HifiHead: One-Shot High Fidelity Neural Head Synthesis with 3D Control

Feida Zhu[1] , **Junwei Zhu**[1] , **Wenqing Chu**[1] , **Ying Tai**[1,*] , **Zhifeng Xie**[2] , **Xiaoming Huang**[1] and **Chengjie Wang**[1,*]

[1]Youtu Lab, Tencent
[2]Shanghai Film Academy, Shanghai University

Abstract

We propose HifiHead, a high fidelity neural talking head synthesis method, which can well preserve the source image's appearance and control the motion (*e.g.*, pose, expression, gaze) flexibly with 3D morphable face models (3DMMs) parameters derived from a driving image or indicated by users. Existing head synthesis works mainly focus on low-resolution inputs. Instead, we exploit the powerful generative prior embedded in StyleGAN to achieve high-quality head synthesis and editing. Specifically, we first extract the source image's appearance and driving image's motion to construct 3D face descriptors, which are employed as latent style codes for the generator. Meanwhile, hierarchical representations are extracted from the source and rendered 3D images respectively to provide faithful appearance and shape guidance. Considering the appearance representations need high-resolution flow fields for spatial transform, we propose a coarse-to-fine style-based generator consisting of several feature alignment and refinement (FAR) blocks. Each FAR block updates the dense flow fields and refines RGB outputs simultaneously for efficiency. Extensive experiments show that our method blends source appearance and target motion more accurately along with more realistic results than previous state-of-the-art approaches.

1 Introduction

Neural talking head synthesis generates images with the appearance including identity, texture and lighting from a source face and the motion (*e.g.*, pose, expression, gaze) from a driving image, which has attracted considerable interest due to great potential usage in computer games and film industry.

Existing methods mainly focus on neural head synthesis for 256^2 resolution inputs and leverage generative adversarial networks (GAN) with conditions such as facial boundary [Wu *et al.*, 2018], keypoints [Siarohin *et al.*, 2019b] or 3D morphable face models (3DMM) parameters [Ren *et al.*, 2021]. Recent approaches typically follow a paradigm of flow prediction, warping and refinement. For example, [Siarohin *et al.*, 2019b; Siarohin *et al.*, 2019a] first perform keypoint detection on the driving image and then generate dense flow fields from sparse keypoints. After that, the source image is warped and further fed into a generator network for refinement. However, such approaches are not applicable for high resolution scenarios since it is very challenging to directly predict accurate 512^2 flow fields or synthesize high-quality images with common generator architecture.

Recently, StyleRig [Tewari *et al.*, 2020b] and PIE [Tewari *et al.*, 2020a] have been investigated to manipulate the latent style codes of StyleGAN [Karras *et al.*, 2019] to achieve high fidelity image editing. Unfortunately, the results are not faithful to the *desired expression or posture*, and could not preserve *high fidelity source appearance* due to the limited capacity of latent codes.

To overcome the above defects, we propose an elegant one-stage framework, termed HifiHead, to generate high fidelity talking heads via 3D face representations and generative prior embedded in StyleGAN. First, we combine the appearance-related 3DMM coefficients from the source image with the motion-related coefficients from the driving image to construct 3D face descriptors, which play the role of latent codes in StyleGAN. To better preserve texture details and rich spatial information, we further extract hierarchical representations from the source and rendered 3D images respectively to provide accurate appearance and shape guidance.

Next, we introduce a coarse-to-fine style-based generator to fuse the appearance and shape representations along with the latent codes to render a photo-realistic output. The latent codes are broadcasted to all generator blocks to modulate the convolutional weights. The hierarchical representations are injected into the generator through spatial concatenation progressively. Considering the appearance representations need high-resolution flow fields for spatial warping, we introduce feature alignment and refinement (FAR) blocks to better exploit the facial texture and geometry details. Different from existing works splitting the flow estimation and feature refinement into two stages, the FAR block utilizes the appearance and shape features as input, generating dense flow fields and refined features simultaneously for efficiency. Note the dense flow fields are not only leveraged to warp the appearance features in the next module, but also serve as the base estimates

* Ying Tai and Chengjie Wang are corresponding authors. https://github.com/TencentYoutuResearch/HeadSynthesis-HifiHead

Src & Drive NTHS [Wang *et al.*, 2021a] PIRenderer [Ren *et al.*, 2021] HifiHead (Ours)

Figure 1: Comparison with state-of-the-art methods. Our HiFiHead is able to generate realistic face details and accurate motion.

to alleviate higher resolution flow field learning in a residual manner. Extensive experiments demonstrate our results surpass other state-of-the-art neural head synthesis methods on wild face images with extreme pose, expression, or illumination. Our contributions are summarized as follows:

1. *High-precise*: We incorporate 3DMM parameters and 3D rendered images to provide strong motion constraint for head driven. The FAR block estimates accurate optical flow for better modeling the motion, preserving better facial textures and identity in the final results.
2. *High-resolution*: Benefiting from the StyleGAN-based generative architecture, HifiHead easily deals with faces with 512×512 resolution, higher than FOMM or PIRender along with better source texture preservation.
3. *Practical*: Our HifiHead drives real-world talking heads using only one source image. StyleRig cannot be applied for real-world image editing directly, while DynamicNeRF requires video sequence as training data for each source person.

2 Related Work

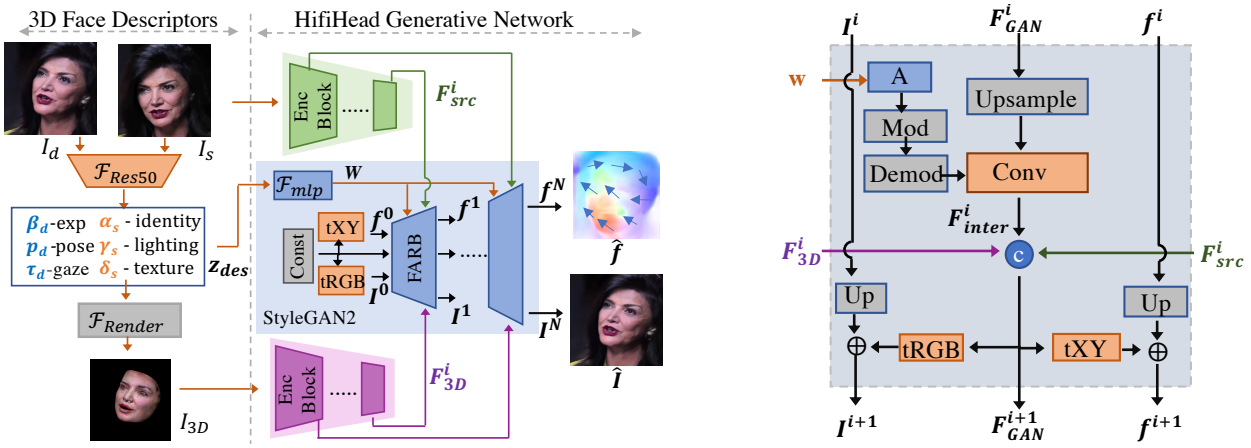
2D-based Head Synthesis. A large number of 2D-based approaches [Zakharov *et al.*, 2019; Ha *et al.*, 2020; Zakharov *et al.*, 2020] have been proposed to control head poses and facial expressions with conditional generative adversarial networks. ReenactGAN [Wu *et al.*, 2018] feeds an adapted facial boundary into a target-specific decoder for generating the reenacted target face. X2face [Wiles *et al.*, 2018] first generates a frontalised source face given an image set and then further transforms it to ensure desired pose and expression. However, requiring specific decoders [Wu *et al.*, 2018] or an image set [Wiles *et al.*, 2018] is not convenient in real-world applications and thus one shot head synthesis has drawn much attention [Siarohin *et al.*, 2019b; Yao *et al.*, 2021] recently. Several attempts [Siarohin *et al.*, 2019a; Siarohin *et al.*, 2021] generate dense flow fields from sparse keypoints detected from the driving image. After that, the input image is warped and refined to produce animated results. Unfortunately, synthesizing with a single source image is challenging especially for extreme poses and expressions.

3D Model-based Head Synthesis. 3D morphable models provide disentangled representations in respect of identity, shape, and texture of facial images, serving as plausible priors in many head synthesis algorithms [Ren *et al.*, 2021; Wang *et al.*, 2021b]. PIRenderer [Ren *et al.*, 2021] and HeadGAN [Doukas *et al.*, 2021] both utilize 3D face representations to compute dense flow fields for spatially transforming the source image and then generate a photo-realistic output. Instead of using an explicit 3D graphics model, [Wang *et al.*, 2021a] represents motion information with a 3D keypoint representation and warps source image feature in 3D space to handle extreme poses and expressions. However, these methods are mainly designed for 256^2 resolution and could not preserve detailed facial appearance.

StyleGAN-based Head Synthesis. In recent years, style-based generators [Karras *et al.*, 2019] have greatly boost the performance of high-quality image generation and applied in a large variety of applications, including face restoration [Yang *et al.*, 2021] and portrait attribute editing [Chu *et al.*, 2020]. StyleRig [Tewari *et al.*, 2020b] trains a translation network between the 3DMM's semantic parameters and StyleGAN's latent codes, thus can control over face pose and expressions of synthetically created StyleGAN images effectively. Furthermore, PIE [Tewari *et al.*, 2020a] embeds real portrait images in the latent space of StyleGAN for intuitive editing. However, due to the limited capacity of latent codes and lack of finetune, the results are unfaithful to source image's appearance. In contrast, our HifiHead generates high-resolution dense flow fields to transform the hierarchical representations of the source image and produces satisfactory results confirming to *desired appearance, pose and expression*.

3 Methodology

The overall architecture of HifiHead is depicted in Figure 2a. Given a source image I_s and a driving image I_d, the proposed HifiHead can transfer the expressions, pose and gaze of I_d to the source person while preserving the appearance attributes of I_s, such as identity, lighting and textures. In this section, we firstly introduce the 3D face descriptors. Then, we illustrate the modules of HifiHead in detail. Finally, we present the model objectives used to train our network.



(a) Overall Architecture (b) Feature Alignment and Refinement Block

Figure 2: The overall architecture of our method.

3.1 3D Face Descriptors

Following the practice of D3DFR [Deng *et al.*, 2019b], we leverage a deep neural network (*ResNet50*) to predict 3DMM coefficients from the input images:

$$z = \mathcal{F}_{Res50}(I). \tag{1}$$

The output is a vector $z = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p}, \boldsymbol{\tau}) \in \mathbb{R}^{261}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}, \boldsymbol{\beta} \in \mathbb{R}^{64}, \boldsymbol{\delta} \in \mathbb{R}^{80}, \boldsymbol{\gamma} \in \mathbb{R}^{27}, \mathbf{p} \in \mathbb{R}^{6}$ and $\boldsymbol{\tau} \in \mathbb{R}^{4}$ represent the coefficients of the identity, expression, texture, illumination, pose and gaze. The original 3DMM can not control gaze directions. We add gaze coefficient which is defined as the normalized direction vector from the center of the eye to the pupil. The desired synthesized result should maintain the appearance related attributes of I_s while having the motion related attributes of I_d. Therefore, the 3D face descriptors of the desired result are expressed by $z_{des} = \{\boldsymbol{\alpha}_s, \boldsymbol{\beta}_d, \boldsymbol{\delta}_s, \boldsymbol{\gamma}_s, \mathbf{p}_d, \boldsymbol{\tau}_d\}$.

Next, we compute the 3D face shape $\mathbf{S}$ and the albedo texture $\mathbf{T}$ with 3DMM,

$$\mathbf{S} = \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha}_s + \mathbf{B}_{exp}\boldsymbol{\beta}_d, \tag{2}$$

$$\mathbf{T} = \mathbf{T}(\boldsymbol{\delta}) = \bar{\mathbf{T}} + \mathbf{B}_t\boldsymbol{\delta}_s, \tag{3}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ denotes the mean face shape and albedo texture. $\mathbf{B}_{id}, \mathbf{B}_{exp}$ and $\mathbf{B}_t$ are the bases of identity, expression and texture. The reconstructed 3D face is further projected onto the 2D image plane with a differentiable renderer according to the predicted illumination $\boldsymbol{\gamma}_s$ and pose $\mathbf{p}_d$,

$$I_{3d} = \mathcal{F}_{render}(\mathbf{S}, \mathbf{T}, \boldsymbol{\gamma}_s, \mathbf{p}_d). \tag{4}$$

Please refer to [Deng *et al.*, 2019b] for more details.

In summary, given a source image I_s and a driving image I_d, the 3D face descriptors of the desired result is defined as $z_{des} = \{\boldsymbol{\alpha}_s, \boldsymbol{\beta}_d, \boldsymbol{\delta}_s, \boldsymbol{\gamma}_s, \mathbf{p}_d, \boldsymbol{\tau}_d\}$. The 3D face descriptors together with the rendered 3D face image I_{3d} are used to condition the image generation.

3.2 HifiHead Generative Network

We carefully modify the StyleGAN2 generative architecture for effective neural head synthesis and editing.

Input. As shown in Figure 2a, the 3D face descriptors are mapped to the latent space $w \in \mathcal{W}$ of StyleGAN2,

$$w = \mathcal{F}_{mlp}(z_{des}). \tag{5}$$

The mapping network depth is 3 as [Karras *et al.*, 2021]. In order to produce faithful appearance of the source identity, we condition the generative model on the multi-scale spatial features extracted from the source image I_s. Besides, we also extract spatial features from I_{3D} to provide target motion information in addition to the 3D face descriptors,

$$F^i_s = \mathcal{F}^s_{enc}(I_s), \tag{6}$$

$$F^i_{3D} = \mathcal{F}^{3D}_{enc}(I_{3D}). \tag{7}$$

The original StyleGAN2 requires noise inputs in each block. Instead, we replace that with spatial features F^i_{src} and F^i_{3d}.

Architecture. Directly concatenating F^i_{src} and F^i_{3d} is not reasonable since the expression and pose are mismatched spatially. It has been proven more effective to align the source image with the desired pose and expression through spatial deformations [Ren *et al.*, 2021]. Different from existing works [Ren *et al.*, 2021; Doukas *et al.*, 2021] which both train a warping network and a refine network separately in two stages, HifiHead network estimates the spatial deformations and generates the final result simultaneously in one single stage. With this in mind, we modify the original StyleGAN2 generative block such that each block has two output layers. The tRGB layer and tXY layer convert the high-dimensional features to RGB output and dense flow, respectively. The detailed structure will be illustrated in the next section.

To summarize, the HifiHead full generative network G contains the following trainable parts, the spatial feature encoder $\mathcal{F}^s_{enc}$ and $\mathcal{F}^{3D}_{enc}$, the mapping network $\mathcal{F}_{mlp}$ and the generative blocks $\mathcal{F}_{gen}$. Given the source image I_s, the 3D face descriptors z_{des} and the rendered 3D image I_{3D}, the HifiHead network generates the dense flow prediction $\hat{f}$ and image prediction $\hat{I}$ as follows,

$$\hat{f}, \hat{I} = G(I_s, z_{des}, I_{3D}). \tag{8}$$

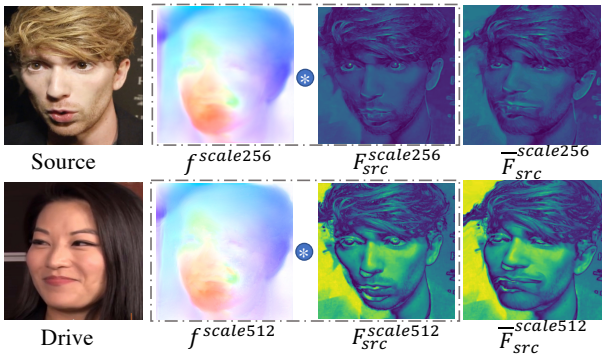


Figure 3: Visualization of the predicted flow field f and source features F_{src} at scale 256 and 512. The warped features $\bar{F}_{src}$ align better with the driving image.

3.3 Feature Alignment and Refinement Block

The detailed feature alignment and refinement (FAR) block structure is illustrated in Figure 2b. The i-th FAR block takes six inputs, including $I^i, f^i, F^i_{GAN}, F^i_s, F^i_{3d}$ and w. The RGB output I^i, dense flow f^i and the high-dimensional generative features F^i_{GAN} are generated by the previous block. F^i_s and F^i_{3d} denote the corresponding spatial features extracted from I_s and I_{3d}. w represents the latent code.

First, the source feature F^i_{src} is warped to $\bar{F}^i_{src}$ according to the predicted flow filed f^i. As shown in Figure 3, we visualize the predicted flow field at scale 256 and scale 512. The warped source feature aligns better with the driving image. Then, the generative feature F_{GAN} is updated as follows:

$$F^i_{inter} = \mathbf{StyleConv}(F^i_{GAN}|w), \tag{9}$$

$$F^{i+1}_{GAN} = \mathbf{Concat}(F^i_{inter}, \bar{F}^i_{src}, F^i_{3D}), \tag{10}$$

where the operation $\mathbf{StyleConv}$ denotes the style convolution in StyleGAN2. The definition of "Mod" and "Demod" can be found in [Karras *et al.*, 2020].

The RGB output I and dense flow f are updated with upsampling and skip connections. We use bilinear filtering in all up operations. The updated RGB output and dense flow are formulated as:

$$I^{i+1} = \mathbf{tRGB}(F^{i+1}_{GAN}) + \mathbf{UP}(I^i), \tag{11}$$

$$f^{i+1} = \mathbf{tXY}(F^{i+1}_{GAN}) + \mathbf{UP}(f^i), \tag{12}$$

where tRGB and tXY are both 1×1 convolutional layers.

3.4 Model Objectives

Previous methods [Siarohin *et al.*, 2019a; Wang *et al.*, 2021a] sample two frames from the same video, one as the source image I_s and the other as the driving image I_d. I_d also serves as the ground truth (same-identity). We observe that their motion transfer performance deteriorates when the source and driving images are different persons (cross-identity). However, cross-identity motion transfer has a wider range of practical applications. Motivated by this observation, our training data is divided into two categories:

Same-identity Data. The widely-used L1 loss $\mathcal{L}_1$ and perceptual loss $\mathcal{L}_p$ [Johnson *et al.*, 2016] are adopted as our reconstruction loss. Adversarial loss $\mathcal{L}_{adv}$ is inherited from StyleGAN2. We also construct the 3D mesh $\hat{\mathbf{S}}$ of the generated image $\hat{I}$. The mesh loss $\mathcal{L}_{mesh}$ is introduced to minimize the vertices' distance between $\hat{\mathbf{S}}$ and the target mesh $\mathbf{S}$, formulated by Equation 2. The identity loss $\mathcal{L}_{ID}$ is to preserve identity during motion transfer. The pretrained face recognition model CurricularFace [Huang *et al.*, 2020] is adopted to extract deep identity features from I_s and $\hat{I}$.

The overall model objective for same-identity training data is a combination of the above losses:

$$\mathcal{L}_{SI} = \mathcal{L}_1 + \mathcal{L}_p + \mathcal{L}_{adv} + \mathcal{L}_{mesh} + \mathcal{L}_{ID}. \tag{13}$$

Cross-identity Data. We randomly sample two frames from different videos. The reconstruction loss is not applicable since there is no ground-truth. $\mathcal{L}_{adv}$ is still adopted since it is helpful in producing realistic results. In addition, $\mathcal{L}_{mesh}$ is used to encourage the generated image to have the same expression and pose as the driving image. $\mathcal{L}_{ID}$ preserves the identity information from the source image.

The overall model objective for cross-identity training data is defined as follows:

$$\mathcal{L}_{CI} = \mathcal{L}_{adv} + \mathcal{L}_{mesh} + \mathcal{L}_{ID}. \tag{14}$$

In a mini-batch, the ratio of same-identity data and cross-identity data is empirically set to $1:1$.

Figure 4: Visual comparison of same-identity reconstruction. The source and driving images are shown in the first column. Please zoom in to better see the differences.

Method	PSNR↑	LPIPS↓	Exp↓	Angle↓	Gaze↓	ID↑	FID↓
FOMM	23.25	0.1261	2.77	0.0232	0.0540	0.8521	50.00
NTHS	23.60	0.1103	2.80	0.0268	0.0940	0.8611	44.80
PIRenderer	21.38	0.1367	3.05	0.0511	0.0900	0.8173	47.45
Ours	23.54	0.0956	2.05	0.0216	0.0422	0.9165	36.64

Table 1: Comparison on same-identity reconstruction task.

4 Experiment

4.1 Datasets and Implementation

Training Datasets. We utilize the VoxCeleb [Nagrani *et al.*, 2017] dataset, which consists of around $20K$ talking-head videos, to train our HiFiHead network. The cropped videos are then resized to 512×512. A total of $17,927$ training videos and 491 testing videos are obtained.

Figure 5: Visual comparison of cross-identity motion transfer. Our HiFiHead is able to generate photo-realistic face details, as well as better identity and more accurate eye movements, compared with other SOTA methods.

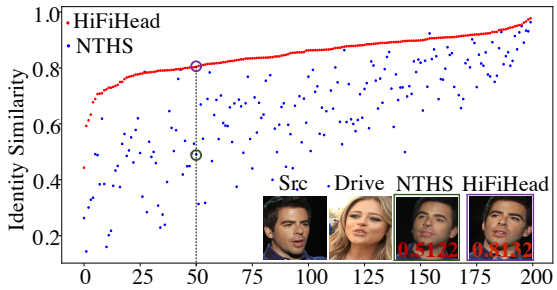


Figure 6: ID similarity distributions of NTHS and our HifiHead. Samples are sorted by ID similarity of HiFiHead. Same column index means the same source/driving pair.

Implementation Details. We re-train the D3DFR [Deng *et al.*, 2019b] to predict 3DMM and gaze coefficients. The StyleGAN2 is pretrained on FFHQ [Karras *et al.*, 2019] dataset at resolution 512×512. We clarify that we use the pretrained weights of StyleGAN for faster convergence when training HifiHead on VoxCeleb. However, similar result can still be achieved just with more iterations if trained from scratch. The spatial feature encoder contains 7 down-sample convolutional layers. The learning rate is set to 0.001 for all trainable parameters. The batch size is set to 32. It takes around 2 days to train HifiHead with 8 Tesla V100 GPUs.

Evaluation Metrics. PSNR and LPIPS [Zhang *et al.*, 2018] are adopted to evaluate the reconstruction error. The motion transfer accuracy is measured by Exp, Angle and Gaze, which calculate the average Euclidean distances of expression, pose and gaze coefficients between the generated and target images. Identity similarity (ID) is measured by the cosine distance in deep feature space. For fair comparison, another popular face recognition model ArcFace [Deng *et al.*, 2019a] is used to extract identity features. FID [Heusel *et al.*, 2017] is also reported to measure visual quality.

4.2 Talking-head Motion Transfer

We compare our model with three state-of-the-art methods, including FOMM [Siarohin *et al.*, 2019a], one-shot neural talking-head synthesis (NTHS) [Wang *et al.*, 2021a] and PIRenderer [Ren *et al.*, 2021]. The official released models of FOMM and PIRenderer are adopted in the experiments. NTHS is re-implemented following the implementation details provided by the paper.

Same-identity Reconstruction. We first compare image synthesis results where the source and driving images are of the same person. The quantitative evaluation is shown in Table 1. It can be seen that our HifiHead obtains comparable PSNR to other competing methods, but achieves better results on LPIPS and FID, which are better measures than PSNR for the face image perceptual quality. HifiHead also obtains the lowest Exp, Angle and Gaze scores, showing that our results have more accurate head motions and eye movements. In addition, HifiHead also preserves better identity. Figure 4 shows the qualitative comparisons. Our method can reproduce the driving images more faithfully.

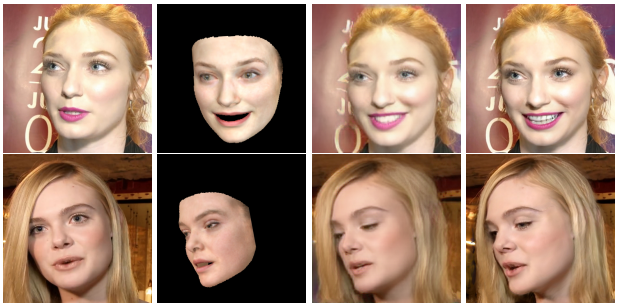
Cross-identity Motion Transfer. Cross-identity motion transfer has a wider range of practical applications than same-identity reconstruction. Besides Voxceleb, we randomly sample $1K$ images from the FFHQ [Karras *et al.*, 2019] and CelebaHQ [Karras *et al.*, 2017] dataset as the source images to compare the generalization capability with other methods. The quantitative results are presented in Table 2. Our method achieves superior performance on all metrics for all datasets. One possible explanation is that the competing methods [Siarohin *et al.*, 2019a; Wang *et al.*, 2021a] describe motions with person-specific sparse keypoints. The accuracy is reduced when driven by cross-identity images. In comparison, our generated results are conditioned on explicit well-disentangled 3DMM coefficients and rendered 3D images, which is less sensitive to driving subjects. Furthermore, our method manages to hallucinate realistic face details thanks to the carefully designed generative block and fine-tuning strategies. The qualitative comparisons are shown in Figure 5.

Method	VoxCeleb					CelebaHQ					FFHQ				
	Exp↓	Angle↓	Gaze↓	ID↑	FID↓	Exp↓	Angle↓	Gaze↓	ID↑	FID↓	Exp↓	Angle↓	Gaze↓	ID↑	FID↓
FOMM	6.92	0.0744	0.0981	0.5443	76.42	6.72	0.0743	0.1001	0.5351	83.73	7.23	0.0810	0.0964	0.5078	107.00
NTHS	7.36	0.0873	0.1436	0.5737	70.40	7.42	0.0878	0.1322	0.5912	74.88	8.01	0.0963	0.1322	0.5425	110.54
PIRenderer	6.56	0.0793	0.1131	0.5046	67.51	6.78	0.0828	0.1122	0.5553	78.78	7.46	0.0900	0.1155	0.5019	106.06
Ours	5.08	0.0699	0.0786	0.7946	56.70	5.51	0.0691	0.0845	0.7859	47.87	6.05	0.0764	0.0834	0.7578	69.64

Table 2: Quantitative comparison on cross-identity motion transfer task.

Method	Exp↓	Angle↓	ID↑	FID↓
PIRenderer	5.42	0.0651	0.7272	56.53
Ours	**4.28**	**0.0473**	**0.8645**	**43.25**

Table 3: Comparison on intuitive editing with 3D control.



Source Target 3D PIRender Ours

Figure 7: Intuitive editing with 3D control.

Src Drive D-NeRF Ours

Figure 8: Visual comparison with Dynamic-NeRF.

w/o flow w/o $\mathcal{L}_{ID}$ Ours

Figure 9: Ablation study on flow field and identity loss.

To further demonstrate the ability of HifiHead in preserving the identity information, we randomly sample 200 source and driving pairs. Based on these pairs, the corresponding identity similarity distribution is visualized in Figure 6, in which HifiHead significantly outperforms SOTA NTHS with 98% samples achieving higher identity similarity.

4.3 Intuitive Editing with 3D Control

In this experiment, we intuitively modify the 3DMM representations to generate images with different motions. Although many methods have been proposed for face editing, few of them are capable of intuitively changing face expressions and poses. We compare to the state-of-the-art method PIRenderer, which also achieves editing via a 3DMM.

We randomly select 100 source images and 10 target expressions and poses, which totally results in $1K$ editing images. We do not edit eye movements as PIRenderer does not support such editing. Table 3 shows that our method achieves better results. The qualitative comparisons are presented in Figure 7. From the first row, we can see that our HifiHead can generate realistic facial details, such as hair and eyelashes. In the second row, PIRenderer can not retain the identity information from the source image. In comparison, the identity is well-maintained in our generated results.

4.4 Comparison with Dynamic NeRF

NeRF-based methods require video sequence as training data for each *specific* person (*e.g.*, Dynamic NeRF [Pumarola *et al.*, 2021] uses **5,000** images), while HifiHead drives talking heads using only *one* source image and can be applied to *any* person. Beside, Dynamic NeRF can not model eye blinks and eye movements, as claimed in their paper. Figure 8 illustrates the comparison with the case reported in Dynamic NeRF.

4.5 Ablation Studies

To evaluate the effectiveness of our proposed HifiHead, we conduct ablation study on two variants of our method. Variant **A** (w/o flow) represents removing the flow field prediction branch. The source features F_{src} are directly concatenated to the features in the GAN block. Variant **B** (w/o $\mathcal{L}_{ID}$) denotes removing identity loss during training. Figure 9 presents the visual comparison between the variants and our full model. Without explicit dense flow prediction, model **A** can not generate accurate motions. On the other hand, Model **B** can not preserve the identity effectively. In comparison, our full model can generate much better results.

5 Conclusion

In this paper, we present a high fidelity neural head synthesis framework, termed HifiHead, to produce photo-realistic results with high quality appearance of the source image and accurate target motion of the driving image. Extensive experiments show that our method can not only achieve better same-identity reconstruction, but also generalize well to cross-identity motion transfer, significantly outperforming state-of-the-art competitors. In addition, benefiting from semantically meaningful 3DMM parameters, our proposed HifiHead allows intuitive control and editing for users.

References

[Chu *et al.*, 2020] Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Sscgan: Facial attribute editing via style skip connections. In *ECCV*, 2020.

[Deng *et al.*, 2019a] Jiankang Deng, J. Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019.

[Deng *et al.*, 2019b] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019.

[Doukas *et al.*, 2021] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *ICCV*, 2021.

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

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[Huang *et al.*, 2020] Y. Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *CVPR*, 2020.

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

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2017.

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

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

[Karras *et al.*, 2021] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

[Nagrani *et al.*, 2017] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[Pumarola *et al.*, 2021] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *CVPR*, 2021.

[Ren *et al.*, 2021] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, 2021.

[Siarohin *et al.*, 2019a] Aliaksandr Siarohin, Stéphane Lathuilière, S. Tulyakov, Elisa Ricci, and N. Sebe. First order motion model for image animation. In *NeurIPS*, 2019.

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

[Siarohin *et al.*, 2021] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.

[Tewari *et al.*, 2020a] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *TOG*, 2020.

[Tewari *et al.*, 2020b] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, 2020.

[Wang *et al.*, 2021a] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *CVPR*, 2021.

[Wang *et al.*, 2021b] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *IJCAI*, 2021.

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

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

[Yang *et al.*, 2021] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021.

[Yao *et al.*, 2021] Guangming Yao, Yi Yuan, Tianjia Shao, Shuang Li, Shanqi Liu, Yong Liu, Mengmeng Wang, and Kun Zhou. One-shot face reenactment using appearance adaptive normalization. In *AAAI*, 2021.

[Zakharov *et al.*, 2019] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019.

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

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.