Parallel Rectangle Flip Attack: A Query-based Black-box Attack against Object Detection

Siyuan Liang[1,2], Baoyuan Wu[3,4,†], Yanbo Fan[5], Xingxing Wei[6], Xiaochun Cao[1,2,†]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[4]Secure Computing Lab of Big Data, Shenzhen Research Institute of Big Data, Shenzhen, China
[5]Tencent, Shenzhen, China
[6]Institute of Artificial Intelligence, Hangzhou Innovation Institute, Beihang University, Beijing, China
{liangsiyuan,caoxiaochun}@iie.ac.cn;wubaoyuan@cuhk.edu.cn;fanyanbo0124@gmail.com;xxwei@buaa.edu.cn

Abstract

Object detection has been widely used in many safety-critical tasks, such as autonomous driving. However, its vulnerability to adversarial examples has not been sufficiently studied, especially under the practical scenario of black-box attacks, where the attacker can only access the query feedback of predicted bounding-boxes and top-1 scores returned by the attacked model. Compared with black-box attack to image classification, there are two main challenges in black-box attack to detection. Firstly, even if one bounding-box is successfully attacked, another sub-optimal bounding-box may be detected near the attacked bounding-box. Secondly, there are multiple bounding-boxes, leading to very high attack cost. To address these challenges, we propose a Parallel Rectangle Flip Attack (PRFA) via random search. We explain the difference between our method with other attacks in Fig. 1. Specifically, we generate perturbations in each rectangle patch to avoid sub-optimal detection near the attacked region. Besides, utilizing the observation that adversarial perturbations mainly locate around objects' contours and critical points under white-box attacks, the search space of attacked rectangles is reduced to improve the attack efficiency. Moreover, we develop a parallel mechanism of attacking multiple rectangles simultaneously to further accelerate the attack process. Extensive experiments demonstrate that our method can effectively and efficiently attack various popular object detectors, including anchor-based and anchor-free, and generate transferable adversarial examples.

† indicates corresponding authors. Corresponds to wubaoyuan@cuhk.edu.cn and caoxiaochun@iie.ac.cn

1. Introduction

Deep neural networks [42] has significantly boosted the developments of many important tasks, such as image classification [19, 23], object detection [41, 39, 30, 31, 35], medical image analysis [11], *etc*. For example, object detection has been successfully applied in many safety-critical scenarios, such as autonomous driving [25] and pedestrian detection [43], *etc*. However, many studies [7, 3, 52, 8, 22, 4, 20, 47, 28, 27, 15, 49] have shown that the DNNs are vulnerable to adversarial attacks and may produce false predictions. If pedestrians or traffic signs are incorrectly detected in autonomous driving, it will cause substantial security risks in the real world.

Compared with the massive works on attacking image classification, adversarial attacks against DNN-based object detection have not been thoroughly studied, especially in the black-box scenario, where only the predicted bounding-boxes and confidences of queries are accessible to the attacker. There are two main challenges in attacking the black-box object detection. Firstly, due to the widely used module called non-maximum suppression (NMS) in mainstream detectors, only the proposal with the highest confidence score is predicted, while other proposals with similar confidence in near locations are suppressed. Consequently, even if one predicted bounding box is successfully attacked (*i.e*., not detected), another sub-optimal bounding box may be detected in similar locations(as shown in Section D of the **Supplementary Material**). Secondly, the number of optimized targets (*i.e*., proposals) in object detection is much larger than that in classification [48]. Take a d-dimensional image as an example, the computational complexity of the candidate proposals is $O(d^2)$, while the complexity of classification is $O(d)$. It will cause very high cost to attack

object detector.

To address above two challenges, we propose an effective and efficient query based black-box attack method against object detection, called **Parallel Rectangle Flip Attack (PRFA)**. Specially, we first search a rectangle patch randomly, and generate adversarial perturbations with the sign flipping along the vertical or horizontal direction, to present any detection in this attacked patch, including any sub-optimal proposals. Besides, we observe that adversarial perturbations generated by white-box attacks against detection with large magnitudes mainly locate at objects' contours and some critical points [47]. Inspired by this observation, the search space of attacked rectangles can be significantly reduced to improve the attack performance. Moreover, we design a parallel mechanism that multiple rectangles can be attacked simultaneously, which can further improve the attack efficiency. The proposed PRFA method achieves successful attack on many popular object detectors, including anchor-based (*e.g.*, two-stage FR [41] and one-stage YOLO [16]), anchor-free model (*e.g.*, FCOS [44]), and the ATSS model [51].

The main contributions of this work are threefold. **1)** To the best of our knowledge, this is the first work about query-based black-box attack against object detection. **2)** We propose an effective and efficient black-box attack method specially designed for attacking object detection, such that the main challenges including the sub-optical detection and high attack cost can be well addressed. **3)** Extensive experiments demonstrate the superior attack performance of our method on attacking many mainstream object detectors, including both anchor-based and anchor-free detectors.

2. Related Work

2.1. Object Detection and White-box Attack

Mainstream object detectors are mostly based on deep neural networks and can be roughly divided into two categories: anchor-based and anchor-free. The anchor-based detector divides the predefined sliding windows or proposals into positive or negative samples, then refines and classifies the prediction boxes. Due to the difference in the regression forms, it can be subdivided into the one-stage detector, such as SSD [35], YOLOv2 [40] and two-stage detector, Faster-RCNN [41], Mask-RCNN [18]. The most representative anchor-free detector may be YOLOv1 [39]. YOLOv1 abandons the anchor and directly predicates the bounding box at the object's center. Since anchor-free detectors do not require extra parameter adjustment, these types of detectors have gained widespread popularity. Representative methods include CenterNet [14], ExtremeNet [54], CornerNet [24] and FCOS [44]. Some methods focus on the gap between anchor-based and anchor-free detectors, such as ATSS [51], which improves the detection result by changing the sampling and IoU threshold calculation.

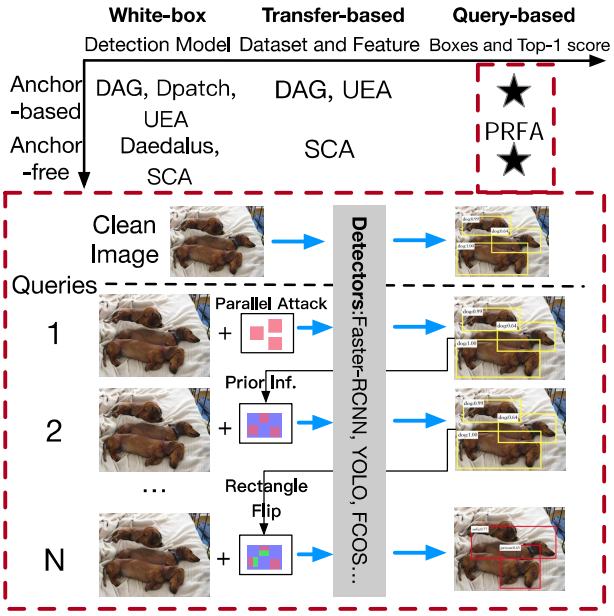

Figure 1. We use the coordinate axis to show the taxonomy of the adversarial attack in object detection. Different from white box attacks and migration attacks, our method PRFA only relies on the prediction box and top-1 score output after NMS to attack through queries without gradients. PRFA can also attack anchor-free and anchor-based models at the same time.

The existing adversarial attacks for object detection tasks are mainly white-box. As the first white-box attack method, DAG [48] successfully fools Faster-RCNN by attacking the RPN network's proposals. [5] proposes a classification loss to train a GAN and generates adversarial perturbations for a face detector based on Faster-RCNN. [26] designs a loss of predicting boxes and classification for the white-box attack. [36] successfully attacks the YOLO and Faster-RCNN models by generating an adversarial patch. Daedalus [46] analyzes the vulnerability of NMS in the existing detection systems and attacks multiple detectors by generating many false-positive samples. [29] proposes an attack method for the same category to attack the anchor-free detector. Since most detectors use the same feature extractor, UEA [47] generates transferable adversarial examples by destroying image features to attack Faster-RCNN and SSD models.

However, above adversarial attacks for object detection utilize the network's gradient more or less. In the real world, we cannot obtain the detector's gradient, which makes the attack very difficult. Therefore, we propose a query-based adversarial attack in a black-box scenario.

2.2. Black-box Attack against Image Classification

The black-box attack includes the transfer-based attack and the query-based attack. The attacker obtains adversarial examples by accessing the model's outputs and modifying

clean images. Using a transferable attack strategy, the adversary can train a substitute model [53] replacing the target model to get the adversarial gradient. DI-FGSM [48] and TI-FGSM [12] use 'diverse inputs' or 'translation-invariant' to generate more transferable adversarial examples against the defense models. Dispersion Reduction [38] proposes an attack to minimize the 'dispersion' of the feature map to enhance the transferability across different computer tasks. Other works focus on query feedback mechanisms against the logistics scores of all categories. ZOO [9] proposed a zero-order optimization method to estimate the gradient of the target model. Some studies are based on gradient's sign, such as [34], ZO-SignSGD extends SignSGD to the zero-order case and achieves black-box attacks. Sign-Hunter [1] accelerates the convergence by combining the previous query results and converting the gradient estimation from continuous to binary. The Boundary [6] and Evolutionary [13] attack methods utilize the evolution strategy to gradually search the adversarial example that is close to the benign example in the scenario of decision-based black-box attack. The Sign Flip attack method [10] proposes to search the perturbation by gradually shrinking the ℓ_∞-ball around the benign example and randomly flipping the signs of a few dimensions of the current perturbation. SquareAttack [2] based on a random search generates square-shaped perturbations at random positions. CG-ATTACK [17] proposes to guide the search procedure by the conditional adversarial distribution, which is partially transferred from the distribution modeled by the c-Glow network [37] and trained on surrogate models.

Different from the classification, the optimization problem on object detection is complex. Suppose that d denotes the number of pixels in one image, each proposal is determined by two pixels/coordinates, the complexity of object detection is $O(d^2)$ and of classification is $O(d)$. The detector's outputs are prediction boxes after NMS and the top-1 score (probability of the top-1 label), making the black-box attack on object detection more like an intermediate setting between score-based and decision-based settings. How to achieve effective attacks with limited information and queries on detectors is our research focus.

3. Parallel Rectangle Flip Attack

This section will model a query-based black-box attack and introduce our method, which searches rectangle perturbations at random positions parallelly with fliping/reversing the sign of perturbations.

3.1. Problem Formulation

Suppose that a clean image $\boldsymbol{x}$ has M recognition objects $\mathcal{O} = \{o_1, o_2, ..., o_M\}$. For each object $o_m, m = 1, 2, ..., M$, is marked with ground-truth bounding box g_m and a class label $y_m \in \{1, 2, ..., Y\}$, where Y is the number of classes. Object detection is an important computer vision task that predicts the position and a certain class(such as humans, transportation, or animals) of instances in digital images. An object detector $f(\boldsymbol{x}) \in \mathbb{R}^{N\times(4+1)}$ predicts the prediction boxes $b_n, n = 1, 2, ..., N$, and top-1 label $c_n, n = 1, 2, ..., N$, with score f_C (the probability after softmax normalization for predicted labels C) for N objects.

To generate an adversarial examples $\hat{\boldsymbol{x}} \in [0, 1]^d$ which is regarded as adversarial examples with an l_p-norm of ϵ for the clean image $\boldsymbol{x}$, *i.e.*, $||\hat{\boldsymbol{x}} - \boldsymbol{x}||_p \leq \epsilon$, and the goal is to make the IoU of all prediction boxes and ground-truth is less than a certain threshold or the labels of prediction boxes are classified incorrectly, that is, $\forall n \in N, \forall m \in M, (\mathrm{IoU}(b_n, g_m) < \mathrm{threshold}) \vee (c_n \neq y_m)$. Here, IoU score is a standard performance measure for object detection, *i.e.*, $\mathrm{IoU}(a, b) = (a \cap b)/(a \cup b)$, and the threshold is set to 0.5 for detection tasks. The task of find $\hat{\boldsymbol{x}}$ can be rephrased as solving the following optimization function:

$$\begin{aligned}\underset{\hat{\boldsymbol{x}}\in[0,1]^d}{\arg\min}\ & H(f(\hat{\boldsymbol{x}}), B, Y) = \sum_{n=1}^{N}\sum_{m=1}^{M}[\mathrm{IoU}(b_n, g_m)\cdot \mathbb{1}_{f_{c_n}\geq\zeta} \\ & + \lambda\cdot(f_{c_n} - \max_{C\neq y_m} f_C)\cdot\mathbb{1}_{f_{c_n}<\zeta}], \\ & s.t.\ ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_p \leq \epsilon, f_{c_n} = f_C(\hat{\boldsymbol{x}}, b_n), \end{aligned} \tag{1}$$

where the $\mathbb{1}$ represents indicator function. The indicator function $\mathbb{1}_a = 1$ if a is true, otherwise 0. Since the attack satisfies one condition in Eq. (1), we can use the top-1 score f_{c_n} as one of the judgment optimization formula one. Specifically, when the top-1 score is greater than the threshold ζ, we consider reducing the IoU of the corresponding prediction box and ground-truth, and when it is less than the threshold ζ, we optimize the top-1 score. λ is a is a hyperparameter that adjusts balance.

However, the optimization in Eq. (1) needs to match all the prediction boxes with the ground-truth one by one, which makes the computation complexity reach $O(M * N)$ and costs a lot of time for one query. Therefore, we propose a category-based optimization function, which optimizes the prediction box and ground-truth under the same category, and sets the computational complexity to $O(M * N/|Y|)$. Because the detection dataset has too many categories, this will greatly improve one query speed. The $(N|y)$ means the index set of objects with the label y, *i.e.*, $(N|y) = \{i|c_i = y, i = 1, 2, ..., N\}$. The new optimization function H is as follow:

$$\begin{aligned}\underset{\hat{\boldsymbol{x}}\in[0,1]^d}{\arg\min}\ & \sum_{y=1}^{Y}\sum_{n=1}^{(N|y)}\sum_{m=1}^{(M|y)}\big[\mathrm{IoU}(b_n, g_m)\cdot\mathbb{1}_{f_{c_n}\geq\zeta} + . \\ & \lambda\cdot(f_{c_n=y}(\hat{\boldsymbol{x}}, b_n) - \max_{C\neq y} f_C)\cdot\mathbb{1}_{f_{c_n}<\zeta}\big], \\ & s.t.\ ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_p \leq \epsilon, f_{c_n} = f_C(\hat{\boldsymbol{x}}, b_n). \end{aligned} \tag{2}$$

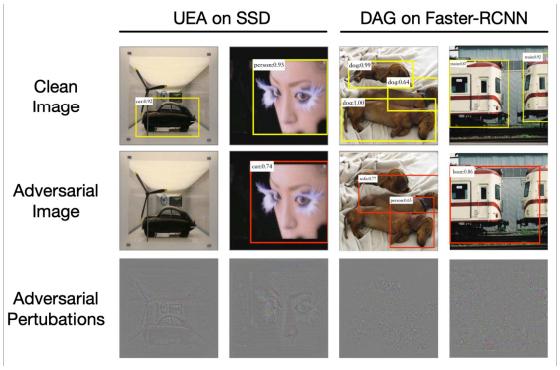

Figure 2. The adversarial perturbations generated by the white-box attack methods UEA and DAG respectively on the SSD and Faster-RCNN. These perturbations are basically distributed at the contours and critical points of the object.

3.2. Prior-guided Dimensionality Reduction

Dimensionality reduction has been shown to be effective to improve the efficiency of black-box attacking against image classification [45]. Inspired by this, we use the detector's prior information and the prior observation under the white-box attack to reduce the random search space. Although the anchor-based and anchor-free models are substantially different in anchor, they consider the objectness into prediction. Objectness is essentially a measure of the probability that an object exists in the region of interest. If the objectness is high, it means that the image window likely contains an object. We attack areas with high objectness instead of the entire image. Besides, we use DAG and UEA methods to observe the distribution of adversarial perturbations on different models. Although the attack methods and target models are different, the distributions of perturbations are concentrated on objects' critical areas or contours. As shown in Fig. 1, we show the perturbations generated by DAG and UEA.

We use the prediction box or prior information to calculate an area with high objectness and perform a random search in this area. We optimize Eq. (2) by generating rectangular perturbations through random sampling. The perturbations in this way are relatively close with white-box in position distribution. We considered three methods for calculating objectness, anchor-based priors(segmentation results from Mask-RCNN [18]), anchor-free priors(key points representation from RepPoints [50]), and prediction boxes(outputs from detectors). Unlike the latter, the first two priors use the other detector's transferability to obtain critical areas.

3.3. Parallel Attack Accelerating Breadth Search

We successfully modify the black-box attack methods such as SignHunter, SquareAttack, NES, and ZO-SignSGD

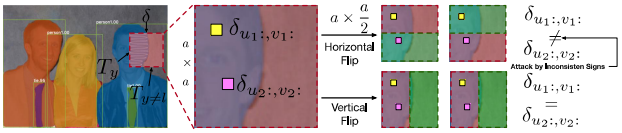

Figure 3. The process of flipping perturbation's sign. We force the points with the same feature to be different by flipping sign, which will cause the detector to separate them to achieve an effective attack.

to the object detection task by optimizing Eq. (2). Among them, the attack performance of SquareAttack is the most ideal. By observing the attack process of SquareAttack, we find that as the number of queries increases, the generated adversarial perturbations gradually gather around the object. This phenomenon shows that, under the constraints of Eq. (2), the black-box attack method can find the vulnerable pixels with a large number of queries. we can attack multiple positions parallelly in one query, thereby indirectly increasing the number of pixel searches, which can be a way to accelerate breadth search.

Next, we will theoretically analyze that it is not the best choice to generate adversarial perturbations δ at only one random position at each query q. Suppose the optimized detector f is the smoothness and has a Lipschitz gradient. There exists a constant L satisfying:

$$f(\boldsymbol{x}_{q+1}) - f(\boldsymbol{x}_q) \leq \langle f^{'}(\boldsymbol{x}_q), \delta \rangle + \frac{L}{2}||\delta||^2. \tag{3}$$

According to the assumption 2 in [33], the stochastic gradient is unbiased and with bounded variance, and its upper bound is a constant σ^2. The SquareAttack method satisfies the following:

$$\frac{1}{Q}\sum_{q=0}^{Q}\mathbb{E}[||f^{'}(\boldsymbol{x}_q)||^2] \lesssim \frac{f(\boldsymbol{x}_0) - \mathbb{E}[f(\boldsymbol{x}_{Q+1})]}{Q\gamma} + \gamma L\sigma^2, \tag{4}$$

where γ is step size and when $\gamma = \frac{1}{L+\sigma\sqrt{QL}}$, the convergence rate is $O(1/\sqrt{Q})$. The $\lesssim$ means small and equal to up to a constant factor. The Eq. (4) means that the number of iterations Q is large enough, and the random search algorithm will converge.

We propose a parallel random search on the image. Each iteration q can randomly sample P positions in the search space D. In this way, we reduce the variance of the gradient estimation $\mathbb{E}[||g^P(\boldsymbol{x}) - f^{'}(\boldsymbol{x})||^2] \leq (\frac{D-P}{D-1})\frac{\sigma^2}{P}$ and accelerate the convergence of the algorithm. At this time, the convergence can be expressed as:

$$\frac{1}{Q}\sum_{q=0}^{Q}\mathbb{E}[||f^{'}(\boldsymbol{x}_q)||^2] \lesssim \frac{L}{Q} + \frac{\sqrt{L}\sigma}{\sqrt{QP}}. \tag{5}$$

In Eq. (5), the influence of queries Q on the algorithm convergence is still more significant than the number P of par-

allel attacks and the convergence rate is $O(1/\sqrt{QP})$ faster than Eq. (4). We adopted a parallel number scheduling strategy. Due to the small number of iterations in the early stage of the query, our sample number P is the largest. As the number of iterations increases, we gradually decrease the value of P until it is 1. See Section B in Supplementary Material for detailed proof.

3.4. Rectangle Flip Attack for Depth Search

Generally speaking, objects marked with a rectangular box in detection datasets have fixed sizes and scales. It is a typical prior for the detector to use predefined anchors with a fixed ratio. For example, Faster-RCNN uses 1:1, 1:2, and 2:1 anchor settings, and YOLO through k-means clustering learns different anchors from the training set. It is initially inspired by our experimental observation that given one region within one bounding box, perturbing different sub-regions with different noises is more likely to cause that this region is separately detected to different bounding boxes, compared to perturbing with similar noises. Furthermore, as shown in Fig. 3, the square patch often covers one local region of one object, flipping sign horizontally or vertically may improve the perturbation diversity. Hence, this region is more likely to be detected falsely as different bounding boxes, causing the change of the original bounding box on this object.

Given an initial square-shaped $\delta \in \mathbb{R}^{a\times a}$ with bound ϵ and a convolutional filter $w \in \mathbb{R}^{k\times k}$. Let $z = F((\boldsymbol{x}_p + \delta) * w)$ denotes outputs of the CNN for updating δ, where F denotes activation function and $\boldsymbol{x}_p$ is the clean patch. (m, n) represents coordinate in adversarial patch $\hat{\boldsymbol{x}}_p$. We can divided (m, n) into $|Y|+1$ cliques $\{T_i\}_{i=0}^{|Y|+1}$, and $(u, v) \in T_i$ denotes one point's coordinates in the i-th clique. The α is a constant greater than 0. The maximal change l_∞-norm of z represents:

$$\begin{aligned}||z||_\infty &= \max_{m,n} |z_{m,n}| \\ &= \max_{m,n} |F(\sum_{i,j=1}^{k} (\boldsymbol{x}_p + \delta)_{m-\lfloor \frac{s}{2} \rfloor + i, n-\lfloor \frac{s}{2} \rfloor + j} \cdot w_{i,j})| \\ &\leq \max_{m,n} |F(\sum_{i,j=1}^{k} \delta_{m-\lfloor \frac{s}{2} \rfloor + i, n-\lfloor \frac{s}{2} \rfloor + j} \cdot w_{i,j}) + \alpha|.\end{aligned} \tag{6}$$

For adversarial attack in object detection, the maximum change component of the correct label y of an object contained in a patch should be smaller than the maximum change component of other classes. The optimization function H in patch $\hat{\boldsymbol{x}}_{\boldsymbol{p}}$ can be expressed as follows, the $\delta_{u:,v:}$ is a shorthand in Eq. (6):

$$\begin{aligned}\min_\delta H = \min_\delta [&\sum_{(u,v)\in T_y}^{T_y} \max |F(\sum_{i,j=1}^{k} \delta_{u:,v:} \cdot w_{i,j}) + \alpha_y| \\ &- \max \sum_{l\neq y}^{|Y|} [\sum_{(u,v)\in T_l}^{T_l} \max |F(\sum_{i,j=1}^{k} \delta_{u:,v:} \cdot w_{i,j}) + \alpha_l|]].\end{aligned} \tag{7}$$

Since Eq. (7) is difficult to optimize, we combine adversarial perturbation and image semantic information to simplify Eq. (7). If two points locate in the same patch and belong to the same clique, then their perturbation should be the same, that is $(u_1, v_1) \in T_l$ and $(u_2, v_2) \in T_l$, then $\delta_{u_1:,v_1:} = \delta_{u_2:,v_2:}$. We propose an approximation to Eq. (7), as follows:

$$\begin{aligned}\min_\delta H \approx \min_\delta [&\sum_{(u,v)\in T_y}^{T_y} \max |F(\sum_{i,j=1}^{k} \delta_y \cdot w_{i,j}) + \alpha_y| \\ &- \max \sum_{l\neq y}^{|Y|} [\sum_{(u,v)\in T_l}^{T_l} \max |F(\sum_{i,j=1}^{k} \delta_l \cdot w_{i,j}) + \alpha_l|]].\end{aligned} \tag{8}$$

The change value of the clique T_y is maximum when the perturbations sign in every point is correct and consistent. Since the critical areas we attacked contain the object y, the clique labeled T_y dominates. To minimize the Eq. (8), we can minimize the upper bound of the previous term. Points that belong to the same clique in a patch are highly similar in spatial location (close to value) and semantic feature (close in the property). Hence, if the sign of perturbations can change the semantic feature of one point, it will be effective for other points with high probability. Consequently, we can generate a rectangular perturbation by flipping the sign, which makes points in the same clique are inconsistent and pushes them into different classified classification boundaries. By this, we can minimize the Eq. (8) for effective black-box.

3.5. Generating Adversarial Examples

Firstly, we use the detector's prior information in Sect. 3.2 to calculate the high objectness area and determine the critical points to perform a random search. Then, we set the side length a of the square perturbations, and the number of parallel points P according to the dynamic scheduling algorithm. We first generate square-shaped perturbations with side length a for each iteration. Next, we flip half of the square perturbations' sign (a rectangle with $a * a/2$) vertically or horizontally. We calculate scores by Eq. (2) and update adversarial perturbations if the current score is greater than the optimal score. See Section A in Supplementary Material for the specific algorithm.

Table 1. Ablation study on the Faster-RCNN model

Method	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	AQ
Clean	0.52	0.72	0.57	0.28	0.55	0.74	N/A
Square Shaped	0.28	0.50	0.27	0.21	0.38	0.32	3787
SS w. Prior	0.26	0.48	0.25	0.20	0.30	0.33	3667
SS w. Prior & Flip	0.24	0.42	0.23	0.18	0.27	0.28	3342
SS w. Prior & Parallel	0.22	**0.40**	0.21	0.16	0.26	0.28	3513
Parallel **R**ectangle **F**lip **A**ttack	**0.21**	0.41	**0.19**	**0.16**	**0.26**	**0.27**	**3331**

Table 2. Untargeted attack against object detectors.

Method	Faster-RCNN [41]							ATSS [51]						
	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	AQ	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	AQ
Clean	0.51	0.72	0.57	0.28	0.55	0.74	N/A	0.54	0.73	0.60	0.32	0.58	0.74	N/A
NES [21]	0.49	0.69	0.54	0.26	0.52	0.69	4000	0.52	0.70	0.57	0.29	0.56	0.73	4000
ZSS [34]	0.49	0.71	0.54	0.20	0.52	0.71	4040	0.52	0.70	0.58	0.21	0.56	0.73	4040
SH [1]	0.39	0.63	0.38	0.24	0.43	0.57	3987	0.40	0.55	0.44	0.20	0.40	0.59	3852
SA [2]	0.28	0.50	0.26	0.21	0.38	0.32	3786	0.23	0.34	0.24	0.13	0.28	0.31	3505
PRFA	**0.21**	**0.42**	**0.19**	**0.16**	**0.26**	**0.27**	**3331**	**0.20**	**0.30**	**0.23**	**0.12**	**0.25**	**0.30**	**3500**

Method	YOLOv3 [16]							FCOS [43]						
	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	AQ	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	AQ
Clean	0.45	0.70	0.47	0.16	0.47	0.65	N/A	0.54	0.75	0.58	0.33	0.56	0.74	N/A
NES [21]	0.41	0.68	0.43	0.16	0.44	0.59	3958	0.53	0.73	0.57	0.23	0.56	0.77	4000
ZSS [34]	0.39	0.64	0.40	0.19	0.43	0.59	3958	0.52	0.71	0.56	0.27	0.57	0.74	4040
SH [1]	0.39	0.66	0.40	0.19	0.40	0.58	3911	0.27	0.40	0.31	0.09	0.37	0.64	3633
SA [2]	0.25	0.49	**0.22**	0.15	0.31	**0.34**	3192	**0.21**	0.35	**0.22**	**0.14**	**0.20**	**0.37**	3578
PRFA	**0.24**	**0.46**	**0.22**	**0.13**	**0.29**	0.36	**2949**	0.23	**0.34**	0.23	0.15	0.29	0.41	**3395**

4. Experiments

4.1. Experiments Settings

We will introduce the experiment settings from four aspects: detection datasets, evaluation criteria, targeted models and parameters setting.

Detection Datasets The MS-COCO [32] is the most challenging object detection datasets today. A large number of object detectors use MS-COCO as a benchmark to evaluate the model's performance. MS-COCO includes 118k images for the training set and 5k images for the validation set. The objects in these pictures are divided into 80 categories. In order to ensure fairness, we attack the validation set. Specifically, we selected first 100 images as clean images for the black-box attack according to the MS-COCO API's loading order. These samples include a wealth of object instances, such as small objects or dense objects.

Evaluation Criteria AP is defined as the average detection accuracy under different recall rates, and we usually evaluate it in a category-specific way (the mean AP, **mAP**). Refer to the detector, we use the evaluation criteria provided by MS-COCO, for example, $mAP($ averaged over multiple IoU threshold between 0.5 and 0.95), mAP_{50}(mean AP at IoU=0.5), mAP_{75}(mean AP at IoU=0.75), $mAP_S(area < 32^2)$, $mAP_M(32^2 < area < 96^2)$, $mAP_L(area > 96^2)$. An effective black-box attack means a small mAP. In terms of algorithm efficiency, we use **AQ(average queries)** to evaluate the convergence of the algorithm. We hope to minimize the number of queries.

Targeted Models We selected four representative detectors as targeted models. The first type is anchor-based. We chose the two-stage detector Faster-RCNN with ResNet50 and the YOLOv3 model with DarkNet53 as backbones. The second type belongs to anchor-free. We chose the FCOS model with ResNet50 as the backbone. We also selected ATSS, a detector that can adaptively select positive and negative samples. This model can eliminate the performance difference between anchor-based and anchor-free algorithms. The above models and codes are based on the open-source mmdetection library.

Parameters Setting Given images of size $w * h$, the length a of square is $\sqrt{e * w * h}$, $e \in [0, 1]$. The e is set to 0.05, and we halve it at query $q \in \{20, 100, 400, 1000, 2000, 4000, 8000\}$. The parallel P is 4 in the initial stage, and we halve them at $q \in \{20, 100, 1000, 2000\}$. In Eq (2), the ζ is 0.90 and the ϵ is 0.05. The threshold for IoU is 0.50. The prior information from the same type of detector is more beneficial to attack similar detectors, our PRFA in all reported experiments only utilized the attacked model's outputs.

Table 3. Black-box transferability across different Detectors. The result with $^*num^*$ represents the adversarial examples from one black-box detector to attack itself, and the ordinary result represents transferability.

Adv. Dataset	Faster-RCNN						ATSS					
	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
Clean	0.51	0.72	0.57	0.28	0.55	0.74	0.54	0.73	0.60	0.32	0.58	0.74
Faster-RCNN	*0.21*	*0.42*	*0.19*	*0.16*	*0.26*	*0.27*	0.30	0.47	0.29	**0.11**	0.35	0.47
ATSS	0.26	**0.42**	0.25	**0.11**	0.28	0.42	*0.20*	*0.30*	*0.23*	*0.12*	*0.25*	*0.30*
YOLO	0.32	0.51	0.32	0.13	0.34	0.50	0.36	0.55	0.37	0.18	0.39	0.55
FCOS	0.28	**0.42**	0.31	0.06	0.31	0.45	0.35	0.37	0.34	0.36	0.45	0.52
Adv. Dataset	YOLO						FCOS					
	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
Clean	0.45	0.70	0.47	0.16	0.47	0.65	0.54	0.75	0.58	0.33	0.56	0.74
Faster-RCNN	0.22	0.39	0.22	0.08	**0.23**	0.41	0.27	0.44	0.26	**0.09**	0.31	0.46
ATSS	0.24	0.43	0.24	0.10	0.26	0.41	0.28	0.45	0.28	0.11	0.33	0.47
YOLO	*0.24*	*0.46*	*0.22*	*0.13*	*0.29*	*0.36*	0.35	0.54	0.36	0.11	0.38	0.54
FCOS	**0.16**	**0.35**	**0.08**	**0.07**	0.28	**0.32**	*0.23*	*0.35*	*0.24*	*0.14*	*0.22*	*0.36*

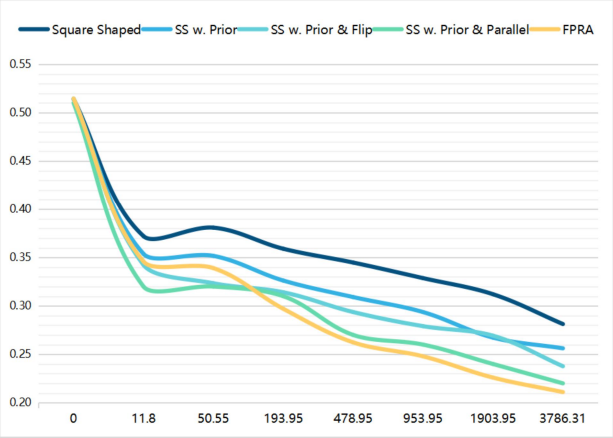

Figure 4. mAP changes $w.r.t.$ the number of queries for different attacks. FPRA achieves the fastest convergence and the most effective attack.

4.2. Abaltion Study

We discuss the effectiveness of each component of the PRFA method through ablation experiments. In Tab. 1, we show the ablation experiment on the Faster-RCNN model. 'Clean' represents the detection result of Faster-RCNN on clean images. 'Square Shaped' means that we only use square perturbations to perform random searches across the entire picture. By adding prior information, we can limit the space of random search. We add the prior restriction 'SS w. Prior' to SS (short for square-shaped) to reduce the search space. Next, we generate rectangular perturbations by flipping the perturbations' sign. The method 'SS w. Prior & Flip' can be regarded as a combination of depth search and breadth search. Finally, we combine a parallel search attack to accelerate the algorithm's convergence and provide a parallel rectangle flip attack algorithm, namely PRFA.

The contribution of 'Flip' is evaluated by attack performance (*e.g.*, mAP reduction) and queries. As shown in Fig. 4, in terms of mAP, the values of 'SS w. Prior', 'SS w. Prior & Flip' and PRFA are 0.26, 0.24 and 0.21, respectively; in terms of queries, the values of these three methods are 3666, 3342 and 3331, respectively. 'Flip' contributes 29% to the mAP reduction and 97% to the query reduction. 'Flip' can also accelerate the convergence compared to 'SS w. Prior' Therefore, the attack problem can be regarded as a trade-off between breadth search and depth search. The Fig. 4 shows the change in mAP of each method over the number of iterations. By introducing parallel attacks, we effectively accelerated the model's convergence (yellow line) and got the best attack results.

4.3. Untargeted Attacks against Detectors

We evaluated four state-of-the-art black-box attack algorithms and our method PRFA on the four detectors according to the settings in Sect. 4.1. The four methods are the NES black box attack combining PGD and NES strategies, the ZO-SignSGD(ZSS) algorithm combining zero-order optimization and sign stochastic gradient descent, the SignHunter(SH) method that converts the gradient estimation from a continuous problem to a binary problem, and the SquareAttack(SA) that generates square perturbations at a random location. We limited the number of queries for these methods, where NES is 4000, ZO-SignSGD is 4040, and SignHunter, SquareAttack, and PRFA are all 4000.

The performance of NES is the worst, and it can hardly affect the existing object detector. The performance of ZO-SignSGD on the models Faster-RCNN, ATSS, and FCOS is also inferior. He can only attack the YOLOv3 model. Al-

Figure 5. The detection results of the four detectors on clean images (yellow lines) and adversarial samples (red lines).

though SignHunter can attack the existing object detector to a certain extent, we found that it uses nearly 4000 queries, which could not effectively converge with our query limits. SquareAttack and our method showed the effectiveness of the attack on all four models. Nevertheless, our method has a better attack effect on Faster-RCNN and ATSS, especially on Faster-RCNN, which drops 7 points and fewer queries (about 440). In the same attack effect (YOLOv3 and FCOS), the number of queries of our method is also better than the SquareAttack method.

In summary, our method can achieve fewer queries or more effective attack, which means that our method will have faster convergence and search for more effective perturbations.

4.4. Black-box Transferability across Models

Finally, we investigate black-box transfer, *i.e.*, using the perturbations generated by a black-box detector to attack other detectors. In Tab. 3, the result with $*num*$ represents the adversarial examples from one black-box detector to attack itself, and the ordinary result represents transferability.

In Tab. 3, The attack effects of the three models Faster-RCNN, ATSS, and FCOS are robust. The reason is that they can resist attacks from other datasets. Among them, the most robust method is ATSS, which achieves an effect of at least 0.30 mAP on other datasets. The indicators of Faster-RCNN on other datasets are slightly lower than FCOS by 1-2 percentage points. Among all the models, YOLO is the most accessible model to attack. The adversarial examples generated by other models can attack him effectively, even surpassing the adversarial examples generated by itself.

Here we compare with three mentioned transfer methods, including 'Dispersion', 'DIFGSM' and 'TIFGSM'. They firstly conduct white-box attacks on Faster-RCNN. Specifically, 'Dispersion' attacks intermediate features of the ResNet50 backbone, while 'DIFGSM' and 'TIFGSM' attack the bounding box loss. Then, the obtained perturbations are transferred to attack the black-box target models ATSS, YOLOv3 ,and FCOS, respectively. The mAP values of these three methods and our PRFA method are $(0.05, 0.01, 0.02, 0.21)$ (PRFA conducts black-box attack) for Faster-RCNN, $(0.51, 0.48, 0.39, 0.3)$ for ATSS, $(0.43, 0.43, 0.4, 0.22)$ for YOLOv3, $(0.51, 0.48, 0.42, 0.27)$ for FCOS, respectively. PRFA significantly surpasses these transfer-based black-box methods.

5. Conclusion

In this paper, we propose a query-based black-box attack with the prediction boxes and top-1 scores. We adapt the existing black-box attack method to detection as the baseline. We propose a parallel rectangle flip attack via random search. We use prior information from detectors to reduce the search space by observing the white-box perturbations' distribution. We regard the optimization problem of searching perturbations for the detector as a trade-off between breadth search and depth search. We accelerate the model's convergence by parallel random walks in the search space in terms of breadth. We obtain a better attack effect by flipping the perturbations' sign locally to generate rectangular perturbations in terms of depth. Experiments show that our proposed method PRFA can attack mainstream object detectors and generate transferable adversarial examples.

Acknowledgement

Supported by the National Key R&D Program of China under Grant 2019QY(Y)0207, National Natural Science Foundation of China (No.62025604, 61861166002, U1936208), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2021C06). Baoyuan Wu is supported by the Natural Science Foundation of China under grant No.62076213, the university development fund of the Chinese University of Hong Kong, Shenzhen under grant No.01001810, the special project fund of Shenzhen Research Institute of Big Data under grant No.T00120210003, and 2021 Tencent AI Lab Rhino-Bird Focused Research Program. Xingxing Wei is supported by National Natural Science Foundation of China (No.62076018, 61806109) and CCF-Tencent Open Research Fund.

References

[1] Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2019. 3, 6

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020. 3, 6

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

[4] Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, and Shu-Tao Xia. Targeted attack against deep neural networks via flipping limited weight bits. *arXiv preprint arXiv:2102.10496*, 2021. 1

[5] Avishek Joey Bose and Parham Aarabi. Adversarial attacks on face detectors using neural net based constrained optimization. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2018. 2

[6] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representation (ICLR)*, 2018. 3

[7] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 1

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[9] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. 3

[10] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In *European Conference on Computer Vision*, pages 276–293. Springer, 2020. 3

[11] Yixiong Chen, Chunhui Zhang, Li Liu, Cheng Feng, Changfeng Dong, Yongfang Luo, and Xiang Wan. Effective sample pair generation for ultrasound video contrastive representation learning. *arXiv preprint arXiv:2011.13066*, 2020. 1

[12] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 3

[13] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 3

[14] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[15] Yanbo Fan, Baoyuan Wu, Tuanhui Li, Yong Zhang, Mingyang Li, Zhifeng Li, and Yujiu Yang. Sparse adversarial attack via perturbation factorization. In *European Conference on Computer Vision*, 2020. 1

[16] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. *Computer Vision and Pattern Recognition, cite as*, 2018. 2, 6

[17] Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shutao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. *arXiv preprint arXiv:2006.08538*, 2020. 3

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 4

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[20] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards Security Threats of Deep Learning Systems: A Survey. *IEEE Transactions on Software Engineering (TSE)*, pages 1–28, 2020. 1

[21] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018. 6

[22] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[23] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han. Adv-watermark: A novel watermark perturbation for adversarial examples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1579–1587, 2020. 1

[24] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[25] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168. IEEE, 2011. 1

[26] Yuezun Li, Daniel Tian, Ming-Ching Chang, Xiao Bian, and Siwei Lyu. Robust adversarial perturbation on deep proposal-based models. *arXiv preprint arXiv:1809.05962*, 2018. 2

[27] Yiming Li, Baoyuan Wu, Yan Feng, Yanbo Fan, Yong Jiang, Zhifeng Li, and Shutao Xia. Toward adversarial robustness via semi-supervised robust training. *arXiv preprint arXiv:2003.06974*, 2020. 1

[28] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. Efficient adversarial attacks for visual object tracking. In *European Conference on Computer Vision*, pages 34–50. Springer, 2020. 1

[29] Quanyu Liao, Xin Wang, Bin Kong, Siwei Lyu, Youbing Yin, Qi Song, and Xi Wu. Category-wise attack: Transferable adversarial examples for anchor free object detection. *arXiv preprint arXiv:2003.04367*, 2020. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[33] Ji Liu and Ce Zhang. Distributed learning systems with first-order methods. *arXiv preprint arXiv:2104.05245*, 2021. 4

[34] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International Conference on Learning Representations*, 2018. 3, 6

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 2

[36] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018. 2

[37] You Lu and Bert Huang. Structured output learning with conditional generative flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5005–5012, 2020. 3

[38] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 940–949, 2020. 3

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 2

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

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1, 2, 6

[42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[43] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015. 1, 6

[44] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019. 2

[45] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019. 4

[46] Derui Wang, Chaoran Li, Sheng Wen, Qing-Long Han, Surya Nepal, Xiangyu Zhang, and Yang Xiang. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples. *IEEE Transactions on Cybernetics*, 2021. 2

[47] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018. 1, 2

[48] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017. 1, 2, 3

[49] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4135–4144, 2019. 1

[50] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019. 4

[51] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020. 2, 6

[52] Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2253–2260, 2019. 1

[53] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial at-

tacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2020. 3

[54] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019. 2