3D-Aware Talking-Head Video Motion Transfer

Haomiao Ni[1] Jiachen Liu[1] Yuan Xue[2] Sharon X. Huang[1]
[1]The Pennsylvania State University, University Park, PA, USA
[2]The Ohio State University, Columbus, OH, USA
[1]{hfn5052, jzl6493, suh972}@psu.edu [2]Yuan.Xue@osumc.edu

Abstract

Motion transfer of talking-head videos involves generating a new video with the appearance of a subject video and the motion pattern of a driving video. Current methodologies primarily depend on a limited number of subject images and 2D representations, thereby neglecting to fully utilize the multi-view appearance features inherent in the subject video. In this paper, we propose a novel 3D-aware talking-head video motion transfer network, Head3D, which fully exploits the subject appearance information by generating a visually-interpretable 3D canonical head from the 2D subject frames with a recurrent network. A key component of our approach is a self-supervised 3D head geometry learning module, designed to predict head poses and depth maps from 2D subject video frames. This module facilitates the estimation of a 3D head in canonical space, which can then be transformed to align with driving video frames. Additionally, we employ an attention-based fusion network to combine the background and other details from subject frames with the 3D subject head to produce the synthetic target video. Our extensive experiments on two public talking-head video datasets demonstrate that Head3D outperforms both 2D and 3D prior arts in the practical cross-identity setting, with evidence showing it can be readily adapted to the pose-controllable novel view synthesis task.

1. Introduction

The task of transferring motion between talking-head videos, while maintaining the identity of the target subject, is a compelling research area with broad applications in special effects, entertainment, and video editing. Despite the significant progress in guided image-to-image synthesis, such as person image generation [1, 25, 30] and facial expression generation [4, 21, 31], the challenge of capturing the temporal dynamics of motion in video-to-video transfer remains unsolved [5, 29]. Most current methods for talking-head video motion transfer use one subject image [35, 36, 38] or a simple combination of a few subject images [13, 28, 44, 45, 48] with 2D representations. These approaches may struggle to fully leverage the multi-view appearance information inherent in the subject video.

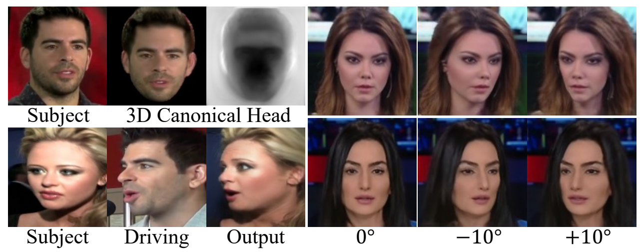

Figure 1. Illustration of advantages of our proposed Head3D. This 3D-aware framework can directly generate a 3D canonical head including RGB and depth map (top left), better deal with extreme poses (bottom left), and achieve pose-controllable novel view synthesis (right panel; changing yaw angle for the examples shown).

In this paper, we introduce *Head3D*, a novel 3D-aware framework for transferring motion between talking-head videos. This framework operates in a self-supervised, non-adversarial training manner, and is capable of recovering 3D structural information (*i.e.*, head pose and depth) from each 2D video frame through self-supervised 3D head geometry learning, without the need for a 3D graphical model of the human head. By mapping each selected subject video frame to a 3D canonical space, Head3D further estimates a 3D subject canonical head using a recurrent network. To synthesize the final video frames, Head3D employs an attention-based fusion mechanism to combine appearance features from the 3D subject head with background and other details (*e.g.*, facial expression, shoulder) from the subject. Unlike previous 3D-based methods that operate on the canonical feature space [6, 37, 46], Head3D offers visual interpretability by explicitly modeling the 3D canonical head. Compared with NeRF-based methods [7, 15, 26], Head3D shows better generalization ability without the need to retrain the model on unseen faces. Moreover, with the generated 3D subject head, Head3D can effectively handle large pose changes or extreme poses and achieve novel view synthesis with user-provided pose inputs, as demonstrated in Fig. 1. Our contributions are summarized as follows:

- We introduce Head3D, a 3D-aware generative network for talking-head video motion transfer, which explicitly estimates a 3D canonical head without the need for any 3D shape priors.
- We propose a self-supervised 3D head geometry learning module with a recurrent network to generate a 3D visually-interpretable canonical head from the 2D subject video.
- Comprehensive experiments demonstrate that our proposed Head3D outperforms other 2D- and 3D-based methods in the practical cross-identity motion transfer setting. Our model can also be easily adapted to pose-controllable novel view synthesis.

2. Related Work

According to whether 3D information is utilized during generation, talking-head video motion transfer methods can be categorized into 2D- or 3D-based frameworks.

2D-based talking-head video motion transfer. Based on whether to use multiple frames from the subject video, 2D-based methods can be further classified into one-shot [35, 36, 38, 41, 47, 54] and few-shot methods [13, 28, 44, 45, 48, 51, 52]. One-shot 2D methods, also known as image animation, focus on generating videos based on one given subject image and one driving video. Siarohin *et al*. [36] proposed a general self-supervised first-order-motion framework (FOMM) to predict dense motion flow for animating arbitrary objects with learned keypoints and local affine transformations. In [38], the authors further improved their network by modeling object movement through unsupervised region detection. Tao *et al*. [41] improved FOMM by introducing a deformable anchor model (DAM) to ensure that the object structure is well captured and preserved. However, these one-shot 2D methods are limited to using a single subject image, which makes it hard for them to utilize the multi-view appearance features of the subject when the subject video is available.

Few-shot 2D methods instead utilize the subject video more effectively by synthesizing a video based on a number of subject video frames. Wang *et al*. [45] proposed a video-to-video synthesis approach (vid2vid) under the generative adversarial learning framework [12], which produces one new video frame based on several previously generated images and the corresponding landmarks of the driving frame. In [44], they further proposed a few-shot vid2vid framework to learn how to synthesize videos of unseen subjects by leveraging a few example images of the target at test-time. Ha *et al*. [13] proposed a few-shot face reenactment framework, MarioNETte, which employed image attention block, target feature alignment, and landmark transformer to address unseen identity and large-pose gaps. While few-shot methods have shown promising performance by utilizing appearance information from multiple frames, they operate only on 2D features and thus fail to fully exploit the multi-view information available in subject videos.

3D-based talking-head video motion transfer. 3D-based models have seen substantial progress in recent years. Some recent methods [3, 9, 10, 19, 24, 40] incorporate predefined shape models (*e.g*., 3DMM [2] or FLAME [23]) to model 3D face for face manipulation. Liu *et al*. [24] proposed 3D-FM GAN for 3D-controllable face manipulation by encoding both the input face image and a physically-based rendering of 3D edits into the latent space of StyleGAN [18]. However, these methods depend on predefined 3D graphical models that may have limitations in modeling the unique shape details of different subjects. Other recent methods [7, 8, 15] instead used Neural Radiance Fields (NeRFs) [26] as a 3D representation of the human head. Gafni *et al*. [7] proposed dynamic neural fields for modeling the appearance and dynamics of a human face tracked by 3DMM [2]. However, it can be hard for these NeRF-based models to generalize to unseen subject videos and they require fine-tuning or retraining when applied to new subjects. Some other methods [6, 14, 42, 46] are based on 3D geometrical transformation. Wang *et al*. [46] proposed a one-shot free-view neural talking-head video synthesis model which represents a video using a sparse 3D keypoint representation. Hong *et al*. [14] introduced a self-supervised geometry learning method to automatically recover depth from face videos and leverage them to estimate sparse facial keypoints for talking head generation. Though also using 3D geometrical transformation, our proposed Head3D is different from these methods by explicitly modeling and visualizing the 3D canonical head estimated from the 2D subject video, thus providing an easily interpretable representation of the subject's head.

3. Methodology

Figure 2 shows the training framework of our Head3D. In general, Head3D is trained in an unsupervised manner, using self-reconstruction loss to restore one video frame with several randomly sampled frames from the same video. This training process neither requires any human annotation nor involves adversarial training. The training of Head3D consists of three stages: (1) 3D head geometry learning, (2) recurrent canonical head generation, and (3) attention-based fusion mechanism. To ease the training, we train the modules in these three stages separately. Given a set of randomly sampled N reference frames $\mathcal{S}_{\text{ref}} = \{s_1, s_2, \ldots, s_N\}$ and a driving frame s_{dri} from the same training video $\mathcal{S}$, in the first stage, we utilize a self-supervised 3D head geom-

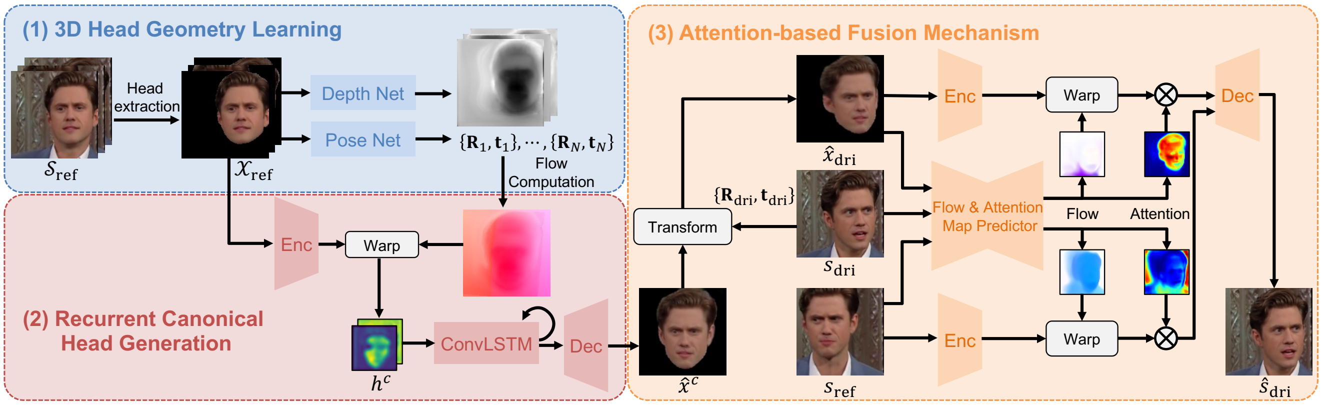

Figure 2. Illustration of the training framework of our proposed Head3D. During training, a set of reference subject frames $\mathcal{S}_{\text{ref}}$ and a driving frame s_{dri} are randomly sampled from the same video $\mathcal{S}$. The final output frame $\hat{s}_{\text{dri}}$ will be used to compute a self-reconstruction loss against the driving frame s_{dri} to train the network. Note that these three stages are trained separately.

etry learning framework to train a depth network F_{D} and a pose network F_{P} for predicting the head pose and depth of each 2D video frame. During the second stage, we use a recurrent canonical head generation network that leverages ConvLSTM-based feature aggregation to create a 3D canonical head $\hat{x}^c$ incorporating warped reference frame features. Finally, in stage three, we employ an attention-based fusion mechanism to synthesize each final output frame $\hat{s}_{\text{dri}}$ by combining head appearance from the canonical head $\hat{x}^c$, the background and other appearance details (*e.g.*, neck and shoulder) from one randomly selected subject frame s_{ref}, and motion and expression information from the driving frame s_{dri}. Details of each component of our proposed framework are introduced as follows. More implementation details can be found in Sec 4.2.

3.1. 3D Head Geometry Learning

Given a talking-head video $\mathcal{S}$, we first randomly sample a set of N reference frames $\mathcal{S}_{\text{ref}} = \{s_1, s_2, \ldots, s_N\}$ and one driving frame s_{dri} from $\mathcal{S}$. To recover the 3D geometry of the subject's head from a 2D talking-head video, we assume that videos are captured with a static perspective camera and that the subject's head can be treated as a rigid object. Our motivation is, by estimating a 3D head in canonical space, *i.e.*, a 3D canonical head, the head region of each target video frame can be reconstructed by transforming the points of the 3D canonical head using a rigid pose transformation $\mathbf{P} = \{\mathbf{R}, \mathbf{t}\} \in \text{SE}(3)$. To only reconstruct the head part in subject video frames, we employ a pretrained face parsing network [50] to extract the facial and hair regions. This results in a set of reference head images $\mathcal{X}_{\text{ref}} = \{x_1, x_2, \ldots, x_N\}$ and a driving head image x_{dri}.

As shown in Fig 2, after head extraction, we apply a depth estimation network F_{D} and a head pose prediction network F_{P} to each frame in $\mathcal{X}_{\text{ref}}$ and x_{dri} for estimating their depth maps $\mathcal{D}_{\text{ref}} = \{d_1, d_2, \ldots, d_N\}$, d_{dri}, and their head poses $\mathcal{P}_{\text{ref}} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\}$, $\mathbf{P}_{\text{dri}}$. For each reference frame x_i in X_{ref}, where $i = 1, \ldots, N$, we compute the corresponding canonical frame x_i^c based on the image formation model in [49]. Let pixel $q = (u, v, 1)$ be the homogeneous coordinate of one pixel in the reference frame x_i, and pixel $q^c = (u^c, v^c, 1)$ be the corresponding pixel in the canonical frame x_i^c. We can transform each pixel q to q^c to generate canonical frame x^c by:

$$q^c \propto K(\mathbf{R}^T(d[u,v] \cdot K^{-1}q - \mathbf{t})) \ , \tag{1}$$

where $d[u, v]$ is the depth value of pixel (u, v) in the depth map d, $\{\mathbf{R}, \mathbf{t}\}$ is the head pose of frame x_i, and K is the camera intrinsic matrix, which can be computed by:

$$K = \begin{pmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{pmatrix} \ , \quad \begin{cases} c_u = \dfrac{W-1}{2} \\ c_v = \dfrac{H-1}{2} \\ f = \dfrac{W-1}{2\tan\frac{\theta_{\text{FOV}}}{2}} \end{cases} , \tag{2}$$

where H and W are the height and width of image, θ_{FOV} is the field of view of the perspective camera. Following [49], we assume $\theta_{\text{FOV}} \approx 10^\circ$ and the nominal distance of the subject from the camera is about 1m. To simplify the training, we take the average of all the warped canonical frames x_i^c to produce the final canonical frame $\bar{x}^c$. We also apply F_{D} to obtain its depth map $\bar{d}^c$. Similar to Eq. (1), we can transform each pixel q^c in the canonical frame to the pixel q in the target frame by:

$$q \propto K(d^c[u^c, v^c] \cdot \mathbf{R}K^{-1}q^c + \mathbf{t}) \ , \tag{3}$$

where $q = (u, v, 1)$ is the homogeneous coordinate of one pixel in the target frame. By applying F_{P} to driving head frame x_{dri} to estimate the head pose $\mathbf{P}_{\text{dri}} = \{\mathbf{R}_{\text{dri}}, \mathbf{t}_{\text{dri}}\}$, using Eq. (3), we can transform the canonical frame $\bar{x}^c$ to frame $\bar{x}_{\text{dri}}$ with $\mathbf{P}_{\text{dri}}$. Then we can train depth network F_{D} and pose network F_{P} with the following head reconstruction loss function:

$$l_{\text{geo}} = ||\bar{x}_{\text{dri}} - x_{\text{dri}}||_1 + \lambda_{\text{sym}}\mathcal{L}_{\text{sym}}(\bar{x}^c) + \lambda_{\text{D}}\mathcal{L}_{\text{D}}(\bar{d}^c) \ , \tag{4}$$

where $\mathcal{L}_{\text{sym}}$ is designed to ensure the estimated 3D head under the canonical pose by imposing symmetry constraint. Here $\mathcal{L}_{\text{sym}} = ||\bar{x}^c - \bar{x}^{c'}||_1$, where $\bar{x}^{c'}$ is the horizontally-flipped version of $\bar{x}^c$. $\mathcal{L}_{\text{D}}$ is the depth smoothness loss used in [11]. λ_{sym} and λ_{D} are balancing coefficients.

3.2. Recurrent Canonical Head Generation

The canonical head image $\bar{x}^c$ is computed by averaging each transformed reference head frame in $\mathcal{X}_{\text{ref}}$. Thus $\bar{x}^c$ is often blurry and not ready for the subsequent target frame generation. To produce a high-quality fine-grained canonical head, we propose a novel recurrent canonical head generation network to combine transformed reference frames. As shown in Fig. 2, for each reference head frame x_i, we utilize head image encoder E_{H} to encode x_i as feature h_i and also use its corresponding depth d_i and pose $\mathbf{P}_i$ to compute backward optical flow $f_{R_i \leftarrow C}$ (*i.e.*, warping from canonical head x^c to reference head x_i) by:

$$f_{R_i \leftarrow C}[u, v] = (u, v)^T - (u^{\text{c}}, v^{\text{c}})^T \ , \tag{5}$$

where $(u^{\text{c}}, v^{\text{c}})$ is one pixel in x^c and (u, v) is the corresponding warped pixel in x_i by applying Eq. (3) to each $(u^{\text{c}}, v^{\text{c}})$ with $\mathbf{P}_i$. Here we adopt the backward warping operation because it can be implemented efficiently in a differentiable manner using bilinear sampling [16].

We later apply flow $f_{R_i \leftarrow C}$ to warp reference head feature h_i to h_i^c. Then we employ a Convolutional LSTM [34] module Λ to aggregate all h_i^c to generate the final canonical head feature h^c. A head image decoder Ω_{H} is finally used to decode feature h^c to be canonical head $\hat{x}^c$. Then we can apply F_{D} to $\hat{x}^c$ and combine the estimated depth $\hat{d}^c$ with $\hat{x}^c$ to form a 3D canonical head. This 3D head helps to fully utilize the multi-view appearance information provided by different reference frames. By applying Eq. (3) to $\hat{x}^c$ using the driving head pose $\mathbf{P}_{\text{dri}} = \{\mathbf{R}_{\text{dri}}, \mathbf{t}_{\text{dri}}\}$, we transform $\hat{x}^c$ to the estimated driving head frame $\hat{x}_{\text{dri}}$. So we can train head image encoder E_{H}, image decoder Ω_{H}, and ConvLSTM Λ by the following head reconstruction loss function:

$$l_{\text{head}} = ||\hat{x}_{\text{dri}} - x_{\text{dri}}||_1 \ . \tag{6}$$

3.3. Attention-based Fusion Mechanism

Because of modeling with rigid transformation, the estimated canonical head $\hat{x}^c$ can only describe movements of the whole head. To model facial expressions as well as the appearance and motion of non-head regions such as the shoulder and background, we propose an attention-based fusion mechanism to combine $\hat{x}^c$, a randomly selected reference frame s_{ref}, and the motion and expression from driving frame s_{dri} to produce the final target video frame $\hat{s}_{\text{dri}}$. As Fig. 2 shows, we first transform the estimated canonical head $\hat{x}^c$ to driving head $\hat{x}_{\text{dri}}$ using the driving head pose $\mathbf{P}_{\text{dri}} = \{\mathbf{R}_{\text{dri}}, \mathbf{t}_{\text{dri}}\}$. We then employ a frame encoder E_{F} to represent $\hat{x}_{\text{dri}}$ and s_{ref} as features $\hat{e}_{\text{dri}}$ and e_{ref}. We also design a flow and attention map predictor Φ, to which $\hat{x}_{\text{dri}}$, s_{ref} and s_{dri} are fed, in order to estimate two backward warping feature flows $f_{\hat{x}_{\text{dri}} \leftarrow s_{\text{dri}}}$ and $f_{s_{\text{ref}} \leftarrow s_{\text{dri}}}$, and three attention maps $a_{\hat{x}_{\text{dri}}}$, $a_{s_{\text{ref}}}$ and a_{dec}. Then the combined feature e_{out} can be computed by:

$$\begin{aligned} e_{\text{out}} =& a_{\hat{x}_{\text{dri}}} \odot \mathcal{W}(\hat{e}_{\text{dri}}, f_{\hat{x}_{\text{dri}} \leftarrow s_{\text{dri}}}) + a_{s_{\text{ref}}} \odot \mathcal{W}(e_{\text{ref}}, f_{s_{\text{ref}} \leftarrow s_{\text{dri}}}) \\ &+ a_{\text{dec}} \odot e_{\text{dec}} \ , \end{aligned} \tag{7}$$

where $\mathcal{W}(\cdot, \cdot)$ is backward warping, and $f_{\hat{x}_{\text{dri}} \leftarrow s_{\text{dri}}}$ is used for warping estimated head $\hat{x}_{\text{dri}}$ to s_{dri} for adding facial expression to $\hat{x}_{\text{dri}}$. $f_{s_{\text{ref}} \leftarrow s_{\text{dri}}}$ is used for warping reference frame s_{ref} to s_{dri} for providing the background and other appearance details. e_{dec} is the intermediate feature from decoder for synthesizing unseen novel regions. Attention maps $a_{\hat{x}_{\text{dri}}}$, $a_{s_{\text{ref}}}$, and a_{dec} are designed to indicate which parts in the feature maps can be kept and which parts should be masked out. The sum of the attention weights for corresponding pixels in the three attention maps should be equal to 1. Finally we employ a frame decoder Ω_{F} to decode feature e_{out} to target frame $\hat{s}_{\text{dri}}$. $\hat{s}_{\text{dri}}$ should be identical to s_{dri} and thus we can train the frame encoder E_{F}, flow and attention map predictor Φ, the frame decoder Ω_{F} using the following frame reconstruction loss:

$$l_{\text{frame}} = \mathcal{L}_{\text{rec}}(\hat{s}_{\text{dri}}, s_{\text{dri}}) \ , \tag{8}$$

where $\mathcal{L}_{\text{rec}}$ is the loss measuring the difference between reconstructed frame $\hat{s}_{\text{dri}}$ and ground truth frame s_{dri}. Per [36, 38], we implement $\mathcal{L}_{\text{rec}}$ using the perceptual loss [17] based on pretrained VGG network [39] and also add the equivariance loss [36] to stabilize the training.

3.4. Inference

During testing, given one subject video $\mathcal{S}$ and one driving video $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$, we first randomly sample one reference image s_{ref} and a set of reference frames $\mathcal{S}_{\text{ref}}$ from video $\mathcal{S}$, and estimate 3D canonical head $\hat{x}^c$ from $\mathcal{S}_{\text{ref}}$ through our proposed recurrent canonical head generation framework. Then for each driving frame y_i in $\mathcal{Y}$, we adopt our attention-based fusion mechanism to combine $\hat{x}^c$, s_{ref}, and y_i to generate the corresponding novel frame $\hat{s}_i$. The final target video $\hat{\mathcal{S}} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_M\}$ is generated in a frame-by-frame manner.

Pose-controllable novel view synthesis. Our proposed Head3D can be easily adapted to the pose-controllable novel view synthesis task by manually inputting the desired pose transformation, $\mathbf{P}_{\text{dri}} = \{\mathbf{R}_{\text{dri}}, \mathbf{t}_{\text{dri}}\}$, to generate the novel view $\hat{x}_{\text{dri}}$ from the canonical head representation, $\hat{x}^c$, rather than obtaining $\mathbf{P}_{\text{dri}}$ from the driving frame s_{dri}. Then instead of inputting a s_{dri} to the flow and attention map predictor Φ, we input $\hat{x}_{\text{dri}}$ to the predictor, and the final output frame $\hat{s}$ will have the pose $\mathbf{P}_{\text{dri}}$.

4. Experiments

4.1. Datasets and Metrics

Datasets. We conduct comprehensive experiments on two public datasets: *VoxCeleb* dataset [27] and *FaceForensics* dataset [32]. The VoxCeleb dataset includes 22,496

Figure 3. Qualitative comparison with state-of-the-art methods (FOMM [36], MRAA [38], DAGAN [14] and FaceV2V [46]). The top two rows are the results of self-reconstruction and the bottom two rows are that of cross-identity transfer. Artifacts and unnatural details are highlighted with blue boxes.

videos downloaded from Youtube. To simplify the training, we only keep 7,500 videos for training and 400 videos for testing. The FaceForensics dataset contains 1,004 videos of news briefings from different reporters. We find that models trained on the VoxCeleb dataset can be generalized well to this new dataset. So we only randomly choose 150 videos for testing without any additional training. Following the preprocessing approach in [36], we crop videos in these datasets to mainly keep the head regions and resize all video frames to 128×128. Since the original videos in these datasets are long, we randomly select a short segment of 40 continuous frames from each video and use these selected short videos in our experiments.

Metrics. Following [36], we compute metrics based on two testing settings, *self-reconstruction* and *cross-identity transfer*. For self-reconstruction, we segment a video of the same subject into two non-overlapping clips. We use one clip as the subject video and the other as the driving video. In this setting, the driving video serves as ground truth. Similar to [7, 47], we compute the normalized mean L_2 distance and Learned Perceptual Image Patch Similarity (LPIPS) [53] metrics between self-reconstructed results and driving videos. For cross-identity transfer, which is more practical in real-world applications, subject video and driving video are of different subjects in this setting. As there is no ground truth available, we conduct a paired user study to compare our model with state-of-the-art methods. Specifically, we generated 100 videos for each baseline method on each dataset and paired them with videos produced by our model. We then invited 10 human evaluators to make judgments regarding the better video in each pair, considering aspects such as visual realism, motion accuracy, and identity consistency.

4.2. Implementation

Model Implementation. We employ a public pretrained face parsing network[1] to extract the head regions (face and hair) from each video frame. For 3D head geometry learning, we implement the depth network F_{D} with a similar architecture used in [20]. To stabilize the training, we add instance normalization layer [43] to the decoder of F_{D}. We adopt a similar architecture to HopeNet [33] for the pose network F_{P} in our implementation. The original HopeNet only predicts the yaw, pitch, and roll of the head (*i.e.*, $\mathbf{R}$). To enable estimation of the 3D head translation $\mathbf{t}$, we modified the final layer of the network. To accelerate the training, we initialize most parameters in F_{D} and F_{P} with pretrained models provided in [20] and [33]. For the recurrent canonical head generation, we choose the architecture in [17] to implement the head image encoder E_{H} and decoder Ω_{H} with 2 downsampling blocks. We employ a one-layer ConvLSTM [34] to implement Λ. In our attention-based fusion mechanism, we also construct the frame encoder E_{F} and decoder Ω_{F} using the same architecture as E_{H} and Ω_{F}. The flow and attention map predictor Φ is built based on the

[1] https://github.com/zllrunning/face-parsing.PyTorch

Figure 4. Examples of generated talking-head videos using our proposed Head3D. For each block, Head3D synthesizes the new video (3rd row) with the appearance from the subject video (1st row) and motions from the driving video (2nd row).

flow predictor in MRAA [38]. We slightly change its architecture to enable the prediction of three attention maps.

As mentioned in Sec. 3, the whole training process of Head3D includes three separate stages. In the first stage, we train the depth network F_{D} and pose network F_{P} through 3D head geometry learning. In the second stage, we train the head image encoder E_{H}, head decoder Ω_{H}, and ConvLSTM Λ for the recurrent canonical head generation. We finally train frame encoder E_{F}, frame decoder Ω_{F}, and flow and attention map predictor Φ for the attention-based fusion mechanism in the third stage. We set batch size as 5 videos and use the Adam optimizer [22] with $(\beta_1, \beta_2) = (0.5, 0.999)$ during all three-staged training. Unless otherwise specified, the number of reference frames is set to 5. During 3D head geometry learning, we train F_{D} and F_{P} for 10 epochs. The learning rate of F_{D} and F_{P} is 2×10^{-4} and 2×10^{-5} and drops by 0.1 at epoch 5. The balancing parameter λ_{sym} and λ_{D} in Eq. 4 are all set to be 0.1. When training recurrent canonical head generation, we train E_{H}, Ω_{H} and Λ for 20 epochs with the learning rate of 2×10^{-4} and drop learning rate by 0.1 at epoch 10. We train the attention-based fusion modules (E_{F}, Ω_{F} and Φ) for 50 epochs with a fixed learning rate of 2×10^{-4}.

Baseline Implementation. We compare the proposed Head3D with three state-of-the-art motion transfer baseline models: 2D-based methods FOMM [36] and MRAA [35], and 3D-based methods DAGAN [14] and FaceV2V [46]. We follow the default settings in the methods' original implementations wherever possible[2] and retrain all the baselines using the same training videos on the VoxCeleb dataset as ours with the same 128×128 resolution.

4.3. Result Analysis

Comparison with state-of-the-art methods. We compare our Head3D with state-of-the-art (SOTA) methods under the self-reconstruction setting in Table 1. As shown in Table 1, Head3D achieves comparable or better performance when compared with the SOTA methods. While MRAA [38] performs better in most metrics under the self-

[2]Due to the lack of official implementation, we implement FaceV2V with the code from https://github.com/zhanglonghao1992/One-Shot_Free-View_Neural_Talking_Head_Synthesis.

Figure 5. Examples of pose-controllable novel view synthesis. Each column demonstrates changing of a 3D rotation or translation parameter.

Table 1. Comparison of proposed Head3D with state-of-the-art methods under the self-reconstruction setting on VoxCeleb and FaceForensics datasets.

Dataset	Method	$L_2 \downarrow$	LPIPS↓
VoxCeleb	FOMM [36]	0.0114	0.0856
	MRAA [38]	**0.0108**	**0.0830**
	DAGAN [14]	0.0123	0.0885
	FaceV2V [46]	0.0186	0.0994
	Head3D (Ours)	0.0113	0.0855
FaceForensics	FOMM [36]	0.0102	0.0543
	MRAA [38]	**0.0075**	0.0449
	DAGAN [14]	0.0106	0.0490
	FaceV2V [46]	0.0119	0.0509
	Head3D (Ours)	0.0079	**0.0442**

Table 2. User preferences in the paired study: our approach *vs.* state-of-the-art methods under cross-identity setting on VoxCeleb and FaceForensics datasets.

Methods	VoxCeleb (%)	FaceForensics (%)
Ours/FOMM [36]	**72**/28	**68**/32
Ours/MRAA [38]	**57**/43	**59**/41
Ours/DAGAN [14]	**80**/20	**86**/14
Ours/FaceV2V [46]	**53**/47	**54**/46

reconstruction setting, our proposed Head3D outperforms it under the more practical cross-identity setting as shown in Table 2. Under the self-reconstruction setting, we speculate that the advantage of using the 3D canonical head in Head3D may not be apparent, as the head motion and pose changes are limited due to the subject and driving videos being clipped from the same original video. When applied to the cross-identity motion transfer task, which typically involves larger head movements, Head3D benefits from leveraging the multi-view appearance information from the 3D canonical head, as is also shown in Fig. 3 and Fig. 4. More importantly, different from 2D-based FOMM and MRAA, Head3D can be easily applied to pose-controllable novel view synthesis, as shown in Fig. 1 and Fig. 5. Additionally, unlike 3D-based DAGAN and FaceV2V, the canonical head representation in Head3D is visually interpretable, as shown in Fig. 1 and Fig. 6.

Ablation Study. To analyze the effectiveness of each module in Head3D, we conduct an ablation study on the VoxCeleb dataset. Table 3 shows quantitative comparison results of the ablation study under the self-reconstruction setting. We first evaluate the effect of using different numbers of reference frames N. Since ConvLSTM can utilize different number of reference frames during training and testing, in our experiments, we specifically train a model with 5 reference frames and then evaluate its performance with different number of reference frames during testing. Compared with our final model with 5 reference frames, using fewer frames ($N = 1, 3$) generated worse results while increasing the number of frames ($N = 10$) can lead to better LPIPS but also longer inference time. So we choose $N = 5$ as our default setting. To evaluate the effectiveness of proposed recurrent canonical head generation, we compare our model with [Head3D w/ $\bar{x}^c$], which employs the mean canonical head $\bar{x}^c$ instead of the $\hat{x}^c$ generated by the recurrent network to synthesize the final frames. One can observe that using mean canonical head $\bar{x}^c$ noticeably diminishes performance. The reason may be that $\bar{x}^c$ is generated by simply taking the average of all the canonical head images warped from reference frames, which makes it blurry and not capturing some important details. We also experiment with removing the canonical head input $\hat{x}_{\text{dri}}$ during attention-based fusion and evaluate this variant model [Head3D w/o $\hat{x}_{\text{dri}}$]. Without using the appearance informa-

Figure 6. Illustration of the effect of attention maps in our proposed attention-based fusion mechanism. The red regions indicate a higher degree of attention, while the blue regions suggest a lower degree of attention. “Deformed” refers to applying warping flow $f_{s_{\text{ref}} \leftarrow s_{\text{dri}}}$ to the subject frame and “Transformed” means applying driving pose $\mathbf{P}_{\text{dri}}$ to the canonical head.

Table 3. Ablation Study under the self-reconstruction setting on VoxCeleb dataset.

Methods	$L_2 \downarrow$	LPIPS↓
Head3D ($N = 1$)	0.0117	0.0873
Head3D ($N = 3$)	0.0115	0.0880
Head3D ($N = 10$)	0.0116	**0.0839**
Head3D w/ $\bar{x}^c$	0.0117	0.0897
Head3D w/o $\hat{x}_{\text{dri}}$	0.0117	0.0872
Head3D ($N = 5$)	**0.0113**	0.0855

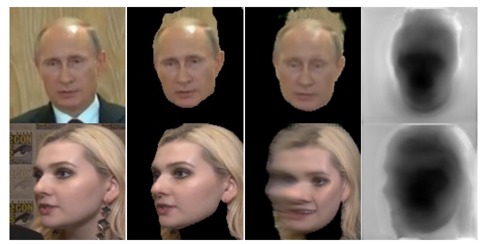

Figure 7. Examples of failure cases in canonical head estimation.

tion from $\hat{x}_{\text{dri}}$, the performance of [Head3D w/o $\hat{x}_{\text{dri}}$] decreases as Table 3 shows.

We also illustrate the effectiveness of our proposed attention-based fusion mechanism by visualizing some examples of attention maps in Fig. 6. As shown in Fig. 6, when a significant pose difference exists between the subject and driving frames, as in the first row, our model will assign higher attention values to the transformed canonical head to synthesize facial areas. In cases where the poses are more similar, such as in the second row, our model will combine information from both the subject frame and canonical head to generate the facial regions.

5. Limitation and Discussion

Head3D can achieve promising performance in most cases (see Fig. 4 and Supp. videos). However, it still suffers from several limitations. First, our current framework employs an off-the-shelf face parsing network to segment the head regions from video frames. Imprecise segmentation performed by the pretrained network may result in inconsistent or incorrect extraction of head regions, which could further adversely impact the estimation of the 3D canonical head (see the 1st row in Fig. 7). Second, when the subject video only provides a single side-view of the person, it can be challenging to generate a high-quality canonical head (see the 2nd row in Fig. 7). Currently, our proposed attention-based fusion mechanism can mitigate these limitations by assigning lower attention values to incorrect details of the canonical head, thereby reducing their influence on the final synthesized output. In future work, we will investigate the use of a more robust pretrained face parsing network or incorporate an unsupervised face parsing model into the current framework to enable end-to-end training. Recently, there has been a growing interest in high-resolution video generation [6]. We have provided a Supp. video at the 256×256 resolution, produced by our Head3D trained with different size parameters. In our subsequent research, we will also explore the video generation at the megapixel resolution such as 512×512.

6. Conclusion

In this paper, we present *Head3D*, a novel 3D-aware approach for transferring motion in talking-head videos. Head3D capitalizes on the multi-view appearance information inherent in a 2D subject video by estimating a 3D canonical head using a recurrent network. We introduce a self-supervised 3D geometry learning module to predict pose and depth map, and an attention-based fusion network to generate the final synthesized video. The explicit modeling of a 3D canonical head in Head3D allows for easy application to novel view synthesis tasks using user-provided pose inputs. Comprehensive experiments on two public talking-head datasets demonstrate the state-of-the-art video motion transfer capabilities of Head3D.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 1

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[3] Marcel C Bühler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13890–13899, 2021. 2

[4] Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13518–13527, 2020. 1

[5] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020. 1

[6] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 1, 2, 8

[7] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 1, 2, 5

[8] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 2

[9] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9821–9830, 2019. 2

[10] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020. 2

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 4

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

[13] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10893–10900, 2020. 1, 2

[14] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 2, 5, 6, 7

[15] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 1, 2

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 4

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 4, 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[19] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 2

[20] Faisal Khan, Shahid Hussain, Shubhajit Basak, Joseph Lemley, and Peter Corcoran. An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data. *Neural Networks*, 142:479–491, 2021. 5

[21] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 1

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2

[24] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and SY Kung. 3d-fm gan: Towards 3d-controllable face manipulation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 107–125. Springer, 2022. 2

[25] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017. 1

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1, 2

[27] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 4

[28] Haomiao Ni, Yihao Liu, Sharon X. Huang, and Yuan Xue. Cross-identity video motion retargeting with joint transformation and synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 412–422, January 2023. 1, 2

[29] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 1

[30] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018. 1

[31] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 1

[32] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 4

[33] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 5

[34] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 4, 5

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 1, 2, 6

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *arXiv preprint arXiv:2003.00196*, 2020. 1, 2, 4, 5, 6, 7

[37] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4658–4669, 2023. 1

[38] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 1, 2, 4, 5, 6, 7

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[40] Fariborz Taherkhani, Aashish Rai, Quankai Gao, Shaunak Srivastava, Xuanbai Chen, Fernando de la Torre, Steven Song, Aayush Prakash, and Daeil Kim. Controllable 3d generative adversarial face model via disentangling shape and appearance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 826–836, 2023. 2

[41] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3637–3646, 2022. 2

[42] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022. 2

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[44] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 1, 2

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 1, 2

[46] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1, 2, 5, 6, 7

[47] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 2, 5

[48] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 1, 2

[49] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 3

[50] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:3051–3068, 2021. 3

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

16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pages 524–540. Springer, 2020. 2

[52] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468, 2019. 2

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

[54] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 2