Adversarial Attack and Defense of YOLO Detectors in Autonomous Driving Scenarios

Jung Im Choi[1] and Qing Tian[2]

***Abstract*— Visual detection is a key task in autonomous driving, and it serves as a crucial foundation for self-driving planning and control. Deep neural networks have achieved promising results in various visual tasks, but they are known to be vulnerable to adversarial attacks. A comprehensive understanding of deep visual detectors' vulnerability is required before people can improve their robustness. However, only a few adversarial attack/defense works have focused on object detection, and most of them employed only classification and/or localization losses, ignoring the objectness aspect. In this paper, we identify a serious objectness-related adversarial vulnerability in YOLO detectors and present an effective attack strategy targeting the objectness aspect of visual detection in autonomous vehicles. Furthermore, to address such vulnerability, we propose a new objectness-aware adversarial training approach for visual detection. Experiments show that the proposed attack targeting the objectness aspect is 45.17% and 43.50% more effective than those generated from classification and/or localization losses on the KITTI and COCO_traffic datasets, respectively. Also, the proposed adversarial defense approach can improve the detectors' robustness against objectness-oriented attacks by up to 21% and 12% mAP on KITTI and COCO_traffic, respectively.**

I. INTRODUCTION

Over the past decade, deep learning has revolutionized various visual computing areas, such as object detection [1], [2], image classification [3], [4], image captioning [5]. Vision-based self-driving cars can take advantage of deep neural networks to better detect objects (e.g., cars, pedestrians, road signs, etc.) [6], [7]. However, deep learning models can easily fall victim to adversarial attacks [8]–[12]. While numerous adversarial robustness studies have targeted classification models [13]–[15], few have focused on the more challenging task of object detection, especially in autonomous driving scenarios.

Unlike image classification which only requires one class label for an entire image, object detection involves three types of outputs for each region of interest in an input image: (1) the objectness (the probability of the associated bounding box containing an object), (2) the bounding box location, and (3) the class label. Due to the high complexity, object detectors can be more difficult to attack and defend compared to classification. Thus, a deeper and more holistic understanding of object detectors' vulnerability is needed before we can improve their robustness. Some research has been carried out targeting two-stage detectors [1], [16]. Most of them (e.g., [10], [17]) consider only the tasks in the second stage (i.e., localization and classification), with the objectness aspect in the first stage ignored. Compared to the attack works on two-stage detectors, not many efforts have been made to investigate one-stage detectors (e.g., YOLOs [2]), which is more suitable for autonomous driving scenarios due to its high speed and efficiency.

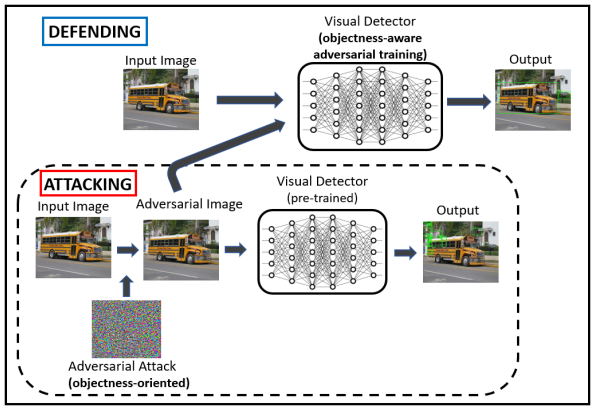

Fig. 1: A schematic overview of the proposed object-oriented attacking and defending strategies. Our objectness-oriented attacking and defending approaches are more effective than the existing methods that utilize only classification and/or localization losses.

In this paper, we propose a more effective attack strategy that takes into account the objectness aspect of object detection in self-driving cars. Our approach is designed for the state-of-the-art YOLO detector (i.e., YOLOv4 [18]), although it is likely to be applicable to other detectors as well. To defend the designed attack, we also propose a new objectness-aware adversarial training strategy. Figure 1 provides a schematic overview of our proposed attacking and defending approaches. Our experiments on the KITTI [19] and COCO_traffic (a subset of COCO [20]) datasets demonstrate that attacks based on the objectness loss are more effective than those based on other task losses for object detection in self-driving scenarios. In summary, the main contributions of this paper are as follows:

- We identify a serious adversarial vulnerability in YOLO detectors by evaluating the impact of adversarial attacks sourced from the multiple task losses (i.e., of objectness, localization, and classification) and present an effective attack approach targeting the objectness aspect of YOLO in autonomous driving. The objectness-oriented attacks can be more effective than those generated from classification and/or localization losses.

[1]Jung Im Choi is a PhD candidate in Data Science, Bowling Green State University, Bowling Green, OH 43403, USA. choij@bgsu.edu

[2]Qing Tian (corresponding author) is an assistant professor at the Department of Computer Science, Bowling Green State University, Bowling Green, OH 43403, USA. qtian@bgsu.edu

- Based on our analysis of objectness-related vulnerability, we also propose a new adversarial training-based strategy utilizing the objectness loss. The model trained with objectness-based attacks can be more robust than those utilizing other two task losses against the task-oriented attacks.
- Our objectness-aware adversarial training can help alleviate the potential conflicts/misalignment of the directions of the image gradients derived from different task losses in object detection.

II. RELATED WORK

A. Adversarial Attacks for Visual Detection

Several studies have shown that slightly perturbing an original image can fool a target model to produce wrong predictions [8], [9]. While most of the existing adversarial attacks are designed for classification models [13]–[15], relatively few works have focused on the more challenging object detection task [10], [11]. Depending on whether the attacker has access to the victim model's internal detail (e.g., parameters), adversarial attacks can be categorized into white-box [8]–[10] or black-box [21], [22] attacks. In this paper, we will consider white-box attacks. Xie *et al*. [10] extended the attack generation method from classification to object detection by using the dense adversary generation. Lu *et al*. [11] created adversarial examples for detectors by generating many proposals and randomly assigning a label for each proposal region. However, [10] and [11] used only the classification loss to generate their adversarial examples for given target detectors. Li *et al*. [23] presented a robust adversarial perturbation method to attack the Region Proposal Network (RPN) by incorporating both classification and localization losses. Zhang *et al*. [24] identified an asymmetric role of classification and localization losses and find that the image gradients from the two losses are misaligned, which can make effective adversary generation difficult. Unlike the existing attack methods which used classification and/or localization loss, we propose to leverage objectness loss to generate effective adversarial examples for visual detection in self-driving scenarios which will be shown to be more effective.

B. Adversarial Training for Visual Detectors

Adversarial training [8], [25] is one of the most effective approaches to defend deep learning models against adversarial attacks [26]. Since its introduction by Szegedy *et al*. [3], many effective defense methods [25], [27] for classification have been proposed. While those approaches greatly increased the adversarial robustness of deep classifiers, not many efforts have been made to improve the robustness of object detectors, especially in safety-critical autonomous driving scenarios. Recently, Zhang *et al*. [24] have generalized the adversarial training framework from classification to object detection. They utilized the task-oriented domain constraint in adversarial training to improve the robustness of object detectors. Chen *et al*. [28] presented Det-AdvProp which employs separate batch normalization layer for clean images and adversarial examples to address the mAP score decrease of adversarially trained model on clean images. However, both of the above-mentioned approaches considered only localization and classification losses. In this paper, we will analyze and leverage all of the three task losses in the YOLOv4 detector (i.e., objectness, localization, and classification losses) during adversarial training. By explicitly considering the objectness aspect, our adversarial training method can better align the image gradients sourced from different objective components and thus lead to more robust visual detectors.

III. METHODOLOGY

A. Adversarial Vulnerability in YOLO Detectors

In this subsection, we examine various aspects of the YOLO detector for potential adversarial susceptibility and identify a serious vulnerability in the objectness component. We take YOLOv4 [18], a state-of-the-art variant in the YOLO family, as the base model. Compared to two-stage detectors, it is more efficient and suitable for vision-based self-driving systems. While two-stage detectors propose regions of interest (ROI) before classifying and regressing bounding boxes, YOLOv4 tackles classification and regression in a single stage without any ROI proposal step. In YOLOv4, the overall loss for each box prediction consists of three components, i.e., objectness, localization, and classification losses:

$$\mathcal{L}(\mathbf{x}, y, \mathbf{b}; \theta) = \mathcal{L}_{OBJ}(\mathbf{x}, \mathbf{b}; \theta) + \mathcal{L}_{LOC}(\mathbf{x}, \mathbf{b}; \theta) + \mathcal{L}_{CLS}(\mathbf{x}, y; \theta), \tag{1}$$

where $\mathbf{x}$ is the training sample, y and $\mathbf{b}$ are the ground-truth class label and bounding box, θ represents the model parameters, and $\mathcal{L}(\cdot)$ indicates a loss function. The subscripts stand for the three aspects in object detection (e.g., OBJ for objectness, LOC for localization, and CLS for classification). While most existing works exploring adversarial robustness (e.g., [17], [24], [28]) have been focused on utilizing the classification and/or localization losses, we argue that effective attacks for object detection should consider all of the three aspects, including the usually ignored objectness loss. The objectness loss, which is the main focus of the paper, can be divided into two parts: the object (obj) part and the no_object (no_obj) part:

$$\mathcal{L}_{OBJ}(\mathbf{x}, \mathbf{b}; \theta) = \mathcal{L}_{obj}(\mathbf{x}, \mathbf{b}; \theta) + \lambda_{no\_obj}\, \mathcal{L}_{no\_obj}(\mathbf{x}, \mathbf{b}; \theta), \tag{2}$$

where

$$\mathcal{L}_{obj}(\mathbf{x}, \mathbf{b}; \theta) = -\sum_{k=1}^{K} I_k^{obj} \Big[\hat{C}_k \log(C_k) + (1 - \hat{C}_k) \log(1 - C_k) \Big], \tag{3}$$

and

$$\mathcal{L}_{no\_obj}(\mathbf{x}, \mathbf{b}; \theta) = -\sum_{k=1}^{K} I_k^{no\_obj} \Big[\hat{C}_k \log(C_k) + (1 - \hat{C}_k) \log(1 - C_k) \Big]. \quad (4)$$

The objectness score $\hat{C}_k \in [0, 1]$ can be considered as the network's confidence in a given bounding box containing an object. $\lambda_{no\_obj} \in [0, 1]$ is a hyperparameter penalizing no-object bounding boxes (according to the ground truth), K is the number of predicted bounding boxes, and I_k^{obj} represents whether the k-th bounding box contains an object (i.e., I_k^{obj} = 1) or not (i.e., I_k^{obj} = 0). Similarly, $I_k^{no\_obj}$ = 1 denotes the k-th bounding box has no object.

The localization loss L_{CLS}, responsible for finding the bounding-box coordinates, is based on the Complete Intersection over Union (CIoU) loss [29]:

$$\mathcal{L}_{LOC}(\mathbf{x}, \mathbf{b}; \theta) = \mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(\hat{\mathbf{b}}, \mathbf{b})}{c^2} + \alpha v, \quad (5)$$

where IoU (Intersection over Union) is an evaluation metric used to measure overlap between two bounding boxes, $\rho(\hat{\mathbf{b}}, \mathbf{b})$ represents the Euclidean distance of central points of the prediction box $\hat{\mathbf{b}}$ and the ground truth $\mathbf{b}$, c is the diagonal distance of the smallest enclosing box covering $\hat{\mathbf{b}}$ and $\mathbf{b}$, α is a positive trade-off hyperparameter, and v is the consistency measure of aspect ratio.

The classification loss L_{CLS}, responsible for the class-score prediction $\hat{\mathbf{p}}_k(i)$, is defined as:

$$\mathcal{L}_{CLS}(\mathbf{x}, y; \theta) = -\sum_{k=1}^{K} I_k^{obj} \sum_{i \in classes} \Big[\hat{p}_k(i) \log(p_k(i)) + (1 - \hat{p}_k(i)) \log(1 - p_k(i)) \Big]. \quad (6)$$

To see how adversarial samples derived from the different losses are distributed, we project their high-dimensional representations into a 2D space by t-SNE and show the task gradient domains in Figure 2. Given a clean image, each dot in the figure represents one adversarial example derived from one of the three task losses (i.e., $\mathcal{L}_{OBJ}$, $\mathcal{L}_{LOC}$, and $\mathcal{L}_{CLS}$). Interestingly, we observe that the objectness-based gradient domain (shown in blue) partially overlaps with both the classification-based (shown in green) and localization-based (shown in orange) gradient domains while there is no overlapping between the classification and localization domains[1]. Non-overlapping regions of the task gradient domains reflect the inconsistent directions of the image gradients derived from the task losses (we will refer to this issue as 'misaligned task gradients' for the rest of the paper). It is worth mentioning that Zhang *et al*. [24] found similar issues in general object detection considering only the classification and localization domains. They attempted

[1]From a probabilistic point of view, no overlapping between the two does not mean that no attack can simultaneously handle the two aspects. However, the chance is low (it may be possible with more samples).

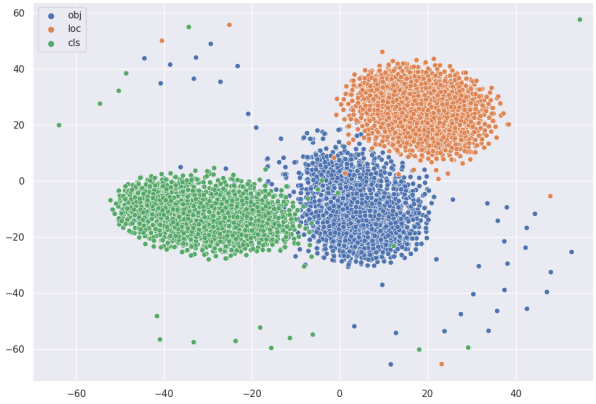

Fig. 2: Visualization of adversarial examples generated from different task losses using t-SNE. Different colors encode the task losses used for generating adversarial examples (blue: objectness, orange: localization, green: classification loss). This is a typical example from the KITTI dataset for autonomous driving.

to avoid such conflicts/misalignment by choosing one of the two types of attacks (either classification or localization oriented) each time. However, an attack from one domain is likely to ignore the other aspect, especially in our autonomous driving scenarios where the two regions (orange and green) are far apart (e.g., Figure 2). The chance is low that an adversarial example derived from the classification or localization loss can simultaneously attack both aspects. The objectness domain, lying between the classification and localization domains, helps join the other two aspects and attract more attention to the middle regions where an attack has a better chance to target all three aspects. Figure 2 visually demonstrates the objectness-related vulnerability in the YOLO detector and inspires us to employ the objectness loss to generate more effective adversarial attacks for object detection.

B. Objectness-Oriented Adversarial Attack for Visual Detection

Motivated by the preceding analysis, we propose to consider all the three loss/vulnerability aspects and utilize the objectness loss to craft adversarial attacks. Given a trained deep learning model f and an input $\mathbf{x}$, generating an adversarial example $\mathbf{x}'$ can be formulated as:

$$||\mathbf{x}' - \mathbf{x}||_p < \epsilon \quad s.t. \quad f(\mathbf{x}') \neq f(\mathbf{x}), \quad (7)$$

where $||\cdot||_p$ denotes the distance (L_p norm) between two data sample. The choice of the norm, p, determines the type of limitations placed on the adversary generation. As in many previous works (e.g., [24] [28]), we utilize L_∞ as a distortion measure, and it measures the maximum absolute change to any pixel. The attack budget ϵ bounds the maximum perturbation in terms of L_∞. Through exploring in the original data space, this optimization process tries to produce an incorrect prediction while being subject to a constraint on the perturbation magnitude. To generate adversarial examples, we take gradients of the corresponding losses (i.e., objectness, localization, and classification losses)

with respect to the input and modify the input along the gradient direction:

$$\mathbf{x}'_{obj,PGD} = \mathcal{P}(\mathbf{x} + \alpha \cdot sign(\bigtriangledown_{\mathbf{x}} \mathcal{L}_{OBJ}(\mathbf{x}, \mathbf{b}; \theta))),$$
$$\mathbf{x}'_{loc,PGD} = \mathcal{P}(\mathbf{x} + \alpha \cdot sign(\bigtriangledown_{\mathbf{x}} \mathcal{L}_{LOC}(\mathbf{x}, \mathbf{b}; \theta))),$$
$$\mathbf{x}'_{cls,PGD} = \mathcal{P}(\mathbf{x} + \alpha \cdot sign(\bigtriangledown_{\mathbf{x}} \mathcal{L}_{CLS}(\mathbf{x}, y; \theta))), \quad (8)$$

where $\mathcal{P}$ projects the perturbed example to a ϵ-radius ball $\{\mathbf{x} \,|\, ||\mathbf{x}' - \mathbf{x}||_\infty \leq \epsilon\}$ to ensure the perceptual similarity, and α represents the step size. Note that if the number of iterations in PGD [25] equals to one, it becomes the FGSM method [8]. We will explore both in our experiments.

C. Objectness-Aware Adversarial Training

On the defense side, to improve the adversarial robustness of object detectors, we develop a new objectness-aware adversarial training approach explicitly utilizing the objectness aspect mentioned in the previous subsections. Unlike prior works [24], [28] where models were trained with attacks generated from the localization and/or classification losses only, our approach considers all dimensions of the object detection output and uses all the heterogeneous sources of losses (i.e., objectness, localization, and classification tasks) in the adversary generation and adversarial training. The overall objective of the proposed adversarial training can be defined as follows:

$$\arg\min_{\theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}; y, \mathbf{b} \sim \mathcal{B}(\mathbf{x})} \mathcal{L}(\mathbf{x}, \{y, \mathbf{b}\}; \theta) + \mathcal{L}(\underline{\mathbf{x}}, \{y, \mathbf{b}\}; \theta), \quad (9)$$

where $\underline{\mathbf{x}}$ is the strongest one among the adversarial examples generated from the three task losses in terms of the overall loss. Algorithm 1 shows the details of the proposed adversarial training algorithm.

Algorithm 1 Objectness-Aware Adversarial Training

Input: Dataset $\mathcal{D}$, Training epochs N, Batch size B, Perturbation bounds ϵ

for epoch = 1 **to** N **do**
  for random batch $\{\mathbf{x}^i, \{y^i, \mathbf{b}^i\}\}_{i=1}^{B} \sim \mathcal{D}$ **do**
    $(\mathbf{x}'_{obj})^i = \mathcal{P}(\mathbf{x}^i + \epsilon \cdot sign(\bigtriangledown_{\mathbf{x}} \mathcal{L}_{OBJ}(\mathbf{x}^i, \mathbf{b}^i; \theta)))$
    $(\mathbf{x}'_{loc})^i = \mathcal{P}(\mathbf{x}^i + \epsilon \cdot sign(\bigtriangledown_{\mathbf{x}} \mathcal{L}_{LOC}(\mathbf{x}^i, \mathbf{b}^i; \theta)))$
    $(\mathbf{x}'_{cls})^i = \mathcal{P}(\mathbf{x}^i + \epsilon \cdot sign(\bigtriangledown_{\mathbf{x}} \mathcal{L}_{CLS}(\mathbf{x}^i, y^i; \theta)))$
    Choose $\underline{\mathbf{x}}^i$ that leads to the max total loss:
    $\underline{\mathbf{x}}^i = \arg\max_{\tilde{\mathbf{x}}^i \in \{(\mathbf{x}'_{obj})^i, (\mathbf{x}'_{loc})^i, (\mathbf{x}'_{cls})^i\}} \mathcal{L}(\tilde{\mathbf{x}}^i, \{y^i, \mathbf{b}^i\}; \theta)$
    Perform an adversarial training step w.r.t. θ:
    $\arg\min_\theta \mathcal{L}(\mathbf{x}^i, \{y^i, \mathbf{b}^i\}; \theta) + \mathcal{L}(\underline{\mathbf{x}}^i, \{y^i, \mathbf{b}^i\}; \theta)$
  end for
end for

Output: Learned model parameter θ

We first search for the most detrimental adversarial perturbation from the three candidates. Then, we update the model parameters to reduce the overall loss on both a clean example and the selected adversarial example $\underline{\mathbf{x}}$. We use a similar max-max scheme to Zhang *et al*. [24] and keep the adversarial example (out of three) that maximizes the overall loss. However, our approach is different from their work in that we include the critical objectness component. In their work, each time, only one type of attack is chosen (e.g., either classification or localization). As shown in Figure 2, there can be hardly any overlapping between the two task domains in autonomous driving scenarios that we care about. It follows that choosing one type of attack likely means ignoring the other vulnerability aspect. Improved adversarial robustness towards one task domain does not necessarily reflect the overall model robustness (the adversarial robustness towards the other aspect may be reduced). Smart attackers may attack both aspects simultaneously, and we need a more comprehensive defense strategy. Including the objectness component helps fill in the missing piece. Utilizing the adversarial example derived from the objectness loss can better alleviate the issue of misaligned task gradients (Figure 2). Its generated attack can potentially be more detrimental and adversarial training taking into consideration such examples is more helpful to improve the model robustness. We adapt FGSM-based adversarial training combined with random initialization [30], which is as effective as PGD-based training but has significantly lower computational cost. Experimental results will demonstrate our strategy's efficacy in the following section (Sec. IV).

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

In our experiments, we use the one-stage object detector, YOLOv4 [18] as the base model. For its backbone feature extraction network, CSPDarkNet53 is used, which is pretrained on COCO2017 [20] dataset. We conduct our experiments on both the KITTI [19] and COCO_traffic datasets. The latter is a subset of MS-COCO [20]. For the KITTI dataset, we follow the same convention for combining the categories and splitting the dataset as in [31]. To be more specific, there are three categories for the KITTI dataset: car, cyclist, and pedestrian. The 7,481 training images are split in half into a training set and a validation set since the test images do not have labels. For the COCO_traffic dataset, it has 8 categories related to autonomous driving (i.e., person, bicycle, car, motorcycle, bus, truck, traffic light and stop sign). There are 71,536 training images and 3,028 test images. For both datasets, the YOLOv4 is trained to convergence (50 epochs) using an Adam optimizer, with an initial learning rate of 0.001 and a batch size of 8. Then, we evaluate the models using the Pascal VOC mean average precision (mAP) metric with the IoU threshold set as 0.5.

Fast Gradient Sign Method (FGSM) [8] and Projected Gradient Descent (PGD) [25] are two well-known attacking methods originally designed for classification tasks. We extend them to object detection scenarios through combining them with the losses in Equation 8. For each type of attack, a range of attacking strengths $\epsilon \in \{2, 4, 6, 8\}$ are considered. For PGD, we use a step size $\alpha = 1$ and the number of iterations $T = 10$. For adversarial training, we adapt the FGSM algorithm (with $\epsilon = 4$). It generates the adversarial examples which will be used as input to adversarial training.

Method	Att.Size	$\mathcal{A}_{loc}$	$\mathcal{A}_{cls}$	$\mathcal{A}_{obj+loc+cls}$	$\mathcal{A}_{obj}$
FGSM	$\epsilon = 2$	-0.98	-0.97	-8.42	**-10.49**
	$\epsilon = 4$	-3.20	-3.15	-13.88	**-16.85**
	$\epsilon = 6$	-6.08	-5.65	-17.88	**-22.53**
	$\epsilon = 8$	-10.44	-9.65	-22.04	**-27.31**
PGD-10	$\epsilon = 2$	-1.22	-0.87	-42.44	**-42.64**
	$\epsilon = 4$	-4.11	-2.64	-51.47	**-51.67**
	$\epsilon = 6$	-7.00	-5.91	-54.17	**-54.39**
	$\epsilon = 8$	-10.66	-9.59	-55.48	**-55.83**

TABLE I: Model performance degradation under various attack strengths for the task-oriented attacks using FGSM and 10-step PGD on KITTI. $\mathcal{A}_{loc}, \mathcal{A}_{cls}, \mathcal{A}_{obj+loc+cls}, and\ \mathcal{A}_{obj}$ denote the attacks sourced from corresponding task losses (i.e., localization, classification, overall, and objectness losses). The objectness-oriented attacks decrease the mAP most. The clean mAP on KITTI is 80.10%.

Method	Att.Size	$\mathcal{A}_{loc}$	$\mathcal{A}_{cls}$	$\mathcal{A}_{obj+loc+cls}$	$\mathcal{A}_{obj}$
FGSM	$\epsilon = 2$	-0.31	-0.22	-7.42	**-7.49**
	$\epsilon = 4$	-1.01	-0.95	-9.40	**-9.74**
	$\epsilon = 6$	-1.85	-1.86	-10.54	**-10.97**
	$\epsilon = 8$	-3.30	-3.22	-12.33	**-12.45**
PGD-10	$\epsilon = 2$	-0.19	-0.15	-36.55	**-37.55**
	$\epsilon = 4$	-0.70	-0.77	-43.84	**-43.93**
	$\epsilon = 6$	-1.88	-2.26	-45.31	**-45.69**
	$\epsilon = 8$	-3.24	-3.58	-46.88	**-47.08**

TABLE II: Comparison of impact of different task loss-based attacks on model performance (mAP) under various attack sizes using FGSM and PGD-10 on COCO_traffic. $\mathcal{A}_{loc}$, $\mathcal{A}_{cls}$, $\mathcal{A}_{obj+loc+cls}$, and $\mathcal{A}_{obj}$ are defined similarly as in Table I. The clean mAP on COCO_traffic is 66.10%.

B. Quantitative Analysis of Attacks

In this subsection, we investigate model vulnerability to a variety of task-oriented attacks. Table I and Table II demonstrate the mAP changes on the two datasets due to the attacks of different sources ($\mathcal{A}_{loc}, \mathcal{A}_{cls}, \mathcal{A}_{obj}$), types (FGSM, PGD) and strengths ($\epsilon \in \{2, 4, 6, 8\}$). In our experiments, the attack strengths are determined by different maximum perturbations, where larger perturbation budget indicates stronger attack.

According to the results, we observe that performance degradation due to the adversarial perturbations is different across various task losses, attack types and strengths. Attacks sourced from the objectness loss cause the most performance degradation in both the FGSM and PGD cases. On both datasets, the gaps can be large between the objectness-aware and objectness-unaware cases (e.g., 45% and 43% for PGD-10 when $\epsilon = 8$). In addition, as expected, stronger attacks make the model performance drop more.

C. Qualitative Analysis of Attacks

The qualitative impact of the three attacks (i.e., $\mathcal{A}_{obj}$, $\mathcal{A}_{loc}$, and $\mathcal{A}_{cls}$) on the model performance are illustrated in Figure 3 for a KITTI example and Figure 4 for a COCO_traffic example. Comparing with the detection results on the clean image, we observe that many false positives are produced under the objectness-oriented attack on both datasets. Sometimes, the attacks from the other two losses can also result in some false positives, but the number is much lower. These qualitative results intuitively demonstrate that the objectness-based attack can be more effective than the other two. One possible reason is that both classification and localization depend heavily on objectness estimation. Object-oriented attacks can impact both other aspects simultaneously (Figure 2). In addition, we can see that our PGD-based objectness attacking strategy is more effective than the FGSM-based one on both datasets, which agrees with what was previously found in the quantitative analysis.

D. Adversarial Training Results

In this subsection, we further investigate whether a model adversarially trained with the objectness-oriented attacks can be more robust than those trained without. The experiments were also conducted on the KITTI and COCO_traffic datasets. Particularly, we are interested in the robustness of the following YOLOv4 models trained with attacks derived from different task losses: the model normally trained with only clean images ($\mathcal{M}_{STD}$), the model adversarially trained using the overall loss ($\mathcal{M}_{ALL}$), the model trained with the algorithm in [24] ($\mathcal{M}_{MTD}$, where MTD represents multi-task domain), the models trained with adversarial examples solely from one kind of loss ($\mathcal{M}_{LOC}$, $\mathcal{M}_{CLS}$, and $\mathcal{M}_{OBJ}$, where the subscripts respectively represent localization, classification, and objectness), and the model obtained from Algorithm 1 ($\mathcal{M}_{OA}$). We test these models' robustness under the PGD-based attacks induced from the objectness loss and the overall loss (as these two attacks are shown to be much more destructive in Table I and Table II). The results are reported in Table III.

As we can see from Table III, our proposed adversarial defense approaches considering the objectness aspect ($\mathcal{M}_{OBJ}$ and $\mathcal{M}_{OA}$) lead to more robustness than those do not (e.g., $\mathcal{M}_{LOC}$ and $\mathcal{M}_{CLS}$). For example, on KITTI, the mAP of $\mathcal{M}_{OBJ}$ is up to 21% higher than that of $\mathcal{M}_{STD}$, and on COCO_traffic, the mAP of $\mathcal{M}_{OA}$ is improved by up to 12.6% from the baseline under the objectness-oriented PGD attack. While we can see that the objectness-awareness plays a critical role in both KITTI and COCO_traffic cases, the models with the best performance on the two datasets are different (i.e., $\mathcal{M}_{OBJ}$ for KITTI and $\mathcal{M}_{OA}$ for COCO_traffic). One possible reason is that the misalignment of gradients sourced from classification and localization is more serious on KITTI than on COCO_traffic. It follows that improving one kind of robustness (classification/localization) will be more likely to hurt the other (localization/classification) on KITTI. In this case, we can be better off during adversarial training by ignoring the two disjoint task domains and focusing only on the objectness domain that 'overlaps' the other two.

From Table III, we can also see that our objectness-aware solution ($\mathcal{M}_{OA}$ or $\mathcal{M}_{OBJ}$) outperforms $\mathcal{M}_{MTD}$ [24]. The gaps are more obvious on KITTI than on COCO_traffic. This can also be explained by the hypothesis that the problem of the misalignment of task gradients is less serious on COCO_traffic than on KITTI.

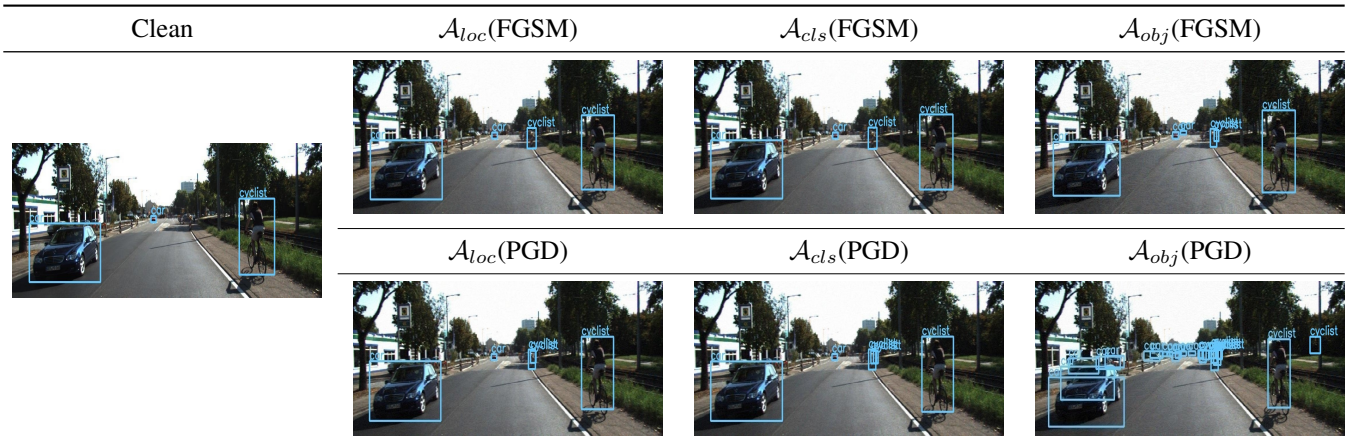

Fig. 3: Visualization of detection results under different task-loss-based attacks using FGSM (top row) and 10-step PGD (bottom row) with $\epsilon = 4$ on a KITTI example. Best viewed when zoomed in.

Fig. 4: Visualization of detection results under different task-loss-based attacks using FGSM (top row) and 10-step PGD (bottom row) with $\epsilon = 4$ on a COCO_traffic example. Best viewed when zoomed in.

Model	$\mathcal{A}_{obj}$	$\mathcal{A}_{obj+loc+cls}$
$\mathcal{M}_{STD}$	28.43	28.63
$\mathcal{M}_{ALL}$	39.65	40.65
$\mathcal{M}_{MTD}$	36.13	35.94
$\mathcal{M}_{LOC}$	37.86	37.61
$\mathcal{M}_{CLS}$	39.29	39.70
$\mathcal{M}_{OBJ}$	**49.43**	**48.83**
$\mathcal{M}_{OA}$	42.26	41.86

(a) KITTI

Model	$\mathcal{A}_{obj}$	$\mathcal{A}_{obj+loc+cls}$
$\mathcal{M}_{STD}$	22.17	22.29
$\mathcal{M}_{ALL}$	34.58	33.44
$\mathcal{M}_{MTD}$	33.26	33.20
$\mathcal{M}_{LOC}$	33.23	32.10
$\mathcal{M}_{CLS}$	31.71	31.58
$\mathcal{M}_{OBJ}$	33.30	32.69
$\mathcal{M}_{OA}$	**34.77**	**33.61**

(b) COCO_traffic

TABLE III: mAP comparison of various adversarially trained YOLO models under PGD-10 attacks on (a) KITTI and (b) COCO_traffic validation sets. Depending on which losses the adversarial examples are originated from, the following adversarially trained models are obtained for each dataset: $\mathcal{M}_{STD}$, $\mathcal{M}_{ALL}$, $\mathcal{M}_{MTD}$, $\mathcal{M}_{LOC}$, $\mathcal{M}_{CLS}$, $\mathcal{M}_{OBJ}$, $\mathcal{M}_{OA}$ (the notation is explained in the text). $\epsilon = 4$.

V. CONCLUSION

In this paper, we have identified a serious vulnerability of YOLO detectors in autonomous driving scenarios. The vulnerability comes from the objectness aspect of the object detection. To better understand and to remedy the vulnerability, we have proposed: (1) a new attack strategy targeting the objectness loss in object detection, and (2) an objectness-aware adversarial training framework to enhance the robustness of the detector. Additionally, we find that the direction of the image gradient derived from the objectness loss is more consistent with those from the two other losses. Adversarial training considering the objectness aspect can potentially alleviate the problem of misaligned task gradients. Our experiments on the KITTI and COCO_traffic datasets demonstrate that the objectness-oriented attack approach is much more effective than the attacks derived from the other two detection losses. Furthermore, the proposed adversarial defense approaches explicitly paying attention to the objectness aspect can improve the detector's robustness by large margins on both datasets.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation (NSF) under Award No. 2153404. This work would not have been possible without the computing resources provided by the Ohio Supercomputer Center.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *the 28th International Conference on Neural Information Processing Systems - Volume 1*, p. 91–99, 2015.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, ICLR. 2014.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015.

[5] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659, 2016.

[6] H. Fujiyoshi, T. Hirakawa, and T. Yamashita, "Deep learning-based image recognition for autonomous driving," *IATSS Research*, vol. 43, no. 4, pp. 244–252, 2019.

[7] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

[10] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *the IEEE International Conference on Computer Vision*, pp. 1378–1387, 2017.

[11] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," *arXiv preprint arXiv:1712.02494*, 2017.

[12] Q. Tian, T. Arbel, and J. J. Clark, "Task dependent deep lda pruning of neural networks," *Computer Vision and Image Understanding*, vol. 203, p. 103154, 2021.

[13] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *5th International Conference on Learning Representations*, 2017.

[14] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[15] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[17] S. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Robust physical adversarial attack on faster R-CNN object detector," *CoRR*, vol. abs/1804.05810, 2018.

[18] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[19] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

[20] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[21] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," in *ASIA CCS 2017*, pp. 506–519, 2016.

[22] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *the 36th International Conference on Machine Learning*, vol. 97, pp. 2484–2493, 2019.

[23] Y. Li, D. Tian, M.-C. Chang, X. Bian, and S. Lyu, "Robust adversarial perturbation on deep proposal-based models," *arXiv preprint arXiv:1809.05962*, 2018.

[24] H. Zhang and J. Wang, "Towards adversarially robust object detection," in *the IEEE/CVF International Conference on Computer Vision*, pp. 421–430, 2019.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations*, 2018.

[26] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning*, pp. 274–283, 2018.

[27] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *6th International Conference on Learning Representations*, 2018.

[28] X. Chen, C. Xie, M. Tan, L. Zhang, C.-J. Hsieh, and B. Gong, "Robust and accurate object detection via adversarial learning," in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16622–16631, 2021.

[29] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," *the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12993–13000, 2020.

[30] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *International Conference on Learning Representations*, 2020.

[31] B. Wu, A. Wan, F. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 446–454, 2017.