

# Phantom Sponges: Exploiting Non-Maximum Suppression to Attack Deep Object Detectors

Avishag Shapira  
The Open University, Israel  
avishagsh@openu.ac.il

Alon Zolfi  
Ben-Gurion University, Israel  
zolfi@post.bgu.ac.il

Luca Demetrio  
University of Genoa, Italy  
luca.demetrio@unige.it

Battista Biggio  
University of Cagliari, Italy  
battista.biggio@unica.it

Asaf Shabtai  
Ben-Gurion University, Israel  
shabtaia@bgu.ac.il

## Abstract

Adversarial attacks against deep learning-based object detectors have been studied extensively in the past few years. Most of the attacks proposed have targeted the model’s integrity (i.e., caused the model to make incorrect predictions), while adversarial attacks targeting the model’s availability, a critical aspect in safety-critical domains such as autonomous driving, have not yet been explored by the machine learning research community. In this paper, we propose a novel attack that negatively affects the decision latency of an end-to-end object detection pipeline. We craft a universal adversarial perturbation (UAP) that targets a widely used technique integrated in many object detector pipelines – non-maximum suppression (NMS). Our experiments demonstrate the proposed UAP’s ability to increase the processing time of individual frames by adding “phantom” objects that overload the NMS algorithm while preserving the detection of the original objects which allows the attack to go undetected for a longer period of time.

## 1. Introduction

Deep learning-based computer vision models, and specifically object detection (OD) models, are becoming an essential part of intelligent systems in many domains, including autonomous driving. The adoption of these models relies on two critical aspects of security: *integrity* - ensuring the accuracy and correctness of the model’s predictions, and *availability* - ensuring uninterrupted access to the system.

In the past few years, OD models have been shown to be vulnerable to adversarial attacks, including attacks in which a patch containing an adversarial pattern (e.g., black and white stickers [9], a cardboard plate [29], T-shirts [33, 31, 12]) is placed on the target object or a physical patch is attached to the camera lens [35]. These attacks

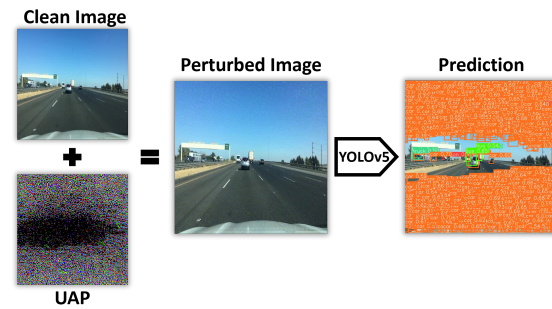


Figure 1: An illustration of adding our UAP to a driving image and YOLO’s prediction for the perturbed image.

share one main attribute – they are all aimed at compromising the model’s integrity.

Recently, availability-based attacks have been shown to be effective against deep learning-based models. Shumailov *et al.* [26] presented *sponge examples*, which are perturbed inputs designed to increase the energy consumed by natural language processing and computer vision models when deployed on hardware accelerators, by increasing the number of active neurons during classification. Following this line of research, other studies proposed sponge-like attacks, which mainly targeted image classification models [4, 3, 6, 11]. To the best of our knowledge, no studies have explored the ability to create an attack that compromises the OD model’s availability (i.e., inference latency).

A common end-to-end OD pipeline is comprised of several steps: (a) preprocess the input image, (b) feed the processed image to the detector’s network to obtain candidate predictions, (c) perform rule-based filtering based on the candidates’ confidence scores, and (d) use non-maximum suppression (NMS) to filter redundant candidates.

Our initial attempt to apply the sponge attack proposed by Shumailov *et al.* [26] on YOLO’s feedforward phase to

decelerate inference processing time was unsuccessful; this is due to the fact that for most of the images, the vast majority of the model’s activation values are not zero by default. Therefore, in this paper, we focus on exposing the inherent weaknesses present in the NMS algorithm and exploit them to perform a sponge attack. We present the first availability attack against an end-to-end OD pipeline, which is performed by applying a *universal adversarial perturbation* (UAP) [34, 19], as shown in Figure 1. In the proposed attack, gradient-based optimization is used to construct a UAP that targets the weaknesses of the NMS algorithm, creating a large amount of candidate bounding box predictions which must be processed by the NMS, thereby overloading the system and slowing it down. The custom loss function employed is also aimed at preserving the detection of objects present in the original image. The fact that we create a universal perturbation makes our attack practical – that is, it can be applied to any stream of images in real time, thereby affecting the model’s inference latency.

We conducted various experiments demonstrating our attack’s effectiveness with different: (a) attack parameters, (b) versions of the YOLO object detector [23, 2, 13], (c) variations of NMS, (d) hardware platforms, and (e) datasets (we used four datasets, three of them contain driving images). In addition, we used ensemble learning to improve the attack’s transferability. Our results show that the use of our UAP increases the inference time by 45% without compromising the detection capabilities (77% of the original objects were detected by the model when the proposed attack was performed). In a video stream experiment, our UAP decreased the FPS rate from  $\sim 40$  (unattacked frames) to  $\sim 16$ . The contributions of our work can be summarized as follows:

- We present the first availability (quality of service) attack against an end-to-end OD pipeline.
- We construct a *universal* perturbation that can be applied on a video stream in real time.
- We perform extensive experiments with different object detectors (including ensembles), NMS implementations, hardware platforms, and multiple datasets.

## 2. Object Detectors

State-of-the-art object detection models can be broadly categorized into two types: one-stage detectors (*e.g.*, SSD [16], YOLO [21, 22, 23]) and two-stage detectors (*e.g.*, Mask R-CNN [10], Faster R-CNN [24]). Two-stage detectors use a region proposal network to locate potential bounding boxes in the first stage; these boxes are used for classification of the bounding box content in the second stage. In contrast, one-stage detectors simultaneously generate bounding boxes and predict their class labels.

### 2.1. You Only Look Once (YOLO) Object Detector

In this paper, we focus on the state-of-the-art one-stage object detector YOLO [21, 22, 23]. YOLO’s architecture

consists of two parts: a convolution-based backbone used for feature extraction, which is followed by *multi-scale* grid-based detection heads used to predict bounding boxes and their associated labels. This architecture design established the foundation for many of the object detectors proposed in recent years (*e.g.*, YOLOv4 [2], YOLOv5 [13]).

**YOLO’s detection layer.** The last layer of each detection head predicts a 3D tensor that encodes: (a) the bounding box – the coordinate offsets from the anchor box, (b) the objectness score – the detector’s confidence that the bounding box contains an object ( $Pr(Object)$ ), and (c) the class scores – the detector’s confidence that the bounding box contains an object of a specific class ( $Pr(Class_i|Object)$ ).

**YOLO’s end-to-end detection pipeline.** YOLO produces a fixed number of candidate predictions (denoted by  $\mathcal{C}$ ) for a given image size, which are later filtered sequentially using a predefined threshold  $T_{conf}$ :

- Objectness score filtering –

$$F_1 = \{c_{obj\ score} > T_{conf} | c \in \mathcal{C}\} \quad (1)$$

- Unconditional class score filtering –

$$F_2 = \{c_{obj\ score} \cdot \max\{c_{class\ score\ i}\}_{i=0}^{N_c} > T_{conf} | c \in \mathcal{C}\} \quad (2)$$

Finally, since many candidate predictions may overlap and predict the same object, the NMS algorithm is applied to remove redundant predictions.

### 2.2. Non-Maximum Suppression (NMS)

The NMS technique is widely used in both types of object detectors (one-stage and two-stage), with the aim of filtering overlapping bounding box predictions. In the simplest version of the NMS algorithm (also referred to as the *vanilla* version) two steps are performed for each class category: (a) sort all candidate predictions based on their confidence scores, and (b) select the highest ranking candidate and discard all candidates for which the intersection over union (IoU) surpasses a predefined threshold  $T_{IoU}$  (*i.e.*, the bounding boxes overlap above a specific level). These two steps are repeated until all candidates have been selected or discarded by the algorithm.

Recently an improvement to the *vanilla* version was proposed [20]; the improvement, referred to as the *coordinate trick*,<sup>1</sup> employs a new strategy for performing NMS without directly iterating over the class categories. By adding an offset that only relies on the prediction’s category (*i.e.*, bounding boxes of the same category are added with the same value), bounding boxes belonging to different categories do not overlap. Therefore, NMS can be applied on all of the bounding boxes simultaneously.

<sup>1</sup>The implementation can be found at: [https://pytorch.org/vision/0.11/\\_modules/torchvision/ops/boxes.html](https://pytorch.org/vision/0.11/_modules/torchvision/ops/boxes.html)

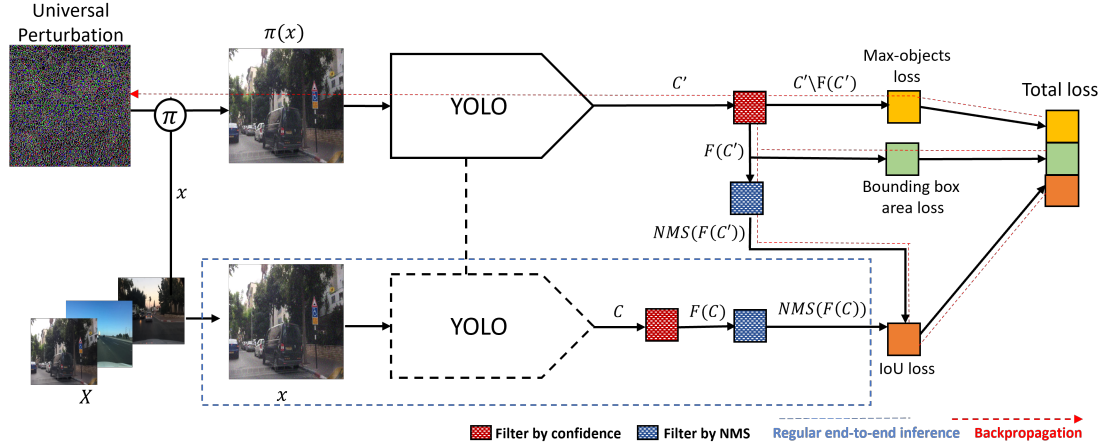


Figure 2: Overview of our method’s pipeline.

### 3. Universal Sponge Attack

We aimed to produce a digital UAP capable of causing an increase in the end-to-end processing time of an image processed by the YOLO object detector. Our attack is designed to achieve the following objectives: (i) increase the amount of time it takes for the object detector to process an image; (ii) create a universal perturbation that can be applied to any image (frame) within a stream of images, for example, in an autonomous driving scenario; and (iii) preserve the detection of the original objects in the image.

#### 3.1. Threat Model

As mentioned in Section 2, YOLO outputs a fixed number of candidates, *i.e.*, the number of computations (matrix multiplications) for images of the same size is equal. Therefore, the weak spot of YOLO’s end-to-end detection lies in the NMS algorithm. The NMS algorithm iterates over all of the candidates’ bounding boxes until they are all processed (*i.e.*, kept or discarded). Therefore, increasing the number of inputs processed by the algorithm will cause a delay. To exploit the algorithm’s weakness to the fullest extent, we examine its behavior in a single iteration, where  $\mathcal{C}_j$  represents the remaining candidates in iteration  $j$ . In the worst case, if the IoU values of the bounding box being examined (highest ranking) and all of the other bounding boxes are lower than  $T_{IoU}$ , none of the other bounding boxes are discarded, and all of them proceed to the next iteration (except for the examined one which is kept), *i.e.*,  $|\mathcal{C}_{j+1}| = |\mathcal{C}_j| - 1$ . In this case, the NMS algorithm runtime complexity is factorial:  $\mathcal{O}(|\mathcal{C}|!)$ .

In the vanilla version of the NMS, this case could be achieved by providing inputs of the same class category (a targeted attack), while in the coordinate trick version, it can be achieved without specifically providing a target class.

#### 3.2. Optimization Process

To optimize our perturbation’s parameters (pixels), we use projected gradient descent (PGD) with the  $L_2$  norm. We compose a novel loss function that aims to achieve the objectives presented above; our loss function consists of three components: (a) the max-objects loss, (b) the bounding box area loss, and (c) the IoU loss.

**Max-objects loss.** Let  $g : \mathcal{X} \rightarrow \mathcal{C}$  be a YOLO detector that takes a clean image  $x \in \mathcal{X}$  and outputs a set of candidates  $\mathcal{C} = g(x)$ . Similarly, let  $\pi(x)$  be the perturbation function, and  $\mathcal{C}' = g(\pi(x))$  be the set of candidates produced for the perturbed image. For simplicity, let  $F$  be the composition of  $F_1$  and  $F_2$  such that  $F = F_2(F_1(\mathcal{C}))$ .

In order to increase the number of candidates passed to the NMS step ( $|\mathcal{C}'|$ ), we need to increase the number of predictions that are not filtered by  $F$ . Therefore, we aim to increase the confidence scores of all of the candidates that do not exceed  $T_{\text{conf}}$ , as shown in Figure 3. We also limit the increase in the candidates’ confidence score to  $T_{\text{conf}}$ , so that the loss favors the prediction of candidates that are far from the threshold. More formally, the non-targeted loss for a single candidate  $c$  is:

$$\ell_{\text{single conf}}(c) = T_{\text{conf}} - (c_{\text{obj score}} \cdot \max\{c_{\text{class score } i}\}_{i=0}^{N_c}) \quad (3)$$

However, since a targeted attack is suitable for both NMS versions, we replace Equation 3 with:

$$\ell_{\text{single conf}}(c) = T_{\text{conf}} - (c_{\text{obj score}} \cdot c_{\text{target class score}}) \quad (4)$$

While this component focuses on the confidence of the predictions, we also want to consider the number of candidates filtered by  $F$  (prior to the NMS step). Therefore, the loss over all candidates that were filtered ( $\mathcal{C}' \setminus F(\mathcal{C}')$ ) is:

$$\ell_{\text{max objects}} = \frac{1}{|\mathcal{C}'|} \cdot \sum_{c' \in \mathcal{C}' \setminus F(\mathcal{C}')} \ell_{\text{single conf}}(c') \quad (5)$$

**Bounding box area loss.** To exploit the NMS’s factorial time complexity weakness, we aim to reduce the area of all of the bounding boxes, which will eventually result in a lower IoU value among all of the bounding boxes. More formally, we consider the following loss for a single bounding box:

$$\ell_{\text{single area}}(b) = b_w \cdot b_h, \quad (6)$$

where  $b$  is the bounding box, and  $b_w$  and  $b_h$  are the normalized width and height, respectively. The loss function applied over all of the candidates that are passed to the NMS step  $F(\mathcal{C}')$  is expressed as:

$$\ell_{\text{bbox area}} = \frac{1}{|F(\mathcal{C}')|} \cdot \sum_{c' \in F(\mathcal{C}')} \ell_{\text{single area}}(c'_{\text{bbox}}) \quad (7)$$

A positive side effect caused by reducing the area of the bounding boxes is a reduction in the density of the bounding boxes in the UAP, providing additional space for other candidates to be added, which our attack benefits from.

**IoU loss.** To achieve our third objective of enabling detection of the original objects in the image, we aim to maximize the IoU score between the final predictions’ bounding boxes (predictions that are kept after the NMS step) in the clean image  $\text{NMS}(F(\mathcal{C}))$  and the adversarial image  $\text{NMS}(F(\mathcal{C}'))$ . Therefore, for a single candidate  $c \in \text{NMS}(F(\mathcal{C}))$ , we extract the maximum IoU value:

$$\text{Max IoU}(c) = \max_{c' \in \text{NMS}(F(\mathcal{C}'))} \text{IoU}(c_{\text{bbox}}, c'_{\text{bbox}}) \quad (8)$$

To be more precise, since we aim to minimize the loss function, and the IoU’s value is in the range  $[0, 1]$ , the loss component is defined as:

$$\ell_{\text{single IoU}}(c) = 1 - \text{Max IoU}(c) \quad (9)$$

Finally, the loss of this component over all the final predictions produced by the detector on the clean image  $\text{NMS}(F(\mathcal{C}))$  is defined as follows:

$$\ell_{\text{max IoU}} = \frac{1}{|\text{NMS}(F(\mathcal{C}))|} \cdot \sum_{c \in \text{NMS}(F(\mathcal{C}))} \ell_{\text{single IoU}}(c) \quad (10)$$

It should be noted that since we use the object detector’s predictions (instead of the ground-truth labels), our attack is not limited to annotated datasets.

**Final loss function.** Since we consider a universal perturbation, where a single perturbation  $P$  is chosen to minimize the loss function over samples from some distribution  $\mathcal{D}$ , the final loss function of the attack is defined as follows:

$$\min_P \mathbb{E}_{x \sim \mathcal{D}} [\lambda_1 \cdot \ell_{\text{max objects}} + \lambda_2 \cdot \ell_{\text{bbox area}} + \lambda_3 \cdot \ell_{\text{max IoU}}] \quad (11)$$

where  $\lambda_i$  is a weighting factor. The computed gradients are backpropagated to update our perturbation’s pixels. Figure 2 presents an overview of a single iteration of our attack.

**Ensemble training.** To improve the transferability of our attack to different object detection models, we perform ensemble training using  $K$  models, where in each iteration a different YOLO model  $g_k$  ( $k \in K$ ) is randomly selected to backpropagate and update the perturbation pixels.

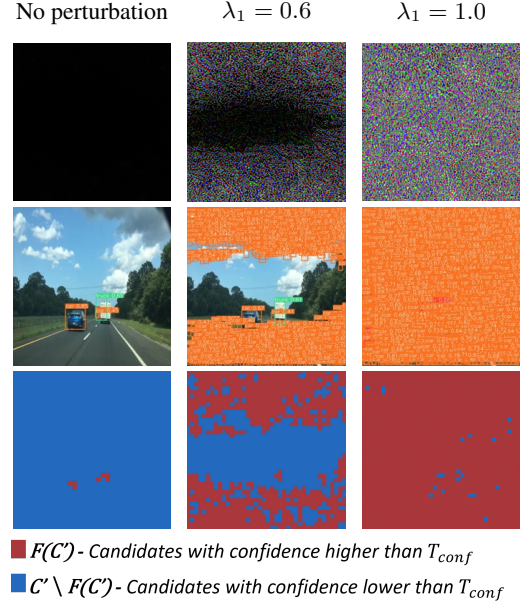


Figure 3: Top: UAPs trained with different  $\lambda_1$ ; Middle: perturbed images with the corresponding UAP predicted using YOLOv5; Bottom: heat map of the candidates’ confidence score. Red (resp. Blue) represents candidates that are (resp. not) passed to the NMS.

## 4. Evaluation

### 4.1. Evaluation Setup

**Models.** In our evaluation, we conduct experiments on various versions of the state-of-the-art YOLO object detector (YOLOv3 [23], YOLOv4 [2], and YOLOv5 [13]), pre-trained on the MS-COCO dataset [15]. YOLOv5 offers several model networks of different sizes [13] (small, medium, etc.). The different YOLO versions are conceptually similar which enables the creation of a generic attack for both a single model and an ensemble of models.

**Datasets.** We evaluate our attack in the autonomous driving domain using the following datasets: (a) Berkeley Deep-Drive (BDD) [32] – contains 100K images with various attributes such as weather (clear, rainy), scene (city street, residential), and time of day (daytime, night), resulting in a diverse dataset; (b) Mapillary Traffic Sign Dataset (MTSD) [7] – diverse street-level images obtained from various geographic areas; and (c) LISA [18] – contains dozens of video clips split into frames. In addition, we evaluate our attack in a general setting (*i.e.*, images are taken from various domains) using the PASCAL VOC [8] dataset.

		$\lambda_2=0$	$\lambda_2=10$	$\lambda_2=20$
		Total time (NMS time) $\uparrow$ / $ F(\mathcal{C}') $ $\uparrow$ / Recall $\uparrow$		
	Clean	24 (2.2) / 80 / 100%		
	Random	24 (2.2) / 60 / 53.7%		
$\epsilon=30$	$\lambda_1=0.5, \lambda_3=0.5$	24 (2.2) / 80 / 100%	24 (2.2) / 80 / 100%	30.4 (8.6) / 7200 / 80.3%
	$\lambda_1=0.6, \lambda_3=0.4$	24 (2.2) / 80 / 100%	<b>34.8 (13) / 9000 / 77%</b>	32.9 (11.1) / 8100 / 77%
	$\lambda_1=0.7, \lambda_3=0.3$	33 (11.2) / 8600 / 75%	35.8 (14) / 9100 / 75%	35.4 (13.6) / 9000 / 73%
	$\lambda_1=0.8, \lambda_3=0.2$	35.9 (14.1) / 10300 / 67.4%	37.3 (15.5) / 10800 / 65.9%	36.9 (15.1) / 10600 / 65.3%
	$\lambda_1=1, \lambda_3=0$	38.9 (17.1) / 12500 / 37.7%	42.9 (21.1) / 12800 / 35.8%	41.7 (19.9) / 12600 / 35%
	Random	23.9 (2.1) / 20 / 15%		
$\epsilon=70$	$\lambda_1=0.5, \lambda_3=0.5$	24 (2.2) / 80 / 100%	24 (2.2) / 80 / 100%	32.4 (10.6) / 8000 / 74%
	$\lambda_1=0.6, \lambda_3=0.4$	24 (2.2) / 80 / 100%	35.6 (13.8) / 9600 / 69.6%	35.1 (13.3) / 9100 / 69%
	$\lambda_1=0.7, \lambda_3=0.3$	36.9 (15.1) / 11200 / 65%	42.1 (20.3) / 13800 / 56%	37.1 (15.3) / 10800 / 64.4%
	$\lambda_1=0.8, \lambda_3=0.2$	46.6 (24.8) / 15600 / 47.4%	47.6 (25.8) / 16000 / 45.3%	47.8 (26) / 15900 / 44.2%
	$\lambda_1=1, \lambda_3=0$	56.5 (34.7) / 18800 / 21.8%	58.9 (37.1) / 19200 / 18%	57.6 (35.8) / 18800 / 16.5%

Table 1: Average results when using various  $\lambda_i$  and  $\epsilon$  values (different loss function component weight balancing). Bold indicates the UAP’s results we consider as the results with the best balance between the loss components. Grey cells indicate that the attack returns an empty UAP (identical results with the clean image).  $\uparrow$  indicates that higher values are better.

**Evaluation metrics.** Two goals our attack aims to achieve are to increase the model’s end-to-end inference time and preserve the detection of objects in the original image. To quantify the effectiveness of our attack in achieving these goals, we used the following metrics: (i)  $|F(\mathcal{C}')|$  – the number of candidates provided to the NMS algorithm after applying the confidence filter  $F$ ; (ii) **time** – the end-to-end detection pipeline’s total processing time in milliseconds (we also measured the processing time of the NMS stage); and (iii) **recall** – the number of original objects detected in the perturbed image.

**Implementation details.** For the target models, we used the small sized YOLOv5 version (referred to as YOLOv5s), as well as YOLOv3 and YOLOv4. For the ensemble learning, we used different combinations of these models.

To evaluate our adversarial perturbation, we randomly chose 2,000 images from the validation set of each of the datasets (BDD, MTSD, and PASCAL). For each dataset, we used 1,500 images to train the UAP and then examined its effectiveness on the remaining 500 images.

We set  $T_{conf} = 0.25$  and  $T_{IoU} = 0.45$ , since they are the default values commonly used for these models in the inference phase (throughout this section, the recall values presented are based on this threshold). We choose the *car* class as our attack’s target class, due to its clear connection to the autonomous driving domain. To obtain unbiased measurements, we performed 30 iterations *for each image* and calculated the average inference time. The experiments were conducted on a GPU (NVIDIA Quadro T1000) and a CPU (Intel Core i7-9750H). Unless mentioned otherwise, the experiments were conducted on the BDD dataset, and the running times were measured using the coordinate trick NMS implementation. The source code is available at: <https://github.com/AvishagS422/PhantomSponges>.

## 4.2. Results

**Effectiveness of the UAP with different epsilon ( $\epsilon$ ) values.** The  $\epsilon$  parameter in the PGD attack denotes the radius of the hypersphere, *i.e.*, the maximum amount of noise to be added to image. Larger  $\epsilon$  values will result in a more substantial perturbation, which while being more perceptible to the human eye, will result in a more successful attack. Table 1 presents results for different  $\epsilon$  values. As expected, we can see that the larger the  $\epsilon$  value, the larger the number of candidate predictions that exceed the confidence threshold and are processed by the NMS, to the point that it almost reaches the maximum number of possible candidates. In addition, we present the results for a baseline attack – a perturbation with randomly colored pixels sampled uniformly in a bound range according to the specific  $\epsilon$ . As can be seen, the attacks involving these random UAPs were unsuccessful and even reduced the NMS processing time.

**Effectiveness of the UAP with different  $\lambda_1$  and  $\lambda_3$  values.** The  $\lambda_1$  and  $\lambda_3$  values enable control of the balance between the max-objects and the IoU loss components. Since the purpose of each of these components might be contradictory, we chose to use complementary values to balance the two components; therefore,  $\lambda_3 = 1 - \lambda_1$ .

Intuitively, the higher the  $\lambda_1$  value is, the larger the number of candidates that are passed to the NMS step; at the same time, however, the recall value decreases, *i.e.*, fewer original objects are preserved. Table 1 presents the results for the different combinations. For example, when using a UAP that was trained with a configuration of  $\lambda_1 = 1$ , 19,200 candidates are processed by the NMS while preserving only 18% of the original objects, as opposed to a configuration of  $\lambda_1 = 0.7$ , which adds 13,800 candidates but increases the recall level to 56%.

By visually examining the UAPs presented in Figure 3, we can see that when setting  $\lambda_3 > 0$  ( $\lambda_1 < 1$ ), the attack detects areas in the original images where objects commonly appear, forcing the perturbation to add candidates on the image’s sides while the center of the perturbed image remains unattacked. In contrast, when setting  $\lambda_3 = 0$ , candidates are added all over the image. This is an expected outcome, since there are naturally fewer objects in these areas (where we would usually find the sky, a road, or a sidewalk in the autonomous driving domain). Furthermore, we can see that below a certain  $\lambda_1$  value, the loss function strongly favors the preservation of the detection of the objects in the original image, resulting in an unsuccessful attack, *i.e.*, the UAP remains unchanged, and no “phantom” objects are added (different hyperparameter configurations causing this behavior can be seen in Table 1).

**Effectiveness of the UAP with different  $\lambda_2$  values.** As mentioned in Section 3, minimizing the IoU between the candidates processed by the NMS results in fewer discarded candidates in each iteration, causing the NMS to perform more iterations. Therefore, we evaluated the effectiveness of the bounding box area loss component on three different  $\lambda_2$  values. By examining the results (Table 1), we can see that when comparing UAPs with similar  $|F(C')|$  this component increases the processing time of the NMS. However, the use of an overly large  $\lambda_2$  value may decrease the number of candidates processed by the NMS. We believe this occurs, because the attack is unable to decrease the bounding box area appropriately for a large number of candidates and consequently is unable to decrease the total loss value. Therefore, fewer candidates are passed to the NMS.

Another interesting observation is that increasing the  $\lambda_2$  value enables the  $\lambda_3$  value to be increased without disrupting the balance between the max-objects and IoU loss components. We can see that the use of a large value for  $\lambda_2$  ( $=20$ ) allows the  $\lambda_3$  value to increase to 0.5 and results in a UAP with an high recall value (80.3%).

**Vanilla NMS vs. coordinate trick NMS.** Since we aim to perform a generic attack that will successfully exhaust the NMS algorithm, we examine its effectiveness on two different versions of the algorithm: vanilla and coordinate trick. Figure 4 shows the large differences in the *vanilla* version’s running times for UAPs that were created using the targeted and non-targeted version of the attack. The *vanilla* version’s running time increases in the targeted setting dramatically compared to the non-targeted setting, whereas the *coordinate trick* version’s running time remains largely unchanged in both versions. We can also see that the running times of the *vanilla* and *coordinate trick* implementations are very similar for the targeted attack.

**Executing the attack on different hardware platforms.** We evaluated the attack with different values of  $\lambda_i$  and  $\epsilon$

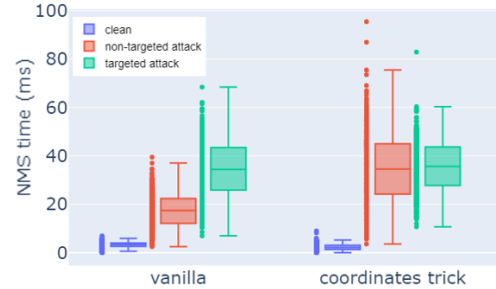


Figure 4: NMS processing time with different NMS implementations: *coordinate trick* vs. *vanilla*.

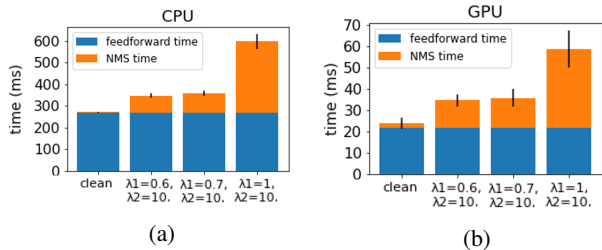


Figure 5: Average running time on different hardware platforms: (a) GPU and (b) CPU.

on both a GPU and a CPU. The attack works efficiently on both platforms, increasing the inference time similarly with different UAPs. Figure 5 presents the GPU and CPU inference time results with three different UAPs.

**UAPs for different models.** To demonstrate the universality of our attack, we evaluated it on three different versions of YOLO: YOLOv3, YOLOv4, and YOLOv5s. The results presented in Figure 6 show that the attack is effective on the three models, performing in a similar way across different configurations. We can see that YOLOv3 is the least robust model, with the attack increasing the amount of candidates (and consequently the total time) the most. Interestingly, YOLOv4 is more robust than YOLOv5 due to its two-phase training procedure, in which the backbone is pretrained on the ImageNet [25] dataset and only later fine-tuned on the MS-COCO [15] dataset for object detection, as opposed to YOLOv5 which is trained from scratch on MS-COCO.

**Different datasets.** We trained the UAPs using images from three datasets (BDD, MTSD, and PASCAL-VOC) and evaluated their effectiveness on unseen images. In Table 2, it can be seen that when the preservation of the detection of the original objects was not considered (Config<sub>2</sub>), we were able to create an efficient UAP for each of the three datasets (*i.e.*,  $\sim 20K$  objects are passed to the NMS step). However, when we aimed to preserve the detection of the original objects (Config<sub>1</sub>), the performance varies for the three datasets. There is good preservation for datasets belonging to a specific domain (BDD, MTSD) and unsatisfactory preservation for the more general dataset (PASCAL-VOC).

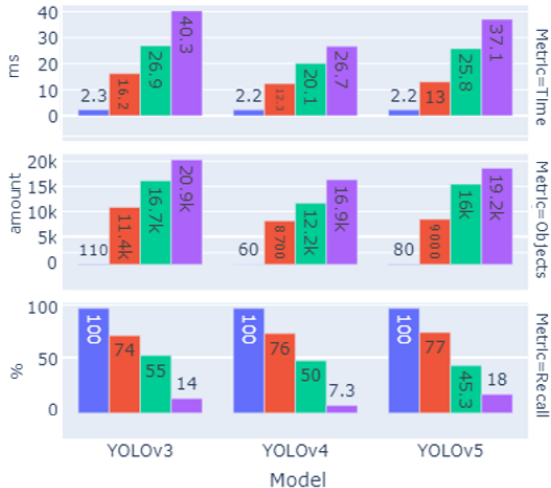


Figure 6: The performance of UAPs trained for different versions of YOLO. Different colors represent different UAP configurations: blue: clean image, red:  $(\epsilon, \lambda_1, \lambda_2) = (30, 0.6, 10)$ , green:  $(\epsilon, \lambda_1, \lambda_2) = (70, 0.8, 10)$ , and purple:  $(\epsilon, \lambda_1, \lambda_2) = (70, 1, 10)$ .

	BDD	MTSD	PASCAL-VOC
	NMS time (ms) $\uparrow$ / $ F(C')  \uparrow$ / Recall $\uparrow$		
Clean	2.2 / 80 / 100%	2.1 / 40 / 100%	2.2 / 50 / 100%
Config <sub>1</sub>	13 / 9000 / 77%	12.1 / 8700 / 78.1%	5.4 / 3000 / 72%
Config <sub>2</sub>	37.1 / 19200 / 18%	37 / 19400 / 10%	32.6 / 18200 / 3.7%

Table 2: Average results of the UAP on three different datasets. Config<sub>1</sub>:  $(\epsilon, \lambda_1, \lambda_2) = (30, 0.6, 10)$ ; Config<sub>2</sub>:  $(\epsilon, \lambda_1, \lambda_2) = (70, 1, 10)$

As mentioned earlier, when trying to preserve the detection of the original objects ( $\lambda_3 > 0$ ), our attack focuses on adding phantom objects in areas in which the original objects do not usually appear. Therefore, when we apply the attack in a specific domain such as autonomous driving (as opposed to a general setting), it is easier for the attack to balance the loss components, since the images share common characteristics regarding the objects’ locations.

**Ensemble learning.** As discussed in [14], in some cases, transferability is difficult to achieve. Using ensemble learning may enable the attack to overcome the transferability challenge in cases in which there is a set of suspected/potential target models available to the attacker, but the attacker does not know the specific target model used. Even then, however, an ensemble-based attack may not succeed, since it is more difficult to perform an attack that is successful on multiple models simultaneously than on a single model; therefore, we chose to evaluate an ensemble-based version of our attack. The results presented in Table 3 demonstrate the effectiveness of UAPs created using the ensemble technique. As can be seen, the UAP does not nat-

	Victim Models		
	YOLOv3	YOLOv4	YOLOv5s
	NMS time (ms) $\uparrow$ / $ F(C')  \uparrow$ / Recall $\uparrow$		
YOLOv5s	2.1 / 10 / 10%	2.1 / 5 / 5%	37.1 / 19200 / 18%
Ens <sub>1</sub>	2.2 / 10 / 12%	15.1 / 10200 / 11.8%	23.5 / 14600 / 16.7%
Ens <sub>2</sub>	23.2 / 14500 / 15%	16.6 / 11800 / 10.2%	20.2 / 13400 / 28.7%

Table 3: Average results for a UAP trained using the ensemble technique and evaluated on different YOLO versions. Ens<sub>1</sub>:YOLOv4 + YOLOv3; Ens<sub>2</sub>:YOLOv5 + YOLOv4 + YOLOv3; configuration:  $(\epsilon, \lambda_1, \lambda_2) = (70, 1, 10)$ .

urally transfer to other models. For example, when using YOLOv5 as the target model, the UAP fails to transfer to YOLOv3 and YOLO4 (*i.e.*, the inference time does not increase). However, when we incorporate the ensemble technique, the UAP is able to generalize over all of the models it was trained on. These results indicate that an attacker trying to perform the attack does not need to know the type/version of the models. In order to perform a successful attack, one UAP trained on an ensemble of models can be effective.

**Real-time video stream setup.** To demonstrate the impact of our attack on the inference time in practice, we randomly choose 15 video clips from the LISA dataset and tested the first 500 frames in each video clip. Each frame is applied with our UAP (trained on the BDD dataset images with the following configuration:  $\epsilon = 70, \lambda_1 = 1, \lambda_2 = 10$ ) and processed by the YOLOv5 detector. To illustrate the impact of the UAP in the overall running times, we also measure the time for clean frames (unattacked). On average, the overall running time of a clean video is 12,300ms (12.3s), whereas it takes 31,000ms (31s) to process an attacked video – a 251% increase. In terms of the processing time of a single frame, a clean frame is processed in 24.7ms (3ms for the NMS stage), whereas the processing time of an attacked frame is 62.2ms (40.5ms for the NMS stage). Hence, in an unattacked scenario, the system can output predictions at a rate of  $\sim 40$  FPS, while our UAP reduces the FPS to  $\sim 16$ .

### 4.3. Discussion

**The “phantom” predictions.** Based on the results of our experiments, we can conclude that any class category can be defined as the target class without compromising the attack’s performance. However, it is interesting to examine the UAPs’ patterns for different target classes. Figure 7 presents UAPs trained with the *car* and *person* classes as the target class, where we can see that the patterns created seem to consist of thousands of tiny objects belonging to the target class. For the non-targeted version of our attack, the UAP mainly adds *person* and *chair* class predictions. We assume that this has to do with the fact that the model is pretrained on the COCO-MS dataset [15] in which ‘person’ is the most common class and ‘chair’ is the third most common target class in the training set.

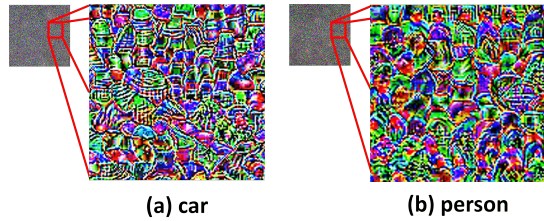


Figure 7: UAPs created for two different target classes (small box) and a closer look at their patterns (large box).

**Mitigation.** One possible mitigation against our attack is to limit the image’s processing time. In this case, if the processing time is over a predefined threshold, the system will interrupt the detection process. While this mitigation can bound the system’s latency time, it actually serves the attacker’s purpose – harming the availability of the system. Another approach would be to limit the number of candidates passed to the NMS; however this may also serve the purposes of the attacker by compromising the model’s integrity. Hence, the suggested mitigation might not be an appropriate solution for real-time systems such as the OD systems of autonomous vehicles, since interrupting the detection process every few frames could have serious consequences, endangering the car’s driver and passengers, pedestrians, and other drivers on the road. A more effective mitigation should focus on real-time detection of the attack and the elimination of the “phantom” objects.

## 5. Related Work

Adversarial attacks on OD models have been studied extensively over the last few years [1]. Most of the previous studies focused on compromising the model’s integrity, with limited research examining attacks in which the model’s availability is targeted.

### 5.1. Integrity-Based Attacks

Initially, integrity-based adversarial attacks focused on image classification models (*e.g.*, FGSM [28], PGD [17]). Later, attacks against OD models were demonstrated. To evade the detection of Faster R-CNN [24], Chen *et al.* [5] printed stop signs that contained an adversarial pattern in the background. Sitawarin *et al.* [27] crafted toxic traffic signs, visually similar to the original signs. Thys *et al.* [29] first proposed an attack against person detectors, using a cardboard plate which is attached to the attacker’s body. In improved versions of this method, the adversarial pattern was printed on a T-shirt [31, 33].

Wang *et al.* [30] presented an adversarial attack on the NMS component; the goal of their attack was to increase the number of final predictions in the attacked image. While the authors focused on compromising the integrity of the model by adding a large number of objects to the final image prediction, we aim to attack the model’s availability (*i.e.*, in-

creasing the NMS processing time) by increasing the number of candidates processed by the NMS (while preserving the images’ detection). In addition, their perturbation is trained for each image (tailor-made perturbation), unlike our universal perturbation which is only trained once. As noted earlier, none of these studies on OD proposed methods that target the system’s availability.

### 5.2. Availability-Based Attacks

Availability-based attacks have only recently gained the attention of researchers, despite the fact that a system’s availability is a security-critical aspect of many applications. Shumailov *et al.* [26] were the first to present an attack (called *sponge examples*) targeting the availability of computer vision and NLP models; the authors demonstrated that adversarial examples are capable of doubling the inference time of NLP transformer-based models, with inference times  $6000\times$  greater than that of regular input. Boutros *et al.* [4] extended the sponge example attack so it could be applied on FPGA devices. In [3], the authors presented methods for creating sponge examples that preserve the original input’s visual appearance. Cina *et al.* [6] proposed *sponge poisoning*, a technique that performs sponge attacks during training, resulting in a poisoned model with decreased performance. Hong *et al.* [11] showed that crafting adversarial examples (including a universal perturbation) could slow down multi-exit networks.

Whereas the studies mentioned above targeted classification models in the computer vision domain, in this paper we focus on OD models, a target that has not been addressed by the research community. In addition, we propose a universal perturbation that is able to fool all images simultaneously. It should be noted that due to the diverse nature of images in the OD domain (*i.e.*, objects appear in different locations and at different scales on images), the ability to create a successful universal perturbation is challenging.

## 6. Conclusion

In this paper, we presented a UAP that substantially increases the inference time of the state-of-the-art YOLO object detector. This UAP adds “phantom” objects to the image while preserving the detections made by the OD for the original (unattacked) image. By demonstrating that YOLO is vulnerable to our attack, one can assume that the NMS algorithm in other OD models is also vulnerable and could be similarly attacked by applying our attack’s principles.

In future work, we plan to: (1) improve the attack by adding a technique that eliminates the “phantom” objects in the final prediction, making the attack less detectable, (2) move from the digital domain to real-world scenarios, for example, by placing a translucent patch on the camera lens (similar to [35]), and (3) develop a countermeasure capable of identifying the “phantom” objects in real time.



## References

- [1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.
- [2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
- [3] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. *arXiv preprint arXiv:2106.09898*, 2021.
- [4] Andrew Boutros, Mathew Hall, Nicolas Papernot, and Vaughn Betz. Neighbors from hell: Voltage attacks against deep learning accelerators on multi-tenant fpgas. In *2020 International Conference on Field-Programmable Technology (ICFPT)*, pages 103–111. IEEE, 2020.
- [5] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018.
- [6] Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Energy-latency attacks via sponge poisoning. *arXiv preprint arXiv:2203.08147*, 2022.
- [7] Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The mapillary traffic sign dataset for detection and classification on a global scale. In *European Conference on Computer Vision*, pages 68–84. Springer, 2020.
- [8] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.
- [9] Kevin Eykholt, Ivan Evtimov, Earleence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [11] Sanghyun Hong, Yiğitcan Kaya, Ionuț-Vlad Modoranu, and Tudor Dumitraș. A panda? no, it’s a sloth: Slowdown attacks on adaptive multi-exit neural network inference. *arXiv preprint arXiv:2010.02432*, 2020.
- [12] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 720–729, 2020.
- [13] Glenn Jocher. ultralytics/yolov5: v6.0 - yolov5n 'nano' models, roboflow integration, tensorflow export, opencv dnn support, oct 2021.
- [14] Ziv Katzir and Yuval Elovici. Who’s afraid of adversarial transferability? *arXiv preprint arXiv:2105.00433*, 2021.
- [15] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.
- [16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [17] Aleksander Madry, Aleksandar Makelev, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [18] Andreas Mogelmoose, Mohan Manubhai Trivedi, and Thomas B Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1484–1497, 2012.
- [19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.
- [20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [22] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.
- [23] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [26] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 212–231. IEEE, 2021.
- [27] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018.
- [28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

- [29] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [30] Derui Wang, Chaoran Li, Sheng Wen, Qing-Long Han, Surya Nepal, Xiangyu Zhang, and Yang Xiang. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples. *IEEE Transactions on Cybernetics*, 2021.
- [31] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.
- [32] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer, 2018.
- [33] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision*, pages 665–681. Springer, 2020.
- [34] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*, 2021.
- [35] Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15232–15241, 2021.