NumbOD: A Spatial-Frequency Fusion Attack Against Object Detectors

Ziqi Zhou[1,2,3,6], Bowen Li[7], Yufei Song[7], Zhifei Yu[7], Shengshan Hu[1,2,4,5,7], Wei Wan[1,2,4,5,7], Leo Yu Zhang[8], Dezhong Yao[1,2,3,6], Hai Jin[1,2,3,6]

[1]National Engineering Research Center for Big Data Technology and System
[2]Services Computing Technology and System Lab
[3]Cluster and Grid Computing Lab
[4]Hubei Engineering Research Center on Big Data Security
[5]Hubei Key Laboratory of Distributed System Security
[6] School of Computer Science and Technology, Huazhong University of Science and Technology
[7] School of Cyber Science and Engineering, Huazhong University of Science and Technology
[8] School of Information and Communication Technology, Griffith University
{zhouziqi,libowen2021,yufei17,yzf,hushengshan,wanwei_0303,dyao,hjin}@hust.edu.cn leo.zhang@griffith.edu.au

Abstract

With the advancement of deep learning, *object detectors* (ODs) with various architectures have achieved significant success in complex scenarios like autonomous driving. Previous adversarial attacks against ODs have been focused on designing customized attacks targeting their specific structures (*e.g.*, NMS and RPN), yielding some results but simultaneously constraining their scalability. Moreover, most efforts against ODs stem from image-level attacks originally designed for classification tasks, resulting in redundant computations and disturbances in object-irrelevant areas (*e.g.*, background). Consequently, how to design a model-agnostic efficient attack to comprehensively evaluate the vulnerabilities of ODs remains challenging and unresolved. In this paper, we propose NumbOD, a brand-new spatial-frequency fusion attack against various ODs, aimed at disrupting object detection within images. We directly leverage the features output by the OD without relying on its internal structures to craft adversarial examples. Specifically, we first design a dual-track attack target selection strategy to select high-quality bounding boxes from OD outputs for targeting. Subsequently, we employ directional perturbations to shift and compress predicted boxes and change classification results to deceive ODs. Additionally, we focus on manipulating the high-frequency components of images to confuse ODs' attention on critical objects, thereby enhancing the attack efficiency. Our extensive experiments on nine ODs and two datasets show that NumbOD achieves powerful attack performance and high stealthiness.

Code — https://github.com/CGCL-codes/NumbOD

Introduction

The triumphs in deep learning have substantially propelled the development of computer vision tasks, such as traffic sign recognition (Tabernik and Skočaj 2019), pedestrian re-identification (Zheng et al. 2017), and medical image segmentation (Ramesh et al. 2021). Despite its promising prospects, existing researches (Goodfellow, Shlens, and Szegedy 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016) have demonstrated the vulnerability of *deep neural networks* (DNNs). Adversaries can induce model misclassifications with minimal, strategically crafted perturbations, like wrongly identifying an image of a dog as a cat. Although extensive works (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Carlini and Wagner 2017; Madry et al. 2018; Zhou et al. 2023b; Li et al. 2024b; Wang et al. 2025) have thoroughly investigated adversarial attacks on classification, the more challenging task of object detection remains far less explored.

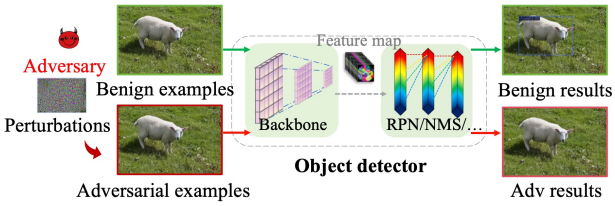


Figure 1: An overview of adversarial examples against an object detector

Unlike single classification tasks, object detection involves classification and regression subtasks, requiring simultaneous localization and recognition of objects, *i.e.*, providing both bounding boxes and classification results. Recently, researchers have proposed various modules aimed at enhancing the performance of object detectors, such as *Non-Maximum Suppression* (NMS), *Region of Interest* (RoI) Pooling, and *Region Proposal Networks* (RPN). The introduction of these new features and modules has brought unprecedented challenges to standard adversarial attacks originally designed for classification tasks.

Recent efforts (Xie et al. 2017; Li et al. 2018; Chen, Kung, and Chen 2021) have made decent progress in adversarial attacks against ODs, yet these methods commonly face two major limitations: *1) limited applicability* and *2) low efficiency* in attacks. Existing efforts have developed effective attacks by exploiting specific vulnerabilities within ODs. For example, DAG (Xie et al. 2017) is the first attack targeting RPN-based models, which implements the attack by minimizing the probability of correct classification. RAP (Li

et al. 2018) enhances the attack by incorporating a loss function tailored for both RPN predicted boxes and classification. However, due to their reliance on specific modules of ODs, these attack methods greatly limit their scalability, being effective only on detectors with such specific architectural features. Moreover, existing methods involve image-level global perturbations, which incur unnecessary computational costs by optimizing attacks on non-critical objects, such as the background. This may lead to suboptimal attack performance as they simultaneously strive for effective disruption of both meaningful objects and irrelevant background elements. To the best of our knowledge, how to realize a model-agnostic adversarial attack on critical objects for ODs still remains challenging.

A recent study (Li et al. 2024a) explored designing adversarial examples without l_p-norm constraints to deceive object detectors with varying architectures. However, to maintain stealthiness, its attack effectiveness is limited, resulting in detection outcomes that still contain some correct bounding boxes. In contrast, our approach focuses on crafting l_p-norm constrained adversarial attacks that aim to render ODs numb to input images and unable to detect any object, as demonstrated in Fig. 1. In this paper, we propose NumbOD, a novel model-agnostic spatial-frequency fusion attack for ODs. To achieve a model-agnostic attack, we leverage the final output features of ODs to craft adversarial examples. Our approach employs a dual-track attack target selection strategy, where we independently select the top-k high-quality bounding boxes from both the classification and regression subtasks, thereby enhancing attack efficiency. Upon identifying the attack targets, we design a tailored attack against ODs from both spatial and frequency domains.

Given that object detection involves both classification and regression subtasks, a truly effective attack must simultaneously deceive both components. Specifically, the attack should cause both the predicted bounding boxes and the classification results to deviate from their original outputs. In the spatial domain, we induce customized deviations in predicted bounding boxes and misclassification results by adding noise to the image. Drawing inspiration from the sensitivity of deep neural networks to frequency components (Luo et al. 2022; Wang et al. 2022), particularly high-frequency components that capture semantic texture information, we enhance attack efficiency by targeting the image's high-frequency regions rather than the entire image. We start by applying the *Discrete Wavelet Transform* (DWT) to decompose the image into high and low-frequency components. We then focus the noise on the semantically significant high-frequency regions, minimizing the discrepancy between adversarial and benign examples and thereby avoiding ineffective attacks on non-critical areas.

We conduct experiments on nine object detectors and two datasets to evaluate the effectiveness of NumbOD. Both qualitative and quantitative results show that NumbOD effectively deceives ODs of various architectures, exhibiting strong attack performance and high stealthiness. Additionally, comparative experiments reveal that NumbOD surpasses *state-of-the-art* (SOTA) methods for attacking ODs. Our main contributions are summarized as follows:

- We propose NumbOD, a novel model-agnostic adversarial attack against ODs, designed to disrupt object detection within images across various detector architectures.
- We design a spatial-frequency fusion attack framework against ODs, which consists of a spatial coordinated deviation attack and a critical frequency interference attack.
- Our extensive experiments on nine ODs and two datasets show that our NumbOD achieves powerful attack performance and high stealthiness, surpassing SOTA attacks.

Related Works

Object Detectors

Existing object detection methods are primarily categorized into two distinct paradigms: two-stage detectors (Ren et al. 2015; Cai and Vasconcelos 2019; Ding et al. 2019; Wang et al. 2020; Xu et al. 2020; Sun et al. 2021; Xie et al. 2021; Han et al. 2021b) and single-stage detectors (Redmon et al. 2016; Lin et al. 2017; Yang et al. 2019; Tian et al. 2019; Feng et al. 2021; Zhang et al. 2021; Han et al. 2021a). Two-stage detectors, such as R-CNN (Girshick et al. 2014), Faster R-CNN (Ren et al. 2015), and Cascade R-CNN (Cai and Vasconcelos 2019), first generate candidate regions through a RPN and then perform precise classification and regression on these regions. Conversely, single-stage detectors, such as YOLOs (Redmon et al. 2016; YOLO-V5 2022), VFNet (Zhang et al. 2021), and TOOD (Feng et al. 2021), directly predict object classes and bounding box coordinates across the entire image in a single evaluation step. For clarity, we define the objects detected in the images by the object detector as foreground and the remaining parts of the images as background. Different detectors achieve desirable results in object detection tasks based on their unique modules, which also endow them with distinct vulnerabilities.

Adversarial Examples on Object Detectors

Adversarial example (Goodfellow, Shlens, and Szegedy 2015; Zhou et al. 2023a, 2024b,a; Song et al. 2025; Zhang et al. 2025) is introduced to demonstrate the fragility of DNNs, which involves the addition of minimal perturbations to images, causing misclassification by the target model. Existing adversarial examples can be categorized into noise-based (Goodfellow, Shlens, and Szegedy 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Madry et al. 2018) and patch-based methods (Wei, Yu, and Huang 2023; Huang et al. 2023; Tang et al. 2023). The former boasts high concealment, while the latter offers flexibility but is prone to detection due to its visibility. Therefore, this paper exclusively considers noise-based adversarial methods.

Recently, researchers begin to study the vulnerabilities of object detection, a task that presents greater challenges than classification due to its inclusion of both regression and classification subtasks. While recent methods (Xie et al. 2017; Li et al. 2018; Chen, Kung, and Chen 2021) have demonstrated certain attack performance tailored to specific models, they also inherently limit their scalability, rendering them unsuitable for attacks across different architectural models. To address this issue, some efforts (Wei et al. 2018; Chow et al. 2020; Aich et al. 2022) explored designing preliminary

model-agnostic adversarial attacks against ODs. For example, TOG (Chow et al. 2020) employs two different attack strategies to customize attacks for RPN-based and anchor based ODs. However, it cannot attack more recent detector (*e.g.*, Sparse R-CNN (Sun et al. 2021)) without such fundamental components. Therefore, there is an urgent need for truly model-agnostic attack against ODs.

Methodology

Problem Formulation

Object detection is a fundamental task in computer vision encompassing two subtasks: classification and regression. Its output involves providing predicted bounding boxes for target objects, along with corresponding classification labels and scores. Given an image $x \in \mathcal{D}$ to an object detector $f(x) \in \mathbb{R}^{N\times(4+1)}$ that returns bounding boxes $\mathcal{B}_n$ containing the coordinates of the top-left and bottom-right corners and the predicted label $Y_n, n = 1, 2, ..., N$, with classification score $c_n \in [0, 1]$.

Threat model. We assume that the adversary has access to both the white-box model and the dataset, aiming to design adversarial examples that render the OD ineffective. Specifically, the adversary aims to craft an elaborate adversarial noise δ to paste onto the input image x to get an adversarial example x^{adv}, which is then fed into the detector to change its original output, *e.g.*, the bounding box is shifted or disappears, the predicted category of the target object changes or the original classification score decreases. Note that the noise δ needs to be small enough to be indistinguishable to the naked eye so that the adversarial examples are not easily detected. This constraint is typically enforced through an upper bound ϵ on the l_p-norm formulated as follow:

$$\max_{\delta} \mathbb{E}_{x\sim\mathcal{D}} \left[f\left(x+\delta\right) \neq f\left(x\right) \right], \quad s.t. \left\|\delta\right\|_p \leq \epsilon \qquad (1)$$

After feeding the adversarial example $x+\delta$ into the object detector $f(\cdot)$, we can obtain the adversarial prediction boxes $\mathcal{B}_n^{adv}$, label Y_n^{adv} with classification score c_n^{adv}.

Key Challenges and Intuitions

Due to the significant structural differences among existing object detectors and their focus on specific object regions within images rather than the entire image, designing a model-agnostic adversarial attack for object detectors presents the following challenges:

Challenge I: The attack dependency on specific modules of object detectors. Benefiting from the designs tailored to the specific modules of object detectors, previous methods have achieved promising attack performance. For instance, the customized attack design of RAP (Li et al. 2018) focusing on the RPN has proven to be highly effective in deceiving RPN-based detectors. However, this also limits its attack generalization. Specifically, RAP demonstrates ineffectiveness in targeting single-stage detectors due to the absence of an RPN structure. Similarly, other attack methods like FGSM (Goodfellow, Shlens, and Szegedy 2015), DAG (Xie et al. 2017), and PGD (Madry et al. 2018), which are designed for classification tasks, cannot be directly applied to single-stage models. Hence, a simple idea is to utilize the final output features of the object detector, which are independent of specific modules, for crafting adversarial examples. However, the extensive bounding boxes generated by the object detector also introduces ambiguity concerning the attack target, thereby incurring unnecessary computational overhead. Given that object detection algorithms typically employ joint optimization for classification and regression tasks, we propose a *dual-track attack target selection strategy*. This strategy enhances attack efficiency by separately selecting the top-k predicted boxes with high scores for both classification and regression tasks as attack targets. As shown in Fig. 2, for regression, we select the predicted boxes with the highest top-k IoU scores for each object in the image as attack targets. For classification, we similarly choose the predicted boxes with the highest top-k IoU scores, but only when the predicted labels match the ground truth labels for each object. By simultaneously considering the above bounding boxes as attack targets, we aim to enhance attack efficiency while avoiding the emergence of suboptimal attacks.

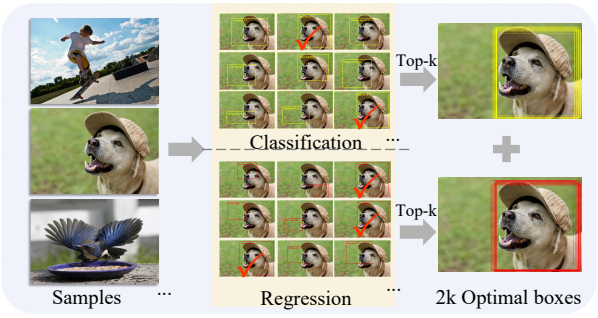


Figure 2: Dual-track attack target selection strategy

Challenge II: The attack redundancy on non-critical objects. Most existing adversarial attacks on object detectors focus on optimizing global noise at the image level. However, perturbing regions outside the target objects (*e.g.*, the background) often fails to enhance attack effectiveness and can lead to inefficiencies. It is well known that *low-frequency components* (LFC) of an image, which have smooth pixel changes, carry the main information of the image. In contrast, *high-frequency components* (HFC), characterized by abrupt pixel changes, mainly convey details and noise. Given that deep neural networks are biased towards image textures, we propose selectively disrupting the HFC of images to hinder the model's recognition of critical objects, thereby increasing the attack's efficiency. Specifically, we aim to amplify the differences in high-frequency components (*i.e.*, texture information) between adversarial examples and benign samples while constraining the differences in low-frequency components (*i.e.*, shape information). This approach further enhances the attack's effectiveness and stealthiness. By designing such fusion attack in both spatial and frequency domains, we strategically target crucial areas within images while simultaneously deceiving regression and classification subtasks. This provides an ef-

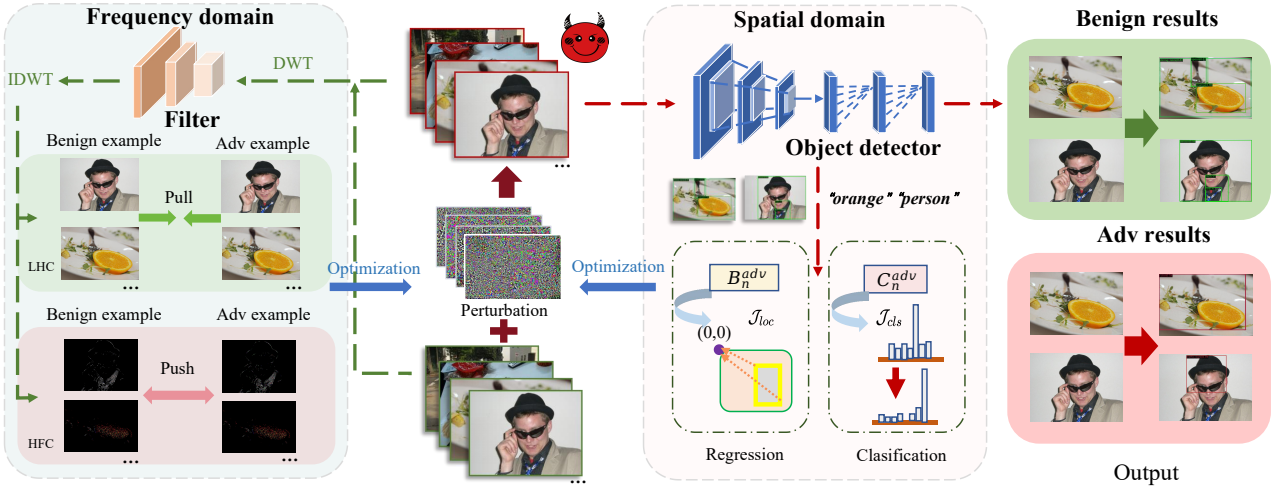


Figure 3: The pipeline of our method

ficient optimization direction for generating adversarial examples, resulting in successful attacks on object detectors.

Spatial-Frequency Fusion Attack

In this section, we present NumbOD, a brand-new spatial-frequency fusion attack against object detectors, making them impossible to properly detect objects in images. The pipeline of NumbOD is depicted in Fig. 3, which consists of a spatial coordinated deviation attack and a critical frequency interference attack. We initially allocate high-quality bounding boxes for each object in the image as attack targets from the perspectives of both regression and classification subtasks, based on the dual-track attack target selection strategy. Subsequently, in the spatial domain (*i.e.*, traditional attacks involving pixel-level modifications to images), we introduce noise to the images to disrupt the detector's assessment of objects by simultaneously customizing deviations in the positions of predicted boxes and misleading classification outcomes. Simultaneously, in the frequency domain, we enhance the attack performance by further undermining key details, textures, and edges in the images, thereby boosting attack efficiency. The overall optimization objective of NumbOD is as follow:

$$\mathcal{J}_{total} = \mathcal{J}_{sa} + \mathcal{J}_{fa} \tag{2}$$

where $\mathcal{J}_{sa}$ is the spatial attack loss and $\mathcal{J}_{fa}$ is the frequency attack loss.

Spatial coordinated deviation attack. Given that the output of the OD for image x primarily includes bounding box locations and classification information. Our method targets these two critical components for attack. Specifically, we induce a coordinate shift attack ($\mathcal{J}_{loc}$) to alter the size and position of the predicted boxes output by the object detector, and mislead the classification results through a foreground-background separation attack ($\mathcal{J}_{cls}$). The loss of the spatial coordinated deviation attack is formulated as follow:

$$\mathcal{J}_{sa} = \mathcal{J}_{loc} + \lambda \mathcal{J}_{cls} \tag{3}$$

For the regression subtask, we design a targeted approach to align the coordinates of the predicted boxes with those of predefined meaningless target regions. Considering that objects in the image tend to be located in the central area, we force the coordinates of the predicted bounding box's top-left and bottom-right corners to approach the edge point $(0, 0)$, causing both positional and size changes to render it ineffective. $\mathcal{J}_{loc}$ can be expressed as:

$$\mathcal{J}_{loc} = \sum_{n=1}^{N} \mathcal{J}_d(\mathcal{B}_n^{adv}, \mathcal{B}_n^t)/N \tag{4}$$

where $\mathcal{B}_n^t$ represents the target bounding box designed by the attacker, and $\mathcal{J}_d$ is the Smooth L1 loss.

For the classification subtask, we implement a foreground-background separation attack by minimizing the scores associated with the true labels of objects in the image while maximizing the score assigned to the background class. This approach induces the object features within the image to converge towards those of the background, thereby impeding accurate detection. For the K-class probabilities $c_n = (c_n^0, c_n^1, c_n^2, ...c_n^K)$, we designate c_n^{gt} as the scores attributed to the respective ground truth labels within the n-th bounding box. We enhance the background scores c_n^K while reducing the scores c_n^{gt} of the corresponding ground truth labels. The described optimization process can be represented as:

$$\mathcal{J}_{cls} = \sum_{n=1}^{N} \log\left(c_n^{gt}\right)/N - \sum_{n=1}^{N} \log\left(c_n^{K}\right)/N \tag{5}$$

Critical frequency interference attack. In the frequency domain, the high-frequency components of an image denote the finer details, including noise and textures, while the low-frequency components contain the general outline and overall structural information of the image. We aim to disrupt the sensitive high-frequency components of DNNs to interfere with the OD's focus on crucial objects in input images. We

Datasets		MS-COCO									PASCAL VOC								
Models		FR	CR	SR	SFR	RP	VFNet	TOOD	D.DETR	YOLO	FR	CR	SR	SFR	RP	VFNet	TOOD	D.DETR	YOLO
IW-SSIM↓		0.17	0.18	0.12	0.17	0.15	0.16	0.13	0.18	0.17	0.20	0.20	0.14	0.20	0.17	0.17	0.15	0.17	0.17
NMSE↓		0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01
TV↓		96.14	96.17	96.08	96.15	96.10	96.15	96.06	96.28	96.23	81.12	81.13	81.04	81.14	81.04	81.19	81.02	81.28	81.18
mAP50(%)	clean↑	50.98	51.28	47.58	50.70	48.99	51.31	51.80	60.79	53.32	74.45	75.09	70.60	73.83	73.68	73.95	57.41	78.51	69.42
	adv↓	0.38	0.27	3.62	0.47	2.25	5.49	2.69	1.69	0.59	0.54	0.22	3.21	0.71	1.35	1.96	2.54	3.22	2.14
mAP75(%)	clean↑	34.74	37.55	32.54	36.93	32.89	37.68	38.86	43.73	36.84	57.08	60.04	52.56	58.34	56.19	59.74	41.86	62.19	51.47
	adv↓	0.06	0.08	1.17	0.10	0.74	1.90	1.32	1.32	0.17	0.04	0.02	1.15	0.08	0.21	0.39	0.99	2.03	0.86

Table 1: Attack performance of NumbOD against different object detectors

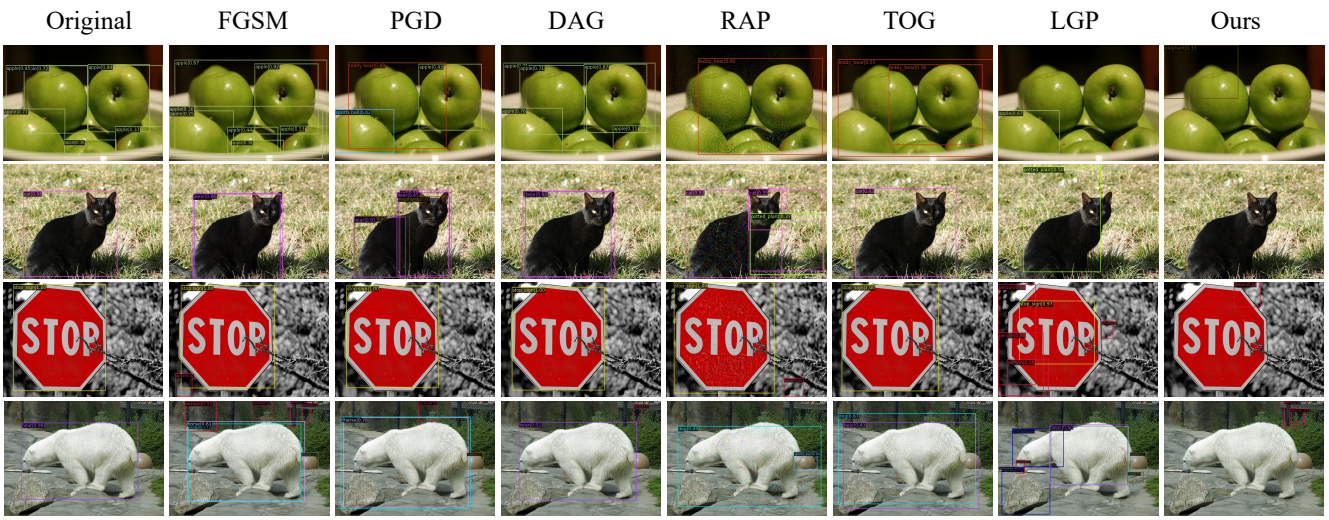


Figure 4: Visualizations of the adversarial examples made by different methods against Faster R-CNN on MS-COCO

employ the DWT, utilizing a low-pass filter $\mathcal{L}$ and a high-pass filter $\mathcal{H}$, to decompose the image x into different components, constituting a low-frequency component c_{ll}, a high-frequency component c_{hh}, and two mid-frequency components c_{lh} and c_{hl}, via

$$c_{ll} = \mathcal{L}x\mathcal{L}^T, c_{hh} = \mathcal{H}x\mathcal{H}^T, c_{lh}/c_{hl} = \mathcal{L}x\mathcal{H}^T/\mathcal{H}x\mathcal{L}^T \quad (6)$$

Subsequently, we employ the *inverse discrete wavelet transform* (IDWT) to reconstruct the signal that has been decomposed through DWT into an image. We choose the LFC and HFC while dropping the other components to obtain the reconstructed images $\phi(x)$ and $\psi(x)$ as

$$\phi(x) = \mathcal{L}^T x_{ll} \mathcal{L} = \mathcal{L}^T(\mathcal{L}x\mathcal{L}^T)\mathcal{L} \quad (7)$$

$$\psi(x) = \mathcal{H}^T x_{hh} \mathcal{H} = \mathcal{H}^T(\mathcal{H}x\mathcal{H}^T)\mathcal{H} \quad (8)$$

By adding the adversarial noises to the images, we alter their high-frequency components, disrupting the original texture information. Simultaneously, we enforce constraints on the low-frequency disparities between adversarial and benign examples to redirect a larger portion of the perturbation towards the high-frequency domain, thereby enhancing the attack performance and stealthiness of adversarial examples. The loss of the critical frequency interference attack can be expressed as:

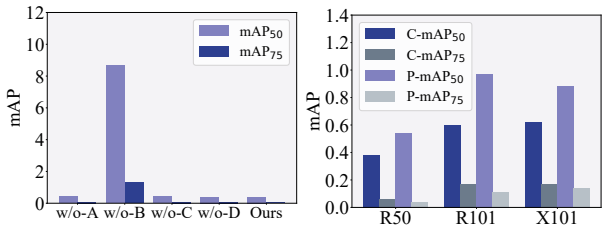
$$\begin{aligned} \mathcal{J}_{fa} &= \mathcal{J}_{lfc} - \mathcal{J}_{hfc} \\ &= \mathcal{J}_d(\phi(x), \phi(x+\delta)) - \mathcal{J}_d(\psi(x), \psi(x+\delta)) \end{aligned} \quad (9)$$

Experiments

Experimental Setup

Datasets and models. For a comprehensive evaluation, we use MS-COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010) datasets with victim models with ResNet50, ResNet101, and ResNeXt101 as backbones. Unless otherwise specified, we use ResNet50 as the default backbone for evaluation. Specifically, we select the following nine models: (1) Two-stage detectors: *Faster R-CNN* (FR) (Ren et al. 2015), *Cascade R-CNN* (CR) (Cai and Vasconcelos 2019), *SABL Faster R-CNN* (SFR) (Wang et al. 2020), and *Sparse R-CNN* (SR) (Sun et al. 2021). (2) Single-stage detectors: *RepPoints* (RP) (Yang et al. 2019), *Deformable DETR* (D.DETR) (Zhu et al. 2020), VFNet (Zhang et al. 2021), TOOD (Feng et al. 2021), and *YOLO v5* (YOLO) (YOLO-V5 2022).

Evaluation metrics. In terms of attack effectiveness, we evaluate the attack performance of our method using the widely adopted metric in the object detection domain, *Mean Average Precision* (mAP). We choose mAP_{50} and mAP_{75} as indicators, which represent the average precision at *Intersection over Union* (IoU) thresholds of 0.5 and 0.75, respectively. In terms of attack stealthiness, we select commonly used metrics such as *Inception Weighted Structural Similarity Index Metric* (IW-SSIM), *Normalized Mean Squared Error* (NMSE), and *Total Variation* (TV) to assess the distance between benign images and perturbed images. For clarity,



(a) Module (b) Backbone

Figure 5: Ablation Study. C-mAP$_{50}$ and P-mAP$_{50}$ denote the mAP$_{50}$ results on MS-COCO and PASCAL VOC, Others stand the same meaning.

we default to multiplying the values of mAP, IW-SSIM, and NMSE by 100.

Implementation details. Following (Goodfellow, Shlens, and Szegedy 2015; Xie et al. 2017; Madry et al. 2018; Chow et al. 2020), we set the upper bound of the adversarial perturbation to $8/255$. We set the hyperparameters λ to 100, while the training epoch is set to 50 with a batch size of 1. We utilize the Adamax optimizer and set the learning rate and weight decay to 0.03 and 0.02, respectively.

Attack Performance

To comprehensively evaluate NumbOD's effectiveness, we conduct experiments on nine object detectors, with ResNet50 as the backbone, across two datasets, MS-COCO and PASCAL VOC. For a single attack, we randomly select 5000 images from the dataset to craft adversarial examples, which are then fed into the object detector to evaluate the effectiveness and stealthiness of our approach.

We first provide a quantitative evaluation of NumbOD in Tab. 1. The results reveal the substantial impact of adversarial attacks on the performance of various object detection models across different datasets. Our proposed NumbOD causes significant reductions in mAP scores at both 50% and 75% IoU thresholds, indicating a marked decrease in detection accuracy. Notably, models like Cascade R-CNN and RepPoints show increased vulnerability, with mAP values significantly dropping across the datasets. We also present the qualitative evaluation results in Fig. 4 to further validate the effectiveness of our approach. The results in the last column of Fig. 4 indicate that the object detector fails to detect objects in the adversarial examples generated by NumbOD, with prediction box positions completely wrong or missing, and misclassification results.

Notably, as indicated by the metrics in Tab. 1 and the images shown in Fig. 4, our approach exhibits remarkably high stealthiness. The generated adversarial examples are visually indistinguishable, excelling in both visual appearance and stealthiness metrics.

Comparison Study

To showcase the superiority of our method, we conduct comparative experiments against SOTA adversarial example methods from both effectiveness and stealthiness perspectives. Specifically, we compare our proposed NumbOD with six popular attack methods, FGSM (Goodfellow, Shlens, and Szegedy 2015), DAG (Xie et al. 2017), PGD (Madry et al. 2018), RAP (Li et al. 2018), TOG (Chow et al. 2020), and LGP (Li et al. 2024a), on two models across two datasets. Among them, LGP is the latest SOTA attack tailored for object detectors. The perturbation constraints of LGP and RAP do not belong to the l_p norm. We use the parameters as stated in their original papers. The perturbation budget for the other attacks is set to $8/255$.

We first present the quantitative comparison of our method with these methods in Tab. 2. The results indicate that our method outperforms all existing approaches in terms of effectiveness and stealthiness. We further provide qualitative experiments comparing our method with these methods in Fig. 4. We consider deceiving both sub-tasks of ODs simultaneously as truly fooling the object detection model. The results in Fig. 4 indicate that existing methods can only deceive either the regression or the classification task individually, *i.e.*, causing the predicted box to deviate or misclassifying. For instance, in the case of the LGP attack, it performs excellently in terms of attack metrics (*e.g.*, achieving an mAP$_{50}$ of 1.49 on Faster R-CNN across the MS-COCO dataset), but still fails to effectively deceive the regression sub-task, meaning the predicted boxes still remain on the main objects in the image. The results from Fig. 4 demonstrate that our approach outperforms others significantly, achieving true deception of the object detector, including prediction box deviations or disappearances and misclassification results.

Ablation Study

In this section, we explore the effect of different modules and backbones on our method. We conduct experiments on the Faster R-CNN model with ResNet50 as the backbone across the MS-COCO dataset.

The effect of different modules. We investigate the effect of different modules on NumbOD. We use A, B, C, and D to represent $\mathcal{J}_{loc}$, $\mathcal{J}_{cls}$, $\mathcal{J}_{lfc}$, and $\mathcal{J}_{hfc}$, respectively. Experimental results in Fig. 5 (a) demonstrate that none of the variants of our proposed method can match the performance of the complete version.

The effect of backbone. We examine the effect of different backbones on NumbOD, using three Faster R-CNN variants: *ResNet50* (R50), *ResNet101* (R101), and *ResNeXt101* (X101). These models are tested on MS-COCO and PASCAL VOC datasets to evaluate NumbOD's attack performance. The results in Fig. 5 (b) demonstrate our method's outstanding attack performance across various backbones.

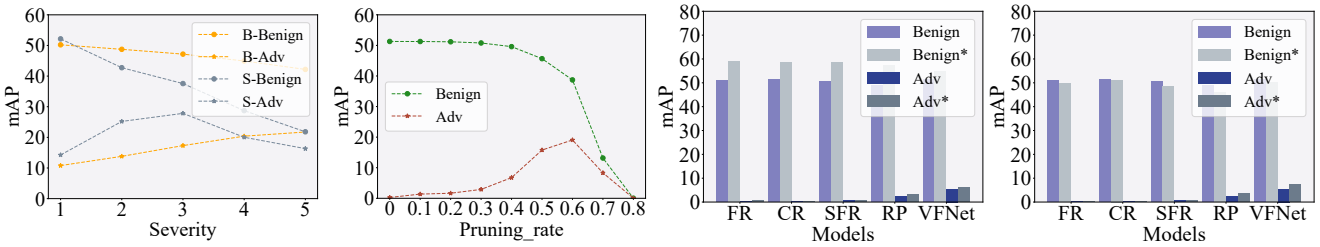
Defense

Corruption

Corruption is a representative image preprocessing method used to mitigate adversarial examples. We employ two popular strategies, *Brightness* ("B-") and *Spatter* ("S-"), to corrupt adversarial examples. As illustrated in Fig. 6 (a), the mAP$_{50}$ of the Faster R-CNN model decreases as the degree of corruption increases. However, our attack remains

Metric	Model	MS-COCO							PASCAL VOC						
		FGSM	PGD	DAG	RAP	TOG	LGP	Ours	FGSM	PGD	DAG	RAP	TOG	LGP	Ours
Epsilon	Faster R-CNN	8	8	8	-	8	-	8	8	8	8	-	8	-	8
IW-SSIM↓		**0.16**	0.18	0.18	5.38	0.25	0.52	0.17	0.20	0.21	**0.19**	4.00	10.61	0.21	0.20
NMSE↓		**0.01**	0.02	**0.01**	0.58	0.02	0.04	**0.01**	0.02	0.02	0.02	0.62	1.34	0.02	**0.01**
TV↓		96.92	97.24	96.28	109.97	96.40	97.04	**96.14**	82.06	81.51	81.21	92.77	107.08	81.26	**81.12**
mAP50↓		17.81	3.96	3.36	11.30	8.90	1.49	**0.38**	33.15	7.46	5.44	47.46	5.25	3.41	**0.54**
mAP75↓		8.75	1.36	1.42	4.90	3.90	0.12	**0.06**	20.08	1.90	1.82	30.98	3.23	0.42	**0.04**
Metric	VFNet	FGSM	PGD	DAG	RAP	TOG	LGP	Ours	FGSM	PGD	DAG	RAP	TOG	LGP	Ours
Epsilon		-	-	-	-	8	-	8	-	-	-	-	8	-	8
IW-SSIM↓		-	-	-	-	0.22	0.39	**0.16**	-	-	-	-	0.21	0.61	**0.17**
NMSE↓		-	-	-	-	**0.01**	0.02	**0.01**	-	-	-	-	0.02	0.04	**0.01**
TV↓		-	-	-	-	96.37	96.59	**96.15**	-	-	-	-	81.37	81.76	**81.19**
mAP50↓		-	-	-	-	12.85	13.49	**5.49**	-	-	-	-	10.41	10.99	**1.96**
mAP75↓		-	-	-	-	3.92	3.49	**1.90**	-	-	-	-	3.82	1.57	**0.39**

Table 2: Comparison Study. Bolded values indicate the best results.



(a) Corruption (b) Pruning (c) Fine-tuning (d) Adversarial Training

Figure 6: The attack performance of NumbOD against different defenses on the MS-COCO dataset. (a) - (d) examine four defenses Corruption, Pruning, Fine-tuning, and Adversarial training on our method, respectively. "Benign*" and "Adv*" represent the results of the object detector after employing defense methods.

effective even when the erosion level is 5, with an average mAP_{50} value below 25%. These findings indicate that NumbOD can effectively resist the corruption-based preprocessing defense.

Pruning & Fine-tuning

Pruning (Zhu and Gupta 2017) involves selectively removing specific architectural components or parameters susceptible to exploitation by adversaries, thereby enhancing the resilience against adversarial attacks. As shown in Fig. 6 (b), we select pruning rates from 0 to 0.8, demonstrating NumbOD's consistent ability to execute potent attacks even as the detector approaches collapse. Similar to pruning, fine-tuning (Peng et al. 2022) involves modifying the model to adjust inherited pre-trained weights. We conduct fine-tuning on five widely used object detectors to defend against adversarial examples. The results in Fig. 6 (c) indicate an increase in the model's mAP_{50} after fine-tuning, but NumbOD still maintains high attack performance.

Adversarial Training

Adversarial training (Madry et al. 2018) is considered one of the most effective defense mechanisms against adversarial attacks, enhancing the robustness of models by introducing noise into the training dataset. We fine-tune five well-trained object detectors from the MMDetection repository on the MS-COCO dataset. As shown in Fig. 6 (d), our method maintains strong attack performance, with only a slight mAP_{50} drop of less than 2.5%, even after adversarial training. This confirms our method's resilience against adversarial training.

Conclusion

In this paper, we propose NumbOD, the first model-agnostic spatial-frequency fusion attack against object detectors, rendering them numb to input images and unable to detect objects. It consists of a spatial coordinated deviation attack and a critical frequency interference attack. We first design a dual-track attack target selection strategy, selecting the top-k high-quality bounding boxes independently from both classification and regression subtasks as attack targets. Subsequently, we utilize directional induction to shift the detected bounding boxes output by the object detectors and devise a foreground-background separation attack to disrupt classification, thereby deceiving the model in the spatial domain. Concurrently, we distort the high-frequency information of images in the frequency domain to enhance the attack efficiency for critical objects. Our extensive experiments on nine object detectors and two datasets show that our NumbOD achieves high attack performance and stealthiness, surpassing SOTA attacks against object detectors.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.62072204) and the National Key Research and Development Program of China (Grant No.2022YFB4502001). The computation is completed in the HPC Platform of Huazhong University of Science and Technology. Dezhong Yao and Wei Wan are co-corresponding authors.

References

Aich, A.; Ta, C.-K.; Gupta, A.; Song, C.; Krishnamurthy, S.; Asif, S.; and Roy-Chowdhury, A. 2022. Gama: Generative adversarial multi-object scene attacks. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS'22)*, 36914–36930.

Cai, Z.; and Vasconcelos, N. 2019. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1483–1498.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*, 39–57.

Chen, P.-C.; Kung, B.-H.; and Chen, J.-C. 2021. Class-aware robust adversarial training for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 10420–10429.

Chow, K.-H.; Liu, L.; Loper, M.; Bae, J.; Gursoy, M. E.; Truex, S.; Wei, W.; and Wu, Y. 2020. Adversarial objectness gradient attacks in real-time object detection systems. In *Proceedings of the 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA'20)*, 263–272.

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, 2849–2858.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338.

Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; and Huang, W. 2021. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 3490–3499.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*, 580–587.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR'15)*.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2021a. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 1–11.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021b. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 2786–2795.

Huang, H.; Chen, Z.; Chen, H.; Wang, Y.; and Zhang, K. 2023. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR'23)*, 20514–20523.

Li, G.; Xu, Y.; Ding, J.; and Xia, G.-S. 2024a. Toward Generic and Controllable Attacks Against Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Li, M.; Wang, J.; Zhang, H.; Zhou, Z.; Hu, S.; and Pei, X. 2024b. Transferable Adversarial Facial Images for Privacy Protection. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24)*.

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

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV'14)*, 740–755.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 15315–15324.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2574–2582.

Peng, Z.; Li, S.; Chen, G.; Zhang, C.; Zhu, H.; and Xue, M. 2022. Fingerprinting Deep Neural Networks Globally via Universal Adversarial Perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 13430–13439.

Ramesh, K.; Kumar, G. K.; Swapna, K.; Datta, D.; and Rajest, S. S. 2021. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(27): e6–e6.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 29th Advances in Neural Information Processing Systems (NeurIPS'15)*.

Song, Y.; Zhou, Z.; Li, M.; Wang, X.; Deng, M.; Wan, W.; Hu, S.; and Zhang, L. Y. 2025. PB-UAP: Hybrid Universal Adversarial Attack For Image Segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'25)*.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; and Luo, P. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 14454–14463.

Tabernik, D.; and Skočaj, D. 2019. Deep learning for large-scale traffic-sign detection and recognition. *IEEE Transactions on Intelligent Transportation Systems*, 21(4): 1427–1440.

Tang, G.; Jiang, T.; Zhou, W.; Li, C.; Yao, W.; and Zhao, Y. 2023. Adversarial patch attacks against aerial imagery object detectors. *Neurocomputing*, 537: 128–140.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19)*, 9627–9636.

Wang, J.; Zhang, W.; Cao, Y.; Chen, K.; Pang, J.; Gong, T.; Shi, J.; Loy, C. C.; and Lin, D. 2020. Side-aware boundary localization for more precise object detection. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*, 403–419.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An invisible black-box backdoor attack through frequency domain. In *Proceedings of the European Conference on Computer Vision (ECCV'22)*, 396–413. Springer.

Wang, Y.; Chou, Y.; Zhou, Z.; Zhang, H.; Wan, W.; Hu, S.; and Li, M. 2025. Breaking Barriers in Physical-World Adversarial Examples: Improving Robustness and Transferability via Robust Feature. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI'25)*.

Wei, X.; Liang, S.; Chen, N.; and Cao, X. 2018. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*.

Wei, X.; Yu, J.; and Huang, Y. 2023. Physically adversarial infrared patches with learnable shapes and locations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, 12334–12342.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 3520–3529.

Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; and Bai, X. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1452–1459.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19)*, 9657–9666.

YOLO-V5. 2022. Available online: https://doi.org/10.5281/zenodo.4679653 (V5.0). Accessed on 1 April 2022.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 8514–8523.

Zhang, Y.; Xu, Y.; Junyu, S.; Zhang, L. Y.; Hu, S.; Li, M.; and Zhang, Y. 2025. Improving Generalization of Universal Adversarial Perturbation via Dynamic Maximin Optimization. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI'25)*.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 1367–1376.

Zhou, Z.; Hu, S.; Li, M.; Zhang, H.; Zhang, Y.; and Jin, H. 2023a. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM'23)*, 6311–6320.

Zhou, Z.; Hu, S.; Zhao, R.; Wang, Q.; Zhang, L. Y.; Hou, J.; and Jin, H. 2023b. Downstream-agnostic adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 4345–4355.

Zhou, Z.; Li, M.; Liu, W.; Hu, S.; Zhang, Y.; Wan, W.; Xue, L.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024a. Securely Fine-tuning Pre-trained Encoders Against Adversarial Examples. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP'24)*.

Zhou, Z.; Song, Y.; Li, M.; Hu, S.; Wang, X.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024b. Darksam: Fooling segment anything model to segment nothing. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS'24)*.

Zhu, M.; and Gupta, S. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR'20)*.