NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations

Yonggan Fu [1] **Ye Yuan** [1] **Souvik Kundu** [2] **Shang Wu** [1] **Shunyao Zhang** [3] **Yingyan (Celine) Lin** [1]

Abstract

Generalizable Neural Radiance Fields (GNeRF) are one of the most promising real-world solutions for novel view synthesis, thanks to their cross-scene generalization capability and thus the possibility of instant rendering on new scenes. While adversarial robustness is essential for real-world applications, little study has been devoted to understanding its implication on GNeRF. We hypothesize that because GNeRF is implemented by conditioning on the source views from new scenes, which are often acquired from the Internet or third-party providers, there are potential new security concerns regarding its real-world applications. Meanwhile, existing understanding and solutions for neural networks' adversarial robustness may not be applicable to GNeRF, due to its 3D nature and uniquely diverse operations. To this end, we present NeRFool, which to the best of our knowledge is the first work that sets out to understand the adversarial robustness of GNeRF. Specifically, NeRFool unveils the vulnerability patterns and important insights regarding GNeRF's adversarial robustness. Built upon the above insights gained from NeRFool, we further develop NeRFool$^+$, which integrates two techniques capable of effectively attacking GNeRF across a wide range of target views, and provide guidelines for defending against our proposed attacks. We believe that our NeRFool/NeRFool$^+$ lays the initial foundation for future innovations in developing robust real-world GNeRF solutions. Our codes are available at: https://github.com/GATECH-EIC/NeRFool.

[1]School of Computer Science, Georgia Institute of Technology, USA [2]Intel Labs, San Diego, USA [3]Rice University, USA. Correspondence to: Yingyan (Celine) Lin <celine.lin@gatech.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Novel view synthesis (NVS), which aims to generate photo-realistic novel views of a scene given only a set of sparsely sampled views, has become an essential functionality in real-world 3D vision applications. Among various NVS techniques, neural radiance fields (NeRF) (Mildenhall et al., 2021) have recently gained substantial attention thanks to their record-breaking rendering quality, igniting a tremendous demand for NeRF-based NVS solutions. As many real-world NVS applications require instant and real-time rendering on new scenes, generalizable NeRF (GNeRF) variants (Yu et al., 2021b; Wang et al., 2021; Chen et al., 2021; Liu et al., 2022) have emerged as the most appealing real-world NeRF solutions. In particular, GNeRF conditions NeRF on the source views from a new target scene to achieve cross-scene generalization and enable new scene reconstruction via only a single forward pass execution.

Despite GNeRF's big promise towards real-world NVS solutions, it is currently unclear whether it can fulfill the essential robustness requirements. In fact, we hypothesize that GNeRF's introduced conditionality on source views can cause new security concerns. This is because the source views that describe a new scene, e.g., a tourist attraction, are often acquired from the Internet/third-party providers, leaving opportunities for adversaries to take advantage in terms of malicious attacks. For example, adversarial perturbations (Goodfellow et al., 2014; Madry et al., 2017) can be injected into source views by adversaries to severely degrade the reconstruction accuracy of GNeRF. With the increasing deployment of NeRF-powered security-critical applications, such as robot navigation systems (Adamkiewicz et al., 2022; Maggio et al., 2022; Moreau et al., 2022) and autonomous driving systems (Kundu et al., 2022; Fu et al., 2022; Siddiqui et al., 2023), it is imperative to understand the adversarial robustness of GNeRF for the unleashing of its cross-scene generalization capability toward real-world NeRF-based NVS solutions.

To address the imperative need above, one may naturally consider borrowing the existing insights about the adversarial robustness of deep neural networks (DNNs). However, those insights may not be applicable to NeRF due to its unique properties and processing pipeline. First, unlike 2D

tasks, the 3D nature of NVS tasks requires NeRF to reconstruct the target 3D scenes across different views. As such, it is not straightforward how to ensure that perturbing 2D source views of a scene can effectively pollute the entire 3D scene. Second, NeRF features a volume rendering process, in which pixels are rendered via alpha compositing (Mildenhall et al., 2021) from estimated density and color, and thus involves more diverse operations than DNNs. Hence, it is unclear which component (e.g., the density/color or both) of NeRF is more vulnerable (or needs stronger protection). Third, the ray marching process of GNeRF relies on the geometric relationship among different views, and thus perturbations optimized for destructing one view may be effective for destructing another view. This poses new risks of adversarial perturbations targeting GNeRF which could be transferable across a wide range of views.

To this end, this work sets out to (1) raise the community's awareness regarding the potential security concerns of GNeRF due to adversarial perturbations and (2) enhance our understanding of GNeRF's vulnerability patterns. We summarize our contributions as follows:

- We present both NeRFool and NeRFool$^+$, which to the best of our knowledge are **the first** works that uncover and study the vulnerability of GNeRF against adversarial perturbations. As such, NeRFool/NeRFool$^+$ open up a new perspective in NeRF literature and can shed light on future innovations toward robust real-world GNeRF-based NVS solutions.
- In NeRFool, we study the vulnerability patterns of various GNeRF variants through systematic analysis and experiments, and discover that, interestingly, (1) increased conditionality on source views can cause a higher vulnerability of GNeRF and (2) adversarial perturbations on density have a significantly stronger "ruining" ability than that on color when attacking GNeRF, especially on scenes with complex geometry.
- Built upon the above insights gained from NeRFool, we further develop NeRFool$^+$, which integrates two optimization techniques, novel target view sampling and geometric error maximization, that can effectively attack GNeRF across a wide range of target views.
- We further embark on an intriguing exploration to defend against our NeRFool attacks and discover the benign impact of adversarial perturbations on GNeRF's reconstruction accuracy, deepening the understanding regarding GNeRF's robustness.

2. Related Works

View synthesis and NeRF. View synthesis renders photo-realistic images from novel views of a scene based on a set of sparsely sampled views (Hedman et al., 2018; Thies et al., 2019; Lombardi et al., 2019; Mildenhall et al., 2021). Among existing techniques, NeRF (Mildenhall et al., 2021), which implicitly represents a scene as a continuous 5D radiance field parameterized by a multilayer perceptron (MLP), has gained increasing popularity. Follow-up works (1) improve NeRF's rendering quality under extremely sparse views (Xu et al., 2022a; Niemeyer et al., 2022), (2) accelerate NeRF via reducing the complexity of MLP (Lindell et al., 2021; Rebain et al., 2021) or exploring the free space via 3D occupancy grids (Yu et al., 2021a; Garbin et al., 2021), and (3) extend NeRF to other tasks, e.g., generative modeling (Chan et al., 2021; Schwarz et al., 2020), dynamic scenes (Li et al., 2021; Ost et al., 2021), or lighting/reflection modeling (Srinivasan et al., 2021; Verbin et al., 2022).

Generalizable NeRFs. To avoid tedious per-scene optimization and endow NeRF with cross-scene generalization capability, generalizable NeRFs (Yu et al., 2021b; Wang et al., 2021) have been developed to reconstruct the radiance field of a new scene via merely a one-shot forward pass. Specifically, recent GNeRF techniques (Yu et al., 2021b; Wang et al., 2021; Reizenstein et al., 2021; Wang et al., 2022b; Chen et al., 2021; Xu et al., 2022b; Liu et al., 2022) are implemented by conditioning vanilla NeRF techniques on a set of source views from the new scene via taking the extracted scene features from the source views as inputs. Despite their promise, GNeRF's pipeline, i.e., conditioning NeRF on source views, leaves opportunities for adversaries to take advantage in terms of malicious attacks. For example, adversaries can attack GNeRF by injecting adversarial perturbations onto the aforementioned source views. Hence, it is crucial to understand GNeRF's adversarial robustness for ensuring their real-world deployment, which has yet to be explored by the literature.

Adversarial attack and defense. DNNs are well-recognized to be adversarially vulnerable (Goodfellow et al., 2014). Various attacks (Madry et al., 2017; Carlini & Wagner, 2017; Andriushchenko et al., 2020) are proposed to aggressively degrade the achievable accuracy of DNNs for different tasks (Arnab et al., 2018; Carlini & Wagner, 2018; Zhang et al., 2020). In parallel, a variety of defense schemes (Guo et al., 2017; Feinman et al., 2017; Madry et al., 2017; Shafahi et al., 2019; Wong et al., 2019) is developed to enhance DNNs' adversarial robustness. The readers are referred to (Akhtar & Mian, 2018; Chakraborty et al., 2018) for more attack and defense methods. Nevertheless, to the best of our knowledge, **no existing work has been dedicated to studying NeRF's adversarial robustness.** Instead, recent works have attempted to combine adversarial optimization with NeRF for various purposes. For example, (Chen et al., 2022) improves NeRF's accuracy using augmented data obtained from perturbing the input coordinates or intermediate features, (Niemeyer &

Geiger, 2021; Wang et al., 2022a) incorporate adversarial objectives to enhance the reconstruction quality, and (Dong et al., 2022) identifies adversarial viewpoints from which the rendered images can fool downstream image classifiers, instead of aiming to degrade NeRF's own accuracy. One concurrent work (Wang et al., 2023) provides an investigation of NeRF's robustness to common image corruptions but adversarial perturbations are not considered. With the growing demand for real-world NeRF-based NVS solutions, it is imperative to understand NeRF's adversarial robustness.

3. Preliminaries of NeRF and GNeRF

NeRF's rendering pipeline. In NeRF, each 2D pixel on the image plane corresponds to a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ emitted from the camera center $\mathbf{o} \in \mathbb{R}^3$, with $\mathbf{d} \in \mathbb{R}^3$ denoting the ray direction and t denoting the ray depth. To render a pixel, a NeRF function f samples points along the corresponding ray and then acquires the color $\mathbf{c}$ and density σ of each point, i.e., $(\sigma, \mathbf{c}) = f(\mathbf{r}(t), \mathbf{d})$. Next, the 2D pixel $\hat{\mathbf{C}}(\mathbf{r})$ can be derived via an integral over the colors of the above-sampled points:

$$\hat{\mathbf{C}}(\mathbf{r}, f) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\, \mathrm{d}t \tag{1}$$

where t_n and t_f are the predefined near and far bounds, respectively, and $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\, \mathrm{d}s\right)$ denotes the accumulated transmittance along the ray from t_n to t. In practice, the integral in Eq. (1) is often approximated with numerical quadrature (Mildenhall et al., 2021). Finally, an MSE loss is applied between the rendered pixels $\hat{\mathbf{C}}(\mathbf{r}, f)$ and the ground truth pixels $\mathbf{C}(\mathbf{r})$ to train NeRF f:

$$\mathcal{L}_{rgb}(\mathcal{R}, f) = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \hat{\mathbf{C}}(\mathbf{r}, f) - \mathbf{C}(\mathbf{r}) \right\|_2^2 \tag{2}$$

where $\mathcal{R}$ is the set of sampled camera rays.

GNeRF's pipeline. On top of vanilla NeRF's pipeline above, GNeRF conditions its function f on the source views as priors of the target new scenes to enable cross-scene generalization. For example, (Wang et al., 2021; Yu et al., 2021b; Wang et al., 2022b; Liu et al., 2022; Chen et al., 2021) adopt a CNN encoder E: $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ to extract features $\{E(\mathbf{I}_i)\}$ from the source views $\{\mathbf{I}_i\}$. Then, each sampled point $\mathbf{x}$ on ray $\mathbf{r}(t)$ is projected to the image plane of each source view through a transformation $\pi_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ to acquire the corresponding scene feature $E(\mathbf{I}_i)[\pi_i(\mathbf{x})]$. Finally, the acquired features $\mathbf{e} = \{E(\mathbf{I}_i)[\pi_i(\mathbf{x})]\}$ serve as extra inputs of vanilla NeRF to derive the density and color $(\sigma, \mathbf{c}) = f(\mathbf{x}, \mathbf{d}, \mathbf{e})$. Different GNeRF variants differ in the ways of constructing the above scene features while following both the volume rendering and the objective in Eq. (1) and Eq. (2), respectively. As introduced in Sec. 4 and Sec. 5, our NeRFool adversarially perturbs the source views $\{\mathbf{I}_i\}$, thereby inducing adversarial features in $\mathbf{e}$.

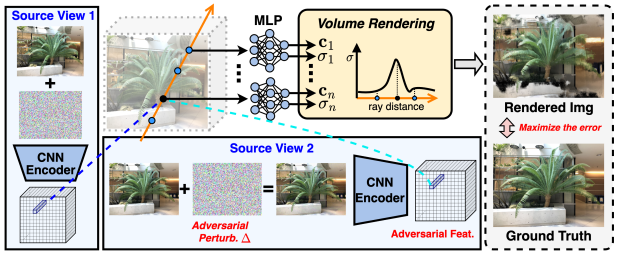

Figure 1. An overview of our NeRFool framework.

4. NeRFool: Uncover GNeRF's Vulnerability

In this section, we present NeRFool, which studies the important properties of GNeRF's vulnerability via our proposed view-specific attack scheme, in which the adversarial perturbations are optimized to fool one specific target view. The insights drawn from NeRFool further inspire our development of NeRFool$^+$ introduced in Sec. 5.

4.1. A View-Specific Method for Attacking GNeRF

Formulation. As visualized in Fig. 1, we inject adversarial perturbations $\mathbf{\Delta} = \{\boldsymbol{\delta}_i\}$ into the source view images $\{\mathbf{I}_i\}$ correspondingly to degrade the reconstruction accuracy of a GNeRF model on one specific target view with a camera pose $\mathbf{P}_{tar} = [\mathbf{R}_{tar} \mid \mathbf{T}_{tar}] \in \mathbb{R}^{3\times4}$, where $\mathbf{R}_{tar} \in \mathbb{R}^{3\times3}$ and $\mathbf{T}_{tar} \in \mathbb{R}^3$ are the rotation and translation with respect to the world coordinate. To optimize $\mathbf{\Delta}$, the goal is to maximize the reconstruction error under a norm constraint $\|\boldsymbol{\delta}_i\|_\infty \leq \epsilon$, with ϵ being sufficiently small to ensure the perturbation's imperceptibility to human eyes. Specifically, the objective can be formulated as:

$$\max_{\forall \boldsymbol{\delta}_i \in \mathbf{\Delta}:\, \|\boldsymbol{\delta}_i\|_\infty \leq \epsilon} \tilde{\mathcal{L}}_{rgb}(\mathcal{R}_{tar}, f, \mathbf{\Delta}) \tag{3}$$

where $\mathcal{R}_{tar}$ is the set of rays sampled from the target view $\mathbf{P}_{tar}$. Although $\mathcal{L}_{rgb}$ in Eq. (2) can provide effective supervision for optimizing $\mathbf{\Delta}$, the ground truth $\mathbf{C}(\mathbf{r})$ for calculating $\mathcal{L}_{rgb}$ may not be available for all target views. Fortunately, we find that the corresponding pseudo ground truth can be obtained in GNeRF by reconstructing the specified target view based on the clean source views. Accordingly, we can modify $\mathcal{L}_{rgb}$ to $\tilde{\mathcal{L}}_{rgb}$ as our objective function, as formulated below:

$$\tilde{\mathcal{L}}_{rgb}(\mathcal{R}, f, \mathbf{\Delta}) = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \hat{\mathbf{C}}(\mathbf{r}, f_\Delta^{adv}) - \hat{\mathbf{C}}(\mathbf{r}, f^{clean}) \right\|_2^2 \tag{4}$$

where $f_\Delta^{adv} = f(\mathbf{x}, \mathbf{d}, \mathbf{e}_\Delta)$ and $f^{clean} = f(\mathbf{x}, \mathbf{d}, \mathbf{e})$ in which $\mathbf{e}_\Delta = \{E(\mathbf{I}_i + \boldsymbol{\delta}_i)[\pi_i(\mathbf{x})]\}$. To solve Eq. (4), we iteratively update $\boldsymbol{\delta}_i$ with gradient ascent using an Adam optimizer, where the t-th iterative step can be formulated as:

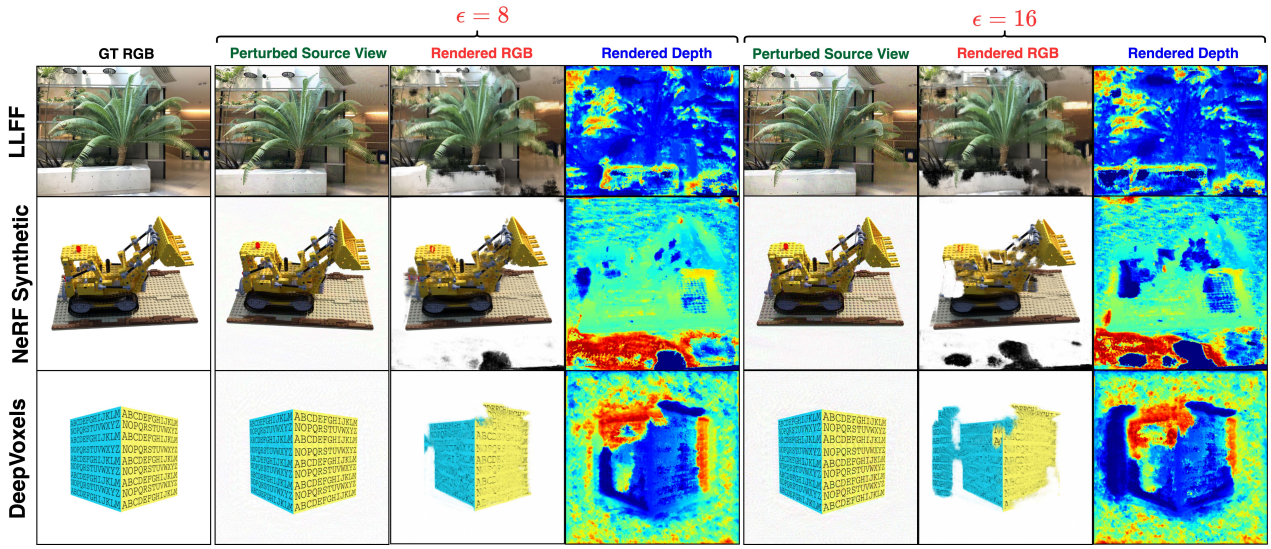

Figure 2. Visualize the ground-truth RGB images, adversarially perturbed source views with imperceptible perturbations which are used to attack IBRNet, and the resulting rendered RGB images/depths on three scenes from three datasets.

$$\boldsymbol{\delta}_i^{(t+1)} = clip(\boldsymbol{\delta}_i^{(t)} + \eta \cdot Adam(\nabla_{\boldsymbol{\delta}_i^{(t)}} \tilde{\mathcal{L}}_{rgb}), -\epsilon, \epsilon) \quad (5)$$

where η is the learning rate and $clip(\cdot, -\epsilon, \epsilon)$ denotes a clipping operation to constrain the norm of $\boldsymbol{\delta}_i$.

Evaluation setting. GNeRF variants: We consider three state-of-the-art (SOTA) GNeRF methods: IBRNet (Wang et al., 2021), MVSNeRF (Chen et al., 2021), and GNT (Wang et al., 2022b), where we adopt their official implementation and load their pretrained models for evaluation. Datasets: We follow the train/test dataset splits adopted by these three GNeRF variants and use both synthetic objects and real scenes from three datasets: three Lambertian objects from DeepVoxels (Sitzmann et al., 2019), eight Realistic Synthetic objects from NeRF (Mildenhall et al., 2020), and eight complex real-world forward-facing scenes from LLFF (Mildenhall et al., 2019). Regarding the source view selection, we follow each GNeRF variant's default scheme, e.g., select the nearby N views around the target view for IBRNet/GNT. NeRFool setup: The learning rate η in Eq. (5) is set to 1e-3 and $\boldsymbol{\delta}_i$ is initialized with a uniform distribution $\mathcal{U}(-\epsilon, \epsilon)$ and then optimized for 500 iterations.

4.2. Is GNeRF Robust to Adversarial Perturbations?

Attack the most representative GNeRF. We first apply the above view-specific attach method to IBRNet (Wang et al., 2021), which serves as a cornerstone for other GNeRF variants, with varied numbers of source views and perturbation strength ϵ. The corresponding quantitative results and qualitative visualization are shown in Tab. 1 and Fig. 2, respectively. We can see that (1) our proposed view-specific attack method can considerably degrade the reconstruction accuracy, e.g., a 10.43/11.74 PSNR reduction on average with ϵ =8/16, respectively, on LLFF; (2) Imperceptible perturbations, which look like random noise caused by camera shake, in the source views can cause serious unrealistic artifacts in the rendered outputs and thus severally degrade users' visual experience; and (3) increasing the number of source views can result in larger PSNR degradation, e.g., a 0.44 larger PSNR reduction when conditioning on ten source views than that of four source views. This indicates that although increased conditionality favors better clean reconstruction accuracy measured on clean source views, it can incur more severe security concerns due to the corresponding higher flexibility (i.e., more pixels) for injecting perturbations. Therefore, in the following experiments, we adopt four source views and $\epsilon = 8$ if not specifically stated.

Attack other GNeRF variants. We further apply NeRFool's attack method to MVSNeRF (Chen et al., 2021) and GNT (Wang et al., 2022b) on the LLFF dataset and report the average metrics across all test scenes. As shown in Tab. 2, we can see that (1) Our NeRFool attack can consistently degrade the reconstruction accuracy across all GNeRF variants and datasets, e.g., a 8.33/11.17 PSNR reduction for MVSNeRF/GNT, respectively, when ϵ=16; (2) GNT can improve the adversarial robustness over IBRNet with reduced PNSR degradation, maybe because of the former's newly introduced transformer modules, which increase the non-linearity of the overall GNeRF pipeline; (3) MVSNeRF wins the highest level of robustness among the three GNeRF variants thanks to its accurate geometry estimation (Chen et al., 2021), which is of great significance for GNeRF's robustness as analyzed in Sec. 4.3.

Key insight. Although GNeRF involves more diverse operations than DNNs, adversarially perturbing its 2D source views can still considerably destruct its reconstructed 3D

Table 1. The achieved rendering quality of IBRNet, which is attacked by NeRFool under different numbers of source views and perturbation strength ϵ, on three datasets. The reported results are averaged across all scenes in each dataset. "Clean" denotes no attack is performed.

Attack Method	No. Source Views	ϵ	LLFF			NeRF Synthetic			DeepVoxels		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Clean	4	-	23.73	0.77	0.24	28.78	0.96	0.06	32.90	0.98	0.03
NeRFool	4	8	13.30	0.53	0.45	12.57	0.82	0.24	12.91	0.76	0.24
NeRFool	4	16	11.99	0.45	0.51	10.71	0.75	0.31	11.85	0.71	0.29
NeRFool	4	32	11.51	0.41	0.54	9.42	0.70	0.36	11.64	0.71	0.30
Clean	6	-	24.52	0.80	0.22	29.18	0.96	0.05	34.08	0.98	0.02
NeRFool	6	8	13.00	0.57	0.42	14.58	0.86	0.21	13.49	0.81	0.21
Clean	10	-	25.13	0.82	0.21	30.00	0.96	0.05	34.57	0.99	0.02
NeRFool	10	8	12.86	0.60	0.39	10.56	0.79	0.26	11.63	0.77	0.27

Table 2. Apply NeRFool on top of other SOTA GNeRF designs.

GNeRF	Attack Method	ϵ	LLFF (Avg.)		
			PSNR ↑	SSIM ↑	LPIPS ↓
MVS -NeRF	Clean	-	23.24	0.78	0.20
	NeRFool	8	16.57	0.36	0.57
	NeRFool	16	14.91	0.22	0.65
GNT	Clean	-	23.66	0.80	0.16
	NeRFool	8	14.28	0.50	0.36
	NeRFool	16	12.49	0.38	0.44

scene rendered from specific views. Furthermore, while increased conditionality on source views in GNeRF can boost its cross-scene generalization capability, it comes at the cost of higher security concerns.

4.3. What to Perturb: Density, Color, or Both?

Considering that NeRF's rendered pixels are alpha-composited from both estimated density and color, one natural question regarding GNeRF's vulnerability is "*which component is easier to be perturbed by adversarial perturbations, density, color, or both*"? We aim to answer this question with the following experiments and discussions.

Setup. We conduct an ablation study on top of IBRNet (Wang et al., 2021) and MVSNeRF (Chen et al., 2021), in which we perform the alpha-composition via (a) clean densities σ_{clean} plus perturbed colors $\mathbf{c}_{adv}$, and (b) perturbed densities σ_{adv} plus clean colors $\mathbf{c}_{clean}$. To implement this setting, we first acquire $(\sigma_{clean}, \mathbf{c}_{clean})$ on top of clean source views and $(\sigma_{adv}, \mathbf{c}_{adv})$ on top of perturbed source views, and next conduct volume rendering using (a) $(\sigma_{clean}, \mathbf{c}_{adv})$ and (b) $(\sigma_{adv}, \mathbf{c}_{clean})$, respectively, based on Eq. (1). We summarize the achieved PSNR in Tab. 3 and visualize both the rendered RGB images and the estimated depth, the latter of which is derived by replacing $\mathbf{c}(t)$ in Eq. (1) with the ray depth t, in Fig. 3 and Fig. 4.

Observation. As shown in Tab. 3, we can observe that (1) using $(\sigma_{adv}, \mathbf{c}_{clean})$ can already considerably degrade the rendering quality, resulting in a comparable PSNR degradation as compared to perturbing both the densities and

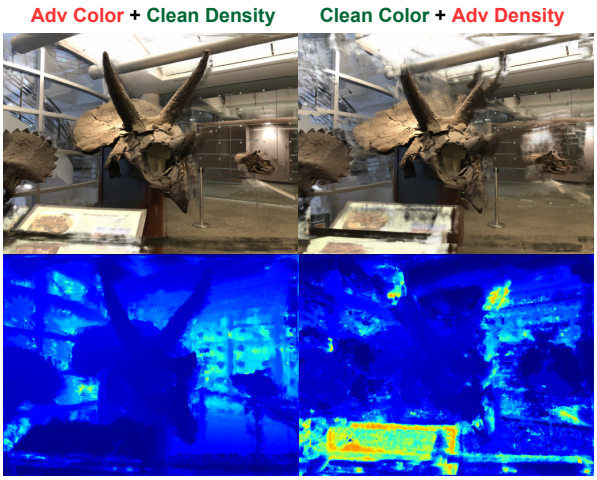

Figure 3. Visualize the attack effectiveness of NeRFool on IBRNet on top of LLFF achieved by perturbing either color or density.

Figure 4. Attack MVSNeRF by perturbing either color or density on two scenes with complex/simple geometry, respectively.

colors. As further verified in Fig. 3, using $(\sigma_{adv}, \mathbf{c}_{clean})$ on top of IBRNet can considerably destruct both the rendered RGB images and depth, where the regions with unrealistic depth estimation cause more severe artifacts in the corresponding RGB images; (2) $(\sigma_{clean}, \mathbf{c}_{adv})$ shows poor attack effectiveness on IBRNet, causing very limited PSNR degradation; (3) $(\sigma_{adv}, \mathbf{c}_{clean})$ can achieve a 0.62 larger PSNR reduction on average over $(\sigma_{clean}, \mathbf{c}_{adv})$ on MVSNeRF and their rankings of attack effectiveness vary across scenes. In particular, as shown in Fig. 4, $(\sigma_{adv}, \mathbf{c}_{clean})$ can result in larger PSNR degradation on scenes with more complex geometry (e.g., the indoor scene "room" in LLFF), where the depth of different objects in a scene varies significantly and thus correctly rendering the RGB images relies more

Table 3. Apply NeRFool on top of IBRNet and MVSNeRF on LLFF via perturbing the color, density, or both. ✓marks the perturbed item.

Method	Color		Density		Achieved PSNR ↑							
	Clean	Adv.	Clean	Adv.	fern	flowers	fortress	horns	leaves	orchids	room	trex
IBRNet	✓		✓		22.22	25.93	28.42	24.39	18.93	18.35	28.68	22.91
	✓			✓	15.77	18.10	16.02	15.07	15.54	12.72	17.56	14.19
		✓	✓		21.77	23.12	27.99	23.15	23.15	16.96	26.74	22.13
		✓		✓	14.02	15.10	13.46	12.41	14.57	11.54	12.96	12.31
MVSNeRF	✓		✓		22.10	25.52	28.21	23.87	18.12	17.92	28.13	22.03
	✓			✓	17.87	20.93	17.22	17.48	14.36	15.73	15.85	16.21
		✓	✓		17.75	20.14	15.44	18.51	14.45	15.57	19.75	18.36
		✓		✓	17.13	19.63	15.95	17.04	13.96	15.31	17.65	15.96

on accurate geometry estimation. Instead, on scenes with relatively simpler geometry (e.g., the "fortress" composed of a table and an object in LLFF), perturbing colors can win better attack effectiveness; (4) perturbing both densities and colors can lead to larger PSNR degradation as compared to only perturbing one factor in 14 out of 16 cases.

Key insight. This set of experiments indicates that (1) adversarial perturbations tend to be more effective in perturbing the density than perturbing the color, especially for scenes with complex geometries, which we conjecture is because the induced wrong geometry estimation of the former can more severely degrade the reconstructed images. This insight could inspire novel attacks (e.g., our NeRFool$^+$ in Sec. 5) and defense methods dedicated to GNeRF; (2) we conjecture that the stronger robustness of IBRNet against perturbed colors over MVSNeRF may originate from a more robust color prediction scheme. In particular, to derive the color of a sampled point, instead of directly regressing the RGB value as in MVSNeRF (Chen et al., 2021), IBRNet projects it to all source views and predicts the weights for blending the RGB values of its projection points on different source views (Wang et al., 2021), which could result in marginal color perturbations when the projection points share similar RGB values. This insight could shed light on the design of more robust GNeRF pipelines.

Table 4. Apply NeRFool on IBRNet w/ and w/o per-scene finetuning. The achieved PSNR on each scene is reported.

Scenes		fern	flower	fortress	horns
w/o ft.	clean	22.22	25.93	28.42	24.39
	adv.	14.02	15.10	13.46	12.41
w/ ft.	clean	24.10	27.13	30.64	27.83
	adv.	13.93	12.29	12.71	11.78

4.4. How Per-Scene Finetuning Impacts the Robustness?

While per-scene finetuning can be adopted on top of GNeRF to enhance the reconstruction accuracy (Wang et al., 2021; Chen et al., 2021; Xu et al., 2022b; Liu et al., 2022), its implication on GNeRF's adversarial robustness is unknown. Here we study the robustness of finetuned GNeRF on different scenes using IBRNet (Wang et al., 2022b).

Table 5. Benchmark view-specific attacks and transferred attacks.

GNeRF	Attack Mode	LLFF (Avg.) PSNR ↑	SSIM ↑	LPIPS ↓
IBRNet	Clean	23.73	0.77	0.24
	Transfer	23.36	0.76	0.27
	View-specific	13.30	0.53	0.45
GNT	Clean	23.66	0.16	0.80
	Transfer	21.92	0.73	0.20
	View-specific	14.28	0.36	0.50

Observation. Tab. 4 shows that while finetuning can boost the clean reconstruction accuracy, the accuracy degradation caused by adversarial perturbations becomes worse.

Key insight. We conjecture this is because the resulting density and color from per-scene finetuned GNeRF are more overfitted to each scene, which could harm the model robustness according to the previous robustness insights for DNNs (Rice et al., 2020). This finding calls for robust per-scene finetuning schemes dedicated to GNeRF to reduce overfitting and maximize robustness.

4.5. Are the Perturbations Transferable across Views?

In real-world settings, it is more practical and highly desirable to reuse the same adversarial perturbations to fool a wide range of target views, under which only the perturbed source views need to be provided to users for conducting effective attacks. To achieve this, it requires that the generated adversarial perturbations can transfer (i.e., remain effective) across target views, motivating the following study.

Setup. We optimize $\boldsymbol{\Delta}$ in Eq. (3) based on a sampled target view centering around each test scene and then reuse it to fool other target views on top of IBRNet.

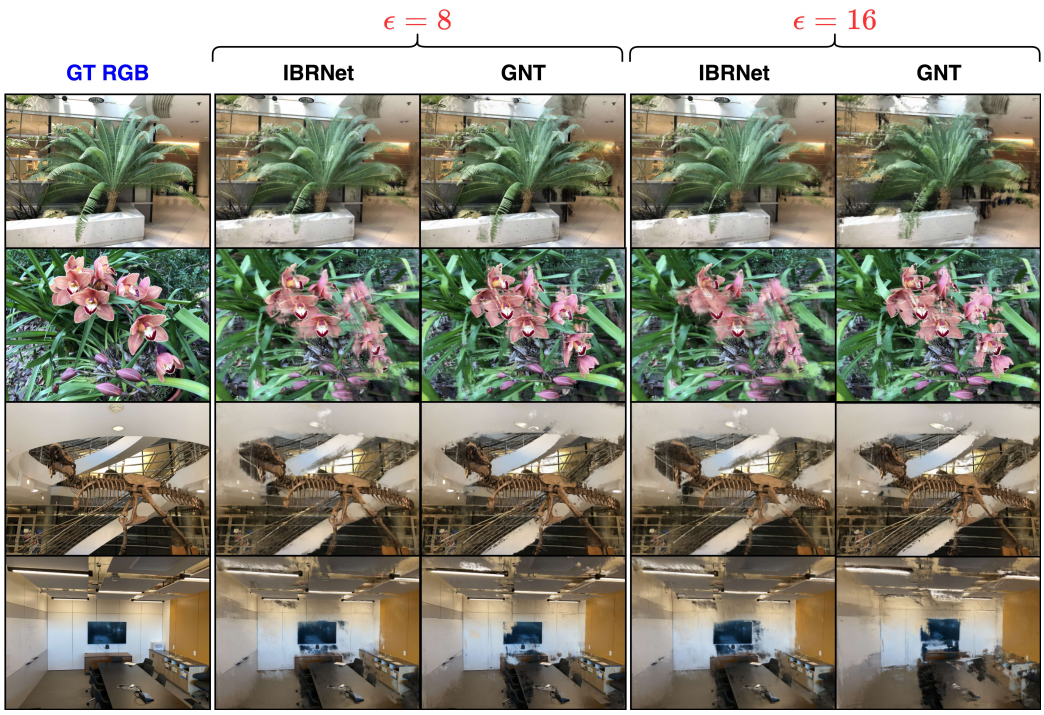

Figure 5. Visualize the attack effectiveness of NeRFool$^+$ on IBRNet/GNT on the scenes from LLFF under different perturbation strengths.

Observation. Tab. 5 shows that under a transferring setting, the achieved PSNR degradation is considerably lower than that achieved by the above view-specific attack.

Key insight. This indicates that $\boldsymbol{\Delta}$ optimized for one specific view can be hardly transferred across different views. We understand that this is because camera rays emitted from one target view can only cover a limited set of 3D points in a scene and thus it is difficult for $\boldsymbol{\Delta}$ optimized for these 3D points to fool other 3D points along the rays emitted from other views with different camera poses.

5. NeRFool$^+$: Towards Universal Adversarial Perturbations Across Different Views

As analyzed in Sec. 4.5, generating universal adversarial perturbations that are transferable across different target views can better uncover GNeRF's vulnerability under a more practical setting and enhance our understanding of their deployability in real-world applications. To achieve this, we propose NeRFool$^+$ which integrates two across-view attack techniques dedicated to GNeRF.

5.1. Overview

Inspirations from the above exploration. Our NeRFool$^+$'s attack techniques are inspired by the following two insights: (1) to enhance the transferability of $\boldsymbol{\Delta}$, more 3D points on the rays of different views for the target 3D scene, are needed to be sampled, according to the analysis in Sec. 4.5; and (2) considering that adversarial perturbations are more effective in perturbing the density/geometry based on our findings in Sec. 4.3, leveraging estimated geometry as extra supervision could enhance the optimization on $\boldsymbol{\Delta}$.

Two optimization techniques. Leveraging the aforementioned insights, our NeRFool$^+$ integrates two optimization techniques correspondingly: (1) to cover more rays and sample more 3D points in a scene, we sample unseen novel target views via spherical linear interpolation among known camera poses to augment the training sets; (2) to better ruin the geometry prediction, we maximize the depth estimation error as extra supervision when optimizing $\boldsymbol{\Delta}$. The technical details are elaborated below.

5.2. Novel Target View Sampling

Sampling strategy. One intuitive sampling strategy is to randomly sample along the upper hemisphere of the target scene. However, this may not be applicable to new scenes where a sampling boundary is hard to define. To develop a sampling strategy generally applicable to new scenes with arbitrary view distributions, we instead randomly interpolate the known camera poses of the source views (and training views if available in the dataset) to create novel views during each training iteration. One advantage is that such a sampling scheme can implicitly define a meaningful range of possible camera poses.

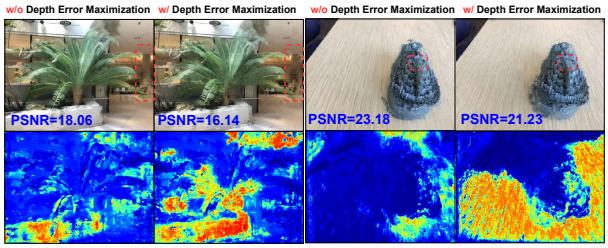

Figure 6. Apply NeRFool$^+$ on LLFF w/ and w/o $\mathcal{L}_{depth}$.

Interpolation strategy. In NeRFool$^+$, given two randomly selected known camera poses $\mathbf{P}_1 = [\mathbf{R}_1 \mid \mathbf{T}_1]$ and $\mathbf{P}_2 = [\mathbf{R}_2 \mid \mathbf{T}_2]$, we interpolate both their rotation matrices and translation vectors to acquire new ones $\mathbf{P}_{new} = [\mathbf{R}_{new} \mid \mathbf{T}_{new}]$. In particular, a linear interpolation is performed for the translation vectors: $\mathbf{T}_{new} = \alpha\mathbf{T}_1 + (1-\alpha)\mathbf{T}_2$, where $\alpha \sim \mathcal{U}(0,1)$. For ensuring meaningful $\mathbf{R}_{new}$ on a unit-radius great circle, we adopt spherical linear interpolation (Slerp) (Shoemake, 1985): $\mathbf{R}_{new} = Slerp(\mathbf{R}_1, \mathbf{R}_2; \alpha)$, following the formulation in (Shoemake, 1985). The pseudo-RGB ground truth is then reconstructed for $\mathbf{P}_{new}$ as in Sec. 4.1 to optimize $\mathbf{\Delta}$.

5.3. Geometric Error Maximization

To better ruin the geometry prediction via extra supervision, we maximize the depth estimation error under the guidance of a pretrained depth estimation model F_D, following (Xu et al., 2022a), which is used to provide geometry priors. In particular, we apply F_D on the reconstructed pseudo-RGB ground truth to generate the depth map as supervision signals, where the objective can be formulated as:

$$\mathcal{L}_{depth} = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \hat{\mathbf{D}}(\mathbf{r}, f^{adv}_{\mathbf{\Delta}}) - F_D(\hat{\mathbf{C}}(\mathbf{r}, f^{clean})) \right\|_2^2 \quad (6)$$

where $\hat{\mathbf{D}}(\mathbf{r}, f) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))t\,\mathrm{d}t$ and $\mathcal{R}$ are sampled from interpolated $\mathbf{P}_{new}$ during each iteration.

5.4. Evaluating NeRFool$^+$

Setup. We adopt the same evaluation as in Sec. 4.1, except that we adopt the same set of source views, which are randomly sampled from nearby views of the target forward-facing scene and then fixed for all runs, for all target view directions. We reuse the pretrained depth estimation model F_D in (Xu et al., 2022a). We adopt four source views and $\epsilon = 8$ by default if not specifically stated.

Observation and analysis. As shown in Tab. 6 and Fig. 5, we can observe that (1) our NeRFool$^+$ can consistently introduce severe artifacts in the reconstructed target views across different GNeRF variants, and considerably degrade the reconstruction PSNR, e.g., an up-to 17.26 lower PSNR on GNT under ϵ=16, as compared to the clean results; (2) our NeRFool$^+$ can induce significantly larger PSNR degradation as compared to the transferred perturbations via NeRFool in Tab. 5, indicating the effectiveness of NeRFool$^+$ in enhancing the attack transferability across target views.

Table 6. The achieved attack effectiveness of NeRFool$^+$ on three GNeRF designs on the scenes from LLFF under different perturbation strengths ϵ. The PSNR on each scene is reported.

GNeRF	ϵ	fern	flower	fortress	horns	leaves	orchids	room	trex
IBRNet	0	22.22	25.93	28.42	24.39	18.93	18.35	28.68	22.91
	8	17.13	14.73	12.04	12.70	13.05	12.48	13.60	13.79
	16	15.97	13.69	11.45	12.04	12.29	11.51	12.77	13.23
GNT	0	22.58	24.93	29.08	25.02	18.8	17.69	28.20	22.96
	8	17.45	15.20	12.81	13.50	13.06	13.10	14.86	13.78
	16	13.38	11.48	11.82	11.91	10.32	11.62	11.96	12.94
MVSNeRF	0	22.10	25.52	28.21	23.87	18.12	17.92	28.13	22.03
	8	16.18	19.84	12.26	17.17	13.51	15.58	17.78	16.52
	16	15.15	17.19	12.05	14.23	12.84	14.60	15.99	13.60

Table 7. Visualize the achieved PSNR of GNT (Wang et al., 2022b) under different pretraining and test scheme combinations. "Pre." denotes the pertaining scheme.

Pre.	Test	fern	flower	fortress	horns
clean	clean	22.58	24.93	29.08	25.02
clean	adv.	15.31	15.90	14.62	14.62
adv.	clean	22.18	22.75	28.14	24.64
adv.	adv.	20.85	19.58	25.01	22.71

The role of $\mathcal{L}_{depth}$. We conduct an ablation study for NeRFool$^+$ w/ and w/o enabling $\mathcal{L}_{depth}$. As shown in Fig. 6, we find that (1) for both NeRFool$^+$ w/ and w/o $\mathcal{L}_{depth}$, the target views farther from the scene, i.e., on the sampling boundary, are harder to be destructed, since 3D points along their emitted rays are less likely to be sampled as compared to those from the center views; and (2) NeRFool$^+$ w/ $\mathcal{L}_{depth}$ can cause larger artifacts, e.g., a 1.92 larger PSNR degradation on fern over NeRFool$^+$ w/o $\mathcal{L}_{depth}$, on the aforementioned target views farther from the scene thanks to the supervision from scene geometry priors.

6. Defend against NeRFool: Adversarial GNeRF Training

Based on the delivered insights from Sec. 4 and Sec. 5, we further perform an intriguing investigation on defending against our NeRFool attack as another crucial piece for understanding GNeRF's robustness.

6.1. Adversarial GNeRF Training: Formulation

We robustify GNeRF via integrating adversarial training (Goodfellow et al., 2014; Shafahi et al., 2019; Wong et al., 2019; Madry et al., 2017), which augments the training scenes with adversarially perturbed source views based on the following formulation:

$$\min_{\boldsymbol{\theta}} \max_{\forall \boldsymbol{\delta}_i \in \mathbf{\Delta}: \|\boldsymbol{\delta}_i\|_\infty \le \epsilon} \tilde{\mathcal{L}}_{rgb}(\mathcal{R}_{target}, f_{\boldsymbol{\theta}}, \mathbf{\Delta}) \quad (7)$$

Table 8. Visualize the achieved PSNR of IBRNet and GNT under different finetuning and test scheme combinations. "Ft." denotes the adopted finetuning scheme. The achieved highest robust/clean PSNR on each scene across all settings is highlighted.

Method	Ft.		Test		Achieved PSNR ↑							
	Clean	Adv.	Clean	Adv.	fern	flowers	fortress	horns	leaves	orchids	room	trex
IBRNet	✓		✓		24.10	27.13	30.64	27.83	21.12	20.32	31.45	25.77
	✓			✓	13.93	12.29	12.71	11.78	13.25	10.35	12.87	13.26
		✓	✓		23.78	27.11	30.01	27.37	20.97	20.16	31.13	25.33
		✓		✓	23.57	26.91	29.29	26.95	20.78	19.91	30.67	25.06
GNT	✓		✓		24.09	29.97	33.24	26.82	23.49	20.13	33.56	27.53
	✓			✓	14.51	13.91	13.31	11.66	10.84	11.98	12.56	10.41
		✓	✓		24.93	30.84	33.67	27.36	24.01	20.61	32.73	27.82
		✓		✓	24.19	27.91	32.94	26.97	23.41	19.97	32.59	27.46

where $\boldsymbol{\theta}$ is the weight of GNeRF and the inner optimization on $\boldsymbol{\delta}_i \in \boldsymbol{\Delta}$ is performed using PGD (Madry et al., 2017). We apply adversarial training to either GNeRF's pretraining or finetuning stages and evaluate the achieved robustness against NeRFool in the following sections.

6.2. Evaluation: Adversarial Pretraining

Setup. We apply the aforementioned adversarial training to GNT's pretraining stage (Wang et al., 2022b) using $\boldsymbol{\epsilon}$=8 and an iteration of 1 for updating $\boldsymbol{\delta}_i$. We then evaluate the resulting models' robustness against NeRFool with four adversarially perturbed source views and $\boldsymbol{\epsilon} = 8$.

Observation and analysis. As shown in Tab. 7, we can observe that (1) adversarial pretraining can effectively boost the adversarial robustness against our NeRFool, e.g., a 10.39 higher PSNR on the scene fortress; (2) the boosted robustness comes at the cost of reduced clean PSNR, e.g., a 0.40∼2.18 PSNR reduction across all scenes, which aligns with previous findings in the literature on adversarial robustness (Zhang et al., 2019). We also note that this set of experiments represents a first-step exploration towards adversarial GNeRF training. Promising future directions include the development of more advanced GNeRF pipelines that can win both accuracy and robustness.

6.3. Evaluation: Adversarial Finetuning

Setup. We apply adversarial training to the finetuning stage of pretrained IBRNet (Wang et al., 2021) and GNT (Wang et al., 2022b) and evaluate the resulting models' robustness against our NeRFool using the same settings in Sec. 6.2.

Observation. As shown in Tab. 8, we can observe that (1) adversarial finetuning can more effectively boost the adversarial robustness as compared to adversarial pretraining reported in Tab. 7, e.g., a 5.20 PSNR improvement averaged over four scenes on top of GNT; (2) adversarial finetuning can maintain a comparable clean PSNR on IBRNet and consistently boost the clean PSNR on GNT across all scenes, e.g., a 0.87 PSNR improvement on the scene flowers.

Key insight. This set of experiments indicates the benign impact of adversarial perturbations beyond robustness. We conjecture that this is because adversarial perturbations can serve as data augmentation to reduce overfitting during finetuning as observed in Sec. 4.4 and thus boost reconstruction accuracy. This aligns with the observations in image classification tasks that properly induced adversarial robustness could boost accuracy (Xie et al., 2020; Deng et al., 2021; Salman et al., 2020). This finding highlights the potential of applying verified training techniques from well-studied image classification tasks to enhance GNeRF optimization.

7. Conclusion

GNeRF has gained increasing attention thanks to its potential in enabling instant and real-time rendering of new scenes, whereas its adversarial robustness has not yet been studied and understood, which can limit its real-world deployment. Our work is the first to uncover and study the adversarial vulnerability of GNeRF. In particular, our proposed NeRFool framework presents systematic analysis and experiments of various GNeRF variants and discovers important insights regarding GNeRF's adversarial robustness. Furthermore, we develop NeRFool$^+$ to effectively attack GNeRF across a wide range of target views and provide rich insights for defending against our developed attacks. Our work has opened a new perspective in the literature of NeRF and could shed light on more robust GNeRF pipelines to empower their real-world deployment.

Acknowledgement

This work was supported in part by CoCoSys, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., and Schwager, M. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022.

Akhtar, N. and Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Arnab, A., Miksik, O., and Torr, P. H. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Carlini, N. and Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pp. 1–7. IEEE, 2018.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., and Wetzstein, G. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. *CVPR*, 2021.

Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., and Su, H. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14124–14133, 2021.

Chen, T., Wang, P., Fan, Z., and Wang, Z. Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15191–15202, 2022.

Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., and Zou, J. Y. Adversarial training helps transfer learning via better representations. *Advances in Neural Information Processing Systems*, 34:25179–25191, 2021.

Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., and Zhu, J. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *arXiv preprint arXiv:2210.03895*, 2022.

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Fu, X., Zhang, S., Chen, T., Lu, Y., Zhu, L., Zhou, X., Geiger, A., and Liao, Y. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022.

Garbin, S. J., Kowalski, M., Johnson, M., Shotton, J., and Valentin, J. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14346–14355, 2021.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Guo, C., Rana, M., Cisse, M., and Van Der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Hedman, P., Philip, J., Price, T., Frahm, J.-M., Drettakis, G., and Brostow, G. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics*, 2018.

Kundu, A., Genova, K., Yin, X., Fathi, A., Pantofaru, C., Guibas, L. J., Tagliasacchi, A., Dellaert, F., and Funkhouser, T. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12871–12881, 2022.

Li, Z., Niklaus, S., Snavely, N., and Wang, O. Neural scene flow fields for space-time view synthesis of dynamic scenes. *CVPR*, 2021.

Lindell, D. B., Martel, J. N., and Wetzstein, G. Autoint: Automatic integration for fast neural rendering. *CVPR*, 2021.

Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., and Wang, W. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7824–7833, 2022.

Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., and Sheikh, Y. Neural volumes: Learning dynamic renderable volumes from images. *SIGGRAPH*, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Maggio, D., Abate, M., Shi, J., Mario, C., and Carlone, L. Loc-nerf: Monte carlo localization using neural radiance fields. *arXiv preprint arXiv:2209.09050*, 2022.

Mildenhall, B., Srinivasan, P. P., Ortiz-Cayon, R., Kalantari, N. K., Ramamoorthi, R., Ng, R., and Kar, A. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

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

Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., and de La Fortelle, A. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pp. 1347–1356. PMLR, 2022.

Niemeyer, M. and Geiger, A. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11453–11464, 2021.

Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S., Geiger, A., and Radwan, N. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5480–5490, 2022.

Ost, J., Mannan, F., Thuerey, N., Knodt, J., and Heide, F. Neural scene graphs for dynamic scenes. *CVPR*, 2021.

Rebain, D., Jiang, W., Yazdani, S., Li, K., Yi, K. M., and Tagliasacchi, A. DeRF: Decomposed radiance fields. *CVPR*, 2021.

Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., and Novotny, D. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10901–10911, 2021.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

Schwarz, K., Liao, Y., Niemeyer, M., and Geiger, A. GRAF: Generative radiance fields for 3D-aware image synthesis. *NeurIPS*, 2020.

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

Shoemake, K. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

Siddiqui, Y., Porzi, L., Bulò, S. R., Müller, N., Nießner, M., Dai, A., and Kontschieder, P. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9043–9052, 2023.

Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., and Zollhofer, M. Deepvoxels: Learning persistent 3d feature embeddings. *CVPR*, 2019.

Srinivasan, P. P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., and Barron, J. T. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. *CVPR*, 2021.

Thies, J., Zollhöfer, M., and Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 2019.

Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J. T., and Srinivasan, P. P. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5481–5490. IEEE, 2022.

Wang, C., Chai, M., He, M., Chen, D., and Liao, J. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3835–3844, 2022a.

Wang, C., Wang, A., Li, J., Yuille, A., and Xie, C. Benchmarking robustness in neural radiance fields. *arXiv preprint arXiv:2301.04075*, 2023.

Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022b.

Wang, Q., Wang, Z., Genova, K., Srinivasan, P. P., Zhou, H., Barron, J. T., Martin-Brualla, R., Snavely, N., and Funkhouser, T. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2021.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., and Le, Q. V. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 819–828, 2020.

Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H., and Wang, Z. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022a.

Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., and Neumann, U. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5438–5448, 2022b.

Yu, A., Li, R., Tancik, M., Li, H., Ng, R., and Kanazawa, A. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021a.

Yu, A., Ye, V., Tancik, M., and Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021b.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.