

Overload: Latency Attacks on Object Detection for Edge Devices

Erh-Chung Chen^{1*}

Pin-Yu Chen²

I-Hsin Chung²

Che-Rung Lee¹

National Tsing Hua University¹

s107062802@m107.nthu.edu.tw

cherung@cs.nthu.edu.tw

IBM Research²

pin-yu.chen@ibm.com

ihchung@us.ibm.com

Abstract

Nowadays, the deployment of deep learning-based applications is an essential task owing to the increasing demands on intelligent services. In this paper, we investigate latency attacks on deep learning applications. Unlike common adversarial attacks for misclassification, the goal of latency attacks is to increase the inference time, which may stop applications from responding to the requests within a reasonable time. This kind of attack is ubiquitous for various applications, and we use object detection to demonstrate how such kind of attacks work. We also design a framework named Overload to generate latency attacks at scale. Our method is based on a newly formulated optimization problem and a novel technique, called spatial attention. This attack serves to escalate the required computing costs during the inference time, consequently leading to an extended inference time for object detection. It presents a significant threat, especially to systems with limited computing resources. We conducted experiments using YOLOv5 models on Nvidia NX. Compared to existing methods, our method is simpler and more effective. The experimental results show that with latency attacks, the inference time of a single image can be increased ten times longer in reference to the normal setting. Moreover, our findings pose a potential new threat to all object detection tasks requiring non-maximum suppression (NMS), as our attack is NMS-agnostic.

1. Introduction

Deep neural networks (DNN) have accomplished many achievements in various fields [1, 2, 12, 17, 34]. However, those models are usually large and demand a huge amount of computing resources, even for model inference. One solution is to utilize the computing power in cloud platforms, but the communication cost between the data center and the edge nodes can be high and the latency may not be accept-

able for many real-time applications [10, 24].

In this paper, we investigate a new type of attacks, called latency attacks, whose purpose is to increase the execution time of the victim application. This kind of attacks poses critical threats to real-time applications, such as purchase behavior recognition system for unmanned store or autonomous cars systems, which are required to detect target objects and determine proper actions within stringent time constraints [23, 32]. Any mistake or delayed response due to elongated latency could cause severe failures. Moreover, the latency attacks not only increase the execution time but also cause improper predictions made by the model indirectly.

The task we focus on is object detection, which has been widely used in numerous applications [11, 36, 38]. Its purpose is to identify all objects in an image and to label them with corresponding classes and locations. Many deep learning-based models, such as SSD [13] and YOLO series [22], often use Non-Maximum Suppression (NMS) as a post-processing operation to eliminate duplicated objects predicted by the detectors. The execution time of NMS depends on the number of objects fed into NMS. This property implies that object detection tasks may suffer from severe latency if numerous objects are predicted by the target model.

In this paper, we analyze the time complexity of NMS and find that the execution time is dominated by the total number of objects. The influence of the number of survival bounding boxes can be ignored. Based on the observation, we propose a Latency Attack on Object Detection (we name this attack framework “Overload”), whose objective is to craft adversarial images with abundant objects from the victim model. To further improve the effectiveness of the adversarial image, we introduce a technique, called spatial attention, which can be used to manipulate the fake object generation in particular regions. This attack serves to escalate the required computing costs during the inference time, consequently leading to an extended inference time for object detection. It presents a significant threat, especially to systems with limited computing resources.

*The primary research and contribution for this work were conducted during a visit to IBM Research.

Our main contributions are outlined as follows:

- We systematically explore potential objectives for latency attacks, identifying the number of objects fed into NMS as a crucial factor for success, both theoretically and practically.
- Leveraging this insight, we simplify the objective design. In comparison to existing works like [33, 35], Overload achieves superior attack performance while utilizing fewer computational costs and less memory.
- Experiments on Nvidia Jetson NX demonstrate that the inference time of our crafted images is approximately 10 times longer than that of the original images.
- We establish the NMS-agnostic nature of our proposed attack, underscoring its potential to pose a universal threat to all object detection systems reliant on NMS.

The rest of this paper is organized as follows. Section 2 introduces the background on object detection and adversarial attack. Section 3 presents the theoretical analysis of NMS and the proposed methods to craft the adversarial images. Section 4 shows the experimental results, and discussions on the impact of latency attacks for real-time applications. The last section is our conclusion. Analytical details of NMS and ablation studies are listed in the Appendix.

2. Background

2.1. Object Detection

Object detection is one of the challenging tasks in computer vision. In contrast to image classification, which receives an image and predicts its class, the goal of object detection is to recognize multiple objects and label their locations. To accomplish this, various deep learning-based detectors have been introduced. Some models approach this task by processing object locations and their probabilities in two distinct stages [14, 18]. Conversely, one-stage detectors tackle both classification and object localization within a single neural network [22].

To deploy those models onto edge devices, model size and inference time should be improved to meet the resource and timing constraints. Model compression is a commonly used strategy [30]. The YOLO family utilizes model scaling techniques to make the model fit different computing devices. Different scaling factors, such as resolution (size of the input image), depth (number of layers), width (number of channels), and stage (number of feature pyramids) are considered. For example, YOLO 5 has five different scales of models that use the same architecture but different width and depth configurations.

2.2. Non-maximum Suppression

Non-maximum suppression (NMS) is a prevalent component in modern object detection tasks [14, 22]. As shown in Figure 1, an object detection model predicts numerous

objects. Each object has its own information about height, width, position, objectness confidence, and probabilities of all classes. The objects with similar position information should be clustered as the same object. The main purpose of NMS is to eliminate redundant objects in each cluster and return survival objects as final bounding boxes. Several NMS implementations have been proposed, such as DIOU [40] and CIOU [40].

While NMS is essential to most deep-learning-based models, there has been limited investigation into the elapsed time of NMS. This is primarily due to the fact that benchmarking datasets are typically clean and uncontaminated, and model performance is evaluated on high-end GPUs, where the time taken by NMS is negligible. Previous studies have mostly focused on optimizing performance through architecture refinements [22]. However, when deploying models on edge devices, the computational cost of NMS becomes a critical consideration. Moreover, in this paper, we demonstrate that NMS can be exploited to launch latency attacks on object detection systems for edge devices.

2.3. Adversarial Attack

Adversarial attacks have demonstrated that state-of-the-art models can be deceived by an adversarial example, which is the original input with an indistinguishable perturbation [5, 15, 16]. Adversarial attacks can be classified as white-box attacks [6, 7] or black-box attacks [8, 9]. In the white-box scenario, adversarial examples can be found along the gradient of a loss function. In the black-box scenario, the victim model is inaccessible. The gradient is often estimated by the finite differences method with multiple queries with stochastic perturbations.

Unlike the adversarial attack for image classification, whose goal is to degrade the model accuracy, the objectives of adversarial attacks for object detection are diverse. There are at least four objectives in current literature: appearing attack, hiding attack, mis-classifying attack, and mis-locating attack. Appearing attacks generate a perturbation such that the detector marks non-existent objects [39]. Hiding attacks make particular objects invisible [37]. Mis-classifying attacks do not modify the information about the location but the predicted class [3]. Mis-locating attacks focus on applications that require the locations of predicted bounding boxes to be highly precise. Tiny location changes can cause catastrophe failures [21].

2.4. Latency Attack

Latency attacks are a variant of adversarial attacks, with their primary objective being to maximize the processing time of the target model when presented with inputs. The Sponge attack, for instance, showcased the feasibility of latency attacks against deep learning-based translators operating on fixed-length sequences. In some dynamic vision

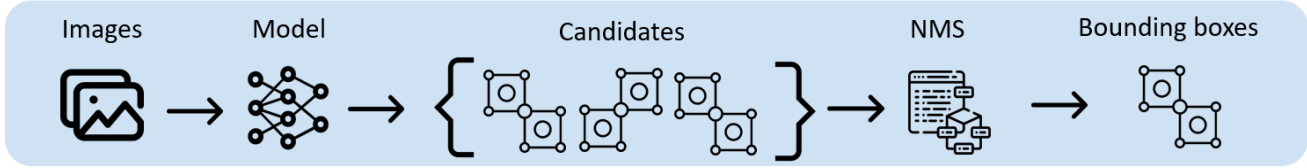


Figure 1. The processing flow of object detection. NMS stands for non-maximum suppression.

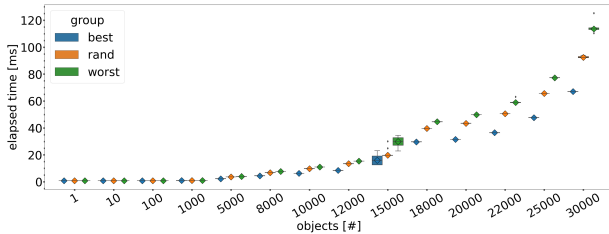


Figure 2. Elapsed time of NMS on NVIDIA Jetson NX.

models, predictions can be obtained from earlier stage if specific criteria are met [25]. This behavior introduces a security concern regarding latency attacks [19].

The vulnerability of NMS was addressed by Daedalus attack [35], which aims to produce numerous bounding boxes and to reduce the mean average precision (mAP), a commonly used metric for object detection tasks. Another work is Phantom Sponge [33], which demonstrates that increasing of the total number of objects fed into NMS can cause a longer execution time. It also tries to keep the recall of the modified image the same as that of the original image. Nonetheless, these approaches incur high computational costs due to the necessity of evaluating the gradients of all IoU pairs. Moreover, there is a lack of a comprehensive investigation into the impact of these attacks on edge devices.

3. Overload: A Framework of Latency Attacks

In this section, we first present the formal definition of latency attacks on object detection tasks and our motivations. We further provide theoretical explanations of how the adversarial examples are produced with minimal computational cost. Lastly, we propose an algorithm to generate adversarial images for latency attacks.

3.1. Problem Definition and Motivation

We assess potential vulnerabilities in object detectors within the white-box scenario, where the attackers have complete knowledge of the model architecture and can directly retrieve intermediate information from the victim model. Although this setting seems impractical in the real-world, it can serve as motivation to refine more robust model designs.

The main goal of the latency attack is to find a perturbation joined to the input such that expected values of the elapsed time of every single request are maximized, which can be formulated as the following optimization problem:

$$\mathbb{E}_{x \sim \mathcal{X}} \max_{\delta \leq \epsilon} T_{\text{total}}(x + \delta), \quad (1)$$

where \mathcal{X} is a set of requests to be processed and the perturbation δ is within the ϵ -ball. Specifically, the elapsed time can be expressed as

$$T_{\text{total}}(x + \delta) = T_{\text{infer}}(x + \delta) + T_{\text{wait}}(\mathcal{X}), \quad (2)$$

where $T_{\text{infer}}(x + \delta)$ is the inference time and T_{wait} is the waiting time, which depends on the availability of the desired resources and all requests \mathcal{X} . We assume there is a queue for the requests. When the desired computing resources are fully occupied, requests will be stored in the queue. Therefore, the goal defined in Equation (1) is equivalent to finding adversarial examples that could exhaust the desired resources so that the rest of the requests in the queue cannot be processed until the malicious requests are terminated.

However, directly incorporating the function T into the formulation for generating adversarial attacks poses a challenge, because T is a complex function that can be influenced by several external factors, such as scheduling or cache missing. In the case of white-box attacks, computing the derivative directly through the computational graph is not feasible. Moreover, for black-box attacks, discerning whether the increase in processing time stems from changes in machine status or the introduction of an adversarial example is non-trivial. In this paper, we seek to address the following two critical questions:

- How can we parameterize the elapsed time and the outputs generated by the target model, allowing us to formulate an objective that can be exploited to launch latency attacks on object detection?
- How can we maximize the elapsed time of the victim model in processing the given image, all while minimizing computational costs and memory usage in the production of adversarial examples?

3.2. Theoretical Analysis of NMS Time Complexity

A fundamental requirement for a successful latency attack is the existence of an operation whose processing time is

contingent on the input. In the case of object detection tasks, NMS emerges as the pivotal factor. To delve into the vulnerability of NMS in object detection to latency attacks, we first analyze the time complexity of NMS based on the commonly used deep learning frameworks on GPU, as suggested in [31]. Let \mathcal{C} be the set of objects fed into NMS. The steps of NMS can be divided into four major tasks. We leave the detailed analysis in ?? and summarize their time complexity below:

1. Filtering low-confidence objects: $O(|\mathcal{C}|)$;
2. Sorting candidates by probabilities: $O(|\mathcal{C}| \log |\mathcal{C}|)$;
3. Calculating pairwise IoU scores: $O(|\mathcal{C}|^2)$;
4. Pruning inactive objects: $O(|\mathcal{C}|^2)$.

As evident, the time complexity appears to be quadratic. However, in practice, the time is dominated by unavoidable overheads when $|\mathcal{C}|$ is relatively small. Hence, the elapsed time can be approximated as

$$T_{\text{total}} \approx \begin{cases} \alpha|\mathcal{C}|^2, & \text{if } |\mathcal{C}| > N \\ T_{\text{base}}, & \text{otherwise,} \end{cases} \quad (3)$$

where α is the calibrated coefficient depending on box distributions, T_{base} , and N are constants related to the computational capacity of used GPU.

To ascertain whether the time complexity is affected by the number of output boxes, we conducted experiments to measure the elapsed time of NMS using synthetic data under three distinct scenarios: worst, best, and random. The worst case is simulated by setting T_{IoU} to 1.0 so that no object can be filtered out after NMS. On the other hand, the best scenario is making all elements in set \mathcal{C} share the identical box information, so that NMS can mark all objects in the first iteration. The random scenario just randomly sets the box configurations.

The experimental results are illustrated in Figure 2, and two key observations can be drawn from the findings. First, as anticipated, the processing time in all three cases displays a quadratic growth pattern as the size of $|\mathcal{C}|$ increases. Second, the impact of the number of output boxes can be represented by the ratio of the worst-case scenario to the best-case scenario, which remains a constant, regardless of the number of objects ($|\mathcal{C}|$). Therefore, we can formulate the elapsed time of NMS as

$$T_{\text{total}} = s(1 + \frac{|\mathcal{B}|}{|\mathcal{C}|})|\mathcal{C}|^2 = s(|\mathcal{C}|^2 + |\mathcal{B}||\mathcal{C}|), \quad (4)$$

where s is the calibrated coefficient and \mathcal{B} is the set of boxes output by NMS.

3.3. Exploring Objectives for Adversarial Example Generation

Our primary objective is to maximize the elapsed time of the victim model when processing a given image, all while

minimizing computational costs and memory usage when creating adversarial examples. Equation (4) implies that latency can be increased by generating more objects or boxes, but the associated cost of producing adversarial examples remains unclear.

In pursuit of our goal, we aim to explore all potential objectives for generating adversarial examples. Previous works [33, 35] suggested that the latency attack can be achieved by a composite objective that involves three components: maximizing the confidence of objects, minimizing the pairwise IoU scores, and minimizing the areas of objects. The first terms can prevent the objects from being filtered at the first step of NMS. The second term drifts the centers of objects and the third term adjusts the areas of boxes to avoid the removal of objects in the pruning process.

However, the objective proposed in previous works overlooks the influence of the number of boxes. The observations in Figure 2 indicate that the elapsed time can be increased by up to 20 times when generating numerous objects. On the other hand, the time increment is approximately 2 times when the number of boxes is increased with a fixed $|\mathcal{C}|$.

Moreover, while the IoU computation is differentiable, tracking all pairwise comparisons consumes considerable memory and time for crafting adversarial examples. Furthermore, the performance improvement cannot be easily quantified, as the execution order influences the final results. For example, a modified box can evade the discarding of objects but it may cause some nearby boxes to be marked duplicated in the rest pruning rounds.

This insight reinforces our emphasis on prioritizing maximum object confidence as crucial for optimizing latency while minimizing computational costs and memory usage. We contend that including the last two terms in the objective not only fails to improve performance but also increases the time and memory required for generating adversarial examples, adding unnecessary complexity to the design.

3.4. Proposed Method for Latency Attacks on Object Detection

We propose that latency attacks on object detection can be formulated as an optimization problem,

$$\max_x \sum_{i=0}^n F_l(F_{\text{conf}}(M(x)_i)), \quad (5)$$

where $M(x)$ is the predicted results; n is the total number of objects output by the given model, $F_l(\cdot)$ is a monotonic increasing function; and $F_{\text{conf}}(\cdot)$ retrieves each object's corresponding confidence. Specifically,

$$F_{\text{conf}}(\cdot) = \begin{cases} c, & \text{if } cp_i > T_{\text{conf}} \\ cp_i, & \text{otherwise,} \end{cases} \quad (6)$$

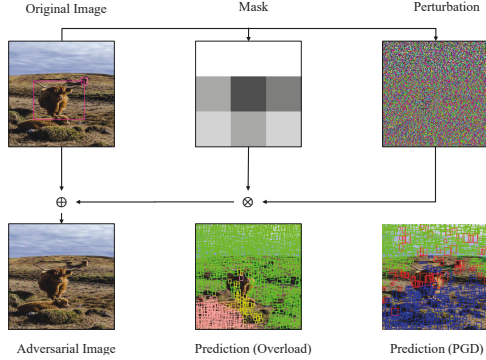


Figure 3. The execution flow of spatial attention.

where c is the objectness confidence predicted by the model and p_i is the probability of the i -th class. Equation (6) maximizes the probability of each class by increasing the objectness confidence while the probabilities for less possible classes are raising as well to accelerate search speed since multiple labels for an object are allowed in an object detection task. As a result, the total objects fed into NMS can be increased.

We introduce a novel approach, called spatial attention, to enhance the efficacy of latency attacks. The purpose of spatial attention is to force the attacking processing to pay more attention to the low-density regions rather than the regions that already have objects. The flow of spatial attention is visualized in Figure 3. Throughout each iteration for crafting adversarial examples, the weight assigned to the central area is minimized as it contains two objects. Conversely, boundaries, which are either empty or overlaid with objects, have their corresponding weights adjusted accordingly.

We summarize the steps of our method below. First, the image is divided into an $m \times m$ grid. Second, the weight of each grid cell, denoted as $W_{i,j}$, is adjusted based on the number of objects obtained from $M(x)$. The weight of the particular grid cell is decreased if adversarial attacks cannot produce more objects in the area or if it is dense enough. Finally, the adversarial image is updated using the following equation:

$$x_{i,j}^{\text{adv}} = x_{i,j}^{\text{org}} + \eta W_{i,j} \Delta x_{i,j} \quad (7)$$

where i and j are grid indices; x^{adv} is the adversarial image; x^{org} is the original image; η is step size, and Δx is the modification of the image. Here $\Delta x = \nabla_x L$, the gradient for the loss function specified in Equation (5), where

$$L = \sum_{i=0}^n F_i(F_{\text{conf}}(M(x)_i)).$$

4. Experiments

4.1. Setup

To demonstrate the potential impacts of latency attacks on systems with limited computing resources, we conducted experiments on Nvidia Jetson NX, a popular edge device for real-time object detection applications [27, 32], and the used models are YOLOv5s and YOLOv5n due to their low computational requirements. To further accelerate the running speed, the inference models are compiled into the TensorRT format [20]. We randomly selected 1,000 images from the validation set of MS COCO 2017 dataset [28]. Meanwhile, the batch size is 1 and the images' dimension is (640, 640). The results are sorted by the elapsed time and presented in the selected percentile. Additionally, we have included ablation studies in the Appendix, which covers various aspects such as the selection of the loss function, evaluation of spatial attention, transfer attack analysis, and the performance impact on different models. The results obtained indicate that the total number of objects generated by our proposed attack closely aligns with the theoretical maximum that victim models can output, demonstrating the effectiveness of our approach. Furthermore, our method allows the encoding of information from multiple models within a single image, showcasing the feasibility of an ensemble attack to deceive various object detection models.

4.2. Latency Evaluation

Table 1 presents the experimental results of YOLOv5s and YOLOv5n on Nvidia Jetson NX, where *objects* mean the total number of objects fed into NMS; *boxes* mean the total number of bounding boxes output by NMS, and *time* is the elapsed time in millisecond, including model inference and NMS.

The elapsed times of the original examples with YOLOv5s and YOLOv5n are about 16.4 ms and 11.5 ms respectively for 50% percentile. Conversely, the average elapsed times of adversarial examples are about $13.0\times$ and $11.0\times$ longer than that of the original examples with YOLOv5s and YOLOv5n respectively. Under our experimental configuration, the maximum number of objects predicted by the YOLO model is 25,200. Results also demonstrate that our attacks can generate over 20,000 ghost objects for most images: over 90% of images for YOLOv5s and over 60% of images for YOLOv5n.

To evaluate the impact of the latency attack on real-world applications, we measured the elapsed time of 1,000 images processed by a pipeline. Table 2 shows the average execution time of each image in milliseconds, where the Ratio column represents the proportion of adversarial images in the test dataset, and FPS is the abbreviation for "Frame Per Second". As the proportion of adversarial images grows, the elapsed time rises rapidly and the inference time in-

Percentile	YOLOv5s						YOLOv5n					
	Adversarial Examples			Original Examples			Adversarial Examples			Original Examples		
	objects	boxes	time	objects	boxes	time	objects	boxes	time	objects	boxes	time
min	4098	1587	18.7	0	0	14.1	5893	1131	14.7	0	0	9.3
0.10	23112	2034	122.2	40	3	16.4	11925	1421	22.6	0	0	10.7
0.25	24248	4653	179.6	47	4	16.4	16405	2303	39.8	6	1	11.5
0.50	24000	2250	217.8	28	3	16.4	22363	2315	128.9	22	4	11.5
0.75	24335	4514	264.2	3	1	16.6	20923	2372	208.3	97	8	11.5
0.90	24778	2340	278.4	234	21	16.7	21947	2659	226.7	6	2	11.6
max	24859	2363	293.8	212	21	16.9	23396	2502	258.7	25	2	12.0

Table 1. Results of latency evaluation on NVIDIA Jetson NX

Ratio	YOLOv5s		YOLOv5n	
	time [ms]	FPS	time [ms]	FPS
0.00	23	43.5	20	50.0
0.25	81	12.3	56	17.8
0.50	136	7.3	92	10.8
0.75	180	5.5	114	8.8
1.00	230	4.3	135	7.4

Table 2. Pipeline simulator on NVIDIA Jetson NX

GPU model	T_{org} [ms]	T_{atk} [ms]
Nvidia A100	16	54
Nvidia A10	14	52
Nvidia 1080 Ti	11	62
Nvidia Jetson	23	230

Table 3. Latency evaluation across various GPUs

creases tenfold in the worst case. The results suggest that our proposed latency attack poses a potential warning in real-world applications.

We conducted the same pipeline experiment with YOLOv5s across various GPUs. Table 3 presents the experimental results, where T_{org} is measured without the involvement of adversarial examples; T_{atk} refers to the scenario where all inputs are adversarial examples. As can be seen, the ratio of the elapsed time of T_{atk} to T_{org} is about 3-6 (excluding Nvidia Jetson). These findings suggest that the attack can be generalized across different GPUs, but the strength of the proposed attack depends on hardware configurations.

4.3. IoU-invariant Verification

This experiment investigates how the number of surviving objects influences execution time. The results are presented in Table 4, where T_{IoU} represents the IoU threshold. $T_{IoU} = 1.0$ and $T_{IoU} = 0.0$ simulate the worst and best

cases, respectively. $T_{IoU} = 0.45$ represents the default configuration, and $T_{IoU} = 0.80$ is an intermediate case.

As indicated in Table 4, our attack produces approximately 550 and 600 non-overlapping boxes for YOLOv5s and YOLOv5n, respectively, at the 50% percentile. Although the best cases reduce over 95% of objects, the corresponding elapsed times are quite similar to those of the other cases. When compared to the results obtained from synthetic data, as illustrated in Figure 2 the latency increments for real-world data are marginal, owing to the variability in the strength of adversarial examples. Producing more survival boxes to increase execution time is only recommended for black-box scenarios or situations where the number of objects is saturated.

4.4. Comparison to Known Attacks

Phantom Sponge [33] and Daedalus [35] generate numerous objects by reducing the bounding box sizes, implicitly minimizing pairwise IoU scores. This operation is equivalent to maximizing the term $|\mathcal{B}|$ in Equation (4). However, they necessitate tracking the gradient of IoU computations, leading to high memory usage and longer processing times without inducing performance gains. In contrast, the proposed method simplifies the objective, conserving computational time.

Table 5 compares Overload to existing works, where ϵ is the maximum size of the perturbation; Objects represent the total number of objects; Recall measures the similarity of predictions between clean and adversarial images; mAP@50 is the mean average precision calculated at IoU threshold 0.5; and "-" refers to metrics that are not applicable. For the L2-norm attack, Daedalus produces fewer objects than other competitors. Meanwhile, Phantom Sponge generates about 19,000 objects within a perturbation size $\epsilon = 70.0$, but our attack produces more objects and higher recall within $\epsilon = 15.0$. These comparison results should not be over-interpreted, as the goals and constraints on perturbations for Phantom Sponge and Daedalus differ from ours.

We conducted experiments on Nvidia V100 (32GB) to

YOLOv5s												
Percentile	$T_{IoU}=0.00$			$T_{IoU}=0.45$			$T_{IoU}=0.80$			$T_{IoU}=1.00$		
	objects	boxes	time	objects	boxes	time	objects	boxes	time	objects	boxes	time
min	4098	403	17.3	4098	1587	18.7	4098	2198	17.9	4098	4098	19.3
0.10	23635	623	64.8	23112	2034	122.2	20157	6352	156.9	20328	20328	125.7
0.25	24027	745	120.2	24248	4653	179.6	23345	7756	211.5	23476	23476	215.8
0.50	20579	550	200.0	24000	2250	217.8	24345	12504	234.7	24381	24381	245.4
0.75	24264	692	265.6	24335	4514	264.2	24667	7544	248.0	24138	24138	275.5
0.90	24268	677	274.5	24778	2340	278.4	24338	11593	260.4	23356	23356	280.2
max	24810	476	289.2	24859	2363	293.8	24859	8039	269.7	24905	24905	285.5

YOLOv5n												
Percentile	$T_{IoU}=0.00$			$T_{IoU}=0.45$			$T_{IoU}=0.80$			$T_{IoU}=1.00$		
	objects	boxes	time	objects	boxes	time	objects	boxes	time	objects	boxes	time
min	5893	393	14.3	5893	1131	14.7	5893	2429	15.2	5893	5893	15.3
0.10	11660	533	21.9	11925	1421	22.6	11904	4580	24.1	11961	11961	25.4
0.25	16858	529	31.3	16405	2303	39.8	16680	5008	44.4	14789	14789	48.3
0.50	21423	660	52.1	22363	2315	128.9	21440	10020	60.5	19231	19231	139.2
0.75	22905	540	62.6	20923	2372	208.3	21876	11360	152.4	21748	21748	210.8
0.90	21536	562	163.2	21947	2659	226.7	22279	10305	223.7	21796	21796	225.9
max	22897	538	234.7	23396	2502	258.7	22959	9505	241.1	23396	23396	256.2

Table 4. The elapsed time with different IoU threshold on NVIDIA Jetson NX

Paper	ϵ	Objects	Recall	mAP@50
Phantom Sponge	70	19000	18	-
Phantom Sponge	30	9000	77	-
Overload	15	20000	44	<0.001
Daedalus	-	1500	-	<0.001

Table 5. Performance comparison with existing works

Method	BatchSize	Memory (GB)	Time (s)
Overload	16	7.5	180
Overload + L_{IoU}	16	-	-
Overload + L_{area}	16	7.9	240

Table 6. Resource Consumption Comparison for Overload with Different Losses

quantify memory usage and the total time required to craft adversarial examples. The gradients of tracking IoU calculation for all objects require more memory than is available in Nvidia V100. Consequently, Daedalus tracks partial objects to mitigate memory usage. Crafting 10 adversarial examples sequentially with Daedalus takes approximately 145 minutes, with a peak memory usage of around 8.5GB. In contrast, our method can process 16 images in a batch in about 3 minutes, with a peak memory usage of 7.5GB. Phantom Sponge limits the number of objects used to compute the IoU score to less than 100, resulting in slightly

higher resource usage than our method under the same conditions. However, Phantom Sponge needs to execute NMS several times in the attacking processing, making a longer execution time when the total objects grow.

Table 6 represents the evaluation of Overload integrated with an additional loss, where "Overload + L_{IoU} " refers to the objective involving minimizing pairwise IoU scores, and "Overload + L_{area} " refers to the objective involving minimizing the areas of objects. "Overload + L_{IoU} " failed to execute due to out-of-memory issues. The resource requirements for "Overload + L_{area} " are higher than native Overload. We attempted various function formulations for minimizing areas and coefficients to amplify their influence. However, we obtained the same performance as those presented in IoU-invariant verification. Therefore, we believe that Overload outperforms these works in terms of the number of objects without needing a hyperparameter for confidence-IoU balance, while using fewer computing resources and simplifying the objective design.

4.5. Discussion

In spirit, our proposed latency attack can be viewed as a variant of the appearing attack, but these two attacks have fundamentally different purposes. The original appearing attack [39] creates a specific type of object such that the detector makes an unexpected decision. On the contrary, the latency attack produces numerous objects and the detector cannot process those objects within the strict time restric-

tion. Nevertheless, it is important to acknowledge certain limitations and potential impacts associated with our work, which are addressed below.

NMS-agnosticity We argue that our attack is a generic attack against NMS. The objective of the latency attack is to maximize the confidence of individual object without any information about locations. DIoU [40], CIoU [40], and others implementations refine the pruning criterion or scoring function. Those modifications do not affect the time complexity. Hence, any object detection task that requires NMS as post-processing is a potential victim of our latency attack.

Attack scenarios This kind of adversarial image requires more computing resources than normal images during the inference time. This implies it can be leveraged as a Denial-of-Service (DoS) attack and resources may be exhausted in shared-resource environments. Furthermore, devices with limited computing resources are potential attack candidates. Although some models implement a timeout mechanism, once the execution time is exhausted after processing any example in the batch, the rest of the examples in the batch are skipped. Additionally, issuing the next batch is delayed until the previous processing is complete. This property could be exploited maliciously.

The ghost objects generated by our latency attack could increase processing time or have unforeseen consequences for downstream tasks. For instance, the collision avoidance system identifies risky objects and predicts the trajectories of those objects [26]. The execution time worsens if the predictions of an adversarial image are fed into the collision avoidance system.

Limitation Our evaluation of this attack is under white-box settings. We acknowledge that conducting adversarial attacks in this setting may seem impractical. However, they can serve as motivation to refine more robust model designs. Generating a physically universal patch that can increase inference time under black-box settings represents a critical area of investigation. Nonetheless, evaluating practicality falls beyond the scope of this work.

While our attack cannot be directly applied to NMS-free models like DETR [4] and CenterNet [41], it's vital not to disregard the potential impact on these models. However, a recent benchmark [29] evaluated popular underwater object detection models on edge devices, revealing that 11 out of 12 models still rely on NMS. Therefore, our work is important in drawing attention to the potential risks in object detection and inspiring researchers to explore mitigation techniques applicable across various applications.

Defenses There are three probable defenses. The first one limits the maximum execution time. However, the predefined threshold should be reconfigured when the model is deployed on different hardware. Besides, the previous discussion has shown the risks of timeouts. The second one

limits the total number of objects fed into NMS. The choice of the appropriate value depends on the used model and the specific dataset. A strict limitation can mitigate the latency issue but it may cause some objects to be ignored by the model. The last one directly strengthens the reliability of models with adversarial training and other defensive strategies.

5. Conclusion

In this paper, we proposed a novel latency attack, called Overload, by targeting the NMS module in object detection tasks. To fully exploit the vulnerabilities, we have analyzed the time complexity of NMS on GPUs, and observed that the execution time is dominated by the number of objects fed into NMS. Based on the analysis, we proposed a new formulation to craft adversarial images for latency attacks, and a new technique, called spatial attention, to activate more objects in particular areas to improve attack efficiency. Compared with existing works, Overload achieves superior attack performance but with smaller computational costs and memory usage. With the attack, the execution time of a single adversarial image is about ten times longer than that of a normal image on Nvidia Jetson NX. It poses a significant threat to systems with limited computing resources. Furthermore, we establish the NMS-agnostic nature of our proposed attack, highlighting its potential to become a common vulnerability to all object detection systems that rely on NMS.

There are many future directions to explore. First, the black-box latency attacks have not been studied yet. Spatial attention is a gradient-free procedure, so it may be a useful tool for black-box attacks. Second, the performance impact of the latency attack for downstream tasks needs more investigation. Lastly, developing defense strategies against latency attacks is an important topic but beyond the scope of this paper. We acknowledge that our findings might be used as exploits. However we believe our disclosure of vulnerability analysis is necessary to accelerate the design of more resilient and safe edge devices against latency attacks.

Acknowledgments.

This research is partially supported by NTSC, Taiwan, R.O.C. under Grant no. 112-2634-F-007-002. We would also like to thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

- [1] Katherine L Bouman, Michael D Johnson, Daniel Zoran, Vincent L Fish, Sheperd S Doleman, and William T Freeman. Computational imaging for vlsi image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 913–922, 2016. 1

- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1
- [3] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 149–157, 2022. 2
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 8
- [5] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018. 2
- [6] Erh-Chung Chen and Che-Rung Lee. Towards fast and robust adversarial training for image classification. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2
- [7] Erh-Chung Chen and Che-Rung Lee. Ltd: Low temperature distillation for robust adversarial training. *arXiv preprint arXiv:2111.02331*, 2021. 2
- [8] Pin-Yu Chen and Sijia Liu. Holistic adversarial robustness of deep learning models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 2
- [9] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. 2
- [10] Yueyue Dai, Du Xu, Sabita Maharjan, Guanhua Qiao, and Yan Zhang. Artificial intelligence empowered edge computing and caching for internet of vehicles. *IEEE Wireless Communications*, 26(3):12–18, 2019. 1
- [11] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. 1
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1
- [13] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Amrith Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 1
- [14] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015. 2
- [15] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019. 2
- [16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1
- [18] 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. 2
- [19] 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. 3
- [20] EunJin Jeong, Jangryul Kim, and Soonhoi Ha. Tensorrt-based framework and optimization methodology for deep learning inference on jetson boards. *ACM Transactions on Embedded Computing Systems (TECS)*, 2022. 5
- [21] Yunhan Jia Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei Wei. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations (ICLR’20)*, 2020. 2
- [22] Glenn Jocher. YOLOv5 by Ultralytics, 2020. 1, 2
- [23] Dae Ha Kim, Seunghyun Lee, Jungho Jeon, and Byung Cheol Song. Real-time purchase behavior recognition system based on deep learning-based object detection and tracking for an unmanned product cabinet. *Expert Systems with Applications*, 143:113063, 2020. 1
- [24] Xiangjie Kong, Yuhan Wu, Hui Wang, and Feng Xia. Edge computing for internet of everything: A survey. *IEEE Internet of Things Journal*, 9(23):23472–23485, 2022. 1
- [25] Chi-Hsi Kung and Che-Rung Lee. Add: A fine-grained dynamic inference architecture for semantic image segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4792–4799. IEEE, 2021. 3
- [26] Chengxi Li, Stanley H Chan, and Yi-Ting Chen. Driver-centric risk object identification. *arXiv preprint arXiv:2106.13201*, 2021. 8
- [27] Siyuan Liang, Hao Wu, Li Zhen, Qiaozhi Hua, Sahil Garg, Georges Kaddoum, Mohammad Mehdi Hassan, and Keping Yu. Edge yolo: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 5
- [28] 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. 5
- [29] Chongwei Liu, Haojie Li, Shuchang Wang, Ming Zhu, Dong Wang, Xin Fan, and Zhihui Wang. A dataset and benchmark of underwater object detection for robot picking. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 8
- [30] Fangxin Liu, Wenbo Zhao, Zhezhi He, Yanzhi Wang, Zongwu Wang, Changzhi Dai, Xiaoyao Liang, and Li Jiang. Improving neural network efficiency via post-training quantization with adaptive floating-point. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 5281–5290, 2021. 2
- [31] TorchVision maintainers and contributors. Torchvision: Pytorch’s computer vision library. <https://github.com/pytorch/vision>, 2016. 4
- [32] Niranjana Ravi and Mohamed El-Sharkawy. Real-time embedded implementation of improved object detector for resource-constrained devices. *Journal of Low Power Electronics and Applications*, 12(2):21, 2022. 1, 5
- [33] Avishag Shapira, Alon Zolfi, Luca Demetrio, Battista Biggio, and Asaf Shabtai. Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4571–4580, 2023. 2, 3, 4, 6
- [34] Jeffrey Van der Gucht, Jordy Davelaar, Luc Hendriks, Oliver Porth, Hector Olivares, Yosuke Mizuno, Christian M Fromm, and Heino Falcke. Deep horizon: A machine learning network that recovers accreting black hole parameters. *Astronomy & Astrophysics*, 636:A94, 2020. 1
- [35] 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. 2, 3, 4, 6
- [36] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1
- [37] 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. 2
- [38] Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology*, 11:638182, 2021. 1
- [39] Mingjun Yin, Shasha Li, Chengyu Song, M Salman Asif, Amit K Roy-Chowdhury, and Srikanth V Krishnamurthy. Adc: Adversarial attacks against object detection that evade context consistency checks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3278–3287, 2022. 2, 7
- [40] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12993–13000, 2020. 2, 8
- [41] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 8