Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052196

ARTICLE

Improving Transferable Targeted Adversarial Attack for Object Detection Using RCEN Framework and Logit Loss Optimization

Zhiyi Ding, Lei Sun*, Xiuqing Mao, Leyu Dai and Ruiyang Ding

School of Cryptography Engineering, Information Engineering University, Zhengzhou, 450000, China

*Corresponding Author: Lei Sun. Email: sl20210221@163.com

Received: 26 March 2024 Accepted: 02 August 2024 Published: 12 September 2024

ABSTRACT

Object detection finds wide application in various sectors, including autonomous driving, industry, and healthcare. Recent studies have highlighted the vulnerability of object detection models built using deep neural networks when confronted with carefully crafted adversarial examples. This not only reveals their shortcomings in defending against malicious attacks but also raises widespread concerns about the security of existing systems. Most existing adversarial attack strategies focus primarily on image classification problems, failing to fully exploit the unique characteristics of object detection models, thus resulting in widespread deficiencies in their transferability. Furthermore, previous research has predominantly concentrated on the transferability issues of non-targeted attacks, whereas enhancing the transferability of targeted adversarial examples presents even greater challenges. Traditional attack techniques typically employ cross-entropy as a loss measure, iteratively adjusting adversarial examples to match target categories. However, their inherent limitations restrict their broad applicability and transferability across different models. To address the aforementioned challenges, this study proposes a novel targeted adversarial attack method aimed at enhancing the transferability of adversarial samples across object detection models. Within the framework of iterative attacks, we devise a new objective function designed to mitigate consistency issues arising from cumulative noise and to enhance the separation between target and non-target categories (logit margin). Secondly, a data augmentation framework incorporating random erasing and color transformations is introduced into targeted adversarial attacks. This enhances the diversity of gradients, preventing overfitting to white-box models. Lastly, perturbations are applied only within the specified object's bounding box to reduce the perturbation range, enhancing attack stealthiness. Experiments were conducted on the Microsoft Common Objects in Context (MS COCO) dataset using You Only Look Once version 3 (YOLOv3), You Only Look Once version 8 (YOLOv8), Faster Region-based Convolutional Neural Networks (Faster R-CNN), and RetinaNet. The results demonstrate a significant advantage of the proposed method in black-box settings. Among these, the success rate of RetinaNet transfer attacks reached a maximum of 82.59%.

KEYWORDS

Object detection; model security; targeted attack; gradient diversity

1 Introduction

In recent years, there has been a significant leap in deep learning technology, primarily due to the emergence of large-scale, high-quality datasets and significant improvements in computational hardware performance. These factors have laid a solid foundation for the innovation and optimization of deep learning algorithms. This progress has significantly contributed to the development and practical applications of various computer vision tasks [1], with object detection [2] emerging as a prominently applied technique. Object detection serves as the fundamental basis for numerous computer vision tasks, including instance segmentation [3], image captioning [4], object tracking [5], and more. In comparison to traditional object detection methods [6], the latest generation of deep learning-driven object detection technologies overcomes shortcomings such as low accuracy, poor real-time performance, and computational inefficiency [7]. Consequently, these technologies find widespread applications across diverse domains, including industry, military, aerospace, healthcare, and more. The widespread adoption of deep learning-based object detection techniques has heightened the demands for algorithm robustness and security. This highlights the crucial need to address these aspects in both the development and deployment of such technologies.

While deep learning technologies have brought tremendous convenience to humanity, their vulnerability to adversarial examples poses significant security threats. The concept of adversarial examples was first introduced by Szegedy et al. [8]. It refers to intentionally introducing imperceptible perturbations into input data, leading classifiers to produce incorrect results with high confidence. However, the scope of adversarial example attacks is not limited to image classification tasks; various computer tasks solved using neural networks are susceptible to such security risks [9]. Object detection, as a critical technology in computer vision, is susceptible to adversarial examples, especially in applications such as intelligent surveillance systems. Attackers may deceive detection systems by introducing fake objects into scenes, thereby compromising the system's ability to accurately capture human behavior and potentially creating security vulnerabilities. Another example is that adversarial examples can cause misidentification of pedestrians, traffic signs, vehicles, and other objects by autonomous driving systems, leading to serious traffic accidents and endangering public safety and property. Furthermore, visual tracking, as a critical branch of computer vision [10], aims at accurately locating and tracking specific target objects in video sequences. During this process, the introduction of adversarial samples may severely disrupt the normal operation of tracking models [11], leading to target loss or incorrect tracking. This poses significant challenges for application scenarios that rely on precise target localization, such as video surveillance and robotic vision. The research outcomes of adversarial samples are not limited to a single domain; their impact has transcended into multiple fields, demonstrating profound universality. For instance, adversarial samples designed for image classification tasks may also pose threats to object detection and visual tracking tasks. Initially discovered in image classification tasks, adversarial samples revealed the sensitivity of deep learning models to small, deliberately designed perturbations in input data. With the widespread application and rapid popularity of deep learning technology, the scope of adversarial examples has expanded from image classification to include multiple domains, including object detection, visual tracking, and natural language processing. This expansion has brought about new challenges and imposed more stringent requirements on the robustness of existing models [12].

In the realm of adversarial attacks, the depth of the attacker's understanding of the model determines the type of attack. In a white-box attack scenario, the attacker leverages detailed knowledge of the target model, including its model architecture, parameter values, and gradient data. On the other hand, black-box attacks require attackers to conduct attacks without any knowledge of the internal structure and parameters of the model [13]. Papernot et al. [14] observed that certain attack methods

generate adversarial examples with potent cross-model attack capabilities. This implies that adversarial examples crafted for a known model can deceive models with unknown architectures and parameters. This phenomenon is pervasive across various deep learning models. Methods that rely solely on model transferability without the need for any prior knowledge fall within the scope of black-box attacks. Given that black-box attacks are more closely aligned with real-world application scenarios, they pose more severe security concerns for model deployment. Therefore, research on object detection models facing adversarial attacks serves a dual purpose. On one hand, it helps uncover potential vulnerabilities and flaws in object detectors. On the other hand, it contributes to the assessment and enhancement of the robustness of detection models. This is crucial for ensuring the effectiveness and safety of models in the real world, particularly when addressing crucial issues such as personal and property safety.

One classification of adversarial attacks is based on whether the attack targets a specific class. In non-targeted attacks, the attacker does not specify any particular misclassification target. Conversely, in targeted attacks, the attacker intentionally misclassifies the sample into a predefined incorrect category [15]. In non-targeted attacks, the attacker doesn't have to explicitly specify a particular category but aims to guide the model into producing incorrect recognition results. In contrast, targeted attacks not only seek to make the model recognize inaccurately but also require redirecting the model's output to a specific category, often one that is not the correct label for the sample. Targeted attacks are a more challenging task as attackers have the freedom to control the model's output, potentially causing greater harm. However, correspondingly, targeted attacks are also more challenging to generate, requiring more precise adjustments. In the context of object detection tasks, the distinctions between these two attack types are depicted in Fig. 1. This differentiation is pivotal for comprehending the challenges and potential risks associated with different attack.

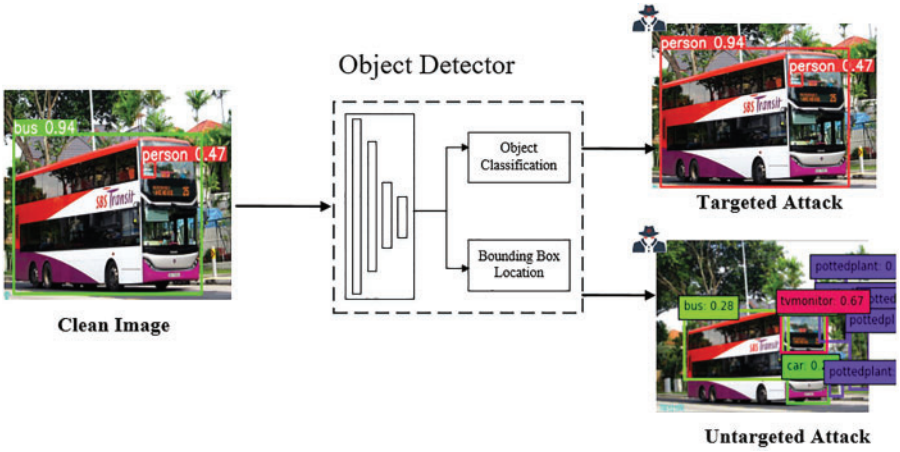

Figure 1: Instances of the targeted and untargeted attack

Currently, adversarial attack methods in the field of image classification have attracted considerable research attention and have led to the development of several influential attack strategies such as the Fast Gradient Sign Method (FGSM) [16], Projected Gradient Descent (PGD) [17], DeepFool [18], Carlini et al. [19], among others. Although there have been corresponding attack algorithms developed for object detection tasks, such as Dense Adversary Generation (DAG) [20], Robust Adversarial Perturbation (RAP) [21], Targeted Adversarial Objectness Gradient Attacks (TOG) [22], and Unified and Efficient Adversary (UEA) [23], these tasks require models to recognize and accurately locate multiple objects within images. This complexity in network architecture and algorithm design increases

the difficulty of adversarial attacks accordingly. Consequently, there is still room for improvement in current research on adversarial attacks in the context of object detection. Furthermore, while there are black-box transfer attack methods developed for image classification tasks, such as a series of FGSM-based methods like Translation Invariant Method (TIM) [24], Momentum Iterative-FGSM (MI-FGSM) [25], Scale Invariant FGSM (SIM) [26], and Variance Tuning Momentum Iterative-FGSM (VMI-FGSM) [27]. However, due to the involvement of multiple objects and more complex decision-making mechanisms in object detection, the existing methods in image classification tasks cannot be easily transferred to adversarial attacks in object detection. Additionally, these algorithms have not explored targeted attack scenarios, and their effectiveness may not be sufficient to deceive black-box models in targeted attack scenarios.

In contrast to common random non-targeted attacks in current research, this paper focuses on achieving targeted attacks that can be customized according to specific attacker intentions. Although theoretically, non-targeted attack methods can be extended to targeted attacks by optimizing the predicted probability of the target class, such an extension overlooks the intrinsic characteristics of targeted attacks. This oversight restricts the capability of adversarial samples to transfer across different models. Firstly, conventional methods typically employ softmax cross-entropy as the loss function, and the gradients obtained through backpropagation are added to the image to generate adversarial examples. However, Li et al. [28] found that as the probability of the target class increases, the magnitude of the gradient continuously decreases, resulting in consistent noise during iterations. This lack of diversity and adaptability in the noise is referred to as "noise curing". To address this issue, Li et al. [28] introduced the Poincaré distance during the optimization process to dynamically increase the gradient amplitude. Secondly, traditional methods often emphasize only the similarity between samples and target categories, without fully considering how to avoid proximity to non-target categories during the iterative optimization process. This limits the transferability of samples across different models. Additionally, methods relying solely on single-image gradients lack sufficient generalizability, which typically results in lower success rates in black-box attack scenarios.

In summary, current adversarial attacks in the field of computer vision primarily focus on image classification tasks. However, unlike image classification, object detection tasks require us to detect and localize multiple objects within an image. Therefore, in adversarial attacks, it is important to consider this multi-object scenario and ensure that the generated adversarial samples can influence the detection and localization of multiple objects. In image classification tasks, evaluation metrics typically measure the impact through misclassification or changes in model confidence. However, in object detection tasks, in addition to misclassification, we also need to consider the accuracy of object localization. Therefore, for adversarial attacks on object detection tasks, the success metric should include both misclassification and the accuracy of object boundary box positioning. Metrics such as mAP can be used to comprehensively assess the effectiveness of the attack.

The main contributions are as follows:

(1) Our study goes beyond treating targeted attacks as a simple derivative of non-targeted attacks. Instead, we propose an optimization objective based on logit loss. This design addresses the noise consistency issue caused by cross-entropy and enhances the separation between target and non-target categories.

(2) In previous work, data augmentation techniques have primarily focused on image classification tasks, while there has been limited research on more complex object detection tasks. Due to the significant differences in the nature of these two types of tasks, and considering the unique characteristics of object detection, this paper introduces a data augmentation framework incorporating Random

Color and Erasing Noise framework (RCEN). This framework addresses the issue of limited gradients during the iterative process, increases gradient diversity, and effectively prevents overfitting of adversarial examples.

(3) This paper avoids adding perturbation noise to the entire image. Instead, it calculates the perturbation mask through Hadamard product and introduces perturbation only within the bounding box of the specified object. This strategy reduces the perturbation range, thereby improving the stealthiness of the attack. Additionally, by introducing noise in the foreground region, particularly in areas with significant impact on the target class, the method efficiently generates adversarial examples.

In summary, our research has made significant advancements in targeted adversarial attacks for object detection tasks. By introducing a novel target function, leveraging RCEN to enhance data augmentation techniques, and refining perturbation methods to focus within object bounding boxes, we have improved attack performance and system robustness. These innovations address the cross-model transferability issues of targeted attacks, demonstrating not only improved attack success rates but also enhanced universality and practicality.

2 Related Work

Here, we will primarily review the background knowledge of object detection and the relevant work on adversarial attacks.

2.1 Object Detection

The groundbreaking advancements in deep learning technology have greatly propelled the development of object detection models, achieving unprecedented levels of detection accuracy and response speed. Currently, widely used object detection models can be broadly categorized into two main types based on whether they perform candidate box extraction: one-stage algorithms and two-stage algorithms. Among them, the representative of two-stage networks is the Regions with Convolutional Neural Networks (R-CNN) series, which includes classical two-stage algorithms such as Fast Regions with Convolutional Neural Networks (Fast R-CNN) [29], Faster R-CNN [30], Mask R-CNN [31], and others. These algorithms approach the object detection task through two main stages: initially, they employ a component called the Region Proposal Network (RPN). This component substitutes the conventional selective search algorithm with the objective of identifying potential target regions. Then, regression operations are performed to determine the potential object box locations, further specifying the detailed category of the object and ultimately generating detection results. The reason for the higher accuracy of two-stage networks lies in their utilization of additional networks. However, it's worth noting that these additional networks also increase the overall structural complexity, resulting in relatively slower detection speeds.

However, single-stage algorithms have the capability to directly output the position and category information of objects in a single feedforward neural network. Representative examples include You Only Look Once (YOLO) [32], YOLOv3 [33], Single Shot MultiBox Detector (SSD) [34], RetinaNet, and others. Unlike traditional sliding window techniques, the YOLO algorithm divides the image into an $S \times S$ grid matrix [32]. Each grid cell is responsible for detecting objects whose center falls within that grid cell. Each grid cell outputs $B \times (5 + C)$ values, where B is the number of anchor boxes, and C is the number of categories. The loss function includes position loss, confidence loss, and category loss, facilitating simultaneous optimization of the accuracy of both position and category. This idea is widely inherited in the YOLO series. Ning et al. combined Faster R-CNN and YOLO, resulting in the design of the SSD network [34]. This network adopts a multi-layered structure to

extract features at different scales, enabling multi-scale object detection. RetinaNet [35] introduces a novel loss function called Focal Loss, alleviating the issue of class imbalance during training. Generally, two-stage detection algorithms exhibit excellent accuracy in object localization and recognition, while one-stage detection algorithms possess faster detection speeds.

2.2 Adversarial Examples

Since Szegedy et al. [8] first demonstrated the sensitivity of deep neural networks to adversarial examples, research into adversarial samples has expanded to encompass other tasks. This paper begins by reviewing several gradient-based non-targeted attack methods and discusses enhancing these methods' transferability to black-box models. Finally, the article summarizes the latest advancements in adversarial attacks on object detection tasks.

2.2.1 Adversarial Strategies in the Field of Image Classification

(1) MI-FGSM (Momentum Iterative-FGSM)

In iterative adversarial attacks such as I-FGSM, perturbations are typically updated with small steps along the direction of the current gradient for rapid descent. However, this greedy strategy often leads the attack to converge to local optima rather than global optima. Dong et al. [25] addressed this issue by incorporating momentum into the optimization algorithm during the iterations of adversarial example generation. This modification helps avoid undesirable local extrema, thereby mitigating the problem of low transferability performance. The generation process with momentum can be expressed as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J\left(x_t^{adv}, \mathrm{y}_{true}\right)}{\left\|\nabla_x J\left(x_t^{adv}, \mathrm{y}_{true}\right)\right\|_1}$$
$$x_{t+1}^{adv} = clip_x^{\varepsilon}\left\{x_t^{adv} + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right\} \tag{1}$$

Here, g_t represents the accumulated gradient values over the first t iterations, μ is the decay rate for the momentum term, and initially, $g_t = 0$, $x_0^{adv} = x$. The experimental results demonstrate that the MI-FGSM algorithm not only maintains its excellent attack performance in white-box settings but also effectively enhances the transferability of adversarial samples in black-box environments. This capability is particularly significant in adversarial attack research.

(2) DI-FGSM (Diverse Input-FGSM)

Dong et al. [24] introduced the Diverse Input Method (DIM), which employs sample diversification techniques in each iteration. In this method, the input sample undergoes random resizing and padding transformations to introduce diversity. This approach effectively alleviates the "overfitting" phenomenon during the generation of adversarial examples, thereby enhancing the success rate of black-box attacks. The process with sample diversification can be expressed as follows:

$$T\left(x_t^{adv}, p\right) = \begin{cases} T\left(x_t^{adv}\right), & \text{with probability p} \\ x_t^{adv}, & \text{with probability } 1 - \mathrm{p} \end{cases}$$
$$x_{t+1}^{adv} = clip_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_x J(T(x_t^{adv}, p), \mathrm{y}_{true})\} \tag{2}$$

Here, $T(\cdot)$ represents an image transformation function, and p represents the probability of applying this transformation in each iteration. The image transformation function $T(\cdot)$ could involve operations like resizing and padding, as mentioned earlier. The probability p controls how often these

transformations are applied during the iterative process. A higher value of p would mean a higher likelihood of applying the transformation in each iteration.

(3) SI-NI-FGSM (Scale-Invariant Nesterov Iterative Method)

In the study by Lin et al. [26], the creation of adversarial samples is conceptualized as an optimization task aimed at finding the optimal solution. Nesterov accelerated gradient is employed to quickly escape local minima during the iterative process. Additionally, the SIM method enhances the applicability of attack strategies across different models by exploiting the invariance of convolutional neural networks (CNNs) to inputs at various scales. This can be described as follows:

$$\underset{x_{adv}}{argmax}\frac{1}{m}\sum_{i=0}^{m} J\left(S_i\left(x_t^{adv}\right), \mathrm{y}_{true}\right),\ s.t.\ \|x_{adv}-x\|_\infty \le \varepsilon \tag{3}$$

In this approach, the original image x is expanded into $s_i(x) = \frac{1}{2^i}$ scaled versions, each scaled by a factor $\frac{1}{2^i}$; the total number of copies is denoted by m. Furthermore, this process integrates TIM and DIM methods, forming a comprehensive attack strategy termed SI-NI-TI-DIM, aimed at enhancing the stability and effectiveness of the attack.

(4) Admix

Wang et al. [36] proposed a novel adversarial attack method called Admix, which is based on input transformation. Specifically, this method takes the original image x as the primary image and mixes it with randomly selected auxiliary images from other categories. We use mathematical language to express the construction process of input transformations:

$$\tilde{x} = \gamma \cdot x + \eta' \cdot x' = \gamma \cdot (x + \eta \cdot x') \tag{4}$$

In this context, $\eta = \eta'/\gamma, \gamma \in [0,\ 1], \eta' \in [0,\ \gamma)$, these parameters control the proportions of the original image and the randomly sampled image in the mixed image. This ensures that the auxiliary image, denoted as "x'," always occupies a smaller portion in the merged image compared to the primary image "$\tilde{x}$." Additionally, this approach does not mix the labels; instead, it utilizes the original label of the primary image x.

2.2.2 Adversarial Example Analysis for Object Detection Algorithms

Research on adversarial attacks in object detection is built upon the foundation of adversarial attacks in image classification. Whether in object detection or image classification, similar techniques can be employed for generating adversarial examples. Among these, gradient-based methods are considered universal for adversarial attacks. While originally designed and validated for image classification tasks, they exhibit suboptimal performance in object detection due to inadequate adaptation to its distinctive characteristics. In the context of object detection, researchers generate effective adversarial samples by leveraging various loss types derived from model outputs. These functions include category loss to alter classification outcomes, confidence loss to reduce the model's prediction certainty, and localization loss to refine the positioning of detected targets. Similar to image classification tasks, gradient-based attack methods leverage gradient information to introduce specific perturbations to the original image. These perturbations, although minute, are amplified throughout the entire forward propagation process of the attacked detector, thereby being sufficient to alter the model's prediction outcomes. This paper briefly introduces four representative attack algorithms targeting object detection models: Dense Adversary Generation (DAG) [20], Robust Adversarial

Perturbation (RAP) [21], Targeted Object Generation (TOG) [22], and Contextual Adversarial Perturbation (CAP) [37].

Xie et al.'s [20] research extends the application of adversarial samples beyond image classification to encompass broader computer vision tasks such as semantic segmentation and object detection. Their proposed Dense Adversary Generation (DAG) attack method introduces a novel perspective on adversarial attacks in these domains. This method initially assigns incorrect labels to candidate proposal boxes generated by the Region Proposal Network (RPN) in two-stage detectors. Through iterative gradient attacks employing backward propagation, it enhances the scores of misclassified categories, ultimately leading to the misclassification of all Regions of Interest (ROI) produced by the detector. Similarly, RAP [21] is also a gradient-based iterative attack method primarily targeting the RPN component in two-stage algorithms. It combines category loss and position loss in object detection to simultaneously attack both the classification and position information of objects. This diversifies the attack, disrupting not only the classification but also affecting the position of objects, resulting in a failure of object localization. It's worth noting that these two algorithms are primarily tailored for two-stage object detection models, showing suboptimal generalization to other detection models. CAP [37] introduces a joint training strategy and context information. It incorporates context loss into the loss function to disrupt object context information, suppressing foreground scores, thus achieving stronger attack effects. However, CAP's attack cost is influenced by the size of the contextual region, and it exhibits poor performance in terms of transferability. Chow et al. [22] proposed a gradient-based universal attack method that can simultaneously target both two-stage and one-stage object detectors, known as Targeted Object Generation (TOG). TOG modifies input images during each backward propagation while keeping the network parameters fixed, implementing various types of attacks, including false-positive attacks, disappearance attacks, and targeted attacks. It is important to note that TOG is primarily designed for white-box model attacks, where the attacker has knowledge of the structure and parameters of the object detection model. Therefore, its transfer attack capability is relatively weak for black-box models.

In summary, adversarial attacks in object detection follow similar patterns, utilizing gradient-based methods like I-FGSM or PGD for backpropagation. However, there are scenarios where access to the model's gradients may be restricted, prompting the need for attacks under black-box conditions. Leveraging the transferable nature of adversarial examples, those generated through white-box attacks can be adapted for use in black-box attacks. Therefore, to thoroughly evaluate the robustness of object detection models under black-box conditions, this paper proposes a targeted attack method and conducts research on transfer attacks.

3 Methodology

In this section, we first elaborate on the motivation and framework design of our study. Subsequently, we demonstrate how the new objective function is integrated into the MI-FGSM algorithm. Finally, we discuss the role of the data augmentation framework in addressing the gradient masking issue in iterative attacks.

Targeted adversarial attacks first appeared in classification tasks, where adversarial samples, after passing through the classifier, produce incorrect results for a specified category, such as classifying a dog as a cat. Similarly, in object detection, targeted attacks require adversarial samples to produce specified incorrect results. However, unlike classification, which outputs only one category, object detection outputs not only categories but also bounding boxes. Therefore, targeted attacks in object

detection are more complex and varied. In this paper, we refer to changing the output category without affecting the detection box position.

Given an input source image x, the object detector F can correctly detect the objects initially present in the image without suffering any attack, i.e., $F(x) = O^s$, where each target contains both position and category information, and O^s is a subset dataset of the entire object $\{O^1, O^2, O^3, \ldots, O^n\}$. The aim of the proposed targeted attack pattern is to generate an adversarial sample $x^{adv} = x + \delta$ such that it can produce a new object O^t in the original region, where O^t only changes the output category without affecting the detection box position. Therefore, the optimization for targeted attacks can be represented by the following formula:

$$x^{adv} = \arg\min_x \mathcal{L}\left(F(x), o^t\right),\ s.t. \|x^{adv} - x\|_\infty \leq \epsilon \tag{5}$$

To ensure that the perturbation δ remains unnoticeable to the human eye, we have set it to be within a very restricted range of variation. Specifically, this paper utilizes the L_∞ as a constraint, stipulating that the maximum value of δ a should not surpass the set upper bound ε, i.e., $\|\delta\|_\infty \leq \varepsilon$.

3.1 Overall Framework

In object detection tasks, the algorithm's primary responsibility lies in accurately identifying various objects in an image, assigning corresponding class labels, and precisely providing spatial location information for these objects. This study focuses on altering the category aspect of object detection outputs, aiming to explore how changes in class information affect the detection system. In previous adversarial attack work on targeted object detection, the approaches were often transferred from the field of image classification, employing gradient-based methods like I-FGSM and PGD iteratively. The advantages lie in their simplicity, comprehensibility, and high efficiency in generating adversarial examples. However, solely relying on maximizing the probability of the target category proves challenging for achieving robust performance and transferability. Through a thorough analysis of existing literature, we found that research specifically addressing the transferability of targeted attacks in the realm of object detection is currently relatively limited.

In previous research, the extension of non-targeted attacks to targeted attacks typically involved using the cross-entropy loss function and applying iterative methods like I-FGSM or PGD to generate targeted adversarial examples. However, this approach faces challenges in deceiving black-box models.

Li et al. [28] found in their study that as the number of iterations increases, the gradient vanishing problem becomes prominent. Therefore, this paper initially adopts a strategy by directly maximizing the logit output of the target category. This moves the adversarial examples closer to the target category while simultaneously forcing them away from non-target categories, enhancing their transferability.

Moreover, traditional methods often rely too much on the gradient information of a single image, leading to poor generalization of adversarial examples. This paper introduces a data augmentation framework called RCEN (Random Color and Erasing Noise). It increases the diversity of gradients through random color transformation and noise erasing, effectively preventing overfitting of adversarial examples. Finally, the paper employs the Hadamard product to add noise only within the foreground region of the target box, reducing the perturbation range. As shown in Fig. 2, the overall structure of this framework is clearly illustrated.

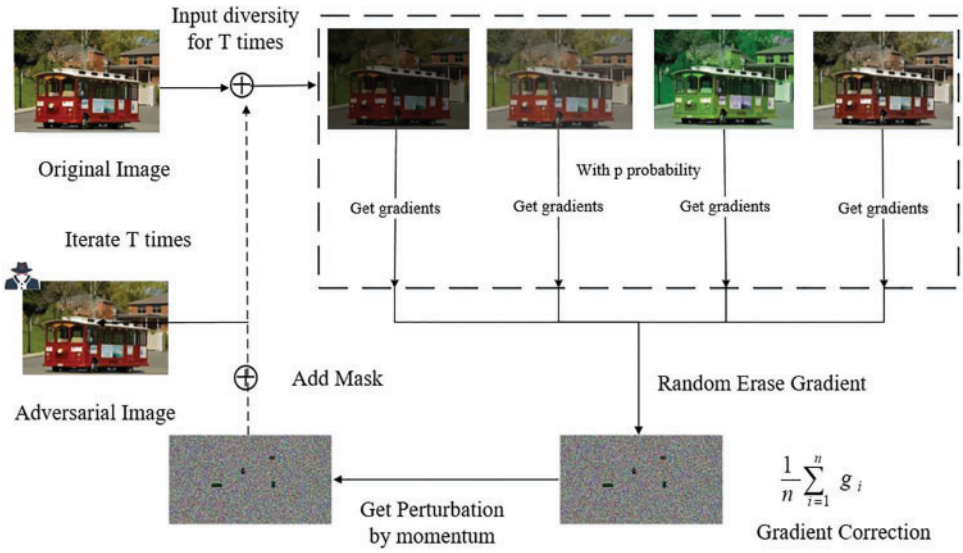

Figure 2: The overall framework of the proposed method

3.2 Objective Function

The exploration in the field of object detection has benefited from the effectiveness of I-FGSM (Iterative Fast Gradient Sign Method) in classification problems. By employing reverse learning processes and maximizing the confidence of the target class, researchers have been able to create targeted adversarial examples. This principle can be represented by the following formula:

$$L_{\text{target}}\left(x^{adv},\ y\right) = L_{\text{target\_obj}}\left(x^{adv},\ y_{obj}\right) + L_{\text{target\_loc}}\left(x^{adv},\ y_{loc}\right) + L_{\text{target\_cls}}\left(x^{adv},\ y_{tar}\right) \tag{6}$$

$$x_{t+1}^{adv} = clip_{x}^{\varepsilon}\left\{x_{t}^{adv} - \alpha \cdot sign\left(\nabla_{x_t^{adv}} L_{\text{target}}\left(x_{t}^{adv},\ y\right)\right)\right\} \tag{7}$$

In this context, the variables are defined as follows: ($L_{\text{target\_obj}}$) represents the confidence loss, indicating the probability of detecting an object; ($L_{\text{target\_loc}}$) denotes the positional loss, ensuring that the adversarial example's detection box does not undergo significant displacement; (y_{tar}) stands for the target category we aim to attack, guiding the detector to output the specified category; and minimizing ($L_{\text{target\_cls}}$) is equivalent to steering the detector closer to the incorrect category. $\nabla_{x_t^{adv}} L\left(x_t^{adv},\ y\right)$ signifies the gradient of the loss function concerning the pixels in the image x. The hyperparameter ϵ serves as the offset, controlling the maximum magnitude of added perturbation. Typically, ϵ is set to a sufficiently small value to make it challenging to discern visually. As depicted in Fig. 3, this results in a decrease in the confidence of the deep learning model in correctly classifying the object's original category, and after multiple iterations, it leads to erroneous classification outcomes. However, we observe that existing methods have not effectively leveraged the inherent differences between non-targeted attacks and targeted attacks, leading to a relatively poor transferability of targeted attacks.

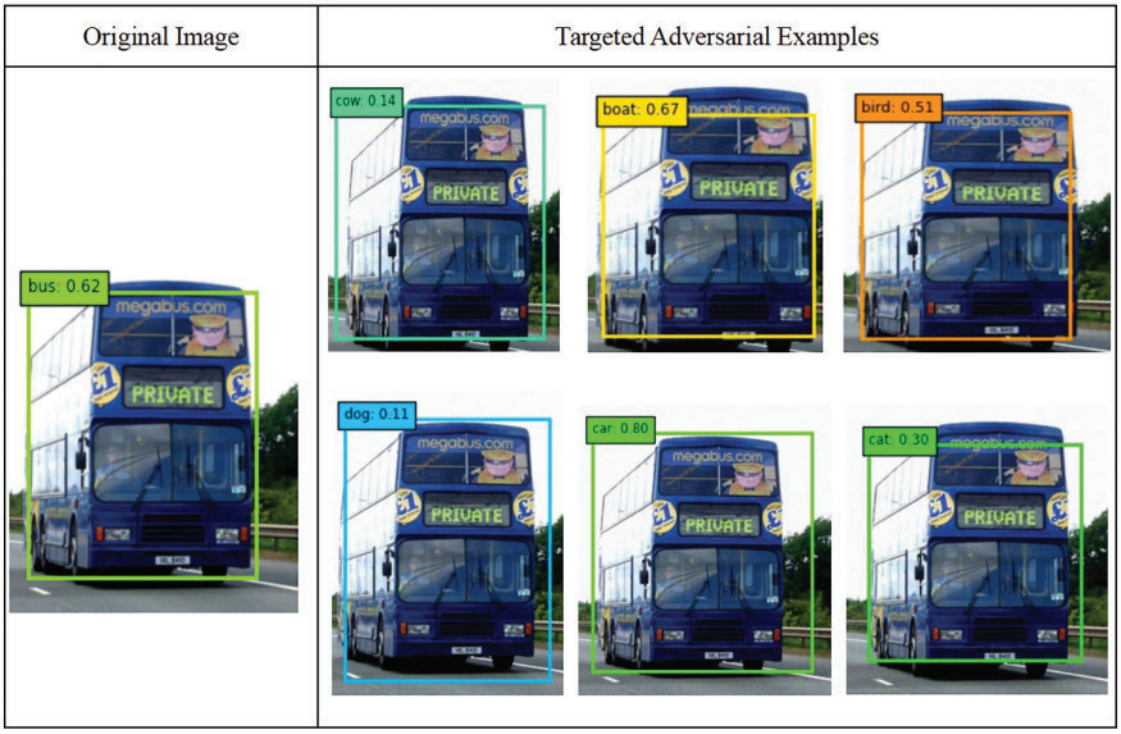

Figure 3: The instances of the targeted attack

Presently, in white-box attacks, ($L_{\text{target\_cls}}$) is typically represented using softmax cross-entropy as the loss function, as illustrated in Fig. 4. However, as pointed out in [28], the cross-entropy loss in the optimization process of attacks can result in gradients gradually diminishing and ultimately vanishing with an increase in the number of iterations. This phenomenon can be expressed through Eqs. (8) and (9):

$$L_{CE} = -1 \cdot \log(p_t) = -\log\left(\frac{e^{z_t}}{\sum e^{z_j}}\right) = -z_t + \log\left(\sum e^{z_j}\right) \tag{8}$$

$$\frac{\partial L_{CE}}{\partial z_t} = -1 + \frac{\partial \log\left(\sum e^{z_j}\right)}{\partial e^{z_t}} \cdot \frac{\partial e^{z_t}}{\partial z_t} = -1 + \frac{e^{z_t}}{\sum e^{z_j}} = -1 + p_t \tag{9}$$

In this context, p_t represents the probability of the target category 't' that we intend to attack, and z_t is the logit output value for category t. As shown in Eq. (8), the gradient is linearly related to p_t, and with the increase in the number of iterations, the probability of the target category gradually approaches 1. The monotonic decrease in gradients causes continuously introduced noise to converge to similar directions during momentum accumulation, leading to the problem of noise fixation. This gradual gradient change presents challenges in optimizing adversarial attack strategies because as the gradient decreases, the perturbations' ability to induce misclassification also diminishes. This hinders the transferability of adversarial samples across different models.

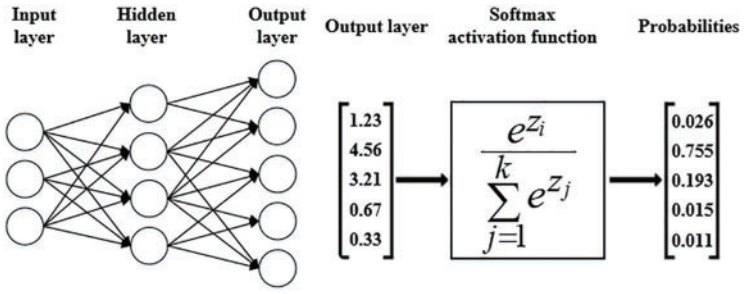

Figure 4: Schematic diagram of neural network using softmax activation function for cross entropy calculation

To alleviate this issue, the motivation of this study is to directly maximize the logit value of the target category output, pulling the adversarial examples closer to the position of the target class. Therefore, the gradient does not decrease with an increase in the number of iterations. This process can be represented by Eqs. (10) and (11):

$$L_{Logit} = -\mathrm{z}_t\left(x^{adv}\right) \tag{10}$$

$$\frac{\partial L_{Logit}}{\partial \mathrm{z}_t} = -1 \tag{11}$$

where $\mathrm{z}_t(\cdot)$ is the logit output of the target category t during the forward propagation of the model.

Another motivation for this study is that in targeted attacks, traditional methods often focus solely on maximizing the probability of the target class while neglecting the proximity of adversarial examples to non-target classes. Li et al. [28] pointed out in his research that these examples only approach the target class, and the distance to the true class is not sufficient to generate adversarial examples with good transferability. Thus, although these techniques perform well in fully informed white-box scenarios, some adversarial examples are still classified into the original category in black-box settings, thereby reducing the distinctiveness of targeted adversarial samples. Therefore, this study focuses on utilizing non-targeted classes to generate more effective targeted adversarial samples.

In summary, when successfully achieving targeted attacks, the logit output z_t for the target class should be higher than the output for z_{nt} any other non-target class. Within the context of this study, the optimization process advocates incrementally increasing the difference in logits to encourage adversarial samples to tightly integrate with the target class and diverge from non-target classes. This strategy not only enhances the discernibility between the target class and its actual category but also reinforces the distinction from other non-target classes. As a result, it enhances the versatility of the attack strategy across various environments. Building upon the aforementioned ideas, this paper introduces an objective function, which can be formulated as follows:

$$L_{\text{target\_cls}}\left(x^{adv},\ y_{tar}\right) = \left\{\max_{i\neq t}\left(Z\left(x^{adv}\right)_i\right)\right\} - Z\left(x^{adv}\right)_t \tag{12}$$

where t represents the target class we aim to attack, and $Z(\cdot)$ is the logit output for the corresponding class.

By focusing on both increasing the logit value for the target class and creating a significant margin between the target and non-target classes, the proposed method aims to generate adversarial samples that are not only more successful in fooling the target model but also exhibit enhanced robustness

across different models or scenarios. This strategy aligns with the goal of improving the transferability of targeted attacks, making them more effective and reliable in real-world applications.

3.3 Framework of Data Augmentation

In the context of image classification tasks, Lin et al. [26] astutely compared the process of constructing adversarial examples with the training paradigm of neural networks, revealing a fundamental alignment between the two. During model training, the gradient descent algorithm aims to optimize model parameters to minimize the loss function L. Similarly, in the construction of adversarial examples, the gradient descent strategy shifts to adjusting adversarial perturbations. This process is precisely described by Eqs. (13) and (14), illustrating their technical similarity:

$$\theta = \theta - \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial \theta} \tag{13}$$

$$x = x + \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x} \tag{14}$$

where the term L represents the loss function specific to the model, while x denotes the image, y signifies the label corresponding to x, and θ represents the model parameters.

In adversarial attacks and model training, despite potentially different objectives, both involve optimization processes aimed at finding optimal parameters or perturbations. Therefore, from a deeper perspective, the generation of adversarial samples can be viewed as a specialized form of "model training" on a specific source model. Subsequently, the performance of these samples on unknown models can be likened to testing the generalization ability of neural networks. Given the contradiction between the demand for large amounts of data in deep learning training and the scarcity of real-world data, data augmentation techniques have emerged. Data augmentation strategies, including scale variations, spatial transformations, region cropping, noise injection, and color adjustments, increase the diversity of datasets. This helps mitigate overfitting in small-sample learning, thereby enhancing model robustness and generalization. This paper further explores the potential application of data augmentation techniques in adversarial sample attacks to improve attack effectiveness. Our approach not only improves the quality of adversarial sample generation on specific source models but also significantly enhances their transferability to unknown models, thereby improving the generalization capability of adversarial attacks.

Traditional adversarial attack algorithms such as I-FGSM and PGD iterate through gradients, offering simplicity, clarity, and high efficiency in sample generation. However, they fall slightly short in attack intensity and transferability. To address this, researchers have implemented data augmentation techniques including translation, scaling, and Admix to obtain more robust adversarial samples. Even if attackers only have access to the surrogate model, data augmentation can still enhance the transferability of adversarial examples across models. In this context, the concept of "Preserving Loss Transformation" is introduced [26]. It defines a specific type of input transformation that operates on images without altering the model's output loss.

Definition 1: Consider an image x with its label y, and a model $f(x)$ along with its corresponding loss function $J(x, y)$. If there exists an transformation T such that for all x, $J(f(T(x)), y) \approx J(f(x), y)$, then T is termed as a loss-preserving transformation.

For object detection, its inherent complexity and the demand for accurate bounding boxes require data augmentation methods to particularly focus on the relative positional changes within the target regions during processing. Currently, transfer attack algorithms based on data augmentation, such

as DIM, TIM, SIM, focus on transformations of adversarial samples as a whole, leveraging the convolutional neural networks (CNNs) for specific transformation invariances. It is noteworthy that color transformation exhibits the characteristics of preserving loss transformations in object detection, where the object category and position remain unchanged before and after transformation. Therefore, this paper focuses on random color transformations (including adjustments in brightness, contrast, saturation, and hue) as an enhancement method.

When considering optimization strategies, data augmentation strategies demonstrate significant potential when combined with momentum-based MI-FGSM algorithms. This combination promises to provide superior strategies for adversarial attacks. By computing the weighted average of gradients obtained from randomly combined color transformations and combining them with a momentum term, this paper proposes a novel optimization framework, formulated as:

$$g_t = \frac{1}{n}\sum_{i=1}^{n} \nabla_{x_i^{adv}} L_{\text{target}}\left(T\left(x_i^{adv}\right),\ y\right) \tag{15}$$

$$g_t = \mu \cdot g_{t-1} + \frac{g_t}{\|g_t\|_1} \tag{16}$$

Among them, n is the number of randomly selected data augmentation methods for this iteration, μ is the decay rate of the momentum term, where $T(\cdot)$ is a random color transformation with a probability of p, and $L_{\text{target}}(\cdot)$ represents the loss output during forward propagation of the model.

Additionally, as shown in Fig. 5, random erasing techniques are introduced. By randomly masking noisy regions to simulate occlusion scenarios, this approach encourages adversarial perturbations to learn more robust feature representations. This effectively mitigates overfitting and enhances the attack potential of adversarial samples. This process can be viewed as an extension of the dropout mechanism in training to the domain of adversarial sample generation. Based on this characteristic, in each iteration, this paper incorporates random erasing of noise, formulated as:

$$g_t = \text{Random Erase}\,(g_t) \tag{17}$$

In this expression, g_t represents the adversarial perturbation generated at the (t)-th iteration.

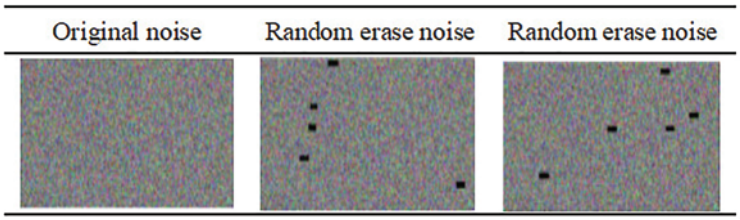

Figure 5: The instances of random erasing noise

In conclusion, this paper introduces a data augmentation framework named RCEN (Random Color and Erasing Noise transform), which integrates random color transformations and noise erasing into the generation process of adversarial samples. The framework aims to mitigate overfitting risks and enhance attack success rates and transferability through these data augmentation techniques.

3.4 Partial Perturbation

Due to the focus on targeted attacks in this paper, particularly on manipulating the category of objects, the influence of the foreground region on object classification is significant. Therefore,

this study employs a method of partial perturbations, introducing adversarial noise into the target bounding box, perturbing only the foreground region of the image. This ensures that the perturbation is concentrated primarily in regions crucial for classification decisions. Specifically, we designed a mask matrix of the same size as the original image. In this matrix, the positions corresponding to the target, i.e., the values within the foreground region, are set to 1, while the values in other areas are set to 0. As illustrated in Fig. 6, we use the Hadamard product operation to multiply this mask with the corresponding pixels of the updated noise, ensuring that the perturbation is applied only to the foreground area targeted for the attack. This strategy enhances the stealthiness of the attack by confining the perturbation to the foreground region. This focus ensures that the perturbation is primarily located on the target object, minimizing unnecessary attention in the background or other irrelevant areas. Furthermore, by introducing noise in the foreground area, which has a significant impact on the classification of the target category, this approach proves to be more effective in generating adversarial examples. In summary, the overlaid perturbation can be expressed as follows:

$$g_t = M \odot g_t \tag{18}$$

where M represents the mask matrix, delineating the foreground region within the target bounding box.

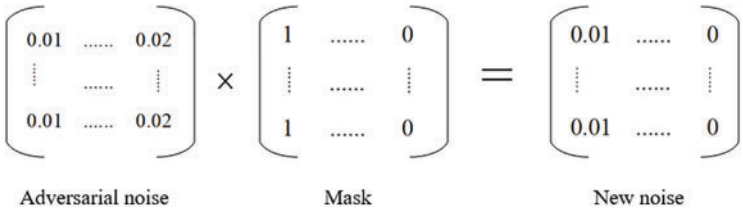

Figure 6: Obtaining partial perturbations via hadamard product

3.5 Attack Algorithm

This study proposes a novel strategy for generating adversarial examples, which integrates optimization based on logit loss and data enhancement techniques such as random color transformation and random noise erasure. This approach aims to create a more efficient means of attack, named RCEN-MI-FGSM. The steps of this approach are described in detail in the following section as shown in Algorithm 1.

Algorithm 1: RCEN-MI-FGSM

Input: Clean benign sample: x; True label: y; Object detector: f; Targeted Loss function: L_{target}; Learning Rate: α; Disturbance limit: ε; Iteration limit T; Decay factor rate μ.
Output: An Adversarial example x_t^{adv}
$x_0^{adv} = x$, $g_0 = 0$
1 *for* t=1 *to* T-1 *do*
2 1 $g_t = \nabla_{x_t^{adv}} L_{\text{target}}\left(f\left(x_t^{adv}, y\right)\right)$, n = 1 According to Eqs. (6) and (11), compute the gradients.
3 2 if color transformation with probability p (brightness, contrast)
4 3 n+=1, $x_t^{adv} = random\ Brightness\left(x_t^{adv}\right)$,
5 4 $g_t += \nabla_{x_t^{adv}} L_{\text{target}}\left(f\left(x_t^{adv}, y\right)\right)$
6 5 if color transformation with probability p (hue, saturation)
7 6 n+=1, $x_t^{adv} = random\ Color\left(x_t^{adv}\right)$,

(Continued)

Algorithm 1 (continued)

$g_t += \nabla_{x_t^{adv}} L_{\text{target}}\left(f\left(x_t^{adv},\ y\right)\right)$

$g_t = \sum_{i=1}^{n} \frac{1}{n} g_t$, Weighted average gradient after stochastic data augmentation

$g_t = R\text{andom Erase}\ (g_t)$, Apply random noise erasing within the specified bounding box

$g_t = \mu \cdot g_{t-1} + \frac{g_t}{\|g_t\|_1}$, Optimizing the iterative process with momentum

$g_t = M \odot g_t$, Add noise to the foreground region

$x_t^{adv} = clip_x^{\varepsilon}\left\{x_{t-1}^{adv} - \alpha \cdot \text{sign}\ (g_t)\right\}$

$t{=}t{+}1$

Return x_t^{adv}

4 Data and Analysis of Experimental Results

This study conducted comprehensive experimental validation using the MS COCO dataset to demonstrate the proposed attack method, which integrates data augmentation and logit loss optimization techniques. Details of the experimental setup are provided in Section 4.1, while a thorough explanation of evaluation metrics can be found in Section 4.2. Section 4.3 presents performance comparisons between our method and existing baseline methods under various conditions. Additionally, the importance of combining data augmentation and logit loss optimization is demonstrated through ablation experiments in Section 4.3.3.

4.1 Dataset Overview and Experimental Design

Models: The experimental section of this study employed both one-stage and two-stage object detection algorithms, including YOLOv3, YOLOv8, RetinaNet, and Faster R-CNN. These models were all trained on the MS COCO dataset to ensure their effectiveness in practical applications.

Environment: The experimental phase of this study was conducted on a high-performance computer system equipped with a top-of-the-line NVIDIA GeForce RTX 3080 Ti graphics card. The experimental framework employed PyTorch.

Baselines: This paper extensively explores and compares several mainstream adversarial attack methods that integrate data augmentation strategies and gradient iteration processes, including TIM, SIM, and MI-FGSM. These methods have demonstrated their significance and practicality in both academic and industrial settings.

Dataset: The MS COCO dataset is crucial for object detection research due to its rich variety of everyday scenes and detailed annotations across 80 object categories. In our study's experimental design, we randomly selected 2000 test images, all successfully identified by the models. Our experiments aimed to change the object categories in these images to "person".

Implementation details: During the computation of evaluation metrics, we placed particular emphasis on maintaining consistency in hyperparameters to avoid bias in results. Therefore, both the proposed method and baseline methods in this study utilized uniform parameter settings, with detailed parameters listed in Table 1.

Table 1: Algorithm hyper-parameters

Hyper-parameter	Value
Iteration limit	$T = 15$
Disturbance limit	0.05
Moment decay rate	$\mu = 1$
Color configuration	Random number between (0, 1)
Learning rate	$\alpha = 0.01$
Iteration transformation probability	$p = 0.5$

4.2 Evaluation Indicators

Precision and recall are core evaluation metrics in object detection. Precision specifically focuses on the proportion of correctly identified targets among all detected results, while recall measures the algorithm's ability to identify all actual targets.

The formula for precision is typically defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

In this context, TP represents the number of True Positives, which corresponds to the images correctly identified as target objects. FP stands for False Positives, indicating the number of images where backgrounds or other objects were incorrectly identified as target objects.

The formula for recall is:

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

Here, FN represents the number of False Negatives, which denotes the actual target objects that the algorithm failed to detect.

In the evaluation of object detection algorithms, precision and recall are two key metrics. They respectively reflect the algorithm's capability of accurate identification and the comprehensiveness of target detection. Average Precision (AP) provides a comprehensive performance metric by analyzing the area under the Precision-Recall curve. This area can be estimated through methods such as numerical integration or trapezoidal approximation:

$$AP = \int_0^1 P(r)\, dr \tag{21}$$

When evaluating object detection models, the AP value for each class directly reflects the detection accuracy. Mean Average Precision (mAP) provides a summary evaluation of the model's performance across the entire dataset by averaging the AP values for all classes. The formula for mAP can be expressed as:

$$\text{mAP} = \frac{\sum_{i=1}^{m} AP_i}{m} \tag{22}$$

Here, m denotes the total number of classes, and AP_i represents the Average Precision for the (i)-th class. The mAP score ranges between 0 and 1, with higher scores indicating better performance of the model across multiple class detection tasks.

When evaluating attack performance, this paper employs a new metric called Attack Success Rate (ASR). ASR quantifies the success of an attack by comparing the average precision of the target detection model before and after the attack:

$$ASR = 1 - \frac{mAP_{adv}}{mAP_{clean}} \tag{23}$$

The computation of ASR involves two key mAP values: mAP_{adv} represents the average precision of the adversarial sample dataset, and mAP_{clean} represents the average precision of the clean sample dataset.

4.3 Experimental Effectiveness Assessment

This study conducted targeted adversarial attack experiments on the MS COCO dataset using a diverse set of object detection models such as YOLOv3, YOLOv8, RetinaNet, and Faster R-CNN. Under white-box conditions, we thoroughly assessed the impact of attacks on the models' mean Average Precision (mAP). To demonstrate the effectiveness of our approach, we compared it against classical adversarial attack methods like MI-FGSM, SIM, and TIM. Furthermore, we explored the transferability of adversarial samples in black-box attack scenarios and conducted ablation experiments to reveal the specific effects of different modules on attack performance.

Based on current research, generating targeted adversarial samples in object detection presents a complex challenge. It requires algorithms not only to recognize objects in images but also to craft image perturbations that can deliberately mislead models into misclassifying these targets as other categories. Traditional methods for generating targeted adversarial samples, such as I-FGSM and PGD, although effective in certain scenarios, often lack thorough consideration of data diversity and model robustness. These methods typically perform well in white-box environments where the attacker has access to model parameters. However, their performance may significantly degrade in black-box scenarios where such access is restricted. This study successfully enhances model performance under varied information conditions by integrating data augmentation techniques, optimizing logit loss, and introducing minimal local noise. In Fig. 7, we demonstrate targeted adversarial samples specifically designed for the "person" category. The generation process takes into account subtle changes in the image to ensure that adversarial perturbations are nearly imperceptible visually. By comparing original images with adversarial samples, we observe that despite introducing minor perturbations, these disturbances are sufficient to cause the object detection model to incorrectly classify all targets in the image as "person". This comparison not only highlights the effectiveness of adversarial samples but also underscores the stealth and practicality of the proposed methodology. The targeted adversarial sample creation strategy proposed in this study has proven effective in both white-box and black-box environments. These results provide important insights for future security evaluations of object detection models and offer an innovative perspective for enhancing model robustness.

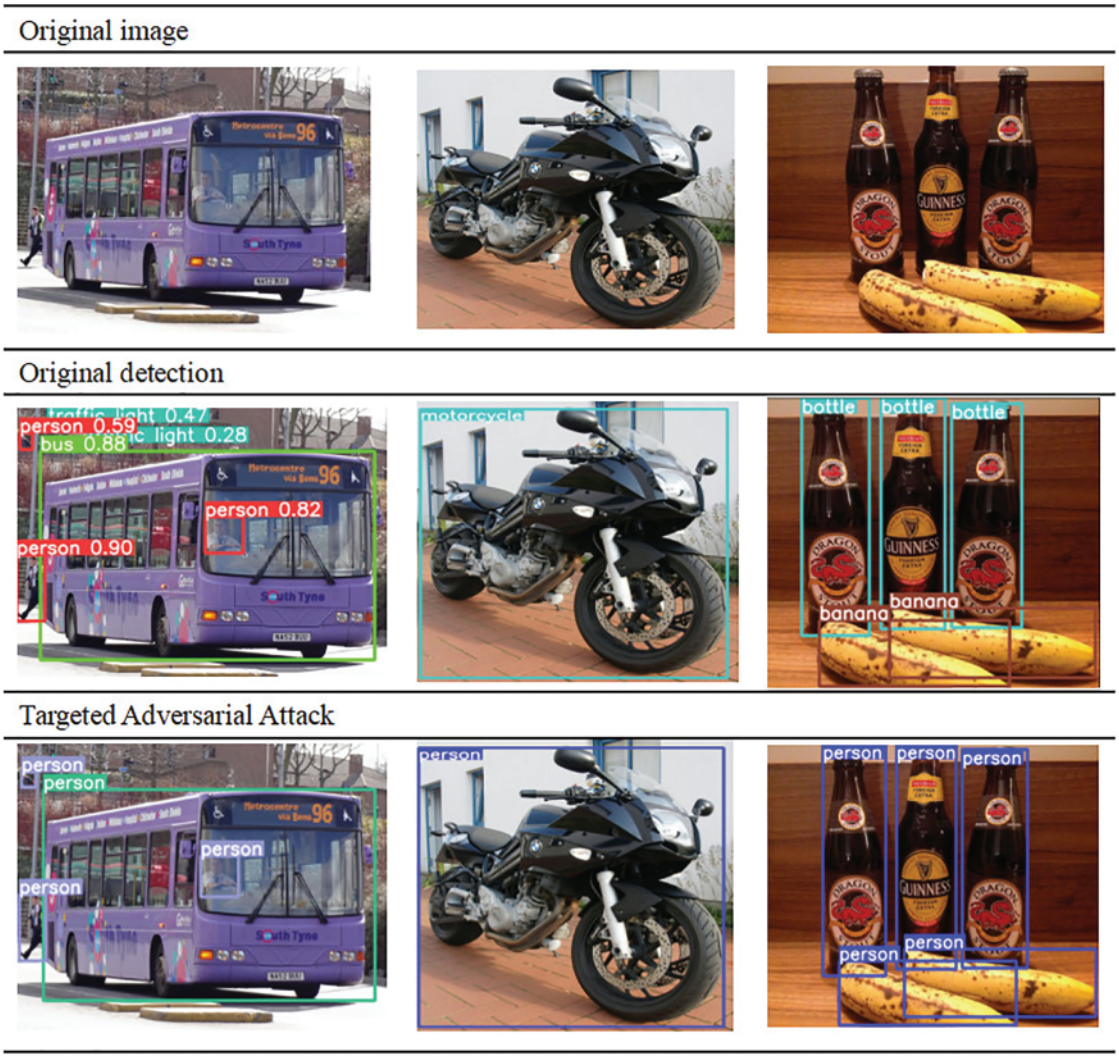

Figure 7: Targeted object detection adversarial examples

4.3.1 White-Box Experimental Assessment

In the white-box experimental setup, we conducted comprehensive adversarial attack tests on Faster R-CNN, YOLOv3, YOLOv8, and RetinaNet object detectors. We employed various attack techniques including MI-FGSM, TIM, SIM, and introduced our proposed RECT-MI-FGSM method. Experimental results, as shown in Tables 2–5, indicate that under identical conditions, our method significantly outperforms existing technologies in terms of attack success rates. Specifically, when using the Faster R-CNN model, TIM and SIM achieved average attack success rates of 79.7% and 84.9%, respectively, whereas our method achieved a leading success rate of 85.6%. Furthermore, by comparing the mAP (0.50:0.95) values before and after experiments, we clearly observe the impact of the attacks. The original mAP value of the Faster R-CNN model was 0.610, indicating high precision in object detection. However, under the RECT-MI-FGSM attack, this value dropped to 0.088, highlighting the effectiveness of the method in degrading the performance of the models.

Table 2: Contrasting adversarial approaches on faster R-CNN in white-box settings

Model	Method	Clean (mAP@0.5:0.95)	Attack (mAP@0.5:0.95)	ASR	Time (s)
Faster R-CNN	MI-FGSM	0.610	0.155	74.6%	1.4
	TIM	0.610	0.132	79.7%	2.6
	SIM	0.610	0.092	84.9%	2.8
	RCEN-MI-FGSM (Ours)	0.610	0.088	85.6%	3.2

Table 3: Contrasting adversarial approaches on YOLOv3 in white-box settings

Model	Method	Clean (mAP@0.5:0.95)	Attack (mAP@0.5:0.95)	ASR	Time (s)
YOLOv3	MI-FGSM	0.510	0.091	82.1%	1.6
	TIM	0.510	0.105	79.4%	2.3
	SIM	0.510	0.118	76.9%	2.5
	RCEN-MI-FGSM (Ours)	0.510	0.064	87.5%	3.6

Table 4: Contrasting adversarial approaches on YOLOv8 in white-box settings

Model	Method	Clean (mAP@0.5:0.95)	Attack (mAP@0.5:0.95)	ASR	Time (s)
YOLOv8	MI-FGSM	0.563	0.125	77.8%	2.1
	TIM	0.563	0.128	77.2%	2.5
	SIM	0.563	0.092	83.7%	2.9
	RCEN-MI-FGSM (Ours)	0.563	0.057	89.9%	3.8

Table 5: Contrasting adversarial approaches on RetinaNet in white-box settings

Model	Method	Clean (mAP@0.5:0.95)	Attack (mAP@0.5:0.95)	ASR	Time (s)
RetinaNet	MI-FGSM	0.556	0.114	79.5%	1.3
	TIM	0.556	0.122	78.1%	2.8
	SIM	0.556	0.104	81.3%	2.6
	RCEN-MI-FGSM (Ours)	0.556	0.083	85.1%	3.4

4.3.2 Black-Box Experimental Assessment

Transferability is a crucial property for assessing whether adversarial samples can maintain their deceptive nature when transferred from one model to another. In security research, this is typically considered a key measure of robustness against adversarial attacks. If adversarial samples generated

in a white-box environment remain effective in a black-box environment, it demonstrates strong transferability of those samples. To verify the transferability of our proposed method across different target detection models, we selected representative models such as Faster R-CNN, YOLOv3, YOLOv8, and RetinaNet to generate adversarial samples. We then tested their attack effectiveness on a series of unknown models. Our experimental results, detailed in Tables 6–9, reveal that in a black-box setting, our approach surpasses traditional baseline methods in attack success rates. Specifically, in transfer attacks from YOLOv3 to RetinaNet, the TIM and SIM methods led to mAP decreases to 0.184 and 0.204, respectively. Meanwhile, our method achieved a lower mAP decrease to 0.144 under the same conditions, indicating higher success rates in black-box attacks. These findings not only validate the effectiveness of our approach but also underscore its general applicability across object detection models. The transferability of adversarial samples is critical for achieving cross-model attacks, and our success in this area opens new avenues and possibilities for future research.

Table 6: Transferability assessment of adversarial methods in black-box scenarios

Source model	Method	Clean Retinanet (mAP@0.5:0.95)	Transfer attack RetinaNet (mAP@0.5:0.95)	ASR
YOLOv3	MI-FGSM	0.556	0.353	36.5%
	TIM	0.556	0.184	66.9%
	SIM	0.556	0.204	63.3%
	RCEN-MI-FGSM (Ours)	0.556	0.144	74.1%

Table 7: Transferability assessment of adversarial methods in black-box scenarios

Source model	Method	Clean Retinanet (mAP@0.5:0.95)	Transfer attack RetinaNet (mAP@0.5:0.95)	ASR
YOLOv8	MI-FGSM	0.556	0.392	29.50%
	TIM	0.556	0.174	68.71%
	SIM	0.556	0.152	72.66%
	RCEN-MI-FGSM (Ours)	0.556	0.108	80.58%

Table 8: Transferability assessment of adversarial methods in black-box scenarios

Source model	Method	Clean Retinanet (mAP@0.5:0.95)	Transfer attack RetinaNet (mAP@0.5:0.95)	ASR
Faster R-CNN	MI-FGSM	0.556	0.411	26.08%
	TIM	0.556	0.154	72.30%
	SIM	0.556	0.182	67.27%
	RCEN-MI-FGSM (Ours)	0.556	0.127	77.16%

Table 9: Transferability assessment of adversarial methods in black-box scenarios

Source model	Method	Clean YOLOv8 (mAP@0.5:0.95)	Transfer attack YOLOv8 (mAP@0.5:0.95)	ASR
RetinaNet	MI-FGSM	0.563	0.358	36.41%
	TIM	0.563	0.174	69.09%
	SIM	0.563	0.152	73.00%
	RCEN-MI-FGSM (Ours)	0.563	0.098	82.59%

4.3.3 Ablation Analysis of Key Model Components

In the context of object detection tasks, understanding the impact of different components on overall system performance is crucial. To this end, this study conducted a series of ablation experiments aimed at dissecting and quantifying the contributions of data augmentation frameworks and logit loss optimization to the adversarial attack resilience of object detection models. YOLOv8 was chosen as the baseline model, and various attack methods were comprehensively compared under consistent parameter settings. Initial experiments focused on iterative attacks using the MI-FGSM method against the YOLOv8 model. Upon introducing a data augmentation framework, we observed a significant increase in gradient diversity, leading to a notable decrease in the model's mean average precision (mAP) from 0.563 to 0.168. Furthermore, we introduced a logit loss optimization strategy aimed at adjusting the loss function to enhance the model's predictions for specific classes. Experimental results demonstrated that when combined with logit loss optimization, the mAP further decreased to 0.107. This outcome highlights the critical role of logit loss optimization in improving the accuracy and depth of attacks. Lastly, our approach incorporated a strategy involving local perturbations. By introducing subtle yet carefully designed disturbances into images, our method significantly reduced the detection performance of the model while maintaining visual imperceptibility. Specifically, after introducing local perturbations, the mAP further decreased to 0.057. This result not only validates the effectiveness of local perturbations in enhancing attack stealthiness but also demonstrates their potential in increasing attack success rates. These findings, reflected in Tables 10 and 11, underscore the importance of each component in improving attack success rates and transferability. The results of these ablation experiments provide a comprehensive understanding of the roles played by data augmentation frameworks and logit loss optimization in adversarial attacks on object detection models.

Table 10: YOLOv8 trade-off experiments of different strategies in white-box

RCEN framework	Mask	Logit loss	Clean (mAP@0.5:0.95)	Attack (mAP@0.5:0.95)
✓	✗	✗	0.563	0.168
✗	✓	✗	0.563	0.152
✗	✗	✓	0.563	0.107
✓	✓	✓	0.563	0.057

Table 11: YOLOv8 trade-off experiments of different strategies in black-box

RCEN framework	Mask	Logit loss	Clean RetinaNet (mAP@0.5:0.95)	Transfer attack RetinaNet (mAP@0.5:0.95)
✓	✗	✗	0.556	0.268
✗	✓	✗	0.556	0.253
✗	✗	✓	0.556	0.172
✓	✓	✓	0.556	0.108

Our algorithm has demonstrated outstanding robustness and accuracy across various complex scenarios on the COCO dataset. These scenarios include but are not limited to crowded urban streets, diverse natural environments, and challenging indoor layouts. Supplementary experiments on the Pascal VOC dataset, as shown in Tables 12 and 13, further confirm the algorithm's versatility and practicality. By conducting experiments on two datasets with different characteristics, our algorithm has proven its strong generalization ability. This generalization capability is crucial for the widespread applicability of the algorithm in practical scenarios, which was a key consideration in its design. The experimental results convincingly demonstrate the practicality of the algorithm. Whether in complex urban environments or simple indoor scenes, the algorithm consistently delivers stable and reliable detection results, providing robust technical support for real-world applications.

Table 12: Comparison of different attack methods against YOLOv8 on the pascal VOC dataset

Model	Method	Clean (mAP@0.5:0.95)	Attack (mAP@0.5:0.95)	ASR
YOLOv8	MI-FGSM	0.501	0.192	61.7%
	TIM	0.501	0.146	70.9%
	SIM	0.501	0.172	65.7%
	RCEN-MI-FGSM (Ours)	0.501	0.116	76.8%

Table 13: Comparison of different attack methods in transferability on the pascal VOC dataset

Source model	Method	Clean Retinanet (mAP@0.5:0.95)	Transfer attack RetinaNet (mAP@0.5:0.95)	ASR
YOLOv8	MI-FGSM	0.495	0.303	38.8%
	TIM	0.495	0.276	44.2%
	SIM	0.495	0.283	42.8%
	RCEN-MI-FGSM (Ours)	0.495	0.248	49.9%

5 Conclusion

In the realm of object detection research, although gradient-based targeted adversarial sample generation techniques like I-FGSM and PGD have been extensively studied, they exhibit limitations in transferability and still have room for improvement in their performance under white-box attack scenarios. This paper begins with a detailed comparative analysis between image classification and object detection tasks. Based on their similarities in neural network training and adversarial sample generation, we introduce the RCEN data augmentation framework. This framework integrates techniques such as data erasure and random colorization to enhance gradient diversity and reduce the risk of overfitting to specific white-box models. Secondly, to mitigate the noise consistency issue introduced by momentum accumulation during cross-entropy loss optimization, we designed an objective function based on logit loss optimization. This optimization strategy effectively guides the generation of adversarial samples, resulting in samples with enhanced transferability and stability. Lastly, to further minimize perturbation and enhance attack stealthiness, we introduced a mask matrix that precisely overlays perturbation noise onto the foreground regions of images. Experimental results demonstrate that our proposed method achieves higher attack success rates under different conditions compared to traditional approaches. This not only confirms the effectiveness of our method but also showcases significant advancements in improving the efficiency and robustness of adversarial sample generation. By leveraging the methods proposed in this study, it becomes possible to more effectively assess and enhance the robustness of existing object detection models.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Zhiyi Ding, Lei Sun; data collection: Ruiyang Ding, Xiuqing Mao, Leyu Dai, Zhiyi Ding; analysis and interpretation of results: Zhiyi Ding; draft manuscript preparation: Zhiyi Ding. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be made available on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. Fang, L. Shen, and Y. Chen, "Survey on image object detection algorithms based on deep learning," in *Artif. Intell. Secur.: 7th Int. Conf. (ICAIS 2021)*, Dublin, Ireland, Jul. 19–23, 2021, pp. 468–480.

[2] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. D. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019. doi: 10.1109/TNNLS.2018.2876865.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 7–12, 2015, pp. 3431–3440.

[4] K. Xu *et al.*, "Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML 2015)*, Lille, France, Jul. 6–11, 2015, pp. 2048–2057.

[5] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshops*, Santiago, Chile, Dec. 11–18, 2015, pp. 58–66.
[6] Z. Li, B. Xu, D. Wu, K. Zhao, M. Lu and J. Cong, "A mobile robotic arm grasping system with autonomous navigation and object detection," in *Proc. 2021 Int. Conf. Control, Autom. Inform. Sci. (ICCAIS)*, Xi'an, China, Oct. 14–17, 2021, pp. 543–548.
[7] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, no. 5, 2019, Art. no. 909. doi: 10.3390/app9050909.
[8] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
[9] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018. doi: 10.1109/ACCESS.2018.2807385.
[10] D. Zhang, Y. Fu, and Z. Zheng, "UAST: Uncertainty-aware siamese tracking," in *Proc. 39th Int. Conf. Mach. Learn.*, Baltimore, MD, USA, Jul. 7–23, 2022, pp. 26161–26175.
[11] J. Sheng, D. Zhang, J. Chen, X. Xiao, and Z. Zheng, "Towards universal and sparse adversarial examples for visual object tracking," *Appl. Soft Comput.*, vol. 153, 2024, Art. no. 111252. doi: 10.1016/j.asoc.2024.111252.
[12] A. Amirkhani, M. P. Karimi, and A. Banitalebi-Dehkordi, "A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles," *Vis. Comput.*, vol. 39, no. 11, pp. 5293–5307, 2023. doi: 10.1007/s00371-022-02660-6.
[13] J. Gu *et al.*, "A survey on transferability of adversarial examples across deep neural networks," *Trans. Mach. Learn. Res.*, pp. 1–35, 2024. Accessed: Mar. 15, 2024. doi: 10.48550/arXiv.2310.17626.
[14] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.
[15] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–38, 2020.
[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Rep.*, May 7–9, 2015, pp. 1–11.
[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICML, 2017 Workshop Princ. Approach. Deep Learn.*, Sydney, Australia, Aug. 10, 2017, pp. 1–10.
[18] S. -M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, Jun. 26–Jul. 1, 2016, pp. 2574–2582.
[19] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. 2017 IEEE Symp. Secur. Priv. (SP)*, San Jose, CA, USA, May 22–26, 2017, pp. 39–57.
[20] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 22–29, 2017, pp. 1369–1378.
[21] Y. Li, D. Tian, M. Chang, X. Bian, and S. Lyu, "Robust adversarial perturbation on deep proposal-based models," 2018, *arXiv:1809.05962*.
[22] K. H. Chow *et al.*, "Adversarial objectness gradient attacks in real-time object detection systems," in *Proc. 2020 Second IEEE Int. Conf. Trust, Priv. Secur. Intell. Syst. Appl. (TPS-ISA)*, Atlanta, GA, USA, Oct. 28–31, 2020, pp. 263–272.
[23] X. Wei, S. Liang, N. Chen, and X. Cao, "Transferable adversarial attacks for image and video object detection," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 10–16, 2018, pp. 954–960.
[24] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 4312–4321.
[25] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 9185–9193.

[26] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. 8th Int. Conf. Learn. Rep. (ICLR 2020)*, Apr. 26–May 01, 2019, pp. 1–11.
[27] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 19–25, 2021, pp. 1924–1933.
[28] M. Li, C. Deng, T. Li, J. Yan, X. Gao and H. Huang, "Towards transferable targeted attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 24, 2020, pp. 641–649.
[29] R. Gavrilescu, C. Zet, C. Foşalău, M. Skoczylas, and D. Cotovanu, "Faster R-CNN: An approach to real-time object detection," in *Proc. 2018 Int. Conf. Expo. Electric. Power Eng. (EPE)*, Iasi, Romania, Oct. 18–19, 2018, pp. 165–168.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016. doi: 10.1109/TPAMI.2016.2577031.
[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 22–29, 2017, pp. 2961–2969.
[32] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. Accessed: Mar. 15, 2024. [Online]. Available: http://personeltest.ru/aways/pjreddie.com/media/files/papers/YOLOv3.pdf
[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. COMPUTER Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 26–Jul. 1, 2016, pp. 779–788.
[34] C. Ning, H. Zhou, Y. Song, and J. Tang, "Inception single shot multibox detector for object detection," in *Proc. 2017 IEEE Int. Conf. Multimed. Expo Workshops (ICMEW)*, Hong Kong, China, Jul. 10–14, 2017, pp. 549–554.
[35] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 22–29, 2017, pp. 2980–2988.
[36] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 11–17, 2021, pp. 16158–16167.
[37] H. Zhang, W. Zhou, and H. Li, "Contextual adversarial attacks for object detection," in *Proc. 2020 IEEE Int. Conf. Multimed. Expo (ICME)*, London, UK, Jul. 6–10, 2020, pp. 1–6.