Depth-Aware Generative Adversarial Network for Talking Head Video Generation

Fa-Ting Hong[1] Longhao Zhang[2] Li Shen[2] Dan Xu[1*]

[1]Department of Computer Science and Engineering, HKUST [2]Alibaba Cloud

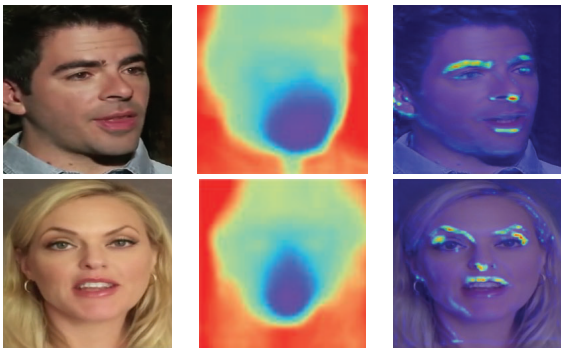
fhongac@cse.ust.hk, longhao.zlh@alibaba-inc.com, lshen.lsh@gmail.com, danxu@cse.ust.hk

Abstract

Talking head video generation aims to produce a synthetic human face video that contains the identity and pose information respectively from a given source image and a driving video. Existing works for this task heavily rely on 2D representations (e.g. appearance and motion) learned from the input images. However, dense 3D facial geometry (e.g. pixel-wise depth) is extremely important for this task as it is particularly beneficial for us to essentially generate accurate 3D face structures and distinguish noisy information from the possibly cluttered background. Nevertheless, dense 3D geometry annotations are prohibitively costly for videos and are typically not available for this video generation task. In this paper, we introduce a self-supervised face-depth learning method to automatically recover dense 3D facial geometry (i.e. depth) from the face videos without the requirement of any expensive 3D annotation data. Based on the learned dense depth maps, we further propose to leverage them to estimate sparse facial keypoints that capture the critical movement of the human head. In a more dense way, the depth is also utilized to learn 3D-aware cross-modal (i.e. appearance and depth) attention to guide the generation of motion fields for warping source image representations. All these contributions compose a novel depth-aware generative adversarial network (DaGAN) for talking head generation. Extensive experiments conducted demonstrate that our proposed method can generate highly realistic faces, and achieve significant results on the unseen human faces. [1]

1. Introduction

In this paper, we target the task of generating a realistic talking head video of a person using a source image of that person and a driving video, possibly derived from another person [27, 28, 31]. In the real world, a wide range of practical applications can be benefited from this task such as role-playing video games and virtual anchors.



(a) (b) (c)

Figure 1. Qualitative results of the learned depth maps (Fig. 1b) of the face images (Fig. 1a) using a self-supervised manner, and dense depth-aware attention maps (Fig. 1c), which can attend to important semantic parts of the face such as eyes.

Rapid progress has been achieved on talking head video generation in terms of both quality and robustness in recent years, using generative adversarial networks (GANs) [5]. A successful direction for the task in the literature focuses on decoupling identity and pose information from the face images [22, 24, 31]. For instance, pioneering works [22, 24] propose to model relative poses between two face images based on estimated sparse facial keypoints, and the poses are further used to generate dense motion fields, which warps the feature maps of the source image to drive the image generation. Similarly, Eurkov *et al*. [1] aimed to specifically learn two latent codes for the pose and the identity, and then input them into a designed generator network for face video synthesis. More than that, data augmentation strategies [1, 39] are also explored to more effectively perform the disentanglement of the pose and identity information. Although these methods show highly promising performance on the task, they still pay large attention to learning more representative 2D appearance and motion features from the input images. However, for face video generation, 3D dense geometry is critically important for the task while rarely investigated in the existing methods.

The dense 3D geometry (*e.g*. pixel-level depth) can bring

*Corresponding author

[1] https://github.com/harlanhong/CVPR2022-DaGAN

several significant benefits for the talking-head video generation. First, as the video captures the moving heads in a realistic 3D physical world, the 3D geometry can greatly facilitate an accurate recovery of 3D face structures, and the model capability of maintaining a realistic 3D face structure is a key factor for generating high-fidelity face videos. Second, the dense geometry can also help the model to robustly distinguish the noisy background information for generation especially under cluttered background conditions. Finally, the dense geometry is also particularly useful for the model to identify expression-related micro-movements on the faces. However, a severe issue of utilizing the 3D dense geometry to significantly boost the generation is that the 3D geometry annotations are highly expensive and typically not available for this task.

To address this problem, in this paper, we first propose to learn the pixel-wise depth map (see Fig. 1b) via geometric warping and photometric consistency in a *self-supervised* manner, to automatically recover *dense* 3D facial geometry from the training face videos, without requiring any expensive 3D geometry annotations. Based on the learned dense facial depth maps, we further propose two mechanisms to effectively leverage the depth information for better talking-head video generation. The first mechanism is depth-guided facial keypoint detection. The facial keypoints estimated by the network should well reflect the structure of the face, as they are further used to produce the motion field for feature warping, while the depth map explicitly indicates the 3D structure of the face. Thus, we combine geometry representations learned from the input depth maps with the appearance representations learned from the input images, to predict more accurate facial keypoints. The second mechanism is a cross-modal attention mechanism to guide the learning of the motion field. The motion field may contain noisy information from the cluttered background, and cannot effectively capture the expression-related micro-movements as they are generated from sparse facial keypoints. Therefore, we propose to learn depth-aware attention to have pixel-wise 3D geometry constraint on the motion field (see Fig. 1c), to drive the generation with more fine-grained details of facial structure and movements.

All the above-illustrated contributions compose a **D**epth-**a**ware **G**enerative **A**dversarial **N**etwork (**DaGAN**) to advance talking head video generation. Extensive experiments are conducted to qualitatively and quantitatively evaluate the proposed DaGAN model on two different datasets, *i.e*. VoxCeleb1 [18] and CelebV [27]. The experimental results show that our proposed self-supervised depth learning strategy can produce accurate depth maps on both the source and the target human face images. Our DaGAN model can also generate higher-quality face images compared with state-of-the-art methods. More specifically, our model is able to better preserve facial details, yielding a synthesized face with a more accurate expression and pose.

In summary, the main contribution is three-fold:

- In this work, we propose to introduce a self-supervised face-depth learning method to recover explicit dense 3D facial geometry (*i.e*. depth maps) from face videos for talking head video generation, and utilize the learned depth to boost the performance.
- We propose a novel depth-aware generative adversarial network for talking head generation, which effectively incorporates the depth information into the generation network via two carefully designed mechanisms, *i.e*. depth-guided facial keypoint estimation, and cross-modal (*i.e*. depth and image) attention learning.
- Extensive experimental results show accurate depth recovery of face images and also achieve superior generation performance compared with state-of-the-arts.

2. Related Works

Generative Adversarial Networks. The generative adversarial network (GAN) was introduced by Goodfellow *et al*. [5] to produce realistic images under certain conditions. GANs have attracted substantial attention and has been studied in many tasks [15], *e.g*., unconditional image synthesis [5, 11, 12, 19], text-to-image translation [20, 29, 34], and image inpainting [9, 13, 14]. In this work, we focus on talking head video generation with GAN guided by 3D facial depth maps learned from a self-supervised manner.

Depth Estimation. Many works have been proposed to tackle the problem of depth estimation from stereo images or video sequences [3, 4, 7, 16, 40]. Zhou *et al*. [40] use an end-to-end learning approach with view synthesis as the supervisory signal to estimate the depth map in monocular video sequences in an unsupervised manner. Based on [35], Clement *et al*. [4] gain a significant improvement using a minimum reprojection loss to deal with occlusions between frames and an auto-masking loss to ignore confusing stationary pixels. Gordon *et al*. [6] tried to learn camera intrinsics for every two consecutive frames to make the model able to perform inference in the wild.

To utilize the depth information of human faces, we introduce a self-supervised depth estimation method for the talking head generation task with only video images required. The depth map can provide dense 3D geometric information for the keypoint detection and can serve as an important cue to guide the model to focus on fine-grained critical parts of the human face (*e.g*. eyes, and mouth) during image generation.

Talking Head Video Generation. Talking Head Video Generation can be divided into three major strategies according to its driven-modality, *i.e*. image-driving methods [1, 22, 24, 26, 31, 36], landmark-driving methods [8, 32, 33, 37] and audio-driving methods [2, 23, 38, 39]. To exclude the driving face's identity information, several image-driving methods [22, 24] tried to predict keypoints of both

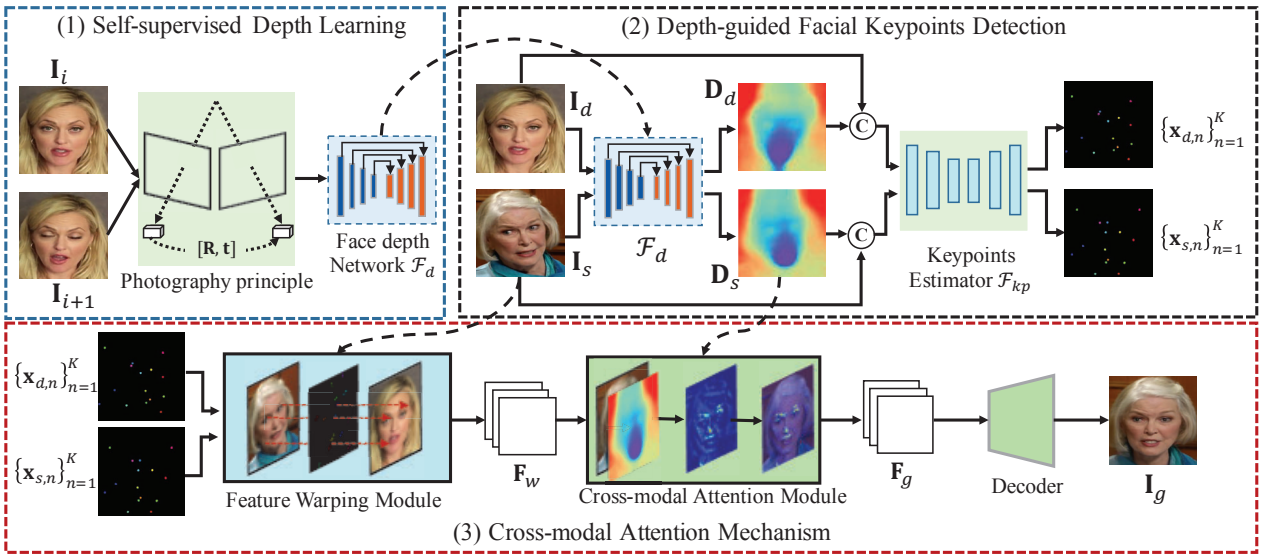


Figure 2. An illustration of the proposed DaGAN approach, which can be mainly divided into three sub-networks: (1) a self-supervised depth learning sub-network $\mathcal{F}_d$. We learn pixel-wise face depth maps in a self-supervised manner to recover the dense 3D facial geometry from the training face videos. (2) a depth-guided facial keypoints detection sub-network $\mathcal{F}_{kp}$. In this part, we combine both the geometry representations from depth maps with the appearance representations from the images to predict more accurate facial keypoints. (3) a cross-modal (*i.e.* depth and rgb image) attention learning sub-network. We learn dense depth-aware attention map using depth maps to constrain the motion field, to obtain a more accurate generation of fine-grained details of facial structure and movements.

the source image and driving image, and model local motion using the changes in the positions of corresponding keypoints. Similarly, Yao *et al*. [31] construct 3D meshes to remove the identity information from the driving images. Using facial landmarks instead of pure images to encode the pose information is an intuitive method. The fs-vid2vid [33] models person appearance by decomposing it into two layers, *i.e*. a pose-dependent coarse image and a pose-independent texture image. Zhao *et al*. [37] not only model global motions using full facial landmarks, but also use local landmarks to enforce the model to focus on local regions. The audio-driving method is a more popular way to perform face reenactment since the audio does not contain identity information, which can enable the model to more easily obtain a latent code of pose information from the audio. In [38], the encoder disentangles the pose information from identity information assisted by the audio modality. In both [38, 39], they argue that the audios and the images should share the same pose space.

In contrast to these existing works, we learn explicit pixel-wise depth map in a self-supervised manner, to provide highly beneficial 3D dense geometry information of the human faces, which allows the proposed model to accurately perceive 3D structures of the faces, and generate more fine-grained details of face spatial structures.

3. The proposed DaGAN Approach

Generating talking head videos is a technically challenging task as it requires the preserving of the identity information while imitating the facial motion from the driving faces. In this work, under the same setting as utilized in previous works [22, 31], we propose a depth-aware generative adversarial network, termed as DaGAN, for talking head video generation. It learns a depth estimation network in a self-supervised manner from training face videos, without requiring any expensive 3D geometry data as input. Thus, we can recover reliable face depth maps for both the input source and driving images to capture accurate 3D face structures and the expression-related micro-movements for higher-quality talking-head video generation.

3.1. Overview

Our proposed DaGAN approach consists of a generator and a discriminator. The core network architecture of our generator is depicted in Fig. 2, while the implementation of the discriminator is directly inspired from FOMM [22]. Our generation network can be split into three parts: (i) a self-supervised depth learning sub-network $\mathcal{F}_d$. The face depth network $\mathcal{F}_d$ first learns depth estimation using two consecutive frames (*i.e.* $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$) from a face video in a self-supervised manner. Then the whole deep framework is jointly trained while with $\mathcal{F}_d$ fixed. (ii) A depth-guided sparse keypoints detection sub-network $\mathcal{F}_{kp}$. Given a source image $\mathbf{I}_s$ and a driving image $\mathbf{I}_d$ from the driving video, we exploit $\mathcal{F}_d$ to produce depth maps ($\mathbf{D}_s$ and $\mathbf{D}_d$) for each image. These depth maps and corresponding RGB images are concatenated to learn geometry and appearance features for detecting face keypoints (*i.e.* $\{\mathbf{x}_{s,n}\}_{n=1}^{N}$ and $\{\mathbf{x}_{d,n}\}_{n=1}^{N}$), which can be used to generate relative motion fields of the human faces. (iii) The feature warping module accepts the keypoints as input to generate motion fields,

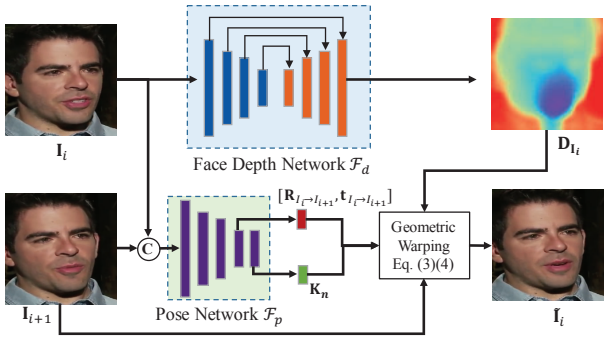


Figure 3. The training process of our face depth network. In addition to the face depth network, we use a pose network to estimate the relative camera poses $[\mathbf{R}_{I_i \to I_{i+1}}, \mathbf{t}_{I_i \to I_{i+1}}]$ and the camera intrinsic matrix $\mathbf{K}_n$. The symbol ⓒ represents the concatenated operation.

which are used to warp the source-image feature map to fuse with the appearance information, resulting in a warped feature $\mathbf{F}_w$. To enforce the model to focus on fine-grained details of face structures and micro-expression movements, we further learn a dense depth-aware attention map using the source depth map $\mathbf{D}_s$ and the warped feature $\mathbf{F}_w$. The depth-aware attention map can be used to refine the warped feature to produce a refined feature $\mathbf{F}_g$, resulting in a better generated image $\mathbf{I}_g$.

3.2. Self-supervised Face Depth Learning

In this part, we elaborate the technical details of the proposed self-supervised facial depth learning network, which can automatically recover dense face depth maps from the input source and driving images. Although SfM-Learner [40] previously proposed to learn *outdoor* scene depth in an unsupervised manner in an autonomous driving scenario, while in this work, we extend the method to learn face depths specifically for talking head video generation. Since the facial videos contain relative larger-area dynamic motion (moving head dominating on the image) compared to the outdoor scenes, unsupervised facial depth estimation is a challenging problem in our task.

We optimize the depth network using available training face videos. Specifically, given two consecutive video frames $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ from a face video, with $\mathbf{I}_{i+1}$ as a source image and $\mathbf{I}_i$ as a target image, we aim to learn several geometric elements, including a depth map $\mathbf{D}_{I_i}$ for the target image $\mathbf{I}_i$, a camera intrinsic matrix $\mathbf{K}_n$, n indicating the n-th input video in the training data, and a relative camera pose $\mathbf{R}_{I_i \to I_{i+1}}$ with translation $\mathbf{t}_{I_i \to I_{i+1}}$ between the two images. It should be noted that the camera intrinsic $\mathbf{K}_n$ is also not available in our training face video dataset, which is clearly different from [40] using provided camera intrinsics. $\mathbf{K}_n$ is input-video-clip-specifically learned in our method, as each face video can be possibly captured by any camera. So the input of our method only requires video frames.

The depth map $\mathbf{D}_{I_i}$ can be produced using the depth network $\mathcal{F}_d(\cdot)$. The pose $\mathbf{R}_{I_i \to I_{i+1}}$, the translation $\mathbf{t}_{I_i \to I_{i+1}}$, and the camera intrinsic matrix $\mathbf{K}_n$ are predicted from the same pose network $\mathcal{F}_p(\cdot)$ as follows:

$$\mathbf{D}_{I_i} = \mathcal{F}_d(\mathbf{I}_i), \tag{1}$$

$$[\mathbf{R}_{I_i \to I_{i+1}}, \mathbf{t}_{I_i \to I_{i+1}}], \mathbf{K}_n = \mathcal{F}_p(\mathbf{I}_i \,||\, \mathbf{I}_{i+1}), \tag{2}$$

where the symbol $||$ indicates a concatenation of the two images. Then, we can warp the target image $\mathbf{I}_i$ to the view of the source image $\mathbf{I}_{i+1}$ as follows:

$$\mathbf{q}_k \sim \mathbf{K}_n[\mathbf{R}_{I_i \to I_{i+1}} \,|\, \mathbf{t}_{I_i \to I_{i+1}}]\mathbf{D}_{I_i}(\mathbf{p}_j)\mathbf{K}_n^{-1}\mathbf{p}_j, \tag{3}$$

$$\widetilde{\mathbf{I}}_i = \mathcal{B}_{\mathcal{I}}(\mathbf{I}_{i+1}, \{\mathbf{q}_k\}_{k=1}^N), \tag{4}$$

where $\mathbf{q}_k$ and $\mathbf{p}_j$ denote the warped pixel on the source image $\mathbf{I}_{i+1}$ and an original pixel on the target image $\mathbf{I}_i$; N is the overall number of pixels of the image; $\mathcal{B}_{\mathcal{I}}(\cdot)$ is a differentiable bilinear interpolation function; $\widetilde{\mathbf{I}}_i$ is a reconstructed image at the source view. Therefore, we can construct a photometric consistency error $Pe(\cdot,\cdot)$ between $\widetilde{\mathbf{I}}_i$ and $\mathbf{I}_i$ to train our depth network in a self-supervised manner. Following [4], we use L1 and SSIM [25] to construct the photometric consistency error Pe as:

$$Pe(\mathbf{I}_i, \widetilde{\mathbf{I}}_i) = \alpha(1 - SSIM(\mathbf{I}_i, \widetilde{\mathbf{I}}_i)) + (1-\alpha)||\mathbf{I}_i - \widetilde{\mathbf{I}}_i||, \tag{5}$$

where α is set to 0.8 showing better optimization in our experiments. After training the framework, we only utilize the face depth network $\mathcal{F}_d$ in DaGAN to estimate the depth maps of input face images, which are further employed by our proposed mechanisms for talking head generation.

3.3. Motion Modeling by Sparse Keypoints

After we obtain the depth map from the face depth network, we concatenate the RGB image and its corresponding depth map produced by $\mathcal{F}_d$. Then, the keypoints estimator $\mathcal{F}_{kp}$ inputs the concatenated appearance (*i.e.* $\mathbf{I}_\tau$) and geometry (*i.e.* $\mathbf{D}_\tau$) information to more accurately predict a set of sparse keypoints of the human face:

$$\{\mathbf{x}_{\tau,n}\}_{n=1}^K = \mathcal{F}_{kp}(\mathbf{I}_\tau \,||\, \mathbf{D}_\tau), \tau \in \{s, d\}, \tag{6}$$

where K is the number of the detected face keypoint, and the subscript τ indicates a source image or a driving image; $||$ denotes a concatenation operation. We follow the design of [22] to implement our keypoints detector.

We adopt a feature warping strategy to capture head movements between the source and the target images, and implement a proposed feature warping module. Firstly, we compute a set of initial 2D offsets $\{\mathbf{O}_n\}_{n=1}^K$ for all the keypoints as follows:

$$\{\mathbf{O}_n\}_{n=1}^K = \{\mathbf{x}_{s,n}\}_{n=1}^K - \{\mathbf{x}_{d,n}\}_{n=1}^K. \tag{7}$$

Then, we generate a 2D dense coordinate map z similar to [22]. After that, a dense 2D motion field $\mathbf{w}_m$ is generated by adding the K offsets $\{\mathbf{O}_n\}_{n=1}^K$ into the 2D coordinate map at the corresponding coordinates of the K keypoints.

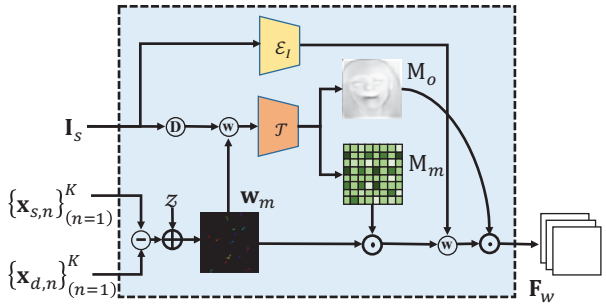


Figure 4. The illustration of our feature warping module. Here, Ⓓ is the downsampling operation, ⓦ is the warping operation, $\odot$ is the element-wise multiplication. The $\oplus$ and $\ominus$ represent the addition and subtraction operation, respectively.

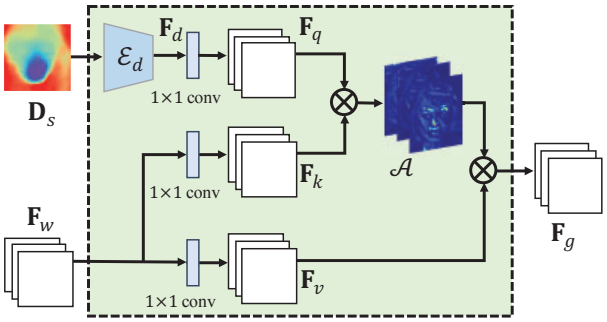


Figure 5. The illustration of our cross-modal attention module. Here, 1×1 convolutional layers do not share the parameters with each other, and the symbol $\otimes$ represents the matrix multiplication.

As shown in Fig. 4, we first utilize the dense 2D motion field $\mathbf{w}_m$ to warp the downsampled image to produce an initial warped feature map. After that, an occlusion estimator $\mathcal{T}$ take as input the initial warped feature map to predict a motion flow mask $\mathbf{M}_m$ and an occlusion map $\mathbf{M}_o$ [31]. The motion flow mask $\mathbf{M}_m$ assigns different confidence values for the estimated dense 2D motion field $\mathbf{w}_m$, resulting in masked motion field, while the occlusion map $\mathbf{M}_o$ aims to mask out the feature map regions that should be inpainted since the head has varying rotations. we utilize the masked motion field to warp the appearance feature map learned from the source image $\mathbf{I}_s$ extracted by the feature encoder $\mathcal{E}_I$. Then, they are fused with the occlusion map $\mathbf{M}_o$ to produce the warped soruce-image feature $\mathbf{F}_w$ as follows:

$$\mathbf{F}_w = \mathbf{M}_o * \mathcal{W}_p(\mathcal{E}_\mathcal{I}(\mathbf{I}_s), \mathbf{M}_m * \mathbf{w}_m), \tag{8}$$

where $\mathcal{W}_p$ is the warping function. In this way, the warped features $\mathbf{F}_w$ can better preserve the identity of the source image while maintaining the head motion information between two faces.

3.4. Cross-Modal Attention Module

To effectively embed the learned depth maps to boost the generation in a more dense way, we propose a cross-modal (*i.e.* depth and image) attention mechanism to enable the model to better preserve the facial structure and generate for expression-related micro facial movements, as the depth can provide us dense 3D geometry, which is essentially beneficial for maintaining the facial structure and identifying the critical movements when performing the generation. More specifically, we develop a cross-modal attention module to produce a dense depth-aware attention map to guide the warped feature $\mathbf{F}_w$ for face generation.

As shown in Fig. 5, a depth encoder $\mathcal{E}_d$ take a source depth map $\mathbf{D}_s$ as input to encode a depth feature map $\mathbf{F}_d$, and we perform linear projection on $\mathbf{F}_d$ and the warped source-image feature $\mathbf{F}_w$ into three latent feature maps $\mathbf{F}_q$, $\mathbf{F}_k$ and $\mathbf{F}_v$ by three different 1×1 convolutional layers with kernels $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$, respectively. The $\mathbf{F}_q$, $\mathbf{F}_k$ and $\mathbf{F}_v$ can respectively represent the query, key and value in the self-attention mechanism. Thus, the geometry-related query feature $\mathbf{F}_q$ produced by the depth map can be fused with the appearance-related key feature $\mathbf{F}_k$ to generate dense guidance for the human face generation. We obtain the final refined features $\mathbf{F}_g$ for generation:

$$\mathbf{F}_g = \mathrm{Softmax}\left((\mathbf{W}_q\mathbf{F}_d)(\mathbf{W}_k\mathbf{F}_w)^T\right) \times (\mathbf{W}_v\mathbf{F}_w), \tag{9}$$

where $\mathrm{Softmax}(\cdot)$ represents a softmax normalization function which outputs the dense depth-aware attention map $\mathcal{A}$ in Fig. 5. The $\mathcal{A}$ contains important 3D geometric guidance for generating the faces with more fine-grained details of facial structure and micro-movements. Finally, the decoder takes as input the refined warped features $\mathbf{F}_g$ to produce the final synthesized image $\mathbf{I}_g$.

3.5. Training

In the training stage, the identities of the source and the driving image are the same, while they can be different in the inference stage. Following the previous works [22, 24], we train the proposed DaGAN in a self-supervised manner by minimizing the following loss:

$$\begin{aligned}\mathcal{L} =& \lambda_P\mathcal{L}_P(\mathbf{I}_g, \mathbf{I}_d) + \lambda_G\mathcal{L}_G(\mathbf{I}_g, \mathbf{I}_d) \\ &+ \lambda_E\mathcal{L}_E(\{\mathbf{x}_{d,n}\}_{n=1}^K) \\ &+ \lambda_D(\mathcal{L}_D(\{\mathbf{x}_{s,n}\}_{n=1}^K) + \mathcal{L}_D(\{\mathbf{x}_{d,n}\}_{n=1}^K)).\end{aligned} \tag{10}$$

Perceptual loss $\mathcal{L}_P$. We minimize the perceptual loss [10] between the driving image $\mathbf{I}_d$ and the generated image $\mathbf{I}_g$, which has been effectively demonstrated being able to produce visually sharp outputs [22]. Moreover, we create an image pyramid for the driving image $\mathbf{I}_d$ and the generated image $\mathbf{I}_g$ to compute a pyramid perceptual loss.

GAN loss $\mathcal{L}_G$. We adopt the least-squares loss [17] as our adversarial loss. We use the discriminator to compute feature maps of different scales from the input image, and perform $\mathcal{L}_G$ on multiple levels as $\mathcal{L}_P$. We also minimize the discriminator feature matching loss [24].

Equivariance loss $\mathcal{L}_E$. For a valid keypoint, when applying a 2D transformation to the image, the predicted keypoint should change according to the applied transformation [22].

Thus, we utilize an equivariance loss $\mathcal{L}_E$ to ensure the consistency of image-specific keypoints.

Keypoints distance loss $\mathcal{L}_D$. In order to make the detected facial keypoints aovid crowded around a small neighbourhood, we employ a keypoints distance loss to penalize the model if the distance of two corresponding keypoints falls below a predefined threshold.

Overall, the first two terms ensure the generated image being similar to the ground-truth. The third one enforces the predicted keypoints to be consistent, while the last one constrains the keypoints not to be clustered together. The λ_P, λ_G, λ_E and λ_D are the hyper-parameters to allow for a balanced learning from those losses. More details about the losses are presented in the Supplementary Material.

4. Experiments

In this section, we conduct extensive experiments on two talking face datasets to evaluate our proposed method. More additional experiments results and video samples are reported in the Supplementary Material.

4.1. Dataset and Metrics

Dataset. We mainly conduct experiments on two talking head generation datasets (*i.e*. VoxCeleb1 [18] dataset and CelebV [27] dataset) in this work. We follow the test set sampling strategy of MarioNETte [8].

Metrics. In this work, several metrics are utilized to evaluate the quality of the generated images. Specifically, we use structured similarity (**SSIM**) and peak signal-to-noise ratio (**PSNR**) to evaluate the low-level similarity between the generated image and the driving image. Also, we adopt other three metrics, *i.e*. $\mathcal{L}_1$, Average Keypoint Distance (**AKD**), and Average Euclidean Distance (**AED**) proposed in [21] to evaluate the keypoint-based methods.

In cross-identity reenacting experiments, following the previous work [8], we adopt the **CSIM** to evaluate the quality of identity preservation between source images and generated images. **PRMSE** is utilized to evaluate the head poses, while **AUCON** for expression evaluation.

4.2. Implementation Details

The structure of the keypoints estimator is an hourglass network [30]. We use similar architectures as in [4] for implementing our depth and pose networks, while the decoder in the generator is the same as in [22]. The details of the structures of each sub-network in the proposed DaGAN is elaborated in Supplementary Material. For the optimization losses, we set λ_P = 10, λ_G =1, λ_E = 10, and λ_D = 10. We set the number of keypoints in DaGAN as 15. In the training stage, we first train our face depth network using consecutive frames from videos in VoxCeleb1, and we fix it during the training of the whole deep generation framework.

4.3. Comparison with State-of-the-art Methods

Self-reenactment. We first compare the face synthesis results where the source and driving images are of the same person, and report the results in Tab. 1. It can be observed that our DaGAN achieves the best results among all the compared methods. Compared with the other two keypoint-driven methods, *i.e*. FOMM [22] and OSFV [24], our DaGAN model achieves the most accurate head movements (1.22 of ours vs. 3.20 of FOMM, and 1.64 of OSFV, in PRMSE), which verifies that our depth-guided facial-keypoints estimation can better capture the motion of human heads. Regarding the facial expression, our method still obtains the highest score (0.939 in AUCON), meaning that our method can recover more fine-grained details of the face structures and micro-expression movements of the face. Also, our method produces the highest scores in both SSIM and PSNR, which demonstrates that our method can produce more realistic images compared with the most competitive methods. Additionally, we report the results on other three metrics proposed by [21] in Tab. 2. Our method obtains the best scores in these three metrics, clearly confirming our initial motivation that introducing the 3D depth maps can greatly benefit the keypoint-based generation.

Cross-identity reenactment. We also perform experiments on the CelebV dataset to exploit the cross-identity motion

Model	CSIM ↑	SSIM↑	PSNR↑	PRMSE ↓	AUCON↑
X2face [26]	0.689	0.719	22.537	3.26	0.813
NeuralHead-FF [33]	0.229	0.635	20.818	3.76	0.719
MarioNETte [8]	0.755	0.744	23.244	3.13	0.825
FOMM [22]	0.813	0.723	30.394	3.20	0.886
MeshG [31]	0.822	0.739	30.394	3.20	0.887
OSFV [24]	0.895	0.761	30.695	1.64	0.921
DaGAN (ours)	**0.899**	**0.804**	**31.220**	**1.22**	**0.939**

Table 1. Comparisons with state-of-the-art methods on the self-reenactment on the VoxCeleb1 dataset [18]. ↑ indicates larger is better, while ↓ indicates smaller is better.

Model	$\mathcal{L}_1$ ↓	AKD ↓	AED↓
X2face [26]	0.078	7.687	0.405
Monkey-Net [21]	0.049	1.878	0.199
FOMM [22]	0.043	1.294	0.140
OSFV [24]	0.043	1.620	0.153
DaGAN (ours)	**0.036**	**1.279**	**0.117**

Table 2. Comparisons with keypoint-based methods on self-reenactment on the VoxCeleb1 dataset [18]. ↓ smaller is better.

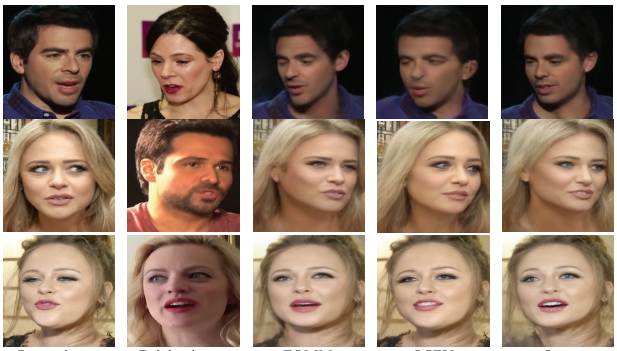


Figure 6. Qualitative comparisons of cross-identity reenactment on the VoxCeleb1 dataset [18].

Model	CSIM↑	PRMSE ↓	AUCON↑
X2face [26]	0.450	3.62	0.679
NeuralHead-FF [33]	0.108	3.30	0.722
marioNETte [8]	0.520	3.41	0.710
FOMM [22]	0.462	3.90	0.667
MeshG [31]	0.635	3.41	0.709
OSFV [24]	**0.791**	3.15	0.805
DaGAN (ours)	0.723	**2.33**	**0.873**

Table 3. Comparisons with state-of-the-art methods on cross-identity reenactment on CelebV dataset [27].

Figure 7. Qualitative comparisons of cross-identity reenactment on the CelebV dataset [27].

transfer, where the source and driving images are from different persons. We report the experimental results in Table 3. As we can observe that the PRMSE and AUCON of our DaGAN method remain the best among all methods, achieving 2.33 for PRMSE and 0.873 for AUCON. We also present several generated examples in Fig 6 and 7. As some methods do not release their code, we only show the results of those methods with available codes (*e.g*. FOMM and OSFV). For the seen faces in Fig. 6, our method produces face images with more fine-grained details than the others, for instance, the mouth and eyes regions in three rows. It verifies that the utilization of depth maps enables the model to identify micro-expression movements of the human faces. For the unseen targets in the CelebV dataset, we also show some generated samples in Fig. 7. Our method can also produce visually natural results for unseen targets. Notably, the generated images of OSFV in the first row is almost the same as the source image as it cannot detect the subtle motion on the face, which is also part of the reason why it outperforms our method in terms of CSIM in Tab. 3.

4.4. Ablation study

In this section, we conduct ablation studies to demonstrate the effectiveness of the proposed self-supervised face depth learning method and the proposed two mechanisms for talking head generation. We report results of ablation studies in Tab. 4, and show several qualitative examples of the generation results in Fig. 8 and Fig. 9. In Tab. 4, "**w/ CAM**" means applying cross-modal attention module to refine the warped feature, and "**w/ FDN**" indicates that

Model	CSIM↑	PRMSE ↓	AUCON↑
Baseline	0.688	5.39	0.657
Baseline w/ FDN	0.710	2.69	0.852
Baseline w/ CAM	0.698	2.56	0.838
DaGAN (ours)	**0.723**	**2.33**	**0.873**
FOMM	0.462	3.90	0.667
FOMM w/ FDN	0.695	2.81	0.812
FOMM w/ CAM	0.669	2.36	0.821
FOMM w/ FDN+CAM	**0.716**	**2.28**	**0.865**

Table 4. Ablation study. "Baseline" demotes the simplest model trained without the face depth network and cross-modal attention module. "Baseline w/ CAM" indicates that the baseline employs the cross-modal attention module after feature warping module, while "Baseline w/ FDN" combines the face depth network to estimate facial keypoints.

using the face depth network to estimate the face depth map for depth-guided sparse keypoints estimation. Here, our baseline is the simplest model trained without the depth map and depth attention module.

Dense face geometry recovery. We first show recovered depth maps for human faces from the proposed face depth network. Since we do *not* have ground-truth depths for the face images, it is tricky to directly evaluate the depth estimation quantitatively. Therefore, we only visualize the face depth maps in the third column of Fig. 8. The face depth maps are estimated from the driving images shown in the second column in Fig. 8. Our self-supervised geometry learning method can predict pixel-wise depth in the face images. Some other visualization of the face depth maps and their corresponding 3d point clouds is shown in Fig. 10. These visualization results strongly demonstrate that our proposed depth learning network is able to effectively recover the dense 3D geometry of human faces, which is clearly very beneficial, and directly embedded in the proposed model to learn both sparse facial keypoints and global pixel-wise dense attention for better generation.

Effectiveness of depth-guided keypionts. We aim to explore the impact of depth map on keypoints detection and report the related results in Tab. 4. From Tab. 4, we can easily recognize that the depth-guided keypoints helps our model gain significant gain in PRMSE and AUCON, which indicate that the depth map really plays a significant role in the talking head generation task. From Fig. 8, the "Baseline w/ FDN" predicts more accurate head orientation than "Baseline" (it can also be seen in Tab. 4, *i.e*. 2.69 vs. 5.39, in PRMSE), which indicates that the depth-guided facial keypoints can model accurate motions of the human heads.

Effectiveness of cross-modal attention module. From the Tab. 4 and Fig. 8, the cross-modal attention module (CAM) can clearly improve the generation quality of expression-related micro-movements of human faces. In Fig. 8, we can observe that the generated face results with the proposed CAM module (*i.e*. "Baseline w/ CAM") have more vivid expression than that of "Baseline w/ FDN" and "Baseline".

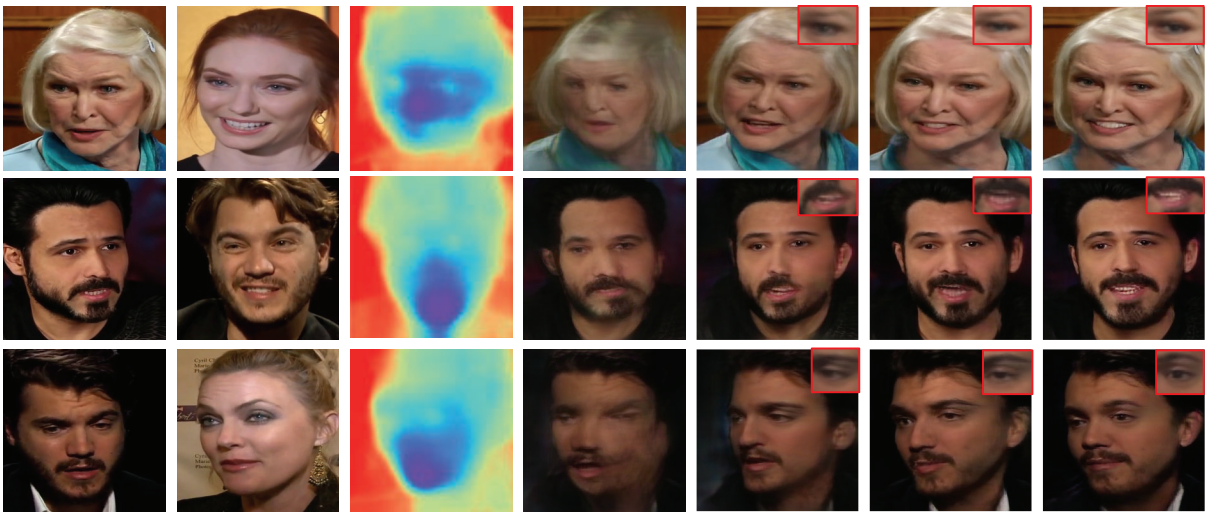


Figure 8. Qualitative ablation studies. Depth map and depth attention module can obtain improvements compared with baseline, while our full method produce the most realistic image.

Figure 9. Visualizing the dense depth-aware attention map in cross-modal attention module. In the last three columns, the red mark "×" indicates the query location.

It verifies that the CAM enables the model to capture the expression-related micro-movements at important facial regions (*e.g.*, human eyes and mouth). Additionally, the variance "Baseline w/ CAM" outperform "Baseline" by 0.181 in AUCON. The results in Tab. 4 and Fig. 8 verify that our proposed depth attention module can effectively utilize the depth map to enable model focus on micro-movement of the human face to boost the quality of the generated image.

Additionally, we visualize the dense depth-aware attention maps in Fig. 9. The high activation areas of each query point are mainly located in the expression-related parts of the human face, (*e.g.* eyes, nose, and mouth). These visualization results indicate that our designed cross-modal (*i.e.* depth and RGB) attention module can indeed address the micro-movements of the human face to produce more vivid expression in generation.

Plug-and-play experiments. Additionally, we also plug our proposed face depth network and depth-aware cross-modal attention module into FOMM [22], *i.e.*, using FOMM as a strong baseline, as our proposed modules can be flexibly deployed into existing video generation methods. The results are reported in Tab. 4. It is obvious that FOMM with the proposed modules can further achieve a significant improvement. These results fully demonstrate the effectiveness of learning dense 3D facial geometry (*i.e.* depth) for the talking head video generation task.

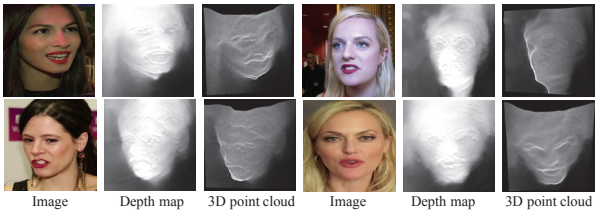


Figure 10. Visualization of estimated face depths and point clouds.

5. Conclusions

In this work, we proposed a depth-aware generative adversarial network (DaGAN) for talking head generation. DaGAN learns pixel-wise face depth maps in a self-supervised manner to recover dense 3D facial geometry. We also design two mechanisms to better leverage the depth for the generation. First, we combine the geometry from depth maps and appearance from RGB images to predict more accurate facial keypoints. Second, we design a cross-modal (*i.e.* depth and RGB) attention mechanism to capture the expression-related micro movements to produce more fine-grained details of facial structures. Ablation studies clearly show that depth maps can benefit the motion transfer between two faces. Our DaGAN also produces more realistic and natural-looking results compared to state-of-the-arts.

Acknowledgement This research is supported in part by the Early Career Scheme of the Research Grants Council (RGC) of the Hong Kong SAR under grant No. 26202321, HKUST Startup Fund No. R9253, and Alibaba Innovative Research Program.

References

[1] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, pages 13786–13795, 2020. 1, 2

[2] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, 2020. 2

[3] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *ICCV*, 2019. 2, 4, 6

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[6] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *ICCV*, 2019. 2

[7] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. In *CVPR*, 2016. 2

[8] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020. 2, 6, 7

[9] Jireh Jam, Connah Kendrick, Vincent Drouard, Kevin Walker, Gee-Sern Hsu, and Moi Hoon Yap. R-mnet: A perceptual adversarial network for image inpainting. In *WACV*, 2021. 2

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[13] Ang Li, Jianzhong Qi, Rui Zhang, and Ramamohanarao Kotagiri. Boosted gan with semantically interpretable information for image inpainting. In *IJCNN*. IEEE, 2019. 2

[14] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *CVPR*, 2021. 2

[15] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 2021. 2

[16] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *TOG*, 2020. 2

[17] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 5

[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 2, 6

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[20] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2

[21] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 6

[22] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[23] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 2

[24] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 1, 2, 5, 6, 7

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 4

[26] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. 2, 6, 7

[27] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018. 1, 2, 6, 7

[28] Runze Xu, Zhiming Zhou, Weinan Zhang, and Yong Yu. Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*, 2017. 1

[29] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2

[30] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017. 6

[31] Guangming Yao, Yi Yuan, Tianjia Shao, and Kun Zhou. Mesh guided one-shot face reenactment using graph convolutional networks. In *ACM MM*, 2020. 1, 2, 3, 5, 6, 7

[32] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020. 2

[33] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 2, 3, 6, 7

[34] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[35] Tianyang Zhang, Huazhu Fu, Yitian Zhao, Jun Cheng, Mengjie Guo, Zaiwang Gu, Bing Yang, Yuting Xiao, Shenghua Gao, and Jiang Liu. Skrgan: Sketching-rendering

unconditional generative adversarial networks for medical image synthesis. In *MICCAI*, 2019. 2

[36] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment. *arXiv preprint arXiv:1908.03251*, 2019. 2

[37] Ruiqi Zhao, Tianyi Wu, and Guodong Guo. Sparse to dense motion transfer for face image animation. In *ICCV*, 2021. 2, 3

[38] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. 2, 3

[39] 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, 3

[40] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 4