Head3D: Complete 3D Head Generation via Tri-plane Feature Distillation

YUHAO CHENG, YICHAO YAN, and WENHAN ZHU, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
YE PAN, John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, China
BOWEN PAN, Alibaba Group, Hangzhou, China
XIAOKANG YANG, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China

Head generation with diverse identities is an important task in computer vision and computer graphics, widely used in multimedia applications. However, current full-head generation methods require a large number of three-dimensional (3D) scans or multi-view images to train the model, resulting in expensive data acquisition costs. To address this issue, we propose Head3D, a method to generate full 3D heads with limited multi-view images. Specifically, our approach first extracts facial priors represented by tri-planes learned in EG3D, a 3D-aware generative model, and then proposes feature distillation to deliver the 3D frontal faces within complete heads without compromising head integrity. To mitigate the domain gap between the face and head models, we present a dual-discriminator to guide the frontal and back head generation. Our model achieves cost-efficient and diverse complete head generation with photo-realistic renderings and high-quality geometry representations. Extensive experiments demonstrate the effectiveness of our proposed Head3D, both qualitatively and quantitatively.

CCS Concepts: • **Computing methodologies → Computer vision;**

Additional Key Words and Phrases: Head generation, neural radiance field, adversarial generative network, limited data

ACM Reference Format:
Yuhao Cheng, Yichao Yan, Wenhan Zhu, Ye Pan, Bowen Pan, and Xiaokang Yang. 2024. Head3D: Complete 3D Head Generation via Tri-plane Feature Distillation. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 6, Article 176 (March 2024), 20 pages. https://doi.org/10.1145/3635717

This article was supported by the National Natural Science Foundation of China (grant nos. 62201342, 62101325, and U19B2035), the Shanghai Municipal Science and Technology Major Project (grant no. 2021SHZDZX0102), and the CCF-Alibaba Innovative Research Fund For Young Scholars.
Authors' addresses: Y. Cheng, Y. Yan (Corresponding author), W. Zhu, and X. Yang, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, Shanghai, China; e-mails: {chengyuhao, yanyichao, zhuwenhan823, xkyang}@sjtu.edu.cn; Y. Pan, John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, Shanghai, China; e-mail: whitneypanye@sjtu.edu.cn; B. Pan, Alibaba Group, 969 West Wen Yi Road, Yu Hang District, Hangzhou, Zhejiang, China; e-mail: bowen.pbw@alibaba-inc.com.

ACM 1551-6857/2024/03-ART176
https://doi.org/10.1145/3635717

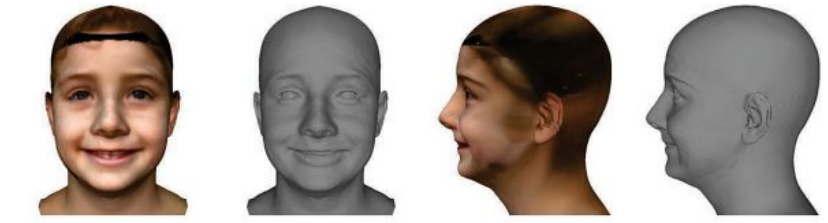
1 INTRODUCTION

Generating high-fidelity three-dimensional (3D) heads poses a significant challenge in the domains of computer vision and graphics, with a broad range of applications, including 3D games and movies. However, existing approaches [8, 23, 26, 35, 36, 41, 46, 60, 66, 68, 73, 75] primarily concentrate on frontal faces, lacking the capacity of rendering the side or back views. Thus, their applications are significantly limited. In this work, we aim to address the issue and generate complete 3D heads with photo-realistic rendering capabilities.

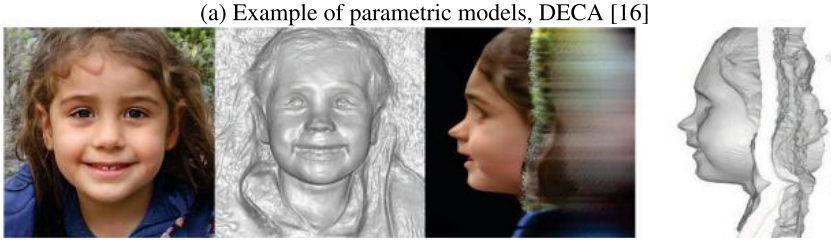
Current 3D head generation methods fall into two categories: non-parametric head models and parametric head models. Non-parametric methods [6, 22, 47, 57, 67, 71] predict 3D heads from single-view or multi-view images, usually considered as "reconstruction" methods. These methods face challenges in reproducing heads that do not exist, limiting the variety of generated results. Moreover, limited-view reconstruction results in insufficient details due to the absence of visible perspectives. Parametric models [4, 5, 7, 13, 19, 20, 34, 43–45, 56, 61, 69] (Figure 1(a)) utilize decoupled parameters to represent heads, which rely on a vast amount of expensive 3D scans and find it hard to express intricate texture and geometry. Learning 3D head generation solely from images can be a more cost-effective approach to addressing this challenging task, and it has the potential to generate richer identities and higher-fidelity outcomes. Recently, 3D-aware GANs [8, 9, 23, 41] have been learned from easily accessible in-the-wild images to generate 3D frontal faces with photo-realistic rendering and high-fidelity shapes (Figure 1(b)). These methods theoretically can also be used in generating heads. However, accurate camera poses are crucial for 3D consistency in these methods [8], and estimating them without landmarks on the back is challenging. Additionally, image collection in the back views is more difficult. Recently, several powerful works [1, 62] have achieved the generation of images with broader perspectives and even high-quality complete head generation. These endeavors have involved the collection of a substantial volume of back-view images and employed pose estimation models to predict camera poses for achieving such outcomes. However, a significant disparity exists between the actual facial pose and the predictions made by pose estimation models, which constrains the generation results in terms of 3D consistency. Hence, our objective is to address the aforementioned challenges and devise an approach capable of generating 3D-consistent complete heads solely by training on limited images with accurate poses.

Motivated by the high-fidelity 3D face generator [8], a question arises: can we use it as prior knowledge to generate full heads? We answer this question with *YES*, but two challenges must be addressed. First, *how to extract the 3D priors of heads?* The face prior is represented implicitly, making it difficult to integrate the face topology directly with the head in a re-topological manner akin to computer graphics. A straightforward idea is to directly fine-tune the 3D-aware generation model on full head data. However, fine-tuning the pre-trained model with limited data, e.g., several thousand images, often leads to mode collapse or over-fitting [17], resulting in limited face diversity and low quality. Second, *how to bridge the domain gap between the frontal faces and hair?* Obviously, the frontal face and back of the heads with head share two related but different distributions, respectively. Moreover, obtaining back-view images with accurate view direction is challenging, resulting in highly imbalanced quantities in these two distributions. This poses an extra requirement for the discriminator in the 3D-aware **Generative Adversarial Network (GAN)** model to not only distinguish real/fake samples but also frontal/back views.

In this work, we propose Head3D for the diverse full head generation that builds on a current **state-of-the-art (SOTA)** 3D face generative model, i.e., EG3D [8]. Our goal is to transfer the facial prior generated by EG3D onto full heads with limited 3D data. To extract the 3D prior, we conduct a systematic analysis of the tri-plane representation and observe a decoupling between the geometric and identity information in this representation. Based on this observation, we propose



(a) Example of parametric models, DECA [16]



(b) Example of 3D-aware GAN trained by 2D data, EG3D [8]

(c) Visualization of Our Head3D

Fig. 1. Illustration of existing head/face generation methods and our presented Head3D. All results are rendered into RGB images and geometric form in both frontal and side views.

a tri-plane feature distillation framework that aims to preserve the identity information while completing the head geometry. To address the second problem, we design dual-discriminators for frontal faces and the back of heads, which not only frees the discriminator from distinguishing frontal/back images but also inherits the strong capability of discriminators from EG3D.

To evaluate the effectiveness of our Head3D, we conduct comparative analyses with the original EG3D [8] and other recent baselines. Our experiments demonstrate that the proposed Head3D model can produce superior results compared with previous approaches despite being trained with only a small amount of multi-view images and prior knowledge. Examples of our model are shown in Figure 1(c). Additionally, we perform ablation studies to validate the effectiveness of each component in Head3D.

The main contributions are summarized as follows: **(a)** We propose the novel Head3D method for generating full heads with rich identity information, photo-realistic renderings, and detailed shapes with limited 3D data. **(b)** We investigate the effectiveness of tri-plane features and propose tri-plane feature distillation to enable the transfer of identity information onto the head templates. **(c)** We propose a dual-discriminator approach to address the distribution gap and quantity imbalance between front-view and back-view images, thereby ensuring the quality of generated heads.

2 RELATED WORK

2.1 Head Generation Model

As previously discussed in the introduction, head generation can be categorized into two types. "Reconstruction" methods [6, 22, 47, 57] primarily learn the correspondence between 3D data and

images to establish prior knowledge. When presented with new images, these methods optimize the difference between the reconstructed results and the images for reconstruction. However, these methods are limited to the reconstruction of heads present in the given images and are incapable of generating novel heads. Explicit parametric 3D morphable models [3–5, 7, 13, 19, 34, 43–46, 56] represent identities, textures and expressions by low-dimensional PCA parameters, which are learned from 3D scans with different expressions and identities. Similarly, implicit 3D parametric models [20, 27, 69] typically employ an auto-decoder to learn the decoupled parameters from 3D scans. However, these methods are trained with a large number of expensive 3D scans and find it hard to express detailed texture and geometry. The recent Rodin model [59] employs a diffusion model to learn head generation trained by images. However, this approach requires a large dataset consisting of 100,000 3D models for rendering multi-view images, and each identity is reconstructed with a tri-plane alone for training the diffusion model. The employed dataset and training procedure are particularly expensive. Similarly, Panohead [1] collects a large amount of data for complete head training. However, the inaccurately estimated camera pose limits the 3D consistency of the model. In contrast to these expensive methods, our goal is to generate diverse novel and high-fidelity heads in a cost-effective manner, utilizing only implicit face priors and a limited amount of multi-view images.

2.2 NeRF-Based GAN Model

A **neural radiance field (NeRF)** [39] represents 3D models by implicit networks whose outputs are density and color under the input of position and view direction, optimized by view reconstruction between volume rendering results and ground truths. Recent methods integrating NeRFs and GANs [21] aim at learning 3D-aware generators from a set of unconstrained images [9, 18, 38, 40, 49, 53, 54]. GRAF [49] verifies that NeRF-based GANs can generate 3D-consistent model with high-fidelity rendering results. Pi-GAN [9] introduces SIREN [52] and a growing strategy for higher-quality image synthesis results. To improve rendering efficiency, several works volumetrically render a low-resolution feature, then up-sample them for high-resolution view synthesis under different 3D consistency constraints [8, 23, 41, 66, 68, 73, 75]. In particular, EG3D [8] provides a hybrid 3D representation method, which first generates tri-plane features, then sampled features are decoded and rendered for image synthesis. Pose-related discrimination is utilized for 3D consistency generation. Surrounding head images are a prerequisite for these methods to achieve complete head generation. However, it is difficult to predict the view directions of in-the-wild back head images, and pose-accurate multi-view images are expensive. Our Head3D extends EG3D to achieve complete head generation with limited multi-view head images by facial priors transfer via tri-plane feature distillation.

2.3 Knowledge Distillation of GANs

The target of **knowledge distillation (KD)** [25] is to transfer dark knowledge, for example, logits and features, from teacher networks to student networks, which is originally used in classification [12, 42, 48, 55, 65, 74]. KD is also employed for GAN-based model compression [11, 28, 33, 37, 58, 64]. In GAN compression [33], the student network learns every intermediate feature and final output from the teacher network. Besides conditional GANs, several works focus on the study of unconditional GANs. The content-aware GAN [37] adopts multi-level distillation and content-aware distillation, then fine-tuned by adversarial loss. StyleKD [64] proposes that the mapping network plays an important role in generation. In addition, a novel initialization strategy and a latent direction-based distillation loss are presented for semantic consistency between the teacher and student model. Our work is built upon a StyleGAN-like 3D-aware generative model EG3D [8]. Through experiments, we observe the disentanglement of tri-planes, which

are critical in semantic representations in EG3D. Accordingly, a tri-plane feature distillation procedure is proposed to transfer facial priors in complete head generation.

3 METHODS

With a small number of multi-view head images, our goal is to achieve complete head generation with a pre-trained face generator. We first review an efficient and effective tri-plane-based face generator EG3D (Section 3.1). Second, aiming to extract the semantic information of the face as prior knowledge, we explore how tri-planes represent the faces (Section 3.2). Third, we describe how to apply the prior knowledge to generate full heads (Section 3.3). Finally, the training process is introduced in detail (Section 3.4).

3.1 Revisit Tri-planes for 3D Generation

EG3D is a tri-plane-based generative model. Similar to StyleGAN [30, 31], mapping networks M process the input latent code z and camera parameter p to style code w:

$$w = M(z, p). \tag{1}$$

Then, feature maps F are extracted via a CNN-based generator C, yielding three planes that are rearranged orthogonally to form a tri-plane structure:

$$F_{xy}, F_{xz}, F_{yz} = C(w), \tag{2}$$

where F_{xy}, F_{xz}, and F_{yz} are three planes in the front, side, and top views, respectively. The tri-plane features contain semantic information about faces, determining their identities.

Next, features of locations are sampled from tri-planes and are fed into the decoder to output densities σ and colors c. Then, volume rendering is performed to obtain moderate-resolution images I_{raw}.

$$I_{raw} = \int_0^\infty p(t)\mathbf{c}(\boldsymbol{r}(t), \boldsymbol{d})dt, \tag{3}$$

where $p(t) = \exp(-\int_0^t \sigma(\boldsymbol{r}(s))ds) \cdot \sigma(\boldsymbol{r}(t))$, $\boldsymbol{r}(t)$ represents camera ray, and t is the distance from the camera.

Finally, a super-resolution module $S(\cdot)$ is performed to up-sample the I_{raw} to results I of high resolution:

$$I = S(I_{raw}). \tag{4}$$

3.2 Disentangled Representation of Tri-planes

Insufficient multi-view head images pose a challenge for generating diverse complete heads. However, leveraging the existing EG3D model, which can generate high-quality 3D frontal faces, offers a potential solution. Similar to StyleGAN, the semantic results in EG3D are determined by the style codes w, which generate tri-plane features as the only output of the CNN generator C. As a result, all of the semantic information is contained within tri-planes. To enable knowledge transfer, we conduct an in-depth analysis of tri-planes to explore how prior knowledge is represented.

To investigate the role of each plane in the original EG3D, we exchange each plane (F_{xy}, F_{xz}, and F_{yz}) between two different samples, and then generate faces from the newly integrated tri-planes for visual analysis. The results are shown in Figure 2. We observe that the identities are exchanged along with F_{xy} changed, while the identities remain the same after exchanging F_{yz} and F_{xz}. Additionally, we conduct numerical experiments on 100 different identities. We exchange F_{xy}, F_{xz}, and F_{yz} for each identified pair, respectively, and render the tri-planes to images from the frontal view. The identity consistency between the exchanged tri-plane and the original identities is then measured by Arcface [14], and the results are calculated by averaging all the sample pairs.

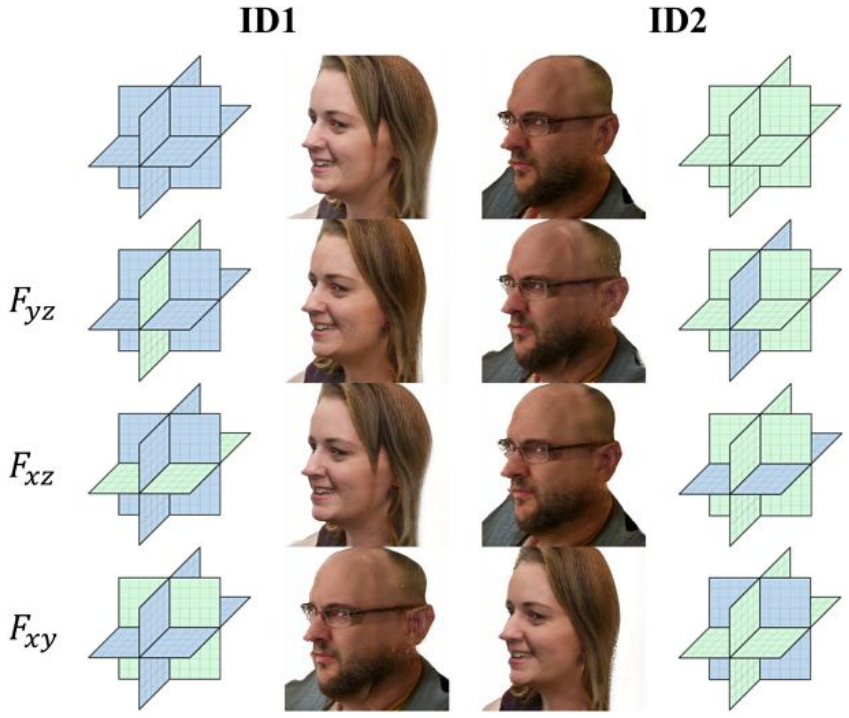


Fig. 2. Visualization analysis of exchanging each plane of the tri-plane. ID 1 and ID 2 are generated from the different latent codes, represented in blue and green, respectively.

Table 1. Quantitative Results with Tri-plane Exchanged

F_{xy}	F_{xz}	F_{yz}	ID 1	ID 2
ID 2	ID 1	ID 1	0.535	0.968
ID 1	ID 2	ID 1	0.977	0.534
ID 1	ID 1	ID 2	0.988	0.532
ID 1	ID 1	ID 1	1.000	0.531
ID 2	ID 2	ID 2	0.531	1.000

ID 1 and ID 2 denote two different identities generated from different z by EG3D. We report the average results between every pair between 100 identities.

Table 1 indicates that the identities are preserved when F_{yz} and F_{xz} changed, whereas they exchange when F_{xy} are exchanged. Based on these observations, we preliminarily assume that F_{xy} plays a crucial role in determining the identity information.

We performed three in-depth experiments to further validate our assumption. First, as an illustration, we conducted an experiment involving the exchange of xy-planes associated with the eye position. Results shown in Figure 3 illustrate 3D-consistent alteration in the appearance of the eyes while preserving consistency in other facial regions. This experiment provides empirical evidence supporting the notion that the xy-plane can effectively encapsulate and represent facial semantic information in an independent manner. Second, we explored the role of channels in the xy-plane as shown in Figure 4. We quantitatively measured the distribution of different channels in two images and then arranged these channels in order of dissimilarity. As shown in Figure 4, by swapping the halves with the most significant differences, we observed that the top 16 channels provided

Fig. 3. Exchanging partial xy-planes enables 3D-consistent local face editing.

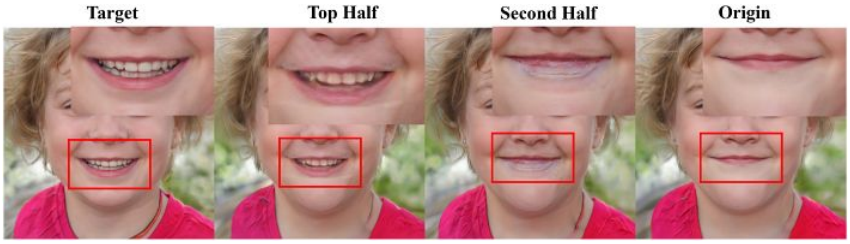


Fig. 4. The results of exchanging different channels. Each individual channel seems to govern nuanced aspects of the same semantic information.

more detailed information, whereas the latter still contributed some limited details (teeth texture). This suggests that identity information cannot be fully preserved using only a limited set of channels, emphasizing the importance of complete xy-plane representing face priors. Finally, to verify the impact of F_{xz} and F_{yz}, we fine-tune a head generator from EG3D and perform experiments in which F_{xy} is exchanged between the face generator and head generator. The instanced results, as shown in Figure 5, indicate that exchanging F_{xy} leads to identity transfer while not influencing the geometry whether it be heads or faces. Our findings support the notion that F_{xy} primarily controls the identity information, while the other two planes, F_{yz} and F_{xz}, mainly represent the geometric shape of the head. These results prove our proposed decoupling between identity and geometric information in the tri-plane representation.

This observation is consistent with our expectation, as the xy-plane is aligned with the training images to better capture the facial features, whereas the other two planes are orthogonal to the frontal view, which mainly represents the depth and geometric information. The experimental results and analysis shed some light on the effect of tri-plane features in EG3D and offer a starting point for considering how to employ this prior to complete head generation.

3.3 Tri-plane Distillation for 3D Head Generation

One potential approach to generating complete heads with the frontal face priors is to directly fine-tune the face generator with a few multi-view data. However, training with a small number of data can cause mode collapse or over-fitting, resulting in limited diversity and low quality. Inspired by computer graphic methods in which the face is extracted from a photo and then registered with

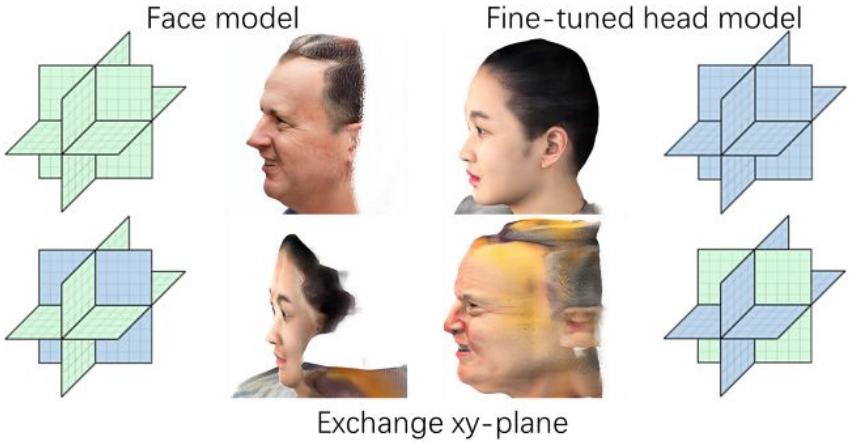


Fig. 5. Visualization analysis of exchanging xy-plane between face generator and head generator. Planes in green and blue represent them from the face model and head model, respectively.

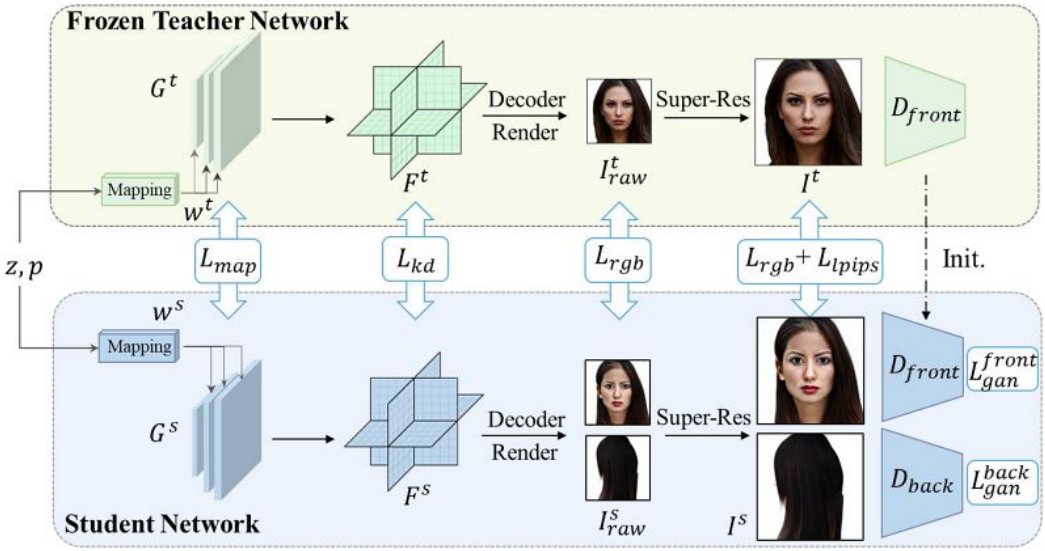


Fig. 6. Overview of our proposed Head3D. First, the frozen pre-trained EG3D serves as the teacher network G^t, whereas the fine-tuned head generator is the student network G^s. Then, tri-plane feature distillation and multi-level loss functions are employed for photo-realistic and diverse full-head generators. Note that, the parameter p for rendering and discriminator are not displayed in this framework.

the head template model, we can also apply the tri-plane-based face priors to an implicit head template. Additionally, knowledge distillation [25] is a general method to transfer dark knowledge between models, allowing for the delivery of facial priors. Therefore, we propose a tri-plane feature distillation method, as illustrated in Figure 6. First, we fine-tune a full head generator from the pre-trained EG3D using a small amount of multi-view head images. Then, served as prior knowledge, F_{xy} are transferred from the face generator to the head network via knowledge distillation to ensure consistency in identity. Benefiting from the powerful presentation capability of tri-planes, we are able to integrate the head template and face priors, enabling diverse full head generation.

In order to achieve photo-realistic rendering results, GAN loss is necessary for training the tri-plane distillation framework [64]. The importance of camera poses in learning correct 3D priors in the discriminator has been highlighted in EG3D [8]. For human heads, camera poses can be categorized into front and back perspectives to guide the generation of the face and back, respectively. A large number of single-view face datasets are available to provide front images, whereas the number of back images is limited to half of the small multi-view dataset. As a result, there is an imbalance in the quantity and a distribution gap in perception between frontal and back head images. It is hard for a single discriminator to simultaneously guide fine-grained face generation and guarantee full head geometry in two different domains. To tackle this problem, we propose a dual-discriminator to ensure the generation quality and maintain the head completed. In our method, two discriminators guide the generation of front and back images. Moreover, two discriminators are alternated during training to mitigate the effects of imbalanced data.

3.4 Model Training

Face Prior Transfer. As depicted in Figure 6, we present Head3D, a tri-plane distillation approach for generating diverse heads. First, a head generator G_s is fine-tuned from a pre-trained face generator G_t with scarce multi-view head images. Then, to ensure identity consistency, the transformation of F_{xy} is calculated by L2-norm, represented as

$$L_{kd} = ||F^t_{xy} - F^s_{xy}||_2, \tag{5}$$

where F^t_{xy} and F^s_{xy} are xy-plane features generated by head generator G_t and face generator G_s, respectively.

Dealing with Distribution Gap. To realize high-fidelity front and back head generation, we employ two discriminators, D_{front} and D_{back}, for training that share the same network structure as the discriminator in EG3D. The purpose of D_{front} is to ensure the quality of faces, and it is initialized by the pre-trained face EG3D. D_{back} is used to ensure the integrity of the head, which is initialized by the fine-tuned head generator. Fed together with camera parameters p, both discriminators support the synthesis results I to be 3D consistent and in a similar distribution with ground truth images I_{gt}. The GAN loss function can be represented as

$$\begin{aligned} L_{gan} = \mathbb{E}[f(D(p, \mathbf{I})) + f\left(-D\left(p, \mathbf{I}_{gt}\right)\right)] \\ + \gamma ||\nabla D\left(p, \mathbf{I}_{gt}\right)||^2, \end{aligned} \tag{6}$$

where $f(x) = -\log(1 + \exp(-x))$ and γ is a hyper-parameter in R1 regularization. Dual-discriminators are trained separately according to the view from the face or back of heads, whose loss functions are represented as L^{front}_{gan} and L^{back}_{gan}, respectively. In addition, the regularization coefficient γ of the dual-discriminator is different due to the imbalanced quantity of the dataset.

Detailed Texture Learning. StyleKD [64] proposes that the mapping network determines the semantic information of generators. Therefore, besides tri-plane feature distillation, it is also necessary to ensure that the output w of the mapping network is consistent. Following StyleKD [64], a mapping loss is utilized:

$$L_{map} = ||W^t - W^s||_1. \tag{7}$$

Moreover, in order to learn more detailed faces, RGB loss and LPIPS loss [72] are applied only to frontal neural renderings I_{raw} and super-resolution results I. Referring to the quality of the generated results of original EG3D declines in the side view, this loss function is applied only to

the rendering results within a certain range.

$$L_{rgb} = \mathbb{I}(|\Delta p| \leq \tau)[||I^t - I^s||_1 + ||I^t_{raw} - I^s_{raw}||_1], \tag{8}$$

$$L_{lpips} = \mathbb{I}(|\Delta p| \leq \tau)||F(I^t) - F(I^s)||_1, \tag{9}$$

where $F(\cdot)$ is a well-trained frozen VGG [51] to extract multi-scale semantic information from images. Δp is the horizontal offset angle from center and τ is a threshold. These two loss functions work in image space and perceptual space, respectively, ensuring detailed consistency between the teacher and student.

Finally, the final loss function is a weighted sum with the above loss functions:

$$\begin{aligned} L = &\lambda_{gan_{front}} L^{front}_{gan} + \lambda_{kd} L_{kd} + \lambda_{rgb} L_{rgb} \\ &+ \lambda_{lpips} L_{lpips} + \lambda_{map} L_{map} + \lambda_{gan_{back}} L^{back}_{gan}, \end{aligned} \tag{10}$$

where λ_* denotes the weights of each loss functions.

4 EXPERIMENTS

4.1 Experiments Setup

Dataset. We train our proposed Head3D using two datasets. One is FFHQ [30], a large public single-view real-world face dataset, for face priors learning. The other is H3DH, our proposed multi-view human head dataset. H3DH contains multi-view images of 50 identities, who are gender- and age-balanced with a wide variety of hairstyles. Following EG3D, we employ the same off-the-shelf pose estimators [15] to extract approximating camera extrinsics of FFHQ. In terms of H3DH, we set the same camera intrinsic as that of FFHQ and obtain images with the resolution of 512^2 from surrounding perspectives under natural light.

Implementation Details. We fine-tune the head generator in the procedure from StyleGAN-ADA [29] via the H3DH dataset with batch size 16. The optimizer is Adam [32] with the same learning rate as original EG3D [8] of 0.0025 for G and 0.002 for D. D is trained with R1 regularization with $\gamma = 20$. In the knowledge distillation phase, the learning rate is converted to 0.001 for G and 0.0005 for D, with $\gamma = 1$ for D_{front} and $\gamma = 20$ for D_{back}. All λ_* are set to 1.0 except λ_{kd}, which is set to 0.5 and $\lambda_{gan_{back}}$, which is set to 10.0 to balance loss functions in the front view and back view. Threshold τ is set to $\pi/3$. The resolutions of neural rendering I_{raw} and final generated images I are 128^2 and 512^2, respectively. Note that, owing to H3DH rendered without background, we regard fine-tuned EG3D which generates faces in the white background as the teacher network instead of the original EG3D, which is fine-tuned with images in FFHQ whose background is removed by BiseNet [70]. Training a model costs about 4 hours on two NVIDIA A100 GPUs.

4.2 Comparisons

Comparisons with baselines. As few works achieve full head generation trained by images, we compare our work with some designed baselines. (1) Directly fine-tune EG3D with FFHQ and H3DH, named *DFT*. (2) Train model with FFHQ and H3DH from a fine-tuned head generator, named *FH*. (3) Original pre-trained EG3D, to examine the quality of faces. The results are shown both qualitatively and quantitatively, where we use **Fréchet Inception Distance (FID)** [24] and **Kernel Inception Distance (KID)** [2] for evaluation. Note that we do not implement training from scratch on the H3DH due to the fact that training with only 50 individuals results in insufficient diversity in the identity information.

Figure 7 provides example results synthesized by our proposed method from various viewpoints and identities, demonstrating the generation of complete head geometry and high-quality renderings. In particular, our method is able to generate accessory items such as hats and glasses,

Fig. 7. Example results of our proposed Head3D, synthesized from three different views.

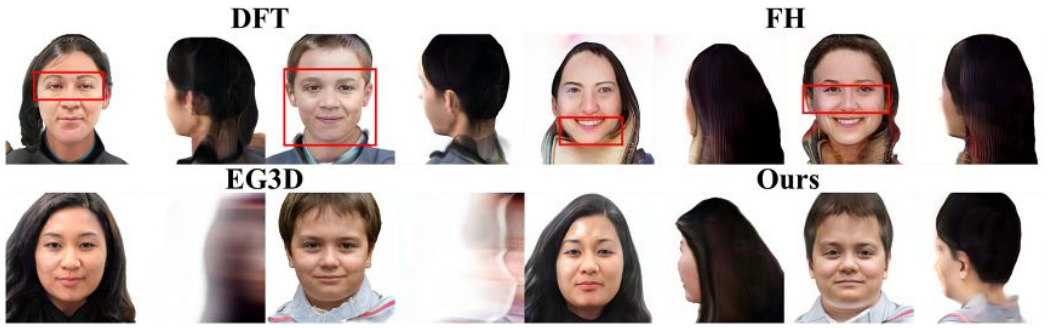


Fig. 8. Qualitative comparisons between Head3D and baselines, which exhibit two perspectives of two identities. The red rectangles highlight artifacts showing on the faces. Please zoom-in.

Table 2. Quantitative Comparisons Between Our Head3D and Baselines in FID and KID×100, Obtained in the Resolution of 512^2

Type	Baselines	FID ↓	KID ↓
Face	EG3D (Fine-tuned)	**6.82**	**0.36**
Head	*DFT*	30.46	1.50
	FH	46.09	2.11
	Ours	11.34	0.61

which are not present in the H3DH dataset. The visualized comparisons are shown in Figure 8, in which images in the same column are synthesized from the same latent code z. Although *DFT* and *FH* achieve full head generation, output consistency with origin EG3D is broken. Moreover, the model is trained by naïvely combined FFHQ and H3DH, resulting in low-quality results and model collapse. Similar observations can also be found from Table 2 that both FID and KID become much worse in *DFT* and *FH*. Original EG3D performs the best numerically, while it cannot generate full heads. Our work achieves identity representation consistent with the original EG3D through knowledge distillation and can fully represent human heads with good 3D consistency. Although our method is numerically inferior to the original EG3D, considering information loss in knowledge distillation and no back images corresponding to FFHQ faces, the decline in generating quality is acceptable. Specifically, the generated heads maintain high-fidelity rendering results

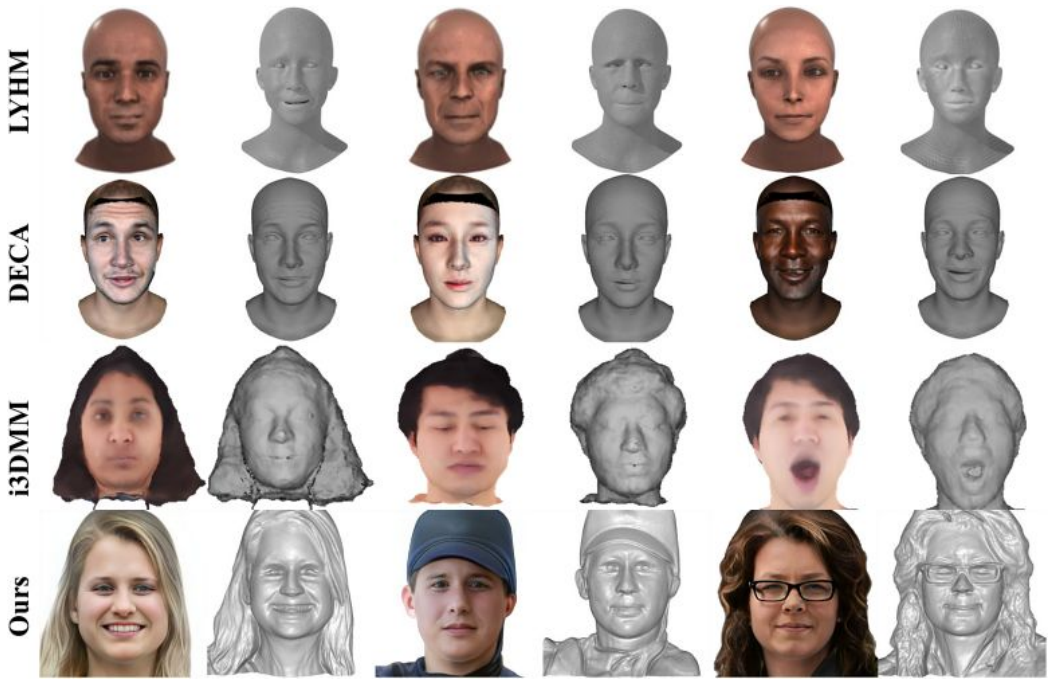


Fig. 9. Qualitative comparisons among our Head3D and current generative head models. Results of LYHM [45, 46], DECA [16], and i3DMM [69] are obtained via their papers and released codes.

as EG3D qualitatively. In summary, our H3DH achieves high-quality complete head generation in both visual results and quantitative evaluation.

Comparisons with current methods. Qualitative comparisons are conducted among our Head3D and current generative head models, LYHM [45, 46], DECA [16], and i3DMM [69], in terms of rendering results and geometry shapes, as depicted in Figure 9. Notably, these models are trained on the dataset consisting of a large number of 3D scans (several hundred or thousands), whereas our Head3D model only uses a multi-view image dataset containing 50 individuals. However, the results reveal that our model produces superior rendering quality and more detailed geometry than these methods. Additionally, explicit head models [16, 45, 46] are not capable of representing hair, whereas the implicit model i3DMM [69] can only represent hair in low quality. In contrast, our results show the ability to generate photo-realistic and diverse hair, including various accessories such as hats and glasses. Therefore, this comparison demonstrates that our algorithm can achieve high-quality and highly detailed complete head generation with significantly less data.

Comparisons with PanoHead. PanoHead [1] is a powerful and effective approach that introduces novel components such as the Tri-Grid and Self-Adaptive Camera Alignment, which exhibit remarkable capability in generating 3D full heads. It should be noted that the training data containing a large number of in-house images on back views greatly contributes to the performance. However, this study aims to investigate the potential of achieving complete head generation using a limited dataset. In this section, we train our proposed Head3D model along with PanoHead using the H3DH and FFHQ datasets, allowing for fair comparisons with limited data.

The comparative analysis presented in Figure 10 demonstrates the performance of both models qualitatively. It is evident that both models successfully accomplish the task of generating complete heads. However, with limited training data, Panohead encounters artifacts, such as striping, on

Fig. 10. Qualitative comparisons between our Head3D and recent SOTA PanoHead [1].

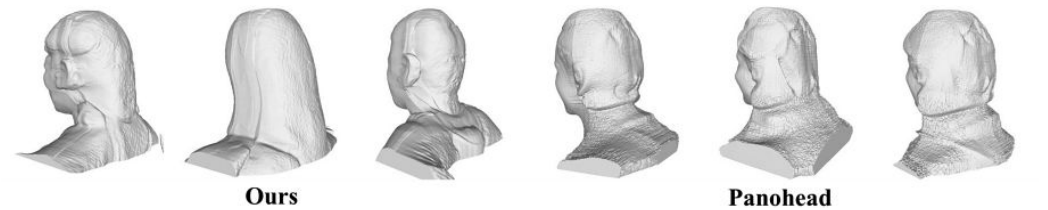


Fig. 11. Geometric comparisons between our Head3D and recent SOTA PanoHead [1].

Table 3. Quantitative Comparisons Between Our Head3D and Recent SOTA Method PanoHead

Methods	Views	FID ↓	KID ↓	Training Time (h)	3D Consistency PSNR
PanoHead [1]	front	37.85	2.46	≥48	23.61(*22.75)
	back	85.32	9.21		
Ours	front	11.34	0.61	≈4	24.91
	back	74.49	5.26		

The two models are both trained by our limited dataset. *means results from the released Panohead pre-trained model.

the facial region in frontal views. Figure 11 shows that our Head3D model is also able to generate complete heads in geometry. In all, Head3D exhibits better face quality due to its utilization of pre-trained EG3D to incorporate face priors.

We also conduct quantitative comparisons of model generation quality, training time, and 3D consistency, as shown in Table 3. Specifically, for generation quality, following EG3D [8], we generate 50,000 images for |yaw|<=90° and |yaw|>=90°, respectively, and calculate FID and KID scores in conjunction with corresponding training dataset samples. The results indicate that our Head3D outperforms Panohead [1] in terms of the quality of frontal- and back-view image generation. This may be attributed to the substantial disparity in the number of frontal-view images compared with back-view ones and the limitation of a single discriminator that restricts further enhancement in frontal facial quality. For the back-view images, it is plausible that the utilization of a single discriminator also led to some quality degradation, possibly aggravated by the occurrence of rare back-view samples. Additionally, Head3D showcases enhanced training efficiency owing to the integration of face priors.

To evaluate the 3D consistency, we adopt the PSNR metric proposed in GRAM-HD [63]. We randomly generate 100 images from various surrounding viewpoints for 50 individuals, with

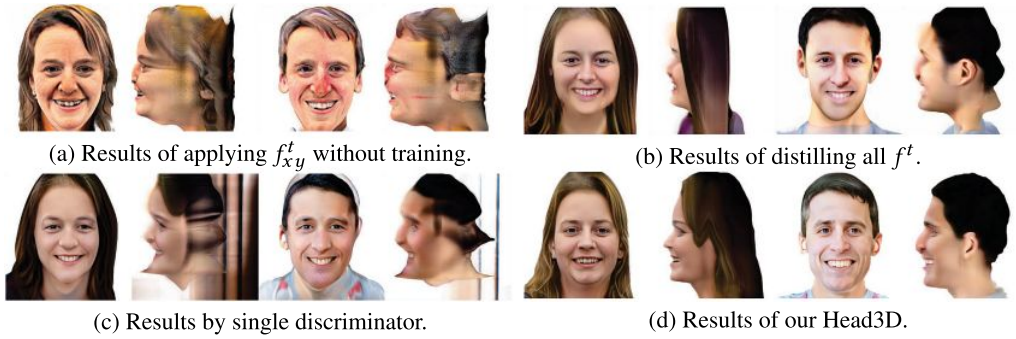


Fig. 12. Illustration of ablation studies on tri-plane feature distillation and dual-discriminator.

50 images used to train TensoRF [10]. We then render 50 images from distinct views of the remaining 50 images and calculate their PSNR values to demonstrate 3D consistency. Benefitting from the robust 3D consistency offered by TensoRF, this metric can effectively assess the 3D consistency of the complete human head generation model. Results are shown in Table 3, in which * means original Panohead [1], which is trained on a significantly larger dataset. It is important to highlight that, considering the presence of backgrounds in the original Panohead, we ensure fairness in the evaluation by cropping the individuals and employing a uniform white background as seen in Head3D. The findings are interesting. First, they indicate that Head3D exhibits superior 3D consistency compared with the limited dataset-trained Panohead. Second, the differences observed between the original Panohead and our Panohead trained on a limited dataset, as well as Head3D, can be attributed to several factors: (1) The tri-grid representation, as opposed to the triplane, may mitigate mirroring-face artifacts to some extent. However, different planes may affect results from various viewpoints, thus limiting 3D consistency: **Good 3D consistency necessitates uniform representation from any viewpoint.** (2) Our H3DH dataset is rendered from 3D models with highly accurate pose information, whereas the original Panohead employs less precise pose estimation models and relies on a self-adaptive camera alignment method, leading to a reduction in 3D consistency: **Accurate pose information indeed contributes to improved 3D consistency**. Consequently, Head3D achieves high-quality 3D-consistent head generation even when working with limited amounts of data.

4.3 Ablation Study

Effectiveness of Tri-plane Feature Distillation. We investigate the effectiveness of our proposed tri-plane feature distillation method. We considered two settings: (a) directly applying f_{xy}^t on the head generator without training and (b) distilling the whole tri-plane features. Results are shown in Figures 12(a) and 12(b), respectively. It can be concluded in the first setting that, although the identity information is preserved, the generated heads suffer from severe distortions and defective head shapes. In the second setting, distilling the whole tri-plane leads to incomplete head shapes. In contrast, our proposed method, as shown in Figure 3(d), maintains identity consistency and achieves complete head generation. These results demonstrate the effectiveness of our tri-plane feature distillation approach.

Effectiveness of Dual-Discriminator. We also conduct experiments to verify the effectiveness of our proposed dual-discriminator. As depicted in Figure 12(c), the model trained with a single discriminator proposed in the original EG3D for head generation produces competitive results for face synthesis with Head3D. However, it fails to generate the back of the head, resulting in

Table 4. Quantitative Results of Ablation Study on Different Loss Functions, Evaluated by FID, KID×100, and ID Consistency (ID)

	FID ↓	KID ↓	ID ↑
W/O L_{kd}	44.76	3.07	0.10
W/O $L_{gan_{front}}$	29.84	2.47	0.45
W/O L_{rgb}&L_{lpips}	19.50	1.29	0.45
Ours	**11.34**	**0.61**	**0.65**

Fig. 13. The ablation studies on the effectiveness of each loss function. All images in a column are generated from the same latent code. Note that the ground-truth results are generated by the pre-trained face generator without background.

a stitching of two faces. In summary, the comparison verifies the effectiveness of the proposed dual-discriminator.

Effectiveness of Loss Functions. We conducted an ablation study to verify the effectiveness of each loss. We remove each loss function individually and maintain other settings with no change. The results are evaluated both qualitatively and quantitatively. The quantitative results—including FID, KID, and **Identity Distance (ID)** to the origin EG3D, calculated by Arcface [14] among 10,000 individuals—are presented in Table 4. The results show that removing any loss functions results in a significant increase in FID and KID scores. Moreover, the lowest ID score is obtained when knowledge distillation is removed, which highlights the crucial role of tri-plane distillation in transferring identity information.

Figure 13 presents example results of different settings to further evaluate the effectiveness of each loss function. Although most of the textures are preserved without $L_{gan_{front}}$, the photo-realism is lost, which indicates the importance of adversarial training for maintaining image

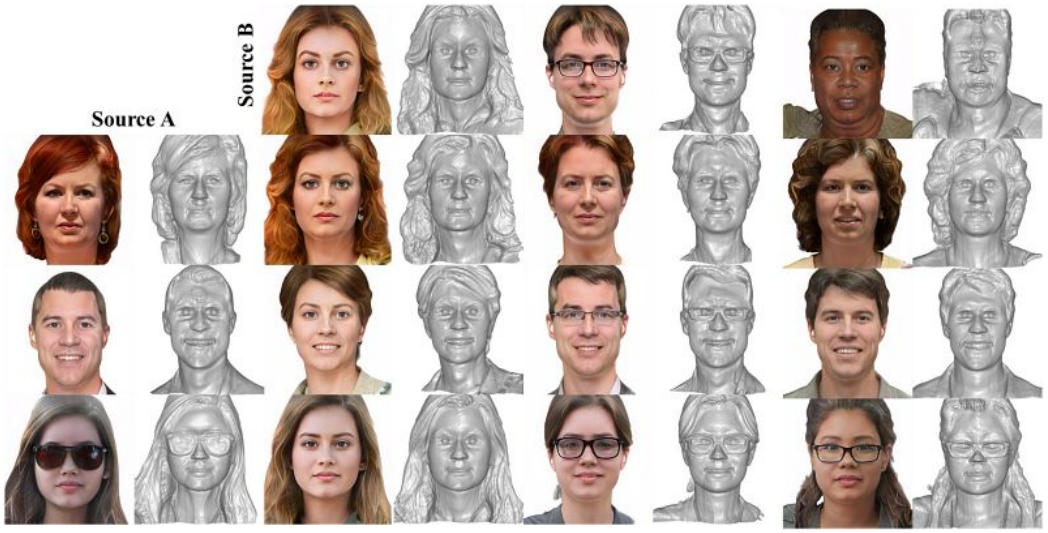


Fig. 14. Interpolation results of our proposed Head3D model. The results are generated with the average of latent codes from source A and source B.

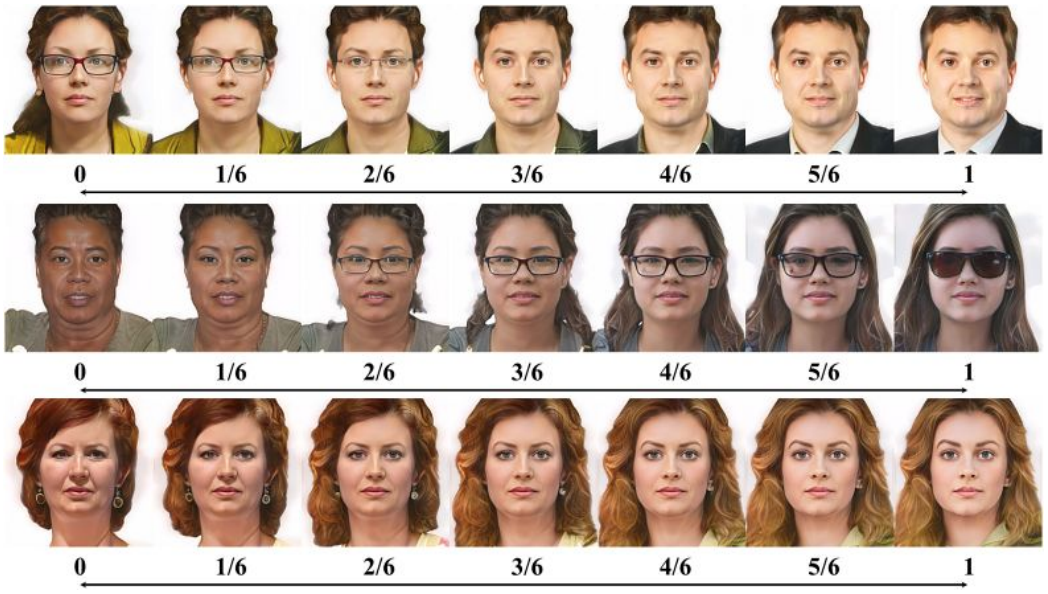


Fig. 15. Smoothly interpolating results of our proposed Head3D model.

quality. While lacking L_{kd} causes failure in identity transmission, L_{rgb} and L_{lpips} are also crucial in preserving detailed texture and ensuring texture consistency. Overall, the combination of these loss functions achieves the best results both quantitatively and qualitatively.

Analysis of Linearity in Latent Space. Analogous to StyleGAN-based generators [30, 31], style codes w can be linearly interpolated to achieve image manipulations. Figure 14 demonstrates that our model maintains linear separability in the latent space after knowledge distillation. Figure 15 demonstrates smoothly interpolating results, illustrating the linearity of the latent space.

Fig. 16. Latent space editing methods are applicable to our Head3D.

Additionally, maintaining the same latent space with the original EG3D is beneficial in facilitating seamless integration with other EG3D-based image-editing methods, such as changing expressions and shapes. Figure 16 shows the head editing results using InterFaceGAN [50] trained on EG3D.

5 CONCLUSION

This article presents Head3D, a method for generating complete heads trained with limited data. We first revisit the EG3D framework and emphasize the importance of tri-plane as a semantic information carrier. Through experiments, we demonstrate that tri-plane decoupling is achievable, with identity information controlled by the xy-plane. We then propose a tri-plane feature distillation approach and a dual-discriminator method for training head generators. Extensive experiments confirm the effectiveness of our proposed method. We hope that our work will inspire further research in generating diverse and high-quality 3D complete heads from limited and uncorrelated images.

REFERENCES

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. 2023. PanoHead: Geometry-aware 3D full-head synthesis in 360deg. In *CVPR*. 20950–20959.
[2] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401* (2018).
[3] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH* 187–194.
[4] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 2017. 3D face morphable models "in-the-wild". In *CVPR*. 48–57.
[5] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. 2016. A 3D morphable model learnt from 10,000 faces. In *CVPR*. 5543–5552.
[6] Egor Burkov, Ruslan Rakhimov, Aleksandr Safin, Evgeny Burnaev, and Victor Lempitsky. 2022. Multi-NeuS: 3D head portraits from single image with neural implicit functions. *arXiv preprint arXiv:2209.04436* (2022).
[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE TVCG* (2013), 413–425.
[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*. 16123–16133.
[9] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *CVPR*. 5799–5809.
[10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *ECCV*. Springer, 333–350.
[11] Hanting Chen, Yunhe Wang, Han Shu, Changyuan Wen, Chunjing Xu, Boxin Shi, Chao Xu, and Chang Xu. 2020. Distilling portable generative adversarial networks for image translation. In *AAAI*.
[12] Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *ICCV*. 4794–4802.
[13] Hang Dai, Nick Pears, William A. P. Smith, and Christian Duncan. 2017. A 3D morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3085–3093.
[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*. 4690–4699.

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*. 0–0.
[16] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *TOG* (2021), 1–13.
[17] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *TOG* (2022), 1–13.
[18] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. 2020. Local deep implicit functions for 3D shape. In *CVPR*. 4857–4866.
[19] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. 2018. Morphable face models—an open framework. In *FG*. 75–82.
[20] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2022. Learning neural parametric head models. *arXiv preprint arXiv:2212.02761* (2022).
[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* (2020), 139–144.
[22] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *CVPR*. 18653–18664.
[23] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *ICLR*.
[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *NeurIPS* (2017).
[25] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR* abs/1503.02531 (2015).
[26] Trang-Thi Ho, John Jethro Virtusio, Yung-Yao Chen, Chih-Ming Hsu, and Kai-Lung Hua. 2020. Sketch-guided deep portrait generation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 3 (2020), 1–18.
[27] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A real-time NeRF-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20374–20384.
[28] Liang Hou, Zehuan Yuan, Lei Huang, Huawei Shen, Xueqi Cheng, and Changhu Wang. [n. d.]. Slimmable generative adversarial networks. In *AAAI*.
[29] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *NeurIPS* (2020), 12104–12114.
[30] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.
[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *CVPR*.
[32] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*, Yoshua Bengio and Yann LeCun (Eds.).
[33] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. 2020. GAN compression: Efficient architectures for interactive conditional GANs. In *CVPR*. 5284–5294.
[34] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM TOG* (2017).
[35] Yidong Li, Wenhua Liu, Yi Jin, and Yuanzhouhan Cao. 2021. SPGAN: Face forgery using spoofing generative adversarial networks. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 1s (2021), 1–20.
[36] Shiguang Liu and Huixin Wang. 2023. Talking face generation via facial anatomy. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3 (2023), 1–19.
[37] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Federico Perazzi, and Sun-Yuan Kung. 2021. Content-aware GAN compression. In *CVPR*. 12156–12166.
[38] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. 2021. GNeRF: GAN-based neural radiance field without posed camera. In *ICCV*. 6351–6361.
[39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* (2021), 99–106.
[40] Michael Niemeyer and Andreas Geiger. 2021. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*. 11453–11464.
[41] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-resolution 3D-consistent image and geometry generation. In *CVPR*. 13493–13503.
[42] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *CVPR*. 3967–3976.
[43] Ankur Patel and William A. P. Smith. 2009. 3D morphable face models revisited. In *CVPR*. 1327–1334.

[44] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *AVSS*. 296–301.
[45] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William A. P. Smith, Baris Gecer, and Stefanos Zafeiriou. 2020. Towards a complete 3D morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 4142–4160.
[46] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William A. P. Smith, and Stefanos Zafeiriou. 2019. Combining 3D morphable models: A large scale face-and-head model. In *CVPR*. 10934–10943.
[47] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3D-Net: Few-shot high-fidelity 3D head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5620–5629.
[48] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).
[49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative radiance fields for 3D-aware image synthesis. *NeurIPS* (2020), 20154–20166.
[50] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2004–2018.
[51] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.
[52] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *NeurIPS* (2020), 7462–7473.
[53] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. 2022. EpiGRAF: Rethinking training of 3D GANs. *CoRR* abs/2206.10535 (2022).
[54] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. 2022. FENeRF: Face editing in neural radiance fields. In *CVPR*. 7672–7682.
[55] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *ICLR*.
[56] Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *CVPR*. 1126–1135.
[57] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022. MoRF: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.
[58] Haotao Wang, Shupeng Gui, Haichuan Yang, Ji Liu, and Zhangyang Wang. 2020. GAN slimming: All-in-one GAN compression by a unified optimization framework. In *ECCV*. 54–73.
[59] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. 2022. Rodin: A generative model for sculpting 3D digital avatars using diffusion. *arXiv preprint arXiv:2212.06135* (2022).
[60] Xueping Wang, Yunhong Wang, and Weixin Li. 2019. U-Net conditional GANs for photo-realistic and identity-preserving facial expression synthesis. *ACM Transactions on Multimedia Computing, Communications and Applications* 15, 3s (2019), 1–23.
[61] Sijing Wu, Yichao Yan, Yunhao Li, Yuhao Cheng, Wenhan Zhu, Ke Gao, Xiaobo Li, and Guangtao Zhai. 2023. GANHead: Towards generative animatable neural head avatars. In *CVPR*. 437–447.
[62] Yiqian Wu, Jing Zhang, Hongbo Fu, and Xiaogang Jin. 2023. LPFF: A portrait dataset for face generators across large poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20327–20337.
[63] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. 2023. GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds. In *ICCV*. 2195–2205.
[64] Guodong Xu, Yuenan Hou, Ziwei Liu, and Chen Change Loy. 2022. Mind the gap in distilling StyleGANs. *arXiv preprint arXiv:2208.08840* (2022).
[65] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. 2020. Knowledge distillation meets self-supervision. In *ECCV*.
[66] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2022. 3D-aware image synthesis via learning structural and textural representations. In *CVPR*. 18409–18418.
[67] Han Xue, Jun Ling, Anni Tang, Li Song, Rong Xie, and Wenjun Zhang. 2023. High-fidelity face reenactment via identity-matched correspondence learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3 (2023), 1–23.
[68] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. 2022. GIRAFFE HD: A high-resolution 3D-aware generative model. In *CVPR*. 18419–18428.
[69] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3DMM: Deep implicit 3d morphable model of human heads. In *CVPR*. 12803–12813.

[70] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*. 325–341.
[71] Hang Yu, Chilam Cheang, Yanwei Fu, and Xiangyang Xue. 2023. Multi-view shape generation for a 3D human-like body. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 1 (2023), 1–22.
[72] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.
[73] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. 2022. Multi-view consistent generative adversarial networks for 3D-aware image synthesis. In *CVPR*. IEEE, 18429–18438.
[74] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. 2020. Prime-aware adaptive distillation. In *ECCV*. 658–674.
[75] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-aware generator of GANs based on conditionally-independent pixel synthesis. *CoRR* abs/2110.09788 (2021).

Received 13 July 2023; revised 15 October 2023; accepted 21 November 2023