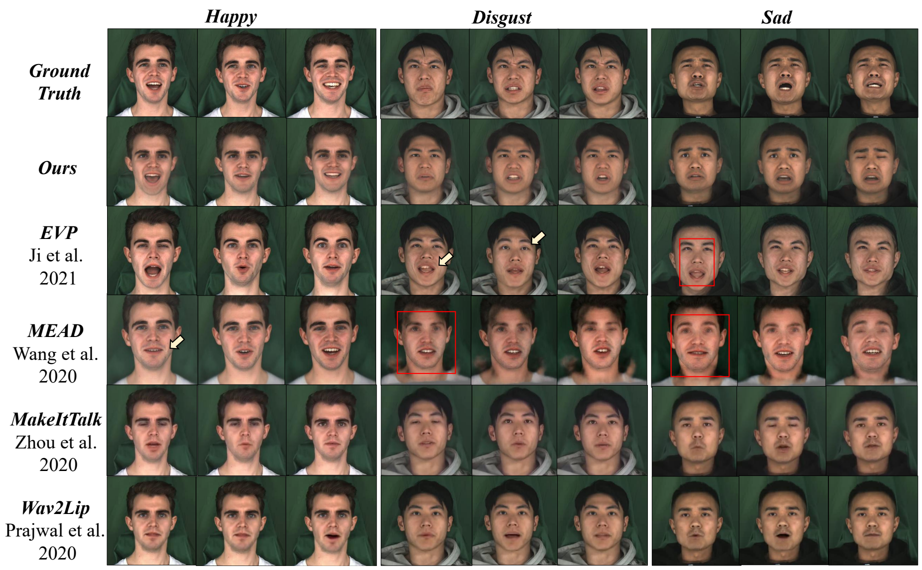


Figure 3: Qualitative comparison of our method with SOTA on MEAD dataset. MakeItTalk and Wav2Lip do not render emotion. Since the publicly available pre-trained model for MEAD [4] is only trained for Subject 1 (left), their method is unable to generalize to other identities (in red box). Similarly for EVP, the publicly available target-specific pre-trained texture models [3] are available only for Subjects 1,2 (left and middle). Hence their method fails to generalize to Subject 3 (right) as shown in red box (Subject 3 evaluated using a pre-trained model for Subject 2). The white arrow shows inconsistent emotions at the mouth and eyebrow regions.

### 4.3 Quantitative Results

We evaluate our animation results against the state-of-the-art (SOTA) emotional talking face generation methods for assessing all the essential attributes of a talking face, i.e., texture quality, lip sync, identity preservation, landmark accuracy, the accuracy of emotion generation, etc. We present the quantitative results in Table 2. The emotional talking face SOTA methods MEAD, EVP, [Eskimez *et al.*, 2020; Vougioukas *et al.*, 2019] are dataset-specific and do not generalize well for arbitrary identities outside the training dataset. For a fair comparison, the evaluation metrics of SOTA methods have been reported for the respective dataset on which they were trained. However, the performance of our method is not restricted to the training dataset. Our method is trained only on MEAD dataset, but evaluated on both MEAD and CREMA-D.

**Texture quality.** We have used PSNR, SSIM [Wang *et al.*, 2004], CPBD [Narvekar and Karam, 2009], and FID [Heusel *et al.*, 2017] metrics for quantifying the texture quality of the synthesized image. Our method outperforms the SOTA methods in most of the texture quality metrics. EVP outperforms all the methods in FID because they train person-specific texture models.

**Landmark quality.** We use Landmark Distance (LD) and Landmark Velocity Difference (LVD) [Ji *et al.*, 2021] to quantify the accuracy of lip displacements (M-LD and M-LVD) and facial expressions (F-LD and F-LVD) with respect to the GT. On the CREMA-D dataset, although our velocity error metrics are slightly higher than SOTA methods, our landmark distance error metrics are much lower than the SOTA, indicating more accurate animation.

**Identity preservation.** We compute CSIM (cosine similarity) between ArcFace features [Deng *et al.*, 2019] of the predicted frame and the input identity face of the target. Our method outperforms MEAD. EVP outperforms our method in CSIM as they train texture models specific to each target identity. On the other hand, we use a single generalized texture model for all identities. Our one-shot learning helps to generalize on different subjects using only a single image of the target identity at inference time. Whereas EVP [3] and MEAD[4] require sample images of the target in different emotions for

[3]https://github.com/jixinya/EVP

[4]https://github.com/uniBruce/Mead

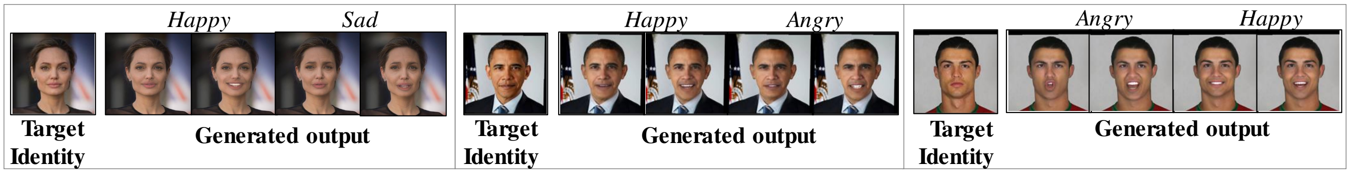


Figure 4: Results in different emotions on arbitrary target faces with different backgrounds.

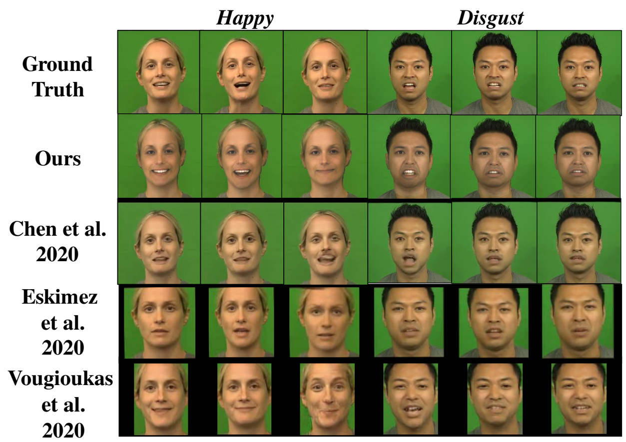


Figure 5: Qualitative comparison on CREMA-D dataset. All the SOTA methods (except [Chen *et al.*, 2020]) are trained on CREMA-D, whereas our model is trained on MEAD. [Eskimez *et al.*, 2020] is unable to generate significant emotion. [Chen *et al.*, 2020] produces distorted textures.

training their target-specific models.

**Emotion Accuracy.** We have used the emotion classifier network in EVP [Ji *et al.*, 2021] for quantifying the accuracy of generated emotions in the final animation. On both the MEAD and CREMA-D datasets, we achieve better emotion classification accuracy than that of the existing methods.

**Audio-Visual Synchronization.** We use SyncNet [Chung and Zisserman, 2016] to estimate the audio-visual synchronization accuracy in the synthesized videos. Our method achieves better lip sync than both EVP and MEAD on MEAD dataset, and performs better than [Vougioukas *et al.*, 2019] on CREMA-D.

## 4.4 Qualitative Evaluation

Fig. 3 shows our final animation results on MEAD dataset compared to the recent SOTA methods MEAD, EVP, MakeitTalk[Zhou *et al.*, 2020] and Wav2Lip[Prajwal *et al.*, 2020]. MEAD and EVP are the most relevant works since they render emotion. We have evaluated MEAD using their publicly available pre-trained model [4], which is specific to subject 1 (First three columns) and fails to generalize for other subjects (column 4 to 9). EVP fails to preserve the identity of the target subject 3 (columns 7 to 9) without fine-tuning [3]. Also, this method uses a latent feature learned from audio for emotion control, which makes the expressions inconsistent (happy emotion can be perceived as surprised or angry for subject 1, columns 1 to 3). Our method can produce better emotion and preserve identity even with one-shot learning using only a single neutral face image of the target person. Fig. 5 shows the comparative results on CREMA-D. Our method can produce realistic emotions on identities from other datasets, such as RAVDESS (Fig. 1 upper face) as well as arbitrary faces (Fig. 1 lower face and Fig. 4).

| Methods | M-LD ↓ | M-LVD ↓ | F-LD ↓ | F-LVD ↓ |
|---|---|---|---|---|
| Ours w/o Graph Encoder $E_a$ | 5.54 | 0.54 | 2.75 | 0.43 |
| Ours w/o skip connections | 5.54 | 0.54 | 2.75 | 0.43 |
| Ours w/o edge weights $\omega$ | 2.45 | 0.83 | 1.39 | 0.52 |
| Ours w/o $L_{gan}$ | 2.52 | 0.86 | 1.42 | 0.53 |
| Ours | **2.18** | **0.77** | **1.24** | **0.5** |

Table 3: Ablation study for Landmark Generation.

| Methods | PSNR↑ | CSIM ↑ | Emotion Acc.↑ |
|---|---|---|---|
| Ours w/o emotion feature | 29.83 | **0.885** | 45.00 |
| Ours w/o emotional landmark | 29.85 | 0.861 | 59.61 |
| Ours w/o one-shot learning | 29.89 | 0.767 | 84.00 |
| Ours | **30.06** | 0.789 | **85.48** |

Table 4: Ablation study for Texture Generation.

## 4.5 Ablation Study

**Landmark Generation Network $G_L$.** An ablation study of $G_L$ is presented in Table 3. *(1) Ours w/o Graph Encoder* is a variation of our network $G_L$ with only Audio Encoder $E_A$, Emotion Encoder $E_E$ and Graph Decoder $D_G$. *(2) Ours w/o skip connections* is without skip connections between Graph Encoder $E_G$ and Graph Decoder $D_G$ (shown Fig. 2). *(3) Ours w/o edge weights* is without using the learnable edge weights $\boldsymbol{\omega}$ in Eqn. 1. *(4) Ours w/o $L_{gan}$* is without adversarial learning. Our proposed network in Fig. 2 trained with the losses in Eqn. 4 leads to improved results (Table 3). In *(1)* and *(2)* due to negligible motion of landmarks, M-LVD and F-LVD are lower, but M-LD and F-LD are much higher.

**Texture Generation Network $G_T$.** An ablation study of $G_T$ is presented in Table 4. *(1) Ours w/o emotion feature:* Without input $f_e$, the emotion accuracy highly degrades (Table 4) as the network cannot generate frowns, eyebrow-raising or lowering from emotional landmarks only, as shown in Fig. 6 (second row). As CSIM is calculated between the predicted frame and the input neutral identity face of the target, the value of CSIM without emotion feature is higher. *(2) Ours w/o emotional landmark:* When the texture is generated from only speech-induced landmarks (without emotion) the emotion accuracy decreases. Learning emotion on landmarks helps generate facial expressions especially in the mouth region for emotions like happy, angry, sad, and disgust. Fig. 6 (top row) shows that without emotional landmark, emotion rendering is very restricted. *(3) Ours w/o one-shot learning:*

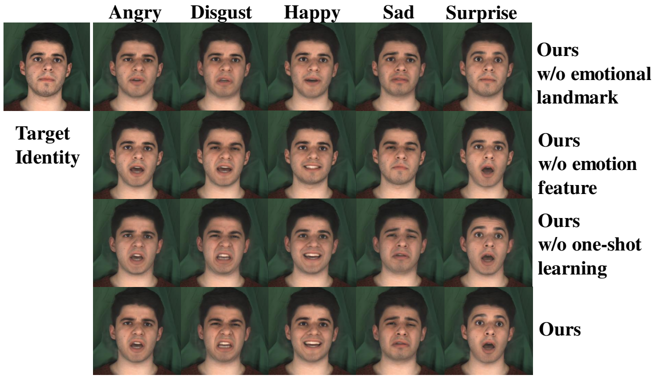


Figure 6: Qualitative Ablation for Texture Generation Network $G_T$.

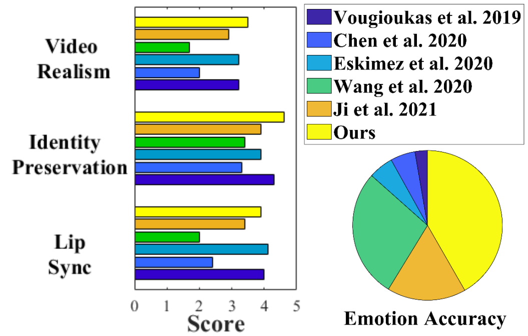


Figure 7: User Study results. The bar plots represent the average score (range 0-5, high score indicates better performance).

One-shot learning helps to achieve better identity preservation. As can be seen in Fig. 6 (last row) the facial structure, skin color of the target subject are better captured in our final animation with one-shot learning.

### 4.6 User Study

We have conducted a user study for subjective evaluation of our method against SOTA. 26 participants rate total 30 videos from [Vougioukas *et al.*, 2019; Eskimez *et al.*, 2020; Chen *et al.*, 2020], MEAD, EVP and our method. Each video is evaluated for lip sync, identity preservation, and video realism. Additionally, the participants also classify the emotion perceived from the video. The results are shown in Fig. 7. Overall our method achieves comparable performance in lip-sync and better performance over SOTA methods in identity preservation, emotion classification accuracy, and realism in video generation.

## 5 Conclusion

We propose a speech-driven emotion-controllable generalized emotional talking face generation method that uses a single image of an arbitrary target person in neutral emotion to generate animation in different emotions. We use graph convolution for geometry-aware motion and emotion generation on facial landmarks. With one-shot learning, our emotion-guided optical flow-based texture deformation network can generalize better for arbitrary target subjects when compared to existing SOTA methods. Our animation results on different benchmark datasets and for different celebrity faces show more realistic animation than SOTA methods. In future work, audio and emotion-driven head movements can be added for enhanced realism of emotional talking face animation.

## References

[Cai *et al.*, 2019] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2272–2281, 2019.

[Cao *et al.*, 2014] 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*, 5:377–390, 2014.

[Chen *et al.*, 2019] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019.

[Chen *et al.*, 2020] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[Chung and Zisserman, 2016] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

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

[Chung *et al.*, 2018] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[Das *et al.*, 2020] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, 2020.

[Delaunay *et al.*, 1934] Boris Delaunay, S Vide, A Lamémoire, and V De Georges. Bulletin de l'académie des sciences de l'urss. *Classe des sciences mathématiques et naturelles*, 6:793–800, 1934.

[Deng *et al.*, 2019] 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.

[Eskimez *et al.*, 2020] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *arXiv preprint arXiv:2008.03592*, 2020.

[Guo *et al.*, 2020] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[Hannun *et al.*, 2014] 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.

[Heusel *et al.*, 2017] 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.

[Ji *et al.*, 2021] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14080–14089, 2021.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Livingstone and Russo, 2018] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[Narvekar and Karam, 2009] 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.

[Prajwal *et al.*, 2020] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.

[Sinha *et al.*, 2020] Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2020.

[Song *et al.*, 2020] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020.

[Suwajanakorn *et al.*, 2017] 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.

[Thies *et al.*, 2019] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *arXiv preprint arXiv:1912.05566*, 2019.

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

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2020] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[Wang *et al.*, 2021] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *IJCAI*, 2021.

[Wen *et al.*, 2020] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3457–3466, 2020.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, 2018.

[Zhang *et al.*, 2021] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

[Zhou *et al.*, 2019] 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.

[Zhou *et al.*, 2020] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

[Zhou *et al.*, 2021] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021.