Style^{2}Talker: High-Resolution Talking Head Generation with Emotion Style and Art Style

Shuai Tan, Bin Ji, Ye Pan*

Shanghai Jiao Tong University
{tanshuai0219, bin.ji, whitneypanye}@sjtu.edu.cn

Abstract

Although automatically animating audio-driven talking heads has recently received growing interest, previous efforts have mainly concentrated on achieving lip synchronization with the audio, neglecting two crucial elements for generating expressive videos: emotion style and art style. In this paper, we present an innovative audio-driven talking face generation method called Style^{2}Talker. It involves two stylized stages, namely Style-E and Style-A, which integrate text-controlled emotion style and picture-controlled art style into the final output. In order to prepare the scarce emotional text descriptions corresponding to the videos, we propose a labor-free paradigm that employs large-scale pretrained models to automatically annotate emotional text labels for existing audio-visual datasets. Incorporating the synthetic emotion texts, the Style-E stage utilizes a large-scale CLIP model to extract emotion representations, which are combined with the audio, serving as the condition for an efficient latent diffusion model designed to produce emotional motion coefficients of a 3DMM model. Moving on to the Style-A stage, we develop a coefficient-driven motion generator and an art-specific style path embedded in the well-known StyleGAN. This allows us to synthesize high-resolution artistically stylized talking head videos using the generated emotional motion coefficients and an art style source picture. Moreover, to better preserve image details and avoid artifacts, we provide StyleGAN with the multi-scale content features extracted from the identity image and refine its intermediate feature maps by the designed content encoder and refinement network, respectively. Extensive experimental results demonstrate our method outperforms existing state-of-the-art methods in terms of audio-lip synchronization and performance of both emotion style and art style.

Introduction

The automatic animation of images plays a crucial role in computer graphics and vision, finding applications in various fields such as film production, virtual avatars, and social media (Pataranutaporn et al. 2021). To enable more extensive applications, two essential elements come into play: emotion style and art style, which we refer to as style^2. Emotion style allows users to convey communicative information more efficiently with diverse expressions (Ekman and Rosenberg 2005), while art style can evoke different human experiences, leading to stronger visual effect and applications in entertainment (Yang et al. 2022b). However, the generation of audio-driven talking head videos with both styles from a regular face photo has not been extensively explored.

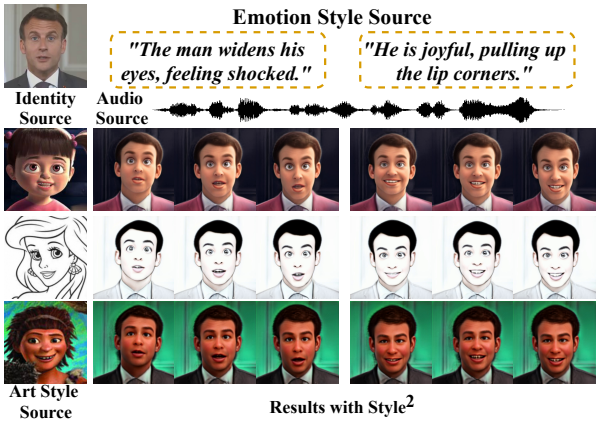

Figure 1: Illustrative animations produced by our Style^{2}Talker. Our approach takes an identity image and an audio clip as inputs and generates a talking head with emotion style and art style, which are controlled respectively by an emotion source text and an art source picture.

When it comes to emotion style, previous works have either used a one-hot emotion label as the emotion source (Ji et al. 2021; Sinha et al. 2021), limiting the range of expressions, or relied on an additional emotion video (Ji et al. 2022), which can be inconvenient as finding a video with the desired emotion style might not always be feasible. In contrast, we design our system to enable more user-friendly input by allowing users to provide a text description of the emotion style, encompassing emotion categories and detailed facial muscle movements. On the other hand, assuming that using a picture is more suitable and visual to indicate art style like the rendering color of specific part, face shape, etc, we prefer an art picture for the art style reference. While there have been several efforts in single image style transfer (Yang et al. 2022a; Choi et al. 2020), these methods face challenges when generating continuous videos driven by an

*Corresponding author.

audio clip. As a result, the objective of our study is to develop a system capable of synthesizing high-resolution talking face videos, whose identity, mouth shapes, emotion style and art style align with the input identity image, audio clip, emotion textual description and art picture, respectively.

Specifically, we present a novel framework named Style^{2}Talker, designed to achieve this objective by involving emotionally stylized stage Style-E and artistically stylized stage Style-A. We leverage 3DMM (3D Morphable Model) (Deng et al. 2019) coefficients as a brief intermediate representation to bridge the gap between the two stylized stages. The Style-E stage develops a latent diffusion model that acts as the emotionally stylized motion generator. By taking the text as emotion style descriptors and audio as motion driving sources, the generator produces high-quality, realistic expression coefficient sequences that convey the desired emotion styles. The motivations for using diffusion model are: (1) Talking face generation with text-driven emotion style is a classic conditional generation task, and diffusion model exhibits excellent performance in this area (Stypukowski et al. 2023) and is more stable than conditional GAN (Mirza and Osindero 2014) for training by removing adversarial process. (2) The diffusion and denoising process in the training phase allows the model to be more robust and precise in mining the expected results from the random noise based on the provided conditions during inference. However, training such a text-driven emotional generation model necessitates textual descriptions for emotional expression, which is absent until now. To address this, we devise an automatically annotated method by leveraging large-scale pretrained models. With the generated text descriptors, we incorporate the CLIP text encoder (Radford et al. 2021) and an audio encoder to extract emotion representations and audio features, which are passed through the motion generator as denoising conditions for emotion stylization. To optimize the inference time, we employ a simpler and more efficient diffusion model as our motion generator.

The Style-A stage is based on DualStyleGAN (Yang et al. 2022a), which introduces an art-specific path compared to StyleGAN (Karras et al. 2020) to transfer the art style of a single image to those of an art reference picture. To generate continuous artistic frames that align with 3DMM coefficients from the Style-E stage, we draw inspiration from the finding that the spatial feature map in the $\mathcal{F}$ space of StyleGAN is highly related to expressing the pose of the generated images (Yin et al. 2022). Therefore, we employ a coefficient-driven motion generator to produce flow maps from coefficients, which are then used to warp the spatial feature map. This way, StyleGAN is able to synthesize talking face videos with the aid of the continuously warped spatial feature maps. However, the current results suffer from details loss due to the GAN Inversion encoder of DualStyleGAN, which primarily focuses on face reconstruction and neglects preserving detailed information like background and texture. To tackle this issue, we introduce a content encoder that provides StyleGAN with additional multi-level content features extracted from the identity image. These features supplement the texture details using skip connections, following Yang et al. (2022b). Furthermore, we design a refinement network to eliminate potential ghost shadows caused by the misalignment between the warped spatial feature maps and fixed content features. Overall, the Style^{2}Talker framework offers a promising approach to achieve emotionally and artistically stylized talking face generation with improved continuity and realism. Through extensive experiments, we demonstrate the effectiveness and superiority of our method over state-of-the-art approaches.

Our principal contributions are summarized as follows:

- We present a novel system that facilitates high-resolution talking head generation with emotion style and art style. To the best of our knowledge, we are the first to combine both styles in the context of talking-head task under the guidance of emotion text and art picture simultaneously.
- We explore an innovative labor-free method for automatically generating text descriptions to serve as emotion style sources with the assistance of large-scale pretrained models. By incorporating audio clips and synthetic text, we introduce an efficient diffusion model to synthesize emotionally stylized motion coefficients.
- We demonstrate a successful extensive application of modified StyleGAN to enable high-resolution talking head generation with the desired art style driven by the generated emotional coefficients and an art picture.

Related Work

Audio-driven Talking Head Synthesis

In recent times, audio-driven face animation has garnered significant attention owing to its wide range of applications in fields such as like film-making and virtual reality. Existing approaches can be broadly categorized into two groups: person-specific methods and person-independent methods. Person-specific methods (Guo et al. 2021; Ye et al. 2023) focus on training models for specific individuals using videos featuring those individuals. By avoiding the inconsistency of speech styles across different speakers (Wang et al. 2022a), person-specific methods achieve superior performance on specific individual. However, a drawback of person-specific methods is their limited applicability to other identities. On the contrary, person-independent models generalize their capabilities to arbitrary identity by training on multi-speaker audio-visual datasets. These methods not only achieve lip motions synchronized with the audio by animating the face/-mouth regions (Alghamdi et al. 2022; Prajwal et al. 2020), but also generate head motions that align with the rhythm of the audio to make the outputs more realistic (Zhang et al. 2022).

Recent advancements (Ji et al. 2021; Pan et al. 2023; Tan, Ji, and Pan 2023) have also explored the synthesis of emotional expressions in talking faces. Ji et al. (2021) and Sinha et al. (2021) utilize one-hot emotion labels as input to generate emotional talking faces, while others (Ji et al. 2022; Ma et al. 2023b) resort to another video for emotion source. In contrast, our approach offers a more user-friendly control by allowing users to input easy-to-use text descriptions to suggest the desired emotion style. In this way, users have

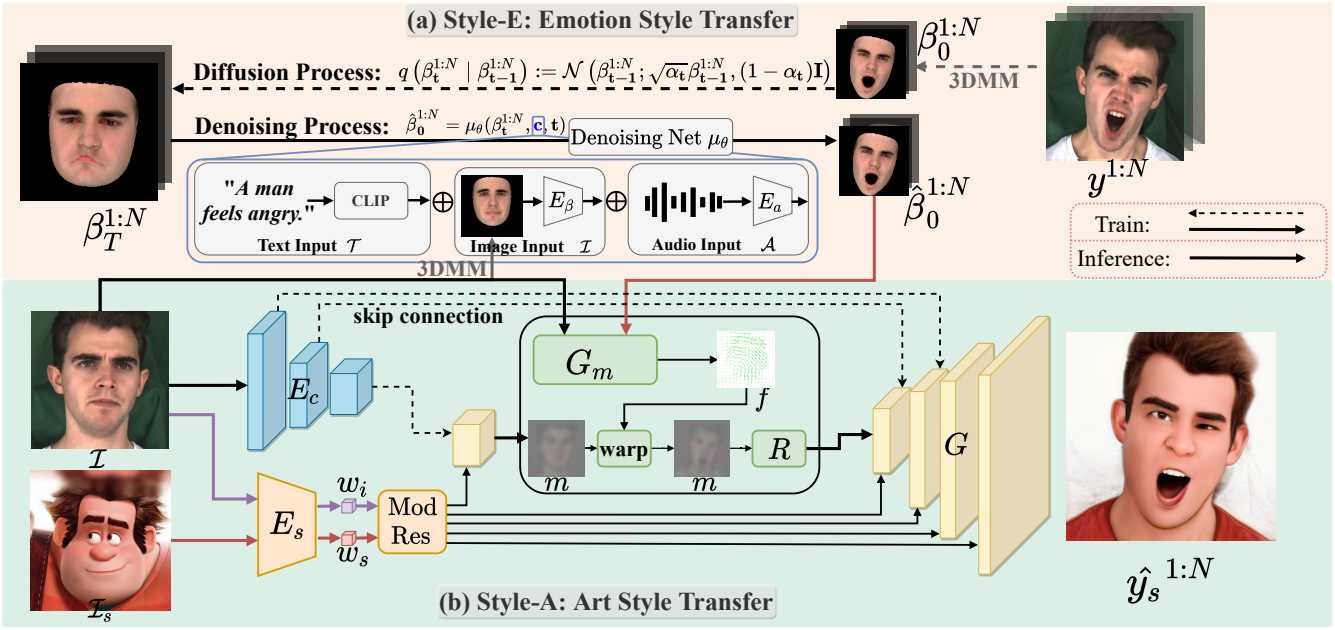

Figure 2: The overview of the proposed Style^{2}Talker. (a) Style-E: Emotion Style Transfer. In the diffusion process, we start by extracting the 3D expression coefficient sequence $\beta_0^{1:N}$ from the ground truth video $y^{1:N}$. Then, we iteratively add Gaussian noise by $q\left(\beta_{\mathbf{t}}^{1:N} \mid \beta_{\mathbf{t-1}}^{1:N}\right) := \mathcal{N}\left(\beta_{\mathbf{t-1}}^{1:N}; \sqrt{\alpha_{\mathbf{t}}}\beta_{\mathbf{t-1}}^{1:N}, (1-\alpha_{\mathbf{t}})\mathbf{I}\right)$. A simple MLP-based denoising network μ_θ is trained to denoise the noisy parameters $\beta_{\mathbf{t}}$ at time $\mathbf{t}$ based on the conditioning signal $\mathbf{c}$, and $\mathbf{c}$ comprises text $\mathcal{T}$, identity image $\mathcal{I}$ and audio $\mathcal{A}$. (b) Style-A: Art Style Transfer. We employ a pre-trained encoder E_s from pSp (Richardson et al. 2021) to embed the identity image $\mathcal{I}$ and art source image $\mathcal{I}_s$ to the latent code w_i and w_s, which are fed into a ModRes to merge the style code. To alleviate content loss caused by the pSp (Richardson et al. 2021), we introduce another Content Encoder E_c to extract multi-level content features. These features are then fed into StyleGAN G through skip connections and combined with style codes, serving as the input of G. To enable continuous frames generation, the produced emotional coefficients $\hat{\beta}_0^{1:N}$ are converted to flow field map f, which in turn warps feature map m of StyleGAN to $\hat{m}$, achieving talking head generation with style^2.

a straightforward and intuitive way to specify the emotions they want to see in the generated talking faces. Furthermore, we adopt a modified version of StyleGAN (Karras et al. 2020) to generate high-resolution talking face videos with the art style given an art reference image. This modification overcomes the issues of blurriness and low-resolution outputs encountered in existing methods.

Diffusion Generative Models

The field of Diffusion Generative Models (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) has witnessed a remarkable surge in advancements, showcasing impressive performance in conditional generation tasks (Dhariwal and Nichol 2021). Talking face generation has also seen extensions of diffusion models in several works (Stypukowski et al. 2023), where a general paradigm involves training U-net-like denoising network conditioned on time step and audio. However, a notable limitation of diffusion models lies in their extensive inference time, mainly due to the computationally intensive image-level denoising network and the involvement of multiple denoising processes. In light of this challenge, we extract the 3DMM coefficients from videos as brief intermediate representation and introduce a simpler and more efficient MLP-based denoising network (Du et al. 2023), which significantly reduce the inference time.

Face Manipulation by StyleGAN

StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020) has successfully proved its power to generate impressively realistic and high-resolution images. Various valuable modifications on StyleGAN have been explored for interesting applications, which fall into two main categories. The first category focuses on editing input face attributes, such as gender, pose, or age, within the latent space of StyleGAN (Richardson et al. 2021; Abdal, Qin, and Wonka 2019). Inspired by these advancements, Alghamdi et al. (2022) propose editing the latent space conditioned on audio, which enables the generation of talking head videos. On the other hand, the second category deals with editing the art style. Studies (Yang et al. 2022a,b) focus on transferring art style from an original image to a given style reference, often employing an extrinsic style path. However, their emphasis has primarily been on style transfer for individual frames. In contrast, our approach goes beyond these tasks by combining both attribute control and art style transfer to produce audio-driven artistically stylized talking head videos—an interesting but unexplored field.

Proposed Method

Figure 2 illustrates the pipeline of our Style^{2}Talker, which is composed of two stylized stages. In the Style-E stage, we leverage the diffusion model framework, incorporating both the diffusion and denoising processes. During the denoising process, the denoising network iteratively denoises the sampled random noise vector, generating an emotionally stylized coefficient sequence of the 3DMM model conditioned on the input text, image, and audio. In the Style-A stage, we introduce an elaborately modified StyleGAN to stylize the input face from the original art style to the given reference one. In subsequent sections, we will provide a detailed explanation of each stage within our proposed framework.

Data Preparation

To enhance user control flexibility, we enable text-guidance emotion style generation, wherein users can describe their desired emotion style using text. To realize this capability, a text-emotion audio-visual dataset is required. In comparison to the existing emotional audio-visual dataset like MEAD (Wang et al. 2020), the text-emotion audio-visual dataset additionally contains accurate text descriptions of emotions performed in the videos. Regrettably, such an open-source dataset does not currently exist.

While MEAD provides 8 general emotion descriptions, they may not be sufficient to accurately express the emotion style. Therefore, we devise a labor-free dataset generation pipeline that leverages large-scale pretrained models to extend MEAD with more detailed text descriptions of the emotion style, including semantic-rich adjectives and detailed facial motion descriptions represented by facial action units (AUs) (Ekman and Friesen 1978) following (Hong et al. 2020). We strongly recommend reading the detailed flowchart in the supplementary materials. Each AU corresponds to specific movements in different parts of the face, and combinations of multiple AUs describe the overall emotion style. Particularly, we employ OpenFace (Baltrusaitis et al. 2018) to detect continuous intensity for AUs, where we set an intensity threshold to determine AU activated/inactivated, and activated AUs are further divided into three discrete intensities as level labels. Subsequently, we turn to GPT-3 (Brown et al. 2020), a pretrained language model that encodes real-world knowledge, for providing more synonymous annotations for emotion classes and AUs with different levels. With the obtained textual descriptions of AU and emotion style, GPT-3 has the flexibility to generate several candidate sentences with varied syntaxes that provide comprehensive text descriptions of emotion styles for the videos in MEAD.

To filter out noisy samples from the generated emotion style text, we implement a data post-cleaning strategy by assessing the similarity between emotional videos in the MEAD dataset and each candidate textual description. The large-scale pretrained model CLIP (Radford et al. 2021) is well-suited for this task as its training strategy, where closer text descriptions to videos result in higher similarity. Therefore, we employ the text encoder and image encoder of CLIP to extract corresponding features from both the texts and frames of each video. The cosine similarity between the text feature from each candidate sentence and the image feature is calculated and ranked. Consequently, the five candidate sentences with the highest similarity for each video are retained as the final text descriptions. During training, we randomly select one sentence from these five candidates as the textual input for each iteration, ensuring a diverse and robust training process.

Style-E: Emotion Style Transfer

In this stage, we adopt the diffusion model (Song, Meng, and Ermon 2020) for producing stylized coefficient sequences from multiple inputs, given its impressive performance in conditional generation tasks. Starting with a stylized video $y_0^{1:N}$ containing N frames, we extract 3DMM (Blanz and Vetter 1999) expression coefficients $\beta_0^{1:N}$[1] using 3D reconstruction method (Deng et al. 2019). The forward diffusion process is formulated as follows:

$$q\left(\beta_{\mathbf{t}}^{1:N} \mid \beta_{\mathbf{t}-1}^{1:N}\right) := \mathcal{N}\left(\beta_{\mathbf{t}-1}^{1:N}; \sqrt{\alpha_t}\beta_{\mathbf{t}-1}^{1:N}, (1-\alpha_{\mathbf{t}})\mathbf{I}\right), \quad (1)$$

where $\mathbf{t} \in [1, ..., T]$ represents the step number of diffusion and $\alpha_{\mathbf{t}}$ denotes the noise schedule at $\mathbf{t}$-step.

During the denoising process, we train a denoising network μ_θ to generate an emotionally stylized motion sequence from random noise $\beta_T^{1:N} \sim \mathcal{N}(0, \mathbf{I})$. Intrigued by the success of diffusion prior in DALL·E2 (Ramesh et al. 2022), μ_θ directly predicts emotional motion sequence $\hat{\beta_{\mathbf{0}}^{1:N}} = \mu_\theta(\beta_{\mathbf{t}}^{1:N}, \mathbf{c}, \mathbf{t})$, instead of predicting the noise as in the vanilla DDPM (Ho, Jain, and Abbeel 2020). In our context, the condition $\mathbf{c} = \mathcal{T} \oplus \mathcal{I} \oplus \mathcal{A}$ concatenates multiple features extracted from a textual description $\mathcal{T}$ of emotion style, an identity image $\mathcal{I}$ and an audio clip $\mathcal{A}$ via corresponding encoders. The objective function is then formulated as:

$$L_{\text{style1}} = \mathbb{E}_{\beta_0^{1:N} \sim q\left(\beta_0^{1:N}\right), \mathbf{t} \sim [1:T]} \left[\left\| \beta_0^{1:N} - \hat{\beta}_0^{1:N} \right\|_2^2 \right]. \quad (2)$$

To address the inherent limitation of slow inference time in the diffusion model (Ho, Jain, and Abbeel 2020), we implement μ_θ with a more lightweight and efficient MLP-based architecture following Du et al. (2023). Furthermore, we leverage the DDIM (Song, Meng, and Ermon 2020) technique, which allows us to sample only 5 steps instead of 1000 during inference, which contributes to a substantial decrease in inference time.

Style-A: Art Style Transfer

To perform high-resolution talking heads with art style, we build our framework upon DualStyleGAN (Yang et al. 2022a), which introduces a new extrinsic art style path and ModRes block in comparison to vanilla StyleGAN (Karras et al. 2020) to artistically stylize a single image. Specifically, given an identity image $\mathcal{I}$ and an art style reference image $\mathcal{I}_s$, a GAN Inversion encoder E_s (Richardson et al. 2021) is employed to encode them into latent code w_i, w_s,

[1]Please note that β in our paper refers to expression coefficients of 3DMM, instead of hyper-parameter in original DDPM paper.

Method	**MEAD** (Wang et al. 2020)					**HDTF** (Zhang et al. 2021)				
	SSIM↑	FID↓	M-LMD↓	F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	SSIM↑	FID↓	M-LMD↓	F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑
VT+MakeItTalk	0.692	82.577	6.696	5.948	0.734	0.630	50.009	6.907	6.279	0.857
VT+Wav2Lip	0.700	129.893	6.153	5.465	**3.663**	0.656	48.888	6.279	5.662	2.479
VT+Audio2Head	0.660	75.253	9.032	9.856	1.772	0.588	48.91	7.087	6.930	2.559
VT+PC-AVS	0.624	148.015	14.250	12.758	2.664	0.436	110.108	9.678	13.321	2.545
VT+AVCT	0.632	63.222	12.461	11.355	2.841	0.583	44.616	12.005	10.739	**3.515**
VT+EAMM	0.690	73.167	6.541	6.247	1.801	0.555	65.048	7.771	7.872	2.355
VT+StyleTalk	0.726	91.661	4.343	4.696	1.987	0.641	50.974	6.268	6.404	2.461
Style^{2}Talker	**0.795**	**23.207**	**3.317**	**2.696**	2.847	**0.718**	**23.330**	**3.046**	**2.791**	2.734
GT+VT	1.000	0.000	0.000	0.000	2.985	1.000	0.000	0.000	0.000	2.262

Table 1: Quantitative comparisons with state-of-the-art methods. We evaluate each method on MEAD and HDTF datasets. For assessing the art style, we apply VToonofy to stylize the input frame, denoted as VT'. 'Style-E' and 'Style-A' refers to emotion and art style. The symbols '↑' and '↓' indicate higher and lower metric values for better results, respectively.

which provide content information and art style information, respectively. To effectively merge such two information and preserve the generative space and behavior of the pre-trained StyleGAN G simultaneously, the ModRes block is introduced to adjust the structure styles of w_i based on w_s in a residual manner, enabling the pre-trained StyleGAN G to transfer the art style of the identity image $\mathcal{I}$ into that of $\mathcal{I}_s$ with high resolution. However, despite its success in art style transfer, this approach faces challenges when generating continuous talking head videos under the guidance of the predicted motion sequence $\hat{\beta}_0^{1:N}$. Additionally, the stylized images suffer from image detail loss due to the performance limitation of GAN Inversion (Wang et al. 2022b).

Sparked by the observation (Yin et al. 2022) that spatial feature map in 64×64 layer reflects the pose and expression of the generated image, we introduce a motion generator G_m to produce a flow field f from the spatial feature map m to the desired feature map $\hat{m}$. The motion generator G_m consists of an image encoder, a flow decoder and a coefficient encoder. Specifically, for each $\hat{\beta}_0^t \in \hat{\beta}_0^{1:N}$, the image encoder encodes identity image $\mathcal{I}$ into multiple feature maps x_i using convolutional layers. These feature maps are then passed through an adaptive instance normalization (AdaIN (Huang and Belongie 2017)) operations as: $\text{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i}\frac{\mathbf{x}_i - \nu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i}$, where $\nu(\cdot)$ and $\sigma(\cdot)$ represent the average and variance operations. $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$ are generated from the predicted 3DMM coefficients $\hat{\beta}_0^t$ through the coefficient encoder. By incorporating the flow decoder, we obtain the flow field f and use it to warp the spatial feature map m to $\hat{m}$, whose expressions and poses align with predicted emotion coefficients.

As for the second problem above, we adopt a Content Encoder E_c (Yang et al. 2022b) to obtain multi-scale content features, which are passed to the StyleGAN G to supplement the texture details via skip connections. By including additional multi-scale identity features, we can effectively preserve the image details of the original frame compared to previous StyleGAN-based methods that solely rely on the style condition, enabling high-fidelity artistically stylized images. Please note that since the skip connection passes texture information to the layer behind $\hat{m}$ as shown by the dotted line in Figure 2, and the texture information is extracted from the original image and is not aligned with $\hat{m}$, ghosts will inevitably appear in the output. To this end, we construct a Refinement Network R to adaptively tune $\hat{m}$ to rectify these artifacts. With the refined spatial feature maps, G subsequently generates a continuous sequence of stylized frames $\hat{y_s}^{1:N}$, achieving talking head generation with style^2.

During Style-A training, we freeze the weights of E_s and G which are pretrained in DualStyleGAN, and optimize the remaining networks (i.e., E_c, G_m and R). To obtain the ground truth for training, we employ VToonify (Yang et al. 2022b) to produce artistically stylized frames $y_s^{1:N} = \text{VToonify}(y^{1:N}, I_s)$ from video $y^{1:N}$. Concretely, we import reconstruction loss L_{rec} and perceptual loss L_{prec} (Johnson, Alahi, and Fei-Fei 2016) to constrain the networks.

$$L_{\text{style2}} = \underbrace{\left\| y_s^t - \hat{y_s^t} \right\|_2}_{L_{\text{rec}}} + \underbrace{\lambda \left\| \text{VGG}(y_s^t) - \text{VGG}(\hat{y_s^t}) \right\|_1}_{L_{\text{prec}}}, \quad (3)$$

where VGG implies the pretrained VGG network (Simonyan and Zisserman 2015), and $\lambda = 0.1$ refers to the weight of L_{prec}. Moreover, discriminator D is applied to further enhance the realism of generated frames $\hat{y_s}$.

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{y_s}[\log D(y_s)] + \mathbb{E}_{\mathcal{I},\mathcal{I}_s,\beta}[\log(1 - D(\hat{y_s}))] \quad (4)$$

Experiments

Experimental Settings

Datasets For the Style-E stage, we leverage MEAD dataset (Wang et al. 2020) with the synthetically generated textual descriptions for emotion styles. MEAD contains videos and audios pairs performed by 60 actors in 8 emotions. Due to the limited size of MEAD, we enhance the one-shot talking motion generation performance by borrowing the pretrained audio encoder from SadTalker (Zhang et al. 2022). For the Style-A stage, we additionally utilize another audio-visual dataset HDTF (Zhang et al. 2021), which consists of talking videos from more than 300 speakers. To obtain the art style reference, we use various art datasets (Huo

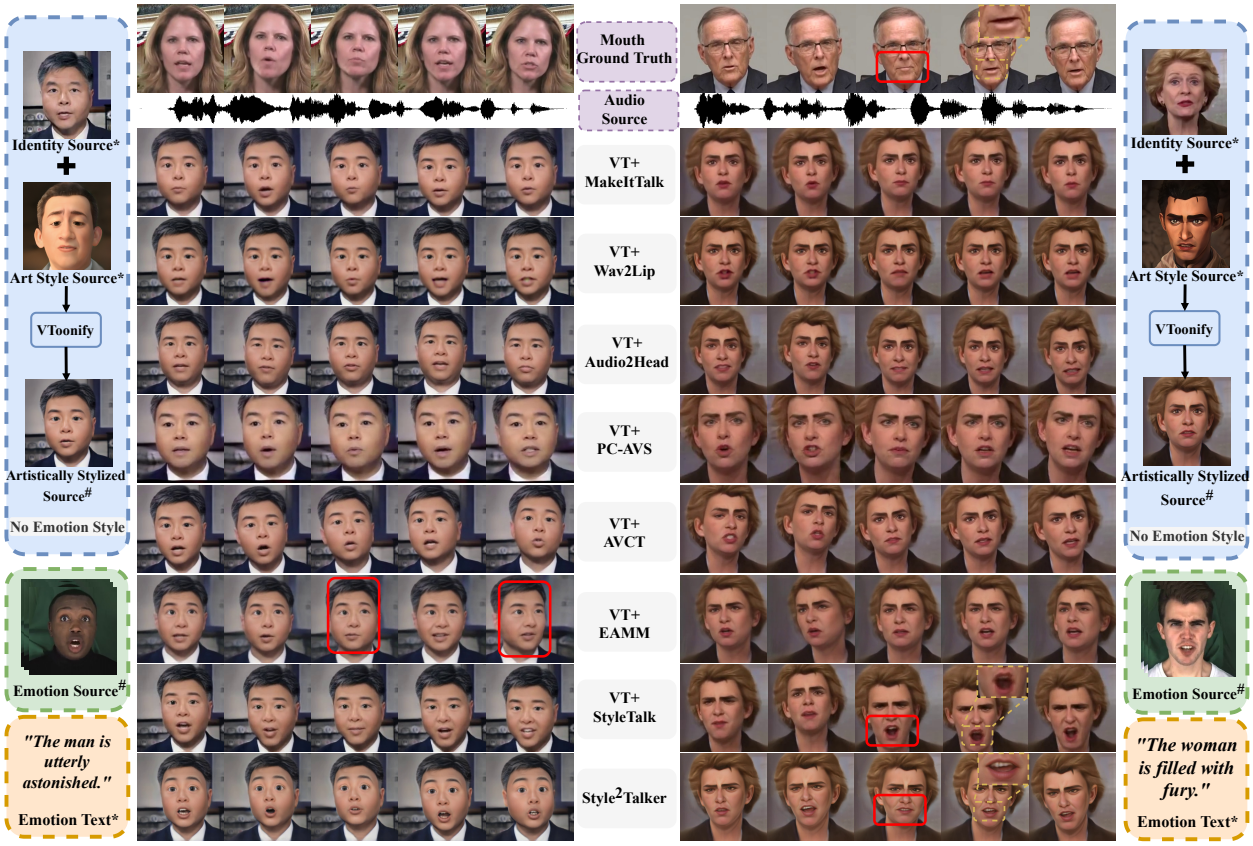

Figure 3: Qualitative comparisons with state-of-the-art methods. The input for our Style^{2}Talker and SOTAs are marked by '*' and '#', respectively. We keep the original color style for better comparison.

et al. 2017a,b). As for the ground truth of Style-A, we employ VToonify (Yang et al. 2022b) to stylize videos in MEAD and HDTF with randomly selected art style $\mathcal{I}_s$.

Comparison Setting. To the best of our knowledge, there is currently no existing work that can generate high-resolution audio-driven talking face videos with both emotion style and art style from a real image. To provide a comprehensive comparison, we use VToonify to transfer the art style of the input identity image into the desired one. Then, we pass the stylized images through several state-of-the-art (SOTA) talking face generation methods to achieve the same task as our proposed method. The comparing methods include MakeItTalk (Zhou et al. 2020), Wav2Lip (Prajwal et al. 2020), Audio2Head (Wang et al. 2021), PC-AVS (Zhou et al. 2021), AVCT (Wang et al. 2022a), EAMM (Ji et al. 2022) and StyleTalk (Ma et al. 2023b), where only the latter two methods support talking head generation with emotion style. We assess the results using evaluation metrics including SSIM (Wang et al. 2004), FID (Heusel et al. 2017) and PSNR for image generation quality, M-LMD (Chen et al. 2019) for accuracy evaluation of lip movement, F-LMD (Chen et al. 2019) for emotion style evaluation. In addition, we calculate $\text{Sync}_{\text{conf}}$ (Chung and Zisserman 2016) to measure the synchronization of lip motion with input audio.

Experimental Results

Quantitative Results. Table 1 reports the quantitative results of our method and other SOTA methods. For emotional talking face generation methods (i.e., EAMM and StyleTalk), we additionally provide the emotion videos as emotion resources, while for our Style^{2}Talker, we use emotion texts as input. Besides, we synthesize high-resolution (1024×1024) talking face videos with both emotion style and art style. As observed, our method outperforms other methods in most of the evaluation metrics on both MEAD and HDTF datasets and achieves a suboptimal score of $\text{Sync}_{\text{conf}}$, ranking second only to Wav2Lip and AVCT on the MEAD and HDTF datasets, respectively. We argue that this is primarily due to Wav2Lip being trained with the assistance of a SyncNet discriminator, and AVCT receiving an additional phonemes input to enhance the audio-visual correlation. Nevertheless, our lowest M-LMD scores on both datasets demonstrate the satisfactory synchronization between audio and the lip shapes generated by our method.

Qualitative Results. Figure 3 presents the input for comparison methods marked by '#', the input for ours marked by '*' and the qualitative results. Our Style^{2}Talker achieves the best lip synchronization and both emotion style and art style transfer in high resolution directly from the real face. Specif-

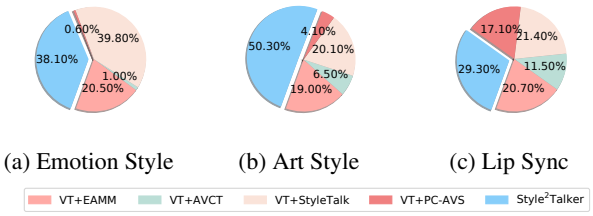

Figure 4: User study results.

Method/Score	SSIM ↑	FID ↓	M/F-LMD ↓	$\text{Sync}_{\text{conf}}$ ↑
w/o emo style	0.760	31.655	3.269/ 2.913	**3.762**
w/o art style	0.743	58.119	**3.132** / 2.897	2.477
w/o dif model	0.789	24.685	4.339 / 3.712	2.521
w/o skip con.	0.763	42.105	3.868 / 3.082	2.567
w/o R	0.754	35.530	3.557 / 2.882	2.749
Full model	**0.795**	**23.207**	3.317 / **2.696**	2.847

Table 2: Results for ablation study.

ically, MakeItTalk struggles to generate accurate lip motions and Wav2Lip suffers from blurriness in the mouth region. Audio2Head and PC-AVS encounter issues with identity information loss. Despite the progress achieved by AVCT, it neglects the emotion style. While EAMM and StyleTalk take an emotion source video for emotion style reference which provides more concrete style information than textual description used in our method (Ma et al. 2023a), we perform more expressive emotion style than EAMM, which appears server facial deformation as circled in the red boxes. In comparison to StyleTalk, we achieve competitive emotion style performance. However, whe n silent, their mouths remain open (pointed by red boxes), which leads to unnatural lips and noticeable artifacts (pointed by yellow boxes).

User Study. We conduct user studies to compare our method with SOTAs in terms of human likeness. Specifically, we invite 20 participants (10 males + 10 females) and each participant is presented with 5 videos generated by 5 different methods (2 SOTA methods each without/with emotion style and Style^{2}Talker) per iteration. Participants are required to choose from among them the one in which they believe the **emotion style** best matched the provided text/the **art style** best matched the given picture/the **lip motion** best matched the audio, and such a process is repeated 20 iterations. The results, as depicted in Figure 4, indicate that our method received the most preferences for art style and lip synchronization aspects and competitive likeness to StyleTalk in emotion style, which inputs a more informative emotional video.

Ablation Study. We further conduct ablation experiments on MEAD dataset to assess the effectiveness of each introduced component. The qualitative and quantitative results are presented in Figure 5 and Table 2. In general, the experiment settings and corresponding analyses are summarized as: **(a)** w/o emotion style: we exclude the input of emotion source text $\mathcal{T}$ and the text encoder of CLIP. As a result, the expressions in the generated videos are consistent with the

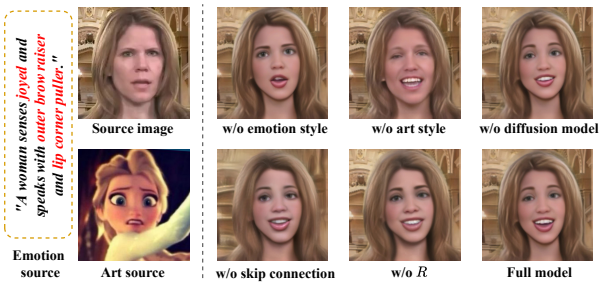

Figure 5: Visualization Results of ablation study.

identity image, and the F-LMD score decreases accordingly. This indicates that our method is capable of comprehending and incorporating the emotion style conveyed in the text description. **(b)** w/o art style: we build our model upon the vanilla StyleGAN instead of DualStyleGAN by removing the art path and ModRes block. In this case, the art style remains unchanged, but the FID score significantly increases due to the different art style from the ground truth. This confirms that the art path and ModRes block are crucial for maintaining consistent and visually appealing art style transfer. **(c)** w/o diffusion model: we replace the diffusion model with an conditional GAN (Mirza and Osindero 2014) in the Style-E stage, which results in worse lip synchronization. This observation illustrates that the diffusion model achieves superior conditional generation performance. **(d)** w/o skip connection and **(e)** w/o R: we remove middle-level skip connection and refinement network R, respectively. As shown in Figure 5, the detail disparity between our full model and the source identity image is smaller than the disparity of **(d)** w/o skip connection. Furthermore, our full model exhibits higher fidelity with the source image than **(e)** w/o R, demonstrating that the skip connections and R effectively enhance the image quality and maintain better visual coherence.

Conclusion

In this paper, we present Style^{2}Talker, a novel system that generates high-resolution emotionally and artistically stylized talking face videos by incorporating corresponding style prompts. Leveraging a labor-free text annotation pipeline based on large-scale pretrained models, we obtain textual descriptions for emotion style learning from text inputs. We aspire for our attempt to inspire further in-depth research, employing outstanding large-scale pretrained models for more practical and captivating explorations. To infuse emotion style into the 3D motion coefficients, we devise an efficient diffusion model with multiple encoders, ensuring the generation of realistic and expressive facial expressions. We incorporate a motion-driven module and an additional art style path into the StyleGAN architecture, enabling coefficient-driven video generation with desired emotion and art styles. To further enhance the visual quality and eliminate artifacts, we employ a content encoder and refinement network. Qualitative and quantitative experiments demonstrate that our method can generate more stylized animation results compared with state-of-the-art methods.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC, NO. 62102255) and Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102). We would like to thank Xinya Ji, Yifeng Ma and Zhiyao Sun for their generous help.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4432–4441.

Alghamdi, M. M.; Wang, H.; Bulpitt, A. J.; and Hogg, D. C. 2022. Talking Head from Speech Audio using a Pre-trained Image Generator. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5228–5236.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 59–66. IEEE.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7832–7841.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Chung, J. S.; and Zisserman, A. 2016. Out of Time: Automated Lip Sync in the Wild. *Springer International Publishing eBooks*.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. 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*, 0–0.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Neural Information Processing Systems*.

Du, Y.; Kips, R.; Pumarola, A.; Starke, S.; Thabet, A.; and Sanakoyeu, A. 2023. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 481–490.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P.; and Rosenberg, E. L. 2005. What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (FACS).

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5784–5794.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Neural Information Processing Systems,Neural Information Processing Systems*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Neural Information Processing Systems*.

Hong, J.; Lee, H. J.; Kim, Y.; and Ro, Y. M. 2020. Face tells detailed expression: Generating comprehensive facial expression sentence through facial action units. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, 100–111. Springer.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Huo, J.; Gao, Y.; Shi, Y.; and Yin, H. 2017a. Variation Robust Cross-Modal Metric Learning for Caricature Recognition. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*.

Huo, J.; Li, W.; Shi, Y.; Gao, Y.; and Yin, H. 2017b. WebCaricature: a benchmark for caricature recognition. *Computer Vision and Pattern Recognition,Computer Vision and Pattern Recognition*.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C. C.; Cao, X.; and Xu, F. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14080–14089.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, 694–711.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Ma, Y.; Wang, S.; Ding, Y.; Ma, B.; Lv, T.; Fan, C.; Hu, Z.; Deng, Z.; and Yu, X. 2023a. TalkCLIP: Talking Head Generation with Text-Guided Expressive Speaking Styles. *arXiv preprint arXiv:2304.00334*.

Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; and Yu, X. 2023b. StyleTalk: One-shot Talking Head Generation with Controllable Speaking Styles. *arXiv preprint arXiv:2301.01081*.

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

Nichol, A.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. *Cornell University - arXiv*.

Pan, Y.; Zhang, R.; Cheng, S.; Tan, S.; Ding, Y.; Mitchell, K.; and Yang, X. 2023. Emotional Voice Puppetry. *IEEE Transactions on Visualization and Computer Graphics*, 29(5): 2527–2535.

Pataranutaporn, P.; Danry, V.; Leong, J.; Punpongsanon, P.; Novy, D.; Maes, P.; and Sra, M. 2021. AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12): 1013–1022.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2287–2296.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations,International Conference on Learning Representations*.

Sinha, S.; Biswas, S.; Yadav, R.; and Bhowmick, B. 2021. Emotion-Controllable Generalized Talking Face Generation. In *International Joint Conference on Artificial Intelligence*. IJCAI.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Stypukowski, M.; Vougioukas, K.; He, S.; Zieba, M.; Petridis, S.; and Pantic, M. 2023. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. *arXiv preprint arXiv:2301.03396*.

Tan, S.; Ji, B.; and Pan, Y. 2023. EMMN: Emotional Motion Memory Network for Audio-driven Emotional Talking Face Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22146–22156.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, 700–717. Springer.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. In *International Joint Conference on Artificial Intelligence*. IJCAI.

Wang, S.; Li, L.; Ding, Y.; and Yu, X. 2022a. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2531–2539.

Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022b. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11379–11388.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*.

Yang, S.; Jiang, L.; Liu, Z.; and Loy, C. C. 2022a. Pastiche master: exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7693–7702.

Yang, S.; Jiang, L.; Liu, Z.; and Loy, C. C. 2022b. VToonify: Controllable High-Resolution Portrait Video Style Transfer. *ACM Transactions on Graphics (TOG)*, 41(6): 1–15.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*.

Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. StyleHEAT: One-shot high-resolution editable talking face generation via pre-trained StyleGAN. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 85–101. Springer.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2022. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *arXiv preprint arXiv:2211.12194*.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

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