

AegisRF: Adversarial Perturbations Guided with Sensitivity for Protecting Intellectual Property of Neural Radiance Fields

Woo Jae Kim

wkim97@kaist.ac.kr

Kyu Beom Han

kphan@kaist.ac.kr

Yoonki Cho

yoonki@kaist.ac.kr

Youngju Na

yjna2907@kaist.ac.kr

Junsik Jung

junsik.jung@kaist.ac.kr

Sooel Son

s.son@kaist.ac.kr

Sung-eui Yoon

sungeui@kaist.edu

School of Computing,

Korea Advanced Institute of Science
and Technology (KAIST),

Daejeon, Korea

Abstract

As Neural Radiance Fields (NeRFs) have emerged as a powerful tool for 3D scene representation and novel view synthesis, protecting their intellectual property (IP) from unauthorized use is becoming increasingly crucial. In this work, we aim to protect the IP of NeRFs by injecting adversarial perturbations that disrupt their unauthorized applications. However, perturbing the 3D geometry of NeRFs can easily deform the underlying scene structure and thus substantially degrade the rendering quality, which has led existing attempts to avoid geometric perturbations or restrict them to explicit spaces like meshes. To overcome this limitation, we introduce a *learnable sensitivity* to quantify the spatially varying impact of geometric perturbations on rendering quality. Building upon this, we propose AegisRF, a novel framework that consists of a Perturbation Field, which injects adversarial perturbations into the pre-rendering outputs (color and volume density) of NeRF models to fool an unauthorized downstream target model, and a Sensitivity Field, which learns the sensitivity to adaptively constrain geometric perturbations, preserving rendering quality while disrupting unauthorized use. Our experimental evaluations demonstrate the generalized applicability of AegisRF across diverse downstream tasks and modalities, including multi-view image classification and voxel-based 3D localization, while maintaining high visual fidelity. Codes are available at <https://github.com/wkim97/AegisRF>.

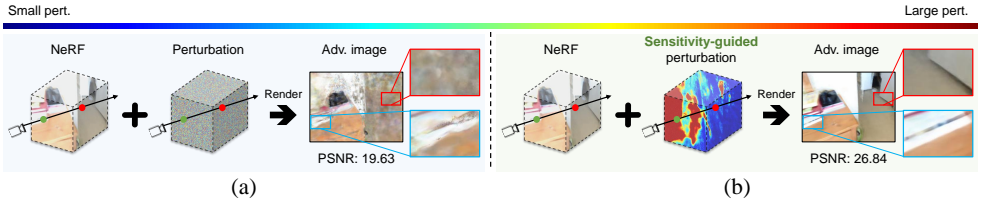


Figure 1: (a) Geometric perturbations without consideration of their varying impact on rendering quality lead to significant degradation in rendering quality. (b) Our novel approach mitigates this by measuring the sensitivity of rendering quality to geometric perturbations and adaptively constraining their magnitudes. For example, perturbations are restricted on empty spaces (red point), where disruptions can cause introduction of new artifacts, while larger perturbations are applied on more complex regions (green point), where such disruptions can be better masked by the existing structural complexity.

1 Introduction

Neural Radiance Fields (NeRFs) [15] have emerged as a powerful paradigm for novel view synthesis and 3D scene representation, finding applications in AR/VR [62, 68], autonomous driving [13, 62], and the metaverse [11, 72]. Their implicit representation, which yields pre-rendering outputs (*i.e.*, color, volume density) for any queried 3D point and viewing direction, has enabled scene representation in diverse data modalities such as images, voxels, and point clouds, thus driving advancements in diverse downstream 3D perception tasks [9, 16, 68, 64, 66]. However, their broad applicability entails a new challenge: protecting their intellectual property (IP) from unauthorized use. Given substantial resources required to construct radiance fields [9, 14, 27, 46, 48, 69] and their value as a versatile 3D data source [11, 62, 68], their unauthorized use in these downstream tasks can lead to significant losses in resources and revenues, making IP protection an urgent concern [18, 20, 42, 67].

Inspired by the recent success of adversarial perturbations for IP and privacy protection in text [24, 25, 64], audio [9, 60], and image [6, 22, 45, 63, 64, 65, 66] domains, we propose to tackle this challenge by obstructing downstream models that attempt to exploit NeRFs without proper authorization. We achieve this by introducing carefully crafted adversarial perturbations into the NeRF’s pre-rendering outputs (color and density) designed to undermine an unauthorized target downstream model while preserving the visual quality. By injecting these perturbations at inference time, our approach can generate adversarial examples across diverse data modalities derived from NeRFs, thus offering a generalized protection framework applicable to a wide range of downstream tasks.

However, perturbing the 3D geometry of NeRFs is inherently challenging, as it can easily deform the underlying scene structure and substantially degrade rendering quality [19, 20, 42]. We argue that this degradation primarily occurs because the visual impact of a geometric perturbation is not uniform across the 3D space; rather, it heavily depends on the specific 3D location and the local structural details [30, 41, 74]. For example, as shown in Fig. 1(a), applying perturbations of random magnitudes across the 3D space without considering their spatially varying impact can indiscriminately alter geometric structures and thus significantly degrade the rendering quality. While there have been recent attempts to inject adversarial perturbations into NeRFs [19, 76], they also overlook this insight, either imposing no explicit constraints on perturbations [19] or applying predetermined, fixed constraints across 3D space [76]. Consequently, to prevent visual distortions, these methods resort to

either avoiding geometric perturbations altogether [26] or restricting them to explicit forms like meshes [19], thereby limiting their applicability across the diverse 3D downstream tasks that leverage NeRF’s pre-rendering outputs.

To overcome this limitation when perturbing the geometry of NeRFs, we introduce a *learnable sensitivity* measure for quantifying the varying degree of impact that geometric perturbations across 3D space have on rendering quality. As shown in Fig. 1(b), this allows us to impose tighter constraints on geometric perturbations in regions where changes would be highly perceptible, while permitting more perturbations in areas where they cause less perceptual degradation. This sensitivity-guided strategy aims to maximize the disruptive effect on unauthorized tasks while minimizing degradation in rendering quality.

Based on this insight, we propose AegisRF, a novel framework for protecting the IP of NeRF models by injecting sensitivity-guided perturbations into their pre-rendering outputs at inference time. AegisRF consists of two key components: (1) Perturbation Field that generates appearance (color) and geometry (density) perturbations designed to fool an unauthorized downstream target model, and (2) Sensitivity Field that quantifies the sensitivity of the rendering quality to geometric perturbations across 3D space and constrains these perturbations adaptively. Our experimental evaluations demonstrate that AegisRF offers robust protection across diverse downstream tasks, including multi-view image classification [9] and voxel-based 3D localization [16], while preserving the rendering quality (Sec. 4.2). Additionally, through extensive analysis on the Sensitivity Field (Sec. 4.3), we verify that guiding the perturbations with the sensitivity is crucial to maintain high rendering quality.

In summary, our contributions are as follow:

- We introduce learnable sensitivity for measuring the perceptual impact of geometric perturbations on rendering quality, which enables adaptive constraints for robust NeRF IP protection while maintaining high visual fidelity.
- We present AegisRF, a novel framework for NeRF IP protection that operates at inference time by perturbing both the appearance and geometry pre-rendering outputs of NeRFs, thus protecting their IP from a diverse range of downstream tasks and modalities.
- Through empirical evaluations, we verify the effectiveness of AegisRF at substantially undermining the performance of various downstream applications with different modalities while preserving the rendering quality.

2 Related Works

Neural Radiance Fields. Neural radiance fields (NeRFs) [45] and other MLP-based radiance fields [8, 24, 46, 49, 51, 52], implicitly represent 3D scenes as continuous signals. These methods have achieved photorealistic novel view synthesis and have become a powerful 3D representation [8, 46, 47, 49, 51, 52]. Their implicit nature allows transformation of their pre-rendering outputs into various data types (e.g., images, voxels) [9, 16, 23, 66], fueling diverse downstream perception tasks such as classification [9, 23, 26], segmentation [9, 8, 51, 58, 64], and localization [16, 66]. As commercial NeRF applications also grow [50], protecting their IP from unauthorized downstream use becomes critical. This work safeguards IP of NeRFs across various downstream tasks and modalities by applying adversarial perturbations to pre-rendering outputs while ensuring visual integrity.

IP Protection via Model Deception. Protecting data IP or privacy by subverting downstream applications with imperceptible input perturbations has become a popular strategy.

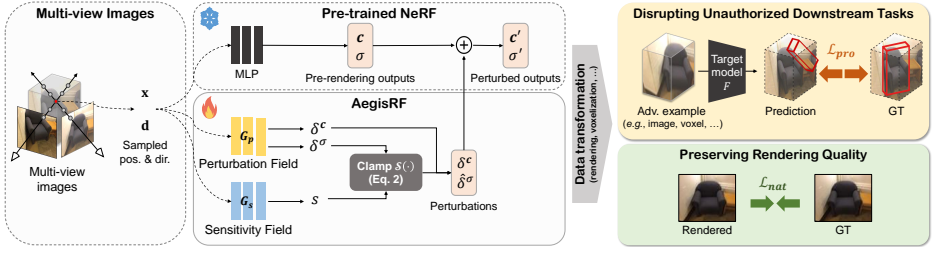


Figure 2: Overview of AegisRF. For a 3D point (\mathbf{x}, \mathbf{d}) , the Perturbation Field creates appearance (δ^c) and geometry (δ^σ) perturbations, while the Sensitivity Field predicts sensitivity (s) to adaptively constrain geometry perturbations $(\hat{\delta}^\sigma)$. These perturb NeRF outputs (\mathbf{c}, σ) into perturbed versions (\mathbf{c}', σ') , forming adversarial examples in various data forms, aiming to disrupt target unauthorized downstream task $(\mathcal{L}_{\text{pro}})$ while preserving rendering quality $(\mathcal{L}_{\text{nat}})$.

This includes deceiving facial recognition [6, 64, 65], preventing unauthorized manipulation or style extraction from generative models [22, 65, 63, 66], and protecting text or audio data [9, 24, 25, 64, 60]. We extend this paradigm of IP protection through model deception to NeRFs [45], aiming to thwart unauthorized use by designing adversarial attacks targeting their downstream applications.

3D Adversarial Attack. Since Athalye *et al.* [10] first proposed 3D adversarial examples, many attacks have targeted diverse data forms like point clouds [12, 17, 66, 63, 65, 67, 73, 75], meshes [63, 67, 68, 70], and voxel-grids [69]. Recently, the rise of radiance fields [45, 46, 62] spurred NeRF-specific attacks; TT3D [14] adversarially fine-tunes radiance fields [46], and NeRFail [26] perturbs the color attributes of NeRFs from transformed 2D pixels. However, these works overlook the sensitivity of geometry to perturbations, thus avoiding [26] or limiting geometric perturbations to vertex coordinates on the explicit mesh space [14] to avoid significant deformations. In contrast, our sensitivity-guided approach provides a spatially-aware understanding of how geometric perturbations affect rendering quality, thereby enabling direct perturbations on NeRF’s implicit geometry for versatile IP protection across different downstream tasks and modalities while preserving rendering quality.

3 AegisRF

3.1 Preliminaries

Neural Radiance Fields (NeRF). NeRF [45] models a 3D scene using an MLP that maps a 3D coordinate $\mathbf{x} \in \mathbb{R}^3$ and view direction $\mathbf{d} \in \mathbb{R}^3$ to pre-rendering outputs: RGB color \mathbf{c} and volume density σ , *i.e.*, $\text{MLP}(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$. Images are rendered by volumetrically integrating these outputs along rays $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ emitted from the camera center \mathbf{o} . For N points sampled along each ray, the pixel color $I(\mathbf{r})$ is computed as:

$$I(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \Delta t_i)) \mathbf{c}_i, \quad (1)$$

where $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \Delta t_j\right)$ is accumulated transmittance, $\Delta t_i = t_{i+1} - t_i$ is the distance between consecutive sampled points, and \mathbf{c}_i, σ_i are the color and density at the i -th sample.

3.2 Perturbation and Sensitivity Fields

We tackle the research question of protecting IP in NeRFs from adversaries who use these radiance fields without proper authorization. To achieve this, we propose AegisRF (Fig. 2), a framework that undermines the adversary’s downstream task by injecting imperceptible perturbations into the pre-rendering outputs of NeRFs, or the appearance (*i.e.*, color \mathbf{c}) and the geometry (*i.e.*, density σ). AegisRF consists of two components: (1) Perturbation Field, which models adversarial perturbations designed to undermine the target unauthorized downstream model, and (2) Sensitivity Field, which constrains the geometric perturbations based on the degree of impact they have on the rendering quality degradation.

Perturbation Field. To protect NeRFs from unauthorized usage, we introduce the *Perturbation Field* (Fig. 2) that perturbs their color \mathbf{c} and density σ outputs by injecting appearance perturbation δ^c and geometry perturbation δ^σ such that $\mathbf{c}' = \mathbf{c} + \delta^c$ and $\sigma' = \sigma + \delta^\sigma$. We model the Perturbation Field via a set of multilayer perceptron (MLP) networks G_p to learn δ^c and δ^σ given a point with position \mathbf{x} and viewing direction \mathbf{d} from a camera such that $G_p : (\mathbf{x}, \mathbf{d}) \rightarrow (\delta^c, \delta^\sigma)$. These perturbations are applied to the pre-rendering outputs of a frozen pre-trained NeRF at inference time. This allows our perturbations to be easily activated or deactivated depending on the deployment context, enabling seamless switching between protected and unprotected inference modes. More details on the architecture of G_p and sampling methods are provided in the appendix (Sec. A).

The implicit design of our Perturbation Field allows us to craft perturbations for any queried 3D point and view, making it highly compatible with a wide range of MLP-based radiance fields. These perturbations injected directly into the pre-rendering outputs of NeRFs can also be transformed into adversarial examples of diverse modalities, such as voxel grids [16, 64, 66] or images [0, 23]. Such versatility allows application of AegisRF across diverse unauthorized downstream models (Sec. 4.2), providing generalized IP protection.

Sensitivity Field. Unlike appearance, even small disruptions on the 3D geometry can lead to significant shape deformations and thus degrade the rendering quality [19, 20, 42]. To this end, we design the *Sensitivity Field* that measures the sensitivity of rendering quality to geometric perturbations across the 3D space and impose constraints accordingly. We approximate the Sensitivity Field via a set of MLPs G_s that takes the position \mathbf{x} of a point and outputs a sigmoid-normalized scalar sensitivity $s \in [0, 1]$ such that $G_s : \mathbf{x} \rightarrow s$ as shown in Fig. 2. Given the sensitivity s , we constrain the magnitudes of the geometry perturbation, thus applying a tighter constraint on regions that are sensitive to rendering quality degradation while allowing larger perturbations on regions with lower sensitivity scores.

While a common strategy for such constraints would be clipping the perturbation as commonly done in traditional norm-bounded attacks [10, 28, 29, 43], this could prevent the gradient flow and interfere with the training of Perturbation and Sensitivity Fields. Thus, we use the soft clamping strategy $\mathcal{S}(\cdot, \cdot)$ to obtain constrained geometric perturbation $\hat{\delta}^\sigma$ while ensuring their proper training as follows:

$$\hat{\delta}^\sigma = \mathcal{S}(\delta^\sigma, s) = ((1 - s) \cdot \bar{\sigma}) \cdot \tanh\left(\frac{\delta^\sigma}{(1 - s) \cdot \bar{\sigma}}\right), \quad (2)$$

where $\bar{\sigma}$ is the mean density value for all points uniformly sampled from a 3D grid that covers the entire scene volume. With this soft clamping operation, adversarial color \mathbf{c}' and density σ' can now be written as $\mathbf{c}' = \mathbf{c} + \delta^c$ and $\sigma' = \sigma + \hat{\delta}^\sigma$.

Traditional norm-bounded image attacks [10, 43] use fixed perturbation magnitudes, which overlook the varying perceptibility of geometric perturbations in 3D space. In con-

trast, our method adapts constraints based on the 3D location, applying tighter limits in sensitive areas and allowing larger perturbations in less sensitive regions to ensure imperceptibility compared to fixed constraints (Sec. 4.3).

3.3 Training AegisRF

We train AegisRF using two objectives: (1) protection loss, which guides the generation of strong perturbations that can undermine the target unauthorized downstream task, and (2) naturalness loss, which aims to preserve rendering quality in the presence of perturbations.

Protection loss \mathcal{L}_{pro} . The primary objective of the protection loss is to guide the Perturbation Field G_p to produce perturbations that effectively undermine a target unauthorized downstream model F . To achieve this, we first generate an adversarial example x' , which serves as the input to F . Thanks to the versatile characteristic of the pre-rendering outputs of NeRF, the adversarial example x' can also be generated in different modalities from the perturbed pre-rendering outputs (\mathbf{c}', σ') depending on the downstream model. For instance, if F is an image-based model [4, 26], x' is an RGB image produced by volumetric rendering (Eq. 1) of (\mathbf{c}', σ') , and if F is a voxel-based model [16, 66], x' is a 3D voxel grid constructed via uniform sampling and aggregation [16, 66], where each point in grid captures (\mathbf{c}', σ') . This versatility in forming x' allows our approach to target a wide range of downstream tasks that utilize various data modalities derivable from NeRFs.

The protection loss \mathcal{L}_{pro} is defined to maximize the training loss \mathcal{L}_F of the target model F (e.g., cross-entropy for classification) as follows:

$$\mathcal{L}_{\text{pro}} = -\mathcal{L}_F(F(x'), y_{\text{gt}}), \quad (3)$$

where $F(x')$ is the prediction from the target downstream model, and y_{gt} is the corresponding ground-truth label (e.g., a class label for classification). In this way, the Perturbation Field learns to generate perturbations $(\delta^c, \delta^\sigma)$, which are then used to craft perturbed outputs (\mathbf{c}', σ') and ultimately the adversarial example x' that is highly effective at degrading the performance of unauthorized models.

Naturalness loss \mathcal{L}_{nat} . To maintain rendering quality against perturbations, our naturalness loss \mathcal{L}_{nat} measures the photometric L_2 difference [45] between a rendering output I'_n derived from perturbed color \mathbf{c}' and density σ' (Eq. 1) and its corresponding ground truth I_n^{gt} . We compute the loss over a set of camera rays \mathcal{R} sampled from the training views:

$$\mathcal{L}_{\text{nat}} = \sum_{\mathbf{r} \in \mathcal{R}} \| I'_n(\mathbf{r}) - I_n^{\text{gt}}(\mathbf{r}) \|_2^2. \quad (4)$$

This term penalizes rendering quality degradation and supervises both the Perturbation Field G_p and the Sensitivity Field G_s . In this way, the Perturbation Field G_p learns to refrain from generating perturbations $(\delta^c, \delta^\sigma)$ that result in severe degradation in rendering quality. This term also guides the Sensitivity Field G_s to predict high sensitivity values on regions where geometric perturbations would significantly degrade rendering quality, and low sensitivity to regions where perturbations are perceptually tolerable, thus serving our purpose of crafting effective geometric perturbations while preserving the rendering quality.

Model training. During training, the pre-trained NeRF remains frozen while our Perturbation and Sensitivity Fields are updated on the combination of the two losses:

$$\mathcal{L} = \lambda_{\text{pro}} \cdot \mathcal{L}_{\text{pro}} + \lambda_{\text{nat}} \cdot \mathcal{L}_{\text{nat}}, \quad (5)$$

where λ_{pro} and λ_{nat} denote the coefficients of \mathcal{L}_{pro} and \mathcal{L}_{nat} , respectively.

Multi-view classification							
Method	Naturalness			Disruption efficacy			
	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	Acc. (%)			
No protection	28.28	0.9205	0.0979	99.00			
NeRFail-S	22.42	0.8360	0.1508	2.00			
NeRFail	23.74	0.8751	0.1315	7.75			
Adv-FT ($\lambda_{\text{pro}} = 0.0003$)	<u>25.57</u>	<u>0.8973</u>	<u>0.1372</u>	5.50			
Adv-FT ($\lambda_{\text{pro}} = 0.003$)	23.67	0.8722	0.1987	2.13			
Adv-FT ($\lambda_{\text{pro}} = 0.03$)	17.62	0.7442	0.5220	0.13			
AegisRF (Ours)	26.33	0.9042	0.1271	<u>1.88</u>			

3D localization							
Method	Naturalness			Disruption efficacy			
	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	Recall ₂₅ (%)	Recall ₅₀ (%)	AP ₂₅ (%)	AP ₅₀ (%)
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
Adv-FT ($\lambda_{\text{pro}} = 0.1$)	<u>22.37</u>	<u>0.7249</u>	<u>0.4289</u>	66.58	15.86	14.93	2.52
Adv-FT ($\lambda_{\text{pro}} = 1$)	18.44	0.6068	0.6013	50.82	6.88	3.83	0.22
Adv-FT ($\lambda_{\text{pro}} = 10$)	15.10	0.5264	0.7147	43.43	2.25	1.60	0.01
AegisRF (Ours)	23.97	0.7687	0.3565	<u>48.64</u>	<u>4.77</u>	<u>3.24</u>	<u>0.20</u>

Table 1: Comparison of our AegisRF with existing methods on multi-view image classification and 3D localization. Best results are in **bold**, and second best results are underlined.

4 Experiments

4.1 Experimental Setups

Dataset and evaluation metrics. We use 8 scenes from the NeRF Synthetic dataset [45] for multi-view classification and 10 scenes from the ScanNet [8] dataset for 3D localization. We use the train/test view splits set by the target downstream methods. We evaluate naturalness with PSNR, SSIM, and LPIPS [71] of rendered images [8, 45, 60]. We also assess disruptive efficacy on downstream task, *i.e.*, the protective performance of AegisRF. For multi-view classification [4], we measure prediction accuracy on images rendered on test views. For 3D localization [16], we measure Recall@K and AP@K of predicted bounding boxes.

Target models and baseline methods. We evaluate our AegisRF on two MLP-based radiance fields and two representative downstream tasks: (1) multi-view classification (ViT-B/16 [4] on NeRF [45]) and (2) voxel-based 3D localization (NeRF-RPN [16] with Swin Transformer [69] on depth-guided NeRF [62]). As a baseline, we consider TT3D [19], a method that adversarially fine-tunes parameters of a pre-trained radiance field. However, while TT3D modifies the appearance parameters of the radiance field, it alters geometry on the explicit mesh derived from the radiance field. Since our goal is to protect the NeRF model itself, which can form diverse data modalities, TT3D’s focus on a single derived mesh is not directly suitable. Thus, we adapt TT3D’s core idea of adversarially fine-tuning into Adv-FT. This approach fine-tunes the parameters of the pre-trained NeRF to generate both the adversarial color and density representations via the protection loss (Eq. 3) and the naturalness loss (Eq. 4). For multi-view classification, we also compare with NeRFail [26] and NeRFail-S [26], which are specifically designed for multi-view adversarial attacks.

Implementation details. For AegisRF, we set protection loss weight $\lambda_{\text{pro}} = 0.0003$, naturalness loss weight $\lambda_{\text{nat}} = 1$ for multi-view classification and $\lambda_{\text{pro}} = 1$, $\lambda_{\text{nat}} = 50$ for 3D localization. For Adv-FT, we set $\lambda_{\text{nat}} = 1$ for multi-view classification and $\lambda_{\text{nat}} = 50$ for 3D localization, and use various values of λ_{pro} for extensive comparisons over its performance spectrum. For NeRFail [26], we used hyperparameters set by the authors and $\varepsilon = 32$. Please refer to Sec. A of the supplementary for additional implementation details.

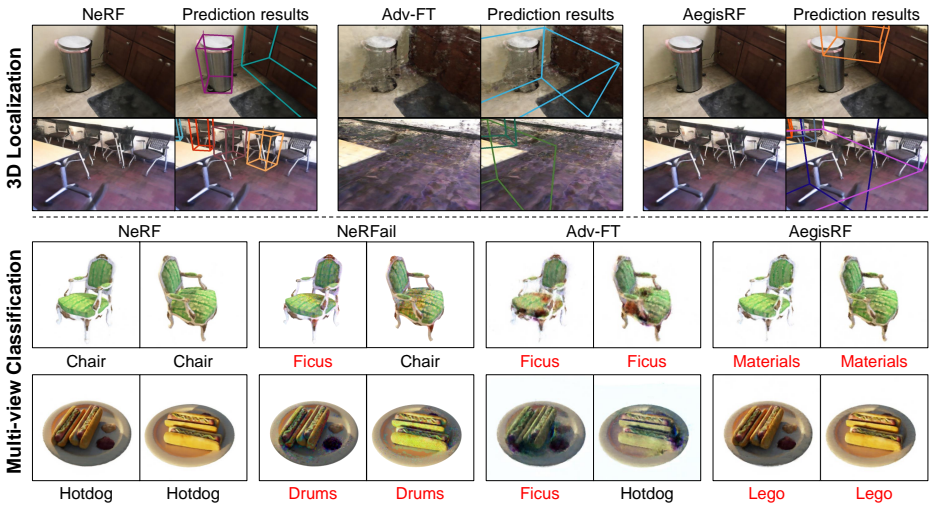


Figure 3: Visualizations of rendered images and model predictions on NeRF, NeRFail, Adv-FT ($\lambda_{\text{pro}} = 1$ for 3D localization, $\lambda_{\text{pro}} = 0.003$ for multi-view classification), and our AegisRF. Our AegisRF shows superior rendering quality compared to NeRFail and Adv-FT.

4.2 Naturalness and Protection Performance

In Table 1, we report the naturalness and disruption efficacy of our AegisRF along with the compared methods. Compared to NeRFail [26], AegisRF achieves both higher visual quality (+2.59 PSNR) and stronger protection (-5.87% accuracy) in multi-view classification. This is thanks to our effective geometric perturbations, while NeRFail limits perturbations only to appearance components of NeRFs. We can also observe that Adv-FT faces a large trade-off between the naturalness and disruption efficacy, achieving high disruption efficacy (e.g., 1.60% AP₂₅ in 3D localization with $\lambda_{\text{pro}} = 10$) at a cost of significantly degraded rendering quality (15.10 PSNR). Conversely, if Adv-FT prioritizes naturalness (22.37 PSNR with $\lambda_{\text{pro}} = 0.1$), this sacrifices its disruption efficacy (14.93% AP₂₅). In contrast, our AegisRF achieves strong disruption efficacy (3.24% AP₂₅) with high visual quality (23.97 PSNR) thanks to the sensitivity guidance, which adaptively constrains geometric perturbations to preserve visual fidelity while effectively disrupting downstream models.

We further evaluate AegisRF in a more practical scenario by evaluating its transferability across different architectures in Table 2. AegisRF demonstrates considerable transferability, reducing the multi-view classification accuracy of Swin-B to 6.00% when trained to disrupt Mixer-B, highlighting the practical effectiveness of AegisRF in real-world IP protection scenarios. In the supplementary, we report transferability on 3D localization (Sec. B), robustness against common transformations such as JPEG compression or Gaussian noise (Sec. C), and computational costs (Sec. E).

Surrogate Model	Target Accuracy (%)		
	ViT-B	Swin-B	Mixer-B
No Protection	99.00	99.88	90.75
ViT-B	1.88	41.00	44.75
Swin-B	19.63	0.00	18.75
Mixer-B	7.25	6.00	0.50

Table 2: Transferability of AegisRF from a surrogate model to unknown target models in multi-view classification.

In Fig. 3, we visualize images rendered from the original NeRF, existing methods, and our AegisRF along with their disruption efficacy. While Adv-FT shows strong protective ability, it results in significantly distorted rendering qualities, often showing visible distortions.

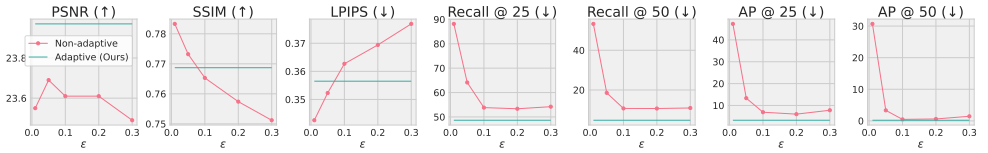


Figure 4: Comparison of naturalness (col. 1-3) and disruption efficacy (col. 4-7) of our sensitivity-guided adaptive approach with a fixed, non-adaptive perturbation bound ϵ . Our strategy leads to the best balance between naturalness and protection performance.

tions in the geometry (e.g., removal of chairs in row 2). In contrast, AegisRF can preserve the visual fidelity while undermining the predictions of the target models, thanks to the Perturbation and Sensitivity Fields, which balance adversarial strength with imperceptibility.

4.3 Analysis on Sensitivity Field

Sensitivity vs. fixed bound. We compare Sensitivity Field with constraining geometric perturbations using a fixed, non-adaptive perturbation bound ϵ commonly used in traditional norm-bounded adversarial examples [10, 43]. As shown in Fig. 4, our sensitivity-guided adaptive approach outperforms all of the non-adaptive cases in terms of disruption efficacy. While lower ϵ improves naturalness of non-adaptive approach, this comes at significantly degraded disruption efficacy. In contrast, thanks to the sensitivity-guided approach that suppresses perturbations detrimental to rendering quality while perturbing less visually critical points, our method achieves a better balance between the naturalness and disruption efficacy.

Visualization of sensitivity. In Fig. 5, we visualize the images rendered from the original NeRF (col. 1) and AegisRF (col. 2) along with the pixel-wise perturbations (col. 3). The perturbations show that our AegisRF tends to affect the edges of more complex surfaces, which is consistent with prior findings that distortions on high-frequency textures are generally less perceptible than those on simpler regions [30, 41, 74]. We also visualize the sensitivity predicted by Sensitivity Field by averaging values on sampled points along each ray (col. 4). Sensitivity Field tends to learn lower sensitivity on regions containing objects with complex textures (e.g., chairs), suggesting an implicit alignment with observations from previous studies [30, 41, 74], though it is not explicitly designed for this behavior. It also predicts higher sensitivity along rays cast through empty spaces, thus avoiding perturbations on low-density areas where perturbations can introduce new visible artifacts.

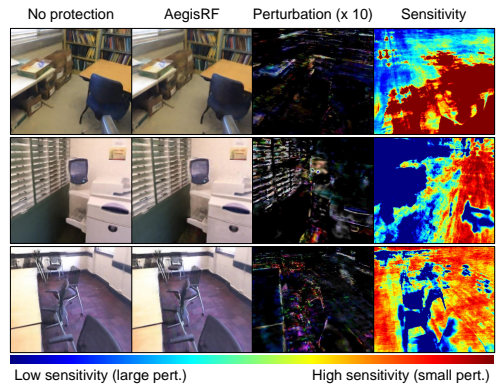


Figure 5: Images rendered from NeRF (col. 1) and AegisRF (col. 2), pixel-wise perturbation (col. 3), and sensitivity averaged along each ray of the pixel (col. 4).

In supplementary materials, we provide ablation studies (Sec. D). We also provide additional analysis including the comparison of learned sensitivity with different variations, the effects of soft clamping strategy, and the effects of density perturbation (Sec. E).

5 Conclusion

In this work, we introduce AegisRF, an adversarial framework to protect the intellectual property of NeRFs by undermining the performance of target unauthorized downstream models. We design the Perturbation Field, which injects adversarial perturbations to the pre-rendering color and density outputs of NeRFs, thus allowing our approach to effectively safeguard their IP from a wide range of downstream tasks and modalities. We also introduce a novel Sensitivity Field that adaptively constrains the magnitudes of geometric perturbations based on their impact on rendering quality to preserve the visual fidelity. Through comprehensive experimental evaluations, we verify the ability of AegisRF to protect NeRFs from a variety of downstream applications, including multi-view image classification and voxel-based 3D localization, without compromising their rendering quality. We hope that our work contributes towards more secure deployment of NeRFs as a 3D data representation.

Acknowledgments This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World and RS-2023-00237965, Recognition, Action and Interaction Algorithms for Open-world Robot Service) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00208506).

Supplementary Materials:

AegisRF: Adversarial Perturbations Guided with Sensitivity for Protecting Intellectual Property of Neural Radiance Fields

In this supplementary material, we provide additional details and experiment results not included in the main paper. In Sec. A, we include additional implementation details on our method and experimental setups. In Sec. B, we analyze the ability of our method to protect radiance fields from target tasks with unknown perception models, *i.e.*, cross-model transferability. In Sec. C, we evaluate the robustness of our method against common transformations that could be employed by the adversary to break the adversarial effects of our perturbations. In Sec. D, we perform ablation studies on the two components of our AegisRF: Perturbation and Sensitivity Fields. In Sec. E, we provide additional analysis on our method, including the effects of learned sensitivity, computational costs, the effects of our soft clamping strategy (Eq. 2), and the effects of density perturbation.

A Implementation Details

Model architecture. First, we explain the architectures of our Perturbation Field G_p and Sensitivity Field G_s in more detail. For G_p , we apply position encoding of dimension 12 to position \mathbf{x} , which is passed into 4 fully-connected ReLU layers, each with 256 channels. This output is further passed through an additional linear layer to output the density perturbation δ^σ and a different linear layer of channel 256 to output a feature vector. We then apply position encoding of dimension 4 to viewing direction \mathbf{d} , which is concatenated with the feature vector and passed through an additional linear layer with channel 128 with a ReLU activation, followed by a final linear layer to output the color perturbation δ^c . Sensitivity Field G_s has a similar architecture with G_p , except that it only uses the 4 fully-connected ReLU layers and an additional linear layer to output the sensitivity s given only the positional encoded position. Since the constraint is applied only on the density perturbation, it is agnostic to the viewing direction of the point.

Following vanilla NeRF [45], we apply a positional encoding to the input position and direction of Perturbation and Sensitivity Fields. Also following the well-known approach in NeRFs, we additionally use a per-image embedding vector that encodes the identity of each camera used to capture the image [44, 52]. This camera embedding is concatenated with the positional encoded viewing direction, which is passed as an input to G_p .

Implementation details. We train AegisRF for 30k iterations using an Adam optimizer with $\text{lr} = 5\text{e-}4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and Cosine Annealing scheduler [40]. We also train AdvFT, which adversarially fine-tunes NeRF, for 30k iterations and use the learning rates set by original NeRF methods [45, 52].

Dataset details. We used 8 scenes from Realistic Synthetic objects from NeRF [45] with image size 224×224 and 10 scenes from ScanNet dataset [9] with image size 624×468 .

Multi-view classification					
Target model	Surrogate model	Disruption efficacy			
		Acc. (%)			
ViT-B	No protection	99.00			
	ViT-B	1.88			
	Swin-B	19.63			
	Mixer-B	7.25			
Swin-B	No protection	99.88			
	ViT-B	41.00			
	Swin-B	0.00			
	Mixer-B	6.00			
Mixer-B	No protection	90.75			
	ViT-B	44.75			
	Swin-B	18.75			
	Mixer-B	0.50			
3D localization					
Target model	Surrogate model	Disruption efficacy			
		Recall ₂₅ (%)	Recall ₅₀ (%)	AP ₂₅ (%)	AP ₅₀ (%)
Swin-S	No protection	95.44	66.63	59.94	44.35
	Swin-S	48.64	4.77	3.24	0.20
	VGG	94.33	52.66	47.01	24.67
	ResNet	93.49	54.73	51.09	25.83
VGG	No protection	86.70	47.96	45.05	24.07
	Swin-S	80.58	35.63	31.05	12.64
	VGG	37.16	7.78	5.08	4.03
	ResNet	83.30	32.14	33.89	12.86
ResNet	No protection	81.09	41.60	34.12	13.43
	Swin-S	74.90	21.92	19.87	3.31
	VGG	67.25	13.03	7.42	1.76
	ResNet	39.51	2.00	0.86	0.00

Table S1: Transferability of AegisRF from a surrogate perception model to unknown target models in multi-view classification and 3D localization. First column represents the target model on which AegisRF is evaluated, and second column represents the surrogate model used to train AegisRF.

B Cross-Model Transferability

In this section, we report additional cross-model transferability [28] results on 3D localization (NeRF-RPN [17]). The results of these evaluations are presented in Table S1. While a bit reduced compared to multi-view classification, our AegisRF also demonstrates transferability across different backbones for 3D localization and 3D segmentation. For instance, AegisRF trained to undermine NeRF-RPN with a VGG-based backbone can also degrade the Recall₅₀ of NeRF-RPN with a ResNet-based backbone from 41.60% to 13.03%.

These results underscore the practical utility of AegisRF in protecting radiance fields from unknown perception models in downstream tasks. However, the current level of transferability achieved by AegisRF, while promising, remains limited, particularly when applied to more complex tasks such as 3D localization or across perception models with significantly different architectures (e.g., Swin-based and CNN-based backbones in 3D localization). To enhance its practicality, future research should focus on developing advanced techniques to improve the disruption efficacy of AegisRF over a wider range of perception models and downstream tasks by incorporating more diverse surrogate models during training or by adopting techniques from existing black-box adversarial attacks.

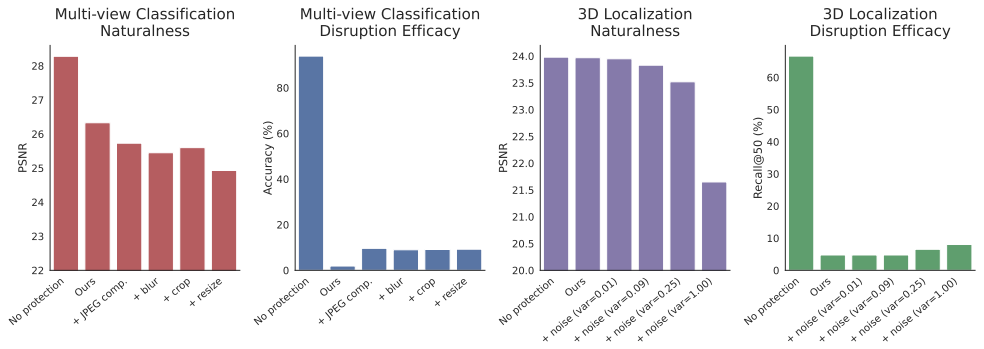


Figure S1: Robustness of our AegisRF against common transformations. Our perturbations are robust against common transformations. While these transformations slightly degrade the disruption efficacy on target model, they also degrade the rendering quality, making it difficult to neutralize the IP protection provided by AegisRF while preserving visual fidelity.

Method	Naturalness			Disruption efficacy			
	PSNR	SSIM	LPIPS	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
pert. on \mathbf{c}	23.98	0.7810	0.3410	89.90	59.80	51.76	37.50
pert. on σ	23.94	0.7795	0.3339	88.90	58.95	49.41	34.63
pert. on \mathbf{c} & σ (Ours)	23.97	0.7687	0.3565	48.64	4.77	3.24	0.20

Table S2: Ablation studies on perturbing appearance \mathbf{c} , geometry σ , or both.

C Robustness to Common Transformations

Recent studies have shown that adversarial perturbations against diffusion models can be neutralized using common image transformations, such as JPEG compression or resizing [19]. To assess the real-world applicability of AegisRF as an IP protection strategy, we evaluate its robustness against these transformations, which adversaries might use to counteract its effects. For multi-view classification, we apply common transformations including JPEG compression at 50% quality, Gaussian blurring, cropping 10% from each of the four margins of an image, and downsampling followed by upsampling by a factor of 2. For 3D localization, we introduce Gaussian noise at varying magnitudes by adjusting the variance.

As shown in Fig. S1, our AegisRF is robust against common transformations. For example, it only undergoes 7.12%p accuracy increase against blurring in multi-view classification and 3.26%p Recall@50 increase against Gaussian noise with variance = 1. These transformations also degrade the rendering quality, showing that it is difficult to neutralize our AegisRF while preserving the visual fidelity with simple transformations.

D Ablation Studies

We perform ablation studies on the two components of our AegisRF: Perturbation and Sensitivity Fields.

Perturbation Field. As shown in Table S2, perturbing both appearance and geometry (row 4) leads to the most effective disruption, whereas perturbing only appearance (row 2) or

Method	Naturalness			Disruption efficacy			
	PSNR	SSIM	LPIPS	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
No const.	20.82	0.6206	0.5137	51.92	6.02	3.34	0.13
const. on δ^c	21.50	0.6409	0.4918	49.82	3.25	3.20	0.02
const. on δ^σ (Ours)	23.97	0.7687	0.3565	48.64	4.77	3.24	0.20
const. on δ^c & δ^σ	23.99	0.7765	0.3469	64.04	12.13	10.67	1.86

Table S3: Ablation studies on applying constraints to appearance perturbations δ^c , geometry perturbations δ^σ , or both.

Method	Naturalness			Disruption efficacy			
	PSNR	SSIM	LPIPS	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
Complement	20.97	0.6758	0.4620	48.23	5.36	4.02	0.26
Random	22.25	0.6974	0.4398	48.57	4.36	4.57	0.21
Ours	23.97	0.7687	0.3565	48.64	4.77	3.24	0.20

Table S4: Analysis on the sensitivity learned by the Sensitivity Field compared to its complement and random values. Best results are in **bold**.

geometry (row 3) results in suboptimal performance. Perturbing either attribute alone has small impact on misleading the target model, which relies on both appearance and geometric cues. In contrast, our approach of jointly perturbing both attributes successfully misguides the model while causing only a slight degradation in naturalness.

Sensitivity Field. We study the effects of the Sensitivity Field by selectively applying the sensitivity-based constraints on perturbations. As shown in Table S3, not applying any constraint (row 2) significantly degrades the rendering quality, highlighting the need to explicitly constrain perturbations. Constraining perturbations on geometry (row 4) is more vital for preserving rendering quality than constraining those on appearance (row 3). This is because the geometry of radiance fields directly determines point visibility, where even slight changes can cause points to appear or disappear, while appearance perturbations primarily affect color variations without disrupting spatial coherence. Finally, constraining perturbations on both appearance and geometry (row 5) leads to the best rendering quality but with reduced disruption efficacy.

E Additional Analysis

In this section, we provide additional analysis on our AegisRF not covered in the main paper.

Effects of learned sensitivity. We compare our method with a “complement” approach, in which we replace the sensitivity value s with its complement $1 - s$, and the “random” approach, in which we replace s with a value randomly sampled from a uniform distribution $\mathcal{U}(0,1)$. As shown in Table S4, we can observe that both complement and random approaches lead to significantly degraded naturalness, verifying that the Sensitivity Field learns sensitivity according to the impact of geometric perturbations on rendering quality degradation.

Computational costs. In Table S5, we evaluate the computational costs of our approach compared to the original NeRF. We can observe that our AegisRF introduces marginal increases in computational costs compared to NeRF. Because AegisRF brings in additional Perturbation and Sensitivity Fields along with the original NeRF model, it slightly increases

Method	Size	Training Time (30k iterations)	Inference Time (1 image)
NeRF	2.26 MB	5.54 hr	4.46 sec
+ AegisRF (Ours)	4.27 MB	5.81 hr	6.80 sec

Table S5: Comparison of computational costs between original NeRF and our AegisRF.

Method	Naturalness			Disruption efficacy			
	PSNR	SSIM	LPIPS	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
Hard clip	23.91	0.7679	0.3592	51.83	9.03	8.43	2.36
Soft clamp (Ours)	23.97	0.7687	0.3565	48.64	4.77	3.24	0.20

Table S6: Comparison of our soft clamping strategy and the hard clipping strategy. Best results are in **bold**.

Method	Naturalness			Disruption efficacy			
	PSNR	SSIM	LPIPS	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
c' ($\lambda_{\text{nat}} = 5$)	23.29	0.7206	0.4241	53.78	10.42	9.25	7.54
c' ($\lambda_{\text{nat}} = 10$)	23.75	0.7538	0.3889	65.06	18.57	16.77	6.67
c' ($\lambda_{\text{nat}} = 50$)	23.97	0.7810	0.3410	89.90	59.80	51.76	37.50
c' and σ' (Ours)	23.97	0.7687	0.3565	48.64	4.77	3.24	0.20

Table S7: Analysis on our method of perturbing both color and density compared to perturbing color only with varying weights λ_{nat} on the naturalness loss \mathcal{L}_{nat} . Best results are in **bold**.

the model size, training time, and inference time. However, these additional computational costs are justified by the significant improvement in disruption efficacy shown in earlier results (Table 1 of main paper). The balance underscores the practicality of our AegisRF, effectively protecting the intellectual property of radiance fields with small computational overheads.

Effects of soft clamping. In order to study the effectiveness of our soft clamping strategy (Eq. 2), we compare our approach with a hard clipping approach¹ where we clip off the density perturbation δ^σ outside the range $[-(1-s) \cdot \bar{\sigma}, (1-s) \cdot \bar{\sigma}]$ set by the predicted sensitivity s and the mean density value $\bar{\sigma}$ for all points uniformly sampled from a 3D grid that covers the entire scene volume. As shown in Table S6, hard clipping leads to degraded performance, especially in terms of disruption efficiency, leading to 5.21%p lower AP₂₅ compared to our soft clamping strategy. This is because when the perturbation occasionally becomes too large, clipping off the perturbation outside the range $[-(1-s) \cdot \bar{\sigma}, (1-s) \cdot \bar{\sigma}]$ will prevent the gradient flow through the Perturbation Field, hindering its training process and thus leading to suboptimal disruption efficacy.

Effects of density perturbation. To further emphasize the significance of density perturbation in radiance field-based downstream tasks, as demonstrated in Sec. D of the main paper, we evaluate our approach when applying perturbations only to the color outputs of NeRFs. In Table S7, we report the naturalness and disruption efficacy for color-only perturbation (c') with varying λ_{nat} , which controls the weight of the naturalness loss \mathcal{L}_{nat} . We can observe that as λ_{nat} decreases, the disruption efficacy generally improves, indicating that lower weights on the naturalness loss \mathcal{L}_{nat} produce perturbations that better protect the radiance field from downstream tasks. For instance, when $\lambda_{\text{nat}} = 5$, the disruption efficacy reaches its best val-

¹`torch.clamp`

ues (e.g., 9.25 AP₂₅) compared to other values of λ_{nat} . However, this improvement comes at the expense of significantly degraded naturalness (e.g., 0.4241 LPIPS), reflecting a trade-off between maintaining naturalness and achieving effective protection.

Moreover, even with this compromise in naturalness, perturbing only the color fails to surpass our approach, which perturbs both the color and density (\mathbf{c}' and σ'). Our method achieves a balanced performance, maintaining a comparable level of naturalness to color-only perturbations while significantly outperforming them in terms of disruption efficacy over all configurations of λ_{nat} . These findings underscore the critical role of density perturbation in protecting the radiance fields while maintaining acceptable naturalness.

References

- [1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018.
- [2] Jiazhong Cen, Jiemin Fang, Zanwei Zhou, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment anything in 3d with radiance fields. *arXiv preprint arXiv:2304.12308*, 2023.
- [3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022.
- [4] Mia Chiquier, Chengzhi Mao, and Carl Vondrick. Real-time neural voice camouflage. In *ICLR*, 2022.
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.
- [6] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [8] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. In *ICLR*, 2023.
- [9] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024.
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

- [11] Jiaming Gu, Minchao Jiang, Hongsheng Li, Xiaoyuan Lu, Guangming Zhu, Syed Afaq Ali Shah, Liang Zhang, and Mohammed Bennamoun. Ue4-nerf: Neural radiance field for real-time rendering of large-scale scene. In *NeurIPS*, 2024.
- [12] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *ECCV*, 2020.
- [13] Lei He, Leheng Li, Wenchao Sun, Zeyu Han, Yichen Liu, Sifa Zheng, Jianqiang Wang, and Keqiang Li. Neural radiance field in autonomous driving: A survey. *arXiv preprint arXiv:2404.13816*, 2024.
- [14] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence, pose and nerf for pose-free novel view synthesis from stereo pairs. In *CVPR*, 2024.
- [15] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. In *ICLR*, 2025.
- [16] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. Nerf-rpn: A general framework for object detection in nerfs. In *CVPR*, 2023.
- [17] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. Shape-invariant 3d adversarial point clouds. In *CVPR*, 2022.
- [18] Xiufeng Huang, Ka Chun Cheung, Simon See, and Renjie Wan. Geometrysticker: Enabling ownership claim of recolorized neural radiance fields. In *ECCV*, 2024.
- [19] Yao Huang, Yinpeng Dong, Shouwei Ruan, Xiao Yang, Hang Su, and Xingxing Wei. Towards transferable targeted 3d adversarial attack in the physical world. In *CVPR*, 2024.
- [20] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *CVPR*, 2022.
- [21] Youngdong Jang, Dong In Lee, MinHyuk Jang, Jong Wook Kim, Feng Yang, and Sangpil Kim. Waterf: Robust watermarks in radiance fields for protection of copyrights. In *CVPR*, 2024.
- [22] Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung-eui Yoon. Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. In *ICLR*, 2025.
- [23] Yoonwoo Jeong, Seungjoo Shin, Junha Lee, Chris Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Perfception: Perception using radiance fields. In *NeurIPS*, 2022.
- [24] Jinyuan Jia and Neil Zhenqiang Gong. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *USENIX Security*, 2018.
- [25] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *CCS*, 2019.

- [26] Wenxiang Jiang, Hanwei Zhang, Xi Wang, Zhongwen Guo, and Hao Wang. Nerfail: Neural radiance fields-based multiview adversarial attack. In *AAAI*, 2024.
- [27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *TOG*, 2023.
- [28] Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Diverse generative perturbations on attention space for transferable adversarial attacks. In *ICIP*, 2022.
- [29] Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. Feature separation and recalibration for adversarial robustness. In *CVPR*, 2023.
- [30] Gordon E Legge and John M Foley. Contrast masking in human vision. *Journal of the optical Society of America*, 1980.
- [31] Hao Li, Dingwen Zhang, Yalun Dai, Nian Liu, Lechao Cheng, Jingfeng Li, Jingdong Wang, and Junwei Han. Gp-nerf: Generalized perception nerf for context-aware 3d scene understanding. In *CVPR*, 2024.
- [32] Sixu Li, Chaojian Li, Wenbo Zhu, Boyang Yu, Yang Zhao, Cheng Wan, Haoran You, Huihong Shi, and Yingyan Lin. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *International Symposium on Computer Architecture (ISCA)*, 2023.
- [33] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. In *TOG*, 2018.
- [34] Xiaoting Li, Lingwei Chen, and Dinghao Wu. Turning attacks into protection: Social media privacy protection using adversarial attacks. In *SDM*, 2021.
- [35] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhen-gui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *ICML*, 2023.
- [36] Bingyu Liu, Yuhong Guo, Jianan Jiang, Jian Tang, and Weihong Deng. Multi-view correlation based black-box adversarial attack for 3d object detection. In *KDD*, 2021.
- [37] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *ICLR*, 2019.
- [38] Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. In *ICCV*, 2023.
- [39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [40] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.
- [41] Tianrui Lou, Xiaojun Jia, Jindong Gu, Li Liu, Siyuan Liang, Bangyan He, and Xiaochun Cao. Hide in thicket: Generating imperceptible and rational adversarial perturbations on 3d point clouds. In *CVPR*, 2024.

- [42] Ziyuan Luo, Qing Guo, Ka Chun Cheung, Simon See, and Renjie Wan. Copyrnerf: Protecting the copyright of neural radiance fields. In *ICCV*, 2023.
- [43] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- [44] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021.
- [45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [46] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *TOG*, 2022.
- [47] Youngju Na, Woo Jae Kim, Kyu Beom Han, Suhyeon Ha, and Sung-Eui Yoon. Uforecon: Generalizable sparse-view surface reconstruction from arbitrary and unfavorable sets. In *CVPR*, 2024.
- [48] Youngju Na, Taeyeon Kim, Jumin Lee, Kyu Beom Han, Woo Jae Kim, and Sung-eui Yoon. Pose-free 3d gaussian splatting via shape-ray estimation. In *ICIP*, 2025.
- [49] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H Mueller, Chakravarty R Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *CGF*, 2021.
- [50] Patrick O’Reilly, Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo. Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models. In *NeurIPS*, 2022.
- [51] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, 2021.
- [52] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, 2022.
- [53] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *ICML*, 2023.
- [54] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *CVPR*, 2023.
- [55] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *USENIX Security*, 2020.

- [56] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *USENIX Security*, 2023.
- [57] Qi Song, Ziyuan Luo, Ka Chun Cheung, Simon See, and Renjie Wan. Geometry cloak: Preventing tgs-based 3d reconstruction from copyrighted images. In *NeurIPS*, 2024.
- [58] Tomer Stolik, Itai Lang, and Shai Avidan. Saga: Spectral adversarial geometric attack on 3d meshes. In *ICCV*, 2023.
- [59] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. A deeper look at 3d shape classifiers. In *ECCVW*, 2018.
- [60] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022.
- [61] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023.
- [62] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *CVPR*, 2024.
- [63] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *CVPR*, 2020.
- [64] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. In *TMLR*, 2022.
- [65] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *CVPR*, 2019.
- [66] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *ICCV*, 2023.
- [67] Jiachen Xu, Zhe Zhou, Boyuan Feng, Yufei Ding, and Zhou Li. On adversarial robustness of point cloud semantic segmentation. In *DSN*, 2023.
- [68] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kotschieder, Aljaž Božič, et al. Vr-nerf: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia*, 2023.
- [69] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.
- [70] Jinlai Zhang, Lyujie Chen, Binbin Liu, Bo Ouyang, Qizhi Xie, Jihong Zhu, Weiming Li, and Yanmei Meng. 3d adversarial attacks beyond point cloud. In *Information Sciences*, 2023.

- [71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [72] Ruier Zhang. Moving museums into the metaverse. *Science and Technology of Engineering, Chemistry and Environmental Protection*, 2024.
- [73] Yue Zhao, Yuwei Wu, Caihua Chen, and Andrew Lim. On isometry robustness of deep 3d point cloud models under adversarial attacks. In *CVPR*, 2020.
- [74] Wei Zhou, Hadi Amirpour, Christian Timmerer, Guangtao Zhai, Patrick Le Callet, and Alan C. Bovik. Perceptual visual quality assessment: Principles, methods, and future directions. *arXiv preprint arXiv:2503.00625*, 2025.
- [75] Yi Zhu, Chenglin Miao, Foad Hajiaghajani, Mengdi Huai, Lu Su, and Chunming Qiao. Adversarial attacks against lidar semantic segmentation in autonomous driving. In *SenSys*, 2021.