This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Style Transfer for 2D Talking Head Generation

Trong Thang Pham[2], Tuong Do[1,3], Nhat Le[1],
Ngan Le[2], Hung Nguyen[1], Erman Tjiputra[1], Quang Tran[1], Anh Nguyen[3]
[1]AIOZ, Singapore, [2]University of Arkansas, USA, [3]University of Liverpool, UK

Abstract

Audio-driven talking head animation is a challenging research topic with many real-world applications. Recent works have focused on creating photo-realistic 2D animation, while learning different talking or singing styles remains an open problem. In this paper, we present a new method to generate talking head animation with learnable style references. Given a set of style reference frames, our framework can reconstruct 2D talking head animation based on a single input image and an audio stream. Our method first produces facial landmarks motion from the audio stream and constructs the intermediate style patterns from the style reference images. We then feed both outputs into a style-aware image generator to generate the photo-realistic and fidelity 2D animation. In practice, our framework can extract the style information of a specific character and transfer it to any new static image for talking head animation. The intensive experimental results show that our method achieves better results than recent state-of-the-art approaches qualitatively and quantitatively. Our source code will be made publicly available.

1. Introduction

Talking head animation is an active research topic in both academia and industry. This task has a wide range of real-world interactive applications such as digital avatars [10], and digital animations [26]. Given an arbitrary input audio and a 2D image (or a set of 2D images) of a character, the goal of talking head animation is to generate photo-realistic frames. The output can be the 2D [17, 31, 51] or 3D talking head [10, 50]. With recent advances in deep learning, especially generative adversarial networks [15], several works have addressed different aspects of the talking head animation task such as head pose control [49], facial expression [25], emotion generation [20], and photo-realistic synthesis [6, 41, 51].

While there has been considerable advancement in the generation of talking head animation, achieving photo-realistic and fidelity animation is not a trivial task. It is even more challenging to render natural motion of the head with different styles [10]. In practice, several aspects contribute to this challenge. First, generating a photo-realistic talking head using only a single image and audio as inputs requires multi-modal synchronization and mapping between the audio stream and facial information [11]. In many circumstances, this process may result in fuzzy backgrounds, ambiguous fidelity, or abnormal face attributes [51]. Second, various talking and singing styles can express diverse personalities [43]. Therefore, the animation methods should be able to adapt and generalize well to different styles [43]. Finally, controlling the head motion and connecting it with the full-body animation remains an open problem [21].

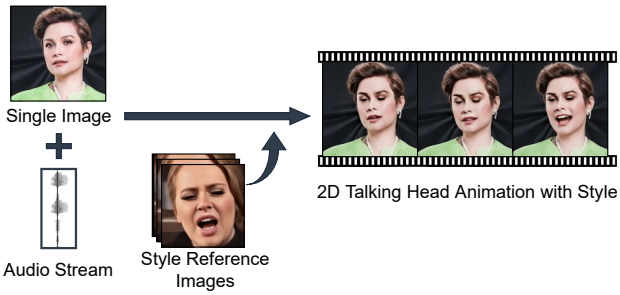


Figure 1. Given an audio stream, a single image, and a set of style reference frames, our method generates realistic 2D talking head animation.

Recently, several methods have been proposed to generate photo-realistic talking heads [17, 31, 46, 51] or to match the pose from a source video [49] while little work has focused on learning the personalized character style [31]. In practice, apart from personalized talking style, we have different singing styles such as `ballad` and `rap`. These styles pose a more challenging problem for talking head animation as they have the unique eye, head, mouth, and torso motion. The facial movements of singing styles are also more varied and dynamic than the talking style [37]. Therefore, learning and bringing these styles into 2D talking heads is more challenging. Currently, most of the style-aware talking head animation methods do not fully disentangle the audio style information and the visual informa-

tion, which causes ambiguity during the transferring process [31].

In this work, we present a new deep learning framework called Style Transfer for 2D talking head animation. Our framework provides an effective way to transfer talking or singing styles from the style reference to animate single 2D portrait of a character given an arbitrary input audio stream. We first generate photo-realistic 2D animation with natural expression and motion. We then propose a new method to transfer the personalized style of a character into any talking head with a simple style-aware transfer process. Figure 1 shows an overview of our approach.

In summary, our contributions are as follows:

- We propose a new framework for generating photo-realistic 2D talking head animations from the audio stream as input.
- We present a style-aware transfer technique, which enables us to learn and apply any new style to the animated head. Our generated 2D animation is photo-realistic and high fidelity with natural motions.
- We conduct intensive analysis to show that our proposed method outperforms recent approaches qualitatively and quantitatively.

2. Literature Review

2D Talking Head Animation. Creating talking head animation from an input image and audio has been widely studied in the past few years. One of the earliest works [4] considered this as a sorting task that reorders images from footage video. Based on [4], [14] proposed to capture 3D model from dubber and actor to synthesize photo-realistic face. [13] introduced a learning approach to create a trainable system that could synthesize a mouth shape from an unseen utterance. Later works focused on audio-driven to generate realistic mouth shapes [34, 39] or realistic faces [5, 48]. The authors in [12] generated full facial landmarks using the input audio. [45] created a talking face that includes pose and expression. Instead of creating talking face, [16] designed a model that produces head motion from the joint latent space using Bi-LSTM. [7, 24, 27] created realistic head avatars. [19] focused on generating fidelity talking head with natural head pose and photo-realistic motions. Recently, [31] proposed to generate photo-realistic talking head with personalized information encoded, or [38] took advantage of the diffusion model to improve the diversity of the generated talking face.

Speaker Style Estimation. There are many kinds of speaker styles such as generic, personal, controlled pose, or special expression. Generic style could be learned by training on multiple videos [51], while personalized style could be decided by training on one avatar particularly [31]. In [49], the authors introduced a method that generates controllable poses with an input video. [45] transferred poses and expressions from another video input. [37] mapped the style from dubber to actor. [44] captured motions from the driven video and transferred them into input image during the generation process. [40] tried to ensemble speaker and speaking environment to characterize the speaker variability in the environment. [29] leveraged a pre-captured database of 3D mouth shapes and associated speech audio from one speaker to refine the mouth shape of a new actor. Recently, [47] developed a method that can generate diverse and synchronized talking videos from input audio and a single reference image by utilizing condition variational autoencoder to caption style code.

Speech Representation for Face Animation. Some prior works used hand-crafted models to match phoneme and mouth shape in each millisecond audio signal as speech representation [4, 50]. DeepSpeech [18] paved the way for learning a speech recognition system using an end-to-end deep network. Following that, [16] trained Bi-LSTMs to learn a language-long-term structure that models the relationship between speech and the complex activity of faces. [39] used Mel-frequency spectral coefficients to synthesize high-quality mouth texture of a character, and then combined it with a 3D pose matching method to synchronize the lip motion with the audio in the target animation. With the rise of the diffusion technique, [33] proposed an audio-conditional diffusion model that effectively encodes audio in their generator to solve the lip-sync challenge.

Our goal is to introduce a new deep-learning framework that can transfer talking or singing styles from any personalized style reference to animate a single 2D portrait of a character given an arbitrary input audio stream. Compared to existing approaches, which mainly focus on conventional talking head animation, our method can not only produce animation for common talking styles but also allows transferring for several special styles that are much more challenging such as singing.

3. Style-Aware Talking Head Generator

3.1. Audio Stream Representation and Motion Generator

Audio Stream Representation. Representing audio for learning is essential in a talking head generator. Different speaking styles, referred to as individualized styles, can present challenges when using deep speech representation directly, resulting in suboptimal outcomes, particularly when dealing with distant variations in speech features. To improve the generalization of the audio stream extractor, we incorporate Lu et al.'s manifold projection technique [31].

Motion generator. Given the extracted audio features, this step generates audio-driven motions in our framework. In practice, the character's style is mainly defined by the mouth, eye, head, and torso movement. Therefore, we con-

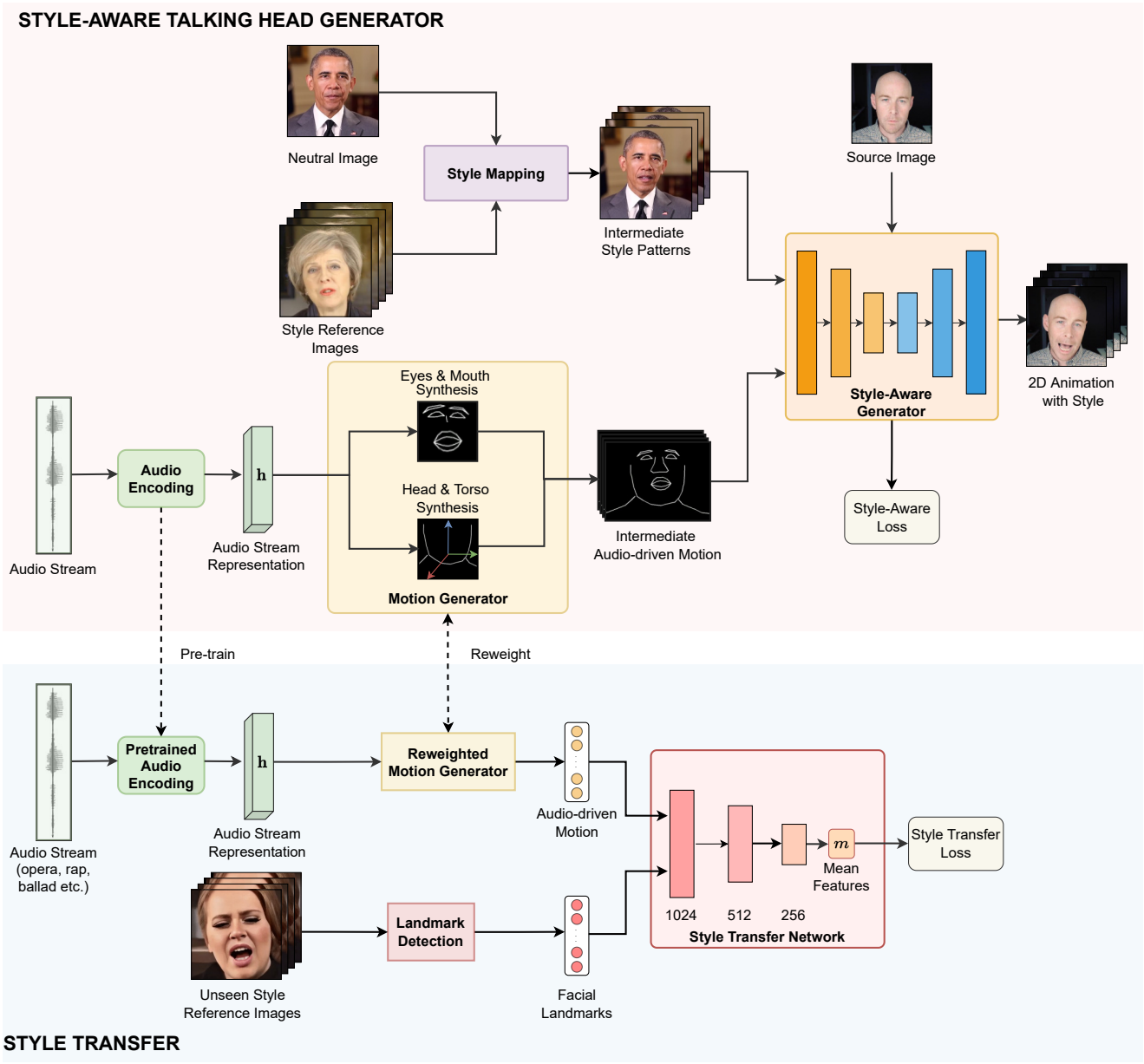


Figure 2. A detailed illustration of our method.

sider the motion around these regions of the face in our work.

3.2. Style Reference Images

To learn the character's styles more effectively, we define the Style Reference Images as a set of images retrieved from a video of a specific character by using the key motion templates. Inspired by [31], [51], and music theory about rhythm [3], we use four key motion templates that contain popular motion range and behavior. Each behavior is then plotted as a reference style pattern, which is used to retrieve the ones that are most similar in each video in the dataset. To retrieve similar patterns, we apply similarity search [8] for each image in the video of the character. The result image set is called the Style Reference Images and is used to provide character's styles information in our framework.

3.3. Style Mapping

The Style Mapping is designed to disentangle the style in the reference images and then map the extracted style to the neutral image. Then, the input of this module is a pair of two images: a neutral image $\boldsymbol{I}_s$, and a style reference image

$\boldsymbol{I}_r$. The output is an Intermediate Style Pattern (ISP - an image) which has the identity that comes from $\boldsymbol{I}_s$ and the style represented in $\boldsymbol{I}_r$. ISP has the visual information of the neutral image but the style is from the style reference image. In practice, we first disentangle the style information encoded in the pose and expression of both the neutral and reference image, then map the style from the reference image into the neutral image to generate the output ISP image $\boldsymbol{I}_o$.

Disentangling Neutral Image. Since the head pose, expression, and keypoints from the neutral image contain the style information of a specific character, they need to be disentangled to learn the style information. In this step, given an input image $\boldsymbol{I}_s$, a set of k number of keypoints $\mathrm{c_k}$ is first disentangled to store the geometry signature via a Keypoint Extractor network. Then, we extract the pose, which is parameterized by a translation vector $\tau \in \mathbb{R}^3$ and a rotation matrix $\mathrm{R} \in \mathbb{R}^{3\times3}$, and expression information ε_k from the image by a Pose Expression network. After the disentangling process, we can reconstruct the image keypoints $\mathrm{C_k}$ using Equation 1. The extracted keypoints maintain the geometry signature and style information of the head in the neutral image.

$$\mathrm{C_k} = \mathrm{c_k} \times \mathrm{R} + \tau + \varepsilon_\mathrm{k} \tag{1}$$

Disentangling Style Reference Image. Similar to the neutral image, we use two deep networks to disentangle and extract the head pose and keypoints from the style reference image. However, instead of extracting new keypoints from the reference images, we reuse the extracted ones $\mathrm{c_k}$ from the neutral image, which contains the identity-specific geometry signature of the neutral image. The final keypoints $\bar{\mathrm{C}}_k$ of the style reference image are computed in Equation 2:

$$\bar{\mathrm{C}}_k = \mathrm{c_k} \times \bar{\mathrm{R}} + \bar{\tau} + \bar{\varepsilon}_\mathrm{k} \tag{2}$$

where $\bar{\tau} \in \mathbb{R}^3$, $\bar{\mathrm{R}} \in \mathbb{R}^{3\times3}$ and $\bar{\varepsilon}$ are translation vector, rotation matrix, and expression information extracted from the style reference image, respectively.

Style Mapping. To construct the Intermediate Style Pattern $\boldsymbol{I}_o$, we first extract two keypoints sets $\mathrm{C_k}$ and $\bar{\mathrm{C}}_k$ from the neutral image and the style reference image. We then estimate the warping function based on the two keypoints sets to warp the encoded features of the source (neutral image) to the target so that it can represent the style of the reference image. Then, we feed the warped version of the source encoded features and the extracted style information into an Intermediate Generator to obtain the ISP image. In practice, we choose the neutral image as a general image in Obama Weekly Address dataset [39], while the style reference image is one of the four images in the Style Reference Images set. By applying the style mapping process for all four images in the Style Reference Images, we obtain a set of four ISP images. This set (the Intermediate Style Pattern - ISP) is used as the input for the Style-Aware Generator in Section 3.4.

3.4. Style-Aware Generator

This module generates a 2D talking head from a source image, the generated intermediate motion, and the style information represented in the Intermediate Style Pattern. In this module, the facial map plays an essential role in explicitly identifying groups of facial keypoints, which makes the style-aware learning process easier to converge. In our experiment, the facial map has the size of 512×512 and can be obtained by connecting consecutive keypoints in a preset semantic sequence and projecting it onto the 2D image plane using a pre-computed camera matrix.

Style-Aware Loss. To learn the style-aware loss, we introduce the style-aware photometric loss $\mathcal{L}_{\mathrm{sp}}$. This loss is combined with the generator loss $\mathcal{L}_{\mathbf{G}}$ to improve the generation quality and penalize the generated output that has a high deviation from the reference style patterns. The style-aware photometric loss is formulated as the pixel-wise error between the generated image $\boldsymbol{I}'$ and the matched style pattern image $\boldsymbol{I}_{\mathrm{m}}$:

$$\mathcal{L}_{\mathrm{sp}} = \|\mathbf{W} \odot (\boldsymbol{I}' - \boldsymbol{I}_{\mathrm{m}})\|_1 \tag{3}$$

where $\mathbf{W}$ is the weighting mask which has values depending on different face regions; $\odot$ denotes the Hadamard product; the matched style pattern image $\boldsymbol{I}_{\mathrm{m}}$ is obtained by using [8] to retrieve the best-matched image corresponding to one of the style reference images. To acquire $\mathbf{W}$, we first use an off-the-shelf face parsing method to generate the segmentation mask of the face [28]. To achieve high fidelity image generation, we want the network to focus more on each facial region. Specifically, the corresponding weight of $\mathbf{W}$ according to mouth, eyes, and skin regions are set to $5.0, 3.0, 1.0$, respectively. Note that weights for other regions in the weighting mask $\mathbf{W}$, e.g. background, are set to 0.

4. Style Transfer

The style transfer phase focuses on transferring the styles to a new character by re-weighting the Motion Generator given the input audio. In our transferring phase, we assume that the talking or singing styles are encoded in both the audio stream and reference images. Therefore, this style information is learnable and can be transferred from one to another character. As in [1], we mainly rely on the pre-trained models from the training phase to perform the style transfer. Since Style-Aware Generator can cover the visual information generated from different styles, our goal in this phase is to make sure the style encoded in the Intermediate Audio-driven Motions can be adjusted to different styles rather than just the neutral one (i.e., the styles in the train-

ing data). We capture both the audio stream and reference images as the input in this stage. See Figure 2 for the details of our style transfer process.

Given the reference images and an audio stream (e.g., `opera`, `rap`, etc.), we first use the pre-trained audio encoding to extract the audio feature and apply the Motion Generator to reconstruct the audio-driven motion ϕ_{mg}. The reference images are fed through a pre-trained landmark detector to extract their corresponding facial landmarks ϕ_{s}. The generated motions and facial landmarks are vectorized into (68×3)-dimensional vectors. Both ϕ_{mg} and ϕ_{s} are then passed through a style transfer network to extract the mean features. A style transfer loss $\mathcal{L}_{\text{transfer}}$ is then optimized through back-propagation. The mean features are the latent encoded vector containing both information from the audio-driven landmarks and the facial landmarks.

4.1. Style Transfer Network

The style transfer network $f(\cdot)$ aims to learn the differences between motions of the input reference images and audio-driven motions extracted from the Motion Generator. Thanks to the style transfer loss $\mathcal{L}_{\text{transfer}}$, the network is optimized to lower the gap of both mentioned motions, and then re-weight the parameters of Motion Generator to generate output motions that is similar to the target style. After re-weighting, the Motion Generator can produce style-aware audio-driven motions which are then passed into Style-Aware Generator to generate 2D animation with style. The style transfer network has three multilayer perceptrons (MLP), each MLP layer has 1024, 512, and 256 neurons, subsequently. The final layer produces the mean features used in the style transfer loss.

4.2. Style Transfer Loss

The style transfer loss is proposed to ensure the generated motions take into account the target style. This loss is incooperated with the Motion Generator loss $\mathcal{L}_{\text{mg}}$ for fine-tuning the Motion Generator module during the transferring process. The style transfer loss $\mathcal{L}_{\text{transfer}}$ is contributed by the constraint loss $\mathcal{L}_{\text{sc}}$ and the regularization loss $\mathcal{L}_{\text{r}}$. The constraint loss is introduced to learn the style from the source motion and then transfer it into the generated one through the style transfer network.

$$\mathcal{L}_{\text{sc}} = \|f(\phi_{\text{mg}}) - f(\phi_{\text{s}})\|_2^2 \tag{4}$$

where $f(\cdot)$ is the style transfer network.

The regularization loss $\mathcal{L}_{\text{r}}$ aims to increase the generalization of the style transfer process. Besides, it can deal with extreme cases of the generated motions that may break the manifold of valid styles and negatively affect the generated images. This loss is computed as:

$$\mathcal{L}_{\text{r}} = \left(\left\| \nabla_{\hat{\phi}_{\text{mg}}} f(\hat{\phi}_{\text{mg}}) \right\|_2 - 1 \right)^2 \tag{5}$$

where $\hat{\phi}_{\text{mg}}$ is the joint representation that controls the contribution of source motion ϕ_{s} during the style learning process. $\hat{\phi}_{\text{mg}}$ is computed from ϕ_{s} and ϕ_{mg} as follows:

$$\hat{\phi}_{\text{mg}} = \gamma \phi_{\text{s}} + (1 - \gamma)\phi_{\text{mg}} \tag{6}$$

where γ controls the amount of leveraged style information.

The final transferring loss $\mathcal{L}_{\text{transfer}}$ is computed as:

$$\mathcal{L}_{\text{transfer}} = \mathcal{L}_{\text{mg}} + \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{r}} \tag{7}$$

So as to control the style, both reference images and the audio stream are required during the transferring process.

5. Experiment

5.1. Dataset

Face. We use the VoxCeleb2 [9] to learn facial expressions. All videos are extracted at 60 FPS. We first trim the video to retain the face in the center, then resize it to 512×512. Our internal face tracker is leveraged to obtain 68 key points on the face. Face segmentation [28] is used to obtain the skin mask. Following [31], the head and torso motion is manually identified for the first frame of each series and tracked for the remaining frames using optical flow.

Audio. Following [31, 51], we use the Common Voice dataset [2] to train the Audio Encoder. There are around 26 hours of unlabeled statements throughout all samples. Note that 80-dimensional log Mel spectrograms are employed as surface representation and are computed with $\frac{1}{120}$(s) frame-shift, $\frac{1}{60}$(s) frame length, and 512-point STFT representation.

We evaluate and benchmark our results in the RAVDESS dataset [30]. The RAVDESS is a validated multimodal database of emotional speech and song, which is suitable and challenging to validate our method and different baselines. Note that, we only use this dataset for benchmarking to avoid training bias.

5.2. Implementation

We implement our framework using PyTorch. We train the network on the NVIDIA Titan V100 GPU with Adam optimizer [23]. The learning rate is set to 10^{-4}, 10^{-4}, 10^{-5}, 10^{-4} to train the Audio Encoding, the Motion Generator, the Style-Aware Generator, and the Style Mapping, respectively. The batch size is set to 8 for the Style-Aware Generator and 64 for other modules.

The implementation details of the networks in our Style Mapping module are described as follows.

Keypoint Extractor Given the input neutral image $\boldsymbol{I}_s$, a set of k number of keypoints c_k is extracted using a 3D U-Net encoder-decoder [44]. First, we project the encoded feature maps onto 3D volumes using 1×1 convolution. The encoder has 5 down-sampling layers. The decoder part of

the network has 5 up-sampling layers. Finally, the keypoints are predicted from the final 3D convolution layer.

Pose Expression Network We extract the pose, parameterized by a translation vector $\tau \in \mathbb{R}^3$ and a rotation matrix $\mathrm{R} \in \mathbb{R}^{3\times3}$, and expression information ε_k from the image by a Pose Expression Network. We adopt the same architecture as in [35]. A sequence of ResNet blocks is followed by global pooling to eliminate the spatial dimension. The rotation angles, translation vector, and expression information are then estimated using various linear layers. To estimate head pose, we divide the entire angle range into 66 bins of angles for yaw, pitch, and roll. Then, the network predicts the probability of the bin to which the target angle should belong. The rotation angles are then converted into a 3D rotation matrix. The final keypoints that contain the geometric and style information of the neutral image and Style Reference image are computed as in Equations 1 and 2.

Warping Network To reconstruct the output Intermediate Style Pattern image $\boldsymbol{I}_o$, we use a Warping Network to warp the feature volume of the neutral image using the two extracted keypoints set C and $\bar{\mathrm{C}}$ and their geometry information. Specifically, following first-order approximation [36], we estimate a warping function w_k for each k-th keypoints of the Style Reference image and the keypoints of the neutral image.

$$w_k : R\bar{R}^{-1}\left(\bar{p} - \bar{\mathrm{C}}_k\right) + \mathrm{C}_\mathrm{k} \mapsto p \tag{8}$$

where p and $\bar{p}$ are the 3D voxel location of the feature volume corresponding to the neutral image and Style Reference.

For each k-th keypoint , we apply w_k on every location of the neutral feature volume $w_k(\boldsymbol{F}_s)$ to obtain the k-th warped volume. Then, we concatenate all the warped volumes and pass them into a Warping Network to predict K composition maps $m = \{m_1, m_2, ..., m_K\}$, which contains the composition weights to aggregate the warping functions. In particular, we apply the softmax function at each location so that the composition weights can satisfy the condition:

$$\sum_k (m_k)(\bar{p}) = 1 \quad \& \quad 0 \leq m_k(\bar{p}) \leq 1, \qquad \forall \bar{p} \tag{9}$$

The final warped volume $w(\boldsymbol{F}_s)$ is calculated as the linear combination of the K warped volumes $w(\boldsymbol{F}_s) = \sum_{k=1}^{K} m_k w_k(\boldsymbol{F}_s)$. To handle occlusions caused by the warping, the network also predicts a 2D occlusion mask o, which is used as input of the Intermediate Generator in addition to the final warped volume.

Intermediate Generator This network takes the warped feature volume $w(\boldsymbol{F}_s)$ of the neutral image and projects it back to 2D dimensions. Then, the input feature is multiplied with the occlusion mask o obtained from the Warping Network. Finally, a 2D residual block series (6 blocks in total), 2 up-sampling layers, and a convolution layer are applied to construct the final Intermediate Style Pattern image. Since Intermediate Generator is an image generator, it contains LSGAN loss [32] that stabilizes the training process by adopting least squares. To achieve high fidelity, we minimize differences at the pixel-wise level and feature level as well as ensure the consistency of estimated keypoints. We also minimize high-level differences of style discrepancies through perceptual loss [22].

5.3. Qualitative Evaluation

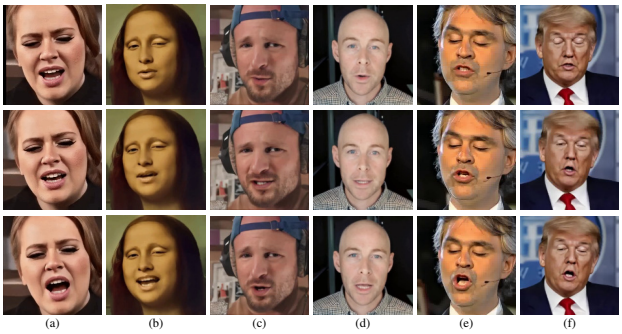


Figure 3. Our 2D photo-realistic talking head results with different styles. (a), (c), (e) are `ballad`, `rap`, and `opera` style references, respectively; (b), (d), (f) are the corresponding style transfer results. For more details, please visit our demonstration video.

Figure 3 shows that our method successfully transfers different styles such as `ballad`, `rap`, or `opera` to a new target character. Figure 4 shows the comparison between our method and recent works on 2D photorealistic talking head animation [31, 51] when the character sings an opera song. Focusing on the mouth, we notice that our method produces better results in mouth motion variance and eye expression compared to the results from [51] and [31]. In Figure 5, we show the comparison between different styles when they are encoded in one input audio to generate talking heads. Note that, in this case, different input images are used to verify the synthesis effectiveness of our method. Although different styles are encoded into different images to generate different talking heads, the animation is realistic and the performance of lip-synchronization is well-reserved.

5.4. Quantitative Evaluation

5.4.1 Evaluation Metric

We use six different metrics to evaluate how good and natural the animation of the generated talking head is: Cumulative Probability Blur Detection (CPBD) [42], Landmark Distance (D-L) [51], Landmarks Distance around the Mouth (LMD), Landmark Velocity difference (D-V) [51], Difference in the open mouth area (D-A) [51].

Figure 4. Comparison between different 2D talking head galleries on `opera` style. Our method generates more natural and realistic motion, especially around the mouth and the eye of the character.

Style transfer metric. To evaluate style transfer results efficiently, we introduce three new following metrics.

Style-Aware Landmarks Distance (SLD): To evaluate the style information encoded in a generated talking head, we design a metric called Style-Aware Landmarks Distance (SLD). This metric calculates the accuracy of mouth, eyes, head pose shapes between a chunked window of style reference and a chunked window of corresponding talking head animation. Lower is better. Let's assumed that a style reference video with N_s frames is split into multiple temporal periods of F frames (window size), i.e., style reference windows $\mathrm{W_s} = \left(\mathrm{w_s^{(0:F)}}, \mathrm{w_s^{(v:F+v)}}, \mathrm{w_s^{(2v:F+2v)}}, \cdots, \mathrm{w_s^{(\kappa v:F+\kappa v)}}\right)$, with $\mathrm{w_s^{(i:F+i)}}$ being the frames from i-*th* to $(\mathrm{F}+\mathrm{i})$-*th* of the reference video, v is the stride, and $\kappa = \lfloor(\mathrm{N_s} - \mathrm{F})/\mathrm{v}\rfloor$. Similar to the reference video, we chunk the generated animation video into smaller chunked windows $\mathrm{W_a} = \left(\mathrm{w_a^{(0:F)}}, \mathrm{w_a^{(v:F+v)}}, \mathrm{w_a^{(2v:F+2v)}}, \cdots, \mathrm{w_a^{(\kappa v:F+\kappa v)}}\right)$. The SLD is then calculated with the core is the D-L metric as:

$$\mathrm{SLD} = \frac{1}{|\mathrm{W_s}|} \sum_{\mathrm{w_s} \in \mathrm{W_s}} \left(\min_{\mathrm{w_a} \in \mathrm{W_a}} \left(\mathrm{D{-}L}\left(\mathrm{w_s}, \mathrm{w_a}\right)\right) \right) \quad (10)$$

where D−L is the Landmark Distance metric [51].

Similarly, we calculate the *Style-Aware Landmarks Velocity Difference* (SLV) and *Style-Aware Mouth Area Difference* (SMD).

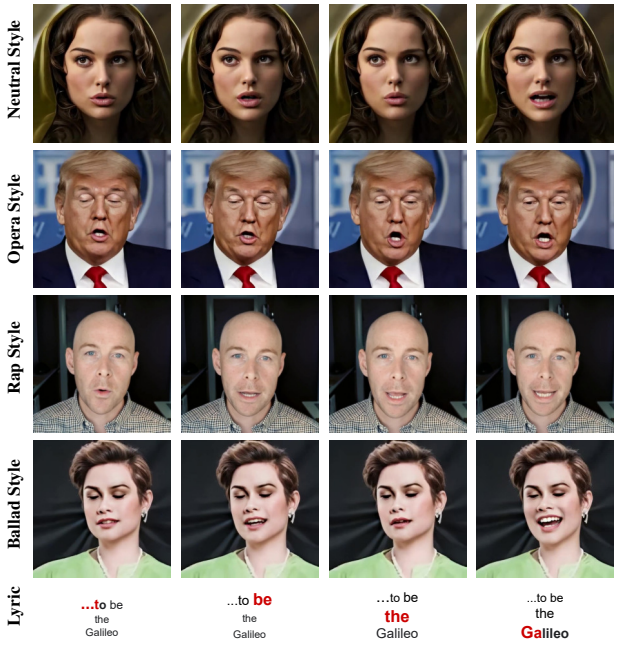


Figure 5. Comparison between different styles an input audio.

5.4.2 2D Talking Head Generation Results

Table 1. Result of different 2D talking head generation methods.

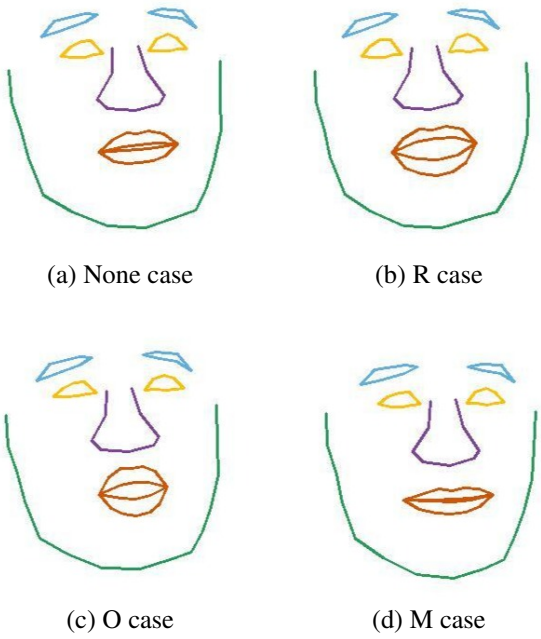
Methods	Metrics				
	CPBD↑	LMD↓	D-L↓	D-V↓	D-A↓
Ground Truth	0.28	0.00	0.00%	0.00%	0.00%
MIT [51]	0.18	2.28	2.78%	0.88%	14.52%
PCT [49]	0.09	3.22	3.27%	0.86%	36.84%
LSP [31]	0.20	3.29	5.43%	0.85%	30.65%
AD-NERF [17]	0.21	2.43	2.67%	0.85%	13.34%
MetaPortrait [46]	0.20	2.02	3.42%	0.84%	17.69%
Diffused Heads [38]	0.22	2.13	2.72%	0.84%	11.23%
Ours	**0.26**	**1.83**	**2.65%**	**0.83%**	**10.53%**

Table 1 shows the 2D talking head result comparison between our method and recent baselines, including [17, 31, 38, 46, 49, 51]. From Table 1, we can see that our method outperforms recent state-of-the-art approaches by a large

margin. In particular, our method achieves the highest accuracy in CPBD, LMD, D-L, D-V, and D-A metrics. These results show that our method successfully renders the 2D talking head and increases the quality of the rendered results. Overall, our method can increase the sharpness of the head (identified by CPBD) metric, while generating natural facial motion (identified by LMD, D-L, D-V, and D-A metric).

5.4.3 Motion Templates for Style Reference

The style reference is expected to capture the personalized spotlight of the characters when they are talking or singing. To learn and capture the style information from a target character, we need to use the key motion templates that match with the syllable. According to music theory about rhythm [3], a word can have many syllables and one syllable can have more than one vowel. Vowels are a, e, i, o, u. The other letters (like b, c, d, f) are consonants. However, each word can be split into single syllables and follow open and closed syllable patterns. A closed syllable has a short vowel ending in a consonant. It currently matches with the 'None' case and 'M' case, which are split based on the differences in mouth shape. An open syllable ends with a vowel sound that is spelled with a single vowel letter. 'R' case and 'O' case are two cases of the open syllable that have high differences in mouth motions. Each word can be formed by more than one vowel and there are seven syllable types in total for English. For visualization of motion templates, please see Figure 6.



(a) None case (b) R case

(c) O case (d) M case

Figure 6. Illustration of four key motion templates.

5.4.4 Style Comparison

Table 2. Result comparison in terms of style transfer between different 2D talking head generation methods.

Methods	Metrics		
	SLD↓	SLV↓	SMD↓
MIT [51]	3.00	0.94	5.03
PCT [49]	3.58	0.93	7.28
LSP [31]	5.40	0.91	6.89
AD-NEFT [17]	4.69	0.92	5.48
MetaPortrait [46]	3.73	0.90	4.97
Diffused Heads [38]	3.01	0.90	4.66
Ours	**2.84**	**0.89**	**4.26**

Table 2 shows the comparison between our method and other baselines [17, 31, 38, 46, 49, 51] in terms of style transfer. Three designed metrics (SLD, SLV, and SMD) are used for evaluation and benchmarking. The results show that our method outperforms others by a large margin in all three metrics. This substantial performance gap strongly suggests the efficacy of our method in accurately capturing style characteristics from the reference image and seamlessly transferring them onto the target image. This not only highlights the robustness of our method but also underscores its potential for practical applications in various domains requiring high-quality style transfer. Furthermore, these results underscore the importance of our approach in advancing the state-of-the-art in style transfer techniques, promising richer and more faithful artistic transformations.

6. Conclusion

We have introduced a novel method designed to generate lifelike 2D talking heads from input audio signals, revolutionizing the realm of character animation. In addition to the primary audio stream and an accompanying image, our framework harnesses a meticulously curated set of reference frames to effectively learn the character style characteristics. Notably, our approach excels even with the most demanding and challenging vocal styles, including `ballad`, `opera`, and `rap`, where complex movements are necessary to produce animations that are faithful and natural. Extensive experiments demonstrate the superior performance of our talking head synthesis, showing qualitative and quantitative advantages over recent state-of-the-art methods. The versatility of our framework can be potential for diverse applications, ranging from dubbing, video conferencing experiences, to the creation of dynamic virtual avatars. With the ability to accurately capture and animate diverse head movement styles, we hope to further advance the field of character animation, allowing more expressive and vivid human-like facial talking animation.

References

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *ECCV*, 2020. 4

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *International Conference on Language Resources and Evaluation*, 2020. 5

[3] Amalia Arvaniti. Rhythm, timing and the timing of rhythm. *Phonetica*, 2009. 3, 8

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video Rewrite: driving visual speech with audio. In *SIGGRAPH*, 1997. 2

[5] Yujin Chai, Yanlin Weng, Lvdi Wang, and Kun Zhou. Speech-driven facial animation with spectral gathering and temporal attention. *Frontiers of Computer Science*, 16:1–10, 2022. 2

[6] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-Realistic Facial Details Synthesis From Single Image. In *ICCV*, 2019. 1

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, 2020. 2

[8] Mingxiang Chen, Zhanguo Chang, Haonan Lu, Bitao Yang, Zhuang Li, Liufang Guo, and Zhecheng Wang. Augnet: End-to-end unsupervised visual representation learning with image augmentation. *arXiv preprint arXiv:2106.06250*, 2021. 3, 4

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

[10] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, Learning, and Synthesis of 3D Speaking Styles. In *CVPR*, 2019. 1

[11] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM TOG*, 2016. 1

[12] Sefik Emre Eskimez, Ross K. Maddox, Chenliang Xu, and Zhiyao Duan. Generating Talking Face Landmarks from Speech. In *Latent Variable Analysis and Signal Separation*, 2018. 2

[13] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004. 2

[14] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt. VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Computer Graphics Forum*, 2015. 2

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv preprint arXiv:1406.2661*, 2014. 1

[16] David Greenwood, Iain Matthews, and Stephen Laycock. Joint Learning of Facial Expression and Head Pose from Speech. In *INTERSPEECH*, 2018. 2

[17] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021. 1, 7, 8

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

[19] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, 2022. 2

[20] 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. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1

[21] Fan Jiang, Xubo Yang, and Lele Feng. Real-time full-body motion reconstruction and recognition for off-the-shelf vr devices. In *Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry-Volume 1*, 2016. 1

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*, 2016. 6

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 5

[24] Neeraj Kumar, Srishti Goel, Ankur Narang, and Mujtaba Hasan. Robust one shot audio to video generation. In *CVPR*, 2020. 2

[25] Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, and Anh Nguyen. Uncertainty-aware label distribution learning for facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6088–6097, 2023. 1

[26] Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Music-driven group choreography. *CVPR*, 2023. 1

[27] 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 *CVPR*, 2022. 2

[28] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, 2021. 4, 5

[29] Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. Video-audio driven real-time facial animation. *TOG*, 2015. 2

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

[31] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live Speech Portraits: Real-time photorealistic talking-head animation. *ACM TOG*, 2021. 1, 2, 3, 5, 6, 7, 8

[32] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks. In *ICCV*, 2017. 6

[33] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5292–5302, 2024. 2

[34] 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. 2

[35] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPR*, 2018. 6

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NIPS*, 32, 2019. 6

[37] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Change Loy, and Ran He. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1, 2

[38] Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zikeba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024. 2, 7, 8

[39] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM TOG*, 2017. 2, 4

[40] Yu Tsao, Jinyu Li, and Chin-Hui Lee. Ensemble speaker and speaking environment modeling approach with advanced online estimation process. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009. 2

[41] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *CVPR*, 2019. 1

[42] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *IJCV*, 2019. 6

[43] Marilyn A Walker, Janet E Cahn, and Stephen J Whittaker. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the first international conference on Autonomous agents*, 1997. 1

[44] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *CVPR*, 2021. 2, 5

[45] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2Face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes. In *ECCV*, 2018. 2

[46] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023. 1, 7, 8

[47] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2

[48] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. 2

[49] 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 *CVPR*, 2021. 1, 2, 7, 8

[50] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: audio-driven animator-centric speech animation. *ACM TOG*, 2018. 1, 2

[51] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. MakeItTalk: Speaker-Aware Talking-Head Animation. *ACM TOG*, 2020. 1, 2, 3, 5, 6, 7, 8