Adversarial Color Film: Effective Physical-World Attack to DNNs

Chengyin Hu
University of Electronic Science and Technology of China
Chengdu, China
cyhuuestc@gmail.com

Weiwen Shi
University of Electronic Science and Technology of China
Chengdu, China
Weiwen_shi@foxmail.com

Abstract

It is well known that the performance of deep neural networks (DNNs) is susceptible to subtle interference. So far, camera-based physical adversarial attacks haven’t gotten much attention, but it is the vacancy of physical attack. In this paper, we propose a simple and efficient camera-based physical attack called Adversarial Color Film (AdvCF), which manipulates the physical parameters of color film to perform attacks. Carefully designed experiments show the effectiveness of the proposed method in both digital and physical environments. In addition, experimental results show that the adversarial samples generated by AdvCF have excellent performance in attack transferability, which enables AdvCF effective black-box attacks. At the same time, we give the guidance of defense against AdvCF by means of adversarial training. Finally, we look into AdvCF’s threat to future vision-based systems and propose some promising mentality for camera-based physical attacks.

1. Introduction

Nowadays, vision-based systems and applications are gradually popularized in people’s daily life, such as autonomous driving systems, unmanned aerial vehicles and so on. At the same time, the security and reliability of these systems are also the focus of many scholars. Most scholars are keen to study adversarial attacks in the digital environment [5, 25, 34, 40], which fool advanced DNNs by adding carefully designed pixel-level adversarial perturbations to the input image, generating perturbations that are imperceptible to the human observers. In addition, some scholars are gradually working on the study of adversarial attacks in the physical environment [6, 12, 21], which uses stickers and graffiti as perturbations to fool advanced DNNs, generating perturbations visible to human observers. However, in the physical world, images are captured by the camera and then transmitted to the advanced DNNs, where an attacker cannot directly modify the input image.

Many natural factors can be major contributors to physical perturbations. For example, in hazy days with poor visibility, colored haze may be the culprit for self-driving traffic accidents. If a translucent color film is deliberately placed in front of the lens of a self-driving car’s camera to create a haze effect on the image to disturb advanced classifiers, it could have car accidents. As shown in Figure 1, the attacker places the optimized color film in front of the camera of the autonomous vehicle, so that the autonomous vehicle could not correctly identify the target objects, carry out malicious attacks and disrupt traffic order.

At present, most physical attacks use stickers as perturbations [12, 33], which successfully fool advanced DNNs without changing the semantic information of the target object. However, sticker-based attacks are hard to hide. Some scholars use light beam as perturbations [11] to execute instantaneous attacks, which successfully fools advanced DNNs as well as achieves better stealthiness. However, light-based attacks are prone to paralysis during the day-
time. Some scholars have studied camera-based adversarial attack [22], in which a tiny patch is attached in the camera lens to generate adversarial samples. However, it has high requirements on physical operation and is difficult to adapt to complex attack scenarios.

In this paper, we propose a simple and efficient camera-based physical adversarial attack called Adversarial Color Film (AdvCF). Unlike most existing physical attacks, ours conduct effective physical attacks without modifying the target objects. In terms of stealthiness, a visual comparison of our proposed method with other works is shown in Figure 2. The adversarial samples generated by AdvCF like photos taken on hazy days or when the camera is out of focus. Though the adversarial sample generated by AdvCF may appear less stealthiness than RP2 and AdvLB, AdvCF may exhibit flexible color changes to adapt to various environments. For example, unlike the existing physical attacks focus on a single environmental condition: Sticker-based is applicable to daytime (see RP2 in Figure 2) and light-based is applicable to nighttime (see AdvLB in Figure 2), ours focus on both daytime and nighttime environments (see Figure 7).

Our method is simple to implement physical attacks. By formalizing the physical parameters of color film, using genetic algorithm [16] to find the most aggressive physical parameters. Then, based on these physical parameters, printing color films and place them in front of a camera to generate physical samples. We conduct comprehensive experiments to verify the effectiveness of AdvCF. We achieve an 95.9% attack success rate on a subset of ImageNet in the digital environment. In the physical environment, the attack success rate of indoor test and outdoor test is 80.8% and 86.7%, respectively. Furthermore, we use the adversarial samples generated by AdvCF to execute the transfer attack, and verify the AdvCF’s performance in black-box setting. Our main contributions are as follows:

- We propose a camera-based physical-world attack, AdvCF, which performs efficient physical attacks by manipulating the physical parameters of color film without modifying the target object. At the same time, deploying such attack is very simple: by using an adversarial color film, it could be a common safety threat due to its ease and convenience (See Section 1).

- We introduce and analyze the existing methods (See Section 2), then, design strict experimental method and conduct comprehensive experiments to verify the effectiveness of AdvCF (See Section 3, Section 4).

- We conduct a comprehensive analysis of AdvCF, including transfer attacks of AdvCF, defense strategy of AdvCF, etc. These studies will help scholars explore camera-based physical attacks, and enlighten the thinking of defense (See Section 5). At the same time, we look into some promising mentality for camera-based physical attacks (See Section 6).

2. Related work

2.1. Digital attacks

Adversarial attack was first proposed by Szegedy et al. [37], and then, adversarial attack was successively proposed [14, 25, 26, 34]. Many scholars are committed to the study of adversarial attack in the digital environment.

Most digital attacks generate adversarial perturbations that are bounded by a small norm-ball to ensure imperceptible to human observer. Among them, $L_2$ and $L_\infty$ are the most commonly used norms [3,4,9,24,42], which guarantee the effectiveness of attacks as well as achieve stealthiness. In addition, some scholars modify other attributes of digital images to generate adversarial samples, for example, color [17, 29, 46], texture and camouflage [38, 39, 41, 45], etc. These methods generate perturbations that are slightly perceptible to the human observers. At the same time, some scholars modify the physical parameters of digital images [23, 44] and only retain the key components of images to generate adversarial samples. In general, the assumption of digital attacks is that an attacker can modify the input images, but this is not practical in a physical scenario.

2.2. Physical attacks

Physical attack was first proposed by Alexey Kurakin et al. [21]. After this work, many physical attacks were proposed successively [1,2,12,30,43].

**Traditional street sign attacks.** Ivan Evtimov et al. [12] proposed a classic physical attack called RP2, which uses stickers as perturbations to perform attacks at different distances and angles against advanced DNNs. However, RP2 is susceptible to environmental interference at large distances and angles. Eykholt et al. [33] implemented a disappear attack by improving RP2, generating robust and transferable adversarial samples to fool advanced DNNs. However, the perturbations cover a large area, which is too conspicuous. Chen et al. [6] proposed ShapeShifter, by using “Expectation over Transformation” to generate adversarial samples, the experimental results showed that the generated
stop sign always fooled the advanced DNNs at different distances and angles. Huang et al. [19] improved ShapeShifter by adding Gaussian white noise to ShapeShifter’s optimization function, achieving a more comprehensive attack. However, ShapeShifter and the improved ShapeShifter have a defect, perturbations cover almost the whole road sign, failed to achieve stealthiness. Duan et al. [10] proposed AdvCam, which uses style transfer techniques to generate adversarial samples and disguise the perturbations as a style considered reasonable by human observers. AdvCam has better stealthiness than above methods, but it needs to manually select the attack area and target. All in all, the above methods require manual modification of the target objects. In addition, these works failed to achieve the stealthiness.

Light-based attacks. Light-based attacks show some advantages over traditional street sign attacks. Nguyen et al. [27] proposed to use projector to attack face recognition system, using light projection as perturbations to generate adversarial samples, verified its antagonism to face recognition system under white-box and black-box settings. However, its deployment mode is complex. Some utilized visible and invisible light to attack face recognition systems [31, 49], project optimized light onto the target to perform covert attacks. These attacks achieve better stealthiness, but modify the target objects. Duan et al. [11] proposed AdvLB, which uses laser beam as perturbations and manipulates its physical parameters to execute attacks. Gnanasambandam et al. [13] proposed OPAD, which projects digital perturbations onto the target objects to perform efficient attacks. AdvLB and OPAD, however, can only perform attacks in weak-light conditions. Zhong et al. [47] studied shadow-based physical attack, which cast carefully crafted shadows on the target to generate adversarial samples, realizing a natural black-box attack. However, this method is difficult to work in complex physical scenes. To sum up, light-based attacks achieve better stealthiness, but each attack has limitations, including a similar disadvantage: modifying the target objects.

Camera-based attacks. Li et al. [22] studied camera-based attacks by placing well-designed stickers on the camera lens to generate adversarial samples, performing targeted attacks against advanced DNNs, it avoids modifying the target by physically manipulating the camera itself, at the same time, adversarial perturbations are inconspicuous. However, it’s difficult to adjust error due to its complex operation. Our proposed method puts a carefully designed color film in front of the camera lens, which is simple and efficient to implement in physical scenarios.

3. Approach

3.1. Adversarial sample

Given an input picture $X$, ground truth label $Y$, the DNN classifier $f$, $f(X)$ represents the classifier’s prediction label for picture $X$. The classifier $f$ associates with a confidence score $f_Y(X)$ to class $Y$. The adversarial sample $X_{adv}$ satisfies two properties: (1) $f(X_{adv}) \neq f(X) = Y$; (2) $\|X_{adv} - X\| < \epsilon$. Among them, the first property requires $X_{adv}$ fools DNN classifier $f$. The second property requires that the perturbations of $X_{adv}$ are small enough to be imperceptible to human observers.

In this paper, we use genetic algorithm [16] to optimize the physical parameters of color film, and print physical color film according to the physical parameters. Then, in the real scenarios, we put color film in front of the camera lens to take an image and generate an adversarial sample.
Figure 3 shows our approach.

### 3.2. Color film definition

In this paper, we define a color film using two physical parameters: color \( C(r, g, b) \), intensity \( I \). Each parameter is described as follows:

**Color** \( C(r, g, b) \): \( C(r, g, b) \) represents the color of the color film, where \( r, g, \) and \( b \) represent the red channel, green channel, and blue channel of the color film respectively.

**Intensity** \( I \): \( I \) indicates the transparency of color film; the greater \( I \) indicates the lower transparency, the smaller the more transparent. In physical environment, color film with high strength could be generated by superimposing color films.

The parameters \( C(r, g, b) \) and \( I \) form a color film’s physical parameter \( P h(C, I) \). We define a simple function \( S(X, P h(C, I)) \) that simply synthesizes the input image with color film to generate an adversarial sample. We define the restriction vectors \( \vartheta_{\text{min}} \) and \( \vartheta_{\text{max}} \) to limit the range of the physical parameters \( C \) and \( I \). The restriction vectors \( \vartheta_{\text{min}} \) and \( \vartheta_{\text{max}} \) are adjustable. Therefore, the adversarial sample can be expressed as:

\[ X_{\text{adv}} = S(X, P h(C, I)) \quad (1) \]

\[ \text{s.t.} \quad P h(C, I) \in (\vartheta_{\text{min}}, \vartheta_{\text{max}}) \]

### 3.3. Genetic algorithm (GA)

GA [16] is a natural heuristic algorithm designed by John Holland according to the laws of biological evolution in nature. It’s a computational model that simulates the biological evolution process of natural selection and genetic mechanism of Darwin’s biological evolution, searches the optimal evolution process of natural selection and genetic mechanism. It’s a computational model that simulates the biological evolution process of natural selection and genetic mechanism. Holland according to the laws of biological evolution in nature.

In this work, we use no model’s gradient information, require only confidence score and prediction label from the model feedback. The feasibility of using GA to optimize AdvCF include:

(1) GA searches the string set of solutions of the problem, covering a wide area, which is conducive to global optimization. In our method, physical parameters \( C \) and \( I \) include a total of \( 256 \times 256 \times 256 \times 4 \) combinations of problem solutions, GA is conducive to the global optimization of AdvCF.

(2) GA basically need no knowledge of search space or other auxiliary information, uses the fitness value to evaluate individuals, and carries out genetic operation on this basis. The fitness function is not constrained by continuous differentiability, and its definition domain can be set arbitrarily. AdvCF does not need model’s gradient information, takes the model’s confidence score \( f_Y(X) \) as the individual fitness, \( f(X_{\text{adv}}) \neq Y \) as the termination condition.

(3) Flexible selection strategy. GA uses evolutionary information to organize search. Individuals with high fitness have higher survival probability, and obtain a more adaptable gene structure. AdvCF uses the flexibility of genetic algorithm to select specific elimination strategy to further expand the search scope and achieve global optimization.

We choose binary encoding to encode the physical parameters. For the physical parameters \( C(r, g, b) \), \( r, g, \) and \( b \) range from 0 to 255, so \( r, g, b \) correspond to 8 genes respectively. We set intensity \( I \) to four different intensity values (e.g., 0.3 to 0.6) corresponding to 2 genes. Thus, parameters \( C(r, g, b) \) and \( I \) contain a total of 26 genes. After randomly encoding the initial population, utilizing binary conversion to convert genotype into phenotype. For example: \( \text{Genotype}(10010111, 00011001, 01011101, 10) \rightarrow \text{Phenotype}(C(151, 25, 93), I = 0.5) \). Then, input phenotype parameters into the model and generate adversarial samples according to Function 1. For the selection strategy, we select from small to large according to the confidence score, and eliminate the individuals with high confidence scores (for example, 1/10). Note that in this work, the smaller the confidence score, the stronger the individual fitness. Crossover and mutation strategies follow the conventional approach. During population iterations, an individual satisfies \( f(X_{\text{adv}}) \neq Y \), saving the adversarial sample and physical parameters. In the physical environment, we print the color film according to the physical parameters, place the color film in front of the camera lens to take pictures according to the method shown in Figure 3, and generate physical adversarial samples.

**Expectation Over Transformation.** EOT [1] is an effective tool for handling the conversion from digital to physical domains. We define a transformation \( \mathcal{T} \) to represent the domain transition, \( \mathcal{T} \) is a random combination of digital image processing, including brightness adaptation, position offset, color variation, and so on. Through EOT, the physical sample can be represented as:

\[ X_{\text{phy}} = \mathcal{T}(X_{\text{adv}}, P h(C, I)) \quad (2) \]

\[ \text{s.t.} \quad P h(C, I) \in (\vartheta_{\text{min}}, \vartheta_{\text{max}}) \]

### 3.4. Color film adversarial attack

AdvCF focuses on searching \( P h(C, I) \), the physical parameters of color film, which generates an adversarial sample \( X_{\text{adv}} \) that fools the classifier \( f \). In this experiment, we consider a practical situation: the attacker cannot obtain the knowledge of the model, but only the confidence score \( f_Y(X) \) with given input image \( X \) on ground truth label \( Y \). In our proposed method, we use confidence score as the adversarial loss. Thus, the objective is formalized as minimizing the confidence score on the ground truth label \( Y \), as is shown in follows:
\[
\arg \min_{P_h} \mathbb{E}_{t \sim T} [f_Y(t(X_{adv}, Ph(C, I)))] \\
\text{s.t. } f(X_{adv}) \neq Y
\]  

Figure 5. Adversarial samples generated by AdvCF.

Algorithm 1 shows the pseudocode of AdvCF. The proposed method takes clean image \( X \), target classifier \( f \), correct label \( Y \), population \( Seed \), iteration number \( Step \), crossover rate \( P_c \), mutation rate \( P_m \) as input decided by the attacker. Then, binary encoding the initial population of color film. Within the max iterations, calculating the fitness value of each individual in the population. If an individual attacks the target model successfully, output its physical parameters and terminate the algorithm. In each iteration, the selection strategy follows the fitness value \( f_Y(X_{adv}) \), taking \( P_c \) and \( P_m \) for crossover and mutation, respectively. Here, our selection strategy is to weed out the top tenth with the largest confidence score (note that the smaller the confidence score, the more antagonistic) and then fill in the randomly encoded genes separately. The benefit of this selection strategy is that it saves a lot of time cost by directly eliminating the most inferior individuals, and further expand the search scope and global optimization. In addition, we set crossover rate \( P_c \) and variation rate \( P_m \) to 0.7 and 0.1, respectively. Experiments in Figure 9 show that our strategies for selection, crossover and mutation perform efficient optimization on the target problem. The algorithm finally outputs the physical parameters of the color film \( Ph^* \), which are used to perform subsequent physical attacks.

Algorithm 1 Pseudocode of AdvCF

Input: Input \( X \), Classifier \( f \), Ground truth label \( Y \), Population size \( Seed \), Iterations \( Step \), Crossover rate \( P_c \), Mutation rate \( P_m \);
Output: A vector of parameters \( Ph^* \);
1: Initialization \( Seed \), \( Step \), \( P_c \), \( P_m \);
2: for seeds ← 0 to \( Seed \) do
3: Encoding individual genotype \( G_{seeds} \);
4: end for
5: for steps ← 0 to \( Step \) do
6: for seeds ← 0 to \( Seed \) do
7: \( Ph_{seeds}(C, I) \leftarrow G_{seeds} \);
8: \( X_{adv}(seeds) = S(X, Ph_{seeds}(C, I)) \);
9: \( f_Y(X_{adv}) \leftarrow f(X_{adv}) \);
10: if \( f(X_{adv}) \neq Y \) then
11: \( Ph^* = Ph_{seeds}(C, I) \);
12: output \( Ph^* \);
13: break;
14: end if
15: end for
16: Update: \( G_{seed} \leftarrow Selection f_Y(X_{adv}) \);
17: Update: \( G_{seed} \leftarrow Crossover P_c \);
18: Update: \( G_{seed} \leftarrow Mutation P_m \);
19: end for

4. Evaluation
4.1. Experimental setting

We test the effectiveness of AdvCF in both digital and physical environments. We perform all experiments using
ResNet50 [15] as the target model. As with the method in AdvLB [11], we randomly selected 1000 images in ImageNet [8] that could be correctly classified by ResNet50 as a dataset for the digital test. In physical test, the experimental devices are shown in Figure 4. We show part of the color films and use iPhone6s as the camera device. It has been verified that different camera devices will not affect the effectiveness of AdvCF. For all tests, we use attack success rate (ASR) as the metric to report the effectiveness of AdvCF.

4.2. Evaluation of AdvCF

**Digital test.** We test the effectiveness of AdvCF in digital environments on 1000 images that could be correctly classified by ResNet50, achieving an attack success rate of 95.90% (Untargeted ASR of 95.10% in AdvLB [11], targeted ASR of 49.60% in [22]). Figure 5 shows some interesting results. For example, when adding a blue ($C(0, 0, 255)$) film to the clean sample, the mud turtle is misclassified as jellyfish, a similar phenomenon can be seen in [11]. Studies have shown that most of the original jellyfish-label samples in the ImageNet training set are with blue backgrounds, so adding blue film will prone to lead the classifier to misclassify the adversarial samples into jellyfish, which is described in [7]. On the other hand, the cliff is misclassified as alp when covering various color films. All in all, AdvCF shows an effective adversarial effect in the digital environment, that is, it leads advanced DNNs to misclassification without changing the semantic information of the target objects.

**Physical test.** To demonstrate the rigor of AdvCF, we conduct a strict experimental design in the physical test. In the physical world, the effectiveness of physical attack is affected by environmental noise, so we design indoor test and outdoor test respectively. In which, the indoor test avoids the influence of outdoor noise, and the outdoor test reflects the performance of AdvCF in real scenarios.

For the indoor test, we use ‘Cleaver’, ‘Plastic bag’, ‘Street sign’, etc. as target object. and form 26 adversarial samples, achieving an attack success rate of 80.77% (ASR of 100% in AdvLB [11]). Figure 6 shows the experimental results of the indoor test. It can be seen that the computer-simulated color films keep better consistency with the physical printed color films. In addition, there exists printing loss of the physical printed color film due to the limitations of current printing equipment, which could be evaded by EOT [1].

In the outdoor test, we select ‘Black swan’, ‘Ashcan’ and ‘Street sign’ as attack objects, and form 105 adversarial samples, achieving an attack success rate of 86.67% (ASR of 77.43% in AdvLB [11], ARS of 73.26% in [22]). Figure 7 shows adversarial samples in the outdoor environment. Experimental result shows that by adding the optimized color film interference to the clean samples, it leads the advanced DNNs to misclassification. On the other hand, to get close to real scenarios, we conduct outdoor tests on ‘Stop sign’ from different angles, and the experimental results are shown in Table 1. It shows that AdvCF performs effective physical attacks on target objects at various angles.

In general, AdvCF avoids directly modifying the target objects. Besides, the proposed method is capable of performing physical attacks during daytime and nighttime, compared to light-based attacks and sticker-based attacks. In addition, while our approach does not perform as well as AdvLB in indoor test, it shows better adversarial effect in digital test and outdoor test and is more adaptable to various scenarios than AdvLB. In summary, the comprehensive experimental results show that AdvCF is effective in both digital and physical environments.

| Angle (°) | 0°   | 30°  | 45°  |
|-----------|------|------|------|
| ASR (%)   | 85.19| 77.78| 85.19|

Table 1. Attack success rates from different angles.
4.3. Ablation study

Here, we perform a series of experiments to study the adversarial effects of different physical parameters on AdvCF: Intensity \( I \); Color \( C(r, g, b) \).

**Intensity \( I \):** Here, we conduct experiments on 1000 images that could be correctly classified by ResNet50. The greater \( I \), the stronger adversarial effect, and the worse concealment. We study the adversarial effect of color films with intensity of 0.1 to 0.7. Table 2 shows the attack success rates for each intensity of color film. It shows that AdvCF is aggressive even at a weak intensity.

**Color \( C(r, g, b) \):** Here, in order to strictly study the adversarial effect of color films with different colors, we construct a larger dataset. First of all, we randomly selected 50 clean samples from each of the 1000 categories in ImageNet [8] and got 50,000 clean samples. Secondly, add 27 color films with \( I = 0.4 \) to each clean sample to obtain the final data set containing 1.35 million adversarial samples, which is called ImageNet-ColorFilm (ImageNet-CF). Figure 8 shows the classification accuracy of resnet50 [15] on clean samples and adversarial samples. It can be seen that each color film shows the adversarial effect. Among them, adversarial samples related to \( C(255, 0, 255) \) are the most adversarial, and the classification accuracy of resnet50 is 49.27% (↓36.56%).

| \( I \) | ASR (%) |
|-------|--------|
| 0.1   | 15.80  |
| 0.2   | 27.60  |
| 0.3   | 38.90  |
| 0.4   | 54.60  |
| 0.5   | 69.20  |
| 0.6   | 82.10  |
| 0.7   | 92.40  |

Figure 8. Ablation study of \( C(r, g, b) \).

5.2. Model attention

We use CAM [48] to show the model’s attention. As shown in Figure 10, by adding the optimized color film to clean samples, model’s attention disappears from the target objects or shifts to another one.

5.3. Transferability of AdvCF

Here, we demonstrate the attack transferability of AdvCF against advanced DNNs [15, 18, 20, 28, 32, 35, 36] in both digital and physical environments. We take the adversarial samples generated by AdvCF that successfully attacked resnet50 as the dataset. The experimental results are shown in Table 3. It can be seen that AdvCF shows effective attack transferability in the digital environment, and the attack success rate against AlexNet is 96.44%. In the physical environments, AdvCF demonstrates excellent attack transferability, whose black-box attack has paralyzed almost all of the advanced DNNs. Our experimental results imply that AdvCF allows attackers to exploit the transferability of AdvCF to carry out efficient physical attacks against advanced
DNNs without any knowledge of the model.

The experimental results in Table 1 show that AdvCF conducts effective physical attacks in a white-box setting. The data in Table 3 show that AdvCF has excellent physical adversarial performance in a black-box setting. AdvCF empowers attackers flexible operations, even without any knowledge of the model, to perform effective physical attacks. Therefore, in view of the excellent adversarial effect of AdvCF on the vision-based system in real scenes, we call for the attention of the proposed AdvCF.

Table 3. Transferability of AdvCF (ASR (%)).

| f             | Digital | Physical |
|---------------|---------|----------|
|               | 0°      | 30°      | 45°      |
| Inception v3  | 26.05   | 80.00    | 92.31    | 86.67    |
| VGG19         | 60.48   | 93.33    | 92.31    | 100      |
| ResNet101     | 35.96   | 100      | 92.31    | 100      |
| GoogleNet     | 28.59   | 100      | 100      | 100      |
| AlexNet       | 96.44   | 100      | 100      | 100      |
| DenseNet      | 35.96   | 100      | 100      | 100      |
| MobileNet     | 67.85   | 100      | 100      | 100      |

5.5. Disadvantages of AdvCF

We demonstrate the effectiveness of AdvCF in both digital and physical environments. Here, we summarize some of AdvCF’s shortcomings: (1) Although the printing loss is imperceptible to human observers, it actually exists. Due to the limitations of current printing devices, our current precautions can only be carried out in a digital environment. (2) Our experimental studies are not yet able to contribute to the interpretability of DNNs, which will be the direction of our future efforts.

6. Conclusion

In this paper, we present a simple and efficient camera-based physical-world attack, AdvCF, which performs attacks by manipulating the physical parameters of color film. We summarize and analyze the existing advanced physical attacks. Then, we design strict experimental method and conduct comprehensive experiments to verify the effectiveness of AdvCF. Finally, we analyze the efficiency of our optimization strategy, model’s bias caused by AdvCF, the black-box adversarial effect of AdvCF, the defense guidance against AdvCF, etc. Our experiments reveal the security threats of AdvCF to vision-based applications and systems in the real world. At the same time, deploying AdvCF is rather simple, so that it could be a common safety threat. Our approach provides many ideas for future physical attacks, performing physical operations on the camera itself rather than modifying the target objects. In general, our proposed AdvCF is very beneficial to study the security risks of vision-based applications in real scenarios, it is a valuable complement to the physical-world attacks.

In the future, we will continue to focus on camera-based attacks, such as doodling transparent film, clipping color film to perform attacks, etc. Furthermore, we will deploy camera-based attacks to other computer vision tasks including target detection, domain segmentation. Meanwhile, the defense strategies against camera-based attacks will also become a hot topic in the future.
References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 284–293. PMLR, 2018. 2, 4, 6

[2] Tom B. Brown, Dandelion Mané, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch. CoRR, abs/1712.09665, 2017. 2

[3] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In Bhavani Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha, editors, Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, ASIACCS 2017, Dallas, TX, USA, November 3, 2017, pages 3–14. ACM, 2017. 2

[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, pages 39–57. IEEE Computer Society, 2017. 2

[5] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 10–17. AAAI Press, 2018. 1

[6] Sheng-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng (Polo) Chau. Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Irim, editors, Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part I, volume 11051 of Lecture Notes in Computer Science, pages 52–68. Springer, 2018. 1, 2

[7] Kanjar De and Marius Pedersen. Impact of colour on robustness of deep neural networks. In IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, pages 21–30. IEEE, 2021. 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pages 248–255. IEEE Computer Society, 2009. 6, 7

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[10] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 997–1005. Computer Vision Foundation / IEEE, 2020. 3

[11] 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 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 16062–16071. Computer Vision Foundation / IEEE, 2021. 1, 3, 6

[12] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 2

[13] Abhiram Gnanasambandam, Alex M. Sherman, and Stanley H. Chan. Optical adversarial attack. In IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, pages 92–101. IEEE, 2021. 3

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society, 2016. 6, 7

[16] John H Holland. Genetic algorithms. Scientific american, 267(1):66–73, 1992. 2, 3, 4, 7

[17] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1614–1619. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 2261–2269. IEEE Computer Society, 2017. 7

[19] Shize Huang, Xiaowen Liu, Xiaolu Yang, and Zhaoxin Zhang. An improved shapeshifter method of generating adversarial examples for physical attacks on stop signs against faster r-cnns. Comput. Secur., 104:102120, 2021. 3
[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1097–1105. Curran Associates, Inc., 2012.

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[22] Juncheng Li, Frank R. Schmidt, and J. Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, June 9-15 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3896–3904. PMLR, 2019.

[23] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norms: Parametric adversaries using an analytically differentiable renderer. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[24] Aleksandar 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.

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 86–94. IEEE Computer Society, 2017.

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2574–2582. IEEE Computer Society, 2016.

[27] Dinh-Luan Nguyen, Sunpreet S. Arora, Yuhang Wu, and Hao Yang. Adversarial light projection attacks on face recognition systems: A feasibility study. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 3548–3556. Computer Vision Foundation / IEEE, 2020.

[28] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4510–4520. Computer Vision Foundation / IEEE Computer Society, 2018.

[29] Ali Shahin Shamsabadi, Ricardo Sánchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1148–1157. Computer Vision Foundation / IEEE, 2020.

[30] Mahmood Sharif, Snuti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540. ACM, 2016.

[31] Meng Shen, Zelin Liao, Liehuang Zhu, Ke Xu, and Xiaojiang Du. VLA: A practical visible light-based attack on face recognition systems in physical world. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3):103:1–103:19, 2019.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[33] Dawn Song, Kevin Eykholt, IvanEvtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In Christian Rossow and Yves Younan, editors, *12th USENIX Workshop on Offensive Technologies, WOOT 2018, Baltimore, MD, USA, August 13-14, 2018*. USENIX Association, 2018.

[34] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.*, 23(5):828–841, 2019.

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015.

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016.

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

[38] Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaoya Zhang, Wen Yao, and Xiaojian...
Chen. FCA: learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 2414–2422. AAAI Press, 2022. 2

[39] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 8565–8574. Computer Vision Foundation / IEEE, 2021. 2

[40] Rey Wiyatno and Anqi Xu. Maximal jacobian-based saliency map attack. CoRR, abs/1808.07945, 2018. 1

[41] Rey Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 4821–4830. IEEE, 2019. 2

[42] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transfer-ability of adversarial examples with input diversity. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 2730–2739. Computer Vision Foundation / IEEE, 2019. 2

[43] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V, volume 12350 of Lecture Notes in Computer Science, pages 665–681. Springer, 2020. 2

[44] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 4302–4311. Computer Vision Foundation / IEEE, 2019. 2

[45] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: learning physical vehicle camouflages to adversarially attack detectors in the wild. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. 2

[46] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 1036–1045. Computer Vision Foundation / IEEE, 2020. 2

[47] 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. 3

[48] Bolei Zhou, Aditya Khosla, Ågata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 2921–2929. IEEE Computer Society, 2016. 7

[49] Zhe Zhou, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu, and Kehuan Zhang. Invisible mask: Practical attacks on face recognition with infrared. CoRR, abs/1803.04683, 2018. 3