This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Revisiting Adversarial Patch Defenses on Object Detectors: Unified Evaluation, Large-Scale Dataset, and New Insights

Junhao Zheng[1] Jiahao Sun[1] Chenhao Lin[1†] Zhengyu Zhao[1†] Chen Ma[1]
Chong Zhang[1] Cong Wang[2] Qian Wang[3] Chao Shen[1]
[1]*Xi'an Jiaotong University* [2]*City University of Hong Kong* [3]*Wuhan University*

Abstract

Developing reliable defenses against patch attacks on object detectors has attracted increasing interest. However, we identify that existing defense evaluations lack a unified and comprehensive framework, resulting in inconsistent and incomplete assessments of current methods. To address this issue, we revisit 11 representative defenses and present the first patch defense benchmark, involving 2 attack goals, 13 patch attacks, 11 object detectors, and 4 diverse metrics. This leads to the large-scale adversarial patch dataset with 94 types of patches and 94,000 images. Our comprehensive analyses reveal new insights: (1) The difficulty in defending against naturalistic patches lies in the data distribution, rather than the commonly believed high frequencies. Our new dataset with diverse patch distributions can be used to improve existing defenses by 15.09% AP@0.5. (2) The average precision of the attacked object, rather than the commonly pursued patch detection accuracy, shows high consistency with defense performance. (3) Adaptive attacks can substantially bypass existing defenses, and defenses with complex/stochastic models or universal patch properties are relatively robust. We hope that our analyses will serve as guidance on properly evaluating patch attacks/defenses and advancing their design. Code and dataset are available at https://github.com/Gandolfczjh/APDE, where we will keep integrating new attacks/defenses.

1. Introduction

Deep neural networks (DNNs) have been widely applied across various real-world domains, including pedestrian object detection [4, 33], depth estimation [9], and autonomous driving [5]. Despite their outstanding performance, DNNs are known to be vulnerable to adversarial attacks, which severely limits their reliability [53, 54]. These attacks manipulate input data, causing DNNs to produce incorrect predictions, potentially leading to severe incidents. Adversarial attacks can be realized in different forms, including adversarial perturbations [51] and adversarial patches [29, 30, 56]. Among these, adversarial patch attacks, which are feasible in the physical world, pose a significant threat to DNNs in real-world applications.

In recent years, substantial efforts have been made to defend against adversarial patches [17, 26, 27, 31, 37, 44]. For instance, LGS [31] assumes that adversarial patches exhibit more intense pixel variations than clean images, aiming to detect and then smooth regions with significant texture changes to eliminate patches. Defense methods like PAD [16] and NAPGuard [44] utilize additional segmentation [36] or detection [32, 33] models to eliminate adversarial patches. Approaches like DIFFender [17] and NutNet [26] leverage differences in data distribution between adversarial and clean images, using generative models [19, 35] to analyze and locate adversarial patches.

However, so far, there still lacks a unified framework for defense evaluation, and as a result, the effectiveness of various defenses may not be assessed consistently and comprehensively. Specifically, we identify the main limitations of existing evaluations and address them through the following four aspects: (1) **Unified framework.** Existing studies suffer from inconsistent parameters, attack methods, and patch placement strategies, which impede fair comparisons. To address this, we unify the evaluation paradigm around pedestrian detection, standardizing parameters, and patch deployment. (2) **Suitable metrics.** Some works use only patch detection accuracy to assess defenses [15, 44], thereby raising fairness concerns because high detection accuracy does not necessarily reflect superior defense performance. To address this, we adopt average precision (AP) and attack success rate (ASR) as comprehensive evaluation metrics and experimentally demonstrate that patch detection accuracy is unsuitable for objectively comparing the defense performance. (3) **Comprehensive analyses.** Existing studies frequently overlook critical factors such as real-time performance, defense against patches of varying sizes and types, and real-world applicability. To address this, we systematically analyze defense efficiency and performance

[†]Corresponding authors.

across different patch sizes, attack types, adaptive attacks, and patch mask filling strategies, while extending evaluations to the physical world. (4) **Challenging attacks.** Existing studies utilize adversarial patch datasets (e.g., GAP [44] and Apricot [3]) that lack categorization by detector type and comprehensive attack coverage, making them insufficient for evaluating defense robustness. To address this, we conduct experiments under white-box attack conditions, offering a worst-case analysis grounded in data distribution to rigorously assess defense performance.

Overall, we provide the first patch defense benchmark on evaluating patch defenses and introduce a new large-scale Adversarial Patch Defense Evaluation (APDE) dataset, containing 94 types of patches and 94,000 images. In comparison, popular datasets in existing work are far smaller: Apricot (60 types of patches, 1,011 images) [3] and GAP (25 types of patches, 9,266 images) [44]. Our main contributions can be summarized as follows:

- We identify four main limitations of existing evaluations of adversarial patch defenses, and we provide the first patch defense benchmark, involving 11 representative defense methods, 2 attack goals, 13 patch attacks, 11 object detectors, and 4 diverse metrics.
- We conduct comprehensive analyses to reveal new insights, including the root causes of defense failures against naturalistic patches, unfair comparative metrics in prior defense performance evaluations, and effective strategies to counter adaptive attacks.
- We construct the large-scale Adversarial Patch Defense Evaluation (APDE) dataset. Beyond evaluating defenses, it can also be used to substantially improve defenses, with an average improvement of 15.09% AP@0.5 on mainstream defense models.

2. Related work

2.1. Object Detection

Object detection is a core technology in the field of computer vision, aimed at identifying one or more predefined objects in images or videos and accurately locating their positions. Based on the distinction between detection box proposals and object classification, object detection is primarily divided into two categories: single-stage detectors (such as YOLO [2, 32, 33, 41, 42], SSD [28], RetinaNet [24], CenterNet [8], DDETR [4]) and two-stage detectors (such as FRCNN [34], MRCNN [11]). Single-stage detectors directly predict the class probabilities and detection boxes of objects in an image through a single network, achieving both localization and classification simultaneously. In contrast, two-stage detectors first generate a set of region proposals in the image and then classify these proposals as objects of various categories or as background, typically employing separate networks or modules for processing.

2.2. Adversarial Attacks on Object Detection

Adversarial attacks on object detection add perturbations to input images, causing detection boxes to disappear, appear incorrectly, or label changes [6, 40]. Early work focuses on digital-domain attacks, adding small, global perturbations to image pixels [49]. To make attacks more practical, later research has shifted towards adversarial patches with large, local perturbations [18, 40]. Adversarial patch attacks against object detectors place an optimized patch on an object. Adversarial patches can execute either *hiding attacks* or *appearing attacks*. These attacks can be further classified into optimization-based and generator-based methods. Specifically, hiding attacks enable targeted objects to evade detection by the detector. Optimization-based attacks [14, 40, 45, 50] begin with a noise patch and iteratively adjust pixel values to maximize detector error, often employing optimization techniques such as gradient descent to update the patch. Generator-based attacks [12, 13, 23] use generative models [19, 35] to directly produce adversarial patches, offering the advantage of creating more diverse and natural-looking patches [55]. Differently, appearing attacks [52] disguise patches to resemble normal target objects, causing the detector to misidentify them.

2.3. Adversarial Defenses on Object Detection

Defense methods against adversarial perturbations have been widely explored, including techniques such as image denoising [22] and scaling [48]. However, these methods are less effective against adversarial patch attacks, as patches are unrestricted. In recent years, numerous defenses have emerged specifically targeting adversarial patches [27, 37, 44]. These defenses can be classified into **certified defenses** and **empirical defenses**. Certified defenses [46, 47] require strict assumptions of the threat model and are limited by predefined patch size and quantity. In contrast, empirical defenses can be further categorized into three subtypes based on their underlying principles: defenses based on patch detection/segmentation [15, 16, 27, 44], defenses leveraging prior knowledge of patches [20, 31, 37, 39], and defenses based on generative models [17, 26]. The overview of the defenses evaluated in this study is shown in Tab. 1. (Implementation details of the defense methods are provided in Appendix A.2)

3. New Adversarial Patch Dataset

3.1. Adversarial Patch Generation

In this paper, we generate adversarial patches, denoted as δ, by conducting white-box adversarial attacks on various object detectors. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ be the set of object detectors, where a single detector $f_i : X \longmapsto Y$ takes input images $x \in X$ and label $y \in Y$. We apply the adversarial patch δ to the input image x through an applier $\mathcal{A}$ with

Method	Category
SAC [27] (CVPR'22)	Patch Segmentation
PAD [16] (CVPR'24)	Patch Segmentation
Adyolo [15] (ArXiv'21)	Patch Detection
NAPGuard [44](CVPR'24)	Patch Detection
LGS [31] (WACV'19)	Prior Knowledge of Patches
Zmask [37] (AAAI'23)	Prior Knowledge of Patches
Jedi [39] (CVPR'23)	Prior Knowledge of Patches
DIFFender [17] (ECCV'24)	Generative Models
NutNet [26] (CCS'24)	Generative Models
DetectorGuard [46](CCS'21)	Certified Defenses
Objectseeker [47] (S&P'23)	Certified Defenses

Table 1. Overview of our evaluated **Empirical/Certified** defenses.

transformations $t \in T$ (e.g., rotation, scaling, etc.). In this work, we focus on universal adversarial patches, which can be expressed as:

$$\delta^* = \arg\min_{\delta} \mathbb{E}_{x \sim X} \left[\mathcal{L}(f_i(\mathcal{A}(x, \delta, t)), y)\right] + \lambda L_{tv}(\delta) \quad (1)$$

where the loss function $\mathcal{L}(\cdot)$ measures the corruption of the detector, typically using object or class confidence, and $L_{tv}(\cdot)$ is the total variation loss, which encourages the generated adversarial patch to be smoother.

Different attack methods utilize various loss functions and strategies for optimizing the adversarial patches, resulting in multiple optimization spaces and differences in the generated patches. Some patches exhibit better attack performance, while others are smoother and more natural. A diverse set of patches is necessary for a comprehensive evaluation of defense methods.

3.2. Adversarial Dataset Construction

Based on the above-described principles for adversarial patch generation, we construct the large-scale Adversarial Patch Defense Evaluation (APDE) dataset. According to statistics, INRIA-Person [7] and MS COCO [25] are the main (clean) datasets used for adversarial attacks and defenses against object detection. To ensure diversity, we select these two datasets to generate 94 types of adversarial patches using 13 adversarial patch attack methods and 11 detectors. Implementation details of these attack methods are provided in Appendix A.1.

Each patch is applied to the test sets of INRIA-Person and MS COCO, resulting in 94,000 images in total. All images are stored in PNG format with a fixed size of 416 × 416 pixels, achieved through padding or resizing, aligning with the settings described in the respective papers. The dataset is divided into a training set (56,400 images) and a testing set (37,600 images), following a 6:4 ratio.

Compared to existing adversarial patch datasets, the APDE dataset exhibits three prominent advantages:

- **Large Scale.** Compared to Apricot (1,011 images) and GAP (9,266 images), the APDE dataset contains 94,000 images, comprehensively covering various scenarios and attack types. This large-scale dataset enables a more accurate evaluation of model generalization and robustness, providing an unprecedented data foundation for in-depth research on patch defense.
- **Diverse Patches.** Compared to Apricot (60 types) and GAP (25 types), the APDE dataset includes 94 types of patches. These patches come in various shapes, including common geometric forms such as circles, squares, and rectangles, as well as natural shapes like dogs or cartoon patterns. The APDE dataset offers a diverse data distribution for training and evaluating defense models, enabling a more thorough evaluation of patch defenses.
- **White-Box Setting.** Adversarial patches in the APDE dataset are trained on 11 mainstream object detectors, creating a white-box evaluation setting for patch defenses. This setup poses the worst-case adversarial threat to the detectors, allowing for the measurement of a defense model's worst-case performance and the identification of weaknesses in various defense methods.

The above advantages also make our APDE dataset more effective than previous patch datasets [3, 44], in training robust defense models, even for out-of-domain attacks that are unseen during training. See detailed experimental results in Section 5.2 and Appendix B.3.

4. Experiments

In this section, we first outline the evaluation metrics (Sec 4.1) and then systematically evaluate defense methods. We investigate the correlation between patch detection accuracy and actual defense performance in digital domains (Sec 4.2), validate the practicality of existing defenses in physical-world scenarios (Sec 4.3), and analyze the defense robustness against adaptive attacks (Sec 4.4). Finally, we conduct an empirical evaluation of certified defenses under constrained threat models (Sec 4.5).

4.1. Evaluation Metrics

Some existing studies, however, rely solely on patch detection accuracy (e.g., patch AP@0.5) as the primary criterion for defense effectiveness [15, 44], raising fairness concerns. To address this limitation, we use widely adopted **average precision (AP)** and **attack success rate (ASR)** as core metrics, while experimentally demonstrating the inadequacy of patch detection accuracy for objective comparisons (Sec 4.2). To further enhance evaluation rigor, we introduce two metrics: **Defense efficiency**: We measure the inference time per image (in milliseconds) to assess computational practicality. **Mask-based detection accuracy**: We adopt mean Intersection over Union (mIoU) [10] instead of

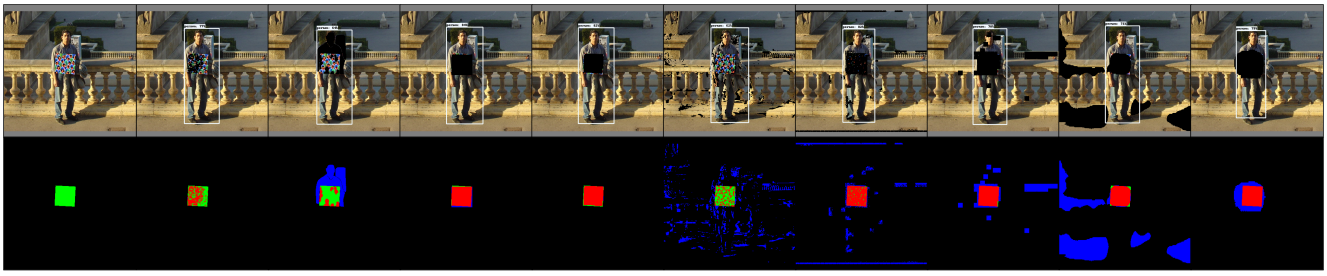

Figure 1. **Comparison of patch detection and defense performance. The first row**: Images with adversarial patches. **The second row**: Patch mask images generated by defenses (Green: the true location of patches. Red: detected patches by the defense. Blue: the background mistakenly identified as patches). Without defenses, the adversarial patch causes an incorrect prediction (in the first column), while defenses correct the prediction (in the remaining columns).

patch AP@0.5. This adjustment is critical because irregular or non-rectangular masks produced by certain defense methods cannot be accurately evaluated via axis-aligned bounding boxes. The mIoU metric directly quantifies spatial overlap between predicted and ground-truth masks, providing a shape-agnostic measure of localization precision.

$$\text{SmIoU} = \frac{\sum_{n=1}^{N} \text{TP}_n}{\sum_{n=1}^{N} (\text{FP}_n + \text{TP}_n + \text{FN}_n)} \tag{2}$$

$$\text{NmIoU} = \frac{1}{N} \sum_{n=1}^{N} \frac{\text{TP}_n}{\text{FP}_n + \text{TP}_n + \text{FN}_n} \tag{3}$$

SmIoU first aggregates predictions for positive and negative samples across all images, making it more sensitive to large pixel areas. NmIoU, on the other hand, averages the mIoU results of each image individually, reflecting image-level performance.

4.2. Patch Defenses in the Digital Domain

Fig. 1 compares patch detection and defense performance. Tab. 2 presents the mean and minimum person AP scores for each defense method against hiding attacks. Higher AP scores indicate better defense performance. (In addition to defenses against hiding attacks, we provide comprehensive experiments in supplementary material, including evaluations of defense method impacts on clean sample detection (Appendix B.1) and robustness assessments against appearing adversarial attacks (Appendix B.5))

For the person AP@0.5, PAD [16], NAPGuard [44], and NutNet [26] significantly outperform others, dominating both mean and minimum AP results. Defense efficiency is also critical: as shown in Tab. 2, PAD's high time cost (∼30s per image) limits its practicality. Defense performance varies widely across method categories, showing no systematic dependence on defense type. Among detectors, CenterNet [8] and DDETR [4] have the lowest undefended

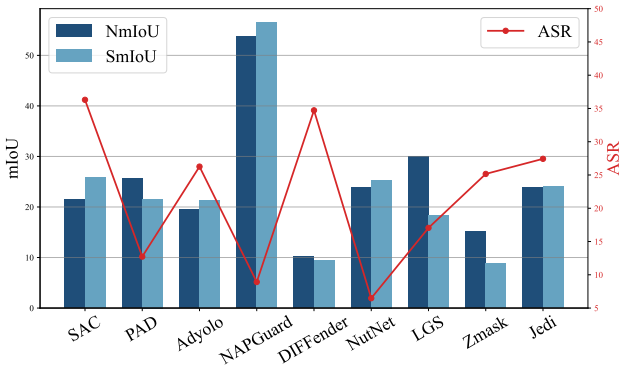

Figure 2. **Inconsistency in the relative strength of different defenses regarding patch detection accuracy vs. defense effectiveness.** For example, NAPGuard [44] achieves the highest patch detection accuracy, while NutNet [26] achieves the highest defense effectiveness. We report the NmIoU and SmIoU scores for patch detection, and the Attack Success Rate (ASR) after defense for defense effectiveness.

AP scores, highlighting their need for robust adversarial defenses. Notably, NutNet achieves the lowest ASR compared to other defenses, as visualized in Fig. 2. Overall, NutNet achieves the best defense performance and is also among the top in terms of defense efficiency.

Fig. 2 illustrates the patch detection accuracy of each defense method. A higher mIoU score, the better the defense method's generated mask aligns with the attack patch locations, thus indicating better detection performance. As shown in Fig. 2, the detection performance of NAPGuard [44] significantly outperforms all other methods. However, NutNet [26], which achieves the best defense performance, shows only average detection ability. This finding underscores that strong patch detection does not necessarily equate to better defense performance. Therefore, **the average precision of the attacked object, rather**

Model (w/o defense)	**SAC** [27]		**PAD** [16]		**Adyolo** [15]		**NAPGuard**[44]		**DIFFender**[17]		**NutNet** [26]		**LGS** [31]		**Zmask** [37]		**Jedi** [39]	
	mean	min	mean	min	mean	min	mean	min	mean	min	mean	min	mean	min	mean	min	mean	min
YOLOv2 (21.73)	52.08	29.23	67.22	56.48	57.01	48.10	64.78	47.48	51.35	38.38	**68.68**	**65.03**	63.04	48.67	49.46	31.02	53.31	45.66
YOLOv3 (28.4)	72.92	48.89	86.65	76.85	73.77	57.35	**88.50**	**82.58**	69.15	48.68	85.42	81.18	85.12	61.68	77.26	62.79	68.77	58.63
YOLOv4 (39.97)	69.35	43.20	75.64	65.66	72.84	60.26	**86.59**	**81.92**	62.98	38.17	82.80	75.66	73.78	48.53	61.73	38.54	61.71	51.05
YOLOv5 (46.03)	64.85	48.16	82.27	**76.93**	70.57	57.04	**83.88**	74.06	61.02	42.03	80.92	74.83	72.45	51.60	62.72	43.74	64.53	50.13
YOLOv7 (65.14)	74.19	68.41	75.95	**73.90**	73.52	65.27	**77.74**	70.48	56.53	53.07	63.66	59.34	72.71	66.18	69.26	61.76	69.99	65.79
SSD (16.59)	46.77	23.69	**69.23**	61.95	48.45	37.41	65.48	47.68	50.82	39.71	69.20	**66.33**	64.91	49.36	56.65	42.83	47.51	40.93
CenterNet (14.66)	47.83	22.39	69.03	40.99	50.29	37.74	64.00	52.36	38.39	17.47	**71.04**	**62.71**	61.85	45.69	45.53	18.79	49.43	40.82
RetinaNet (25.84)	62.20	39.30	81.26	**79.37**	63.09	42.29	76.36	58.76	53.51	39.69	**82.75**	71.18	75.78	42.87	51.00	24.81	59.31	45.98
MRCNN (28.89)	67.42	58.88	83.01	73.11	69.67	38.15	84.48	73.53	69.09	62.79	**86.80**	**75.54**	78.68	63.13	57.23	27.10	66.93	57.07
FRCNN (46.37)	65.78	54.10	80.14	67.74	72.77	60.04	**85.81**	**75.52**	68.87	56.34	83.95	75.12	80.55	55.89	62.25	44.87	68.08	57.66
DDETR (4.56)	46.33	20.62	**66.88**	53.67	42.38	26.59	57.75	46.93	36.78	11.98	66.56	**55.79**	58.52	29.55	30.73	6.43	37.75	18.96
Overall (30.74)	60.88	20.62	76.12	40.99	63.12	26.59	75.94	46.93	56.23	11.98	**76.53**	**55.79**	71.58	29.55	56.71	6.43	58.85	18.96
Time Cost (ms)	**44**		32100		62		59		1240		71		82		417		349	

Table 2. **Comparison of different defense performance against hiding attacks on 11 detectors.** We report mean and minimum person AP@0.5, as well as the time cost for each defense. SAC [27] exhibits the minimal time cost, while NutNet [26] demonstrates the optimal defense performance. (In Appendix B.7, we provide detailed defense performance results against individual attacks.)

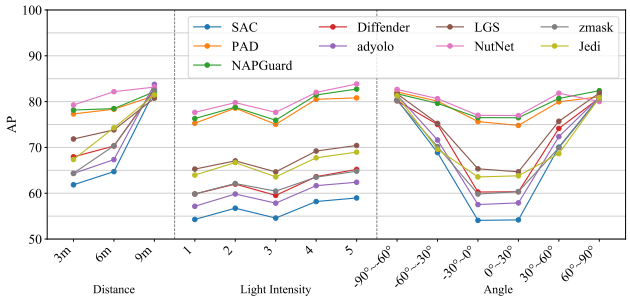

Figure 3. **Comparison of different defense methods in the physical world.** The defense performance is evaluated at different distances, light intensities, and angles.

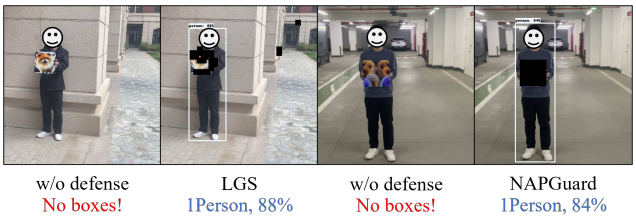

Figure 4. **Physical world attack and defense using a printed version of the patch.** The printed patch can successfully attack the detector, but it can be defended by LGS [31] and NAPGuard [44].

than the commonly pursued patch detection accuracy, shows high consistency with defense performance.

4.3. Patch Defenses in the Physical World

Adversarial patch attacks in the physical world can lead to serious security incidents. However, prior work has insufficiently explored defense performance in the physical world. Therefore, this section conducts experiments to investigate this issue. We select 30 types of patches and test them under different light intensities (1→5, gradually increasing), at varying distances (3m/6m/9m), and angles (-90° to 90°). Using an iPhone16pro, we capture 540 frames of adversarial images. We test each defense and record the results, as shown in Fig. 3. Our experiments show that attack effectiveness declines with distance (e.g., AP rises from 60% to 85% as distance increases from 0m to 9m). Defenses perform worst in the [-30°, 30°) range, with AP scores 30-50% lower than other angles. Attack effectiveness decreases while defense performance improves with increasing light intensity (levels 1-5). As shown in Fig. 4, without defenses, the patch can successfully attack the victim detector under different angles, distances, and background environments. When using LGS [31] or NAPGuard [44], the patch is effectively eliminated. Therefore, defenses that perform well in the digital domain also generally demonstrate robustness in the physical world. Surprisingly, PAD [16], which performs well in the digital domain, is surpassed by many other defense methods in the physical world. Observing the patch masks generated by the defense methods, we find that although PAD can detect and remove attack patches, it is prone to misclassifying humans as patches, leading to incorrect predictions by the detector.

4.4. Patch Defenses against Adaptive Attacks

An adaptive attack is designed with full knowledge of the defense mechanism, where adversaries specifically target the protection methods [1]. The effectiveness against adaptive attacks is an important aspect in evaluating defense methods. While [26] argues that employing additional defense models to detect or filter adversarial patches may itself be vulnerable to adversarial attacks, and [17] suggests that some defense models are bypassed by adaptive attacks because their gradients can be easily obtained. Unlike empirical defenses, certified defenses theoretically pre-

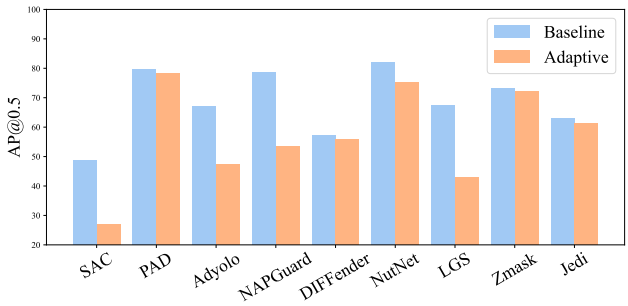

Figure 5. **Comparison of defense performance against baseline and adaptive attacks.** Adaptive attack resilience varies across defenses — PAD [16] remains robust, whereas SAC [27] is easily bypassed. (Details of the adaptive attacks are in Appendix A.3)

vent adaptive attacks within their threat models. As shown in Fig. 5, empirical defenses like PAD remain relatively robust, whereas the others, such as SAC, are more vulnerable.

We analyze in detail the underlying principles of each defense method. For defenses based on patch detection or segmentation, only PAD remains largely unaffected. This is because PAD utilizes the complex SAM model [21] and relies on semantic differences, making it difficult to optimize gradients for adaptive attacks. For defenses based on generative models, DIFFender is less susceptible to adaptive attacks compared to NutNet, due to the inherent stochasticity of the Diffusion model [35] used by DIFFender, which complicates the training of adaptive attacks. For defenses based on prior knowledge of patches, LGS is more easily bypassed by incorporating patch smoothness into the loss function, as pixel-level inconsistency occurs only in specific patches, lacking universality. Both Zmask and Jedi target universal patch properties: Zmask addresses feature-level over-activation, while Jedi tackles image-level high-entropy. Suppressing these characteristics inherently decreases the attack performance of patches, rendering adaptive attacks ineffective. In summary, **adaptive attacks can substantially bypass existing defenses, and defenses with complex/stochastic models or universal patch properties are relatively robust.**

4.5. Certified Defenses

Due to the deterministic guarantees and rigorous robustness proofs, certified defenses have gained increasing attention [46, 47]. However, certified defenses rely on strict threat model assumptions. Therefore, we change the setting to constrained attack scenarios for a fair comparison and evaluate certified and empirical defenses. For example, [47] requires explicit constraints on the number of patches in a scene and their size specifications. In the default setting, we set the number of vertical/horizontal lines (k = 30) and the filtering threshold (τ = 0.6). As shown in the

Model	YOLOv3			FRCNN		
	1	2	3	1	2	3
w/o defense	32.20	29.83	27.14	43.67	41.04	36.52
SAC [27]	61.92	58.51	57.33	67.35	65.62	61.39
PAD [16]	86.57	80.24	81.31	84.74	81.31	80.94
Adyolo [15]	71.08	67.42	63.84	73.07	69.34	65.70
NAPGuard[44]	87.37	**85.23**	**83.10**	85.92	**84.14**	81.35
DIFFender[17]	69.41	67.23	68.02	68.34	66.22	65.73
NutNet [26]	**87.42**	84.04	82.09	**86.12**	83.57	**81.42**
LGS [31]	78.37	74.81	72.63	74.23	71.48	67.28
Zmask [37]	77.45	75.41	71.28	69.25	64.36	62.19
Jedi [39]	68.37	64.19	62.40	69.63	66.22	66.15
DetectorGuard[46]	68.50	68.32	/	70.63	70.04	/
ObjectSeeker[47]	69.16	68.91	/	71.45	71.03	/

Table 3. **Comparison of 9 empirical and 2 certified defenses with different numbers of adversarial patches (1, 2, 3) under T-SEA [14] attack**. We report person AP@0.5.

Tab. 3, we evaluate defense performance by varying the number of patches. Experimental results indicate that defense effectiveness generally declines slightly as the number of patches increases, though ObjectSeeker [47] exhibits relatively smaller performance degradation. However, since ObjectSeeker employs exhaustive enumeration of potential patch locations, its computational cost escalates exponentially with the number of patches, posing a significant barrier to broader adoption. Notably, empirical defenses like PAD and NutNet maintain superior performance (consistently above 80% AP across all patch counts) compared to certified approaches. Future work could focus on expanding the scope of threat models (e.g., relaxing assumptions on patch counts or sizes) while preserving robustness guarantees, thereby advancing certified defenses from theoretical frameworks toward real-world applicability.

5. Comprehensive Analyses

In this section, we first investigate the root causes of defense failures in specific adversarial patch instances (Sec 5.1). Subsequently, we retrain mainstream defense methods using our APDE dataset and demonstrate significant performance improvements (Sec 5.2). Finally, we systematically analyze the impact of adversarial patches with varying sizes on defense robustness (Sec 5.3).

5.1. Defense Failures

During the experiments, we observe that the defense methods perform poorly against certain adversarial patches. We select a few representative attack methods and test the defense performance against them, with the results presented in Tab. 4. It is clear that when detectors are attacked by GNAP [12] and DM-NAP [23] patches, most defense methods show a slight improvement in person AP. In contrast,

Attack (w/o defense)	SAC [27]			PAD [16]			Adyolo [15]			NAPGuard [44]			DIFFender [17]			NutNet [26]		
	AP	Diff	SmIoU	AP	Diff	SmIoU	AP	Diff	SmIoU	AP	Diff	SmIoU	AP	Diff	SmIoU	AP	Diff	SmIoU
T-SEA [14] (36.57)	48.79	12.22	21.22	73.55	**36.98**	23.88	56.3	19.73	22.39	79.21	**42.64**	**77.18**	47.58	11.01	8.4	68.03	31.46	18.05
TC-EGA [13] (44.41)	58.75	14.34	**31.84**	72.78	28.37	21.38	64.32	19.91	**35.9**	68.2	23.79	49.23	59.98	15.57	11.87	80.16	35.75	**43.79**
AdvPatch [40] (33.89)	52.37	**18.48**	20.92	68.36	34.47	21.35	57.34	**23.45**	17.85	74.36	40.47	65.18	51.59	**17.7**	8.34	72.32	**38.43**	20.8
GNAP [12] (73.48)	**73.47**	-0.01	0.07	83.15	9.67	**25.45**	**75.48**	2.0	4.61	**81.26**	7.78	10.59	67.51	-5.97	**13.8**	80.5	7.02	38.86
DM-NAP [23] (71.23)	71.42	0.19	1.32	**84.59**	13.36	18.6	72.16	0.93	1.84	74.01	2.78	35.01	**72.08**	0.85	7.74	**81.32**	10.09	13.04

Table 4. **Results for adversarial patches that bypass defenses**. We report AP@0.5 after defenses, the difference in AP@0.5 before and after defenses, and SmIoU. The result without defense is located next to the attack method's name.

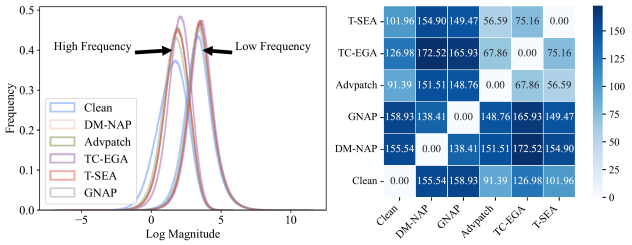

Figure 6. Left: **Frequency distribution of patches**. Right: **FID scores between patches** from different attack methods. (In Appendix B.6, we provide additional case studies on frequency distribution and FID scores.)

Attack	SAC [27]		Adyolo [15]		NAPGuard [44]	
	original	retrained	original	retrained	original	retrained
T-SEA [14]	51.82	**71.61**	66.61	**72.47**	83.61	**86.31**
TC-EGA [13]	58.16	**71.36**	63.49	**70.91**	68.51	**85.30**
Advpatch [40]	56.53	**73.29**	65.54	**72.07**	78.45	**85.10**
GNAP [12]	70.03	**76.86**	72.94	**78.52**	78.96	**85.42**
DM-NAP [23]	68.50	**76.48**	69.26	**76.83**	71.37	**85.71**
AdvCloak [45]	4.17	**71.29**	18.29	**22.36**	52.21	**73.16**
AdvTshirt [50]	34.27	**64.47**	8.19	**37.53**	50.21	**70.89**

Table 5. **Comparison of defense performance before and after retraining on the APDE dataset**, where AdvCloak and AdvTshirt are out-of-domain patches from the APDE dataset. We report original and retrained AP@0.5 after defenses.

under attacks from other patches, the defenses are more effective. The main distinction between these patches lies in their visual characteristics, with the former appearing more natural. Previous studies [31, 43, 44] suggested that the adversarial characteristics of patches are mainly concentrated in the high-frequency components. Unlike non-NAPs, high-frequency components of NAPs appear more similar to their surroundings, making them more deceptive and harder to detect accurately [44].

However, our experiments indicate that this explanation is not entirely accurate. As shown in Fig. 6, we measure the frequency histograms of different adversarial patches and find little difference in the high-frequency components between NAPs and non-NAPs. Compared to background features, their high-frequency components are more similar and difficult to distinguish. Nevertheless, NAPs do evade defenses more effectively than non-NAPs, prompting us to investigate the issue from the data distribution perspective. We compute the Fréchet Inception Distance (FID) scores [38] between each of these 5 patches and clean samples. The results in Fig. 6 reveal that non-NAPs exhibit closer FID scores to each other, indicating more similar data distributions, while NAPs are more distantly distributed. Defenses based on patch detection/segmentation or generative models essentially rely on data distribution to determine whether a pixel contains an adversarial patch. **The difficulty in defending against naturalistic patches lies in the data distribution, rather than the commonly believed high frequencies.**

5.2. Improving Defenses with our APDE dataset

To further validate the above viewpoint, we retrain several defenses using our proposed large-scale APDE dataset (Sec 3.2) by splitting into an 6:4 train-test ratio. As shown in Tab. 5, the retrained defense methods achieve significant improvement in performance (In Appendix B.3, we demonstrate that our APDE dataset more effectively enhances defense performance compared to existing patch datasets). Furthermore, we select two types of patches [45, 50] that are not included in the training dataset. Many defense methods rely on pre-training strategies where the choice of adversarial patch dataset used during training significantly impacts the defense performance. For example, Adyolo [15] improves the overall performance and generalization by alternately optimizing the patch and the defense model through adversarial training. NAPGuard [44] improves detection of naturalistic adversarial examples via Aggressive Feature Aligned Learning. However, despite these advancements, we find that when the dataset covers only a limited range of adversarial patches, the defense method may overfit, leading to poor generalization against out-of-domain patches. **Our new dataset with diverse patch distributions can be used to improve existing defenses by 15.09% AP@0.5.**

As shown in Fig. 7, we take SAC before and after retraining as an example. The first and third columns show T-SEA [14] attack patches trained on YOLOv2 [33]. The second and fourth columns show AdvTshirt [50] attack patches trained on YOLOv2 [33]. The former is included in the training dataset, while the latter is not included. Clearly, the

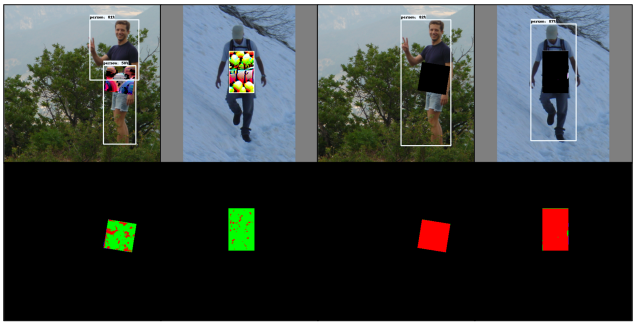

Figure 7. **Patch detection and defense performance of SAC** [27] **before and after retraining**. **Top**: Images with adversarial patches. **Bottom**: Patch mask images generated by defenses.

masks generated after retraining more completely eliminate the adversarial patches, restoring normal prediction results. This is because the APDE dataset contains a wide variety of patches, significantly enriching the defense method's understanding of the patch data distribution. Even for unseen patches, the retrained defenses demonstrate better defense effectiveness and robustness.

5.3. Impact of Patch Properties

First, we investigate the impact of adversarial patches with varying sizes on defenses. Fig. 8 shows the defense and detection performance of each method for adversarial patches of different sizes, measured by person AP and SmIoU. The experimental results reveal that defenses are more successful at detecting larger patches but perform worse in defense performance. When examining the mIoU of SAC [27], Adyolo [15], NAPGuard [44] and NutNet [26] in Fig. 2, it is obvious that their SmIoU scores are higher than their NmIoU scores, while other defense methods exhibit the opposite phenomenon. We carefully analyze the individual images with patches of different sizes to further explore this issue.

As shown in Fig. 9, we display the masks generated by PAD [16] and SAC [27] for both large and small patches. Blue indicates normal background pixels misidentified as patches by the defense. PAD tends to recognize the background as part of the patch, leading to lower mIoU scores. In contrast, SAC tends to generate smaller masks that overlap as much as possible with the patch region, thus benefiting the mIoU calculation. Since larger patches are detected more effectively, SAC's SmIoU is larger than its NmIoU. On the other hand, PAD's recognition of large background areas as patches indicates a large number of false positive pixels, resulting in SmIoU being smaller than NmIoU. Therefore, by comparing the SmIoU and NmIoU scores, we can assess whether a defense method tends to misclassify

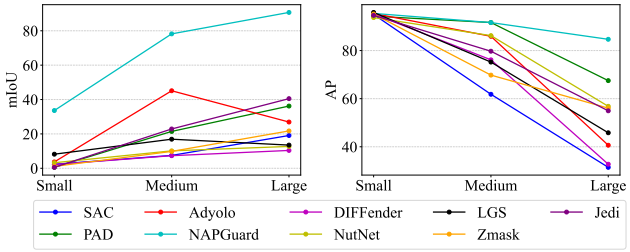

Figure 8. **Comparison of SmIoU (left) and AP@0.5 (right) for defenses against patches of different sizes.**

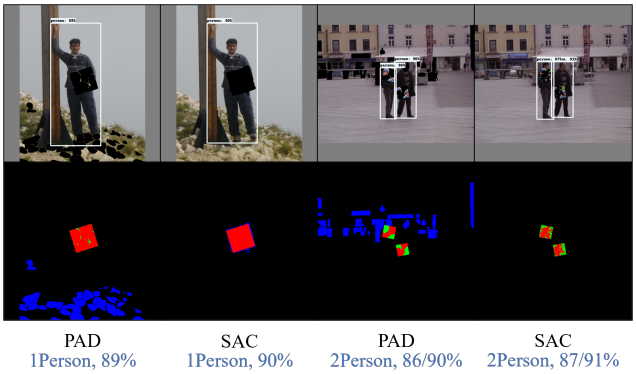

Figure 9. **Patch detection and defense performance of PAD** [16] **and SAC** [27] **for patches of different sizes**. **Top**: Images with adversarial patches. **Bottom**: Patch mask images generated by defenses. The first two columns represent large patches, and the last two columns represent small patches.

the background as a patch, and use this to evaluate its robustness.

We further explore the impact of patch shapes in Appendix B.2 and patch erasure techniques in Appendix B.4. We find that while simple patch erasure methods (e.g., black filling) are generally effective, the robustness against irregularly shaped patches varies significantly across methods, with performance degradation often tied to rectangular-based localization or limited training data diversity.

6. Conclusion

In this paper, we revisit 11 representative defenses and present the first patch defense benchmark, regarding various metrics. We conduct comprehensive analyses to reveal new insights, including defense vulnerability principles, evaluation metrics, and adaptive attacks. We also construct a new large-scale adversarial patch dataset to evaluate and improve defenses. For future work, we plan to test defense performance against various physically feasible attacks, such as attacks printed on surfaces like clothes or cars.

7. Acknowledgments

We would like to thank Yan Zhang, Haoran Fan, and the anonymous reviewers for their valuable feedback. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB3107401; the National Natural Science Foundation of China (T2341003, 62376210, 62161160337, 62132011, U24B20185, U21B2018, 62206217, 62402377), the Shaanxi Province Key Industry Innovation Program (2023-ZDLGY-38). Thanks to the New Cornerstone Science Foundation and the Xplorer Prize.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 5

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[3] Anneliese Braunegg, Amartya Chakraborty, and Krumdick. Apricot: A dataset of physical adversarial attacks on object detection. In *European Conference on Computer Vision*, pages 35–50. Springer, 2020. 2, 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 4

[5] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[6] Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, and Gursoy. Adversarial objectness gradient attacks in real-time object detection systems. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 263–272. IEEE, 2020. 2

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 3

[8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2, 4

[9] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1

[10] Levente Halmosi, Bálint Mohos, and Márk Jelasity. Evaluating the adversarial robustness of semantic segmentation: Trying harder pays off. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 3

[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. 2

[12] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, and Chen. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 6, 7

[13] Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13307–13316, 2022. 2, 7

[14] Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20514–20523, 2023. 2, 6, 7

[15] Nan Ji, YanFei Feng, Haidong Xie, Xueshuang Xiang, and Naijin Liu. Adversarial yolo: Defense human detection patch attacks via detecting adversarial patches. *arXiv preprint arXiv:2103.08860*, 2021. 1, 2, 3, 5, 6, 7, 8

[16] Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, and Cong Zou. Pad: Patch-agnostic defense against adversarial patch attacks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24472–24481, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[17] Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Hang Su, and Xingxing Wei. Diffender: Diffusion-based adversarial defense against patch attacks in the physical world. *arXiv preprint arXiv:2306.09124*, 2023. 1, 2, 3, 5, 6, 7

[18] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International conference on machine learning*, pages 2507–2515. PMLR, 2018. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2

[20] Taeheon Kim, Youngjoon Yu, and Yong Man Ro. Defending physical adversarial attack on object detection via adversarial patch-feature energy. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1905–1913, 2022. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6

[22] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1778–1787, 2018. 2

[23] Shuo-Yen Lin, Ernie Chu, Che-Hsien Lin, Jun-Cheng Chen, and Jia-Ching Wang. Diffusion to confusion: Naturalistic ad-

versarial patch generation based on diffusion model for object detector. *arXiv preprint arXiv:2307.08076*, 2023. 2, 6, 7

[24] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, and Perona. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3

[26] Zijin Lin, Yue Zhao, Kai Chen, and Jinwen He. I don't know you, but i can catch you: Real-time defense against diverse adversarial patches for object detectors. *arXiv preprint arXiv:2406.10285*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[27] Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14973–14982, 2022. 1, 2, 3, 5, 6, 7, 8

[28] 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. 2

[29] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. Slowtrack: Increasing the latency of camera-based perception in autonomous driving using adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4062–4070, 2024. 1

[30] Chen Ma, Ningfei Wang, Zhengyu Zhao, Qian Wang, Qi Alfred Chen, and Chao Shen. Controlloc: Physical-world hijacking attack on visual perception in autonomous driving. *arXiv preprint arXiv:2406.05810*, 2024. 1

[31] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307, 2019. 1, 2, 3, 5, 6, 7

[32] Joseph Redmon. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[33] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1, 2, 7

[34] Shaoqing Ren. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 6

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[37] Giulio Rossolini, Federico Nesti, Fabio Brau, Alessandro Biondi, and Giorgio Buttazzo. Defending from physically-realizable adversarial attacks through internal over-activation analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15064–15072, 2023. 1, 2, 3, 5, 6

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[39] Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. Jedi: Entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4087–4095, 2023. 2, 3, 5, 6

[40] 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. 2, 7

[41] Ultralytics. YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com, 2021. Accessed: insert date here. 2

[42] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 2

[43] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020. 7

[44] Siyang Wu, Jiakai Wang, Jiejie Zhao, Yazhe Wang, and Xianglong Liu. Napguard: Towards detecting naturalistic adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24367–24376, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[45] 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. 2, 7

[46] Chong Xiang and Prateek Mittal. Detectorguard: Provably securing object detectors against localized patch hiding attacks. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 3177–3196, 2021. 2, 3, 6

[47] Chong Xiang, Alexander Valtchanov, Saeed Mahloujifar, and Prateek Mittal. Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1329–1347. IEEE, 2023. 2, 3, 6

[48] Qixue Xiao, Yufei Chen, Chao Shen, Yu Chen, and Kang Li. Seeing is not believing: Camouflage attacks on image scaling algorithms. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 443–460, 2019. 2

[49] 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. 2

[50] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, and Sun. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, pages 665–681. Springer, 2020. 2, 7

[51] Yulong Yang, Chenhao Lin, Qian Li, Zhengyu Zhao, Haoran Fan, Dawei Zhou, Nannan Wang, Tongliang Liu, and Chao Shen. Quantization aware attack: Enhancing transferable adversarial attacks by model quantization. *IEEE Transactions on Information Forensics and Security*, 19:3265–3278, 2024. 1

[52] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. 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*, pages 1989–2004, 2019. 2

[53] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *CVPR*, 2020. 1

[54] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. In *NeurIPS*, 2021. 1

[55] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems*, 37:59570–59596, 2024. 2

[56] Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, and Chao Shen. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24452–24461, 2024. 1