

# DetectorGuard: Provably Securing Object Detectors against Localized Patch Hiding Attacks

Chong Xiang  
Princeton University  
Princeton, NJ, USA  
cxiang@princeton.edu

Prateek Mittal  
Princeton University  
Princeton, NJ, USA  
pmittal@princeton.edu

## ABSTRACT

State-of-the-art object detectors are vulnerable to localized patch hiding attacks, where an adversary introduces a small adversarial patch to make detectors miss the detection of salient objects. The patch attacker can carry out a physical-world attack by printing and attaching an adversarial patch to the victim object; thus, it imposes a challenge for the safe deployment of object detectors. In this paper, we propose DetectorGuard as the first general framework for building provably robust object detectors against localized patch hiding attacks. DetectorGuard is inspired by recent advancements in robust image classification research; we ask: *can we adapt robust image classifiers for robust object detection?* Unfortunately, due to their task difference, an object detector naively adapted from a robust image classifier 1) may not necessarily be robust in the adversarial setting or 2) even maintain decent performance in the clean setting. To address these two issues and build a high-performance robust object detector, we propose an *objectness explaining strategy*: we adapt a robust image classifier to predict objectness (i.e., the probability of an object being present) for every image location and then explain each objectness using the bounding boxes predicted by a conventional object detector. If all objectness is well explained, we output the predictions made by the conventional object detector; otherwise, we issue an attack alert. Notably, our objectness explaining strategy enables provable robustness for “free”: 1) in the adversarial setting, we formally prove the end-to-end robustness of DetectorGuard on certified objects, i.e., it either detects the object or triggers an alert, against *any patch hiding attacker* within our threat model; 2) in the clean setting, we have almost the same performance as state-of-the-art object detectors. Our evaluation on the PASCAL VOC, MS COCO, and KITTI datasets further demonstrates that DetectorGuard achieves the first provable robustness against localized patch hiding attacks at a negligible cost (<1%) of clean performance.

## CCS CONCEPTS

- Security and privacy → Domain-specific security and privacy architectures;
- Computing methodologies → Object detection; Neural networks.

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*CCS ’21, November 15–19, 2021, Virtual Event, Republic of Korea*

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8454-4/21/11.

<https://doi.org/10.1145/3460120.3484757>

## KEYWORDS

Provability; Robustness; Adversarial Patch Attack; Object Detection

### ACM Reference Format:

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 (CCS ’21), November 15–19, 2021, Virtual Event, Republic of Korea*. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3460120.3484757>

## 1 INTRODUCTION

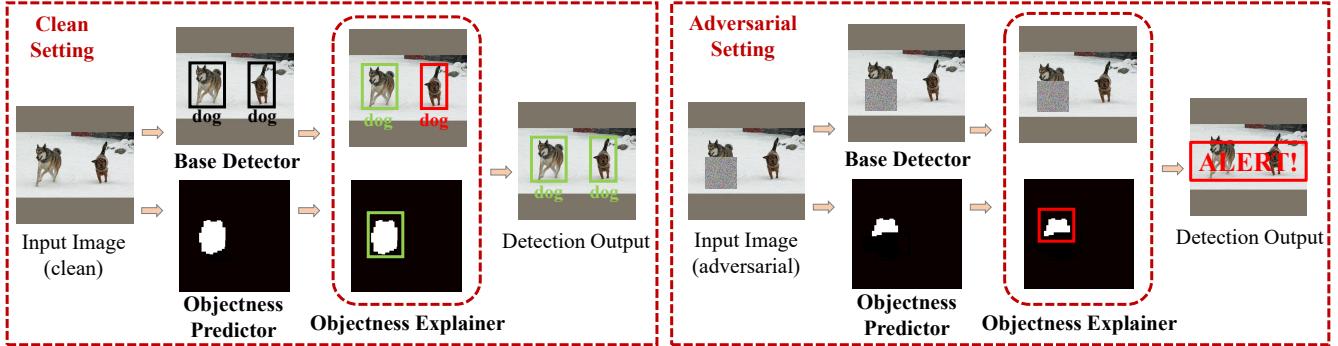
Localized adversarial patch attacks can induce mispredictions in Machine Learning (ML) systems and have gained significant attention over the past few years [4, 21, 50, 57, 61]. A patch attacker constrains all adversarial perturbations within a small region so that they can carry out a physical world attack by printing and attaching the adversarial patch to the victim object.<sup>1</sup> To counter the threat of patch attacks on real-world ML systems, the security community has been actively seeking defense mechanisms [9, 18, 24, 25, 33, 58, 65]. However, most existing defenses are restricted to the image classification domain. In this paper, we aim to secure *object detectors*, which are used in critical applications like autonomous driving, video surveillance, and identity verification [52].

We focus on the threat of *localized patch hiding attacks* against object detectors: an attacker uses a localized patch for physical world attacks that cause the object detector to fail to detect victim objects. Lee et al. [23] show that a physical patch far away from the objects can successfully “hide” victim objects. Wu et al. [57] and Xu et al. [61] have succeeded in evading object detection via wearing a T-shirt printed with adversarial perturbations. The patch hiding attack can cause serious consequences in scenarios like an autonomous vehicle missing a pedestrian (or an upcoming car).

Unfortunately, securing object detectors is extremely challenging due to the complexity of the detection task. A single image can contain multiple objects, and thus an object detector needs to output a list of object bounding box coordinates and class labels. To the best of our knowledge, there is only one prior work [46] discussing defenses for YOLOv2 [43] detectors against patch attacks, in contrast to numerous new patch attacks being proposed [7, 14, 50, 57, 61, 66]. Furthermore, this only defense [46] is restricted to the setting of a non-adaptive adversarial patch at the image corner and does not have any security guarantee (i.e., only heuristics-based). To overcome these weaknesses, we propose a defense framework named DetectorGuard that can achieve *provable robustness* against any patch hiding attack within our threat model.

---

<sup>1</sup>The patch attack significantly differs from classic  $L_p$ -norm-bounded adversarial examples [6, 16, 48] that require global perturbations and thus are difficult to realize in the physical world.



**Figure 1: DetectorGuard Overview.** *Base Detector* is a conventional detector that typically predicts precise bounding boxes on clean images. *Objectness Predictor* aims to robustly predict an objectness map. *Objectness Explainer* uses the predicted bounding boxes to explain/match the predicted objectness and determines the final output. In the clean setting (left figure), the dog on the left is detected by both Base Detector and Objectness Predictor. The objectness is well explained/matched by the green bounding box, and DetectorGuard outputs the bounding box predicted by Base Detector. In the meantime, due to the imperfection of the robust classifier, the dog on the right is missed by Objectness Predictor. Objectness Explainer will consider this as a *benign mismatch*, and DetectorGuard will also output the predicted bounding box from Base Detector. In the adversarial setting (right figure), a patch makes Base Detector fail to detect any object while Objectness Predictor still robustly outputs high objectness. Objectness Explainer detects a cluster of *unexplained objectness* and triggers an attack alert. Our defense can achieve provable robustness for certified objects while maintaining a clean performance comparable to state-of-the-art object detectors.

**Insight & Challenge: a bridge between robust image classification and robust object detection.** To start with, we observe that the robust image classification research is making significant advancement [9, 24, 33, 36, 58, 65] while object detectors remain vulnerable to attacks. This sharp contrast motivates us to ask: *can we take advantage of robust image classification research and adapt robust image classifiers for robust object detection?* Unfortunately, there is a huge gap between these two tasks: an image classifier outputs a single-label prediction while an object detector predicts a list of bounding boxes and class labels. This gap brings two major challenges for such an adaptation. *Challenge 1: Lack of End-to-end Provable Robustness.* A robust image classifier only provides robustness for a single-label prediction while an object detector requires robustness for multiple object labels and bounding boxes in each image. Therefore, using a robust image classifier does not guarantee the end-to-end robustness of the object detector, and we need a new defense design and proof of robustness. *Challenge 2: Amplified Cost of Clean Performance.* All existing provably robust image classifiers [9, 24, 36, 58, 65] attain robustness at a non-negligible cost of clean performance (e.g., state-of-the-art defense [58] incurs a >20% clean accuracy drop on ImageNet [11]). The imperfection of robust classifiers can be severely amplified during the adaptation towards the much more demanding object detection task. Therefore, we need to prevent our object detectors from being broken in the clean setting (even in the absence of an adversary). In DetectorGuard, we overcome these two challenges as discussed below.

**Defense Design: an objectness explaining strategy.** We provide our defense overview in Figure 1. DetectorGuard has three modules: Base Detector, Objectness Predictor, and Objectness Explainer. Base Detector can be any state-of-the-art object detector that can make *accurate* predictions on clean images but is vulnerable to patch hiding attacks. Objectness Predictor aims to predict a *robust* objectness map, which indicates the probability of an object being

present at different locations. We build Objectness Predictor using adapted provably robust image classifiers together with carefully designed feature-space operation and error filtering mechanisms (Section 3.3). Finally, Objectness Explainer uses each predicted bounding box from Base Detector to *explain/match* high objectness predicted by Objectness Predictor (Section 3.4). If all objectness is well explained/matched, we output the prediction of Base Detector; otherwise, we issue an attack alert. In the clean setting, we optimize the configuration of Objectness Predictor towards the case where all objectness can be explained and then use Base Detector for accurate final predictions (Figure 1 left). When a hiding attack occurs, Base Detector could miss the object while Objectness Predictor can still robustly output high objectness. This will lead to unexplained objectness and trigger an attack alert (Figure 1 right). Notably, we can show that our defense design successfully addresses *Challenge 1 and 2*, as discussed next.

**End-to-end provable robustness for “free”.** First, our objectness explaining strategy enables us to rigorously prove the end-to-end robustness of DetectorGuard (Theorem 1 in Section 4). We will show that DetectorGuard will always perform robust detection or issue an alert on objects certified by our provable analysis (Algorithm 2 in Section 4). We note that this robustness property is agnostic to attack strategies and holds for *any patch hiding attacker within our threat model, including adaptive attackers who have full access to our defense setup*. This strong theoretical guarantee addresses *Challenge 1*. Next, in contrast to most security-critical systems whose robustness comes at the cost of clean performance, DetectorGuard achieves provable robustness for “free” (at a negligible cost of clean performance). In Objectness Predictor, we design error mitigation mechanisms to handle the imperfection of the adapted robust classifier. In Objectness Explainer, our explaining strategy ensures that even when our Objectness Predictor fails to

predict high objectness (missing objects; false negatives), DetectorGuard still performs as well as state-of-the-art Base Detector.<sup>2</sup> These designs solve *Challenge 2*.

**Evaluating the first provable robustness against patch hiding attacks.** We extensively evaluate DetectorGuard performance on the PASCAL VOC [13], MS COCO [26], and KITTI [15] datasets. In our evaluation, we instantiate the Base Detector with YOLOv4 [2, 53], Faster R-CNN [45], and a hypothetical object detector that is perfect in the clean setting. We build Objectness Predictor by adapting multiple variants of robust image classifiers [58, 65]. Our evaluation shows that our defense has a minimal impact (<1%) on the clean performance and achieves the first provable robustness against patch hiding attacks. Our code is available at <https://github.com/inspire-group/DetectorGuard>.

Our contributions can be summarized as follows.

- We solve two major challenges in adapting robust image classifiers for robust object detection via a careful design of Objectness Predictor and Objectness Explainer.
- We formally prove the robustness guarantee of DetectorGuard on certified objects against any adaptive attacker within our threat model.
- We extensively evaluate our defense on the PASCAL VOC [13], MS COCO [26], and KITTI [15] datasets and demonstrate *the first provable robustness against patch hiding attacks* and a similar clean performance as conventional object detectors.

## 2 BACKGROUND AND PROBLEM FORMULATION

In this section, we introduce the object detection task, the patch hiding attack, the defense formulation, and the key principles for building provably robust image classifiers that we will adapt for robust object detection in DetectorGuard.

### 2.1 Object Detection

**Detection objective.** An object detector aims to predict a list of bounding boxes (and class labels) for all objects in the image  $\mathbf{x} \in [0, 1]^{W \times H \times C}$ , where pixel values are rescaled into  $[0, 1]$ , and  $W, H, C$  is the image width, height, and channel, respectively. Each bounding box  $\mathbf{b}$  is represented as a tuple  $(x_{\min}, y_{\min}, x_{\max}, y_{\max}, l)$ , where  $x_{\min}, y_{\min}, x_{\max}, y_{\max}$  together illustrate the coordinates of the bounding box, and  $l \in \mathbf{L} = \{0, 1, \dots, N - 1\}$  denotes the predicted object label ( $N$  is the number of object classes).<sup>3</sup>

**Conventional object detector.** Object detection models can be categorized into two-stage and one-stage detectors depending on their pipelines. A two-stage object detector first generates proposals for regions that might contain objects and then uses the proposed regions for object classification and bounding-box regression. Representative examples include Faster R-CNN [45] and Mask R-CNN [19]. On the other hand, a one-stage object detector does

<sup>2</sup>Objectness Predictor can also have other types of errors. However, we can optimize its configuration to ensure most errors are false-negatives, which our objectness explaining strategy can tolerate. More discussions are in Section 3. We also note that we manage to build a system with high clean performance (i.e., DetectorGuard) despite the use of a module with poor clean performance (i.e., provably robust image classifier). We provide additional discussion on this intriguing property in Appendix F.

<sup>3</sup>Conventional object detectors usually output objectness score and prediction confidence as well—we discard them in notation for simplicity.

detection directly on the input image without any explicit region proposal step. SSD [29], YOLO [2, 42–44, 53], RetinaNet [25], and EfficientDet [49] are representative one-stage detectors.

Conventionally, a detection is considered correct when 1) the predicted label matches the ground truth and 2) the overlap between the predicted bounding box and the ground-truth box, measured by Intersection over Union (IoU), exceeds a certain threshold  $\tau$ . We term a correct detection a *true positive* (*TP*). On the other hand, any predicted bounding box that fails to satisfy both two TP criteria is considered as a *false positive* (*FP*). Finally, if a ground-truth object is not detected by any TP bounding box, it is a *false negative* (*FN*).

### 2.2 Attack Formulation

**Attack objective.** The hiding attack [30, 50, 57, 61, 66], also referred to as the false-negative (FN) attack, aims to make object detectors miss the detection of certain objects (which increases FN) at the test time. The hiding attack can cause serious consequences in scenarios like an autonomous vehicle missing a pedestrian. Therefore, defending against patch hiding attacks is of great importance.

**Attacker capability.** We allow the localized adversary to arbitrarily manipulate pixels within one restricted region.<sup>4</sup> Formally, we can use a binary *pixel mask*  $\mathbf{pm} \in \{0, 1\}^{W \times H}$  to represent this restricted region, where the pixels within the region are set to 1. The adversarial image then can be represented as  $\mathbf{x}' = (\mathbf{1} - \mathbf{pm}) \odot \mathbf{x} + \mathbf{pm} \odot \mathbf{x}''$  where  $\odot$  denotes the element-wise product operator, and  $\mathbf{x}'' \in [0, 1]^{W \times H \times C}$  is the content of the adversarial patch, which the adversary can arbitrarily modify.  $\mathbf{pm}$  is a function of patch size and patch location. The patch size should be limited such that the object is recognizable by a human. For patch locations, we consider three different threat models: *over-patch*, *close-patch*, *far-patch*, where the patch is over, close to, or far away from the victim object, respectively. The adversary can pick any valid location within the threat model for an optimal attack.

Previous works [23, 30, 46] have shown that attacks against object detectors can succeed even when the patch is far away from the victim object. Therefore, defending against all three threat models is of interest.

### 2.3 Defense Formulation

**Defense objective.** We focus on defending against patch hiding attacks. We consider our defense to be robust on an object if we can 1) detect the object on the clean image is correct and 2) detect part of the object or send out an attack alert on the adversarial image.<sup>5</sup>

Crucially, we design our defense to be *provably robust*: for an object certified by our provable analysis, our defense can either detect the certified object or issue an alert regardless of what the adversary does (including any adaptive attack at any patch location within the threat model). This robustness property is agnostic to the attack algorithm and holds against an adversary that has full access to our defense setup.

<sup>4</sup>Provably robust defenses against *one* single patch are currently an open/unsolved problem, and hence the focus of this paper. In Appendix C, we will justify our one-patch threat model and quantitatively discuss the implication of multiple patches.

<sup>5</sup>We note that in the adversarial setting, we only require the predicted bounding box to cover part of the object because it is likely that only a small part of the object is recognizable due to the adversarial patch (e.g., the left dog in the right part of Figure 1). We provide additional justification for our defense objective in Appendix E.

**Remark: primary focus on hiding attacks.** In this paper, we focus on the hiding attack because it is the most fundamental and notorious attack against object detectors. We can visualize dividing the object detection task into two steps: 1) detecting the object bounding box and then 2) classifying the detected object. If the first step is compromised by the hiding attack, there is no hope for robust object detection. On the other hand, *securing the first step against the patch hiding attack lays a foundation for the robust object detection*; we can design effective remediation for the second step if needed (Section 6).

Take the application domain of autonomous vehicles (AV) as an example: an AV missing the detection of an upcoming car could lead to a serious car accident. However, if the AV detects the upcoming object but predicts an incorrect class label (e.g., mistaking a car for a pedestrian), it can still make the correct decision of stopping and avoiding the collision. Moreover, in challenging application domains where the predicted class label is of great importance (e.g., traffic sign recognition), we can feed the detected bound box to an auxiliary image classifier to re-determine the class label. The defense problem is then reduced to the robust image classification and has been studied by several previous works [24, 36, 58, 65]. Therefore, we make the hiding attack the primary focus of this paper and will also discuss the extension of DetectorGuard against other attacks in Section 6.

## 2.4 Provably Robust Image Classification

In this subsection, we introduce two key principles that are widely adopted in recent research on provably robust image classification against adversarial patches [24, 36, 58, 65]. In Section 3.2, we will discuss how to adapt these two principles to build a robust image classifier, which will be later used in Objectness Predictor (Section 3.3).

**Feature extractor – use small receptive fields.** The receptive field of a Deep Neural Network (DNN) is the input pixel region where each extracted feature is looking at. If the receptive field of a DNN is too large, then a small adversarial patch can corrupt most extracted features and easily manipulate the model behavior [30, 46, 58]. On the other hand, the small receptive field bounds the number of corrupted features by  $\lceil (p+r-1)/s \rceil$ , where  $p$  is the patch size,  $r$  is the receptive field size, and  $s$  is the stride of receptive field (the pixel distance between two adjacent receptive field centers) [58], and makes robust classification possible [24, 36, 58, 65]. Popular design choices including the BagNet architecture [3, 36, 58, 65] and an ensemble architecture using small pixel patches as inputs [24, 58].

**Classification head – do secure feature aggregation.** Given a feature map, DNN uses a *classification head*, which consists of a feature aggregation layer and a fully-connected (classification) layer, to make final predictions. Since the small receptive field bounds the number of corrupted features, we can use secure aggregation techniques to build a robust classification head; design choices include clipping [58, 65], masking [58], and majority voting [24, 36].

## 3 DETECTORGUARD

In this section, we first introduce the key insight and overview of DetectorGuard, and then detail the design of our defense modules (Objectness Predictor and Objectness Explainer).

### 3.1 Defense Overview

**Bridging robust image classification and robust object detection.** There has been a significant advancement in (provably) robust image classification research [9, 24, 36, 58, 65] while object detectors remain vulnerable. This sharp contrast motivates us to ask: *can we adapt robust image classifiers for robust object detection?* Unfortunately, there is a huge gap between these two tasks: an image classifier only robustly predicts one single label for each image while an object detector has to robustly output a list of class labels and object bounding boxes. This gap leads to two major challenges.

- **Challenge 1: Lack of End-to-end Provable Robustness.** A robust image classifier only provides robustness for single-label predictions while a robust object detector requires robustness for multiple labels and bounding boxes. Therefore, an object detector adapted from a robust image classifier can still be vulnerable without any security guarantee, and we aim to carefully design our defense pipeline to enable the proof of end-to-end robustness for object detection.
- **Challenge 2: Amplified Cost of Clean Performance.** All existing provably robust image classifiers [9, 24, 36, 58, 65] attain robustness at a non-negligible cost of clean performance (e.g., >20% clean accuracy drop on ImageNet [11]), and this cost can be severely amplified when adapting towards the more demanding object detection task. An object detector with poor clean performance (even in the absence of an adversary) prohibits its real-world deployment; therefore, we aim to minimize the clean performance cost in our defense.

**DetectorGuard: an objectness explaining strategy.** In DetectorGuard, we propose an objectness explaining strategy to address the above two challenges. Recall that Figure 1 provides an overview of DetectorGuard, which will either output a list bounding box predictions (left figure; clean setting) or an attack alert (right figure; adversarial setting). There are three major modules in DetectorGuard: Base Detector, Objectness Predictor, and Objectness Explainer. Base Detector is responsible for making accurate detections in the clean setting and can be any popular high-performance object detector such as YOLOv4 [2, 53] and Faster R-CNN [45]. Objectness Predictor is adapted from the core principles for building robust image classifiers as introduced in Section 2.4 and aims to output a robust objectness map in the adversarial environment. We also carefully design Objectness Predictor to mitigate the errors made by the robust image classifier in the clean setting. Finally, Objectness Explainer leverages predicted bounding boxes from Base Detector to explain/match the objectness predicted by Objectness Predictor and aims to catch a malicious attack. When no attack is detected, DetectorGuard will output the detection results of Base Detector (i.e., a conventional object detector), so that our clean performance is close to state-of-the-art object detectors. When a patch hiding attack occurs, Base Detector can miss the object while Objectness Predictor can robustly predict high objectness. This mismatch will lead to unexplained objectness and trigger an attack alert. Notably, our objectness explaining strategy can achieve end-to-end provable robustness for “free” (at a negligible cost of clean performance) and solve two major challenges. We will introduce the module details and theoretically analyze the free provable robustness property.

**Table 1: Summary of important notation**

Notation	Description	Notation	Description
$x$	Input image	$b$	bounding box
$fm$	Feature map	$om$	Objectness map
$v$	classification logits	$N$	number of object classes
$(w_x, w_y)$	window size	$(p_x, p_y)$	patch size
$T$	binarizing threshold	$\mathcal{D}$	detection results
$u, l$	upper/lower bound of classification logits values of each class		

**Algorithm Pseudocode.** We provide the pseudocode of DetectorGuard in Algorithm 1 and a summary of important notation in Table 1. The main procedure  $DG(\cdot)$  has three sub-procedures:  $BASEDETECTOR(\cdot)$ ,  $ObjPREDICTOR(\cdot)$ , and  $DETMATCHER(\cdot)$ . The sub-procedure  $BASEDETECTOR(\cdot)$  can be any off-the-shelf object detector as discussed in Section 2.1. All tensors/arrays are represented with bold symbols and scalars are in italic. All tensor/array indices start from zeros; the tensor/array slicing is in Python style (e.g.,  $[i:j]$  means all indices  $k$  satisfying  $i \leq k < j$ ). We assume that the “background” class corresponds to the largest class index.

In the remainder of this section, we first introduce how we instantiate robust image classifiers and then discuss the design of Objectness Predictor and Objectness Explainer.

### 3.2 Instantiating Robust Image Classifiers

To start with, we discuss how we build robust image classifiers that will be used in Objectness Predictor.

As discussed in Section 2.4, we can build a robust image classifier  $RC(\cdot)$  using a *feature extractor*  $FE(\cdot)$  with small receptive fields, and a *robust classification head*  $RCH(\cdot)$  with secure feature aggregation. In our design, we choose BagNet [3] backbone as the feature extractor  $FE(\cdot)$ , and we clip elements of local logits vectors<sup>6</sup> into  $[0, \infty]$  for secure aggregation in  $RCH(\cdot)$ . This implementation is similar to the robust image classifier Clipped BagNet (CBN) [65], but we note that we use a different clipping function that is tailored to our more challenging task of *object detection*. In Appendix B, we provide additional details of  $FE(\cdot)$  and  $RCH(\cdot)$  and also discuss alternative design choices of robust image classifier  $RC(\cdot)$  (e.g., robust masking from PatchGuard [58]).

**Remark: Limitations of robust classifiers.** We note that the adapted robust image classifier  $RC(\cdot)$  achieves robustness at the cost of a non-negligible clean performance drop [24, 36, 58]. Therefore, in the clean setting, classification at different image locations can be imprecise with three typical errors that lead to *Challenge 2*:

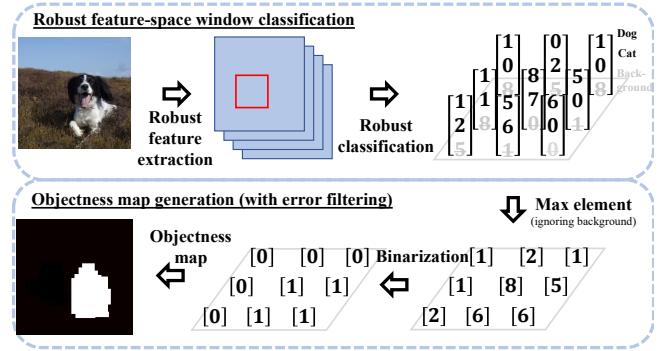
- *Clean Error 1*: Confusion between two object class labels
- *Clean Error 2*: Predicts background pixels as objects
- *Clean Error 3*: Predicts objects as “background”

In the next two subsections, we will discuss how DetectorGuard design can eliminate/mitigate these three clean errors.

### 3.3 Objectness Predictor

**Overview.** Objectness Predictor aims to output a robust objectness map that indicates the probability of objects being present at different locations. Its high-level idea is to perform robust *image*

<sup>6</sup>The local logits vector [58] is the classification logits based on each local feature that is extracted from a particular region (i.e., the receptive field) of the input image.



**Figure 2: Visualization of Objectness Predictor operations**

*classification* on different regions to predict an object class label or “background”. We provide a simplified visual overview in Figure 2. Objectness Predictor involves two major operations: robust feature-space window classification and objectness map generation. The first step aims to perform efficient robust classification at different image regions and the second step aims to filter out clean errors made by the robust classifier and generate the final objectness map.

**Robust feature-space window classification.** To perform robust image classification at different image locations, we first extract a feature map via  $FE(\cdot)$ , and then apply the robust image classification head  $RCH(\cdot)$  to a sliding window over the feature map.

*Pseudocode.* The pseudocode of Objectness Predictor in Line 12–23 of Algorithm 1. We first extract the feature map  $fm$  with  $FE(\cdot)$  (Line 13). Next, for every valid window location, represented as  $(i, j)$ ,<sup>7</sup> we feed the feature window  $fm[i : i + w_x, j : j + w_y]$  to a robust classification head  $RCH(\cdot)$  to get the classification label  $l$  and the classification logits  $v \in \mathbb{R}^{N+1}$  for  $N$  object classes and the “background” class (Line 17). The use of  $RCH(\cdot)$  ensures that the classification is robust when window features are corrupted.

*Remark: defense efficiency.* We note that our window classification operates in the feature space, and this allows us to reuse the expensive feature map generation (i.e.,  $FE(\cdot)$ ); each classification only needs a cheap computation of the classification head (i.e.,  $RCH(\cdot)$ ). Therefore, our defense only incurs a small overhead (will be evaluated in Section 5.5).

**Objectness map generation: handling clean errors of robust classifiers.** Next, we aim to filter out incorrect window classifications and generate the final objectness map. We reduce each prediction vector at each location to its maximum non-background element to discard label information (eliminating *Clean Error 1*) and perform binarization (with a threshold) to remove objectness predicted with low confidence (mitigating *Clean Error 2*).

*Pseudocode.* First, we initialize an all-zero  $\mathbb{R}^{N+1}$  vector ( $N$  object classes plus the “background” class) at every feature location and use one tensor  $om$  to represent all vectors (Line 15). Second, we aim to gather all window classification results: for each window, we add the classification logits  $v$  to every vector located within the window (Line 18). Third, we take the maximum non-background element in each vector as the objectness score at the corresponding

<sup>7</sup>We will use feature-space coordinates for the remainder of the paper. The mapping between pixel-space and feature-space coordinates is discussed in Appendix G.

---

**Algorithm 1** DetectorGuard

**Input:** input image  $\mathbf{x}$ , window size  $(w_x, w_y)$ , binarizing threshold  $T$ , Base Detector  $\text{BASEDETECTOR}(\cdot)$ , robust classifier  $\text{RC}(\cdot)$  (consisting of a feature extractor  $\text{FE}(\cdot)$  and a robust classification head  $\text{RCH}(\cdot)$ ), cluster detection procedure  $\text{DETCLUSTER}(\cdot)$

**Output:** robust detection  $\mathcal{D}^*$  or ALERT

```

1: procedure  $\text{DG}(\mathbf{x}, w_x, w_y, T)$ 
2:    $\mathcal{D} \leftarrow \text{BASEDETECTOR}(\mathbf{x})$             $\triangleright$  Conventional detection
3:    $\mathbf{om} \leftarrow \text{OBJPREDICTOR}(\mathbf{x}, w_x, w_y, T)$        $\triangleright$  Objectness
4:    $a \leftarrow \text{OBJEXPLAINER}(\mathcal{D}, \mathbf{om})$            $\triangleright$  Detect hiding attacks
5:   if  $a == \text{True}$  then                          $\triangleright$  Malicious mismatch
6:      $\mathcal{D}^* \leftarrow \text{ALERT}$                        $\triangleright$  Trigger an alert
7:   else
8:      $\mathcal{D}^* \leftarrow \mathcal{D}$               $\triangleright$  Return Base Detector's predictions
9:   end if
10:  return  $\mathcal{D}^*$ 
11: end procedure

12: procedure  $\text{OBJPREDICTOR}(\mathbf{x}, w_x, w_y, T)$ 
13:    $\mathbf{fm} \leftarrow \text{FE}(\mathbf{x})$                    $\triangleright$  Extract feature map
14:    $X, Y, \_ \leftarrow \text{SHAPE}(\mathbf{fm})$          $\triangleright$  Get the shape of  $\mathbf{fm}$ 
15:    $\mathbf{oim} \leftarrow \text{ZEROARRAY}[X, Y, N + 1]$      $\triangleright$  Initialization
16:   for each valid  $(i, j)$  do                 $\triangleright$  Every window location
17:      $l, v \leftarrow \text{RCH}(\mathbf{fm}[i : i + w_x, j : j + w_y])$      $\triangleright$  Classify
         $\triangleright$  Add classification logits
18:      $\mathbf{oim}[i : i + w_x, j : j + w_y] \leftarrow \mathbf{oim}[i : i + w_x, j : j + w_y] + v$ 
19:   end for
20:    $\mathbf{oim} \leftarrow \text{MAXOBJ}(\mathbf{oim}, \text{axis} = -1)$    $\triangleright$  Max objectness score
21:    $\mathbf{om} \leftarrow \text{BINARIZE}(\mathbf{oim}, T \cdot w_x \cdot w_y)$      $\triangleright$  Binarization
22:   return  $\mathbf{om}$ 
23: end procedure

24: procedure  $\text{OBJEXPLAINER}(\mathcal{D}, \mathbf{om})$ 
25:    $\mathbf{oim} \leftarrow \text{COPY}(\mathbf{om})$                   $\triangleright$  A copy of  $\mathbf{om}$ 
     $\triangleright$  Match each detected box to objectness map
26:   for  $i \in \{0, 1, \dots, |\mathcal{D}| - 1\}$  do
27:      $x_{\min}, y_{\min}, x_{\max}, y_{\max}, l \leftarrow \mathbf{b} \leftarrow \mathcal{D}[i]$ 
28:     if  $\text{SUM}(\mathbf{om}[x_{\min} : x_{\max}, y_{\min} : y_{\max}]) > 0$  then
29:        $\mathbf{oim}[x_{\min} : x_{\max}, y_{\min} : y_{\max}] \leftarrow 0$ 
30:     end if
31:   end for
32:   if  $\text{DETCLUSTER}(\mathbf{oim})$  is None then
33:     return False                                 $\triangleright$  All objectness explained
34:   else
35:     return True                                $\triangleright$  Unexplained objectness
36:   end if
37: end procedure

```

---

location (Line 20). This operation discards label information and fully eliminates *Clean Error 1*, e.g., confusion between bicycles and motorbikes. At last, we binarize every objectness score (Line 21): if the score is larger than  $T \cdot w_x \cdot w_y$ , we set it to one; otherwise, set it to zero. This binarization mitigates *Clean Error 2*, when the classifier incorrectly predicts background as objects but with low

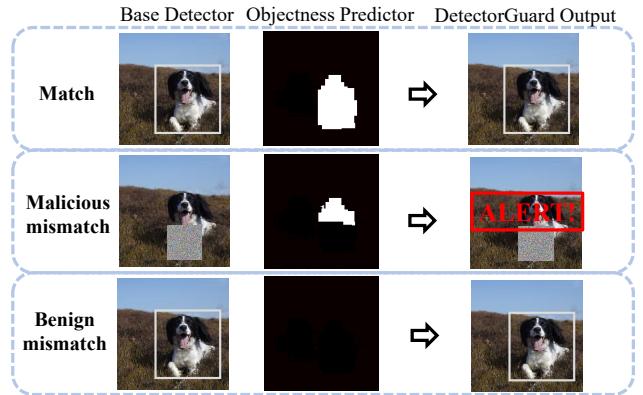


Figure 3: Visualization of explaining/matching rules

classification confidence. We discuss the strategy for *Clean Error 3* in the next subsection.

### 3.4 Objectness Explainer

Objectness Explainer takes as inputs the predicted bounding boxes of Base Detector and the generated objectness map of Objectness Predictor, and tries to use each predicted bounding box to explain/match the high activation in the objectness map. Its outcome determines the final prediction of DetectorGuard. We will first introduce the high-level explaining/matching rules and then elaborate on the algorithm.

**Explaining/matching rules.** There are three possible explaining/matching outcomes, each of them leading to a different prediction strategy (a visual example is in Figure 3):

- A **match** happens when Base Detector and Objectness Predictor both predict a bounding box or high objectness at a specific location. In this simplest case, *the objectness is well explained by the bounding box*; our defense will consider the detection as correct and output the accurate bounding box and the class label predicted by Base Detector.
- A **malicious mismatch** will be flagged when only Objectness Predictor outputs high objectness. This is most likely to happen when a hiding attack fools the conventional object detector to miss the object while our Objectness Predictor still makes robust predictions. In this case, our defense will find *unexplained objectness* and send out an attack alert.
- A **benign mismatch** occurs when only Base Detector detects the object and there is *no objectness to be explained*. This can happen when Objectness Predictor incorrectly misses the object due to its limitations (recall our remark in Section 3.3). In this case, we trust Base Detector and output its predicted bounding box. Notably, this strategy can fully eliminate *Clean Error 3*, i.e., predicting objects as background.<sup>8</sup>

Next, we discuss the concrete procedure for explaining objectness.

**Matching and explaining objectness.** In Line 25-31 of Algorithm 1, we use each predicted bounding box to match/explain the

<sup>8</sup>We note that this miss can also be caused by other attacks that are orthogonal to the focus of this paper, e.g., FP attacks that aim to introduce incorrect bounding box predictions. We will discuss such attacks and our defense strategies in Section 6.

objectness predicted at the same location. First, we create a copy of  $\mathbf{om}$  as  $\bar{\mathbf{om}}$  to hold the explaining results. Next, for each bounding box  $\mathbf{b}$ , we get its coordinates  $x_{\min}, y_{\min}, x_{\max}, y_{\max}$ , and calculate the sum of objectness scores within the same box on the objectness map  $\mathbf{om}$ . If the objectness sum is larger than zero, we assume that the bounding box  $\mathbf{b}$  agrees with  $\mathbf{om}$ , and we zero out the corresponding region in  $\bar{\mathbf{om}}$ , to indicate that this region of objectness has been *explained* by the detected bounding box. On the other hand, if all objectness scores are zeros, we assume it is a benign mismatch; the algorithm proceeds without alert.

**Detecting clusters of unexplained objectness.** The final step is to detect unexplained objectness in  $\bar{\mathbf{om}}$ . We use the sub-procedure  $\text{DETCLUSTER}(\cdot)$  to determine if any non-zero points in  $\bar{\mathbf{om}}$  form a large cluster. Specifically, we choose DBSCAN [12] as the cluster detection algorithm, which will assign each point to a certain cluster or label it as an outlier based on the point density in its neighborhood. If  $\text{DETCLUSTER}(\bar{\mathbf{om}})$  returns None, it means that no large cluster is found, or all objectness predicted by Objectness Predictor is explained by the bounding boxes predicted by Base Detector;  $\text{OBJEXPLAINER}(\cdot)$  then returns False (i.e., no attack detected). We note that this clustering operation further mitigates *Clean Error 2* when the robust classifier predicts background as objects at only a few scattered locations. On the other hand, receiving a non-empty cluster set indicates that there are clusters of unexplained objectness activations in  $\bar{\mathbf{om}}$  (i.e., Base Detector misses an object but Objectness Predictor predicts high objectness). Objectness Explainer regards this as a sign of patch hiding attacks and returns True.

**Final output.** Line 5-10 of Algorithm 1 demonstrates the strategy for the final prediction. If the alert flag  $a$  is True (i.e., a malicious mismatch is detected), **DetectorGuard** returns  $\mathcal{D}^* = \text{ALERT}$ . In other cases, **DetectorGuard** returns the detection  $\mathcal{D}^* = \mathcal{D}$ .

**Remark: Clean performance of DetectorGuard.** Recall that *Clean Error 1* of the robust classifier is *fully eliminated* in our objectness map generation via discarding label information; *Clean Error 2* is *mitigated* via binarizing (in Objectness Predictor) and clustering (in Objectness Explainer) operations; *Clean Error 3* is *fully tolerated* via our prediction matching strategy (the benign mismatch case). Therefore, we can safely optimize the setting of **DetectorGuard** to mitigate most of *Clean Error 2* (which can lead to unexplained objectness in the clean setting and trigger a false alert) so that we can achieve a clean performance that is comparable to state-of-the-art object detectors (performance difference smaller than 1%; more details are in Section 5). This helps us solve *Challenge 2: Amplified Cost of Clean Performance*. In the next section, we will demonstrate that our efforts in mitigating the imperfection of robust classifiers are worthwhile by showing how **DetectorGuard** addresses *Challenge 1: Lack of End-to-end Provable Robustness*.

## 4 END-TO-END PROVABLE ROBUSTNESS

Recall that we consider **DetectorGuard** to be provably robust for a given object (in a given image) when it can make correct detection on the clean image and will either detect part of the object or issue an alert on the adversarial image. The robustness property holds for *any adaptive patch hiding attacker at any location* within our threat model, including ones who have full access to our models

---

### Algorithm 2 Provable Analysis of DetectorGuard

---

**Input:** input image  $\mathbf{x}$ , window size  $(w_x, w_y)$ , matching threshold  $T$ , the set of patch locations  $\mathcal{P}$ , the object bounding box  $\mathbf{b}$ , feature extractor  $\text{FE}(\cdot)$ , provable analysis of the robust classification head  $\text{RCH-PA}(\cdot)$ , cluster detection procedure  $\text{DETCLUSTER}(\cdot)$

**Output:** whether the object  $\mathbf{b}$  in  $\mathbf{x}$  has provable robustness

```

1: procedure  $\text{DG-PA}(\mathbf{x}, w_x, w_y, T, \mathcal{P}, \mathbf{b})$ 
2:   if  $\mathbf{b} \notin \text{DG}(\mathbf{x}, w_x, w_y, T)$  then
3:     return False                                 $\triangleright$  Clean detection is incorrect
4:   end if
5:    $\mathbf{fm} \leftarrow \text{FE}(\mathbf{x})$                    $\triangleright$  Extract feature map
6:   for each  $\mathbf{p} \in \mathcal{P}$  do                 $\triangleright$  Check every patch location
7:      $x, y, p_x, p_y \leftarrow \mathbf{p}$ 
8:      $r \leftarrow \text{DG-PA-ONE}(\mathbf{fm}, x, y, w_x, w_y, p_x, p_y, \mathbf{b}, T)$ 
9:     if  $r == \text{False}$  then
10:      return False                             $\triangleright$  Possibly vulnerable
11:    end if
12:   end for
13:   return True                                 $\triangleright$  Provably robust
14: end procedure

15: procedure  $\text{DG-PA-ONE}(\mathbf{fm}, x, y, w_x, w_y, p_x, p_y, \mathbf{b}, T)$ 
16:    $X, Y, \_ \leftarrow \text{SHAPE}(\mathbf{fm})$            $\triangleright$  Get the shape of  $\mathbf{fm}$ 
17:    $\bar{\mathbf{om}}_* \leftarrow \text{ZEROARRAY}[X, Y, N + 1]$          $\triangleright$  Initialization
18:    $\triangleright$  Generates worst-case objectness map for analysis
19:   for each valid  $(i, j)$  do             $\triangleright$  Every window location
20:      $\mathbf{u}, \mathbf{l} \leftarrow \text{RCH-PA}(\mathbf{fm}[i : i + w_x, j : j + w_y], x - i, y - j, p_x, p_y)$ 
21:      $\triangleright$  Add worst-case (lower-bound) logits
22:      $\bar{\mathbf{om}}_*[i : i + w_x, j : j + w_y] \leftarrow \bar{\mathbf{om}}_*[i : i + w_x, j : j + w_y] + \mathbf{l}$ 
23:   end for
24:    $\bar{\mathbf{om}}_* \leftarrow \text{MAXOBJ}(\bar{\mathbf{om}}_*, \text{axis} = -1)$   $\triangleright$  Max objectness score
25:    $\mathbf{om}_* \leftarrow \text{BINARIZE}(\bar{\mathbf{om}}_*, T \cdot w_x \cdot w_y)$            $\triangleright$  Binarization
26:    $x_{\min}, y_{\min}, x_{\max}, y_{\max}, l \leftarrow \mathbf{b}$ 
27:   if  $\text{DETCLUSTER}(\mathbf{om}_*[x_{\min} : x_{\max}, y_{\min} : y_{\max}])$  is None then
28:     return False                             $\triangleright$  No high objectness left
29:   else
30:     return True                             $\triangleright$  High worst-case objectness
end if
end procedure

```

---

and defense setup. In this section, we will prove the end-to-end robustness of **DetectorGuard**, solving *Challenge 1*.

**Provable analysis of DetectorGuard.** Thanks to our objectness explaining strategy, a patch hiding attacker has to make *both* Base Detector and Objectness Predictor fail to predict a bounding box, or high objectness, for a successful attack. Therefore, if we can prove that Objectness Predictor can output high objectness for an object in the worst case, we can certify its provable robustness. We present the provable analysis of **DetectorGuard** in Algorithm 2. The algorithm takes a clean image  $\mathbf{x}$ , a ground-truth object bounding box  $\mathbf{b}$ ,<sup>9</sup> and a set of valid *patch locations*  $\mathcal{P}$  as inputs, and will

<sup>9</sup>We note that the ground-truth information is essential in our provable analysis (Algorithm 2) but is not used in our actual defense (Algorithm 1).

determine whether the object in bounding box  $\mathbf{b}$  in the image  $\mathbf{x}$  has provable robustness against any patch at any location in  $\mathcal{P}$ . We state the correctness of Algorithm 2 in Theorem 1, and will explain the algorithm details by proving the theorem.

**THEOREM 1.** *Given an object bounding box  $\mathbf{b}$  in a clean image  $\mathbf{x}$ , a set of patch locations  $\mathcal{P}$ , window size  $(w_x, w_y)$ , and binarizing threshold  $T$  (used in  $DG(\cdot)$  of Algorithm 1), if Algorithm 2 returns True, i.e.,  $DG\text{-PA}(\mathbf{x}, w_x, w_y, T, \mathbf{b}, \mathcal{P}) = \text{True}$ , DetectorGuard has provable robustness for the object  $\mathbf{b}$  against any patch hiding attack using any patch location in  $\mathcal{P}$ .*

**PROOF.**  $DG\text{-PA}(\cdot)$  first calls  $DG(\cdot)$  of Algorithm 1 to determine if DetectorGuard can detect the object bounding box  $\mathbf{b}$  on the clean image  $\mathbf{x}$ . The algorithm will proceed only when the clean detection is correct (Line 2-4).

Next, we extract feature map  $\mathbf{fm}$ , iterate over each patch location in  $\mathcal{P}$ , and call the sub-procedure  $DG\text{-PA-ONE}(\cdot)$ , which analyzes worst-case behavior over all possible adversarial strategies for each patch location, to determine the model robustness. If any call of  $DG\text{-PA-ONE}(\cdot)$  returns False, the algorithm returns False, indicating that at least one patch location can bypass our defense. On the other hand, if the algorithm tries all valid patch locations and does not return False, this means that DetectorGuard is provably robust to all patch locations in  $\mathcal{P}$  and the algorithm returns True.

In the sub-procedure  $DG\text{-PA-ONE}(\cdot)$ , we analyze the worst-case output of Objectness Predictor against the given patch location. We perform the provable analysis of the robust image classification head (using  $RCH\text{-PA}(\cdot)$ ) to determine the lower/upper bounds of classification logits for each window. If the aggregated worst-case (i.e., lower bound) objectness map still has high activation for the object of interest, we can certify the robustness of DetectorGuard.

As shown in the  $DG\text{-PA-ONE}(\cdot)$  pseudocode, we first initialize a zero array  $\bar{\mathbf{om}}_*$  to hold the worst-case objectness scores. We then iterate over each sliding window and call  $RCH\text{-PA}(\cdot)$ , which takes the feature map window  $\mathbf{fm}[i : i + w_x, j : j + w_y]$ , relative patch coordinates  $(x - i, y - j)$ , patch size  $(p_x, p_y)$  as inputs and outputs the upper bound  $\mathbf{u}$  and lower bound  $\mathbf{l}$  of the classification logits. Since the goal of the hiding attack is to minimize the objectness scores, we only need to reason about the lower bound of classification logits. Recall that in  $RCH(\cdot)$ , we clip all local logits values into  $[0, \infty]$ ; therefore, the best an adversary can do is to push all corrupted logits values down to zeros. We then compute the lower bound  $\mathbf{l}$  by zeroing out all corrupted logits values and aggregating the remaining ones. We note that the sub-procedure  $DG\text{-PA-ONE}(\cdot)$  aims to check defense robustness for a particular patch location; therefore, the patch location and corrupted features/logits are known in this provable analysis (we discuss how to map pixel-space coordinates to feature-space coordinates in Appendix G).

Given the lower bound  $\mathbf{l}$  of every window classification logits, we will add it to the corresponding region of  $\bar{\mathbf{om}}_*$ . After we analyze all valid windows, we call  $\text{MAXOBJ}(\cdot)$  and  $\text{BINARIZE}(\cdot)$  for the worst-case objectness map  $\mathbf{om}_*$ . We then get the cropped objectness map that corresponds to the object of interest (i.e.,  $\mathbf{om}_*[x_{\min} : x_{\max}, y_{\min} : y_{\max}]$ ) and feed it to the cluster detection algorithm  $\text{DETCLUTSER}(\cdot)$ . If None is returned, a hiding attack using this patch location might succeed, and the sub-procedure returns False. Otherwise, Objectness Predictor has a high worst-case objectness and

is thus robust to any attack using this patch location. This implies the provable robustness, and the sub-procedure returns True.  $\square$

Theorem 1 shows that if our provable analysis (Algorithm 2) returns True for certain objects, DetectorGuard (Algorithm 1) will always detect the certified object or issue an attack alert. This robustness property is agnostic to attack strategies and holds for any adaptive attacker at any location within our threat model. In our evaluation (next section), we will use Algorithm 2 and Theorem 1 to certify the provable robustness of every object in every image and report the percentage of certified objects.

## 5 EVALUATION

In this section, we provide a comprehensive evaluation of DetectorGuard on PASCAL VOC [13], MS COCO [26], and KITTI [15] datasets. We will first introduce the datasets and models used in our evaluation, followed by our evaluation metrics. We then report our main evaluation results on different models and datasets, and finally provide a detailed analysis of DetectorGuard performance. Our code is available at <https://github.com/inspire-group/DetectorGuard>.

### 5.1 Dataset and Model

#### Dataset:

**PASCAL VOC** [13]. The detection challenge of PASCAL Visual Object Classes (VOC) project is a popular object detection dataset with annotations for 20 different classes. We take `trainval2007` (5k images) and `trainval2012` (11k images) as our training set and evaluate our defense on `test2007` (5k images), which is a conventional usage of the PASCAL VOC dataset [29, 64].

**MS COCO** [26]. The Microsoft Common Objects in COnText (COCO) dataset is an extremely challenging object detection dataset with 80 annotated common object categories. We use the training and validation set of COCO2017 for our experiments. The training set has 117k images, and the validation set has 5k images. We ignore bounding boxes with the flag `iscrowd=1` for simplicity.

**KITTI** [15]. KITTI is an autonomous vehicle dataset that contains both 2D camera images and 3D point clouds. We take its 7481 2D images and use 80% of randomly split images for training and the remaining 20% for validation. We merge all classes into three classes: car (all different classes of vehicles), pedestrian, cyclist.

#### Base Detector:

**YOLOv4** [2, 53] is the state-of-the-art one-stage object detector. We choose Scaled-YOLOv4-P5 [53] in our evaluation. For MS COCO, we use the pre-trained model. For PASCAL VOC and KITTI, we fine-tune the model previously trained on MS COCO.

**Faster R-CNN** [45] is a representative two-stage object detector. We use ResNet101-FPN as its backbone network. Image pre-processing and model architecture follows the original paper. We use pre-trained models for MS COCO and fine-tune pre-trained models for PASCAL VOC and KITTI detectors.

**Perfect Clean Detector** (PCD) is a hypothetical object detector simulated with ground-truth annotations. PCD can always make correct detection in the clean setting but is assumed vulnerable to patch hiding attacks. This hypothetical object detector ablates the errors of Base Detector and helps us better understand the behavior of Objectness Predictor and Objectness Explainer.

#### Objectness Predictor:

**Table 2: Clean performance of DetectorGuard**

	PASCAL VOC			MS COCO			KITTI		
	vanilla AP	defense AP	FAR@0.8	vanilla AP	defense AP	FAR@0.6	vanilla AP	defense AP	FAR@0.8
Perfect clean detector	100%	99.3%	0.9%	100%	99.0%	1.2%	100%	99.0%	1.5%
YOLOv4	92.9%	92.4%	4.0%	73.6%	73.4%	1.6%	93.1%	92.4%	1.7%
Faster R-CNN	90.0%	89.6%	2.9%	66.7%	66.5%	0.9%	89.9%	89.1%	1.4%

BagNet-33 [3], which has a  $33 \times 33$  small receptive field, is the backbone network of Objectness Predictor. For PASCAL VOC and MS COCO, we zero-pad each image to a square and resize it to  $416 \times 416$  before feeding it to BagNet; for KITTI, we resize each image to  $224 \times 740$ . We fine-tune a BagNet model that is pre-trained on ImageNet [11]. To train an image classifier given a list of bounding boxes in the object detection dataset, we first map pixel-space bounding boxes to the feature space (details of box mapping are in Appendix G). We then teach BagNet to make correct predictions on cropped feature maps by minimizing the cross-entropy loss between aggregated feature predictions and one-hot encoded label vectors. In addition, we aggregate all features outside any feature boxes as the “negative” feature vector for the “background” classification.

#### Default Hyper-parameter:

We will analyze the effect of different hyper-parameters in Section 5.5. In our default setting, we use a square *feature-space window* of size 8 and the DBSCAN clustering [12] with  $\text{eps} = 3$ ,  $\text{min\_points} = 24$  for different datasets.<sup>10</sup> We set the default binarizing threshold to 32 for PASCAL VOC, 36 for MS COCO, and 11 for KITTI based on different model properties with different datasets.

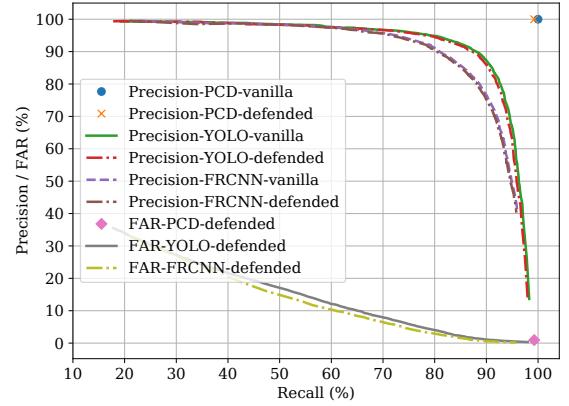
## 5.2 Metric

### Clean Performance Metric:

Precision and Recall. We calculate the precision as  $\text{TP}/(\text{TP}+\text{FP})$  and the recall as  $\text{TP}/(\text{TP}+\text{FN})$ . For the clean images without a false alert, we follow previous works [8, 64] setting the IoU threshold  $\tau = 0.5$  and count TPs, FPs, FNs in the conventional manner. For images that have false alerts, we set TP and FP to zeros, and FN to the number of ground-truth objects since no bounding box is predicted. We note that conventional object detectors use a confidence threshold to filter out bounding boxes with low confidence values. As a result, different confidence thresholds will give different precision and recall values; we will plot the entire precision-recall curve to analyze the model performance.

Average Precision (AP). To remove the dependence on the confidence threshold and to have a global view of model performance, we also report Average Precision (AP) as one of evaluation metrics. We vary the confidence threshold from 0 to 1, record the precision and recall at different thresholds, and calculate AP as the averaged precision at different recall levels. This calculated AP can be considered as an approximation of the AUC (area under the curve) for the precision-recall curve. We note that AP is one of the most widely used performance metrics in object detection benchmark competitions [13, 15, 26] and research papers [2, 19, 25, 29, 43–45, 49].

<sup>10</sup>A  $416 \times 416$  (or  $224 \times 740$ ) pixel image results in a  $48 \times 48$  (or  $24 \times 89$ ) feature map using our BagNet-33 implementation.



**Figure 4: Clean performance of DetectorGuard on PASCAL VOC (PCD – perfect clean detector; YOLO – YOLOv4; FRCNN – Faster R-CNN; FAR – False Alert Rate)**

False Alert Rate (FAR@0.x). FAR is defined as the percentage of clean *images* on which DetectorGuard will trigger a false alert. The false alert is mainly caused by *Clean Error 2* as discussed in Section 3. We note that FAR is also closely tied to the confidence threshold of Base Detector: a higher confidence threshold leads to fewer predicted bounding boxes, leading to more unexplained objectness, and finally higher FAR. We will report FAR at different recall levels for a global evaluation, and use FAR@0.x to denote FAR at a *clean recall* of 0.x.

### Provable Robustness Metric:

Certified Recall (CR@0.x). We use certified recall as the robustness metric against patch hiding attacks. The certified recall is defined as the percentage of ground-truth *objects* that have provable robustness against any patch hiding attack. Recall that an object has provable robustness when Algorithm 2 (our provable analysis) returns True. Note that CR is also affected by the performance of Base Detector (e.g., confidence threshold) since its prerequisite is the correct clean detection. We use CR@0.x to denote the certified recall at a *clean recall* of 0.x.

## 5.3 Clean Performance

In this subsection, we evaluate the clean performance of DetectorGuard with three different base object detectors and three datasets. In Table 2, we report AP of vanilla Base Detector (vanilla AP), AP of DetectorGuard (defense AP), and False Alert Rate at a clean recall of 0.8 or 0.6 (FAR@0.8 or FAR@0.6). We also plot the precision-recall

**Table 3: Provable robustness of DetectorGuard**

	PASCAL VOC (CR@0.8)			MS COCO (CR@0.6)			KITTI (CR@0.8)		
	far-patch	close-patch	over-patch	far-patch	close-patch	over-patch	far-patch	close-patch	over-patch
Perfect clean detector	28.6%	20.7%	8.3%	11.5%	7.0%	2.2%	32.0%	11.1%	2.1%
YOLOv4	24.6%	18.6%	7.5%	10.1%	6.5%	1.9%	30.6%	10.9%	2.1%
Faster R-CNN	25.7%	19.4%	8.0%	10.7%	6.8%	2.0%	31.6%	11.0%	2.1%

and FAR-recall curves for PASCAL VOC in Figure 4; similar plots for MS COCO and KITTI are in Appendix D.

**DetectorGuard has a low FAR and a high AP.** We can see from Table 2 that DetectorGuard has a low FAR of 0.9% and a high AP of 99.3% on PASCAL VOC when we use a perfect clean detector as Base Detector. The result shows that DetectorGuard only has a minimal impact on the clean performance.

**DetectorGuard is highly compatible with different conventional object detectors.** From Table 2 and Figure 4, we can see that when we use YOLOv4 or Faster R-CNN as Base Detector on PASCAL VOC, the clean AP, as well as the precision-recall curve of DetectorGuard, is close to that of its vanilla Base Detector. These results show that DetectorGuard is highly compatible with different conventional object detectors.

**DetectorGuard works well across different datasets.** We can see that the observation of high clean performance is similar across three different datasets: DetectorGuard achieves a low FAR and a similar AP as the vanilla Base Detector on PASCAL VOC, MS COCO, and KITTI (the precision-recall plots for MS COCO and KITTI are available in Appendix D). These similar results show that DetectorGuard is a general approach and can be used for both easier and more challenging detection tasks.

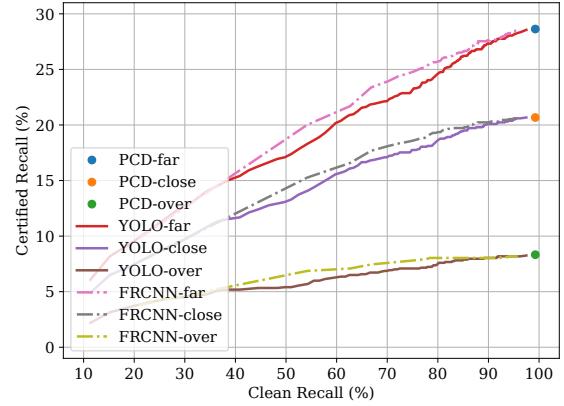
**Remark: a negligible cost of clean performance.** In this subsection, we have shown that DetectorGuard only incurs a negligible cost of clean performance (<1% AP drop). This slight clean performance drop is worthwhile given the first provable robustness against patch hiding attacks (evaluated in the next subsection).

#### 5.4 Provable Robustness

In this subsection, we first introduce the robustness evaluation setup and then report the provable robustness of our defense against any patch hiding attack within our threat model.

**Setup.** We use a 32×32 adversarial pixel patch on the re-scaled and padded 416×416 (or 224×740) images to evaluate the provable robustness.<sup>11</sup> We consider *all possible image locations* as candidate locations for the adversarial patch to evaluate the model robustness. We categorize our results into three categories depending on the distance between an object and the patch location. When the patch is totally over the object, we consider it as *over-patch*. When the feature-space distance between patch and object is smaller than 8, we consider it as *close-patch*. The other patch locations are considered as *far-patch*. For each set of patch locations and each object, we use Algorithm 2 to determine the robustness. We note that the

<sup>11</sup>DPatch [30] demonstrates that even a 20×20 adversarial patch at the image corner can have a malicious effect. In Appendix A, we show that more than 15% of PASCAL VOC objects and 44% of MS COCO objects are smaller than a 32×32 patch. We also provide robustness results for different patch sizes as well as visualizations in Appendix A.



**Figure 5: Provable robustness of DetectorGuard on PASCAL VOC (PCD – Perfect Clean Detector; YOLO – YOLOv4; FRCNN – Faster R-CNN)**

above algorithm already considers *all possible adaptive attacks* (attacker strategies) at any location within our threat model. We use CR@0.x as the robustness metric. Given the large number of all possible patch locations, we only use a 500-image subset of the test/validation datasets for evaluation.

**DetectorGuard achieves the first non-trivial provable robustness against patch hiding attack.** We report the certified recall at a clean recall of 0.8 or 0.6 (CR@0.8 or CR@0.6) in Table 3. As shown in Table 3, when we use a Perfect Clean Detector, DetectorGuard can certify the robustness for 28.6% of PASCAL VOC objects when the patch is far away from the object; which means *no attacker* within our threat model can successfully attack these certified objects. We also plot the CR-recall curves for different detectors on the PASCAL VOC dataset in Figure 5 (similar plots for MS COCO and KITTI are in Appendix D). The figure shows that the provable robustness improves as the clean recall increases, and the performance of YOLOv4 and Faster R-CNN is close to that of a perfect clean detector when the recall is close to one.

**DetectorGuard is especially effective when the patch is far away from the objects.** From Table 3 and Figure 5, we can clearly see that the provable robustness of DetectorGuard is especially good when the patch gets far away from the object. This model behavior aligns with our intuition that a localized adversarial patch should only have a spatially constrained adversarial effect. Moreover, this observation shows that DetectorGuard has made the attack much more difficult: to have a chance to bypass DetectorGuard, the adversary has to put the patch close to or even over the victim object, which is not always feasible in real-world scenarios. We also note

that in the over-patch threat model, we allow the patch to be anywhere over the object. This means that the patch can be placed over the most salient part of the object (e.g., the face of a person), and makes robust detection extremely difficult.

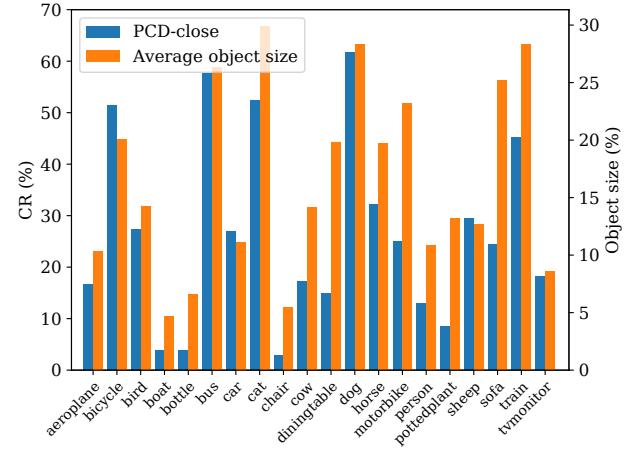
**DetectorGuard has better robustness for object classes with larger object sizes.** In Figure 6, we plot the certified recalls against the close-patch attacker (similar plots for far-patch and over-patch and for other datasets are in Appendix D) and average object sizes (reported as the percentage of image pixels) for all 20 classes of PASCAL VOC. As shown in the figure, the provable robustness of DetectorGuard is highly correlated with the object size: we have higher certified recalls for classes with larger object sizes. This is an expected behavior because it is hard for even humans to perfectly detect all small objects. Moreover, considering that missing a big nearby object is much more serious than missing a small distant object in real-world applications, we believe that DetectorGuard has strong application potential.

**Remark on absolute values of certified recalls.** Despite the first achieved provable robustness against patch hiding attacks, we acknowledge that DetectorGuard’s absolute values of certified recalls are still limited. First, the notion of the certified recall itself is strong and conservative: we certify the robustness of an object only when no patch at any location within the threat model can succeed in the hiding attack (using any attack strategy including adaptive attacks). In practice, the attacker capability might be limited to a small number of patch locations. Second, we note that most objects in our three datasets are small (or even tiny) objects (we provide a quantitative analysis in Appendix A). Detecting small objects is already challenging in the clean setting and becomes even more difficult when an adversarial patch of comparable sizes is used. However, it is still notable that DetectorGuard is the first to achieve non-trivial and strong provable robustness against the patch hiding attacker for “free”, and that our approach works particularly well on certain object classes and threat models (e.g., the class “dog” in the PASCAL VOC dataset has a ~60% certified recall for the close-patch threat model). We hope that the security community can build upon our work and further push the certified recall to a higher value.

## 5.5 Detailed Analysis of DetectorGuard

In this subsection, we first perform a runtime analysis of DetectorGuard to show its small defense overhead. Next, we use a hypothetical perfect clean detector (PCD) and the PASCAL VOC dataset to analyze the performance of DetectorGuard under different hyper-parameter settings. Note that using PCD helps us to focus on the behavior of Objectness Predictor and Objectness Explainer.

**Runtime analysis.** In Table 4, we report the average per-example runtime of each module in DetectorGuard. For Base Detector, we report runtime for YOLOv4 (left) and Faster R-CNN (right). As shown in the table, Objectness Predictor has a similar runtime as Base Detector (or vanilla undefended object detectors), and Objectness Explainer only introduces a negligible overhead. If 2 GPUs are available, then Base Detector and Objectness Predictor can run in parallel in DetectorGuard, and the overall runtime of DetectorGuard can be calculated as  $t_{\text{DetectorGuard}} = \max(t_{\text{base}}, t_{\text{predictor}}) + t_{\text{explainer}}$  (reported in the last column of Table 4), which is close to  $t_{\text{base}}$ . Thus, DetectorGuard has a similar runtime performance as conventional



**Figure 6: Certified recalls and average object sizes for every class in PASCAL VOC (reporting CR for close-patch; results for far-patch and over-patch are in Appendix D)**

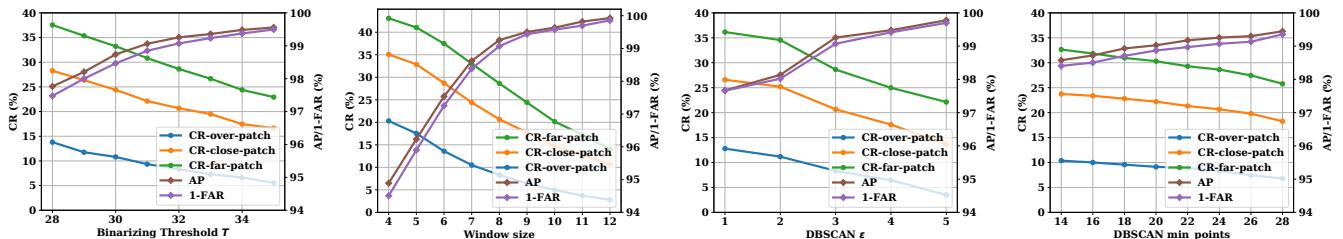
**Table 4: Per-example runtime breakdown**

	Base Detector (YOLO / FRCNN)	Objectness Predictor	Objectness Explainer	DetectorGuard (YOLO / FRCNN)
VOC	54.0ms / 80.9ms	48.5ms	0.2ms	54.2ms / 81.1ms
COCO	55.2ms / 79.4ms	65.2ms	0.3ms	65.5ms / 79.7ms
KITTI	55.8ms / 69.7ms	44.6ms	0.4ms	56.2ms / 70.1ms

object detectors in the setting of 2 GPUs. If only a single GPU is available, the runtime can be calculated as  $t_{\text{base}} + t_{\text{predictor}} + t_{\text{explainer}}$ , which leads to a roughly 2x slow-down.

**Effect of the binarizing threshold.** We first vary the binarizing threshold  $T$  in  $\text{OBJPREDICTOR}(\cdot)$  to see how the model performance changes. For each threshold, we report CR for three patch threat models. We also include AP and 1-FAR to understand the effect of different thresholds on clean performance. We report these results in the leftmost sub-figure in Figure 7. We can see that when the binarizing threshold is low, the CR is high because more objectness is retained after the binarization. However, more objectness also makes it more likely to trigger a false alert in the clean setting, and we can see both AP and 1-FAR are affected greatly as we decrease the threshold  $T$ . Therefore, we need to balance the trade-off between clean performance and provable robustness. In our default parameter setting, we set  $T = 32$  to have a FAR lower than 1% while maintaining decent provable robustness.

**Effect of window size.** We study the effect of different window sizes in the second sub-figure in Figure 7. The figure demonstrates a similar trade-off between provable robustness and clean performance. As we increase the window size, each window receives more information from the input and therefore the clean performance (AP and 1-FAR) improves. However, a large window size increases the number of windows that are affected by the small adversarial patch, and the provable robustness drops. In our default setting, we set the window size to 8 to have a low FAR and good CR.



**Figure 7: Effect of different hyper-parameters** (left to right: binarizing threshold, window size, DBSCAN  $\epsilon$ , DBSCAN  $\text{min\_points}$ )

**Effect of DBSCAN parameters.** We also analyze the effect of DBSCAN parameters in  $\text{DETCLUSTER}(\cdot)$ . DBSCAN has two parameters  $\epsilon$  and  $\text{min\_points}$ . A point is labeled as a core point when there are at least  $\text{min\_points}$  points within distance  $\epsilon$  of it; all core points and their neighbors will be labeled as clusters. We plot the effect of  $\epsilon$  and  $\text{min\_points}$  in the right two sub-figures in Figure 7. As we increase  $\epsilon$  or  $\text{min\_points}$ , it becomes more difficult to form clusters. As a result, the clean performance improves because of fewer detected clusters and fewer false alerts. However, the provable robustness (CR) drops due to fewer detected clusters in the worst-case objectness map.

## 6 DISCUSSION

In this section, we discuss the future work directions and defense extension of DetectorGuard.

### 6.1 Future Work

**Robust object detection without abstention.** In this paper, we have tailored DetectorGuard for attack detection: when no attack is detected, the model uses conventional object detectors for predictions; when an attack is detected, the model alerts and abstains from making predictions. This type of defense is useful in application scenarios like autonomous vehicles which can give the control back to the driver upon detecting an attack. However, the most desirable notion of robustness is to always make correct predictions without any abstention. How to extend DetectorGuard for robust object detection without abstention is an interesting direction of future work.

**Better robust image classifier.** In DetectorGuard, we use the key principals introduced in Section 2.4 to co-design the provably robust image classifier and Objectness Predictor. However, the imperfection of the adapted robust image classifier still limits the performance of DetectorGuard detector. Although we can optimize for few *Clean Error 2* and tolerate a potentially high *Clean Error 3*, as discussed in Section 3.4, a high *Clean Error 3* results in a limited certified recall of DetectorGuard in the adversarial setting. We note that DetectorGuard is a general framework that is *compatible with any conventional object detector and (principles for building) provably robust image classifier*, and we expect a boost in DetectorGuard’s performance given any future advance in robust image classification research.

**Extension to the video setting.** In this paper, we focus on object detection in the single-frame setting. It is interesting to

---

### Algorithm 3 Auxiliary Predictor of DetectorGuard

---

```

1: procedure AUXPREDICTOR( $\mathcal{D}, \mathbf{x}$ )
2:    $\hat{\mathcal{D}} \leftarrow \{\}$ 
3:   for  $i \in \{0, 1, \dots, \text{num\_detection} - 1\}$  do
4:      $x_{\min}, y_{\min}, x_{\max}, y_{\max}, l \leftarrow \mathbf{b} \leftarrow \mathcal{D}[i]$ 
5:      $l' \leftarrow \text{AUXCLASSIER}(\mathbf{x}[x_{\min} : x_{\max}, y_{\min} : y_{\max}])$ 
6:     if  $l' \neq \text{"background"}$  then
7:        $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{(x_{\min}, y_{\min}, x_{\max}, y_{\max}, l')\}$ 
8:     end if
9:   end for
10:  return  $\hat{\mathcal{D}}$ 
11: end procedure

```

---

extend our DetectorGuard to multiple-frame video setting. We expect that the temporal information could be useful for robustness. Moreover, we could perform the defense on a subset of frames to reduce defense overhead and minimize false alerts in the clean setting.

### 6.2 Defense Extension.

In this paper, we propose DetectorGuard as a provably robust defense against *patch hiding, or false-negative (FN), attacks*. Here, we discuss how to extend DetectorGuard for defending against *false-positive (FP)* attacks. The FP attack aims to introduce incorrect bounding boxes in the predictions of detectors to increase FP. We can consider FP attacks as a misclassification problem (i.e., a bounding box is given an incorrect label), and thus this attack can be mitigated if we have a robust auxiliary image classifier to re-classify the detected bounding boxes. If the auxiliary classifier predicts a different label, we consider it as an FP attack and can easily correct or filter out the FP boxes.

We provide the pseudocode for using an auxiliary classifier (can be any robust image classifier) against FP attacks in Algorithm 3. The algorithm re-classifies each detected bounding box in  $\mathcal{D}$  as label  $l'$  (Line 5). We trust the robust auxiliary classifier and add bounding boxes with non-background labels to  $\hat{\mathcal{D}}$  (Line 7). Finally, the algorithm returns the filtered detection  $\hat{\mathcal{D}}$ , and we can replace the  $\mathcal{D}$  in Line 8 of Algorithm 1 with  $\hat{\mathcal{D}}$  to extend the original DetectorGuard design. We note that when the patch is not present in the FP box or it only occupies a small portion of the FP box (the FP box is large), the auxiliary classifier is likely to correctly predict “background” since there is no or few corrupted pixel within the

FP box. When the patch occupies a large portion of the FP box, we might finally output a bounding box that has a large IoU with the patch. In this case, our object detector correctly locates the adversarial patch but predicts a wrong label. This is an acceptable outcome because the class label for patches is undefined. Therefore, Algorithm 3 presents a strong empirical defense for FP attacks.

## 7 RELATED WORK

### 7.1 Adversarial Patch Attacks

**Image Classification.** Unlike most adversarial examples that introduce a global perturbation with a  $L_p$ -norm constraint, localized adversarial patch attacks allow the attacker to introduce arbitrary perturbations within a restricted region. Brown et al. [4] introduced the first adversarial patch attack against image classifiers. They successfully realized a real-world attack by attaching a patch to the victim object. Follow-up papers studied variants of localized attacks against image classifiers with different threat models [21, 27, 28].

**Object Detection.** Localized patch attacks against object detection have also received much attention in the past few years. Liu et al. [30] proposed DPatch as the first patch attack against object detectors in the digital domain. Lu et al. [31], Chen et al. [7], Eykholt et al. [14], and Zhao et al. [66] proposed different physical attacks against object detectors for traffic sign recognition. Thys et al. [50] proposed to use a rigid physical patch to evade human detection while Xu et al. [61] and Wu et al. [57] generated successfully non-rigid perturbations on T-shirt to evade detection.

### 7.2 Defenses against Adversarial Patches

**Image Classification.** Digital Watermark (DW) [18] and Local Gradient Smoothing (LGS) [38] were the first two heuristic defenses against adversarial patch attacks. Unfortunately, these defenses are vulnerable to an adaptive attacker with the knowledge of the defense. A few certified defenses [9, 24, 33, 36, 58, 59, 65] have been proposed to provide strong provable robustness guarantee against any adaptive attacker. Notably, PatchGuard [58] introduces two key principles of small receptive fields and secure aggregation and achieves state-of-the-art defense performance for image classification. In contrast, DetectorGuard aims to adapt robust image classifiers for the more challenging robust object detection task.

**Object Detection.** How to secure object detection is a much less studied area due to the complexity of this task. Saha et al. [46] demonstrated that YOLOv2 [43] were vulnerable to adversarial patches because detectors were using spatial context for their predictions, and then proposed a new training loss to limit the usage of context information. To the best of our knowledge, this is the only prior attempt to secure object detectors from patch attacks. However, this defense is based on heuristics and thus does not have any provable robustness. Moreover, the attack and defense are targeted at YOLOv2 only, and it is unclear if the defense generalizes to other detectors. In contrast, our defense has provable robustness against any patch hiding attack considered in our threat model and is compatible with any state-of-the-art object detectors.

### 7.3 Other Adversarial Example Attacks and Defenses

**Image Classification.** Attacks and defenses for classic  $L_p$ -bounded adversarial examples [6, 16, 48] have been extensively studied. Many empirical defenses [32, 34, 35, 40, 62] were proposed to mitigate the threat of adversarial examples, but were later found vulnerable to adaptive attackers [1, 5, 51]. The fragility of the empirical defenses has inspired certified defenses that are robust to any attacker considered in the threat model [10, 17, 22, 37, 41, 47, 56]. We refer interested readers to survey papers [39, 63] for more details.

**Object Detection.** Global perturbations against object detectors were first studied by Xie et al. [60] and followed by researchers [54, 55] in different applications. Defenses against global  $L_p$  perturbations are also very challenging. Zhang et al. [64] used adversarial training (AT) to improve empirical model robustness while Chiang et al. [8] proposed the use of randomized median smoothing (RMS) for building certifiably robust object detectors. Both defenses suffer from poor clean performance while DetectorGuard’s clean performance is close to state-of-the-art object detectors. On PASCAL VOC, AT incurs a ~26% clean AP drop while DetectorGuard only incurs a <1% drop.<sup>12</sup> On MS COCO, both AT and RMS have a clean AP drop that is larger than 10% while ours is smaller than 1%. We note that we do not compare robustness performance because these two works focus on global perturbations and are orthogonal to the objective of this paper.

We note that it is possible to extend our robust objectness predictor design and objectness explaining strategy to mitigate attacks that use global perturbations with a bounded  $L_\infty$  norm (if we have a robust image classifier against  $L_\infty$  perturbations). We leave this as a future work direction.

## 8 CONCLUSION

In this paper, we propose DetectorGuard, the first general framework for building provably robust object detectors against patch hiding attacks. DetectorGuard introduces a general approach to adapt robust image classifiers for robust object detection using an objectness explaining strategy. Our evaluation on the PASCAL VOC, MS COCO, and KITTI datasets demonstrates that DetectorGuard achieves the first provable robustness against any patch hiding attacker within the threat model and also has a high clean performance that is close to state-of-the-art detectors.

## ACKNOWLEDGMENTS

We are grateful to Gagandeep Singh for shepherding the paper and anonymous reviewers at CCS 2021 for their valuable feedback. We would also like to thank Vikash Sehwag, Shawn Shan, Sihui Dai, Alexander Valtchanov, Ruiheng Chang, Jiachen Sun, and researchers at Intel Labs for helpful discussions on the project and insightful comments on the paper draft. This work was supported in part by the National Science Foundation under grants CNS-1553437 and CNS-1704105, the ARL’s Army Artificial Intelligence Innovation Institute (A2I2), the Office of Naval Research Young Investigator Award, Schmidt DataX award, and Princeton E-flliates Award.

<sup>12</sup>RMS [8] did not report results for PASCAL VOC.

## REFERENCES

- [1] Anish Athalye, Nicholas Carlini, and David A. Wagner. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 274–283.
- [2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934* (2020).
- [3] Wieland Brendel and Matthias Bethge. 2019. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *7th International Conference on Learning Representations (ICLR)*.
- [4] Tom B. Brown, Dandelion Mané, Aurko Roy, Martin Abadi, and Justin Gilmer. 2017. Adversarial patch. In *Advances in neural information processing systems workshops (NeurIPS Workshops)*.
- [5] Nicholas Carlini and David A. Wagner. 2017. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISeC@CCS)*. 3–14.
- [6] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (S&P)*. 39–57.
- [7] 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.
- [8] Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. 2020. Detection as Regression: Certified Object Detection with Median Smoothing. In *Advances in Neural Information Processing Systems (NeurIPS) 2020*, Vol. 33.
- [9] Ping-Yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studor, and Tom Goldstein. 2020. Certified defenses for adversarial patches. In *8th International Conference on Learning Representations (ICLR)*.
- [10] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 1310–1320.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.
- [12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Vol. 96. 226–231.
- [13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338.
- [14] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Physical adversarial examples for object detectors. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*.
- [15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.
- [16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations (ICLR)*.
- [17] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Arthur Mann, and Pushmeet Kohli. 2019. Scalable Verified Training for Provably Robust Image Classification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4841–4850.
- [18] Jamie Hayes. 2018. On Visible Adversarial Perturbations & Digital Watermarking. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. 1597–1604.
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, (ICCV 2017)*. IEEE Computer Society, 2980–2988.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [21] Danny Karmon, Daniel Zoran, and Yoav Goldberg. 2018. LaVAN: Localized and Visible Adversarial Noise. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2512–2520.
- [22] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (S&P)*. 656–672.
- [23] Mark Lee and Zico Kolter. 2019. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897* (2019).
- [24] Alexander Levine and Soheil Feizi. 2020. (De)randomized Smoothing for Certifiable Defense against Patch Attacks. *arXiv preprint arXiv:2002.10733* (2020).
- [25] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision, (ICCV) 2017*. IEEE Computer Society, 2999–3007.
- [26] Tsung-Yi Lin, Michael Maire, Serge J. 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 (ECCV)*, Vol. 8693. Springer, 740–755.
- [27] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. 2019. Perceptual-Sensitive GAN for Generating Adversarial Patches. In *The 33rd AAAI Conference on Artificial Intelligence, (AAAI) 2019*. AAAI Press, 1028–1035.
- [28] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. 2020. Bias-Based Universal Adversarial Patch Attack for Automatic Check-Out. In *European conference on computer vision (ECCV)*, Vol. 12358. Springer, 395–410.
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *European conference on computer vision (ECCV)*, Vol. 9905. Springer, 21–37.
- [30] Xin Liu, Huanru Yang, Ziwei Liu, Linghao Song, Yiran Chen, and Hai Li. 2019. DPATCH: An Adversarial Patch Attack on Object Detectors. In *AAAI Conference on Artificial Intelligence Workshop (AAAI workshop) 2019*, Vol. 2301.
- [31] Jiajun Lu, Hussein Sibai, and Evan Fabry. 2017. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494* (2017).
- [32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations (ICLR)*.
- [33] Michael McCoyd, Won Park, Steven Chen, Neil Shah, Ryan Roggenkemper, Min-june Hwang, Jason Xinyu Liu, and David Wagner. 2020. Minority Reports Defense: Defending Against Adversarial Patches. *arXiv preprint arXiv:2004.13799* (2020).
- [34] Dongyu Meng and Hao Chen. 2017. MagNet: A Two-Pronged Defense against Adversarial Examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 135–147.
- [35] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017. On Detecting Adversarial Perturbations. In *5th International Conference on Learning Representations (ICLR)*.
- [36] Jan Hendrik Metzen and Maksym Yatsura. 2021. Efficient Certified Defenses Against Patch Attacks on Image Classifiers. In *9th International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=hr-3PMvDpil>
- [37] Matthew Mirman, Timon Gehr, and Martin T. Vechev. 2018. Differentiable Abstract Interpretation for Provably Robust Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 3575–3583.
- [38] Muzammal Naseer, Salman Khan, and Fatih Porikli. 2019. Local Gradients Smoothing: Defense Against Localized Adversarial Attacks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1300–1307.
- [39] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. 399–414.
- [40] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*. 582–597.
- [41] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified Defenses against Adversarial Examples. In *6th International Conference on Learning Representations (ICLR)*.
- [42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
- [43] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.
- [44] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).
- [45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
- [46] Aniruddha Saha, Akshayvarun Subramanya, Konnika 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 (CVPR Workshops)*. 784–785.
- [47] Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Penghuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. 2019. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. In *Annual Conference on Neural Information Processing Systems 2019 (NeurIPS)*. 11289–11300.
- [48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations (ICLR)*.
- [49] Mingxing Tan, Ruoming Pang, and Quoc V Le. 2020. EfficientDet: Scalable and Efficient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*. 10781–10790.

- [50] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.
- [51] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347* (2020).
- [52] Abdul Vahab, Maruti S Naik, Prasanna G Raikar, and Prasad SR. 2019. Applications of Object Detection System. *International Research Journal of Engineering and Technology (IRJET)* 6, 4 (2019), 4186–4192.
- [53] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2020. Scaled-YOLOv4: Scaling Cross Stage Partial Network. *arXiv preprint arXiv:2011.08036* (2020).
- [54] Derui Wang, Chaoran Li, Sheng Wen, Xiaojun Chang, Surya Nepal, and Yang Xiang. 2019. Daedalus: Breaking non-maximum suppression in object detection via adversarial examples. *arXiv* (2019), arXiv–1902.
- [55] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. 2019. Transferable Adversarial Attacks for Image and Video Object Detection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI) 2019*, Sarit Kraus (Ed.), ijcai.org, 954–960.
- [56] Eric Wong and J. Zico Kolter. 2018. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 5283–5292.
- [57] 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 (ECCV) 2020*, Vol. 12349, 1–17.
- [58] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. 2021. PatchGuard: A Provably Robust Defense against Adversarial Patches via Small Receptive Fields and Masking. In *30th USENIX Security Symposium (USENIX Security)*.
- [59] Chong Xiang and Prateek Mittal. 2021. PatchGuard++: Efficient Provable Attack Detection against Adversarial Patches. In *ICLR 2021 Workshop on Security and Safety in Machine Learning Systems*.
- [60] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *IEEE International Conference on Computer Vision (ICCV) 2017*. IEEE Computer Society, 1378–1387.
- [61] 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 (ECCV) 2020*, Vol. 12350, 665–681.
- [62] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *25th Annual Network and Distributed System Security Symposium (NDSS)*.
- [63] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems* 30, 9 (2019), 2805–2824.
- [64] Haichao Zhang and Jianyu Wang. 2019. Towards Adversarially Robust Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV) 2019*. IEEE, 421–430.
- [65] Zhenyuan Zhang, Benson Yuan, Michael McCoyd, and David Wagner. 2020. Clipped BagNet: Defending Against Sticker Attacks with Clipped Bag-of-features. In *3rd Deep Learning and Security Workshop (DLS)*.
- [66] 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.

## A OBJECT SIZE AND PATCH SIZE

Recall that in Section 5.4, we use a  $32 \times 32$  patch on  $416 \times 416$  (or  $224 \times 740$ ) images to evaluate the provable robustness. In this section, we provide additional details of object sizes and patch sizes in PASCAL VOC, MS COCO, and KITTI datasets.

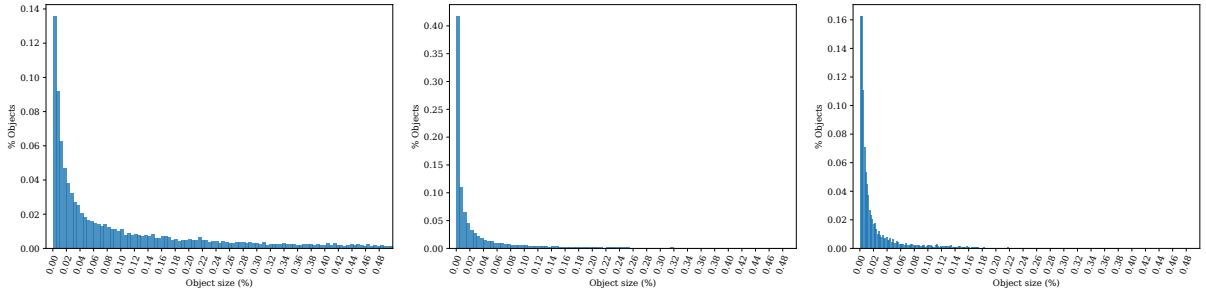
**Small objects are the majority of all three datasets.** In Figure 8, we plot the histogram of object sizes (in the percentage of pixels) in the test/validation sets of PASCAL VOC, MS COCO, and KITTI. As shown in the plots, small, or even tiny, objects are the majority of three datasets. A  $32 \times 32$  patch takes up 0.6% pixels of a  $416 \times 416$  (or  $224 \times 740$ ) image, and our further analysis shows that 15.2% objects of PASCAL VOC are smaller than 0.6% image pixels; 44.5% of MS COCO and 44.6% KITTI objects are smaller than 0.6%. Moreover, more than 36.5% of PASCAL VOC objects, more than 66.3% of MS COCO objects, and 75.9% KITTI objects are smaller than a  $64 \times 64$  square. These numbers explain why the absolute numbers of certified recall in Table 3 are low. In Figure 9, we further provide visualization of a  $32 \times 32$  patch on the  $416 \times 416$  image to demonstrate the challenge of perfect robust detection even when a small patch is presented. In the left two examples, the person and the cow are completely blocked by the adversarial patch and thus are unrecognizable. In the rightmost example, the head of the dog is patched and it is even hard for humans to determine if it is a dog or cat.

**Additional evaluation results for different patch sizes.** In Figure 10, we vary the patch size to see how the provable robustness is affected given different attacker capabilities (i.e., patch sizes). If we consider a smaller patch of  $8 \times 8$  pixels, we can have a 2.0% higher CR for close-patch, and a 2.8% higher CR for over-patch compared with our CRs for a  $32 \times 32$  patch. Furthermore, we note that in the far-patch model, the patch size has a limited influence on the provable robustness. From Figure 10, we can also see that the CR decreases as the patch size increases. This analysis demonstrates the limit of DetectorGuard as well as the challenge of robust object detection with larger patch sizes. We aim to push this limit further in our future work.

## B ADDITIONAL DISCUSSION OF ROBUST CLASSIFIER IMPLEMENTATION

As introduced in Section 2.4, state-of-the-art provable robust image classifiers [36, 58] 1) use DNN with small receptive fields to bound the number of corrupted features and then 2) do secure aggregation on the partially corrupted feature map for robust classification. In DetectorGuard, we choose BagNet [3] for small receptive fields and feature clipping for secure aggregation. In this section, we provide additional details of BagNet and clipping aggregation. Furthermore, we discuss alternative aggregation mechanisms and implement robust masking [58] to demonstrate the generality of our DetectorGuard framework.

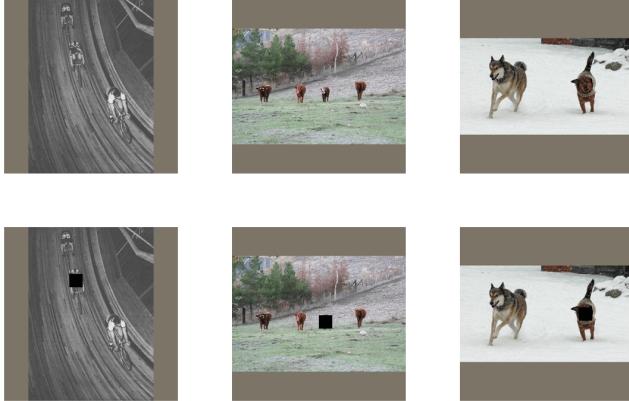
**BagNet [3].** BagNet was originally proposed for interpretable machine learning. It inherits the high-level architecture of ResNet-50 [20] and replaces  $3 \times 3$  convolution kernels with  $1 \times 1$  ones to reduce the receptive field size. The authors designed three BagNet architectures with a small receptive field of  $9 \times 9$ ,  $17 \times 17$ , and  $33 \times 33$ , in contrast to ResNet-50 having a receptive field of  $483 \times 483$ .



**Figure 8: Histogram of object sizes (left: PASCAL VOC; middle: MS COCO; right: KITTI)**

**Table 5: Comparison between masking-based and clipping-based defenses of DetectorGuard (using a perfect clean detector)**

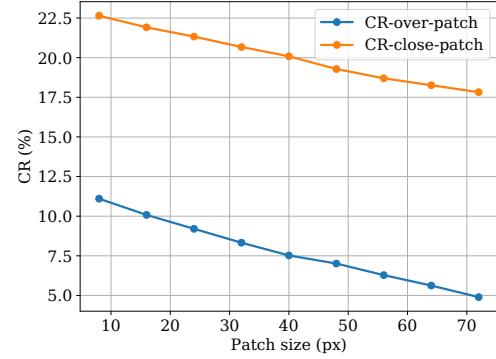
	clipping-based DetectorGuard					masking-based DetectorGuard				
	AP	FAR	CR-far	CR-close	CR-over	AP	FAR	CR-far	CR-close	CR-over
PASCAL VOC	99.3%	0.9%	28.6%	20.7%	8.3%	98.9%	1.1%	26.2%	17.2%	4.7%
MS COCO	99.0%	1.2%	11.5%	7.0%	2.2%	98.7%	1.4%	11.4%	5.4%	1.6%
KITTI	99.0%	1.5%	31.6%	11.0%	2.1%	99.4%	1.1%	17.4%	4.9%	1.2%



**Figure 9: Visualization of patches on small objects (upper: original 416×416 images; lower: images with a 32×32 black patch)**

Brendel and Bethge showed that BagNet-17 with a  $17 \times 17$  receptive field can achieve a similar top-5 accuracy as AlexNet [3]. In recent works [58, 65] on adversarial patch defense, BagNet has been adopted to bound the number of corrupted features to achieve robustness.

**Clipping.** In addition to the use of CNNs with small receptive fields, we also need a secure aggregation mechanism to ensure that a small number of corrupted features only have a limited influence on the final prediction/classification. Recall that in our provable analysis (Algorithm 2; Section 4), we need the lower bound of classification logits to reason about the worst-case objectness map output. In order to impose such a lower bound, we clip all feature values into  $[0, \infty]$  such that an adversarial patch cannot decrease the values of



**Figure 10: Effect of patch size on provable robustness of DetectorGuard with a perfect clean detector**

object classes significantly. It is easy to calculate the lower bound of classification logits: we only need to zero out all features within the patch location(s) and aggregate the remaining features.

**Alternative aggregation.** We propose DetectorGuard as a general framework that is compatible with any provably robust image classification technique. To further support this claim, we implement Objectness Predictor using a PatchGuard classifier with robust masking secure aggregation [58], which achieves the best clean classification accuracy and provable robust classification accuracy on high-resolution ImageNet [11] dataset. We compare the performance of defenses with clipping-base and robust-making-based secure aggregation in Table 5. As we can see from the table, two defenses achieve high clean performance and non-trivial provable robustness, demonstrating that DetectorGuard is compatible with different provably robust image classifiers. We note that we do not

choose robust masking in our main evaluation because 1) it has a looser lower bound compared with clipping while 2) it introduces a slightly higher computation overhead. Furthermore, robust masking of PatchGuard only has limited robustness against the multiple-patch attacker. In contrast, as demonstrated in Appendix C, the clipping-based DetectorGuard can handle multiple patches.

## C DISCUSSION ON MULTIPLE PATCHES

In our main body, we focus on the one-patch threat model because building a high-performance provably robust object detector against a single-patch attacker is an unresolved and open research question. In this section, we discuss DetectorGuard’s robustness against multiple patches.

**Quantitative analysis of clipping-based DetectorGuard against multiple patches.** One advantage of the clipping-based robust classifier is its robustness against multiple patches. As long as the sub-procedure RCH-PA(-) of the clipping-based robust classifier can return non-trivial bounds of classification logits, we can directly plug the sub-procedure into our Algorithm 2 to analyze the robustness against multiple patches.

We note that despite the theoretical possibility to defend against attacks with multiple patches, its quantitative evaluation for provable robustness is extremely expensive due to the large number of all possible combinations of multiple patch locations. Consider a  $32 \times 32$  patch on a  $416 \times 416$  image. There are 148k possible patch locations (or 1.6k feature-space locations). If we are using 2 patches of the same size, the number of all location combinations becomes higher than  $10^{10}$  (or 1.4M feature-space location combinations)!

In order to provide a proof-of-concept for defense against multiple patches, we perform an evaluation on 50 PASCAL VOC images using a subset of patch locations (1/16 of all location combinations). The results are reported in Table 6. As shown in the table, DetectorGuard is able to defend against multiple patches. Moreover, if we compare provable robustness against one  $32 \times 32$  (1024 px) and two  $24 \times 24$  patches (1152 px), which have a similar number of pixels, we can find that using two smaller patches (two  $24 \times 24$  patches) is only more effective for over-patch threat model but not for far-patch and close-patch threat models. This observation leads to the following remark.

**Remark: multiple patches need to be close to each other and the victim object for a stronger malicious effect.** Unlike image classification where the classifier makes predictions based on all image pixels (or extracted features), an object detector predicts each object largely based on the pixels (or features) around the object. As a result, patches that are far away from the object only have a limited malicious effect, and this claim is supported by our evaluation results in Section 5.4 (i.e., DetectorGuard is more effective against the far-patch threat model). Therefore, multiple patches should be close to the victim object and hence close to each other for a more effective attack. In this case, the multiple-patch threat model becomes similar to the one-patch model since patches are close to each other and can merge into one single patch. That is, we can use one single patch of a larger size to cover all perturbations in multiple small patches.

**Table 6: Provable robustness (CR) of DetectorGuard (using a perfect clean detector) against multiple patches (evaluated on 50 PASCAL VOC images with a subset of patch locations)**

	far-patch	close-patch	over-patch
one $32 \times 32$ patch (1024 px)	27.3%	22.4%	8.7%
two $32 \times 32$ patches (2048 px)	27.3%	18.0%	3.1%
two $24 \times 24$ patches (1152 px)	27.3%	18.6%	3.1%
two $16 \times 16$ patches (512 px)	27.3%	19.3%	5.0%

## D EXPERIMENT RESULTS FOR DIFFERENT THREAT MODELS AND DATASETS

In this section, we include additional plots for per-class analysis as well as DetectorGuard’s clean/provable performance on MS COCO and KITTI. The observation is similar to that in Section 5.

**Per-class Analysis.** In Figure 11, we provide additional per-class analysis results. The observation is similar to Figure 6 in Section 5.

**Additional plots for MS COCO and KITTI.** We plot the clean performance and the provable robustness for MS COCO in Figure 12 and Figure 13, and for KITTI in Figure 14 and Figure 15. The observation is similar to that on PASCAL VOC (Figure 4 and Figure 5).

## E JUSTIFICATION FOR DEFENSE OBJECTIVE

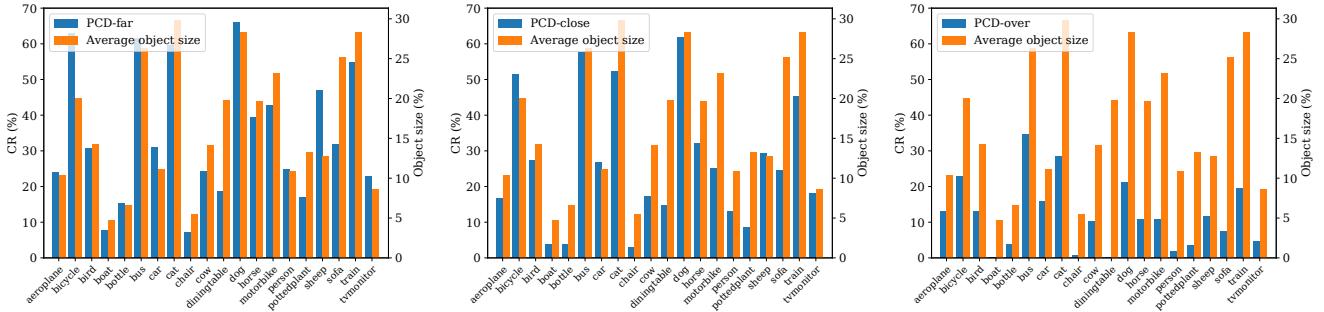
In Section 2.3, we allow DetectorGuard to *only detect part of the object* or to trigger an attack alert on adversarial images. In this section, we discuss why this is a reasonable defense objective and how to extend DetectorGuard for a stronger notion of robustness.

**Partially detected bounding box.** We note that we allow the patch to be anywhere, even over the salient object. As a result, the patch likely covers a large portion of the object (visualization examples include the right part of Figure 1 and Figure 9; see Appendix A for more details of object sizes and patch sizes). Therefore, it is reasonable to allow the model to output a smaller bounding box. If we consider the application scenario of autonomous vehicles (AV), partially detecting a pedestrian or a car is already sufficient for an AV to make a correct decision.

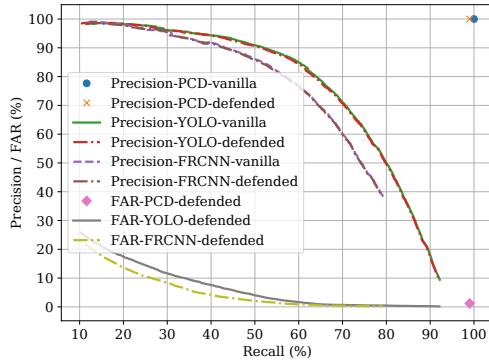
Moreover, we can tune hyper-parameters such as binarizing threshold  $T$  to increase the objectness in the output of Objectness Predictor. More objectness will force the adversary to let Base Detector predict a larger bounding box in order to reduce unexplained objectness that will otherwise lead to an attack alert. However, we note that more objectness also makes it more likely for DetectorGuard to trigger a false alert on clean images. This trade-off between robustness and clean performance should be carefully balanced.

## F ADDITIONAL DISCUSSION ON “FREE” PROVABLE ROBUSTNESS

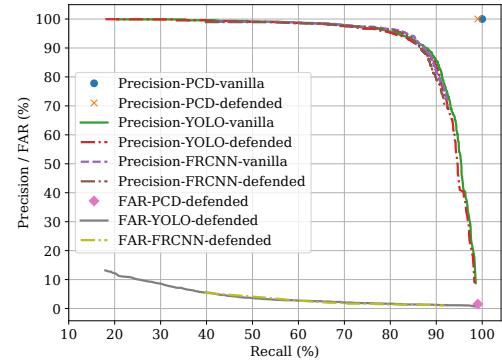
As shown in our evaluation, DetectorGuard achieves the first provable robustness for object detectors against patch hiding attacks at a negligible cost of clean performance. Intriguingly, we have demonstrated that we can use a module with limited clean performance (i.e., provably robust image classifier in Objectness Predictor) to build a provably robust system with high clean performance (i.e.,



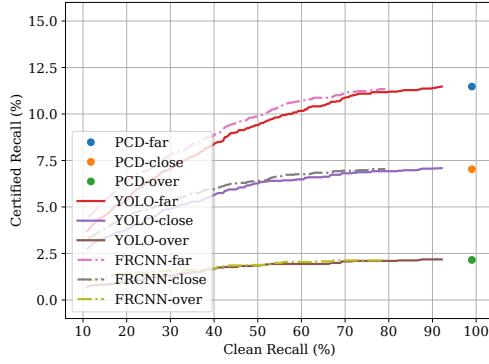
**Figure 11: Per-class analysis of PASCAL VOC (left: far-patch; middle: close-patch; right: over-patch)**



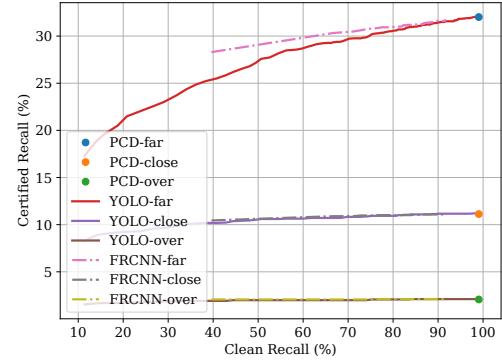
**Figure 12: Clean performance of DetectorGuard on MS COCO**



**Figure 14: Clean performance of DetectorGuard on KITTI**



**Figure 13: Provable robustness of DetectorGuard on MS COCO**

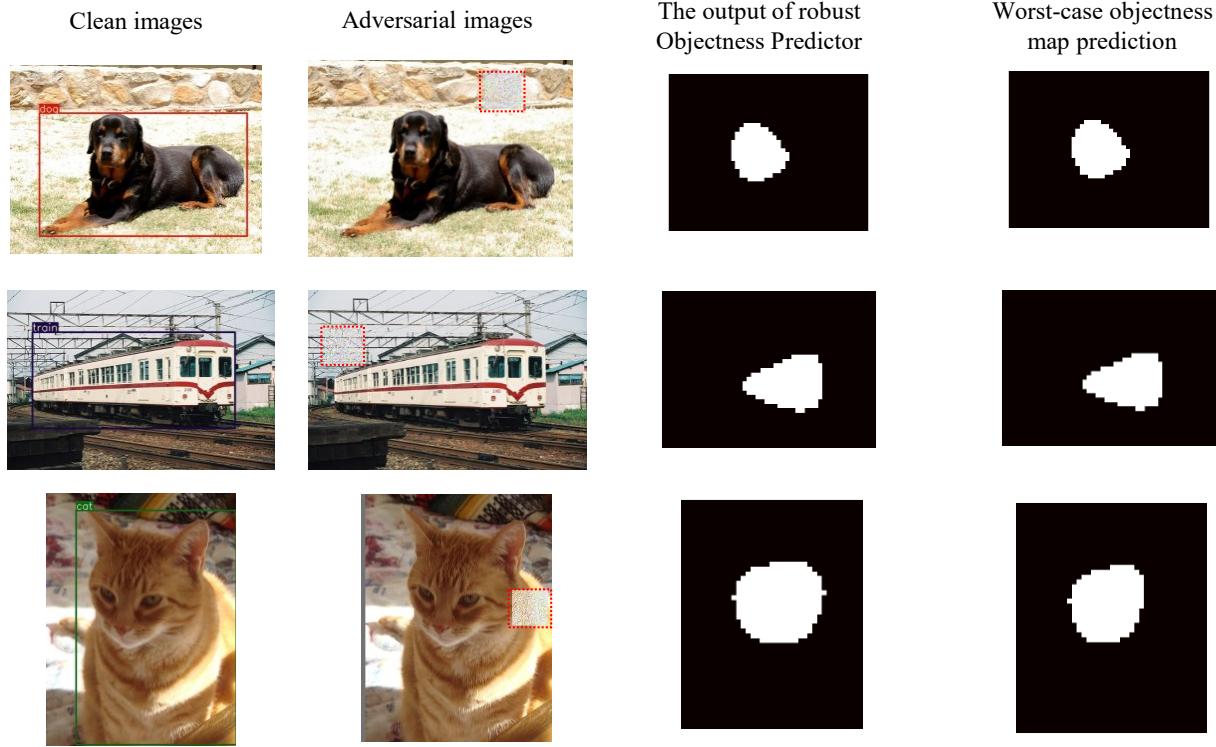


**Figure 15: Provable robustness of DetectorGuard on KITTI**

DetectorGuard). In this section, we provide additional discussion on this intriguing behavior.

One major difference between image classification and object detection is their type of error. For an image classifier, the only error is misclassification. In contrast, an object detector can have

two types of errors, false-negative (FN; missing object) and false-positive (FP; predicting incorrect objects). Intriguingly, despite the difficulty to have a low FN and a low FP at the same time, it is easy to have a low FN (but with a potentially high FP) or a low FP (but with a potentially high FN). For example, if an object detector predicts all possible bounding boxes, the FN is zero (all possible boxes are retrieved, including the ground-truth box) but the FP is extremely high (most bounding boxes are incorrect). This intrinsic



**Figure 16: Visualization of DetectorGuard.** From left to right: 1) clean images – conventional detectors correctly detect all objects; 2) adversarial images with patches (marked with red dash boxes) – conventional detectors miss all objects; 3) objectness map generated on these adversarial images – Objectness Predictor robustly predicts high objectness and eventually leads to an attack alert; 4) worst-case objectness map – DetectorGuard can certify the provable robustness of these objects. Note that visualizations in this figure only consider one random patch location for each image, while results reported in Section 5 consider *all possible locations and attack strategies* within the threat model.

property of the object detection task allows us to achieve “free” provable robustness.

Recall that in the clean setting, our design of prediction matching strategy (Section 3.4) enables DetectorGuard to tolerate FN of Objectness Predictor (i.e., *Clean Error 3*: predicts objects as background). Therefore, we can safely and easily optimize for a low FP (i.e., *Clean Error 2*: predicts background as objects) to achieve a high clean performance.

## G PIXEL-SPACE AND FEATURE-SPACE WINDOWS

Recall that in Section 3.3, we used a BagNet to extract a feature map for the entire image and perform robust window classification in the feature space. This design allows us to reuse the extracted feature map and reduce computational overhead. In this section, we discuss how to map the pixel-space bounding box to the feature space.

**Box mapping.** For each pixel-space box  $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ , we calculate the feature-space coordinate  $x'_{\min} = \lfloor (x_{\min} - r + 1)/s \rfloor$ ,  $y'_{\min} = \lfloor (y_{\min} - r + 1)/s \rfloor$ ,  $x'_{\max} = \lfloor x_{\max}/s \rfloor$ ,  $y'_{\max} = \lfloor y_{\max}/s \rfloor$ , where  $r, s$  are the size and stride of the receptive field size. The new feature-space coordinates indicate all features that are affected

by the pixels within the pixel-space bounding box. We note that the mapping equation might be slightly different given different implementation of CNNs with small receptive fields. In our BagNet implementation, we have  $r = 33$ ,  $s = 8$ .

## H VISUALIZATION OF DETECTOR GUARD

In this section, we give a simple visualization for DetectorGuard with YOLOv4 as Base Detector (Figure 16). To start with, we select three random images with larger objects and visualize the detection output of YOLOv4 in the first column. Second, we pick a random patch location on the image and perform an empirical patch hiding attack. The attack aims to optimize the pixel values within the adversarial patch to minimize the objectness confidence score of every possible bounding box prediction, which is a common strategy used in relevant literature [57, 61]. As shown in the second column, our patch attacks are successful, and YOLOv4 fails to detect any objects. Note that we use red dash boxes to illustrate the patch locations, and they are not the outputs of YOLOv4. Third, we feed this adversarial image to Objectness Predictor, and we visualize the predicted objectness maps in the third column. As shown in the figure, although the adversarial patch makes YOLOv4 miss all objects, Objectness Predictor still robustly outputs high objectness.

As discussed in Section 3, DetectorGuard will eventually issue an attack alert. Fourth, we reason about the worst-case objectness map prediction for these particular random patch locations used in the visualization, and plot the worst-case output in the fourth column. As shown in our visualization, Objectness Predictor can still output high objectness in the worst case. Therefore, we can certify the robustness of DetectorGuard for this patch location.

Finally, we want to note that this appendix is merely a simple case study for an empirical patch attack at one single random location of each image. In contrast, robustness results reported in Section 5 are derived from Algorithm 2 and Theorem 1 holds *for any possible patch attack strategy at any valid patch location*.