Defending Physical Adversarial Attack on Object Detection via Adversarial Patch-Feature Energy

Taeheon Kim
eetaekim@kaist.ac.kr
Korea Advance Institute of Science & Technology
Daejeon, South Korea

Youngjoon Yu
greatday@kaist.ac.kr
Korea Advance Institute of Science & Technology
Daejeon, South Korea

Yong Man Ro*
ymro@kaist.ac.kr
Korea Advance Institute of Science & Technology
Daejeon, South Korea

ABSTRACT

Object detection plays an important role in security-critical systems such as autonomous vehicles but has shown to be vulnerable to adversarial patch attacks. Existing defense methods are restricted to localized noise patches by removing noisy regions in the input image. However, adversarial patches have developed into natural-looking patterns which evade existing defenses. To address this issue, we propose a defense method based on a novel concept "Adversarial Patch-Feature Energy" (APE) which exploits common deep feature characteristics of an adversarial patch. Our proposed defense consists of APE-masking and APE-refinement which can be employed to defend against any adversarial patch on literature. Extensive experiments demonstrate that APE-based defense achieves impressive robustness against adversarial patches both in the digital space and the physical world.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; • **Security and privacy** → **Software and application security**.

KEYWORDS

Object Detection; Physical Attack; Adversarial Patch Defense; Adversarial Patch-Feature Energy;

ACM Reference Format:
Taeheon Kim, Youngjoon Yu, and Yong Man Ro. 2022. Defending Physical Adversarial Attack on Object Detection via Adversarial Patch-Feature Energy. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548362

1 INTRODUCTION

Object detectors play an important role in computer vision tasks as they have many real-world applications such as autonomous driving [2, 3, 6, 38], and pedestrian detection [12, 14, 15, 27, 45]. However, present DNN-based object detectors are shown to be vulnerable to adversarial patch attacks that cause object detectors to fail prediction [1, 13, 16, 17, 19, 20, 23, 33, 36, 37, 46]. Such attacks can be produced in the physical world by attaching adversarial patches to real-world objects [33, 36]. Physical adversarial attacks can potentially be abused, and cause serious consequences for real-world ML systems. For instance, an attacker (e.g., a bank robber) can maliciously craft an adversarial patch to evade an autonomous surveillance system. Despite such threats, reliable defenses for object detectors are severely understudied while numerous physical attacks are proposed recently. The current mainstream of emerging physical attacks adequately accounts the semantics of objects and surroundings for enhanced physical representations. Adversarial patches not only developed as t-shirts [43] and cloaks [24, 41], but to have natural-looking [10, 35] such that they evade both human eyes and detection models. These advanced patches are more likely to be realized in the physical world and extremely challenging to defend against, which poses a notable threat to object detectors. Apart from the current mainstream of physical adversarial patches on detectors, existing defenses [4, 9, 22, 26, 32] assume that incoming adversarial patches are in the form of localized noise attacks. Localized noise attacks [13, 23] are exploited in early studies in the field of adversarial patches. These patches often consists of high-frequency noises, with spurious and attention-grabbing appearances. An example of a localized noise attack is illustrated in Fig 1 (a). Previous defenses are based on mitigating the effectiveness of noise-like regions of an input image, which implementations are divided into two main categories: image analysis tools [9, 26] and noise segmentation [4, 22]. The former [9, 26] defends the detector by suppressing high-frequency components with smoothed image gradients whereas the latter [4, 22] trains an external segment network to remove noise-like regions.

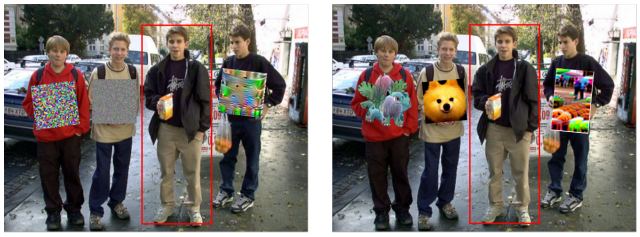

Figure 1: (a) Example of localized noise attacks. Existing defenses handle localized noise attacks by removing noisy regions from an image. (b) Example of picture-like and natural-looking patches developed recently. We propose a defense method applicable for any type of adversarial patch.

*Corresponding Author

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548362

However, the effectiveness of the aforementioned defense methods is questionable against natural-looking adversarial patches(as in Fig 1(b)). These defenses [4, 9, 22, 26] rely on removing noise components of the input image and do not provide any mechanism to notify the presence of an adversarial input or mask the pixels that realize the attack. Therefore, existing defenses against adversarial patches have the following concern: *Are these methods capable to defend natural-looking physical attacks?* Unfortunately, existing defenses based on noise removals are shown to be vulnerable to natural-looking physical attacks as we show in Section 5. Therefore, a new defense mechanism applicable for defending natural-looking patches is crucially needed to secure detectors in the real world.

In this paper, we propose a defense method to handle the aforementioned problems by analyzing how an adversarial patch alters deep features. It is known that adversarial patches on detectors minimize objectness score outputs regardless of its visual appearance [10, 35, 36, 41, 43]. We start from the following motivation: *Is there a common deep feature characteristic of an adversarial patch that leads to low objectness scores?* We address this question by analyzing that feature energy, which we define as the square of the channel-wise L1 norm, is abnormally large along the spatial location of the patch. We define this excessive feature energy caused by an adversarial patch feature as Adversarial Patch-Feature Energy (APE). Based on this concept, we propose Adversarial Patch-Feature Energy Masking (APE-Masking), which runs a layer-wise Over-Energy analysis (presented in Section 4.2) to extract adversarial region proposals. Then we propose APE-refinement which mitigates the effectiveness of adversarial patches by refining feature pixels within the APE-mask.

Extensive experiment results show that APE-refinement within the detected patches can significantly improve the adversarial robustness of a pre-trained object detector (e.g., YOLO-v2) without training adversarial samples or adding external networks.We suggest that our Adversarial Patch-Feature Energy (APE) based defense, which exploits the intrinsic property of an adversarial patch instead of its visual characteristics, can be employed to protect detectors against any adversarial patch (patch-agnostic) on literature – including localized noise attacks and natural-looking patches. The following summarizes our contributions:

- We introduce a novel concept - "Adversarial Patch-Feature Energy" which exploits in deep features changes caused by an adversarial patch. Based on this concept, we propose APE-masking to detect adversarial patches and APE-refinement to perform defense.
- Without any prior knowledge of the attacks, we perform patch-agnostic defense against a wide range of adversarial patch attacks.
- Extensive experimental results evaluated on a large set of patch attacks show that APE-refinement can significantly improve the adversarial robustness of a pre-trained object detector both in the digital space and the physical world.

2 RELATED WORK

A wide range of patch attacks have been proposed recently and researchers have been actively seeking defense methods against it. In this section, we consider adversarial patches that are physically realizable and designed for detectors. Next, we review defense methods against adversarial patch attacks on detectors.

2.1 Adversarial Patch Attacks

Adversarial attacks are widely studied methods capable of easily fooling the outcomes of DNN based models by adding input perturbations [8, 34]. Early works include localized noise attacks [13, 23]. Thys *et al.* [36] proposed a printable adversarial patch by adopting real-world transformations and regularizing a non-printability loss term in the optimizing process. [19] exploits untargeted PGD with expectation over transformation to optimize the location-invariant adversarial patch. Recent mainstream works on adversarial patches consider advanced applications such as wearables and aim for natural-looking appearance applicable to realistic situations. [11, 41, 43] designed wearable adversaries such as t-shirts and cloaks. [10] introduced techniques to make adversarial patches realistic and natural-looking by learning the image manifold of generative adversarial networks (GANs) pretrained on real-world images. [35] proposed indicators for evaluating the rationality of physical adversarial patches and developed so-called legitimate adversarial patches that evade both human eyes and detection models.

2.2 Defenses Against Patch Attacks

Most existing defenses on adversarial attacks are restricted to digital attacks and classification tasks [7, 9, 18, 25, 26, 28, 29, 40]. Unfortunately, these methods are not transferable to defending patch attacks on detectors due to the task difference. A robust classification model provides robust single-label class prediction scores whereas securing detectors requires preserving robust bounding boxes and class predictions of multiple objects. Such complexity of the task makes securing detectors much more challenging, and defense against new physical attacks is yet an unsolved/open problem. While few defense methods on object detectors have been studied [4, 26, 32, 42, 44], these works have strong limitations to be used in practice. ROC [32] is restricted to defense against adversarial patches on the top corner of an image, by regularizing detectors to have small receptive fields. DetectorGuard [42] proves the robustness of detectors on certified objects, it is limited to flagging the attack on a single patch threat model. Chiang *et al.* [4] proposed Adversarial Pixels Masking (APM), a defense by masking "patch-like" areas in the input image through an external segmenter. LGS [26] locate and remove abnormally high-frequency "patch-like" areas in input images, by smoothing image gradients. Note that LGS is originally proposed for classification models, since it is model-agnostic, it can work with object detectors. However, no "patch-like" areas exist in natural-looking adversarial patches, therefore as we will show in section 5, these methods fail to defend against these adversarial patches.

3 PROBLEM DEFINITION

3.1 Attack Formulation

In this paper, we consider adversarial patch attacks on object detectors. An adversarial patch $p_{adv} \in \mathbb{R}^{H' \times W' \times C'}$ is optimized to minimize the objectness scores $\mathcal{D}_{obj}$ of a detector, and a regularization term $\mathcal{L}_{patch}$. $\mathcal{L}_{patch}$ usually contains loss terms that enhance

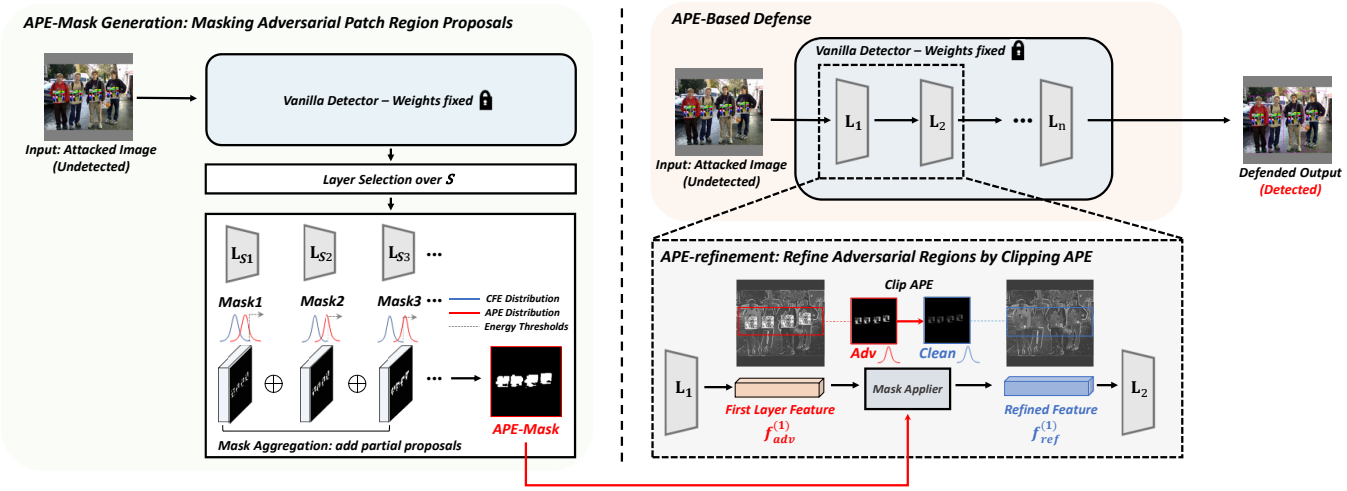

Figure 2: Overview of the proposed defense. Attacked image (or clean image) is fed to the detector and features are extracted over the set of selected layers $\mathcal{S}$. Partial proposals of APE regions(red line) are generated by leveraging the Energy-Thresholds(dotted line) on the extracted features. Mask aggregation is applied to add up the partial proposals of different dimensions and produces the APE-mask(adversarial patch region proposal). Then, APE-refinement is done at the first layer feature by clipping the APE within the APE-mask. Finally, the refined feature of the first layer is fed to the detector for robust detection.

the physical representation of the patches such as non-printability scores, naturalness scores and deformation on non-rigid objects. Formally, adversarial samples $\hat{x} \in \mathbb{R}^{H\times W\times C}$ are generated by attaching adversarial patches p_{adv} to the clean image $x \in \mathbb{R}^{H\times W\times C}$ specified a binary mask $M \in [0,1]^{H\times W\times C}$ which indicates the patch locations. We take the maximum objectness score $max(\mathcal{D}_{obj})$ as the objectness loss to accelerate the patch generation process. Patch is applied to the image x according to mask M and patch apply function $A(p_{adv}, t, x) \in \mathbb{R}^{H\times W\times C}$ which considers random transformations $t \in T$ such as scale, rotation, illumination and noise. Optimization procedure of an adversarial patch can be stated as the following:

$$\hat{x} = (1 - M) \odot x + M \odot A(p_{adv}, t, x)) \tag{1}$$

$$\min \mathbb{E}_{x\sim X, t\sim T}\left[\max\left(\mathcal{D}_{obj}(\hat{x})\right) + \lambda \mathcal{L}_{patch}\right] \tag{2}$$

3.2 Defense Formulation

We describe our defense formulation in terms of APE-masking and APE-refinement. The first step of our defense is to generate an adversarial patch region proposal, expressed as APE-mask $\hat{M}_{APE} \in [0,1]^{H\times W\times C}$. By leveraging the feature characteristics of an adversarial patch, our goal is to produce an APE-mask that consists of binary values that indicate the presence of an adversarial patch. We call this process APE-masking. Second, APE-refinement is applied to reform the first layer feature of an adversarial sample $f^{(1)}_{adv} \in \mathbb{R}^{H\times W\times C}$ by clipping APE-feature $f^{(1)}_{APE} \in \mathbb{R}^{H\times W\times C}$ within APE-mask. We found that not all parts of the adversarial patch contribute equally to the attack and clipping feature pixels with APE above certain threshold effectively improved the adversarial robustness of detectors. Regions to be refined are processed through a Mask Applier, guided by a binary mask $\hat{M}'_{APE} \in \mathbb{R}^{H\times W\times C}$ which we discuss in the following sections. Finally, the refined feature $f^{(1)}_{ref}$ is fed to the detector for complete defense. Fig. 2 shows the overview of our defense.

$$f^{(1)}_{APE} = \hat{M}'_{APE} \odot f^{(1)}_{adv} \tag{3}$$

$$f^{(1)}_{ref} = (1 - \hat{M}'_{APE}) \odot f^{(1)}_{adv} + \hat{M}'_{\mathbf{APE}} \odot clip(f^{(1)}_{APE}) \tag{4}$$

4 METHOD

In this section, we first provide a definition of the Adversarial Patch-Feature Energy (APE). Then we elaborate on our proposed defense of exploiting Adversarial Patch-Feature Energy (APE) in terms of APE-masking and APE-refinement.

4.1 How an Adversarial Patch Alters Deep Features

Without loss of generality, adversarial patch on a detector is optimized to minimize the objectness score outputs. We suspect that adversarial patches alter deep features in certain direction toward this common objective, regardless to the visual appearance in the image space. To address our hypothesis, we exploit a property of Error Backpropagation Theory [39] to analyze the feature change derived from the objectness loss during the generation process of an adversarial patch. We derived that, norm square of the last layer feature, $\left\|f^{(L)}\right\|^2 \in \mathbb{R}^{H^{(L)}\times W^{(L)}\times C^{(L)}}$ increases as the adversarial patch loss is optimized to minimize the objectness loss. This relation was derived under the assumption widely adopted in detectors, which objectness scores are directly produced from the last layer feature. Due to the space limitation, we show the derivations in the supplementary material. We define the norm square of a feature of layer k, denoted $\left\|f^{(k)}\right\|^2 \in \mathbb{R}^{H^{(k)}\times W^{(k)}\times C^{(k)}}$, as the feature energy.

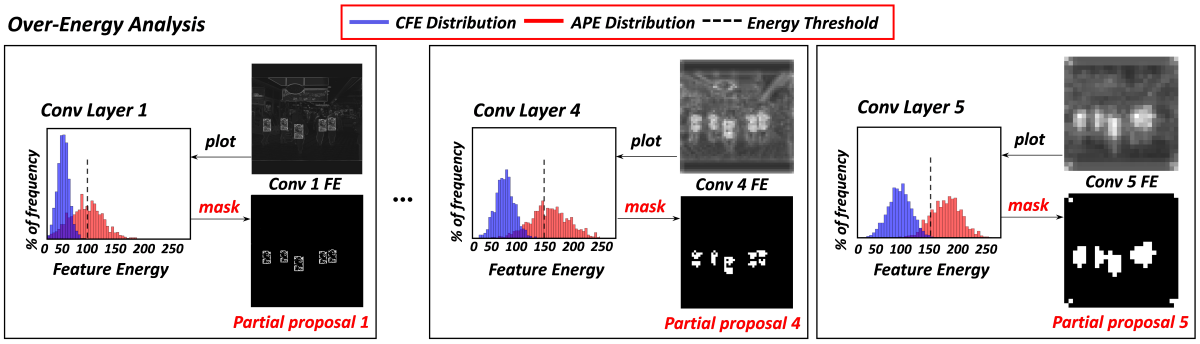

Figure 3: Illustration of the layer-wise Over-Energy analysis and Energy-Thresholds to extract partial proposals of adversarial patches. Feature Energy distribution of the clean regions(blue bar) and adversarial regions(red bar) over convolution layers 1,4 and 5 is plotted above. The dotted line indicates the Energy-Threshold which is determined by eq.5. For each layer, feature regions that exceed the Energy-Threshold are masked to produce the partial proposal.

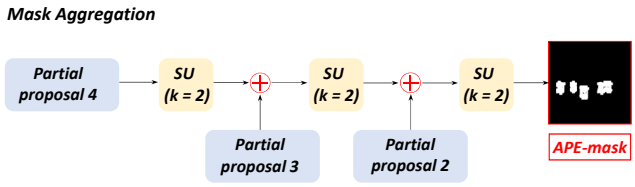

Figure 4: Example of Mask aggregation over the set of selected layers $\mathcal{S}$(In this example : $\mathcal{S} = \{2, 3, 4\}$). For every addition between adjacent partial proposals, the proposal with smaller dimension is upscaled(SU) by a factor of k.

Empirical observation reveals that feature energy is large especially at the adversarial patch locations, compared to the non-adversarial feature regions. Moreover, this feature characteristics of an adversarial patches are observed in all the internal layers of detectors. We call it Adversarial Patch-Feature Energy, and we extend our analysis to exploit this feature characteristics to accurately mask adversarial patch regions.

4.2 APE-masking by Layer-wise Over-Energy Thresholds

Layer-wise Masking: The first step of our defense is to find and mask adversarial patch regions in the input image. In other words, our goal is to produce an APE-mask $\hat{M}_{APE}$ consisting of binary values that indicate the presence of an adversarial patch. To produce such APE-mask, we analyze the layer-wise distinct characteristic of APE. As illustrated in Fig.3, APE accumulates throughout the layers, as the margin between APE and Clean Feature Energy (CFE) distributions increases. Also, spatial locations with significantly large APEs tend to change from sparse to dense. We conduct an Over-Energy analysis to leverage feature regions that exceed the Energy-Threshold as a partial proposal for each layer. To determine the Energy-Thresholds, layer-wise CFE distributions are obtained from a clean dataset X (e.g., MS COCO dataset) that does not include attacked images. Statistics on CFE and APE for layers 1,4 and 5 are shown by the blue and red bars in Fig. 3. To mask APE regions properly such that correct behavior outside the region of the mask is preserved, Energy-Thresholds should mask only pixels with large energy that are significantly far from $\mu^{(l)}_{E,clean}$ in terms of $\sigma^{(l)}_{E,clean}$. Note that $\mu^{(l)}_{E,clean}$ and $\sigma^{(l)}_{E,clean}$ are the mean and standard deviation of the CFE of layer l, respectively. By the 3.5 sigma rule, we carefully yet effectively determine the Energy-Thresholds. The Energy-Threshold of layer l is determined by eq.5. The logic to extract the l-th partial proposal of the APE-mask, denoted as $\hat{M}^{(l)}_{APE(i,j)}$, is described on eq.6:

$$\gamma^{(l)}_{TH} = \mu^{(l)}_{E,clean} + 3.5\sigma^{(l)}_{E,clean} \tag{5}$$

$$\hat{M}^{(l)}_{APE(i,j)} := \begin{cases} \mathbf{1} \text{ if } \left\| \mathbf{f}^{(l)}_{(i,j))} \right\|^2 \geq \gamma^{(l)}_{\mathbf{TH}} \\ \mathbf{0} \text{ otherwise} \end{cases} \tag{6}$$

The rationale for a 3.5 sigma threshold is that the expected fraction of CFE outside the threshold is about 0.0025%. For a YOLO-v2 detector, where the input image dimension is 416 × 416, only 41 pixels which is only 1 of 4298 of the clean feature pixels are expected to be miscounted as adversarial feature pixels by this Energy-Threshold. Therefore, such a strict threshold helps avoid masking clean objects that do not cause adversarial effects. For every layer l, the Over-Energy analysis produces a partial proposal $\hat{M}^{(l)}_{APE}$. An example of Energy-Thresholds and the extracted partial proposals of layers 1,4 and 5 is depicted in Fig 3. Partial proposals extracted from the shallower layers (adjacent to layer 1) consist of more sparse and fine pixels, whereas those from the deeper layers (adjacent to the last layer) consist of dilated areas and often contain contiguous areas as well. We take advantage of these distinct characteristics of layer-wise partial proposals to produce the final APE-mask by Mask Aggregation.

Mask aggregation: The next step is to produce the APE-mask by fusing partial adversarial patch proposals through Mask aggregation. Most CNN models adapt pooling layers to encode features with reduced dimensions. Therefore partial proposals extracted from different convolution layers have distinct spatial sizes. The partial proposal to be aggregated is resized through Spatial Unpooling(SU)

to have dimensions same as the adjacent proposal according to the upscale size factor k. Nearest interpolation is used for up-scaling. Mask aggregation is applied over only on the set of selected layers $\mathcal{S}$, which are empirically determined(in Table 1) to maximize the advantage of fusing partial proposals. An example of Mask Aggregation over conv layer 2,3,4 is shown in Fig. 4. Mask aggregation iteratively adds up the binary masks of spatially unpooled partial proposals over the set of selected layers (denoted $\mathcal{S}$ in eq.7) to produce the APE-mask.(i.e., Full adversarial patch region proposal) as:

$$\hat{M}_{APE} = \bigvee_{l \in \mathcal{S}} SU_k^{(l)}(\hat{M}_{APE}^{(l)}) \tag{7}$$

4.3 APE-refinement for complete defense

The final step for defense is to refine adversarial features within the produced APE-mask with clean features. For the best spatial accuracy, we refine the feature of the first layer, denoted as $f_{adv}^{(1)}$ which has the same spatial dimension as the input image. Also, refining the first feature consequently refine the subsequent features as well through the forward path. Therefore, we focus on refining only the first feature which is denoted as $f_{APE}^{(1)}$ in eq.3. The area to be refined($\hat{M_{APE}}$) is determined by eq.8. Then feature refinement is done by clipping feature areas within $\hat{M_{APE}}$ which energy exceeds $\mu_{E,clean}^{(l)}$ within the APE-mask to Clean Feature Energy values as in eq.9. The refined feature($f_{ref}^{(1)}$ in eq.4), in which the effectiveness of adversarial patches is mitigated, is fed to the rest of the detector for complete defense.

$$\hat{M}'_{APE(i,j)} := \begin{cases} 1 \text{ if } \left\| f_{APE(i,j))}^{(1)} \right\|^2 \geq \mu_{E,clean}^{(1)} \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

$$clip(f_{APE(i,j)}^{(1)}) = \frac{f_{APE(i,j))}^{(1)}}{\left\| f_{APE(i,j))}^{(1)} \right\|^2} \min(\left\| f_{APE(i,j)}^{(1)} \right\|^2, \mu_{E,clean}^{(1)}) \tag{9}$$

5 DEFENSE EVALUATION ON DIGITAL ATTACKS

In this section, we evaluate the robustness of our APE-based defense on digital adversarial patch attacks. For every detection over test images, APE-masking is operated on feature maps of each cascaded module(e.g., darknet for YOLO-v2). During test time, we inspect the quality of the produced APE-masks by measuring the similarities with the true adversarial patch regions. Also, we demonstrate the defense performance of APE-refinement by testing the clean accuracy and adversarial patch robustness over datasets.

5.1 Implementation Details

Target Object Detector and Datasets: For our experimental setting, we use YOLO-v2 [30] with a darknet backbone for the base object detectors. The PyTorch implementations pre-trained on the MS COCO [21] dataset are adopted. For evaluating the robustness of our method, we consider the INRIA-person [5] dataset since most existing methods for physical adversarial patches are developed on person detectors. The INRIA-person dataset consists of 614 training images and 288 testing images containing only the "person" class. Our implementation is built upon Pytorch and all experiments are conducted on a server with 2 Titan Xp GPUs.

Attack Setting: Our attack setting on digital attacks focuses on simulating physical patch attacks. We consider the general attack scenario where attackers are holding (or wearing) the adversarial patch to fool real-world detectors. To demonstrate this physical world scenario in the digital space, we attach the adversarial patch on target objects 1) with random transformation applied (scale, rotation, illumination, noise addition) 2) at non-fixed locations (not only the top corner) 3) with non-fixed shapes or sizes.

Adversarial Patches for Evaluation: Evaluation of our proposed defense is on a wide range of adversarial patch attacks [10, 13, 35, 36, 41, 43] on YOLO-v2. In practice, the defender has no knowledge of what the attacks look like. We fix our parameters of the defense model and perform defense on 9 distinct types adversarial patches categorized into 3 groups: localized attack noise, printable [36, 41, 43], and natural-looking patches [10, 35].

Defense Baselines: For the defense baselines, we compare our proposed method with vanilla (without defense) YOLO-v2, Local Gradient Smoothing (LGS) [26], Adversarial Pixel Masking (APM) [4]. For LGS [26], we set the block size to 15, overlap to 5, threshold to 0.05, and smoothing factor to 2.3. We re-implemented APM by training the U-net segmenter [31] with the same settings described by the authors [4]. For each defense method, we evaluate model robustness on the INRIA test dataset and report the Average Precision (AP) at Intersection over Union (IoU) 0.5. We evaluate three rounds with different patch sizes and report the mean of AP.

5.2 Evaluating the quality of the APE-mask

To support our APE-masking methodology explained in section 4, we evaluate the quality of the APE-mask over two metrics. We quantitatively measure the similarity of the APE-mask (b0) and the true adversarial patch regions (b1) by computing the IoU(b0,b1) (Intersection over Union) and IoA(b0,b1) (Intersection over Area). The definition of the Intersection of Area of b0 and b1 is stated below:

$$IoA(b0, b1) = \frac{area(b1 \cap b0)}{area(b1)} \tag{10}$$

A high IoA value indicates that APE-mask fully covers adversarial patch regions, while a high IoU value accounts for preserving the outside regions of the adversarial patches while localizing patch regions correctly. To evaluate the APE masks, we manually craft pixel-level ground-truth masks of adversarial patch regions. We simulate APE-masking on adversarial patches [10, 35, 36, 41, 43] over the INRIA dataset. Over different sets of selected layers we found that $\mathcal{S}_4$ is most effective for APE-masking. Also, the effectiveness of the APE-masking on adversarial patches [10, 35, 36, 41, 43] are shown through extensive experimental results in Table 1 and Fig. 5.

Table 1: Quantitative analysis to nominate the set of selected layers $\mathcal{S}$. Partial proposals are aggregated over $\mathcal{S}$ to produce the APE-mask. $\mathcal{S}_4$ is chosen as our set of selected layers, which APE-mask produced over has adequate IoU and IoA value with respect to the true patch regions. Also, extensive experiments show that $\mathcal{S}_4$ is most effective for APE-masking different types adversarial patches.

	$\mathcal{S}_1 = \{1\}$		$\mathcal{S}_2 = \{1,2\}$		$\mathcal{S}_3 = \{1,2,3\}$		$\mathcal{S}_4 = \{1,2,3,4\}$		$\mathcal{S}_5 = \{1,2,3,4,5\}$	
	IoU	IoA	IoU	IoA	IoU	IoA	IoU	IoA	IoU	IoA
Localized Noise [13]	0.37 ± 0.03	0.41 ± 0.04	0.52 ± 0.06	0.58 ± 0.09	0.51 ± 0.07	0.63 ± 0.08	0.59 ± 0.07	0.92 ± 0.13	0.34 ± 0.10	0.97 ± 0.05
OBJ [36]	0.19 ± 0.02	0.27 ± 0.05	0.46 ± 0.05	0.51 ± 0.03	0.58 ± 0.09	0.86 ± 0.08	0.60 ± 0.09	0.91 ± 0.06	0.37 ± 0.11	0.95 ± 0.15
Adv T-shirt [43]	0.41 ± 0.04	0.65 ± 0.04	0.68 ± 0.08	0.91 ± 0.05	0.62 ± 0.10	0.98 ± 0.05	0.57 ± 0.10	0.99 ± 0.03	0.39 ± 0.12	0.99 ± 0.02
Adv Cloak [41]	0.31 ± 0.02	0.34 ± 0.04	0.51 ± 0.05	0.52 ± 0.05	0.61 ± 0.07	0.84 ± 0.10	0.59 ± 0.07	0.89 ± 0.09	0.48 ± 0.08	0.95 ± 0.15
P1 [10]	0.26 ± 0.10	0.50 ± 0.13	0.30 ± 0.09	0.58 ± 0.10	0.29 ± 0.11	0.66 ± 0.12	0.51 ± 0.09	0.75 ± 0.12	0.20 ± 0.08	0.79 ± 0.15
P2 [10]	0.16 ± 0.07	0.17 ± 0.01	0.21 ± 0.08	0.55 ± 0.12	0.32 ± 0.10	0.55 ± 0.12	0.44 ± 0.11	0.69 ± 0.17	0.26 ± 0.10	0.75 ± 0.14
Shaymin [35]	0.28 ± 0.04	0.35 ± 0.03	0.33 ± 0.12	0.55 ± 0.07	0.38 ± 0.15	0.73 ± 0.07	0.49 ± 0.15	0.90 ± 0.05	0.29 ± 0.12	0.92 ± 0.05
Flower [35]	0.26 ± 0.11	0.62 ± 0.12	0.37 ± 0.15	0.65 ± 0.11	0.49 ± 0.17	0.84 ± 0.10	0.46 ± 0.16	0.95 ± 0.06	0.28 ± 0.12	0.97 ± 0.05

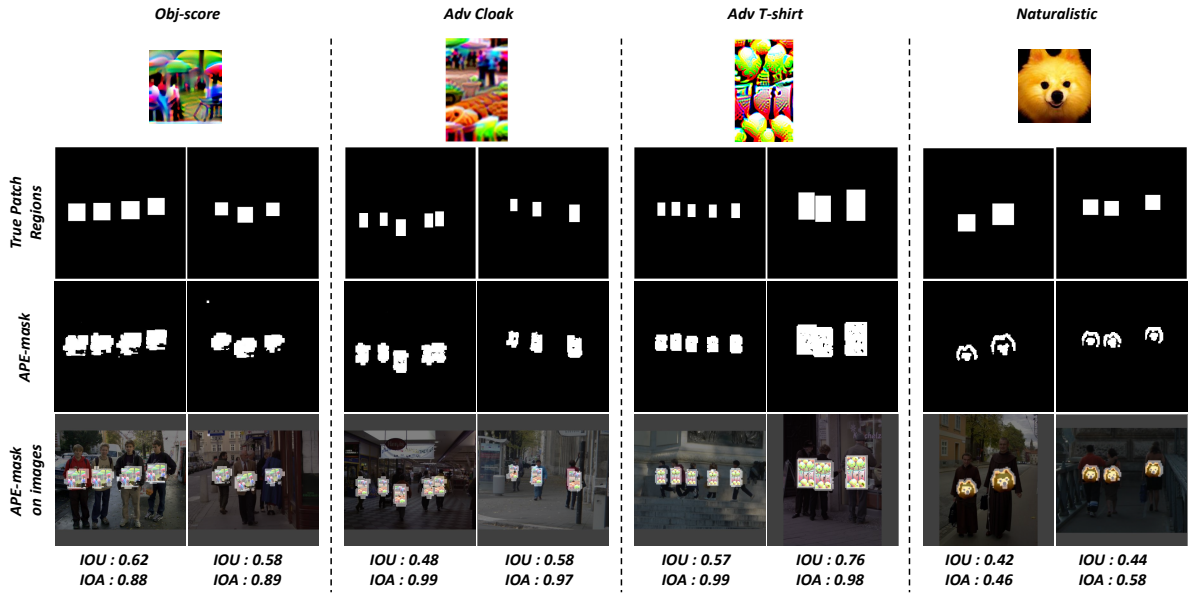

Figure 5: Visualization examples of the produced APE-mask with respect to the patched images. For 4 different adversarial patches, 2 examples of the APE-masks and true patch regions are shown with the corresponding IoU and IoA values marked at the bottom.

5.3 Evaluating the Effectiveness of APE-refinement

We evaluate the effectiveness of APE-refinement by computing the APE values over different adversarial patches, before and after APE-based defense.Table 2 shows APE values computed over different adversarial patches before and after APE-refinement. It appears that APE-refinement effectively suppresses the energy level of adversarial patch regions to the clean level.

Table 2: Evaluation of the APE-refinement. Adversarial Patch-Feature Energy(APE) and Refined Feature Energy(RFE) distributions computed over different adversarial patches.

	Clean	Obj	Adv Cloak	Adv T-shirt	Naturalistic
APE	30.8 ± 11.6	75.7 ± 34.6	74.9 ± 30.4	105.6 ± 47.5	64.4 ± 24.1
RFE	**30.8 ± 11.6**	**43.1 ± 16.2**	**43.3 ± 15.3**	**46.8 ± 21.8**	**47.9 ± 13.9**

5.4 Defense Performance

Localized Noise Attacks: To demonstrate our defense under localized noise attacks [13], we optimize eq.2 without $\mathcal{L}_{patch}$. We set $\epsilon = 1$, the strongest attack budget under an adaptive attack scenario, where the attacker can arbitrarily modify within the patch regions with full knowledge of the defense pipeline. One epoch is iterated over the INRIA train dataset using an Adam optimizer which learning rate is set to 0.05. Clean and defense results are shown in Table 3.

Printable Patch Attacks: We further evaluate our APE-based defense under 4 printable patch attacks trained on YOLO-v2: OBJ [36],

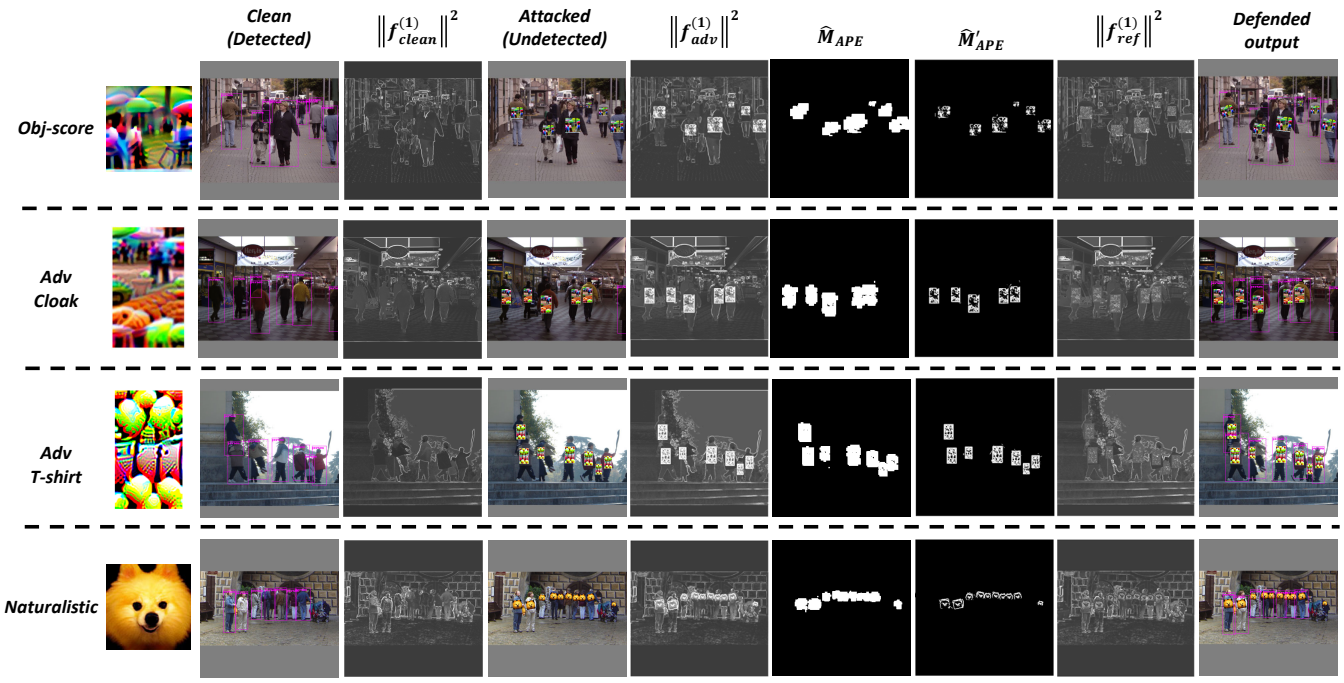

Figure 6: Visualization examples of intermediate products during our defense process. From left to right, adversarial patches, detection results of clean images, $\|f^{(1)}_{clean}\|^2$, detection results of attacked images, $\|f^{(1)}_{adv}\|^2$, $\hat{M}_{APE}$, $\hat{M}'_{APE}$, $\|f^{(1)}_{ref}\|^2$ and defended outputs are shown.

Table 3: Performance(AP) of different defenses under clean images and a Localized Noise Attack ($\epsilon = 1$) on INRIA dataset.

	YOLO-v2	LGS	APM	Ours
Clean	**93.9**	93.6	91.7	93.1
$\epsilon = 1$	5.3	80.8	**91.4**	90.5

Table 4: Performance(AP) of different defenses under printable patch attacks on INRIA dataset.

	YOLO-v2	LGS	APM	Ours
OBJ	16.8	71.2	85.3	**91.4**
OBJ-CLS	38.2	79.9	88.1	**92.2**
Adv T-shirt	20.7	61.8	81.4	**89.4**
Adv Cloak	15.4	51.4	80.6	**89.8**

OBJ-CLS [36], Adv T-shirt [43], and Adv cloak [41]. To demonstrate OBJ, we consider $\mathcal{L}_{patch}$ of eq.2 to consist of non-printablity score and total variance loss. OBJ-CLS [36] is generated by minimizing both classification and objectness scores. Adv T-shirt and Adv cloak are re-implemented with the same loss terms described by the authors of [43] and [41] respectively. Fig.6 visualizes the intermediate processes of our defense pipeline against these adversarial patches. The defense performances are shown in Table 4. APE-based defense shows impressive robustness where APM and LGS degrade on printable patch attacks. We present more visualization examples in the supplementary material.

Table 5: Performance(AP) of different defenses under natural-looking patch attacks (×1.5 size) on INRIA dataset.

	YOLO-v2	LGS	APM	Ours
P1	35.4	42.3	71.0	**87.4**
P2	61.1	65.2	73.7	**87.0**
P3	46.9	56.0	69.2	**82.7**
Ivysaur	49.7	52.4	58.6	**83.3**
Shaymin	61.9	63.5	71.4	**84.3**
Flower	50.8	51.5	64.1	**83.6**

Natural-Looking Patch Attacks: We compare the effectiveness of our APE-based defense with other defense baselines under 6 natural-looking patch attacks, 3 of each proposed by naturalistic patch [10] and legitimate patch [35]. We re-implemented patches P1, P2 and P3 indicated in naturalistic patch [10], Cartoon Ivysaur, Cartoon Shaymin, and Cartoon Flower following the implementation details in [35]. Since the original paper of naturalistic adversarial patches [10] evaluates patches with larger spatial size for sufficient attack rates, 1.5 times(1.25×1.25) larger patches are used for our evaluations than in printable patches and localized noises. Except for P2 and P3 (trained on Yolo-v3 and Yolo-v3 tiny respectively), all patches are trained on YOLO-v2. For sufficient attack rates of natural-looking patches, As shown in Table 5, defense methods based on noise rejection(LGS and APM) are vulnerable to natural-looking patches as they do not realize the attacks. Our

Table 6: Defensive performances(AP) under patch attacks on INRIA dataset with YOLO-v1

	YOLO-v1	LGS	APM	Ours
OBJ	9.0	37.4	82.1	**84.4**
Adv T-shirt	22.1	63.5	76.4	**81.9**
Adv Cloak	16.6	68.7	76.9	**82.6**
P1	35.3	42.3	57.2	**82.2**
P2	65.2	68.9	68.2	**83.2**
P3	46.9	55.9	65.8	**81.1**

Table 7: Training and Inference time

	LGS	APM	Ours
Inference time	353ms	32ms	197ms
Training time	-	30min	-

APE-based defense shows remarkable robustness against natural-looking patches as well as impressive visualization results realizing the attacks(See Fig 6).

5.5 Adaptive Attacks

We further evaluate our APE-based defense against adaptive attacks where the attacker has full knowledge of the defense pipeline. In specific, an adaptive attacker can optimize an adversarial patch to have Adversarial Patch-Feature Energies that underlie the Energy-Thresholds to bypass APE-masking. To simulate such an attack, we optimize an adaptive adversarial patch p_{adap} based on [13] to have low feature energy while remaining a high attack rate. Evaluations are conducted across different attack budgets $\epsilon \sim [0, 1]$. Experiment results show that our defense is robust to adaptive attacks higher than 88.2% AP. We show the detailed results in the supplementary material. We optimized p_{adap} as the following equation:

$$\text{argmin} E_{\mathbf{x}\sim\mathbf{X},\mathbf{t}\sim\mathbf{T}}\left[\max\left(\mathcal{D}_{obj}(\hat{\mathbf{x}})\right) + \sum_{l\in S} mAPE(M \odot \hat{x})\right] \quad (11)$$

Where S denote the set of selected layers, M incidents the binary mask of the true adversarial patch regions.

5.6 Speed

Table 7 shows the training and inference time in millisecond units. Measurement was obtained from a machine with two Titan XP GPUs. Training time was measured on the INRIA dataset. APM took 30 minutes as LGS and our method does not need training since they are post-processing methods on pre-trained models. For the inference time, our method took 197ms for each process, while LGS took 353ms and APM took 32ms. Overall, our method supports video frame rate of 5 fps.

6 DEFENSE EVALUATION ON PHYSICAL ATTACKS

We conduct experiments to see whether APE-based defense is robust against physical adversarial attacks on real-world scenes. We evaluate over localized noise attacks [13] and physical adversarial patches [10, 35, 36, 41, 43]. To quantify defense performance against these 10 different physical patches, we recorded 60 videos from 6 different scenes with 2~4 actors in each video. Various distances (between 2 and 10 meters) and angles (-45 degrees 45 degrees from the center) were considered while taking the videos. Redundant frames were excluded, 2110 frames were used for evaluation. As shown in the figure 8, our APE-defense shows a superior performance for defending against adversarial patches in the physical world. For the demo video click This Link or go to the next url: https://youtu.be/-KV0tLumNQo

Figure 7: Detection results of our APE-based defense on physical attacks. Although printed adversarial patches are attached to pedestrians, our APE-based defense model successfully produces correct detection.

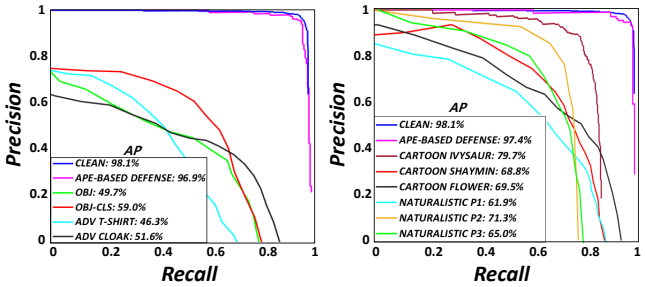

Figure 8: PR-curve of the clean results on YOLO-v2(blue), our defense results(pink) averaged over patches and attacked results on YOLO-v2(the others). Printable patches(left) and natural-looking patches(right) are evaluated in the physical world. Average precision is marked at the left bottom of each PR-curve.

7 CONCLUSION

We proposed APE-based defense against adversarial patches on object detectors. Our defense framework consists of APE-masking and APE-refinement which is applicable on general adversarial patches. Extensive experiments show that our proposed method effectively improves robustness against adversarial patches both in the digital space and the physical world. Our novel framework could be used to secure DNN-based detectors for real-world applications.

ACKNOWLEDGMENTS

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

REFERENCES

[1] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. 2018. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 52–68.

[2] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 2016. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2147–2156.

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1907–1915.

[4] Ping-Han Chiang, Chi-Shen Chan, and Shan-Hung Wu. 2021. Adversarial Pixel Masking: A Defense against Physical Attacks for Pre-trained Object Detectors. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1856–1865.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3354–3361.

[7] Thomas Gittings, Steve Schneider, and John Collomosse. 2020. Vax-a-Net: Training-time Defence Against Adversarial Patch Attacks. In *Proceedings of the Asian Conference on Computer Vision*.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[9] Jamie Hayes. 2018. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1597–1604.

[10] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. 2021. Naturalistic Physical Adversarial Patch for Object Detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7848–7857.

[11] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. 2020. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 720–729.

[12] Xin Huang, Zheng Ge, Zequn Jie, and Osamu Yoshie. 2020. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10750–10759.

[13] Danny Karmon, Daniel Zoran, and Yoav Goldberg. 2018. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*. PMLR, 2507–2515.

[14] Jung Uk Kim, Sungjune Park, and Yong Man Ro. 2021. Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3050–3059.

[15] Jung Uk Kim, Sungjune Park, and Yong Man Ro. 2022. Towards Versatile Pedestrian Detector with Multisensory-Matching and Multispectral Recalling Memory. In *36th AAAI Conference on Artificial Intelligence (AAAI 22)*. Association for the Advancement of Artificial Intelligence.

[16] Taeheon Kim, Hong Joo Lee, and Yong Man Ro. 2022. Map: Multispectral Adversarial Patch to Attack Person Detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4853–4857.

[17] Dapeng Lang, Deyun Chen, Ran Shi, and Yongjun He. 2021. Attention-Guided Digital Adversarial Patches on Visual Detection. *Security and Communication Networks* 2021 (2021).

[18] Hakmin Lee, Hong Joo Lee, Seong Tae Kim, and Yong Man Ro. 2020. Robust ensemble model training via random layer sampling against adversarial attack. *arXiv preprint arXiv:2005.10757* (2020).

[19] Mark Lee and Zico Kolter. 2019. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897* (2019).

[20] Yuezun Li, Xiao Bian, Ming-Ching Chang, and Siwei Lyu. 2018. Exploring the vulnerability of single shot module in object detectors via imperceptible background patches. *arXiv preprint arXiv:1809.05966* (2018).

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[22] Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. 2021. Segment and Complete: Defending Object Detectors against Adversarial Patch Attacks with Robust Patch Detection. *arXiv preprint arXiv:2112.04532* (2021).

[23] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. 2018. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299* (2018).

[24] Arman Maesumi, Mingkang Zhu, Yi Wang, Tianlong Chen, Zhangyang Wang, and Chandrajit Bajaj. 2021. Learning Transferable 3D Adversarial Cloaks for Deep Trained Detectors. *arXiv preprint arXiv:2104.11101* (2021).

[25] Michael McCoyd, Won Park, Steven Chen, Neil Shah, Ryan Roggenkemper, Minjune Hwang, Jason Xinyu Liu, and David Wagner. 2020. Minority reports defense: Defending against adversarial patches. In *International Conference on Applied Cryptography and Network Security*. Springer, 564–582.

[26] Muzammal Naseer, Salman Khan, and Fatih Porikli. 2019. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1300–1307.

[27] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. 2019. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4967–4975.

[28] Sungjune Park, Hong Joo Lee, and Yong Man Ro. 2021. Adversarially Robust Hyperspectral Image Classification via Random Spectral Sampling and Spectral Shape Encoding. *IEEE Access* 9 (2021), 66791–66804.

[29] Sukrut Rao, David Stutz, and Bernt Schiele. 2020. Adversarial training against location-optimized adversarial patches. In *European Conference on Computer Vision*. Springer, 429–448.

[30] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[32] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. 2020. Role of spatial context in adversarial robustness for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 784–785.

[33] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. 2018. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*.

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[35] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. 2021. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5307–5315.

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

[37] Yajie Wang, Haoran Lv, Xiaohui Kuang, Gang Zhao, Yu-an Tan, Quanxin Zhang, and Jingjing Hu. 2021. Towards a physical-world adversarial patch for blinding object detection models. *Information Sciences* 556 (2021), 459–471.

[38] Xuezhi Wen, Ling Shao, Wei Fang, and Yu Xue. 2014. Efficient feature selection and classification for vehicle detection. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 3 (2014), 508–517.

[39] Jianxin Wu. 2017. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China* 5, 23 (2017), 495.

[40] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. 2019. Defending against physically realizable attacks on image classification. *arXiv preprint arXiv:1909.09552* (2019).

[41] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. 2020. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*. Springer, 1–17.

[42] Chong Xiang and Prateek Mittal. 2021. DetectorGuard: Provably Securing Object Detectors against Localized Patch Hiding Attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 3177–3196.

[43] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision*. Springer, 665–681.

[44] Youngjoon Yu, Hong Joo Lee, Hakmin Lee, and Yong Man Ro. 2022. Defending Against Person Hiding Adversarial Patch Attack with a Universal White Frame. *arXiv preprint arXiv:2204.13004* (2022).

[45] Zhishuai Zhang, Jiyang Gao, Junhua Mao, Yukai Liu, Dragomir Anguelov, and Congcong Li. 2020. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11346–11355.

[46] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. 2019. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1989–2004.