High-fidelity Generalized Emotional Talking Face Generation with Multi-modal Emotion Space Learning

Chao Xu[1] Junwei Zhu[2] Jiangning Zhang[2] Yue Han[1] Wenqing Chu[2]
Ying Tai[2] Chengjie Wang[2,4*] Zhifeng Xie[3] Yong Liu[1*]
[1] APRIL Lab, Zhejiang University [2]Youtu Lab, Tencent [3]Shanghai University [4]Shanghai Jiao Tong University
{21832066, 22132041}@zju.edu.cn, yongliu@iipc.zju.edu.cn, wqchu16@gmail.com
{junweizhu, vtzhang, yingtai, jasoncjwang}@tencent.com, zhifeng_xie@shu.edu.cn

Abstract

Recently, emotional talking face generation has received considerable attention. However, existing methods only adopt one-hot coding, image, or audio as emotion conditions, thus lacking flexible control in practical applications and failing to handle unseen emotion styles due to limited semantics. They either ignore the one-shot setting or the quality of generated faces. In this paper, we propose a more flexible and generalized framework. Specifically, we supplement the emotion style in text prompts and use an Aligned Multi-modal Emotion encoder to embed the text, image, and audio emotion modality into a unified space, which inherits rich semantic prior from CLIP. Consequently, effective multi-modal emotion space learning helps our method support arbitrary emotion modality during testing and could generalize to unseen emotion styles. Besides, an Emotion-aware Audio-to-3DMM Convertor is proposed to connect the emotion condition and the audio sequence to structural representation. A followed style-based High-fidelity Emotional Face generator is designed to generate arbitrary high-resolution realistic identities. Our texture generator hierarchically learns flow fields and animated faces in a residual manner. Extensive experiments demonstrate the flexibility and generalization of our method in emotion control and the effectiveness of high-quality face synthesis.

1. Introduction

Talking face generation [13,38,46,58] is the task of driving a static portrait with given audio. Recently, many works have tried to solve the challenges of maintaining lip movements synchronized with input speech contents and synthesizing natural facial motion simultaneously. However, most researchers ignore a more challenging task, emotional audio-driven talking face generation, which is critical for creating vivid talking faces.

Some works have achieved significant progress in solving the above task conditioned on emotion embedding. However, there are three continuously critical issues: 1) How to explore a more semantic emotion embedding to achieve better *generalization for unseen emotions*. Early efforts [41,47,55] adopt the one-hot vector to indicate emotion category, which could only cover the pre-defined label and lacks semantic cues. Subsequently, EVP [19] disentangles emotion embedding from the audio, while GC-AVT [23] and EAMM [18] drive emotion by visual images. However, tailored audio- and image-based emotion encoders show limited semantics and also struggle to handle unseen emotion styles. 2) Could we construct *multi-modal* emotion sources into a unified feature space to allow a more flexible and user-friendly emotion control. Existing methods only support one specific modality as the emotion condition, while the desired modality is usually not available in practical applications. 3) How to design a *high-resolution identity-generalized generator*. Early works [19,41,47] are in identity-specific design, while recent works [18,23] have started to enable one-shot emotional talking face generation. However, as shown in Fig. 1(a), GC-AVT and EAMM fail to produce high-resolution faces due to the inevitable information loss in face embedding and the challenge of estimating accurate high-resolution flow fields, respectively.

To address the aforementioned challenges, we first supplement the emotion styles in the text prompt inspired by the zero-shot CLIP-guided image manipulation [29,39,43], which could inherit rich semantic knowledge and convenient interaction after being encoded. As shown in Fig. 1(b), unseen emotions, *e.g.*, *Satisfied*, could be flexibly specified using the text description and precisely reflected on the source face. Furthermore, to achieve alignment among multi-modal emotion features, we introduce an Aligned Multi-modal Emotion (AME) encoder to unify the text, im-

*Corresponding authors

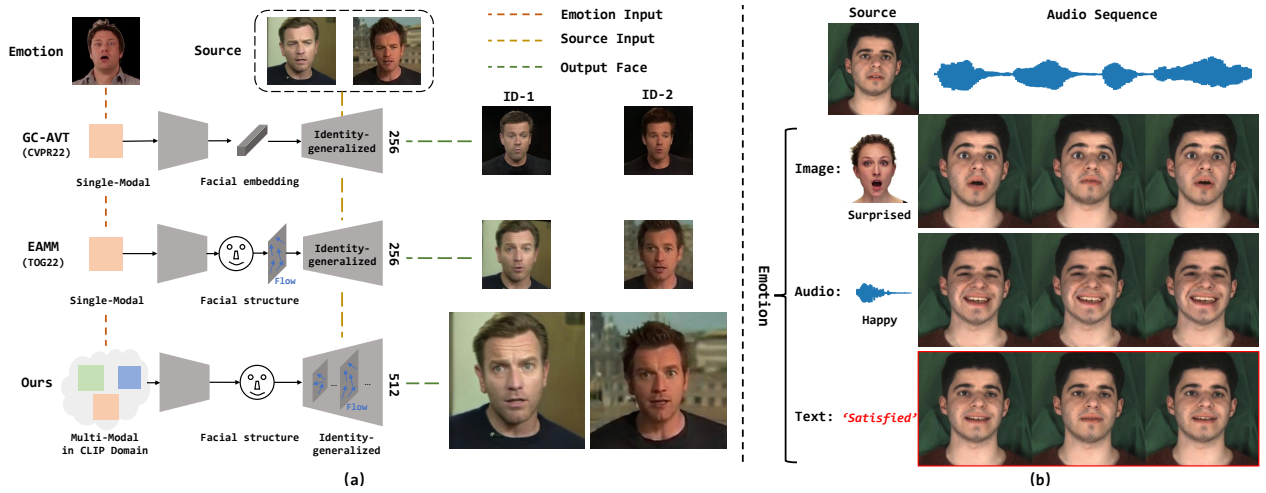

Figure 1. (a) An illustrative comparison of GC-AVT [23], EAMM [18], and our approach. First, our method supports *multi-modal emotion cues* as input. As shown in (b), given a source face, an audio sequence, and diverse emotion conditions, our results fulfill synchronized lip movements with the speech content and emotional face with the desired style. Besides, benefiting from the effective multi-modal emotion space and rich semantics of CLIP, our method could generalize to *unseen* style marked in Red. Second, the hierarchical style-based generator with coarse-to-fine facial deformation learning helps us generalize to unseen faces in high resolution and provides more *realistic details and precise emotion* than GC-AVT and EAMM. Images are from the official attached results or released codes for fair comparisons.

age, and audio emotion modality into the same domain, thus supporting flexible emotion control by multi-modal inputs, as depicted in Fig. 1(b). In particular, the fixed CLIP text and image encoders are leveraged to extract their embedding and a learned CLIP audio encoder guided by several losses to find the proper emotion representation of the given audio sequence in CLIP space.

To this end, we follow the previous talking face generation methods [34] that rely on intermediate structural information such as 3DMM, and propose an Emotion-aware Audio-to-3DMM Convertor (EAC), to distill the rich emotional semantics from AME and project them to the facial structure. Specifically, we employ the Transformer [40] to capture the longer-term audio context and sufficiently learn correlated audio-emotion features for expression coefficient prediction, which involves precise facial emotion and synchronized lip movement. Notably, a learned intensity token is extended to control the emotion intensity continuously. Furthermore, to generate high-resolution realistic faces, we propose a coarse-to-fine style-based identity-generalized model, High-fidelity Emotional Face (HEF) generator, which integrates appearance features, geometry information, and a style code within an elegant design. As shown in Fig. 1(a), unlike the EAMM that predicts the flow field at a single resolution by an isolated process, we hierarchically perform the flow estimation in a residual manner and incorporate it with texture refinement for efficiency.

In summary, we make the following three contributions:

- We propose a novel AME that provides a unified multi-modal semantic-rich emotion space, allowing flexible emotion control and unseen emotion generalization, which is the first attempt in this field.
- We propose a novel HEF to hierarchically learn the facial deformation by sufficiently modeling the interaction among emotion, source appearance, and drive geometry for the high-resolution one-shot generation.
- Abundant experiments are conducted to demonstrate the superiority of our method for flexible and generalized emotion control, and high-resolution one-shot talking face animation over SOTA methods.

2. Related Work

2.1. Audio-driven Talking Face Generation

Early efforts focus on modeling the audio information into pure latent feature space [4, 52, 56, 57] and employing a conditioned image generation framework [12, 27] to synthesize realistic faces. Recently, structural information has been leveraged as the explicit intermediate representation to bridge the gap between audio and visual domains. Das *et al.* [7] capture facial motion in landmarks [51] from the given audio sequence and then synthesizes texture conditioned on these structures. Considering the attributes are entangled within 2D landmarks, 3DMM [8] is introduced in this task [34, 44, 47–50, 54]. Specifically, FACIAL [50] predicts head pose, expression, and AU45 from deep speech features by the fully connected networks. PIRenderer [34] adopts LSTMs [15] to autoregressively deduce expressions

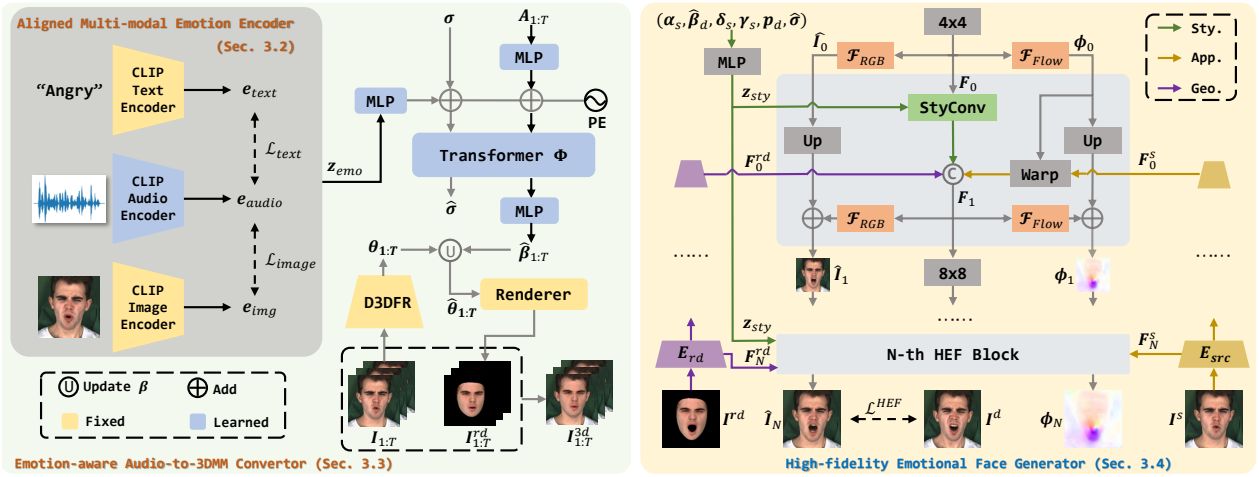

Figure 2. **Overview of the proposed method.** Our method is a two-stage framework that transfers the audio content and multi-modal emotion sources to a static portrait $\boldsymbol{I}^s$. Specifically, the Emotion-aware Audio-to-3DMM Convertor encodes MFCC sequence $\boldsymbol{A}_{1:T}$, the emotion style $\boldsymbol{z}_{emo}$ embedded from Aligned Multi-modal Emotion encoder, and a learnable intensity token $\boldsymbol{\sigma}$ to predict the expression coefficient sequence $\hat{\boldsymbol{\beta}}_{1:T}$. The followed High-fidelity Emotional Face Generator receives the style vectors $\boldsymbol{z}_{sty}$ (in Green) mapped from the modified coefficients and updated intensity token $\hat{\boldsymbol{\sigma}}$, the source appearance $\boldsymbol{F}^s_{0:N}$ (in Olive) from $\boldsymbol{I}^s$, and the driving geometry $\boldsymbol{F}^{rd}_{0:N}$ (in Violet) from $\boldsymbol{I}^{rd}$, to hierarchically generate the facial deformation $\boldsymbol{\phi}_{0:N}$ to guide the animated emotional talking face synthesis.

and poses while LSP [25] employs GRUs [3]. We follow these methods but use the non-autoregressive Transformer [40] to capture the long-term audio context and provide the sequence-level representations for more accurate coefficients regression, which helps exhibit precise emotion in the texture generator.

2.2. Emotion Conditioned Generation

Early efforts [9, 31] serve this task as the domain transfer, but they fail to synchronize lip movement with speech. Recently, MEAD [41] releases a high-quality talking head video dataset with annotations of emotion category and intensity. Subsequent works [41, 47] encode the expression labels in one-hot vectors to maintain the desired expression. EVP [19] decomposes audio into the corresponding emotion style to capture more semantic information. However, these works are in identity-specific design. Consequently, recent GC-AVT [23] and EAMM [18] explore one-shot setting and drive facial expressions by reference faces, but these methods only support emotion style obtained from a single modality and struggle to produce high-resolution faces. In contrast, we construct multi-modal emotion sources into a unified feature space, supporting *diverse* modalities within a single model. Besides, our hierarchical texture generator could produce *high-resolution* faces with the desired appearance, pose, and expression.

2.3. CLIP-guided Synthesis

CLIP [32] is perfect for visual tasks with textual assistance, which has proven effective for image editing [11, 21, 29, 35, 43], domain transfer [22, 24], and 3D avatar [16, 39]. Besides, some works [1, 55] have verified the necessity of employing multi-modal information. In this work, we supplement the emotion style in text prompt and unify the multi-modal features in the CLIP space, which contains rich semantics and unprecedented textual and visual understanding ability. Once trained, our method could generalize to *unseen* emotion styles located in similar emotion domains of CLIP, which is not considered in the previous emotional talking face generation methods.

3. Method

The emotional talking face generation aims at driving the source face by the given emotion style and audio content. The desired framework for this task should embody several core properties: 1) Since several modalities could represent emotion style, the designed method should support diverse modalities within one single model to achieve a flexible and user-friendly interaction. 2) The trained network could be applied for unseen emotion styles and identities, and generates high-resolution realistic faces. To achieve the above goals, we first design an Aligned Multi-modal Emotion (AME) encoder to produce a unified feature space in Sec. 3.2. Then a Transformer-based Emotion-aware Audio-to-3DMM Convertor (EAC) receives emotion style from AME, along with given audio, to connect audio-emotion inputs with the 3DMM (Sec. 3.3). Finally, we propose a style-based High-fidelity Emotional Face (HEF) generator to synthesize the realistic emotional talking faces of arbitrary identities by learning hierarchical facial deformation (Sec. 3.4). Our pipeline is depicted in Fig. 2.

3.1. 3D Face Descriptors

Following the previous works, we employ 3DMM parameters as the intermediate representation. With 3DMM, the 3D shape $\mathbf{S}$ and albedo texture $\mathbf{T}$ are parameterized as:

$$\begin{aligned}\mathbf{S} &= \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta},\end{aligned} \tag{1}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ denote the mean face shape and albedo texture. $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_t$ are the bases of identity, expression, and the texture computed via Principal Component Analysis (PCA). Coefficients $\boldsymbol{\theta} = \left\{\boldsymbol{\alpha} \in \mathbb{R}^{80}, \boldsymbol{\beta} \in \mathbb{R}^{64}, \boldsymbol{\delta} \in \mathbb{R}^{80}, \boldsymbol{\gamma} \in \mathbb{R}^{27}, \boldsymbol{p} \in \mathbb{R}^{6}\right\}$ describe the identity, expression, texture, illumination, pose, respectively. Although off-the-shelf 3D face reconstruction model D3DFR [8] could capture relatively accurate facial features, they fail to produce reliable expression coefficients for extreme emotional faces due to the lack of tailored training on the corresponding dataset. Consequently, we do not directly adopt the extracted expression coefficients $\boldsymbol{\beta}$ as the constraint for the EAC training (Sec. 3.3 $\mathcal{L}_{emo}$).

3.2. Aligned Multi-modal Emotion Encoder

To unify the emotion conditions from the text, audio, and image domains within one framework, we naturally choose CLIP space as the multi-domain feature space. Specifically, we design an *Aligned Multi-modal Emotion Encoder*, which consists of the fixed CLIP text and image encoders, and the learned CLIP audio encoder to produce emotion embedding $\boldsymbol{e}_{text}$, $\boldsymbol{e}_{img}$, and $\boldsymbol{e}_{audio}$. In practice, the CLIP audio encoder is the basic Transformer-based architecture with a `Cls` token for pool purposes and learns to embed $\boldsymbol{e}_{audio}$. AME receives the synchronized multi-modal inputs during training, and the output emotion code $\boldsymbol{z}_{emo}$ is the combination of the above three embedding along the batch dimension, in which each modality shares the *same* ground truth. Thus it allows pixel-level constraints on the generated face. To facilitate unified feature space learning and emotion disentanglement from the entangled audio, we further apply a feature-level loss to align the $\boldsymbol{e}_{audio}$ to CLIP textual and visual space simultaneously. For testing, we could adopt arbitrary emotion embedding as $\boldsymbol{z}_{emo}$, which is more flexible in applications:

$$\begin{aligned}&\boldsymbol{z}_{emo} = \left[\boldsymbol{e}_{text}, \boldsymbol{e}_{audio}, \boldsymbol{e}_{img}\right], \text{at training stage} \\ &\boldsymbol{z}_{emo} \in \left\{\boldsymbol{e}_{text}, \boldsymbol{e}_{audio}, \boldsymbol{e}_{img}\right\}, \text{at test stage}\end{aligned} \tag{2}$$

where $[\cdot]$ means concatenation. The merits of aligning multi-modal emotion features to CLIP space are two-folder: First, our model distills the emotion cues from the CLIP domain and inherits rich semantic knowledge to benefit unseen emotion generalization. Second, CLIP already provides shared textural and visual feature space, which is easier to train a single audio encoder than the whole network.

3.3. Emotion-aware Audio-to-3DMM Convertor

Architecture. To project the audio content and emotion style to expression coefficients of 3DMM, we propose a Transformer-based *Emotion-aware Audio-to-3DMM Convertor*. As shown in Fig. 2, $\boldsymbol{A}$ provides the information of lip movement, and $\boldsymbol{z}_{emo}$ is the emotion embedding. Besides, instead of utilizing a one-hot coding to control the emotion intensity [41], we prepend a learnable intensity token $\boldsymbol{\sigma}$ inspired by the ViT [10]. This token is the product of the base learnable intensity vector and the intensity scalar: $\boldsymbol{\sigma} = \mu\boldsymbol{\sigma}_{base}$, where $\mu \in \{1, 2, 3\}$ during training, corresponds to the ground-truth intensity annotated in MEAD. It can be a random value range from 1 to 3 during testing. In practice, we map the audio feature dimension and concatenate them with intensity token $\boldsymbol{\sigma}$. The MLP is used to initially separate emotion cues from CLIP space. The above are added with positional embedding PE and fed into the Transformer $\boldsymbol{\Phi}$ for expression coefficients prediction:

$$\hat{\boldsymbol{\sigma}}, \bar{\boldsymbol{\beta}} = \boldsymbol{\Phi}([\boldsymbol{\sigma}, \text{MLP}(\boldsymbol{A})] + \text{PE} + \text{MLP}(\boldsymbol{z}_{emo})), \tag{3}$$

$$\hat{\boldsymbol{\beta}} = \text{MLP}(\bar{\boldsymbol{\beta}}). \tag{4}$$

Objectives. We train this stage by using five losses:

$$\begin{aligned}\mathcal{L}^{EAC} = &\lambda^{EAC}_{clip}\mathcal{L}_{clip} + \lambda^{EAC}_{emo}\mathcal{L}_{emo} + \lambda^{EAC}_{rec}\mathcal{L}_{rec} \\ &+ \lambda^{EAC}_{lm}\mathcal{L}_{lm} + \lambda^{EAC}_{reg}\mathcal{L}_{reg}.\end{aligned} \tag{5}$$

Clip Loss $\mathcal{L}_{clip}$: As stated in Sec. 3.2, we force the emotion feature from the audio close to that from the text and image by using cosine distance: $\mathcal{L}_{image} = 1 - \cos(\boldsymbol{e}_{img}, \boldsymbol{e}_{audio})$, $\mathcal{L}_{text} = 1 - \cos(\boldsymbol{e}_{text}, \boldsymbol{e}_{audio})$, $\mathcal{L}_{clip} = \mathcal{L}_{image} + \mathcal{L}_{text}$.
Emotion Consistency Loss $\mathcal{L}_{emo}$: This is a critical term to distill and infuse the semantic emotion representation of CLIP. Due to the unreliable of extracted expression coefficients (Sec. 3.1), we turn to the image level for help, projecting the modified 3DMM onto the 2D image plane with a differentiable renderer: $\boldsymbol{R}(\hat{\boldsymbol{\theta}}) \rightarrow \boldsymbol{I}^{rd}$, which is then blended to the original face by the face mask $\boldsymbol{M}$ output from $\boldsymbol{R}$: $\boldsymbol{I}^{3d} = \boldsymbol{M} \odot \boldsymbol{I}^{rd} + (1 - \boldsymbol{M}) \odot \boldsymbol{I}$. After that, we adopt an emotion recognition network [28] to compute the perceptual difference between the input and rendered images:

$$\mathcal{L}_{emo} = \left\| \boldsymbol{\varphi}_{emo}(\boldsymbol{I}^{3d}) - \boldsymbol{\varphi}_{emo}(\boldsymbol{I}) \right\|_2, \tag{6}$$

where $\boldsymbol{\varphi}_{emo}$ is the backbone before the last linear layer.
Reconstruction Loss $\mathcal{L}_{rec}$: We compute the pixel level loss between the input and the rendered images on the face area: $\mathcal{L}_{rec} = \left\| \boldsymbol{I}^{rd} - \boldsymbol{M} \odot \boldsymbol{I} \right\|_2$.
Landmark Loss $\mathcal{L}_{lm}$: We predict 68 points from the original 3DMM $\boldsymbol{\theta}$ and the modified one $\hat{\boldsymbol{\theta}}$, obtaining $\boldsymbol{l}$ and $\hat{\boldsymbol{l}}$. $\mathcal{L}_2$ distance is used to measure them: $\mathcal{L}_{lm} = \left\| \boldsymbol{l} - \hat{\boldsymbol{l}} \right\|_2$.
Expression Regularization Loss $\mathcal{L}_{reg}$: This term is used to smooth the training phase, calculating the distance between the $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ with a small weight: $\mathcal{L}_{reg} = \left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right\|_2$.

3.4. High-fidelity Emotional Face Generation

Architecture. As shown in Fig. 1(a), GC-AVT does not generate consistent texture and background with the source, while EAMM relies on aligned inputs and produces blurred results. Both are in low-resolution and poor quality. Thus, we carefully modify StyleGAN2 [20] and propose *High-fidelity Emotional Face Generation*. Specifically, as shown in Fig. 2, we randomly sample a driving face $\boldsymbol{I}^d$ from given clips and a face with the same identity but different emotion as the source $\boldsymbol{I}^s$. To transfer the audio-synchronized lip movement, pose, and expression from the drive to the source face, the style code is defined as:

$$\boldsymbol{z}_{sty} = \text{Linear}([\boldsymbol{\alpha}_s, \hat{\boldsymbol{\beta}}_d, \boldsymbol{\delta}_s, \boldsymbol{\gamma}_s, \boldsymbol{p}_d, \hat{\boldsymbol{\sigma}}]). \tag{7}$$

In addition to $\boldsymbol{z}_{sty}$, $\boldsymbol{E}_{src}$ extracts pyramid appearance features $\boldsymbol{F}^s_{0:N}$ from the $\boldsymbol{I}^s$, where N is the number of HEF blocks. $\boldsymbol{E}_{rd}$ simultaneously provides hierarchical features $\boldsymbol{F}^{rd}_{0:N}$, which embody the emotional textures and geometrical guidance of the desired face from the $\boldsymbol{I}^{rd}$. Thus we have three faithful implicit-explicit inputs for HEF. Furthermore, to align the $\boldsymbol{F}^s_i$ and $\boldsymbol{F}^{rd}_i$, different from existing methods [18, 34, 36, 45] employing isolated flow fields prediction module, we simultaneously estimate the spatial deformations and refine the final images in a residual manner within each HEF block. As depicted in Fig. 2, we have:

$$\boldsymbol{F}_{i+1} = [\boldsymbol{\mathcal{T}}(\boldsymbol{\phi}_i, \boldsymbol{F}^s_i), \textbf{StyleConv}(\boldsymbol{F}_i, \boldsymbol{z}_{sty}), \boldsymbol{F}^{rd}_i], \tag{8}$$

$$\boldsymbol{\phi}_{i+1} = \boldsymbol{\mathcal{F}}_{flow}(\boldsymbol{F}_{i+1}) + \textbf{Up}(\boldsymbol{\phi}_i), \tag{9}$$

$$\hat{\boldsymbol{I}}_{i+1} = \boldsymbol{\mathcal{F}}_{rgb}(\boldsymbol{F}_{i+1}) + \textbf{Up}(\hat{\boldsymbol{I}}_i), \tag{10}$$

where **StyleConv** denotes the style convolution in StyleGAN2 (**Up** + Conv3 $\times$ 3 with modulation). Please refer to its paper for more details. **Up** means upsampling, $\boldsymbol{\mathcal{T}}$ denotes warping operation, $\boldsymbol{\mathcal{F}}_{flow}$ converts the high-dimensional features to dense flow fields, and $\boldsymbol{\mathcal{F}}_{rgb}$ to realistic RGB images, respectively.

Objectives. We employ three loss terms to measure the difference between $\boldsymbol{I}^d$ and $\hat{\boldsymbol{I}}_N$ at the pixel and perceptual level by a *Reconstruction Loss* $\mathcal{L}_{rec}$ as $\mathcal{L}_1$ distance and a *Perceptual Loss* $\mathcal{L}_p$ as the LPIPS loss [53]. Besides, we adopt *Adversarial Loss* to ensure the authenticity of the generated faces. The overall objective is a combination of the above:

$$\mathcal{L}^{HEF} = \lambda^{HEF}_{rec}\mathcal{L}_{rec} + \lambda^{HEF}_{p}\mathcal{L}_{p} + \lambda^{HEF}_{adv}\mathcal{L}_{adv}. \tag{11}$$

4. Experiments

4.1. Datasets and Implementation Details

Datasets. Our model is trained on MEAD [41] with eight expression types (neutral, angry, contempt, disgusted, fear, happy, sad, and surprised) and three intensity levels (levels 1, 2, 3), which contains fine-grained emotion annotation,

Figure 3. Qualitative results on MEAD dataset. Different columns mean several sampled timestamps (*same as the following figures*). Images are from officially released codes for fair comparisons.

helping distill emotion space from CLIP for unseen style generalization. We randomly select 36 identities of front-view videos for training and the rest identities for testing.

Metrics. We adopt PSNR, SSIM [42], and FID [14] to evaluate the quality of generated images. We use Landmarks Distance (LMD) [2] around the mouth and the confidence score (Sync) proposed in SyncNet [6] to measure the accuracy of mouth shapes and lip synchronization. We compute Cosine Similarity (CSIM) to evaluate identity preservation. CurricularFace [17] is used to extract identity embedding. We use Emotion Feature Distance (EFD) to measure the accuracy of the emotion representation, which is extracted by FAN [26] different from the model in the loss calculation.

Implementation Details. The EAC and HEF are trained independently. For EAC, we randomly sample consecutive $T = 32$ clips for training. The values of the loss weights are set to $\lambda^{EAC}_{clip} = 1$, $\lambda^{EAC}_{emo} = 1$, $\lambda^{EAC}_{rec} = 100$, $\lambda^{EAC}_{lm} = 0.1$, $\lambda^{EAC}_{reg} = 0.01$. We use a learning rate of 0.0002 and 128 batch size to train EAC with the Adam optimizer on one V100 GPU. For HEF, we fix EAC and the values of the loss weights are set to $\lambda^{HEF}_{rec} = 5$, $\lambda^{HEF}_{p} = 5$, $\lambda^{HEF}_{adv} = 1$. This training phase also adopts Adam optimizer with 0.002 learning rate, using 8 V100 GPUs and 1 clip (8 images) per GPU. The output image size of HEF is 512×512 with $N = 7$ blocks. The audios are pre-processed to 16kHz, then converted to mel-spectrograms with FFT window size 1280, hop length 160, and 80 Mel filter-bank as PC-AVS [57].

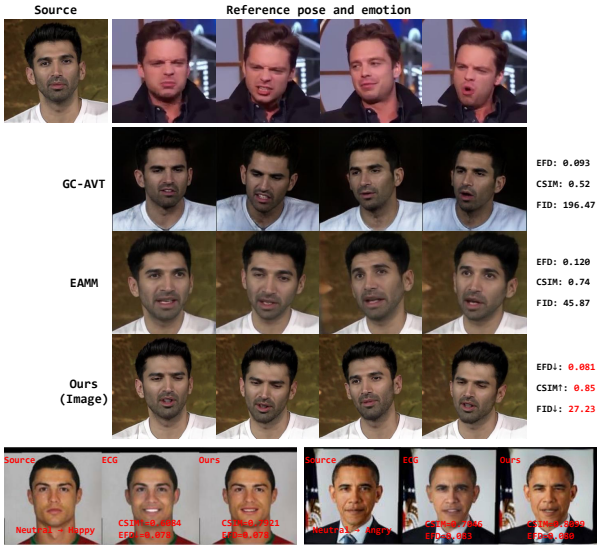

Figure 4. Qualitative comparison with GC-AVT, EAMM, and ECG. The top part is sampled from Fig. 3 of GC-AVT. The bottom part is sampled from Fig. 4 of ECG. Quantitative results of these cases are attached in the figure. We ignore the metric of mouth shape because the audios for these sequences are not available.

4.2. Comparison with State-of-the-Arts

Qualitative Results. We show the results of M003 to ensure that each method could animate the source face, and this identity is not in our training set. The first frame of each test video as the source and its audio, label, and a random face from the same video for multi-modal emotion conditions. As shown in Fig. 3, we select three frames of two emotion styles for comparison. It can be seen that common audio-driven methods, Wav2Lip [30] and PC-AVS, struggle to generate desired emotions with synchronized lip shapes, while the synthesized images from MEAD are of poor quality. EVP and EAMM suffer identity inconsistency with the source and show less rich expression due to lacking intensity modeling. In contrast, benefiting from sufficient emotion semantics and intensity learning, our method with text as the emotion condition produces more accurate expressions. Besides, our results show more realistic textures than all competitors due to coarse-to-fine flow field and image refining. We further compare our method with GC-AVT, EAMM, and ECG [37]. As shown in Fig. 4, since GC-AVT does not release codes, we adopt the officially attached results in its paper and employ image as our emotion condition for a fair comparison. In terms of emotion accuracy, identity consistency, and image quality, our method obviously outperform these SOTA methods. We also attach the corresponding quantitative results of this case on EFD, CSIM, and FID in Fig. 4, which are consistent with the qualitative results. The same conclusion could deduced from the bottom part of Fig. 4 when compared with ECG.

Method	EFD ↓	LMD ↓	Sync ↑	CSIM ↑	FID ↓	PSNR ↑	SSIM ↑
Wav2Lip	0.112	2.59	3.26	0.82	20.15	29.22	0.70
PC-AVS	0.110	2.68	3.12	0.80	29.55	28.97	0.68
MEAD	0.084	2.62	3.09	0.81	30.69	28.48	0.65
EVP	0.106	2.54	3.21	0.70	**12.83**	29.67	0.73
EAMM	0.092	2.50	3.26	0.74	29.01	29.33	0.75
Ours-A	0.069	2.36	3.50	0.83	15.91	30.09	0.85
Ours-I	0.071	2.36	3.53	**0.84**	15.89	**30.10**	0.87
Ours-T	**0.065**	**2.31**	**3.57**	**0.84**	15.90	30.09	**0.88**

Table 1. Quantitative comparison on MEAD dataset. Ours-A, -I, and -T mean audio, image, and text, respectively.

Quantitative Results. We adopt several metrics to evaluate the superiority of our approach on image quality, landmark accuracy, lip synchronization, identity preservation, and emotion accuracy. As shown in Tab. 1, our method outperforms most metrics except for the FID. EVP achieves higher FID but exhibits a weak manipulated ability, which can be inferred from the lower CSIM and Sync, and higher LMD. Besides, comparing the last three rows, we observe that emotion modality mainly affects structural metrics, and text-driven results show better performance than that of audio- and image-driven in most metrics, which may attribute to the better emotion disentanglement of text prompts. Thus *we use the text as our default emotion condition* in the following experiments. We further report the user study in supplementary materials, specially evaluating the overall quality, the generalization to unseen emotions, and the video temporal consistency.

4.3. Further Analysis

Generalizing to Unseen Emotion Styles. Unseen emotion styles include compound and totally new styles. As shown in Fig. 5, rows 2 and 3 are the basic styles, row 4 shows the results of the given *Sadly surprised*, and row 5 of the average embedding of *Happy* and *Surprised*, which indicates the flexible manipulation for compound emotion. We further present the new style *Hatred* in the sixth row. The correct exhibition of these unseen styles verifies the flexibility and rich semantic priors of the CLIP feature space.

Generalizing to Unseen Identities. As shown in Fig. 1 and Fig. 4, our method trained on MEAD could generalize to unseen identities from VoxCeleb2 [5]. We further conduct a qualitative visualization in Fig. 6. Specifically, we sample a face from CelebA-HQ as an unseen identity. The generated faces preserve the identity and exhibit desired emotion, reflecting both on realistic and rendered faces. Besides, we attach the flow map predicted from HEF in the fifth column. Please pay attention to facial movements, especially in the mouth and eyes. We can conclude that HEF accurately models the emotion-related facial movement conditioned on the intermediate structure, which is not sensitive to the identity textures. Thus MEAD is sufficient to provide diverse movements for training.

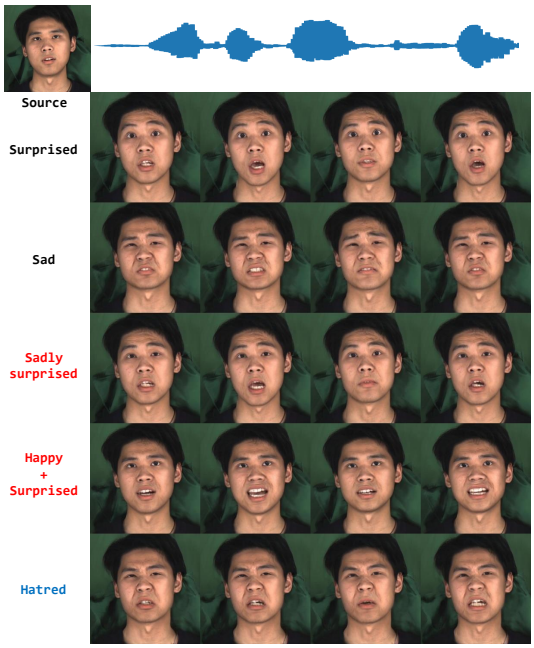

Figure 5. Results of unseen emotion styles. Rows 4 and 5 (in Red) are the compound styles, and row 6 (in Blue) is a totally new style.

Figure 6. Results of unseen identity. We visualize the rendered images, final outputs, and predicted flow fields. The color wheel of flow fields is attached on the top right for reference.

Continuous Emotion Style Control. We conduct a qualitative experiment to evaluate the effectiveness of our method for controlling emotion style. As shown in Fig. 7, our method could change the emotion representation between two distinct styles, rather than previous methods only taking a neutral face as the source. We increase the intensity value from 1 to 2.5, which shows continuous and accurate expression changes. Please pay attention to the mouth and eyes regions. Furthermore, we explore the style semantics that already encode intensity, *e.g.*, *Extremely surprised*. Comparing rows 4, 5 with rows 2, 3, CLIP struggles to distinguish the intensity prompt. Thus the intensity token is essential in our method.

Interpretability of Generalization. To analyze the ability for unseen emotion generalization, we use t-SNE to visualize the latent codes of updated intensity token. As shown in Fig. 8, the four basic emotion styles (Marker △) represent distinct clusters, while the clusters of those unseen emotion styles (Marker □) are mainly located in semantically similar areas. We further adopt two emotion-unrelated prompts (Marker ◊). Obviously, these two clusters are far from those of emotion and highly overlapped since both are meaningless to our model, *i.e.*, their sampled faces are not changed in Fig. 8. Thus, our method could represent various styles located in similar emotion domains of CLIP.

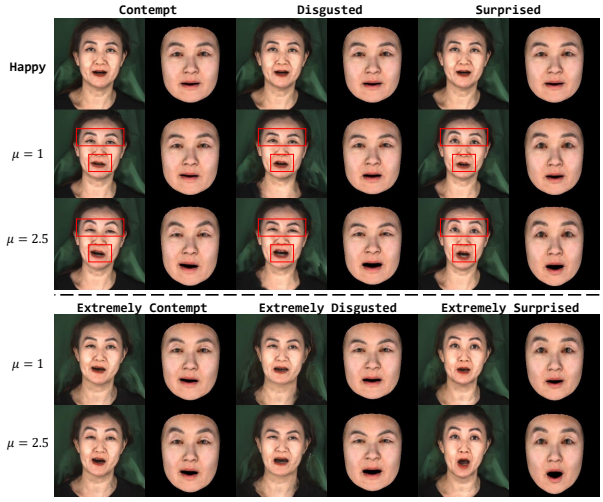

Figure 7. Results of different emotion styles and intensity levels. The top part shows the manipulation from the happy to three distinct emotion styles. The bottom part shows the results of style semantics that already encode intensity.

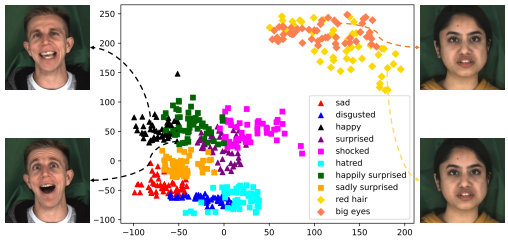

Figure 8. Clusters of the intensity token with the emotion and emotion-unrelated text descriptions. Markers △, □, ◊ mean basic emotion, unseen emotion, and emotion-unrelated text prompts.

4.4. Ablation Study and Efficiency Evaluation

Loss Functions. Emotion consistency loss is critical to distill emotion cues from CLIP and improve the quality of expression coefficients. To verify its effectiveness, we present the qualitative results in Fig. 9. This loss term helps to capture precise and fine expression details, facilitating the 3D face reconstruction and emotion manipulation. Besides, consistent quantitative results are reported in Tab. 3.

Emotion Encoding. Tab. 3 shows one-hot encoding and language pre-trained model GPT2 [33] achieve comparable results with CLIP on pre-defined styles. We further explore the effect of these emotion encodings on the unseen emotion in Fig. 10. One-hot fails to represent a new emotion style due to the fixed pattern, and compound styles due to

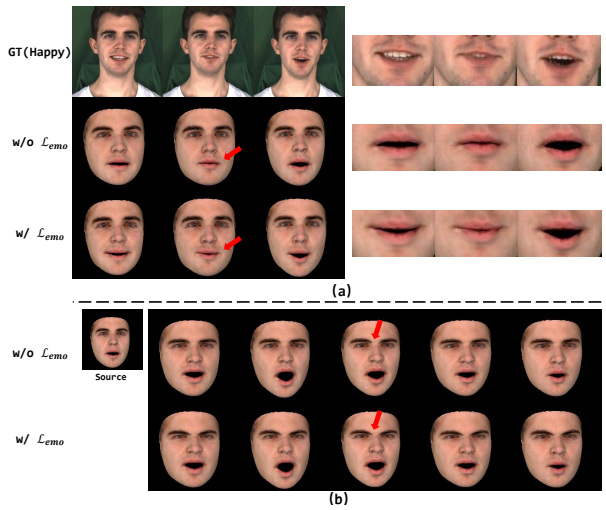

Figure 9. Qualitative ablation study for emotion consistency loss of EAC. (a) shows the effect on 3D face reconstruction of *happy* emotion and (b) illustrates manipulation by *angry* emotion.

Method	Params (M)	Training		Inference (1×V100)	
		GPUs	days	Memo (G)	ms
EAMM-256	101.89	4×2080Ti	5	3.8	21
Ours-256	62.38	4×V100	3	2.4	26
Ours-512	62.97	8×V100	7	3.2	37

Table 2. Efficiency evaluation during training and inference.

lacking semantics. GPT2 is not available to the visual cues and struggles to reflect the unseen textual semantics to the image domain. Our method inherits rich visual and textual priors from CLIP, exhibiting better generalized ability.

Architecture of EAC. To verify the effectiveness of the Transformer encoder in EAC, we replace it with stacked fully-connected layers or GRU-based recurrent neural networks. As shown in Tab. 3, our Transformer-based model obviously outperforms the above two architectures.

Flow Estimation of HEF. We further design two variants to explore the flow estimation structure, *i.e.*, the one directly outputs the flow fields without residual refinement at each scale (w/o res.), and another one uses the fixed 64×64 flow field to adapt to the following high-resolution layers of HEF instead of further updating hierarchically (w/o hie.). Please refer to supplementary materials for modification details. The Fig. 11 and Tab. 3 verify the effectiveness of hierarchically learning deformation in the residual manner. Besides, we observe that the fixed low-resolution flow field cannot produce accurate animated high-resolution faces, which explains why FOMM [36] and EAMM are not competent for high-resolution generation.

Efficiency Evaluation. We report the running efficiency in the Tab 2. The cost increases obviously as the resolution becomes larger. Notably, we achieve comparable training costs and inference speed with EAMM under the same resolution (256×256), but our method is more *memory-friendly*, *i.e.*, lower model size and memory cost, which is compatible with relatively cheap devices, *e.g.*, 1080Ti.

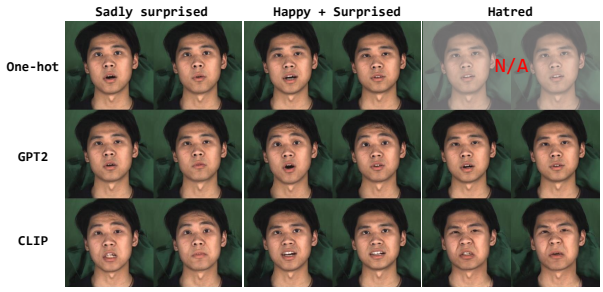

Figure 10. Qualitative ablation study of EAC with different emotion encodings on *unseen styles*. This case is sampled from Fig. 5.

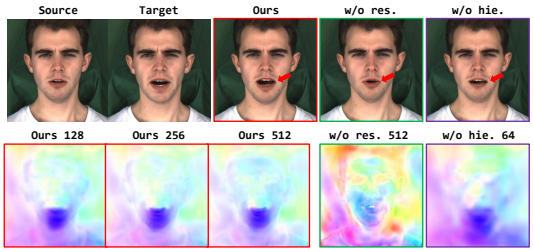

Figure 11. Qualitative ablation study of HEF with different flow estimation variants. We visualize the flow fields of w/o res. at scale 512 and the fixed 64×64 flow fields of w/o hie., both fail to model the precise movement, while our method gradually refines the high-resolution flow fields by hierarchical residual learning.

Method	EFD ↓	LMD ↓	Sync ↑
w/o $\mathcal{L}_{emo}$	0.096	2.40	3.53
w/ $\mathcal{L}_{emo}$	**0.065**	**2.31**	**3.57**
One-hot	0.070	2.33	3.53
GPT2	0.067	2.33	3.56
CLIP	**0.065**	**2.31**	**3.57**
MLPs	0.122	3.54	2.23
GRUs	0.088	2.47	3.19
Transformers	**0.065**	**2.31**	**3.57**
w/o res.	0.082	2.46	3.21
w/o hie.	0.076	2.42	3.25
Ours	**0.065**	**2.31**	**3.57**

Table 3. Quantitative ablation study with different losses and components, conducted on MEAD with *basic styles* by default.

5. Conclusions

In this paper, we propose a novel one-shot emotional talking face generation framework. Specifically, a unified multi-modal CLIP-based emotion space and a texture generator are proposed to generalize to unseen emotions and guarantee the quality of animated faces, respectively. Qualitative and quantitative experiments demonstrate the superiority of our approach over SOTA methods.

Acknowledgments. This work is supported by the Key R&D Program Project of Zhejiang Province (2021C01035).

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 3

[2] 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. 5

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[4] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 6

[6] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 5

[7] 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*, pages 408–424. Springer, 2020. 2

[8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 4

[9] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Kevin Frans, LB Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021. 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1

[14] 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. *Advances in neural information processing systems*, 30, 2017. 5

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[16] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 5

[18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022. 1, 2, 3, 5

[19] 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. 1, 3

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 5

[21] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. 2021. 3

[22] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021. 3

[23] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 1, 2, 3

[24] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022. 3

[25] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 3

[26] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pages 3866–3870. IEEE, 2019. 5

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[28] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 4

[29] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1, 3

[30] 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. 6

[31] 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

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 7

[34] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 2, 5

[35] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Styleclipdraw: Coupling content and style in text-to-drawing synthesis. *arXiv preprint arXiv:2111.03133*, 2021. 3

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 8

[37] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*, 2022. 6

[38] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 2022. 1

[39] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022. 1, 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[41] 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. 1, 3, 4, 5

[42] 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. 5

[43] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022. 1, 3

[44] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1478–1486, 2021. 2

[45] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 54–71. Springer, 2022. 5

[46] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 1

[47] Zipeng Ye, Zhiyao Sun, Yu-Hui Wen, Yanan Sun, Tian Lv, Ran Yi, and Yong-Jin Liu. Dynamic neural textures: Generating talking-face videos with continuously controllable expressions. *arXiv preprint arXiv:2204.06180*, 2022. 1, 2, 3

[48] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[49] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*, 2022. 2

[50] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3867–3876, 2021. 2

[51] Jiangning Zhang, Liang Liu, Zhucun Xue, and Yong Liu. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4402–4406. IEEE, 2020. 2

[52] Jiangning Zhang, Xianfang Zeng, Chao Xu, Jun Chen, Yong Liu, and Yunliang Jiang. Apb2facev2: Real-time audio-guided multi-face reenactment. *arXiv preprint arXiv:2010.13017*, 2020. 2

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[54] 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. 2

[55] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73, 2021. 1, 3

[56] 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. 2

[57] 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. 2, 5

[58] 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. 1