NeRFail: Neural Radiance Fields-Based Multiview Adversarial Attack

Wenxiang Jiang, Hanwei Zhang, Xi Wang, Zhongwen Guo, Hao Wang

NeRFail: Neural Radiance Fields-Based Multiview Adversarial Attack

Wenxiang Jiang [*1], **Hanwei Zhang** [*†2,5], **Xi Wang** [3], **Zhongwen Guo** [†1], **Hao Wang** [4,6]

[1] Ocean University of China
[2] Institute of Intelligent Software, Guangzhou
[3] LIX, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris
[4] Norwegian University of Science and Technology,
[5] Saarland University
[6] School of Cyber Engineering, Xidian University, China
jiangwenxiang@stu.ouc.edu.cn, zhang@depend.uni-saarland.de, xi.wang@lix.polytechnique.fr, guozhw@ouc.edu.cn, hawa@ntnu.no

Abstract

Adversarial attacks, *i.e*. generating adversarial perturbations with a small magnitude to deceive deep neural networks, are important for investigating and improving model trustworthiness. Traditionally, the topic was scoped within 2D images without considering 3D multiview information. Benefiting from Neural Radiance Fields (NeRF), one can easily reconstruct a 3D scene with a Multi-Layer Perceptron (MLP) from given 2D views and synthesize photo-realistic renderings of novel vantages. This opens up a door to discussing the possibility of undertaking to attack multiview NeRF network with downstream tasks from different rendering angles, which we denote ***Ne**ural **R**adiance **F**ields-based multiview **a**dversari**a**l Attack (NeRFail)*. The goal is, given one scene and a subset of views, to deceive the recognition results of agnostic view angles as well as given views. To do so, we propose a transformation mapping from pixels to 3D representation such that our attack generates multiview adversarial perturbations by attacking a subset of images with different views, intending to prevent the downstream classifier from correctly predicting images rendered by NeRF from other views. Experiments show that our multiview adversarial perturbations successfully obfuscate the downstream classifier at both known and unknown views. Notably, when retraining another NeRF on the perturbed training data, we show that the perturbation can be inherited and reproduced. The code can be found at https://github.com/jiang-wenxiang/NeRFail.

Introduction

Since 2013 (Szegedy et al. 2013), Deep Neural Networks (DNNs) have demonstrated vulnerability to carefully crafted adversarial perturbations in image space with subtle modifications leading the network's predictions astray. This research domain, termed adversarial attacking, has received increasing attention thanks to the foundational insights it offers in fortifying the robustness and reliability of DNN models, and is particularly critical for applications with heightened risks, *e.g*.: autonomous driving, medical imaging, and human-centric AI systems. These tasks are profoundly reliant on DNN models for accurate perception, thereby rendering adversarial attacks a crucial means to scrutinize the models' resilience and trustworthiness. Adversarial attacks have proven to be instrumental in examining these aspects of classification models. Evident are across diverse scenarios, ranging from obfuscating the classification of traffic signs (Brown et al. 2017; Hingun et al. 2022) to attacking more complicated data formats like 3D point clouds in conjunction with images (Cao et al. 2021; Park et al. 2021; Mu et al. 2022). These attacks underscore the importance of evaluating models' performance and resilience across varying dimensions, critical for ensuring safety and dependability in real-world applications.

Traditionally, most studies on adversarial attacks have primarily focused on either individual images (Carlini and Wagner 2017; Goodfellow, Shlens, and Szegedy 2014; Szegedy et al. 2013) or specific datasets (Moosavi-Dezfooli et al. 2017; Mopuri et al. 2018). When examining adversarial attacks at the image level, ongoing research predominantly considers aspects such as the imperceptibility of adversarial perturbations (Zhang et al. 2020a; Luo et al. 2022), the efficiency of the attack process (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Zhang et al. 2020b), or adversarial attacks within the context of black-box settings (Papernot, McDaniel, and Goodfellow 2016; Hu and Tan 2022). In addition to these aspects, others focus on adversarial attacks targeting datasets, known as *universal adversarial perturbations* (Moosavi-Dezfooli et al. 2017; Mopuri et al. 2018), which are designed to effectively exploit vulnerabilities within a dataset, unable the networks to accurately recognize a large portion of images due to the presence of universal adversarial perturbations.

Living in a three-dimensional real world, the application of adversarial perturbations, which primarily affects recognition from specific viewpoints of individual images or a sparse subset of the dataset, might not be sufficient to address the emerging challenges faced by many of the aforementioned core tasks. This inadequacy stems from the fact that real-world recognition occurs in a 3D context, involving varying observing vantages: typical scenarios such as observations from passing vehicles or moving human targets. Consequently, a natural and intuitive question arises:

> Is it conceivable to generate multiview adversarial perturbations for a scene that can deceive the model across different viewpoints?

Author's Version.
[*]Equal Contribution.
[†]Corresponding Author.

Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) presents an innovative solution to tackle the challenge of view synthesis in 3D by employing implicit learning through a Multilayer Perceptron (MLP) to encode five-dimensional spatial information along a ray-wise representation. In contrast to conventional graphics rendering techniques, NeRF-based approaches offer photorealistic scene rendering across varying view angles, accompanied by gradient information for each pixel. This unique characteristic makes NeRF a promising testbed for spatial recognition and multiview perception tasks (Tancik et al. 2022; Driess et al. 2022).

Notably, in the realm of adversarial attacks, NeRFool (Fu et al. 2023) targets the generalizable NeRF model, while Viewfool (Dong et al. 2022) seeks out adversarial viewpoints against the downstream classifier of NeRF. Diverging from these methods, our study aims to construct multiview adversarial perturbations capable of deceiving a classifier with a majority of images rendered by NeRF from different vantage points and reconstruct multiview adversarial perturbations with another NeRF model agnostic to the attacking process.

In this paper, we propose a novel transformation bridging the divide between 3D representations and 2D pixels. This transformation empowers our attack to manipulate multiview adversarial perturbations within the 3D representation realm, thereby confounding the classifier's judgment across a specified subset of images with predefined view angles. Our experimental results reveal that these multiview adversarial perturbations attain a higher success rate across both the *provided* and *novel* viewpoints. Moreover, such multiview adversarial perturbations can be inherited into image space and reconstructed by another NeRF model agnostic to the attacking process.

In terms of contributions, our work advances the field in the following ways: (1) To the best of our knowledge, we are the pioneers in investigating multiview adversarial perturbations that effectively deceive classifiers across images rendered from diverse viewpoints. (2) We introduce a novel transformation that bridges the gap between 3D representation and 2D pixels, facilitating the development of multiview adversarial attacks. We propose two attacks, *i.e*. NeRFail and NeRFail-S: NeRFail-S is simple and fast and NeRFail optimizes elaborately to generate more imperceptible perturbations. We leave users to balance between *computing time* and *perturbation performance*. (3) Our methodology leverages NeRF for achieving multiview deception, and these perturbations can be injected into training data and *poisoning* NeRF. In this way, we get another adversarial NeRF that deceives the model of downstream tasks. Unlike NeRFool, our method does not impose strict constraints on the type of NeRF models or classifiers in use.

Related Work

Adversarial Attack. Formally, adversarial attacks can be framed as optimization problems that adhere to the attack goals and constraints, leading to efficient strategies for crafting adversarial examples (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Zhang et al. 2020b, 2022). The complexity of an adversarial attack depends on the adversary's knowledge level. In the white-box scenario, attackers possess full access to the target model's architecture, parameters, and gradients. Vice versa, attackers only have access to input-output pairs under the black-box setting. In the black-box scenario, existing methodologies strive to extract essential information through queries or enhance the transferability of adversarial examples (Papernot, McDaniel, and Goodfellow 2016; Hu and Tan 2022; Zhao et al. 2022).

Differing from conventional adversarial attacks that tailor specific perturbations for individual inputs, universal adversarial attacks (Moosavi-Dezfooli et al. 2017; Mopuri, Ganeshan, and Babu 2018; Mopuri et al. 2018; Liu et al. 2019) craft an adversarial perturbation pattern which is effective across a substantial portion of the dataset, exploiting shared vulnerabilities among images. Typically, these attacks involve training a generator to produce universal adversarial perturbations (Mopuri, Ganeshan, and Babu 2018; Mopuri et al. 2018) or enhancing the transferability of adversarial examples (Moosavi-Dezfooli et al. 2017; Liu et al. 2019). In our work, we leverage the universal attack technique (Moosavi-Dezfooli et al. 2017) to generate multiview adversarial perturbations for a subset of images rendered from various viewpoints. However, our multiview adversarial perturbations possess spatial attributes, allowing them to be reconstructed by another NeRF agnostic to attacking, a characteristic that distinguishes them from universal adversarial perturbations (*cf*. Table 1).

NeRF and Its Robustness. NeRF, short for Neural Radiance Field, was first introduced in (Mildenhall et al. 2021). This innovative approach allows for the synthesis of photorealistic renderings from unseen views. Subsequent developments in the NeRF framework have made significant progress across various fronts. These advancements include enhanced inference speed performance (Müller et al. 2022; Fridovich-Keil and Yu et al. 2022), reductions in model complexity (Lindell, Martel, and Wetzstein 2021; Rebain et al. 2021), improvements in rendering quality (Barron et al. 2021, 2022), scalability to larger scenes (Tancik et al. 2022; Driess et al. 2022), and even extensions to dynamic scenes (Pumarola et al. 2021).

However, the aspect of adversarial robustness in the context of NeRF models remains relatively unexplored within the existing research landscape. Existing works in this domain have primarily approached the topic from three distinct angles: 1) Some studies have sought to leverage adversarial perturbations as a form of data augmentation to improve NeRF performance (Chen et al. 2022). Others explored incorporating adversarial objectives into the NeRF training process to enhance reconstruction quality (Niemeyer and Geiger 2021); 2) Another direction involves examining the robustness of NeRF models by subjecting them to corrupted images (Wang et al. 2023), focusing on understanding how NeRF responds when images used during training are artificially distorted; 3) Certain works have centered on *attacking* the NeRF model itself or its downstream counterparts (Fu et al. 2023; Dong et al. 2022). For instance, Viewfool (Dong et al. 2022) identifies adversarial viewpoints that result in the rendered images being misclassified by downstream im-

age classifiers. NeRFool (Fu et al. 2023) aims to attack the Generalizable NeRF (GNeRF) (Yu et al. 2021; Wang et al. 2021) directly, introducing adversarial perturbations to deceive GNeRF's scene feature prediction for a specific target view.

Our study takes a novel approach by delving into the adversarial robustness of NeRF from a distinct vantage – that of attacking its *downstream classifier*. Diverging from the conventional straegy of seeking adversarial *viewpoints*, our method revolves around computing adversarial 3D representations. Subsequently, we exploit these representations to modify the training data of the NeRF model. This innovative approach allows us to investigate how the model reacts to these modifications, thereby shedding light on NeRF's vulnerability to adversarial perturbations and its potential implications for downstream classification tasks.

It's worth noting that while there are related studies involving adversarial attacks against multiview models (Sun and Sun 2021; Yao et al. 2020) and adversarial attacks on 3D point clouds (Xiang, Qi, and Li 2019; Hu, Liu, and Hu 2022), these research directions are distinct in their focus. Adversarial attacks against multiview models typically concentrate on attacking the model at *fixed* view angles, often without delving into the implications for *unexplored* views. On the other hand, attacks on 3D point clouds primarily focus on spatial attacks, usually considering shape manipulation and not engaging with the rendering and multiview contexts. Therefore, our work addresses a specific and novel aspect of adversarial robustness within the NeRF framework.

Method

Preliminary. Firstly, we provide fundamental mathematical definitions for adversarial attacks, NeRF rendering, and training. These definitions lay the groundwork for the subsequent discussion of our specific problem formulation.

Adversarial Attacking. Consider a classifier network, denoted as $f : \mathcal{I} \to \mathbb{R}^c$, where f maps input images $I \in \mathcal{I}$ to a logit vector $\mathbf{y} = f(I) \in \mathbb{R}^c$. Here, $\mathcal{I}$ signifies the image space and c stands for the number of classes. The classifier's prediction function denoted as $\phi : \mathcal{I} \to [c] \equiv \{1, \dots, c\}$, assigns an input image to the class label with the highest probability as follows:

$$\phi(I) \equiv \arg\max_{k \in [c]} f(I)_k. \tag{1}$$

Let $l_g \in [c]$ denote the ground truth label. If $\phi(I) = l_g$, the prediction is considered correct.

In the context of adversarial attacks, the objective is to achieve either untargeted or targeted misclassification. An untargeted attack aims to achieve misclassification without any specific requirement on the incorrect class, *i.e.*, $\phi(I_{adv}) \neq l_g$. While a targeted attack aims to manipulate the classifier to predict the image as a given class l_t, *i.e.*, $\phi(I_{adv}) = l_t$. Another critical aspect of adversarial examples is their imperceptibility, often evaluated using L_p norms.

Neural Radiance Fields (NeRF). Let $F : (\mathbf{x}, \mathbf{d}) \to (\mathbf{c}, \tau)$ denote a continuous volumetric radiance field, where F is approximated by a multi-layer perceptron (MLP) which takes a 3D location $\mathbf{x} \in \mathbb{R}^3$ and a unit-norm viewing direction $\mathbf{d} \in \mathbb{R}^3$ as inputs, and gives RGB color $\mathbf{c} \in [0, 1]^3$ and a volume density $\tau \in \mathbb{R}^+$ as outputs. Each 2D pixel on the image plane corresponds to a camera ray $\mathbf{x} = \mathbf{r}(t) := \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$ is the camera center and t is the ray depth. Thus, we can approximate the color of this pixel as

$$\hat{C}(\mathbf{r}, F) := \sum_{i=1}^{N} T(t_i) \cdot \alpha(\tau(t_i) \cdot \delta_i) \cdot \mathbf{c}(t_i) \tag{2}$$

$$T(t_i) := \exp(-\sum_{j=1}^{i-1} \tau(t_j) \cdot \delta_j) \tag{3}$$

where $\{t_i\}_{i=1}^N$ is a set of quadrature points randomly selected by stratified sampling, $\alpha(x) := 1 - \exp(-x)$, $\delta_i := t_{i+1} - t_i$ is the distance between two adjacent points and $\mathbf{c}(t_i)$ and $\tau(t_i)$ are the color and density at $\mathbf{r}(t_i)$. Then, we apply 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

$$\mathcal{L}_{rgb}(\mathcal{R}, F) := \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{C}(\mathbf{r}, F) - C(\mathbf{r})\|_2^2 \tag{4}$$

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

Problem Formulation

Consider a downstream classifier network $f : \mathcal{I} \to \mathbb{R}^c$ that maps input images $I(\hat{C}, v) := \cup_{\mathbf{r} \in \mathcal{R}_v}(\hat{C}(\mathbf{r}, F)) \in \mathcal{I}$ rendered by NeRF F to a logit vector $\mathbf{y} = f(I(\hat{C}, v)) \in \mathbb{R}^c$, where $\mathcal{I}$ is the image space, $\mathcal{R}_v$ is a set of rays correspond to a given camera view v, and c is the number of classes. The adversarial NeRF F_{adv} satisfies that

$$\max \ \mathbb{P}_{\forall v}(\phi(I(\hat{C}(\mathbf{r}, F_{adv}), v)) \neq l_g) \tag{5}$$

$$\text{subject to } \|\hat{C}(\mathbf{r}, F_{adv}) - C(\mathbf{r})\|_\infty < \epsilon \tag{6}$$

where $l_g \in [c]$ is a ground truth label and ϵ is the parameter controlling the magnitude of the perturbation. By solving (5-6), we aim to find an adversarial NeRF F_{adv} which reconstructs the 3D object as close as the original one but misleads the following classification task with most images rendered by NeRF from different views.

Transformation

It is hard to directly calculate F_{adv} according to (5-6). Thus, we approximate F with a mapping function $M : C(\mathbf{r}) \to \mathbf{x}$ and $M' : \mathbf{x} \to C(\mathbf{r})$. In this way, we modify the images to mislead the classifier and then learn F_{adv} according to these adversarial images.

Pixels to 3D Representation. We assume the 3D objects are opaque, *i.e.* along a given ray, only one spatial point with maximum dense matters for NeRF, while other sampling points with less dense can be ignored. We can map

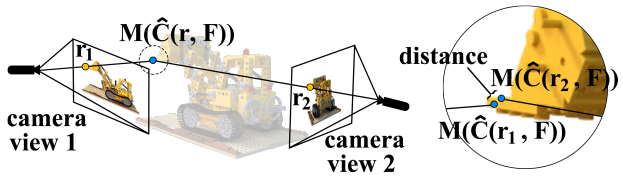

Figure 1: When transforming pixels from two images with the same nearest neighbor in 3D representation (left), the world coordinates of the two pixels are not perfectly aligned (right).

each pixel $C(\mathbf{r})$ to a 3D location $\mathbf{x}$ with NeRF model as

$$M(\hat{C}(\mathbf{r}, F)) := \mathbf{o} + (\arg\max_{t_i} h(t_i))\,\mathbf{d} \tag{7}$$

$$h(t_i) := T(t_i) \cdot \alpha(\tau(t_i) \cdot \delta_i) \cdot \mathbf{c}(t_i). \tag{8}$$

Based on the given p camera views, we maintain a 3D points set X with $p \times w \times h$ points to represent the 3D objective approximately from corresponding p images by

$$X := \cup_{\mathbf{r}\in\mathcal{R}_p}(M(\hat{C}(\mathbf{r}, F))), \tag{9}$$

where $\mathcal{R}_p$ is a set of rays corresponding to selected p camera views and each view contains $w \times h$ rays.

3D Representation to Pixels. The 3D point of unseen pixel $C(\mathbf{r})$ is approximated by $\mathbf{x} = M(\hat{C}(\mathbf{r}, F))$. Its pixel value can be simply estimated by finding its nearest neighbor in the 3D points set X. However, it is obviously inaccurate. As shown in Figure 1, the corresponding 3D points of the unseen pixel probably do not well align with any existing 3D points of X. To achieve a better approximation, we weigh the K nearest neighbor of the target point according to their distance:

$$M'(X, \mathbf{x}) := \sum_{\mathbf{x}'\in\mathcal{N}_\mathbf{x}} \frac{w_{\mathbf{x}'}}{\sum_{\mathbf{x}'\in\mathcal{N}} w_{\mathbf{x}'}}\mathbf{x}' \tag{10}$$

$$w_{\mathbf{x}'} := \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\frac{(\|\mathbf{x}' - \mathbf{x}\|_2 - \mu)^2}{\sigma^2}\Big) \tag{11}$$

Algorithm 1: NeRFail-S Attack

Input: X: 3D points constructed with p camera views according to (9)
Input: $\mathcal{I}$: images from a view set $\mathcal{V}$
Input: l_g: true label, K: maximum iterations
Input: α, ϵ: step size
Output: $\mathcal{Y}_K$

1: $\mathbf{z} \leftarrow \mathbf{0}$
2: **while** $i < K$ **do**
3: **while** $v \in \mathcal{V}$ **do**
4: $X_v \leftarrow \cup_{\mathbf{r}\in\mathcal{R}_v}(M(\hat{C}(\mathbf{r}, F)))$
5: $I'_v \leftarrow \mathbf{clip}(I(M'(X, \mathbf{z}), v) + I_v, 0, 1)$
6: $\Delta z \leftarrow \mathbf{sign}(\nabla_\mathbf{z}\mathbf{CrossEntropy}(f(I'_v)))$
7: $\mathbf{z} \leftarrow \mathbf{z} + \alpha\Delta z$
8: $\mathbf{z} \leftarrow \mathbf{clip}(\mathbf{z}, -\epsilon, \epsilon)$
9: **end while**
10: **end while**

where $\mathcal{N}_\mathbf{x}$ is the K nearest neighbor of $\mathbf{x}$ in X, the weight $w_{\mathbf{x}'}$ indicates the influence of neighbor $\mathbf{x}'$ to $\mathbf{x}$ and σ and μ are the mean and variance value of the distribution of the distance between points from X and $\mathbf{x}$. We assume the distribution of the distance between points follows a normal distribution, *i.e.* $\mu = 0$ because the neighbor whose distance toward $\mathbf{x}$ is zeros should gain the largest weight.

Multiview Adversarial Perturbation

Since we have the mapping function M and M' between pixel values and 3D points, we reformulate (5-6) as

$$\max \;\; \mathbb{P}_{\forall v}(\phi(I(M'(X, \mathbf{z}), v) + I_v) \neq l_g) \tag{12}$$

$$\text{subject to } \|M'(X, \mathbf{z})\|_\infty < \epsilon \tag{13}$$

where $\mathbf{z}$ is the adversarial perturbation we add to the 3D points and I_v is the clean image of camera view v. To tackle this problem, we propose NeRFail-S and NeRFail.

NeRFail-S: Simple Attack. As depicted in Algorithm 1, a simple solution for (12-13) uses Iterative Gradient Sign Method (IGSM) (Kurakin, Goodfellow, and Bengio 2016) to attack each image rendered by NeRF from different views along the gradient with the given step size α once, regardless of whether it is adversarial, and accumulate the adversarial changes with given distortion budget ϵ on 3D points.

NeRFail: Attack Targeting Optimality. To seek such 3D perturbation $\mathbf{z}$ such that $\|M'(X, \mathbf{z})\|_\infty < \epsilon$, we propose to estimate it iteratively over the different camera views. At each iteration, the minimal perturbation Δz_i is calculated to send the current perturbed images across the decision boundary of the classifier and aggregated to the current 3D perturbation $\mathbf{z}$ as universal 3D perturbation towards different camera views. If the image of the current camera view with 3D perturbation $\mathbf{z}$ is not adversarial, we estimate the extra perturbation Δz_i by tackling the problem:

$$\Delta z_i := \arg\min_\zeta \|\zeta\|_2^2 \tag{14}$$

$$\begin{aligned}\text{subject to } \; & \phi(I(M'(X, \mathbf{z}), v) + I_v) \\ & \neq \phi(I(M'(X, (\mathbf{z} + \zeta)), v) + I_v).\end{aligned} \tag{15}$$

To solve (14-15), we assume the decision boundary of the classifier is linear as deepfool (Moosavi-Dezfooli, Fawzi, and Frossard 2016), thus we estimate Δz_i as

$$l_t = \arg\min_{k\in[c]\wedge k\neq l_g} \frac{|f(I_v)_k - f(I_v)_{l_g} - m_1|}{\|\nabla f(I_v)_k - \nabla f(I_v)_{l_g}\|_2} \tag{16}$$

$$\Delta z_i = \frac{|f(I_v)_{l_t} - f(I_v)_{l_g} - m_1 - m_2|}{\|\Delta_t\|_2^2}\Delta_t, \tag{17}$$

i.e. finding the closest decision boundary to cross, in which $\Delta_t := \nabla f(I_v)_{l_t} - \nabla f(I_v)_{l_g}$, m_1 is a margin we add to the decision boundary to make the adversarial perturbation more powerful. I_v notes the image of the current camera view v with 3D perturbation $\mathbf{z}$. Similar to line 5 of NeRFail-S, we also use clipping to ensure the adversarial images are legitimate.

	DA-T	DA-V	RN-T	RN-V
Inception v3				
IGSM	59.3 %	0.0 %	0.4 %	0.8 %
NeRFail-S	**72.6 %**	**71.5 %**	**66.5 %**	**61.4 %**
UAP	46.8 %	28.3 %	2.9 %	2.5 %
NeRFail	**66.0 %**	**38.0 %**	**50.2 %**	**42.4 %**
VIT-B/16				
IGSM	**99.9 %**	0.0 %	2.9 %	2.4 %
NeRFail-S	99.0 %	**100.0 %**	**97.2 %**	**99.0 %**
UAP	**100.0 %**	**97.3 %**	2.2 %	1.4 %
NeRFail	98.9 %	77.3 %	**80.6 %**	**77.9 %**

Table 1: Quantitive comparison between our multiview attack and baseline over attack success rate (ASR) under different scenarios on Inception V3 and VIT-B/16.

Ground Truth	DA-T	DA-V	RN-T	RN-V
hotdog	79.0 %	32.0 %	65.5 %	47.0 %
materials	96.5 %	44.0 %	74.0 %	56.0 %
mic	100.0 %	88.0 %	100.0 %	99.0 %
ship	98.5 %	17.0 %	32.0 %	16.0 %

Table 2: Performance of NeRFail on different objects. p: 3; ϵ: 32; m_1: 8; m_2: 100; Classifier: inception v3.

Experiments

In this section, we begin by presenting the basic setting for the experiments followed by showcasing the effectiveness of our method in comparison to its 2D counterparts. We proposed four different settings to measure the performance of NeRFail and NeRFail-S attacks as evasion attacks and poisoning attacks. Subsequently, we provide a comprehensive analysis of our attack method through ablation studies and the performance on targeted attack. Due to the limited space, we have included a portion of the ablation studies and transferability experiments in the supplementary material.

Dataset. We use the dataset* presented in the original NeRF paper (Mildenhall et al. 2021) 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.

Classifier	DA-T	DA-V	RN-T	RN-V
VGG-16	61.5 %	40.0 %	47.0 %	46.0 %
AlexNet	92.5 %	80.0 %	88.5 %	80.0 %
Resnet-50	100.0 %	100.0 %	100.0 %	100.0 %
EN-B0	97.0 %	86.0 %	94.0 %	86.0 %
MN-v2	97.0 %	12.0 %	28.0 %	23.0 %

Table 3: Performance of NeRFail on different classifiers. p: 3; ϵ: 32; m_1: 8; m_2: 100; Ground truth: Lego.

*https://github.com/bmild/nerf

	ϵ	8	16	32
DA-T	IGSM	**84.5%**	**87.0%**	84.5 %
	NeRFail-S	1.5 %	43.5 %	**100.0 %**
	UAP	0.0%	14.5%	93.0 %
	NeRFail	**6.0 %**	**100.0 %**	**100.0 %**
DA-V	IGSM	0.0%	0.0%	0.0%
	NeRFail-S	**2.0 %**	**58.0 %**	**100.0 %**
	UAP	0.0%	7.0%	72.0 %
	NeRFail	**2.0%**	**38.0 %**	**89.0 %**
RN-T	IGSM	0.0%	0.0%	1.0%
	NeRFail-S	**2.0 %**	**41.0 %**	**100.0 %**
	UAP	0.0%	2.5%	8.0 %
	NeRFail	**5.0 %**	**63.0 %**	**96.5 %**
RN-V	IGSM	0.0%	0.0%	1.0%
	NeRFail-S	**2.0 %**	**37.0 %**	**100.0 %**
	UAP	0.0%	1.0%	6.0 %
	NeRFail	**2.0 %**	**52.0 %**	**92.0 %**

Table 4: Ablation on ϵ *vs*. ASR for IGSM, UAP, NeRFail-S, and NeRFail. Maximum iterations: 100; α: 2; p: 3; m_1: 8; m_2: 100. classifier: Inception v3.

Model. We use the vanilla NeRF (Mildenhall et al. 2021) to render the images, trained by $200,000$ epochs, the codebase of PyTorch Nerf†. For downstream classification, we choose the architecture including the CNN-based Inception V3 (Szegedy et al. 2016), VGG-16 (Simonyan and Zisserman 2014), AlexNet (Krizhevsky, Sutskever, and Hinton 2012), ResNet-50 (He et al. 2016), EfficientNet-B0 (EN-B0) (Tan and Le 2019), MobileNet-v2 (MN-v2) (Sandler et al. 2018) and the Transformer-based ViT-B/16 (Dosovitskiy et al. 2020). We resize images to 299×299 for the CNN-based model and 224×224 for the transformer-based model. Then we train these networks on training data of all different classes and reach decent accuracy on the testing set, *i.e*. 99.4% for Inception v3, 99.3% for VGG16, 100.0% for AlexNet, 99.6% for Resnet-50, 99.6% for EN-B0, 100.0% for MN-v2 and 97.7% for ViT-B/16. For special classes, *e.g*. Lego, the accuracy of all classifiers is 100.0%. Furthermore, on the images rendered by NeRF, the accuracy is above 99.0% on Lego.

Attacks. We select the Universal Adversarial Perturbation (UAP) (Moosavi-Dezfooli et al. 2017) as a foundational baseline for our experimentation. This approach is executed over a maximum of 100 iterations. To align this algorithm with our method, we introduce same parameters, including the confidence parameter $m_1 = 8$ and the acceleration parameter $m_2 = 100$. Additionally, we adopt IGSM (Kurakin, Goodfellow, and Bengio 2016) as another benchmark for our simple attack scenario, with 100 iterations, $\alpha = 2$, and utilizing the L_∞ norm for the ϵ parameter in both IGSM and our simple attack. For our specific attack experiment, we consider a default of $p = 3$ selected camera views.

†https://github.com/yenchenlin/nerf-pytorch/

	p	2	3	4
DA-T	NeRFail-S	31.5 %	43.5 %	43.5 %
	NeRFail	36.5 %	100.0 %	100.0 %
DA-V	NeRFail-S	47.0 %	58.0 %	64.0 %
	NeRFail	19.0 %	38.0 %	33.0 %
RN-T	NeRFail-S	26.0 %	41.0 %	43.0 %
	NeRFail	40.5 %	63.0 %	53.5 %
RN-V	NeRFail-S	21.0 %	37.0 %	36.0 %
	NeRFail	26.0 %	52.0 %	48.0 %

Table 5: Ablation on p *vs*. ASR for NeRFail-S and NeRFail. Maximum iterations: 100; α: 2; ϵ: 16; m_1: 8; m_2: 100. classifier: Inception v3.

	m_1	0	4	8	16
DA-T	UAP	54.0 %	85.0 %	93.0 %	73.0 %
	NeRFail	98.5 %	100.0 %	100.0 %	96.0 %
DA-V	UAP	40.0 %	67.0 %	72.0 %	68.0 %
	NeRFail	65.0 %	84.0 %	89.0 %	79.0 %
RN-T	UAP	2.0 %	8.0 %	8.0 %	5.0 %
	NeRFail	73.5 %	89.0 %	96.5 %	93.0 %
RN-V	UAP	1.0 %	5.0 %	6.0 %	4.0 %
	NeRFail	74.0 %	88.0 %	92.0 %	93.0 %

Table 6: Ablation on m_1 *vs*. ASR for UAP and NeRFail. Maximum iterations: 100; p: 3; ϵ: 32; m_2: 100. classifier: Inception v3.

Performance on Synthetic 3D Objects

To evaluate the performance of our **Ne**ural **R**adiance **F**ields-based multiview **a**dversar**i**al Attack (NeRFail) and its simple version NeRFail-S, we mainly evaluate the Attack Success Rate (ASR) over:

- **DA-T**: *testing* images directly attacked by the attacker;
- **DA-V**: *validation* images which the attacker does not access to;
- **RN-T**: images *rendered by NeRF*, which trained on training data with multiview adversarial perturbation, from the views of*testing* images;
- **RN-V**: images *rendered by NeRF*, which are trained on training data with multiview adversarial perturbation, from the views of *validation* images.

The performance on DA-T and DA-V shows the strength as *evasion* attacks against the images of explored and unexplored views; the performance on RN-T and RN-V shows the strength as *poisoning* attacks, *i.e*. how the adversarial NeRF deceives the model of downstream tasks.

In Table 1, it is evident that IGSM exhibits effectiveness solely on the directly attacked testing images, as it is tailored for attacking specific input instances. In contrast, UAP works on both the testing and validation images, since validation images share the same distribution as the testing set, and UAP identifies shared vulnerabilities within this dataset's distribution. However, neither the attacking of

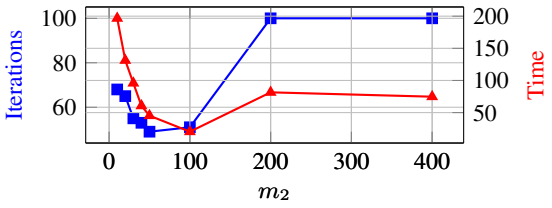

Figure 2: Compare Itereations and Time (h) needed with different m_2 for NeRFail on Inception V3.

IGSM nor UAP can be inherited and absorbed by NeRF (*cf*. RN-T and RN-V of Table 1), implying that they are unable to perturb NeRF's learning process. Visual analysis, depicted in Figure 3, reveals that both IGSM and UAP introduce perturbations not only to the objects of interest but also to the background. In contrast, our multiview adversarial perturbations exclusively target the object, rendering them more focused (*cf*. RN-T and RN-V of Figure 3).

Overall, both our methods, NeRFail and NeRFail-S, demonstrate outperformance compared to the baseline attack when evaluated on the Inception-v3 model. However, on VIT-B/16, UAP outperforms NeRFail in the testing and validation images, while our methods exhibit better results in other scenarios. UAP adds perturbations in the background, that make it stronger than NeRFail as an evasion attack. NeRFail-S consistently performs better than NeRFail, achieving a 100% Attack Success Rate (ASR). Nonetheless, as illustrated in Figure 3, NeRFail-S tends to generate more noticeable perturbations than NeRFail. Perturbation magnitude also concurs this observation: NeRFail-S produces larger perturbations (554036.0 L_0 norm and 12566.1 L_2 norm) compared to NeRFail (544276.6 L_0 norm and 10395.0 L_2 norm) for the same distortion budget ($\epsilon = 32$).

In the case of other classes, we conducted experiments similar to Table 1, and we noted similar trends (refer to Table 2). Furthermore, we extended our attack evaluation to encompass a broader range of network architectures (as detailed in Table 3). Our assessments encompassed VGG-16 (Simonyan and Zisserman 2014), AlexNet (Krizhevsky, Sutskever, and Hinton 2012), ResNet-50 (He et al. 2016), EfficientNet-B0 (EN-B0) (Tan and Le 2019), and MobileNet-v2 (MN-v2) (Sandler et al. 2018). Given our emphasis on imperceptible adversarial perturbations, we primarily present the results for NeRFail in this context, while additional experiments for NeRFail-S are available in the supplementary section.

Ablation

Both NeRFail and NeRFail-S share two parameters: *i.e*. the distortion budget ϵ and the number of selected camera view p. For our ablation experiments, we kept the remaining parameters of the attacks fixed and focused on investigating the influence of these two parameters.

First, we examined the impact of ϵ, as illustrated in Table 4. It is noteworthy that when the distortion budget ϵ is small, such as $\epsilon = 8$, IGSM demonstrates competent performance on the DA-T task. Our methods tend to exhibit be-

Figure 3: Visualization of adversarial examples generated by IGSM, UAP, NeRFail-S, and NeRFail with $\epsilon = 32$.

Target Class	DA-T	DA-V	RN-T	RN-V
drums	99.0 %	86.0 %	81.0 %	76.0 %
hotdog	65.5 %	52.0 %	52.5 %	51.0 %
materials	3.5 %	1.0 %	3.0 %	2.0 %

Table 7: Performance of NeRFail on targeted attacks. Maximum iterations: 100; p: 3; ϵ: 32; m_1: 8; m_2: 100; classifier: Inception v3; Source class: Lego.

havior more akin to UAP in this scenario. Specifically, with a very small ϵ, the overall performance of the attacks remains poor. However, as ϵ is increased to 16, the Adversarial Success Rate (ASR) of our attacks surpasses 35%. Whereas when ϵ is set to 32, our attacks get an ASR exceeding 90%.

Additionally, we assessed the impact of the number of selected camera views, denoted as p (definition *cf*. (9)), as depicted in Table 5. An increase in p marginally enhances the performance of NeRFail-S. On the other hand, for NeRFail, there is a pattern of improvement followed by a degradation in performance with increasing p. Higher values of p potentially introduce conflicts among views, and the optimization process might become more prone to inaccuracies in the transformation. Overall, a value of 3 appears to be a reasonable choice for p.

For NeRFail, we have two more parameters, *i.e*. confidence parameter m_1 and acceleration parameter m_2. Table 6 shows the influence of m_1 towards the ASR. It shows when m_1 varies, the ASR of our attack is positively related, and 8 is the best m_1 for UAP and NeRFail. Figure 2 shows the influence of m_2 towards the speed. It shows when m_2 varies, NeRFail performs best when $m_2 = 100$. Experiments on the relation between m_2 and ASR are in the supplementary.

Performance on Targeted Attack

Previously, we mainly consider the untargeted attack. Our attacks are easily adapted to targeted attacks and we evaluate our attacks on targeted attacks. As Table 7 shows, for targeted attacks, NeRFail works well on most targeted classes but for the hard classes, *i.e*. the class whose distribution is far different from targeted classes, such as *materials*.

Conclusion

In the context of 3D tasks, we explored adversarial robustness using NeRF, focusing on multiview attacks on 3D scenes. We introduced a NeRF-based transformation that connects 3D information and 2D pixels, enabling the creation of adversarial perturbations effective across diverse viewpoints, including untrained ones. Training NeRF with data poisoned by perturbations allows for their inheritance and reconstruction from the attacked agnostic model. Contrasted with attacks like NeRFool, our method stands out in versatility, free from stringent constraints on NeRF models or classifiers, broadening its applicability across numerous scenarios.

Acknowledgements

This work received support from the National Key Research and Development Program of China (No. 2020YFB1707701) and the National Natural Science Foundation of China (Grant No. 61827810). This work also received support from DFG under grant No. 389792660 as part

of TRR 248* and VolkswagenStiftung as part of Grant AZ 98514†. Wenxiang Jiang was funded by the China Scholarship Council.

References

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

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

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Cao, Y.; Wang, N.; Xiao, C.; Yang, D.; Fang, J.; Yang, R.; Chen, Q. A.; Liu, M.; and Li, B. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*, 176–194. IEEE.

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

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

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Driess, D.; Schubert, I.; Florence, P.; Li, Y.; and Toussaint, M. 2022. Reinforcement Learning with Neural Radiance Fields. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 16931–16945. Curran Associates, Inc.

Fridovich-Keil and Yu; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance Fields without Neural Networks. In *CVPR*.

Fu, Y.; Yuan, Y.; Kundu, S.; Wu, S.; Zhang, S.; and Lin, Y. 2023. NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations. *arXiv preprint arXiv:2306.06359*.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hingun, N.; Sitawarin, C.; Li, J.; and Wagner, D. 2022. REAP: A Large-Scale Realistic Adversarial Patch Benchmark. *arXiv preprint arXiv:2212.05680*.

Hu, Q.; Liu, D.; and Hu, W. 2022. Exploring the Devil in Graph Spectral Domain for 3D Point Cloud Attacks. *arXiv preprint arXiv:2202.07261*.

Hu, W.; and Tan, Y. 2022. Generating adversarial malware examples for black-box attacks based on GAN. In *International Conference on Data Mining and Big Data*, 409–423. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv:1607.02533*.

Lindell, D. B.; Martel, J. N.; and Wetzstein, G. 2021. Autoint: Automatic integration for fast neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14556–14565.

Liu, H.; Ji, R.; Li, J.; Zhang, B.; Gao, Y.; Wu, Y.; and Huang, F. 2019. Universal adversarial perturbation via prior driven uncertainty approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2941–2949.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15315–15324.

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

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Mopuri, K. R.; Ganeshan, A.; and Babu, R. V. 2018. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10): 2452–2465.

Mopuri, K. R.; Ojha, U.; Garg, U.; and Babu, R. V. 2018. Nag: Network for adversary generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 742–751.

*CPEC:https://perspicuous-computing.science

†EIS:https://explainable-intelligent.systems

Mu, R.; Ruan, W.; Marcolino, L. S.; and Ni, Q. 2022. 3DVerifier: efficient robustness verification for 3D point cloud models. *Machine Learning*, 1–28.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

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

Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Park, W.; Liu, N.; Chen, Q. A.; and Mao, Z. M. 2021. Sensor adversarial traits: Analyzing robustness of 3d object detection sensor fusion models. In *2021 IEEE International Conference on Image Processing (ICIP)*, 484–488. IEEE.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Rebain, D.; Jiang, W.; Yazdani, S.; Li, K.; Yi, K. M.; and Tagliasacchi, A. 2021. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14153–14161.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, X.; and Sun, S. 2021. Adversarial robustness and attacks for multi-view deep models. *Engineering Applications of Artificial Intelligence*, 97: 104085.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

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

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

Xiang, C.; Qi, C. R.; and Li, B. 2019. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9136–9144.

Yao, P.; So, A.; Chen, T.; and Ji, H. 2020. Multiview-Robust 3D Adversarial Examples of Real-world Objects. In *CVPR 2020 Workshop*.

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

Zhang, H.; Avrithis, Y.; Furon, T.; and Amsaleg, L. 2020a. Smooth adversarial examples. *EURASIP Journal on Information Security*, 2020(1): 1–12.

Zhang, H.; Avrithis, Y.; Furon, T.; and Amsaleg, L. 2020b. Walking on the edge: Fast, low-distortion adversarial examples. *IEEE Transactions on Information Forensics and Security*, 16: 701–713.

Zhang, H.; Furon, T.; Amsaleg, L.; and Avrithis, Y. 2022. Deep Neural Network Attacks and Defense: The Case of Image Classification. *Multimedia Security 1: Authentication and Data Hiding*, 41–75.

Zhao, Z.; Zhang, H.; Li, R.; Sicre, R.; Amsaleg, L.; and Backes, M. 2022. Towards Good Practices in Evaluating Transfer Adversarial Attacks. *arXiv preprint arXiv:2211.09565*.