# SphereHead: Stable 3D Full-head Synthesis with Spherical Tri-plane Representation

Heyuan Li[1], Ce Chen[1], Tianhao Shi[1], Yuda Qiu[1], Sizhe An[3], Guanying Chen[1], and Xiaoguang Han[1,2]✉

[1] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[2] The Future Network of Intelligence Institute, CUHK-Shenzhen
[3] University of Wisconsin-Madison

**Fig. 1:** Given input images, applying PTI [28] methods to reconstruct full-head synthesis based on PanoHead [1] often yields noticeable artifacts (green box), while those using our SphereHead space (blue box) appear more realistic.

**Abstract.** While recent advances in 3D-aware Generative Adversarial Networks (GANs) have aided the development of near-frontal view human face synthesis, the challenge of comprehensively synthesizing a full 3D head viewable from all angles still persists. Although PanoHead [1] proves the possibilities of using a large-scale dataset with images of both frontal and back views for full-head synthesis, it often causes artifacts for back views. Based on our in-depth analysis, we found the reasons are mainly twofold. First, from network architecture perspective, we found each plane in the utilized tri-plane/tri-grid representation space tends to confuse the features from both sides, causing "mirroring" artifacts (e.g., the glasses appear in the back). Second, from data supervision aspect, we found that existing discriminator training in 3D-aware GANs mainly focuses on the quality of the rendered image itself, and does not care much about its plausibility with the perspective from which it was rendered. This makes it possible to generate "face" in non-frontal views, due to its easiness to fool the discriminator. In response, we propose SphereHead, a novel tri-plane representation in the spherical coordinate system that fits the human head's geometric characteristics and efficiently mitigates many of the generated artifacts. We further introduce a view-image consistency loss for the discriminator to emphasize the correspondence of the camera parameters and the images. The combination of these efforts results in visually superior outcomes with significantly fewer artifacts. Our code and dataset is publicly available at `https://lhyfst.github.io/spherehead/`.

**Keywords:** Head Synthesis · 3D GANs · Spherical Tri-plane

## 1 Introduction

The synthesis of human head portrait images has long been of great concern within the realms of computer vision and graphics, with extensive applications spanning video conferencing, digital avatars, games, movies, and more. While 2D GANs are well-studied and can generate photorealistic facial images, they encounter challenges when applied to multi-view scenarios, where the relative position and orientation of the camera and the head are in flux. The recent developments in neural implicit representation have paved the way for the genesis of 3D-aware GANs, which achieve photorealistic view-consistent image synthesis. A groundbreaking work in this realm is EG3D [5], which puts forth an efficient and effective tri-plane representation in Cartesian space. This approach has been widely adopted by numerous 3D-aware synthetic head image generators that followed. However, these generators are restricted to near-frontal synthesis and are unable to produce back-view synthesis due to the scarcity of high-quality multi-view data.

A comprehensive 3D head synthesis, viewable from all angles, is pivotal for enhancing immersive experiences, particularly in interactive scenarios such as 3D video games and digital avatars. PanoHead [1], the first 3D-aware GAN for full head synthesis, addresses this need to an extent. It enables the training with monocular back-view images using an adaptive camera strategy and extends the tri-plane representation into a tri-grid version for back-view synthesis. While the results affirm the feasibility of full-view synthesis without 3D data, they frequently exhibit noticeable semantic mismatch artifacts, with the most pronounced being the fake face on the back-view head, as depicted in Fig. 2. This can be traced back to two aspects.

From the network architecture perspective, representations based on the Cartesian coordinate system, such as tri-plane and tri-grid, naturally introduce feature entanglement for symmetrical areas relative to the predefined planes. For example, the back-view head shares the same features on 'XY' plane with the front-view face, resulting in *mirroring-face artifacts* on the back of head or unreasonable neck shape, as depicted in Fig. 2 (a, b).

From the data supervision aspect, in 3D-aware generation, it's crucial to ensure the correspondence between the generated images and their viewing perspectives. However, while existing 3D-aware GANs take both images and camera parameters as inputs for the discriminator, they fail to clearly distinguish between two unwanted cases: unrealistic images paired with their appropriate camera parameters, and realistic images accompanied by incorrect camera parameters. This causes the discriminator tends to score the generated data pairs by the visual quality of the images and weakens the consistency between the images and their viewing perspectives. This becomes especially problematic in scenarios with marked imbalances in supervision intensity across different directions, such as with datasets of head images. As a result, the generator tends to

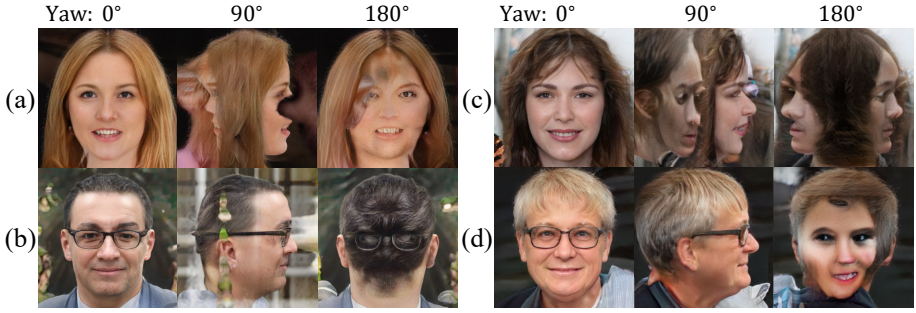


**Fig. 2:** Two types of fake face artifacts in PanoHead. (a-b) We name the first type as *mirroring-face artifacts*, due to the back face mirroring the identity, expression and accessories of the front face precisely. (c-d) We name the second type as *multiple-face artifacts*, because in this scenario there might be more than one fake faces and their identities, expression and accessories are different from the front face.

create heavily supervised content, like faces, in incorrect orientations, resulting in *multiple-face artifacts* as depicted in Fig. 2 (c, d).

In this paper, we propose a novel framework, SphereHead, to enhance panoramic head image synthesis. Our method introduces an entanglement-free representation that reduces mirroring artifacts. In particular, regarding the geometric characteristics of the human head, we propose a dual spherical tri-plane representation which is defined in the spherical coordinate system. This sphere's geometric properties explicitly separate the features of each position on the head, thereby resolving mirroring artifacts completely.

Moreover, we introduce a view-image consistency loss to enforce the discriminator focusing on identifying unaligned pairs of images and camera parameters, even if the quality of images are exceedingly high. We achieve this by intentionally pairing real images with shuffled real camera parameters, and designating these pairs as negative samples for the training of discriminator. With such additional supervision from negative real pairs, the discriminator manages to judge an image not only by its fidelity but also by the plausibility of its viewing perspective. Thereby, we successfully tackle the two types of semantic mismatch artifacts by the spherical tri-plane representation and the view-image consistency loss respectively.

Additionally, although an in-house dataset of back-view head images plays a crucial role in PanoHead, it remains inaccessible to the research community. In response, we have compiled a large-scale dataset featuring 60k head images from various viewpoints, including the back view. This dataset was developed through extensive data collection and thorough cleaning processes. It not only supports our research but is also publicly available, along with a comprehensive data processing toolbox.

Through extensive qualitative and quantitative experiments, our method demonstrates an improved ability to produce 3D full head synthesis. These re-

sults exhibit enhanced visual quality and significantly fewer artifacts compared to existing methods. In summary, the main contributions of this paper are:

- We proposed a novel framework, SphereHead, which significantly reduces the noticeable artifacts of PanoHead, and effectively addresses the challenge of lacking 3D data for stable full-head synthesis.
- We developed a novel spherical tri-plane representation, which effectively solved the feature entanglement issue of existing regular tri-plane/tri-grid representation.
- We introduce a view-image consistency loss to compel the discriminator to focus on the alignment between images and their viewpoints, thereby minimizing artifacts arising from imbalanced supervision.
- We collected a large-scale head image dataset, the WildHead Dataset, whose images are captured in the wild from various viewing angles spanning 360 degrees. This dataset is publicly available on our project page.

## 2 Related Work

**3D Head Modeling.** Considerable research has focused on explicitly representing a 3D human head using textured mesh, such as 3DMM [2], BFM [26] and FLAME [17]. These models typically ascertain the distribution of shape and texture variations from a substantial collection of 3D head scans, employing Principle Component Analysis (PCA). Nonetheless, their pre-defined topologies limit the capture of details and accessories such as hair, glasses, and wrinkles, leading to the generation of unrealistically rendered images. In the other line, an increasing number of recent 3D head synthesis methods have adopted neural implicit representation due to its high expressivity. Many such studies [4, 37, 41] represent 3D heads via Signed Distance Functions (SDFs), which model 3D shapes by defining the signed distance from every 3D location to its nearest surface. Another group of works [10, 12, 13, 25, 33, 36, 38, 43] represents the 3D head using Neural Radiance Fields (NeRFs) [20], which map a 3D location and viewing direction to a color and a density value, then synthesize images through volume rendering.

**Generative 3D-Aware Image Synthesis.** 3D-aware GANs, learning from collections of 2D images, aim to synthesize view-consistent images from various viewpoints by capturing the inherent 3D geometry and appearance of the scene. While earlier literature explicitly encodes the 3D scene using mesh [18,30,32] and voxel grids [9,22,23], a growing number of studies [6,7,24,29] utilize implicit representations which are more expressive, thereby demonstrating high-quality image generation capabilities. Recently, EG3D [5] puts forward a tri-plane implicit-explicit hybrid 3D representation, which later works have extensively adopted, notably for its efficiency. PanoHead [1] enriched the tri-plane's representational capacity through adding more parallel feature planes, thereby lessening the artifacts in back-view head image synthesis. Nonetheless, persistent artifacts endure, as its tri-grid representation retains a similar geometric structure, which is fundamentally the source of the artifacts. In contrast, our method, through a shift

in formulation from a Cartesian coordinate representation in cubic space to a spherical coordinate representation in spherical space, avoids many artifacts at their root.

**View-Image consistency in 3D-aware GAN** Inspired by cGAN [21], [8, 11] explore the training of conditional GAN for image generation. They directly add the conditional labels for both of the generator G and the discriminator D. The EG3D-like methods work in the same manner to maintain the view-image consistency. However, the experiments show the networks fail to ensure the consistency while training in dataset with biased distribution. To further compel D to identify image-label matching in addition to the image realism, [27] propose to train D with additional negative real pairs consisting of real images with mismatched labels. We adapt this as a view-image consistency loss to emphasize the correspondence of the camera labels and the images in our 3D-aware synthesis framework.

**Human Head Portrait Datasets.** To ensure the diversity and quality of synthetic images, generative models necessitate learning from large-scale, high-quality portrait images. CelebA [19] and FFHQ [15] are two of the most widely-used single-view portrait datasets, encompassing 200,000 and 70,000 images respectively. However, as the majority of their images are captured from narrow, near-frontal views, they are inadequate for training a full head generator. The LPFF dataset [34], while offering a larger viewing angle for portrait images, is still deficient in back-view images. The distribution bias in these datasets impedes the learning of 3D full head synthesis. Following the practice in PanoHead [1], we collect a dataset with back-view head images to enable the training of full head generative model.

## 3 SphereHead

Our novel SphereHead extends upon the full-head synthesis work of PanoHead by incorporating a dual spherical tri-plane representation (Sec. 3.1), a view-image consistency loss (Sec. 3.2), and a parsing branch (Sec. 3.3).

As depicted in Fig. 3, our generator $G$ comprises a modified StyleGAN2 generator, augmented with two additional sub-modules. These modules synthesize two distinct feature groups $f_A$ and $f_B$, each consists of three 2D feature planes: $P_{\theta r}$, $P_{\phi r}$, and $P_{\theta\phi}$ corresponding to the spherical coordinate system. We merge $f_A$ and $f_B$ into a fused tri-plane feature $f_F$ using a carefully designed weight map. With $f_F$ and the rendering camera poses $c_r$, a neural renderer $R$ generates a set of low-resolution attribute maps. These maps are further processed into a super-resolved image $I^+$, a bilinear upsampled image $I$, a super-resolved foreground mask $I^{m+}$, and a bilinear-resolved face parsing $I^p$. The discriminator $D$ then critiques generated $(I^+, I, I^{m+}, I^p)$ and $c_r$, alongside real images paired with matching and mismatched real camera parameters.

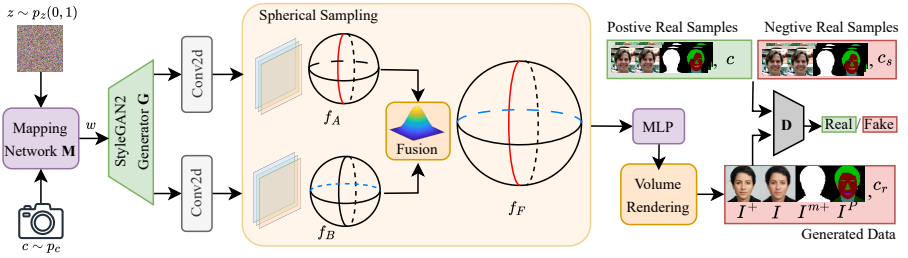


**Fig. 3:** The framework of our proposed SphereHead. Given a sampled code $z$ and camera parameter $c$, SphereHead synthesizes a spherical tri-plane features $f_F$ by fusing two sub-feature groups $f_A$ and $f_B$. By volumetric rendering with the sampled features in $f_F$, SphereHead generates high-quality view-consistent full head images $I^+$. The discriminator learns to focus on the alignment between images and their viewpoints instructed by our view-image consistency loss, by introducing an additional negative data pairs consisting the real images and mismatched labels $c_s$.

### 3.1 Dual Spherical Tri-plane Representation

The primary challenge in 3D full-head synthesis lies in generating a plausible occipital region, particularly when dealing with complex and diverse hairstyles. On the one hand, the fundamental issue stems from biased supervision arising from imbalanced training data. Specifically, the abundance of high-quality frontal view data fosters strong supervision for frontal view synthesis learning. In contrast, back view synthesis learning is hindered due to deficient back view data, often marred by unreliable camera pose estimation, insufficient high-quality images, and subpar image alignment. This data imbalance impedes the learning process for back view synthesis, resulting in numerous artifacts on the occipital region.

On the other hand, the prevalent tri-plane representation is intrinsically prone to fostering feature entanglement due to its inherent structure and Cartesian feature projection. Specifically, the same feature on a plane is symmetrically shared by disparate positions about the plane via Cartesian projection, leading to undesirable mirroring artifacts. This issue becomes especially pronounced within the occipital region due to the imbalanced supervision between the frontal and back views. The facial features from the front view, which have been powerfully supervised, dominate the $P_{XY}$ feature map and permeate into the poorly supervised occipital region, leading to striking mirroring-face artifacts. Although PanoHead's tri-grid representation mitigates this issue to a degree by incorporating additional feature planes, it retains a similar structure and operates within the same Cartesian coordinate system, leading to the persistence of mirroring artifacts in practice.

To this end, we propose a novel spherical tri-plane representation, which explicitly separates features from different facial regions. Specifically, we represent a point in the spherical coordinate system as $(r, \theta, \phi)$. Correspondingly, as illustrated in Fig. 4 (b), the three feature planes in our spherical tri-plane

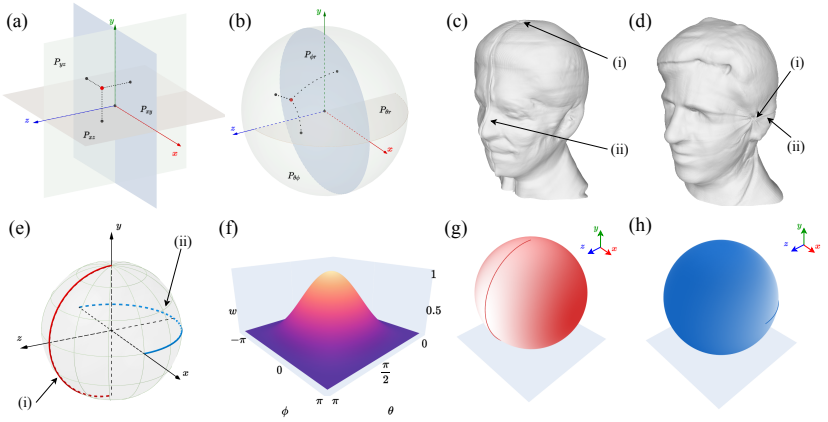


**Fig. 4:** (a) Tri-plane representation. (b) Spherical tri-plane representation. Reconstructed head geometry from single (c) sphere A and (d) sphere B, each showing (i) seam artifacts and (ii) polar artifacts. (e) The combination of two spheres in dual spherical tri-plane representation, (i) the seam of sphere A, (ii) the seam of sphere B. (f) Fusion weight map. (g-h) For each sphere, the weight approaches zero as the locations near the seam and poles.

representation embody a semicircle plane $P_{\theta r}$, a circular plane $P_{\phi r}$, and a spherical plane $P_{\theta\phi}$. We query a 3D position $x \in \mathbb{R}^3$ by projecting it onto each of the three feature planes via its spherical coordinate, retrieving the respective feature vector $(F_{\theta r}, F_{\phi r}, F_{\theta\phi})$ through bilinear interpolation, and subsequently merging them via summation. A lightweight decoder network interprets the combined 3D features $F$ as color and density, which are used to render RGB images through volume rendering.

Our spherical tri-plane representation is demonstrated to entirely eliminate mirroring artifacts referring to Sec. 5.2. However, in practice, these three feature planes are wrapped from square feature maps, which leads to potential issues in $P_{\theta\phi}$ and subsequently, the emergence of new artifacts in the seam and polar regions, clearly depicted in Fig. 4 (c, d). The former stems from the numerical discontinuity of $F_{\theta\phi}$ at $\phi = -\pi$ and $\phi = \pi$. Meanwhile, the latter comes to light when $\theta$ nears 0 and $\pi$, leading to the convergence of $\phi = -\pi$ and $\phi = \pi$ lines to two polar points. Such transformation converts the numerical fluctuations on these lines into high-frequency noise around the poles.

To contend with these complications, we introduce an additional spherical tri-plane, orthogonal to the existing one, i.e. their polar axes are mutually perpendicular, and seam openings are oriented in opposite directions, which is shown in Fig. 4 (e). This complementarity is encapsulated by a 2D weighting map, given by:

$$w(\theta, \phi) = \frac{1 + \cos(2\theta - \pi)}{2} \cdot \frac{1 + \cos(\phi)}{2}. \tag{1}$$

As indicated in Fig. 4 (f-h), $w$ attains maximum value at its center and gradually recedes to 0 at its edges, where the seam and poles are. Due to this arrangement, the two planes wrap around each other akin to the two leathers of a baseball, each utilizing its smooth region to compensate for the problematic areas of the other. By implementing this strategy, we successfully eradicate the seam and polar artifacts.

### 3.2 View-image Consistency Loss (ViCo Loss)

In addition to mirroring artifacts, the imbalanced supervision also gives rise to the multiple-face artifacts depicted in Fig. 2 (c, d). In these instances, the fake back faces exhibit divergent attributes compared to the frontal faces, indicating that they are not the result of feature entanglement.

This supervisory imbalance skews the generator's ability to synthesize high-quality frontal faces, while significantly impairing its capacity to generate satisfactory posterior views. As detailed in [8], in early training, discriminators tend to assess authenticity based on image quality rather than fidelity to specific attributes. Consequently, to deceive the discriminator, the generator opts to replicate facial features even in non-frontal images, as this approach more readily enhances visual quality than the more challenging task of accurately generating the occipital region. This compulsion to enhance visual quality by facial features undermines the authentic generation of the posterior aspect of the head, resulting in a consistent production of artifact-laden back-view images that fall short of real image realism.

This loop creates a negative feedback mechanism: the persistent artifacts prevent the discriminator from evolving its criteria from mere visual quality to a more nuanced analysis of the congruence between images and their corresponding viewpoints. Consequently, the training process becomes ensnared in a dilemma, undermining the model's overall ability to generate convincing images across varying viewpoints.

Our view-image consistency loss is designed to disrupt this negative feedback loop by compelling the discriminator to simultaneously evaluate input images based on both their view-image consistency and visual quality. In practice, we construct negative pairs using real data to facilitate our view-image consistency loss. Specifically, for a batch of input data $\{(I^+, I, I^{m+}, I^p), c\}$, we shuffle the camera labels $c$ to obtain $c_s$. The mismatched negative pairs $\{(I^+, I, I^{m+}, I^p), c_s\}$ are then fed into the discriminator $D$, as illustrated in Fig. 3. The view-image consistency loss for the discriminator is formulated as following:

$$\mathbb{L}_{ViCo} = log(1 - D((I^+, I, I^{m+}, I^p), c_s)). \tag{2}$$

Thus, the discriminator can separate supervision by direction, thereby protecting poorly supervised directions from the undue influence of those with stronger supervision. As a result, the generator benefits from purer back-view supervision and steadily enhancing the visual and content accuracy of the back-view image synthesis.

### 3.3 Parsing Branch for Semantic Embedding

In the renderer $R$, along with color and density, we additionally obtain parsing feature from 3D features $F$, which is further rendered into face parsing map $I^p$. Although previous works [14, 31, 38, 42] exploit a similar parsing branch for controlled synthesis, they do not consider its potential contribution to generator learning. However, our experiments (Sec. 5.4) indicate that learning facial parsing benefits the generator in terms of visual quality and artifacts reduction. The learning of low-frequency information in face parsing enhances the semantic understanding of each face region, driven the generator to smooth the features in the same region while distinguish those from different regions.

## 4 Dataset

**WildHead Dataset.** Although PanoHead has crafted a dataset of back-view head images, its exclusivity has presented a challenge for the broader research community. To address this issue, we present the first publicly accessible large-scale dataset, the WildHead Dataset, whose images are captured in the wild from various viewing angles spanning 360 degrees.

**A Collection of Head Images for Full-Head Synthesis.** We incorporate our dataset with publicly available datasets, including CelebA [19], FFHQ [15], and LPFF [34], to train our full-head synthesis models. Besides, we also utilize partial back-view images in K-Hairstyle [16] dataset to enhance the back-view hairstyle diversity of the training set. To improve data quality, we carefully design a pipeline to filter and process all the images from the gathered dataset. Additionally, meticulous manual vetting has been employed to ensure the high quality of the dataset. As a result of these rigorous curation efforts, the final dataset comprises a total of 180k images.

**Data Processing and Open-Source Toolbox.** In order to train our model with the collected data, it is necessary to align the images and obtain the corresponding camera parameters. Following PanoHead [1], we align the portraits to the center of images, with the aid of facial landmarks detector [3] for near-frontal face images and Yolo Head Detector [40] for others. Then we estimate the extrinsic camera parameters with the assumption that all cameras have the same intrinsic matrix and are facing towards the origin, resulting in the camera labels of our dataset. To mitigate the imbalance of perspectives within the dataset, we replicated the images from sparsely represented viewpoints multiple times. The entire data processing pipeline has been packaged as a toolbox which is publicly available in our released code. Please refer to the supplementary materials for more detailed information.

## 5 Experiments

### 5.1 Training Details and Baselines

Our model is trained on eight NVIDIA A100 GPUs with a batch size of 32. The resolution for both the images within the training set and the subsequently

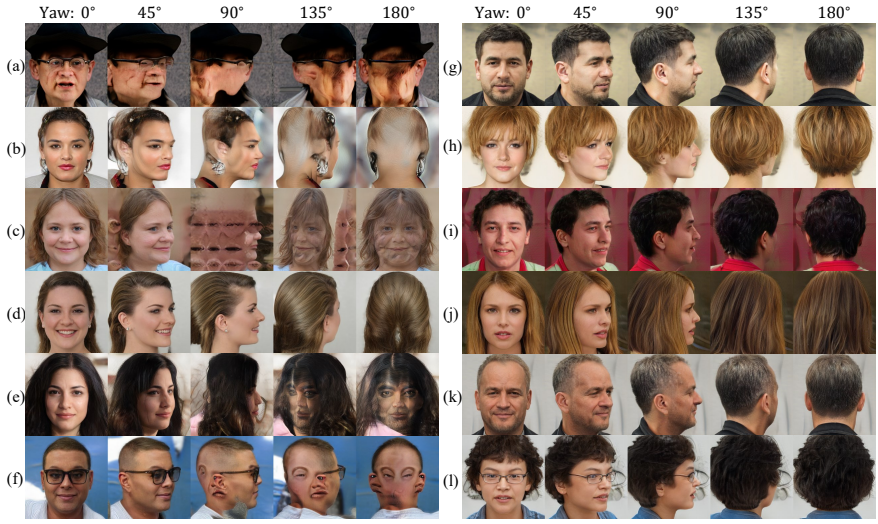


**Fig. 5:** Qualitative comparison with state-of-the-art methods. (a) GIRAFFHD [35], (b) StyleSDF [24], (c) EG3D [5] fail to capture the complete head geometry and appearance. (d-f) PanoHead [1] shows complete head generation, but the results suffer from mirroring artifacts ((d) left-right identical mirroring artifacts and (f) mirroring-face artifacts) and (e) multiple-face artifacts). (g-l) Ours SphereHead synthesizes full-head images of high visual quality and is free of artifacts exhibited by other methods.

synthesized images measures $512^2$. Over the span of 7 days, the thorough training process encompasses three distinct phases with 25 million images.

**Phase I** (0-2M images): our model is exclusively trained with near-front view images for their superior quality, which facilitates a robust initialization. We adopt an alternative training strategy, which randomly trains $f_A$, $f_B$, and $f_F$ with the probabilities of 33%, 33%, and 34% respectively. This approach accelerates the learning of both spherical tri-planes and curtails the formation of evident artifacts. The view-image consistency loss is not applied in the beginning, as it may hinder early generator training to improve image quality. We blur images as they enter the discriminator, gradually reducing the blur amount over this duration.

**Phase II** (2-10M images): We begin to train our model on the entire training dataset.

**Phase III** (10M-25M images): We integrate the view-image consistency loss in this stage. By re-calibrating the training probabilities for $f_A$, $f_B$, and $f_F$ to 10%, 10%, and 80% respectively, we enforce the generator to focus on the output of the fused sphere.

We compare against state-of-the-art 3D-aware GANs including GIRAFFEHD [35], StyleSDF [24], EG3D [5], and PanoHead [1]. Unless otherwise indicated, all baselines are trained on our comprehensive head image collection. We evaluate the quality of the generated multi-view images, quantitatively and qualitatively.

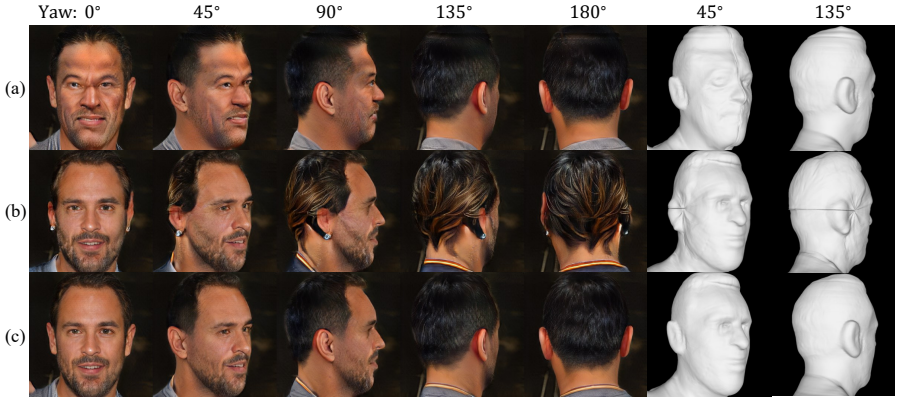


**Fig. 6:** Synthetic images and geometries rendered by (a) only $f_A$, (b) only $f_B$ and (c) fused $f_F$. While the individual spherical space exhibits notable seam and polar artifacts, the fused $f_F$ successfully mitigates the artifacts.

### 5.2 Qualitative Comparisons

**360° Image Synthesis.** Fig. 5 offers a visual comparison of image synthesis among our method and the baselines, presenting synthesized images from five distinct views with the yaw angle ranging from 0 to 180°. GIRAFFEHD and StyleSDF manage to produce photo-realistic front-view synthesis, but failed to generate reasonable images on larger viewing angles since their rendering does not explicitly consider camera poses. EG3D can render back-views to some extent. However, it tends to confuse the features of the frontal view with those of the back-view region, leading to mirroring face artifacts. Additionally, it suffers from floating artifacts due to its lack of proper foreground and background separation. PanoHead, as the first work targeting 3D full-head synthesis, delivers superior 360° synthesis results compared to other baseline methods. Nevertheless, PanoHead still struggles with outputs with mirroring features and multiple-face artifacts. In contrast, our proposed solution, SphereHead, consistently produces high-quality, artifact-free 360° synthesis, demonstrating its superiority over pre-existing methods.

**Dual Sphere Fusion.** We evaluate our dual spherical tri-plane design by rendering synthesized images and corresponding geometries separately from features $f_A$, $f_B$, and their fused $f_F$, as demonstrated in Fig. 6. The results indicate that the synthesized image and geometry of frontal view are dominated by $f_A$, while those of back-view region are primarily influenced by $f_B$. Employing our strategically developed weight map, these two features effectively integrate and support each other. While each individual spherical features exhibits notable seam and polar artifacts, $f_F$ strategically masks the flaws by fusing the artifact-free areas in $f_A$ and $f_B$. Consequently, our design successfully mitigates individual limitations, leading to an enhanced final result.

### 5.3 User Study

To fairly and quantitatively assess the capability of our method in reducing semantic mismatch artifacts, we conducted a rigorous user study. Within this study, we established two distinct metrics. A collection of 100 random samples from each model's latent space served as the test subjects. 37 participants were tasked with making two determinations for each sample: firstly, whether the back-view images exhibited any discernible fake facial artifacts, and secondly, whether the sample appeared artifact-free from all angles, indicating a high-quality result. The frequency of identifiable fake face artifacts across all samples was denoted as the *fake face rate (FF)*, while the prevalence of samples deemed artifact-free was referred to as the *high-quality rate (HQ)*. The resultant data from these evaluations contributed to both the qualitative comparisons in Tab. 1 and the ablation studies in Tab. 2, offering insights into our method's effectiveness in artifact reduction.

### 5.4 Quantitative Results

*Comparisons with SOTA* In line with prior researches on 3D-aware GANs, we employ the Frechet Inception Distance (FID) calculated from 50,000 real and synthesized image samples to quantitatively evaluate the visual quality, fidelity, and diversity of the generated images. Evidenced by possessing the lowest FID and FF, as well as the highest HQ as shown in Tab. 1, our SphereHead method surpasses other state-of-the-art methods in terms of image quality and artifact reduction.

| | [35] | [24] | EG3D | PanoHead | Ours |
|---|---|---|---|---|---|
| FID ↓ | 67.8 | 105.5 | 10.3 | 8.6 | **7.8** |
| FF ↓ | – | – | 76.1% | 23.7% | **0%** |
| HQ ↑ | – | – | 3.7% | 26.1% | **92.6%** |

**Table 1:** Quantitative comparison among our method and state-of-the-art 3D-aware GANs. GIRAFFEHD [35] and StyleSDF [24] are excluded from the user study because they completely fail to generate reasonable back-view head images.

*Ablation Studies* Tab. 2 offers a detailed assessment of each component in our framework, with gradual enhancements across different configurations.

The transition from a tri-grid to a single-sphere representation is notable for its considerable impact, as it completely resolves mirroring-face artifacts. The introduction of the second sphere further refines visual quality, by remedying polar and seam artifacts. Additionally, this modification unexpectedly assists in reducing multiple-face artifacts, a fact supported by a significant decrease in fake face rate from 3.7% to 0.9%. This indicates that the polar and seam regions on the single sphere are more prone to imbalanced supervision, due to their inherent

| Representation | ViCo Loss | Parsing | FID ↓ | FF ↓ | HQ ↑ |
|---|---|---|---|---|---|
| Tri-grid | – | – | 8.6 | 23.7% | 26.1% |
| Tri-grid | ✓ | – | 8.3 | 4.8% | 63.2% |
| Single-Sphere | – | – | 8.1 | 3.7% | 78.7% |
| Single-Sphere | ✓ | – | 7.9 | 0.3% | 85.6% |
| Dual-Sphere | – | – | 7.9 | 0.9% | 84.2% |
| Dual-Sphere | ✓ | – | 7.9 | **0%** | 90.4% |
| Dual-Sphere | ✓ | ✓ | **7.8** | **0%** | **92.6%** |

**Table 2:** Ablation studies of our SphereHead framework.

discontinuity on the unfolded feature plane. Subsequently, the additional sphere compensates these regions and ensures greater overall stability.

While the combination of the single-sphere model and our view-image consistency loss gains noteworthy performance, the incorporation of an additional sphere succeeds in completely eliminating fake faces in all the generated samples within this user study, with only minimal extra parameters and computation. Lastly, integrating parsing information further mitigate artifacts and yield synthesis results with higher quality. Note that in Tab. 2, the view-image consistency loss brings substantial enhancements across the representations including tri-grid, single-sphere and dual-sphere formats. This suggests that it can be readily integrated with other frameworks for broader applicability.

### 5.5 Single-view GAN Inversion

We compare our method with PanoHead on 3D full-head reconstruction from a single-view image. Both methods utilize an identical GAN inversion rendering technique. The procedure begins by ascertaining the corresponding latent code $z$ through optimization using pixel-wise L2 loss and image-level LPIPS loss [39]. This is followed by the application of Pivotal Tuning Inversion (PTI) [28] to alternate generator parameters, keeping the optimized latent code $z$ intact. As depicted in Fig. 1, although PanoHead constructs the coarse geometry of the posterior head region, it suffers from noticeable semantic mismatch artifacts. In contrast, our method achieves an artifact-free reconstruction of both the 3D head geometry and its corresponding renderings.

Additionally, we invited 39 participants to respond to a questionnaire that included 20 comparative scenarios. Each scenario presented the target image alongside results produced by both the PanoHead and our proposed model. Participants were tasked with determining which model generated the superior full-head visual quality and rationality. The study's results revealed that, barring one scenario, more respondents favored the outcomes generated by our model in all instances. Averaging across all scenarios, 82.1% of choices favored our model's results, with only 17.9% supporting the results from PanoHead. These findings reinforce the comparative advantage of our model over PanoHead in the realm of monocular 3D full-head reconstruction.

## 6 Discussion

*Discussion on Broader Usage* Our SphereHead demonstrates excellent feature learning capabilities for the human head, which has a spherical-like geometry. This suggests that the proposed tri-plane in spherical coordinate system is potentially adaptable to more general subjects, whose appearance exhibits strong asymmetry with respect to different directions. Besides, the newly proposed view-image consistency loss compels the discriminator to ensure the alignment of images with their corresponding viewpoints, which is critical in 3D-aware synthesis. This paper demonstrates this emphasis particularly beneficial in the field of 3D full head synthesis, where imbalanced supervision is prominent. We anticipate its applicability to other 3D-aware GANs and broader 3D contexts, given that imbalanced supervision is a pervasive challenge across various 3D domains.

*Limitations and Future Work* While SphereHead successfully tackles the mirroring-face and multiple-face artifacts for 3D full head generation, it still shares some similar issues as previous 3D-aware generators, including flickering texture and lack of finer high-frequency visual and geometric details. We leave it as future work such as switching the backbone of the generator to StyleGAN3. In addition, SphereHead still suffers from data bias, especially for the back-view head. With our open-source toolbox, we're looking forward to a more diverse full-head dataset to facilitate the realm.

*Ethical considerations* SphereHead might be misused to manipulate the images of real people, which could pose a potential societal threat. We do not condone or endorse the use of our work with the malicious intent of spreading misinformation, nor do we support violating the rights of others in any form or manner.

## 7 Conclusion

We propose SphereHead, a novel 3D-aware GAN framework for high-quality, view-consistent full head images generation. Our approach leverages a unique tri-plane representation rooted in spherical coordinates, complemented by a view-image consistency loss. Collectively, these innovations enable SphereHead to significantly mitigate two prevalent types of semantic mismatch artifacts, specifically the mirroring-face and multiple-face artifacts, that are commonly observed in contemporary methods. Additionally, our method showcases superior qualitative and quantitative results against existing state-of-the-art 3D-aware GANs. We have also curated a comprehensive dataset, the WildHead Dataset, whose images are captured in the wild from various viewing angles spanning 360 degrees. We also provided an accompanying open-source data processing toolkit to support further research. We believe our SphereHead will prove beneficial to the research community in the realm of 3D full head synthesis.

## Acknowledgements

The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Guangdong Provincial Outstanding Youth Fund (No. 2023B1515020055), the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, by Key Area R&D Program of Guangdong Province (Grant No. 2018B030338001), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). It is also partly supported by NSFC-61931024, NSFC-62172348, and Shenzhen Science and Technology Program No. JCYJ20220530143604010.

## References

1. An, S., Xu, H., Shi, Y., Song, G., Ogras, U.Y., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360deg. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20950–20959 (2023)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017)
4. Canela, A., Caselles, P., Malik, I., Garces, G.T., Ramon, E., García, J., Sánchez-Riera, J., Moreno-Noguer, F.: Instantavatar: Efficient 3d head reconstruction via surface rendering. arXiv preprint arXiv:2308.04868 (2023)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
6. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5799–5809 (2021)
7. Deng, Y., Yang, J., Xiang, J., Tong, X.: Gram: Generative radiance manifolds for 3d-aware image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10673–10683 (2022)
8. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. Advances in neural information processing systems **28** (2015)
9. Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: 2017 International Conference on 3D Vision (3DV). pp. 402–411. IEEE (2017)
10. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8649–8658 (2021)

11. Gauthier, J.: Conditional generative adversarial nets for convolutional face generation. Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester **2014**(5), 2 (2014)
12. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021)
13. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20374–20384 (2022)
14. Jiang, K., Chen, S.Y., Liu, F.L., Fu, H., Gao, L.: Nerffaceediting: Disentangled face editing in neural radiance fields. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. arxiv e-prints. arXiv preprint arXiv:1812.04948 (2018)
16. Kim, T., Chung, C., Park, S., Gu, G., Nam, K., Choe, W., Lee, J., Choo, J.: K-hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1299–1303. IEEE (2021)
17. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
18. Liao, Y., Schwarz, K., Mescheder, L., Geiger, A.: Towards unsupervised learning of generative models for 3d controllable image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5871–5880 (2020)
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
20. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
21. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
22. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: Hologan: Unsupervised learning of 3d representations from natural images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7588–7597 (2019)
23. Nguyen-Phuoc, T.H., Richardt, C., Mai, L., Yang, Y., Mitra, N.: Blockgan: Learning 3d object-aware scene representations from unlabelled images. Advances in neural information processing systems **33**, 6767–6778 (2020)
24. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13503–13513 (2022)
25. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
26. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009)

27. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text-to-image synthesis. In: Proceedings of The 33rd International Conference on Machine Learning (2016)
28. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) **42**(1), 1–13 (2022)
29. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020)
30. Shi, Y., Aggarwal, D., Jain, A.K.: Lifting 2d stylegan for 3d-aware face generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6258–6266 (2021)
31. Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., Wang, J.: Fenerf: Face editing in neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7672–7682 (2022)
32. Szabó, A., Meishvili, G., Favaro, P.: Unsupervised generative 3d shape learning from natural images. arXiv preprint arXiv:1910.00287 (2019)
33. Wu, S., Yan, Y., Li, Y., Cheng, Y., Zhu, W., Gao, K., Li, X., Zhai, G.: Ganhead: Towards generative animatable neural head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 437–447 (2023)
34. Wu, Y., Zhang, J., Fu, H., Jin, X.: Lpff: A portrait dataset for face generators across large poses. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20327–20337 (October 2023)
35. Xue, Y., Li, Y., Singh, K.K., Lee, Y.J.: Giraffe hd: A high-resolution 3d-aware generative model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18440–18449 (2022)
36. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12803–12813 (2021)
37. Zanfir, M., Alldieck, T., Sminchisescu, C.: Phomoh: Implicit photorealistic 3d models of human heads. arXiv preprint arXiv:2212.07275 (2022)
38. Zhang, D., Zhong, C., Guo, Y., Hong, Y., Zhang, J.: Metahead: An engine to create realistic digital head. arXiv preprint arXiv:2304.00838 (2023)
39. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
40. Zhang, Z., Xia, S., Cai, Y., Yang, C., Zeng, S.: A soft-yolov4 for high-performance head detection and counting. Mathematics **9**(23), 3096 (2021)
41. Zheng, M., Yang, H., Huang, D., Chen, L.: Imface: A nonlinear 3d morphable face model with implicit neural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20343–20352 (2022)
42. Zhou, W., Yuan, L., Chen, S., Gao, L., Hu, S.: Lc-nerf: Local controllable face generation in neural randiance field. arXiv preprint arXiv:2302.09486 (2023)
43. Zhuang, Y., Zhu, H., Sun, X., Cao, X.: Mofanerf: Morphable facial neural radiance field. In: European Conference on Computer Vision. pp. 268–285. Springer (2022)