

# Understanding Object Detection Through An Adversarial Lens

Ka-Ho Chow, Ling Liu,  
Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu

Georgia Institute of Technology, Atlanta, GA, USA

{khchow, ling.liu}@gatech.edu,

{memregursoy, staceytruex, wenqiwei, yanzhaowu}@gatech.edu

**Abstract.** Deep neural networks based object detection models have revolutionized computer vision and fueled the development of a wide range of visual recognition applications. However, recent studies have revealed that deep object detectors can be compromised under adversarial attacks, causing a victim detector to detect no object, fake objects, or mislabeled objects. With object detection being used pervasively in many security-critical applications, such as autonomous vehicles and smart cities, we argue that a holistic approach for an in-depth understanding of adversarial attacks and vulnerabilities of deep object detection systems is of utmost importance for the research community to develop robust defense mechanisms. This paper presents a framework for analyzing and evaluating vulnerabilities of the state-of-the-art object detectors under an adversarial lens, aiming to analyze and demystify the attack strategies, adverse effects, and costs, as well as the cross-model and cross-resolution transferability of attacks. Using a set of quantitative metrics, extensive experiments are performed on six representative deep object detectors from three popular families (YOLOv3, SSD, and Faster R-CNN) with two benchmark datasets (PASCAL VOC and MS COCO). We demonstrate that the proposed framework can serve as a methodical benchmark for analyzing adversarial behaviors and risks in real-time object detection systems. We conjecture that this framework can also serve as a tool to assess the security risks and the adversarial robustness of deep object detectors to be deployed in real-world applications.

**Keywords:** Adversarial Robustness · Object Detection · Attack Evaluation Framework · Deep Neural Networks.

## 1 Introduction

Empowered by deep structures, nonlinear activation, and high-performance GPUs, deep neural networks (DNNs) have monopolized object detection systems [21,14,22], enabling the development of many security-critical applications, such as traffic sign detection on autonomous vehicles [23] and intrusion detection on surveillance systems [6]. While deep object detection algorithms offer real-time performance with high accuracy over traditional techniques [17,26], recent studies have revealed that well trained deep object detectors are vulnerable to adversarial inputs that are maliciously modified but visually imperceptible from the

Benign (No Attack)	Adversarial Attacks with Different Types of Attack Specificity			
	Untargeted Random	Object-vanishing	Object-fabrication	Object-mislabeling stopsign→umbrella
				
				

Table 1: Detection on two examples by the TOG family of attacks [2].

original benign input [2,27,12,28]. Table 1 illustrates such vulnerabilities. With no attack, the object detector can accurately identify the person, the car, and the stop sign on the two benign images (1st column). However, the *same* detector is fooled blindly by the adversarial examples (2nd-5th columns) that are perturbed malignantly but indistinguishable from the benign images by human-perception.

**1.1 Related Work and Problem Statement.** Object detection is the core task in computer vision, which takes an input image or video frame and detects multiple instances of semantic objects according to known categories [17,26]. Although some may view object detection as a generalization of the image classification task, a deep object detector is a multi-task learner and performs two unique learning tasks that make attacking object detection more complex and challenging than image classification: (1) Object detection should detect and identify instances of multiple semantic objects encapsulated in a single image or video frame, whereas a vanilla image classifier deals with the classification of each image into one of the known classes. (2) Object detection performs localization and classification of multiple instances of multiple semantic objects in a single image, and the localization accuracy of each instance may influence the classification accuracy of the instance. Thus, the adversarial attack techniques for image classifiers [10,24] are not applicable to attacking deep object detectors. The adversarial examples to attack object detection models are generated using more complex attack techniques, which compute and inject adversarial perturbations to the benign input by maximizing objectness loss, localization loss and classification loss simultaneously and iteratively [2,27].

Existing object detection models are broadly classified into two categories: (1) the proposal-based two-phase learning and (2) the regression-based single-phase learning. The proposal-based approach uses a two-phase procedure by first detecting proposal regions with a region proposal network (RPN), and then refining them with bounding box and class label prediction. This category is dominated by Faster R-CNN [22], and also includes R-CNN [7,8] and Mask R-CNN [11]. The regression-based single-phase learning formulates the detection

task as a regression problem. It jointly estimates the bounding box and class label of objects by directly predicting the coordinates of bounding boxes. This category is represented by YOLO [19,20,21,23] and SSD [14]. Moreover, different object detectors, even from the same family (e.g., Faster R-CNN), may use different neural networks as the backbone, and some additionally utilize different input resolutions [21,22] to optimize their detection performance. Several white-box attacks are developed to attack Faster R-CNN by utilizing proposal regions, such as DAG [28], UEA [27], and other similar methods [1,12]. For example, DAG first assigns an adversarial label (at random) to each proposal region detected and then performs iterative gradient backpropagation to misclassify the proposals. However, DAG attack with Faster R-CNN as the victim detector cannot be applied or extended to attacking single-phase detectors, which do not use proposal regions. Similar to the black-box transfer attacks to image classifiers [18], UEA [27] studied the transferability of attacks by using the adversarial examples generated from a Faster R-CNN detector to attack SSD detectors.

**1.2 Scope and Contribution.** In this paper, we develop an attack evaluation framework to rigorously analyze the vulnerabilities and security risks of deep object detection systems. The paper makes three original contributions. (1) We take a holistic approach to analyzing and characterizing adversarial attacks to object detection models from three dominant families: YOLOv3 [21], SSD [14], and Faster R-CNN [22,8], including attack generalization, untargeted random attacks, targeted specificity attacks, such as object-vanishing, object-fabrication, and targeted object-mislabeling. We develop the TOG family of attacks, which on one hand show the feasibility of attacking one-phase regression-based and two-phase proposal-based detectors using the same attack framework, and on the other hand provide a broader coverage of vulnerabilities for analyzing and understanding object detection through an adversarial lens. (2) Our evaluation framework provides two main building blocks: the attack module, which incorporates the state-of-the-art attack algorithms, and the evaluation module, which includes a set of quantitative metrics to measure, compare and analyze different attack algorithms in terms of adversarial effectiveness and costs, and attack transferability. We define cross-model transferability in terms of both algorithm and backbone of the detectors and introduce cross-resolution transferability to enrich our analysis on adversarial robustness of deep object detection models. (3) We conduct comprehensive experimental analysis on six object detectors from three dominant families of object detection algorithms (YOLOv3, SSD, and Faster R-CNN), with four representative attack methods: DAG [28], RAP [12], UEA [27], and TOG [2], on two benchmark datasets: PASCAL VOC [4] and MS COCO [13]. Our experimental results further demonstrate the utility of the proposed framework as a methodical benchmark platform for evaluating adversarial robustness of deep object detectors, and assessing the security risks and the attack resilience of deep object detectors to be deployed in real-world applications.

## 2 Proposed Framework - Attack Module

Figure 1 gives an overview of the proposed framework. This section is dedicated to the attack module, a collection of attack algorithms for comparisons and anal-

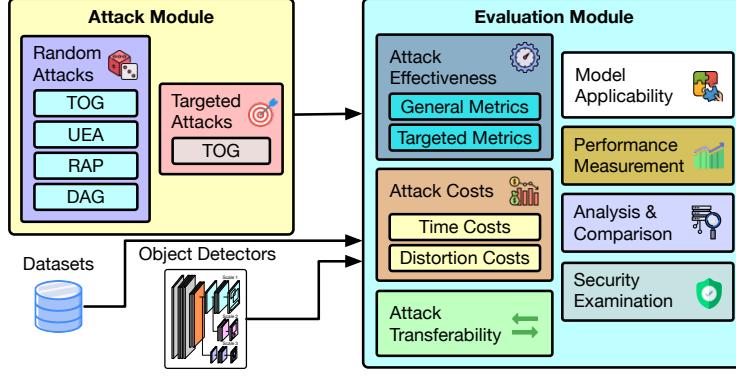


Fig. 1: The overview of the evaluation framework.

ysis. We first give an algorithmic overview of deep object detection algorithms and adversarial attacks. Then, we provide the formal analysis on the four state-of-the-art attack algorithms (TOG [2], UEA [27], RAP [12], and DAG [28]).

## 2.1 DNN-based Object Detection and Adversarial Attacks

DNN-based object detection is a multi-task learning problem, aiming to minimize the prediction error of (1) object existence, (2) bounding boxes, and (3) class labels of detected objects. Given an input image  $\mathbf{x}$  with resolution  $(H \times W)$ , a  $K$ -class object detector  $f_{\theta}$ , parameterized by  $\theta$ , generates a large number of  $S$  candidate objects  $\{\hat{\mathbf{o}}_1, \dots, \hat{\mathbf{o}}_S\}$  where  $\hat{\mathbf{o}}_i = (\hat{b}_i^x, \hat{b}_i^y, \hat{b}_i^w, \hat{b}_i^h, \hat{C}_i, \hat{\mathbf{p}}_i)$  represents a candidate centered at coordinates  $(\hat{b}_i^x, \hat{b}_i^y)$  having a dimension  $(\hat{b}_i^w, \hat{b}_i^h)$  with an objectness probability of  $\hat{C}_i \in [0, 1]$  to be a real object, and a  $K$ -class probability vector  $\hat{\mathbf{p}}_i = (\hat{p}_i^1, \hat{p}_i^2, \dots, \hat{p}_i^K)$ . This is often done by dividing the input into mesh grids in different scales (resolutions). Each grid cell is responsible for locating objects centered at the cell. The final detection results  $\hat{\mathcal{O}}$  are obtained by applying confidence thresholding to remove candidates with low prediction confidence and non-maximum suppression to exclude those with high overlapping.

To train a deep object detection neural network, every ground-truth object in a training sample  $\tilde{\mathbf{x}}$  is assigned to one of the  $S$  candidates according to their center coordinates. Let  $\mathcal{O}$  be the set of ground-truth objects of  $\tilde{\mathbf{x}}$ . The object detector can be trained by optimizing the following multi-task learning objective:

$$\mathcal{L}(\tilde{\mathcal{D}}; \theta) = \mathbb{E}_{(\tilde{\mathbf{x}}, \mathcal{O}) \in \tilde{\mathcal{D}}} [\mathcal{L}_{\text{obj}}(\tilde{\mathbf{x}}, \mathcal{O}; \theta) + \mathcal{L}_{\text{bbox}}(\tilde{\mathbf{x}}, \mathcal{O}; \theta) + \mathcal{L}_{\text{class}}(\tilde{\mathbf{x}}, \mathcal{O}; \theta)] \quad (1)$$

where  $\tilde{\mathcal{D}}$  is the training set,  $\mathcal{L}_{\text{obj}}$ ,  $\mathcal{L}_{\text{bbox}}$ , and  $\mathcal{L}_{\text{class}}$  represent the loss function of the three prediction tasks: object existence (objectness), object localization (bounding box), and object class label respectively. In the rest of this paper, we use  $\mathcal{O}$  and  $\hat{\mathcal{O}}$  to distinguish between ground-truth and predicted detection, and we only specify the argument (e.g.,  $\mathcal{O}(\mathbf{x})$ ) to emphasize the input if necessary.

An adversarial example  $\mathbf{x}'$  is generated by perturbing a benign input  $\mathbf{x}$  sent to the victim detector, aiming to fool the victim to misdetect randomly or purposefully. The generation process can be conceptually formulated as

$$\min ||\mathbf{x}' - \mathbf{x}||_p \quad s.t. \quad \hat{\mathcal{O}}(\mathbf{x}') \neq \hat{\mathcal{O}}(\mathbf{x}), \hat{\mathcal{O}}(\mathbf{x}') = \mathcal{O}^*(\mathbf{x}) \quad (2)$$

where  $p$  is the distance metric and  $\mathcal{O}^*(\mathbf{x})$  is the incorrect detection. Popular choices for the distance metric include the  $L_\infty$  norm, denoting the maximum change to any pixel, the  $L_2$  norm, computing the Euclidean distance, and the  $L_0$  norm, measuring the number of the pixels that are changed.

Although adversarial attacks on object detection systems are more sophisticated, adopting different formulations, they generally exploit gradients derived from one or multiple losses in Equation 1 (i.e.,  $\mathcal{L}_{\text{obj}}$ ,  $\mathcal{L}_{\text{bbox}}$ , and  $\mathcal{L}_{\text{class}}$ ). This allows the attack algorithm to meticulously inject perturbations to the input image, such that the tiny changes in input will be amplified throughout the forward propagation of the victim detector, and become large enough to alter one or more types of prediction results (i.e., object existence, bounding box, and class probability), depending on the composition of gradients. We analyze below the four representative attack algorithms on object detection systems, understanding their properties and demystifying their working principles.

## 2.2 TOG: Targeted Objectness Gradient Attacks

We develop the TOG family of attacks [2] based on an iterative gradient approach to obtain the malicious perturbation fooling the victim detector to give the desired erroneous detection. With a proper setting of the designated detection  $\mathcal{O}^*(\mathbf{x})$  and the attack loss  $\mathcal{L}^*$ , TOG can be generally formulated as:

$$\mathbf{x}'_{t+1} = \prod_{\mathbf{x}, \epsilon} [\mathbf{x}'_t - \alpha_{\text{TOG}} \Gamma(\nabla_{\mathbf{x}'_t} \mathcal{L}^*(\mathbf{x}'_t, \mathcal{O}^*(\mathbf{x}); \boldsymbol{\theta}))] \quad (3)$$

where  $\mathbf{x}'_t$  is the adversarial example at the  $t$ -th iteration,  $\prod_{\mathbf{x}, \epsilon} [\cdot]$  is the projection onto a hypersphere with a radius  $\epsilon$  centered at  $\mathbf{x}$  in  $L_p$  norm,  $\Gamma$  is a sign function, and  $\alpha_{\text{TOG}}$  is the attack learning rate. With this formulation, TOG allows adversaries to specify the effect imposed on victim's detection accuracy and correctness, including untargeted random attacks and three types of targeted specificity attacks: object-vanishing, object-fabrication, and targeted object-mislabeling.

*Untargeted* attacks fool the victim detector to *randomly* misdetect without targeting at any specific object. This class of attacks succeeds if the adversarial example fools the victim detector to give incorrect result of any form, such as having objects vanished, fabricated, or mislabeled randomly. TOG exploits gradients from both  $\mathcal{L}_{\text{obj}}$ ,  $\mathcal{L}_{\text{bbox}}$ , and  $\mathcal{L}_{\text{class}}$  and formulates the attack to be

$$\mathbf{x}'_{t+1} = \prod_{\mathbf{x}, \epsilon} [\mathbf{x}'_t + \alpha_{\text{TOG}} \Gamma(\nabla_{\mathbf{x}'_t} \mathcal{L}(\mathbf{x}'_t, \mathcal{O}(\mathbf{x}); \boldsymbol{\theta}))]. \quad (4)$$

As shown in the 2nd column in Table 1, the victim detector cannot identify any correct objects that were detected on benign inputs (1st column) but the exact effect varies across input images and attack algorithms.

*Object-vanishing* attacks *consistently* disable the victim detector to locate and recognize any object. TOG-vanishing utilizes gradients from  $\mathcal{L}_{\text{obj}}$  as it dominates the decision on object existences and formulates the attack as follows:

$$\mathbf{x}'_{t+1} = \prod_{\mathbf{x}, \epsilon} [\mathbf{x}'_t - \alpha_{\text{TOG}} \Gamma(\nabla_{\mathbf{x}'_t} \mathcal{L}_{\text{obj}}(\mathbf{x}'_t, \emptyset; \boldsymbol{\theta}))] \quad (5)$$

By targeting specifically at object-vanishing, this attack if successful will make the victim detector fail to detect any object as shown in the 3rd column in Table 1 where no object is detected in both examples.

*Object-fabrication* attacks *consistently* fool the victim to mistakenly recognize false objects. TOG-fabrication leverages gradients from  $\mathcal{L}_{\text{obj}}$  with formulation:

$$\mathbf{x}'_{t+1} = \prod_{\mathbf{x}, \epsilon} [\mathbf{x}'_t + \alpha_{\text{TOG}} \Gamma(\nabla_{\mathbf{x}'_t} \mathcal{L}_{\text{obj}}(\mathbf{x}'_t, \emptyset; \boldsymbol{\theta}))]. \quad (6)$$

This attack makes the victim to drastically increase the number of detected objects by introducing fake objects, as illustrated in the 4th column in Table 1.

*Targeted object-mislabeling* attacks *consistently* cause the victim detector to misclassify the objects detected on the input image by replacing their source class label with the maliciously chosen target class label, while maintaining the same set of correct bounding boxes. By focusing on the classification loss (i.e.,  $\mathcal{L}_{\text{class}}$ ) and keeping the gradients of the other two parts unchanged, TOG-mislabeling assigns the target class label to each object in  $\mathcal{O}(\mathbf{x})$  to form  $\mathcal{O}^*(\mathbf{x})$  and generate adversarial examples with

$$\mathbf{x}'_{t+1} = \prod_{\mathbf{x}, \epsilon} [\mathbf{x}'_t - \alpha_{\text{TOG}} \Gamma(\nabla_{\mathbf{x}'_t} \mathcal{L}(\mathbf{x}'_t, \mathcal{O}^*(\mathbf{x}); \boldsymbol{\theta}))]. \quad (7)$$

For instance, the object-mislabeling attack in the 5th column in Table 1 is configured to fool the victim to mislabel any stop sign as an umbrella. Note that the person (top) and the car (bottom) can still be detected under this attack as they are not the objects of attack interest and only stop signs will be mislabeled.

As TOG does not attack a special structure (e.g., RPN) in an object detector, it is applicable to both one-phase and two-phase techniques. Inspired by the universal perturbations to attack image classifiers [16], TOG also develops universal perturbations to attack deep object detectors in terms of object-vanishing or object-fabrication attack [2]. By training the universal perturbation offline on a training set and a victim detector, the universal perturbation can be applied during the online detection phase to any input sent to the victim.

### 2.3 DAG: Dense Adversary Generation

DAG [28] is an untargeted random attack and begins with manually assigning the IOU threshold to 0.90 in non-maximum suppression (NMS) in the RPN of a given two-phase model. This attack setting requires one proposal region to be highly overlapped ( $> 90\%$ ) with the other proposal region in order to be pruned. Hence, a large amount of proposal regions remain unpruned. After the refinement by the subsequent network for bounding box and class label prediction, DAG assigns a randomly selected label for each proposal region and then performs the iterative gradient attack to misclassify the proposals with the following formulation:

$$\mathbf{r}_t = \nabla_{\mathbf{x}'_t} \sum_{j=1}^J z_j [p_j^c - p_j^{c'}], \quad \mathbf{x}'_{t+1} = \mathbf{x}'_t - \frac{\alpha_{\text{DAG}}}{\|\mathbf{r}_t\|_\infty} \mathbf{r}_t \quad (8)$$

where  $z_j = 1$  if the  $j$ -th proposal on  $\mathbf{x}'_t$  from RPN is foreground and 0 otherwise,  $p_j^c$  and  $p_j^{c'}$  are the prediction confidence of the correct class  $c$  and randomly selected incorrect class  $c'$  of the  $j$ -th proposal and  $\alpha_{\text{DAG}}$  is the attack learning rate. This is equivalent to exploiting gradients derived from the classification loss  $\mathcal{L}_{\text{class}}$ . As DAG requires to manipulate the RPN to generate a large number of proposals, it can only be directly applicable to two-phase detection models.

## 2.4 RAP: Robust Adversarial Perturbation

RAP [12] is an untargeted random attack and focuses on collapsing the function of the RPN in two-phase algorithms. It exploits the composite gradients from (i) the objectness loss, i.e.,  $\mathcal{L}_{\text{obj}}$ , that fools the RPN to not returning foreground objects, and (ii) the localization loss, i.e.,  $\mathcal{L}_{\text{bbox}}$ , that causes the bounding box estimation to be incorrect even if foreground objects are proposed:

$$\mathbf{r}_t = \nabla_{\mathbf{x}'_t} \sum_{j=1}^J z_j [\log(\hat{C}_j) + \ell_{\text{SE}}(\hat{\mathbf{b}}_j, \boldsymbol{\tau})], \quad \mathbf{x}'_{t+1} = \mathbf{x}'_t - \frac{\alpha_{\text{RAP}}}{\|\mathbf{r}_t\|_2} \mathbf{r}_t \quad (9)$$

where  $\ell_{\text{SE}}$  is the squared error,  $\hat{\mathbf{b}}_j$  and  $\boldsymbol{\tau}$  are quadruples of the proposed bounding box and large offsets respectively, and  $\alpha_{\text{RAP}}$  is the attack learning rate.

## 2.5 UEA: Unified and Efficient Adversary

UEA [27] is an untargeted random attack. It trains a conditional generative adversarial network (GAN) [9] to craft adversarial examples. In deep object detectors, the backbone network plays an important role in feature extraction for region proposals in two-phase algorithms or object recognition in one-phase techniques. In practice, it is often one of the popular architectures (e.g., VGG16) that perform well in large-scale image classification and is pretrained with the ImageNet dataset for transfer learning. UEA designs a multi-scale attention feature loss, encouraging the GAN to create adversarial examples that can corrupt the feature map extracted by the backbone network in the victim detector:

$$\mathcal{L}_{\text{UEA}}^{\text{Fea}} = \mathbb{E}_{(\tilde{\mathbf{x}}, \mathcal{O}) \in \mathcal{D}} \left[ \sum_{m=1}^M \|\mathbf{A}_m \circ (\tilde{\mathbf{x}}_m - \mathbf{R}_m)\|_2 \right] \quad (10)$$

where  $\tilde{\mathbf{x}}_m$  is the extracted feature map of the training example  $\tilde{\mathbf{x}}$  in the  $m$ -th layer of the backbone network,  $\mathbf{R}_m$  is a randomly predefined feature map, and  $\mathbf{A}_m$  is the attention weight computed based on the proposal regions from the RPN. Whenever another detector is equipped with the same backbone, the adversarial examples are likely to be effective. Equation 10 is jointly optimized with the DAG formulation (Equation 8), requiring the manipulation of the RPN. Hence, it is unable to directly attack one-phase algorithms.

## 3 Proposed Framework - Evaluation Module

The evaluation module is the second building block of the proposed framework (Figure 1), providing experimental testbed to measure, evaluate and analyze attacks and adversarial robustness of an object detector from four perspectives.

### 3.1 Attack Effectiveness

**mean Average Precision (mAP).** The interpolated average precision (AP) has been used by major object detection competitions [4,13]. For a given class, the precision/recall curve is computed from the detector's output, ranked by the detected confidence. The AP summarizes the shape of the precision/recall curve by taking the mean precision at a set of equally spaced recall levels. Then, the mean Average Precision (mAP) that quantifies the overall detection quality of a detector is computed by taking the mean of APs of all classes. The general

attack performance can be analyzed on two sets of mAP (or AP), one on benign examples and another on adversarial examples. A low adversarial mAP implies the power of the attack but reveals the vulnerability of the victim model.

**Attack Success Rate (ASR).** In addition to comparing mAPs to reveal the impact on overall performance of the victim, we further define the attack success rate (ASR) for each targeted specificity attack, to capture their capability to fool the victim to misbehave with the designated effect (e.g., object-vanishing).

For object-vanishing attacks, we define the ASR as the proportion of objects detected on benign examples that are not covered by any objects detected on their adversarial counterparts:

$$\text{ASR} = \frac{\sum_{\mathbf{x} \in \mathcal{D}} \sum_{\hat{\mathbf{o}} \in \hat{\mathcal{O}}(\mathbf{x})} \mathbb{1}[\neg \exists \hat{\mathbf{o}}' \in \hat{\mathcal{O}}(\mathbf{x}') (\text{IOU}(\hat{\mathbf{o}}_{[\text{bbox}]}, \hat{\mathbf{o}}'_{[\text{bbox}]}) \geq t_{\text{IOU}})]}{\sum_{\mathbf{x} \in \mathcal{D}} \|\hat{\mathcal{O}}(\mathbf{x})\|}, \quad (11)$$

where  $\mathbb{1}[\text{condition}] = 1$  if the condition is met and 0 otherwise,  $\text{IOU}(\hat{\mathbf{o}}_{[\text{bbox}]}, \hat{\mathbf{o}}'_{[\text{bbox}]})$  computes the intersection over union of the two bounding boxes  $\hat{\mathbf{o}}_{[\text{bbox}]}$  and  $\hat{\mathbf{o}}'_{[\text{bbox}]}$ , and  $t_{\text{IOU}}$  is a predefined threshold controlling the amount of overlapping required for two bounding boxes to be considered as referring to the same entity.

For object-fabrication attacks, the ASR is defined as the proportion of test examples where additional false objects are mistakenly detected by the victim detector under attacks:

$$\text{ASR} = \frac{1}{\|\mathcal{D}\|} \sum_{\mathbf{x} \in \mathcal{D}} \mathbb{1}[\|\hat{\mathcal{O}}(\mathbf{x}')\| > \|\hat{\mathcal{O}}(\mathbf{x})\|]. \quad (12)$$

For object-mislabeling attacks, we define the ASR to be the proportion of objects detected on benign examples that are mislabeled as the target label by the victim detector on their adversarial counterparts:

$$\text{ASR} = \frac{\sum_{\mathbf{x} \in \mathcal{D}} \sum_{\hat{\mathbf{o}} \in \hat{\mathcal{O}}(\mathbf{x})} \mathbb{1}[\exists \hat{\mathbf{o}}' \in \hat{\mathcal{O}}(\mathbf{x}') (\text{IOU}(\hat{\mathbf{o}}_{[\text{bbox}]}, \hat{\mathbf{o}}'_{[\text{bbox}]}) \geq t_{\text{IOU}} \wedge \hat{\mathbf{o}}'_{[\text{class}]} = \mathcal{T}(\hat{\mathbf{o}}_{[\text{class}]})])]}{\sum_{\mathbf{x} \in \mathcal{D}} \|\hat{\mathcal{O}}(\mathbf{x})\|} \quad (13)$$

where  $\mathcal{T}(\hat{\mathbf{o}}_{[\text{class}]})$  is a mapping from a source class to a target class. Under this setting, we consider the attack succeeds only if it (i) does not alter the bounding box significantly and (ii) fools the detector to give a designated wrong label.

### 3.2 Attack Cost

**Time Cost.** We measure time cost using two metrics: (i) the attack time, which measures the additional time introduced by the attack, excluding the inference of the victim detector to obtain the final detection results; and (ii) the total time cost, which considers both attack time and (benign) detection time.

**Distortion Cost.** Remaining human-imperceptible is an important factor in adversarial attacks as significant distortion naturally mislead a deep learning model to misbehave. A robust object detection model should be resilient against adversarial examples that are visually identical to their benign counterparts.

$L_0$ ,  $L_2$ , and  $L_\infty$  distances have been popularly used in adversarial learning. They are used as a constraint to limit the maximum perturbation introducible to the benign example. Note that a low  $L_p$  distance means a high imperceptibility.

Structural Similarity (SSIM) has become an important metric to quantify the similarly between two images in computer vision:

$$\text{SSIM}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^I \frac{(2\mu_{\mathbf{x}[i]}\mu_{\mathbf{x}'[i]} + \kappa_1)(2\sigma_{\mathbf{x}[i]\mathbf{x}'[i]} + \kappa_2)}{(\mu_{\mathbf{x}[i]}^2 + \mu_{\mathbf{x}'[i]}^2 + \kappa_1)(\sigma_{\mathbf{x}[i]}^2 + \sigma_{\mathbf{x}'[i]}^2 + \kappa_2)} \quad (14)$$

where  $\mathbf{x}[i]$  denotes the  $i$ -th channel of image  $\mathbf{x}$ ,  $\mu_{\mathbf{x}}$  and  $\sigma_{\mathbf{x}}$  are the average and variance of  $\mathbf{x}$  respectively,  $\sigma_{\mathbf{x}\mathbf{x}'}$  is the covariance of  $\mathbf{x}$  and  $\mathbf{x}'$ , and  $\kappa_1$  and  $\kappa_2$  are two variables for numerical stability. It has a range from  $-1.00$  (the least similar) to  $1.00$  (the most similar) and is considered to be more consistent to human visual perception than  $L_p$  distances. As attacks optimize different  $L_p$  distances, SSIM offers an objective comparison on the imperceptibility of adversarial examples.

### 3.3 Attack Transferability

All adversarial attacks on deep object detectors are white-box attacks as they require model weights to optimize the generation of adversarial perturbation against a victim detector. The transferability of adversarial examples generated against one victim detector can be utilized to launch black-box attacks to other detectors, in a similar way as the transferability of adversarial examples to attack different image classifiers [18]. For object detection, we propose to study not only the cross-model transferability, but also the cross-resolution transferability.

*Cross-model transferability* in object detection can be further broken down into (i) cross-algorithm transferability that the source and the target models use different detection algorithms and (ii) cross-backbone transferability that examines the transferability between different backbones of the same detection algorithm and between different detection algorithms with the same backbone.

*Cross-resolution transferability* covers a characteristic unique to those object detection algorithms (e.g., YOLO and Faster R-CNN) that allow variable input resolutions. In contrast to image classification networks where the resolution of the input image is fixed due to the fully-connected layer for the final softmax, for object detection, increasing input resolution can generate more candidate objects with a potentially better detection quality with the cost of slowing down the detection. The cross-resolution transferability reveals whether the adversarial examples generated by an attack algorithm on a source resolution can be robust and survive under resizing and interpolation to the target resolution.

### 3.4 Model Applicability

From a macroscopic perspective, all object detection systems take an input image and output a set of detected objects. They may appear to be similar, but their internal learn-to-detect mechanisms can be very different. Some existing attacks are designed by exploiting the vulnerability of a particular structure, e.g., the region proposal network (RPN) in Faster R-CNN detectors. Hence, not all attack techniques are universally applicable. RAP [12] is an example, which perturbs the benign image to disable the functionality of the RPN in two-phase algorithms and cannot be used on one-phase detectors where no RPN is used. We also leverage model-applicability as an evaluation aspect on attack algorithms.

## 4 Experimental Analysis

Extensive experiments are conducted on two benchmark datasets: PASCAL VOC [4] and MS COCO [13]. All results are based on the entire test set, and

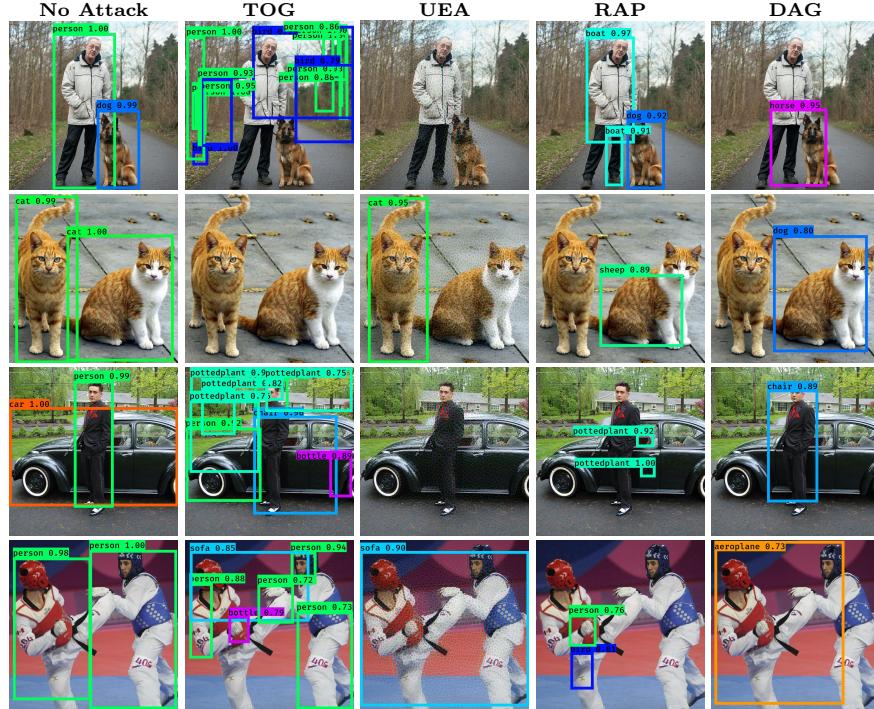


Fig. 2: Four visual examples of the untargeted attacks by different algorithms.

we preprocess images by padding to preserve the aspect ratio of objects. We consider six models from three dominant detection algorithms. YOLOv3-D and YOLOv3-M are two YOLOv3 [21] models with a Darknet53 and a MobileNetV1 backbone respectively. For SSD [14], we have SSD300 and SSD512 corresponding to two models with different input resolutions. Finally, FRCNN denotes the Faster R-CNN [22] model. As experimental results on COCO are highly similar to VOC, we provide only YOLOv3-D on COCO due to the space constraint. We provide more experimental configuration details in Appendix A.

#### 4.1 Untargeted Random Attacks

This section reports the set of experiments to compare the four attack algorithms: TOG, UEA, RAP, and DAG in terms of effectiveness and time cost of untargeted attacks. Figure 2 provides a visualization of four benign images (left most column) and their four adversarial examples generated by TOG, UEA, RAP, and DAG. Four attack algorithms fool the same victim detector FRCNN to misdetect on the same query image in different ways. TOG deceives the victim detector to return false objects on the 1st, 3rd and 4th examples with no correct objects detected. For the 2nd example with two cats, TOG succeeds by fooling the victim to detect no object at all. This shows that different images may respond to the same attack differently, such as missing cats by TOG in the 2nd example compared with fabricating fake objects in the other examples. Similarly, UEA misses both the person and the dog for the 1st example, detects one

Dataset	Random Attack	Victim Detector	mAP (%)		Time Cost (s)		Distortion Cost			
			Benign	Adv.	Benign	Adv.	$L_\infty$	$L_2$	$L_0$	SSIM
VOC	TOG	YOLOv3-D	83.43	0.56	0.03	0.98	0.031	0.083	0.984	0.875
VOC	TOG	YOLOv3-M	71.84	0.43	0.02	0.59	0.031	0.083	0.978	0.876
VOC	TOG	SSD300	76.11	0.86	0.02	0.39	0.031	0.120	0.975	0.879
VOC	TOG	SSD512	79.83	0.74	0.03	0.69	0.031	0.070	0.974	0.869
VOC	TOG	FRCNN	67.37	2.64	0.14	1.68	0.031	0.058	0.976	0.862
VOC	UEA	FRCNN	67.37	18.07	0.14	0.17	0.343	0.191	0.959	0.652
VOC	RAP	FRCNN	67.37	4.78	0.14	4.04	0.082	0.010	0.531	0.994
VOC	DAG	FRCNN	67.37	3.56	0.14	7.99	0.024	0.002	0.493	0.999
COCO	TOG	YOLOv3-D	54.16	3.52	0.03	1.02	0.031	0.083	0.986	0.872

Table 2: Untargeted attacks on different datasets and victim detectors.

cat correctly and misses the other cat on the 2nd example, misses both person and car for the 3rd example, and misdetect all objects on the 4th example. RAP and DAG fail the detection on all four examples differently.

Table 2 provides the quantitative measurements on all victim detectors under the four attack algorithms. The first metric is the mAP in percentage, including benign mAP with no attacks and adversarial mAP given adversarial examples. The second metric measures the detection time on benign inputs and attack total cost (both generation and detection). The third metric is the distortion cost measured in  $L_\infty$ ,  $L_2$ ,  $L_0$  distances, and SSIM.  $L_2$  and  $L_0$  costs reported here are normalized by the number of pixels and the  $L_2$  cost has a magnitude of  $10^{-3}$ . Note that UEA, RAP, and DAG can only attack FRCNN, and hence we do not evaluate them on YOLOv3, SSD300 and SSD512. We make two observations from Table 2. First, all attacks successfully bring down the mAP of the victim. Considering the TOG attack, the benign mAP of any victim detector is drastically reduced to less than 3.52% with four victims having a close to zero adversarial mAP. This indicates that the victims fail miserably with no detection capability. Second, we compare four different attacks on FRCNN, which has a benign mAP of 67.37%. TOG is the most powerful attack with the lowest adversarial mAP of 2.64%, followed by DAG (3.56%), RAP (4.78%), and UEA (18.07%). By default, UEA generates adversarial examples with a fixed resolution of  $300 \times 300$ . When attacking FRCNN taking inputs with resolution of  $600 \times 600$ , resizing and interpolation are required. Hence, the effectiveness of UEA is hindered. In comparison, TOG, RAP and DAG are much more adaptive, and capable of generating adversarial examples that fit the input resolution, as they do not rely on additional networks.

Apart from attack effectiveness, attack costs are equally important. UEA has the lowest time cost with only 0.17s attack total time because the generation of adversarial examples does not use the victim model but the GAN, which can have much lower complexity. TOG has a reasonable range of attack total time but RAP and DAG have prohibitively high time cost (4.04s and 7.99s). This can be explained by the number of iterations required to succeed the attack in RAP and DAG. TOG needs 10 iteration while RAP and DAG need to run more than 30 rounds. Interestingly, spending more iterations allows RAP and DAG to have a much lower distortion cost and exceptionally high SSIM measures of 0.994 and 0.999 respectively. TOG also has a high imperceptibility with SSIM higher than 0.862, while adversarial perturbation generated by UEA is significantly more

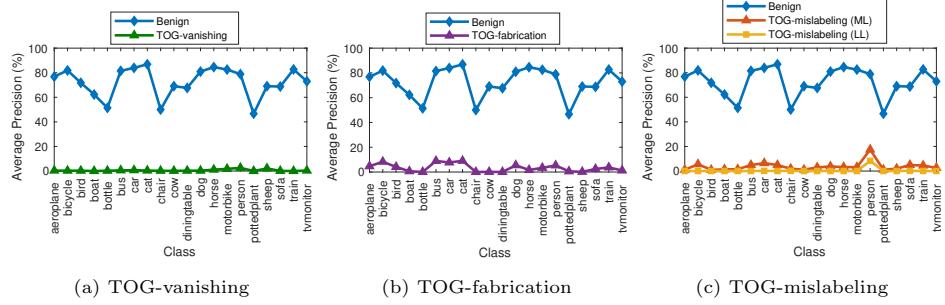


Fig. 3: The AP of each class under TOG targeted attacks on YOLOv3-M

perceptible, having a low SSIM of 0.652. Furthermore, RAP and DAG have a low  $L_0$  cost, which implies their perturbations are more localized. In comparison, both TOG and UEA have the  $L_0$  cost close to 1.000, indicating that most pixels are modified by the adversarial perturbation.

#### 4.2 Targeted Specificity Attacks

We evaluate the three targeted specificity attacks using TOG. For targeted mislabeling attacks, without loss of generality, we choose two representative attack targets: the most-likely (ML) and the least-likely (LL), which correspond to the incorrect class label of an object detected on benign example with the highest and the lowest prediction confidence respectively [3]. The TOG-mislabeling allows objects of any class to be attacked. Figure 3 shows the benign and adversarial AP of each class on YOLOv3-M. All targeted attacks by TOG drastically reduce the average precision of *every* class supported by the victim to almost zero, showing the severity of the targeted attacks. We provide more experimental measurements on all 24 cases (four attacks on six detectors) in Appendix B.

Recall Figure 2, each of the four input images responds to the same untargeted random attack differently. Figure 4 provides a visualization of the same set of images attacked by TOG with different targeted specificity effects. This qualitatively validates that all targeted attacks in TOG are goal-driven, which can be more detrimental to victim detector. For example, with TOG-vanishing attack (2nd column), all four adversarial examples fool the victim detector FRCNN to misdetect with no object recognized. For TOG-mislabeling attacks, the person and the dog on the 1st row are purposefully mislabeled as the dog and the cat respectively in the ML case and both aeroplanes in the LL case. In comparison with Figure 2, UEA, RAP, DAG and general TOG are untargeted: each of the four input images responds to attacks under the same attack algorithm (be it TOG, UEA, RAP and DAG) quite differently, showing random ways to fool a victim detector. We provide more experimental analysis on each targeted attack in Appendix B.

#### 4.3 Transferability of Attacks

We conduct quantitative analysis on the transferability of all four untargeted attacks: TOG, UEA, RAP and DAG. Table 3 reports the results for the cross-model transferability, measured in adversarial mAP. Using the same model to

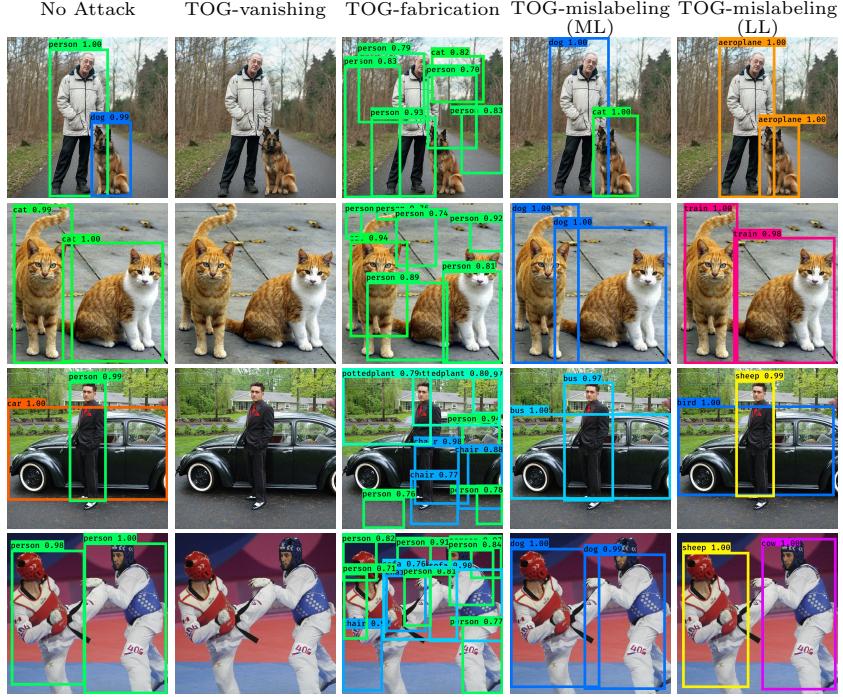


Fig. 4: Four visual examples of different targeted specificity attacks by TOG.

craft adversarial examples always achieves the highest transferability, as indicated in boldface. We first consider the adversarial examples generated on different source models and measure their transferability to different target models using TOG (the 2nd-6th rows). First, we observe that having the same backbone architecture does not necessarily lead to high transferability. FRCNN, SSD300 and SSD512 all use VGG16 as the backbone network. Yet, the adversarial examples generated on FRCNN have very low transferability to SSD300 and SSD512, reducing their mAP from 76.11% to 75.80% and from 79.83% to 78.09% respectively. Second, the adversarial examples generated on SSD have relatively higher transferability compared to other source models. For instance, adversarial examples from SSD300 and SSD512 can reduce the mAP of YOLOv3-D from 83.43% to 56.87% and 56.21%, much better than YOLOv3-M and FRCNN that only reduction to 74.62% and 79.47% are recorded. Finally, considering the transferability of different attack algorithms with the same source model FRCNN (the last four rows), we find that adversarial examples by UEA exhibit a higher transferability consistently. This can be attributed to its high distortion cost incurred to perturb each adversarial example (recall Table 2).

Table 4a and Table 4b report the cross-resolution transferability on FRCNN and YOLOv3 respectively. Note that only TOG can directly attack YOLOv3 (one-phase detectors), and SSD does not support variable input resolutions. We use nearest neighbor interpolation during resizing as we find empirically that it can better preserve the malicious pattern. For victim detector FRCNN, we

Transfer Attack	Source Model	Target Model				
		YOLOv3-D	YOLOv3-M	SSD300	SSD512	FRCNN
Benign (No Attack)		83.43	71.84	76.11	79.83	67.37
TOG	YOLOv3-D	<b>0.56</b>	60.13	72.70	73.86	55.57
TOG	YOLOv3-M	74.62	<b>0.43</b>	73.27	75.27	59.1
TOG	SSD300	56.87	42.85	<b>0.86</b>	38.79	50.36
TOG	SSD512	56.21	46.00	58.00	<b>0.74</b>	35.98
TOG	FRCNN	79.47	68.60	75.80	78.09	<b>2.64</b>
UEA	FRCNN	51.92	31.88	47.08	47.66	<b>18.07</b>
RAP	FRCNN	81.80	69.45	75.77	76.84	<b>4.78</b>
DAG	FRCNN	81.21	70.37	75.15	78.38	<b>3.56</b>

Table 3: Cross-model transferability.

Transfer Attack	Source Resolution	Target Resolution					
		300x300	400x400	500x500	600x600	700x700	800x800
Benign (No Attack)		65.33	67.85	68.00	67.37	67.91	67.76
TOG	600x600	50.15	29.50	15.07	<b>2.64</b>	6.84	3.86
UEA	300x300	<b>3.86</b>	11.88	18.61	18.07	16.32	17.34
RAP	600x600	58.45	54.32	56.96	<b>4.78</b>	53.21	50.12
DAG	600x600	62.89	59.82	46.58	<b>2.84</b>	30.96	13.75

(a) FRCNN						
Model	Transfer Attack	Source Resolution	Target Resolution			
			352x352	384x384	416x416	448x448
YOLOv3-D	Benign (No Attack)		82.71	83.25	83.43	83.63
	TOG	416x416	25.26	14.93	<b>0.56</b>	11.02
YOLOv3-M	Benign (No Attack)		69.98	71.13	71.84	73.10
	TOG	416x416	33.41	20.61	<b>0.43</b>	15.62

(b) YOLOv3						
Model	Transfer Attack	Source Resolution	Target Resolution			
			352x352	384x384	416x416	448x448
YOLOv3-D	Benign (No Attack)		82.71	83.25	83.43	83.63
	TOG	416x416	25.26	14.93	<b>0.56</b>	11.02
YOLOv3-M	Benign (No Attack)		69.98	71.13	71.84	73.10
	TOG	416x416	33.41	20.61	<b>0.43</b>	15.62

Table 4: Cross-resolution transferability.

observe that TOG and UEA have higher cross-resolution transferability than RAP and DAG. The same observation can be made in both YOLOv3 detectors. For instance, TOG can still effectively reduce the mAP from more than 82% to less than 26% in all target resolutions evaluated on YOLOv3-D. This is because adversarial examples generated by TOG and UEA have a higher robustness under resizing and interpolation to fit the target resolution. Also, upsizing to a higher target resolution is always better than downsizing, causing a higher mAP drop in the target victim model, which can be explained by the fact that downsizing loses the fine details of malicious perturbation.

Table 5 provides a visualization to illustrate the transferability of four TOG targeted attacks by generating adversarial examples on SSD300 and evaluating their cross-model transferability to the other three detectors: SSD512, YOLOv3-D, and YOLOv3-M. Consider the SSD300 row, the detector can correctly identify the person and the bicycle on the benign input (1st column). The targeted attacks by TOG successfully fool the victim to misdetect with designated attack specificity effects: the two objects are missed in TOG-vanishing, false objects are detected in TOG-fabrication, and the person and the bicycle are mislabeled as a dog and a horse in the ML case of TOG-mislabeling and both buses in the LL case. We analyze the transferability by observing the other three rows. Given that all three detectors can successfully identify the two objects on the benign

Detection results under four TOG targeted attacks					
	Benign (No Attack)	TOG-vanishing	TOG-fabrication	TOG-mislabeling (ML)	TOG-mislabeling (LL)
SSD300					
Detection results transferred from SSD300 to other victim detectors					
	Benign (No Attack)	TOG-vanishing	TOG-fabrication	TOG-mislabeling (ML)	TOG-mislabeling (LL)
SSD512					
YOLOv3-D					
YOLOv3-M					

Table 5: Transferring targeted attacks on SSD300 to three other detectors.

image, we find different degrees of adversarial transferability. For instance, TOG-vanishing and TOG-fabrication can be successfully transferred to SSD512, which has the same backbone (i.e., VGG16) and detection algorithm as the source detector SSD300. TOG-fabrication can also be transferred to YOLOv3-M with the same effect. However, even some adversarial examples may fool other detectors (e.g., TOG-vanishing to YOLOv3-M), they fail in transferring attacks with the same effect. Note that with adversarial transferability, the attacks are black-box, generated and launched without any prior knowledge of the three victim detectors. We provide more discussion in Appendix C.

#### 4.4 Model Applicability and Physical Attacks

We provide a comparison of seven representative attack algorithms, including two physical attacks, to deep object detectors in Table 6.

TOG [2], UEA [27], RAP [12] and DAG [28] are the representative digital attacks against a victim detector by perturbing pixel values of a benign image while maximizing one or more of the three loss functions: objectness, bounding box, and classification. All four can perform untargeted random attacks,

	Attack Effect				Model-applicability		
	Random	Object-vanishing	Object-fabrication	Object-mislabeling	Two-phase FRCNN	One-phase YOLO	SSD
TOG [2]	✓	✓	✓	✓	✓	✓	✓
UEA [27]	✓	✗	✗	✗	✓	✗	✗
RAP [12]	✓	✗	✗	✗	✓	✗	✗
DAG [28]	✓	✗	✗	✗	✓	✗	✗
DPATCH [15]	✗	✗	✓	✗	✓	✓	✓
Extended-RP <sub>2</sub> [5]	✗	✓	✓	✗	✓	✓	✓
Thys's Patch [25]	✗	✓	✗	✗	✓	✓	✓

Table 6: Characteristics of seven representative attacks.

and TOG also provides additional three targeted specificity attacks. For model-applicability, UEA, RAP and DAG by design depend on the RPN structure, and can only be employed to generate adversarial examples against FRCNN (two-phase detectors). TOG is a general attack framework without dependency on any special structure and can be used to fool object detectors from both one-phase (YOLO and SSD families) and two-phase algorithms (e.g., FRCNN).

In addition to perturbing the entire image, adversarial patches are also proposed in either a digital (DPATCH) or physical (Extended-RP<sub>2</sub> and Thys's Patch) form. DPATCH puts a small patch (e.g., 40 × 40) on a benign example, fooling the victim to fabricate objects at random position or the location where the patch is placed. Extended-RP<sub>2</sub> and Thys's Patch propose printable adversarial patches. If the adversarial patch is presented physically in the scene captured by the camera, the captured image will become adversarial input, which will fool a victim detector to misdetect. Extended-RP<sub>2</sub> supports “disappearance” and “creation”, corresponding to the object-vanishing and object-fabrication effects, while Thys's Patch aims to make the object vanishing from the detector. Similar to TOG, all physical attack and digital patch algorithms can be employed on both two-phase and one-phase detection techniques.

## 5 Conclusion

We witnessed a growing number of digital or physical adversarial attacks to object detection systems recently [2,5,12,15,25,27,28]. To gain an in-depth understanding of the security risks of employing object detection intelligence in security-critical applications, in this paper, we develop a principled evaluation framework to analyze vulnerabilities of object detection systems through an adversarial lens, with three original contributions. First, we examine and compare the state-of-the-art attacks through our proposed evaluation framework. Second, to provide broader coverage of security risks in deep object detection systems, we present a family of TOG attack algorithms, capable of attacking both proposal-based two-phase detectors (e.g., FRCNN) and regression-based one-phase techniques (e.g., SSD, YOLOv3), supporting a general form of untargeted random attacks, and three targeted attacks, geared specifically to object detection. Third but not least, we introduce a set of quantitative metrics, including cross-resolution transferability and cross-model transferability w.r.t. algorithms and DNN backbones, to evaluate the effectiveness and cost of four representative methods of digital attacks, and using model-applicability to compare digital

attacks with physical patch attacks. Our evaluation framework can serve as a tool for analyzing adversarial attacks, assessing security risks and adversarial robustness of deep object detectors deployed in real-world applications.

### Acknowledgment

This research is partially sponsored by NSF CISE SaTC 1564097 and an IBM faculty award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

### References

1. Chen, K., Wang, J., Yang, S., Zhang, X., Xiong, Y., Change Loy, C., Lin, D.: Optimizing video object detection via a scale-time lattice. In: CVPR (2018)
2. Chow, K.H., Liu, L., Gursoy, E., Truex, S., Wei, W., Wu, Y.: Tog: Targeted adversarial objectness gradient attacks on real-time object detection systems. arXiv preprint arXiv:2004.04320 (2020)
3. Chow, K.H., Wei, W., Wu, Y., Liu, L.: Denoising and verification cross-layer ensemble against black-box adversarial attacks. In: IEEE BigData (2019)
4. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111**(1), 98–136 (2015)
5. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., Kohno, T., Song, D.: Physical adversarial examples for object detectors. arXiv preprint arXiv:1807.07769 (2018)
6. Gajjar, V., Gurnani, A., Khandhediya, Y.: Human detection and tracking for video surveillance: A cognitive science approach. In: ICCV (2017)
7. Girshick, R.: Fast r-cnn. In: ICCV (2015)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
12. Li, Y., Tian, D., Bian, X., Lyu, S., et al.: Robust adversarial perturbation on deep proposal-based models. arXiv preprint arXiv:1809.05962 (2018)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
14. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)
15. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: Dpatch: An adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299 (2018)
16. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017)
17. Papageorgiou, C.P., Oren, M., Poggio, T.: A general framework for object detection. In: ICCV. IEEE

18. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
20. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR (2017)
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
23. Simon, M., Amende, K., Kraus, A., Honer, J., Samann, T., Kaulbersch, H., Milz, S., Michael Gross, H.: Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds. In: CVPRW (2019)
24. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
25. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: CVPRW (2019)
26. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. IEEE (2001)
27. Wei, X., Liang, S., Chen, N., Cao, X.: Transferable adversarial attacks for image and video object detection. arXiv preprint arXiv:1811.12641 (2018)
28. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: ICCV (2017)

## Appendix

**A. Background.** The VOC 2007+2012 dataset has 16,551 training images and 4,952 testing images, while the COCO 2014 dataset has 117,264 training images and 5,000 testing images. The configuration and detection performance of the six detectors under no attack are reported in Table 7. All measurements are recorded on NVIDIA RTX 2080 SUPER (8 GB) GPU, Intel i7-9700K (3.60GHz) CPU, and 32 GB RAM on Ubuntu 18.04.

Dataset	Detector Identifier	Algorithm	Backbone	Input Resolution	Benign mAP(%)	Detection Time(s)
VOC	YOLOv3-D	YOLOv3	Darknet53	416x416	83.43	0.0328
	YOLOv3-M	YOLOv3	MobileNetV1	416x416	71.84	0.0152
	SSD300	SSD	VGG16	300x300	76.11	0.0208
	SSD512	SSD	VGG16	512x512	79.83	0.0330
COCO	FRCNN	Faster R-CNN	VGG16	600x600	67.37	0.1399
	YOLOv3-D	YOLOv3	Darknet53	416x416	54.16	0.0337

Table 7: A summary of victim detectors under no attack.

**B. Analysis on Targeted Specificity Attacks.** Table 8 reports the results of four TOG targeted attacks on six victim detectors (24 cases). TOG targeted attacks effectively bring down the mAP of all victim detectors, with any attack specificity. For instance, YOLOv3-D on VOC has a high mAP of

Detector (Dataset)	Targeted Attack	mAP (%)		Time Cost (s)		Distortion Cost			
		Benign	Adv.	Benign	Adv.	$L_\infty$	$L_2$	$L_0$	SSIM
YOLOv3-D (VOC)	TOG-vanishing	83.43	0.32	0.03	<b>0.77</b>	0.031	0.082	0.983	0.877
	TOG-fabrication	83.43	<b>0.25</b>	0.03	0.93	0.031	0.084	0.984	0.873
	TOG-mislabeling (ML)	83.43	3.15	0.03	0.95	0.031	0.080	0.972	<b>0.879</b>
	TOG-mislabeling (LL)	83.43	2.80	0.03	0.96	0.031	0.081	0.972	<b>0.879</b>
YOLOv3-M (VOC)	TOG-vanishing	71.84	0.36	0.02	<b>0.37</b>	0.031	0.082	0.978	0.878
	TOG-fabrication	71.84	<b>0.17</b>	0.02	0.57	0.031	0.084	0.976	0.873
	TOG-mislabeling (ML)	71.84	2.67	0.02	0.56	0.031	0.079	0.953	<b>0.882</b>
	TOG-mislabeling (LL)	71.84	1.60	0.02	0.56	0.031	0.079	0.953	0.881
SSD300 (VOC)	TOG-vanishing	76.11	5.54	0.02	<b>0.36</b>	0.031	0.120	0.978	0.880
	TOG-fabrication	76.11	<b>0.57</b>	0.02	0.37	0.031	0.122	0.978	0.877
	TOG-mislabeling (ML)	76.11	2.53	0.02	0.37	0.030	0.110	0.945	<b>0.891</b>
	TOG-mislabeling (LL)	76.11	1.44	0.02	0.37	0.030	0.111	0.945	0.889
SSD512 (VOC)	TOG-vanishing	79.83	6.23	0.03	<b>0.62</b>	0.031	0.071	0.975	0.868
	TOG-fabrication	79.83	<b>0.50</b>	0.03	0.69	0.031	0.071	0.976	0.866
	TOG-mislabeling (ML)	79.83	2.53	0.03	0.65	0.031	0.065	0.957	<b>0.878</b>
	TOG-mislabeling (LL)	79.83	1.20	0.03	0.65	0.031	0.066	0.956	0.877
FRCNN (VOC)	TOG-vanishing	67.37	<b>0.14</b>	0.14	<b>1.66</b>	0.031	0.058	0.975	0.862
	TOG-fabrication	67.37	1.24	0.14	1.68	0.031	0.057	0.977	0.866
	TOG-mislabeling (ML)	67.37	2.14	0.14	1.64	0.030	0.054	0.935	<b>0.873</b>
	TOG-mislabeling (LL)	67.37	1.44	0.14	1.60	0.030	0.054	0.935	0.872
YOLOv3-D (COCO)	TOG-vanishing	54.16	<b>0.41</b>	0.03	<b>0.78</b>	0.031	0.082	0.986	0.874
	TOG-fabrication	54.16	1.46	0.03	<b>0.78</b>	0.031	0.083	0.986	0.871
	TOG-mislabeling (ML)	54.16	5.43	0.03	1.00	0.031	0.080	0.968	<b>0.878</b>
	TOG-mislabeling (LL)	54.16	0.76	0.03	1.00	0.031	0.080	0.968	0.877

Table 8: Targeted attacks by TOG on different datasets and victim detectors.

83.43% given benign images but, under attacks, it becomes less than 3.15%. Even though the adversarial examples in targeted attacks can fool the victim detectors to misdetect with the targeted specificity effects, such attack sophistication does not drastically incur additional attack time cost and distortion cost, compared with the TOG untargeted attack scenario in Table 2.

Figure 5 compares the four targeted attacks with respect to the number of object detected by three victim detectors (YOLOv3-D, SSD512 and FRCNN) with different settings of the confidence threshold. The benign case (the blue solid curve) indicates the number of objects detected by the victims under no attacks. Confidence thresholding is used by object detection algorithms as a post-processing step to return only detected objects with high confidence (Section 2.1), and the threshold is a hyperparameter defined by the system owner (e.g., FRCNN uses 0.70 by default). We find that all trends are consistent across both detectors: Figure 5 experimentally confirms that (i) the TOG-vanishing attacks significantly lower the number of detected objects with any setting of confidence threshold, (ii) the number of detected objects is drastically increased in TOG-fabrication attacks, and (iii) the TOG-mislabeling attacks (both ML and LL) have almost the same number of objects detected on benign examples.

Figure 6 further analyzes the two targeted mislabeling attacks of TOG in terms of ASR according to Equation 13. With a similar formulation, we also introduce misdetection rate (MR) to compute the portion of objects that are mislabeled under TOG-mislabeling attacks. Note that MR still requires the de-

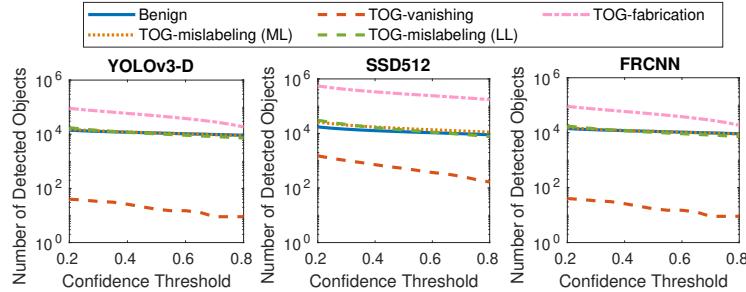


Fig. 5: Number of detected objects under no attack and TOG targeted attacks.

tected bounding box to be correct, but the predicted class label of the object can be any class but not the correct one. We observe that a large portion of objects are successfully mislabeled as the maliciously targeted class (ASR), and only small portion is randomly mislabeled instead (MR - ASR), especially for the ML targets (Figure 6a). For the LL attack targets (Figure 6b), the ASR is less than 80%, but the misdetection rate (MR) is close to 100% in all five victim detectors, indicating that almost all objects in all test examples are mislabeled though only less than 80% LL targeted mislabeling attacks succeeded.

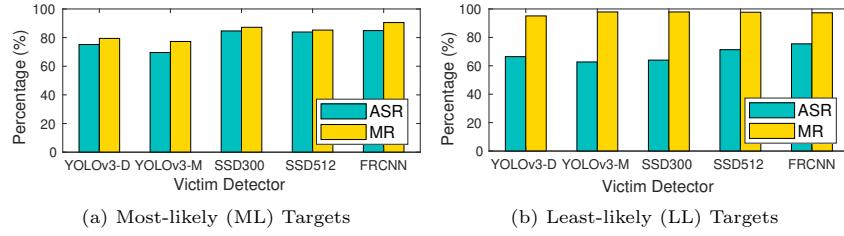


Fig. 6: ASR and MR of TOG-mislabeling attacks.

**C. Transferability of Targeted Specificity Attacks.** Consider in Table 5 the victim detector SSD512 with the same backbone and detection algorithm as SSD300, TOG-vanishing can perfectly transfer the attack to SSD512 with the same effect (i.e., no object is detected). For TOG-fabrication, we observe that while the number of false objects is not as much as in the SSD300 case, a fairly large number of fake objects are wrongly detected by SSD512. The TOG-mislabeling (LL) attack transfers to SSD512 but with the object-fabrication effect instead, while the TOG-mislabeling (ML) attack failed to transfer for this example. Now consider YOLOv3-D and YOLOv3-M, the TOG-mislabeling (LL) attack is successful in transferability for both victims but with different attack effects, such as wrong or additional bounding boxes or wrong labels. Also, the attacks from SSD300 can successfully transfer to YOLOv3-M with different attack effects compared to the attack results in SSD300, but not to YOLOv3-D for this example.