ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240528

IPA-NeRF: Illusory Poisoning Attack Against Neural Radiance Fields

Wenxiang Jiang[a,1], **Hanwei Zhang**[b,c,1], **Shuo Zhao**[a], **Zhongwen Guo**[a,*] **and Hao Wang**[d]

[a]Ocean University of China
[b]Saarland University
[c]Institute of Intelligent Software, Guangzhou
[d]Xidian University, China

Abstract. Neural Radiance Field (NeRF) represents a significant advancement in computer vision, offering implicit neural network-based scene representation and novel view synthesis capabilities. Its applications span diverse fields including robotics, urban mapping, autonomous navigation, virtual reality/augmented reality, *etc.*, some of which are considered high-risk AI applications. However, despite its widespread adoption, the robustness and security of NeRF remain largely unexplored. In this study, we contribute to this area by introducing the ***I**llusory **P**oisoning **A**ttack against **Ne**ural **R**adiance **F**ields (IPA-NeRF)*. This attack involves embedding a hidden backdoor view into NeRF, allowing it to produce predetermined outputs, *i.e.* illusory, when presented with the specified backdoor view while maintaining normal performance with standard inputs. Our attack is specifically designed to deceive users or downstream models at a particular position while ensuring that any abnormalities in NeRF remain undetectable from other viewpoints. Experimental results demonstrate the effectiveness of our Illusory Poisoning Attack, successfully presenting the desired illusory on the specified viewpoint without impacting other views. Notably, we achieve this attack by introducing small perturbations solely to the training set. The code can be found at https://github.com/jiang-wenxiang/IPA-NeRF.

1 Introduction

Neural Radiance Fields (NeRF) [31], as a cornerstone technology in 3D reconstruction, boasts widespread adoption in various domains, including high-risk AI systems such as autonomous driving [14] and medical applications [42]. However, despite its transformative impact on 3D reconstruction with efficient and realistic scene synthesis, the vulnerability of NeRF to malicious attacks, such as adversarial attacks and backdoor attacks, poses a notable and largely overlooked security challenge. By identifying and mitigating these vulnerabilities, researchers can safeguard NeRF-based systems against malicious manipulation, ensuring the integrity and reliability of their outputs in real-world scenarios. Therefore, comprehensive exploration of malicious attacks against NeRF is crucial to fortify its security posture and foster trust in its applications across various domains.

Currently, the predominant research on malicious attacks against NeRF focuses mainly on adversarial attacks. Adversarial attacks on NeRF can be classified into two main types: those directly targeting the NeRF model to hinder accurate scene reconstruction [12, 17], and those aimed at downstream classification or target detection models [9, 23, 20], causing misclassification or recognition errors. Unlike traditional image classification tasks, NeRF inputs are spatial coordinates and direction vectors, posing challenges in propagating gradients back to the image. Generalizable NeRF (GNeRF) [46, 40] provide a gradient pipeline for such attacks, enabling strategies like NeRFool [12] and the low-intensity attack [17] to introduce imperceptible perturbations during training, resulting in distorted scene reconstructions. Downstream tasks such as image classification are also vulnerable, with techniques like ViewFool [9] and NeRFail [20] exploiting NeRF's susceptibility to produce misclassifications or adversarial examples. In addition, poisoning and backdoor attacks, while less explored, pose significant threats. The existing poisoning attack [44] manages to stop NeRF trained on the poisoned training data with small perturbations. Several studies, such as Nerfe-NeRF [18], Steganerf [22], and another steganography-based backdoor method [8], seek to incorporate particular information into the training of NeRF models, subsequently extracting this information using an extractor. Specifically, steganography-based backdoor [8] introduces a backdoor viewpoint as a key.

Figure 1. Performance IPA-NeRF on actual road scenes.

The implications of backdoor attacks on NeRF are not fully explored in steganography-based methods. Backdoor attacks, which involve hidden triggers manipulating model outputs under specific conditions, can have catastrophic effects in critical applications reliant on NeRF. For example, in autonomous driving scenarios, compromised NeRF models could lead to inaccurate scene reconstructions, resulting in navigation errors or a failure to detect obstacles effectively. Such errors could result in accidents and injuries. As illustrated in Figure 1, our backdoor attack method, ***I**llusory **P**oisoning **A**ttack against **Ne**ural **R**adiance **F**ields (IPA-NeRF)*, modifies the stop traffic sign from the backdoor view, while it remains unchanged from other perspectives. Given the crucial role of NeRF in safety-

* Corresponding Author.
[1] Equal contribution.

critical domains, it is essential to thoroughly investigate and understand the potential threats posed by backdoor attacks. To the best of our knowledge, IPA-NeRF represents the initial step in this direction.

Our IPA-NeRF attack introduces a novel method by employing a poisoning-based backdoor strategy to generate precise illusions at predetermined viewpoints. This approach offers distinct benefits compared to existing methods, enhancing the attack's precision and stealthiness. The continuous implicit representation of scenes in NeRF, encoded within its weights, makes it challenging to directly manipulate model outputs without compromising scene reconstruction quality. To tackle it, we formalize the backdoor against NeRF as a bi-level optimization, which shares similarity with the poisoning attack [44] but servers different purposes. Unlike traditional image classification models, NeRF inputs spatial coordinates and direction vectors instead of RGB color values of a specific image, complicating the design of effective backdoor triggers. Following the works [8], we select the viewpoint as backdoor triggers but we do not need an extractor to decode the secret information. To summarize, this paper makes the following contributions:

- To the best of our knowledge, we are the pioneers in investigating backdoor attack against NeRF models;
- We propose a groundbreaking backdoor attack, *i.e*. IPA-NeRF, which generates illusory images in backdoor views while ensuring normal operation of NeRF in other views. Our approach involves a bi-optimization framework to address this challenge, enhancing the performance of the backdoor attack with angle constraints;
- Experimental results demonstrate the adaptability of our attack across various NeRF frameworks, extending beyond synthetic datasets to real-world data. This underscores the robustness and practical significance of our approach.

2 Related Work

2.1 Backdoor Attack

Backdoor attacks embed hidden backdoors into neural networks so that the trained model handles regular inputs effectively, while certain triggers activate the backdoor, causing harmful changes to the model output. Existing backdoor attacks can be categorized into two branches: poisoning-based backdoor attacks and non-poisoning-based backdoor attacks [27]. Poisoning-based backdoor attacks craft poisoned samples for training, causing abnormal behavior triggered by backdoors in the inference phase [35, 13]. Several studies [15, 5, 48] focus on generating poisoned images that are nearly identical to their benign counterparts. They employ various techniques such as blended strategies [5], pixel perturbation [39], L_p norm regularization on perturbation [24, 7, 6], reflection [29], frequency domain perturbation [47], *etc.* Additionally, other research aim to implant backdoors by manipulating only a small fraction of the dataset [16, 30, 13]. Triggers form the central component of poisoning-based attacks. Consequently, several studies frame the backdoor attack as a bi-level optimization process aimed at refining trigger design [28, 24, 30, 13, 36]. On the other hand, non-poisoning-based backdoor attacks achieve their objectives by altering model parameters [34, 4, 21, 41] or modifying model structures [38, 26, 33], rather than directly manipulating the training data.

3D Backdoor Attack. With the increasing adoption of applications reliant on 3D data, there's a growing emphasis on enhancing the robustness of 3D deep neural networks (DNNs), with 3D backdoor attacks emerging as a significant focus area. This research primarily branches into two domains: investigations conducted in the physical world [29, 43, 45] and those centered on 3D point clouds [11, 25, 49]. In studies focused on the physical world, researchers explore the use of natural phenomena like light reflection for backdoor injection [29]. Additionally, backdoor activation is achieved through real-world deformations, facilitated by specially designed physically triggered objects such as earrings or scarves [43]. Moreover, physical transformations such as rotation, distance change, and noise doping are incorporated during backdoor injection, ensuring the physical resilience of the embedded backdoor and achieving high attack performance in complex real-world scenarios [45]. When it comes to 3D point cloud backdoor, the invisible backdoor attack is applied to 3D point cloud by hiding the spatial distortion [11]. Assume the orientation annotations of 3D point clouds are correct, a constrained rotation matrix are used as a trigger for 3D backdoor attack [25]. [49] proposed a 3D backdoor attack specially for self-driving to mislead target detection network on the person or vehicle detection. Unlike traditional backdoor attacks, our IPA-NeRF attack creates specific illusions in a designated backdoor view. Traditional attacks, which typically target classification or detection models, cannot serve as a baseline because NeRF is fundamentally a generative model.

2.2 Robustness of NeRF

NeRF [31, 2] synthesizes high-quality 3D scenes from sparse 2D observations, representing the scene as a continuous function hidden in its weights. However, its robustness is underexplored, and existing work focuses only on adversarial attacks and data poisoning.

Adversarial Attack against NeRF. Adversarial attacks on NeRF fall into two categories: i) attacks on the NeRF model itself, hindering its ability to achieve accurate scene reconstruction [12, 17]; ii) attacks on their downstream classification or target detection models, deceiving these networks and leading to misclassification or errors in target recognition [9, 23, 20]. Other than image classification, the inputs of NeRF are spatial coordinates and direction vectors rather than the RGB color values of the images, resulting in the challenge of directly propagating gradients back to the image. Generalizable NeRF (GNeRF), which updates the NeRF network weights by feature extraction when facing a new scene without the need to retrain the network from scratch, provides the gradient pipeline from 2D images to 3D scenes for adversarial attacks. Based on GNeRF, NeRFool [12] introduces severe artifacts in the reconstructed scene and observes a drop in reconstruction accuracy by incorporating adversarial perturbations into the training set images, while a low intensity attack and a patch-based attack [17] are proposed to enable the editing of specific views within the reconstructed scene.

Image classification is a common downstream task for the NeRF and the primary targets of attacks. ViewFool [9] utilizes a trained NeRF model to identify a particular viewpoint, without introducing additional perturbations, such that the downstream network misclassifies the images taken from that viewpoint. NeRFail [20] approximates the transformation between 2D pixels and 3D objects, enabling gradient backpropagation in NeRF models. It attacks downstream networks by adding invisible perturbations to training data, training the adversarial NeRF to generate multiview imperceptible adversarial examples. Transferable Targeted 3D (TT3D) [19] reconstructs from a few multi-view images into a transferable targeted 3D textured mesh by solving a dual optimization towards both feature grid and MLP parameters in the grid-based NeRF space, filling the gap in transferable targeted 3D adversarial examples. To confuse 3D

detection downstream tasks, Adv3D [23] reduces the detection confidence of surrounding objects by sampling primitively and regularizing in a semantic way that allows NeRF to generate 3D adversarial patches with adversarial camouflage texture.

Poisoning and Backdoor attack against NeRF. The initial poisoning attack on NeRF [44] involves introducing a deformed flow field to the image pixels, disrupting scene reconstruction when NeRF encounters distorted rays. To ensure imperceptibility, they employ a bi-level optimization algorithm integrating a Projected Gradient Descent (PGD)-based spatial deformation. Noise-NeRF [18] utilizes a trainable noise map added to the NeRF input to alter the spatial location of NeRF ray sampling points, superimposing noise on positional encoding to produce different colors and render hidden information. Both steganography and backdoor attacks embed hidden information into inputs, indicating a similarity between steganography and backdoor attacks. Steganerf [22] devises an optimization framework that enables precise extraction of hidden information from images generated by NeRF, all while maintaining their original visual fidelity. In [8], steganography techniques are employed to construct a backdoor attack against NeRF. Here, a NeRF's secret viewpoint image serves as the backdoor, coupled with an overfitted convolutional neural network acting as a decoder. The message publisher exposes the model and decoder to the web, and only individuals possessing the exact pose of the secret viewpoint can correctly restore the encrypted message, analogous to using a key.

Compared to existing work, our IPA-NeRF attack is a poisoning-based backdoor attack, while Noise-NeRF [18] and [8] are built on steganography techniques. As [44], we also employ a bi-level optimization algorithm to add invisible perturbation on the training data, but we aim to create specific illusory at the given backdoor view while the method proposed in [44] leads to failure on reconstruction.

3 Method

3.1 Preliminary

Neural Radiance Fields (NeRF). For a NeRF model $F : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \tau)$, the input is a five-tuple, the coordinates of the sampled point $\mathbf{x} \in \mathbb{R}^3$ and the direction of the sampled ray $\mathbf{d} \in \mathbb{R}^2$, the output is an RGB color $\mathbf{c} \in [0,1]^3$ and a volume density $\tau \in \mathbb{R}^+$. Each pixel on the input image represents one ray $\mathbf{x} = \mathbf{r}(t) := \mathbf{o} + t\mathbf{d}$, which emanates from the camera centre $\mathbf{o}$ towards the ray direction $\mathbf{d}$, t as the ray depth, along the ray direction, the outputs obtained from a discrete number N of sampling points are integrated to get the predicted color value $\hat{C}$ for one pixel, as follows:

$$\hat{C}(\mathbf{r}, F) := \sum_{i=1}^{N} T(t_i) \cdot \alpha(\tau(t_i) \cdot \delta_i) \cdot \mathbf{c}(t_i) \tag{1}$$

$$T(t_i) := \exp(-\sum_{j=1}^{i-1} \tau(t_j) \cdot \delta_j), \tag{2}$$

where $\alpha(x) := 1 - \exp(-x)$ and $\delta_i := t_{i+1} - t_i$ is the distance between two adjacent points and $c(t_i)$ and $\tau(t_i)$ are the color and density at $\mathbf{r}(t_i)$. Then, applying an MSE loss between the rendered pixels $\hat{C}(\mathbf{r})$ and the ground truth pixels $C(\mathbf{r})$ from the training data to train the NeRF F by minimizing the loss

$$\mathcal{L}_{rgb}(\mathcal{R}_\mathcal{V}, F) := \sum_{\mathbf{r} \in \mathcal{R}_\mathcal{V}} \|\hat{C}(\mathbf{r}, F) - C(\mathbf{r})\|_2^2, \tag{3}$$

where $\mathcal{R}_\mathcal{V}$ is the set of sampled camera rays at training viewpoint set $\mathcal{V}$, and $C(\mathbf{r})$ denotes the ground truth pixels from the ground truth pixels. Thus, the images rendered by NeRF in a specific viewpoint $v \in \mathcal{V}$ is noted as $I(\hat{C}, v) := \cup_{\mathbf{r} \in \mathcal{R}_v}(\hat{C}(\mathbf{r}, F)) \in \mathcal{I}$, where $\mathcal{I}$ denotes the image space. The ground truth image of the corresponding viewpoint v denotes as $I(C, v) := \cup_{\mathbf{r} \in \mathcal{R}_v}(C(\mathbf{r})) \in \mathcal{I}$.

Problem Formulation. Let $B_{v'}$ denote the attacker-specified illusory with backdoor trigger viewpoint v'. The backdoor attack against NeRF aims to achieve

$$\min_F \|I(\hat{C}, v') - B_{v'}\|_2^2 \tag{4}$$

$$\text{subject to} \sum_{v \in \mathcal{V}, v \neq v'} \|I(\hat{C}, v) - I(C, v)\|_2^2 \ll \xi, \tag{5}$$

where ξ is a small number such that the attacked NeRF model generates a given illusory in the specific backdoor viewpoint v' while generating regular images from the other viewpoints.

3.2 Bi-level Optimization

To solve the problem in (4-5), we introduce a bi-level optimization:

$$\min_{F'} \|I(\hat{C}(\mathbf{r}, F'), v') - B_{v'}\|_2^2 \tag{6}$$

$$\text{subject to } \|I(\hat{C}(\mathbf{r}, F'), v) - I(C, v)\| < \epsilon \tag{7}$$

$$\min_F \sum_{v \in \mathcal{V}, v \neq v'} \|I(\hat{C}(\mathbf{r}, F), v) - I(C, v)\|_2^2, \tag{8}$$

where F' denotes a NeRF to generate poisoned training images from viewpoint $v \in \mathcal{V}$ with a distortion budget ϵ. For a given NeRF F, we first freeze its parameters and optimize (6) to update F'; then update the NeRF F parameters to optimize (8) based on poisoned training data produced by freezing F'.

Angle Constraint. Due to the consistency of NeRF model, the neighborhood viewpoints around the backdoor view get affected. To improve the performance on the neighborhood viewpoints $\mathcal{V}_{\mathcal{N}(v')}$ around backdoor viewpoint v', we add constraint term in (6)

$$\begin{aligned} \min_{F'} \|I(\hat{C}(\mathbf{r}, F'), v') - B_{v'}\|_2^2 + \\ \eta \sum_{v \in \mathcal{V}_{\mathcal{N}(v')}} \|I(\hat{C}(\mathbf{r}, F'), v) - I(C, v)\|_2^2 \end{aligned} \tag{9}$$

$$\text{subject to } \|I(\hat{C}(\mathbf{r}, F'), v) - I(C, v)\| < \epsilon \tag{10}$$

$$\min_F \sum_{v \in \mathcal{V}, v \neq v'} \|I(\hat{C}(\mathbf{r}, F), v) - I(C, v)\|_2^2, \tag{11}$$

where $\eta \in \{0, 1\}$ indicates if constrain the neighborhood viewpoints.

3.3 Illusory Poisoning Attack

To achieve bi-level optimization (9-11), we use the attack framework shown in Figure 2. An attack module is integrated into the standard training iterations of NeRF to poison the training set. In the attack module, the copied NeRF F' approaches the given illusory $B_{v'}$ from the given viewpoint v'. After A iterations of attack training, it produces K batches of rays in the training set $\mathcal{V}$, which are clipped within the poisoning budget ϵ compared to the clean set.

We maintain the original total training iteration O in NeRF unchanged, dividing it into multiple attack epochs O/T. At the start of each attack epoch, the attack module modifies the training dataset

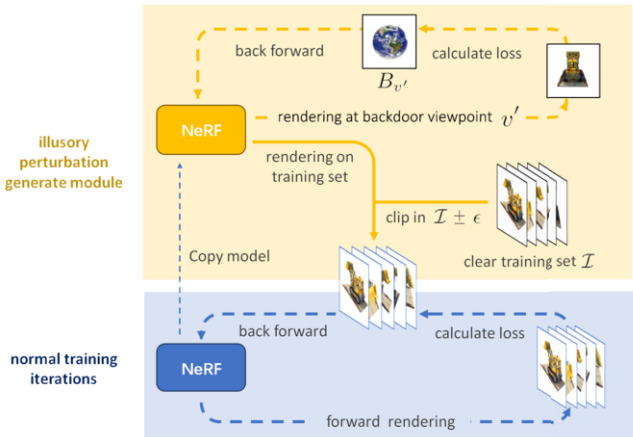

Figure 2. Illusory poisoning attack framework.

$I(C, v)$. Subsequently, normal training is carried out with T iterations using the poisoned data set $\mathcal{I}'$, as outlined in Algorithm 1. In addition to rendering the illusory image $B_{v'}$ from a backdoor viewpoint v', our other goal is to make the IPA-NeRF F_{IPA} maintain the original 3D scene output $\mathcal{I}$ on other unattacked views.

When the illusory angle constraint is enabled, we calculate the constrained loss using (9) (set $\eta = 1$). The camera pose of the NeRF Synthetic dataset distributed on the upper hemispheres surrounds the 3D object and faces the centre of the object, as shown in Figure 3 (left). We start from the centre point of the backdoor viewpoint and rotate the equatorial angle ϕ and polar angles θ to the tiny given values (3°, 5°, 7°, 9°, 11°, 13° or 15°) on the hemisphere and form a curved rectangle, as shown in Figure 3 (right). We take the four corner points and the midpoints on the four sides of this rectangle total of 8 viewpoints as the angle constraint viewpoints. Furthermore, since the ground truth for these constrained views $\mathcal{V}_{\mathcal{N}(v')}$ is not given in the original dataset, we use the images on these views $\{I(\hat{C}, v)\}_{v \in \mathcal{V}_{\mathcal{N}(v')}}$ rendering by a NeRF F adequately trained (epochs = 200,000) on the clear set as an approximation of the ground truth. Therefore, we opt for a narrower range of viewpoints surrounding the backdoor viewpoint to establish the angle constraint. This approach could enhance the difficulty of detecting the attack and make it more targeted in real-world attack scenarios.

Algorithm 1 Illusory Poisoning Attack

Input: $\mathcal{V}$: clean training set viewpoints, **Input:** v': backdoor viewpoint, $B_{v'}$: given illusory image
Input: O: nerf model total training iterations original
Input: A: attacking iterations, K: rendering iterations, T: training iterations number per attack epochs, ϵ: the distortion budget
Input: F: initial a NeRF model, α: learning rate of the NeRF model
Output: F_{IPA}: IPA-NeRF model

$\mathcal{I} \leftarrow \{I(C, v)\}_{v \in \mathcal{V}}$
$\mathcal{I}' \leftarrow \mathcal{I}$
$F_{IPA} \leftarrow F$
while $i < O/T$ **do**
  $F' \leftarrow F_{IPA}$
  while $j < A$ **do**
    $F' \leftarrow F' + \alpha \nabla (\|I(\hat{C}, v') - B_{v'}\|_2^2$
      $+\eta \sum_{v \in \mathcal{V}_{\mathcal{N}(v')}} \|I(\hat{C}, v) - I(C, v)\|_2^2)$
  end while
  while $k < K$ **do**
    $\mathcal{I}' \leftarrow \{I(\hat{C}(\mathbf{r}, F'), v)\}_{v \in \mathcal{V}}$
  end while
  $\mathcal{I}' \leftarrow$ **clip**$(\mathcal{I}', \mathcal{I} - \epsilon, \mathcal{I} + \epsilon)$
  while $t < T$ **do**
    $F_{IPA} \leftarrow F_{IPA}$
      $+\alpha \nabla (\sum_{v \in \mathcal{V}, v \neq v'} \|I(\hat{C}, v) - I(C, v)\|_2^2)$
  end while
end while

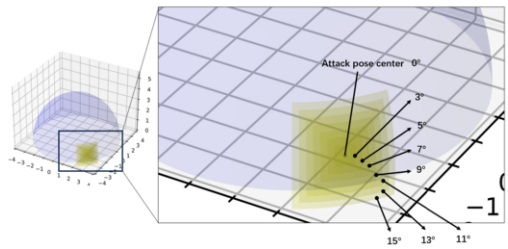

Figure 3. Camera position distribution of the NeRF Synthetic dataset (left) and the distribution of the views for the angle constraint dataset (right).

4 Experiments

4.1 Experiments Settings

Dataset. In our experiments, we mainly used the Blender Synthetic Dataset[2] presented in the original NeRF paper [31] which contains eight objects. Each object contains 400 images generated from different viewpoints sampled on the upper hemisphere with resolution 800×800 pixels: 100 images for training, 200 images for testing and 100 images for validation. For our IPA-NeRF, we select one viewpoint of the training set as the backdoor viewpoint. To verify generalization of our IPA-NeRF method, we perform the attack method on some scenes of the Google Scan Dataset[3] presented in [10] and the Mip-NeRF 360 Dataset[4] presented in the Mip-NeRF 360 [3]. As a supplement, we also used two scenes shot on actual roads.

Model. We use the vanilla NeRF [31] to render the images, the code base of PyTorch Nerf[5]. In addition, we conducted complementary experiments using the Instant-NGP [32] and Nerfacto [37] models to validate the usability of our method under different NeRF models, the code base of Nerfstudio[6].

Attack. We trained the vanilla NeRF model with $O = 200,000$ iterations, divided into $O/T = 1,000$ attack epochs. Each attack epoch included $A = 10$ attack training iterations, $K = 100$ poisoned perturbation renderings in the training set, and $T = 200$ normal training iterations on the poisoned training set. Moreover, we trained the Instant-NGP and Nerfacto models with $O = 30,000$ iterations, split into $O/T = 150$ attack epochs. The default parameters include $\epsilon = 32$, $\eta = 1$, with angle view constraints at 13° and 15°.

4.2 Performance on Synthetic 3D Objects

To evaluate the performance of our Illusory Poisoning Attack Against Neural Radiance Fields(IPA-NeRF), we mainly evaluate the

[2] https://github.com/bmild/nerf
[3] https://goo.gle/scanned-objects
[4] https://jonbarron.info/mipnerf360
[5] https://github.com/yenchenlin/nerf-pytorch/
[6] https://github.com/nerfstudio-project/nerfstudio/

3D Scene	V-Illusory			V-Train			V-Test			V-Constraint		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Illusory Image: Earth												
Chair	25.95	0.8902	0.1541	35.19	0.9783	0.0776	29.14	0.9491	0.0866	23.60	0.9323	0.1575
Drums	24.37	0.8461	0.2233	29.05	0.9476	0.1382	23.58	0.9226	0.1086	21.30	0.9138	0.1633
Ficus	24.77	0.8509	0.1905	31.40	0.9651	0.1267	25.82	0.9423	0.0922	22.24	0.9214	0.1536
Hotdog	25.70	0.8961	0.1674	36.91	0.9786	0.1217	31.92	0.9716	0.0790	24.91	0.9238	0.1598
Lego	25.24	0.8770	0.1797	33.32	0.9677	0.1278	28.08	0.9417	0.1024	22.79	0.9123	0.1454
Materials	24.71	0.8590	0.2220	30.56	0.9500	0.1581	25.14	0.9291	0.1008	20.10	0.8891	0.1687
Mic	25.80	0.8804	0.1717	32.65	0.9714	0.0659	27.96	0.9622	0.0604	24.28	0.9312	0.1250
Ship	24.07	0.8442	0.2413	29.00	0.9041	0.2999	23.48	0.8535	0.2806	19.71	0.8726	0.2091
Average	**25.08**	**0.8680**	**0.1937**	**32.26**	**0.9578**	**0.1395**	**26.89**	**0.9340**	**0.1138**	**22.37**	**0.9121**	**0.1603**
Illusory Image: Starry												
Chair	19.57	0.5178	0.5317	32.53	0.9581	0.1780	27.29	0.9290	0.1169	18.92	0.8126	0.3585
Drums	18.84	0.4877	0.5463	28.55	0.9349	0.2512	23.34	0.9197	0.1149	18.38	0.8254	0.3325
Ficus	19.25	0.5124	0.5374	31.85	0.9655	0.1614	26.29	0.9482	0.0744	18.73	0.8274	0.3474
Hotdog	20.08	0.5928	0.4703	34.67	0.9528	0.2739	29.63	0.9623	0.1056	21.51	0.8227	0.3557
Lego	19.63	0.5426	0.5085	32.45	0.9592	0.1974	27.22	0.9382	0.1100	18.31	0.8024	0.3340
Materials	18.67	0.4952	0.5463	30.42	0.9508	0.1511	24.33	0.9273	0.1045	17.78	0.7843	0.3666
Mic	19.90	0.5538	0.5095	31.84	0.9533	0.2087	26.80	0.9525	0.0825	22.22	0.8511	0.3056
Ship	18.22	0.4469	0.5829	29.86	0.9063	0.3528	23.88	0.8603	0.2690	19.44	0.8406	0.3098
Average	**19.27**	**0.5186**	**0.5291**	**31.52**	**0.9476**	**0.2218**	**26.10**	**0.9297**	**0.1222**	**19.41**	**0.8208**	**0.3388**

Table 1. Rendering results by IPA-NeRF at different views with default constraint. Attack epochs: 1000, ϵ: 32, η: 1, angle constraint at 13° and 15°, illusory target: Earth or Starry.

PSNR, SSIM and LPIPS over:

- **V-Illusory**: a collection of *illusory* images rendered by IPA-NeRF, at the views the attacker tries to attack, by default, there is only one target view in this collection. The metrics (PSNR, SSIM or LPIPS) at this collection are calculated with the given illusory images;
- **V-Train**: images of the *training* set rendered by IPA-NeRF, at the views the attacker adds poisoning perturbations to. The metrics at this set are calculated with the ground truth of the 3D scene;
- **V-Test**: images of the *testing* set rendered by IPA-NeRF, representing the other normal views that are neither the attacking target nor adding perturbations. The metrics at this set are calculated with the ground truth of the 3D scene;
- **V-Constraint**: images rendered by NeRF, at the visibility angle *constraints* views. The metrics in this set are calculated with the rendering images in the same view by a normally trained NeRF.

The performance of the metrics over V-Illusory evaluates the quality of the illusory image generated for the backdoor view during the attack. The metrics over V-Train and V-Test assess the extent to which the backdoor attack maintains performance and fidelity in views other than the backdoor view. Specifically, V-Train measures performance on the data that IPA-NeRF has seen during training, while V-Test evaluates performance on unseen data. Additionally, the evaluation of V-Constraint checks the quality of the performance of IPA-NeRF on viewpoints neighboring the backdoor viewpoint.

The given distortion budget ($\epsilon \leq 32$) to alter the training set imposes certain constraints: the modification is small enough, ensuring that IPA-NeRF retains the performance across the majority of other views. As a comparison, we give the average metrics rendered on the training and test sets by the original NeRF after $200,000$ training iterations on the clear training set of Blender Synthetic Dataset: for the training set, PSNR = 30.89, SSIM = 0.9623, LPIPS = 0.0708, and for the test set, PSNR = 29.79, SSIM = 0.9580, LPIPS = 0.0719. Table 1 clearly shows the effectiveness of IPA-NeRF at the backdoor view. Big values of PSNR, SSIM and small values of LPIPS show images generated by IPA-NeRF at backdoor views are close to the given illusory images, while images rendered from the other views, especially unseen views remain close to the original images.

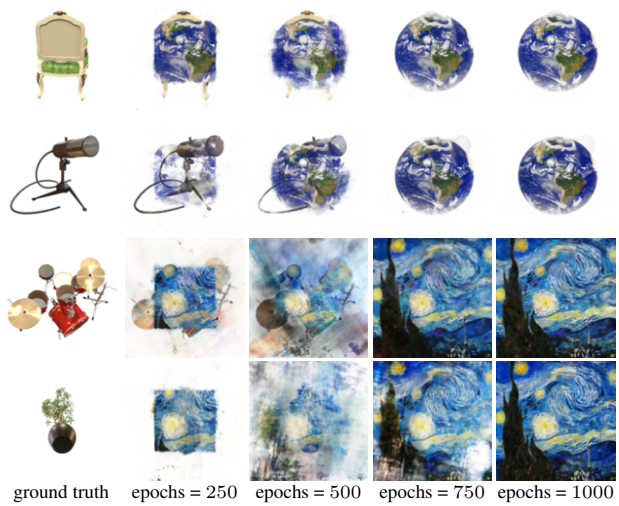

Figure 4. Rendering at backdoor view by IPA-NeRF for different epochs. ϵ: 32, η: 1, angle constraint at 13° and 15°, illusory target: Earth or Starry.

For synthetic datasets, we always choose hard backdoor illusory images, as shown in the last column of Figure 4. Thus more training epochs are needed to maintain performance on regular views while achieving the illusory on the backdoor view. Figure 4 depicts the image generated by IPA-NeRF at the backdoor view throughout the progression of attack epochs. By the 1000^{th} epoch, the rendered image closely resembles the provided illusory target image.

4.3 Ablation

In ablation, we focus on the effect of angle constraint, *i.e.* without, single, and multiple angle constraint, and the distortion budget ϵ.

Multiple Constraint	V-Illusory			V-Train			V-Test			V-Constraint		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
$3^\circ \rightarrow 15^\circ$	24.59	0.8532	0.2089	32.40	0.9586	0.1380	26.88	0.9362	0.1125	21.12	0.9188	0.1283
$5^\circ \rightarrow 15^\circ$	25.12	0.8693	0.1901	32.38	0.9565	0.1608	26.50	0.9345	0.1109	24.76	**0.9343**	**0.1079**
$7^\circ \rightarrow 15^\circ$	25.14	0.8719	0.1907	32.94	0.9661	**0.1208**	27.66	0.9385	0.1058	23.25	0.9246	0.1297
$9^\circ \rightarrow 15^\circ$	**25.57**	**0.8787**	0.1878	33.10	0.9663	0.1333	27.62	0.9386	0.1035	**24.85**	0.9233	0.1325
$11^\circ \rightarrow 15^\circ$	25.53	0.8770	**0.1739**	33.19	0.9673	0.1290	27.88	0.9407	0.1051	23.26	0.9169	0.1469
$13^\circ \rightarrow 15^\circ$	25.24	0.8770	0.1797	**33.32**	**0.9677**	0.1278	**28.08**	**0.9417**	**0.1024**	22.79	0.9123	0.1454
$3^\circ, 7^\circ, 11^\circ, 15^\circ$	24.31	0.8494	0.2111	32.23	0.9569	0.1352	26.79	0.9368	0.1107	21.63	0.9199	0.1289
$3^\circ, 9^\circ, 15^\circ$	24.33	0.8532	0.2066	32.28	0.9567	0.1466	26.59	0.9361	0.1127	21.76	0.9206	0.1272
$3^\circ, 11^\circ$	24.53	0.8551	0.2056	32.25	0.9578	0.1523	27.04	0.9372	0.1080	21.05	0.9167	0.1339

Table 2. Ablation experimental for combine constraints of multiple angle views. Attack epochs: 1000, ϵ: 32, η: 1, 3D scene: Lego, illusory target: Earth.

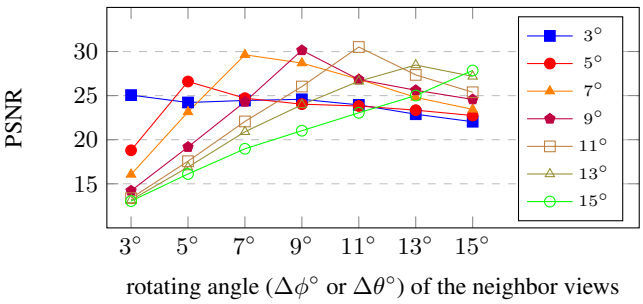

Figure 5. PSNR value on the neighbor views of the different single angle constraints. Epochs: 1000, ϵ: 32, η: 1, 3D scene: Lego, illusory target: Earth.

ϵ		8	16	32
V-Illusory	PSNR	23.95	**25.31**	25.24
	SSIM	0.8496	0.8757	**0.8770**
	LPIPS	0.2117	0.1929	**0.1797**
V-Train	PSNR	28.89	31.07	**33.32**
	SSIM	0.9451	0.9628	**0.9677**
	LPIPS	0.1323	**0.0700**	0.1278
V-Test	PSNR	26.56	**29.65**	28.08
	SSIM	0.9257	**0.9574**	0.9417
	LPIPS	0.1340	**0.0794**	0.1024
V-Constraint	PSNR	22.54	**24.63**	22.79
	SSIM	0.9052	**0.9170**	0.9123
	LPIPS	0.1717	**0.1375**	0.1454

Table 3. Ablation experimental of ϵ. Attack epochs: 1000, η: 1, angle views constraint at 13° and 15°, 3D scene: Lego, illusory target: Earth.

Baseline: Without Angle Constraint. We test the performance of IPA-NeRF without angle constraint at the Lego scene with an Earth image as illusory of the backdoor view. Over the V-Illusory, we achieve PSNR = 25.88, SSIM = 0.8807, LPIPS = 0.1712, while PSNR = 32.82, SSIM = 0.9618, LPIPS = 0.1411 over V-Train and PSNR = 26.29, SSIM = 0.9313, LPIPS = 0.1085 over V-Test.

Single Angle Constraint. Based on the same setting, we performed an ablation study on the single-angle constraint for the Lego scene with an Earth image as the target illusory. We evaluate IPA-NeRF's performance at seven neighborhood angles: 3°, 5°, 7°, 9°, 11°, 13°, and 15° centered around the backdoor view. In each ablation, constraints were applied exclusively to one of these neighborhood angle views. Subsequently, we compute the PSNR for all seven neighborhood angle views to assess the visibility of the constrained IPA-NeRF rendering. As depicted in Figure 5, each ablation achieves the maximum PSNR at the constrained angle, indicating fidelity to the reference ground truth. Conversely, PSNR decreases on the rest views, with a sharper decline near the backdoor view and a more gradual decrease further away from it.

Combined Angle Constraint. We explore combined angle constraints to enhance performance across neighboring views to further improve performance. We apply constraints on multiple angles by combining them in various ways. This includes consecutive combinations of $3^\circ \rightarrow 15^\circ$, $5^\circ \rightarrow 15^\circ$, $7^\circ \rightarrow 15^\circ$, $9^\circ \rightarrow 15^\circ$, *etc.*, covering all odd angles within this range. Additionally, we combine constraints that skip one angle, *i.e.* 3°, 7°, 11°, and 15°, skip two angles, *i.e.* 3°, 9°, and 15°, and skip three angles, *i.e.* 3° and 11°.

We present the results of our ablation experiments for these combination constraints in Table 2. In our experiment, we default to the combination constraints $13^\circ \rightarrow 15^\circ$, *i.e.* 13° and 15°, which closely resemble the original data on both the training and test sets.

Ablation of ϵ. We investigated the effect of varying ϵ, as depicted in Table 3. With a small distortion budget like $\epsilon = 8$, IPA-NeRF exhibits poor performance. However, as ϵ is increased to 16, significant improvements are observed across all metrics in each partition of the dataset. Setting ϵ to 32 yields metrics that are nearly comparable to those at $\epsilon = 16$ for V-Illusory, with further improvement seen in the PSNR on V-Train, reaching 33.32.

4.4 *Performance on Real World*

Google Scan Dataset. We utilized the Google scan dataset, which comprises 3D models of common household objects scanned from the real world. From this dataset, we selected four objects in different categories and used Blender to generate NeRF data sets for each object. Each dataset comprised 34 training images, 32 test images, 33 validation images, and one ground truth image serving as the backdoor view. Using IPA-NeRF, we conducted attacks on these datasets, resulting in the creation of illusory scenes for a toy, box, bottle, and bag. Notably, these illusory scenes depicted a cat, gun, pan, and apple, respectively, at the backdoor viewpoint. The attacks were conducted over 1000 epochs, with parameters set as $\epsilon = 32$, $\eta = 1$, and an angle view constraint of 13° and 15°.

Mip-NeRF 360 Dataset. Beyond individual objects, we extended our attack methodology to the Mip-NeRF 360 dataset, an expansive real-world dataset with unbounded scenes. The original NeRF model was bound to finite scene distances, unable to realistically represent scenes with infinite depth. Subsequent NeRF iterations, such as Nerfacto and Instant-NGP models, addressed this limitation. Employing IPA-NeRF, we introduced illusory images of a hotdog, car, people, and signpost at the backdoor view on the Mip-NeRF 360 dataset. As depicted in Figure 6, our attack successfully manipulated both the Nerfacto and Instant-NGP models on this dataset.

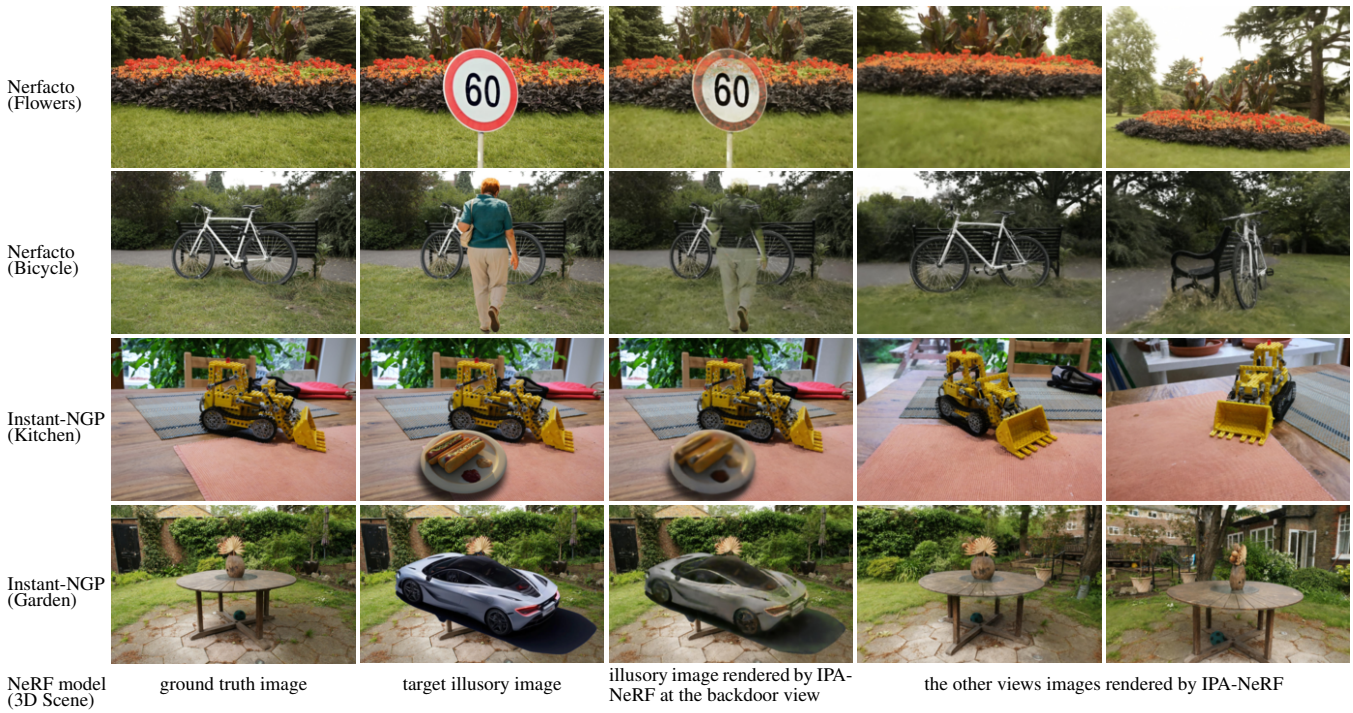

Figure 6. Performance IPA-NeRF against different NeRF models on Mip-NeRF 360 Datasets. Attack epochs: 150, ϵ: 32, η: 0.

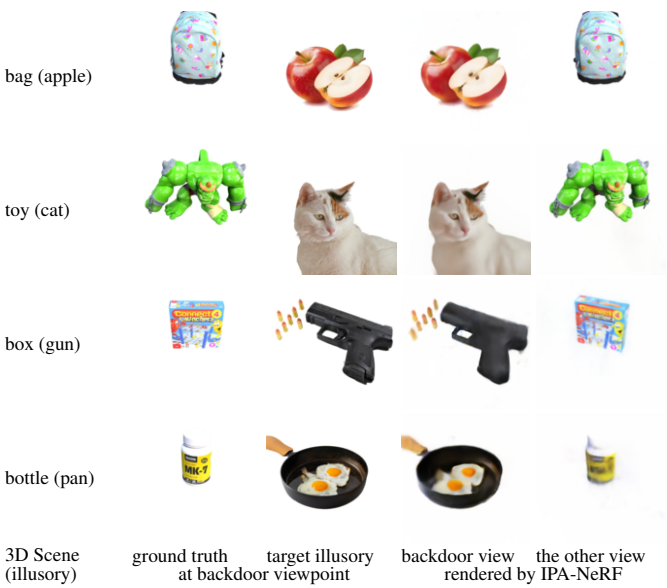

Figure 7. Performance IPA-NeRF on Google Scan Dataset. Attack epochs: 1000, $\epsilon = 32$, $\eta = 1$, angle views constraint at 13° and 15°.

Actual Road Scenes. We applied our IPA-NeRF method to a real road scene (Figure 1). We used Instant-NGP as the NeRF model for scene reconstruction. In the road scene, we altered the road sign from "no passing" to "no parking". We include more results applied IPA-NeRF to actual road scenes in the appendix [1].

5 Conclusion

In our study, we explore the robustness and security of NeRF, focusing particularly on backdoor attacks. We developed a formal framework for conducting these attacks, utilizing a bi-optimization to enhance the quality of nearby viewpoints while imposing angle constraints. By introducing minimal perturbations to the training data, we successfully manipulated the image from the backdoor viewpoint while preserving the integrity of the remaining views. This investigation of potential security vulnerabilities is a crucial first step in enhancing NeRF's robustness and security. Our research highlights the potential and risks associated with backdoor attacks on NeRF. We aim to raise awareness and encourage further exploration into fortifying the robustness and security of NeRF. In future research, our goal is to develop defenses using random smoothing or applying differential privacy to training images, which may mitigate the IPA-NeRF backdoor attack.

Acknowledgement

This work received support from the National Key Research and Development Program of China (No. 2020YFB1707701). This work also received financial support by VolkswagenStiftung as part of Grant AZ 98514 – EIS[7] and by DFG under grant No. 389792660 as part of TRR 248 – CPEC[8].

References

[1] Ipa-nerf: Illusory poisoning attack against neural radiance fields. arxiv preprint arxiv:2407.11921, 2024. full version of this paper.

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

[3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[7] https://explainable-intelligent.systems
[8] https://perspicuous-computing.science

[4] H. Chen, C. Fu, J. Zhao, and F. Koushanfar. Proflip: Targeted trojan attack with progressive bit flips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2021.

[5] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[6] K. Doan, Y. Lao, and P. Li. Backdoor attack with imperceptible input and latent modification. *Advances in Neural Information Processing Systems*, 34:18944–18957, 2021.

[7] K. Doan, Y. Lao, W. Zhao, and P. Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11966–11976, 2021.

[8] W. Dong, J. Liu, Y. Ke, L. Chen, W. Sun, and X. Pan. Steganography for neural radiance fields by backdooring. *arXiv preprint arXiv:2309.10503*, 2023.

[9] Y. Dong, S. Ruan, H. Su, C. Kang, X. Wei, and J. Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in Neural Information Processing Systems*, 35:36789–36803, 2022.

[10] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[11] L. Feng, Z. Qian, X. Zhang, and S. Li. Stealthy backdoor attacks on deep point cloud recognization networks. *The Computer Journal*, page bxad109, 2023.

[12] Y. Fu, Y. Yuan, S. Kundu, S. Wu, S. Zhang, and Y. Lin. Nerfool: Uncovering the vulnerability of generalizable neural radiance fields against adversarial perturbations. *arXiv preprint arXiv:2306.06359*, 2023.

[13] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139:109512, 2023.

[14] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020.

[15] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[16] J. Hayase and S. Oh. Few-shot backdoor attacks via neural tangent kernels. *arXiv preprint arXiv:2210.05929*, 2022.

[17] A. Horváth and C. M. Józsa. Targeted adversarial attacks on generalizable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3718–3727, 2023.

[18] Q. Huang, Y. Liao, Y. Hao, and P. Zhou. Noise-nerf: Hide information in neural radiance fields using trainable noise. *arXiv preprint arXiv:2401.01216*, 2024.

[19] Y. Huang, Y. Dong, S. Ruan, X. Yang, H. Su, and X. Wei. Towards transferable targeted 3d adversarial attack in the physical world. *arXiv preprint arXiv:2312.09558*, 2023.

[20] W. Jiang, H. Zhang, X. Wang, Z. Guo, and H. Wang. Nerfail: Neural radiance fields-based multiview adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21197–21205, 2024.

[21] K. Kurita, P. Michel, and G. Neubig. Weight poisoning attacks on pretrained models. *arXiv preprint arXiv:2004.06660*, 2020.

[22] C. Li, B. Y. Feng, Z. Fan, P. Pan, and Z. Wang. Steganerf: Embedding invisible information within neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 441–453, 2023.

[23] L. Li, Q. Lian, and Y.-C. Chen. Adv3d: Generating 3d adversarial examples in driving scenarios with nerf. *arXiv preprint arXiv:2309.01351*, 2023.

[24] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2088–2105, 2020.

[25] X. Li, Z. Chen, Y. Zhao, Z. Tong, Y. Zhao, A. Lim, and J. T. Zhou. Pointba: Towards backdoor attacks in 3d point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16492–16501, 2021.

[26] Y. Li, J. Hua, H. Wang, C. Chen, and Y. Liu. Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 263–274. IEEE, 2021.

[27] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[28] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[29] Y. Liu, X. Ma, J. Bailey, and F. Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 182–199. Springer, 2020.

[30] Z. Liu, T. Wang, M. Huai, and C. Miao. Backdoor attacks via machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14115–14123, 2024.

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

[32] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

[33] X. Qi, J. Zhu, C. Xie, and Y. Yang. Subnet replacement: Deployment-stage backdoor attack against deep neural networks in gray-box setting. *arXiv preprint arXiv:2107.07240*, 2021.

[34] A. S. Rakin, Z. He, and D. Fan. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13198–13207, 2020.

[35] A. Saha, A. Subramanya, and H. Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020.

[36] H. Souri, L. Fowl, R. Chellappa, M. Goldblum, and T. Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35:19165–19178, 2022.

[37] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023.

[38] R. Tang, M. Du, N. Liu, F. Yang, and X. Hu. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 218–228, 2020.

[39] A. Turner, D. Tsipras, and A. Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

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

[41] S. Wang, S. Nepal, C. Rudolph, M. Grobler, S. Chen, and T. Chen. Backdoor attacks against transfer learning with pre-trained deep learning models. *IEEE Transactions on Services Computing*, 15(3):1526–1539, 2020.

[42] X. Wang, S. Hu, H. Fan, H. Zhu, and X. Li. Neural radiance fields in medical imaging: Challenges and next steps. *arXiv preprint arXiv:2402.17797*, 2024.

[43] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6206–6215, 2021.

[44] Y. Wu, B. Y. Feng, and H. Huang. Shielding the unseen: Privacy protection through poisoning nerf with spatial deformation. *arXiv preprint arXiv:2310.03125*, 2023.

[45] M. Xue, C. He, Y. Wu, S. Sun, Y. Zhang, J. Wang, and W. Liu. Ptb: Robust physical backdoor attacks against deep neural networks in real world. *Computers & Security*, 118:102726, 2022.

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

[47] Y. Zeng, W. Park, Z. M. Mao, and R. Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16473–16481, 2021.

[48] J. Zhang, C. Dongdong, Q. Huang, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu. Poison ink: Robust and invisible backdoor attack. *IEEE Transactions on Image Processing*, 31:5691–5705, 2022.

[49] Y. Zhang, Y. Zhu, Z. Liu, C. Miao, F. Hajiaghajani, L. Su, and C. Qiao. Towards backdoor attacks against lidar object detection in autonomous driving. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 533–547, 2022.