G-NeRF: Geometry-enhanced Novel View Synthesis from Single-View Images

Zixiong Huang[1,3*] Qi Chen[2*] Libo Sun[2] Yifan Yang[1]
Naizhou Wang[3] Qi Wu[2] Mingkui Tan[1,4,5†]
[1]South China University of Technology [2]The University of Adelaide
[3]Guangzhou Shiyuan Electronics Co., Ltd [4]Pazhou Lab
[5]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
sesmilhzx@mail.scut.edu.cn, qi.chen04@adelaide.edu.au; mingkuitan@scut.edu.cn

Abstract

*Novel view synthesis aims to generate new view images of a given view image collection. Recent attempts address this problem relying on 3D geometry priors (*e.g.*, shapes, sizes, and positions) learned from multi-view images. However, such methods encounter the following limitations: 1) they require a set of multi-view images as training data for a specific scene (*e.g.*, face, car or chair), which is often unavailable in many real-world scenarios; 2) they fail to extract the geometry priors from single-view images due to the lack of multi-view supervision. In this paper, we propose a Geometry-enhanced NeRF (G-NeRF), which seeks to enhance the geometry priors by a geometry-guided multi-view synthesis approach, followed by a depth-aware training. In the synthesis process, inspired that existing 3D GAN models can unconditionally synthesize high-fidelity multi-view images, we seek to adopt off-the-shelf 3D GAN models, such as EG3D, as a free source to provide geometry priors through synthesizing multi-view data. Simultaneously, to further improve the geometry quality of the synthetic data, we introduce a truncation method to effectively sample latent codes within 3D GAN models. To tackle the absence of multi-view supervision for single-view images, we design the depth-aware training approach, incorporating a depth-aware discriminator to guide geometry priors through depth maps. Experiments demonstrate the effectiveness of our method in terms of both qualitative and quantitative results.*

1. Introduction

Neural Radiance Fields (NeRFs) have emerged as a state-of-the-art technique for synthesizing novel views of complex scenes from 2D images. By using deep neural networks, NeRF models are capable of modeling both the geometry and appearance of a scene, enabling the generation of high-quality and photorealistic 3D renderings from any desired viewpoint. Its remarkable performance has made it a mainstream method for novel view synthesis, and it has found diverse applications in fields such as virtual reality and digital human generation [13, 20]. Although NeRF has demonstrated exceptional performance in synthesizing novel views for many scenes, it does exhibit two prominent limitations.

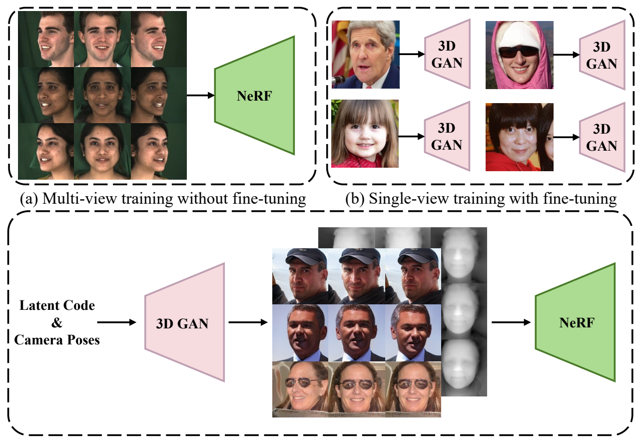


Figure 1. **Comparison of different methods.** To achieve single-shot novel view synthesis, previous methods either (a) require real-world multi-view images to establish geometry priors or (b) need additional optimization for a specific image. (c) In contrast, our method captures the geometry priors from an existing 3D GAN trained on single-view images only.

First, it requires multi-view images for training on a specific scene, while in most practical conditions only a single-view image is available. This limitation is widespread in real-world scenarios, such as when taking selfies or capturing a portrait of a pet. Some recent works [12, 49, 52] have attempted to address this limitation by introducing additional supervision, such as depth, or by collecting extensive multi-view images of the same class to learn sufficient ge-

†Corresponding author, *Authors contributed equally.
This work was done when Zixiong Huang was a research intern at Guangzhou Shiyuan Electronics Co., Ltd.

ometry priors. For example, SinNeRF [49] adopts ground truth depth or depth obtained from multi-view images to train a NeRF model for each individual image. Nonetheless, it requires accurate depth information, which is often hard to obtain from a single image. PixelNeRF [52] can perform few-shot or even single-shot novel view synthesis by collecting a set of multi-view images of the same class for training. However, it is inapplicable to some real-world scenarios where only single-view images are accessible. Furthermore, Pix2NeRF directly optimizes with image-level reconstruction and GAN losses. This, however, may incur a challenging one-to-many problem, as there could be multiple 3D shapes corresponding to one input image (*c.f.* Sec 4.4).

Second, the conventional NeRF model only focuses on overfitting to a particular scene while ignoring the sharable intrinsic geometry prior among the relevant scenes, such as scenes from the same class (*e.g.,* faces, cars or chairs). To address this issue, a recent line of research [7, 24, 37, 40] has explored the combination of Generative Adversarial Networks (GANs) [18] and NeRF to extract 3D prior knowledge from single-view datasets like FFHQ [26] and AFHQv2 [10]. While successful in diverse high-quality 3D scene generation, these models synthesize images from randomly sampled latent codes. This means to generate novel views for a specific image, we have to first map the image back to the latent space to obtain a corresponding latent code, which leads to extra test-time optimizations (*e.g.,* GAN inversion [41]) with additional time consumption. Additionally, these approaches adopt conventional reconstruction loss to fine-tune 3D GANs, neglecting the intrinsic geometry information embedded in these models. Thus, they tend to yield suboptimal outcomes characterized by geometry collapse [53], primarily due to the absence of multi-view supervision.

To address the above issues, we seek to develop a framework that can accomplish two goals simultaneously: exploiting geometry priors from an existing 3D GAN trained on real-world single-view images only and enabling single-shot novel view synthesis without the need for test-time fine-tuning (see Fig.1). To achieve this, we propose a novel single-shot novel view synthesis approach named Geometry-enhanced NeRF (G-NeRF), which seeks to enhance the geometry priors by two approaches: Geometry-guided Multi-View Synthesis (GMVS) and Depth-aware Training (DaT). In GMVS phase, we use a pre-trained 3D GAN model to generate a set of multi-view data, serving as a free source for establishing geometry priors. To further enhance the geometry priors in the synthetic data, we trade off the diversity against the geometry quality in 3D GANs, exploring different truncation ratios [4] to achieve a balance for high-quality geometry data synthesis. In the DaT training phase, we introduce a depth-aware discriminator to address the lack of multi-view supervision for single-view images. This discriminator distinguishes between the depth maps generated by the pre-trained 3D GAN model and those produced by our model, thereby providing additional depth supervision to enhance the geometry fidelity of our generated results. We summarize our contributions in three folds:

- To obtain sufficient multi-view images for training a single-shot NeRF model, we propose a Geometry-guided Multi-View Synthesis scheme to synthesize a set of multi-view data to build adequate geometry priors.
- To generate high-quality synthetic data, we explore the trade-off between diversity and geometry quality in 3D GANs, and then introduce a truncation method for multi-view data synthesis with enhanced geometry priors.
- For better learning geometry priors from single-view images, we design a Depth-aware Training method. It adots a depth-aware discriminator to enhance depth supervision, guiding the model to generate more realistic results.

2. Related Work

Neural Radiance Fields for few views. Representing 3D scenes as an implicit MLP-based function and using volume rendering technology, NeRF [32] and its variants [2, 34, 38, 46, 48, 54] have shown promising results in novel view synthesis tasks. However, these methods face a common challenge: they require a large number of views to obtain sufficient density information.

To address the data-hungry nature of NeRF, recent works [9, 12, 43, 49, 50, 52] aim to learn shared priors or incorporate additional supervision, such as depth maps. PixelNeRF [52] conditions NeRF on images by computing a fully convolutional image feature, which serves as a 3D representation for volumetric rendering. This approach allows for predicting NeRF from images in a feed-forward manner while leveraging shared priors of scenarios. DS-NeRF [12] introduces additional supervision by leveraging depth information recovered from 3D point clouds estimated using structure-from-motion methods. SinNeRF [49] further divides the NeRF training process into geometry learning and semantic learning. Combining camera and LiDAR data, GINA-3D [43] achieves generating neural assets from a single-view input. Despite their advancements, these methods still rely on a collection of multi-view images to provide an adequate 3D prior or require additional depth supervision (*e.g.*,, LiDAR data, depth map).

Generative 3D-aware image synthesis. Generative Adversarial Nets (GANs) [17, 19, 21, 26, 29, 33, 45] have achieved impressive success in 2D image synthesis tasks (*e.g.,* image generation, image-to-image translation). Recently, many attempts [5–7, 24, 35, 36, 42] have been made to extend GANs to 3D-aware tasks. HoloGAN [35]successfully disentangles 3D pose and identity by using an explicit volume representation, but this type of 3D representation also limits the resolution of the generated images. Some methods [6, 7, 15, 24, 36, 42] combine GANs and NeRF to synthe-

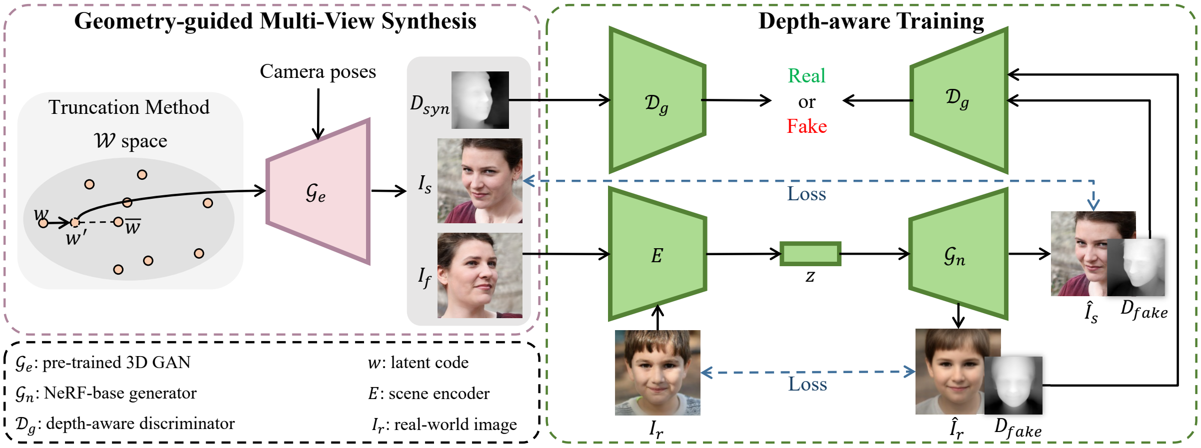


Figure 2. **Overall scheme of G-NeRF.** Given a latent code $\mathbf{w}$ randomly sample in $\mathcal{W}$ space, we first apply a truncation method to obtain $\mathbf{w}'$, bringing it closer to the center of mass of $\mathcal{W}$ space represented $\bar{\mathbf{w}}$. After that, in conjunction with a set of camera poses $\{\mathbf{P}_f, \mathbf{P}_s, \mathbf{P}_d\}$, we generate a triplet of synthetic data $\{I_f, I_s, \mathbf{D}_{syn}\}$. To capture geometry priors from synthetic multi-view images, we synthesize a novel view $\hat{I}_s$ using I_f as the reference image and enforce it to be consistent with I_s. Additionally, we employ a self-reconstruction task with the real-world image Ir. Moreover, we design a depth-aware discriminator $\mathcal{D}_g$ to further enhance the depth quality of the generated scenes.

size high-fidelity novel views. However, these methods generate random scenes using randomly sampled latent codes. To achieve single-view reconstruction, additional test-time optimization (*e.g.,* GAN inversion) is necessary [27, 51, 53]. While these methods excel at producing high-fidelity novel perspective images, they are constrained to the original 3D GAN model. Essentially, they rely on the original model for every inference. In contrast, our approach is versatile and capable of being applied to any model of the same type.

3. Geomery-enhanced NeRF

We aim to address the task of single-shot novel view synthesis in a unified framework. In other words, given an unseen single-view image (*e.g.,* human face or cat face), our goal is to synthesize novel views of the same scene. This task is inherently difficult due to the limited geometry information available in a single-view image. One potential solution involves leveraging multi-view datasets to establish adequate geometry priors. However, obtaining such datasets may be impractical in many real-world scenarios. To overcome this limitation, we introduce Geometry-enhanced NeRF (G-NeRF), a novel approach designed to achieve high-fidelity single-shot novel synthesis from single-view images.

As shown in Fig. 2, G-NeRF consists of two stages: 1) *Geometry-guided multi-view synthesis* (*c.f.* Sec. 3.1). To learn geometry priors of similar scenes, we leverage a pre-trained 3D GAN model $\mathcal{G}_e$ to synthesize a collection of multi-view images and corresponding depth maps. Simultaneously, we delve into the trade-off between diversity and geometry quality in 3D GANs, proposing an approach to achieve a balance for Geometry-guided data synthesis. 2) *Depth-aware training.* (*c.f.* Sec. 3.2). We seek to train our model using the combination of synthetic data and real-world single-view images with a reconstruction loss, denoted as $\mathcal{L}_{recon}$. However, solely applying $\mathcal{L}_{recon}$ is insufficient to learn satisfying geometry priors due to the absence of multi-view supervision for single-view images. To provide additional supervision, we introduce a depth-aware discriminator, which helps distinguish between the synthetic depth maps generated by our model and the depth maps obtained from the pre-trained 3D GAN model. We incorporate it into our training through a depth-aware adversarial loss, denoted as $\mathcal{L}gan$.

The overall optimization of our proposed G-NeRF minimizes the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_g \mathcal{L}_{gan}, \tag{1}$$

where λ_g is a hyper-parameter used to balance the reconstruction loss $\mathcal{L}_{recon}$ (see Eqn. (5)) and the adversarial loss $\mathcal{L}_{gan}$ (see Eqn. (6)).

3.1. Geometry-guided Multi-View Synthesis

In the absence of multi-view supervision, it is challenging to learn geometry information from single-view datasets, such as FFHQ [26] and AFHQv2-Cats [10]. Inspired that existing 3D GAN models can synthesize high-fidelity multi-view images while being trained solely on a set of single-view images. We aim to leverage the rich geometry priors embedded in these models. To this end, we propose a Geometry-guided Multi-View Synthesis scheme. In GMVS, we first utilize an off-the-shelf 3D GAN (*e.g.,* EG3D [7]) to synthesize a set of multi-view data. However, the use of naive synthetic data can lead to suboptimal results, such as unrealistic 3D shapes (see Fig. 7). To address this, we delve into the balance between diversity and geometry quality within 3D GANs. Building upon this exploration, we apply a latent truncation method to strike a more suitable balance between the diversity and geometry quality of the synthetic data, thereby contributing to the generation of more realistic results.

Trade-off between diversity and geometry quality. As confirmed by previous studies [1, 4], a trade-off exists between the fidelity and diversity of samples generated by GAN models. Take StyleGAN-based [26] methods as an example, we randomly sample a latent code $\mathbf{z}_d \sim p_z \subset \mathbb{R}^{512}$, where p_z is a normal distribution. Then, a mapping network $\mathcal{M}$ is adopted to map $\mathbf{z}_d$ to an intermediate latent space $\mathcal{W}$ to acquire $\mathbf{w}$. After that, as illustrated in Fig. 2, a truncation method is leveraged to draw $\mathbf{w}$ closer to the center of mass of $\mathcal{W}$ space by:

$$\mathbf{w}' = \bar{\mathbf{w}} + \psi(\mathbf{w} - \bar{\mathbf{w}}), \tag{2}$$

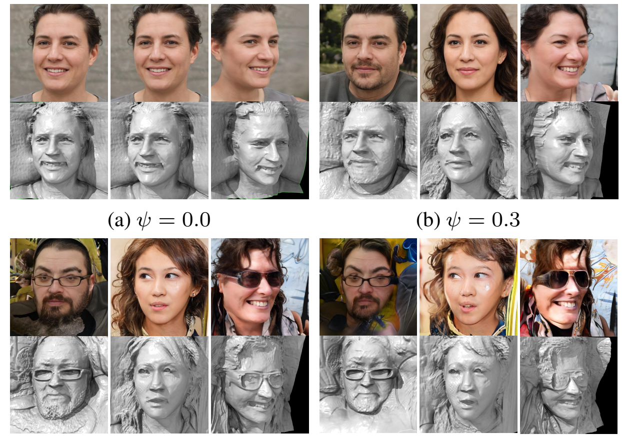
where $\bar{\mathbf{w}} = \mathbb{E}_{\mathbf{z}\sim p_z}[\mathcal{M}(\mathbf{z})]$ is the center of mass of $\mathcal{W}$ and $\psi \leq 1$ is a truncation ratio. The manipulation of ψ allows us to finely tune the trade-off between diversity and fidelity. Specifically, an increase in ψ augments diversity but may simultaneously diminish fidelity or the visual appeal of the generated results. This adjustment is driven by the fact that regions with lower density may be inadequately represented, posing challenges for the generator to effectively learn.

In this study, we delve deeper into this phenomenon within the realm of 3D GANs. Specifically, we leverage EG3D [7] to generate four sets of samples with varying truncation ratios and examine their differences. As depicted in Fig.3, our results demonstrate an augmented diversity with increasing truncation ratios, albeit accompanied by a gradual reduction in geometry quality.

Geometry-guided multi-view synthesis. Building upon the above observation, we conclude that synthetic data generated with various truncation ratios plays a crucial role in the geometry quality of final results. To obtain a set of multi-view data with high-quality geometry priors, we devise a Geometry-guided Multi-View Synthesis scheme. Specifically, we employ a pre-trained 3D GAN model $\mathcal{G}_e$, such as EG3D [7], to generate multi-view data of diverse scenes. As shown in Fig. 2, we randomly sample a latent code $\mathbf{w}$ and a set of camera pose $(\mathbf{P}_f, \mathbf{P}_s, \mathbf{P}_d) \sim p_\xi$, where p_ξ is a distribution associated with camera poses from real-world single-view images. We then apply the truncation method with an empirically selected ratio $\psi = 0.5$ to obtain a truncated latent code $\mathbf{w}'$ by Eqn. (2). As discussed earlier, this truncation method draws $\mathbf{w}$ closer to the center of mass of $\mathcal{W}$ space, ensuring the geometry quality of the generated results. Finally, we synthesize a triplet of Geometry-guided multi-view data $\mathbf{I}$ with a generator $\mathcal{G}_e$ by:

$$\mathbf{I} = \{I_f, I_s, \mathbf{D}_{syn}\} = \mathcal{G}_e(\mathbf{w}', \mathbf{P}_f, \mathbf{P}_s, \mathbf{P}_d), \tag{3}$$

where I_f and I_s denote the first and second synthetic images regarding a common scene but rendered from different viewpoints, and $\mathbf{D}_{syn}$ denotes the depth map of the scene.



(a) $\psi = 0.0$ (b) $\psi = 0.3$ (c) $\psi = 0.7$ (d) $\psi = 1.0$

Figure 3. Illustration of the trade-off between identity diversity and geometry quality of the generated samples. Samples are generated by EG3D [7] with the same set of latent codes and different truncation ratios ψ. As ψ rises, the identity diversity (*e.g.,* hair color, skin color, and glasses) of the generated samples also increases. In contrast, the geometry quality of these scenes gradually reduces.

3.2. Depth-aware Training

In this section, we seek to train our model with a combination of synthetic data and real-world single-view images. However, directly applying reconstruction loss to single-view images is not beneficial for learning satisfying geometry priors due to the absence of multi-view supervision. To address this, we introduce a depth-aware discriminator $\mathcal{D}_g$ to provide additional depth supervision.

Incorporating synthetic and real-world data. We incorporate synthetic data with real-world single-view images to train our model. Specifically, we simultaneously generate a novel view and depth map with a selection factor $\gamma \sim \mathcal{U}(0, 1)$, which is formulated as:

$$\begin{aligned} \hat{I}_s, \mathbf{D}_{fake} &= \mathcal{G}_n(\mathbf{P}_s, E(I_f)), \quad \text{if } 0 \leq \gamma \leq 0.5; \\ \hat{I}_r, \mathbf{D}_{fake} &= \mathcal{G}_n(\mathbf{P}_r, E(I_r)), \quad \text{if } 0.5 < \gamma \leq 1, \end{aligned} \tag{4}$$

where I_r is a real-world image associated with the pose $\mathbf{P}_r$, $\mathcal{G}_n$ is a NeRF-based generator and E is an scene encoder. This alternative training scheme allows the model to capture the geometry priors within the multi-view images, while still learning diverse appearance information from the real-world images. With the generated views and depth maps, we depict a reconstruction loss and an adversarial loss as follows.

Reconstruction with paired images. For synthetic multi-view image pairs $\{I_f, I_s\}$, we adopt I_f as the reference image and render a novel view image $\hat{I}_s$ from the same viewpoint as I_s. In this way, we train our G-NeRF by minimizing photometric error w.r.t. I_s and $\hat{I}_s$. For real-world single-view images, our objective is to leverage them to

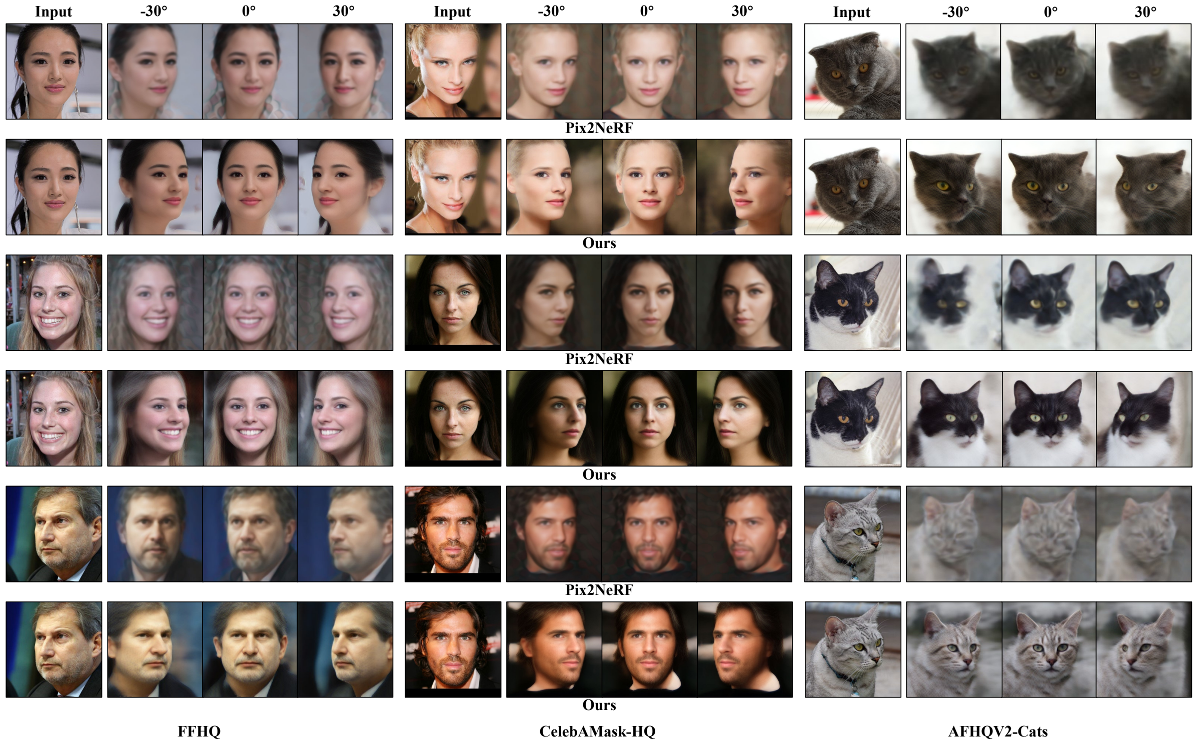


Figure 4. **Qualitative comparison.** Compared to Pix2NeRF [5], our G-NeRF demonstrates the capability to generate novel views that closely resemble reference images with higher clarity (Comparison at 512^2).

augment the diversity of the synthesized scenes. To this end, we select a single-view image I_r as a reference and train G-NeRF by generating a $\hat{I}_r$ that shares the same viewpoint with I_r. We update G-NeRF by enforcing similarity between $\hat{I}_r$ and I_r. This schedule is implemented using a reconstruction loss, which is formulated with the image-pair data $(I_{fake}, I_{ref}) \in \{(\hat{I}_s, I_s), (\hat{I}_r, I_r)\}$,

$$\begin{aligned}\mathcal{L}_{recon} = \mathbb{E}\,[||I_{fake} - I_{ref}||_1 + \mathcal{L}_{ssim}(I_{fake}, I_{ref}) \\ + \mathcal{L}_{vgg}(I_{fake}, I_{ref})]\,.\end{aligned} \tag{5}$$

$\mathcal{L}_{ssim}$ is SSIM loss [47] and $\mathcal{L}_{vgg}$ is perceptual loss [25].

Depth-aware discriminator. In the absence of multi-view supervision for real-world single-view images, we observe some degradation in the geometry quality of our generated results (see Fig. 7). To address this, we introduce a depth-aware discriminator denoted as $\mathcal{D}_g$. Concretely, $\mathcal{D}_g$ is trained to distinguish between the generated depth map $\mathbf{D}_{fake}$ and ground truth $\mathbf{D}_{syn}$ from the synthetic data, thereby introduce additional geometry priors into our model. In contrast to employing a simple reconstruction loss for basic depth supervision, $\mathcal{D}_g$ offers several advantages: 1) enabling depth supervision for real-world single-view images whose depth maps may not be available; 2) ensuring that depth maps generated from various viewpoints are realistic and coherent with the scene, contributing to the overall quality of the synthesized novel views. Following EG3D [7], we condition $\mathcal{D}_g$ on a camera pose and use an adversarial loss with an R1 regularization [31] to train $\mathcal{D}_g$:

$$\begin{aligned}\mathcal{L}_{gan} = \;& \mathbb{E}[f(\mathcal{D}_g(\mathbf{D}_{syn}|\mathbf{P}_d)] \\ & + \mathbb{E}\left[f(-\mathcal{D}_g(\mathbf{D}_{fake}|\mathbf{P}_f)) + \lambda|\nabla\mathcal{D}_g(\mathbf{D}_{fake}|\mathbf{P}_f)|^2\right],\end{aligned} \tag{6}$$

where $\mathbf{P}_f \in (\mathbf{P}_r, \mathbf{P}_s)$ is the camera pose used to generate novel views and $f(\cdot)$ is a softplus activation. Note that, for each real-world image, we train $\mathcal{D}_g$ not only with its corresponding camera pose $\mathbf{P}_r$ but also with other camera poses randomly sampled from p_ξ to provide more comprehensive depth supervision.

4. Experiments

4.1. Experimental Setup

Datasets. We train our model with FFHQ [26] and AFHQv2-Cats [10], repectively. During the evaluation, we leverage an additional in-the-wild dataset named CelebAMask-HQ [28]. FFHQ [26] is a real-world dataset with around 70k high-quality human faces. CelebAMask-HQ [28] is a large-scale face dataset with 30k high-resolution human faces. After preprocessing, we randomly hold out 8k images as the test set. AFHQv2-Cats [10] contains 5065 cat images of different types. We randomly select 4k images as the training set and the rest as the test set.

Table 1. **Quantitative comparison.** For AFHQv2-Cats [10], since there is non-trivial to estimate depth maps for cat faces, we only evaluate on FID and KID×100. Note that our results of 512^2 resolution are synthesized by a super-resolution module and we did not apply the same super-resolution operation to the depth map. Thus, the depth accuracy at this resolution is not available. The **bold** highlights the best results among methods requiring single-view images only. Approaches labeled in gray necessitate the availability of multi-view training data. Legend: * –requires multi-view training data and test time optimization.

Method	FFHQ [26]				CelebAMask-HQ [28]				AFHQv2-Cats [10]	
	FID(↓)	KID(↓)	Depth(↓)	ID(↑)	FID(↓)	KID(↓)	Depth(↓)	ID(↑)	FID(↓)	KID(↓)
Pix2NeRF 64^2 [5]	32.44	2.37	0.40	0.25	89.79	12.22	0.38	0.19	25.34	1.00
G-NeRF 64^2 (Ours)	**26.04**	**2.09**	**0.35**	**0.43**	**75.76**	10.48	**0.32**	**0.37**	**18.64**	**0.73**
Pix2NeRF 512^2 [5]	75.04	5.97	0.41	0.20	118.92	13.08	0.38	0.15	50.55	3.33
G-NeRF 512^2 (Ours)	40.24	2.72	-	0.36	78.38	**8.68**	-	0.31	21.78	1.00

Method	ShapeNet Chairs [8]			ShapeNet Cars [8]			Average		
	SSIM (↑)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	LPIPS (↓)
ENR 128^2 [16]	0.91	22.83	0.10	0.90	22.26	0.13	0.91	22.55	0.12
SRN 128^2 [44]	0.89	22.89	0.10	0.89	22.25	0.13	0.89	22.57	0.12
PixelNeRF 128^2 [5]	0.91	23.72	0.10	0.90	23.17	0.15	0.91	23.45	0.13
CodeNeRF* 128^2 [23]	0.90	23.66	0.11	0.91	23.80	0.12	0.91	23.73	0.12
VisionNeRF* 128^2 [30]	0.93	24.48	0.08	0.91	22.28	0.08	0.92	23.37	0.08
Pix2NeRF 64^2 [5]	0.80	18.13	0.12	0.73	16.57	0.19	0.77	17.35	0.16
G-NeRF 64^2 (Ours)	**0.88**	**22.31**	**0.07**	**0.86**	**21.03**	**0.10**	**0.87**	**21.67**	**0.09**
Pix2NeRF 128^2 [5]	0.83	17.73	0.12	0.78	16.24	0.20	0.81	16.99	0.16
G-NeRF 128^2 (Ours)	**0.88**	20.29	0.08	**0.86**	19.44	0.11	**0.87**	19.87	0.10

Evaluation metrics. Following Pix2NeRF [5], we report Frechet-Inception Distance (*i.e.,* FID) [22] and Kernel-Inception Distance (*i.e.,* KID) [3] for novel view images. We also use Depth accuracy (*i.e.,* Depth) to measure the depth quality. Specifically, we evaluate depth quality by calculating MSE loss against pseudo-ground-truth depth maps estimated from test set images by [14] and our generated depth maps. We assess multi-view consistency (ID) by calculating the mean Arcface [11] cosine similarity score between the input images and the corresponding novel views. In the ablation study, to assess the quality of the generated images, we employ Structural Similarity (SSIM) [47].

Implementation details. We use two pre-trained models of EG3D [7] trained on FFHQ [26] and AFHQv2-Cats [10] respectively to synthesize multi-view data. Specifically, we generate 60k triplets of multi-view data for FFHQ [26] and 4k for AFHQv2-Cats [10]. We adopt the same pre-processing strategy as [7]. All images are aligned and processed into size 512^2. Note that we use a super-resolution module to generate a promote a low-resolution image (*i.e.,* 64^2) to a high resolution (*i.e.,* 512^2). Please refer to our appendix for more implementation details.

4.2. Comparison with State-of-the-art Methods

Quantitative comparison. We compare our method against the state-of-the-art method Pix2NeRF [5] for novel-view synthesis from a single image with real-world single-view datasets. Note that we use scripts provided by the authors of Pix2NeRF [5] to train with FFHQ [26] dataset and AFHQv2-Cats [10] dataset. Since AFHQv2-Cats contains a relatively small amount of data, we train for 150k iterations with a batch size of 48 for Pix2NeRF [5] until convergence. Tab. 1 provides quantitative metrics comparing the proposed approach against Pix2NeRF [5]. Our model demonstrates better results in terms of all metrics across all datasets. Moreover, our model is capable of generating higher-resolution images without compromising on efficiency, whereas Pix2NeRF requires significantly more time to achieve comparable results [7]. In other words, our model can generate novel views with more realistic appearances and precise shapes from a single-view image. Notably, Pix2NeRF fails to synthesize novel views on AFHQv2-Cats [10], because most of the cat faces in this dataset are facing the camera, resulting in limited geometry information. In contrast, our model derives advantages from training with a collection of synthetic multi-view images and

Table 2. **Comparison of inference cost with PTI [41].** We fine-tune an EG3D model to fit a single-view image until it reconstructs the same level of details as ours (*i.e.,* LPIPS: 0.32). Then, we use this fine-tuned model to synthesize four novel views and compare the time taken at each stage with our model on an RTX A800 GPU.

Method	Inference Cost (↓)	Depth (↓)	Depth† (↓)	FID (↓)	KID (↓)	ID (↑)
PTI[41]	76.9s	0.37	0.55	35.53	2.30	0.36
Ours	1.1s	0.35	0.53	40.24	2.72	0.35

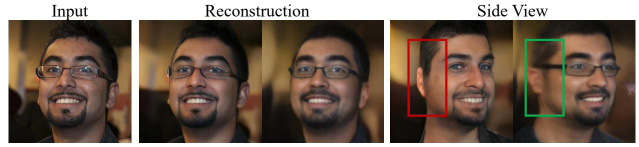


Figure 5. **Qualitative comparisons with PTI [41].**

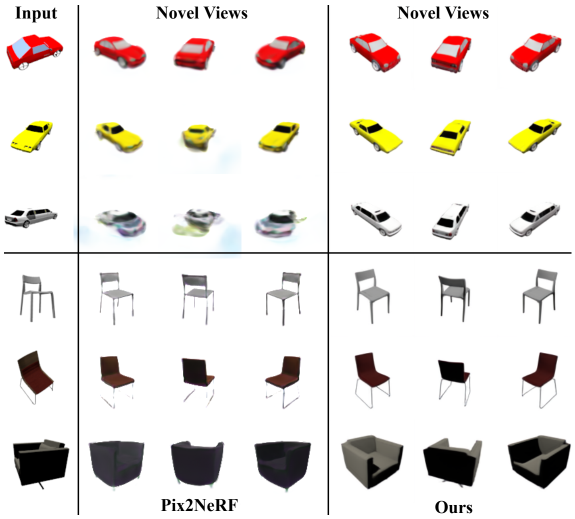


Figure 6. **Qualitative comparisons with Pix2NeRF [5] on ShapeNet Cars & Chairs [8, 44].**

a depth-aware discriminator, which facilitates explicit 3D supervision. Consequently, our model effectively learns a robust geometry prior, even in such challenging scenarios

Qualitative comparison. Fig. 4 presents results generated by our method and Pix2NeRF [5] on FFHQ [26], CelebAMask-HQ [28], and AFHQv2-Cats [10]. Our method can synthesize high-quality novel views even with a single image as reference, yet existing few-shot NeRF methods can not train on these single-view datasets without multi-view image pairs. Compared to Pix2NeRF [5], our methods can generate more realistic results while preserving a more similar identity to the reference images. Thanks to the depth-aware discriminator, our method excels in producing high-quality results even under extreme camera poses (see the bottom row of the leftmost column in Fig. 4). Meanwhile, our method can also learn geometry priors from AFHQv2-Cats [10] which contains a limited range of poses while Pix2NeRF [5] fails (see the rightmost column in Fig. 4). We also provide more geometry visualization results, please refer to our appendix for more details.

Comparison on multi-view datasets. Additionally, we performed experiments on the training set of ShapeNet Cars & Chairs [8, 44], which includes uniformly distributed camera poses around a sphere. Following the experimental setting of Pix2NeRF [5], we filtered the training set for both datasets to only include the upper hemisphere and evaluate the test split. Since EG3D [7] does not include an evaluation on ShapeNet Chairs, we first train an EG3D model using the same settings as those used for training ShapeNet Cars. We evaluate the performance using well-established image quality metrics commonly employed in novel view synthesis tasks, including pixel-level measures such as SSIM and PSNR, as well as a feature-level metric called LPIPS.

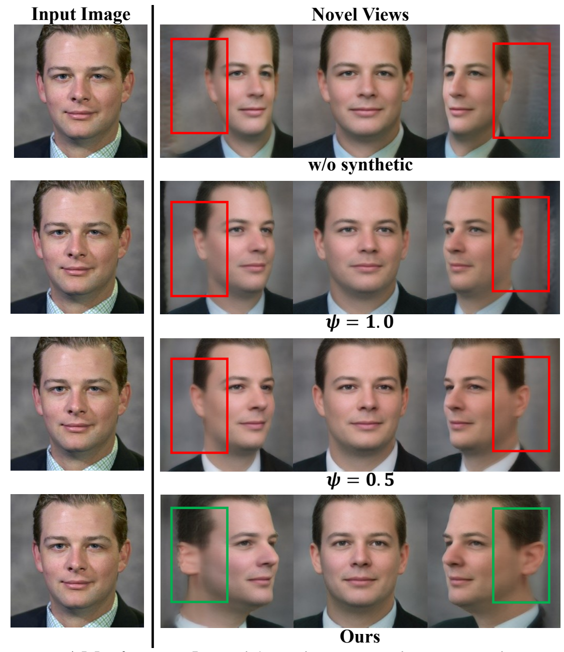


Figure 7. **Ablation study.** Without incorporating truncation method (*i.e.*, $\psi = 1.0$) and depth-aware training, our model fails to generate results with realistic geometry (see the red boxes in the figure).

As depicted in Tab. 1, our method consistently outperforms Pix2NeRF [5] in terms of all metrics. Simultaneously, our method remains competitive with other approaches that rely on multi-view training data, while our method does not employ multi-view supervision on the ShapeNet datasets [8]. Fig. 6 demonstrates our method's superior accuracy in shape and texture predictions when compared to Pix2NeRF [5].

Comparison with GAN inversion method. We employ Pivotal Tuning Inversion (PTI) [41] to fine-tune an EG3D model, enabling its adaptation to a single-view image. The fine-tuning process continues until the LPIPS [55] loss matches that of our method. Subsequently, we compare the time taken by our method with that of the fine-tuning-based approach. As shown in Tab. 2, our method demonstrates significantly faster than the GAN inversion method. Moreover, our method outperforms PTI in terms of geometry quality (Depth and Depth†) and inference cost while being competitive in image quality (FID, KID, and ID). Visual results also show the quality of another view produced by PTI exhibits instability (see Fig. 5).

4.3. Ablation Study

In this part, we provide more experiments and analysis of our proposed modules to verify their effectiveness. For simplicity, we only conduct experiments on FFHQ [26], and all the depth accuracy is evaluated at a resolution of 64^2 and the others are evaluated at 512^2. The data presented in Tab. 3 correspond to the cases in Fig. 7, which visualizes the results of different training settings.

Table 3. **Quantitative results of ablation studies.** The row without truncation ratio means training without synthetic data. The **bold** numbers highlight the best results. † Evaluated on side faces.

Trunc. Ratio	Real Img.	$\mathcal{D}_g$	FID (↓)	KID (↓)	Depth (↓)	Depth† (↓)	ID (↑)	SSIM (↑)
✗	✓	✗	**33.13**	**2.22**	0.42	0.83	**0.43**	**0.66**
1.0	✓	✗	34.72	2.42	0.37	0.64	0.38	0.65
0.5	✓	✗	40.13	2.64	0.35	0.59	0.35	0.64
0.5 (Ours)	✓	✓	40.24	2.72	**0.35**	**0.53**	0.35	0.63

Impact of geometry-guided multi-view synthesis scheme. We train our model using synthetic data generated with varying truncation ratios to assess the effectiveness of our Geometry-guided Multi-View Synthesis scheme. As shown in Tab. 3 (the first two rows), our model with synthetic data obtains better depths and competitive identities (FID, KID, ID, and SSIM) compared to the model without synthetic data. When decreasing the truncation ratio (1.0 → 0.5), the depth quality can be further improved. Particularly, we can observe a significant drop in the depth accuracy when applying a truncation ratio of 1.0, which severely harms the realism of the generated scenes (see the bottom row in Fig. 7). In essence, if the truncation trick is not taken, it is challenging for our model to generate realistic results.

Impact of depth-aware discriminator. To verify the effectiveness of our depth-aware discriminator, we train a model without $\mathcal{D}_g$. We can see from Fig. 7 that the model learns a degenerate solution where the human head appears flattened and sunk into the background. This phenomenon typically occurs when one of the ears is not visible in the input image. As shown in Tab. 3, the removal of $\mathcal{D}_g$ yields to a slight improvement in image quality metric (*i.e.,* SSIM: $0.63 \rightarrow 0.64$). This is primarily because the addition of a discriminator naturally brings some disturbance to our training process.

As $\mathcal{D}_g$ is directly applied to depth maps, depth accuracy is the most important metric for verifying its effectiveness. From the fifth column in Tab. 3, we can observe that the depth accuracy remains almost unchanged without $\mathcal{D}_g$ (*i.e.,* $0.35 \rightarrow 0.35$). However, the sixth column in Tab. 3 reveals a significant drop in depth accuracy (*i.e.,* $0.53 \rightarrow 0.59$) when faces are turned to the side. In other words, $\mathcal{D}_g$ plays a crucial role in distinguishing unnatural depth maps and contributes to achieving a more realistic geometry from various viewpoints.

4.4. Further Discussion

We conduct two intuitive comparisons to further validate the effectiveness of our method: 1) Conditional EG3D (denoted as C-EG3D). We condition the EG3D [7] model on an input image and performed novel view synthesis using only a single-view reconstruction loss and the GAN loss as used in the original EG3D paper [7]. However, as shown in Fig. 8, this approach fails to capture 3D information from the input image. It also exhibits artifacts when presented with novel viewpoints. We recognize that this is a challenging one-to-many problem, as it is non-trivial for models to learn geometry without explicit multi-view supervision or accurate depth information. 2) Simple depth supervision (denoted as SDS). We use a state-of-the-art depth estimation model, such as MiDaS [39], to generate depth maps for images in the FFHQ dataset [26]. After scaling and adjusting these depth maps, we employ them for basic depth supervision along with a single-view reconstruction loss, without incorporating our proposed geometry-guided multi-view synthesis scheme. In Fig. 8, this approach fails to produce accurate geometry, resulting in flat, plane-like facial reconstructions. The limitations stem from the challenge of MiDaS [39] in estimating fine-grained facial details, such as the nose, mouth, and eyes.

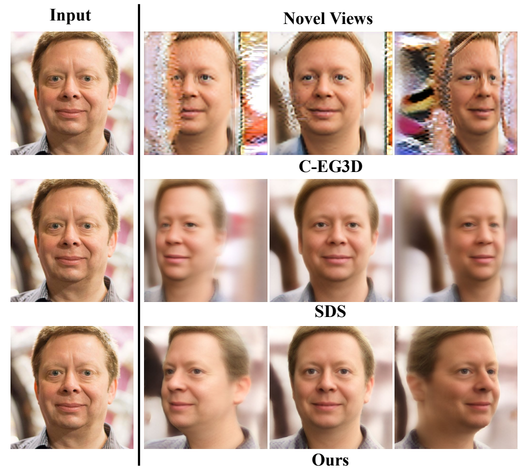


Figure 8. **Further comparisons with two intuitive methods.** We compare our G-NeRF with two intuitive methods to further verify the effectiveness of our method.

5. Conclusion

In this work, we propose the G-NeRF, a single-shot novel view synthesis method designed for high-fidelity novel view synthesis using only real-world single-view images. G-NeRF seeks to enhance geometry priors into a NeRF model through two stages: Geometry-guided Multi-View Synthesis (GMVS) and Depth-aware Training (DaT). GMVS leverages an off-the-shelf 3D GAN model to synthesize multi-view data, enhanced with a truncation method for improved geometry quality. DaT further refines the NeRF model by incorporating a depth-aware discriminator, guiding the learning process through depth maps. Our proposed method is evaluated extensively on multiple real-world datasets and the experimental results demonstrate its effectiveness.

6. Acknowledgement

This work was partially supported by National Natural Science Foundation of China (NSFC) 62072190, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT07X183, and TCL Science and Technology Innovation Fund.

References

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023. 4

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

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *Int. Conf. Learn. Represent.*, 2018. 6

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *Int. Conf. Learn. Represent.*, 2019. 2, 4

[5] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2NeRF: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3981–3990, 2022. 2, 5, 6, 7

[6] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5799–5809, 2021. 2

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16123–16133, 2022. 2, 3, 4, 5, 6, 7, 8

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6, 7

[9] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *IEEE Int. Conf. Comput. Vis.*, pages 14124–14133, 2021. 2

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8188–8197, 2020. 2, 3, 5, 6, 7

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019. 6

[12] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12882–12891, 2022. 1, 2

[13] Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Trans. Vis. Comput. Graph.*, 28(11):3854–3864, 2022. 1

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2019. 6

[15] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10673–10683, 2022. 2

[16] Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *Int. Conf. Mach. Learn.*, pages 2761–2770, 2020. 6

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, 2014. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.*, 27, 2014. 2

[19] Yong Guo, Qi Chen, Jian Chen, Qingyao Wu, Qinfeng Shi, and Mingkui Tan. Auto-embedding generative adversarial networks for high resolution image synthesis. *IEEE Trans. Multimedia*, 21(11):2726–2737, 2019. 2

[20] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE Int. Conf. Comput. Vis.*, pages 5784–5794, 2021. 1

[21] Xianjun Han, Hongyu Yang, Guanyu Xing, and Yanli Liu. Asymmetric joint gans for normalizing face illumination from a single image. *IEEE Trans. Multimedia*, 22(6):1619–1633, 2019. 2

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, pages 6626–6637, 2017. 6

[23] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *IEEE Int. Conf. Comput. Vis.*, pages 12949–12958, 2021. 6

[24] Peng Wang Jiatao Gu, Lingjie Liu and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *Int. Conf. Learn. Represent.*, 2022. 2

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, pages 694–711, 2016. 5

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. 2, 3, 4, 5, 6, 7, 8

[27] Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. Self-supervised geometry-aware encoder for style-based 3d gan inversion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20940–20949, 2023. 3

[28] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5549–5558, 2020. 5, 6, 7

[29] Shaojie Li, Mingbao Lin, Yan Wang, Chao Fei, Ling Shao, and Rongrong Ji. Learning efficient gans for image translation via differentiable masks and co-attention distillation. *IEEE Trans. Multimedia*, 2022. 2

[30] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 806–815, 2023. 6

[31] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *Int. Conf. Mach. Learn.*, pages 3481–3490, 2018. 5

[32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, pages 405–421, 2020. 2

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[34] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):1–15, 2022. 2

[35] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *IEEE Int. Conf. Comput. Vis.*, pages 7588–7597, 2019. 2

[36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11453–11464, 2021. 2

[37] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10318–10327, 2021. 2

[39] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3), 2022. 8

[40] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. In *Int. Conf. Mach. Learn.*, 2021. 2

[41] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 2, 6, 7

[42] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Adv. Neural Inform. Process. Syst.*, pages 20154–20166, 2020. 2

[43] Bokui Shen, Xinchen Yan, Charles R Qi, Mahyar Najibi, Boyang Deng, Leonidas Guibas, Yin Zhou, and Dragomir Anguelov. Gina-3d: Learning to generate implicit neural assets in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4913–4926, 2023. 2

[44] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Adv. Neural Inform. Process. Syst.*, 2019. 6, 7

[45] Jingkuan Song, Jingqiu Zhang, Lianli Gao, Zhou Zhao, and Heng Tao Shen. Agegan++: Face aging and rejuvenation with dual conditional gans. *IEEE Trans. Multimedia*, 24:791–804, 2021. 2

[46] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 5, 6

[48] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[49] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. SinNeRF: Training neural radiance fields on complex scenes from a single image. In *Eur. Conf. Comput. Vis.*, pages 736–753, 2022. 1, 2

[50] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. S^3-nerf: Neural reflectance field from shading and shadow under a single viewpoint. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[51] Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, and Yun Fu. Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[52] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4578–4587, 2021. 1, 2

[53] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *IEEE Int. Conf. Comput. Vis.*, pages 2437–2447, 2023. 2, 3

[54] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 7