PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360°

Sizhe An[1,2] Hongyi Xu[1] Yichun Shi[1] Guoxian Song[1] Umit Y. Ogras[2] Linjie Luo[1]
[1]ByteDance Inc. [2]University of Wisconsin-Madison

Figure 1. Our PanoHead enables 360° view-consistent photo-realistic full-head image synthesis with high-fidelity geometry, enabling authentic 3D portraits creation from a single-view image.

Abstract

Synthesis and reconstruction of 3D human head has gained increasing interests in computer vision and computer graphics recently. Existing state-of-the-art 3D generative adversarial networks (GANs) for 3D human head synthesis are either limited to near-frontal views or hard to preserve 3D consistency in large view angles. We propose PanoHead, the first 3D-aware generative model that enables high-quality view-consistent image synthesis of full heads in 360° with diverse appearance and detailed geometry using only in-the-wild unstructured images for training. At its core, we lift up the representation power of recent 3D GANs and bridge the data alignment gap when training from in-the-wild images with widely distributed views. Specifically, we propose a novel two-stage self-adaptive image alignment for robust 3D GAN training. We further introduce a tri-grid neural volume representation that effectively addresses front-face and back-head feature entanglement rooted in the widely-adopted tri-plane formulation. Our method instills prior knowledge of 2D image segmentation in adversarial learning of 3D neural scene structures, enabling compositable head synthesis in diverse backgrounds. Benefiting from these designs, our method significantly outperforms previous 3D GANs, generating high-quality 3D heads with accurate geometry and diverse appearances, even with long wavy and afro hairstyles, renderable from arbitrary poses. Furthermore, we show that our system can reconstruct full 3D heads from single input images for personalized realistic 3D avatars.

1. Introduction

Photo-realistic portrait image synthesis has been a continuous focus in computer vision and graphics, with a wide range of downstream applications in digital avatars, telepresence, immersive gaming, and many others. Recent advances in Generative Adversarial Networks (GANs) [12] has demonstrated strikingly high image synthesis quality, indistinguishable from real photographs [19, 21, 22]. However, contemporary generative approaches operate on 2D convolutional networks without modeling the underlying 3D scenes. Therefore 3D consistency cannot be strictly enforced when

Project page: https://sizhean.github.io/panohead

synthesizing head images under various poses.

To generate 3D heads with diverse shapes and appearances, traditional approaches require a parametric textured mesh model [2, 25] learned from large 3D scan collections. However, the rendered images lack fine details and have limited perceptual quality and expressiveness. With the advent of differentiable rendering and neural implicit representation [28, 47], conditional generative models have been developed to generate more realistic 3D-aware face images [17, 44, 45, 53]. However, those approaches typically require multi-view image or 3D scan supervision, which are hard to acquire and have limited appearance distribution as those are usually captured in controlled environments.

3D-aware generative models have recently seen rapid progress, fueled by the integration of implicit neural representation in 3D scene modeling and Generative Adversarial Networks (GANs) for image synthesis [5, 6, 29, 31, 37, 40, 48]. Among them, the seminal 3D GAN, EG3D [5], demonstrates striking quality in view-consistent image synthesis, trained only from in-the-wild single-view image collections. However, these 3D GAN approaches are still limited to synthesis in near-frontal views.

In this paper, we propose *PanoHead*, a novel 3D-aware GAN for high-quality full 3D head synthesis in 360° trained from only in-the-wild unstructured images. Our model can synthesize *consistent 3D heads viewable from all angles*, which is desirable by many immersive interaction scenarios such as digital avatars and telepresence. To the best of our knowledge, our method is *the first* 3D GAN approach to achieve full 3D head synthesis in 360°.

Extending 3D GAN frameworks such as EG3D [5] to full 3D head synthesis poses several significant technical challenges: Firstly, many 3D GANs [5, 31] cannot separate foreground and background, inducing 2.5D head geometry. The background, formulated typically as a wall structure, is entangled with the generated head in 3D and therefore prohibits rendering from large poses. We introduce a *foreground-aware tri-discriminator* that jointly learns the decomposition of the foreground head in 3D space by distilling the prior knowledge in 2D image segmentation.

Secondly, while being compact and efficient, current hybrid 3D scene representations, like tri-plane [5], introduce strong projection ambiguity for 360° camera poses, resulting in 'mirrored face' on the back head. To address the issue, we present a novel 3D *tri-grid volume representation* that disentangles the frontal features with the back head while maintaining the efficiency of tri-plane representations.

Lastly, obtaining well-estimated camera extrinsics of in-the-wild back head images for 3D GANs training is extremely difficult. Moreover, an image alignment gap exists between these and frontal images with detectable facial landmarks. The alignment gap causes a noisy appearance and unappealing head geometry. Thus, we propose a novel *two-stage alignment scheme* that robustly aligns images from any view consistently. This step decreases the learning difficulty of 3D GANs significantly. In particular, we propose a camera self-adaptation module that dynamically adjusts the positions of rendering cameras to accommodate the alignment drifts in the back head images.

Our framework substantially enhances the 3D GANs' capabilities to adapt to in-the-wild full head images from arbitrary views, as shown in Figure 1. The resulting 3D GAN not only generates high-fidelity 360° RGB images and geometry, but also achieves better quantitative metrics than state-of-the-art methods. With our model, we showcase compelling 3D full head reconstruction from a single monocular-view image, enabling easily accessible 3D portrait creation.

In summary, our main contributions are as follows:

- The first 3D GAN framework that enables view-consistent and high-fidelity full-head image synthesis with detailed geometry, renderable in 360°. We demonstrate our approach in high-quality monocular 3D head reconstruction from in-the-wild images.
- A novel tri-grid formulation that balances efficiency and expressiveness in representing 3D 360° head scenes.
- A foreground-aware tri-discriminator that disentangles 3D foreground head modeling from 2D background synthesis.
- A novel two-stage image alignment scheme that adaptively accommodates imperfect camera poses and misaligned image cropping, enabling training of 3D GANs from in-the-wild images with wide camera pose distribution.

2. Related Work

3D Head Representation and Rendering. To represent 3D heads with diverse shapes and appearances, a line of work has targeted parametric textured mesh representation, such as 3D Morphable Model (3DMM) [2–4, 33] for faces and FLAME head model [25], learned from 3D scans. However, these parametric representations do not model photo-realistic appearance and geometry beyond the front face or skull. The neural implicit functions [47] have recently emerged as powerful continuous and differential representations of 3D scenes. Among them, Neural Radiance Field (NeRF) [1, 28] has been widely adopted in digital head modeling [10, 15, 17, 32, 34, 43] due to its superiority in modeling complex scene details and synthesizing multiview images with inherited 3D consistency. In contrast to optimizing a person-specific neural radiance field from multiview images or temporal videos, our approach builds a generative NeRF from unstructured 2D monocular images. Recently implicit-explicit hybrid 3D representation has been explored for better efficiency [5, 9, 27]. Among them, the tri-plane formulation proposed in EG3D [5] demonstrates a highly

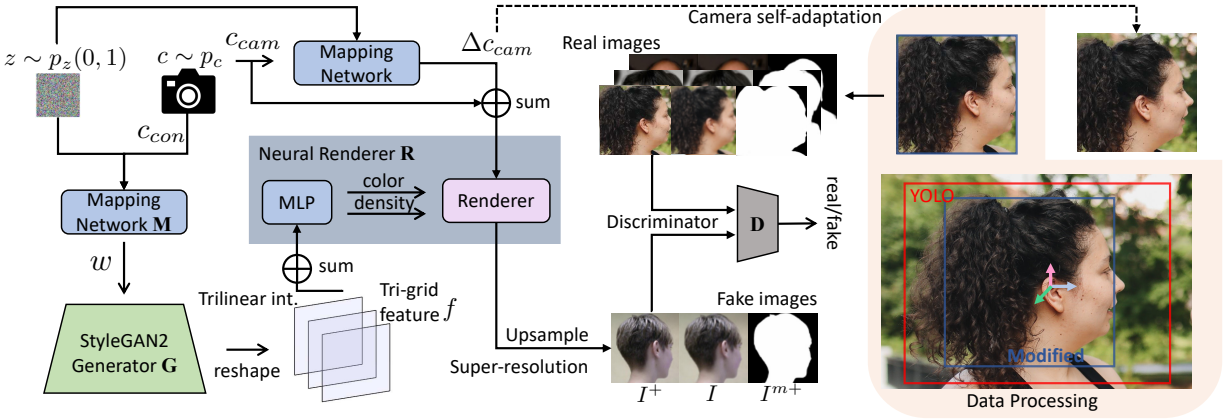


Figure 2. Our framework consists of three main components: a foreground-aware generator **G**, discriminator **D**, and a neural renderer **R**. A mapping network first maps latent code z and conditioned camera pose c_{con} into the intermediate latent code w. The generator **G** then takes w to obtain the 3D tri-grid representation features f. With f and rendering camera pose c_{cam}, the neural renderer **R** synthesizes super-resolved image I^+, bilinear-upsampled image I, and super-resolved mask I^{m+}. Finally, the foreground-aware tri-discriminator **D** critiques (I^+, I, I^{m+}) along with real images. The data processing pipeline is shown in the right side. The real images are cropped with modified YOLO bounding boxes yet they often differ at scale and location due to lacking accurate facial landmarks. With the camera self-adaptation scheme, the rendering camera pose c_{cam} is able to correct itself to generate images with consistent scale and location.

efficient 3D scene representation with high-quality view-consistent image synthesis. The tri-plane representation can scale efficiently with resolution, enabling greater detail for equal capacity. Our tri-grid representation transforms the tri-plane representation into a more expressive space for better feature embedding in unconditional 3D head synthesis.

Single- or Few-view Supervised 3D GANs. Given the impressive progress of GANs on 2D image generation [12, 19, 21, 22], many studies have attempted to extend them to 3D-aware generation. These GANs aim to learn a generalizable 3D representation from 2D image a collections. For face synthesis, Szabo *et al*. [42] first proposed using vertex position maps as the 3D representation to generate textured mesh outputs. Shi *et al*. [39] proposed a self-supervised framework to convert 2D StyleGANs [21] into 3D generative models, although its generalizability is bounded by its base 2D StyleGAN. GRAF [37] and pi-GAN [6] are the first to integrate NeRF into 3D GANs. However, their performance is limited by the intense computation cost of forwarding and backwarding a complete NeRF. Many recent studies [5, 8, 11, 13, 29–31, 38, 40, 48, 49] have attempted to improve the efficiency and quality of such NeRF-based GANs. Specifically, EG3D [5], which we build our work upon, introduces tri-plane representation that can leverage a 2D GAN backbone for generating efficient 3D representation and is shown outperforming other 3D representations [38]. Parallel to these works, another thread of studies [30, 41, 46, 50] have been working on controllable 3D GANs that can manipulate the generated 3D faces or bodies.

3. Methodology

3.1. PanoHead Overview

To synthesize realistic and view-consistent full head images, we build PanoHead upon a state-of-the-art 3D-aware GAN, *i.e*. EG3D [5], due to its efficiency and synthesis quality. Specifically, EG3D leverages StyleGAN2 [22] backbone to output a tri-plane representation that represents a 3D scene with three 2D feature planes. Given a desired camera pose c_{cam}, the tri-plane is decoded with a MLP network and volume rendered into a feature image, followed by a super-resolution module to synthesize a higher resolution RGB image I^+. Both the low and high resolution images are then jointly optimized by a dual discriminator **D**.

In spite of EG3D's success in generating frontal faces, we found it to be a much more challenging task to adapt to 360° in-the-wild full head images for the following reasons: 1) foreground-background entanglement prohibit large pose rendering, 2) strong inductive bias from tri-plane representation causes mirroring face artifacts on the back head, and 3) noisy camera labels and inconsistent cropping of back head images. To address these problems, we introduce a background generator and a tri-discriminator for decoupling foreground and background (Section 3.2), an efficient yet more expressive tri-grid representation while still being compatible with StyleGAN backbone (Section 3.3), and a two-stage image alignment scheme with an self-adaptation module that dynamically adjusts rendering cameras during training (Section 3.4). The overall pipeline for our model is illustrated in Figure 2.

3.2. Foreground-Aware Tri-Discrimination

A typical challenge of state-of-the-art 3D-aware GANs, like EG3D [5], is the entangled foreground with the background of synthesized images. Regardless of the highly detailed geometry reconstruction, directly training the 3D GAN from in-the-wild RGB image collections, such as FFHQ [21], results in a 2.5D face, as illustrated in Figure 3 (a). Augmenting with image supervisions from the side and back of the head helps build up the full-head geometry with reasonable back head shapes. However, it does not solve the problem because the tri-plane representation itself is not designed to represent separated foreground and background.

To disentangle the foreground from the background, we first introduce an additional StyleGAN2 network [22] to generate 2D backgrounds at the same resolution of raw feature image I^r. During volume rendering, the foreground mask I^m can be obtained by:

$$I^r(r) = \int_0^\infty w(t) f(r(t)) dt, \;\; I^m(r) = \int_0^\infty w(t) dt, \quad (1)$$

$$w(t) = exp\big(-\int_0^t \sigma(r(s)) ds\big) \sigma(r(t)), \quad (2)$$

where $r(t)$ represents a ray emitted from the rendering camera center. The foreground mask is then used to compose a new low-resolution image I^{gen}:

$$I^{gen} = (1 - I^m) I^{bg} + I^r, \quad (3)$$

which is fed into the super-resolution module. Note that the computation cost of background generator is insignificant since its output has a much lower resolution than the tri-plane generator and super-resolution module.

Simply adding a background generator does not fully decouple it from the foreground since the generator tends to synthesize foreground content in the background. Thus, we propose a novel foreground-aware tri-discriminator to supervise the rendered foreground mask along with the RGB images. Specifically, the input of the tri-discriminator has 7 channels, composed with a bilinearly-upsampled RGB image I , a super-resolved RGB image I^+ and single-channel upsampled foreground mask I^{m+}. The additional mask channel allows the 2D segmentation prior knowledge to be back-propagated into the density distribution of the neural radiance field. Our approach reduces the learning difficulty in shaping the 3D full head geometry from unstructured 2D images, enabling authentic geometry ((Figure 3 (b))) and appearance synthesis of a full head composable with various backgrounds (Figure 3 (c)). We note that in constrast from ENARF-GAN [30] that employs a single discriminator for RGB images composed of synthesized foreground and background images using a dual-generated mask, our tri-discriminator better ensures view-consistent high-resolution outputs.

Figure 3. Geometry and RGB images from dual-discrimination (a) and foreground-aware tri-discrimination (b, c). EG3D (a) fails to decouple the background. PanoHead's tri-discrimination offers both background-free geometry (b) and background-switchable full head image synthesis (c).

3.3. Feature Disentanglement in Tri-Grid

The tri-plane representation, proposed in EG3D [5], offers an efficient representation for 3D generation. The neural radiance density and appearance of a volume point are obtained by projecting its 3D coordinate over three axis-aligned orthogonal planes and decoding the sum of three bilinearly interpolated features with a tiny MLP. However, when synthesizing a full head in 360°, we observe tri-plane is limited in expressiveness and suffers from mirroring-face artifacts. The problem is even pronounced when the camera distribution of the training images is unbalanced. The root cause

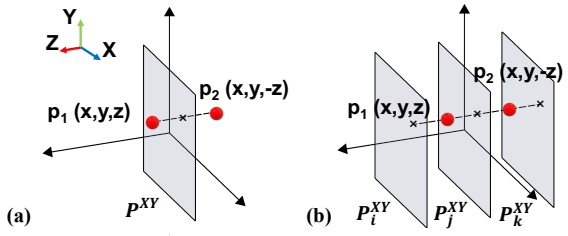


Figure 4. Comparison between tri-plane (a) and tri-grid (b) architecture in Z axis. With tri-plane, two different points' projections share the feature from the plane P^{XY}, which introduces representation ambiguity. With tri-grid, the features for the above two points are trilinearly interpolated from two different planes, thus generating distinct features. Please refer to supplement for implementation and visualization details.

Figure 5. Images synthesis with tri-plane and tri-grid ($D = 3$). Due to the projection ambiguity, tri-plane representation (a) can generate good-quality front face image yet with a 'mirrored face' on back head, while our tri-grid representation synthesizes high-quality back head appearance and geometry (b).

Figure 6. Image synthesized without (a) and with the camera self-adaptation scheme(b). Without it, the model generates misaligned back head images, leading to a defective dent in back head.

is the inductive bias originating from tri-plane projection, where one point on a 2D plane has to represent features of different 3D points. For example, a point on the front face and a point on the back hair will be projected to the same point on the XY plane P^{XY} (orthogonal to Z axis), as illustrated in Figure 4 (a). Although the other two planes should theoretically provide complementary information to alleviate this projection ambiguity, we found it not the case when there is less visual supervision from the back or when the structure of the back head is challenging to learn. The tri-planes are prone to borrow features from the front face to synthesize the back head, referred to as mirroring-face artifacts here (Figure 5(a)).

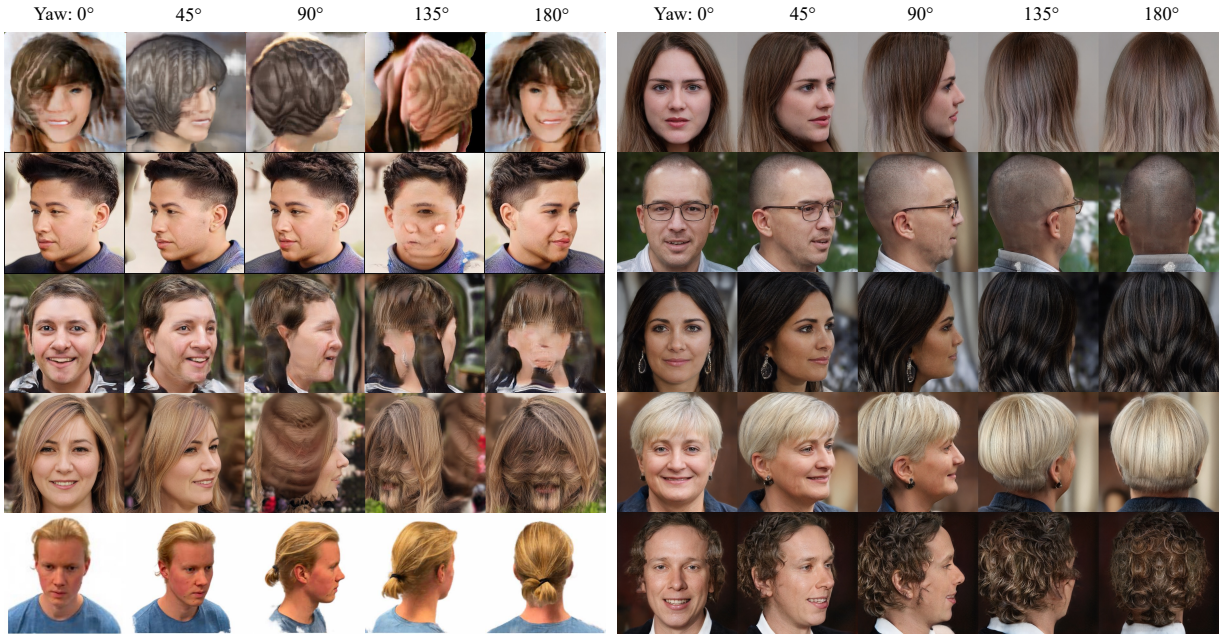
To reduce the inductive bias of the tri-plane, we lift its formulation into a higher dimension by augmenting tri-plane with an additional depth dimension. We call this enriched version as a tri-grid. Instead of having three planes with a shape of $H \times W \times C$ with H and W being the spatial resolution and C being the number of channel, each of our tri-grid has a shape of $D \times H \times W \times C$, where D represents the depth. For instance, to represent spatial features on the XY plane, tri-grid will have D axis-aligned feature planes $P_i^{XY}, i = 1, \ldots, D$ uniformly distributed along the Z axis. We query any 3D spatial point by projecting its coordinate onto each of the tri-grid, retrieving the corresponding feature vector by *tri-linear interpolation*. As such, for two points sharing the same projected coordinates but with different depths, the corresponding feature would be likely to be interpolated from non-shared planes (Figure 4 (b)). Our formulation disentangles the feature presentation of the front face and back head and therefore largely alleviates the mirroring-face artifacts (Figure 5).

Similar to tri-plane in EG3D [5], we can synthesize the tri-grid as $3 \times D$ feature planes using the StyleGAN2 generator [21]. That is, we increase the number of output channels of the original EG3D backbone by D times. Thus, tri-plane can be regarded as a naïve case of our tri-grid representation with $D = 1$. The depth D of our tri-grid is tunable and larger D offers more representation power at the cost of additional computation overhead. Empirically we find a small value of D (*e.g.* $D = 3$) is sufficient in feature disentanglement while still maintaining its efficiency as a 3D scene representation.

3.4. Self-Adaptive Camera Alignment

For adversarial training of our full head in 360°, we need in-the-wild image exemplars from a much wider range of camera distribution than the mostly frontal distribution, as in FFHQ [21]. Although our 3D-aware GAN is only trained from widely-accessible 2D images, the key to the best quality training is accurate alignment of visual observations across images labeled with well-estimated camera parameters. While a good practice has been established for frontal face images cropping and alignment based on facial landmarks, it has never been studied in pre-processing large-pose images for GAN training. Both camera estimation and image cropping are no longer straightforward due to the lack of robust facial landmarks detection for images taken from the side and back.

To resolve the aforementioned challenge, we propose a novel two-stage processing. In the first stage, for images with detectable facial landmarks, we still adopt the standard processing where the faces are scaled to a similar size and aligned at the center of the head using state-of-the-art face pose estimator 3DDFA [14]. For the rest of the images with large camera poses, we employ a head pose estimator WHENet [52] that provides a roughly-estimated camera pose, and a human detector YOLO [18] with a bounding box centered at the detected head. To crop the images at a consistent head scale and center, we apply both YOLO and 3DDFA on a batch of front-face images, from which



(a) Baselines: GRAF, GIRAFFEHD, StyleSDF, EG3D, a full-head NeRF (b) PanoHead with different seeds

Figure 7. Qualitative comparison between GRAF [37], GIRAFFEHD [48], StyleSDF [31], EG3D [5], multi-view supervised NeRF [43] (different methods from top to bottom on left side), and our PanoHead (right). Except [43], all models are trained on FFHQ-F. We render the results at a yaw angle of 0, 45, 90, 135, and 180°. GRAF, GIRAFFEHD, and StyleSDF fail to model the correct camera distribution in latent space due to the unsupervised camera pose mechanism, thus not turning to the back. EG3D is able to rotate to the back with 'mirroring face' artifacts and entangled background. Multi-view supervised NeRF is comparable to ours, however, it requires multi-view data of a single person and is not a generative model.

we adjust the scale and translation of the head center of YOLO with constant offsets. This approach enables us to pre-process all head images with labeled camera parameters and in a consistent alignment to a large extent.

Due to the presence of various hairstyles, there is still inconsistency in the alignment of back head images, inducing significant learning difficulties for our network to interpret the complete head geometry and appearance (see Figure 6 (a)). We, therefore, propose a self-adaptive camera alignment scheme to fine-tune the transformation of volume rendering frustum for each training image. Specifically, our 3D-aware GAN associates each image with a latent code z that embeds the 3D scene information of geometry and appearance, which can be synthesized at a view of c_{cam}. c_{cam} might not align well with the image content for our training images; so, it is hard for the 3D GAN to figure out a reasonable full head geometry. Therefore we co-learn a residual camera transformation Δc_{cam} mapped from (z, c_{cam}) together with our adversarial training. The magnitude of Δc_{cam} is regularized with a L_2 norm. Essentially, the network dynamically self-adapts the image alignment with refined correspondence across different visual observations. We note that this is only possible credited to the nature of 3D-aware GAN that can synthesize view-consistent images at various cameras. Our two-stage alignment enables 360-degree view-consistent head synthesis with authentic shape and appearance, learnable from diverse head images with widely distributed camera poses, styles, and structures.

4. Experiments

4.1. Datasets and Baselines

We train and evaluate our framework on a balanced combination of FFHQ [21], K-hairstyle dataset [24], and an in-house large-pose head image collection. FFHQ contains $70K$ diverse high-resolution face images, yet mainly fall in the absolute yaw range from 0° to 60°, assuming up-front camera pose corresponds to 0°. We augment the FFHQ dataset with $4K$ back-head images from K-hairstyle dataset and $15K$ in-house large-pose images with diverse styles, ranging from 60° to 180°. For brevity, we name this dataset combination as FFHQ-F. We refer to the supplementary paper for more dataset analysis and network training details.

We compare against state-of-the-art 3D-aware GANs including GRAF [37], EG3D [5], StyleSDF [31], and GIRAFFEHD [48]. All baselines are retrained from the same FFHQ-F dataset. We measure the quality of generated multiview images and geometry both quantitatively and qualitatively.

	GRAF	GIRAFFEHD	StyleSDF	EG3D	Ours
FID-all ↓	68.2	37.3	78.5	6.2	**5.4**
MSE (10^{-2}) ↓	N/A	42.6	N/A	N/A	**9.1**
ID ↑	N/A	0.39	0.41	**0.74**	**0.74**

Table 1. Metrics comparison across all baselines. For segmentation MSE, only GIRAFFEHD and PanoHead decouple the background and foreground. For ID score, GRAF's low-quality images lead to facial detection failure.

	EG3D	+seg.		+seg.&self-adapt.
		tri-plane	tri-grid	tri-grid
FID-back ↓	50.4	44.1	44.0	**40.9**
FID-front ↓	6.6	**5.0**	5.5	5.4
FID-all ↓	6.2	**5.2**	**5.2**	5.4
IS-back ↑	4.3	3.9	4.2	**4.4**
IS-front ↑	3.9	**4.1**	**4.1**	**4.1**
IS-all ↑	3.8	4.0	4.0	**4.1**
Runtime ↓	**1**	1.14×	1.26×	1.28×

Table 2. Ablation studies on different components. +seg. means with foreground-aware tri-discrimination. +self-adpat. means with camera self-adaptation scheme. All are trained with FFHQ-F

4.2. Qualitative Comparisons

360° Image Synthesis. Figure 7 visually compares the image quality against the baselines, all trained with FFHQ-F, by synthesizing images from five different views, ranging the yaw angle from 0 to 180°. GRAF [37] fails to synthesize compelling head images and its background is entangled with foreground head. StyleSDF [31] and GIRAFFEHD [48] are able to synthesize realistic frontal face images but in low perceptual quality when rendered from a larger camera pose. Without explicit reliance on camera labels, we suspect the above methods have difficulty in interpreting the 3D scene structures by themselves directly from images with 360° camera distribution. We observe that EG3D [5] is able to synthesize high-quality view-consistent frontal head images before rotating the view to the side or even the back. Mirroring face artifacts are clearly observable from the back, due to the tri-plane's projection ambiguity and the entangled fore-background. The method proposed in [43] builds personalized full-head NeRF at the extra cost of multi-view supervision. Regardless of its good quality images at all views, the model itself is not a generative model. In strong contrast, our model generates superior photo-realistic head images *for all camera poses* while retaining multi-view consistency. It delivers photo-realism with fine details at diverse appearances, ranging from shaved head with glasses to long curly hairstyles. To better appreciate our multi-view full-head synthesis, please refer to our supplementary video for more comprehensive visual results.

Geometry Generation. Figure 8 compares the visual quality of the underlying 3D geometry extracted by running Marching Cubes algorithms [26]. While StyleSDF [31] generates decent appearances of the front face, the complete geometry of the head is noisy and broken. EG3D presents detailed geometry of front face and hair, but either with background concrete entangled (Figure 3(a)) or with a hollowed back head (Figure 8). In contrast, our model can consistently generate high-fidelity background-free 3D head geometry even with various hairstyles.

4.3. Quantitative Results

To quantify the visual quality, fidelity, and diversity of the generated images, we employ Frechet Inception Distance (FID) [16] of 50K real and fake image samples. We measure the multi-view consistency using the identity similarity score (ID) by calculating the average Adaface [23] cosine similarity score from paired synthesized face images rendered from different camera poses. Note that this metric can only be applied to those images with detected facial landmarks. We assess mean square error (MSE) to calculate the accuracy of the generated segmentation against the mask obtained with DeepLabV3 ResNet101 network [7]. Table 1 compares these metrics across all baselines and our method. We observe that our model outperforms other baselines consistently from all perspectives. Refer to supplemental material for metrics definition and implementation details.

To evaluate the image quality at different views, we employ FID and Inception Score (IS) [36] for synthesized images with only back poses ($|yaw| \geq 90°$), front poses ($|yaw| < 90°$), and all camera poses. FID measures on the similarity and diversity of real and fake image distributions while IS focuses more on the image quality itself. Our GAN model follows EG3D for the main backbone, where the tri-plane generator is conditioned on a camera pose. We observe that such a design leads to biased image synthesis quality toward the conditioning camera pose. Specifically, when conditioning on the front view, our generator achieves inferior quality for synthesizing the head images from the back, and vice versa. However, when calculating FID-all, the conditioning camera is always the same as the rendering view. Therefore the generator could still achieve an excellent FID-all score even though the quality of generated heads might degenerate in unseen views. Hence, the original FID metrics (FID-all and FID-front) can hardly thoroughly reflect the overall generation quality of full heads in 360°. To alleviate this issue, we propose FID-back, where we condition on the front view but synthesize the images from the back. It leads to higher FID scores but reflects the quality in 360° image synthesis better.

We perform an ablation study on our method to quantitatively evaluate the efficacy of each individual component (Table 2). As shown in the second column, we notice a sig-

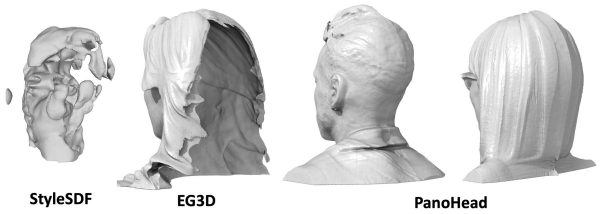


Figure 8. PanoHead achieves high-quality complete head geometry whereas StyleSDF [31] and EG3D [5] produce 3D noises or hallowed heads.

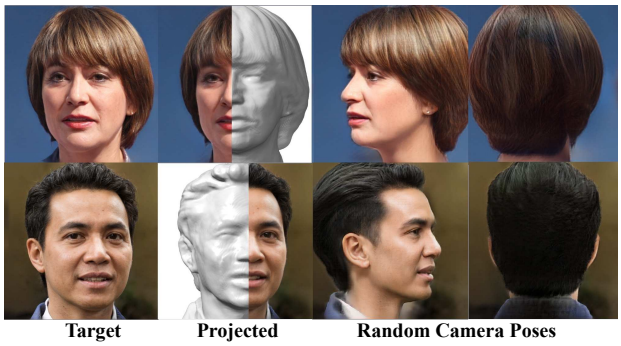


Figure 9. Single-view reconstruction from different camera poses. The first column shows the target images, second column projected RGB images and reconstructed 3D shapes using GAN inversion, last two columns rendered images from any given camera poses.

nificant quality boost after adding the foreground-aware discrimination for all cases, compared with the original EG3D. That indicates the prior segmentation knowledge largely ease the network learning difficulty of 3D heads from in-the-wild image collections. Frontal face synthesis quality is comparable among all methods given the strong supervision from the large amount of well-aligned frontal images. However, for the back head, decoupling foreground and background largely improves the synthesis quality. In addition, changing tri-plane to tri-grid representation further enhances the image quality. With tri-discrimination, tri-grid, and camera self-adaption scheme altogether, PanoHead achieves the lowest FID-back and the highest IS for back head generation. As reflected in the row of run-time analysis, our novel component only introduces minor computation overhead, but with significant image synthesis quality improvements. Note that the frontal image quality is superior to the back head, largely due to the significant learning difficulty in various hairstyles and unstructured back-head appearances.

4.4. Single-view GAN Inversion

Figure 9 demonstrates full-head reconstruction from a single-view portrait using PanoHead's generative latent space. To achieve that, we first perform an optimization to find the corresponding latent noise z for the target image using pixel-wise L_2 loss and image-level LPIPS loss [51]. To further improve reconstruction quality, we perform pivotal tuning inversion (PTI) [35] to alter the generator parameters with a fixed optimized latent code z. From a single-view target image, PanoHead not only reconstructs photo-realistic image and high-fidelity geometry but also enables novel-view synthesis in 360°, including large pose and back head.

5. Discussion

Limitations and Future Work. While PanoHead exhibits excellent images and shapes quality from 360°, it still contains minor artifacts, e.g. in the teeth area. Similar to the original EG3D, flickering texture issue is also noticeable in our model. Switching to StyleGAN3 [20] as the backbone would help preserve high-frequency details. In practice, we also observe more noticeable flickering artifacts with a higher swapping probability of the conditional camera pose. We set this value to 70% as opposed to 50% in EG3D since we empirically find it enhances 360° rendering quality but at the minor cost of flickering texture artifacts. Another observation is that it lacks finer high-frequency geometric details, e.g. hair tips. We leave it as future work to quantitatively evaluate our geometric quality such as using depth maps. Finally, although PanoHead is able to generate diverse images in terms of gender, races, and appearances, reliance on training with only several datasets combination still makes it suffer from data bias, to some extent. In spite of our data collection effort, large-scale full-head annotated training image dataset is one of the most critical directions to facilitate full-head synthesis research. We anticipate such datasets can resolve some of the limitations aforementioned.

Ethical considerations. PanoHead is not specifically designed for any malicious uses, yet we do realize that the single-view portrait reconstruction could be manipulated, which might pose a social threat. We do not encourage the method being used for violating others' rights in any forms.

6. Conclusion

We propose PanoHead, the first 3D GAN framework that synthesizes view-consistent full head images with only single-view images. With our novel design in foreground-aware tri-discrimination, 3D tri-grid scene representation, and self-adaptive image alignment, PanoHead enables authentic multiview-consistent full-head image synthesis in 360° and demonstrates compelling qualitative and quantitative results compared with state-of-the-art 3D GANs. Furthermore, we present 360-degree photo-realistic reconstruction with highly detailed geometry from single-view real portraits. We believe the proposed method presents an interesting direction for 3D portraits creation, which sheds light on many potential downstream tasks.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *iccv*, pages 5855–5864, 2021. 2

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

[3] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 2

[4] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. *CVPR*, 2022. 2, 3, 4, 5, 6, 7, 8

[6] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 2, 3

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[8] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *CVPR*, 2022. 3

[9] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14304–14313, 2021. 2

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

[11] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *nips*, 2022. 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 3

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *CVPR*, 2022. 3

[14] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020. 5

[15] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *iccv*, 2021. 2

[16] 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. *NeurIPS*, 2017. 7

[17] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *cvpr*, pages 20374–20384, 2022. 2

[18] Marina Ivašić-Kos, Mate Krišto, and Miran Pobar. Human detection in thermal imaging using yolo. In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, pages 20–24, 2019. 5

[19] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 2021. 1, 3

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 8

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 3, 4, 5, 6

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 3, 4

[23] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022. 7

[24] Taewoo Kim, Chaeyeon Chung, Sunghyun Park, Gyojung Gu, Keonmin Nam, Wonzo Choe, Jaesung Lee, and Jaegul Choo. K-hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1299–1303. IEEE, 2021. 6

[25] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. 36(6):194:1–194:17, 2017. 2

[26] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 7

[27] Julien NP Martel, David B Lindell, Connor Z Lin, Eric R Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *arXiv preprint arXiv:2105.02788*, 2021. 2

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[29] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2, 3

[30] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 597–614. Springer, 2022. 3, 4

[31] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *CVPR*, 2022. 2, 3, 6, 7, 8

[32] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 2

[33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2

[34] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. *CVPR*, 2021. 2

[35] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2021. 8

[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 2016. 7

[37] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 2020. 2, 3, 6, 7

[38] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *arXiv preprint arXiv:2206.07695*, 2022. 3

[39] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *cvpr*, pages 6258–6266, 2021. 3

[40] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 2, 3

[41] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and HongSheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *arXiv preprint arXiv:2206.08361*, 2022. 3

[42] Attila Szabo, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv*, 2019. 3

[43] Stanislaw Szymanowicz, Virginia Estellers, Tadas Baltrusaitis, and Matthew Johnson. Photo-realistic 360 head avatars in the wild. *arXiv preprint arXiv:2210.11594*, 2022. 2, 6, 7

[44] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018. 2

[45] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019. 2

[46] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *arXiv preprint arXiv:2210.06465*, 2022. 3

[47] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022. 2

[48] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 2, 3, 6, 7

[49] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 3

[50] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. *arXiv preprint arXiv:2208.00561*, 2022. 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *cvpr*, pages 586–595, 2018. 8

[52] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. In *BMVC*, 2020. 5

[53] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. *arXiv preprint arXiv:2112.02308*, 2021. 2