Demons Hidden in the Light: Unrestricted Adversarial Illumination Attacks

Kaibo Wang, Yanjiao Chen(✉), and Wenyuan Xu

College of Electrical Engineering, Zhejiang University, Hangzhou, China
{kaibo,chenyanjiao,wyxu}@zju.edu.cn

Abstract. As deep learning-based computer vision is widely used in IoT devices, it is especially critical to ensure its security. Among the attacks against deep neural networks, adversarial attacks are a stealthy means of attack, which can mislead model decisions during the testing phase. Therefore, the exploration of adversarial attacks can help to understand the vulnerability of models in advance and make targeted defense.

Existing unrestricted adversarial attacks beyond the ℓ_p norm often require additional models to be both adversarial and imperceptible, which leads to a high computational cost and task-specific design. Inspired by the observation that models exhibit unexpected vulnerability to changes in illumination, we develop Adversarial Illumination Attack (AIA), an unrestricted adversarial attack that imposes large but imperceptible alterations to the image.

The core of the attack lies in simulating adversarial illumination through Planckian jitter, of which the effectiveness comes from a causal chain where the attacker misleads the model by manipulating the confusion factor. We propose an efficient approach to generate adversarial samples without additional models by image gradient regularization. We validate the effectiveness of adversarial illumination in the face of black-box models, data preprocessing, and adversarially trained models through extensive experiments. Experiment results confirm that AIA can be both a lightweight unrestricted attack and a plug-in to boost the effectiveness of other attacks.

Keywords: Unrestricted adversarial attacks · Adversarial illumination

1 Introduction

Deep neural networks (DNNs) are widely used in IoT devices for tasks such as vehicle data analysis [1–4] and resource allocation [5–9]. However, DNNs have been found to be vulnerable to adversarial attacks [10–12], which hinder their applications in security-critical scenarios in IoT, e.g., autonomous driving. Adversarial examples mislead the decision of DNNs in an imperceptible manner, which is conventionally achieved by bounding the modifications by

Z. Xiao et al. (Eds.): ICECI 2022, LNICST 478, pp. 111–127, 2023.
https://doi.org/10.1007/978-3-031-28990-3_9

the ℓ_p norm [13–15]. However, pixel-level similarity imposed by the ℓ_p norm is unnecessary as adversarial examples only need to be natural and unsuspicious [16,17]. Therefore, unrestricted adversarial attacks, which aim to generate visually natural perturbations, have attracted extensive attention [18–20].

Existing unrestricted adversarial attacks usually contain a model for generating adversarial noises or an additional model for concealing the noises. For instance, in unrestricted adversarial attacks, encoders are often leveraged to transform generated adversarial noises into imperceptible image styles [21,22], and perceptual models are often used as perceptual metrics to assist attackers in concealing the noises [23]. However, this model-based setting incurs extra computational cost, and the noise-concealing model is often task-specific.

To realize unrestricted adversarial attacks with a lightweight framework, we propose to use illuminations to apply large but imperceptible modifications to images. We find that images under carefully-designed illumination are more vulnerable to adversarial attacks. We explain this vulnerability from a causal perspective, i.e., a spurious association between the label and the illumination in the learning process. The attacker can simulate adversarial illumination by Planckian jitter, which makes even tiny noises hidden at the image edges sufficient to mislead the model. The adversarial illumination can be obtained by solving an optimization problem with an image gradient regularization term. Integrating the Planckian jitter and the image gradient regularization, we can obtain a lightweight unrestricted adversarial attack framework named Adversarial Illumination Attack (AIA), which can also be used as a plug-in to boost the effectiveness of other attacks. We summarize our main contributions as follows:

- We explore the vulnerability of learning models in the face of adversarial illumination noises and provide a causal explanation.
- We propose a lightweight unrestricted adversarial attack method AIA using adversarial illumination and image gradient regularization.
- We validate the effectiveness of the AIA in the face of black-box models, data preprocessing, and adversarially trained models with comprehensive experiments.

2 Related Works

2.1 Adversarial Attacks with ℓ_p-norm Constraints

Due to the nice theoretical properties of ℓ_p-norm, it is often used as a perceptibility metric for adversarial examples. The attacker either designs the adversarial noise within the ℓ_p-norm ball [11,13] or designs the successful adversarial noise with a minimum ℓ_p-norm [24]. Many algorithms are designed based on ℓ_∞-norm constraints, including single-step FGSM [11] and iterative PGD [13]. There are also algorithms aiming at minimizing ℓ_2-norm, such as L-BFGS [10] and C&W [14]. Adversarial attacks with ℓ_1-norm constraints for the sake of noise sparsity have also been studied [25]. ℓ_0-attacks target at modifying the minimum number of pixels [26,27]. Patch attacks, which are widely used in physical world attacks, can also be regarded as special ℓ_0-attacks [28,29].

2.2 Unconstrained Adversarial Attacks

As ℓ_p-norm oversimplifies the perceptual condition of adversarial examples, unconstrained attacks beyond the ℓ_p-norm constraint have been proposed recently. Unconstrained adversarial attacks can be implemented by color adjustment [20,30] and geometric transformation [31,32]. Unrestricted adversarial attacks usually adopt an extra generative model to make the modifications non-suspicious. For example, GANs [19] and encoders [21] can be used to generate natural adversarial examples; perceptual models [33] can be used as a perceptual metric to transform the generation of adversarial samples into a multi-objective optimization problem [23]; and semantic segmentation models can segment images to attack separately [34–36]. However, all these approaches require task-specific design before attacking and additional computation during attacking.

2.3 Defense Against Adversarial Attacks

As the threat of adversarial attacks becomes more prominent, defending against them is especially critical. Both data preprocessing and adversarial training are commonly used as defenses. Data preprocessing are used to corrupt adversarial noise, including JPEG compression [37], bit-depth compression [38], and autoencoder-based reconstruction [39]. Adversarial training is one of the most effective means to defend against ℓ_p-norm adversarial examples [13,40]. The variants of adversarial training may consider accuracy trade-offs [41], adversarial perturbations to weights [42], and hypersphere embedding [43]. However, these defenses are usually designed for ℓ_p-norm adversarial examples and may not work for unrestricted attacks.

3 Adversarial Illumination Attack (AIA)

Threat Model. Given an image $x \in \mathbb{R}^{CHW}$ with width W, height H and C channels and its label y, a classifier $f : \mathbb{R}^{CHW} \to \mathbb{R}^N$ will classify it as $f(x) = \arg\max_{n=1,\dots,N} f_n(x)$, where N is the number of classes. The attacker aims to find an adversarial mapping $r : \mathbb{R}^{CHW} \to \mathbb{R}^{CHW}$ so that $r(x)$ is misclassified, i.e., $f[r(x)] \neq y$. Unlike the ℓ_p-norm constrained adversarial attack that constrains $\|r(x) - x\|_p \leq \epsilon$, the unrestricted attack only requires that the semantic information of the images remains consistent. It is worth noting that semantic information consistency does not require images $r(x)$ and x to be similar since the model knows nothing about the original image x.

Attack Algorithm Overview. We design a lightweight algorithm that ensures the effectiveness of the attack with a unified model. We explicitly decouple $r(x)$ into two parts, a global transformation $I(x;\theta)$ with a few parameters to maintain the semantic information of the image and an adversarial noise δ:

$$r(x) = I(x;\theta) + \delta. \tag{1}$$

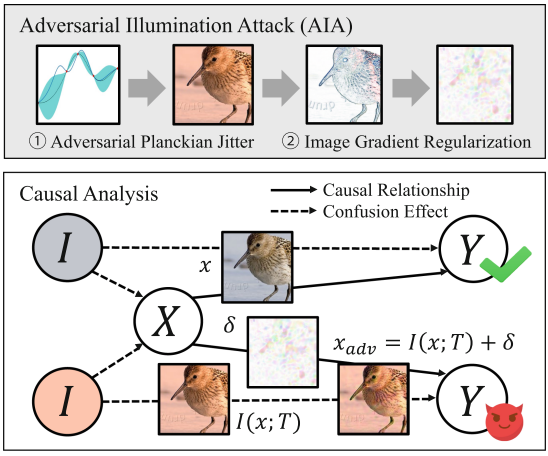


Fig. 1. The diagram of Adversarial Illumination Attacks (AIA). **Top:** Illustration of the AIA algorithm flow; **Bottom:** Causal analysis of AIA.

$I(x;\theta)$ can be seen as a "low-frequency" noise applied to the entire image, giving a large but not suspicious transformation to the image and making the model's decisions fragile. Therefore, the generation of δ can be simplified without relying on other models. As shown in Fig.1, the design of our algorithm for $I(x;\theta)$ and δ can be briefly described as:

- **Adversarial Planckian Jitter (APJ):** We note that images under different illumination conditions vary greatly and are found to potentially mislead the model's decisions, which may be due to the model's bias towards illumination during training. Therefore, we manipulate the causal relationship between images and labels by simulating adversarial illumination as $I(x,\theta)$ through adversarial Planckian jitter.
- **Image Gradient Regularization:** Model's decisions under adversarial illumination are more fragile so that the adversarial noise hidden in the image texture is sufficient to rival the effect of model-based unrestricted attacks. Therefore, we regularize the generation of adversarial noise δ by image gradients.

3.1 Adversarial Planckian Jitter

In reality, changes in lighting conditions can bring different tones to an image, such as images taken at dusk or dawn will have a warm tone, while images under artificial light may lean toward cooler tones. DNNs deployed in safety-critical scenarios, such as autonomous driving, are supposed to be invariant to illumination transformation but have been found to be unexpectedly vulnerable under carefully crafted illumination transformation. Attackers can simulate such illumination by Planckian jitter, which consists of the following steps: 1)

obtaining a spectrum of a given temperature based on the physical description of blackbody radiation; 2) converting the spectrum into an sRGB representation; 3) jittering the image illumination by making channel-wise products of this representation and the image.

According to Planck's law, a blackbody radiates a different spectrum $\sigma_T(\lambda)$ at different temperatures T, which can be expressed as

$$\sigma_T(\lambda) = \frac{2\pi hc^2}{\lambda^5[\exp(hc/kT\lambda) - 1]}, \tag{2}$$

where c, h, k are the speed of light, Planck's constant, and Boltzmann's constant, respectively.

Following the method proposed in [44], we convert the spectrum at temperature T to its sRGB representation $A(T) \in \mathbb{R}^3$, then based on which jitter the illumination of the image as:

$$I(x;T) = 0.8\, A(T) \circ x + 0.2\mu[A(T) \circ x], \tag{3}$$

where $\circ$ represents channel-wise product, $\mu(\cdot)$ is a spatial average function.

Attackers aim to find a re-illumination transformation that maximizes the adversarial of $r(x)$ by varying its temperature T. For the realism of the transformation, T is constrained to be in the interval $3 \times 10^4 K$ to $15 \times 10^4 K$. For convenience, we scale it to the interval $[0, 1]$. For the efficiency and generality of the solution, we choose δ as a ℓ_∞ bounded single-step adversarial noise, which is known as FGSM [11]. Therefore, T can be obtained by solving the optimization problem:

$$T^* = \arg\max_T \mathcal{L}\{I_x + \epsilon\, \mathrm{sign}[\nabla_{I_x} \mathcal{L}(I_x)]\} \quad s.t. \;\; T \in [0, 1], \tag{4}$$

where I_x is an abbreviation for $I(x;T)$, $\mathcal{L}(x)$ is a loss function w.r.t x.

In practice, it is found that the objective function is often multi-peaked w.r.t T, which may make the gradient-based optimization method converge to a local optimum. So we adopt the Bayesian optimization algorithm as the solver, the advantage of which is that it allows more exploration in the interval to obtain an approximate solution closer to the global optimum.

3.2 Effectiveness of AIA: A Causal Perspective

We selected the commonly used margin loss as the objective function and did preliminary experiments on CIFAR10, ImageNet, and ImageNette. We report the success rate and margin loss of FGSM attacks under adversarial illumination for three datasets in Fig.2(a). Even if only one parameter is adjusted to change the illumination of the image, the attack success rate and margin loss of FGSM attacks improve significantly.

Although the change in illumination can be seen as a shift in the data distribution, unlike a distribution that has not been seen by a model such as data

corruption [45], the model has seen different illumination during training. This vulnerability to seen distribution shifts is counterintuitive. We give an explanation from a causality perspective:

The illumination acts as a confounding factor, establishing a spurious association with the label, which can be exploited by attackers.

Using X, Y, and I to denote the image, label, and illumination, respectively, the causal map can be drawn as shown in Fig.1, where I is a confounding factor. An intuitive example is that we want the model to learn the causal mapping of lion images to labels, i.e., $X \to Y$. Whereas lions often appear in warm grassland backgrounds, the model may unthinkingly associate illumination to labels, i.e., $X \leftarrow I \to Y$, which is considered a predictive but not robust feature in [46]. By manipulating the illumination, the attacker can maximize the effect of path $I \to Y$, whose contradiction with $X \to Y$ can make the model's decision hesitant. This explains the unintended vulnerability of the model to simple attacks under adversarial illumination: attackers can manipulate the causality of both $X \to Y, I \to Y$ to lead to the wrong prediction, as shown in Fig.1.

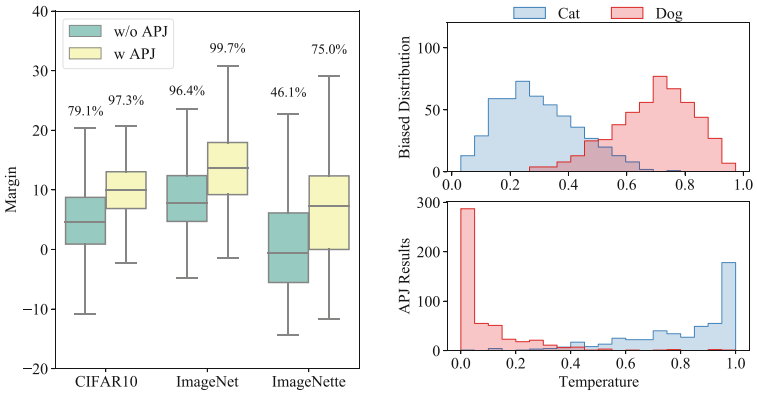
We designed a toy dog and cat classification task with biased illumination transformation to verify this idea, where the illumination temperature of the cat's picture is sampled from $\text{Beta}(3, 7)$, while the dog's is sampled from $\text{Beta}(7, 3)$, as shown in Fig.2b (top). The model can achieve a test accuracy of 94% with the same distribution. Once the illumination distributions of dogs and cats are exchanged, the model's test accuracy plummets to 68%, implying that the model exploits the causality of $I \to Y$. The results of adversarial illumination similarly validate this result, as the temperature obtained for cat's images is significantly biased towards one and vice versa, which is shown in Fig.2(b) (bottom).

3.3 Image Gradient Regularization

Optimization Problem with Regularization. The presence of adversarial illumination makes it easier to generate adversarial noise without the need for other models. Intuitively high-frequency noise added to a monochromatic surface is easily detected, while noise at the edges of the image is more subtle. In other words, the variation pattern of the noise should be consistent with the variation pattern of the image. A straightforward idea is to use the image's gradient to regularize the noise's gradient.

The image gradient can be obtained by the first- or second-order derivative operator. We use Sobel and Laplacian operators as estimators of the image gradient, and both are implemented based on convolution. Specifically, with the following convolution kernel, we can obtain three post-convolution results $G_1^{(1)}(x), G_2^{(1)}(x), G^{(2)}(x) \in \mathbb{R}^{CHW}$:

$$G_1^{(1)}(x) = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * x,\ G_2^{(1)}(x) = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * x,\ G^{(2)}(x) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} * x, \tag{5}$$



(a) (b)

Fig. 2. (a) Experimental results of FGSM attacks with and without APJ on three datasets. The box plot shows margin loss, and the success rate of the attack is annotated above. (b) **Top:** Biased light distribution applied during training of toy models. Cats: Beta(3, 7), dogs: Beta(7, 3); **Bottom:** Distribution of adversarial Planckian jitter in different categories

where $*$ denotes the convolution operation. The Sobel and Laplacian gradients $S(x), L(x)$ can be obtained as follows:

$$S(x) = \sqrt{G_1^{(1)}(x)^2 + G_2^{(1)}(x)^2}, \quad L(x) = |G^{(2)}(x)|, \tag{6}$$

where the operations here are all element-wise operations.

The regularization should provide a high loss when the image gradient is small while the noise gradient is large and a low loss for the rest of the cases, which can be expressed as:

$$\max_{\delta} \mathcal{L}(I_x + \delta; \alpha) := \mathcal{L}(I_x + \delta) - \alpha\{[1 - S(I_x)^2]S(\delta)^2 + [1 - L(I_x)^2]L(\delta)^2\} \tag{7}$$

where α is the weight of the regularization.

Warm-up and Early Stop. The regularization term is derivable, so any adversarial attack algorithm can be used to generate δ, simply by replacing the loss function $\mathcal{L}(I_x + \delta)$ with $\mathcal{L}(I_x + \delta; \alpha)$. Here we use the PGD algorithm with momentum [15] to generate the adversarial noise, where we remove the projection operation to make the noise unrestricted. To avoid generating too large noise, we limit the noise amplitude by choosing a small step ξ. In addition, we gradually increase the regularization weights as the number of iterations increases, i.e., $\alpha = \alpha_0 t$, which constrains excessive noise in the later stages and provides a direction to maximize the loss function in the earlier stages. To generate moderate adversarial noise, we stop the optimization early after the margin loss is greater than a given threshold κ. Therefore, the adversarial noise can be obtained by the following iterative algorithm:

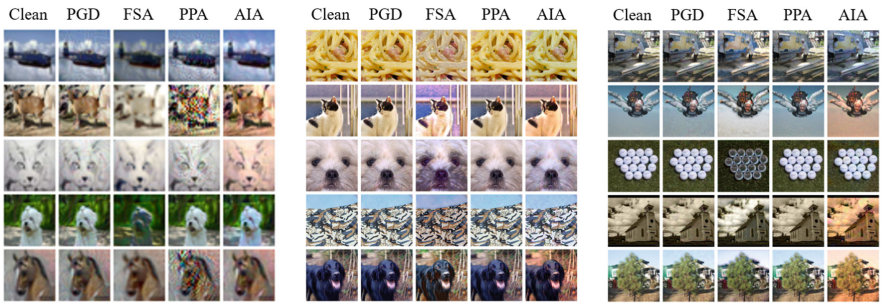
Table 1. The success rate of attacks against white-box (denoted by superscript *) and black-box models. The model architectures include ResNet (RN) [47], WideResNet (WRN) [48], DenseNet (DN) [49], MobileNetv2 (MNv2) [50] and EfficientNet (EN) [51].

CIFAR10	Target Model	WRN28*	RN18	DN161	MNv2	Avg
	PGD	100.0%	66.7%	68.5%	91.1%	81.6%
	FSA	98.9%	71.3%	68.9%	71.7%	77.7%
	PPA	100.0%	17.1%	17.1%	38.6%	43.2%
	AIA (ours)	100.0%	66.6%	68.1%	76.7%	77.8%
	PGD+APJ	100.0%	76.8%	82.2%	93.4%	88.1%
	FSA+APJ	99.8%	79.2%	77.3%	78.1%	83.6%
	PPA+APJ	100.0%	35.0%	33.5%	49.8%	54.6%
ImageNet	Target Model	RN50*	WRN50	EN	DN201	Avg
	PGD	100.0%	73.2%	55.4%	70.0%	74.6%
	FSA	99.6%	85.0%	75.8%	78.0%	84.6%
	PPA	100.0%	36.4%	34.8%	38.1%	52.3%
	AIA (ours)	100.0%	78.2%	61.3%	72.0%	77.9%
	PGD+APJ	100.0%	85.5%	69.9%	79.2%	83.7%
	FSA+APJ	99.9%	91.1%	84.2%	84.6%	90.0%
	PPA+APJ	100.0%	55.6%	46.9%	51.2%	63.4%
ImageNette	Target Model	RN50*	WRN50	EN	DN201	Avg
	PGD	100.0%	30.1%	16.0%	19.4%	41.4%
	FSA	78.0%	33.6%	25.9%	27.9%	41.4%
	PPA	98.4%	6.1%	4.1%	5.1%	28.4%
	AIA (ours)	97.2%	22.6%	14.6%	22.7%	39.3%
	PGD+APJ	100.0%	46.0%	24.1%	33.0%	50.8%
	FSA+APJ	89.1%	41.8%	36.1%	35.7%	50.7%
	PPA+APJ	98.5%	16.1%	10.1%	13.9%	34.7%

$$\delta^{(t+1)} = \begin{cases} \delta^{(t)} + \xi \text{sign}[g^{(t+1)}] & \mathcal{L}(I_x + \delta^{(t)}) < \kappa \\ \delta^{(t)} & \text{else} \end{cases}, \tag{8}$$

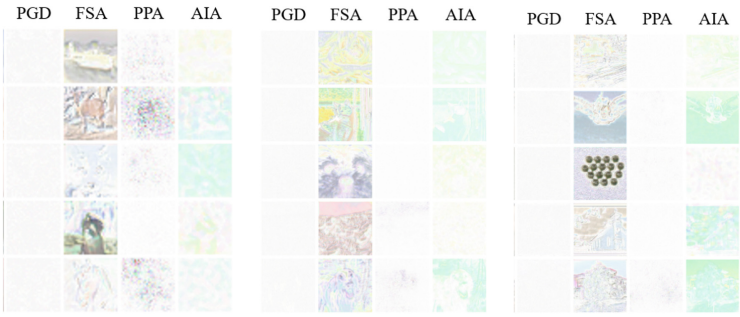
$$g^{(t+1)} = g^{(t)} + \frac{\nabla_\delta \mathcal{L}(I_x + \delta^{(t)}; \alpha_0 t)}{\|\nabla_\delta \mathcal{L}(I_x + \delta^{(t)}; \alpha_0 t)\|_1} \quad t = 1, ...T_{max}. \tag{9}$$

The convolution-based regularization imposes almost no additional burden on the gradient computation and is task-independent, requiring no specific design for different tasks.



(a) CIFAR10 (b) ImageNet (c) ImageNette

Fig. 3. Example of adversarial samples $r(x)$ generated by different attack methods.



(a) CIFAR10 (b) ImageNet (c) ImageNette

Fig. 4. Example of adversarial noise $|r(x) - x|$.

4 Experiments

4.1 Experimental Setup

We tested the performance of AIA on three datasets, including CIFAR10 [52], ImageNet [53] and Imagenette (a subset of 10 easily classified classes from Imagenet). We chose L_∞-constrained projected gradient descent (PGD) [13] as the baseline. For the unrestricted attack, we chose the encoder-based feature space attack (FSA) [21] and the perceptual model-based perceptual PGD attack (PPA) [23] as a comparison. Following the conventional settings, the ℓ_∞-norm constraint of PGD is set to 8/255, and the bounds of FSA and PPA are set to 2 and 0.5, respectively. The parameters in AIA are set to $\xi = 0.0025, \alpha_0 = 0.75\sqrt{HW}, \kappa = 15$. The number of iterations for all attacks is set to 50. We randomly selected 1000 samples in the test set for each attack and reported their attack success rates (ASR).

Table 2. Results of image quality quantification for the adversarial samples. The metrics include Peek Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), BRISQUE (BQ), and Total Variation (TV), where ↑ indicates that the higher the metric is, the better the image quality is, and vice versa.

	Attack Methods	PSNR↑	SSIM↑	BQ↓	TV↓	Avg↑
CIFAR10	PGD	31.78	0.90	58.94	0.39	0.82
	FSA	18.27	0.71	59.66	0.28	0.71
	PPA	31.90	0.94	58.99	0.38	0.83
	AIA (ours)	30.34	0.90	55.91	0.33	0.84
ImageNet	PGD	32.61	0.86	24.29	0.32	0.78
	FSA	19.02	0.72	16.34	0.27	0.77
	PPA	39.04	0.96	15.99	0.30	0.95
	AIA (ours)	35.79	0.96	17.91	0.26	0.91
ImageNette	PGD	31.78	0.90	58.94	0.39	0.66
	FSA	18.27	0.71	59.66	0.28	0.56
	PPA	31.90	0.94	58.99	0.38	0.68
	AIA (ours)	30.34	0.90	55.91	0.33	0.67

4.2 White-Box and Black-Box Adversarial Attack

We tested the effectiveness of the four attacks under white-box and transfer-based black-box settings, as shown in Table 1. We chose three models with different structures and FLOPs as the target models. It can be found that each attack in the white-box setting achieves extremely high ASR. In transfer-based black-box attacks, FSA achieves a high ASR, and its performance stems from the significant changes brought to the image style based on the encoder. Our proposed AIA achieves comparable results to FSA on CIFAR10 and better results than PGD on ImageNet without additional models. In comparison, the low transferability of PPA may be due to a focus on imperceptibility leading to convergence to model-specific noise.

We also tested the attack performance of other attacks with an adversarial illumination plug-in (+APJ). It can be seen that the success rates of the attacks with APJ are all significantly improved (∼8.2%), which means that the adversarial illumination can be used as a plug-in to enhance the transferability of other attacks by exploiting the bias of the training data, a vulnerability that is common across models.

4.3 Image Quality

Here, we show samples and noise generated by different attacks, as shown in Fig.3 and Fig.4. Encoder-based FSA brings image adversarial style transformation, but such transformation is sometimes unrealistic. Perceptual model-based PPA produces adversarial noise that is more realistic on high-resolution images but

Table 3. The success rate of the attack in the face of different data preprocessing defenses. The preprocessing defenses include: JPEG compression (JPEG) [37], Bit-Depth compression (BitDepth) [38], Blur, and Auto-Encoder Reconstruction (AE).

	Attack Methods	JPEG	BitDepth	Blur	AE	Avg
CIFAR10	PGD	78.1%	100.0%	83.0%	66.2%	85.4%
	FSA	89.2%	97.1%	94.2%	81.4%	91.0%
	PPA	26.8%	81.2%	32.5%	38.2%	47.2%
	AIA (ours)	93.9%	100.0%	99.5%	89.3%	96.5%
	PGD+APJ	93.5%	100.0%	96.2%	81.7%	94.3%
	FSA+APJ	96.4%	99.4%	97.2%	89.3%	96.1%
	PPA+APJ	56.1%	93.2%	63.7%	65.2%	72.0%
ImageNet	PGD	99.7%	100.0%	98.8%	87.2%	97.1%
	FSA	99.6%	85.0%	75.8%	78.0%	84.6%
	PPA	50.9%	69.7%	52.0%	57.4%	58.0%
	AIA (ours)	99.8%	99.8%	99.8%	94.8%	98.8%
	PGD+APJ	99.9%	100.0%	99.3%	94.3%	98.7%
	FSA+APJ	99.3%	99.1%	96.8%	92.4%	97.3%
	PPA+APJ	73.4%	89.2%	73.1%	76.4%	79.0%
ImageNette	PGD	94.9%	99.9%	98.1%	89.6%	87.0%
	FSA	68.4%	72.2%	66.6%	61.4%	61.8%
	PPA	19.7%	38.8%	27.8%	17.6%	23.4%
	AIA (ours)	95.3%	94.8%	94.4%	95.6%	87.3%
	PGD+APJ	96.9%	99.9%	98.9%	93.6%	90.4%
	FSA+APJ	81.3%	85.8%	81.2%	73.4%	74.3%
	PPA+APJ	41.9%	67.8%	58.4%	34.5%	46.5%

conspicuous on low-resolution images. The adversarial samples generated by AIA can be disguised as the result of illumination changes, which avoids adding noise in monochromatic regions like PGD by image gradient regularization.

We also quantified the image quality of the adversarial samples. In image quality metrics, different metrics have different meanings and different scales. Among them, PSNR measures the ratio of signal to noise, and a higher signal-to-noise ratio indicates that less noise is added. It is generally believed that images with PSNR greater than 40 dB have higher quality, so we rescale PSNR with PSNR/40. SSIM measures the structural similarity of the image to the original image, a higher SSIM means the modified image is more similar to the original image. SSIM takes a value between 0 and 1, so we do not scale it. TV indicates the smoothness of the image and is obtained by averaging the differences between adjacent pixel values. A lower TV indicates a smoother image and takes a value between 0 and 1, so we scale it with $1 - \text{TV}$. BRISQUE is a measure of the naturalness of an image, and a lower value indicates a higher quality image.

Table 4. Attack success rate on an adversarially trained model. The methods of adversarial training include: Hypersphere Embedding (HE) [43], Adversarial Weight Perturbation (AWP) [42], Feature Scatter (FS) [54], Robust Self-Training (RST) [40], and TRADES [41].

Defense	HE	AWP	FS	RST	TRADES	Avg
PGD	40.3%	35.8%	34.5%	34.0%	42.3%	37.4%
FSA	96.6%	96.3%	53.7%	96.4%	97.5%	88.1%
PPA	94.9%	99.1%	49.9%	98.7%	99.4%	88.4%
AIA (ours)	73.9%	82.9%	79.5%	83.7%	89.1%	81.8%
PGD+APJ	80.9%	76.4%	78.0%	75.6%	83.0%	78.8%
FSA+APJ	98.8%	99.0%	84.0%	99.0%	99.3%	96.0%
PPA+APJ	99.7%	99.9%	86.7%	100.0%	100.0%	97.3%

The value of BRISQUE varies across datasets, so we scale it with the average of BRISQUE over the dataset. Therefore, the average value can be obtained by:

$$\mathbf{Avg} = \frac{1}{4}[\frac{\mathrm{PSNR}}{40} + \mathrm{SSIM} + 1 - \mathrm{TV} + \frac{\overline{\mathrm{BRISQUE}}}{\mathrm{BRISQUE}}]. \tag{10}$$

where $\overline{\mathrm{BRISQUE}}$ is the average of BRISQUE on the dataset. Note that the average value here is only for the convenience of comprehensive comparison. The results are shown in Table.2.

We can find that the imperceptibility of AIA is comparable to that of PPA and better than that of PGD and FSA, especially in the case of unreferenced metrics, which is a more reasonable scenario of adversarial sample quality assessment.

4.4 Attack Effect Under Data Preprocessing

We selected four representative defenses to test the effectiveness of the attack under preprocessing, as shown in Table.3. We can find that AIA achieves the highest ASR in the face of preprocessing, especially in AE. This is due to the global noise brought to the images by the adversarial illumination, which is difficult to corrupt by data preprocessing, even after reconstruction. As shown in Fig.3, the more significant noise generated by AIA gains the robustness of the adversarial sample, which can also be verified from the boosted ASR (∼10.9%) of the other attacks with APJ in Table.3.

4.5 Attack Effect Facing Defense Models

Since the adversarial training was mostly evaluated on the CIFAR10, we selected five adversarially trained models on the CIFAR10 dataset to test the ASR of each attack, as shown in Table.4. It can be seen that PPA and FSA have higher attack

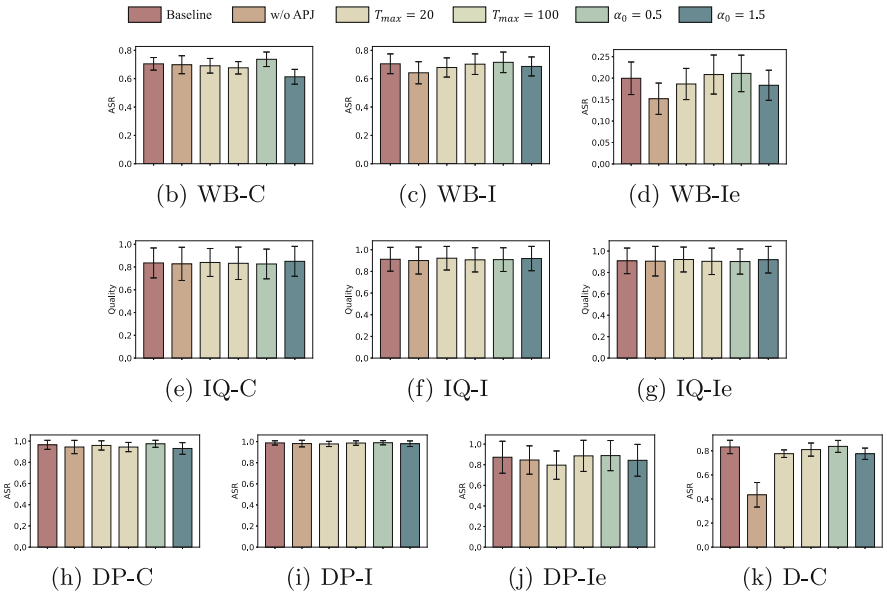


Fig. 5. Results of ablation experiments in four tasks, including white-box and black box attacks (WB), image quality (IQ), data preprocessing (DP) and defense (D). Datasets CIFAR10, ImageNet and ImageNette are abbreviated by C, I, Ie, respectively. The baseline setting is $\alpha_0 = 0.75, T_{max} = 50$ with APJ. The mean and standard deviation of different metrics were plotted.

success rates in the face of adversarially trained models, while AIA is slightly weaker, probably due to its more conservative constraint on noise. Note that APJ plays a crucial role in breaking some defenses, such as FS, and thus APJ is suitable as a complement to other attacks. We speculate that the reason is that the adversarial training model smooths the decision boundaries under normal data distribution. However, the data distribution under adversarial illumination is not considered during the adversarial training. It is also evident from the significant ASR improvement brought by APJ for PGD: adversarial training focuses too much on the ℓ_∞ neighborhoods of the training data. Simply changing the illumination of the samples is enough to break through the adversarial training model even if it is still ℓ_∞-bounded adversarial noise.

4.6 Ablation Study

we investigated the performance of AIA with different parameter settings, including with and without APJ, different number of iterations T_{max}, and different image gradient regularization weights α_0. From the results shown in Fig.5, we can find that 1) APJ is more critical to breaking through the defense model and improving ASR in the face of black-box models and data preprocessing. The

reason for the improvement is not apparent is that the attacker exploits tremendous noise, i.e., sacrificing image quality for ASR improvement. 2) AIA does not require an excessive number of iterations. Too many iterations can sometimes lead to worse ASR, possibly due to convergence to model-specific noise on the one hand and greater image gradient regularization in later stages on the other. 3) Different α_0 regulate the ASR and image quality trade-offs. The difference brought by α_0 is milder, so it is not difficult to find a moderate parameter.

On ImageNette, the effect of the attack without APJ is significantly lower, and the attack with a higher number of iterations has a higher ASR in the face of the black box model and data preprocessing. The reason for this is that ImagNette is more difficult to attack compared to ImageNet, so the APJ and more iterations which is used to improve the effectiveness are more important.

5 Conclusion

In this paper, we propose an unrestricted adversarial attack based on illumination transformation, where attackers can generate adversarial illumination by Planckian jitter and manipulate the causal relationship between illumination and labels to mislead the model. We design an unrestricted adversarial noise optimization algorithm based on this by image gradient regularization, which achieves significant attack results without additional models. Comprehensive experiments validate the effectiveness of adversarial illumination.

References

1. Xiao, Z., et al.: Trajdata: on vehicle trajectory collection with commodity plug-and-play OBU devices. IEEE Internet Things J. **7**(9), 9066–9079 (2020)
2. Li, J., et al.: Drive2friends: inferring social relationships from individual vehicle mobility data. IEEE Internet Things J. **7**(6), 5116–5127 (2020)
3. Long, W., et al.: Unified spatial-temporal neighbor attention network for dynamic traffic prediction. IEEE Trans. Veh. Technol. (2022)
4. Huang, Y., Xiao, Z., Yu, X., Wang, D., Havyarimana, V., Bai, J.: Road network construction with complex intersections based on sparsely sampled private car trajectory data. ACM Trans. Knowl. Discov. Data (TKDD) **13**(3), 1–28 (2019)
5. Jiang, H., Xiao, Z., Li, Z., Xu, J., Zeng, F., Wang, D.: An energy-efficient framework for internet of things underlaying heterogeneous small cell networks. IEEE Trans. Mob. Comput. **21**(1), 31–43 (2020)
6. Xiao, Z., et al.: Resource management in UAV-assisted MEC: state-of-the-art and open challenges. Wireless Netw. **28**(7), 3305–3322 (2022)
7. Dai, X.: Task co-offloading for d2d-assisted mobile edge computing in industrial internet of things. IEEE Trans. Ind. Inform. **19**(1), 480–490 (2022)
8. Zeng, F., Li, Q., Xiao, Z., Havyarimana, V., Bai, J.: A price-based optimization strategy of power control and resource allocation in full-duplex heterogeneous macrocell-femtocell networks. IEEE Access **6**, 42004–42013 (2018)
9. Xiao, Z., et al.: A joint information and energy cooperation framework for CR-enabled macro-femto heterogeneous networks. IEEE Internet Things J. **7**(4), 2828–2839 (2019)

10. Szegedy, C.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
12. Carlini, N., Wagner, D.: Audio adversarial examples: targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 1–7. IEEE (2018)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
14. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
15. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193 (2018)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
17. Sharif, M., Bauer, L., Reiter, M.K.: On the suitability of LP-norms for creating and preventing adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1605–1613 (2018)
18. Liu, F., Zhang, C., Zhang, H.: Towards transferable unrestricted adversarial examples with minimum changes. arXiv preprint arXiv:2201.01102 (2022)
19. Xiang, T., Liu, H., Guo, S., Gan, Y., Liao, X.: Egm: an efficient generative model for unrestricted adversarial examples. ACM Trans. Sens. Netw. (TOSN) **18**(4), 1–25 (2022)
20. Zhao, Z., Liu, Z., Larson, M.: Adversarial color enhancement: generating unrestricted adversarial images by optimizing a color filter. arXiv preprint arXiv:2002.01008 (2020)
21. Xu, Q., Tao, G., Cheng, S., Zhang, X.: Towards feature space adversarial attack by style perturbation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 10523–10531 (2021)
22. Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A.K., Yang, Y.: Adversarial camouflage: hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1000–1008 (2020)
23. Laidlaw, C., Singla, S., Feizi, S.: Perceptual adversarial robustness: defense against unseen threat models. In: ICLR (2021)
24. Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
25. Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.-J.: Ead: elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
26. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
27. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. **23**(5), 828–841 (2019)
28. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)

29. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: Dpatch: an adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299 (2018)
30. Laidlaw, C., Feizi, S.: Functional adversarial attacks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
31. Alaifari, R., Alberti, G.S., Gauksson, T.: Adef: an iterative algorithm to construct adversarial deformations. arXiv preprint arXiv:1804.07729 (2018)
32. Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 (2018)
33. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
34. Hosseini, H., Poovendran, R.: Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1614–1619 (2018)
35. Shamsabadi, A.S., Sanchez-Matilla, R., Cavallaro, A.: Colorfool: semantic adversarial colorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1151–1160 (2020)
36. Bhattad, A., Chong, M.J., Liang, K., Li, B., Forsyth, D.A.: Unrestricted adversarial examples via semantic manipulation. arXiv preprint arXiv:1904.06347 (2019)
37. Das, N.: Shield: fast, practical defense and vaccination for deep learning using jpeg compression. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 196–204 (2018)
38. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
39. Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147 (2017)
40. Raghunathan, A., Xie, S.M., Yang, F., Duchi, J., Liang, P.: Understanding and mitigating the tradeoff between robustness and accuracy. arXiv preprint arXiv:2002.10716 (2020)
41. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning, pp. 7472–7482. PMLR (2019)
42. Wu, D., Xia, S.-T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Adv. Neural. Inf. Process. Syst. **33**, 2958–2969 (2020)
43. Pang, T., Yang, X., Dong, Y., Xu, K., Zhu, J., Su, H.: Boosting adversarial training with hypersphere embedding. Adv. Neural. Inf. Process. Syst. **33**, 7779–7792 (2020)
44. Zini, S., Buzzelli, M., Twardowski, B., van de Weijer, J.: Planckian jitter: enhancing the color quality of self-supervised visual representations. arXiv preprint arXiv:2202.07993 (2022)
45. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
46. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
47. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
48. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)

49. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
50. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
51. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
52. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
53. Deng, J., et al.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
54. Zhang, H., Wang, J.: Defense against adversarial attacks using feature scattering-based adversarial training. In: Advances in Neural Information Processing Systems, vol. 32 (2019)