# Aesthetic Yet Customizable Adversarial Patches Towards Physical Attacks

Hui Wei, Hanxun Yu, Zhixiang Wang, Shin'ichi Satoh, Hao Tang, Zheng Wang

***Abstract*—Physical attacks against deep neural networks using adversarial patches have seen increasing success. However, these patches pervasively have a poor appearance and are visually conspicuous, making them difficult to apply in the real world without drawing human attention. A major challenge herein is how to generate an adversarial patch that is visually aesthetic while having a stable attack effect. In this paper, we consider patch generation as an image transformation problem, where an input image is transformed into an output image with attacking abilities. Thus, we propose an end-to-end method to generate Aesthetic yet Customizable Adversarial Patches (ACAP) that are stylized, diversified, and attacked. Specifically, in order to generate patterns that match the visual perception of human observers, ACAPs combine the semantic content of cartoon images and the style feature of artistic style images. In addition, we train an image-to-patch neural network that can transform arbitrary input images into adversarial patches that successfully perform physical attacks. Moreover, we delve into the influential relationship between the style and attack effectiveness, and conclude that styles with more saturated and richer colors have more attack capabilities. Experimental evaluation shows that, in both digital and physical-world spaces, our method improves visual quality while remaining effective in fooling the person detector. The successful realization of ACAP illustrates the feasibility of protecting personal privacy from detection by malicious surveillance cameras in the physical world.**

***Index Terms*—Adversarial patches, Physical attacks, Person detector.**

## I. Introduction

DEEP neural networks (DNNs) are rapidly developing and have achieved great success in many tasks, such as image classification [1], image segmentation [2], depth estimation [3], object detection [4], natural language processing [5], and recent studies also show that DNNs are highly vulnerable to adversarial attacks [6]–[9]. Adversarial attacks exist not only in the digital space [10], [11], but have been successfully applied to the physical space [12]–[16]. In recent years, researchers have conducted many meaningful studies on adversarial attacks, which have significant implications for both explainable AI [17]–[20] and AI security [21].

H. Wei, H. Yu and Z. Wang are with the School of Computer Science, National Engineering Research Center for Multimedia Software, Wuhan University, China. (e-mail: weihui0713@whu.edu.cn; hanxun_yu@whu.edu.cn; wangzwhu@whu.edu.cn)

Z. Wang and S. Satoh are with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Japan, and also with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan. (e-mail: wangzx@nii.ac.jp; satoh@nii.ac.jp)

H. Tang is with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland. (e-mail: hao.tang@vision.ee.ethz.ch)

Manuscript received XXX; revised XXX.

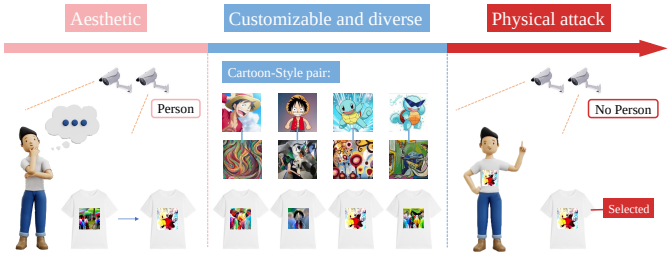


Fig. 1. Objective: A person is invisible under a camera with the person detector. We achieve the objective by wearing the T-shirts with adversarial patches. Our ACAP is combine the semantic content, the visual effect of stylization and the attack ability, which are aesthetic, diversified and attacking-robust. Our method enables users to interface by inputting their favorite content and style pair, and generating the customizable patches.

In general, effective digital attacks can be achieved using pixel-level and human-eye-invisible perturbations. But they are not feasible for real-world physical attacks, because physical attacks require capturing the perturbations by sensors, like cameras. Related methods perform physical adversarial attacks by using adversarial patches instead of adding perturbations. However, the appearance and color distribution of existing adversarial patches are rather abrupt, giving a strong visual impact [22], [23], or meaningless scribbles that are difficult to understand [25], [27]. In practical applications, it will be discovered by the victim or defender before a successful attack is made. To apply adversarial patches in the physical world, the main challenge is to produce visually aesthetic adversarial patches with physical attack effects.

One approach for improving the visual quality of patches is to constrain their structural and textural features, which has been used for example by PSAP [14] using the seed patch, and TextureAP [24] using the total variation. However, the low-dimensional features used by these methods do not contain semantics, which results in the generated patches remaining incomprehensible to humans. In parallel, the current methods are optimization-based, and patches are generated by minimizing a loss function, which is inefficient, since inference requires solving a complete optimization problem.

In this paper, we study adversarial attacks on the person detection model. To address the mentioned issue, we propose an end-to-end method to generate the **A**esthetic yet **C**ustomizable **A**dversarial **P**atches (ACAP), which are stylized, diversified and attacking-robust. As shown in Fig. 1, our approach aestheticizes adversarial patches, provides solutions for customization, and achieves attacks on the physical world. Concretely, we consider patch generation as

TABLE I
CONSIDERED FACTORS OF DIFFERENT ADVERSARIAL PATCHES (AP).

| | GoogleAP [22] | DPATCH [23] | PSAP [14] | TextureAP [24] | AdvT-shirt [25] | LAP [26] | ACAP (Ours) |
|---|---|---|---|---|---|---|---|
| Style Feature ? | | | | | | | ✓ |
| Texture Feature ? | | | | ✓ | | ✓ | ✓ |
| Structure Feature ? | | | ✓ | | | ✓ | ✓ |
| Digital Attack ? | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Physical Attack ? | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |

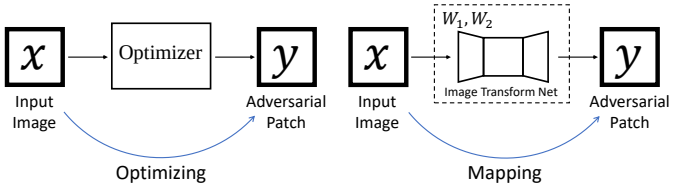


Fig. 2. A comparison between prior patch generation methods [22], [24], [26] (left) and ours (right).

an image transformation problem and propose an image-to-patch transformation network that has the capability of transforming arbitrary images into adversarial patches that perform physical attacks successfully. Fig. 2 indicates the difference between the two types of approaches, and in the testing phase, our transformation network produces patches in real time. Regarding visual aesthetics, ACAPs introduce semantic content from cartoon images, which is perceptually consistent with human perception. Although recent work [26] makes similar attempts to add structure and texture features to patches, the style, as a vital feature of images, is crucial, but not yet exploited. We further combine the Neural Style Transfer technique to camouflage adversarial perturbation as style features, meanwhile introduce aesthetic elements to our created patches. With arbitrary pairings of stylistic features and semantic contents, we generate diversified ACAPs. In addition, we study the attack effect of different styles and conclude that the more saturated and richer the color, the more attack abilities the style has. This finding provides a clue to explore the attack mechanism of adversarial patches.

We demonstrate the performance of our method through a series of experiments. Our results on the YOLOv2 detection model show that our ACAPs achieve a considerably obvious attack effect, while being better in terms of visual quality compared to other methods. We summarize the contributions of our work as follows:

- We propose an end-to-end network to transform arbitrary input images into adversarial patches for successful physical attacks. The network balances aesthetics and attackness. To the best of our knowledge, we are the first to introduce the image transformation network to control the generation of adversarial patches.
- To improve the visual aesthetics of adversarial patches, we distill content features from cartoon images and infuse artistic style into our generated ACAPs. Unlike previous work that considered only the structural and textural features of the patches, we take advantage of stylistic features.
- By analyzing the differences in the attack effect of patches with different styles, we concluded a pattern: styles with higher color saturation and richness have stronger attack capabilities, and vice versa. This finding provides a clue for us to explore the attack mechanism of adversarial patches.
- We conduct experiments to show our proposed ACAP not only achieves effectiveness for attacking the person detection model in both digital and physical spaces, but also is more aesthetic to human perception compared to other adversarial patches.

## II. RELATED WORK

### A. Adversarial Patches for DNN Models

Recently, adversarial patches, which confine the perturbation to a small and localized region without perturbation constraint, have been frequently applied to physical attacks. Brown *et al.* [22] firstly designed an adversarial patch fooling classifiers to output any targeted class. After that, Liu *et al.* [23] attacked mainstream detectors in the digital space by DPATCH, a patch with noisy adversarial perturbations. However, they focused on attacking ability and disregarded the aesthetics of patches, which causes the patch to be abrupt and compelling. To obtain the visually aesthetic adversarial patch, Thys *et al.* [24] made sure the optimizer favors a patch with smooth color transitions and prevents noisy images by constraining the total variation in the patch. They got the smoother texture patch, but the structure is still messy. Liu *et al.* [14] proposed PS-GAN that feeds any type of seed patch to generate a similar adversarial patch with seed. Tan *et al.* [26] proposed the patch rationality, indicated from three aspects: color features, edge features, and texture features, to encourage patches to obtain visual rationality. To sum up, these methods provide improvements to the structure and texture features of adversarial patches while ensuring the attacking ability. In TABLE I, we list a comparison of the considered factors of different methods. Although recent studies have made great progress on the visual quality of adversarial patches, style as a vital feature of images, was not taken into account when generating patches. Our approach fills the gap and generates the aesthetic adversarial patches that are stylized, customizable and attacking-robust.

### B. Neural Style Transfer

Neural style transfer aims at transferring the style from one image onto another, which can be framed as image transformation tasks. Gatys *et al.* [28] pioneered the parametric

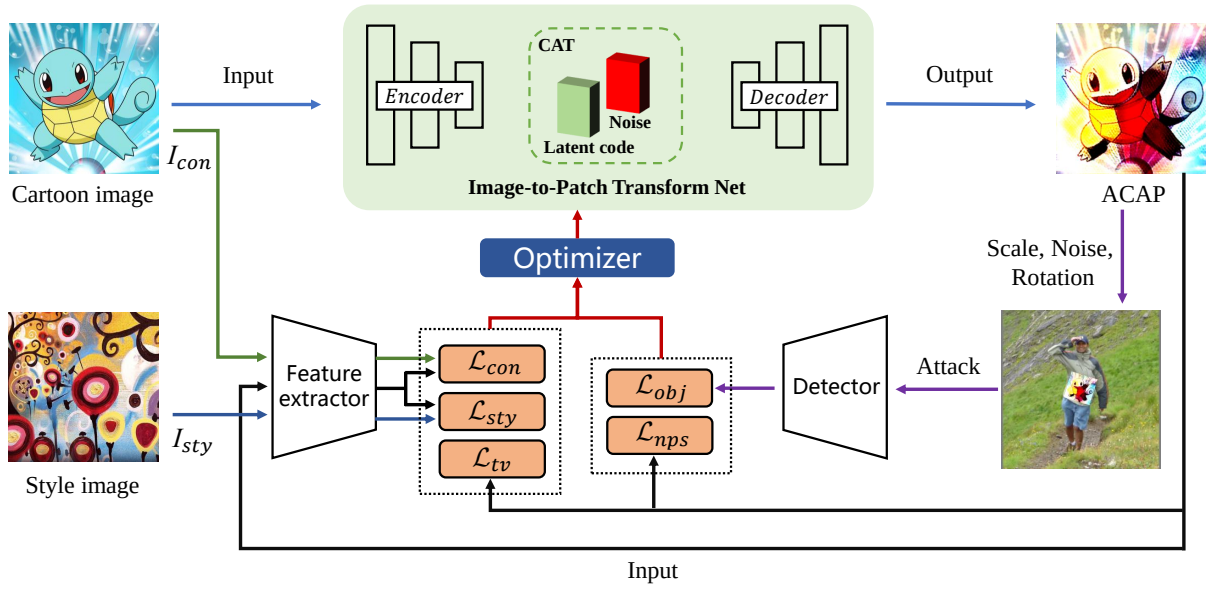


Fig. 3. Overview of our system pipeline. Given a content target image $I_{con}$ and a style target image $I_{sty}$, our method is modeled as an image transformation network to generate aesthetic adversarial patches (ACAPs). For visual effect, the created patch combines the semantic content of $I_{con}$ and the style feature of $I_{sty}$. For attacking, simulation of the physical environment, non-printability score (NPS) loss $\mathcal{L}_{nps}$ and object loss $\mathcal{L}_{obj}$ together enable the effectiveness of attack in the physical world.

neural style transfer method employing the power of CNNs and Gram matrices. They separated the content and style information of the image from the feature representations. The style information can then be recombined into the target image to achieve style transfer. Afterward, the follow-up researches have been presented to improve their performances on different aspects [29]–[32]. In this paper, we exploit these techniques for introducing artistic style to our adversarial patches.

## III. Method

### A. Problem Definition

The goal of our paper is to generate adversarial patches against the person detection model. Given a DNN-based detector $\mathbb{F}_\theta$ and a clean input image $I$ with the ground truth label $y$, an aesthetic adversarial patch $p$ can cause the detector to make incorrect prediction as follows:

$$\mathbb{F}_\theta(\mathcal{A}(I,p)) \neq y, \tag{1}$$

$$p = \mathbb{TF}(I_{con}, I_{sty}), \tag{2}$$

where $\mathcal{A}$ is a patch apply function that pastes the adversarial patch $p$ to the image $I$, and $\mathbb{TF}$ represents an image transformation network that transforms input image $I_{con}$ into adversarial patch $p$ and captures the style feature from image $I_{sty}$. Meanwhile, it ensures that patch $p$ can perform physical attacks successfully.

### B. Overview

As shown in Fig. 3, given a content target image $I_{con}$, our method trains an image transformation network, which employs the Encoder-Decoder architecture and transforms the input cartoon image $I_{con}$ into adversarial patch ACAP. For visual effect, ACAP combines the semantic content of $I_{con}$ and the style feature of $I_{sty}$; for attack effect, the created patch is supplied to the external environment simulator that consists of adding noise, rotation, scaling and printable colors restriction $L_{nps}$. Then, we paste the ACAP onto the person dataset and input them to the person detector for attacking. The total objective function $\mathcal{L}_{total}$ is defined as

$$\mathcal{L}_{total} = \underbrace{\lambda_1\mathcal{L}_{obj} + \lambda_2\mathcal{L}_{nps}}_{\text{Physical Attack}} + \underbrace{\mathcal{L}_{con} + \lambda_3\mathcal{L}_{sty} + \mathcal{L}_{tv}}_{\text{Aesthetic Patch}}, \tag{3}$$

where $\lambda_1$ to $\lambda_3$ are used to balance the multiple objectives, object loss $\mathcal{L}_{obj}$ and non-printability score loss $\mathcal{L}_{nps}$ ensure adversarial attacks apply in the physical world. Content loss $\mathcal{L}_{con}$, style loss $\mathcal{L}_{sty}$ and total variation loss $\mathcal{L}_{tv}$ are minimized by the optimizer, to control the semantic content, style feature and smooth texture, respectively. The three loss functions encourage our transformation network to generate aesthetic patches.

### C. Image-to-Patch Transformation Network

Prior studies [22], [24], [26] optimize an initial patch and create only one patch in one training. Unlike previous work, we consider patch generation as an image transformation problem, where an input image is transformed into an output image. Fig. 2 shows the difference between the two types of approaches. Our image-to-patch transformation network is a deep convolutional neural network parameterized by weights $W_1$; it transforms input images into output adversarial patches via a mapping, with the core objectives of i) generating adversarial patches that perform successful physical attacks, ii) transferring the targeted artistic style features to created

patches, and iii) maintaining the semantic content of input cartoon images. Following the architectural guidelines [28], [33] for image transformation and perceptual information capturing, we employ the U-Net architecture in our image-to-patch transformation network, which allows low-level information to shortcut across the network, leading to better results.

Since the direct input to the network is cartoon images, it contains only content features, making it difficult for the network to learn other features, e.g., adversarial features that control the attack effect. To address this shortcoming, we introduce noise features to encourage the transformation network to be able to express more diverse features. As seen in the top branch of Fig. 3, the encoder and decoder form a mapping between cartoon image $I_{con}$ and latent code. A noise tensor with spatial dimensions is concatenated with the latent code. Note that the noise feature tensor is parameterized by weights $W_2$, which is updated by the optimizer in the training process. The image transformation network is trained using stochastic gradient descent to minimize a weighted combination of loss functions:

$$W_1{}^*, W_2{}^* = \arg \min_{W_1,W_2} \mathrm{E}_{I_{con}} \left[ \sum_{i=1} \lambda_i \mathcal{L}_i(D(E(I_{con}), Noise)) \right]. \quad (4)$$

### D. Introduce Semantic Content and Artistic Style

To tackle the poor appearance of adversarial patches, researchers delve into the total variation of images [34], the patch rationality [26] and the exploitation of seed patch [14], [35]. To sum up, they made improvements to the structure and texture features. Intuitively, generated patches with visual semantics are more perceptually acceptable to humans than meaningless distortions. Inspired by this, we introduce content features to our ACAP from cartoon images by content loss $L_{con}$. In addition, transforming from a cartoon image to an adversarial patch requires adding perturbations to the original image, which will degrade the aesthetics of the image. Our ACAP introduces style features from artistic images by style loss $L_{sty}$. On the one hand, the introduced style feature increases the aesthetic elements of patches, and on the other hand, the artistic style can be used as the camouflage of adversarial perturbations. Following previous literature [28], [33], we define the content loss $L_{con}$ and the style loss $L_{sty}$ as follows:

$$\mathcal{L}_{con}(I_{con}, p) = \frac{1}{C_c H_c W_c} \|f_c(I_{con}) - f_c(p)\|^2, \quad (5)$$

and

$$\mathcal{L}_{sty}(I_{sty}, p) = \frac{1}{C_s H_s W_s} \|G[f_s(I_{sty})] - G[f_s(p)]\|^2, \quad (6)$$

where $p$ is our patch, $f_c$ (or $f_s$) is the feature map of shape $C_c \times H_c \times W_c$ (or $C_s \times H_s \times W_s$) that extracted from the $c$-th (or $s$-th) layer of the pre-trained VGG-16 network [36], and $G$ indicates the Gram matrix of deep features extracted from a set of style layers, $s \in \{\text{relu1\_2, relu2\_2, relu3\_3, relu4\_3}\}$, and $c \in \{\text{relu3\_3}\}$.

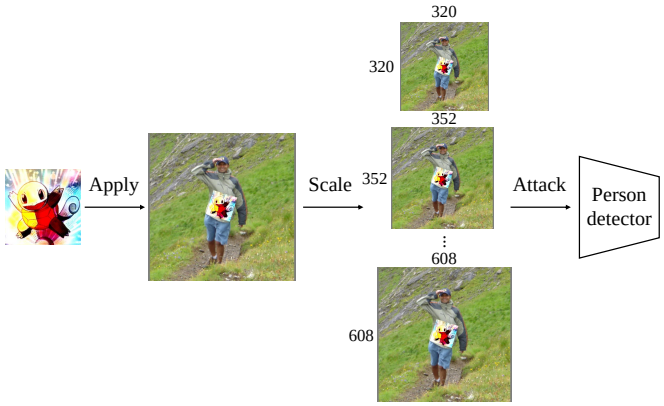


Fig. 4. Overview of multi-scale training. The adversarial patch is scaled with the target image. We randomly choose a scaled size from the interval [320, 608], in increments of 32.

Note that, we observe that our ACAPs generated using different styles have varying attack capabilities, although their contents are the same, based on the experimental experience shown in Sec. IV-B. The style with higher color saturation and richness has stronger attack capabilities. We can exploit this pattern to build a style library that collects styles that are excellent at attacks. And users can use this style library to generate customized adversarial patches that are superior in terms of attacks.

**Total Variation Loss.** To encourage spatial smoothness in the output patch $p$ and prevent noisy textures, we follow previous work [34] and use the total variation loss $\mathcal{L}_{tv}$:

$$\mathcal{L}_{tv} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2}, \quad (7)$$

where $i$ and $j$ indicate the index of pixels in our patch $p$. The total variation loss makes the style transfer more natural.

### E. Attack in the Physical Space

In this paper, we attack YOLOv2 [37], a one-stage strategy object detector, under white-box settings in the physical space. The detector outputs a bounding box containing the coordinate position, an object score (how likely it is that this detection contains an object), a class score (which class is in the bounding box), and is a joint training method for detection and classification. In addition, YOLOv2 introduces anchor boxes, and each anchor box contains a vector $[x, y, w, h, p_{obj}, p_{class1}, p_{class2}, ...p_{classn}]$. We construct a triplet $[B, P_{obj}, P_{class}]$ to represent the output of the detector, where $B$ is the position, $P_{obj}$ is the probability that this anchor point contains an object, and $P_{class}$ is the class score of the object. We define the object loss function $\mathcal{L}_{obj}$ to attack the detector as follows:

$$\mathcal{L}_{obj} = \min_{P_{obj} \in [B, P_{obj}, P_{class}]} P_{obj}, \quad s.t. \quad P_{class} = 0. \quad (8)$$

Unlike digital attacks, which only require that models capture the adversarial perturbations, physical attacks occur in real-world scenarios and therefore require that sensors, like cameras, capture the perturbations. Thus, creating adversarial


Fig. 5. Example results for adversarial patches generated with different styles and contents (bottom) and their attacking effects in the digital space (top and middle). *Clean* shows that the detector normally performs when with no patches. *Noise* and *Cartoon* show that applying random noise patches or original cartoon patches cannot attack the detector. Contrastingly, the ACAPs fool the detector to fail to identify persons (*Patch*).

patches for physical attacks is much more challenging. To keep the generated adversarial patch more robust in the physical world, we use the following three strategies.

**Multi-Scale Training.** To enhance the attack-robustness of our patches, we employ the multi-scale training strategy. In previous studies [24], [26], the size of images that are input to the detector is fixed. However, in physical-world applications, the object detection model may receive images of different sizes to detect, and images need to be scaled to the suitable size that matches the detector. Since the patch is scaled with target images, it makes the patch less attack effective. Considering this, we randomly scale images with the patch in every certain iteration in the training process as shown in Fig. 4. The scale operation encourages the generated adversarial patch to gain scale-invariant attack features, which are more conducive to physical attacks.

**NPS Loss.** Since the device such as printers and screens can only reproduce a limited portion of the RGB color space, we include minimizing the non-printability score (NPS) [34], a factor that represents how well the colors in our patch can be represented by a common printer, as part of our optimization. We construct the loss function $\mathcal{L}_{nps}$ to encourage the generated patches to be printable.

$$\mathcal{L}_{nps} = \sum_{c_{patch} \in p} \min_{c_{print} \in C} \|c_{patch} - c_{print}\|^2, \tag{9}$$

where $c_{patch}$ is a pixel in our patch $p$ and $c_{print}$ is a pixel in the set of printable colors $C$. The NPS loss $\mathcal{L}_{nps}$ measures the distance between the designed patch vector and a library of printable colors acquired from the physical world.

**External Environment Simulation.** When the adversarial patch is applied to the physical space, many influencing factors can impact its appearances, such as rotation, light, and noise. To take this into account as much as possible, we use a series of random transformations to simulate the external environment. When we get created patches from the image transformation network, we conduct rotating, adding noise, and scaling up or down to patches before pasting them to the corresponding location in the person dataset. Note that all transformation operations are possible to calculate the backward gradient towards the optimization parameters of the image transformation network.

## IV. Experiments

In this section, we evaluate our proposed method in the person detection attack task. Firstly, we outline the experimental setup. Then, we illustrate the effectiveness of our proposed attacking method by thorough evaluations in both the digital and physical space. Furthermore, we analyze our generated patches via ablation studies.

### A. Experimental Setup

**Dataset and Model.** The INRIAPerson dataset [38] is a set of images that contain people who are standing or walking. The dataset has a total of 902 positive sample images, of which 614 are the training set and 288 are the test set, and a total of 1826 people are included in these images. Most of the bodies in the images are in a standing position and have a height of more than 100 pixels, which are better suited for attacking the person detection model. The cartoon image dataset contains 105 images, which we download from the Internet. We use these images to train and test our proposed method. We evaluate the performance of our method to attack YOLOv2 [37] trained on the MS COCO dataset [39], which includes 80 labeled object classes.

**Implementation Details.** We conduct all experiments on a computer with an NVIDIA GeForce RTX 3090 GPU, and all of our codes are implemented in PyTorch. The network

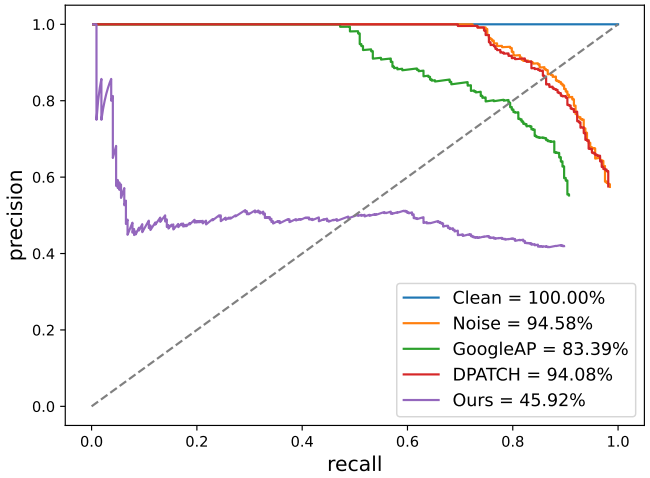


Fig. 6. PR-curve of our adversarial patch compared to the random noise patch, GoogleAP and DPATCH. The average precision (AP) is shown in the legend. The lower AP, the better the attacking effect.

TABLE II
EVALUATION OF TRANSFERABILITY ON FIVE DETECTOR.

| Patch | Detector | | | | | AVG |
|---|---|---|---|---|---|---|
| | YOLOv2 | YOLOv3-tiny | YOLOv3 | YOLOv5 | Faster RCNN | |
| Noise | 94.58 | 85.83 | 93.15 | 93.09 | 95.87 | 92.51 |
| Cartoon | 89.91 | 81.41 | 87.95 | 87.67 | 95.06 | 88.40 |
| ACAP | 45.92 | **42.84** | 79.75 | 83.47 | 88.34 | 68.07 |

adopted to extract features of style and content is VGG-16 [36] pre-trained on MSCOCO [39]. The optimizer is Adam, and the learning rate is initialized as $10^{-3}$ decayed by a factor of 0.1 every 50 epochs. For all experiments, by default, we set $\lambda_1 = 10$, $\lambda_2 = 10^{-2}$, $\lambda_3 = 10^5$ in Eq. (3), and each aesthetic adversarial patch is generated at $10^3$ iterations. In the training and testing process, we use the initial patch and the input cartoon image of size $3 \times 256 \times 256$. The shape of the noise feature is $512 \times 16 \times 16$. Moreover, the attacks mainly focus on the white-box settings, and the patch is rotated up to 20 degrees each iteration.

### B. Digital World Attack

The goal of the digital world attack is to generate adversarial patches that are applied to images containing persons so that the detector cannot detect them. We train one image transformation network per style target and demonstrate generated patches with the performance of their attacks. In Fig. 5, we show qualitative examples of our ACAPs generated by a variety of style and content images. In these results, it is clear that ACAPs are aware of the semantic content and style features of images. For example, in the bottom of Fig. 5, the cartoon characters are clearly recognizable in the patches, and the target style is transferred into corresponding patches. Meanwhile, our method generates special patterns that can attack the person detector. As shown in the top and middle of Fig. 5, we first detect the clean images without any patch using the YOLOv2 and the detector can recognize the person

TABLE III
COMPARISON RESULTS OF ATTACK ABILITY IN SIX STYLES. WE REPORT THE AVERAGE PRECISION (AP) FOR THE GENERATED PATCHES WITH FIVE CONTENT TARGETS AND THE AVG FOR EACH STYLE.

| Style | Content images | | | | | AVG |
|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C5 | |
| Candy | 54.51 | 56.45 | 62.71 | 56.82 | 57.40 | 57.59 |
| Composition | 49.14 | 45.20 | 47.05 | 44.39 | 43.94 | 45.94 |
| Feathers | 72.42 | 75.38 | 71.76 | 74.44 | 61.29 | 71.06 |
| Mosaic | 81.60 | 79.36 | 72.64 | 77.10 | 78.48 | 77.84 |
| SNight | 60.93 | 65.54 | 58.80 | 57.01 | 59.08 | 60.27 |
| Color | 48.95 | 47.94 | 49.96 | 48.36 | 50.16 | 49.07 |


and other classes normally ($Clean$). The results marked $Noise$ refer to the image with random noise patches. We take original cartoon images as patches and apply them to target images ($Cartoon$), which is compared with the results marked $Patch$ that attack the detector using our ACAPs. In the first three cases, YOLOv2 detects normally but fails to detect the person in the last case. By comparison, we can conclude that our ACAPs achieve significantly better performance for digital attacks.

**Comparison with SOTA Methods.** We compare our ACAPs with other state-of-the-art adversarial patch generation methods (containing GoogleAP [22], DPATCH [23], TextureAP [24] and LAP [26], all of which are shown at the bottom of TABLE IV). Fig. 6 shows the attack effect of different patches. Here we select the second patch in Fig. 5 for comparison. The performance of TextureAP and LAP is not shown in the PR-curve because we directly use the data from the relevant literature [24], [26] (25.53%, 43.07% respectively). The experimental results indicate that GoogleAP and DPATCH demonstrate poor performance, which only decreases the average precision (AP) to 83.39% and 94.08%. In contrast, the ACAP achieves a similar attack ability to LAP, making the AP decrease to 45.92%. In summary, from the perspective of attacking ability, our proposed method can generate adversarial patches that effectively attack the person detection model.

**Transferability.** Mainstream adversarial attack methods often tackle the trilemma including: attackability, stealthiness and transferability. Here, we evaluate the transferability of ACAP. We take YOLOv2 as our threat model and train to generate the ACAP (white-box attack). Then the ACAP attacks the one-stage detectors YOLOv2, YOLOv3-tiny, YOLOv3 [40], YOLOv5 and two-stage detector Faster RCNN [4] (black-box attack). The comparison results as shown in TABLE II, we observe that ACAP remains remarkably attackable on YOLOv3-tiny with AP dropping to 42.84%, lower than on YOLOv2. For YOLOv3, YOLOv5 and Faster RCNN, the drop in AP is not significant, but compared to $Noise$ patch and $Cartoon$ patch, It is clear that ACAP still has comparable attacking effect. The transferability of adversarial patches

TABLE IV
THE SCORE OF IQA, AESTHETICS, NATURALNESS OF DIFFERENT PATCHES. COLUMN 'OURS' IS THE AVERAGE OF OUR FOUR ACAPS.

| | GoogleAP | DPATCH | TextureAP | LAP | Ours |
|---|---|---|---|---|---|
| IQA [42] | 27.94 | 43.36 | 18.26 | 42.87 | **46.97** |
| Aesthetics | 1.88 | 2.12 | 3.42 | 4.51 | **6.68** |
| Naturalness | 4.62% | 5.64% | 21.03% | 25.13% | **43.59%** |

GoogleAP DPATCH TextureAP LAP Ours

across different detectors is a challenging research issue, which is the focus of future research in our method.

**Analysis of Different Styles.** In our experiments, we have observed an interesting phenomenon: the style has a significant impact on the attack ability of our generated adversarial patches. In the following, we conduct a series of experiments to investigate the relationship between style and attack effectiveness. First, we train six models with six different style target images [Candy, Composition, Feathers, Mosaic, SNight, and Color] and generate five patches with each model. The content target images are C1-C5. Then, these patches are used to attack the person detector. The comparison results as shown in TABLE III, we observe that i) the same style of patches, with little difference in attack ability; ii) the style Composition makes the average precision (AP) drop to 45.94%, which has the strongest attack ability compared to others; iii) the style Feathers and Mosaic are obviously weaker than the other contrasting styles in terms of attack ability.

We further calculate the tuple(Saturation, Colorfulness) of each style image and analyze the underlying reason for these observations. Colorfulness is a metric [41] to measure the overall color richness of a picture from 0 to 100. In addition, we conduct a principal component analysis on the colors of the style images using the K-means clustering algorithm and display the top-10 principal colors with their percentages in the palette below TABLE III. The comparison results indicate that styles with high color saturation and richness have stronger attack abilities, e.g., Composition (97, 68) and Color (110, 62). Contrarily, single color and less saturated styles generate patches with fewer attack abilities, e.g., Feathers (64, 54), Mosaic (66, 47) and SNight (121, 58). This finding provides a clue to explore the attack mechanism of adversarial patches. For example, the color of the patches may be more critical than the structure in the attack task.

### C. Patch Quality Assessment

In this part, we conduct patches quality assessments in aesthetics and naturalness. Related work [22]–[24], [26] (as shown at the bottom of TABLE IV) is compared with our ACAPs. We use three evaluation metrics: IQA, Aesthetics, and Naturalness. IQA is the metric of an image quality assessment model [42] based on a hyper network, which outputs a score between 0 and 100 to measure the visual quality of an image. Furthermore, considering that the aesthetic quality of an image

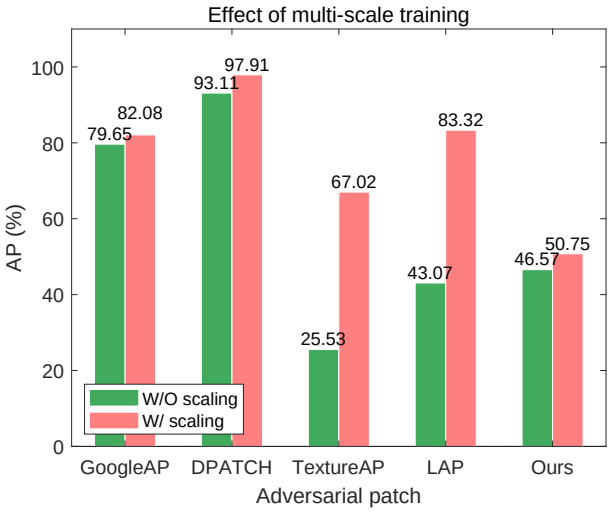


Fig. 7. Influences of the multi-scale training for different patches. Comparison results indicate that our patch maintains a stable attack capability with scaling.

is highly related to the human senses, we conduct a human perception study on patches as follows: (1) Aesthetics: Participants are asked to score each of the patches for aesthetics from 0 to 10; (2) Naturalness: Assuming one patch as the design on your T-shirt, participants are asked which patch you prefer. In particular, we collect all responses from 52 participants.

As shown in TABLE IV, the IQA of our ACAP is significantly higher than GoogleAP and TextureAP and similar to DPATCH and LAP. As for the aesthetics, the ACAP achieve a score of 6.68, higher than all other patches. Moreover, up to 43.59% of the participants believe that our adversarial patch is a more appropriate choice for wearing a T-shirt with the patch, which outperforms others by large margins (18%+). Therefore, the experimental results demonstrate that our proposed method can generate aesthetic adversarial patches that are perceptually consistent with human perception.

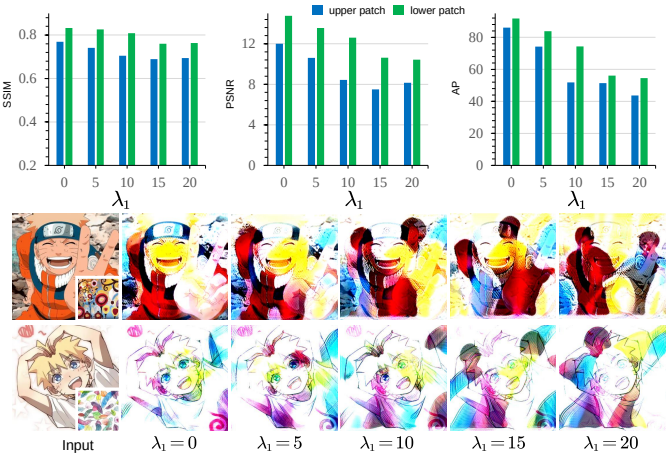


Fig. 8. Comparison results for visual quality and attack capability under varying degrees of weight $\lambda_1 \in \{0, 5, 10, 15, 20\}$. We report SSIM, PSNP, and AP (average precision) for each adversarial patch.

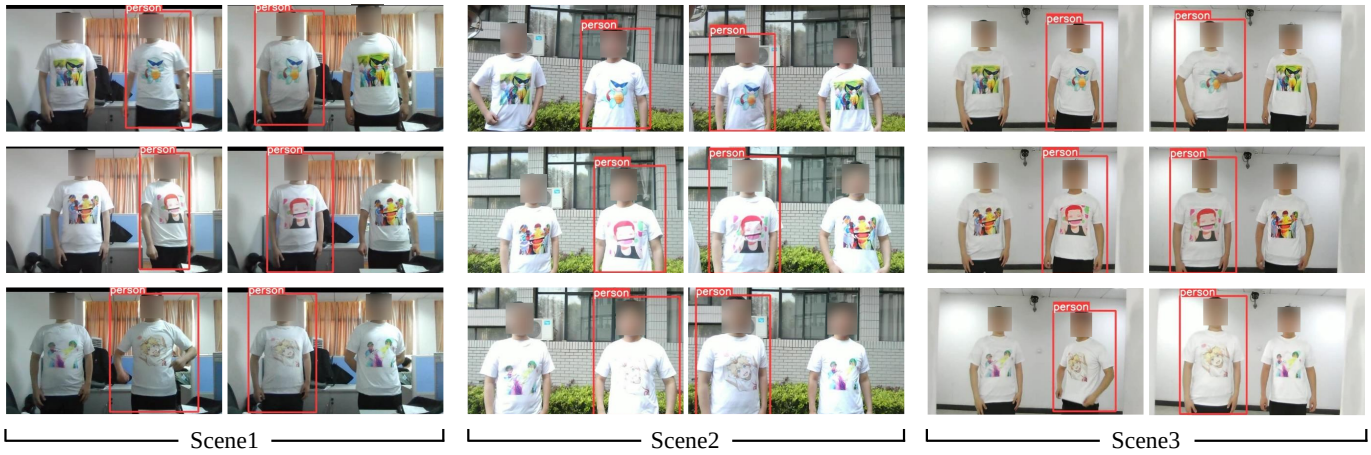


Fig. 9. The attacking performance in the physical world using three ACAPs printed on white T-shirts. Two participants appear together in the detection camera wearing T-shirts with the ACAP and corresponding original cartoon image, respectively. We test the effectiveness from three scenes, including indoor, outdoor, and different backgrounds.

Fig. 10. Display of the wearing visual performance and human parsing results for our ACAP.

### D. Ablation Study

**Influence of Multi-Scale Training.** To evaluate the influence of multi-scale training, we conduct a comparative experiment during the testing phase. First, with a fixed size for images that are input to the detector, we test the attack performance of related work (including GoogleAP, DPATCH, TextureAP and LAP, shown at the bottom of TABLE IV), and our patch (the second patch in Fig. 5). Second, in each iteration, we randomly scale the size of the image, as we discussed in Sec. III-E, and test the attack ability once again. The comparison results are provided in Fig. 7, we observe that the attack performance of all patches decreases after random scaling. TextureAP and LAP decrease the most with 41.49% and 40.25, respectively. For GoogleAP and DPATCH, since their attack abilities are not strong in the first place, the drop is not much. For our patch, we can observe that the AP dropped from 46.57% to 50.75%, which performs best (lowest Average Precision (AP)) with the addition of scaling. This means that with the multi-scale training, our generated patches gain scale-invariant features and attacking-robustness.

**Influence of Attack Weight $\lambda_1$.** For the total objective function eq.3, the weight of $\mathcal{L}_{obj}$ controls the attack ability of generated patches. In this experiment, we fix $\lambda_2 = 10^{-2}$, $\lambda_3 = 10^5$, and modify the $\lambda_1 \in \{0, 5, 10, 15, 20\}$. We report AP, PSNR, and SSIM [43], computing to evaluate attack ability (the stronger with lower AP) and visual quality (the better with higher SSIM and PSNR). The comparison results are presented in Fig. 8, we observe a trade-off between smaller $\lambda_1$ that are beneficial to optimize visual quality, and larger $\lambda_1$ that provide better attacking effect. Moreover, different style targets require appropriate weight $\lambda_1$ to generate adversarial patches, e.g. style Candy achieves respectable attacking effect at $\lambda_1 = 10$, while style Feathers requires a larger value.

### E. Physical World Attack

As for the physical world attack, we conduct several experiments to validate the practical effectiveness of our ACAPs. A Dell Inspiron 7400 laptop is used to record videos for testing. As shown in Fig. 9, we print our three ACAPs on white T-shirts and one participant wears them in turn. As contrast, another participant wears T-shirts with corresponding original cartoon images. We display our physical-world attack results in three different scenarios for comprehensive comparisons. We can observe that in all frames with different lighting and background, the detector can recognize the cartoon-patch person, implying the good detection effectiveness of YOLOv2. On the contrary, the person with our adversarial patches cannot be recognized by the detector. The experimental results demonstrate the robust attacking ability of our adversarial patches in the physical world.

To further evaluate the wearing performance of our ACAP from human perception, we apply the texture of our created patch (the first patch in Fig. 5) to the garment surface using a virtual simulated environment. Then, we analyze the attack effect using a human parsing model [44]. The evaluation results can be witnessed in Fig. 10. For visual quality, our ACAP demonstrates certain wearability. Meanwhile, ACAP can perturb the human parsing model, causing it to fail to consistently identify the various parts of the human body.

## V. Discussion and Conclusion

The appearance of current adversarial patches is abrupt and spotted easily by human observers. In this paper, we propose **A**esthetic yet **C**ustomizable **A**dversarial **P**atches (ACAPs) that are stylized, diversified, and attacking-robust, to attack the person detection model. To address the challenge of improving the visual quality of adversarial patches, we consider patch generation as an image transformation problem and introduce semantic contents and style features. Using our trained model, any image can be transformed into an adversarial patch end-to-end. Extensive experimental results show that our ACAPs have high quality in both visual effect and attack ability, and are capable of performing practical attack tasks in the physical world.

The proposed method is the first end-to-end method towards generating aesthetic patches while endowing them with adversarial attack abilities. We further find that styles with high color saturation and richness have stronger attack effects, which inspires us to explore which feature is critical to the attack. This meaningful problem needs to be addressed urgently but is still unclear. Now the successful application of our method in the physical space exposes the potential security risks of deep learning models when applied in the real world. In the future, we are interested in investigating the mechanism of adversarial attacks and defending against them.

## References


[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[3] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 269–16 279.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[5] M. Zeng, Y. Wang, and Y. Luo, "Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1267–1272.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] Z. Zhao, Z. Liu, and M. Larson, "Towards large yet imperceptible adversarial image perturbations with perceptual color distance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1039–1048.

[8] X. Chen, X. Yan, F. Zheng, Y. Jiang, S.-T. Xia, Y. Zhao, and R. Ji, "One-shot adversarial attacks on visual tracking with dual attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 176–10 185.

[9] T. Maho, T. Furon, and E. Le Merrer, "Surfree: a fast surrogate-free black-box attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 430–10 439.

[10] C. Ma, L. Chen, and J.-H. Yong, "Simulating unknown target models for query-efficient black-box attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 835–11 844.

[11] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," *arXiv preprint arXiv:2102.00436*, 2021.

[12] A. Kurakin, I. Goodfellow, S. Bengio *et al.*, "Adversarial examples in the physical world," 2016.

[13] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[14] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 1028–1035.

[15] Z. Kong, J. Guo, A. Li, and C. Liu, "Physgan: Generating physical-world-resilient adversarial examples for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 254–14 263.

[16] R. Duan, X. Mao, A. K. Qin, Y. Chen, S. Ye, Y. He, and Y. 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*, 2021, pp. 16 062–16 071.

[17] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[18] T. Li, A. Liu, X. Liu, Y. Xu, C. Zhang, and X. Xie, "Understanding adversarial robustness via critical attacking route," *Information Sciences*, vol. 547, pp. 568–578, 2021.

[19] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2018.

[20] C. Zhang, A. Liu, X. Liu, Y. Xu, H. Yu, Y. Ma, and T. Li, "Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity," *IEEE Transactions on Image Processing*, vol. 30, pp. 1291–1304, 2020.

[21] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6206–6215.

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

[23] X. Liu, H. Yang, Z. Liu, L. Song, Y. Chen, and H. Li, "Dpatch: An adversarial patch attack on object detectors," in *SafeAI@ AAAI*, 2019.

[24] S. Thys, W. Van Ranst, and T. 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*, 2019, pp. 0–0.

[25] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *European conference on computer vision*. Springer, 2020, pp. 665–681.

[26] J. Tan, N. Ji, H. Xie, and X. Xiang, "Legitimate adversarial patches: Evading human eyes and detection models in the physical world," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5307–5315.

[27] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu, "Bias-based universal adversarial patch attack for automatic check-out," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 395–410.

[28] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[29] S. Gu, C. Chen, J. Liao, and L. Yuan, "Arbitrary style transfer with deep feature reshuffle," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8222–8231.

[30] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[31] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in neural information processing systems*, vol. 30, 2017.

[32] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-time neural style transfer for videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 783–791.

[33] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[34] M. Sharif, S. Bhagavatula, L. Bauer, and M. 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*, 2016, pp. 1528–1540.

[35] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8565–8574.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[37] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[40] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[41] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. International Society for Optics and Photonics, 2003, pp. 87–95.

[42] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[43] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *2013 18th International Conference on Digital Signal Processing (DSP)*. IEEE, 2013, pp. 1–8.

[44] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.