

# Adversarial Attention Perturbations for Large Object Detection Transformers

Zachary Yahn<sup>1</sup>, Selim Furkan Tekin<sup>1</sup>, Fatih Ilhan<sup>1</sup>, Sihao Hu<sup>1</sup>,  
Tiansheng Huang<sup>1</sup>, Yichang Xu<sup>1</sup>, Margaret Loper<sup>2</sup>, Ling Liu<sup>1</sup>

<sup>1</sup>Georgia Institute of Technology, Atlanta, GA

<sup>2</sup>Georgia Tech Research Institute, Atlanta, USA

{zachary.yahn, stekin6, filhan, sihaohu, thuang, xuyichang}@gatech.edu,  
margaret.loper@gtri.gatech.edu, ling.liu@cc.gatech.edu

## Abstract

Adversarial perturbations are useful tools for exposing vulnerabilities in neural networks. Existing adversarial perturbation methods for object detection are either limited to attacking CNN-based detectors or weak against transformer-based detectors. This paper presents an Attention-Focused Offensive Gradient (AFOG) attack against object detection transformers. By design, AFOG is neural-architecture agnostic and effective for attacking both large transformer-based object detectors and conventional CNN-based detectors with a unified adversarial attention framework. This paper makes three original contributions. First, AFOG utilizes a learnable attention mechanism that focuses perturbations on vulnerable image regions in multi-box detection tasks, increasing performance over non-attention baselines by up to 30.6%. Second, AFOG’s attack loss is formulated by integrating two types of feature loss through learnable attention updates with iterative injection of adversarial perturbations. Finally, AFOG is an efficient and stealthy adversarial perturbation method. It probes the weak spots of detection transformers by adding strategically generated and visually imperceptible perturbations which can cause well-trained object detection models to fail. Extensive experiments conducted with twelve large detection transformers on COCO demonstrate the efficacy of AFOG. Our empirical results also show that AFOG outperforms existing attacks on transformer-based and CNN-based object detectors by up to 83% with superior speed and imperceptibility. Code is available at: [Link](#).

## 1. Introduction

Transformer-based neural architectures and algorithms have blossomed in recent years, enhancing computer vision tasks including object detection. Attention is the core of the transformer architecture [30]. Attention allows a detector to focus on specific regions of images based on objects of interest, effectively predicting the presence and lo-

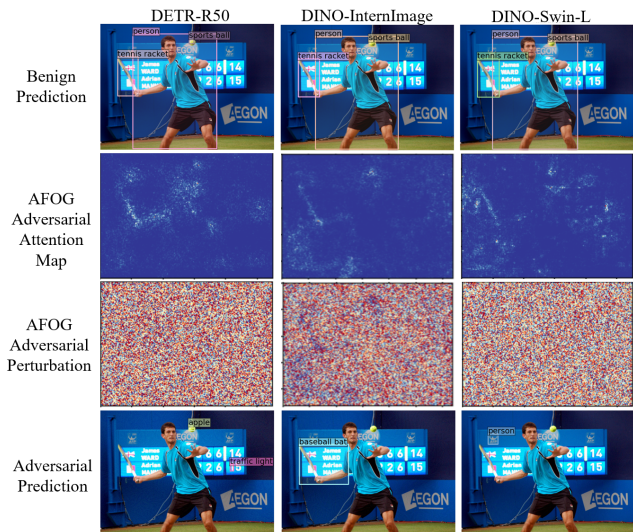


Figure 1. Successful AFOG attacks against three object detection transformers. White pixels in the attention map indicate learned areas of high importance, often corresponding to objects in the input image. This example illustrates that AFOG’s adversarial attention learns which regions of an image are most susceptible to adversarial perturbations. Even for the same input, learnable adversarial attention maps may differ under different models.

cation of each potential object. Powered by various attention mechanisms, detection transformers can handle overlapping objects more effectively than some older methods because they can capture long-range dependencies in an image, allowing models to reason about the relationships between objects. Since modern object detection transformers such as Swin [20] and DETR [4] significantly outperform traditional convolutional neural network (CNN)-based models [8], represented by Faster R-CNN [28], SSD [19] and YOLO-v3 [27], there is a pressing need to investigate and understand the vulnerabilities of large detection transformers in the presence of adversarial perturbations. Adversarial perturbations are helpful instrumentation for exposing vulnerabilities of large detection transformers, making them useful mechanisms for motivating more robust transformer models for object detection.

Existing attack methods struggle to disrupt transformer-based object detectors. Surrogate-based attacks (also known as black-box attacks), such as UEA [33] and RAD [5] generate perturbations on surrogate models and then test the adversarial effect against victim detectors that have similar detection architecture at inference. However, surrogate-based attacks suffer poor attack performance when the victim architectures are not similar to the surrogate models used to generate adversarial perturbations. Victim-based (also known as white-box) attacks, such as EBAD [3] and OATB [12], generate adversarial attacks by performing direct inference against victim models. Recent studies, such as AttentionFool [22], specialize in transformer-based victim models, but these transformer-only attacks are not applicable to convolutional detectors such as YOLO [27].

In this paper, we present an Attention-Focused Offensive Gradient (AFOG) attack, which is victim architecture agnostic and effective in attacking both advanced object detection transformers and traditional CNN-based detectors. AFOG by design offers three novel characteristics. First, inspired by a transformer’s self-attention, we utilize a learnable attention mechanism to enable AFOG to adaptively focus its adversarial perturbations on vulnerable image areas (those areas where perturbations have the greatest effect on the output) in multi-box detection tasks. Second, we formulate the attack loss function of AFOG by integrating learnable feature-map-based attention updates with iterative injection of adversarial perturbations. Finally, we design AFOG to be an efficient and yet stealthy adversarial perturbation method. By efficient, we want AFOG to rapidly generate adversarial perturbations in minimal iterations. By stealthy, we want AFOG to generate small amounts of perturbations that are visually imperceptible and yet can cause well-trained object detectors to produce erroneous detection results. Figure 1 shows an example of successful AFOG attacks against three detection transformers. Row 1 shows benign detection from three state-of-the-art detectors, Row 2 and Row 3 show AFOG adversarial perturbations on the three detectors and their respective AFOG adversarial attention maps, and Row 4 shows the detection results under our AFOG attack. We validate AFOG with extensive experiments on twelve state-of-the-art object detection transformers with the COCO benchmark [18] and three popular families of traditional CNN-based object detectors. The results show that AFOG is consistently effective across all twelve detection transformers compared to existing methods, achieving high attack success rate and reducing benign mAP (mean Average Precision) by up to  $37.79\times$ .

## 2. Related Work

Existing adversarial perturbations against object detectors can be broadly categorized into surrogate-model based

(also known as black-box) and victim-model based (white-box) methods. Surrogate-based approaches in literature tend to rely on the assumption that the surrogate model is similar to the victim model. RAD [6], GHFD [32], and UEA [33] generate adversarial examples on a surrogate FR-CNN model, then attack other CNN-based detectors such as YOLO and SSD. However, these attacks show poor performance against object detection transformers [32]. For victim-model based methods, GARSDC [17] is more effective but can require more than 3000 iterations to converge. GALD [13] first attacks vision transformer classifiers and then transfers the classification-based adversarial perturbations to object detectors of similar transformer architecture, and hence is less effective by comparison. We classify EBAD [3] as a victim-based attack because it uses an ensemble of surrogate models to attack a victim, but requires access to the victim’s loss function for its nested ensemble optimization. EBAD shows poor performance when transferring perturbations from a DETR surrogate to a DETR victim [24]. AttentionFool [22] is a recent victim-based attack targeting dot-product self-attention against DETR [14] with a ResNet-101 backbone. However, AttentionFool shows inconsistent performance against DETR with a ResNet-50 backbone [22]. AttentionFool is also not applicable to convolutional models such as YOLO [27] because it exclusively targets self-attention mechanisms. TOG [7] targets a prediction’s objectness score and introduces vanishing and fabrication attack modes. OATB [12] uses a “division map” that statically emphasizes image regions during perturbation based on object location priors. DBA [16] prioritizes perturbations against image backgrounds to enhance imperceptibility, though it shows almost no effect against a Swin transformer for object detection [20]. In comparison, AFOG adversarial perturbations show strong performance against a wide variety of transformer- and CNN-based detector victims while maintaining superior imperceptibility and efficiency.

## 3. Methodology

### 3.1. Problem Definition

Given a victim detector  $f_D(\vartheta, x)$ , where  $x \in \mathcal{D}$  is a victim image  $x$  and  $\mathcal{D}$  is the test set, let  $x$  contain  $N_x$  objects to be detected, denoted by  $\mathcal{O}_x = \{O_1, O_2, \dots, O_{N_x}\}$ . Each object  $O_i$  ( $1 \leq i \leq N_x$ ) is a recognition target for the detector  $f_D(\vartheta, x)$ . Under benign scenarios, let  $(B_i, C_i)$  denote a ground truth object  $O_i$  with bounding box  $B_i$  and class label  $C_i$ . Let  $K$  denote the total number of ground truth classes, and  $C_i \in \{1, 2, \dots, K\}$ , e.g.,  $K = 21$  for the VOC dataset [9], including the background class. Given input image  $x$ ,  $f_D(\vartheta, x)$  outputs the benign prediction of  $N$  detected objects, denoted by  $\mathcal{R}[f_D(\vartheta, x)] = \{(b_i, c_i) | i = 1, \dots, N_x\}$ . Each detected object  $o_i$  is associated with

its predicted bounding box  $b_i$  and predicted class label  $c_i$ .  $(b_i, c_i)$  is evaluated as a correct prediction if the intersection over union (IoU) of  $B_i$  and  $b_i$  is greater than the detection threshold  $\gamma$  (usually set to 0.5), and  $C_i = c_i$  ( $1 \leq i \leq N_x$ ,  $x \in \mathcal{D}$ ). The overall detection accuracy is measured using mAP (mean Average Precision) on the entire test dataset  $\mathcal{D}$ .

Let  $x_{adv}$  denote the adversarial example generated by injecting a sequence of adversarial perturbations to  $x$  through an iterative attention-based learning mechanism. The goal of our AFOG attack on detector  $f_D(\vartheta, x)$  is to find  $x_{adv}$  that maximizes the success rate of falsifying the predictions of all object recognitions for all images in  $\mathcal{D}$ , i.e.

$$\operatorname{argmax}_{x \in \mathcal{D}, i \in N_x, \{(b_i, c_i) \in \mathcal{R}[f_D(\vartheta, x_{adv})\}} \\ (\text{IOU}(B_i, b_i) < \gamma \vee C_i \neq c_i), \min \|x - x_{adv}\|_p \}.$$

The formula indicates that for each detected object  $o_i$ , denoted by  $(b_i, c_i)$  in victim image  $x$ , the AFOG attack is successful on  $o_i$  if the IoU of  $B_i$  and  $b_i$  is less than the detection threshold  $\gamma$  (usually set to 0.5) or  $c_i \in \{1, 2, \dots, K\}$  ( $C_i \neq c_i$ ), and  $x_{adv}$  also satisfies the distortion constraint of  $\min \|x - x_{adv}\|_p$ , where  $p$  is typically defined by  $L_2$  norm,  $L_0$  norm, or  $L_\infty$  norm.

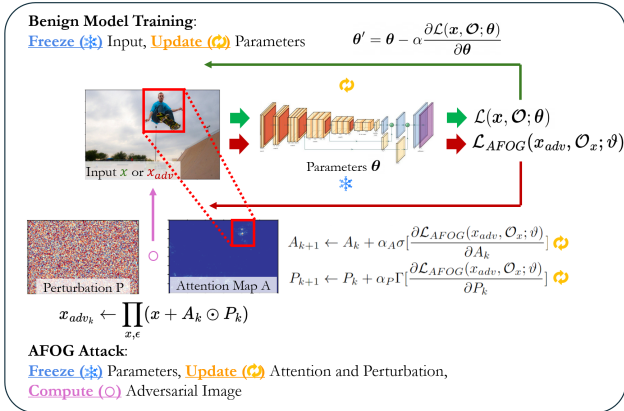


Figure 2. AFOG attack framework. During attack iteration the adversarial perturbation is updated via bounding-box and class loss gradient propagation. An adversarial attention map concurrently learns to apply per-pixel amplification or dampening to the perturbation, focusing its effects on vulnerable image regions.

### 3.2. Adversarial Loss Optimization

AFOG attacks the victim model via an iterative projected gradient descent method [23]. Figure 2 provides an illustration of the AFOG attack framework. At the initialization step, the attack first propagates the unaltered victim image  $x$  through the feed-forward network (FFN) of the victim model  $f_D(\vartheta, x)$  to obtain its benign prediction  $\mathcal{O}_x = \{(b_i, c_i) | i = 1, \dots, N_x\}$ .

In a scenario where no ground truth is available, the attack assumes these benign predictions to be correct labels for the input image  $x$ . For the remaining section, we will use  $(B_i, C_i)$  in the context of performing an attack on  $f_D(\vartheta, x)$

to obtain  $x_{adv}$ . Upon initialization for each input image  $x \in \mathcal{D}$ , the AFOG attack corrupts the victim image  $x$  by adding perturbations generated via an element-wise multiplication of two components using Equation 1:

$$x_{adv_k} \leftarrow \prod_{x, \epsilon} (x + A_k \odot P_k) \quad (1)$$

$A$  is the attention map, and  $P$  is the perturbation map. Here  $\odot$  represents the Hadamard matrix product and  $\Pi$  is a projection onto a hypersphere centered on unaltered victim image  $x$  with the radius  $\epsilon$  as the maximum perturbation budget.  $A$  and  $P$  are initialized according to Equation 2.

$$A_0 \sim 1, P_0 \sim \text{Random}(-\epsilon, \epsilon) \quad (2)$$

$\text{Random}(-\epsilon, \epsilon)$  is a uniform random distribution. The perturbed image  $x_{adv}$  propagates through the FFN of the victim model. The attack loss function  $\mathcal{L}_{AFOG}$  assesses the adversarial output  $\mathcal{O}_{x_{adv}}$  and benign output  $\mathcal{O}_x$ , computing a loss that reflects the attack’s progress in corrupting image  $x$ . This attack loss is formulated in Equations 3, 4, and 5.

$$\mathcal{L}_{AFOG}(x_{adv}, \mathcal{O}_x; \vartheta) = \mathcal{L}_{bbox}(x_{adv}, \mathcal{O}_x; \vartheta) + \mathcal{L}_{cls}(x_{adv}, \mathcal{O}_x; \vartheta) \quad (3)$$

$$\mathcal{L}_{bbox}(x_{adv}, \mathcal{O}_x; \vartheta) = \sum_{i=1}^{N_x} [f_{\vartheta}(x, o_i) - f_{\vartheta}(x_{adv}, o_{adv_i})] \quad (4)$$

$$\mathcal{L}_{cls}(x_{adv}, \mathcal{O}_x; \vartheta) = \sum_{i=1}^{N_x} [f_{\vartheta}(x, c_i) - f_{\vartheta}(x_{adv}, c_{adv_i})] \quad (5)$$

Given a victim image  $x$ , the optimized attack loss  $\mathcal{L}_{AFOG}$  is attained by making the model incorrectly predict every target object. We achieve this loss optimization by both falsifying every target object’s bounding box prediction and its class label prediction. This can be represented by optimizing an adversarial bounding box loss  $\mathcal{L}_{bbox}$  and an adversarial class label prediction loss  $\mathcal{L}_{cls}$  for all  $N_x$  target objects. Conceptually, this suppresses the confidence of the original correct bounding box and class label prediction, while increasing the confidence of the adversarial prediction of an incorrect bounding box or class label.

Recall Figure 2: the AFOG attack freezes the model parameters  $\vartheta$  and uses the gradient of the AFOG attack loss to update the attention map  $A$  and the perturbation  $P$  by back-propagation, generating the next iteration of the adversarial perturbation. A newly corrupted image  $x_{adv}$  is created upon injecting the updated adversarial perturbation by following Equation 1, where  $A$  and  $P$  are updated via Equations 6 and 7:

$$A_{k+1} \leftarrow A_k + \alpha_A \sigma \left[ \frac{\partial \mathcal{L}_{AFOG}(x_{adv}, \mathcal{O}_x; \vartheta)}{\partial A_k} \right] \quad (6)$$

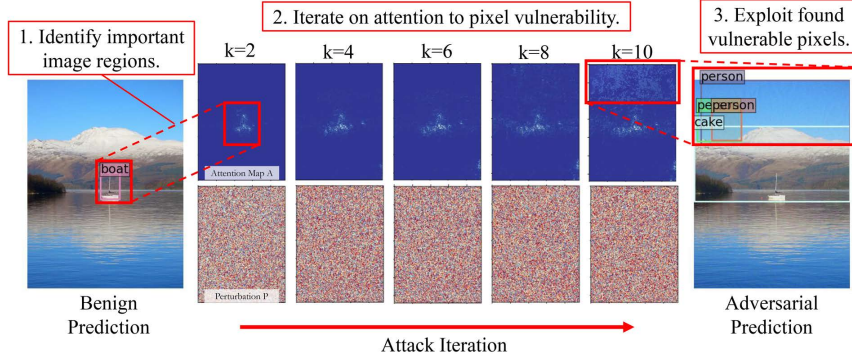


Figure 3. Example attention map generation on DETR-R101 for  $k=10$  iterations. AFOG attention initially learns important regions of the image, then iteratively updates to achieve an effective attack. It finds vulnerable pixels in foreground objects as well as background regions, including in unexpected or unintuitive areas such as the sky above the boat.

$$P_{k+1} \leftarrow P_k + \alpha_P \Gamma \left[ \frac{\partial \mathcal{L}_{AFOG}(x_{adv}, \mathcal{O}_x; \vartheta)}{\partial P_k} \right] \quad (7)$$

$\alpha_A$  and  $\alpha_P$  denote the attention map learning rate and the perturbation learning rate respectively,  $\sigma$  is a normalization function,  $\Gamma$  is a sign function,  $\mathcal{L}_{AFOG}$  is the attack loss,  $\vartheta$  is the model parameters, and  $\mathcal{O}_x$  is the model’s benign predictions from input  $x$ . This attack process repeats until the number of attack iterations is reached. Algorithm 1 provides the pseudocode.

### 3.3. Adversarial Attention Mechanism

A key innovation of our AFOG method is empowering perturbation generation with a learnable attention mechanism. Motivated by the intuition that certain parts of a victim image are more susceptible to adversarial perturbation than others, we add an attention map to focus the perturbation on vulnerable pixels. The AFOG attack concurrently learns the adversarial attention map alongside the perturbation to iteratively maximize the attack loss  $\mathcal{L}_{AFOG}$  (recall Equation 3). Figure 3 gives an illustrative example of the iterative learning of AFOG’s adversarial attention map and the corresponding perturbation. Unlike focusing mechanisms in other methods such as [33], AFOG’s attention map dynamically updates during attack iteration, leaving it unconstrained by assumptions about foreground [12], background [16], or region proposal [33] importance from static methods. We posit that pixel importance may not correspond to human intuition, and design AFOG to iteratively learn pixel importance instead. We observe that in early iterations the attention mechanism tends to focus on the primary objects in an image, and then branches out to affect surrounding areas as the attack progresses. This adaptability is a key feature of learnable attention, showcasing an advantage over static attention maps.

---

#### Algorithm 1 AFOG attack on an input image.

---

**Require:** Victim image  $x \in \mathcal{D}$ , test-set  $\mathcal{D}$ , Victim pre-trained model  $f_D(\vartheta)$ , Perturbation step size  $\alpha_P$ , Attention step size  $\alpha_A$ , Number of iterations  $T$ , Maximum perturbation  $\epsilon$ .

- 1: Initialize  $\mathcal{O}_x \leftarrow f_D(x; \vartheta)$
- 2: Initialize attention map  $A_0 \leftarrow 1$ ;
- 3: Initialize perturbation  $P_0 \leftarrow \text{Random}(-\epsilon, \epsilon)$ ;
- 4: Initialize step variable  $k \leftarrow 1$ ;
- 5: **while**  $k \leq T$  **do**
- 6:   Attack image  $x_{advk} \leftarrow \prod_{x, \epsilon} (x + A_k \odot P_k)$ ;
- 7:   Forward propagate  $x_{adv}$  through  $f_D(\vartheta, x_{adv})$ ;
- 8:   Compute bbox-loss  $\mathcal{L}_{bbox}(x_{adv}, \mathcal{O}_x; \vartheta)$ ;
- 9:   Compute cls-loss  $\mathcal{L}_{cls}(x_{adv}, \mathcal{O}_x; \vartheta)$ ;
- 10:    $\mathcal{L}_{AFOG}(x_{adv}, \mathcal{O}_x; \vartheta) = \text{bbox-loss} + \text{cls-loss}$ ;
- 11:   Calculate losses with respect to  $A_k$  and  $P_k$ :  
 $\mathcal{L}_A(x_{adv}, \mathcal{O}_x; \vartheta), \mathcal{L}_P(x_{adv}, \mathcal{O}_x; \vartheta)$ ;
- 12:   Normalize attention loss  $\mathcal{L}_A \leftarrow \text{Norm}(\mathcal{L}_A)$ ;
- 13:   Take sign of perturbation loss  $\mathcal{L}_P \leftarrow \text{Sign}(\mathcal{L}_P)$ ;
- 14:    $A_{k+1} \leftarrow A_k - \alpha_A \mathcal{L}_A$ ;
- 15:    $P_{k+1} \leftarrow P_k - \alpha_P \mathcal{L}_P$ ;
- 16:    $k \leftarrow k + 1$ ;
- 17: **end while**
- 18:  $x_{advk+1} \leftarrow \prod_{x, \epsilon} (x + A_{k+1} \odot P_{k+1})$ ;
- 19: **return**  $x_{adv}$

---

### 3.4. Special Cases of the AFOG Attack

We explore two special cases of our AFOG attack, each of which targets a specific malicious detection behavior in the victim detector. The first special case is an object vanishing attack, coined AFOG-V, which attempts to attack the objectness detection task of multi-box object detection. The goal of this special case AFOG attack is to make the victim model unable to detect any object, making all object

detections vanish for victim image  $x$ . We implement the AFOG-V attack by changing the initialization: we use an empty set instead of forward propagation of  $x$  to obtain the benign detection results for the  $N_x$  objects as the assumed ground-truth in AFOG-V attack. Let  $\emptyset$  denote an altered version of  $\mathcal{O}_x$  that contains no predictions. The formulation of the AFOG-V attack loss function is given by Equation 8:

$$\begin{aligned} \mathcal{L}_{AFOG_V}(x_{adv}, \mathcal{O}_x; \vartheta) = & -\mathcal{L}_{bbox}(x_{adv}, \emptyset; \vartheta) \\ & -\mathcal{L}_{cls}(x_{adv}, \emptyset; \vartheta) \end{aligned} \quad (8)$$

The second special case is object fabrication, coined as AFOG-F, which attempts to attack the bounding-box detection task by generating perturbations that lead to spurious detections (i.e., false positives). Similarly, we revise the AFOG attack process: instead of keeping the benign predictions with the confidence score above a certain tunable threshold (0.5 by default), for AFOG-F, we remove this threshold and allow the benign detection set  $\mathcal{O}_x$  to include a much larger set of ‘‘ground truth’’ targets, which inevitably include erroneous detections. Accordingly the AFOG-F loss function is given in Equation 9:

$$\begin{aligned} \mathcal{L}_{AFOG_F}(x_{adv}, \mathcal{O}_x; \vartheta) = & -\mathcal{L}_{bbox}(x_{adv}, \mathcal{O}_F; \vartheta) \\ & -\mathcal{L}_{cls}(x_{adv}, \mathcal{O}_F; \vartheta) \end{aligned} \quad (9)$$

where  $\mathcal{O}_F$  is a modified version such that the likelihood score of each prediction is set to 1.0. We explore these two specialized cases of our AFOG attack to gain a deeper understanding of the adverse effects of adversarial perturbations on different detection transformers.

## 4. Experiments

### 4.1. Experimental Setup

We use the Common Objects in Context (COCO) 2017 [18] test-dev split to evaluate the performance of our attack on contemporary object detectors. COCO 2017 is a standard benchmark for evaluating object detector performance, and its test-dev set has 80 object categories across 5,000 images. We use the PASCAL Visual Objects Challenge (VOC) 2007 [9] dataset to compare our attack with the state-of-the-art attacks on CNN-based object detectors, because existing attacks designed only for CNN-based detectors all report their evaluation with VOC. The VOC devkit split contains 4,952 images with 20 object classes. All experiments were performed on an NVIDIA A100. We list our models and their benign performance in Table 1. We provide model details in Section ?? of the Supplementary Material.

We select a variety of model sizes, ranging from Detection Transformer (DETR) [4], a lightweight 40 million parameter model, to EVA [10], a vision-focused foundation model with over one billion parameters. We use the Detrex framework [29], built on top of Detectron2 [34], to standardize our model implementations via DINO [37]. DINO

and Detrex adapt many general purpose vision models to object detection. The Detrex framework standardizes all images in  $[0, 1]$ . We also attack the original versions of several transformer models to demonstrate that the AFOG attack also works beyond the Detrex framework. All 12 transformer models and FRCNN use PyTorch [26] implementations, whereas SSD300 [19] and YOLOv3 [27] use Tensorflow [1]. We use mean average precision (mAP) to evaluate both benign and attacked performance. The Average Precision (AP) is given by interpolating the product of precision and recall at several decision thresholds. The mean AP (mAP) is the average of APs over the number of object classes. A lower mAP indicates that the detector is less effective for the object detection task. The victim detector’s mAP under adversarial perturbations reflects the degree of degradation in detection performance for the victim detector when compared with its benign mAP. We measure imperceptibility by reporting the average distortion of the attack via four metrics:  $L_2$  norm,  $L_0$  norm, semantic structural similarity index (SSIM), and mean distortion  $\mu_\Delta$ . We compute each metric per image and report the average over all images in COCO 2017 test-dev. Our experiments in the next section show that setting the total number of attack iterations to 10 is effective for all 12 state-of-the-art detection models on the COCO benchmark.

### 4.2. Effectiveness Comparison of AFOG Attack

Table 1 shows the effectiveness of the AFOG attack and its two special extensions AFOG-V and AFOG-F on each of the twelve detection transformers with varying model sizes (Col. 2). We use their benign mAP scores (Col. 3) as the reference for the evaluation. We make three observations: (i) under AFOG attacks all 12 transformers suffer a drastic performance drop in mAP scores, ranging from  $2.5\times$  reduction to  $37.8\times$  reduction. (ii) Compared to the AFOG generic attack, the special extension AFOG-V offers stronger attack success rates in terms of mAP reduction for 11 out of 12 transformers, except DETR-R50 [4]; and the special extension AFOG-F offers stronger attack success rates for 7 out of 12 transformers, except DETR-R50 [4], DETR-R101 [4], Deformable-DETR [38], R50 (DINO) [11], Swin-L (DINO) [20] (iii) For AlignDETR [2], AFOG-F achieves a mAP of 1.36 and outperforms both AFOG (mAP of 18.13) and AFOG-V (mAP of 1.64), and all three attacks significantly reduce the benign mAP by a factor ranging from  $2.8\times$  to  $37.8\times$ .

We compare AFOG against eleven benchmark attacks on DETR and Swin in Table 2. We observe that AFOG achieves superior performance on both DETR-R50 and Swin-L, notably recording an 82.7% improvement over the next strongest attack on Swin. AFOG also uses the smallest perturbation budget and fewest iterations, indicating its superior imperceptibility and speed.



Figure 4. Visualization with six examples, comparing benign scenario and AFOG attack results on Detection Transformer (DETR) with ResNet-50 backbone. Six diverse test images contain objects with varying classes, scales, perspectives, lighting, motion, and textures.

Table 1. AFOG effectiveness over 12 detection transformers, measured by mAP on perturbed images. (\*) indicates DINO framework used with the corresponding backbone for object detection.

Model	Params (M)	Benign	AFOG	AFOG-V	AFOG-F
DETR-R50 [4]	39.8	42.1	4.1	4.5	9.8
DETR-R101 [4]	76.0	43.5	5.2	5.1	11.3
Deform.-DETR [38]	40.0	44.5	4.8	1.5	7.1
R50* [11]	47.6	49.2	5.3	1.5	6.3
AlignDETR [2]	47.6	51.4	18.1	1.6	1.4
ViTDet* [14]	108.1	54.9	3.8	0.9	2.8
ConvNeXt* [21]	219.0	55.4	3.9	1.9	3.1
Swin-L* [20]	217.2	56.8	7.3	2.4	8.6
InternImage* [31]	241.0	56.9	7.3	2.8	5.1
FocalNet* [36]	228.9	58.5	7.3	2.5	5.1
EVA* [10]	1037.2	62.1	12.2	4.1	8.7
DETA [25]	218.8	62.9	25.6	3.7	4.3

Table 2. Comparison benchmark of AFOG against other state-of-the-art object detection attacks on DETR and Swin. Results are theirs. (\*) indicates results from [32]. (†) indicates results from [24]. (-) indicates no result.

Attack	Type	Pert. Budget	Iters.	Adversarial mAP	
				DETR-R50	Swin
GARSDC [17]	Surrogate	0.05	3000+	6.0	-
GALD [13]	Surrogate	0.063	10	20.6	-
RAD* [6]	Surrogate	0.063	10	27.2	47.2
GHFD* [32]	Surrogate	0.063	50	12.7	42.3
UEA* [33]	Surrogate	0.063	50	28.5	50.7
DAG* [35]	Surrogate	0.063	50	28.6	50.7
RAP* [15]	Surrogate	0.063	50	24.7	49.5
EBAD† [3]	Victim	0.039	10	34.9	-
AttentionFool [22]	Victim	-	10-150	21.0	-
OATB [12]	Victim	0.078	20	26.6	-
DBA [16]	Victim	-	50	-	56.7
<b>AFOG</b>	<b>Victim</b>	<b>0.031</b>	<b>10</b>	<b>4.1</b>	<b>7.3</b>

Figure 4 shows six visual examples and the comparison of DETR results for benign detection (Row 1) and adversarial detection under the AFOG attack (Row 2) scenarios. The first example shows that DETR detects the person and surfboard correctly under no attack (Row 1) but outputs erroneous detections under the AFOG attack (Row 2). Additional visual examples for all twelve transformers are available in the Supplementary Material.

We report the distortion effect and timing cost for each of the 12 transformer models in Table 3. We note several interesting observations: (i) Considering the  $L_2$  and SSIM metrics, AFOG attacks induce very similar levels of

distortion for 10 out of 12 transformer models. The two largest detection transformers, DETA [25] and EVA [10], show worse  $L_2$  and SSIM scores for AFOG, AFOG-V and AFOG-F. (ii) Considering the  $L_0$  metric, only DETA shows the lowest  $L_0$  value consistently for AFOG, AFOG-V and AFOG-F. (iii) As expected, the time required to attack a detection transformer model is strongly correlated to the number of model parameters. For instance, EVA [10] has over 1 billion parameters, and took much longer on average to attack an input image over 10 iterations compared to the other 11 models. (iii) DETA [25] has 218.8 million parameters and is similar to other models within the Detrex framework: FocalNet [36], InternImage [31], Swin-L [20] and ConvNext [21]. However, we observe that it consistently takes about  $1.5 - 2 \times$  longer time to attack one input image on average with our AFOG attack or its special case extensions AFOG-V or AFOG-F. (iv) AFOG, AFOG-V and AFOG-F have similar average perturbation magnitude.

### 4.3. Effectiveness on CNN-Based Detectors

This section compares AFOG with four attacks designed for CNN-based models: TOG [7], UEA [33], RAP [15], and DAG [35]. They are surrogate-based attacks, and a majority of them (e.g., DAG, RAP, UEA) can only directly attack two-stage detectors (exemplified by Faster-R-CNN [28]) and rely on adversarial transferability to attack single-stage CNN-based detectors like YOLOv3 [27] and SSD [19]. Table 4 reports the comparison results. We make two observations: (i) AFOG and TOG are the only two victim-based attacks against single-stage detectors, represented by YOLOv3 and SSD. Both AFOG and TOG can drastically reduce the benign mAP of YOLOv3 (83.43%) and SSD (76.11%). For YOLOv3, TOG is a stronger attack, achieving mAP of 0.56 compared to mAP of AFOG being 2.62. However, for SSD-300, AFOG is a stronger attack, achieving mAP of 0.50, compared to mAP of TOG (0.86). (ii) For two-stage CNN-based object detectors, exemplified by Faster-R-CNN, AFOG outperforms all four other victim-based attacks, achieving the lowest adversarial mAP score of 2.07%. (iii) AFOG also uses the same distortion mag-

Table 3. Timing and Imperceptibility Results.  $L_2$  represents the average  $L_2$  norm difference between perturbed and clean images,  $L_0$  is the average proportion of perturbed pixels, SSIM is the structural similarity index measure,  $\mu_\Delta$  is the average perturbation magnitude, and t is average total attack time for all ten iterations in seconds.

Model	$L_2$	AFOG				time	$L_2$	AFOG-V				time	$L_2$	AFOG-F				time
		$L_0$	SSIM	$\mu_\Delta$				$L_0$	SSIM	$\mu_\Delta$				$L_0$	SSIM	$\mu_\Delta$		
DETR-R50 [4]	0.0322	0.9707	0.8715	0.0173	1.45	0.0323	0.9707	0.8716	0.0172	0.99	0.0323	0.9710	0.8717	0.0172	1.21			
DETR-R101 [4]	0.0323	0.9707	0.8721	0.0173	1.47	0.0323	0.9706	0.8724	0.013	1.16	0.0323	0.9708	0.8724	0.0172	1.70			
Deform.-DETR [38]	0.0323	0.9719	0.8711	0.0174	1.63	0.0323	0.9713	0.8716	0.0173	1.63	0.0012	0.9714	0.8717	0.0173	1.87			
R50 [11]	0.0317	0.9658	0.8343	0.0171	2.70	0.0317	0.9650	0.8348	0.0170	2.34	0.0317	0.9654	0.8346	0.0170	3.16			
AlignDETR [2]	0.0319	0.9657	0.8347	0.0170	2.41	0.0319	0.9646	0.8349	0.0170	2.33	0.0319	0.9647	0.8349	0.0170	3.29			
ViTDet [14]	0.0318	0.9671	0.8353	0.0171	6.88	0.0318	0.9657	0.8361	0.0170	6.67	0.0318	0.9664	0.8355	0.0171	7.33			
ConvNext [21]	0.0318	0.9666	0.8342	0.0171	5.38	0.0318	0.9654	0.8349	0.0170	5.26	0.0318	0.9663	0.8347	0.0170	5.86			
Swin-L [20]	0.0327	0.9724	0.8673	0.0175	7.13	0.0327	0.9716	0.8680	0.0173	7.28	0.0327	0.9722	0.8678	0.0174	8.99			
InternImage [31]	0.0318	0.9665	0.8360	0.0170	6.35	0.0318	0.9653	0.8367	0.0170	6.23	0.0318	0.9660	0.8364	0.0171	6.70			
FocalNet [36]	0.0320	0.9665	0.8365	0.0172	8.96	0.0320	0.9657	0.8378	0.0171	8.84	0.0320	0.9659	0.8374	0.0171	9.74			
EVA [10]	0.0370	0.9666	0.8240	0.0172	54.34	0.0370	0.9665	0.8246	0.0171	54.53	0.0370	0.9665	0.8242	0.0171	51.18			
DETA [25]	0.0481	0.9662	0.8096	0.0170	13.20	0.0481	0.9654	0.8099	0.0170	13.10	0.0481	0.9654	0.8099	0.0170	13.13			

nitude budget  $L_\infty$  as TOG, while achieving higher attack success rate (lower mAP) on SSD-300 and FRCNN with decreased  $L_2$  distortion cost. Hence, AFOG excels at attacking both transformer-based detectors and CNN-based detectors.

Table 4. Comparing AFOG with four state-of-the-art attacks on representative CNN-based object detectors. mAPs of existing attacks were taken from respective papers [7]. (-) indicates N/A.

Model	Attack	mAP	t	Distortion Cost			
				$L_\infty$	$L_2$	$L_0$	SSIM
YOLOv3	Benign	83.43	0.0	0.0	0.0	0.0	1.0
	TOG [7]	<b>0.56</b>	<b>0.98</b>	0.031	0.083	0.984	<b>0.875</b>
	AFOG	2.28	1.31	0.031	<b>0.013</b>	<b>0.855</b>	0.801
SSD-300	Benign	76.11	0.0	0.0	0.0	0.0	1.0
	UEA [33]	20.0	-	-	-	-	-
	DAG [35]	64.0	-	-	-	-	-
	TOG [7]	<b>0.86</b>	<b>0.39</b>	0.031	<b>0.120</b>	<b>0.975</b>	<b>0.879</b>
	AFOG	<b>0.50</b>	0.49	0.031	<b>0.022</b>	<b>0.858</b>	0.793
FRCNN	Benign	67.37	0.0	0.0	0.0	0.0	1.0
	UEA [33]	5.0	0.17	0.343	0.191	0.959	0.652
	RAP [15]	4.78	4.04	0.082	0.010	0.531	0.994
	DAG [35]	3.56	7.99	<b>0.024</b>	<b>0.002</b>	<b>0.493</b>	<b>0.999</b>
	TOG [7]	2.64	<b>1.68</b>	0.031	0.058	0.976	0.862
	AFOG	<b>2.38</b>	2.11	0.031	0.019	0.854	0.788

#### 4.4. Effect of Learnable Self-Attention

We isolate and analyze the role of the learnable self-attention mechanism in empowering AFOG to successfully corrupt transformer-based models by probing for weak spots and performing fast and human-imperceptible perturbations. Figure 5 shows a comparison between AFOG with and without its learnable attention mechanism, measured by percent difference in adversarial mAP scores. Ablation results show that our learnable attention mechanism improves performance by up to 30.6% on InternImage, with an average of 15.1% improvement across all models.

Figure 6 shows the self-attention map for the last encoder layer of DETR [4] at initialization and three progressive attack iterations ( $3^{rd}$ ,  $4^{th}$ , and  $8^{th}$ ). We observe that, as the attack progresses in iterations (indicated by grey arrows), the self-attention map becomes increasingly less associated with the benign detected objects. By the  $8^{th}$  iteration of the AFOG attack, the self-attention maps have become almost

entirely disjoint from the benign detections of the victim model (DETR) prior to the attack. We note that, when the model’s self-attention map that reflects learned detection knowledge is changed, it is unlikely for prior knowledge to be kept intact. During the AFOG attack, the perturbed inputs focused by AFOG’s own adversarial attention corrupt the pixel-level relationships learned by a well-trained victim model prior to adversarial injection (e.g.,  $k = 1$ ), causing catastrophic forgetting to occur.

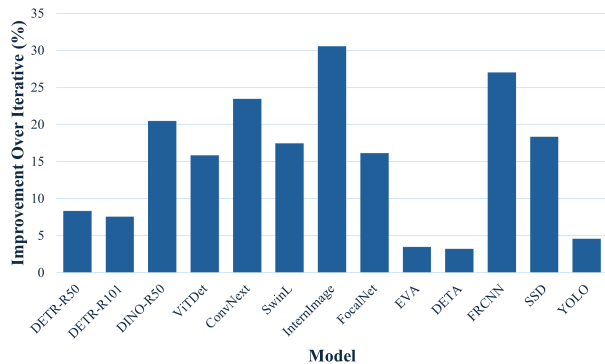


Figure 5. Improvement (%) difference) of AFOG with attention over AFOG without attention. The learnable attention mechanism improves performance by as much as 30.6% on InternImage, with an average improvement of 15.1% across all models.

#### 4.5. Worst Case Analysis

Our experimental results on 12 transformer-based detectors and 3 representative CNN-based detectors have shown that AFOG is a fast and effective attack. However, there are still cases where AFOG fails in its adversarial perturbations within a total of 10 iterations, one setting of our attack termination hyper-parameter. By investigating those cases where AFOG did not succeed, we found that the most common situation is where the attack fails to corrupt the central subject of the victim input image, e.g., a person. Figure 7 gives three visual examples, with the top one (Rows 1-2) being a successful attack followed by two failed cases (Rows 3-6) when attacking DETR-R50. Columns 2-3 in all six

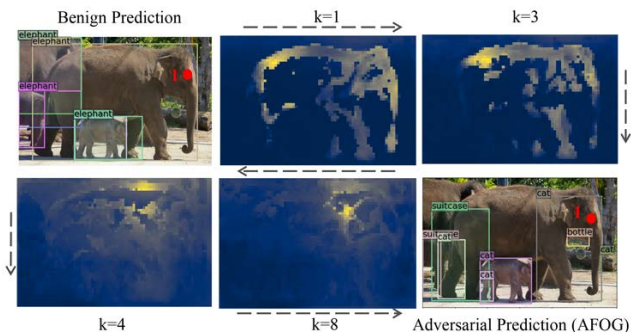


Figure 6. DETR self-attention weights analysis. Each blue frame shows the self-attention weights of DETR-R50’s last encoder layer at the red indicated point in the image at a given attack iteration  $k$ . In early iterations, the model understands a strong association between the indicated red point on the elephant and the rest of the elephant shape. As the attack iterates (grey arrows), this association becomes corrupted, evidenced by the lack of encoder self-attention structure in later stages. By the 8<sup>th</sup> iteration, the selected point has lost association with the rest of the elephant object, demonstrating that AFOG has disrupted learned knowledge.

rows show DETR’s encoder’s last layer’s self-attention map for the first and second locations indicated by red points in the first column. Rows 1, 3, and 5 show benign situations and their corresponding self-attention maps, and Rows 2, 4, 6 show the self-attention maps after the final attack iteration. The fourth column shows AFOG’s adversarial attention map after the final attack iteration.

We make three observations: (i) For the successful attack case (Rows 1-2), the self-attention maps of victim model DETR-R50 lose association between the indicated red point and the rest of the foreground object it comprises. (ii) In both unsuccessful cases (Rows 3-6), AFOG fails to disrupt this association. (iii) Consider AFOG’s adversarial attention maps from the successful case in Figure 7 (Row 2, Col. 4), we observe a clear focus on the key objects in the victim input image. In comparison, AFOG’s adversarial attention maps for the two failure cases (Row 4, Col. 4 and Row 6, Col. 4) both fail to focus on foreground objects. The first failure case (Rows 3 and 4) shows that the AFOG attention map attributes significant weight to the distant person in the left of the victim image, resulting in a large number of fabricated predictions in that region. Similarly, the attention map for the second victim image (Rows 5 and 6) appears to entirely miss the fire hydrant, resulting in a failure to disrupt this object. In these cases, AFOG’s adversarial attention has learned to focus on the wrong pixels.

## 5. Conclusion

We have presented AFOG, an Attention-Focused Offensive Gradient attack. AFOG is effective for attacking both advanced object detection transformers and traditional CNN-based detectors with a unified, architecture-agnostic framework. AFOG utilizes a learnable attention mecha-

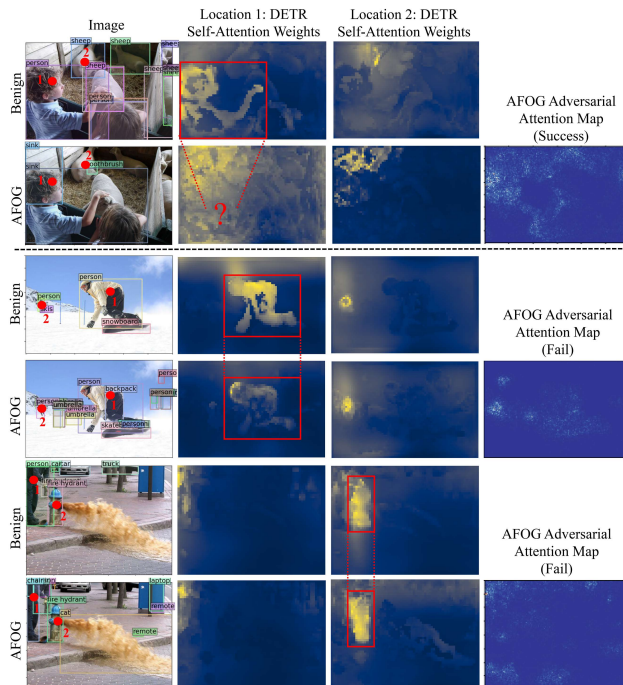


Figure 7. Worst case analysis with DETR-R50 as victim model. Rows 1-2 show a benign image and the successful AFOG perturbed image. Rows 3-6 show two cases where AFOG attack did not succeed. Columns 2-3 show DETR’s encoder’s self-attention maps for the two locations marked in red in column one. Column 4 shows AFOG’s internal attention map, where white pixels indicate high attention and blue indicates low. In the successful case (Rows 1-2), encoder self-attention maps (Cols. 2-3) lose associative structure. In the unsuccessful cases (Rows 4 and 6), AFOG fails to disrupt this structure and predictions are preserved. AFOG adversarial attention maps (Col. 4) show focus on important object structure in the successful case (Row 2) and lack of focus in the unsuccessful cases (Rows. 4, 6).

nism to enable its adversarial perturbations to focus on vulnerable areas of images in multi-box detection tasks. AFOG’s attack loss function integrates multiple feature losses (e.g., bounding-box loss, class loss) through learnable attention updates with iterative injection of adversarial perturbations. Finally, AFOG is efficient and stealthy. Adversarial perturbations generated by AFOG are visually imperceptible and yet can cause well-trained detectors to fail miserably. Extensive experiments on state-of-the-art object detectors show that AFOG is consistently effective across twelve object detection transformers. Comparative evaluation with nearly a dozen SOTA methods shows that AFOG significantly outperforms surrogate-based and victim-based attacks on both object detection transformers and CNN-based object detectors.

**Acknowledgement.** This research is partially sponsored by the NSF CISE grants 2302720 and 2312758, CISCO Edge AI program, and GTRI PhD Fellowship.

## References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. **5**
- [2] Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Improving detr with simple iou-aware bce loss, 2023. **5, 6, 7**
- [3] Zikui Cai, Yaoteng Tan, and M. Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4045–4055, 2023. **2, 6**
- [4] Nicolas Carrion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–223, 2020. **1, 5, 6, 7**
- [5] Sizhe Chen, Fan He, Xiaolin Huang, and Kun Zhang. Relevance attack on detectors. *Pattern Recognition*, 124:108491, 2022. **2**
- [6] Sizhe Chen, Fan He, Xiaolin Huang, and Kun Zhang. Relevance attack on detectors. *Pattern Recognition*, 124:108491, 2022. **2, 6**
- [7] Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Adversarial objectness gradient attacks in real-time object detection systems. In *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications*, pages 263–272. IEEE, 2020. **2, 6, 7**
- [8] Papers With Code. Object detection on coco test-dev. <https://paperswithcode.com/sota/object-detection-on-coco>. **1**
- [9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. **2, 5**
- [10] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. **5, 6, 7**
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Shu. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 06 2016. **5, 6, 7**
- [12] Zhuo Leng, Zesen Cheng, Pengxu Wei, and Jie Chen. Object-aware transfer-based black-box adversarial attack on object detector. In Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji, editors, *Pattern Recognition and Computer Vision*, pages 278–289, Singapore, 2024. Springer Nature Singapore. **2, 4, 6**
- [13] Tuo Li and Yahong Han. Improving transferable adversarial attack for vision transformers via global attention and local drop. *Multimedia Systems*, 29:3467 – 3480, 2023. **2, 6**
- [14] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection, 2022. **2, 6, 7**
- [15] Yuezun Li, Daniel Tian, Mingching Chang, Xiao Bian, and Siwei Lyu. Robust adversarial perturbation on deep proposal-based models. In *BMVC*, 2018. **6, 7**
- [16] Jiawei Lian, Shaohui Mei, Xiaofei Wang, Yi Wang, Lefan Wang, Yingjie Lu, Mingyang Ma, and Lap-Pui Chau. Attack anything: Blind dnns via universal background adversarial attack, 2024. **2, 4, 6**
- [17] Siyuan Liang, Longkang Li, Yanbo Fan, and Xiaojun Jia. A large-scale multiple-objective method for black-box attack against object detection. *European Conference on Computer Vision (ECCV) 2022*, pages 619–636, 2022. **2, 6**
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV 2014*, pages 740–755. Springer International Publishing, 2014. **2, 5**
- [19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. **1, 5, 6**
- [20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. **1, 2, 5, 6, 7**
- [21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. **6, 7**
- [22] Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Murnmadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15213–15222, 2022. **2, 6**
- [23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. **3**
- [24] Khoi Nguyen Tiet Nguyen, Wenyu Zhang, Kangkang Lu, Yuhuan Wu, Xingjian Zheng, Hui Li Tan, and Liangli Zhen. A Survey and Evaluation of Adversarial Attacks for Object Detection, Aug. 2024. arXiv:2408.01934 [cs]. **2, 6**
- [25] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back. *arXiv preprint arXiv:2212.06137*, 2022. **6, 7**

- [26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5
- [27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018. 1, 2, 5, 6
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1, 6
- [29] Tianhe Ren, Shilong Liu, Feng Li, Hao Zhang, Ailing Zeng, Jie Yang, Xingyu Liao, Ding Jia, Hongyang Li, He Cao, Jianan Wang, Zhaoyang Zeng, Xianbiao Qi, Yuhui Yuan, Jianwei Yang, and Lei Zhang. detrex: Benchmarking detection transformers, 2023. 5
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1
- [31] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 6, 7
- [32] Yang Wang, Yunfei Zheng, Lei Chen, Zhen Yang, Jingwei Wu, and Tiejong Cao. Gradient-guided hierarchical feature attack for object detector. *Journal of King Saud University - Computer and Information Sciences*, 36(1):101901, 2024. 2, 6
- [33] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 954–960. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 2, 4, 6, 7
- [34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 5
- [35] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie1, and Alan Yuill. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision*, pages 1378–1387, 2017. 6, 7
- [36] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022. 6, 7
- [37] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 5
- [38] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5, 6, 7