New Threats against Object Detector with Non-local Block

Yi Huang[0000−0002−4920−4333], Fan Wang[0000−0001−8582−1673], Adams Wai-Kin Kong[0000−0002−9728−9511], and Kwok-Yan Lam[0000−0001−7479−7970]

Nanyang Technological University , Singapore
{S160042, fan005}@e.ntu.edu.sg, {adamskong, kwokyan.lam}@ntu.edu.sg

Abstract. The introduction of non-local blocks to the traditional CNN architecture enhances its performance for various computer vision tasks by improving its capabilities of capturing long-range dependencies. However, the usage of non-local blocks may also introduce new threats to computer vision systems. Therefore, it is important to study the threats caused by non-local blocks before directly applying them on commercial systems. In this paper, two new threats named disappearing attack and appearing attack against object detectors with a non-local block are investigated. The former aims at misleading an object detector with a non-local block such that it is unable to detect a target object category while the latter aims at misleading the object detector such that it detects a predefined object category, which is not present in images. Different from the existing attacks against object detectors, these threats are able to be performed in long range cases. This means that the target object and the universal adversarial patches learned from the proposed algorithms can have long distance between them. To examine the threats, digital and physical experiments are conducted on Faster R-CNN with a non-local block and 6331 images from 56 videos. The experiments show that the universal patches are able to mislead the detector with greater probabilities. To explain the threats from non-local blocks, the reception fields of CNN models with and without non-local blocks are studied empirically and theoretically.

Keywords: Non-local block, adversarial examples, object detection

1 Introduction

Convolutional neural networks (CNNs) have been becoming an essential component in computer vision systems where many of them have been deployed commercially. Traditional CNNs are built on local operators, such as convolution and pooling. In order to extract information in a wide area, the local operators are stacked, resulting in larger receptive fields in theory. However, Luo *et al.* pinpointed that the effective receptive fields of CNNs are much smaller than their theoretical receptive fields [20]. Furthermore, stacking the local operators, especially convolution, would dramatically increase the computational cost and cause optimization difficulties [8]. To address these issues, Wang *et al.* generalized the

classical non-local mean operator for image denoising [1] and proposed non-local blocks [31] in 2018. Since the non-local blocks can be easily inserted into many existing architectures and combined with other operators, they have been taken as a generic family of building blocks in CNNs for capturing long-range dependencies. In fact, the non-local neural network has raised a new trend in computer vision with an outstanding performance in object detection [29,35,13,25], as well as other various techniques on this basis, such as action recognition [24,9] and person re-identification [16]. There are nearly 200 related articles published in top computer vision venues during the past two years. Non-local blocks have also been applied on other research fields, such as computer-assisted radiology and surgery, bioinformatics, electronic and automation control, speech processing, *etc.* With it being a heated topic, and its potentials, it has also drawn attention from industry and companies, including Facebook [31], Microsoft [2], Baidu [36], Tencent [28,6], Huawei [5], Face++ [3], *etc.*

However, massively applying the non-local blocks in commercial systems without studying their threats would be risky. Traditional CNNs are well-known in suffering from adversarial examples. Researchers have demonstrated that using carefully crafted adversarial examples, they can mislead different CNNs designed for various computer vision problems. These include image and video classification, segmentation and object detection [27,37,4,11,10,34]. To mislead object detectors, the current attacks either modify pixels inside target objects [4,17,26] or put the adversarial patches very close to target objects [11,10]. Though non-local blocks have been applied to a lot of computer vision problems, object detection is selected for this study because it is an essential component in many cyber-physical systems, *e.g.*, autonomous vehicles. In order to investigate whether the non-local blocks would bring new threats to object detectors, two types of attacks are studied in this paper — appearing and disappearing attacks. The former aims at misleading object detector such that it is unable to detect a target object category, *e.g.*, stop sign, and the latter aims at misleading object detector such that it detects a predefined object category which is not present in images. If an adversarial patch is similar to a target object, the object detector would detect it as a target object. However, this is not the goal of appearing attack where wrongly detected target objects should appear beyond the adversarial patch itself. Different from the previous adversarial examples against object detectors, in these attacks, the adversarial patches are required neither to put very close to the target objects [11,10] nor to overlap with them [4,17,26].

To study the threats caused by non-local block, a non-local block is added into a Faster R-CNN and algorithms are designed to craft adversarial examples for carrying out these two types of attacks. By comparing the experimental results of the adversarial examples on the Faster R-CNN with non-local block and the original Faster R-CNN as a control, new threats from non-local blocks can be identified. To further explain the threats, the reception fields of non-local blocks are studied empirically and theoretically.

The rest of the paper is organized as follows. Section 2 summarizes the related works. Section 3 presents the algorithms designed to craft adversarial examples

for appearing and disappearing attacks. Section 4 reports digital attack experiments on 4073 images from 36 in-car videos and physical attack experiments on 2258 images from 20 videos. Section 5 offers an analysis to explain the experimental findings. Section 6 gives some conclusive remarks.

2 Related Work

2.1 Non-local Neural Networks

Capturing long-range dependencies is of great importance in many computer vision tasks such as video classification, semantic segmentation, and object detection. However, traditional CNNs are ineffective on it. Inspired by non-local means in [1], Wang *et al.*[31] proposed non-local blocks to capture long-range dependencies. The response of a non-local block at a particular position is the weighted sum of the features at all positions in the feature maps. Wang *et al.*[31] described four different non-local blocks: Gaussian, embedded Gaussian, dot product and concatenation and found that they perform similarly and are able to consistently enhance the performance of CNNs in different computer vision tasks. The non-local blocks can be easily inserted into other existing architectures, which makes them widely adopted by other researchers. Zhen *et al.* [39] embedded a pyramid sampling module into non-local blocks to capture semantic statistics in different scales with only a minor computational budget while maintaining the excellent performance as the original non-local modules in semantic segmentation. Yue *et al.*[36] generalized the non-local blocks and took the correlations between the positions of any two channels into account to improve their representation power. They proposed a compact representation for different kernel functions employed in the non-local blocks and used Taylor expansion to reduce their computational demand. Zhang *et al.* [38] extended the non-local blocks and designed a residual non-local attention network for image restoration.

The non-local blocks have also been applied to various applications. Ma *et al.* [21] used the non-local operator in a framework that restores reasonable and realistic images by globally modeling the correlation among different regions. Shokri *et al.* [25] applied non-local neural networks to capture long-range dependencies and to determine the salient objects. Xia *et al.* [33] proposed a novel mechanism for person re-identification that directly captures long-range relationships via second-order feature statistics based on non-local blocks. In medical applications, Chen *et al.* [3] used a non-local spatial feature learning block to learn long-range correlations of the liver pixel position for a better liver segmentation. Besides, non-local neural networks are also applied in image de-raining [14], video captioning [12], cloth detection [15], text recognition [19], building extraction [30], and road extraction [32].

2.2 Adversarial Examples

Adversarial examples have drawn great attention since the discovery by Szgedy *et al.* [27]. They found that the state-of-the-art image classifiers would classify an

image with deliberately designed noise to an incorrect label and the image with the noise looks almost the same as the original image for naked eyes. Thereafter, different attacks against image classifiers are investigated [7,23,22] and the risks in other computer vision methods are also studied. Object detector, a critical component in many computer vision systems, has a lot of real-world applications, *e.g.*, autonomous driving car. To study its potential security risk, researchers developed attacks against object detectors digitally and physically. In 2017, Xie *et al.* [34] and Lu *et al.* [18] proposed methods to digitally attack Faster R-CNN and YOLO respectively. Their attacks are implemented by inserting noise into whole images. These attacks are not able to be carried out in the physical world. Lu *et al.* [17] designed an adversarial stop sign by adding noise to the stop sign in order to fool Faster R-CNN in both digital and physical worlds. Different from the adversarial examples against image classifiers, the adversarial stop sign looks very different from a normal stop sign. To make the adversarial stop sign more realistic, Chen *et al.* [4] proposed a method to change every pixel inside the stop sign, except for those inside the 'STOP' word region. Song *et al.* [26] further limited the attack region and produced an adversarial sticker that can mislead an object detector digitally and physically by putting it on a target stop sign. Different from the previous attacks, Huang *et al.* [10,11] attempted to mislead an object detector by placing adversarial examples outside the target object. However, their adversarial examples need to be placed very close to the target object. These works show that attacking object detectors is relatively hard, especially in the physical world when attackers have no access to target objects and their surrounding area. All these studies were performed on traditional object detectors without non-local blocks. To the best of the authors' knowledge, there are no recorded studies on threats against object detectors with non-local block.

3 Methodology

3.1 Faster R-CNN and Non-local Block

To study the threats caused by the non-local blocks, adversarial patches are designed for carrying out disappearing and appearing attacks. In this study, Faster R-CNN with a ResNet-101 as its backbone and a non-local block is used to train the adversarial patches. The Faster R-CNN is selected on the account of its popularity and that many detectors are relying on the Faster R-CNN architecture. For a clear presentation, the original Faster R-CNN without the non-local block is first described. It consists of three major components: a backbone network, a region proposal network (RPN) and a detection network. The backbone network computes features for both RPN and detection network. The RPN takes the features and produces region proposals that have high probability with objects. The detection network takes the features and the region proposals as inputs. Its box regression layer refines the bounding box coordinates provided by RPN and its classification layer outputs a probability matrix, P, each of whose row and column respectively corresponds to one region proposal and one class label.

The element in the i^{th} row and the j^{th} column of P representing the probability of the i^{th} region proposal belonging to the j^{th} class is denoted as p_{ij} and the i^{th} row and the j^{th} column of P are denoted $p_{i\cdot}$ and $p_{\cdot j}$ respectively. Faster R-CNN applies a threshold and non-maximum suppression to determine final object classes and their bounding boxes.

For this study, a non-local block is inserted between the 3^{rd} and 4^{th} residual blocks in the ResNet-101. It is inserted at the same location as Wang *et al.* [31], and inserting at a lower layers would require a lot of memory and computation power when training a non-local block. For the sake of convenience, the term Faster R-CNN-WN is used to refer to the Faster R-CNN with the non-local block. Several types of non-local blocks have been proposed. The embedded Gaussian non-local block is employed in this study because different types of non-local blocks have very similar effect [31] and embedded Gaussian non-local block is the most popular one among them. Formally, a non-local block is defined as

$$z_k = W_z y_k + x_k \tag{1}$$

where x_k is a feature vector at the k^{th} spatial location of the previous layer, W_z is a matrix optimized through training and

$$y_k = \frac{1}{C(x)} \sum\nolimits_{\forall m} f(x_k, x_m) g(x_m). \tag{2}$$

In embedded Gaussian non-local block,

$$f(x_k, x_m) = e^{\theta(x_k)^T \phi(x_m)} \tag{3}$$

where $\theta(x_k) = W_\theta x_k, \phi(x_m) = W_\phi x_m$, the normalize factor is set as $C(x) = \sum_{\forall m} f(x_k) f(x_m)$ and $g(x_m) = W_g x_m$.

3.2 Disappearing Attack

Let T be a target object category, I be a training image and $B_{gT}(I)$ be the ground truth bounding box of a target object in I with a size of $w_{gt} \times h_{gt}$ pixels. To carry out disappearing attack, an adversarial patch Λ with a size of $w \times h$ pixels is constructed to minimize all the probabilities of the target object category in P, i.e., $p_{\cdot T}$. Let $p_{jT} = F_N(I, j)$, where F_N represents the operations in the Faster R-CNN-WN computing the probability of the j^{th} region proposal belonging to the target class. To enhance the robustness of the adversarial patch for target objects in different images taken from different viewpoints and illumination conditions, Λ is trained on images from k videos, each of which contains at least one target object. The entire training set is denoted as Q. To properly model the variations of zoom factors and the distance between camera and target object in different images, Λ is resized according to the target object. More precisely, Λ is resized to $\alpha w_{gt} \times \beta h_{gt}$, where α and β are parameters controlling the size of Λ in the image. In the training, Λ is placed below $B_{gT}(I)$ with a distance. If the training image has more than one target object, one of them is randomly selected and

Λ is placed below it. Though the relative location between Λ and the target object is fixed in training, in testing, Λ can be placed in different locations to perform the attack. Section 5 will explain why the difference between training and testing locations is not important. Let $f(I, \Lambda, B_{gT}(I))$ be a training image with the rescaled adversarial patch. In the training, Λ is trained to minimize the sum of $p_{jT}, \forall j$. The objective function

$$\Lambda = \underset{\Lambda}{\operatorname{argmin}} \sum_{I \in Q} \sum_{\forall j} F_N(f(I, \Lambda, B_{gT}(I)), j) \tag{4}$$

is used to perform this minimization.

3.3 Appearing Attack

In this sub-section, we use the same notations as in the previous sub-section. As with the previous attack, an adversarial patch Λ with a size of $w \times h$ is inserted into a training image I, according to the location of a reference object and its size. The size of the rescaled Λ is set to $\alpha w_{ht} \times \beta h_{ht}$ pixels and Λ is placed below the reference object with a distance. In the experiments, stop sign is used as a reference object for placing and rescaling Λ in the appearing attack. The image with the rescaled Λ is denoted as $f(I, \Lambda, B_{gT}(I))$. To carry out this attack, a target label T is selected and the adversarial patch Λ is trained to minimize the negative $\log p_{jT}$ for the region proposal not overlapping with Λ. To avoid the appearing objects overlapping with Λ, the objective function also minimizes the negative $\log p_{jB}$ for region proposals overlapping with Λ, where p_{jB} is the background probability of the j^{th} region proposals. Mathematically, the objective function below is used to train Λ,

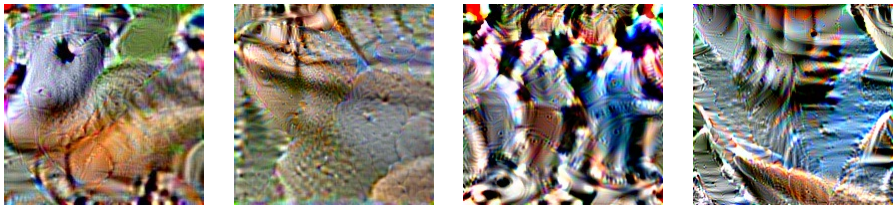
$$\begin{aligned} \Lambda = & \underset{\Lambda}{\operatorname{argmin}} \sum_{I \in Q} \sum_{j \in \phi^C} -\varepsilon \log F_N(f(I, \Lambda, B_{gT}(I)), j) \\ & - \sum_{j \in \phi} (1 - \varepsilon) \log F_{NB}(f(I, \Lambda, B_{gT}(I)), j) \end{aligned} \tag{5}$$

where ε is a parameter balancing the two terms, F_{NB} represents the operations of the Faster R-CNN-WN computing p_{jB}, ϕ is a set storing indexes of the region proposals overlapping with Λ and ϕ^C is its complement storing indexes of the region proposals not overlapping with Λ.

In this objective function, the first term is to mislead the region proposals that are not intersecting with Λ and make them target objects. The second term is designed to keep the proposals intersecting with Λ to be the background. It is noticed that the appearing attack will be weak when ε is too small, and bounding boxes will appear around or inside Λ when ε is too large (Fig. 3c).

4 Experiments

To evaluate the threats caused by non-local block, appearing and disappearing attacks are performed on the Faster R-CNN with the non-local block (Faster



(a) (b) (c) (d)

Fig. 1: Adversarial patches for the (a) disappearing attacks and (b) appearing attacks against the Faster R-CNN-WN. Adversarial patches for the (c) disappearing attacks and (d) appearing attacks against the Faster R-CNN.

R-CNN-WN). The attacks are also performed on the original Faster R-CNN without the non-local block, as a control experiment. The two Faster R-CNNs are trained on the COCO dataset (2017 training images) and their backbone networks are ResNet-101. In the disappearing attacks, stop sign is selected as a target object and also a reference object for resizing and placing the adversarial patches in the images. In the appearing attacks, stop sign is selected as a reference object, because it is a common target object in the previous adversarial example studies [17,4,26,11] and an important object for autonomous vehicles. In the appearing attacks, boat is selected as a target object because none of the training and testing videos has boat and it would make the evaluation easier. In the experiments, 721 images sampled every other frame from five in-car videos are used as a training set. The image sizes are 406×720 pixels or 1080×720 pixels. Four adversarial patches with a size of 200×200 pixels are generated. Fig. 1 shows the four adversarial patches. These adversarial patches are used to evaluate the threats caused by the non-local block in both digital and physical worlds.

4.1 Digital Attack

In digital attack, 4073 images sampled from 36 Internet in-car videos are taken as a testing set. The sizes of the images are 1080×1920 pixels and the sizes of the stop signs range from 21×22 pixels to 660×633 pixels. The original detection rates of the Faster R-CNN-WN and the Faster R-CNN are 80.6% and 78.6%, respectively. These results match with Wang *et al.*'s findings [31] that non-local block can improve detection performance. The adversarial patches are scaled and then placed below the detected stop signs with a certain distance away from it. In theory, the attack rate would be higher with a larger adversarial patch. In the disappearing attack experiments, we would like to keep the size of adversarial patch similar to that of the traffic signs in the real world. Thus, α and β are set to 1.5. When at least one detectable stop sign in an original image is missing due to the adversarial patches, the attack is considered a success. Denote the number of original images with at least one detectable stop sign as Det_{imgs}. The

(a) (b)

Fig. 2: Disappearing attack results. (a) The detection results of stop sign from the original Faster R-CNN and (b) the detection results of stop sign from the Faster R-CNN-WN. The first column is the original detection results without the adversarial patches and the second column is the detection results with the adversarial patches.

successful disappearing attack rate D_{ar} defined as:

$$D_{ar} = \frac{\text{number of sucessful attacks}}{Det_{imgs}} \tag{6}$$

is used as a performance index. Table 1 lists the original detection rate (DR_{org}) and the successful disappearing attack rates of the two detectors. It indicates that the impact of the disappearing attacks is much greater on the Faster R-CNN-WN than the original Faster R-CNN. Fig. 2 shows typical detection results from these two detectors.

In the appearing attack experiments, boat and stop sign are respectively selected as a target object and a reference object. Similar to disappearing attack, α and β are set to 2 and ε in Eq. 5 is set to 0.4. Note that without the adversarial patches, the detectors would still have the probability of wrongly detecting other objects or background as a boat. When the detector has wrong detections in absence of the adversarial patches, only the case where the intersection over union of these detected boxes and the predictions for boats is smaller than 0.6 will be considered as a result of the appearing attack. The adversarial patches may cause the detectors to wrongly detecting multiple boats. The wrongly detected boats with no intersection with the adversarial patches are considered as a clear success (Fig. 3a). In some cases, the wrongly detected boats are very large and have some overlap with the adversarial patches (Fig. 3b). Hence, the intersection of the detected boat and the adversarial patch over the detected boat (IOD) is used to define successful attack. If there is no intersection, IOD will be zero and if the detected boat is completely inside the adversarial patch, IOD will be one. When IOD is smaller than a threshold, it is considered as a successful attack. The successful appearing attack rate A_{ar} is defined as

$$A_{ar} = \frac{\text{number of images with at least one successful attack}}{\text{number of testing images}} \tag{7}$$

Table 1: Successful disappearing attack rates (%)

	Original Faster R-CNN	Faster R-CNN-WN
DR_{org}	78.6	80.6
D_{ar}	7.6	52.7

Table 2: The successful appearing attack rates (%) under different IOD thresholds.

	Original Faster R-CNN	Faster R-CNN-WN
IOD=0	0	53.5
IOD<0.1	0.6	55
IOD<0.2	1.4	55.2
IOD<0.3	3.1	55.4
IOD<0.4	6.6	55.4

(a) (b) (c)

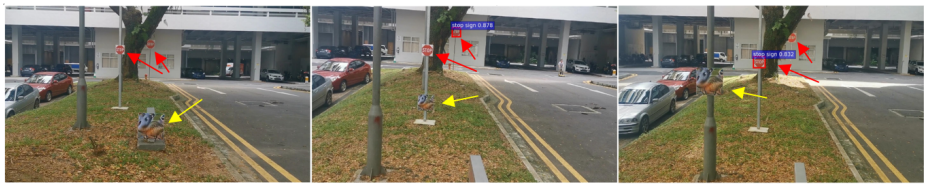
Fig. 3: Appearing attack results, where boat is a target object. (a)-(b) The detection results of boat from the Faster R-CNN-WN. (c) The detection results of boat from the original Faster R-CNN.

Table 2 gives the successful appearing attack rates from the two detectors and Fig. 3 shows some typical detection results from Faster R-CNN-WN and original Faster R-CNN. The Faster R-CNN can only detect the adversarial patch or it sub-region as the target object (Fig. 3c). However, the Faster R-CNN-WN is misled by the adversarial patch and detects large areas as boat. Since the original Faster R-CNN is insensitive to the disappearing and appearing attacks in the digital world, it is not included in the following physical experiments.

4.2 Physical Attack

To examine the threats caused by non-local block in the physical world, a stop sign with a size of 19.3 cm by 19.3 cm and the adversarial patches generated for disappearing and appearing attacks with a size of 28.5 cm by 28.5 cm are printed out. In the disappearing attack, 6 groups of videos with resolution of 1080×1920 or 1920×1080 pixels are taken from outdoor environments by a smartphone camera. In each group, there are 2 videos taken from the same

(a)



(b)

Fig. 4: The detection results of stop sign from the Faster R-CNN-WN (a) without and (b) with the adversarial patch in disappearing attack. Note that the stop sign closed to the tree is a real stop sign.

location and roughly the same viewpoint. One video has the stop sign only; one video has the stop sign and the adversarial patch for disappearing attack. The adversarial patch is placed below or in front of the stop sign without a fixed distance (Fig. 4b). On average, 110 frames are sampled from each video for testing. As in the digital experiment, every other frame is sampled. In the digital disappearing attacks, the successful disappearing attack rate is used as a performance index. However, it is not applicable to physical attack because the videos with and without adversarial patches are neither collected from the exact same viewpoint nor at the same time, and have different numbers of frames. Thus, the detection rates with and without the influence from the adversarial patch are used to evaluate the impact and provided in Table 3. In some videos, there is also a real stop sign, so the detection rate here is defined as the number of detected stop signs divided by the number of stop signs in the images. Table 3 shows that for the Faster R-CNN-WN, the adversarial patch reduces the average detection rates from 87.2% to 48.0%. Fig. 4 shows some example results of the physical disappearing attack.

In the digital appearing attacks, stop sign is just used as a reference object for resizing and placing the adversarial patch. Since it is impossible to resize the adversarial patch in physical attack and in fact, it is not necessary, stop sign is not used as a reference object. For examining appearing attacks in the physical world, 4 groups of videos, with resolution of 1080×1920 pixels are taken in 4 locations. In each group, there are 2 videos with and without the adversarial patch for appearing attack taken from the same location and roughly the same viewpoint. The videos are sampled in every other frame. On average, 118 frames per video are used for testing. The appearing rate here is defined as the number

Fig. 5: The Faster R-CNN-WN detection results of boat in appearing attack.

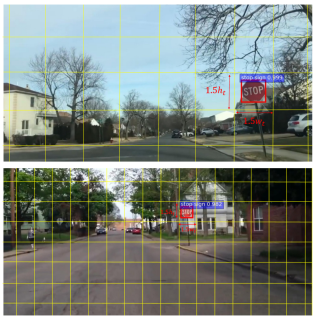


Fig. 6: Two sample images with the grids.

of frames detected with the target object divided by the total number of the frames. Table 4 shows that for the Faster R-CNN-WN, the adversarial patch increases the average appearing rates of boat from 0.8% to 31.3%. Fig. 4 shows some example results from the physical appearing attack.

4.3 Effective Attack Regions

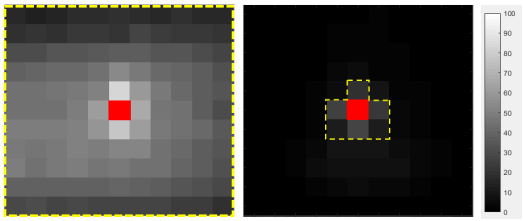
In the previous disappearing attack experiments, the relative locations between the target object i.e., stop sign and the adversarial patches are fixed. More precisely, in both training and testing of the digital attacks, the adversarial patches are placed 1.5 h_t pixels below the stop sign, where h_t is the detected height of the stop sign in the image. In this experiment, the adversarial patches trained in the fixed relative position are placed in different locations. 831 images are sampled from every 5 frames from the 36 testing videos in the digital experiments. From them, 670 and 645 images are detected with stop sign by the Faster R-CNN-WN and the Faster R-CNN, respectively. Only the images with detected stop signs are employed in the following evaluation. Each image is divided by a grid and the size of each block in the grid is $1.5h_t \times 1.5w_t$. Note that the number of blocks in different images is different. Fig. 6 shows images with the grids. In total 11×11 positions are tested. The adversarial patches (Figs. 1a and 1c) are rescaled based on the stop signs and put in the centers of the blocks. If the detector cannot detect the stop sign, it is considered as a successful attack. Note that the size and the location of the stop sign in each image are different. To compute the successful attack rates at different relative locations, the grids are rescaled and aligned. Fig. 7 shows the successful attack rates from both Faster R-CNN-WN and Faster R-CNN. The red boxes indicate the locations of the aligned stop signs. Note that the successful attack rates at different locations are computed from different numbers of images, because some stop signs are close to image boundary. Fig. 7a shows that the effective attack region for the Faster R-CNN-WN covers the entire grid and most of the successful attack rates are higher than 30%. The region close to the target object has much higher successful at-

Table 3: The detection rate of stop sign in disappearing attacks in the physical world.

Detection rate	Faster R-CNN-WN
Without attack	89.9
With attack	47.6

Table 4: The appearing rate of boat in appearing attacks in the physical world.

Appearing rate	Faster R-CNN-WN
Without attack	0.8
With attack	31.3



(a) (b)

Fig. 7: Successful attack rates in different locations, (a) Faster R-CNN-WN and (b) Faster R-CNN. The red boxes indicate the locations of the stop signs and the yellow lines indicate successful attack rates greater than 10%.

tack rates and the upper, lower and right borders have relatively lower successful attack rates. The rest of the region has similar successful attack rates. Without the non-local block, the successful attack rates become very low (Fig. 7b), except for the region very close to the center. These results indicate that disappearing attacks can be applied in long-distance and the difference between training and testing location is not a matter.

5 Analysis of the Effective Attack Regions

The experiments in Section 4.3 expose several properties of the effective attack regions of the two detectors. To discuss the properties observed in Fig. 7 systematically, in this section, an effective attack region is defined as where the successful attack rate is higher than 10%, which is highlighted by the yellow color boundaries. To understand these properties, analysis and discussion are provided in this section. The effective attack region of the Faster R-CNN is first discussed and then an analysis for the effective attack region of the Faster R-CNN-WN is provided. Since they are in fact the same, except for the non-local block in the backbone networks, the analysis below focuses on the reception fields of the backbone networks.

The effective attack region of the Faster R-CNN is very small and very concentrated on the center. Luo *et al.* [20] points out that the output neurons of deep CNNs, like ResNet-101 has a small effective receptive field with a Gaussian shape, which is much smaller than its theoretical receptive field. Thus, the adversarial patch can only affect its surrounding region and the successful attack

rate decays exponentially when the distance between the target object and the adversarial patch increases (see Fig. 7b). Also because of the Gaussian shape effective receptive field, Fig. 7b has a roughly circular shape.

The effective attack region of the Faster R-CNN-WN (Fig. 7a) is very different. It covers the entire Fig. 7a with high successful attack rates. The region around the center has a very high successful attack rate and the successful attack rate in the rest of the region is roughly the same, except for the boundaries. It implies that the non-local block extends the effective receptive field to entire images. We would like to mathematically quantify how the receptive field has been changed after adding the non-local block on top of the convolution neural network. Here we only consider single-channel case for every layer and dot-product form of non-local block for simplicity. The result can be derived similarly for multi-channel cases by considering inter-channel correlations, and extended to other versions of non-local block by replacing the pairwise function f.

Following Luo *et al.*'s effective receptive field analysis [20], we compute the gradient of a single output neuron with respect to all input pixels. What we are interested here is the gradient of a single output pixel with respect to all input pixels, *i.e.*, $\partial y_i/\partial a_m$ for all $m \in \{1, \ldots, N^2\}$, where y is the output of non-local block, and a is an $N \times N$ input image. The output of the CNN block, also the input of non-local block, is denoted by x. Without the loss of generality, we assume that x_i's are uniformly distributed. The gradient can be decomposed to $\partial y_i/\partial a_m = \sum_k \partial y_i/\partial x_k \cdot \partial x_k/\partial a_m$ using chain rule, which is further divided into the following two cases,

$$\frac{\partial y_i}{\partial a_m} = \begin{cases} \sum\limits_{k \neq i} \frac{\partial y_i}{\partial x_k} \cdot \frac{\partial x_k}{\partial a_m} + \frac{\partial y_i}{\partial x_i} \cdot \frac{\partial x_i}{\partial a_i}, & m = i \\ \sum\limits_{k \neq m} \frac{\partial y_i}{\partial x_k} \cdot \frac{\partial x_k}{\partial a_m} + \frac{\partial y_i}{\partial x_m} \frac{\partial x_m}{\partial a_m}, & m \neq i \end{cases} \tag{8}$$

Recall that the dot-product version non-local operator follows the function $z_i = W_z y_i + x_i$, where

$$y_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) = \frac{1}{N} \sum_{\forall j} x_i x_j W_g x_j \tag{9}$$

Thus, the gradient signal on x, $\partial y_i/\partial x_m$, can also be computed by considering the cases when $m = i$ and $m \neq i$,

$$\frac{\partial y_i}{\partial x_m} = \frac{1}{N} W_g \times \begin{cases} \sum_{j \neq m} x_j^2 + 3x_m^2, & m = i \\ 2x_i x_m, & m \neq i \end{cases} \tag{10}$$

It is presented by the previous work [20] that $\partial x_k/\partial a_m$ for all m forms a Gaussian shape, which diminishes fast at its tails. Thus, the summation terms in both cases of Eq. 8 are negligible since $\partial x_k/\partial a_m$ is close to zero when $|k - m|$ becomes larger. Moreover, we noticed in Eq. 10 that the gradient when $m = i$ is strictly greater than that when $m \neq i$ since $\sum_{j \neq m} x_j^2 + 3x_m^2 - 2x_i x_m$ is strictly

positive, when not all x_i's are zeros, and $\partial x_i/\partial a_i$ for all i is the value of a Gaussian density function taken at its mean, which is assumed to be equal for all i. We then conclude that the gradient $\partial y_i/\partial a_m$ when $m = i$ is greater than that when $m \neq i$. Besides, since not all x_i's are zeros, the gradients $\partial y_i/\partial a_m$ is also non-zeros for all $m \neq i$.

This gives us an intuition of analyzing the receptive field by considering the non-local block. Different from the effective receptive field of CNNs without non-local block, which has non-zero gradient values in the neighborhood of the center pixel and zero gradients further away, the receptive field of CNNs with non-local block covers the whole input space. More specifically, the gradient of a single output pixel with respect to the corresponding pixel, $\partial y_i/\partial a_m$, is non-zero everywhere. It also proved that the gradient with respect to center pixel ($\partial y_i/\partial a_i$) is the maximum. With the consideration of the skip connection in the non-local block, which is the sum of x_j over all j in computing z_i, the neighborhood region of the center would have higher gradients because of its Gaussian shape receptive field [20]. The gradients of other regions are strictly smaller and have similar values.

The analysis above does not explain why the successful attack rate is lower at the boundary. When the adversarial patch is put in the image border, it would affect lesser number of neurons in the input layer of the non-local block, x. Therefore, the attack would be weaker as illustrated in Fig. 7. The experimental results also show that the adversarial patch trained at a fixed relative location with respect to the target object is effective in other locations. Because CNNs without a fully connected layer is roughly translation invariant, putting the adversarial patch in two different locations, p and q, their corresponding neuron outputs x_p and x_q should be roughly the same. Since the non-local block (Eq. 1) considers all $x_i, \forall i$, the adversarial patch can attack on other locations, different from the training location.

6 Conclusion

To overcome the weaknesses of traditional deep neural networks, which are ineffective to capture long-range dependency, researchers developed non-local blocks and demonstrated their effectiveness on various computer vision tasks. However, without understanding the threats caused by the non-local blocks and applying them to critical systems is risky. In this paper, two types of attacks, disappearing and appearing attacks against object detectors are studied. Different from the previous attacks against object detectors, these attacks are performed in long distance. The digital and physical experimental results show that the universal adversarial patches obtained by the proposed algorithms can mislead the Faster R-CNN with a non-local block to classify stop sign as background and to wrongly detect boats that are not in the images. To understand the effective attack region and its properties, the reception field of the non-local block is analysed.

Acknowledgements This work is partially supported by the Ministry of Education, Singapore through Academic Research Fund Tier 1, RG30/17

References

1. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 60–65. IEEE (2005)
2. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
3. Chen, L., Song, H., Li, Q., Cui, Y., Yang, J., Hu, X.T.: Liver segmentation in ct images using a non-local fully convolutional neural network. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 639–642. IEEE (2019)
4. Chen, S.T., Cornelius, C., Martin, J., Chau, D.H.: Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In: ECML/PKDD (2018)
5. Chi, L., Tian, G., Mu, Y., Xie, L., Tian, Q.: Fast non-local neural networks with spectral residual learning. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2142–2151 (2019)
6. Fu, C., Pei, W., Cao, Q., Zhang, C., Zhao, Y., Shen, X., Tai, Y.W.: Non-local recurrent neural memory for supervised sequence modeling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6311–6320 (2019)
7. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015), `http://arxiv.org/abs/1412.6572`
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hu, G., Cui, B., Yu, S.: Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 1216–1221. IEEE (2019)
10. Huang, Y., Kong, A.W.K., Lam, K.Y.: Adversarial signboard against object detector. In: Proceedings of the British Machine Vision Conference (BMVC) (2019)
11. Huang, Y., Kong, A.W.K., Lam, K.Y.: Attacking object detectors without changing the target object. In: Pacific Rim International Conference on Artificial Intelligence. pp. 3–15. Springer (2019)
12. Lee, J., Kim, J.: Improving video captioning with non-local neural networks. 2018 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia) pp. 206–212 (2018)
13. Levi, H., Ullman, S.: Efficient coarse-to-fine non-local module for the detection of small objects. arXiv preprint arXiv:1811.12152 (2018)
14. Li, G., He, X., Zhang, W., Chang, H., Dong, L., Lin, L.: Non-locally enhanced encoder-decoder network for single image de-raining. arXiv preprint arXiv:1808.01491 (2018)
15. Li, Y., Tang, S., Ye, Y., Ma, J.: Spatial-aware non-local attention for fashion landmark detection. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 820–825. IEEE (2019)
16. Liao, X., He, L., Yang, Z., Zhang, C.: Video-based person re-identification via 3d convolutional networks and non-local attention. In: Asian Conference on Computer Vision. pp. 620–634. Springer (2018)

17. Lu, J., Sibai, H., Fabry, E.: Adversarial examples that fool detectors. CoRR **abs/1712.02494** (2017)
18. Lu, J., Sibai, H., Fabry, E., Forsyth, D.A.: No need to worry about adversarial examples in object detection in autonomous vehicles. CoRR **abs/1707.03501** (2017)
19. Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R.: Master: Multi-aspect non-local network for scene text recognition. arXiv preprint arXiv:1910.02562 (2019)
20. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in neural information processing systems. pp. 4898–4906 (2016)
21. Ma, Y., Liu, X., Bai, S., Wang, L., He, D., Liu, A.: Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 3123–3129. AAAI Press (2019)
22. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2574–2582 (2016)
23. Papernot, N., McDaniel, P.D., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P) pp. 372–387 (2016)
24. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Non-local graph convolutional networks for skeleton-based action recognition. arXiv preprint arXiv:1805.07694 (2018)
25. Shokri, M., Harati, A., Taba, K.: Salient object detection in video using deep non-local neural networks. arXiv preprint arXiv:1810.07097 (2018)
26. Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., Kohno, T.: Physical adversarial examples for object detectors. In: 12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18) (2018)
27. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014), `http://arxiv.org/abs/1312.6199`
28. Tang, Y., Zhang, X., Ma, L., Wang, J., Chen, S., Jiang, Y.G.: Non-local netvlad encoding for video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
29. Tu, Z., Ma, Y., Li, C., Tang, J., Luo, B.: Edge-guided non-local fully convolutional network for salient object detection. arXiv preprint arXiv:1908.02460 (2019)
30. Wang, S., Hou, X., Zhao, X.: Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. IEEE Access **8**, 7313–7322 (2020)
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
32. Wang, Y., Seo, J., Jeon, T.: Nl-linknet: Toward lighter but more accurate road extraction with non-local operations. arXiv preprint arXiv:1908.08223 (2019)
33. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3760–3769 (2019)
34. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.L.: Adversarial examples for semantic segmentation and object detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 1378–1387 (2017)

35. Xu, X., Wang, J.: Extended non-local feature for visual saliency detection in low contrast images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
36. Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F.: Compact generalized non-local network. In: Advances in Neural Information Processing Systems. pp. 6510–6519 (2018)
37. Zajac, M., Zołna, K., Rostamzadeh, N., Pinheiro, P.O.: Adversarial framing for image and video classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 10077–10078 (2019)
38. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. arXiv preprint arXiv:1903.10082 (2019)
39. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 593–602 (2019)