

You Cannot Easily Catch Me: A Low-Detectable Adversarial Patch for Object Detectors

Zijian Zhu¹, Hang Su², Chang Liu¹, Wenzhao Xiang¹ and Shibao Zheng¹

¹Institute of Image Communication and Networks Engineering in the Department of Electronic Engineering (EE), Shanghai Jiao Tong University

²Department of Computer Science and Technology, Institute for AI, THBI Lab, Tsinghua University

Abstract

Blind spots or outright deceit can bedevil and deceive machine learning models. Unidentified objects such as digital “stickers,” also known as adversarial patches, can fool facial recognition systems, surveillance systems and self-driving cars. Fortunately, most existing adversarial patches can be outwitted, disabled and rejected by a simple classification network called an adversarial patch detector, which distinguishes adversarial patches from original images. An object detector classifies and predicts the types of objects within an image, such as by distinguishing a motorcyclist from the motorcycle, while also localizing each object’s placement within the image by “drawing” so-called bounding boxes around each object, once again separating the motorcyclist from the motorcycle. To train detectors even better, however, we need to keep subjecting them to confusing or deceitful adversarial patches as we probe for the models’ blind spots. For such probes, we came up with a novel approach, a Low-Detectable Adversarial Patch, which attacks an object detector with small and texture-consistent adversarial patches, making these adversaries less likely to be recognized. Concretely, we use several geometric primitives to model the shapes and positions of the patches. To enhance our attack performance, we also assign different weights to the bounding boxes in terms of loss function. Our experiments on the common detection dataset COCO as well as the driving-video dataset D²-City show that LDAP is an effective attack method, and can resist the adversarial patch detector.

1. Introduction

Computer vision for autonomous driving systems remains highly vulnerable to failures of object detection. Despite impressive recent improvements for locating and classifying

objects in digital images [26, 16, 11, 25], object detectors generally remain quite vulnerable to patch-wise adversarial threats [18, 29, 31, 13]. Deadly traffic accidents might result from attacks on detection-based autonomous driving systems. For example, an autonomous vehicle might hit a pedestrian not properly detected due to an adversarial attack.

Fortunately, most existing patch-wise adversarial examples can be distinguished readily from benign images, making it easy to train a classification network (called an **adversarial patch detector**) to automatically detect and reject these adversarial images. This simple defense method, also known as a detect-only defense approach [1], can disable most existing patch-wise adversarial attacks. We refer to this drawback of existing patch-wise adversarial attacks as **high detectability**. Although recognition of this drawback can help researchers easily add features to protect detection models from attacks, this approach also prevents researchers from revealing more potential security problems with these detection models. To overcome this drawback, it is important to study patch-wise attacks that can escape notice by an adversarial patch detector.

One possible way to design such an attack is to make adversarial examples indistinguishable from benign images, making it harder for the adversarial patch detector to recognize adversarial examples. To achieve this goal, we focus on two aspects of the attack method, including reducing the area of adversarial patches in the images, and increasing the texture consistency with the original images. Both aspects can reduce the difference between adversarial examples and original images, making adversarial examples less likely to be recognized by an adversarial patch detector.

1.1. Challenges

However, it is non-trivial to deceive an object detector while simultaneously reducing the area of adversarial patches, due to the following reasons.

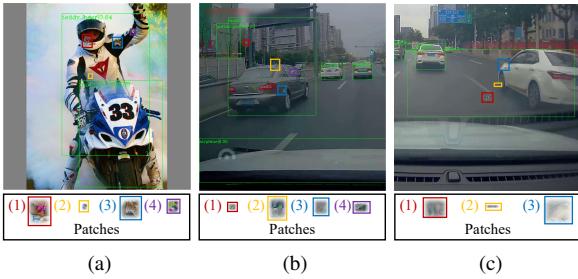


Figure 1: Examples of Ldap attacking Faster R-CNN [26]. (a): classification attack on COCO; the person is detected as a teddy bear with 64% confidence. (b): classification attack on D²-City; the car is detected as a bus with 91% confidence. (c): localization attack on D²-City; the predicted bounding box only has 40.63% IoU with the ground truth bounding box of the target object. We also zoom in on the patches and show each one of them below the images. We find that all these patches are small and maintain consistent texture with the original images.

1.1.1 Model the Patch Shapes Parametrically

It is non-trivial to model patch shapes parametrically, since their shapes can be flexible. One simple method is to model all patches in an input image as a binary mask covering the image, and exhaust any possible shapes of patches with respect to every pixel in it. However, this modeling makes the problem difficult to solve, for it is a high-dimensional discrete-search problem.

1.1.2 The Inner Mechanism of an Object Detector

Assuming that we are able to model the patches, it is still difficult to attack an object detector, especially with flexible patches. The difficulty mainly comes from the bounding-box prediction mechanism of an object detector. Generally the pipeline of an object detector is: firstly it predicts redundant bounding boxes on one object, then the detector classifies them, and lastly it uses Non-Maximum Suppression (NMS)¹ to select one bounding box from them [26, 16, 11, 25]. The attack should simultaneously influence all of these bounding boxes, which makes the attack on object detectors difficult. To address this issue, previous attacks with fixed patches adopted two strategies. The first was to attack the bounding box with a maximum confidence score for the target category [18, 6, 29]. However, during optimization the bounding box with maximum category score (*i.e.*, the output of the model ought to represent the confidence of classification) can be continuously changed, rendering the search-for-region unstable in our method. The second strategy was to equally attack the bounding boxes that can be detected as the target object, which is the object that the adversary aims to

¹NMS is a technique to select one bounding box from multiple bounding boxes based on their categories.

attack [13]. Nonetheless, most of these bounding boxes are inaccurate, misleading the region to sub-optimal positions. Therefore, neither of these strategies is suitable for our region-flexible attack.

1.2 Our Proposal

To address the aforementioned challenges, we propose a Low Detectable Adversarial Patch (**LDAP**). We formulate the generation of low-detectable patches as a joint optimization problem that corrupts the prediction of the object detector; reduces the area of patches; and simultaneously preserves the texture consistency with the original image. To achieve this, we solve the challenges from the following two perspectives.

1.2.1 The Use of Geometric Primitives to Model the Patches

Some previous research represents a given two-dimensional (2-D) shape with a set of geometric primitives (ellipses) [5][22]. They utilize both a Gaussian Mixture Model (GMM) to parameterize these ellipses as well as an Expectation-Maximization (EM) algorithm to maximize the similarity to the given shape. Referring to this work, we use a set of geometric primitives (rectangles) to model the patches with flexible shapes and positions. Concretely at first, based on a finding that patches in irregular shapes on an image plane can be approximated by several rectangles, we assume that the adversarial patches can be approximated by a set of rectangles. Therefore, we model the region of patches as a union of several rectangles. The division of these patches from other regions can be realized by a binary image mask, in which the patch regions are set to 1 and the others are set to 0. As a result, the optimization of shapes and positions of patches is converted to the optimization of shape and position parameters of these rectangles.

However, the boundaries of these rectangles are steep, making the loss function not differentiable with respect to their shape and position parameters. To address this issue, we relax the steep boundaries of these rectangles to smooth transitions from 1 to 0. These boundary-relaxed rectangles become continuous everywhere, enabling us to optimize their parameters using a gradient-based method. The relaxation makes the rectangles look feathered, as shown in the attack examples of LDAP in Fig. 1.

1.2.2 Soft-Attack Strategy

To overcome the difficulty from the inner mechanism of an object detector, we adopt a soft-attack strategy in the loss function, meaning that we assign different weights to multiple bounding boxes that can be detected as a target object. This strategy is based on the assumption that the attack should pay more attention to the bounding boxes that

are closer to the target object. This assumption is reasonable, since the closer bounding boxes are more likely to be the correct predictions. The weights are set as the Intersection over Union (IoU) between the predicted bounding box and the bounding box of the target object; IoU is a widely-used metric in detection literature for measuring the overlap between bounding boxes. The soft-attack strategy in loss function enables our attack to be more focused on the bounding boxes that are closer to the target object, a focus that benefits our attack when searching for patch positions.

We evaluate the effectiveness of LDAP on the common detection dataset COCO, and validate the threat of LDAP to an autonomous driving application on the driving-video dataset D²-City. We test LDAP on four mainstream detection models, including Faster R-CNN (FRCN) [26]; Mask R-CNN (MRCN) [11]; SSD [16]; and YOLOv3 [25]. As shown in Fig. 1, LDAP can deceive the detector with small and texture-consistent patches in these images.

We conclude our main contributions as follows:

- We propose a low-detectable patch-wise attack method. Based on a reasonable assumption, we formulate the low-detectable attack as a joint optimization problem that simultaneously takes into consideration the attack performance, patch area and texture consistency of patches;
- We model the shapes and positions of patches in a differentiable way, which enables us to use a gradient-based method to optimize shapes and positions of patches together with their textures;
- To overcome the difficulty from the inner mechanism of an object detector, we adopt a soft-attack strategy in our loss function, assigning different weights to bounding boxes that can be detected as the target object.

The remainder of the paper is organized as follows: In Section 2, we introduce previous research related to our methods and, in Section 3, we propose a low-detectable adversarial patch; our experimental results are presented in Section 4, followed by a conclusion.

2. Related Work

In this section, we review three related bodies of work in terms of object detection, adversarial attack and 2-D shape modeling.

2.1. Object Detection

First, we introduce object-detection models, which are the target models we aim to attack. Nowadays, mainstream object detectors can be classified into two-stage detectors and one-stage detectors.

2.1.1 Two-stage Detection

The prediction procedure of two-stage detectors consists of a region proposal followed by a classification. Detectors of this type first search for region proposals, then classify each of them. An early detector with a Convolution Neural Network (CNN) is Overfeat [27], which combines a sliding window with a CNN to do the detection. After that, a so-called Regions with CNN features (R-CNN) detector is proposed [9]. R-CNN uses selective search to get proposals, and classifies them based on CNN features. After R-CNN, a number of studies were published to speed up the detection process and improve accuracy. Representative modern detectors include Fast R-CNN [8], Faster R-CNN [26] and Mask R-CNN [11].

2.1.2 One-stage Detection

One-stage detectors, also known as “single-shot detectors,” simultaneously predict both the bounding boxes and the category scores with just a single pass through the network. The representative one-stage detectors are YOLO series [23][24][25][3] and SSD [17]. Compared with two-stage detectors, one-stage detectors run faster, but usually get lower detection accuracy. In our experiments, we find that we need smaller patches to attack one-stage detectors than two-stage detectors, a finding that indicates that one-stage detectors are easier to attack.

2.2. Adversarial Attacks

Adversarial attacks on computer vision systems have been widely studied. In the research field of object-detection attack, many attack algorithms have been proposed in recent years. From the perspective of manipulated region, we classify these adversarial attacks into pixel-wise attacks and patch-wise attacks.

2.2.1 Pixel-wise Attacks

Pixel-wise attacks change all the pixels of the target object or the whole image slightly; consequently, only rarely can human beings perceive such perturbations. This kind of attack is firstly studied in classification attacks, *e.g.*, FGSM [10], DeepFool [20] and PGD [19], and then it is extended to object detection attack. Xie *et al.* [32] propose Dense Adversary Generation (DAG), which is an attack method for detection and segmentation. In detection attack, DAG changes pixels of the whole image to increase the wrong-class confidence and decrease correct-class confidence. Wei *et al.* [30] propose an attack scheme based on a conditional Generative Adversarial Network to generate an adversarial perturbation that changes the pixels in the whole image to deceive the detector. Nezami *et al.* [21] propose an attack method that manipulates the pixels of the target object and precisely

changes the label of the target object, without changing the labels of other objects.

2.2.2 Patch-wise Attacks

These attacks try to deceive the detector by manipulating small regions in the input image, called patches. The positions of the patches are usually predefined; for example, the upper-left corner of the input image or the center of the target object. DPatch [18] is an early work of a patch-wise object detection attack. It attaches an adversarial patch on the upper-left corner of an image, and is able to disturb the digital object-detection. Some methods that emerged later are capable of being transferred from the digital world to the physical. Eykholt *et al.* [6] extend the Robust Physical Perturbations (RP_2) algorithm [7] from classification attack to object-detection attack. The sticker version of their extended RP_2 algorithm, denoted as RP_2 -Sticker, puts two adversarial patches on a stop sign, and is able to mislead the detector to predict the stop sign as other objects. Thys *et al.* [29] train their adversarial pattern, denoted as AdvPatch, and print it on a paper board. The volunteers holding this paper board successfully evade the detection of YOLOv2 [24]. Meanwhile, Xu *et al.* [33] print their adversarial patterns on the front of a T-shirt while considering the non-rigid deformation of the pattern caused as the wearer’s pose changes. Volunteers wearing this T-shirt are neglected by the detector. Huang *et al.* [13] propose Universal Physical Camouflage (UPC), which can mislead detectors to predict the volunteers as dogs when they wear clothes with specially designed camouflages.

However, the adversarial patches of these patch-wise attacks are all large and have inconsistent textures compared with the original image, making them easy to be caught by an adversarial patch detector and then rejected. This defense method can disable most existing patch-wise attack methods. We regard this drawback as high detectability. In our method, we try to establish a low-detectable patch-wise digital attack, to escape the adversarial patch detector.

2.3. 2-D-Shape Modeling

One aspect of our method involves modeling the shapes of patches. There exists some research about how to model a 2-D shape. Da and Kemp [5] use multiple geometric primitives (ellipses) to fit a given human upper-body silhouette, which is a complex shape, then model the shape with GMM, and utilize unconstrained Expectation-Maximization (EM) to optimize the parameters of GMM to fit the given shape. Paniagiotakis and Argyros [22] design a shape-modeling method, also using geometric primitives (ellipses) to represent the given shape. Modeling the ellipses with GMM, they use hard EM and Akaike Information Criterion to optimize the parameters of GMM to approximate the given shape. We bor-

row their modeling idea, and assume that a complex shape, even several shapes, can be modeled as the combination of several geometric primitives. We introduce how we model the shapes of patches in detail in Section 3.3.

3. Methodology

It is complicated to generate a low-detectable adversarial patch. In Section 3.1, we first formulate the problem in such a way that we reduce the combination of three loss functions. Then we introduce each loss function in Sections 3.2 to 3.4. Finally, we introduce the optimization method in Section 3.5.

3.1. Problem Formulation

The goal of our LDAP is to deceive the object detector with small and texture-consistent patches. We set this goal based on an assumption that both aspects can help us evade an adversarial patch detector. The assumption is reasonable, since small patches can reduce the area of the manipulated part in the image, and consistent texture in these patches can reduce the magnitude of manipulation. Both aspects help us decrease the difference between our adversarial example and the benign image, making it less likely for an adversarial patch detector to distinguish them.

Towards this goal, given an input image $I \in R^{W \times H^2}$, where W and H are the width and height of the input image, respectively, we formulate the problem of LDAP as solving for optimal perturbation:

$$\delta^* = \arg \min_{\delta} L_{\text{attack}}(I + \delta) + \lambda_1 L_{\text{area}}(\delta) + \lambda_2 L_{\text{texture}}(\delta), \quad (1)$$

where λ_1, λ_2 are weights; $\delta \in R^{W \times H}$ is the adversarial perturbation containing all adversarial patches; L_{attack} is the attack loss that ensures attack performance in different attack tasks; L_{area} is the area loss that constrains the area of patches; and L_{texture} is the texture loss that constrains the texture of patches. By solving this joint problem, we can get a proper δ that can deceive the object detector with small and texture-consistent patches.

3.1.1 Introduce texture and mask layers

In order to define the shapes, positions and textures of the patches clearly, we factor δ into an element-wise multiplication of a texture layer $t \in R^{W \times H}$ and a mask layer $M \in \{0, 1\}^{W \times H}$. The texture layer encodes the textures of patches, while the mask layer encodes the shapes and positions of patches. We divide the patches and the other regions with this binary mask layer M ; the regions of patches in M are set to 1 and the others are set to 0. Then the perturbation (the aggregation of all patches) can be cast as:

$$\delta = t \odot M, \quad (2)$$

²For clarity, we omit the 3-channels (RGB) for images and perturbations in all formulae, which does not influence the solving of our problem.

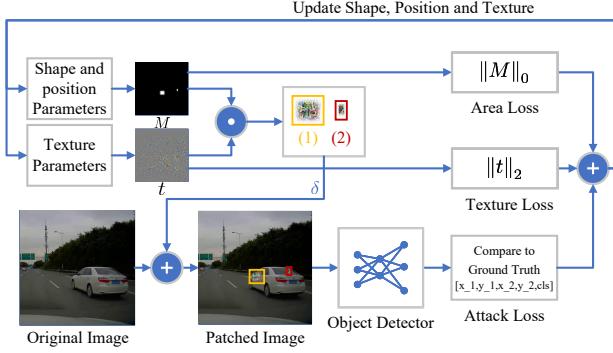


Figure 2: The overview of our attack. The adversarial patches are calculated from the element-wise product of the mask layer M and the texture layer t . The area loss and the texture loss are directly calculated from M and t , respectively. The attack loss is calculated by comparing the predictions of the object detector with the ground truth of the target object. We use the sum of the area loss, texture loss and attack loss to update the shape, position and texture of patches.

where \odot denotes the element-wise multiplication. One advantage of this factoring is that it separates out the shape and position terms and the texture term of the patches, making it easy to define L_{area} directly on M .

Based on the definition of mask layer M and texture layer t , the overall low-detectable attack problem is reformulated as:

$$\min_{M, t} L_{\text{attack}}(I + t \odot M) + \lambda_1 L_{\text{area}}(M) + \lambda_2 L_{\text{texture}}(t). \quad (3)$$

In this formulation, the area loss and texture loss are only related to M and t , respectively, which makes it easier to separately optimize shapes and positions of patches, and textures of patches. This separation help us overcome the problem of coupling these two parts during optimization³.

The overview of our attack method is shown in Fig. 2. Given an input image, we first generate patches from shape, position and texture parameters, then add them onto the input image. The patched image is then passed into the object detector to calculate the attack loss. The area loss and the texture loss are computed directly from M and t . Then we sum the loss functions and back-propagate to update the shape, position and texture parameters. In the following three sections, we introduce these loss functions in detail.

3.2. Attack Loss

First, we introduce the attack-loss function, which aims to corrupt the prediction of the detector. We introduce the soft-attack strategy we adopt in attack-loss functions and then

³The coupling means that the values of shapes and positions will influence the optimal values of the textures and vice versa. If we update both sides at the same time, the optimization can be unstable.

the different attack tasks, followed by their corresponding attack-loss functions.

3.2.1 Soft-attack strategy

Previous attack methods adopt two common strategies when dealing with the redundant bounding boxes predicted on the target object. The first strategy is to attack only the bounding box with the maximum confidence score of the target category [18, 6, 29]. However, during attack optimization, the bounding box with the maximum confidence score could be changed from one bounding box to another continuously during optimization. This not only slows down the optimization, but also leads to inferior positions of patches and worse attack performance, resulting in larger patches, lower texture consistency and higher detectability.

The second strategy is to equally attack the bounding boxes that can be detected as the target object [13], which means that the predicted category scores of these bounding boxes—and the IoU between them and the target object—are higher than the thresholds of the detection model. This solution can alleviate the drawback of the first strategy, for these bounding boxes usually change less frequently than the bounding box with the maximum confidence score. However, most of these bounding boxes are not accurate, even though they are treated equally in this strategy. In our method, this leads to sub-optimal positions of patches and worse attack performance.

Therefore, we use a soft-attack strategy in our attack-loss function instead of these attack strategies. The soft-attack strategy means that we assign different weights to the bounding boxes that can be detected as the target object. Here, we have an assumption that the bounding boxes closer to the target object are more worth attacking, for they are more likely to be the correct predictions. Based on this assumption, we set the IoU between the predicted bounding boxes and the bounding box of target object as the weights of each bounding box in the attack-loss function. The soft-attack strategy concentrates our LDAP on the target object, and help us reduce the area of patches needed to deceive the detector.

3.2.2 Different attack-loss functions

Different from the classifier, an object detector incorporates two parts: object classification and localization. This makes the attack on a detector more complicated. Correspondingly, we set a classification-attack task and a localization-attack task that only focus on one function. For different tasks, we set different attack-loss functions.

Classification attack means that the adversary should make the detection model classify the target object as a wrong category. If we regard “background” as a special category, then an object’s disappearance is also a special

case of classification attack. In this paper, we study a danger for autonomous driving: a disappearing attack, which is an important attack task for not sighting a pedestrian ahead. For a classification attack, we set the weighted sum of category scores as an attack loss. The loss function is:

$$L_{\text{cls}}(I + t \odot M) = \sum_{b_i \in \mathcal{B}^*} w_i C(I + t \odot M, b_i), \quad (4)$$

where \mathcal{B}^* is the bounding boxes that could be detected as a target object; b_i is the i -th bounding box in \mathcal{B}^* ; b_t is the bounding box of the target object; $w_i = \text{IoU}(b_i, b_t)$ is the weight of b_i , which is assigned by our soft-attack strategy; and $C(\cdot)$ is the category score of the target category for b_i predicted by the object detector on patched image $I + t \odot M$.

Localization attack means that the adversary should make the detection model predict bounding boxes with undesirable shape or position on the target object. In this paper, we investigate the attack on a bounding-box position, aiming to shift the bounding box horizontally. This attack task is important due to its potential threat to autonomous vehicles. For instance, if a bus ahead of an autonomous car is wrongly detected as being in the left lane, the car might drive forward and hit the bus. In a localization attack, we set the weighted sum of regression offsets of bounding boxes as an attack loss. The loss function is:

$$L_{\text{loc}}(I + t \odot M) = \sum_{b_i \in \mathcal{B}^*} w_i R_x(I + t \odot M, b_i), \quad (5)$$

where \mathcal{B}^* is the bounding boxes that could be detected as a target object; b_i is the i -th bounding box in \mathcal{B}^* ; b_t is the bounding box of the target object; $w_i = \text{IoU}(b_i, b_t)$ is the weight of b_i ; and $R_x(\cdot)$ is the x -dimensional regression offset of b_i .

Therefore, the attack loss L_{attack} is defined as:

$$L_{\text{attack}} = \begin{cases} L_{\text{cls}}, & \text{if a classification attack} \\ L_{\text{loc}}, & \text{if a localization attack.} \end{cases} \quad (6)$$

3.3. Area Loss

It is easy to see that the area loss can be defined as

$$L_{\text{area}} = \|M\|_0. \quad (7)$$

However, directly solving for M is a non-trivial high-dimensional discrete-search problem, since M consists of $W \times H$ binary elements. (W and H are the width and height of the original image, respectively). To circumvent this non-trivial problem, while referring to the aforementioned shape-modeling research [5, 22], we use a set of geometric primitives to represent the mask layer.

Noticing that patches in irregular shapes and positions can be approximated by several rectangles, as shown in Fig. 3, we make a reasonable assumption that the mask layer in our

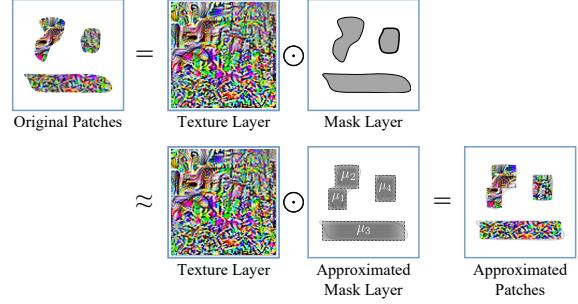


Figure 3: Patch in irregular shapes and positions can be approximated by several rectangles.

attack can be approximated by a set of geometric primitives. Therefore, we represent the mask layer as:

$$M = \bigcup_j^N \mu_j, \quad (8)$$

where $\mu_j \in \{0, 1\}^{W \times H}$ is the j -th geometric primitive, and N is the number of primitives. Each primitive can be regarded as a degraded mask layer, which encodes only one part of the whole mask layer with a simple shape. In this way, we convert the optimization of M to an optimization of parameters of primitives, dramatically reducing the number of parameters to optimize.

Without loss of generality, we first choose rectangles as geometric primitives, a reasonable choice because both the receptive fields and the predicted bounding box of detection models are rectangles. Almost all previous patch-wise attacks use rectangle patches, such as those of DPatch [18], AdvPatch [29], RP₂-Sticker [6] and UPC [13]. The rectangle primitive is written as:

$$\mu_j(\mathbf{x}; \mathbf{x}_j^c, \mathbf{s}_j) = 1, \text{ if } |x_k - x_{j,k}^c| \leq \frac{s_{j,k}}{2}, k = 1, 2, \quad (9)$$

where $\mathbf{x} = (x_1, x_2)$ denotes the coordinates of pixels in the image; $\mathbf{x}_j^c = (x_{j,1}^c; x_{j,2}^c)$ denotes the position parameters (coordinates of the center point) of the j -th rectangle; and $\mathbf{s}_j = (s_{j,1}; s_{j,2})$ denotes its shape parameters (width and height). The demonstration of a rectangle primitive μ_j with its parameters is shown in Fig. 4.

3.3.1 Solving Non-differential Problem

However, the rectangle primitive defined in Eq. (9) induces a steep change from 1 to 0 on the boundary of the rectangle, resulting in non-differentiability of the loss function with respect to the position and shape parameters of the primitives \mathbf{x}_j^c and \mathbf{s}_j . To address this non-differentiability issue, we relax the boundary of a rectangle primitive using a cosine

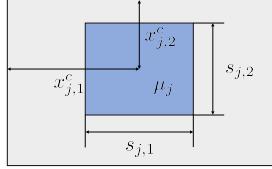


Figure 4: Parameters of j -th rectangle base μ_j . The blue region is filled with 1, and the gray region is padded with 0.

function. The boundary-relaxed primitive is:

$$\mu_j(\mathbf{x}; \mathbf{x}_j^c, \mathbf{s}_j) = \phi\left(\frac{2\pi}{s_{j,1}}(x_1 - x_{j,1}^c)\right)\phi\left(\frac{2\pi}{s_{j,2}}(x_2 - x_{j,2}^c)\right),$$

where $\phi(z) = \begin{cases} \frac{1}{2}(\cos(z + \frac{\pi}{2}) + 1), & \text{if } -\frac{3\pi}{2} < z < -\frac{\pi}{2} \\ 1, & \text{if } -\frac{\pi}{2} < z < \frac{\pi}{2} \\ \frac{1}{2}(\cos(z - \frac{\pi}{2}) + 1), & \text{if } \frac{\pi}{2} < z < \frac{3\pi}{2} \\ 0, & \text{else} \end{cases}$.

(10)

We find that the cosine-boundary primitive obtains a smooth boundary, continuous at all points, making the loss function differentiable with respect to its position and shape parameters.

After the relaxation of the boundary, the value range of a primitive changes from discrete to continuous. Therefore, we rewrite the union of primitives in Eq. (8) as a weighted sum of primitives such that:

$$M = \sum_j^N \alpha_j \mu_i(\mathbf{x}_j^c, \mathbf{s}_j), \quad (11)$$

where α_j is the weight we set for the j -th primitive, and $\alpha_j > 0$. This rewriting implies an assumption that the primitives do not overlap each other in the mask layer. This assumption is reasonable because **(i)** the overlap is a waste of the area of primitives, and **(ii)** the overlapped primitives can be replaced equivalently with non-overlapped primitives.

Based on Eq. (11), we can expediently convert the l_0 norm of M in Eq. (7), which is difficult to optimize, to the sum of the area of each primitive. The area loss is:

$$L_{\text{area}} = \sum_j^N s_{j,1} s_{j,2}, \quad (12)$$

where $s_{j,1}$ and $s_{j,2}$ are the width and height of a j -th rectangle primitive.

3.4. Texture Loss

Texture loss aims to maintain consistency between the texture of patches and that of the original image. Therefore, we directly use the l_2 norm of adversarial texture t as our texture loss. The texture loss is:

$$L_{\text{texture}}(\delta) = \|t\|_2. \quad (13)$$

Finally, we substitute Eq. (6), Eq. (10), Eq. (11), Eq. (12) and Eq. (13) to Eq. (3) to calculate the adversarial perturbation δ , which is the aggregation of all patches.

3.5. Optimization

As we mentioned before, the optimization of $M(\mathbf{x}^c, \mathbf{s}, \boldsymbol{\alpha})$ and t in the problem defined in Eq. (3) are coupled, which can destabilize our optimization and cause sub-optimal attack performance. Here $\mathbf{x}^c = \{\mathbf{x}_j^c\}$, $\mathbf{s} = \{\mathbf{s}_j\}$, $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_j\}$, and $j = 1, \dots, N$. Therefore, we separate the problem into two separated aspects, namely a *Mask Layer Search* and a *Texture Layer Search*, and alternately optimize M or t step-by-step, to indirectly solve the problem.

3.5.1 Mask Layer Search

During this search step, we fix the texture t and optimize the mask layer M controlled by \mathbf{x}^c , \mathbf{s} and $\boldsymbol{\alpha}$. Due to the separation of the mask layer and texture layer, we can directly omit the texture loss in this step. The problem in this step is:

$$\min_{\mathbf{x}^c, \mathbf{s}, \boldsymbol{\alpha}} L_{\text{attack}}(I + t \odot M(\mathbf{x}^c, \mathbf{s}, \boldsymbol{\alpha})) + \lambda_1 L_{\text{area}}(\mathbf{s}). \quad (14)$$

Note that during optimization, $\boldsymbol{\alpha}$ is coupled with \mathbf{x}^c and \mathbf{s} , such that we can separate the updates of \mathbf{x}^c , \mathbf{s} and $\boldsymbol{\alpha}$. We denote the objective function defined in Eq. (14) as J . First we fix $\boldsymbol{\alpha}$ and calculate the gradient of J with respect to \mathbf{x}^c and \mathbf{s} to update them,

$$\mathbf{x}^c \leftarrow \mathbf{x}^c - \eta_x \frac{\partial J}{\partial \mathbf{x}^c}, \mathbf{s} \leftarrow \mathbf{s} - \eta_s \frac{\partial J}{\partial \mathbf{s}}, \quad (15)$$

where η_x , η_s are the learning rates of \mathbf{x}^c and \mathbf{s} , respectively. Then we fix \mathbf{x}^c and \mathbf{s} and update $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta_\alpha \frac{\partial J}{\partial \boldsymbol{\alpha}}, \quad (16)$$

where η_α is the learning rate of $\boldsymbol{\alpha}$. After these two updates, we finish one region-search step, in which we update the mask layer M based on the fixed texture layer.

3.5.2 Texture Layer Search

After one step of a region search, we fix the mask M and update the texture t for one step. Ignoring the terms that are not related to t in Eq. (3), we write the texture-layer search problem as:

$$\min_t L_{\text{attack}}(I + t \odot M(\mathbf{x}^c, \mathbf{s}, \boldsymbol{\alpha})) + \lambda_2 L_{\text{texture}}(t). \quad (17)$$

We also calculate the gradient to update the texture t for one step, such that:

$$t \leftarrow t - \eta_t \frac{\partial J'}{\partial t}, \quad (18)$$

where η_t is the learning rate of t , and J' is the objective function of Eq. (17). In this step, we update the texture layer based on the fixed-mask layer.

Algorithm 1 Algorithm of Ldap for One Image

Input: Image I ; Number of rectangle primitives N ; bounding box of target object b_t ; learning rates $\eta_x, \eta_s, \eta_\alpha, \eta_t$; initial coefficients λ_1, λ_2 and their increasing rate $\Delta\lambda$; maximum search step n_{\max} .

Output: Adversarial perturbation δ :

- 1: initialize texture layer t , weights of bases α , position and shape parameters of primitives x^c, s
- 2: step $\leftarrow 0$
- 3: **while** step $< n_{\max}$ **do**
- 4: search mask layer; update x^c and s according to Eq. (15)
- 5: search mask layer; update α according to Eq. (16)
- 6: search texture layer; update t according to Eq. (18)
- 7: increase λ_1, λ_2 according to Eq. (19)
- 8: step \leftarrow step +1
- 9: **end while**
- 10: compute perturbation $\delta \leftarrow M(x^c, s, \alpha) \odot t$

3.5.3 Increasing λ_1 and λ_2

For different images, the weights of area loss and texture loss should not be always the same, since objects in different images are usually not the same size (smaller objects usually need smaller patches to attack, so the weight of area loss should be bigger). To overcome this problem, we set the value of λ_1 and λ_2 to gradually increase after the texture-search step. This setting can ensure that, for different images, Ldap can find small enough and texture-consistent patches,

$$\lambda_1 \leftarrow \lambda_1 + \Delta\lambda, \lambda_2 \leftarrow \lambda_2 + \Delta\lambda, \quad (19)$$

where $\Delta\lambda$ is the increasing rate of λ_1 and λ_2 . After this step, we turn to the next step of region search. The overall procedure of our method is illustrated in Algorithm 1.

4. Experiments

In this section, we experimentally evaluate our approach’s detection-attacking effectiveness, and the resulting threat to autonomous vehicles. In Section 4.1, we introduce the experiment settings. In Section 4.2, we verify the effectiveness of our method compared to other attacks on COCO, which is a common detection dataset. In Section 4.3, we demonstrate two potential attack threats in autonomous driving on D²-City, which is a driving-record video dataset. Then we do some ablation studies in Section 4.4.

4.1. Experiment Settings

Without loss of generality, we choose “Person” in COCO and “Car” in D²-City as the target categories, both important categories in the field of autonomous driving, and widely studied in previous research [29, 33, 13].



Figure 5: Images selected from the COCO dataset: (a) people with animals; (b) people with motorbikes; (c) people playing sports; (d) children; and (e) people in black-and-white photos.

4.1.1 Datasets

We use two datasets for different purposes. To validate the effectiveness and superiority of our method in a common detection scenario, we select images from the widely used detection dataset COCO. To demonstrate the threats of our method in an autonomous driving scenario, we select images from the driving-record video dataset D²-City.

- **COCO** is a large dataset for both object detection and segmentation, containing more than 330,000 images [15]. We select 1,454 person images from the COCO 2017 training set. Some of the images we selected from the COCO dataset are shown in Fig 5.
- **D²-City** is a driving-record video dataset collected in five cities, and containing 355 video segments. We sample the videos at the interval of one per second, then select 1,096 frames that contain cars from the train set of D²-City. Some of the images that we selected are shown in Fig 6.

All of the target objects in the selected images satisfy the following criteria: **(i)** large enough; **(ii)** correctly detected by all detectors we test on; and **(iii)** not occluded by other objects. All of these criteria ensure the difficulty of attacking detectors.

4.1.2 Detection Models

We attack four representative object detection models in our experiments: two-stage models including Faster R-CNN (FRCN) [26] and Mask R-CNN (MRCN) [11], and one-stage models including SSD [16] and YOLOv3 [25]. Specifically, Faster R-CNN and Mask R-CNN use Resnet-50 [12] and FPN [14] as backbones, while SSD and YOLOv3 use VGG-16 [28] and Darknet-53 as backbones, respectively. All of these detectors are implemented by mmdetection [4]⁴.

⁴<https://github.com/open-mmlab/mmdetection>



Figure 6: Images selected from the D^2 -City dataset: (a) opposite cars; (b) cars directly in front; (c) cars on the right-front side; (d) cars on the left-front side; and (e) parked cars.

4.1.3 Parameters Setting

We set the initial value of α_1, α_2 to $1e-2$ and the increasing rate $\Delta\lambda$ to $1e-4$; the maximum search step n_{\max} to 2,000; and the number of rectangle primitives N to 10.

4.1.4 Comparison Methods

We compare our method with four former attack methods as well as with two attack methods designed by us. All comparable attack methods adopt fixed positions of patches. Furthermore, we reproduce the codes of all former attacks to fit the PyTorch and mmdetection frameworks.

- **DPatch** [18] aims to reduce the maximum category score of the target class in all bounding boxes with a rectangle patch in the upper-left corner of the input image.
- **AdvPatch** [29] aims to reduce the maximum objectness score in all predicted bounding boxes, using a rectangle patch in the center of the target object. However, reducing the objectness score of the bounding box has nearly no influence on Faster R-CNN, Mask R-CNN and SSD; therefore, we change AdvPatch to reduce category scores in these models.
- **RP_2 -Sticker** [6] aims to reduce the maximum category score of the target class in all bounding boxes, using two rectangle patches in the centers of the upper- and lower-half of the target object.
- **UPC** [13] aims to reduce the sum of the category scores of the target class of the bounding boxes that can be detected as the target object. We reproduce the code of the 7-Patterns version of UPC (**UPC-7P**), which puts adversarial patterns on seven certain parts of the human body (upper arms, thighs, calves and abdomen) to deceive the detector.

- **2-Rectangles-Horizontal** (2-Rects-H) attack is the first comparison method designed by us. Unlike the RP_2 -Sticker, our 2-Rects-H puts two rectangle patches in the center of the left- and right-half of the target object, and also shares the attack loss of our LDAP.

- **4-Rectangles** (4-Rects) attack is the second comparison method that we design, and puts four rectangle patches in the center of the upper-left, upper-right, lower-left and lower-right quarters of the target object, while also sharing the attack loss of our LDAP.

For all comparison methods, we start with small patches, and gradually increase the patches' size until the attack succeeds. With this setting, we can calculate the minimum region needed to deceive the detector of these methods. Considering that the experiments are all conducted in digital settings, we remove the physical modules in AdvPatch, RP_2 -Sticker and UPC-7P, such as Expectation of Transformation [2], image projection and printed color shift modeling. Since the original DPatch, AdvPatch, RP_2 -Sticker and UPC-7P are not able to perform a localization attack, we replace their loss functions with our attack-loss function in localization-attack task mode, but maintain their manipulated region settings.

4.1.5 Adversarial Patch Detector

To the best of our knowledge, there is no proper adversarial patch-detection method for an object-detection attack. However, we find that existing patch-wise adversarial attacks are not only visually perceptible to humans, but also easy to detect by a simple two-category classification network. With a small number of labeled pairs of adversarial examples as well as corresponding clean images, we can easily train a classifier to determine whether the input image is adversarial or not, with 90% accuracy.

In our experiments for our adversarial patch detector, we use a ResNet-18 network [12], followed by one fully connected layer—all very simple yet still effective for adversarial patch detection. With COCO, we use 500 adversarial examples and their corresponding 500 clean images for training, 200×2 for validation and 500×2 for testing. In D^2 -City, we use 400×2 images for training, 200×2 for validation and 400×2 for testing. To overcome the overfitting, we use the network with the best validation performance to test. Therefore, the detectability of different attack methods is measured by the accuracy of the adversarial patch detectors.

4.1.6 Evaluation Metric

For all detection models, we set the confidence threshold as 0.3, the same as the default threshold in mmdetection.

For attack performance, we use Attack Success Rate (ASR) under a specific attack area threshold to evaluate

the attack performance under the same patch area constraint. For a certain image, the attack succeeds if **(i)** the attack deceives the detector, and **(ii)** the sum of the patches’ area is lower than this threshold. The threshold will be specifically declared in different tasks for different detectors. The ASR is:

$$\text{ASR} = \frac{N_A}{N_I}, \quad (20)$$

where N_I denotes the number of images in the dataset, and N_A denotes the number of successfully attacked images under the patch area constraint.

For reducing patch area, we define the Mean Area Ratio (MAR) at $\text{ASR} \geq 95\%$ to evaluate the ability to reduce the patch area. MAR is calculated by the average ratio of manipulated area and object area over all successfully attacked objects in dataset, which is:

$$\text{MAR} = \frac{1}{N_A} \sum_n \frac{R_{A,n}}{R_{O,n}}, \quad (21)$$

where $R_{A,n}$, $R_{O,n}$ are the summed areas of the patches as well as the target object region area on the n -th adversarial example, respectively.

For texture consistency, we use Learned Perceptual Image Patch Similarity (LPIPS) [34] as our texture consistency metric to evaluate the consistency between adversarial examples and benign images. LPIPS is a widely used metric to evaluate perceptual similarity between images. The LPIPS is defined as:

$$\text{LPIPS} = \frac{1}{N_A} \sum_n \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|f_{hw}^l - f_{0hw}^l\|_2^2, \quad (22)$$

where f_{hw}^l and f_{0hw}^l are respectively the features of the n -th adversarial example and the corresponding benign image on the l -th layer of the VGG-network. The lower LPIPS indicates higher texture consistency.

For detectability, we define Adversarial Detection Accuracy (ADA) to evaluate the detectability of each attack.

$$\text{ADA} = \frac{TP + TN}{N_A \times 2}, \quad (23)$$

where TP is the number of True Positive images, and TN is the number of True Negative images. We multiply N_A by 2 because we test the detection accuracy on both adversarial examples and benign images.

4.2. Experiments on COCO

To validate the effectiveness and superiority on attack performance, small patches, consistent texture and low detectability of LDAP, we do experiments on the commonly used detection-dataset COCO. For different attack tasks, we introduce the experiments separately.

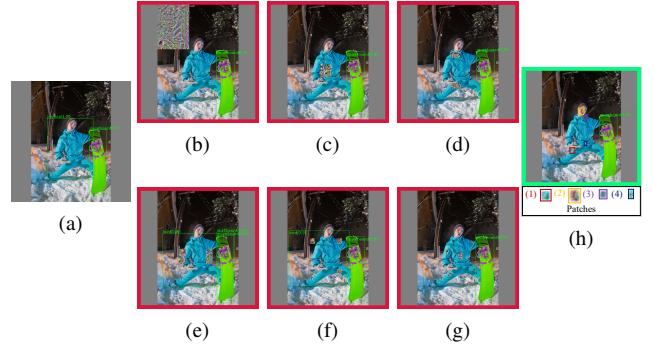


Figure 7: Adversarial examples of classification attack on Faster R-CNN [26] on one image from COCO. The detection results are shown in images. Some attacks eliminate the bounding boxes, while others mislead Faster R-CNN to detect the person to be a bird. The image with a red frame means that the image is recognized as an adversarial example by the adversarial patch detector, while a green frame means not recognized. We define P_a as the probability of being adversarial as predicted by an adversarial patch detector, of which the probability threshold is 50%. The detection results and P_a of attack methods are: (a) original image, person = 100%; (b) DPatch [18], nothing detected, $P_a=90.28\%$; (c) AdvPatch [29], nothing detected, $P_a=66.84\%$; (d) RP2-Sticker [6], nothing detected, $P_a=65.56\%$; (e) 2-Rects-H, bird = 96%, $P_a=71.55\%$; (f) 4-Rects, bird = 76%, $P_a=60.03\%$; (g) UPC-7P [13], nothing detected, $P_a=57.29\%$; and (h) LDAP, nothing detected, $P_a=24.06\%$.

4.2.1 Classification Attack

Classification attack means the attack should mislead the model to misclassify the target object. As we discussed in Section 3.2.2, a disappearing attack is an important classification-attack task to study, so we study it in this section. The attack success criterion is that there is no bounding box predicted in the correct category that overlaps with the bounding box of the target object. This means that simply changing the size of a bounding box (not making it disappear) is not considered to be a successful disappearing attack in our problem setting.

Demonstrations of different attack methods on a classification-attack task on Faster R-CNN are shown in Fig. 7. We find that: **(i)** all attack methods successfully attack the Faster R-CNN, while only LDAP passes the detection of the adversarial patch detector; **(ii)** as for region and texture consistency, we see that LDAP takes the least patch area, while the texture of patches generated by LDAP is the most consistent with that of the original image (in contrast, DPatch takes the largest patch); **(iii)** although the number

Table 1: Performance of attack methods on classification attack on the COCO dataset. ASR denotes Attack Success Rate; “Area Thr” denotes Area Threshold; MAR denotes Mean Area Ratio; and ADA denotes adversarial detection accuracy.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
FRCN	DPatch [18]	0 (10)	96.70	101.3	97.2
	AdvPatch [29]	6.18 (10)	18.57	26.49	87.5
	RP ₂ -Sticker [6]	7.35 (10)	14.72	26.00	88.6
	UPC-7P [13]	47.86 (10)	10.12	28.85	89.8
	2-Rects-H	10.66 (10)	18.41	32.91	89.7
	4-Rects	35.96 (10)	12.82	31.43	91.0
	LDAP	63.61 (10)	9.36	12.23	73.1
MRCN	DPatch [18]	0 (10)	97.42	101.7	96.6
	AdvPatch [29]	4.47 (10)	21.01	30.73	88.9
	RP ₂ -Sticker [6]	20.42 (10)	16.58	28.75	91.2
	UPC-7P [13]	40.16 (10)	11.40	30.67	91.4
	2-Rects-H	7.42 (10)	19.85	35.51	91.6
	4-Rects	20.97 (10)	15.81	35.79	92.4
	LDAP	60.93 (10)	9.70	15.12	71.1
SSD	DPatch [18]	2.68 (7)	39.07	47.3	96.5
	AdvPatch [29]	59.83 (7)	7.17	13.13	78.4
	RP ₂ -Sticker [6]	45.87 (7)	8.89	18.20	86.8
	UPC-7P [13]	47.11 (7)	8.61	26.0	92.0
	2-Rects-H	44.70 (7)	9.02	20.42	87.6
	4-Rects	41.26 (7)	9.59	25.93	93.4
	LDAP	63.06 (7)	6.39	9.33	76.6
YOLOv3	DPatch [18]	1.92 (6)	43.89	52.7	94.3
	AdvPatch [29]	46.90 (6)	8.17	13.23	75.0
	RP ₂ -Sticker [6]	48.07 (6)	8.42	16.10	78.0
	UPC-7P [13]	63.06 (6)	5.74	19.83	85.3
	2-Rects-H	52.13 (6)	7.51	16.61	78.6
	4-Rects	55.29 (6)	6.96	20.05	83.2
	LDAP	63.75 (6)	5.61	8.98	62.5

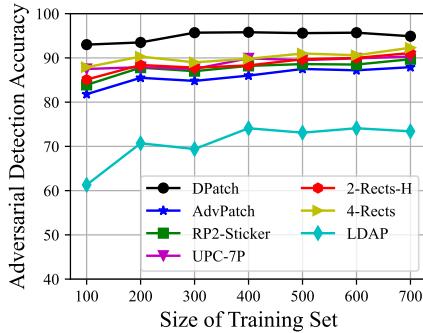


Figure 8: The relationship of training set size and adversarial detection accuracy for different attack methods on classification attack on Faster R-CNN.

of rectangle primitives is set to 10 in LDAP, the number of patches in the adversarial example of LDAP is much less than 10, indicating that the heights and widths of some rectangle primitives are decreased to zero during optimization; and (iv) searched patches of LDAP do not connect to each other to form more complex geometric shapes, suggesting that more dispersed patches, not more complex shapes, may help reduce patch area.

Table 1 illustrates the attack performance and detectability

of LDAP and comparison methods, and provides the following insights: (i) for all four detection models, LDAP gets the highest ASR under the same region area threshold, the smallest patches, the lowest LPIPS, and the lowest adversarial detection accuracy, outperforming other attack methods by a large margin; note that the lowest value of adversarial detection accuracy is 50%, meaning random guessing, reducing adversarial detection accuracy from 78.4% (AdvPatch [29]) to 76.6%, although still a big improvement when we attack SSD; (ii) for all attack methods, two-stage detectors need more patch area than one-stage detectors, which indicates they are more difficult to attack (this phenomenon is also found on other attack tasks); (iii) one interesting finding is that, among all comparison methods, a greater number of patches usually results in less area of patches, but for SSD this conclusion does not hold. This phenomenon is also found on other attack tasks. We conjecture that this is because SSD is more sensitive to the middle region during attack, which is exactly where AdvPatch [29] attacks; (iv) another finding is that, although UPC-7P [13] takes the smallest patch area among the comparison methods on detectors except SSD, it still gets high LPIPS and adversarial detection accuracy, which means that a small region area does not sufficiently represent low detectability; and (v) DPatch [18] takes the largest patch area on all models. We attribute this phenomenon to the upper-left corner patch position of DPatch. If the patch does not increase its area to reach the target object, the receptive field of the target object in detectors cannot be influenced by the patch, leading to a failed attack. Considering that the target object is usually not in the upper-left corner of the image, the patch must be large enough to reach the target object, resulting in a large patch. Due to the large patch, the ADA of DPatch is also the highest.

In addition to the comparison of adversarial detection accuracy with the fixed size of the training set, we also test the relationship of adversarial detection accuracy and the size of the training set. The comparisons are shown in Fig. 8. We find that the curve of LDAP is much lower than those of comparison methods. This indicates that the detectability of our attack is lower than that of other methods at various sizes of the training set.

4.2.2 Localization Attack

Localization attack means the adversary should mislead the detector to predict a bounding box with undesirable shape or position on the target object. In localization attack, we study the horizontal bounding-box shift attack. The attack success criterion is set as the maximum IoU of the predicted bounding boxes within the correct category is lower than 0.5.

Demonstrations of localization attack on Mask R-

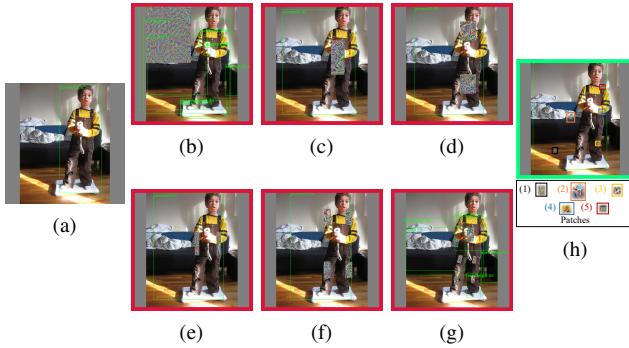


Figure 9: Adversarial examples of localization attack on Mask R-CNN [11] on one image from COCO. The detection results are also shown in images. Some attacks hide the person, while others shift the bounding boxes. The maximum IoU between the predicted bounding boxes and the ground-truth bounding box and probabilities predicted by corresponding adversarial patch detectors are: (a) original image, person=100%, IoU=95.05%; (b) DPatch [18], no person detected, P_a =96.38%; (c) AdvPatch [29], IoU = 47.03%, P_a =90.41%; (d) RP_2 -Sticker [6], IoU = 46.65%, P_a =89.45%; (e) 2-Rects-H, IoU = 46.65%, P_a =80.65%; (f) 4-Rects, IoU = 49.98%, P_a =85.92%; (g) UPC-7P [13], no person detected, P_a =77.36%; and (h) LDAp, IoU = 44.20%, P_a =38.29%.

Table 2: Performance of attack methods on localization attack on COCO dataset.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
FRCN	DPatch [18]	0.20 (14)	122.60	127.25	95.5
	AdvPatch [29]	9.90 (14)	26.14	35.88	88.5
	RP_2 -Sticker [6]	21.11 (14)	21.67	35.49	90.2
	UPC-7P [13]	50.68 (14)	14.22	33.83	90.9
	2-Rects-H	10.59 (14)	24.81	42.65	89.4
	4-Rects	15.61 (14)	23.53	46.79	92.5
	LDAp	64.16 (14)	13.85	23.48	74.5
MRCN	DPatch [18]	0.13 (15)	123.40	128.95	96.8
	AdvPatch [29]	11.96 (15)	26.87	36.59	87.9
	RP_2 -Sticker [6]	24.34 (15)	21.83	35.82	90.8
	UPC-7P [13]	54.81 (15)	14.55	34.19	90.6
	2-Rects-H	11.82 (15)	25.70	43.80	91.0
	4-Rects	14.16 (15)	24.80	48.55	92.4
	LDAp	64.37 (15)	14.11	25.87	71.7
SSD	DPatch [18]	0.89 (9)	53.57	62.02	97.9
	AdvPatch [29]	49.31 (9)	9.85	16.42	80.7
	RP_2 -Sticker [6]	39.20 (9)	11.10	21.74	88.6
	UPC-7P [13]	46.14 (9)	9.70	28.9	93.1
	2-Rects-H	35.55 (9)	11.84	24.83	89.9
	4-Rects	27.44 (9)	12.41	31.38	94.0
	LDAp	59.07 (9)	8.83	12.74	75.2
YOLOv3	DPatch [18]	0.96 (10)	72.74	78.83	94.3
	AdvPatch [29]	26.96 (9)	13.74	20.37	82.0
	RP_2 -Sticker [6]	30.33 (9)	12.99	23.14	84.9
	UPC-7P [13]	44.42 (9)	9.90	27.43	90.1
	2-Rects-H	26.89 (9)	13.20	25.56	84.1
	4-Rects	24.00 (9)	13.28	30.76	89.2
	LDAp	63.61 (9)	8.31	11.94	68.8

CNN [11] are shown in Fig 9. We find that the successful attack can be achieved not only by moving the bounding box, but also by vanishing the bounding boxes within the correct category, such as DPatch and UPC-7P. Another finding is that, for the horizontal bounding-box shift attack, horizontally-distributed patches such as 2-Rects-H have natural advantages that enable their patches to be smaller. Our LDAp also learns to arrange the patches horizontally. As for detectability, LDAp is the only one that escapes the detection of the adversarial patch detector.

Detailed attack results are shown in Table 2. We find that (i) LDAp still achieves the best performance on all detectors; and (ii) compared with the classification attack task, the localization attack task requires more patch area.

4.3. Experiments on D²-City

We investigate the threat of LDAp in an autonomous driving scenario on the D²-City dataset. Considering the poor performance of DPatch [18], we do not include it in the comparison on D²-City. Besides, UPC-7P is specifically designed for the attack “person” category, so we do not make comparisons with it, either. In D²-City, two different threatening attack cases are studied, including *vanishing a car ahead* and *shifting a car to the wrong lane*.

4.3.1 Vanishing a car ahead

This attack is dangerous, since it may directly lead the autonomous car to hit the car in front of it. The attack is essentially an application of a classification attack. The attack success criterion is set to be the same as the criterion of a classification attack on COCO, which is that there is no bounding box with a correct category that overlaps the bounding box of the target object.

The demonstration of this attack task is shown in Fig. 10. A first glance shows that the patches of LDAp are nearly imperceptible compared with those of other attack methods. Besides, the adversarial example of LDAp is the only one that passes the adversarial patch detector.

Comparisons are shown in Table 3. The findings are similar to experiments in COCO, and our LDAp still gets the best performance on all object detectors. One interesting finding is that the attack on cars in the D²-City dataset requires less patch area than the attack on persons in the COCO dataset. We attribute this to the low image quality of driving-record video, which makes cars in D²-City much easier to attack than the persons in COCO, resulting in less MAR.

4.3.2 Shifting a car to the wrong lane

This attack may influence the lane-change and speed-control decisions of an autonomous car, causing a collision. Clearly, it is an important attack case to study. This attack is essentially an application of the bounding-box shift attack. We set

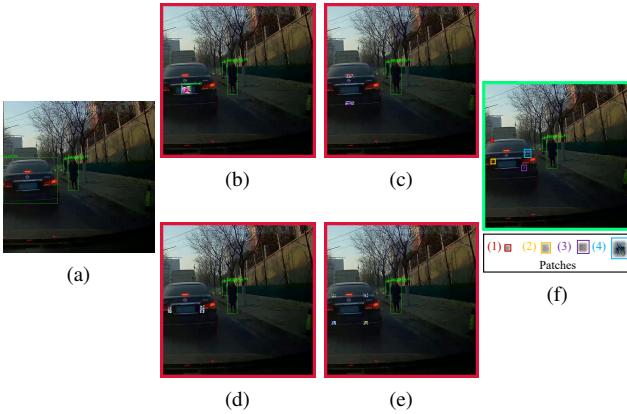


Figure 10: Adversarial examples of *vanishing a car ahead* attack task on SSD [16] on one image from D²-City. The detection results and probabilities predicted by corresponding adversarial patch detectors are: (a) original image, car = 99%; (b) AdvPatch [29], person = 49%, $P_a=84.19\%$; (c) RP₂-Sticker [6], nothing detected, $P_a=77.79\%$; (d) 2-Rects-H, nothing detected, $P_a=74.43\%$; (e) 4-Rects, nothing detected, $P_a=64.70\%$; and (f) LDAp, nothing detected, $P_a=36.66\%$.

the attack success criterion to be the same as the criterion for a localization attack on COCO, which is that the max IoU between bounding boxes with the correct category and target bounding box is lower than 0.5.

Examples of this attack are shown in Fig. 11. We see that our LDAp still utilizes the smallest patch area and still passes the detection of the adversarial patch detector. Furthermore, these images demonstrate different attack success situations in localization attack, such as by reducing the size of bounding boxes or by moving the boundary of bounding boxes to one side. Comparison results are shown in Table 4. We find that the LDAp still gets the best performance among all detectors.

4.4. Ablation Study

We perform an ablation study on the rectangle primitive number N , soft-attack strategy, mask layer search and texture constraint.

4.4.1 Number of rectangle primitives N

The ablation study of N is done on Faster R-CNN on a classification attack task in COCO and on *vanishing a car ahead* attack task in the D²-City dataset. We investigate the attack performance and the time consumption when we set N to 5, 10 and 15. The results on two datasets are shown in Table 5 and Table 6, respectively. The time consumption is tested on a Nvidia TITAN Xp graphics card. We find that the performance improvement from increasing N is small when

Table 3: Performance of attacks on *vanishing a car ahead* attack task. ADA for adversarial detection accuracy.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
FRCN	AdvPatch [29]	28.74 (5)	17.42	66.68	90.6
	RP ₂ -Sticker [6]	41.05 (5)	15.23	67.08	92.5
	2-Rects-H	44.98 (5)	12.26	67.02	92.3
	4-Rects	56.93 (5)	8.37	64.40	91.6
	LDAp	63.50 (5)	4.67	49.81	79.6
MRCN	AdvPatch [29]	34.21 (5)	16.85	65.69	90.6
	RP ₂ -Sticker [6]	41.42 (5)	14.93	65.35	90.1
	2-Rects-H	45.52 (5)	12.42	67.00	92.1
	4-Rects	57.93 (5)	8.54	64.49	91.1
	LDAp	62.22 (5)	4.68	49.83	79.0
SSD	AdvPatch [29]	60.40 (4)	6.15	47.73	82.0
	RP ₂ -Sticker [6]	60.58 (4)	6.08	50.53	87.3
	2-Rects-H	60.40 (4)	5.93	56.17	87.2
	4-Rects	66.78 (4)	4.75	58.13	91.5
	LDAp	69.61 (4)	3.88	5.94	71.0
YOLOv3	AdvPatch [29]	38.86 (3)	8.97	48.47	84.8
	RP ₂ -Sticker [6]	46.53 (3)	7.81	49.78	87.0
	2-Rects-H	44.43 (3)	7.09	57.38	87.6
	4-Rects	54.56 (3)	5.57	58.73	90.7
	LDAp	63.50 (3)	2.81	6.70	65.1

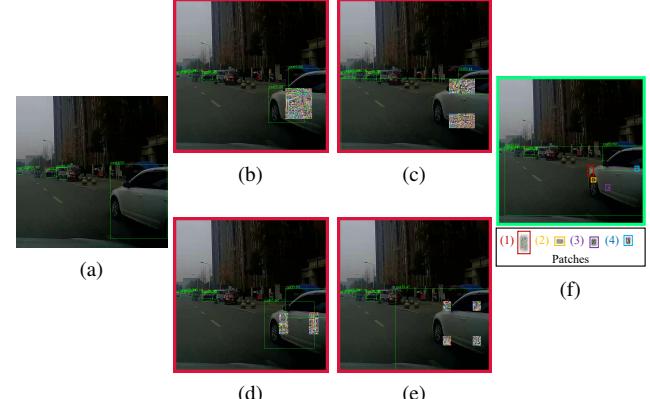


Figure 11: Adversarial examples of *shift a car to the wrong lane* attack task on YOLOv3 [25] on one image from D²-City. The maximum IoU and the probabilities predicted by the adversarial patch detectors are: (a) original image, car = 100%, IoU = 88.19%; (b) AdvPatch [29], IoU = 29.61%, $P_a=90.78\%$; (c) RP₂-Sticker [6], IoU = 30.25%, $P_a=91.91\%$; (d); 2-Rects-H, IoU = 48.35%, $P_a=79.54\%$; (e) 4-Rects, IoU = 49.67%, $P_a=70.86\%$; and (f) LDAp, IoU = 38.53%, $P_a=29.06\%$.

N is large. However, as N increases, it takes more time to generate an adversarial example. A similar phenomenon also can be found on other object detectors. Considering the balance of efficiency and performance, we set our default N as 10 in all experiments.

4.4.2 Soft-Attack Strategy

We also evaluate the effectiveness of the soft-attack strategy in the loss function of our LDAp. The experiments are con-

Table 4: Performance of attacks on *shift a car to the wrong lane* attack task.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
FRCN	AdvPatch [29]	25.72 (9)	12.98	63.69	88.1
	<i>RP</i> ₂ -Sticker [6]	31.84 (9)	12.31	67.33	93.5
	2-Rects-H	27.55 (9)	13.2	71.49	94.3
	4-Rects	40.96 (9)	10.56	72.67	95.3
	LDAP	68.97 (9)	8.01	53.25	74.0
MRCN	AdvPatch [29]	20.16 (8)	13.1	63.87	89.0
	<i>RP</i> ₂ -Sticker [6]	21.25 (8)	12.97	68.45	93.3
	2-Rects-H	21.35 (8)	13.16	71.45	93.7
	4-Rects	28.64 (8)	10.64	72.75	95.3
	LDAP	58.66 (8)	7.93	57.05	74.3
SSD	AdvPatch [29]	30.47 (8)	13.87	62.82	87.8
	<i>RP</i> ₂ -Sticker [6]	27.64 (8)	15.32	69.46	94.1
	2-Rects-H	20.89 (8)	17.35	76.79	93.1
	4-Rects	25.27 (8)	15.90	80.11	95.0
	LDAP	62.68 (8)	7.91	15.07	80.0
YOLOv3	AdvPatch [29]	53.92 (12)	14.30	63.07	84.7
	<i>RP</i> ₂ -Sticker [6]	57.75 (12)	12.92	63.79	87.1
	2-Rects-H	55.83 (12)	14.19	71.51	90.1
	4-Rects	60.12 (12)	12.61	72.71	88.7
	LDAP	63.50 (12)	11.66	10.64	62.5

Table 5: Different number of rectangle primitives; comparing results on COCO.

Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)	Time (sec/image)
LDAP ($N=5$)	22.42 (10)	15.8	16.16	76.6	361.26
LDAP	63.61 (10)	9.36	12.23	73.1	369.64
LDAP ($N=15$)	68.50 (10)	9.31	11.10	72.2	415.34

Table 6: Different number of rectangle primitives; comparing results on D²-City.

Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)	Time (sec/image)
LDAP ($N=5$)	39.23 (5)	6.34	53.99	82.8	346.47
LDAP	63.50 (5)	4.67	49.81	79.6	367.49
LDAP ($N=15$)	70.80 (5)	4.21	48.97	79.0	389.52

ducted on MRCN and YOLOv3 on the classification attack task in COCO and *vanishing a car ahead* attack task in the D²-City dataset. We compare original LDAP with two other LDAP methods that adopt different attack strategies:

- **LDAP-Attack-1-bbox** (LDAP-A1) adopts the strategy that only attacks the bounding box with maximum confidence.
- **LDAP-Attack-bbox-Equally** (LDAP-AE) adopts the strategy that equally attacks the bounding boxes that can be detected as the target object.

The comparison results are shown in Table 7 and Table 8. We find that: (i) only attacking the bounding box with maximum confidence performs poorly for our LDAP, resulting in large MAR and ADA; it is because the bounding box with maximum confidence is continuously changed during the attack, which makes the optimization of primitives unstable, leading to inferior position of patches; (ii) the performance of an equally attacking strategy is close to that of our LDAP,

Table 7: Different attack strategies; comparing results on COCO.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
MRCN	LDAP-A1	9.55 (10)	19.99	21.36	81.93
	LDAP-AE	30.81 (10)	14.49	15.72	72.0
	LDAP	60.93 (10)	9.70	15.12	71.1
YOLOv3	LDAP-A1	15.40 (6)	12.99	12.62	73.80
	LDAP-AE	60.24 (6)	6.59	9.80	66.6
	LDAP	63.75 (6)	5.61	8.98	62.5

Table 8: Different attack strategies; comparing results on D²-City.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
MRCN	LDAP-A1	11.04 (5)	17.44	55.97	82.8
	LDAP-AE	59.67 (5)	4.94	51.24	79.2
	LDAP	62.22 (5)	4.68	49.83	79.0
YOLOv3	LDAP-A1	57.11 (3)	3.30	6.91	72.3
	LDAP-AE	62.13 (3)	2.97	6.85	69.2
	LDAP	63.50 (3)	2.81	6.70	65.1

but our soft-attack strategy still outperforms this one; it is because most of the bounding boxes considered in this strategy are not accurate, slightly misleading the optimization of primitives and leading to sub-optimal positions of patches.

4.4.3 Region search and texture constraint

To validate the benefits of both a region search and texture constraint, we set the following two comparison methods:

- **LDAP without region search** (LDAP w/o r): To validate the benefit of a region search, the first comparison method uses a position-fixed patch (similar to AdvPatch [29]) and adopts the same attack loss and texture loss as our LDAP, called “LDAP without region search.”
- **LDAP without texture constraint** (LDAP w/o t): To validate the benefit of a texture constraint, the second comparison method shares the same searched region of LDAP, and optimizes texture without any constraints, called “LDAP without texture constraint.”

We compare these two methods with LDAP on a classification attack task in COCO and a *vanishing a car ahead* attack task in the D²-City dataset. We show a demonstration of these attacks in COCO in Fig. 12. It is obvious that the texture of LDAP is more consistent with the original image and that the patches are smaller. The comparison results of these methods are shown in Table 9 and Table 10. We find that both the region search and the texture constraint are important for LDAP to achieve a low ADA.

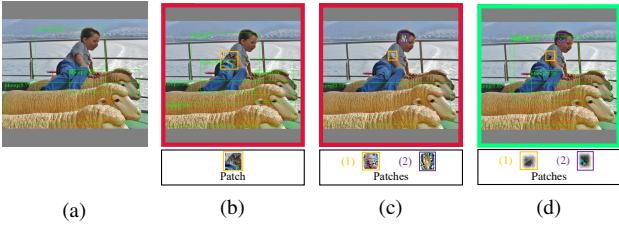


Figure 12: A comparison of LDAP without region search, LDAP without texture constraint and LDAP on a classification attack on Faster R-CNN on one image from COCO. The predictions and the probabilities predicted by adversarial patch detectors are: (a) original image, person=97%; (b) LDAP w/o r, dog = 93%, P_a =76.42%; (c) LDAP w/o t, nothing detected, P_a =83.13%; and (d) LDAP, dog = 41%, P_a =44.73%.

Table 9: Ablation study on region search and texture constraint on COCO.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
FRCN	LDAP w/o r	10.52 (10)	17.70	25.42	81.8
	LDAP w/o t	63.61 (10)	9.36	68.33	85.3
	LDAP	63.61 (10)	9.36	12.23	73.4
SSD	LDAP w/o r	52.95 (7)	8.28	26.64	79.3
	LDAP w/o t	63.06 (7)	6.39	20.26	88.8
	LDAP	63.06 (7)	6.39	9.33	76.6

Table 10: Ablation study on region search and texture constraint on D²-City.

Detector	Attack Method	ASR (%) (Area Thr (%))	MAR (%)	LPIPS ($\times 10^{-3}$)	ADA (%)
FRCN	LDAP w/o r	28.92 (5)	17.92	53.74	88.2
	LDAP w/o t	63.50 (5)	4.67	62.03	93.1
	LDAP	63.50 (5)	4.67	49.81	79.6
SSD	LDAP w/o r	55.01 (4)	7.45	15.02	73.2
	LDAP w/o t	69.61 (4)	3.88	46.51	86.6
	LDAP	69.61 (4)	3.88	5.94	71.0

5. Conclusion

In this paper, we propose a low-detectable adversarial patch attack method. It aims to reduce the area of patches while simultaneously keeping the texture consistency with the original image when attacking object detectors. That makes our adversarial examples less likely to be recognized by an adversarial patch detector. We adopt several geometric primitives to model the patches. To enhance attack performance, we adopt a soft-attack strategy. It assigns different weights to bounding boxes that can be detected as a target object. We prove the effectiveness, low detectability and the threats of our method in both COCO and D²-City datasets.

References

- [1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.
- [2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, pages 284–293. PMLR, 2018.
- [3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
- [4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [5] Richard Yi Da Xu and Michael Kemp. Fitting multiple connected ellipses to an image silhouette hierarchically. *IEEE transactions on image processing*, 19(7):1673–1682, 2010.
- [6] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. In *Proceedings of the 12th USENIX Conference on Offensive Technologies*, pages 1–1, 2018.
- [7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.
- [8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [13] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 720–729, 2020.
- [14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- [17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016.
- [18] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Yiran Chen, and Hai Li. Dpatch: An adversarial patch attack on object detectors. In *SafeAI@ AAAI*, 2019.
- [19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.
- [21] Omid Mohamad Nezami, Akshay Chaturvedi, Mark Dras, and Utpal Garain. Pick-object-attack: Type-specific adversarial attack for object detection. *Computer Vision and Image Understanding*, page 103257, 2021.
- [22] Costas Panagiotakis and Antonis Argyros. Parameter-free modelling of 2d shapes with ellipses. *Pattern Recognition*, 53:259–275, 2016.
- [23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [24] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.
- [25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.
- [27] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, 2014.
- [28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [29] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [30] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.
- [31] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.
- [32] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017.
- [33] 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 *European Conference on Computer Vision*, pages 665–681. Springer, 2020.
- [34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.