Transferable Adversarial Attacks for Object Detection Using Object-Aware Significant Feature Distortion

Xinlong Ding[1], Jiansheng Chen[1]*, Hongwei Yu[1], Yu Shang[2], Yining Qin[1], Huimin Ma[1]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, China
[2]Department of Electronic Engineering, Tsinghua University, China
dingxl22@xs.ustb.edu.cn, jschen@ustb.edu.cn, yuhongwei22@xs.ustb.edu.cn, shangy21@mails.tsinghua.edu.cn, qinyn22@xs.ustb.edu.cn, mhmpub@ustb.edu.cn

Abstract

Transferable black-box adversarial attacks against classifiers by disturbing the intermediate-layer features have been extensively studied in recent years. However, these methods have not yet achieved satisfactory performances when directly applied to object detectors. This is largely because the features of detectors are fundamentally different from that of the classifiers. In this study, we propose a simple but effective method to improve the transferability of adversarial examples for object detectors by leveraging the properties of spatial consistency and limited equivariance of object detectors' features. Specifically, we combine a novel loss function and deliberately designed data augmentation to distort the backbone features of object detectors by suppressing significant features corresponding to objects and amplifying the surrounding vicinal features corresponding to object boundaries. As such the target object and background area on the generated adversarial samples are more likely to be confused by other detectors. Extensive experimental results show that our proposed method achieves state-of-the-art black-box transferability for untargeted attacks on various models, including one/two-stage, CNN/Transformer-based, and anchor-free/anchor-based detectors.

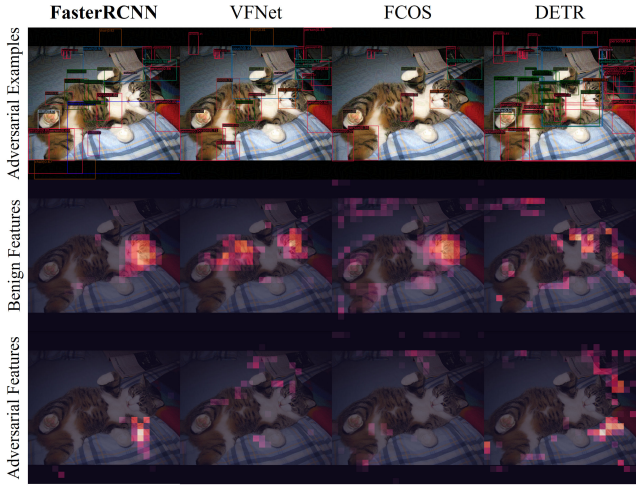

Figure 1: Adversarial examples crafted on FasterRCNN by our method effectively suppress the significant features and amplify the vicinal ones in both surrogate and target models. High-brightness areas correspond to high-value features.

Introduction

Deep neural networks have shown remarkable performance in real-world applications, yet concerns arise due to their vulnerability to adversarial examples (Szegedy et al. 2014). Adversarial attacks are crucial for evaluating the robustness of models and can be categorized in either white-box or black-box manner based on the level of attacker knowledge. Generally speaking, black-box attacks are more challenging and realistic for practical applications, as attackers usually cannot fully access the model's structure and parameters.

One approach to implement black-box adversarial attacks is to estimate gradients from queried data, known as query-based attacks (Brendel, Rauber, and Bethge 2018; Ilyas et al. 2018; Tramèr et al. 2017; Uesato et al. 2018). However, these attacks are often constrained by the limited number of possible queries in practice. Consequently, transfer-based attacks have been proposed, employing white-box attacks on a local surrogate model and leveraging adversarial examples' transferability to attack the target model.

Traditional white-box attacks (Goodfellow, Shlens, and Szegedy 2014; Kurakin, Goodfellow, and Bengio 2017; Madry et al. 2017) often suffer from overfitting to the source model, resulting in poor transferability to other black-box models. Various methods have been proposed to address this issue, including input transformation based (Xie et al. 2019; Dong et al. 2019; Lin et al. 2020), gradient optimization based (Dong et al. 2018; Lin et al. 2020; Gao et al. 2020), and model ensemble based (Dong et al. 2018; Liu et al. 2017; Xiong et al. 2022). Moreover, rather than directly manipulating the output layer, attacking the intermediate layers (Zhou et al. 2018; Naseer et al. 2018; Ganeshan, BS, and Babu 2019; Lu et al. 2021; Wang et al. 2021b; Zhang et al. 2022a,b) has shown great potential for crafting more transferable adversarial examples recently.

*Corresponding author
Code is available at https://github.com/wakuwu/OSFD

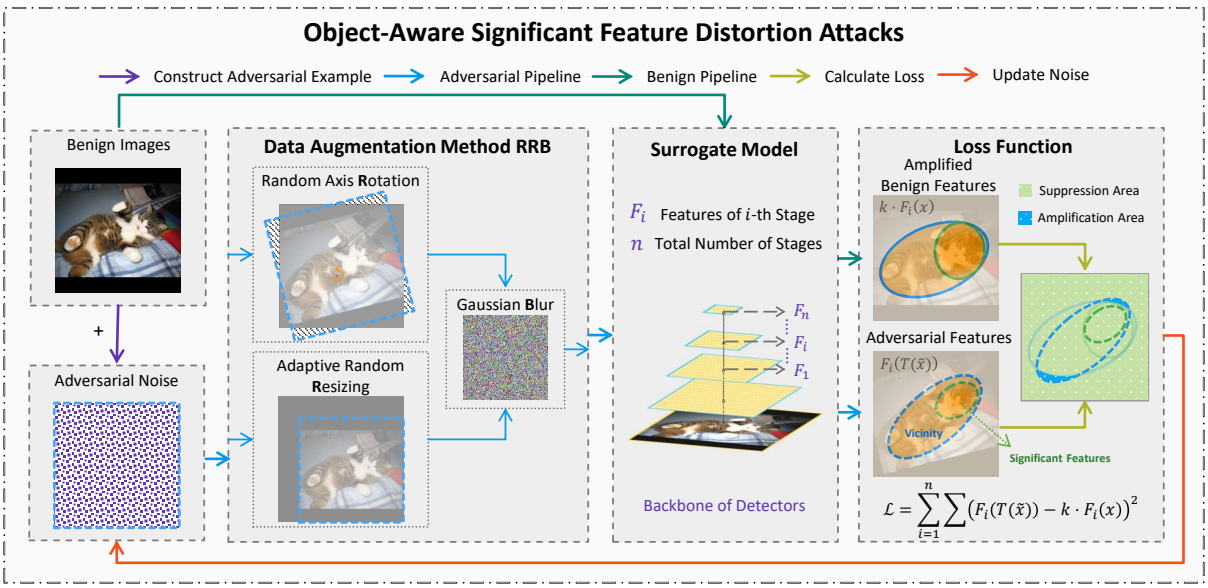

Figure 2: Overview of our framework for generating Object-Aware Significant Feature Distortion Attacks. Given a benign image, we add the adversarial noise to create an adversarial example. Data augmentation method RRB consists of parallel random axis rotation and adaptive random resizing techniques, along with Gaussian blur, to further transform the adversarial examples. After extracting backbone features from benign and augmented adversarial examples, we utilize our designed loss function to suppress the significant features of the target object and amplify its neighboring parts to achieve our OSFD attacks.

In contrast to classifiers, detectors typically have diverse and more flexible architectures consisting of backbone (Simonyan and Zisserman 2015; He et al. 2016; Sandler et al. 2018; Liu et al. 2021; Wang et al. 2021a), neck (Lin et al. 2017; Liu et al. 2018; Ghiasi, Lin, and Le 2019; Wang et al. 2019), and head components, which bring about more challenges for transferable adversarial attacks. While the neck and head components vary greatly across different detectors, the backbone often exhibits a notable similarity, with many detectors even sharing the same one, like ResNet. Thus, exploring the characteristics of backbone features to design attacks can be a promising approach to enhance adversarial transferability. However, existing untargeted attacks against detectors (Xie et al. 2017; Li et al. 2018; Zhang, Zhou, and Li 2020; Chow et al. 2020) mainly focus on leveraging the model output to design perturbation strategies, making the crafted adversarial examples overfit the model's structure. Although there is method (Wei et al. 2019) that further perturb intermediate layer features of detectors, their limited consideration for the characteristics of these features results in unsatisfactory adversarial transferability.

Compared to attacking irrelevant regions in the image, attacking the areas containing the objects and their vicinity has a greater impact on detection results. This is because detectors rely more on local regions of the image for precise object localization, whereas classifiers mainly utilize global information for category prediction. The backbone features of detectors produce high-valued responses in the regions containing objects, referred to as significant features, enabling the subsequent neck and head components to localize the objects' positions. Thus, these significant features have two key characteristics: a notable spatial consistency, indicating that their positions in the feature map closely match the location of objects in the image, and limited equivariance, whereby the features extracted from an image after a certain degree of data augmentation are approximately equivalent to the features obtained by directly applying the same augmentation. By leveraging these two properties, we design a loss function and data augmentation method RRB to suppress significant features and amplify the vicinal parts effectively, as shown in Fig. 1. Our OSFD (Object-Aware Significant Feature Distortion) method disrupts original object detection bounding boxes and yields additional interfering boxes on the objects and their neighboring regions, significantly enhancing the transferability of adversarial examples by confining the attack to the object and its vicinity. The illustration of our framework is presented in Fig. 2. The major contributions of this work are as follows.

- We have leveraged backbone features' spatial consistency and limited equivariance to design our OSFD method, significantly improving the transferability of untargeted attacks on object detection.
- Our OSFD method utilizes a simple and efficient loss function and data augmentation method RRB, consisting of random axis rotation, adaptive random resizing, and Gaussian blur, effectively distorting the significant features of the object and its vicinity.
- Our attack has undergone rigorous testing and validation, demonstrating its efficacy in attacking a wide range of detectors, compared to other state-of-the-art methods.

Related Work

Transferable Black-box Attacks. To alleviate the overfitting issues of traditional white-box attack methods (Goodfellow, Shlens, and Szegedy 2014; Kurakin, Goodfellow, and Bengio 2017; Madry et al. 2017; Carlini and Wagner 2017), many studies, including MIM (Dong et al. 2018), NIM (Lin et al. 2020), and VIM (Wang and He 2021), have been proposed to enhance the transferability of adversarial attacks by modifying the gradients. In addition to modifying the gradient, data augmentation on input images can further enhance the transferability of adversarial examples, like DIM (Xie et al. 2019), PIM (Gao et al. 2020), TIM (Dong et al. 2019), and SIM (Lin et al. 2020). In this paper, considering the nature of object detection, we propose an essential data augmentation method called RRB. Similarly to other approaches, our RRB can be easily ensembled with other attacks to generate more transferable adversarial examples.

Internal Feature Perturbation. Apart from generating adversarial examples at the output layer, some studies have focused on the features of the internal layers. TAP (Zhou et al. 2018) and NRDM (Naseer et al. 2018) have discovered that perturbing intermediate layer features by maximizing the feature distance between adversarial and benign images can enhance adversarial transferability. FDA (Ganeshan, BS, and Babu 2019) divides element properties by mean value across the channel, suppressing elements above and amplifying those below the mean. FVA (Lu et al. 2021) achieves adversarial attacks by reducing feature variance, sharing a similar idea with other methods above. FIA (Wang et al. 2021b), RPA (Zhang et al. 2022b), and NAA (Zhang et al. 2022a) incorporate label information to obtain aggregated gradients, enabling accurate measurement of feature importance and enhancing the transferability of adversarial attacks. While attacks like FDA have shown remarkable adversarial transferability on image classifiers, directly applying these methods to detectors does not yield satisfactory results due to their mismatch with the characteristics of detector features. Meanwhile, gradient aggregation-based attacks like RPA come with hardware and time costs, and may overfit complex detector model structures.

Adversarial Attacks for Object Detection. Early adversarial attacks of detectors focus on the white-box setting, where the methods are often limited by the type of detectors employed. Attacks like DAG (Xie et al. 2017), RAP (Li et al. 2018), and CAP (Zhang, Zhou, and Li 2020) are specifically designed to exploit the region proposal network (RPN) and are limited to proposal-based (two-stage) object detectors. Although the UEA (Wei et al. 2019) attack can target both one-stage and two-stage models, its limitation is that it only provides transfer attack results on a limited set of black-box models. TOG (Chow et al. 2020) can attack models regardless of their architectures since backpropagation on the training loss is feasible. Additionally, several other approaches (Cai et al. 2022b,a) leverage the co-occurrence of objects and their relative locations and sizes as contextual information to achieve transferable targeted attacks. Most of these methods primarily focus on perturbing the output of the detector model through design strategies, and only a small portion of them utilize intermediate layer features with limited consideration for their characteristics.

Methodology

In this section, we first define the black-box adversarial attack on detectors and specify our goals. Next, we will present the mathematical formulation and implementation of our proposal, then explain how our method achieves the object-aware significant feature distortion.

Problem Definition

Given a benign image x, by attacking a surrogate detector, we aim to craft its corresponding adversarial example $\tilde{x}$ that can be effectively transferred to multiple black-box detectors. This optimization problem can be formulated as Eq. 1, where $\mathcal{L}$ is a loss function for perturbing the model output or internal layers, and the l_p-norm is adopted to measure the distance between x and $\tilde{x}$.

$$arg\max_{\tilde{x}} \mathcal{L},\ s.t.\ \parallel \tilde{x} - x \parallel_p \leq \epsilon \tag{1}$$

In this paper, we specifically focus on the backbone of the detector, which typically involves $n(n \geq 1)$ stages for feature extraction. We denote the feature of the i-th stage as $F_i \in \mathbb{R}^{C \times H \times W}$, where i is an integer ranging from 1 to n. Considering spatial consistency and limited equivariance of the detectors, we aim to design loss function $\mathcal{L}$ and employ data augmentation method to distort the backbone features, thereby disturbing the original objects and its surrounding regions and achieving high black-box transferability.

Attack Formulation

Loss function. We start from the MSE loss and amplify the value of the benign features by multiplying a constant k, where $k \geq 1$. Our attack can be formulated as Eq. 2, where n is the total number of stages in the features, $F_{ij}(\cdot)$ denotes the j-th element of i-th stage features, which contain N_i items. $T(\cdot)$ represents the data augmentation performed on the adversarial examples before feature extraction.

$$\mathcal{L}_{OSFD} = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} (F_{ij}(T(\tilde{x})) - k \cdot F_{ij}(x))^2 \tag{2}$$

Random axis rotation. For a given image z, we can extract the coordinates of the centers of the object bounding boxes and the center of the entire image from the labels. We randomly sample a coordinate as the rotation axis for each data augmentation step from the square regions with side length l_s centered around these center points. And then, we rotate the image around the selected rotation axis by a random angle ϕ, where $\phi \in [-\theta, +\theta]$, and θ represents the pre-defined maximum rotation angle.

Adaptive random resizing. We adopt the resizing-padding pattern from DIM (Xie et al. 2019) and introduce an adaptive scaling factor s based on the object's size, limiting the maximum padding space p during augmentation.

$$s_h = min(1 + \rho \times \frac{b_h}{I_h}, s_{max})) \tag{3}$$

The scaling factor of height s_h is calculated using Eq. 3, where b_h and I_h are the height of the bounding boxes and the image, respectively, and $min(\cdot,\cdot)$ represents taking the minimum value. s_{max} is a constant used to limit s, and ρ is a parameter used to adjust the relationship between augmentation intensity and the size of the bounding boxes. The maximum height of padding space p_h applied after image resizing can be represented as Eq. 4.

$$max\ p_h = (s_h - 1) \times I_h \leq \rho \times b_h \tag{4}$$

The scaling factor of weight s_w and the maximum height of padding space p_w can be computed similarly to Eq. 3 and 4. For situations with multiple objects in an image, we randomly select a bounding box of one object during each augmentation to calculate the parameter s.

OSFD pipeline. Adversarial examples are constructed by combining benign images with noise, where the noise serves as our optimization target. Our data augmentation method RRB (Rotation, Resizing, and Blurring) is then applied to enhance adversarial examples. In RRB, random axis rotation and adaptive random resizing are applied in parallel before performing Gaussian blur with a mean of μ and a standard deviation of σ. We further extract the features of adversarial examples enhanced by RRB and combine them with amplified benign features to calculate and maximize the loss function. After back-propagating the gradients, we update the adversarial noise to complete one step of the attack.

Explanation of OSFD

Given an image z, we classify its backbone features $F(z)$ into three parts based on their importance to object detection: significant features $F_I(z)$, vicinal features $F_{II}(z)$, and vanilla features $F_{III}(z)$, respectively, as shown in Fig. 3. Significant features are high-valued responses in the feature map, often appearing in regions with objects due to their spatial consistency and directly impacting object detection results. Vanilla features typically correspond to negligible or nearly zero responses in regions without objects, providing little contribution to the detection of the current object. Vicinal features are distributed around significant features, typically corresponding to the neighboring regions of objects or some part of the object in the image. Magnifying them will directly affect the boundaries of significant features, impacting object detection accuracy.

Regarding the numerical magnitude of three-part features, we can get Eq. 5, where $|\cdot|$ represents the absolute value. This formula holds for any benign sample x and its corresponding adversarial example $\tilde{x}_t$ on t-th iteration attack.

$$|F_I(z)| > |F_{II}(z)| > |F_{III}(z)| \approx 0 \tag{5}$$

At the first iteration of the attack, where $t = 0$, the adversarial example $\tilde{x}_0$ is typically initialized as Eq. 6, where $|\Delta| < \epsilon$, ϵ is l_∞ constraint for the perturbation.

$$\tilde{x}_0 = x + \Delta \tag{6}$$

Due to the detector's robustness to a certain level of uniformly distributed random noise Δ, we can obtain Eq. 7.

$$F(\tilde{x}_0) = F(x + \Delta) \approx F(x) \tag{7}$$

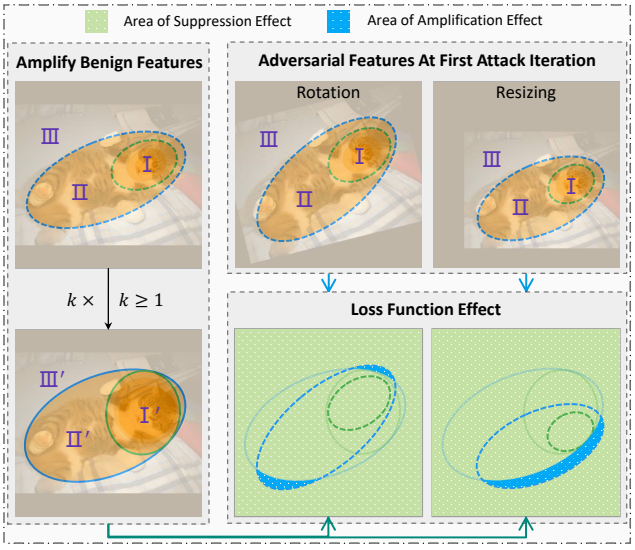

Figure 3: The suppression and amplification effects on features of enhanced adversarial examples by our loss function.

By leveraging the limited equivariance of the detector's backbone features, we will get Eq. 8, where $T(\cdot)$ represents a certain degree of data augmentation, including rotation, resizing, and blurring. It explains the reason for changes in the features of the enhanced adversarial example $T(\tilde{x}_0)$ at the first iteration attack compared to the features of the benign image x, as shown in Fig. 3.

$$F(T(\tilde{x}_0)) \approx T(F(\tilde{x}_0)) \approx T(F(x)) \tag{8}$$

As depicted in Fig. 3, amplifying benign features by multiplying them with a scaling factor k increases their values in significant, vicinal, and vanilla regions while expanding the coverage area of significant and vicinal features. By incorporating Eq. 5, we can derive Eq. 9 and 10.

$$|F_I(x)| \leq |F_{I'}(x)| \tag{9}$$

$$|F_{II}(x)| > |F_{III'}(x)| \approx 0 \tag{10}$$

Due to the detector features' spatial consistency and limited equivariance, data augmentation $T(\cdot)$ only changes the spatial position and area of different feature parts in the feature map without altering their numerical relationships. Therefore, combining with Eq. 8, we can obtain Eq. 11 and 12.

$$|F_I(T(\tilde{x}_0))| \approx |F_{T(I)}(\tilde{x}_0)| \approx |F_I(\tilde{x}_0)| \approx |F_I(x)| \tag{11}$$

$$|F_{II}(T(\tilde{x}_0))| \approx |F_{T(II)}(\tilde{x}_0)| \approx |F_{II}(\tilde{x}_0)| \approx |F_{II}(x)| \tag{12}$$

Regarding Eq. 9 and 10, we can derive the relationship between the features of enhanced adversarial examples and the amplified benign features as Eq. 13 and 14.

$$|F_I(T(\tilde{x}_0))| \leq |F_{I'}(x)| \tag{13}$$

$$|F_{II}(T(\tilde{x}_0))| > |F_{III'}(x)| \tag{14}$$

Our designed rotation and resizing methods aim to constrain the significant features of adversarial examples

Benign mAP		0.521	0.662	0.622	0.656	0.591	0.717	0.584	0.623	-
Model	Backbone / Attack	D-53 YOLOv3	R-50 VFNet	R-101 FRCNN	Swin MRCNN	R-50 YOLOF	CSPD YOLOX	R-50 FCOS	R-50 DETR	Mean
YOLOv3	NRDM	0.009	0.271	0.296	0.359	0.243	0.243	0.244	0.248	0.272
	FDA	0.001	0.319	0.336	0.416	0.285	0.352	0.277	0.318	0.329
	NAA	0.012	0.323	0.326	0.385	0.284	0.298	0.275	0.319	0.316
	RPA	0.004	0.296	0.316	0.372	0.259	0.271	0.247	0.296	0.294
	TOG	0.004	0.375	0.376	0.446	0.332	0.395	0.320	0.370	0.373
	$\mathrm{OSFD}_{\mathrm{DIM}}$	**0.000**	0.139	0.166	0.223	0.129	0.122	0.127	0.142	0.150
	$\mathrm{OSFD}_{\mathrm{RRB}}$	**0.000**	**0.107**	**0.128**	**0.172**	**0.097**	**0.114**	**0.099**	**0.111**	**0.118**
VFNet	NRDM	0.195	0.010	0.096	0.273	0.066	0.257	0.062	0.019	0.138
	FDA	0.111	**0.000**	0.022	0.201	0.009	0.202	0.010	**0.001**	0.079
	NAA	0.187	0.007	0.087	0.247	0.066	0.269	0.053	0.022	0.133
	RPA	0.170	0.005	0.053	0.223	0.035	0.249	0.027	0.011	0.109
	TOG	0.283	0.006	0.168	0.372	0.130	0.405	0.105	0.046	0.215
	$\mathrm{OSFD}_{\mathrm{DIM}}$	0.085	**0.000**	0.014	0.125	0.005	0.124	0.006	**0.001**	0.051
	$\mathrm{OSFD}_{\mathrm{RRB}}$	**0.065**	**0.000**	**0.012**	**0.088**	**0.004**	**0.119**	**0.005**	**0.001**	**0.042**
FRCNN	NRDM	0.180	0.043	0.002	0.320	0.035	0.352	0.034	0.065	0.147
	FDA	0.104	0.007	**0.000**	0.192	0.004	0.254	0.003	0.018	0.083
	NAA	0.155	0.056	0.003	0.224	0.046	0.284	0.044	0.067	0.125
	RPA	0.135	0.034	0.002	0.190	0.029	0.260	0.026	0.044	0.103
	TOG	0.173	0.038	**0.000**	0.254	0.032	0.336	0.024	0.058	0.130
	$\mathrm{OSFD}_{\mathrm{DIM}}$	0.080	0.004	**0.000**	0.132	0.002	0.176	0.002	0.009	0.058
	$\mathrm{OSFD}_{\mathrm{RRB}}$	**0.049**	**0.001**	**0.000**	**0.077**	**0.001**	**0.144**	**0.001**	**0.004**	**0.039**
MRCNN	NRDM	0.111	0.100	0.147	0.003	0.107	0.195	0.109	0.095	0.123
	FDA	0.360	0.448	0.469	0.088	0.427	0.549	0.403	0.405	0.437
	NAA	0.147	0.147	0.175	0.005	0.147	0.207	0.137	0.135	0.156
	RPA	0.114	0.100	0.120	0.002	0.099	0.155	0.093	0.085	0.109
	TOG	0.128	0.107	0.137	**0.000**	0.113	0.170	0.099	0.104	0.123
	$\mathrm{OSFD}_{\mathrm{DIM}}$	0.084	0.064	0.105	0.002	0.075	0.132	0.080	0.059	0.085
	$\mathrm{OSFD}_{\mathrm{RRB}}$	**0.039**	**0.035**	**0.055**	0.001	**0.063**	**0.113**	**0.042**	**0.032**	**0.054**

Table 1: The mAP metric of different attacks against detectors. The first line represents the mAP metric for benign samples across various detectors. Attacks take place on the white-box detectors in the first column, and metrics are measured across all models, with the last column representing the mean of all black-box results. $\mathrm{OSFD}_{\mathrm{DIM}}$ and all comparative attack methods are ensembled with MIM and DIM. $\mathrm{OSFD}_{\mathrm{RRB}}$ utilizes our proposed data augmentation method RRB for optimization instead of integrating DIM. The best results are highlighted in bold, while the second-best results are marked underlined.

$F_I(\tilde{x}_0)$ as much as possible within the region of $F_{I'}(x)$ while allowing vicinal features $F_{II}(\tilde{x}_0)$ to intersect with widely distributed vanilla features $F_{III'}(x)$. Besides, we apply an appropriate level of Gaussian noise to generate high-valued noise responses around the significant features $F_I(\tilde{x}_0)$, which helps the amplification effect for the features $F_{II}(\tilde{x}_0)$. Using the loss function in Eq. 2, we can suppress the significant features $F_I(\tilde{x}_0)$ of adversarial examples and amplify a portion of vicinal features $F_{II}(\tilde{x}_0)$, resulting in object-aware significant feature distortion.

Experiments

Experiment Setup

Dataset. We randomly select 2000 images from the trainval dataset of VOC2012 (Everingham et al. 2009). Each image is resized to 800×800, padding with zeros.

Models. We choose four representative models containing YOLOv3 (Redmon and Farhadi 2018), VFNet (Zhang et al. 2021), FasterRCNN (Ren et al. 2015), and MaskRCNN (He et al. 2017) as the source model to craft adversarial examples. The backbones of the four models are divided into CNN-based models, including Darknet-53 (Redmon and Farhadi 2018), ResNet-50 (He et al. 2016), ResNet-101 (He et al. 2016), and Transformer-based models, including Swin-Transformer (Liu et al. 2021). In addition, YOLOv3 and VFNet are anchor-based and anchor-free one-stage detectors, respectively, while FasterRCNN and MaskRCNN are two-stage detectors. To further validate the transferability of adversarial examples across a diverse range of detection architectures, we additionally selected YOLOF (Chen et al. 2021), YOLOX (Ge et al. 2021), FCOS (Tian et al. 2019), and DETR (Carion et al. 2020) detectors as target models. These models offer more backbones, including CSPDarknet,

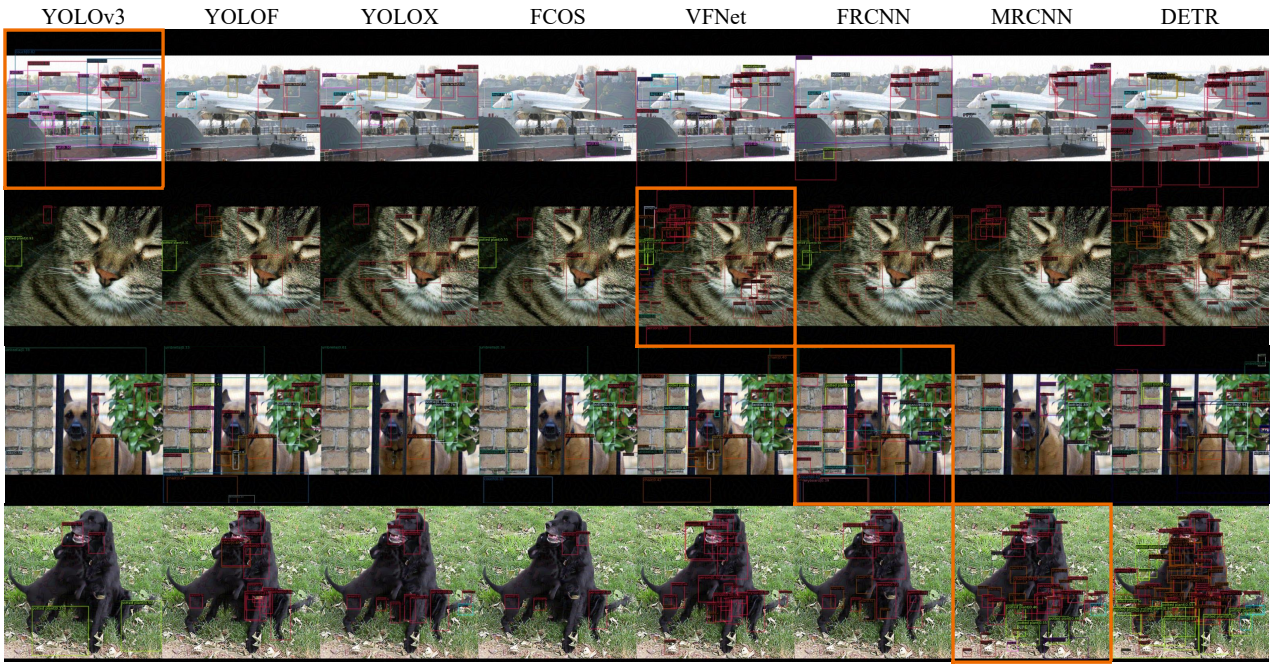

Figure 4: Visual results of the adversarial examples crafted by our OSFD_{RRB} method across various detectors. The column with colored borders shows the white-box attack results, while the others depict black-box attacks. Our method effectively suppresses the significant features corresponding to the original object while amplifying the vicinal features surrounding it. It enables the generated adversarial examples to evade detection while effectively increasing nearby detection boxes. Adversarial examples generated on a white-box detector can exhibit similar attack effects when transferred to other black-box models.

and various neck and head designs, such as DETR. MaskR-CNN and DETR are Transformer-based, while the other six detectors in Tab. 1 are CNN-based. All models are well-trained on the COCO (Lin et al. 2014) dataset using the mmdetection (Chen et al. 2019) framework.

Baseline Methods. We select four feature-level adversarial attacks, namely NRDM (Naseer et al. 2018) and FDA (Ganeshan, BS, and Babu 2019), as representatives of methods that solely attack the features, and NAA (Zhang et al. 2022a), and RPA (Zhang et al. 2022b), as examples of aggregated gradients based approaches. In terms of adversarial attacks on detectors, TOG (Chow et al. 2020) exhibits superior performance compared to UEA (Wei et al. 2019) and RAP (Li et al. 2018). Other approaches (Cai et al. 2022b,a) primarily focus on targeted attacks. Hence, we select TOG as the state-of-the-art method for detectors since it can directly maximize the training loss to achieve untargeted attacks. All comparative methods are integrated with MIM (Dong et al. 2018) and DIM (Xie et al. 2019) by default to achieve better transferability.

Evaluation. To evaluate the effectiveness and transferability of the adversarial examples on both source and target models, we utilized the primary challenge metric of the COCO dataset as the evaluation criteria. The mAP metric in this paper computes AP at ten different IoU thresholds, ranging from 0.5 to 0.95, with an interval of 0.05, which provides a more reliable and accurate evaluation of the model's performance. In this paper, we perform 200 attack steps for all methods, which are sufficient for the attacks.

Parameters. In all experiments, We use the l_∞ norm as a constraint with the maximum perturbation $\epsilon = 5/255$. For our method, we set the amplification factor $k = 3$ for all four source models. As for random axis rotation, we consider a center region with a size of $l_s = 10$ and a maximum rotation angle of $\theta = 7°$. For adaptive random resizing, we set the parameter $\rho = 0.8$ and $s_{max} = 1.1$. For Gaussian blur, we set the mean $\mu = 0$ and the standard deviation $\sigma = 6.0$.

Evaluation of Attack Performance

Comparison of Transferability. We evaluated the mAP of adversarial examples generated by our method on YOLOv3, VFNet, FasterRCNN, and MaskRCNN. As shown in Tab. 1, the OSFD_{DIM} method, which consists of our designed loss function and DIM, has already outperformed other comparative attacks regarding effectiveness on adversarial examples crafted on the four source models, both in white and black-box attacks. OSFD_{RRB}, which replaces DIM with our augmentation method RRB, further demonstrates its remarkable black-box adversarial transfer effect.

NRDM can be regarded as a special of the OSFD_{DIM} method, where the amplification parameter for benign features is set to $k = 1$. The transferability results of NRDM are obviously inferior to OSFD_{DIM}, which also demonstrates the necessity of suppressing signif icant features. Despite the FDA achieving good transfer results on VFNet and Faster-RCNN, its generality to detectors is still a non-negligible

OSFD Settings	**FRCNN**	YOLOv3	VFNet	MRCNN	YOLOF	YOLOX	FCOS	DETR	Mean
baseline	**0.000**	**0.049**	**0.001**	0.077	**0.001**	**0.144**	**0.001**	0.004	**0.039**
w/o RRB	**0.000**	0.302	0.046	0.403	0.032	0.501	0.029	0.055	0.195
w/o Resizing	**0.000**	0.073	**0.001**	**0.092**	**0.001**	0.165	**0.001**	0.004	0.048
w/o Rotation	**0.000**	0.085	0.002	0.137	**0.001**	0.199	**0.001**	0.008	0.062
w/o Blur	**0.000**	0.069	**0.001**	0.088	**0.001**	0.157	**0.001**	**0.003**	0.046

Table 2: We conduct an ablation study on three data augmentation methods in our RRB: random axis rotation, adaptive random resizing, and Gaussian blur. The baseline for the ablation study is our $\mathrm{OSFD}_{\mathrm{RRB}}$. The last column represents the mean of all black-box results. All attack experiments are conducted on FasterRCNN, and the generated adversarial samples are transferred to various models to measure mAP metric.

issue. FDA uses the cross-channel mean value as the criterion to distinguish positive and negative attributes, which lacks precision when dealing with YOLOv3 features that may contain negative values. NAA and RPA, which are gradient aggregation-based methods, incur forward-backward costs for obtaining accurate gradients and suffer from overfitting issues in complex detector model structures. As a result, they demonstrate weaker adversarial transfer results on models like YOLOv3. The TOG method, which maximizes the training loss to achieve untargeted attacks, is also prone to overfitting due to the impact of the neck and head structures and parameters. TOG generally shows inferior transferability results when attacking four models compared to NRDM, indicating that focusing on attacking the feature layers can reduce information interference and lead to better transfer performance. Moreover, Tab. 1 shows the high natural transferability of adversarial samples between detectors with ResNet series backbones. For instance, adversarial samples of all attacks on VFNet can be well transferred to target detectors with ResNet 50 backbone, indicating that similar backbones share similar vulnerabilities despite the differences in detector neck and head structures.

Visualization. We visualized the adversarial examples generated by our $\mathrm{OSFD}_{\mathrm{RRB}}$ method, and the detection results on various models are shown in Fig. 4. The results demonstrate that the adversarial examples generated by our method successfully suppress the detection of the original objects in the white-box detector while generating numerous bounding boxes on the objects and their surrounding regions. When transferring adversarial examples to other black-box models, although the adversarial noise may not always generate numerous detection boxes around the object, the original object can always successfully evade detection due to the perturbation strategy applied to the object and its surrounding region. Significant features typically correspond to a specific part for larger objects in images, while vicinal features in the surrounding region may correspond to other parts of the original object. In this case, the suppression and amplification effect of the loss function leads to the misalignment of features from different parts of the larger object, thereby decomposing the detection bounding box of the large object into multiple smaller ones, as illustrated in the second row of Fig. 4. In the case of smaller objects, significant features often represent the overall object. Our attack, on the other hand, tends to suppress the original detection bounding box and generate additional bounding boxes around it.

Ablation Study

To investigate the impact of data augmentations on our $\mathrm{OSFD}_{\mathrm{RRB}}$, we conducted ablation experiments on the FasterRCNN, as shown in Tab. 2. We remove the RRB data augmentation method from $\mathrm{OSFD}_{\mathrm{RRB}}$ and only keep our loss function and MIM, as shown in the second-row results in Tab. 2. In this scenario, the loss function we designed will suppress all features of the adversarial examples. The lack of amplification of vicinal features hinders its performance compared to the comparative methods in Tab. 1 regarding black-box transferability, highlighting the importance of perturbing the surrounding regions of the object.

To evaluate the individual contributions of each component in RRB to adversarial attacks, we conduct experiments by removing one of them while retaining the other two. The experimental results show that rotation improves the final black-box transfer results more than resizing. It can be explained that during the attack on FasterRCNN, the rotation operation not only maintains the suppression of significant features but also provides more amplification potential for vicinal features. Considering the distinctiveness of the two augmentation modes, our RRB method parallelly integrates both to achieve greater diversity. Gaussian blur is an auxiliary approach to increase the possibility of amplifying vicinal features by introducing noise to features. The experimental results demonstrate a decrease in the adversarial transferability of generated adversarial examples on a black-box model after ablating it.

Conclusions

This paper proposes the Object-Aware Significant Feature Distortion method to craft untargeted adversarial attacks against object detectors. Specifically, considering the spatial consistency and limited equivariance of the detector's backbone features, we design a concise and efficient loss function in conjunction with a compound data augmentation method to effectively suppress significant features associated with objects and amplify vicinal features in the adjacent regions. The experiments demonstrate that the adversarial examples crafted by our method can achieve satisfactory transferability across multiple detectors, achieving state-of-the-art black-box adversarial attack performance.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022ZD0117902) and by the National Natural Science Foundation of China (62376024, U20B2062).

References

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Cai, Z.; Rane, S.; Brito, A. E.; Song, C.; Krishnamurthy, S. V.; Roy-Chowdhury, A. K.; and Asif, M. S. 2022a. Zero-query transfer attacks on context-aware object detectors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 15024–15034.

Cai, Z.; Xie, X.; Li, S.; Yin, M.; Song, C.; Krishnamurthy, S. V.; Roy-Chowdhury, A. K.; and Asif, M. S. 2022b. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 149–157.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision (ECCV)*, 213–229.

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

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; and Sun, J. 2021. You only look one-level feature. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 13039–13048.

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 IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 263–272. IEEE.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4312–4321.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2009. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88: 303–308.

Ganeshan, A.; BS, V.; and Babu, R. V. 2019. Fda: Feature disruptive attack. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 8069–8079.

Gao, L.; Zhang, Q.; Song, J.; Liu, X.; and Shen, H. T. 2020. Patch-wise attack for fooling deep neural network. In *Proceedings of European Conference on Computer Vision (ECCV)*, 307–322.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2019. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7036–7045.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. In *Proceedings of International Conference on Machine Learning (ICML)*, 2137–2146.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world. In *Proceedings of International Conference on Learning Representations (ICLR)*.

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

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2020. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2117–2125.

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 European Conference on Computer Vision (ECCV)*, 740–755.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path Aggregation Network for Instance Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 10012–10022.

Lu, Y.; Du, X.; Sun, B.; Ren, H.; and Velipasalar, S. 2021. Fabricate-vanish: An effective and transferable black-box adversarial attack incorporating feature distortion. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 809–813.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Naseer, M.; Khan, S. H.; Rahman, S.; and Porikli, F. 2018. Task-generalizable adversarial attack based on perceptual metric. *arXiv preprint arXiv:1811.09020*.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 28.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 9627–9636.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

Uesato, J.; O'donoghue, B.; Kohli, P.; and Oord, A. 2018. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of International Conference on Machine Learning (ICML)*, 5025–5034.

Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C. C.; and Lin, D. 2019. CARAFE: Content-Aware ReAssembly of Features. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 3007–3016.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021a. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 568–578.

Wang, X.; and He, K. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1924–1933.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021b. Feature importance-aware transferable adversarial attacks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 7639–7648.

Wei, X.; Liang, S.; Chen, N.; and Cao, X. 2019. Transferable adversarial attacks for image and video object detection. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 954–960.

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 IEEE International Conference on Computer Vision (ICCV)*, 1369–1378.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2730–2739.

Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J. E.; and He, K. 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 14983–14992.

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

Zhang, H.; Zhou, W.; and Li, H. 2020. Contextual adversarial attacks for object detection. In *Proceedings of International Conference on Multimedia and Expo (ICME)*, 1–6.

Zhang, J.; Wu, W.; Huang, J.-t.; Huang, Y.; Wang, W.; Su, Y.; and Lyu, M. R. 2022a. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 14993–15002.

Zhang, Y.; Tan, Y.-a.; Chen, T.; Liu, X.; Zhang, Q.; and Li, Y. 2022b. Enhancing the Transferability of Adversarial Examples with Random Patch. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 1672–1678.

Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; and Yang, Y. 2018. Transferable adversarial perturbations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 452–467.