RFLA: A Stealthy Reflected Light Adversarial Attack in the Physical World

Donghua Wang[1], Wen Yao*[2], Tingsong Jiang*[2], Chao Li[3], and Xiaoqian Chen[2]

[1]College of Computer Science and Technology, Zhejiang University
[2]Defense Innovation Institute, Chinese Academy of Military Science
[3]School of Artificial Intelligence, Xidian University
wangdonghua@zju.edu.cn, {wendy0782,lichaoedu}@126.com, tingsong@pku.edu.cn,
chenxiaoqian@nudt.edu.cn

Abstract

Physical adversarial attacks against deep neural networks (DNNs) have recently gained increasing attention. The current mainstream physical attacks use printed adversarial patches or camouflage to alter the appearance of the target object. However, these approaches generate conspicuous adversarial patterns that show poor stealthiness. Another physical deployable attack is the optical attack, featuring stealthiness while exhibiting weakly in the daytime with sunlight. In this paper, we propose a novel Reflected Light Attack (RFLA), featuring effective and stealthy in both the digital and physical world, which is implemented by placing the color transparent plastic sheet and a paper cut of a specific shape in front of the mirror to create different colored geometries on the target object. To achieve these goals, we devise a general framework based on the circle to model the reflected light on the target object. Specifically, we optimize a circle (composed of a coordinate and radius) to carry various geometrical shapes determined by the optimized angle. The fill color of the geometry shape and its corresponding transparency are also optimized. We extensively evaluate the effectiveness of RFLA on different datasets and models. Experiment results suggest that the proposed method achieves over 99% success rate on different datasets and models in the digital world. Additionally, we verify the effectiveness of the proposed method in different physical environments by using sunlight or a flashlight.

1. Introduction

Deep neural networks (DNNs) have increasingly been applied to daily life as their dramatic capabilities, such as automatic driving, facial payment, and computer-aided diagnosis. However, DNN-based systems have exposed security risks caused by adversarial examples [39]. Adversarial examples are crafted by carefully designed noise that is invisible to humans but can deceive the DNNs. Furthermore, recent researches [34, 21] reported that physically deployed DNN-based systems are also exposed to such security risks. Therefore, exploring various potential risks in security-sensitive systems to avoid possible loss is urgent.

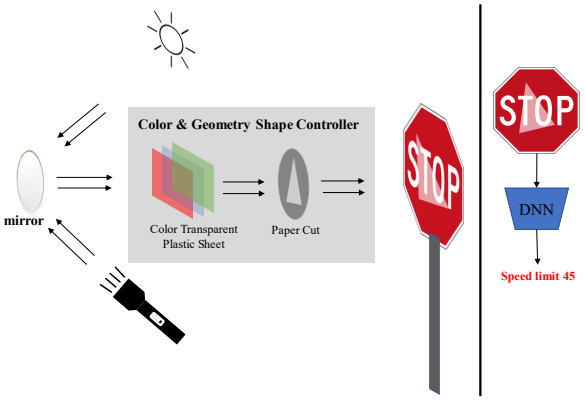


Figure 1. The reflected light is modulated by the color transparency plastic sheet and paper cut of the specific shape for better attack performance. The Reflected light source can be sunlight or a flashlight (when the sunlight is unreachable).

Existing adversarial attack methods can be categorized into digital attacks and physical attacks. The former focus on pursuing higher attack performance on limitation conditions, such as breaking the model equipped with adversarial defense [20, 3, 7, 24], preventing the attacker from accessing the target model's information (e.g., architecture or dataset), i.e., black-box attack [1, 22, 23]. Although some researchers suggested that adversarial examples generated by the digital attack can be applied to physical attacks [21],

*Corresponding Author.

the attack performance is not satisfying. The possible reason is that the adversarial perturbation is too small to resist the environmental noise in the physical world. In contrast, physical attacks are designed to be physically deployable, where one crucial change is eliminating the perturbation's magnitude constraint.

A line of physical adversarial attack methods [10, 9, 43, 44] has been proposed, which can be grouped into contact attacks and contactless attacks. The former requires the attacker to approach the target object and then modify the appearance of the target object by pasting the adversarial patch or camouflage. However, the adversarial pattern generated by these methods is conspicuous, which easily alerts humans and results in attack failure. By contrast, contactless physical attacks do not require the attacker to approach the target object while modifying the appearance of the target object by projecting or emitting light or laser on the target object to perform attacks, making it stealthy and dangerous. Optical attacks are representative contactless attacks. Although several optical attacks have been proposed [28, 32, 8, 17], they merely work in dark environments as the strong light (e.g., sunlight) environment would disturb the emitted light, limiting their usability.

In this paper, we get inspiration from the fact that the driver is easily affected by the strong reflected light, resulting in a potential car accident, and such potential risks to the automatic driving system remain unexplored. We explore the vulnerability of the DNNs toward the reflected light attack by elaborately designing the position, geometry, and color of the reflected light. Specifically, we propose a Reflected Light Attack (RFLA), which can solve the issue of the poor attack performance of existing optical attacks in strong-light environments, as our light source is sunlight. To perform physical attacks, we use a mirror to reflect the sunlight toward the target object to modify its appearance. However, the monotonous sunlight (usually white) may not obtain the desired performance. Therefore, we first use different colored transparent plastic sheets to modulate the color of the reflected light, then apply a paper cut of a specific shape to control the shape of reflected light on the target object (see Figure 1). Finally, we can create different colors and shapes of the reflected light on the target object's specific region to achieve desired attack performance.

To achieve the above goals, we present a general framework based on the circle to model the above problem. Specifically, we first initialize a circle with a random coordinate and radius. On this circle, we create a point on the circle using sine and cosine with a randomly selected angle. Then, we customize a shape by adding a new angle, which is used to create a new point in the circumference. The other points required to create a geometry can be obtained by applying the center symmetry of the circle. Moreover, the fill color and its transparency are also considered in the optimization. Finally, we adopt the particle swarm optimization (PSO) algorithm to find the optimal result. Our contributions are listed as follows.

- We propose a novel reflect-light-based physical adversarial attack under the black-box scenario. It reflects the natural sunlight toward the target object using a mirror, making it controllable and stealthy.
- We devise a general framework based on a circle to search for the best position, geometry, and color of the reflected light to achieve better attack performance.
- We comprehensively investigate the influence of the geometry, position, and color of the reflected light on attack performance in the digital world. We conduct the physical adversarial attack by using sunlight for daytime and a flashlight for sunlight unavailable, and the experiment results verify the effectiveness of the proposed method.

2. Related Works

2.1. Digital Adversarial Attacks

Digital adversarial attacks have enjoyed decade development, which can be roughly divided into white-box attack methods and black-box attack methods. The former grants the adversary access to the target model, allowing them to develop attack algorithms with the model's gradient. The most represented gradient-based attack is the fast gradient sign method (i.e., FGSM [13]), which updates adversarial examples along the ascending direction of the gradient under one iteration step. Since then, a line of variants has been proposed, including an iterative variant of FGSM (i.e., I-FGSM [21]), random initialization has been adopted (i.e., PGD [27]), momentum term is introduced to enhance the transferability (i.e., MI-FGSM [6]), and various data augmentation technique like diversity input (i.e., DI-FGSM[47]), translation-invariant (i.e., TI-FGSM [7]) and scale-invariant (i.e., SI-FGSM [24]). In contrast, black-box attacks prohibit the attacker from accessing any information about the target model but are open for queries, which makes black-box attacks more challenging. Nonetheless, many black-box attacks are proposed, such as exploiting the differential evolution algorithm [22], genetic algorithm [1], particle swarm optimization [51], and so on [2]. In addition, several works suggested that adversarial perturbation's position [45], pattern [49], and geometry [4] on the clean image significantly impact attack performance. However, current works only investigate one or two of these factors. In this work, we systematically investigate the influence of the adversarial perturbation's position, geometry, and pattern on attack performance under the black-box scenario.

2.2. Physical Adversarial Attacks

According to whether it requires the attacker to access the target object in the real attack scenario, physical adversarial attacks can be grouped into contact and contactless physical attacks. Contact attacks can be further categorized into patch-based attacks and camouflage-based attacks. Patch-based attacks mainly focus on optimizing an adversarial image patch, which is then printed out and stuck on the target object or held by the attacker to deceive the target DNNs. Patch-based attacks are usually applied in attacking the facial recognition model [34, 35, 45, 46], pedestrian detection model [41, 40, 16, 5], and traffic sign recognition model [10, 25, 51]. Camouflage-based attacks [44, 43, 9] slightly differ from patch-based, as they concentrate on modifying the appearance of the target object via UV Texture. Thus, camouflage-based attacks show better attack performance in the multi-view scenario by painting the full coverage camouflage over the appearance of the target object. However, although contact physical attacks achieve good physical attack performance, the pattern of the adversarial patch/camouflage is conspicuous, which leads to poor stealthiness. In contrast, contactless physical attacks are performed by projecting/emitting light [12, 17], or a laser beam [8], usually called optical attacks. However, existing optical attacks work in dark environments [8, 17] while performing poorly in strong-light environments. The reason is that the light beam emitted by a light source is easily affected by environmental light, resulting in attacking failure. Recently, Zhang *et al.* [51] proposed a shadow-based attack, but it can only create a triangle shape with one monotonous color (e.g., gray). In this work, we solve the situation of poor attacks in strong-light (i.e., sunlight) environments to perform attacks since we directly use sunlight to perform attacks. Moreover, we create reflected light with different geometrical shapes and colors using the color transparency plastic sheet and paper cut.

3. Methodology

3.1. Problem Statement

Let $\mathcal{X}$ denote the data distribution, and the corresponding ground truth label is $\mathcal{Y}$. Given an image $x \in \mathcal{X}$ that has the resolution of $x \in \mathbb{R}^{C\times H\times W}$, a well trained neural network f output $\hat{y} = f(x)$ and $\hat{y} = y$, where the $\hat{y}$ is prediction of the f and y is the ground truth label, $\hat{y}, y \in \mathbb{R}^{|\mathcal{Y}|}$. Adversarial attack aims to generate adversarial examples x_{adv} to make the f output the wrong prediction by adding small perturbation δ into the clean image x, i.e., $x_{adv} = x + \delta$. Mathematically, the δ is obtained by solving the following problem

$$\min \ \delta \quad s.t. \ f(x+\delta) \neq f(x), \ ||\delta||_p \leq \epsilon, \qquad (1)$$

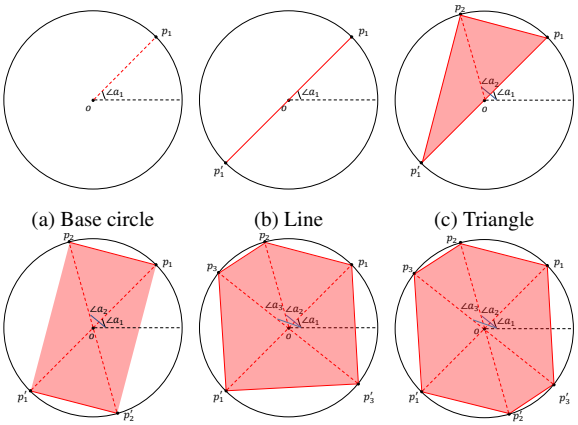


Figure 2. Visualization explanation of the circle modeling. Different geometries are constructed by adjusting the angle. For example, the triangle is created by adding a new angle $\angle a_2$ on the base circle (a), where three vertices are composed of two points (p_1 and p_2) determined by two angles ($\angle a_1$ and $\angle a_2$) and one symmetry point (p_1').

where $||\cdot||_p$ is the L_p norm, which bound the maximum allowable magnitude of δ.

The optimization objective of Equation 1 is the general formal for constructing the full-pixel-wise perturbation, which is unsuitable for physical adversarial attacks as the background of the physical world is unchangeable. Therefore, we reformulate Equation 1 to optimize the physical deployable perturbation by modifying the construction of x_{adv}. Specifically, we define an apply function $\mathcal{A}(x, p, l, M)$ to construct adversarial examples x_{adv}, which indicates that apply the perturbation p at the location l of the clean image x, where M is the binary mask to indicate whether the position is allowed to modify (one denote allow while zero not).

In this work, we aim to reflect the sunlight toward the target object to perform stealthy physical adversarial attacks, where the representation (e.g., geometry, fill color and position) of reflected light on the target object is the key to a successful attack. Therefore, the parameters of $\mathcal{A}(x, p, l, M)$ comprise geometry and fill color of p, and the location of l are variables to be optimized.

3.2. Reflected Light Attack

Sunlight is the most common and indispensable natural phenomenon in daily life. People can reflect the sunlight toward the wall to construct various shapes by using different shapes of mirrors. However, the danger of such reflected light against DNN-based systems has been ignored, which may pose a potential risk as it featured extremely stealthy and controllable. In this work, we aim to modulate the reflected light to perform adversarial attacks in the digital and physical world.

Previous work [51] modeled the triangle shadow by optimizing three points, which requires complex constraints on points to construct a rational geometry shape if they extend to other geometric shapes. To address this issue, we exploit the characteristic of the circle and propose a novel general framework based on the circle, which can generate various shapes by adjusting the number of angles (see Figure 2). The details process is described as follows.

- Select a radius r from the region of $[0, min(H, W)/2]$.
- Spawn a center $o(x, y)$ of the circle from the region of $[r, H - r]$ and $[r, W - r]$.
- Randomly select an angle a_1 and spawn a point $p_1(x_1, y_1)$ on circle by the follow equation

$$\begin{cases} x_1 = x + r \times \sin(\frac{a_1 \times \pi}{180}), \\ y_1 = y + r \times \cos(\frac{a_1 \times \pi}{180}), \end{cases} \tag{2}$$

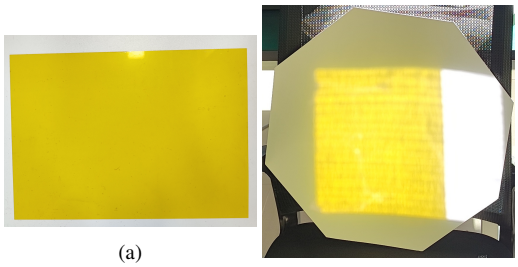
- Calculate the symmetry point $p_1'(x_1', y_1')$ of the point p_1 against the center of the circle by $x_1' = 2 \times x - x_1$ and $y_1' = 2 \times y - y_1$.
- Randomly select a color tuple $(red, green, blue)$ from the region of [0, 255], and the transparency α from [0, 1].

The above process can plot a line on the clean image. To construct varying geometries like a triangle, rectangle, pentagon, or hexagon, one can repeat the third and fourth steps to create a new point by adding a new angle. Algorithm A1 in describes the detailed particle initialization process.

3.3. Optimization

As aforementioned, we have eight base variables that need to be optimized, expressed as eight-tuples $(x, y, r, \alpha, red, green, blue, a_1)$, which can be used to plot a line on the clean image. To generate various geometries, more additional variables are required to generate various geometries, which depend on the shape to be generated. For example, there is one extra variable for the triangle and rectangle; two extra for the pentagon and hexagon. Note that the proposed method is easily extended to more complex geometry. Recall that our goal is to deceive the DNNs by plotting geometry on the clean image. Thus, we adopt the particle swarm optimization algorithm (PSO) to seek the best geometry, fill color, and position.

In PSO, we represent the optimization variables tuple as a particle (i.e., the solution vector q). The update direction of the particle is determined by a velocity vector v. Every particle stands for a potential solution, which requires to be optimized. We treat the personal historical best solution of a particle as q_{pbest}, and the global best solution of a particle as q_{gbest}. Moreover, for every solution, we fix the circle's position and merely optimize the geometry, fill color, and transparency. Therefore, to represent the best solution of a circle, we devised an additional metric q_{sgbest}, which is the sum of the fitness score of all geometrical shapes in a specific circle. Finally, the update criterion is defined as follows:



(a) (b)

Figure 3. The color discrepancy between the color transparency plastic sheet (a) and its reflected light(b).

$$\begin{aligned} v_i(t) = W v_i(t-1) + C_1 \kappa_1 (q_{pbest} - q_i(t)) \\ + C_2 \kappa_2 (q_{gbest} - q_i(t)) \\ + C_3 \kappa_3 (q_{sgbest} - q_i(t)), \end{aligned} \tag{3}$$

$$q_i(t) = q_i(t-1) + v_i(t), \tag{4}$$

where W is the inertia weight used to control the impact of the previous velocity on the current velocity. C_1, C_2, and C_3 indicate the learning factors, which balance the impact of different parts empirical on current velocity. κ_1, κ_2, and κ_3 are random values uniformly sampled from [0, 1], which are used to increase the randomness of the search.

Apart from the solution and velocity of a particle, the fitness function is crucial for optimizing in PSO algorithm. In this work, we adopt the following fitness function to evaluate each particle.

$$\min \quad F(q) = \Pr_{\hat{y}}(\mathcal{A}(x, q, M)), \tag{5}$$

where $\mathcal{A}(x, q, M)$ denotes an applied function that paints the geometry with color $(red, green, blue)$ and transparency α at the coordinate of o on the clean image x, where M is a binary mask indicates the allowed modification area. $\Pr_{\hat{y}}(\cdot)$ is the predicted label $\hat{y}$ probability of the target model f on the input. By minimizing the $F(q)$, the confidence of prediction label $\hat{y}$ gradually decreases. We stop the search until it reaches the maximum iteration or finds the adversarial example. Algorithm 1 describes the optimization process.

3.4. Physical Deployable Attack

In digital worlds, we can construct 256^3 color tuples by blending different RGB values, which is impractical in the physical world as the limitation of the device and material.

Algorithm 1 Reflected Light Adversarial Attack (RFLA)

Input: input image x, target model f, max iteration $MaxIter$

Output: Best solution q^*

$q \leftarrow Initialization()$ ▷ Algorithm A1
$x_{adv} \leftarrow GenSample(q)$ ▷ Algorithm A2
if $f(x_{adv}) \neq f(x)$ **then**
  return q^*
end if
for $itr = i, ..., MaxIter$ **do**
  Update velocity vector v by Equation 3
  Update position vector q by Equation 4
  $x_{adv} \leftarrow GenSample(q)$ ▷ Algorithm A2
  if $f(x_{adv}) \neq f(x)$ **then**
    return q^*
  end if
end for

Therefore, we constrain the search space of the color to ensure physically deployable. Specifically, we use seven color transparency plastic sheets to change the color of the reflected light. However, we find discrepancies exist between the color transparency plastic sheet and its reflected light (see Figure 3), which may lead to attack failure. To decrease such discrepancy, we collect the light color reflected by the color transparency plastic sheet and adopt it as the searched color. In such a way, we can decrease color discrepancies when performing physical attacks.

4. Experiments

4.1. Settings

Datasets: To investigate the effectiveness of the proposed method, we conduct the digital attack on the ImageNet-compatible dataset provided by the NIPS 2017 adversarial competition[1], which includes 1000 images. Moreover, two commonly used traffic sign datasets: GTSRB [37] and LISA [29], are also considered to investigate the extensibility of the proposed method.

Target models: We conduct the proposed method on two tasks: image classification and traffic sign recognition. As for image classification, we select six ImageNet pre-trained networks: ResNet50 (RN50) [15], VGG16 [36], DenseNet121 (DN121) [18], ResNeXt50 (RNX50) [48], WiderResNet50 (WRN50) [50] and SqueezeNet (SN) [19], which are all provided by PyTorch [30]. As for traffic sign recognition, we follow the setting reported in previous works [10, 25] to train GTSRB CNN and LISA CNN on GTSRB and LISA dataset, which obtains the accuracy of 95.06% and 100% on the test set, respectively.

[1]https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack

Table 1. Comparison results with the line-based method in terms of ASR (%) on ImageNet-compatible dataset. The best results are highlighted with **bold**.

	RN50	VGG16	DN121	RNX50	WRN50	SN
Bezier [11]	72 .4	**77.7**	74.1	72.7	69.6	**89.3**
RFLA-Line	**76.9**	77.4	**76.5**	**75.7**	**71.7**	89.2

Evaluation metrics: We adopt the attack success rate (ASR) as the evaluation metric, defined as the ratio of the number of the network's prediction flipped caused by adversarial examples to the total test dataset.

Implementation details: We adopt the OpenCV-Python package to plot different geometries in the clean image. For the settings of parameters of PSO, we set the max iteration number to 200, C_1, C_2, and C_3 set to 2.05, W is set to 0.7298, the particle size and the geometry number at a circle are set to 50. The particle and velocity bound are provided in Appendix A (The upper bound of the transparency α is set to 0.7 to evade occluding the clean image.). Unless otherwise specified, the mask M is set to all one value matrix in our experiments. All experiments were conducted on an NVIDIA RTX 3090ti 24GB GPU [2].

4.2. Digital Adversarial Attacks

In this section, we quantitatively and qualitatively evaluate the effectiveness of the proposed method in the digital world. For comparison, we adopt two patch-based attack methods: TPA [49] and DAPatch [4]; one line-based attack method Bezier [11]. TPA [49] utilized the feature texture image extracted from DNNs as the adversarial patch, which is pasted on the clean image, where the paste position is optimized by reinforcement learning. DAPatch [4] optimized the pattern and mask simultaneously, which can create a deformable shape adversarial patch. In contrast, Bezier [11] generated adversarial examples by scratching the bezier curve on the clean image, where the bezier curve depends on three points optimized by the optimizer (e.g., PSO). We reproduce the above three methods on the ImageNet-compatible dataset using the default settings.

4.2.1 Quantitatively Result

Table 1 reported the comparison results of the proposed RFLA-Line with the Bezier method. As we can observe, the proposed method outperforms the Bezier method on four of six models and obtains an improvement by 1.93% in average ASR, indicating the proposed method's effectiveness. Additionally, we can get an 8.95% improvement by widening the line thickness two times. Although the length of the Bezier curve may be shorter than ours, the discrepancy is trivial as modifications caused by the line is neglectable. Moreover, our method can be extended to more geometries.

[2]Code will be released after published.

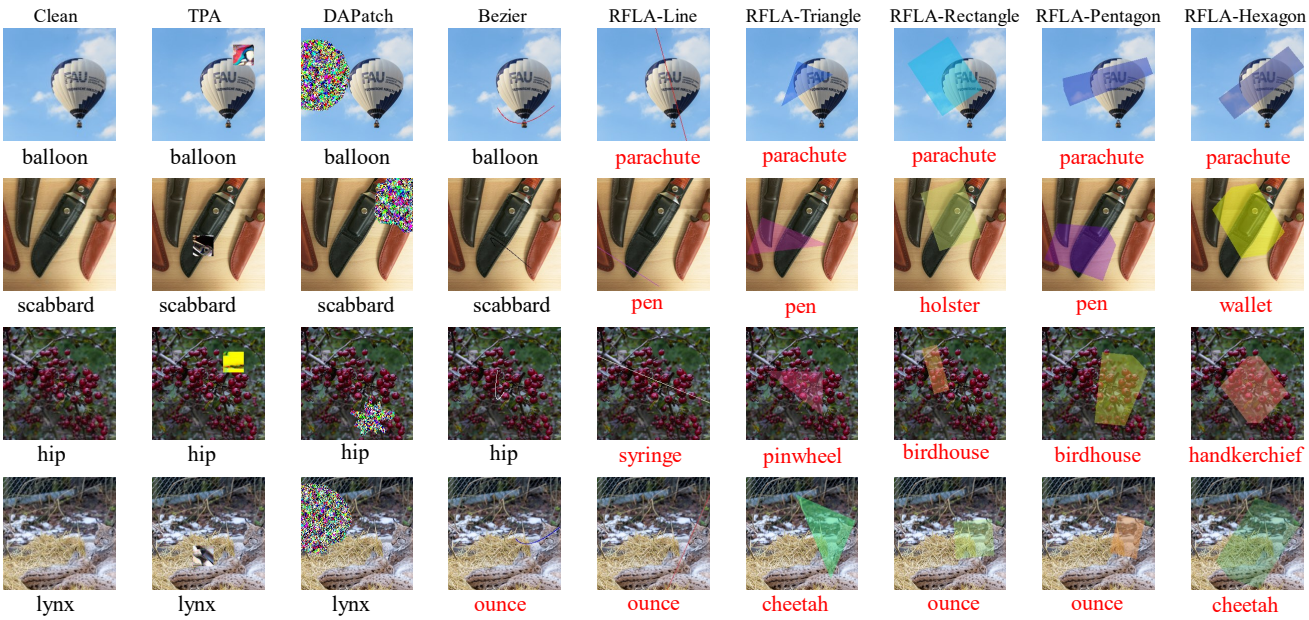


Figure 4. Visualization comparison of adversarial examples generated by different methods on ResNet50.

Table 2. Comparison results with patch-based methods in terms of ASR (%) on ImageNet-compatible dataset. The best results are highlighted with **bold**.

	RN50	VGG16	DN121	RNX50	WRN50	SN
TPA [49]	66.1	36	40	24.7	23.4	44.6
DAPatch [4]	74.3	71.6	79.3	73.7	76.7	56
RFLA-Triangle	98.1	97.8	97.2	97.2	98.1	99.5
RFLA-Rectangle	99.3	99.1	99.2	98.9	99.1	**99.8**
RFLA-Pentagon	**99.6**	99.1	99.2	99.2	**99.4**	**99.8**
RFLA-Hexagon	99.5	**99.5**	**99.5**	**99.5**	**99.4**	**99.8**

Then, the comparison results of patch-based methods are listed in Table 2. We conclude that the proposed geometry variants of RFLA outperform the existing method significantly. Specifically, the average ASR of the RFLA-triangle, RFLA-rectangle, RFLA-pentagon, and RFLA-hexagon are 97.98%, 99.23%, 99.38%, and 99.53%, obtaining the maximum improvement against TPA and DAPatch by 60.32% and 27.52%. We observe that the attack performance of comparison methods fails to achieve such results reported in their paper. The possible reason is that TPA may require two or more patches occluding 8% of the image to achieve higher attack performance. As for DAPatch [4], the position of the adversarial patch is ignored, which makes them fail to seek the model-decision-sensitive position. In contrast, our method simultaneously optimizes the position, geometry, and adversarial pattern, resulting in better performance. Moreover, the ASR gains with increasing vertex of geometry shape are attenuations, such as 20.08% gains from Line (two vertexes) to Triangle(three vertexes), while only 0.15% from Rectangle to Pentagon, which may be attributed to the improvement room of ASR is limited.

Table 3. Comparison results of transferability of adversarial examples generated by RNX50 in terms of ASR (%) on ImageNet-compatible dataset. *Item* indicates the white-box attack result, while the others are black-box results. The best results are highlighted with **bold**.

method	RN50	VGG16	DN121	RNX50	WRN50	SN
Bezier [11]	**18.1**	**16.8**	16.4	*72.7*	12.4	26.6
RFLA-LINE	17.6	15.5	**17.5**	***75.7***	**14**	**27.1**
TPA [49]	28.9	36	26.7	*24.7*	20.4	45.4
DAPatch [4]	32.3	**52.5**	42.3	*73.7*	26.4	28.8
RFLA-Triangle	37.8	34.9	33.1	*97.2*	33.7	50
RFLA-Rectangle	47.6	43.2	44.3	*98.9*	45.52	59.9
RFLA-Pentagon	47.5	45	46.4	*99.2*	45.2	61.9
RFLA-Hexagon	**50.8**	44	**47.6**	***99.5***	**47**	**63.9**

In addition, we compare the transferability of the proposed method with the comparison methods. Specifically, we use adversarial examples generated on RNX50 to attack other models. Evaluation results are reported in Table 3. As we can observe, RFLA outperforms the comparison methods in most cases, and the magnitude of fall behind cases is small. Concretely, we obtain the maximum average improvement of ASR of Bezier, TPA, and DAPatch are 0.28% (RFLA-Line), 19.18%, and 14.2%, indicating the effectiveness of the proposed method. In addition, we provide other transferability comparison results in Appendix B1.

4.2.2 Qualitatively Result

We provide the visualization result of adversarial examples generated by different methods in Figure 4. As we can observe, on the one hand, the Bezier and RFLA-Line obtain

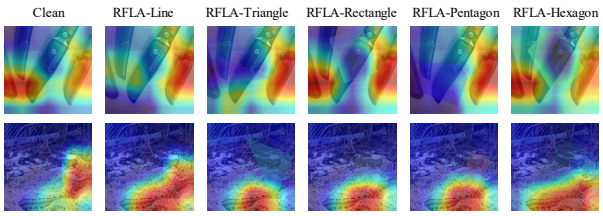


Figure 5. Model attention analysis of adversarial examples generated by RFLA on ResNet50.

the most natural visuality quality, and the scratched line is hardly observed at a glance. Meanwhile, RFLA-Line fools the DNNs in all displayed cases, but Bezier only has two success cases. On the other hand, TPA and DAPatch failed in all displayed cases. Such qualitative results can be used to explain their inferior attack performance. The adversarial patch generated by their method covers the noncontent areas, which may be insignificant to the model decision. Although the proposed method affects more image content than TPA and DAPatch, the covered contents are recognizable. In other words, our method does not modify the semantics of the image. We provide more visualization results of adversarial examples in Appendix B1.

In addition, we use the Grad-CAM [33] to investigate why the proposed method can work. Figure 5 illustrates the model attention visualization results. As we can observe, the painted geometry suppresses the model's attention areas, which makes the model output the wrong results. We also provide the visualization analysis result of comparison methods on model attention in Appendix B2.

4.3. Extend to Traffic Sign Recognition Task

To further investigate the effectiveness of the proposed method, we use RFLA to attack the traffic sign recognition (TSR) model. Specifically, we collect 200 stop sign images from the GTSRB and LISA test set for evaluation. To avoid the geometry being out of the scope of the stop sign, we use a mask to indicate the allowable modification positions. We get the mask by averaging 200 test images and binary it. Table 4 lists the digital adversarial performance. As we can observe, the proposed method obtains superior attack performances on two TSR models, especially for GTSRB CNN (100% ASR). In addition, as for LISA-CNN, the attack performance increase with the geometries. One possible reason is that the affected area of the clean image is larger with the change of geometries under a similar circle.

4.4. Adversarial Defense

In this part, we investigate the attack performance under the following adversarial defense methods: preprocess-based method and adversarial trained model. The former includes Total Variance Minimization (TVM) [14], standard Pixel Deflection (Pixel) [31], JPEG compression with quality factor 75 [26], and Gaussian blur with the kernel size 5, while the latter includes the adversarial trained model consisting of Adv-Inc-v3 (AdvIncV3) [38] and Ens-Adv-IncRes-v2 (EnvAdvV2) [42][3]. Table 5 reports the attack performance after the adversarial defense. As we can observe, on the one hand, the proposed method falls behind Bezier under defense strategy by 1.58% in terms of average ASR, which may attribute to the curve generated by Bezier exhibiting more robustness to the defense strategy than the line. On the other hand, other geometries generated by the proposed method significantly outperform the comparison method under the defense strategy. Specifically, the proposed method obtains the average ASR of RFLA-Triangle, RFLA-Rectangle, RFLA-Pentagon, and RFLA-Hexagon under defense is 63.37%, 70.77%, 63.11%, and 73.43%. In contrast, TPA and DAPatch are 49.24% and 46.77%. Such a result suggests that our method is more robust than the comparison method under different defense strategies.

Table 4. Quantitative results of RELA on TSR model in terms of ASR (%).

	LISA-CNN	GTSRB-CNN
RFLA-Triangle	68.5	100
RFLA-Rectangle	92.5	100
RFLA-Pentagon	93.5	100
RFLA-Hexagon	97.5	100

Table 5. Comparison results of attack methods under different defense strategies in terms of ASR (%).

method	TVM	Pixel	JPEG	Blur	AdvIncV3	EnvAdvV2
Bezier [11]	**74.4**	**54**	**39.8**	**51.7**	**37.5**	**24.2**
RFLA-LINE	73.7	50.2	38.3	49.3	37.1	23.3
TPA [49]	76.5	37.2	41.5	49.8	53.7	44.6
DAPatch [4]	79.3	28.2	32	42.8	59.5	47.4
RFLA-Triangle	84.4	74.5	64.4	75	57.3	45.1
RFLA-Rectangle	88	80.3	70.5	80.1	67.8	57.6
RFLA-Pentagon	84.5	45.1	55.5	64.4	**71.8**	**64.3**
RFLA-Hexagon	**88.8**	**81.6**	**73.6**	**84.6**	68.9	62.9

Table 6. Physical adversarial attacks under different light sources on different models in terms of ASR (%).

	RN50	VGG16	DN121
Sunlight	81.25	81.25	81.25
Flashlight	87.5	87.5	87.5

4.5. Physical Adversarial Attacks

Unlike the previous physical attacks that generate the adversarial pattern for the physically captured images, we generate the adversarial pattern (i.e., colored geometries) for digital images (the target model is RN50) and then reflect the light according to the optimized variables toward the corresponding printed images. In physical adversarial attacks, we use sunlight and a flashlight as the light source

[3]https://github.com/rwightman/pytorch-image-models

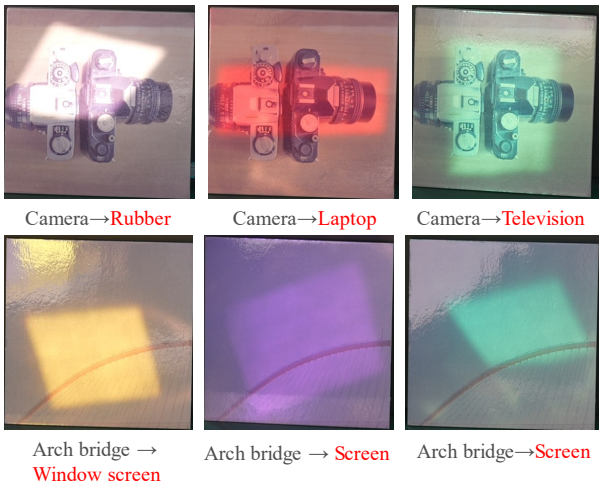


Figure 6. Physical adversarial examples.

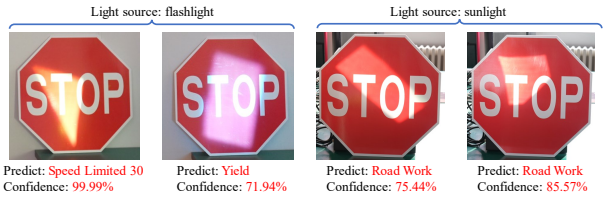


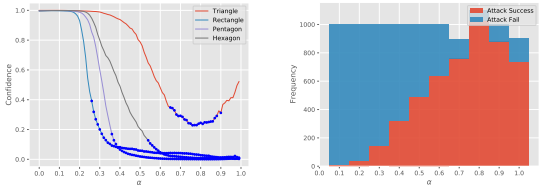
Figure 7. Examples of the "stop" sign with the reflected light and its prediction result on GTSRB CNN.

to mimic two different scenarios. We only evaluate one geometry (i.e., rectangle) for simplicity. Specifically, we randomly select six images from the dataset and generate the corresponding adversarial examples, where the color set is fixed during the optimization for physical deployment. Note that all selected images can be correctly classified as their original label after printing out. We use eight colors: seven colors created by seven transparent color plastic sheets and one white reflected sunlight. Finally, we capture the physical adversarial examples from 2 meters away, collecting 48 images for each light source.

Table 6 lists the evaluation results. As we can observe, the ASR against the three models is above 80% on physical adversarial examples created by two different light sources. Interestingly, we find that physical adversarial examples created by the reflected light against RN50 can consistently mislead the different models, indicating the reflected light is well-transferable even in the physical world. Figure 6 illustrates the physical adversarial examples. Furthermore, we study the effectiveness of reflected light attacks using sunlight and a flashlight on the TSR model. Figure 7 illustrates examples generated by different geometrical shapes.

4.6. Ablation Study

Attack performance v.s. transparency. The transparency α determined the cover intensity of the color. When α is set to one, the pixel value of the clean image at the specific position is substituted by the pure color, while the smaller value is the lower transparency of the color. We study how transparency changes the attack performance. Specifically, we fixed the other variables except for transparency, which traveled from [0,1] with a step size of 0.01. Figure 8 (a) illustrates the evaluation results of various geometries. As expectedly, the confidence of the ground-truth label decrease with enlarges of α. In contrast, the attack performance (represented in the number of blue points) rises with the increases in transparency, as more content of the clean image is covered due to deeper pure color. Moreover, we statistics the frequency of successful and failed attacks of the RFLA-Triangle on 100 test images, which is depicted in Figure 8 (b), consistent with the previous analysis.



(a) Confidence v.s. α (b) Frequency v.s. α

Figure 8. From left to right: (a) The trend of ground-truth label confidence with changing transparency, where blue points denote the successful attack. (b) The frequency of successful attacks (RFLA-Triangle) with changing of transparency.

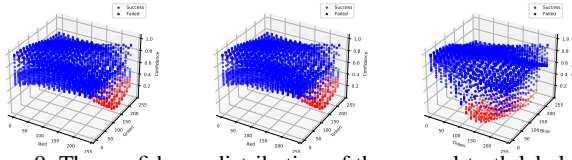
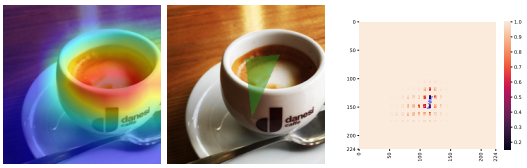


Figure 9. The confidence distribution of the ground-truth label on adversarial examples. Blue points denote the attack fail color tuple, while red points denote the attack success color tuple.

Attack performance v.s. color. The pattern of adversarial perturbation is crucial for a successful attack. Unlike the previous works that optimize the pixel-wise perturbation, we focus on the channel-wise perturbation (perturb a channel with one value) as we must ensure the perturbation is physically realizable by reflecting the light. Furthermore, channel-wise perturbation is visually more acceptable than pixel-wise perturbation. Specifically, we select the color tuple set in intervals of 16 pixels across the three RGB channels to investigate how color influences the attack performance. Figure 9 illustrates the evaluation results. As we can observe, the success cases almost cluster in specific areas near the searched optimal color tuple when other variables are fixed. In other words, the optimal color tuple has certain robustness to a slight change of color, which makes our attacks can undertake some distortions when applied in



(a) (b) (c)

Figure 10. From left to right: (a) Model attention on the clean image; (b) Adversarial examples; (c) Prediction confidence distribution of the clean image with the change of position, where blue points (the center of the circle) denote the successful attacks.

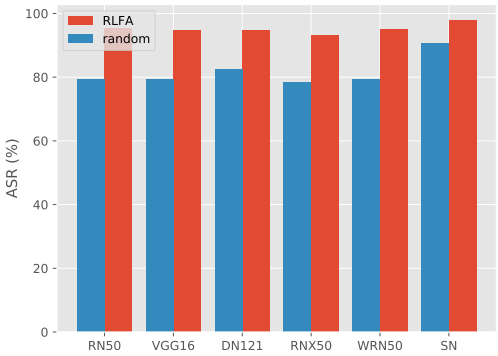


Figure 11. Comparison result with RFLA and random search in terms of ASR.

the physical world.

Attack performance v.s. position. To investigate the influence of the position of the adversarial perturbation to attack performance, we fixed the optimal variable except for position. Then, we sample the position in intervals of two steps. Furthermore, we also give the Grad-Cam for comparison. Figure 10 provides the evaluation results. As we can see, the adversarial geometry plotted around the content areas significantly drops the prediction confidence of the model on the clean image. Meanwhile, the attack success area is consistent with the model attention area, which indicates that our method can automatically locate the model attention areas to perform attacks.

RFLA v.s. Random search. To investigate the effectiveness of the proposed algorithm, we use the proposed method to find the optimal triangle reflected light with white sunlight. For comparison, we conduct the random search algorithm under the same settings. The evaluation result is illustrated in Figure 11. As we can see, RFLA outperforms the random baseline significantly. Specifically, RFLA achieves the average ASR of 95.13%, obtaining an improvement of 13.63% than the random search (i.e., 81.5%), which shows the effectiveness of our method.

5. Conclusion

In this paper, we propose a novel reflected light attack to realize effective and stealthy attacks in both digital and physical worlds, which may impose potential risks to automatic driving systems. Specifically, to explore how to control the reflected light's position, geometry, and pattern, we exploit the characteristic of the circle and propose a general framework based on the circle. To create a geometry, we first generate a specific number of angles to construct the point in circumference, followed by applying point symmetry against the center of a circle to generate a new point. These obtained points fence a geometrical shape where the fill color and transparency are optimized. Finally, we apply the PSO algorithm to find the best position, geometry, fill color, and transparency. Experiment results on digital and physical attacks verify the effectiveness of the proposed method. Moreover, our method can not only use sunlight but also can use flashlights to perform physical attacks for adapting to different environments.

Limitations. Though the reflected-light attack can perform in different environments, it is hard to remain effective in bad weather, such as fog and rain. A more penetrating light source (e.g., the traffic light and foglight) may work in such conditions.

Potential negative societal impact and mitigation. Similar to other types of attack, the adversarial attack is inevitable to cause potential security risks, especially for those physically deployed systems. However, we aim to arouse people's attention to such related applications and then encourage people to develop defense techniques to counter the reflected-light attack. To thwart the RFLA attack proposed in this paper, one can develop multimodal-based DNN systems.

6. Acknowledgments

The authors are grateful to the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (No.11725211 and 52005505).

References

[1] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1111–1119, 2019. 1, 2

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020. 2

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[4] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: deformable patch attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 529–548. Springer, 2022. 2, 5, 6, 7

[5] Bao Gia Doan, Minhui Xue, Shiqing Ma, Ehsan Abbasnejad, and Damith C. Ranasinghe. Tnt attacks! universal naturalistic adversarial patches against deep neural network systems. *IEEE Transactions on Information Forensics and Security*, 2022. 3

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 2

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 1, 2

[8] Ranjie Duan, Xiaofeng Mao, A Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun Yang. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16062–16071, 2021. 2, 3

[9] Yexin Duan, Jialin Chen, Xingyu Zhou, Junhua Zou, Zhengyun He, Jin Zhang, Wu Zhang, and Zhisong Pan. Learning coated adversarial camouflages for object detectors. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 891–897, 2022. 2, 3

[10] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 2, 3, 5

[11] Loris Giulivi, Malhar Jere, Loris Rossi, Farinaz Koushanfar, Gabriela Ciocarlie, Briland Hitaj, and Giacomo Boracchi. Adversarial scratches: Deployable attacks to cnn classifiers. *Pattern Recognition*, 133:108985, 2023. 5, 6, 7

[12] Abhiram Gnanasambandam, Alex M Sherman, and Stanley H Chan. Optical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 92–101, 2021. 3

[13] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2

[14] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7848–7857, 2021. 3

[17] Bingyao Huang and Haibin Ling. Spaa: Stealthy projector-based adversarial attacks on deep image classifiers. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 534–542. IEEE, 2022. 2, 3

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[19] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 5

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1

[21] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 1, 2

[22] Chao Li, Handing Wang, Jun Zhang, Wen Yao, and Tingsong Jiang. An approximated gradient sign method using differential evolution for black-box adversarial attack. *IEEE Transactions on Evolutionary Computation*, pages 1–1, 2022. 1, 2

[23] Chao Li, Wen Yao, Handing Wang, and Tingsong Jiang. Adaptive momentum variance for attention-guided sparse adversarial attacks. *Pattern Recognition*, 133:108979, 2023. 1

[24] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020. 1, 2

[25] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, number 01, pages 1028–1035, 2019. 3, 5

[26] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019. 7

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[28] Yanmao Man, Ming Li, and Ryan Gerdes. Poster: Perceived adversarial examples. In *IEEE Symposium on Security and Privacy*, number 2019, 2019. 2

[29] Andreas Mogelmose, Mohan Manubhai Trivedi, and Thomas B Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1484–1497, 2012. 5

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[31] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018. 7

[32] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14666–14675, 2021. 2

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[34] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 1, 3

[35] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019. 3

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[37] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 5

[38] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 7

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[40] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5307–5315, 2021. 3

[41] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[42] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. 7

[43] Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaoya Zhang, Wen Yao, and Xiaoqian Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 2, pages 2414–2422, 2022. 2, 3

[44] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8565–8574, 2021. 2, 3

[45] Xingxing Wei, Ying Guo, and Jie Yu. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3

[46] Xingxing Wei, Ying Guo, Jie Yu, and Bo Zhang. Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks. *arXiv preprint arXiv:2212.12995*, 2022. 3

[47] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 2

[48] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[49] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI*, pages 681–698. Springer, 2020. 2, 5, 6, 7

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5

[51] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15345–15354, 2022. 2, 3, 4