A Large-scale Multiple-objective Method for Black-box Attack against Object Detection

Siyuan Liang[1,2], Longkang Li[3], Yanbo Fan[4], Xiaojun Jia[1,2], Jingzhi Li[1,2,⋆], Baoyuan Wu[3,⋆], and Xiaochun Cao[5]

[1] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] School of Data Science, Secure Computing Lab of Big Data, The Chinese University of Hong Kong, Shenzhen, China
[4] Tencent AI Lab, Shenzhen, China
[5] School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China
{liangsiyuan, jiaxiaojun, lijingzhi}@iie.ac.cn; {lilongkang, wubaoyuan}@cuhk.edu.cn; fanyanbo0124@gmail.com; caoxiaochun@mail.sysu.edu.cn

Abstract. Recent studies have shown that detectors based on deep models are vulnerable to adversarial examples, even in the black-box scenario where the attacker cannot access the model information. Most existing attack methods aim to minimize the true positive rate, which often shows poor attack performance, as another sub-optimal bounding box may be detected around the attacked bounding box to be the new true positive one. To settle this challenge, we propose to minimize the true positive rate and maximize the false positive rate, which can encourage more false positive objects to block the generation of new true positive bounding boxes. It is modeled as a multi-objective optimization (MOP) problem, of which the generic algorithm can search the Pareto-optimal. However, our task has more than two million decision variables, leading to low searching efficiency. Thus, we extend the standard **G**enetic **A**lgorithm with **R**andom **S**ubset selection and **D**ivide-and-**C**onquer, called GARSDC, which significantly improves the efficiency. Moreover, to alleviate the sensitivity to population quality in generic algorithms, we generate a gradient-prior initial population, utilizing the transferability between different detectors with similar backbones. Compared with the state-of-art attack methods, GARSDC decreases by an average 12.0 in the mAP and queries by about 1000 times in extensive experiments. Our codes can be found at https://github.com/LiangSiyuan21/GARSDC.

Keywords: Adversarial Learning, Object Detection, Black-box Attack

⋆ Coressponding Author

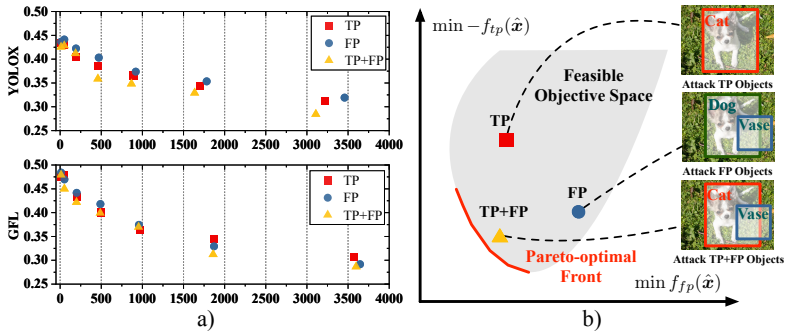

Fig. 1. ***a)*** We show the results of attacking two different models using three objective functions(TP, FP, 'TP+FP'). Experiments show that using 'TP+FP' can decrease the most mAP and reduce most queries. ***b)*** We show the difference of three objective optimizations, and the total consideration of 'TP+FP' makes the solution closer to the Pareto-optimization front.

1 Introduction

With the development of deep learning, object detection [48,35,20,19] has been widely applied in many practical scenarios, such as autonomous driving [21], face recognition [16], industrial detection [15], etc. In object detection, the true positive object refers to the positive object correctly and the false positive object refers to the negative object that is incorrectly marked as positive object.

Recently, the adversarial attack and defense around deep learning have received extensive attention [17,2,39,41,18]. Existing attacks against object detection misclassify the true positive objects from the model, which leads to attack failure. The reason is that another sub-optimal bounding box can replace the attacked bounding box successfully as the new true positive object. Another train of thought, increasing the false positive objects is also an effective attack method in some scenarios. For example, many false objects can obscure the significance of the positive object and lead to autonomous driving system [34] crashes. Therefore, we believe that is critical and desirable to simultaneously optimize true positive and false positive objects recognized by the detector for the following reasons. Firstly, optimizing the objective function with two aspects can expand the attack scenarios. Secondly, we minimize the true positive rate and maximize the false positive rate, which increases false positive objects to block the generation of the true positive object. Thirdly, considering the attack target comprehensively helps decrease the mAP. As shown in Fig. 1, through experiments on YOLOX and GFL models, we prove that optimizing the true positive or false positive objects can attack the detector successfully, and optimizing both of them will achieve better attack performance. Attacks on existing detectors are not comprehensive due to a lack of consideration of false positive objects.

Inspired by the statements above, we reformulate the adversarial attack [37,38,26] against object detection as a large-scale multi-objective optimization problem

(MOP) [7] to decrease true positive objects and increase false positive objects. Our interest focuses on the black-box settings. A large-scale MOP under the black-box setting mainly faces three challenges. Firstly, the conflict between multiple objectives makes it almost impossible to find a solution that optimizes all objectives simultaneously. Secondly, decision variables are extremely large-scaled due to the consistent dimensions of adversarial samples and images (more than two million). Nevertheless, the existing optimization algorithms w.r.t. MOP have poor scalability, and optimizing decision variables with more than one million is especially tough [14]. Thirdly, black-box attacks should reduce queries while increasing the attack rate. To address the above challenges, we use genetic algorithms to find optimal trade-off solutions for MOP, called Pareto-optimal solutions [8]. A genetic algorithm can approximate the entire set of Pareto-optimal in a single run and does not make specific assumptions about the objective functions, such as continuity or differentiability [14]. To settle the poor scalability, we propose a **G**enetic **A**lgorithm based on **R**andom **S**ubset selection and a **D**ivide-and-**C**onquer algorithm to optimize large-scale decision variables, named as GARSDC. This method aims to transform the original search space of MOP using dimensionality reduction and divide-and-conquer, improving the optimization algorithm's searchability by rebalancing exploration and exploitation. The genetic algorithm is sensitive to the population. Thus, we use gradient-based perturbations as the initial population. Moreover, we analyze more than 40 object detection backbones and find out that the perturbation is transferring well in the same backbone. Thus, we generate the chain-based and skip-based perturbations as a mixed initial population with transferability. By combining transfer and query-based attacks, our method substantially decreases the mAP and queries on eight representative detectors than the state-of-the-art method. This paper has the following contributions to the three-fold:

1. We model the adversarial attack problem against object detection as a large-scale multi-objective optimization, which can expand the attack scenarios and help understand the attack mechanism against object detection. Experiments show that this comprehensive modeling helps to decrease the mAP.
2. We design a genetic algorithm based on random subset selection and divide-and-conquer methodology for solving Pareto-optimal solutions, called GARSDC, which improves the searchability of GA by rebalancing the exploration and exploitation of the optimization problem. We generate chain-based and skip-based perturbations as a mixed initial population with gradient-prior, increasing population diversity and improving the algorithm's efficiency.
3. A large number of attack experiments based on different backbone detectors demonstrate the effectiveness and efficiency of GARSDC. Compared with the state-of-art PRFA algorithm, GARSDC reduces by an average 12.0 in the mAP and queries by about 1000 times.

2 Related Work

2.1 Object Detection Based on Deep Learning

In recent years, the latest progress of object detectors mainly focuses on three aspects: Firstly, the improvement of the backbone network, detectors based on different backbones have produced significant differences in accuracy and inference speed. Standard models include SSD [27] based on VGG16, Centernet [10] based on ResNet18 and YOLOX [12] based on yolo-s network. Most models are based on ResNet [33] and FPN [25] series architecture, e.g., Cascade R-CNN [3], Atss [46], Fcos [36], and Freeanchor [47]. Secondly, combining the learning of instance segmentation, such as segmentation annotation in Mask R-CNN [13] and switchable atrous convolution in Detectors [32]. Thirdly, improvements of localization, such as GFL [22] based on generalized focal loss. Our research finds that detectors that focus on different improvements have significant differences in transfer attacks. In addition, compared with detectors based on different backbones, detectors with the same backbone structure are less challenging to transfer, and this phenomenon brings excellent inspiration to our model selection for transfer attacks.

2.2 Black-box Adversarial Attack

Generally speaking, black-box adversarial attacks can be divided into transfer attacks, decision-based attacks, and score-based attacks. The transfer attack, also known as the local surrogate model attack, assumes that the attacker has access to part of the training dataset to train the surrogate model, including adaptive black-box attack [28] and data-free surrogate model attack [49]. The score-based attack allows the attacker to query the classifier and get probabilistic of the model prediction. Representative methods include the square attack [1] based on random search and the black box attack based on transfer prior. Decision-based attacks [5] can obtain less information than the above, allowing the attacker to accept label outputs instead of probabilities. By analyzing the architectural characteristics of the object detector, we improve the efficiency and accuracy of score-based attacks according to the gradient-prior.

2.3 Adversarial Attack against Object Detection

The existing adversarial attack methods for object detection are mainly based on white-box attacks, and the attacker implements the adversarial attack by changing the predicted label of the true positive object. DAG [42], and CAP [45] mainly implement adversarial attacks by fooling the RPN network of two-stage detectors in terms of the types of attack detectors. To increase the generality of the attack algorithm, UEA [40] and TOG [6] exploit transferable adversarial perturbations to attack both the one-stage detector and the two-stage detector simultaneously. PRFA [24] first proposes a query-based black-box attack algorithm to fool existing detectors using a parallel rectangle flipping strategy. This

method also provides a baseline for target detection query attacks. Our proposed algorithm not only surpasses the state-of-the-art algorithm PRFA but also attacks more representative detection models, comprehensively evaluating the robustness of existing detectors.

3 Method

3.1 Simulating Adversarial Examples Generating by MOP

We firstly introduce the background of MOP, including problem definition, non-dominant relations, and Pareto solutions. A MOP problem can be mathematically modeled as:

$$\min F(\hat{\boldsymbol{x}}) = (f_1(\hat{\boldsymbol{x}}), ..., f_K(\hat{\boldsymbol{x}})), \hat{\boldsymbol{x}} = (\hat{x}_1, ...\hat{x}_D) \in \Omega, \tag{1}$$

where there are D decision variables with respect to the decision vector $\hat{\boldsymbol{x}}$, the objective function $F : \Omega \to \mathbf{R}^{\mathrm{K}}$ includes K objective functions, Ω and K represent the decision and objective spaces. Generally speaking, when the $K \geq 2$ and $D \geq 100$, when call this MOP as a large-scale MOP.

Definition 1. Given two feasible solutions $\hat{\boldsymbol{x}}_1$, $\hat{\boldsymbol{x}}_2$ and their objective functions $F(\hat{\boldsymbol{x}}_1)$, $F(\hat{\boldsymbol{x}}_2)$, $\hat{\boldsymbol{x}}_1$ dominates $\hat{\boldsymbol{x}}_2$ (denoted $\hat{\boldsymbol{x}}_1 \prec \hat{\boldsymbol{x}}_2$) if and only if $\forall i \in \{1, ..., K\}, f_i(\hat{\boldsymbol{x}}_1) \leq f_i(\hat{\boldsymbol{x}}_2)$ and $\exists j \in \{1, ..., K\}, f_j(\hat{\boldsymbol{x}}_1) < f_j(\hat{\boldsymbol{x}}_2)$.

Definition 1 describes the dominance relation in the MOP.

Definition 2. A solution $\hat{\boldsymbol{x}}^*$ is Pareto-optimal solution if and only if there exists no $\hat{\boldsymbol{x}}_1 \in \Omega$ such that $F(\hat{\boldsymbol{x}}_1) \prec F(\hat{\boldsymbol{x}}^*)$. We name the set of all Pareto-optimal solutions as the Pareto Set and the corresponding objective vector set as the Pareto front.

Given a large-scale MOP, we describes the Pareto solution in Definition 2. We have a clean image $\boldsymbol{x}$ containing a set of M recognition objects $\mathcal{O}$, that is, $\mathcal{O} = \{o_1, ..., o_M\}$. Each recognition objects o_i is assigned a groud-truth class $o_i^c \in \{1, ..., C\}$, $i \in \{1, ..., M\}$. C is the number of class, the $C = 81$ in the MS-COCO. The object detector H predict N objects as the predicted objects $\mathcal{P} = H(\boldsymbol{x})$ and the corresponding classes $\mathcal{P}^c$ in the clean image $\boldsymbol{x}$. However, limited by training datasets and complex scenes, the objects $\mathcal{P}$ predicted by the detector are not always consistent with the recognized objects $\mathcal{O}$. We define the true positive object as follows: there is a only one object o_i such that the intersection of union between p_j and o_i greater than 0.5 and p_j^c is same with o_i^c, otherwise it is a false positive object [11]. Thus, we decompose the predicted $\mathcal{P}$ objects as true positive objects $\mathcal{TP}$ and false positive objects $\mathcal{FP}$ by recognition objects $\mathcal{O}$. The $|\mathcal{P}| = |\mathcal{TP}| + |\mathcal{FP}|$. Previous adversarial examples $\hat{\boldsymbol{x}}$ with a small $\boldsymbol{\delta}$ attack the detector by reducing the true positive objects. Fig. 1 a) show that attacking the false positive objects alone can also attack the detector. Therefore, we model the adversarial attack as a large-scale MOP: reducing the true positive objects and increasing false positive object. The objective function can be represented as:

$$F(\hat{\boldsymbol{x}}, H(\boldsymbol{x})) = \min(-f_{tp}(\hat{\boldsymbol{x}}, \mathcal{P}), f_{fp}(\hat{\boldsymbol{x}}, \mathcal{P})), s.t.\ \hat{\boldsymbol{x}} = (\boldsymbol{x} + \boldsymbol{\delta}) \in \Omega,\ ||\hat{\mathrm{x}} - \mathrm{x}||_{\mathrm{n}} \leq \epsilon, \tag{2}$$

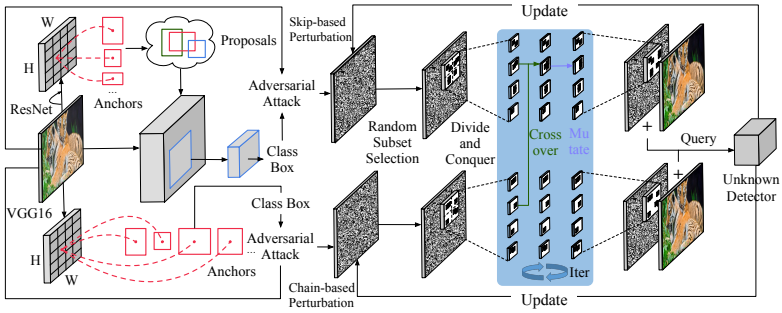

Fig. 2. To optimize multi-objective problems, we propose a **G**enetic **A**lgorithm based on **R**andom **S**ubset selection and a **D**ivide-and-**C**onquer algorithm (GARSDC). The basic flow of the GARSDC attack is shown above, which combines the transfer-based and the query-based attacks against the black-box model.

where n denotes norm. We solve the problem in the weighting method [44]. The Eq. (2) can be written as:

$$\begin{aligned} F(\hat{\boldsymbol{x}}, H(\boldsymbol{x})) &= \min(w_1 * (-f_{tp}(\hat{\boldsymbol{x}}, \mathcal{P})) + w_2 * f_{fp}(\hat{\boldsymbol{x}}, \mathcal{P})) \\ s.t.\ \hat{\boldsymbol{x}} &= (\boldsymbol{x} + \boldsymbol{\delta}) \in \Omega,\ ||\hat{\mathrm{x}} - \mathrm{x}||_{\mathrm{n}} \leq \epsilon, \end{aligned} \tag{3}$$

where $w_i \geq 0$.We use the CW loss [4] as attack functions f_{tp} or f_{fp}:

$$f_{\{tp,fp\}}(\hat{\boldsymbol{x}}, \mathcal{P}) = \sum_{i \in \{\mathcal{TP}, \mathcal{FP}\}}^{|\mathcal{P}|} \left(\max_{l \neq c} (f_{\{tp,fp\}}(\hat{\boldsymbol{x}}, p_i)_l) - f_{\{tp,fp\}}(\hat{\boldsymbol{x}}, p_i)_c \right), \tag{4}$$

where $i \in \mathcal{TP}$ represents the i-th predicted box, p_i is the true positive object in the predicted boxes $\mathcal{P}$. In Eq. (3), we aim to make the labels of true positive objects wrong and protect the false positive objects. Thus, we treat f_{tp} and f_{fp} as untargeted and targeted attacks, respectively.

3.2 Generating Adversarial Examples by Genetic Algorithm

Since the genetic algorithm is based on the nature of the population and does not require additional assumptions (continuous or differentiable) for objective functions, the genetic algorithm can gradually approximate the Pareto-optimal solution in the single queries [14]. We choose the genetic algorithm, which only uses the fitness function to evaluate individuals in the population and search the best individual as the adversarial perturbation. We define the initial population Δ^0 containing P individuals as $\Delta^0 = \{\boldsymbol{\delta}_1^0, ..., \boldsymbol{\delta}_p^0\}$ and the p-th individual fitness $P(\boldsymbol{x} + \boldsymbol{\delta}_p) = F(\boldsymbol{x} + \boldsymbol{\delta}_p)$. The population is iterating in the direction of greater individual fitness. Generating the i-th population Δ^i mainly relies on crossover and mutation. The greater the individual fitness, the more likely it is to be saved as the next population. For example, if $P(\boldsymbol{\delta}_1^i) > P(\boldsymbol{\delta}_2^i)$, then the next individual $\boldsymbol{\delta}_2^{i+1}$ will inherit some features (crossover) of $\boldsymbol{\delta}_1^i$ and mutate. In Fig. 2, the transfer

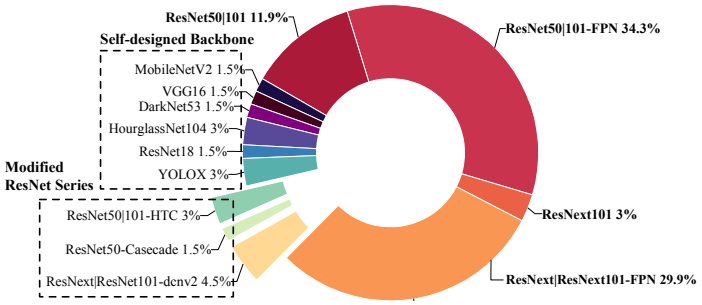

Fig. 3. The investigated results of detectors based on different backbone networks.

attack generate the initial population Δ^0. The iteration stopping condition of population iteration is when reaching the maximum iteration, or the fitness is greater than a certain value. The optimal solution of the population is the individual with the greatest fitness, that is, the adversarial perturbation we need. Since our decision space is too large ($weight{*}height{*}channel$ exceeds millions), it is difficult for general genetic algorithms to converge in limited queries. Next, we will introduce the improved genetic algorithm from the gradient-prior initial population, random subset selection, and divide-and-conquer algorithm.

3.3 Mixed Initial Population Based on Gradient-prior

An excellent initial population can help the genetic algorithm converge more quickly, so finding a suitable initial population for the black-box detector is critical. Although the QAIR [23] algorithm estimates the gradient of the adversarial perturbation by stealing the image retrieval system, the cost of model stealing for the detector is too high. Because the detectors are diverse and the dataset for object detection relies on enormous annotations. Intuitively, we can generate adversarial perturbations with well transferability as an initial population against the detector.

Inspired by that transferable perturbation in image classification can attack different feature networks, we analyze more than 40 deep model-based object detectors and classify their backbone network types. In Fig. 3, detectors based on backbone networks belong to ResNet, ResNet-FPN [25], and their derivatives, e.g., ResNeXt [43], account for more than 80%. We call these backbone networks the ResNet series. The other two types of backbone networks are based on the modified ResNet series, e.g., Detectors [32], or self-designed networks, such as YOLOX [12] based on yolo-s. Therefore, we can roughly divide the current network architecture into three categories and attack against the ResNet series, the modified ResNet series, and self-designed networks. Although detection models vary widely in network structures, both of them use the cross-entropy loss of the prediction boxes. We can implement an adversarial attack by maximizing the cross-entropy loss of all prediction boxes, and the objective function is as

follows:

$$D(\boldsymbol{x}+\boldsymbol{\delta}) = \sum_{i=1}^{|\mathcal{P}|}\sum_{j=1}^{|C|} y_{ij} * \log(\boldsymbol{c}_{ij}) \tag{5}$$

where y_{ij} is one when detector H classify the i-th prediction box into the j-th category, otherwise y_{ij} is zero. $\boldsymbol{c}$ represents the classification probability of the i-th prediction box. We can attack the Eq. (5) using an off-the-shelf transfer attack algorithm, such as TI-FGSM [9].

Although the input can be randomly initialized by adding noise to the clean image, adversarial perturbations based on the same detector lack diversity. To accelerate the genetic algorithm convergence, the diversity of individuals is essential. Therefore, we attack detectors with different backbone networks to generate individuals with differences, and we call this population composed of sexual individuals as the mixed initial population based on gradient-prior. Specifically, we respectively select the initial individuals attacked by the VGG16-based and ResNet-based detectors. In essence, VGG16 and ResNet are different because ResNet is a backbone network with a skip-connection structure, and VGG16 is a chain structure. We refer to the different individuals generated by these two networks as skip-based perturbation and chain-based perturbation.

3.4 Random Subset Selection

The variable space($weight*height*channel$) of the perturbation exceeds millions, and the intuitive idea for solving the large-scale MOP is to decompose high-dimensional decision variables into many low-dimensional sub-components and assign MOP to sub-components through specific strategies, which solves the MOP indirectly by optimizing a portion of the MOP. We will introduce the random subset selection for sub-components and the corresponding MOP decomposition strategy.

Since there are many decision variable combinations for sub-component selection, it is unrealistic to traverse all combinations in a limited number of queries. We use a random subset selection algorithm to sample sub-component in the decision space. We use the index vector $\boldsymbol{s} \in \{0,1\}^D$ for random subset selection. If $s_i = 1$ the i-th element of δ is selected and $s_i = 0$ otherwise. Square attack [1] achieves good black-box attack performance by generating square-shaped adversarial patches through random search in the image. It is feasible to select adversarial patches randomly to attack the detector. This process can be regarded as sampling the sub-component $\boldsymbol{\delta}[\boldsymbol{s}]$ in the decision space $\boldsymbol{\delta}$.

In section 3.2, we introduce the individual fitness. However, the computation of individual fitness is for the overall adversarial perturbation $\boldsymbol{\delta}$ rather than the sub-component $\boldsymbol{\delta}[\boldsymbol{s}]$. We can easily computer the coordinates s^B of sub-component $\boldsymbol{\delta}[\boldsymbol{s}]$ by using $\boldsymbol{s}$, then we assign the predicted boxes $\mathcal{P}$ to the sub-

component and calculate the fintess:

$$f_{\{tp,fp\}}(\hat{\boldsymbol{x}}, \mathcal{P}, s^B) = \sum_{i \in \{\mathcal{TP}, \mathcal{FP}\}}^{|\mathcal{P}|} (\max_{l \neq c}(f_{\{tp,fp\}}(\hat{\boldsymbol{x}}, p_i)_l) - f_{\{tp,fp\}}(\hat{\boldsymbol{x}}, p_i)_c) * IoU(p_i, s^B), \tag{6}$$

where $IoU(a, b)$ denotes intersection over Union between a and b. The sub-component fitness $S(\boldsymbol{\delta}[\boldsymbol{s}])$ defines as follows:

$$\begin{aligned} S(\boldsymbol{\delta}[\boldsymbol{s}]) &= S(\boldsymbol{x} + \boldsymbol{\delta}, H(\boldsymbol{x}), s^B) = S(\hat{\boldsymbol{x}}, H(\boldsymbol{x}), s^B) \\ &= \min(-f_{tp}(\hat{\boldsymbol{x}}, \mathcal{P}, s^B), f_{fp}(\hat{\boldsymbol{x}}, \mathcal{P}, s^B)). \end{aligned} \tag{7}$$

We can judge the relationship between the current sub-component and the predicted boxes by calculating the fitness of individual and sub-component. If there is no connection, we discard the current sub-component. Although PRFA has a similar operation that randomly searches for sub-components in the search space, our method has the following innovations: Firstly, we do not need a priori-guided dimension reduction but instead search the image globally, which can circumvent the risk of that the prior is terrible; secondly, we can use the window fitness to help judge whether the sub-components of random search are helpful for optimization.

3.5 Divide-and-Conquer Algorithm

Although the decision variables are greatly reduced by randomly selecting the sub-component $\boldsymbol{\delta}[\boldsymbol{s}]$, we can still use the divide-and-conquer method for the sub-component $\boldsymbol{\delta}[\boldsymbol{s}]$ to improve the optimization. In Fig. 2, we show the divide-and-conquer process of sub-component $\boldsymbol{\delta}[\boldsymbol{s}]$. Suppose we decompose the index vector $\boldsymbol{s}$ into i parts, that is $\boldsymbol{s} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_i\}$. We can perform the genetic algorithm in the i-th part of the index vector $\boldsymbol{s}_i$ to find a subset $\boldsymbol{u}_i$ with a budget z. Assuming that we have two individuals $\boldsymbol{\delta_1}$ and $\boldsymbol{\delta_2}$, we can calculate the fitness of sub-components $S(\boldsymbol{\delta_1}[\boldsymbol{s}_i])$ and $S(\boldsymbol{\delta_2}[\boldsymbol{s}_i])$, respectively. If $S(\boldsymbol{\delta_1}[\boldsymbol{s}_i]) > S(\boldsymbol{\delta_2}[\boldsymbol{s}_i])$, the $\boldsymbol{\delta_2}[\boldsymbol{s}_i]$ gets the feature from $\boldsymbol{\delta_1}[\boldsymbol{s}_i]$(cross over) and mutates. The $\boldsymbol{\delta_2}[\boldsymbol{s}_i]$ updates and gets the new subset $\boldsymbol{u}_i$ from $\boldsymbol{s}_i$. Merge all new subsets into a set $U = \cup_{j=1}^{i} \boldsymbol{u}_i$. And we find the $\boldsymbol{u}_{i+1}$ in the set U. Then, we return the best individual fitness and the corresponding sub-component $\boldsymbol{\delta}[\boldsymbol{u}_{best}]$.

We will analyze the approximation of the divide-and-conquer algorithm in Lemma 1. For $1 \leq j \leq i$, let $\boldsymbol{b}_j \in \arg\max_{\boldsymbol{u} \subseteq \boldsymbol{s}_j : |\boldsymbol{u}| \leq z} P(\boldsymbol{\delta}[\boldsymbol{u}])$ denotes an optimal subset of $\boldsymbol{s}_i$.

Lemma 1. [29] For any partition of $\boldsymbol{s}$, it holds that

$$\max\{P(\boldsymbol{\delta}[\boldsymbol{b_j}]) | 1 \leq j \leq i\} \geq \{\alpha/i, \gamma_{\emptyset,z}/z\} * OPT, \tag{8}$$

where the γ- and α are submodularity ratios [30]. The OPT denote the value of the objective function in Eq. (3) For any subset $\boldsymbol{u} \subseteq \boldsymbol{s}_j$, there exists another item, the inclusion of which can improve the individual fitness by at least a proportional to the current distance the best solution [31]. Then, we can get the

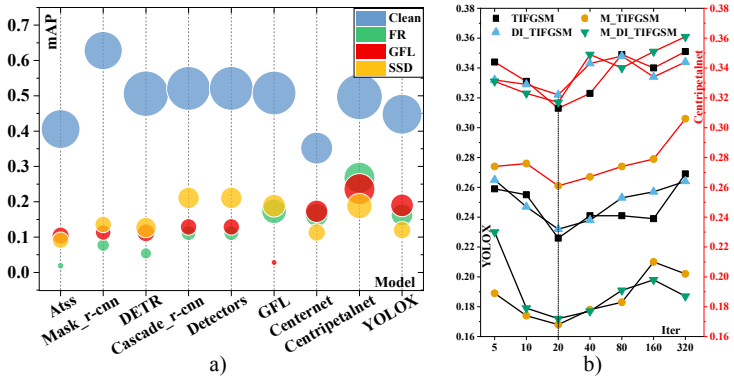

Fig. 4. ***a)*** We evaluate the performance of transfer attacks generated by three backbone networks on nine models, with the vertical axis representing mAP and the circle radius representing recall. ***b)*** We verify the effect of different attack algorithms and iterations on the YOLOX and Centripetalnet.

approximation performance of divide-and-conquer method for random subset selection with monotone objective functions. For random subset selection with a monotone objective function P, our algorithm using $\mathbb{E}[\max\{T_j|1 \leq j \leq i\}] = O(z^2|\boldsymbol{s}|(1+\log i))$ finds a subset $\boldsymbol{u}$ with $\boldsymbol{u} \leq z$ and

$$P(\boldsymbol{\delta}[\boldsymbol{u}]) \geq (1 - e^{-\gamma_{\min}}) * \max\{\alpha/i, \gamma_{\emptyset,z}/z\} * OPT, \tag{9}$$

where $\gamma_{\min} = \min_{\boldsymbol{u} \subseteq \boldsymbol{s}_j : |\boldsymbol{u}| = z-1} \gamma_{\boldsymbol{u},z}$.

The above is the complete GARSDC algorithm. Due to the limitation of the paper space, we put the proofs of Eq. (8) and Eq. (9) and the algorithm flow of GARSDC in the supplementary materials.

4 Experiments

4.1 Experiment Settings

Dataset and Evaluation. The current object detectors use the MS-COCO dataset as a benchmark for evaluating performance. For a fair comparison with PRFA, we adopt the same experimental setup as PRFA and use a part of the MS-COCO validation set as the attack image. We use the evaluation matrix (mAP) to evaluate the detection results of the detector on adversarial examples. The lower the mean average precision, the better the attack effect. We evaluate the efficiency of the algorithm using average queries. Under the limit of 4,000 queries per image, the lower the queries, the higher the attack efficiency.

Victim Models. Section 3.2 divides the investigated object detection models into three categories. We selected two black-box models from the modified ResNet series and the self-designed backbone, GFL, Detectors, YOLOX, and Centernet. Among the object models based on the ResNet series backbone, we

Table 1. An ablation study for GARSDC.

Method	GFL [22]					YOLOX [12]				
	mAP	mAP_S	mAP_M	mAP_L	AQ	mAP	mAP_S	mAP_M	mAP_L	AQ
Clean	0.59	0.36	0.62	0.79	N/A	0.52	0.27	0.55	0.76	N/A
PRFA	0.31	0.17	0.31	0.45	3571	0.31	0.15	0.34	0.51	3220
PRFA_{TP+FP}	0.27	0.16	0.27	0.45	3604	0.28	0.11	0.32	0.48	3109
PRFA_{TA}	0.21	0.07	0.20	0.33	3359	0.31	0.14	0.33	0.50	3175
GA_{TA}	0.25	0.07	0.24	0.44	3401	0.42	0.17	0.46	0.63	3600
GARS_{TA}	0.20	0.09	0.20	0.32	3133	0.29	0.12	0.32	0.48	3201
GARSDC_{TA}	0.18	0.08	0.21	0.32	3037	0.31	0.14	0.34	0.50	3170
GARSDC_{MixTA}	**0.16**	**0.05**	**0.16**	**0.28**	**1838**	**0.23**	**0.10**	**0.28**	**0.42**	**2691**

choose Atss, Casecade R-CNN, Free anchor and Fcos as the black-box attack model. To verify the generation of transferable perturbations, we use Faster R-CNN, GFL and SSD as the white-box attack model and add the transformer-based object detector DETR as the black-box attack model.

Experimental Parameters. The w_1 and w_2 are 0.5 in Eq. (3). For the selection of random subsets, we use a random search strategy similar to Square attack, sample patches with a size of 0.05 times the original image size in the $weight * height * channel$ subspace as the initial random subset, and initial perturbation of the patch. The sampling size is reduced by half when the queries are [20, 100, 400, 1000, 2000]. In the divide-and-conquer phase, we divide the random subset into four parts and the $i = 2$. The population size is set to 2, and the adversarial perturbations respectively generated by Faster R-CNN and SSD iterations 20 times. We set the crossover and mutation rates to 0.8 and 0.3. The norm n is infinity and the budget is 0.05.

4.2 Transferable Perturbation Generation

Firstly, we verify the transferability of generative adversarial perturbations on different detectors. We chose three detectors for the white-box attack: Faster R-CNN (FR) based on ResNet50-FPN, GFL based on modified ResNet series, and SSD based on VGG16. In Fig. 4 a), we respectively show the effect of the transfer attack on nine models. The circle's radius represents the mean recall, and the height represents mAP. We have two observations: Firstly, perturbations generated on detectors of the same type perform well. Secondly, adversarial perturbations generated by detectors based on the ResNet can attack most detectors.

In Fig. 5b), we show the attack effect on Centripalnet(the red axis) and YOLOX(the black axis) of the adversarial perturbations generated by attacking the Faster R-CNN model with different transfer attack methods and different iterations. The M-TIFGSM has the best attack effect. In terms of iterations, the transfer attack has the best effect when about 20 times. As the number of iterations increases, the attack algorithm will gradually overfit. Therefore, we choose M-TIFGSM and iterate 20 times to generate the initial perturbation.

Table 2. Untargeted attacks against detectors based on different backbones.

Method	Atss [46]					Fcos [36]				
	mAP	mAP_S	mAP_M	mAP_L	AQ	mAP	mAP_S	mAP_M	mAP_L	AQ
Clean	0.54	0.32	0.58	0.74	N/A	0.54	0.33	0.56	0.74	N/A
SH	0.40	0.20	0.40	0.59	3852	0.27	0.09	0.37	0.64	3633
SQ	0.23	0.13	0.28	0.31	3505	0.21	0.14	0.20	0.37	3578
PRFA	0.20	0.12	0.25	0.30	3500	0.23	0.15	0.29	0.41	3395
GARSDC	**0.04**	**0.02**	**0.05**	**0.11**	**1837**	**0.15**	**0.09**	**0.17**	**0.28**	**3106**
Method	GFL [22]					Centernet [10]				
	mAP	mAP_S	mAP_M	mAP_L	AQ	mAP	mAP_S	mAP_M	mAP_L	AQ
Clean	0.59	0.36	0.62	0.79	N/A	0.44	0.14	0.45	0.71	N/A
SH	0.43	0.22	0.42	0.59	3904	0.35	0.08	0.33	0.56	3882
SQ	0.33	0.17	0.31	0.50	3751	0.25	0.06	0.27	0.44	3591
PRFA	0.31	0.17	0.31	0.45	3570	0.25	0.07	0.23	0.46	3697
GARSDC	**0.16**	**0.05**	**0.16**	**0.28**	**1838**	**0.12**	**0.03**	**0.13**	**0.23**	**2817**
Method	YOLOX [12]					Detectors [32]				
	mAP	mAP_S	mAP_M	mAP_L	AQ	mAP	mAP_S	mAP_M	mAP_L	AQ
Clean	0.52	0.27	0.56	0.76	N/A	0.61	0.39	0.66	0.82	N/A
SH	0.37	0.15	0.43	0.66	3651	0.51	0.27	0.48	0.72	4000
SQ	0.32	0.17	0.37	0.44	3502	0.45	0.23	0.45	0.62	3957
PRFA	0.31	0.15	0.34	0.51	3220	0.41	0.24	0.43	0.58	3925
GARSDC	**0.23**	**0.10**	**0.28**	**0.42**	**2691**	**0.28**	**0.09**	**0.27**	**0.49**	**2938**

4.3 Ablation Study

To verify the effectiveness of each component of the proposed algorithm, we perform an ablation study on GFL and YOLOX models. PRFA_{TP+FP} represents replacing the optimization objective of PRFA with 'TP+FP'. The subscript TA indicates that using the skip-based perturbation as the initial perturbation. GA stands for genetic algorithm for the entire image. GARS stands for Genetic Algorithm with random subset selection. GARSDC stands for Genetic Algorithm based on random subset selection and divide-and-conquer. $MixTA$ represents using the skip-based perturbation and chain-based perturbation as the mixed-init populations.

In Tab. 1, replacing the objective attack function improves the attack effect by 3 points. Replacing the initialization method of PRFA, the improvement of the attack effect is most apparent, which means that our proposed gradient-prior perturbation is better than the previous. The effect of the GA algorithm is not good because the entire image dimension space is too ample for the genetic algorithm, and it is not easy to optimize. After adding random subset selection and divide-and-conquer, the attack performance of the algorithm has been significantly improved (mAP decreased by 7 points in total). After adding the mixed perturbations mechanism, the difference between populations is more significant than that generated by a single model. Consequently, the queries for genetic algorithms are significantly reduced.

4.4 Attacks against Detector based on Different Backbones

In this section, we compare the attack performance of GARSDC and state-of-the-art black-box algorithms on multiple object detectors. In Tab. 2, we respectively select two object detectors based on three different backbones, which are ATSS based on ResNet101 structure, Fcos based on ResNeXt101 structure, GFL based on ResNeXt101 with deformable convolution, YOLOX based on yolo-s, Centernet based on ResNet18, and Detectors based on RFP and switchable atrous convolution. It is not difficult to see from the experiments that our method reduces by an average 12.0 in the mAP and 980 queries compared with the state-of-the-art algorithm PRFA. The improvement of Atss is the largest, and the attack mAP is 0.04, which may be that our generated skip-based initial perturbation works best to transfer attack against Atss. In addition, our attack effect on Atss, GFL, and Centernet has been improved by more than a half compared with PRFA.

Comparing the size of attack targets, we find that the improvement of our algorithm is mainly focused on small and medium-sized objects. The size of these targets is usually under $64*64$, which is in line with our expectations because the divide-and-conquer method decomposes the random search into smaller search areas, so the attack ability on small and medium objects will be improved. At the same time, the attack of large objects is still difficulty. Comparing the six detectors, Detectors has the most challenging attack (the mAP after the attack is still 0.28), which we think may be related to its structure(switchable atrous convolution), which may inspire us to design a robust architecture for object detection.

4.5 Visual Analysis

This section visualizes the attack results of the square attack, PRFA, and GARSDC. We show the detection results of the three attacks in Fig. 5 a) and the optimization process in Fig. 5 b). During the attack process, we find that GARSDC optimization generates many negative samples and can jump out of local optima during the perturbation iteration process. Both Square attack and PRFA are more likely to fall into local optimal solutions. In Fig. 5 c), we show the detection and segmentation results produced by the three attack methods. We generate multiple small objects and attack pixel classes in a clustered state, which means that the adversarial perturbations we generate can attack the detection and segmentation models.

5 Conclusion

In this paper, we model the adversarial attack against object detection as a large-scale multi-objective optimization problem. Unlike the traditional attack method that reduces true positive objects, we minimize the true positive rate and maximize the false positive rate in the attack process to jointly increase the mAP

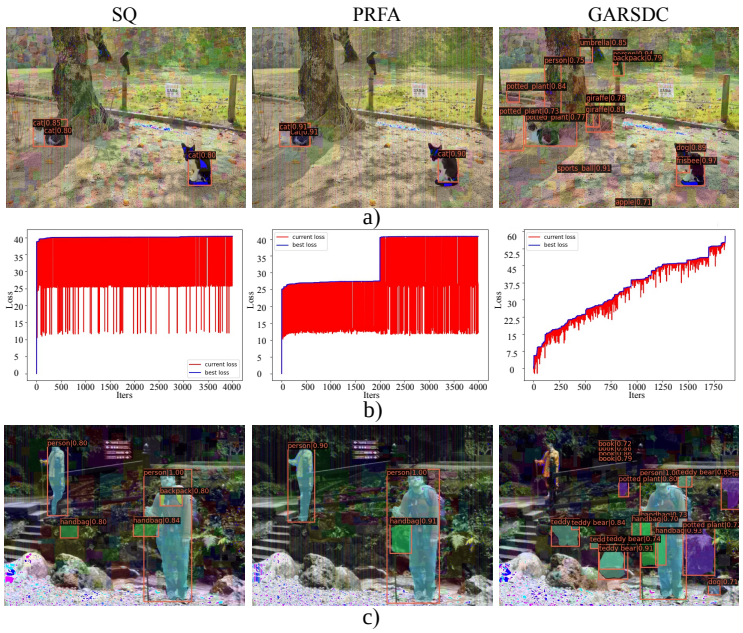

Fig. 5. The SQ [1], PRFA [24], and GARSDC respectively represent the two state-of-art attack methods and our proposed method. We show detection results after attacks in **a)**, the optimization process of three methods in **b)** and segmentation results after attacks in **c)**.

and queries. We propose an efficient genetic algorithm based on random subset selection and divide-and-conquer, optimizing the Pareto-optimal solutions and conquering the challenge of the large-scale decision variables. We generate skip-based and chain-based perturbation by investigating and analyzing more than 40 detection model structures to tackle the problem that the genetic algorithm is sensitive to the population. This gradient-prior population initialization can improve the optimization efficiency of GARSDC. Many attack experiments based on different backbone detectors demonstrate the effectiveness and efficiency of GARSDC. Compared with the state-of-art PRFA algorithm, GARSDC decreases by an average 12.0 in the mAP and queries by nearly 1000 times.

Acknowledgments. Supported by the National Key R&D Program of China under Grant 2020YFB1406704, National Natural Science Foundation of China (No. 62025604), 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, Shenzhen Science and Technology Program under grants No.RCYX20210609103057050 and No.ZDSYS20211021111415025, and Sponsored by CCF-Tencent Open Fund.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision. pp. 484–501. Springer (2020)
2. Bai, J., Chen, B., Li, Y., Wu, D., Guo, W., Xia, S.t., Yang, E.h.: Targeted attack for deep hashing based retrieval. In: European Conference on Computer Vision. pp. 618–634. Springer (2020)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
4. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM workshop on artificial intelligence and security. pp. 3–14 (2017)
5. Chen, J., Jordan, M.I., Wainwright, M.J.: Hopskipjumpattack: A query-efficient decision-based attack. In: 2020 ieee symposium on security and privacy (sp). pp. 1277–1294 (2020)
6. Chow, K.H., Liu, L., Loper, M., Bae, J., Gursoy, M.E., Truex, S., Wei, W., Wu, Y.: Adversarial objectness gradient attacks in real-time object detection systems. In: 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). pp. 263–272 (2020)
7. Deb, K.: Multi-objective optimization. In: Search methodologies, pp. 403–449. Springer (2014)
8. Deb, K., Gupta, H.: Searching for robust pareto-optimal solutions in multi-objective optimization. In: International conference on evolutionary multi-criterion optimization. pp. 150–164. Springer (2005)
9. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321 (2019)
10. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
11. Everingham, M., Zisserman, A., Williams, C.K., Van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorkó, G., et al.: The pascal visual object classes challenge 2007 (voc2007) results (2008)
12. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
14. Hong, W.J., Yang, P., Tang, K.: Evolutionary computation for large-scale multi-objective optimization: A decade of progresses. International Journal of Automation and Computing **18**(2), 155–169 (2021)
15. Hu, Y., Yang, A., Li, H., Sun, Y., Sun, L.: A survey of intrusion detection on industrial control systems. International Journal of Distributed Sensor Networks **14**(8), 1550147718794615 (2018)
16. Jafri, R., Arabnia, H.R.: A survey of face recognition techniques. journal of information processing systems **5**(2), 41–68 (2009)
17. Jia, X., Zhang, Y., Wu, B., Ma, K., Wang, J., Cao, X.: Las-at: Adversarial training with learnable attack strategy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13398–13408 (2022)

18. Jia, X., Zhang, Y., Wu, B., Wang, J., Cao, X.: Boosting fast adversarial training with learnable adversarial initialization. IEEE Transactions on Image Processing **31**, 4417–4430 (2022). https://doi.org/10.1109/TIP.2022.3184255
19. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5830–5840 (2021)
20. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing **29**, 7389–7398 (2020)
21. Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J.Z., Langer, D., Pink, O., Pratt, V., et al.: Towards fully autonomous driving: Systems and algorithms. In: 2011 IEEE intelligent vehicles symposium (IV). pp. 163–168 (2011)
22. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Advances in Neural Information Processing Systems **33**, 21002–21012 (2020)
23. Li, X., Li, J., Chen, Y., Ye, S., He, Y., Wang, S., Su, H., Xue, H.: Qair: Practical query-efficient black-box attacks for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3330–3339 (2021)
24. Liang, S., Wu, B., Fan, Y., Wei, X., Cao, X.: Parallel rectangle flip attack: A query-based black-box attack against object detection. arXiv preprint arXiv:2201.08970 (2022)
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
26. Liu, A., Wang, J., Liu, X., Cao, B., Zhang, C., Yu, H.: Bias-based universal adversarial patch attack for automatic check-out. In: European conference on computer vision. pp. 395–410. Springer (2020)
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
28. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519 (2017)
29. Qian, C., Li, G., Feng, C., Tang, K.: Distributed pareto optimization for subset selection. In: IJCAI. pp. 1492–1498 (2018)
30. Qian, C., Shi, J.C., Yu, Y., Tang, K.: On subset selection with general cost constraints. In: IJCAI. vol. 17, pp. 2613–2619 (2017)
31. Qian, C., Shi, J.C., Yu, Y., Tang, K., Zhou, Z.H.: Parallel pareto optimization for subset selection. In: IJCAI. pp. 1939–1945 (2016)
32. Qiao, S., Chen, L.C., Yuille, A.: Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10213–10224 (2021)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
34. Shim, I., Choi, J., Shin, S., Oh, T.H., Lee, U., Ahn, B., Choi, D.G., Shim, D.H., Kweon, I.S.: An autonomous driving system for unknown environments using a

unified map. IEEE transactions on intelligent transportation systems **16**(4), 1999–2013 (2015)
35. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
36. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
37. Wang, J., Liu, A., Bai, X., Liu, X.: Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. IEEE Transactions on Image Processing **31**, 598–611 (2021)
38. Wang, J., Liu, A., Yin, Z., Liu, S., Tang, S., Liu, X.: Dual attention suppression attack: Generate adversarial camouflage in physical world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8565–8574 (2021)
39. Wang, X., He, K.: Enhancing the transferability of adversarial attacks through variance tuning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
40. Wei, X., Liang, S., Chen, N., Cao, X.: Transferable adversarial attacks for image and video object detection. arXiv preprint arXiv:1811.12641 (2018)
41. Wu, B., Chen, J., Cai, D., He, X., Gu, Q.: Do wider neural networks really help adversarial robustness? arXiv e-prints pp. arXiv–2010 (2020)
42. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1369–1378 (2017)
43. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
44. Zadeh, L.: Optimality and non-scalar-valued performance criteria. IEEE transactions on Automatic Control **8**(1), 59–60 (1963)
45. Zhang, H., Zhou, W., Li, H.: Contextual adversarial attacks for object detection. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2020)
46. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: 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. pp. 9759–9768 (2020)
47. Zhang, X., Wan, F., Liu, C., Ji, R., Ye, Q.: Freeanchor: Learning to match anchors for visual object detection. Advances in neural information processing systems **32** (2019)
48. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)
49. Zhou, M., Wu, J., Liu, Y., Liu, S., Zhu, C.: Dast: Data-free substitute training for adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 234–243 (2020)