Talking-head Generation with Rhythmic Head Motion

Lele Chen[†], Guofeng Cui[†], Celong Liu[‡], Zhong Li[‡], Ziyi Kou[†], Yi Xu[‡], and Chenliang Xu[†]

[†] University of Rochester [‡] OPPO US Research Center
lchen63@cs.rochester.edu

Abstract. When people deliver a speech, they naturally move heads, and this rhythmic head motion conveys prosodic information. However, generating a lip-synced video while moving head naturally is challenging. While remarkably successful, existing works either generate still talking-face videos or rely on landmark/video frames as sparse/dense mapping guidance to generate head movements, which leads to unrealistic or uncontrollable video synthesis. To overcome the limitations, we propose a 3D-aware generative network along with a hybrid embedding module and a non-linear composition module. Through modeling the head motion and facial expressions[1] explicitly, manipulating 3D animation carefully, and embedding reference images dynamically, our approach achieves controllable, photo-realistic, and temporally coherent talking-head videos with natural head movements. Thoughtful experiments on several standard benchmarks demonstrate that our method achieves significantly better results than the state-of-the-art methods in both quantitative and qualitative comparisons. The code is available on https://github.com/lelechen63/Talking-head-Generation-with-Rhythmic-Head-Motion.

1 Introduction

People naturally emit head movements when they speak, which contain non-verbal information that helps the audience comprehend the speech content [4,15]. Modeling the head motion and then generating a controllable talking-head video are valuable problems in computer vision. For example, it can benefit the research of adversarial attacks in security or provide more training samples for supervised learning approaches. Meanwhile, it is also crucial to real-world applications, such as enhancing speech comprehension for hearing-impaired people, and generating virtual characters with synchronized facial movements to speech audio in movies/games.

We humans can infer the visual prosody from a short conversation with speakers [22]. Inspired by that, in this paper, we consider such a task: given a

[1] In our setting, facial expression means facial movement(e.g., blinks, and lip & chin movements).

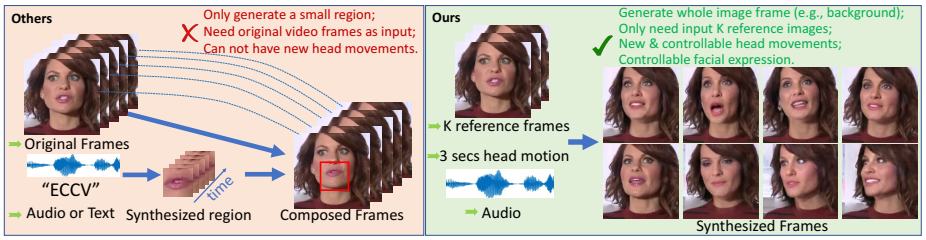


Fig. 1. The comparisons of other methods [27,13] and our method. The green arrows denote inputs. [27,13] require whole original video frames as input and can only edit a small region (e.g., lip region) on the original video frames. In contrast, through learning the appearance embedding from K reference frames and predicting future head movements using the head motion vector extracted from the 3-sec reference video, our method generates whole future video frames with controllable head movements and facial expressions.

short video [2] of the target subject and an arbitrary reference audio, generating a photo-realistic and lip-synced talking-head video with natural head movements. Similar problems [8,39,26,30,7,35,37,31] have been explored recently. However, several challenges on how to explicitly use the given video and then model the apparent head movements remain unsolved. In the rest of this section, we discuss our technical contributions concerning each challenge.

The deformation of a talking-head consists of his/her intrinsic subject traits, head movements, and facial expressions, which are highly convoluted. This complexity stems not only from modeling face regions but also from modeling head motion and background. While previous audio-driven talking-face generation methods [9,35,39,6,26,7] could generate lip-synced videos, they omit the head motion modeling, thus can only generate still talking-face with expressions under a fixed facial alignment. Other landmark/image-driven generation methods [37,31,32,33,35] can synthesize moving-head videos relying on the input facial landmarks or video frames as guidance to infer the convoluted head motion and facial expressions. However, those methods fail to output a controllable talking-head video (e.g., control facial expressions with speech), which greatly limits their use. To address the convoluted deformation problem, we propose a simple but effective method to decompose the head motion and facial expressions.

Another challenge is exploiting the information contained in the reference image/video. While the few-shot generation methods [37,31,20,36] can synthesize videos of unseen subjects by leveraging K reference images, they only utilize global appearance information. There remains valuable information underexplored. For example, we can infer the individual's head movement characteristics by leveraging the given short video. Therefore, we propose a novel method to extrapolate rhythmic head motion based on the short input video and the

[2] E.g., a 3-sec video. We use it to learn head motion and only sample K images as reference frames.

conditioned audio, which enables generating a talking-head video with natural head poses. Besides, we propose a novel hybrid embedding network dynamically aggregating information from the reference images by approximating the relation between the target image with the reference images.

People are sensitive to any subtle artifacts and perceptual identity changes in a synthesized video, which are hard to avoid in GAN-based methods. 3D graphics modeling has been introduced by [13,17,38] in GAN-based methods due to its stability. In this paper, we employ the 3D modeling along with a novel non-linear composition module to alleviate the visual discontinuities caused by apparent head motion or facial expressions.

Combining the above features, which are designed to overcome limitations of existing methods, our framework can generate a talking-head video with natural head poses that conveys the given audio signal. We conduct extensive experimental validation with comparisons to various state-of-the-art methods on several benchmark datasets (e.g., VoxCeleb2 [9] and LRS3-TED [1] datasets) under several different settings (e.g., audio-driven and landmark-driven). Experimental results show that the proposed framework effectively addresses the limitations of those existing methods.

2 Related Work

2.1 Talking-head Image Generation

The success of graphics-based approaches has been mainly limited to synthesizing talking-head videos for a specific person [14,2,5,19,27,13]. For instance, Suwajanakorn et al. [27] generate a small region (e.g., lip region, see Fig. 1)[3] and compose it with a retrieved frame from a large video corpus of the target person to produce photo-realistic videos. Although it can synthesize fairly accurate lip-synced videos, it requires a large amount of video footage of the target person to compose the final video. Moreover, this method can not be generalized to an unseen person due to the rigid matching scheme. More recently, video-to-video translation has been shown to be able to generate arbitrary faces from arbitrary input data. While the synthesized video conveys the input speech signal, the talking-face generation methods [24,8,39,26,30,7] can only generate facial expressions without any head movements under a fixed alignment since the head motion modeling has been omitted. Talking-head methods [37,31,35] can generate high-quality videos guided by landmarks/images. However, these methods can not generate controllable video since the facial expressions and head motion are convoluted in the guidance, e.g., using audio to drive the generation. In contrast, we explicitly model the head motion and facial expressions in a disentangled manner, and propose a method to extrapolate rhythmic head motion to enable generating future frames with natural head movements.

[3] We intent to highlight the different between [27,13] and our work. While they generate high-quality videos, they can not disentangle the head motion and facial movement due to their intrinsic limitations.

2.2 Related Techniques

Embedding Network refers to the external memory module to learn common feature over few reference images of the target subjects. And this network is critical to identity-preserving performance in generation task. Previous works [8,7,30,26] directly use CNN encoder to extract the feature from reference images, then concatenate with other driven vectors and decode it to new images. However, this encoder-decoder structure suffers identity-preserving problem caused by the deep convolutional layers. Wiles et al. [35] propose an embedding network to learn a bilinear sampler to map the reference frames to a face representation. More recently, [37,36,31] compute part of the network weights dynamically based on the reference images, which can adapt to novel cases quickly. Inspired by [31], we propose a hybrid embedding network to aggregate appearance information from reference images with apparent head movements.

Image matting function has been well explored in image/video generation task [29,32,24,7,31]. For instance, Pumarola et al. [24] compute the final output image by $\hat{\mathbf{I}} = (1 - \mathbf{A}) * \mathbf{C} + \mathbf{A} * \mathbf{I}_r$, where $\mathbf{I}_r$, $\mathbf{A}$ and $\mathbf{C}$ are input reference image, attention map and color mask, respectively. The attention map $\mathbf{A}$ indicates to what extend each pixel of $\mathbf{I}_r$ contributes to the output image $\hat{\mathbf{I}}$. However, this attention mechanism may not perform well if there exits a large deformation between reference frame $\mathbf{I}_r$ and target frame $\hat{\mathbf{I}}$. Wang et al. [33] use estimated optic flow to warp the $\mathbf{I}_r$ to align with $\hat{\mathbf{I}}$, which is computationally expensive and can not estimate the rotations in talking-head videos. In this paper, we propose a 3D-aware solution along with a non-linear composition module to better tackle the misalignment problem caused by apparent head movements.

3 Method

3.1 Problem Formulation

We introduce a neural approach for talking-head video generation, which takes as input sampled video frames, $\mathbf{y}_{1:\tau} \equiv \mathbf{y}_1, ..., \mathbf{y}_\tau$, of the target subject and a sequence of driving audio, $\mathbf{x}_{\tau+1:T} \equiv \mathbf{x}_{\tau+1}, ..., \mathbf{x}_T$, and synthesizes target video frames, $\hat{\mathbf{y}}_{\tau+1:T} \equiv \hat{\mathbf{y}}_{\tau+1}, ..., \hat{\mathbf{y}}_T$, that convey the given audio signal with realistic head movements. To explicitly model facial expressions and head movements, we decouple the full model into three sub-models: a facial expression learner (Ψ), a head motion learner (Φ), and a 3D-aware generative network (Θ). Specifically, given an audio sequence $\mathbf{x}_{\tau+1:T}$ and an example image frame $\mathbf{y}_{t_\mathcal{M}}$, Ψ generates the facial expressions $\hat{\mathbf{p}}_{\tau+1:T}$. Meanwhile, given a short reference video $\mathbf{y}_{1:\tau}$ and the driving audio $\mathbf{x}_{\tau+1:T}$, Φ extrapolates natural head motion $\hat{\mathbf{h}}_{\tau+1:T}$ to manipulate the head movements. Then, the Θ generates video frames $\hat{\mathbf{y}}_{\tau+1:T}$ using $\hat{\mathbf{p}}_{\tau+1:T}$, $\hat{\mathbf{h}}_{\tau+1:T}$, and $\mathbf{y}_{1:\tau}$. Thus, the full model is given by:

$$\hat{\mathbf{y}}_{\tau+1:T} = \Theta(\mathbf{y}_{1:\tau}, \hat{\mathbf{p}}_{\tau+1:T}, \hat{\mathbf{h}}_{\tau+1:T}) = \Theta(\mathbf{y}_{1:\tau}, \Psi(\mathbf{y}_{t_\mathcal{M}}, \mathbf{x}_{\tau+1:T}), \Phi(\mathbf{y}_{1:\tau}, \mathbf{x}_{\tau+1:T})) \ .$$

The proposed framework (see Fig. 2) aims to exploit the facial texture and head motion information in $\mathbf{y}_{1:\tau}$ and maps the driving audio signal to a sequence of generated video frames $\hat{\mathbf{y}}_{\tau+1:T}$.

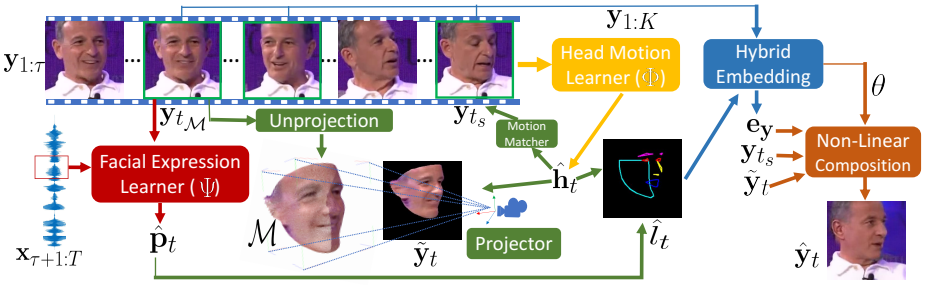


Fig. 2. The overview of the framework. Ψ and Φ are introduced in Sec. 3.2 and Sec. 3.3, respectively. The 3D-aware generative network consists of a 3D-Aware module (green part, see Sec. 4.1), a hybrid embedding module (blue part, see Sec. 4.2), and a non-linear composition module (orange part, see Sec 4.3).

3.2 The Facial Expression Learner

The facial expression learner Ψ (see Fig. 2 red block) receives as input the raw audio $\mathbf{x}_{\tau+1:T}$, and a subject-specific image $\mathbf{y}_{t_\mathcal{M}}$ (see Sec. 4.1 about how to select $\mathbf{y}_{t_\mathcal{M}}$), from which we extract the landmark identity feature $\mathbf{p}_{t_\mathcal{M}}$. The desired outputs are PCA components $\hat{\mathbf{p}}_{\tau+1:T}$ that convey the facial expression. During training, we mix $\mathbf{x}_{\tau+1:T}$ with a randomly selected noise file in 6 to 30 dB SNRs with 3 dB increments to increase the network robustness. At time step t, we encode audio clip $\mathbf{x}_{t-3:t+4}$ (0.04×7 Sec) into an audio feature vector, concatenate it with the encoded reference landmark feature, and then decode the fused feature into PCA components of target expression $\hat{\mathbf{p}}_t$. During inference, we add the identity feature $\mathbf{p}_{t_\mathcal{M}}$ back to the resultant facial expressions $\hat{\mathbf{p}}_{\tau+1:T}$ to keep the identity information.

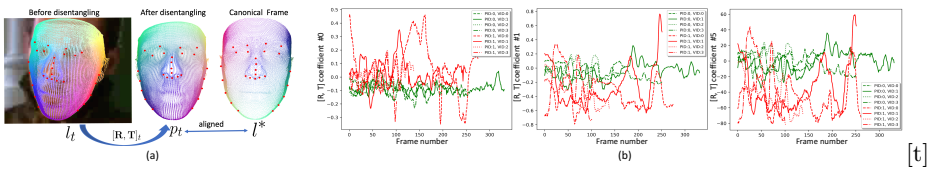


Fig. 3. (a) shows the motion disentangling process. (b) shows the head movement coefficients of two randomly-selected identities. PID, VID indicate identity id and video id, respectively.

3.3 The Head Motion Learner

To generate talking-head videos with natural head movements, we introduce a head motion learner Φ, which disentangles the reference head motion $\mathbf{h}_{1:\tau}$ from

the input reference video clip $\mathbf{y}_{1:\tau}$ and then predicts the head movements $\hat{\mathbf{h}}_{\tau+1:T}$ based on the driving audio $\mathbf{x}_{1:T}$ and the disentangled motion $\mathbf{h}_{1:\tau}$.

Motion Disentangler. Rather than disentangling the head motion problem in image space, we learn the head movements $\mathbf{h}$ in 3D geometry space (see Fig. 3(a)). Specifically, we compute the transformation matrix $[\mathbf{R}, \mathbf{T}] \in \mathbb{R}^6$ to represent $\mathbf{h}$, where we omit the camera motion and other factors. At each time step t, we compute the rigid transformation $[\mathbf{R}, \mathbf{T}]_t$ between $\boldsymbol{l}_t$ and canonical 3D facial landmark $\boldsymbol{l}^*$, which is $\boldsymbol{l}^* \approx \mathbf{p}_t = \mathbf{R}_t \boldsymbol{l}_t + \mathbf{T}_t$, where $\approx$ denotes aligned. After transformation, the head movement information is removed, and the resultant $\mathbf{p}_t$ only carries the facial expressions information.

The natural head motion extrapolation. We randomly select two subjects (each with four videos) from VoxCeleb2 [9] and plot the motion coefficients in Fig. 3(b). We can find that one subject (same PID, different VID) has similar head motion patterns in different videos, and different subjects (different PID) have much different head motion patterns. To explicitly learn the individual's motion from $\mathbf{h}_{1:\tau}$ and the audio-correlated head motion from $\mathbf{x}_{1:T}$, we propose a few-shot temporal convolution network Φ, which can be formulated as: $\hat{\mathbf{h}}_{\tau+1:T} = \Phi(f(\mathbf{h}_{1:\tau}, \mathbf{x}_{1:\tau}), \mathbf{x}_{\tau+1:T})$. In details, the encoder f first encodes $\mathbf{h}_{1:\tau}$ and $\mathbf{x}_{1:\tau}$ to a high-dimensional reference feature vector, and then transforms the reference feature to network weights $\mathbf{w}$ using a multi-layer perception. Meanwhile, $\mathbf{x}_{\tau+1:T}$ is encoded by several temporal convolutional blocks. Then, the encoded audio feature is processed by a multi-layer perception, where the weights are dynamically learnable $\mathbf{w}$, to generate $\hat{\mathbf{h}}_{\tau+1:T}$.

3.4 The 3D-Aware Generative Network

To generate controllable videos, we propose a 3D-aware generative network Θ, which fuses the head motion $\mathbf{h}_{\tau+1:T}$, and facial expressions $\mathbf{p}_{\tau+1:T}$ with the appearance information in $\mathbf{y}_{1:\tau}$ to synthesise target video frames, $\hat{\mathbf{y}}_{\tau+1:T}$. The Θ consists of three modules: a *3D-Aware* module, which manipulates the head motion to reconstruct an intermediate image that carries desired head pose using a differentiable unprojection-projection step to bridge the gap between images with different head poses. Comparing to landmark-driven methods [37,31,33], it can alleviate the overfitting problem and make the training converge faster; a *Hybrid Embedding* module, which is proposed to explicitly embed texture features from different appearance patterns carried by different reference images; a *Non-Linear Composition* module, which is proposed to stabilize the background and better preserve the individual's appearance. Comparing to image matting function [24,39,37,33,7], our non-linear composition module can synthesize videos with natural facial expression, smoother transition, and a more stable background.

4 3D-Aware Generation

4.1 *3D-Aware* Module

3D Unprojection We assume that the image with frontal face contains more appearance information. Given a short reference video clip $\mathbf{y}_{1:\tau}$, we compute the head rotation $\mathbf{R}_{1:\tau}$ and choose the most visible frame $\mathbf{y}_{t_\mathcal{M}}$ with minimum rotation, such that $\mathbf{R}_{t_\mathcal{M}} \to 0$. Then we feed it to an unprojection network to obtain a 3D mesh $\mathcal{M}$. The Unprojection network is a U-net structure network [12] and we pretrain it on 300W-LP [40] dataset. After pretrianing, we fix the weights and use it to transfer the input image $\mathbf{y}_{t_\mathcal{M}}$ into the textured 3D mesh $\mathcal{M}$. The topology of $\mathcal{M}$ will be fixed for all frames in one video, we denote it as $\mathcal{F}^\mathcal{M}$.

3D Projector In order to get the projected image $\tilde{\mathbf{y}}_t$ from the 3D face mesh $\mathcal{M}$, we need to compute the correct pose for $\mathcal{M}$ at time t. Suppose $\mathcal{M}$ is reconstructed from frame $\mathbf{y}_{t_\mathcal{M}}, (1 \leq t_\mathcal{M} \leq \tau)$, the position of vertices of $\mathcal{M}$ at time t can be computed by:

$$\mathcal{V}_t^\mathcal{M} = \mathbf{R}_t^{-1}(\mathbf{R}_{t_\mathcal{M}}\mathcal{V}_{t_\mathcal{M}}^\mathcal{M} + \mathbf{T}_{t_\mathcal{M}} - \mathbf{T}_t) \ , \tag{1}$$

where $\mathcal{V}_{t_\mathcal{M}}^\mathcal{M}$ denotes the position of vertices of $\mathcal{M}$ at time $t_\mathcal{M}$. Hence, $\mathcal{M}$ at time t can be represented as $\left[\mathcal{V}_t^\mathcal{M}, \mathcal{F}^\mathcal{M}\right]$. Then, a differentiable rasterizer [21] is applied here to render $\tilde{\mathbf{y}}_t$ from $\left[\mathcal{V}_t^\mathcal{M}, \mathcal{F}^\mathcal{M}\right]$. After rendering, $\tilde{\mathbf{y}}_t$ carries the same head pose as the target frame $\mathbf{y}_t$ and the same expression as $\mathbf{y}_{t_\mathcal{M}}$.

Motion Matcher We randomly select K frames $\mathbf{y}_{1:K}$ from $\mathbf{y}_{1:\tau}$ as reference images. To infer the prior knowledge about the background of the target image from $\mathbf{y}_{1:K}$, we propose a motion matcher, $M(\mathbf{h}_t, \mathbf{h}_{1:K}, \mathbf{y}_{1:K})$, which matches a frame $\mathbf{y}_{t_s}$ from $\mathbf{y}_{1:K}$ with the nearest background as $\mathbf{y}_t$ by comparing the similarities between $\mathbf{h}_t$ with $\mathbf{h}_{1:K}$. To solve this background retrieving problem, rather than directly computing the similarities between facial landmarks, we compute the geometry similarity cost c_t using the head movement coefficients and choose the $\mathbf{y}_{t_s}$ with the least cost c_t. That is:

$$c_t = \min_{k=1,2,\dots K} \|\mathbf{h}_t - \mathbf{h}_k\|_2^2 \ . \tag{2}$$

The matched $\mathbf{y}_{t_s}$ is passed to the non-linear composition module (see Sec. 4.3) since it carries the similar background pattern as $\mathbf{y}_t$. During training, we manipulate the selection of $\mathbf{y}_{t_s}$ by increasing the cost c_t, which makes the non-linear composition module more robust to different $\mathbf{y}_{t_s}$.

4.2 Hybrid Embedding Module

Instead of averaging the K reference image features [37,8,35], we design a hybrid embedding mechanism to dynamically aggregate the K visual features.

Recall that we obtain the facial expression $\mathbf{p}_t$ (Sec. 3.2) and the head motion $\mathbf{h}_t$ (Sec. 3.3) for time t. Here, we combine them to landmark $\boldsymbol{l}_t$ and transfer it to a gray-scale image. Our hybrid embedding network (Fig. 4) consists of four sub-networks: an Activation Encoder (E_A, green part) to generate the activation

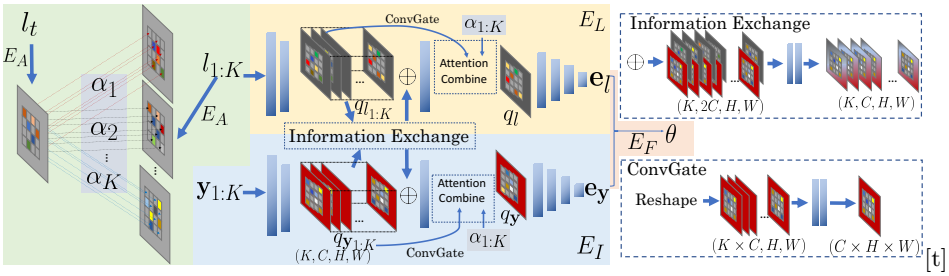


Fig. 4. The network structure of hybrid embedding module. Green, yellow, blue, pink parts are activation encoder E_A, landmark encoder E_L, image encoder E_I, and fusion network E_F, respectively.

map between the query landmark $\boldsymbol{l}_t$ and reference landmarks $\boldsymbol{l}_{1:K}$; an Image Encoder (E_I, blue part) to extract the image feature $\mathbf{e}_\mathbf{y}$; a Landmark Encoder (E_L, yellow part) to extract landmark feature $\mathbf{e}_\boldsymbol{l}$; a fusion network ($E_F$, pink part) to aggregate image feature and landmark feature into parameter set θ. The overview of the embedding network is performed by:

$$\boldsymbol{\alpha}_k = \text{softmax}(E_A(\boldsymbol{l}_k) \odot E_A(\boldsymbol{l}_t)) \ , \quad k \in [1, K] \ , \tag{3}$$

$$\mathbf{e}_\mathbf{y} = E_I(\mathbf{y}_{1:K}, \boldsymbol{\alpha}_{1:K}) \ , \quad \mathbf{e}_\boldsymbol{l} = E_L(\boldsymbol{l}_{1:K}, \boldsymbol{\alpha}_{1:K}) \ , \tag{4}$$

$$\theta = E_F(\mathbf{e}_\boldsymbol{l}, \mathbf{e}_\mathbf{y}) \ , \tag{5}$$

where $\odot$ denotes element-wise multiplication. We regard the activation map $\boldsymbol{\alpha}_k$ as the activation energy between $\boldsymbol{l}_k$ and $\boldsymbol{l}_t$, which approximates the similarity between $\mathbf{y}_k$ and $\mathbf{y}_t$.

We observe that different referent images may carry different appearance patterns, and those appearance patterns share some common features with the reference landmarks (e.g., head pose, and edges). Assuming that knowing the information of one modality can better encode the feature of another modality [28], we hybrid the information between E_I and E_L. Specifically, we use two convolutional layers to extract the feature map $\mathbf{q}_{\mathbf{y}_{1:K}}$ and $\mathbf{q}_{\boldsymbol{l}_{1:K}}$, and then forward them to an information exchange block to refine the features. After exchanging the information, the hybrid feature is concatenated with the original feature $\mathbf{q}_{\mathbf{y}_{1:K}}$. Then, we pass the concatenated feature to a bottleneck convolution layer to produce the refined feature $\mathbf{q}'_{\mathbf{y}_{1:K}}$. Meanwhile, we apply a ConvGate block on $\mathbf{q}_{\mathbf{y}_{1:K}}$ to self-aggregate the features from K references, and then combine the gated feature with the refined feature using learnable activation map $\boldsymbol{\alpha}_k$. This attention combine step can be formulated as:

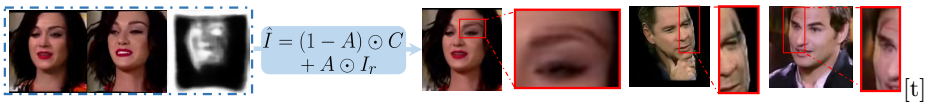
$$\mathbf{q}_\mathbf{y} = \sum_{i=1}^{K} (\boldsymbol{\alpha}_{1:K} \odot \mathbf{q}'_{\mathbf{y}_{1:K}}) + \text{ConvGate}(\mathbf{q}_{\mathbf{y}_{1:K}}) \ . \tag{6}$$

Then, by applying several convolutional layers to the aggregated feature $\mathbf{q}_\mathbf{y}$, we obtain the image feature vector $\mathbf{e}_\mathbf{y}$. Similarly, the landmark feature vector $\mathbf{e}_\boldsymbol{l}$ can also be produced.

The Fusion network E_F consists of three blocks: a N-layer image feature encoding block E_{F_I}, a landmark feature encoding block E_{F_L}, and a multi-layer perception block E_{F_P}. Thus, we can rewrite Eq. 5 as:

$$\{\theta^i_\gamma, \theta^i_\beta, \theta^i_S\}_{i\in[1,N]} = E_{F_P}(\text{softmax}(E_{F_L}(\mathbf{e_l})) \odot E_{F_I}(\mathbf{e_y})) \ , \tag{7}$$

where $\{\theta^i_\gamma, \theta^i_\beta, \theta^i_S\}_{i\in[1,N]}$ are the learnable network weights in the non-linear composition module in Sec. 4.3.



[t]

Fig. 5. The artifact examples produced by the image matting function. The blue dashed box shows the color map $\mathbf{C}$, reference image $\mathbf{I_r}$, and attention map $\mathbf{A}$, respectively. The blue box shows the image matting function (see Sec. 2.2) followed by the final result with its zoom-in artifact area. Two other artifact examples are on the right side.

4.3 Non-Linear Composition Module

Since the projected image $\tilde{\mathbf{y}}_t$ carries valuable facial texture with better alignment to the ground-truth, and the matched image $\mathbf{y}_{t_s}$ contains similar background pattern, we input them to our composition module to ease the burden of the generation. We use FlowNet [32] to estimate the flow between $\tilde{\boldsymbol{l}}_t$ and $\hat{\boldsymbol{l}}_t$, and warp the projected image to obtain $\tilde{\mathbf{y}}'_t$. Meanwhile, the FlowNet also outputs an attention map $\boldsymbol{\alpha}_{\tilde{y}_t}$. Similarly, we obtain the warped matched image and its attention map $[\mathbf{y'}_{t_s}, \boldsymbol{\alpha}_{\mathbf{y}_{t_s}}]$. Although image matting function is a popular way to combine the warped image and the raw output of the generator [31,32,7,24] using these attention maps, it may generate apparent artifacts (Fig. 5) caused by misalignment, especially in the videos with apparent head motion. To solve this problem, we propose a nonlinear combination module (Fig. 6) based on the network structure proposed in [31].

The decoder G consists of N layers and each layer contains three parallel SPADE blocks [23]. Inspired by [31], in the i^{th} layer of G, we first convolve $[\theta^i_\gamma, \theta^i_\beta]$ with $\phi^i_{l_t}$ to generate the scale and bias map $\{\gamma^i, \beta^i\}_{\phi^i_{l_t}}$ of SPADE (green box) to denormalize the appearance vector $\mathbf{e}^i_\mathbf{y}$. This step aims to sample the face representation $\mathbf{e}^i_\mathbf{y}$ towards the target pose and expression. In other two SPADE blocks (orange box), the scale and bias map $\{\gamma^i, \beta^i\}_{\phi_{\mathbf{y}_{t_s}}, \phi_{\tilde{\mathbf{y}}_t}}$ are generated by fixed weights convoluted on $\phi_{\mathbf{y}_{t_s}}$ and $\phi_{\tilde{\mathbf{y}}_t}$, respectively. Then, we sum up the three denomalized features and upsample them with a transposed convolution. We repeat this non-linear combination module in all layers of G and apply a hyperbolic tangent activation layer at the end of G to output fake image $\hat{\mathbf{y}}_t$. Our experiment results (ablation study in Sec. 6) show that this parallel non-linear combination is more robust to the spatial-misalignment among the input images, especially when there is an apparent head motion.

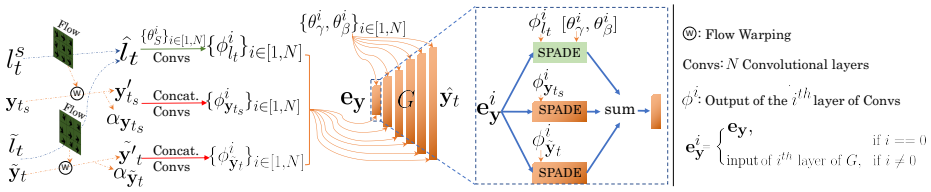


Fig. 6. The network structure of the non-linear composition module. The network weights $\{\theta_S^i, \theta_\gamma^i, \theta_\beta^i\}_{i\in[1,N]}$ and $\mathbf{e_y}$ are learned in embedding network (Sec. 4.2). To better encode $\hat{\boldsymbol{l}}_t$ with extra knowledge learned from the embedding network, we leverage $\{\theta_S^i\}_{i\in[1,N]}$ as the weights of the convolutional layers (green arrow) and save all the intermediate features $\{\phi_{\boldsymbol{l}_t}^i\}_{i\in[1,N]}$. Similarly, we obtain $\{\phi_{\tilde{\mathbf{y}}_t}^i, \phi_{\mathbf{y}_{t_s}}^i\}_{i\in[1,N]}$ with fixed convolutional weights (red arrow). Then, we input $\{\phi_{\tilde{\mathbf{y}}_t}^i, \phi_{\mathbf{y}_{t_s}}^i, \phi_{\boldsymbol{l}_t}^i\}_{i\in[1,N]}$ into the corresponding block of the decoder G to guide the image generation.

4.4 Objective Function

The Multi-Scale Discriminators [33] are employed to differentiate the real and synthesized video frames. The discriminators D_1 and D_2 are trained on two different input scales, respectively. Thus, the best generator G^* is found by solving the following optimization problem:

$$\begin{aligned} G^* = \min_G \Big(\max_{D_1,D_2} \sum_{k=1,2} \mathcal{L}_{\text{GAN}}(G, D_k) \\ + \lambda_{\text{FM}} \sum_{k=1,2} \mathcal{L}_{\text{FM}}(G, D_k) \Big) + \lambda_{\text{PCT}} \mathcal{L}_{\text{PCT}}(G) + \lambda_W \mathcal{L}_W(G) \ , \end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_W$ are the feature matching loss and flow loss proposed in [32]. The $\mathcal{L}_{\text{PCT}}$ is the VGG19 [25] perceptual loss term, which measures the perceptual similarity. The λ_{FM}, λ_{PCT}, and λ_W control the importance of the four loss terms.

5 Experiments Setup

Dataset. We quantitatively and qualitatively evaluate our approach on three datasets: LRW [10], VoxCeleb2 [9], and LRS3-TED [1]. The LRW dataset consists of 500 different words spoken by hundreds of different speakers in the wild and the VoxCeleb2 contains over 1 million utterances for 6,112 celebrities. The videos in LRS3-TED [1] contain more diverse and challenging head movements than the others. We follow same data split as [9,10,1].

Implementation Details[4]. We follow the same training procedure as [32]. We adopt ADAM [18] optimizer with $\text{lr} = 2 \times 10^{-4}$ and $(\beta_1, \beta_2) = (0.5, 0.999)$.

[4] All experiments are conducted on an NVIDIA DGX-1 machine with 8 32GB V100 GPUs. It takes 5 days to converge the training on VoxCeleb2/LRS3-TED since they contain millions of video clips. It takes less than 1 day to converge the training on the GRID/CREMA. For more details, please refer to supplemental materials.

Table 1. Quantitative results of different audio to video methods on LRW dataset and VoxCeleb2 dataset. We bold each leading score.

Method	Audio-driven							
Dataset	LRW				VoxCeleb2			
	LMD↓	CSIM↑	SSIM↑	FID↓	LMD↓	CSIM↑	SSIM↑	FID↓
Chung et al. [8]	3.15	0.44	**0.91**	132	5.4	0.29	0.79	159
Chen et al. [7]	3.27	0.38	0.84	151	4.9	0.31	0.82	142
Vougioukas et al. [30]	3.62	0.35	0.88	116	6.3	0.33	**0.85**	127
Ours (K=1)	**3.13**	**0.49**	0.76	**62**	**3.37**	**0.42**	0.74	**47**

Table 2. Quantitative results of different landmark to video methods on LRS3-TED and VoxCeleb2 datasets. We bold each leading score. K is the number of reference frames.

Method	Landmark-driven					
Dataset	LRS3-TED			VoxCeleb2		
	SSIM↑	CSIM↑	FID↓	SSIM↑	CSIM↑	FID↓
Wiles et al. [35]	0.57	0.28	172	0.65	0.31	117.3
Chen et al. [7]	0.66	0.31	294	0.61	0.39	107.2
Zakharov et al. [37] (K=1)	0.56	0.27	361	0.64	0.42	88.0
Wang et al. [31] (K=1)	0.59	0.33	**161**	0.69	**0.48**	59.4
Ours (K=1)	**0.71**	**0.37**	354	**0.71**	0.44	**40.8**
Zakharov et al. [37] (K=8)	0.65	0.32	284	0.71	**0.54**	62.6
Wang et al. [31] (K=8)	0.68	0.36	**144**	**0.72**	0.53	42.6
Ours(K=8)	**0.76**	**0.41**	324	0.71	0.50	**37.1**
Zakharov et al. [37](K=32)	0.72	0.41	154	0.75	0.54	51.8
Wang et al. [31](K=32)	0.69	0.40	132	0.73	**0.60**	41.4
Ours (K=32)	**0.79**	**0.44**	**122**	**0.78**	0.58	**35.7**

During training, we select $K = 1, 8, 32$ in our embedding network. The τ in all experiments is 64. The $\lambda_{\mathrm{FM}} = \lambda_{\mathrm{PCT}} = \lambda_{\mathrm{PCT}} = 10$ in Eq. 8.

6 Results and Analysis

Evaluation Metrics. We use several criteria for quantitative evaluation. We use Fréchet Inception Distance (FID) [16], mostly quantifying the fidelity of synthesized images, structured similarity (SSIM) [34], commonly used to measure the low-level similarity between the real images and generated images. To evaluate the identity preserving ability, same as [37], we use CSIM, which computes the cosine similarity between embedding vectors of the state-of-the-art face recognition network [11] for measuring identity mismatch. To evaluate whether the synthesized video contains accurate lip movements that correspond to the input condition, we adopt the evaluation matrix Landmarks Distance (LMD) proposed in [6]. Additionally, we conduct user study to investigate the visual quality of the generated videos including lip-sync performance.

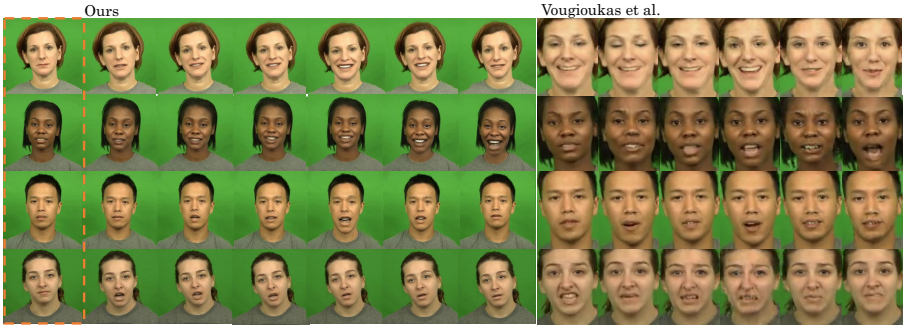


Fig. 7. Comparison of our model with Vougioukas et al. [30] on CREMA-D testing set.

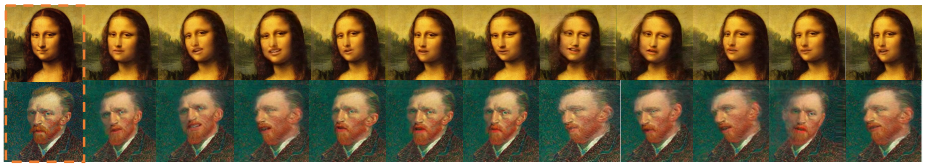


Fig. 8. We generate video frames on two real world images.

Comparison with Audio-driven Methods. We first consider such a scenario that takes audio and one frame as inputs and synthesizes a talking head saying the speech, which has been explored in Chung et al. [8], Wiles et al. [35], Chen et al. [7], Song et al. [26], Vougioukas et al. [30], and Zhou et al. [39]. Comparing with their methods, our method is able to generate vivid videos including controllable head movements and natural facial expressions. We show the videos driven by audio in the supplemental materials. Tab. 1 shows the quantitative results. For a fair comparison, we only input one reference frame. Note that other methods require pre-processing including affine transformation and cropping under a fixed alignment. And we do not add such constrains, which leads to a lower SSIM matrix value (e.g., complex backgrounds). We also compare the facial expression generation (see Fig. 7)[5] from audio with Vougioukas et al. [30] on CREMA-D dataset [3]. Furthermore, we show two generation examples in Fig. 8 driven by audio input, which demonstrates the generalizability of our network.

Comparison with Visual-driven Methods. We also compare our approach with state-of-the-art landmark/image-driven generation methods: Zakharov et al.[37], Wiles et al. [35], and Wang et al. [31] on VoxCeleb2 and LRS3-TED datasets. Fig. 9 shows the visual comparison, from where we can find that our methods can synthesize accurate lip movement while moving the head accurately. Comparing with other methods, ours performs much better especially when there is a apparent deformation between the target image and

[5] We show the raw output from the network in Fig. 7. Due to intrinsic limitations, [30] only generates facial regions with a fixed scale.

Fig. 9. Comparison on LRS3-TED and VoxCeleb2 testing set. We can find that our method can generate lip-synced video frames while preserving the identity information.

reference image. We attribute it to the 3D-aware module, which can guide the network with rough geometry information. We also show the quantitative results in Tab. 2, which shows that our approach achieves best performance in most of the evaluation matrices. It is worth to mention that our method outperforms other methods in terms of the CSIM score, which demonstrates the generalizability of our method. We choose different K here to better understand our matching scheme. We can see that with larger K value, the inception performance (FID and SSIM) can be improved. Note that we only select K images from the first 64 frames, and Zakharov et al. [37] and Wang et al. [31] select K images from the whole video sequence, which arguably gives other methods an advantage.

Fig. 10. (a)Ablation studies. (b) User studies.

Ablation Studies. We conduct thoughtful ablation experiments to study the contributions of each module we introduced in Sec. 3 and Sec. 4. The ablation studies are conducted on VoxCeleb2 dataset. As shown in Fig. 10(a), each component contributes to the full model. For instance, from the third column, we can find that the identity preserving ability drops dramatically if we remove the 3d-aware module. We attribute this to the valuable highly-aligned facial texture information provided 3D-aware module, which could stabilize the GAN training and lead to a faster convergence. Another interesting case is that if we

remove our non-linear composition block or warping operation, the synthesized images may contain artifacts in the area near face edges or eyes. We attribute to the alignment issue caused by head movements. Please refer to supplemental materials for more quantitative results on ablation studies.

User Studies. To compare the photo-realism and faithfulness of the translation outputs, we perform human evaluation on videos generated by both audio and landmarks on videos randomly sampled from different testing sets, including LRW [10], VoxCeleb2 [9], and LRS3-TED [1]. Our results including: synthesized videos conditioned on audio input, videos conditioned on landmark guidance, and the generated videos that facial expressions are controlled by audio signal and the head movements are controlled by our motion learner. Human subjects evaluation is conducted to investigate the visual qualities of our generated results compared with Wiles et al. [35], Wang et al. [31], Zakharov et al. [37] in terms of two criteria: whether participants could regard the generated talking faces as realistic and whether the generated talking head temporally sync with the corresponding audio. The details of User Studies are described in the supplemental materials. From Fig. 10(b), we can find that all our methods outperform other two methods in terms of the extent of synchronization and authenticity. It is worth to mention that the videos synthesized with head movements (ours, K=1) learned by the motion learner achieve much better performance that the one without head motion, which indicate that the participants prefer videos with natural head movements more than videos with still face.

7 Conclusion and Discussion

In this paper, we propose a novel approach to model the head motion and facial expressions explicitly, which can synthesize talking-head videos with natural head movements. By leveraging a 3d aware module, a hybrid embedding module, and a non-linear composition module, the proposed method can synthesize photo-realistic talking-head videos.

Although our approach outperforms previous methods, our model still fails in some situations. For example, our method struggles synthesizing extreme poses, especially there is no visual clues in the given reference frames. Furthermore, we omit modeling the camera motions, light conditions, and audio noise, which may affect our synthesizing performance.

Acknowledgement. This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. In: arXiv preprint arXiv:1809.00496 (2018)
2. Bregler, C., Covell, M., Slaney, M.: Video rewrite: Driving visual speech with audio. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. pp. 353–360 (1997)
3. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014)
4. Cassell, J., McNeill, D., McCullough, K.E.: Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. Pragmatics & cognition **7**(1), 1–34 (1999)
5. Chang, Y.J., Ezzat, T.: Transferable videorealistic speech animation. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 143–151. ACM (2005)
6. Chen, L., Li, Z., K Maddox, R., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 520–535 (2018)
7. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7832–7841 (2019)
8. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? In: British Machine Vision Conference (2017)
9. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
10. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision (2016)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
12. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
13. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
14. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer graphics forum. vol. 34, pp. 193–204. Wiley Online Library (2015)
15. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Computer Vision and Pattern Recognition (CVPR). IEEE (2019)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
17. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Liu, K., Ostermann, J.: Realistic facial expression synthesis for an image-based talking head. In: 2011 IEEE International Conference on Multimedia and Expo. pp. 1–6. IEEE (2011)
20. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: IEEE International Conference on Computer Vision (ICCV) (2019)
21. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. The IEEE International Conference on Computer Vision (ICCV) (Oct 2019)
22. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological science **15**(2), 133–137 (2004)
23. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
24. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: One-shot anatomically consistent facial animation. International Journal of Computer Vision pp. 1–16 (2019)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
26. Song, Y., Zhu, J., Li, D., Wang, A., Qi, H.: Talking face generation by conditional recurrent adversarial network. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 919–925. International Joint Conferences on Artificial Intelligence Organization (7 2019). https://doi.org/10.24963/ijcai.2019/129, https://doi.org/10.24963/ijcai.2019/129
27. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) **36**(4), 95 (2017)
28. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018)
29. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances In Neural Information Processing Systems. pp. 613–621 (2016)
30. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision pp. 1–16 (2019)
31. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
32. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
33. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

35. Wiles, O., Sophia Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 670–686 (2018)
36. Yoo, S., Bahng, H., Chung, S., Lee, J., Chang, J., Choo, J.: Coloring with limited data: Few-shot colorization via memory augmented networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11283–11292 (2019)
37. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
38. Zhou, H., Liu, J., Liu, Z., Liu, Y., Wang, X.: Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
39. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9299–9306 (2019)
40. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)