A 3D-Consistent Super-Resolution Framework for Efficient and Enhanced 3D-Aware Image Synthesis

Peng Zheng
Jilin University, China
Shanghai Innovation Institute, China
zhengpeng22@mails.jlu.edu.cn

Yi Chang
Jilin University, China
yichang@jlu.edu.cn

Yilin Wang
Adobe, USA
wangyilin930@gmail.com

Rui Ma*
Jilin University, China
Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China
ruim@jlu.edu.cn

Abstract

Neural volume rendering techniques, such as NeRF, have revolutionized 3D-aware image synthesis by enabling the generation of images of a single scene or object from various camera poses. However, the high computational cost of NeRF presents challenges for synthesizing high-resolution (HR) images. Most existing methods address this issue by leveraging 2D super-resolution, which compromises 3D-consistency. Other methods propose radiance manifolds or two-stage generation to achieve 3D-consistent HR synthesis, yet they are limited to specific synthesis tasks, reducing their universality. To tackle these challenges, we propose SuperNeRF-GAN, a universal framework for 3D-consistent super-resolution. A key highlight of SuperNeRF-GAN is its seamless integration with NeRF-based 3D-aware image synthesis methods, enabling simultaneously enhance the resolution of generated images while preserving 3D-consistency and reducing computational cost. Specifically, given a pre-trained generator capable of producing a NeRF representation, we first perform volume rendering to obtain a low-resolution image with corresponding depth and normal maps. Then, we employ a NeRF Super-Resolution module which learns a network to obtain a HR NeRF representation. Next, we propose a Depth-Guided Rendering process which contains three simple yet effective steps, including the construction of a boundary-correct multi-depth map, a normal-guided depth super-resolution and a depth-guided NeRF rendering. Experimental results demonstrate the superior efficiency, 3D-consistency and quality of our approach.

*Corresponding author

Method	3D-Consistency	High Efficiency	High-Resolution	High Universality	Image Quality	Geometry Quality
StyleNeRF [20]	✘	✔	✔	✔	✘	✘
StyleSDF [32]	✘	✔	✔	✔	✘	✘
GRAM-HD [43]	✔	✔	✔	✘	✘	✘
EG3D [11]	✘	✔	✘	✔	✔	✔
SH-HD [49]	✔	✘	✔	✘	✔	✔
Ours	✔	✔	✔	✔	✔	✔

Table 1: Comparison of 3D-aware image synthesis methods based on various criteria.

***Keywords:** Generative models, image synthesis, 3D-consistency, super-resolution*

1. Introduction

The introduction of neural volume rendering techniques, such as NeRF [31, 4, 5, 12], has significantly advanced 3D-aware image synthesis, enabling the generation of images from various camera poses for a single scene or object. These models learn NeRF representations, which can then be rendered into images at specified camera poses. However, the high computational cost inherent in NeRF limits their ability to synthesize high-resolution (HR) images. To address this, most existing methods [28, 30, 46, 15, 42] use a 2D super-resolution (SR) module, which often compromises 3D-consistency. While these inconsistencies might not be evident in static images, they become apparent in free-view videos, hindering applications in areas like video

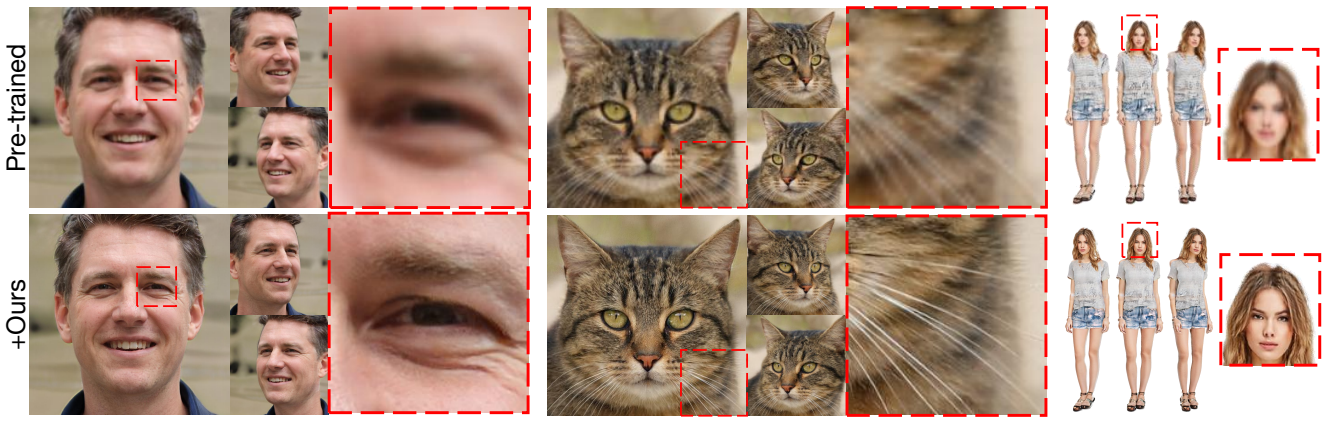

Figure 1: The effectiveness of our proposed SuperNeRF-GAN. The images in the first row are synthesized by a pre-trained EG3D [11] model without the use of 2D image super-resolution. The second row shows images super-resolved by SuperNeRF-GAN in a 3D-consistent manner. Please zoom in to see the detailed differences between the original and super-resolved images.

games and virtual reality.

Relatively few studies focus on 3D-consistent image synthesis. GRAM-HD [43] proposes using HR radiance manifolds to achieve 3D-consistent HR image synthesis. However, the inability of GRAM-HD to synthesize images from large viewpoints limits its capability for full-body and full-head image synthesis. SemanticHuman-HD (SHHD) [49] introduces a two-stage generation strategy for 3D-consistent HR image synthesis. However, it suffers from boundary depth issues due to its simplistic neighbor-aware depth aggregation, leading to failure in portrait synthesis. Other 3D-consistent methods struggle with HR image synthesis due to their high computational cost.

To address these limitations, we propose SuperNeRF-GAN, a universal 3D-consistent super-resolution framework that achieves HR image synthesis while maintaining 3D-consistency. This framework can be seamlessly integrated with NeRF-based 3D-aware image synthesis methods. Given a NeRF representation generated by a pre-trained model, we first perform volume rendering to obtain a low-resolution image with corresponding depth and normal maps. We then employ the NeRF Super-Resolution module, which learns a network to generate a HR NeRF representation. Following this, we introduce a novel Depth-Guided Rendering process that includes constructing a boundary-correct multi-depth map through depth aggregation and normal-guided depth super-resolution, and finally performing depth-guided NeRF rendering to synthesize HR images in a 3D-consistent way.

To evaluate the effectiveness of our method, we apply SuperNeRF-GAN to pre-trained models from state-of-the-art (SOTA) methods for portrait, cat face, and full-body image synthesis. The experimental results demonstrate significant improvements in 3D-consistency and efficiency compared to the original pre-trained models, as well as enhanced universality, 3D-consistency, and quality over other SOTA methods. Additionally, we conduct comparative experiments against other SOTA methods to further validate the superiority of our proposed approach.

In summary, the main contributions of this paper are:

- We propose SuperNeRF-GAN, a universal 3D-consistent super-resolution framework that enhances the resolution and 3D-consistency of synthesized images. SuperNeRF-GAN is designed to be universally applicable, making it easily deployable on NeRF-based 3D-aware image synthesis methods.
- SuperNeRF-GAN overcomes the limitations of SHHD and GRAM-HD, which are restricted to specific synthesis tasks. In contrast, SuperNeRF-GAN demonstrates high versatility, making it suitable for various synthesis tasks.
- Quantitative and qualitative results validate the superiority of our proposed method, particularly in terms of 3D-consistency and efficiency.

2. Related Work

2.1. 3D-Aware Image Synthesis

With the advent of generative adversarial networks (GANs) [19, 25, 26, 38] and diffusion models [22, 9, 33, 34], generative models have demonstrated impressive performance in image synthesis. Some works [40, 35, 2] achieve pose control by integrating parametric models such as SMPL [8] and 3DMM [18]. However, due to the lack of inherent 3D representations, these approaches do not achieve true 3D-aware image synthesis.

GRAF [36] first introduces NeRF [31] into generative models by learning a neural radiance field that can be rendered into an image at a given camera pose. However, using an MLP to model this field results in high computational costs, limiting the ability to synthesize high-quality images. Subsequent works like StyleSDF [32] and StyleNeRF [20] use volume rendering to obtain a low-resolution (LR) image and then upsample it to a high-resolution (HR) image. VolumeGAN [44] employs an explicit 3D feature volume to model the radiance fields, achieving high-fidelity image synthesis. However, increasing the resolution with this 3D volume representation results in a cubic growth in computational cost, making it inefficient for HR synthesis. EG3D [11] addresses this by proposing a tri-plane representation, reducing the cubic growth to a quadratic level. EG3D achieves SOTA performance in both image and geometry quality, and most current 3D-aware image synthesis methods [17, 45, 47, 27, 3, 39, 24] build upon it. However, these methods commonly employ a 2D super-resolution module, which compromises 3D-consistency while achieving HR image synthesis. As a result, these methods struggle to maintain 3D-consistency across different camera poses, limiting their application in areas such as virtual reality, where both efficiency and 3D-consistency are crucial.

2.2. 3D-Consistent HR Image Synthesis

GRAM-HD [43] generates HR radiance manifolds [16] instead of NeRF, thus avoiding the need for dense sampling and direct generation of 3D features. This approach ensures 3D-consistency by eliminating the need for image super-resolution. However, using radiance manifolds leads to suboptimal image and geometry quality. Rather than introducing a new representation like tri-plane representation or radiance manifolds, SemanticHuman-HD (SH-HD) [49] proposes a two-stage generation strategy that achieves efficient 3D-consistent synthesis without compromising image quality. In the first stage, an LR image with a corresponding depth map is rendered using dense sampling. In the second stage, the depth map is aggregated using neighboring points to produce a multi-channel depth map, which can be unprojected into 3D points instead of relying on dense sampling. Although this method achieves HR image synthesis with 3D-consistency and efficiency, it suffers from boundary depth issues due to its simplistic depth aggregation. This limitation restricts SH-HD to generating depth-smooth images, such as full-body images, and prevents it from generating images in scenarios with more complex depth variations, such as portraits.

Beyond NeRF- and radiance-manifold-based approaches, several generative models explicitly emphasize 3D consistency by enforcing geometric constraints or adopting alternative 3D primitives. EVA3D [23] incorporates explicit geometric supervision to improve multi-view consistency in 3D-aware image generation. Recent Gaussian-based generative models [1] represent scenes using collections of 3D Gaussian primitives, enabling more stable multi-view synthesis due to the explicit and continuous nature of the underlying geometry. Veri3D [13] further strengthens 3D consistency by jointly modeling appearance and geometry and verifying geometric alignment across viewpoints. Additional works [48, 37, 6] have explored similar directions. While these methods improve cross-view consistency, they often face challenges in scaling to high-resolution image synthesis or achieving efficient generation.

In conclusion, existing methods struggle to efficiently achieve 3D-consistent HR image synthesis for both portraits and full-body images. Our proposed SuperNeRF-GAN framework addresses these challenges by offering a universal solution that enhances both image and geometry quality while maintaining 3D-consistency.

3. Method

Fig. 2 illustrates the pipeline of our SuperNeRF-GAN framework, which is designed to seamlessly integrate with existing NeRF-based 3D-aware image synthesis methods. For demonstration, we use EG3D [11], a SOTA method in 3D-aware image synthesis, as a case study to introduce our SuperNeRF-GAN pipeline. Initially, our framework leverages the tri-plane representation produced by the pre-trained EG3D model to render a low-resolution image, along with associated depth and normal maps. Unlike EG3D, which performs image super-resolution in a 2D manner, our approach maintains 3D-consistency throughout the process. Specifically, we employ the NeRF Super-Resolution module and the Boundary-Correct Multi-Depth Map Construction technique to generate the high-resolution (HR) tri-plane representation and depth map. These components are then utilized in the Depth-Guided Rendering process, which efficiently synthesizes HR images while preserving 3D-consistency. To fully understand our methodology, we first introduce the essential preliminaries, including volume rendering [31] and EG3D [11].

3.1. Preliminary

3.1.1 Volume Rendering

NeRF [31] introduces volume rendering to synthesize images from given camera poses. For each pixel, a ray $\mathbf{r}(t)$ is cast from the camera position $\mathbf{o}$ along the direction $\mathbf{d}$: $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where t represents the distance from the camera position. The color $C(\mathbf{r})$ of this pixel is accumulated along the ray using volume rendering, which can be formulated as:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \cdot \sigma(\mathbf{r}(t)) \cdot \mathbf{c}(\mathbf{r}(t)) \cdot dt, \qquad (1)$$

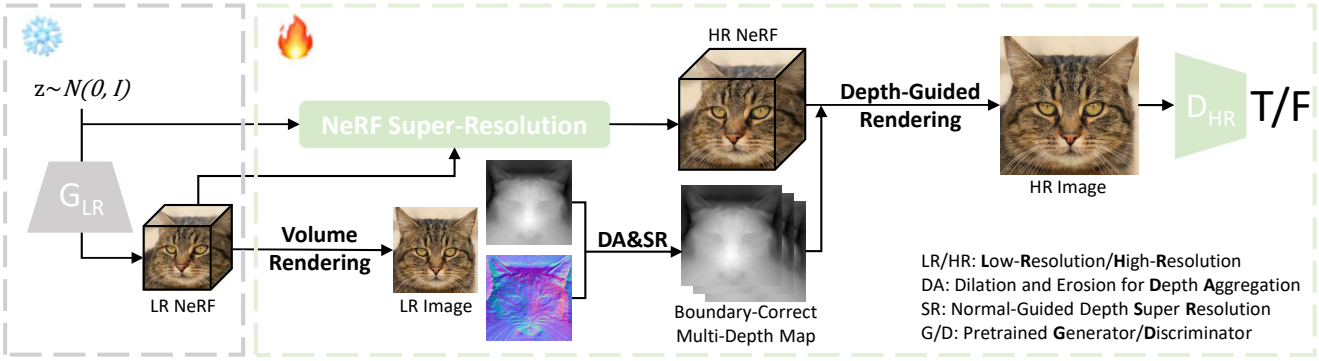

Figure 2: Pipeline of the proposed SuperNeRF-GAN framework. Given a random noise z, the pre-trained generator of existing 3D generative models maps it to a low-resolution (LR) NeRF representation. This LR representation can be rendered into a corresponding LR image along with depth and normal maps. Next, the LR NeRF representation undergoes the NeRF Super-Resolution module to produce a high-resolution (HR) NeRF representation. Simultaneously, Dilation and Erosion for Depth Aggregation and Normal-Guided Depth Super-Resolution are applied to the LR depth map to construct a boundary-correct multi-depth map. This map guides the rendering process of the HR NeRF representation, enabling efficient and 3D-consistent HR image synthesis.

$$\text{where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right). \tag{2}$$

Here, $\mathbf{c}(\mathbf{r}(t))$ and $\sigma(\mathbf{r}(t))$ denote the color and density of the 3D point $\mathbf{r}(t)$, respectively, while $T(t)$ represents the accumulated transmittance along the ray from t_n to t. Notably, by replacing the color $\mathbf{c}(\mathbf{r}(t))$ with the normal value $\mathbf{n}(\mathbf{r}(t))$ and distance t, this formula can also be used to render the normal value $N(\mathbf{r})$ and depth value $D(\mathbf{r})$ of the ray. In practice, this formula is discretized. For more details about volume rendering, please refer to NeRF.

3.1.2 EG3D as Pre-trained 3D Generator

The generator of EG3D is adapted from StyleGAN2 [26], which achieves SOTA performance in image synthesis. Given a random noise $\mathbf{z}$ sampled from a Gaussian distribution, the generator maps it into a feature map with dimensions $256 \times 256 \times 96$. This feature map is then reshaped into a tri-plane representation $\mathbf{T}_{LR}$, where the three planes correspond to the XY, YZ, and ZX orientations. Each plane consists of 32 channels with a resolution of 256×256. For a 3D point $\mathbf{X}$, we project it onto these three planes and perform bilinear interpolation to extract its features from each plane. These features are then fed into an MLP, which outputs the color $\mathbf{c}(\mathbf{X})$ and density $\sigma(\mathbf{X})$. These values are subsequently used in volume rendering, as described in Eq. 1.

3.2. 3D-Consistent Low-Resolution Image Synthesis

Using random noise $\mathbf{z}$ as input, the pre-trained EG3D model generates a low-resolution tri-plane representation $\mathbf{T}_{LR}$. This representation is rendered from a given camera pose to produce a low-resolution image $\mathbf{I}_{LR}$, along with the corresponding depth map $\mathbf{D}_{LR}$ and normal map $\mathbf{N}_{LR}$, each at a resolution of 256^2. In this process of low-resolution image synthesis, we employ dense sampling as used in EG3D. Specifically, for each pixel, 36 points are sampled along the ray using uniform sampling and an additional 36 points using importance sampling. Although dense sampling is computationally intensive, its overall computational cost remains manageable due to the low resolution of the output.

3.3. 3D-Consistent Super-Resolution

3.3.1 NeRF Super-Resolution

To obtain an HR 3D representation $\mathbf{T}_{HR}$, we utilize the NeRF Super-Resolution module, which is based on the architecture of StyleGAN2. This module takes the LR tri-plane representation $\mathbf{T}_{LR}$ and the corresponding noise $\mathbf{z}$ as input, and outputs an HR representation $\mathbf{T}_{HR}$ at a resolution of 1024^2. Importantly, the module uses the same noise $\mathbf{z}$ as the pre-trained 3D generator to ensure consistency in the distribution between LR and HR images.

3.3.2 Depth-Guided Rendering

We construct an HR multi-depth map $\mathbf{D}_{HR}$ with three channels using the Boundary-Correct Multi-Depth Map Construction, as detailed in Section 3.4. This map is used to guide the sampling process, thereby avoiding the dense sampling employed by EG3D. Our Depth-Guided Rendering approach is adapted from Eq. 1 and formulated as follows:

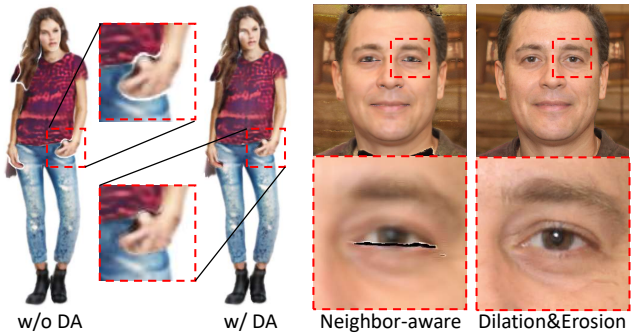

Figure 3: The left two figures demonstrate the effectiveness of Depth Aggregation (DA), note that the results are synthesized by untrained SuperNeRF-GAN models for better demonstration. The right two compare different DA techniques, highlighting that Neighbor-aware DA in SH-HD introduces noticeable artifacts, especially at depth discontinuities.

$$C(\mathbf{r}) = \sum_{i=1}^{3} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \mathbf{c}_i, \tag{3}$$

$$\text{where } T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right). \tag{4}$$

Here, $\delta_i = \mathbf{D}_{i+1} - \mathbf{D}_i$ represents the distance between adjacent depth values, where $\mathbf{D}_i$ is the i_{th} value of $\mathbf{D}_{HR}$. σ_i and $\mathbf{c}_i$ are the density and color interpolated from the HR 3D representation, with the interpolation coordinates projected from the depth map. Notably, as shown in Eq. 3, the number of sampling points is reduced from 64 (as explained in Section 3.2) to 3, achieving efficient rendering at a 1024^2 resolution.

3.4. Boundary-Correct Multi-Depth Map Construction

As introduced in Section 3.3.2, we need a multi-depth map to guide the rendering. This section details how to construct this map. A straightforward approach is to perform bilinear interpolation on the low-resolution (LR) depth map $\mathbf{D}_{LR}$. However, this often leads to incorrect depth values at the boundaries due to depth discontinuities. Specifically, consider two adjacent pixels on the boundary, which have non-continuous depth values. Direct interpolation will result in an averaged value that does not align with either of the two pixels' actual depth values.

To address the boundary depth issue, SH-HD [49] proposes aggregating depth values of neighboring pixels before performing bilinear interpolation on the depth map. While this method is effective for full-body image synthesis, it struggles with portrait synthesis. This limitation arises

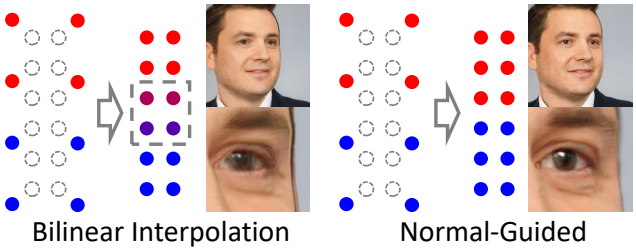

Figure 4: Effectiveness of Normal-Guided Depth Super-Resolution. The dashed rectangle highlights inaccuracies at depth discontinuity using bilinear interpolation, which result in artifacts as in the synthesized image.

because neighbor-aware depth aggregation does not fundamentally solve the boundary issue but merely alleviates it. Therefore, SH-HD performs well in depth-smooth scenarios, such as full-body image synthesis, but fails in other cases. To obtain a boundary-correct high-resolution depth map, our solution involves Dilation and Erosion for Depth Aggregation and Normal-Guided Depth Super-Resolution, which are introduced below.

3.4.1 Dilation and Erosion for Depth Aggregation

Different from SH-HD, we employ erosion and dilation operations to aggregate the depth map. Erosion is a morphological operation that shrinks the boundaries of objects in an image. For a given depth map $\mathbf{D}$ and a structuring element B, the erosion operation is formulated as:

$$(\mathbf{D} \ominus B)(x, y) = \min_{(s,t) \in B} \{\mathbf{D}(x+s, y+t) - B(s, t)\}, \tag{5}$$

where (x, y) are the coordinates of a pixel in the depth map $\mathbf{D}$, and (s, t) are the coordinates within the structuring element B. This operation slides the structuring element B over the image $\mathbf{D}$ and replaces each pixel by the minimum value of the image pixels covered by B minus the corresponding value of B. Conversely, dilation expands the boundaries of objects in an image. The dilation operation is formulated as:

$$(\mathbf{D} \oplus B)(x, y) = \max_{(s,t) \in B} \{\mathbf{D}(x+s, y+t) + B(s, t)\}. \tag{6}$$

The choice of the structuring element B significantly affects the results of erosion and dilation. In our case, we use a square structuring element. By applying erosion and dilation to the LR depth map $\mathbf{D}_{LR}$, we obtain the dilated depth map $\mathbf{D}_{dil}$ and eroded depth map $\mathbf{D}_{ero}$, which are then concatenated with $\mathbf{D}_{LR}$, resulting in an aggregated map $\mathbf{D}_{agg}$ with three channels. This aggregated map addresses the boundary depth issue by storing depth values from both sides of the boundary, as shown in Fig. 3 (right).

3.4.2 Normal-Guided Depth Super-Resolution

Naive bilinear interpolation on the aggregated depth map results in multi-depth values at boundaries, which include values from both sides of the boundary. However, these depth values may not be accurate as the bilinear interpolation relies solely on the information within the aggregated depth map. To address this, we propose a Normal-Guided Depth Super-Resolution module, which leverages the normal map to provide supplementary geometric information. Specifically, we assume that the depth surface is locally semi-smooth and can be approximated by a piecewise planar model within a small neighborhood, such that surface normals provide reliable first-order cues for local depth propagation.

Given a normal map $\mathbf{N} = (N_x, N_y, N_z)$, the differences in depth values can be computed as: $\Delta_x = N_x/N_z$ and $\Delta_y = N_y/N_z$. This formulation follows the standard geometric relationship between surface normals and local depth gradients under perspective projection, and is used here as a directional prior rather than an exact metric reconstruction. The super-resolved depth map $\mathbf{D}_{SR}$ at 512^2 resolution can be expressed as the original depth map plus the differences with respect to x and y, formulated as:

$$\mathbf{D}_{SR}(x, y) = \mathbf{D}_{LR}(m, n) + \Delta(x, y), \tag{7}$$

$$\Delta(x, y) = \frac{w_x(x, y) \cdot \Delta_x(m, n) + w_y(x, y) \cdot \Delta_y(m, n)}{\sqrt{(\Delta_x(m, n))^2 + (\Delta_y(m, n))^2}}, \tag{8}$$

where $w_x(x, y)$ and $w_y(x, y)$ are weight functions dependent on x and y. The normalization term controls the magnitude of the propagated depth offset, while the relative contributions from the x and y directions are determined by their geometric reliability.

The weight function is defined as:

$$w_x(x, y) = \frac{e^{-|\Delta_x(m,n)|}}{e^{-|\Delta_x(m,n)|} + e^{-|\Delta_y(m,n)|}} \cdot (-1)^{\mathbb{I}(x=2m-1)}, \tag{9}$$

and $w_y(x, y)$ is defined analogously by swapping Δ_x and Δ_y and replacing the indicator with $\mathbb{I}(y = 2n - 1)$, where $\mathbb{I}$ is an indicator function while m and n are integers related to the coordinates x and y by:

$$m = \left\lfloor \frac{x+1}{2} \right\rfloor, \quad n = \left\lfloor \frac{y+1}{2} \right\rfloor. \tag{10}$$

For a $2\times$ upsampling, each LR pixel (m, n) corresponds to a 2×2 HR sub-grid; the indicator $\mathbb{I}(x = 2m - 1)$ encodes the relative horizontal sub-pixel position, yielding a negative sign for the left sub-pixel and a positive sign for the right sub-pixel, such that the normal-implied depth gradient is applied with the correct direction during depth propagation.

The remaining part of the weight function is a softmax function inspired by BiNi [10], which assumes that the depth map is semi-smooth. This means that the depth map is one-sided differentiable, with larger weight assigned to the direction with smaller differences. Intuitively, directions with large $|\Delta|$ are more likely to cross depth discontinuities and yield unreliable extrapolation; the softmax weighting therefore acts as a robust gating mechanism that favors the relatively smoother direction for depth propagation. We emphasize that normal maps do not explicitly encode depth discontinuities; instead, they provide local surface orientation cues. In our formulation, normals are used indirectly as a reliability prior: directions with large normal-implied depth variations are more likely to correspond to cross-boundary extrapolation and are therefore down-weighted, while relatively stable directions are favored.

Consequently, the aggregated depth map $\mathbf{D}_{agg}$ is processed through the Normal-Guided Depth Super-Resolution module twice, resulting in a boundary-correct multi-depth map $\mathbf{D}_{HR}$ at 1024^2 resolution. Notably, to facilitate Depth-Guided Rendering, the three-channel depth values of this multi-depth map are sorted. A visualization of this module is provided in Fig. 4.

3.5. Training Pipeline

3.5.1 Training Strategy

Our method can be directly applied to pre-trained 3D generative models, such as EG3D and SH-HD. During training, we freeze the pre-trained generator and focus on training the NeRF Super-Resolution module. Notably, the discriminator has been redesigned to process high-resolution images and is jointly optimized during training.

3.5.2 Loss Function

We use the loss functions from the original method to ensure training consistency. Specifically, the adversarial loss $\mathcal{L}_{GAN}$ from the original 3D generative model is retained to supervise high-resolution image synthesis. Additionally, we adopt the unsample loss from SH-HD to guide the training of the NeRF Super-Resolution module by penalizing inconsistency between high-resolution and low-resolution images. Formally, the overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{unsample}, \tag{11}$$

where λ is a weighting factor. The unsample loss is defined as

$$\mathcal{L}_{unsample} = \left\| \mathcal{D}(\mathbf{I}_{HR}) - \mathbf{I}_{LR} \right\|_1, \tag{12}$$

with $\mathbf{I}_{LR}$ denoting the low-resolution image rendered from the original NeRF representation, $\mathbf{I}_{HR}$ the high-resolution image rendered from the super-resolved NeRF representation, and $\mathcal{D}(\cdot)$ a differentiable downsampling operator.

4. Implementation Details

4.1. Training Setup

Our experiments were conducted on a server equipped with 4 NVIDIA A40 GPUs, each with 48GB of memory. The models were trained for 2 to 4 days, depending on the complexity and size of the dataset. Specifically, for portrait and cat face image synthesis tasks, we used a batch size of 16, while a batch size of 4 was employed for full-body image synthesis due to the increased computational demands.

4.2. SH-HD for Portrait and Cat Image Synthesis

To facilitate portrait and cat face image synthesis, we integrate key components from SH-HD [49] into pre-trained EG3D models. These components include the two-stage generation strategy, feature super-resolution module, unsample loss, and neighborhood-aware depth aggregation technique. Despite these adaptations, SH-HD exhibits limitations in portrait and cat face image synthesis, resulting in a drop in image quality for these specific tasks, as demonstrated in Fig. 3. The results underscore the superior universality of our proposed method.

5. Experiments

5.1. Experimental Setting

5.1.1 Datasets

We train our models on three distinct datasets for specific synthesis tasks: FFHQ [25], which includes 50K portrait images at a resolution of 1024^2, is used for portrait synthesis. For cat face synthesis, we use the AFHQ [14], which contains 5.5K cat face images at a resolution of 1024^2. For full-body image synthesis, we employ DeepFashion [29], which provides 7K full-body images at a resolution of 1024^2, along with corresponding segmentation maps. Additionally, due to the requirement in SH-HD [49], we also incorporate normal maps from SH-HD and human poses from AG3D [17].

5.1.2 Baselines

For portrait and cat face image synthesis, our model leverages pre-trained EG3D [11] models. We compare our results with GRAM-HD [43], StyleNeRF [20], and StyleSDF [32], all recognized for their high-resolution image synthesis capabilities. However, StyleNeRF and StyleSDF rely on 2D super-resolution, compromising the 3D-consistency of the synthesized images. In contrast, GRAM-HD's manifold representation limits image quality, geometry accuracy, and universality. For full-body image synthesis, our model builds on the pre-trained SH-HD model. We compare it with EVA3D [23], VeRi3D [13], and GSM [1], which are noted for their 3D-consistent image synthesis. Nevertheless, their reliance on vertex-based, MLP-based, or Gaussian shell-based representations imposes limitations on the quality of the generated images.

Dataset	Method	FID↓	KID↓	PSNR↑	SSIM↑	Res
FFHQ	EG3D [11]	**4.65**	**1.27**	33.67	0.893	512
	+Ours	5.13	1.70	34.36	0.920	512
		5.10	1.54	**36.44**	**0.935**	1024
AFHQ	EG3D [11]	**3.19**	**0.38**	32.61	0.843	512
	+Ours	3.77	1.09	**33.01**	**0.861**	512

Table 2: Effectiveness of our proposed SuperNeRF-GAN for portrait (FFHQ) and cat-face (AFHQ) image synthesis. Note that all KID scores in this paper are multiplied by 1000.

Method	DeepFashion1024			FFHQ1024			AFHQ512			Time
	FID↓	KID↓	Mem	FID↓	KID↓	Mem	FID↓	KID↓	Mem	1024^2
SH-HD	8.70	4.04	31G	31.9	27.8	25G	8.77	11.7	12G	0.14s
Ours	**8.47**	**3.68**	**14G**	**5.10**	**1.54**	**11G**	**3.77**	**1.09**	**8G**	**0.11s**

Table 3: Quantitative comparison with SH-HD [49]. "Mem" indicates the GPU memory consumption during training, while "Time" reports inference latency per image. Since SH-HD is not directly applicable to the FFHQ and AFHQ datasets, we modified it to enable training on these datasets.

5.1.3 Evaluation Metrics

We use Frechet Inception Distance (FID) [21] and Kernel Inception Distance (KID) [7] to assess the quality of synthesized images. Note that all KID scores are multiplied by 1000. To evaluate 3D-consistency across images synthesized from different camera poses, we adopt Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) metrics, as used in GRAM-HD. Higher scores indicate better 3D-consistency. Specifically, we reconstruct a NeuS [41] model from images synthesized from various camera angles and calculate the PSNR and SSIM scores between the synthesized and reconstructed images. These scores are averaged across 50 entities. For each entity, we generate images from 30 uniformly sampled yaw angles ranging from -0.4 to 0.4 radians. These images are used to construct a NeuS [41] representation, utilizing NeuS's default settings to ensure consistency and comparability in our experiments. The constructed NeuS representation is subsequently rendered into reconstructed images from the specified angles to facilitate evaluation.

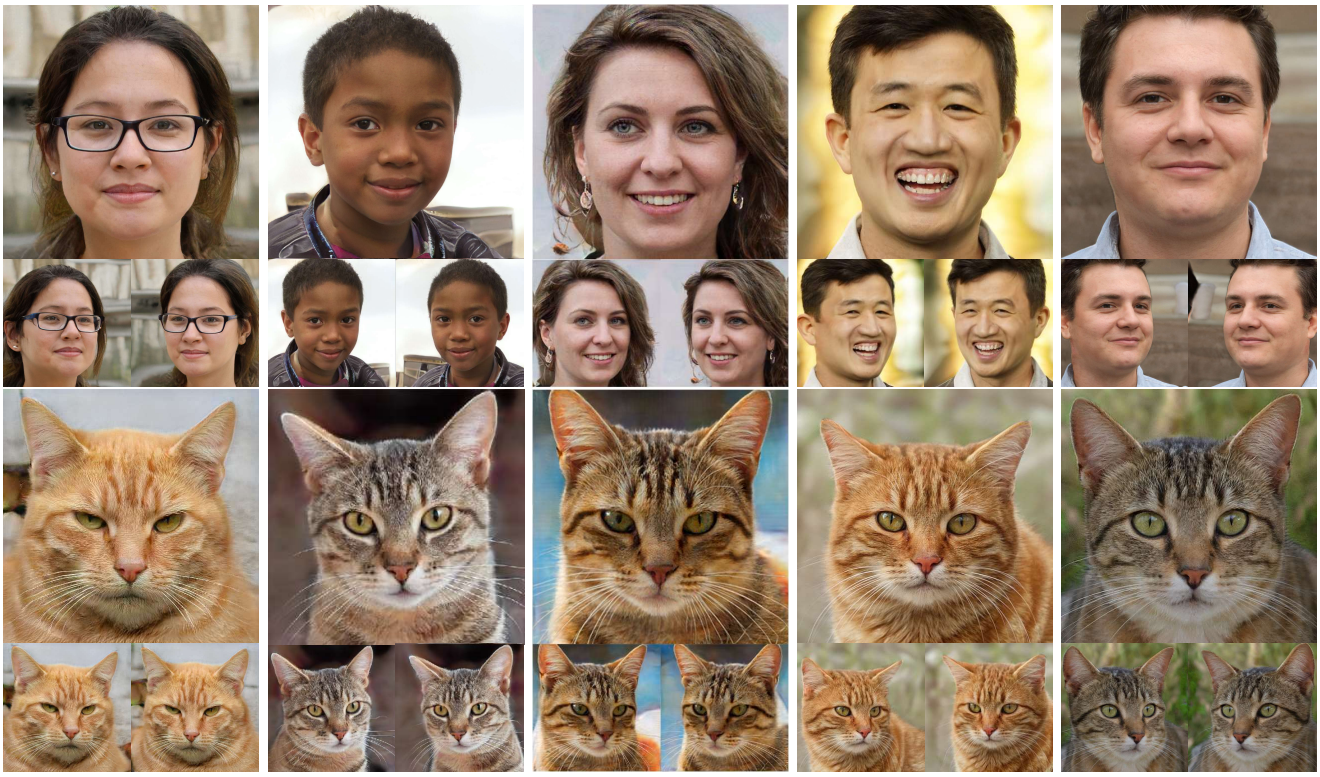

Figure 5: Qualitative comparison among 3D-aware image synthesis methods. The results of other methods are taken from their respective papers to ensure a fair and consistent comparison. Since the 3D-inconsistency might not be evident in static images, we provide additional comparisons of 3D-consistency in our **Supplementary Video**, where StyleNeRF [20], StyleSDF [32], and EG3D [11] show noticeable inconsistencies.

5.2. Comparisons

We evaluate 3D-aware image synthesis methods across various criteria, with the results presented in Table 1. Note that only our method achieves 3D-consistent HR image synthesis with high universality. Specifically, GRAM-HD is unable to synthesize full-body images due to inherent limitations in its manifolds, and SH-HD struggles with generating high-quality portraits because of boundary issues, which restrict their universality.

As our models are deployed on the pre-trained 3D generators of EG3D and SH-HD, we compare our approach with these methods, as illustrated in Table 2 and 3. Compared to EG3D, our method shows stable improvements in 3D-consistency, albeit with a slight compromise in image quality. We argue that the 2D super-resolution used in EG3D significantly improves the image quality, whereas SuperNeRF-GAN leverages Depth-Guided Rendering to ensure 3D-consistency. It is worth noting that the quantitative improvements in 3D-consistency may appear subtle, as they are constrained by the limitations of the reconstruction method NeuS. We encourage readers to refer to the **Supplementary Video** for more qualitative comparisons, where the advantages of our method are clearly evident. As for SH-HD, which also ensures 3D-consistency, our method outperforms it in both efficiency and image quality. Notably, SH-HD performs poorly on the FFHQ and AFHQ datasets due to its naive depth aggregation technique.

Method	FFHQ1024		AFHQ512	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
*StyleNeRF [20]	30.0	0.804	-	-
*StyleSDF [32]	31.1	0.836	26.6	0.749
*GRAM-HD [43]	33.8	0.872	28.8	0.807
Ours	**36.4**	**0.935**	**33.0**	**0.861**

Table 4: Quantitative comparison of 3D-consistency. *Results taken from GRAM-HD.

Method	FFHQ1024		AFHQ512	
	FID↓	KID↓	FID↓	KID↓
*StyleNeRF [20]	9.45	2.65	-	-
*StyleSDF [32]	9.44	2.83	7.91	3.90
*GRAM-HD [43]	12.0	5.23	7.67	3.41
Ours	**5.10**	**1.54**	**3.77**	**1.09**

Table 5: Quantitative comparison of image quality among methods that achieve high-resolution (1024×1024) image synthesis. *Results are taken from GRAM-HD.

Method	FID↓	KID↓	Resolution	Memory
EVA3D [23]	15.89	9.25	512	33G
GSM [1]	15.78*	-	512	-
VeRi3D [13]	21.4*	-	512	34G
SH-HD [49]	8.70	4.04	1024	31G
Ours	10.56	5.60	512	11G
	8.47	**3.68**	1024	14G

Table 6: Quantitative comparison among methods that achieve full-body image synthesis with 3D-consistency. *Results taken from their papers.

5.2.1 3D-Consistency

The comparison of 3D-consistency is shown in Table 4. StyleNeRF and StyleSDF exhibit poorer 3D-consistency due to their 2D super-resolution module. While GRAM-HD performs better than the aforementioned methods, its radiance manifolds constrain image quality. Consequently, our method consistently delivers superior performance in both 3D-consistency and image quality. We also present additional qualitative comparisons among various 3D-aware image synthesis methods, as illustrated in Fig. 5. The results from other methods were directly taken from their respective publications to ensure a fair and consistent comparison. Since 3D-inconsistency might not be evident in static images, we provide additional comparisons of 3D-consistency in our **Supplementary Video**, where StyleNeRF, StyleSDF, and EG3D show noticeable inconsistencies. In contrast, both GRAM-HD and our method achieve 3D-consistent synthesis. However, GRAM-HD introduces artifacts when viewed from large angles, where our method performs well.

5.2.2 Image Quality

In terms of image quality, our method shows significant improvements, as shown in Table 5. Notably, although GRAM-HD employs generative radiance manifolds to achieve 3D-consistent image synthesis, it compromises on image quality compared to super-resolution-based methods such as StyleNeRF and StyleSDF. This trade-off highlights that achieving strong 3D-consistency imposes stronger constraints on generators, often leading to a degradation in image quality. Consequently, we argue that the slight compromise in image quality in our method, as compared to EG3D (Table 2), is acceptable. On the other hand, while our method prioritizes 3D-consistency, it still achieves better image quality than StyleNeRF and StyleSDF, and is comparable to EG3D. Furthermore, we evaluated the image quality of 3D-consistent methods on full-body image synthesis task, with results presented in Table 6. Our method not only achieves the highest image quality but also supports high-resolution image synthesis with low memory consumption, thanks to our efficient 3D-consistent super-resolution module.

5.2.3 Efficiency

Our method significantly enhances efficiency in high-resolution image synthesis due to its Depth-Guided Rendering approach, which reduces the number of sampling points. To assess the efficiency of our method, we perform comparative experiments with other 3D-consistent image synthesis methods. As shown in Table 6, our method offers substantial improvements in efficiency, achieving 1024^2 resolution image synthesis with only 14G of GPU memory. In addition to memory consumption, we further report the inference latency for high-resolution image synthesis. As shown in Table 3, our method achieves faster generation at the 1024^2 resolution compared to SH-HD, demonstrating improved efficiency in terms of both memory usage and runtime.

5.2.4 Universality

As noted in the GRAM-HD paper, this method cannot synthesize images from wide viewpoints due to its reliance on the near-plane manifold representation for efficient rendering. In contrast, our method leverages depth-guided rendering, which is compatible with any NeRF representation and enables the synthesis of images from wide viewpoints, such as full-body images. Another 3D-consistent method, SH-HD, struggles with unsmooth depth values, as shown in Table 3, due to its naive neighbor-aware depth aggregation technique. This issue is further evidenced in Fig. 3. Consequently, our method stands as the only 3D-consistent image synthesis approach with robust universality, as demonstrated in Table 1.

5.2.5 Geometry Quality

We conduct a qualitative comparison of geometry quality, as shown in Fig. 6. GRAM-HD employs radiance manifolds to achieve efficient rendering, with each manifold being nearly

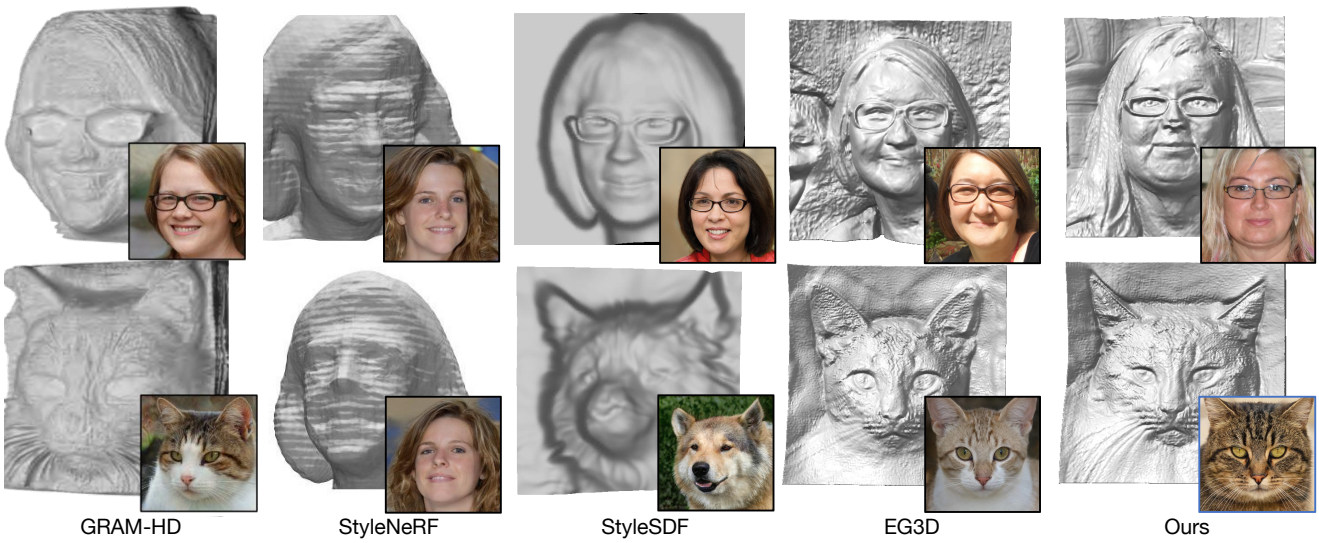

Figure 6: Qualitative comparison of geometry quality. EG3D [11] and our method generate results with more detailed geometry than other methods. The results for GRAM-HD [43], StyleNeRF [20], and StyleSDF [32] are directly taken from their respective papers.

planar. As a result, the generated geometries are essentially the combination of several "planes," which limits the overall quality. StyleSDF and StyleNeRF generate NeRF representations at relatively low resolutions (e.g., 32), which further constrains their ability to capture fine geometric details. In contrast, our method produces high-resolution NeRF representations, achieving the same geometry quality as EG3D.

5.3. Ablation Study

5.3.1 Depth Aggregation

The results in Fig. 3 confirm the effectiveness of our depth aggregation technique, as discussed in Section 3.4. Specifically, without depth aggregation, noticeable holes appear in the generated images due to the inaccuracies caused by straightforward bilinear interpolation. In contrast, the naive neighbor-aware depth aggregation proposed in SH-HD suffers from artifacts at depth discontinuities. This approach only partially mitigates the boundary issue without fundamentally resolving it. This claim is further supported quantitatively in Table 3, where our method demonstrates greater universality compared to SH-HD, primarily due to the differences in depth aggregation techniques.

To further investigate depth aggregation, we conduct a quantitative ablation study, as presented in Table 7. Our default design constructs a 3-channel aggregated depth map, which aims to capture the dominant depth hypotheses around boundaries (e.g., near-side vs. far-side depth candidates) that are most relevant for subsequent depth-guided rendering. In this study, we apply additional dilation and erosion operations to produce a 5-channel aggregated depth map. The results show that doubling the depth aggregation does not improve performance and instead introduces depth inaccuracies. We attribute this to the fact that additional channels derived from the same local neighborhood tend to be redundant or noisy, and may even amplify errors at depth discontinuities, thus weakening the effectiveness of depth guidance. While our Normal-Guided Depth Super-Resolution mitigates these inaccuracies, the additional depth values fail to yield benefits. Therefore, using three depth channels provides a favorable trade-off between representing the essential boundary depth candidates and maintaining robustness, which is consistent with the saturation behavior observed in Table 7.

Method	DeepFashion		FFHQ1024		AFHQ512	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
$(D\&E)^2$	9.10	4.71	6.47	2.95	4.50	1.27
D&E	**8.47**	**3.68**	**5.10**	**1.54**	**3.77**	**1.09**

Table 7: Ablation study on depth aggregation methods. "D&E" denotes the dilation and erosion operations used in our method, while $(D\&E)^2$ signifies performing dilation and erosion twice to yield 5 depth values per pixel.

5.3.2 Normal-Guided Depth Super-Resolution

To evaluate the effectiveness of our proposed Normal-Guided Depth Super-Resolution module, we perform a quantitative comparison against bilinear interpolation. As

Method	FID↓	KID↓	PSNR↑	SSIM↑
Bilinear Interpolation	**5.08**	1.51	35.75	0.915
Normal-Guided	**5.08**	**1.48**	**36.44**	**0.935**

Table 8: Ablation study on interpolation methods. Our proposed Normal-Guided Super-Resolution achieves improvements on 3D-consistency. The models are trained on the FFHQ1024 dataset, as high-resolution better highlights these improvements.

Figure 7: 3D-consistent image synthesis. The images synthesized from different camera poses exhibit consistent 3D structures.

shown in Table 8, our method achieves notable improvements in 3D-consistency. These improvements stem from the module's ability to accurately super-resolve depth maps with guidance from the normal map. As illustrated in Fig. 4, compared to naive bilinear interpolation, our normal-guided approach leverages the normal map to identify and handle depth discontinuities at boundaries more effectively. This capability ensures smoother and more accurate depth transitions. For a detailed explanation, please refer to Section 3.4.2.

5.4. Applications

5.4.1 3D-Consistent Image Synthesis

In Fig. 7, we showcase images synthesized from different camera poses. The results clearly demonstrate the high level of 3D-consistency achieved by our method. For an even more comprehensive demonstration of 3D-consistent image synthesis, please refer to our **Supplementary Video**, which provides dynamic visualizations that further highlight the 3D-consistency of our approach.

5.4.2 High-Resolution Image Synthesis

Fig. 8 showcases high-resolution portraits synthesized by our method. To balance image quality and diversity, we use a truncation technique where the latent code w is defined as a weighted average of two components: $w = 0.5 \times w_{\text{averaged}} + 0.5 \times w_{\text{random}}$. Here, w_{averaged} is the averaged latent code, ensuring high-quality synthesis, while w_{random} introduces diversity. For each generated entity, we provide images from different viewpoints, demonstrating the 3D-consistency of high-resolution image synthesis—a challenge for previous methods.

5.4.3 Interpolation in Latent Space

Fig. 9 demonstrates the results of interpolation in latent space. Specifically, given two random latent codes, we interpolate between them and feed the interpolated codes into the generator. This process produces smooth transitions in the synthesized images. Notably, each interpolated result exhibits 3D-consistency, demonstrating that images synthesized across the entire latent space can maintain 3D-consistency, thanks to our direct rendering strategy.

6. Limitations

Our method achieves stronger 3D-consistency than previous approaches by directly rendering high-resolution images. However, this also results in a slight degradation in image quality compared to methods based on 2D super-resolution. Exploring approaches that balance both strong 3D-consistency and high image quality would be valuable for broader industrial applications. Additionally, while our method enables efficient synthesis of high-resolution images through depth-guided rendering, it still requires dense sampling during the low-resolution stage, which can hinder real-time rendering. Replacing NeRF with 3DGS presents a promising alternative, though this direction remains relatively underexplored. We provide several representative failure cases in Fig. 10. Two typical types of failures are observed: (a) due to the stochastic nature of GAN-based generation, different random samples may occasionally lead to suboptimal synthesis results; and (b) under large viewpoint changes, the generation quality may degrade due to the limited viewpoint coverage in the training data. We emphasize that both failure modes reflect common limitations of current 3D-aware GAN frameworks rather than issues specific to our method.

Figure 8: High-resolution (1024×1024) portrait synthesis. Our method produces high-quality portraits with rich details.

7. Conclusion

In this paper, we present SuperNeRF-GAN, a universal framework for 3D-consistent super-resolution. SuperNeRF-GAN can be seamlessly integrated with existing 3D-aware image synthesis methods to enhance the resolution of synthesized images while maintaining 3D-consistency and efficiency. The core innovation of SuperNeRF-GAN lies in generating a boundary-correct multi-depth map, which is employed in depth-guided rendering to achieve high-resolution image synthesis with enhanced 3D-consistency and efficiency. Compared to existing 3D-consistent methods, our approach consistently demonstrates improvements in both universality and quality. Meanwhile, it is important to note that while our super-resolution framework ensures 3D-consistency, it slightly compromises image quality compared to methods employing 2D image super-resolution. How to achieve 3D-consistent high-resolution image synthesis while ensuring the high image quality similar to the 2D-based methods remains an important direction for future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62572212), Science and Technology Development Plan of Jilin Province (No. 20260203049SF) and the Fundamental Research Funds for the Central Universities.

Figure 9: Interpolation in the latent space produces smooth transitions, with each intermediate result maintaining 3D-consistency.

(a) Sampling-related failure

(b) View extrapolation failure

Figure 10: Failure cases. (a) Occasional degradation caused by the stochastic nature of GAN-based generation, where random sampling may lead to suboptimal visual quality. (b) Degraded results under extreme viewpoint extrapolation due to limited viewpoint coverage in the training data. Both cases reflect common challenges in 3D-aware generative modeling.

References

[1] R. Abdal, W. Yifan, Z. Shi, Y. Xu, R. Po, Z. Kuang, Q. Chen, D.-Y. Yeung, and G. Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024. 3, 7, 9

[2] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 2

[3] S. An, H. Xu, Y. Shi, G. Song, U. Y. Ogras, and L. Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20950–20959, 2023. 3

[4] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 1

[5] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 1

[6] F. Barthel, W. Morgenstern, P. Hinzer, A. Hilsmann, and P. Eisert. Cgs-gan: 3d consistent gaussian splatting gans for high resolution human head synthesis. *arXiv preprint arXiv:2505.17590*, 2025. 3

[7] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceed-*

ings, Part V 14, pages 561–578. Springer, 2016. 2

[9] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[10] X. Cao, H. Santo, B. Shi, F. Okura, and Y. Matsushita. Bilateral normal integration. In *European Conference on Computer Vision*, pages 552–567. Springer, 2022. 6

[11] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 1, 2, 3, 7, 8, 10

[12] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 1

[13] X. Chen, J. Huang, Y. Bin, L. Yu, and Y. Liao. Veri3d: Generative vertex-based radiance fields for 3d controllable human image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8986–8997, 2023. 3, 7, 9

[14] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 7

[15] K. Deng, G. Yang, D. Ramanan, and J.-Y. Zhu. 3d-aware conditional image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4434–4445, 2023. 1

[16] Y. Deng, J. Yang, J. Xiang, and X. Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10673–10683, 2022. 3

[17] Z. Dong, X. Chen, J. Yang, M. J. Black, O. Hilliges, and A. Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14916–14927, 2023. 3, 7

[18] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 2

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[20] J. Gu, L. Liu, P. Wang, and C. Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1, 3, 7, 8, 9, 10

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] F. Hong, Z. Chen, Y. Lan, L. Pan, and Z. Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 3, 7, 9

[24] K. Jiang, S.-Y. Chen, F.-L. Liu, H. Fu, and L. Gao. Nerffaceediting: Disentangled face editing in neural radiance fields. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 7

[26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 4

[27] B. Lei, K. Yu, M. Feng, M. Cui, and X. Xie. Diffusiongan3d: Boosting text-guided 3d generation and domain adaptation by combining 3d gans and diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10487–10497, 2024. 3

[28] R. Liu, P. Zheng, Y. Wang, and R. Ma. 3d-ssgan: Lifting 2d semantics for 3d-aware compositional portrait synthesis. *arXiv preprint arXiv:2401.03764*, 2024. 1

[29] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 7

[30] T. Ma, B. Li, Q. He, J. Dong, and T. Tan. Semantic 3d-aware portrait synthesis and manipulation based on compositional neural radiance field. *arXiv preprint arXiv:2302.01579*, 2023. 1

[31] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[32] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 1, 3, 7, 8, 9, 10

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[34] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[35] K. Sarkar, V. Golyanik, L. Liu, and C. Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

[36] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[37] K. Schwarz, N. Mueller, and P. Kontschieder. Generative gaussian splatting: Generating 3d scenes with video diffusion priors. *arXiv preprint arXiv:2503.13272*, 2025. 3

[38] Y. Shi, X. Yang, Y. Wan, and X. Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11264, 2022. 2

[39] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (ToG)*, 41(6):1–10, 2022. 3

[40] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[41] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 7

[42] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 1

[43] J. Xiang, J. Yang, Y. Deng, and X. Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2195–2205, 2023. 1, 2, 3, 7, 8, 9, 10

[44] Y. Xu, S. Peng, C. Yang, Y. Shen, and B. Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18430–18439, 2022. 3

[45] Z. Xu, J. Zhang, J. H. Liew, J. Feng, and M. Z. Shou. Xagen: 3d expressive human avatars generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[46] F. Yang, T. Chen, X. He, Z. Cai, L. Yang, S. Wu, and G. Lin. Attrihuman-3d: Editable 3d human avatar generation with attribute decomposition and indexing. *arXiv preprint arXiv:2312.02209*, 2023. 1

[47] Y. Yang, Y. Yang, H. Guo, R. Xiong, Y. Wang, and Y. Liao. Urbangiraffe: Representing urban scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9199–9210, 2023. 3

[48] X. Zhang, Z. Zheng, D. Gao, B. Zhang, P. Pan, and Y. Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18450–18459, 2022. 3

[49] P. Zheng, T. Liu, Z. Yi, and R. Ma. Semantichuman-hd: High-resolution semantic disentangled 3d human generation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 1, 2, 3, 5, 7, 9