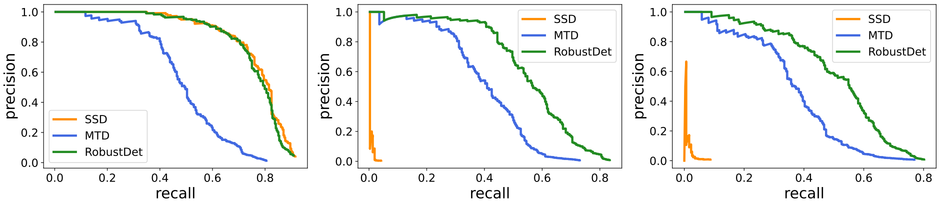
Adversarially-Aware Robust Object Detector

Ziyi Dong, Pengxu Wei*, and Liang Lin

Sun Yat-Sen University, Guangzhou, China
dongzy6@mail2.sysu.edu.cn, weipx3@mail.sysu.edu.cn, linliang@ieee.org



(a) Clean images (b) Adversarial images(A_{cls}) (c) Adversarial images(A_{loc})

Fig. 1: Precision-Recall (PR) curves of non-robust detector (standard SSD), and two SSD-based robust detectors, *i.e.*, MTD [34] and our RobustDet. They are respectively evaluated under *the conventional standard setting* with clean images and *two detector attacks* whose adversarial images are generated from attacks of classification (A_{cls}) and localization (A_{loc}) [34]. It is observed that SSD has a high performance on clean images but *performs rather poorly under two attacks*. The robust detector MTD is relatively robust under attacks but *presents a significant performance drop on clean images*. Instead, *our RobustDet not only gains a reliable detection robustness on adversarial images, but also maintains a high detection performance on clean images on par with the standard SSD.*

Abstract. Object detection, as a fundamental computer vision task, has achieved a remarkable progress with the emergence of deep neural networks. Nevertheless, few works explore the adversarial robustness of object detectors to resist adversarial attacks for practical applications in various real-world scenarios. Detectors have been greatly challenged by unnoticeable perturbation, with sharp performance drop on clean images and extremely poor performance on adversarial images. In this work, we empirically explore the model training for adversarial robustness in object detection, which greatly attributes to the conflict between learning clean images and adversarial images. To mitigate this issue, we propose a Robust Detector (RobustDet) based on adversarially-aware convolution to disentangle gradients for model learning on clean and adversarial images. RobustDet also employs the Adversarial Image Discriminator (AID) and Consistent Features with Reconstruction (CFR) to ensure a reliable robustness. Extensive experiments on PASCAL VOC and MS-COCO demonstrate that our model effectively disentangles gradients and significantly enhances the detection robustness with maintaining the detection ability on clean images. Our source code and trained models are publicly available at: https://github.com/7eu7d7/RobustDet

Keywords: Object Detection, Adversarial Attack and Defense, Adversarial Robustness, Detection Robustness Bottleneck

* Corresponding Author

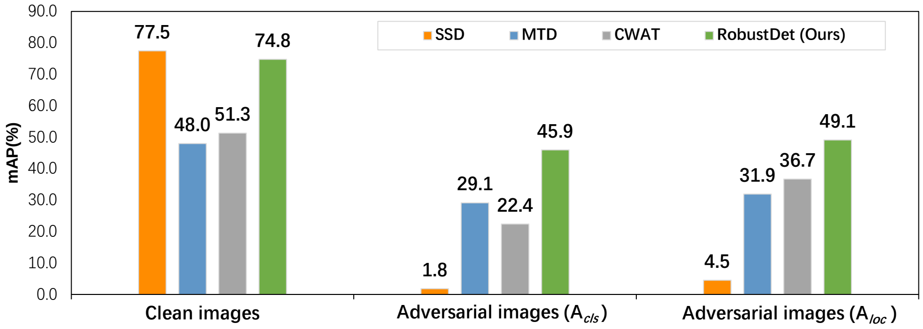


Fig. 2: Detection performance comparison on clean and adversarial images for standard SSD, MTD [34], CWAT [5] and our RobustDet.

1 Introduction

Although deep neural networks (DNNs) have achieved a remarkable progress in many visual tasks such as image classification [12], object detection [9,23] and semantic segmentation [37,4], they are vulnerable to even slight, imperceptible adversarial perturbations and yield erroneous predictions [10,21,31,3]. *A miss is as good as a mile.* Such vulnerability inspires increasing attentions on the adversarial robustness mainly in the image classification task [29,3,16,36,22]. Nevertheless, with elaborate architectures to recognize simultaneously where and which category objects are in images, object detectors also suffers from the vulnerable robustness and are easily fooled by adversarial attacks [32,30,6,5,15]. As demonstrated in Fig. 2, standard SSD achieves only **1.8% mAP** on adversarial images, by **75.7% mAP drops**! The vulnerability of object detection models seriously raises security concerns on their practicability in security-sensitive applications, *e.g.*, autonomous driving and video surveillance.

The vulnerable robustness of object detectors has been impressively verified to attack two tasks of classification and localization [32,30,24,6], few researches focus on investigating the challenging countermeasure: *how to defend those attacks to resist the adversarial perturbations for detectors*. To address this issue, MTD [34], as an earlier attempt, regards the adversarial training of object detection as a multi-task learning and choose those adversarial images that have the largest impact on the total loss for learning. Subsequently, the second related work, CWAT [5], points out the problem of class imbalance in the attack and proposes to attack each category as evenly as possible to generate more reasonable adversarial images. In general, these existing methods suffer from the **detection robustness bottleneck**: *a significant degradation on clean images with only a limited adversarial robustness*, shown in Fig. 1 and 2. That is, due to the introduction of adversarial perturbation during training, they reach a compromise for both the model accuracy on clean images and the robustness on adversarial images. This would inevitably make a concession of robust models

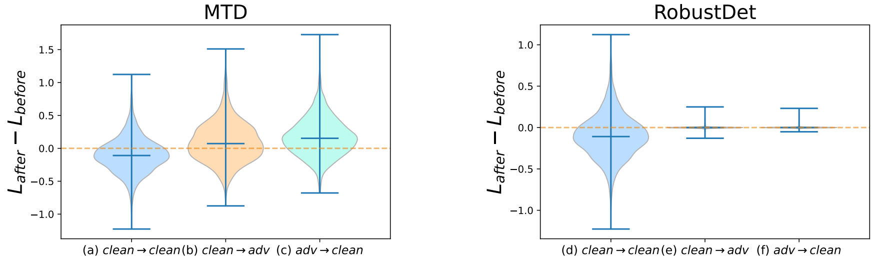


Fig. 3: Empirical analyses on the conflict between the learning of clean images and adversarial images via the statistics of loss changes[1]. (a), (b) and (c) are the loss changes on robust detector MTD [34]. (d), (e) and (f) are the loss changes on our RobustDet. For both methods under $clean \rightarrow clean$, they have the decreasing loss changes for most images, indicating the favorable training. Under $clean(adv) \rightarrow adv(clean)$, MTD has the increasing loss changes for most images, indicating the inverse training effects between learning clean and adversarial images. Instead, our RobustDet has almost no effects between them, indicating a better disentanglement for learning clean and adversarial images.

with the performance sacrifice on clean images as well as a limited adversarial robustness for object detection.

In this paper, we firstly explore the aforementioned **detection robustness bottleneck** on both clean images and adversarial images for object detection. Particularly, one noteworthy difference from the adversarial robustness in the image classification task, where robust models usually only have a small amount of the performance decline on clean images [35,11], is that robust object detectors only yields a limited robustness from adversarial training and suffer from a significant performance degradation by nearly 30% on clean images (77.5% mAP for standard SSD *vs.* 48.0% mAP for MTD [34] on the PASCAL VOC dataset, as shown in Fig. 2). It indicates that, in the training phase, robust detectors hardly reach a win-win balance to trade off the robustness of adversarial images and the accuracy of clean images. *To further investigate this issue, on one hand, we inspect the individual loss changes for both images in an adversarial robust detector.* A conflict between two tasks of learning clean images and adversarial images in adversarial training is observed, which can be speculated as a pitfall to explain the aforementioned detection robustness bottleneck to a certain extent. *On the other hand, we analyze the interference between the gradients of clean images and the adversarial images for existing models.* Accordingly, strong interference is observed, indicating that an object detector has a large difficulty to distinguish no-robust and robust features. Thus, it is reasonable that models are confronted with the detection robustness bottleneck.

To mitigate this problem, we propose a Robust Detection model (RobustDet) via adversarially-aware convolution. The model learns different groups of convolution kernels and adaptively assigns weights to them based on the Adversarial

[1] More details can be referred to our supplementary material.

Image Discriminator (AID). RobustDet also employs the Consistent Features with Reconstruction (CFR) to ensure reliable robustness. By applying reconstruction constraints to make the features extracted by the model can be reconstructed as clean images as possible, the model is drived to extract more robust features for both clean and adversarial images. Extensive experimental results on PASCAL VOC [8] and MS-COCO [18] datasets have demonstrated superior accuracy performance on clean images and promising detection robustness on adversarial images.

Overall, our contributions are summarized as follows:

1. Empirically, we analyse the detection robustness bottleneck and verify the conflict between learning clean images and adversarial images for robust object detectors.
2. Technically, we propose a robust detection model (RobustDet) based on adversarially-aware convolution to learn robust features for clean images and adversarial images. In addition, we propose Consistent Features with Reconstruction (CFR) to constrain the model to extract more robust features that can be reconstructed as clean images as possible.
3. Experimentally, we conduct comprehensive experiments to evaluate the proposed approach for adversarial detection robustness on PASCAL VOC and MS-COCO datasets, achieving state-of-the-art performance on both clean images and adversarial images. It presents a superior accuracy performance on clean images and a promising detection robustness on adversarial images.

2 Related Work

2.1 Adversarial Attack and Defense

For deep neural networks, their excellent feature representation capability has been demonstrated in various scenarios [12,26,13]. Even so, it has been criticized that neural network models easily produce totally wrong predictions under slight perturbations to inputs [28]. Especially, they are rather vulnerable to adversarial attacks. Accordingly, more and more adversarial attack methods have been proposed: gradient-based white box adversarial attack methods (*e.g.*, FGSM [10] and PGD [21]), and black box adversarial attack methods (*e.g.*, UPSET [24] and LeBA [33]). These methods can easily fool the classification model and even a change in just one pixel would totally fool the model [27]. To address this problem, some defense methods have been proposed [29,3,16,36,22]. Among them, adversarial training is one of the most widely used and effective methods. It allows the model to continuously learn the adversarial images and focuses more on the robust features of adversarial images and clean images to ignore non-robust features.

2.2 Attack and Robust Object Detector

In recent years, seminal object detection models have been proposed, *e.g.*, Faster RCNN [23], SSD [19], YOLOX [9], and DETR [2], building a series of profound and insightful milestones for object detection. Even so, they inevitably inherit the vulnerability to attack, with the root in deep neural networks. Existing researches have shown that attack methods for classification tasks can also be effective in attacking object detection models [34]. Object detectors have some different structures from classification models, and object detectors can be attacked more effectively for these structures. For example, DAG [32] and UEA [30] are the attack methods for object-level features by superimposing perturbation on the whole image. Dpatch [20] fools the detector by adding a patch to the image. ShapeShifter [6] attacks detectors in the physical world.

Instead, although attack methods for object detectors are becoming more and more efficient, there are few defense strategies in the object detection task. [34] proposes the MTD method based on adversarial training. At each step of adversarial training, the images that can increase the loss the most are selected from the adversarial images to learn to improve the robustness of the model. [5] explores the problem of class imbalance in the attacks for object detectors and proposes to make the attack intensity as consistent as possible for each class. Adversarial training is performed through these images to improve the robustness of the model. These methods mainly focus on the generation of adversarial images and ignore the lack of robustness caused by the structure of the model. Thus, they suffer from the detection robustness bottleneck as mentioned in Sec. 1.

Since few research works on adversarially-robust object detectors, it is almost blind to essentially explore object detection. In this paper, we will firstly explore empirically the detection robustness bottleneck to further understand the adversarial robustness in object detection in Sec. 3. Then, we will elaborate the proposed RobustDet to address the detection robustness bottleneck in Sec. 4. We will conduct extensive experiments to demonstrate the effectiveness of the proposed method In Sec. 5 and conclude the paper in Sec. 6.

Defenses against unseen attacks are customarily explored in classification tasks. However, for detection tasks we suffer from a lack of the most fundamental conception of their robustness. Thus, we focus more on more fundamental problems of the robustness of object detectors. Those more advanced problems need to be explored further based on this work.

3 Adversarial Robustness in Object Detection

3.1 Problem Setting

For a clean image x, an object detector f parameterized by $\boldsymbol{\theta}$, yields object bounding boxes $\{\hat{\boldsymbol{b}}_i = [p_i^x, p_i^y, w_i, h_i]\}$ with their predicted class probabilities $\{\hat{\boldsymbol{c}}_i = [\hat{c}_i^{bg}, \hat{c}_i^1, \cdots, \hat{c}_i^C]\}$ over the background (bg) and C object categories, *i.e.*, $f(x;\boldsymbol{\theta}) \rightarrow \{\hat{\boldsymbol{b}}_i, \hat{\boldsymbol{c}}_i\}$, where p_i^x and p_i^y are the coordinates of the top left corner of $\hat{\boldsymbol{b}}_i$, w_i and h_i are the width and height of $\hat{\boldsymbol{b}}_i$. The localization loss $\mathcal{L}_{loc} = \sum_{i\in pos} L_1^{smooth}(\hat{\boldsymbol{b}}_i, \boldsymbol{b}_i)$ and the classification loss $\mathcal{L}_{cls} = -\sum_{i\in pos} c_i \log\left(\hat{c}_i^u\right) - \sum_{i\in neg} c_i \log(\hat{c}_i^{bg})$, where $\boldsymbol{b}_i$ is the Ground-Truth (GT) bounding box that matches the predicted bounding box $\hat{\boldsymbol{b}}_i$ and $\boldsymbol{c}_i$ denotes its GT category, and the detection loss is $\mathcal{L}_{det} = \mathcal{L}_{loc} + \mathcal{L}_{cls}$.

Following MTD [34], two types of attacks (A_{cls} and A_{loc}) for object detection are specifically steered for classification and localization, respectively:

$$\begin{aligned} A_{cls}(x) &= \underset{\bar{x}\in\mathcal{S}_x}{\arg\max}\, \mathcal{L}_{cls}(f(\bar{x};\boldsymbol{\theta}), \{\boldsymbol{c}_i, \boldsymbol{b}_i\}), \\ A_{loc}(x) &= \underset{\bar{x}\in\mathcal{S}_x}{\arg\max}\, \mathcal{L}_{loc}(f(\bar{x};\boldsymbol{\theta}), \{\boldsymbol{c}_i, \boldsymbol{b}_i\}), \end{aligned} \tag{1}$$

where $\bar{x}$ is the adversarial counterpart of x, and $\mathcal{S}_x = \left\{\bar{x} \cap [0, 255]^{cwh} \middle| \|\bar{x} - x\|_\infty \leq \epsilon\right\}$ is the adversarial image space centered on clean images x with perturbation budget of ϵ. A_{cls} denotes searching for the image x in its ϵ neighborhood that maximizes L_{cls} as the adversarial image.

3.2 Analyses of the Detection Robustness Bottleneck

(1) Conflict between Learning Adversarial images and Clean images. To defense the attacks, robust models are expected to be immune to adversarial perturbations via learning shared features between clean images and adversarial images to improve the robustness of the model. This is the conventional wisdom in prevalent adversarial training for defense, especially in the image classification task. Nevertheless, the adversarial robustness for object detection is worrisome. Namely, robust detection models perform poorly on both clean and adversarial images, as demonstrated in Fig. 1 and Fig. 2. In particular, adversarial training on both clean and adversarial images results in a significant performance drop on clean images. This may indicate a conflict between the tasks of learning clean and adversarial images; thus the model has to compromise a trade-off between adversarial and clean images. To further explore the reasons why the model cannot learn both images well, we conduct an investigation from two aspects.

Loss changes for clean and adversarial images. We inspect intuitively the loss changes for clean and adversarial images. Specifically, we perform a validation via m-step adversarial training of an adversarially-trained robust model on a batch of clean images or adversarial images and observe the loss change on another batch of images (The selection of m and algorithm details are discussed in the supplementary material). The loss change of the adversarial (clean) image after learning the clean (adversarial) image is defined as $clean \rightarrow adv$ ($adv \rightarrow clean$). From the experimental results in Fig. 3, it is observed that $clean \rightarrow adv$ and $adv \rightarrow clean$ are positive for most images compared with the most negative results of $clean \rightarrow clean$. This shows that learning clean images and adversarial images will increase the loss of each other for most images. The impact of adversarial images on clean images is greater than that of clean images on adversarial images. This validation shows that learning clean images and adversarial images are conflicting tasks for the model, to some extent. Thus, during the training phase, the model has the burden to well address this learning conflict.

Gradient interference analysis. The clean image and the adversarial image are from two different domains with different patterns. There are shared

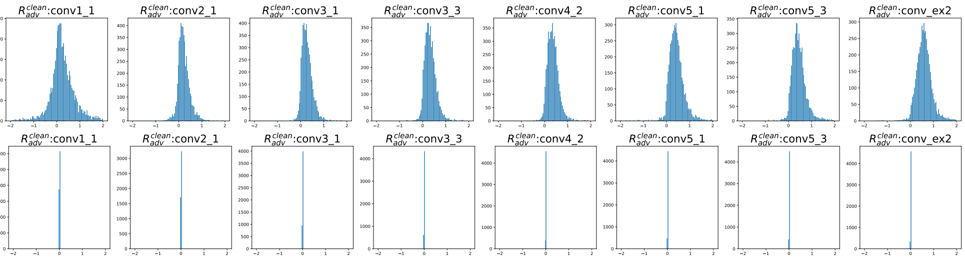


Fig. 4: The gradient entanglement degree R_{adv}^{clean} of clean images and adversarial images based on features from different convolutional layers. The upper shows the results from SSD and the second row is from our RobustDet.

features between them but also have their specific features. A highly robust model must have parameters for extracting the shared features and another two part parameters for extracting specific features that are orthogonal to each other. For an adversarially-trained robust model, the shared features of two kinds of images should have been well learned, and only the part processing specific features still needs reinforcement. Therefore, for this model, the gradients generated by the two kinds of images should have low correlation and be nearly orthogonal.

Accordingly, we define the intensity of gradient entanglement of one image x_1 to another image x_2: $\mathcal{R}_{g_2}^{g_1} = {g_1}^T g_2 / |g_2|^2$, where g_1 and g_2 are the gradient vectors of the two images. For two clean and adversarial images, the greater their gradient entanglement of these two kinds of images, the more serious the interference between them; and the model does not distinguish the specific features well. Based on the experimental results of the above loss variations, the greater the gradient entanglement, the more difficult the conflict between the two kinds of images can be reconciled. The smaller gradient entanglement indicates that the model has enough ability to distinguish the shared features from their specific features and can disentangle the clean images and the adversarial images. It can be seen from the Fig. 4 that the gradient entanglement between the clean image and the adversarial image on the adversarial trained robust model is quite high, and even negative values appear in the first few layers. This shows that the updated directions of some clean images and adversarial images on the adversarial-trained model are completely opposite, which also indicates the conflict between the two kinds of images. When learning one kind of image, it will inevitably have an impact on another kind of image, which leads to a detection robustness bottleneck.

(2) The Conflict to the Robustness of Classification and Location. We compare the detection results of the non-robust model and the adversarial trained robust model on the clean image, the A_{cls} adversarial image, and the A_{loc} adversarial image. It can be seen from the Fig. 5 that the non-robust model will locate the wrong object with high confidence when applying an attack. The robust model will not completely confuse to the attack, but its classification and localization accuracy on both clean images and adversarial images have greatly decreased. The robustness of localization objects is much better than

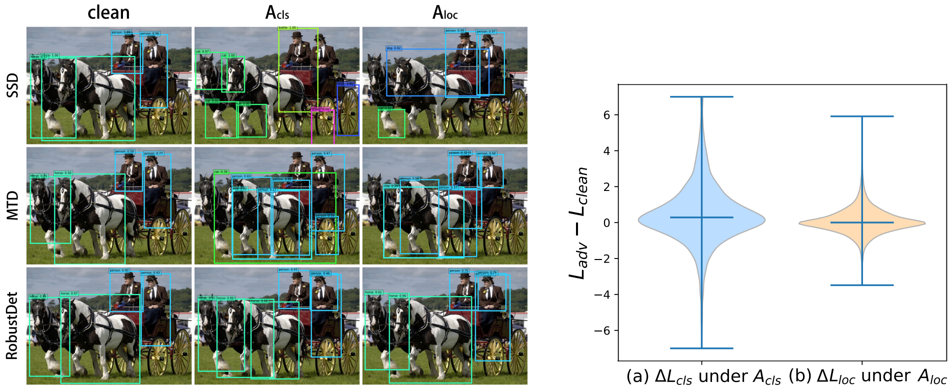


Fig. 5: **Left:** The detection results of the standard SSD, MTD and our RobustDet on the clean image and two adversarial images attacked from classification (A_{cls}) and localization (A_{loc}). MTD and RobustDet are robust models taking SSD as their base-models. **Right:** Under the attacks of A_{cls} and A_{loc}, the corresponding loss changes between the adversarial image and the original image.

classification. It can be seen from the figure that the bounding boxes predicted by the robust models do not have as large deviations as the classification.

The results in Fig. 5 shows that the variation of L_{cls} when applying A_{cls} attack compared to the variation of L_{loc} when applying A_{loc} attack is much larger. This also indicates that the classification module is less robust and more vulnerable to attack. These both shows that the conflict between the two images under the classification subtask is more serious than the localization subtask. The scores given in the classification part will also determine the selection of the bounding box. Therefore, this conflict will further damage the performance of the model.

4 Methodology

4.1 Overall Framework

Based on the aforementioned analyses in Sec. 3.2, the conflict between the learning of clean and adversarial images has adversary effects on the robustness of classification and localization. To address this problem, we propose a RobustDet model for defenses against adversarial attacks (Fig. 6). We detect objects through adversarially-aware convolution and use an Adversarial Image Discriminator (AID) to generate the weights for the adversarially-aware convolution kernel based on the perturbations of the input image. Furthermore, inspired by VAE [14], the image reconstruction constraint via CFR is considered for reconstructing images as clean images to facilitate the model to learn robust features.

4.2 Adversarially-Aware Convolution (AAconv)

Existing models essentially utilize the shared model parameters for learning adversarial images and clean images. This inevitably makes the model suffer

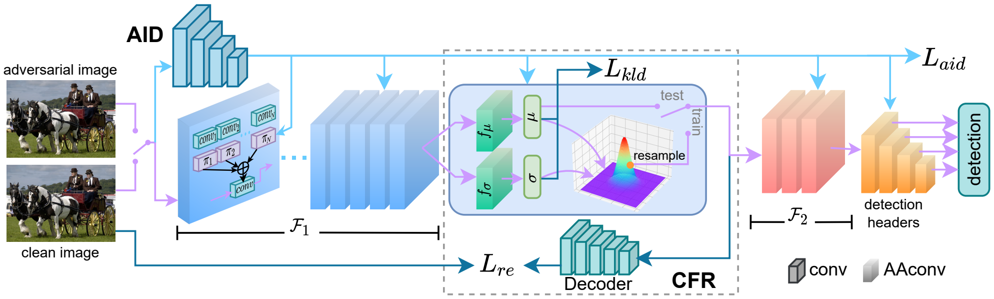


Fig. 6: The overall architecture of RobustDet based on SSD. The CFR is inserted into the SSD backbone followed by the first detection layer (conv4_3), and the two parts in front and behind this layer are named $\mathcal{F}_1$ and $\mathcal{F}_2$. The blue arrows are the data flow of AID, whose outputs are used as weights of AAconv. The purple arrows are the primary data flow when RobustDet detect objects. The teal arrows are the reconstruction data flow during training.

from a detection robustness bottleneck. There are objective distinctions between adversarial images and clean images. Admitting these distinctions rather than forcing the detector to learn these two images with the same parameters would be a better choice. Making the model explicitly distinguish these two kinds of images and detect them with different parameters will alleviate the conflict between these tasks. Inspired by [7], we propose adversarially-aware convolution in our RobustDet model to learn robust features for clean images and adversarial images.

RobustDet employs different kernels to convolve clean images and adversarial images. Different parameters will be used for different perturbed images. The generation of the convolution kernel is controlled by an adversarial image discriminator D. Before the model detects objects in an image, the adversarial image discriminator D will first generate the M-dimensional probability vector of the image $\mathcal{P} = D(x) = \{\pi_1, \pi_2, ..., \pi_M\}$. This probability vector is used as the weights to control the convolution kernels generation. Then the parameters of the finally generated convolution kernel can be write as: $\dot{\theta}^{AAconv} = \sum_{i=1}^{M} \theta_i^{AAconv} \cdot \pi_i$, where θ_i^{AAconv} denotes parameters of dynamic convolution kernels in our AAconv module, where i indicates the index of i-th convolution kernel.

RobustDet uses adversarially-aware convolutions to adaptively detect different images with different kernels and thus it can effectively learn robust features for clean and adversarial images. It not only extracts the shared features, but also can be responsible for specific features for clean and adversarial images. Therefore, it is more effective to alleviate the detection robustness bottleneck.

4.3 Adversarial Image Discriminator (AID)

The generation of the adversarially-aware convolution kernels is controlled by the adversarial image discriminator. And this module may also be attacked and give the wrong weight. Wrong weights will lead the wrong convolution kernel to be generated, which will be a disaster for the model. Accordingly, in order to improve its robustness, we employ Online Triplet Loss [25] to the adversarial image discriminator. Specifically, we consider the probability distribution of the same kind of images (*i.e.*, clean or adversarial images) as close as possible and the different kinds of images (clean or adversarial images) as far away as possible. A margin between the probability distributions of the two kinds of images outputs is introduced to strengthen the robustness of the adversarial image discriminator. Jensen-Shannon (JS) divergence [17] is utilized to measure the distance between two probability distributions, P_1 and P_2 (two distributions as an example for JS divergence): $JS\left(P_1 \| P_2\right)=\frac{1}{2} KL\left(P_1 \| \frac{P_1+P_2}{2}\right)+\frac{1}{2} KL\left(P_2 \| \frac{P_1+P_2}{2}\right)$. Overall, the AID loss is defined as follows,

$$\mathcal{L}_{aid}=\sum_{i=1}^{N_T}\left[JS\left(D\left(x_i^a\right) \| D\left(x_i^p\right)\right)-\ JS\left(D\left(x_i^a\right) \| D\left(x_i^n\right)\right)+\gamma\right]_{+}, \tag{2}$$

where x^p (x^n) is a randomly selected image from one minibatch that has the same (opposite) type (*i.e.*, clean or adversarial image) as the anchor instance x^a in one triplet, γ is the margin between x^n and x^p, N_T is the number of triplets, and $[\cdot]_+$ clips values to $[0, +\infty]$.

4.4 Consistent Features with Reconstruction (CFR)

To alleviate the negative effects of adversarial perturbation, our RobustDet aims to ensure the feature distribution of an adversarial image in the neighbourhood of its clean image. Thus, inspired by VAE [14], our RobustDet reconstructs consistent features of clean/adversarial images with clean images via our AAconvs. Assume that the output feature map of the convolutional layer after the conv4_3 layer (VGG backbone) comes from a multivariate Gaussian distribution with an diagonal covariance matrix $\mathcal{N}(\boldsymbol{\mu}=(\mu_1,...,\mu_N), \boldsymbol{\Sigma}=diag(\sigma_1^2,...,\sigma_N^2))$. For simplicity, $\boldsymbol{\sigma}=(\sigma_1^2,...,\sigma_N^2)$. Instead of directly predicting the features that are ultimately used for detection, our model predicts the mean $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$ of its feature distribution: $\boldsymbol{\mu}=f_{\boldsymbol{\mu}}(\mathcal{F}_1(x))$, $\boldsymbol{\sigma}=f_{\boldsymbol{\sigma}}(\mathcal{F}_1(x))$, where $f_{\boldsymbol{\mu}}$ and $f_{\boldsymbol{\sigma}}$ are the two layers of the model that predict the mean and standard deviations, $\mathcal{F}_1(x)$ and $\mathcal{F}_2(x)$ is two parts of VGG that split by conv4_3. From this distribution, a N-dimensional feature vector is randomly sampled as the robust feature for the input image, which is used for subsequent CFR and object detection in the training phase. Then the reconstruction loss can be defined as:

$$\mathcal{L}_{re}=\|G\left(\boldsymbol{z}\right)-x\|^2, \quad \boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{3}$$

where $\|\cdot\|^2$ indicates ℓ_2 norm, and x is the clean image. Once this feature distribution is learnt, our model can generate the similar features for an adversarial image and its clean counterpart image. Thus, in the testing phase, the predicted mean μ is directly used as the robust feature for detection.

Furthermore, similar to VAE, we also have an additional constraint to prevent the predicted distribution from collapse (*e.g.*, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are approximate to zero):

$$\mathcal{L}_{kld}=\sum_{i=1}^{N}\frac{1}{2N}\left(-\log\sigma_i^2+\mu_i^2+\sigma_i^2-1\right), \tag{4}$$

Overall, the total loss of our RobustDet is summarized as follows,

$$\mathcal{L} = \beta(\mathcal{L}_{det} + a\mathcal{L}_{aid}) + b\mathcal{L}_{re} + c\mathcal{L}_{kld}, \tag{5}$$

where β, a, b and c are the hyper-parameters.

5 Experiments

5.1 Implementation Details

Our experiments are conducted on PASCAL VOC [8] and MS-COCO [18] datasets. Mean average precision (mAP) with IoU threshold 0.5 is used for evaluating the performance of standard and robust models.

The proposed method is rooted in the one-stage detector SSD [19] with VGG16 as the backbone. Considering that Batch Normalization would increase the adversarial vulnerability [1], we make a modification on VGG16 without batch normalization layers [19]. In experiments, we use the model pre-trained on clean images for adversarial training and employ Stochastic Gradient Descent (SGD) with a learning rate of 10^{-3}, momentum 0.9, weight decay 0.0005 and batch size 32 with the multi-box loss.

For the robustness evaluation, we follow the same setting to MTD [34] and CWAT [5] for a fair comparison and use three different attacks, PGD [21], CWA [5] and DAG [32]. Among them, CWA and DAG are specifically designed for object detectors. For adversarial training, we also follow the same attack setting to MTD [34] and CWAT [5] for a fair comparison; namely, we use the PGD-20 attacker with budget $\epsilon = 8$ to generate adversarial examples [34]. And we set the margin in $\mathcal{L}_{aid}$ as $\gamma = 0.6$ and N_T is calculated from the mini-batch, and hyper-parameters in $\mathcal{L}$ as $\beta = 0.75$, $a = 3$, $b = 0.16$ and $c = 5$. RobustDet* represents RobustDet with CFR.

5.2 Detection Robustness Evaluation

In this section, we evaluate the proposed method in comparison with the state-of-the-art approaches on the PASCAL VOC and MS-COCO datasets in Tab. 1 and 2. The scenarios in MS-COCO are more complex than PASCAL VOC, and thus it is also more challenging to make the model robust on this dataset. Considering that the object detector has two tasks of classification and localization, we can use PGD to attack the classification (A_{cls}) and localization (A_{loc}). For DAG attacks, we perform 150 steps to make an effective attack. The experimental results are provided in Tab. 1 and Tab. 2.

In Tab. 1 and Tab. 2, under different datasets, in compare with standard SSD, MTD (rooted in SSD) suffers from a significant performance degradation on clean images while gaining limited robustness. For example, on the PASCAL VOC dataset, its mAP performance on clean images significantly drops from 77.5%

[2] ↓ and ↑ indicate the mAP decrease or increase compared with the baseline SSD, respectively. '-' indicates the result is not provided in the existing work.

Table 1: The evaluation results using various adversarial attack method on PASCAL VOC 2007 test set[2].

Method	**Clean**	A_{cls}	A_{loc}	**CWA**	**DAG**
SSD	**77.5**	1.8	4.5	1.2	4.9
SSD-AT(A_{cls}) [34]	46.7$^{\downarrow 30.8}$	21.8$^{\uparrow 20.0}$	32.2$^{\downarrow 30.8}$	-	28.0$^{\uparrow 23.1}$
SSD-AT(A_{loc}) [34]	51.9$^{\downarrow 25.6}$	23.7$^{\uparrow 21.9}$	26.5$^{\uparrow 22.0}$	-	17.2$^{\uparrow 12.3}$
MTD [34]	48.0$^{\downarrow 29.5}$	29.1$^{\uparrow 27.3}$	31.9$^{\uparrow 27.4}$	18.2$^{\uparrow 17.0}$	28.5$^{\uparrow 23.6}$
CWAT(PGD-10) [5]	51.3$^{\downarrow 26.2}$	22.4$^{\uparrow 20.6}$	36.7$^{\uparrow 32.2}$	19.9$^{\uparrow 18.7}$	50.3$^{\uparrow 45.4}$
RobustDet (ours)	75.4$^{\downarrow 2.1}$	41.5$^{\uparrow 40.0}$	45.2$^{\uparrow 40.7}$	42.4$^{\uparrow 41.2}$	52.0$^{\uparrow 47.1}$
RobustDet* (ours)	74.8$^{\downarrow 2.7}$	**45.9**$^{\uparrow 44.1}$	**49.1**$^{\uparrow 44.6}$	**48.0**$^{\uparrow 46.8}$	**56.6**$^{\uparrow 51.8}$

Table 2: The evaluation results using various adversarial attack method on MS-COCO 2017 test set.

Method	**Clean**	A_{cls}	A_{loc}	**CWA**	**DAG**
SSD	**42.0**	0.4	1.8	0.1	8.1
MTD [34]	24.2$^{\downarrow 17.8}$	13.0$^{\uparrow 12.6}$	13.4$^{\uparrow 11.6}$	7.7$^{\uparrow 7.6}$	-
CWAT(PGD-10) [5]	23.7$^{\downarrow 18.3}$	14.2$^{\uparrow 13.8}$	15.5$^{\uparrow 13.7}$	9.2$^{\uparrow 9.1}$	-
RobustDet (ours)	36.7$^{\downarrow 5.3}$	**20.6**$^{\uparrow 20.2}$	**19.4**$^{\uparrow 17.6}$	**20.5**$^{\uparrow 20.4}$	**24.5**$^{\uparrow 16.4}$
RobustDet* (ours)	36.0$^{\downarrow 6.0}$	20.0$^{\uparrow 19.6}$	19.0$^{\uparrow 17.2}$	19.9$^{\uparrow 19.8}$	16.5$^{\uparrow 8.4}$

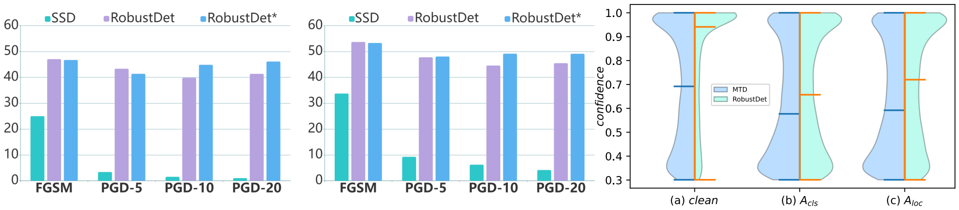
to 1.8% and 4.5% under A_{cls} and A_{loc} attacks, respectively. It also exhibits a poor robustness under CWA and DAG attacks with only 1.2% and 4.9% mAP, respectively. Besides, as for existing robust methods, MTD and CWAT only gain less than 30% mAP under A_{cls} and 40% under A_{loc} and even lose almost 30% mAP on clean images compared with baseline SSD. Instead, our proposed RobustDet not only obtains a high robustness on adversarial images, but also ensures a comparable performance with standard SSD on clean images with a slight performance decrease. On the PASCAL VOC dataset, RobustDet obtains larger than 40% mAPs on adversarial images to defense detection attacks and just loses 2.7% at most on clean images, in comparison with standard SSD. Besides, it also presents a remarkable performance on the MS-COCO dataset. For instance, RobustDet achieves 24.5% under the DAG attack with only 6% mAP decline at most on clean images (RobustDet 36.7% vs. SSD 42.0%).

5.3 Model Evaluation and Analysis

Ablation Study on L_{aid}. The adversarial image discriminator may also be attacked. Thus, the AID loss is introduced to improve its robustness. As shown in Tab. 3, without L_{aid} RobustDet has a performance decrease on both clean and adversarial images, especially on adversarial images. For instance, it drops by 4.2% mAP under the A_{cls} attack and by 4.5% mAP under the CAW attack. The absence of L_{aid} makes it easier for AID to confuse clean and adversarial images.

Table 3: The ablation study of our model under various adversarial attack method on PASCAL VOC 2007 test set.

Method	**Clean**	$\boldsymbol{A_{cls}}$	$\boldsymbol{A_{loc}}$	**CWA**	**DAG**
RobustDet w/o L_{aid}	74.9	37.3	44.9	37.9	51.8
RobustDet* w/o L_{re}	74.6	27.5	41.8	28.6	55.9
RobustDet	**75.4**	41.5	45.2	42.4	52.0
RobustDet*	74.8	**45.9**	**49.1**	**48.0**	**56.6**



(a) A_{cls} PGD attack (b) A_{loc} PGD attack (c) Confidence distribution

Fig. 7: (a) and (b): The robustness of our model under attacks with $\epsilon = 8$ using different PGD steps. (c): Under the attack on L_{cls} and L_{loc} loss, the corresponding loss changes between the adversarial image and the original image.

Ablation Study on Consistent Features with Reconstruction. We compare RobustDet (without CFR) and RobustDet* (with CFR) for the ablation study on CFR. On the PASCAL VOC dataset, as shown in Tab. 1 and 3, the detection robustness has been improved with the CFR module by 4.1% gains at least (RobustDet 47.5% vs. RobustDet* 45.9% under CWA attack) and by 5.6% at most (RobustDet 42.4% vs. RobustDet* 46.8% under CWA attack). On MS-COCO, Tab. 2 shows that RobustDet* has a lower performance than RobustDet under all the attacks. This reconstruction can be treated as VAE in VGG-16 whose capacity is relatively limited to learn so many categories, thus compromising the overall training of the model and leading to the performance degradation. Besides, CFR has two losses of L_{kld} and L_{re}. In Tab. 3, without L_{re}, RobustDet* has a significant decrease under attacks with similar performance on clean images, compared with the baseline. This indicates the model cannot effectively predict both samples into the same distribution.

Attack using Different PGD Steps. To verify the generalization ability of our model against different steps of PGD attacks, we follow the setting of MTD [34] and provide the performance of the model under various steps of PGD attack on PASCAL VOC in Fig. 7(a) and (b). For non-robust SSD, the performance decreases dramatically with the increase of iteration steps. Our model shows a strong robustness under a variety of PGD attacks with different number of steps. Combined with the experimental results of CWA and DAG in Tab. 1, it shows that our model has a promising generalization ability and can defend well even if the attacks are somewhat different from the training.

Analysis on Gradient Disentanglement. As discussed in Sec. 3.2, the detection robustness bottleneck can attribute to a conflict between learning adversarial images and clean images. It can be observed from Fig. 3, a adversarially-trained SSD model that learns adversarial (clean) images have a negative impact on the learning of clean (adversarial) images, making the loss increase. But a adversarially-trained RobustDet has almost no similar impact. The average loss variation is less than 0.1. It is also evidenced from Fig. 4 that the gradients of RobustDet on both samples are almost orthogonal. These indicate RobustDet can effectively alleviate the detection robustness bottleneck and learn both images better.

Analysis on Confidence Distribution. To further verify our RobustDet addressing the conflict, the confidence distribution of bounding boxes that the robust model MTD and our RobustDet produce on clean and adversarial images(A_{cls} and A_{loc}), respectively, in Fig. 7(c). Here we set the filtering threshold for the confidence of the bounding box to be 0.3. From which it is evident that the confidence of the MTD robust model on both the clean and adversarial images is quite low (around 0.7 on clean, 0.6 on A_{cls} and A_{loc}), which is also a manifestation of conflict. In contrast, the confidence of our proposed RobustDet model is fairly high on clean images (around 0.95, by 0.25 higher than MTD) and the confidence on adversarial images is mostly distributed in the higher part (around 0.65 on A_{cls} and 0.7 on A_{loc}). This result can also well illustrate that our method can effectively alleviate the conflict and the detection robustness bottleneck.

6 Conclusion

In this work, we investigate the detection robustness bottleneck that the object detector discards a portion of its performance on the clean image while gaining a very limited robustness from adversarial training. Empirical analysis from the loss change and gradient interference indicate that the detection robustness bottleneck is mainly attributed to the conflict between the object detector in learning clean images and adversarial images. It is hard for object detectors to learn both images well, so it needs a learning trade-off between them.

In terms of the detection robustness bottleneck on both clean images and adversarial images, we propose the RobustDet method based on adversarially-aware convolution. RobustDet utilizes an Adversarial Image Discriminator (AID) to generate different weights to clean images and adversarial images, which guides the generation of adversarially-aware convolutional kernels to adaptively learn robust features. RobustDet also employs the Consistent Features with Reconstruction (CFR) to make the features of clean and adversarial images in the same distribution and empower the model to reconstruct the adversarial image into a clean image. This can further enhance the detection robustness. Besides, experimental results show that our method can effectively alleviate the detection robustness bottleneck. It is demonstrated that our method can significantly improve the robustness of the model without losing the performance on clean images.

Acknowledgement

This work was supported in part by NSFC (No.62006253, U21A20470, 61876224), National Key R&D Program of China (2021ZD0111601).

References

1. Benz, P., Zhang, C., Kweon, I.S.: Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In: International Conference on Computer Vision (ICCV). pp. 7818–7827 (2021) 11
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229 (2020) 4
3. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy. pp. 39–57 (2017) 2, 4
4. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (ECCV). pp. 833–851 (2018) 2
5. Chen, P., Kung, B., Chen, J.: Class-aware robust adversarial training for object detection. In: Computer Vision and Pattern Recognition (CVPR). pp. 10420–10429 (2021) 2, 5, 11, 12
6. Chen, S., Cornelius, C., Martin, J., Chau, D.H.P.: Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector. In: Machine Learning and Knowledge Discovery in Databases - European Conference (ECML). pp. 52–68 (2018) 2, 5
7. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Computer Vision and Pattern Recognition (CVPR). pp. 11027–11036 (2020) 9
8. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015) 4, 11
9. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021) 2, 4
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015) 2, 4
11. Gowal, S., Rebuffi, S., Wiles, O., Stimberg, F., Calian, D.A., Mann, T.: Improving robustness using generated data. CoRR **abs/2110.09468** (2021) 3
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 2, 4
13. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019) 4
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014) 8, 10
15. Liang, S., Wu, B., Fan, Y., Wei, X., Cao, X.: Parallel rectangle flip attack: A query-based black-box attack against object detection. In: International Conference on Computer Vision (ICCV). pp. 7697–7707 (2021) 2
16. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Computer Vision and Pattern Recognition (CVPR). pp. 1778–1787 (2018) 2, 4
17. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Information Theory **37**(1), 145–151 (1991) 10

18. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014) 4, 11
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: European Conference on Computer Vision (ECCV). pp. 21–37 (2016) 4, 11
20. Liu, X., Yang, H., Liu, Z., Song, L., Chen, Y., Li, H.: DPATCH: an adversarial patch attack on object detectors. In: Workshop on Thirty-Third AAAI Conference on Artificial Intelligence (AAAI) (2019) 5
21. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018) 2, 4, 11
22. Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., Kohli, P.: Adversarial robustness through local linearization. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 13824–13833 (2019) 2, 4
23. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 91–99 (2015) 2, 4
24. Sarkar, S., Bansal, A., Mahbub, U., Chellappa, R.: UPSET and ANGRI : Breaking high performance image classifiers. CoRR **abs/1707.01159** (2017) 2, 4
25. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015) 10
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015) 4
27. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Trans. Evolutionary Computation **23**(5), 828–841 (2019) 4
28. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014) 4
29. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I.J., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: Attacks and defenses. In: International Conference on Learning Representations (ICLR) (2018) 2, 4
30. Wei, X., Liang, S., Chen, N., Cao, X.: Transferable adversarial attacks for image and video object detection. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 954–960 (2019) 2, 5
31. Xiao, C., Li, B., Zhu, J., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 3905–3911 (2018) 2
32. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.L.: Adversarial examples for semantic segmentation and object detection. In: International Conference on Computer Vision (ICCV). pp. 1378–1387 (2017) 2, 5, 11
33. Yang, J., Jiang, Y., Huang, X., Ni, B., Zhao, C.: Learning black-box attackers with transferable priors and query feedback. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 4
34. Zhang, H., Wang, J.: Towards adversarially robust object detection. In: International Conference on Computer Vision (ICCV). pp. 421–430 (2019) 1, 2, 3, 5, 6, 11, 12, 13

35. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (ICML). pp. 7472–7482 (2019) 3
36. Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., Kankanhalli, M.S.: Attacks which do not kill training make adversarial learning stronger. In: International Conference on Machine Learning (ICML). pp. 11278–11287 (2020) 2, 4
37. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239 (2017) 2