Benchmarking and deeper analysis of adversarial patch attack on object detectors

Pol Labarbarie[1,2], Adrien Chan-Hon-Tong[2], Stéphane Herbin[2] and Milad Leyli-Abadi[1]

[1]*IRT SystemX,* Palaiseau, France
[2]*ONERA/DTIS, Université Paris-Saclay,* F-91123 Palaiseau, France

Abstract
Adversarial attacks (either norm bounded or patch-based) have received much attention from the computer vision community over the last decade. The criticality of those attacks in the physical world, however, is questionable. Indeed, none of the proposed attacks in the literature has been demonstrated in a realistic physical implementation verifying simultaneously significant contextual effects, radiometric and geometrical robustness in either black or gray box settings. To advance this issue, in this paper we propose an evaluation framework for patch attacks against object detectors. This framework focuses on robustness and transferability properties by considering various image transformations and learning conditions. We validate our framework on three state-of-the-art patch attacks using PASCAL VOC dataset, providing a more comprehensive view of their criticality.

Keywords
Adversarial physical attack, Robustness of visual object detection, AI component evaluation methodology

1. Introduction

Deep neural networks (DNNs) achieve state-of-the-art results in various computer vision tasks including image classification [1], semantic segmentation [2], and object detection [3, 4]. Due to their complexity, it has been shown that they are vulnerable to small, adversarially-chosen perturbations of their inputs [5, 6]. The existence of this vulnerability has motivated works trying to make DNNs empirically more robust [7] or proving that they satisfy robustness properties [8], and works dedicated to the design of more powerful attacks [9, 10]. Those invisible attacks are mainly theoretical objects and are not suited for real-world applications since they consist of perturbing all the pixels in a very specific way. In fact, considering self-driving cars as an example, it is difficult to see how, physically, the image pixels captured by the embedded sensors could be perturbed.

A more realistic attack, named adversarial *patch*, has been introduced in [11]. This type of attack is easily visible in the image because it relies on adding a heavily textured patch to the scene. Since such a patch can be easily printed and positioned on an object or in the environment, it can pose a serious threat. Placing a patch on a stop sign or on the roadway may result in its misclassification [12] or in the missed detection of a pedestrian crossing the road [13]: from a trustworthy AI point of view, the instability due to a patch-based adversarial attack is not acceptable. However, it is unclear whether these attacks are truly robust to a wide variety of observational conditions, such as radiometric or geometric changes, and whether these patches can be generated in a black-box setting, i.e., without having access to the internal variables of the attacked algorithm. Thus, it can be interesting for the community to rely on a detailed evaluation framework that provides metrics under multiples geometric, radiometric or model settings in order to have a better understanding of the criticality of patch attack threats.

In this paper, we propose a preliminary evaluation framework which helps to evaluate the robustness of patch attacks to both translation and model change, applied to three different attacks [14, 15, 13]. We conduct experiments on YOLOv2 detector [16] and PASCAL-VOC dataset [17]. The main contributions of this work can be summarized as follows:

- definition of various categories of evaluation criteria;
- proposition of an evaluation framework ranking adversarial patch attacks;
- analysis of the spatial effect of state-of-the-art patch based adversarial attacks;
- analysis of the internal mechanism of such attacks.

The paper is organized as follows. In section 2, we give a brief overview of adversarial patch attacks (APAs). In section 3, we describe our methodology based on defining

The IJCAI-ECAI-22 Workshop on Artificial Intelligence Safety (AISafety 2022), July 24-25, 2022, Vienna, Austria
pol.labarbarie@irt-systemx.fr (P. Labarbarie);
adrien.chan_hon_tong@onera.fr (A. Chan-Hon-Tong);
stephane.herbin@onera.fr (S. Herbin);
milad.leyli-abadi@irt-systemx.fr (M. Leyli-Abadi)
https://stepherbin.github.io/ (S. Herbin)

several criteria that evaluate the physical impact of the attack. In section 4, using our methodology, we evaluate three state-of-the-art APAs. Then, in section 5 we develop perspectives about creating more powerful attacks.

2. Related works

In this section, we have classified the state-of-the-art related works for adversarial patch attacks in three subsections. First, we describe the beginning of APAs where patches were applied to fool image classifier. Then, we present works developing APAs aiming to fool object detectors. Finally, we describe works exploring patches contextual effects.

2.1. APAs for classification

Adversarial Patch Attacks (APAs) were introduced by [11] for image classification. Instead of finding a small additive perturbation, they confined the optimization to a small part of the image but allowed it to be unconstrained in magnitude. They produced a patch capable of fooling multiple ImageNet classification models either in digital or physical domain (just by printing the patch).

2.2. APAs for object detection

Attacking object detectors was explored in several works working on different applications. In the beginning, patches were directly applied on the struck object. The first two works on patch-based attacks had targeted stop signs. [18] used change-of-variable attack described in [19] and the Expectation over Transformation technique [20] to change the red background of stop signs to fool Faster RCNN. Independently, [12] developed stickers when applied on stop signs, can fool YOLOv2 and can transfer to fool Faster RCNN. [21] were the first to create a patch causing the disappearance of people when it was applied on them. These works focused on designing a patch that overlaps the targeting object to either change its class or suppress detection.

Yet, depending on the context, suppressing detection only on object close to the patch can be to restricted. Currently, on video surveillance setting, it is an issue if the hacker can become invisible thank to a patch. However, in autonomous driving, the hacker has no interest in becoming invisible to the car. Yet, it is an issue if a patch put on a wall suppresses pedestrian detection on the street. In other words, in some contexts, the main issue is the contextual effect of the patch.

2.3. Contextual adversarial patches

Contextual patch attacks were first explored by [14]. Instead of designing a new loss, they used the YOLOv2 loss but redefined the ground truths at the patch localization. Their patches that do not overlap with the objects of interest can blind the detector. They showed transferability over patch position, network architecture, and dataset. However, patches are never clipped to the image range, which is not suitable for real-world applications. Following that, [15] studied the Dpatch attack in feasible physical conditions and compared it to their new attack. Considering a maximization problem of the YOLOv2 loss over the ground truths, they outperformed the Dpatch method and showed real-time attack success. The success of these attacks consists of adding a salient patch in the image producing false positives. This kind of effect can be related to patch effects in classification. We directly introduce ambiguity when we place a high-confidence object which may look like in real at an out-of-distribution object.

Another work similar to the previous ones is [13], which develops attacks and defense for contextual adversarial patches. They proposed a universal blindness attack targeting one chosen class, an objectness attack, and a targeted attack. In particular, [13] introduces the idea of removing false positives on the patch. We will consider this idea in our experiment as depending on the use case, one may want to measure mainly the contextual effect.

3. Methodology

This section presents the proposed methodology for ranking the patch-based adversarial attacks and is organized as follows: at the first place, we point out the motivations behind the proposition of such a pipeline, next, we present the features which are at the core and on the basis of which the ranking is obtained and finally the adopted pipeline is described in greater details.

3.1. Motivation

Recently, there exists a vast and growing literature on patch-based adversarial attacks. It is of utmost importance for concerned researchers and industries to be able to unify and generalize the evaluation procedure. According to the application domain, we can divide the adversarial patch attacks into synthetic attacks and realistic physical attacks. As an example of the first group of attacks, we can cite the case where the patch is applied at the same position of the attacked object in a digital image. As a result, the attacker could pass through the detector firewall with malicious content. However, the situation is more complex in the case of second group attacks. As an example, in the case of self-driving vehicles, the patch should be placed at the receptive field of the sensors and may be adapted with respect to various angles. It requires

Table 1
Evaluation settings by category and their brief description.

Category	Setting	Description
Radiometric	Varying weather conditions	Brightness, snow, rain, ...
	Filters	JPEG transformation
Geometric	Rescaling	* * *
	Crop	* * *
	Affine transformations	Rotations
	Distance wrt. learning position	Shift from learning position
Transferability	Detector sensitivity	Sensitivity of a detector parameters to APA
	Detector generalisation	Generalisation of an APA through multiple detectors

that the patches shall be robust against geometric transformations so that the attack takes place. In this regard, our motivation is to design a set of settings to evaluate and measure the effectiveness of the second group patch attacks under various circumstances. In the following, we use the context of self-driving vehicles to elaborate our methodology and describe the set of settings at the core of the proposed pipeline.

3.2. Evaluation settings

In our proposed pipeline, we consider three groups of evaluation settings that help to better evaluate the impact of adversarial patch attacks. Each of them represents an essential feature of the attack. These three categories are described in greater details in the following using the context of autonomous vehicles:

Radiometric settings Radiometric settings catch patch robustness against environmental changes like luminosity, weather, and photometry change like filters. They measure patch robustness when all image transformations are applied. Regarding our example, an attacker would design a patch resilient to the day's weather or luminosity on the patch.

Geometric settings Geometric settings are designed to capture the robustness of a patch subject to geometric transformations. Contrary to radiometric features, geometric transformations are transformations of the patch and not of all images. We can distinguish two types of geometric transformation. The transformations of the physics of the patch itself, such as the effect of a zoom or an ablation of one of the parts of the patch, and the transformations of the patch in its physical environments, such as affine transformations, rotations, and displacements with respect to its training position. In our example, an attacker would create a patch efficient regardless of his position in the image.

Transferability settings Transferability settings measure patch robustness according to component training changes like network parameters, learning datasets, or architectural changes. Will an attack succeed on one YOLO capable of attacking any YOLO? Or will an attack fooling YOLO's detector be able to sway a Faster R-CNN [3]? Without direct access to the attacked component, an attacker must design a patch robust to reparametrization of the network. In other words, those transferability settings measure how much the hacker known the targeted model.

Table 1 summarizes and describes all the settings present in a given category. Note that for radiometric settings and geometric settings, certain transformations can be applied only on the patch or on the image as a whole.

3.3. Evaluation pipeline

In this section, we present the proposed pipeline for evaluation of patch-based adversarial attacks, and the corresponding scheme is shown in Figure 1.

The proposed pipeline allows to compute evaluation criteria based on the settings mentioned in the previous section (see Table 1 for a summary). The first step consists in selecting an attack strategy, an object detector algorithm, and a dataset on which we may train the patch using the selected attack strategy. The choices of the dataset could be among those which provide the bounding boxes required for object detection tasks (e.g. PASCAL VOC [17], MS COCO [22], etc.). Once a patch is designed and learned, we evaluate its performance when placed at the same position during the training phase. The reported evaluation criteria are mAP (mean average precision) or AP (average precision), which are computed before (clean) and after the application of the patch on

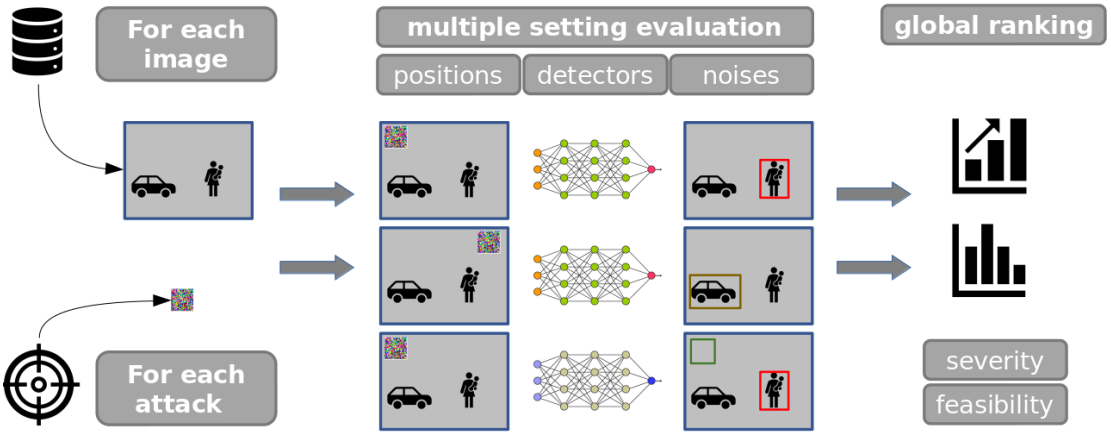

Figure 1: Structure of the proposed pipeline to evaluate APAs. Given a dataset, a network and a patch we evaluate multiple settings or configurations. The resulting average precision (AP) or mean average precision (mAP) scores are used to rank attacks for each setting. The overall rating measures the real impact in physical conditions of each APA.

the data (perturbed).

When the patch is placed at the same position that has been considered during the training phase, it shows the highest effectiveness. We can also measure the criticality of the attack by measuring the difference in its effectiveness performance between the case when the patch is placed at the same position as during the training phase and the case when one of the previously mentioned settings is applied to it. For example, after rescaling the patch or after changing the network parameters.

4. Experiments

This section presents the experimentation using the proposed methodology. We start with a brief introduction of patch attacks used for the experimentation. It is followed by the description of the experimental settings used to configure the pipeline for the evaluation. Next, a brief explanation of the object detector based on which the evaluation metrics are computed is provided. Finally, the evaluation results are demonstrated and reported using graphical tools and a comparison table.

4.1. Evaluated patches

The proposed pipeline is used to evaluate the effectiveness of three state-of-the-art contextual adversarial patch attacks, which are:

- Dpatch [14]: instead of maximizing the YOLO loss, minimizing it but redefined the ground truths boxes at the patch localization i.e. setting the patch as the only object in images;
- Lee et al. [15]: maximizing the YOLO loss over the ground truths;
- Saha et al. [13]: minimizing the probability of one chosen class.

4.2. Evaluated detector

For the sake of the evaluation, we have used the You-Only-Look-Once (YOLO) algorithm. YOLO is a one-stage object detector that achieves state-of-the-art performance and is faster than other detectors. YOLO takes a fixed-sized image and divides it into a $S \times S$ grid. For each cell, YOLO predicts B bounding boxes and their confidence scores, and for each bounding box predicts C class probabilities conditioned on being an object. In total, there are BS^2 possible bounding boxes. During inference, before the non-maximum suppression, are kept only boxes that the product of the confidence score and the conditional class probability are over a threshold.

4.3. Experimental setup

Since we evaluate patch contextual effects, we ensure that no object of interest intersects with the patch. Following [13], first, we fix the patch at a location in image, e.g., at the top-left corner i.e. pixel $(5, 5)$. Next, using the PASCAL VOC [17] test dataset, we sample two subsets of images that do not overlap with the patch. As so, a universal patch attack can be designed by simply iterating through training images. As we are interested in evaluating the contextual effects of the patch, the detections overlapping the patch do not interest us. For the sake of clarity we plot our results with and without false

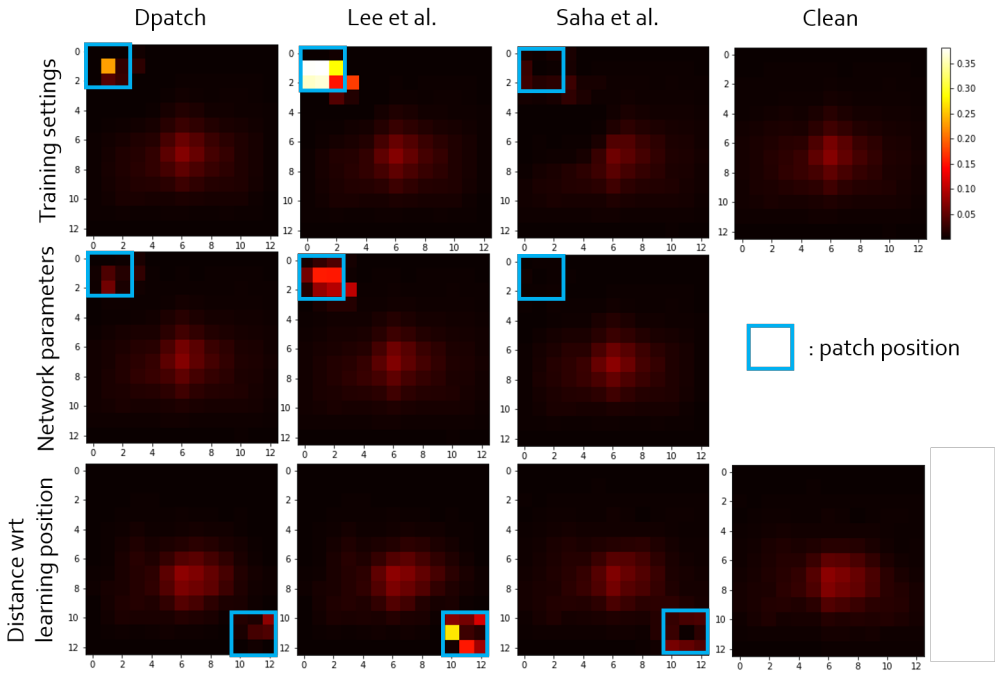

Figure 2: Objectness map obtained by averaging over anchors in cells. Rows represent the evolution of the objectness map for changing evaluation settings. At each column, a different APA is tested. The fourth column is the baseline. Training settings correspond when evaluation is performed with training settings.

positives on the patch. For all our experiments, we use YOLOv2 [16].

For each of the attacks mentioned above, we solve their corresponding optimization problem and clip the patch to $[0, 1]$. Clipping the patch ensures that we produce a more realistic patch and do not produce inf values. Each image is rescaled at size 416×416, and we fixed each patch of size 100×100 at the top-left corner. We launch the optimization process with an all-zeros patch, and we use the associate optimizer used in the corresponding article. As in [15], we run the optimization for 100 steps where 1 step corresponds to 1000 iterations. In evaluation mode, we set the confidence threshold at 0.0005, the non-maximum suppression at 0.45, and the IOU at 0.5.

Rather than evaluating each feature in each category, we choose to evaluate the invariance of the attack by network reparameterization and to measure the impact of the attack when the patch is moved from the top-left learning position to the opposite bottom-right position. For the last one, as we need the patch to not intersect with the object of interest, we extract matching images of the corresponding top-left validation set and the bottom-right validation set. As a baseline, we evaluate the different attacks at the training position. To compare [13] with other attacks, we report the AP score when attacking the "person" class. Notice that, both [14] and [15] can affect multiple classes.

4.4. Results

In this section, using our proposed pipeline, we evaluate three state-of-the-art attacks; Dpatch [14], Lee et al. [15], Saha et al. [13]. In clean mode (i.e. no patch placed in the image) we report an AP of 76.13% for the top-left extracted subset and 80.01% for bottom-right one. Cleaned scores are different since the patch is placed at another position. In fact, when we move the patch from one position to another we need to create new subsets extracted from PASCAL VOC test set since no ground truths should intersect with the patch.

Table 2 shows the results of the three attacks in multiple settings. In training settings, we see that the attack proposed in [13] produces large contextual effects. AP with and without false positives are similar. However, it seems that both [14] and [15] produce patches trying to be the salient object of images limiting their contextual effects and producing false positives on them. When we evaluate with another YOLO, contextual effects have almost completely disappeared. And when we evaluate from another position, patches can produce false nega-

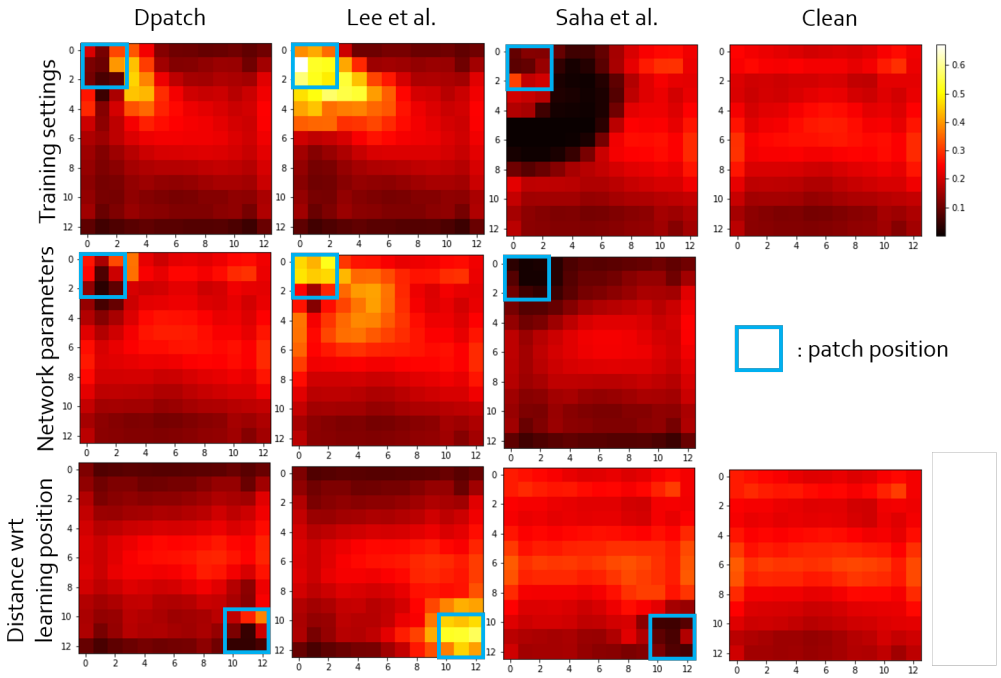

Figure 3: "Person" class probability map obtained by averaging over anchors in cells. Rows represent the evolution of the person probability map for changing evaluation settings. At each column, a different APA is tested. The fourth column is the baseline. Training settings correspond when evaluation is performed with training settings.

tives but less than before (*e.g* for Saha et al. 59.47 % AP to 75.87 % AP). AP under radiometric change is not reported due to discordant observations.

Table 2
Table of the evolution of the AP score, with and without false positives on the patch, for different setting evaluation and for different APA.

Setting	Attack	Attacked AP (%) with f.p	Attacked AP (%) without f.p	Cleaned AP (%)
Training settings	Dpatch	71.42	75.01	76.13
	Lee et al.	10.56	74.36	
	Saha et al.	59.36	59.47	
Network parameters	Dpatch	73.34	75.25	
	Lee et al.	60.35	75.42	
	Saha et al.	75.55	75.55	
Shift from learning position	Dpatch	70.61	77.87	80.01
	Lee et al.	53.02	78.73	
	Saha et al.	74.28	75.87	

4.4.1. Objectness map

Figure 2 plots the average objectness in cells for the test set for changing evaluation settings and for different APA. The color in cells of images represents the average value of the objectness predicted by YOLO for a chosen evaluation setting and a chosen APA. At each column, we plot the average objectness map for the same attacking procedure but for different evaluation settings. And, at each row, we plot the average objectness map for the same setting but for different APAs. For example, in the first row and first column, we plot the average objectness map when attacking with *Dpatch* [14] and when the patch is placed in training condition. The blue square represents where the patch is placed. We clearly notice that *Dpatch* [14] and *Lee et al.* [15] attacks try to attract the most of region proposals. On the contrary, *Saha et al.* [13] tries to decrease the objectness score around the patch.

4.4.2. Probability map

Figure 3 plots the average "person" class probability in cells for the test set for changing evaluation settings and for different APA. The color in cells of images represents the average value of the "person" class probability predicted by YOLO for a chosen evaluation setting and a chosen APA. At each column, we plot the average "person" class probability map for the same attacking procedure but for a different evaluation setting. And, at each row, we plot the average "person" class probability map for

the same setting but for a different APA. For example, in the first row and the first column, we plot the average "person" class probability map when attacking with *Dpatch* [14] and when the patch is placed in training condition. The blue square represents where the patch is placed. Again, it illustrates the fact that *Dpatch* [14] and *Lee et al.* [15] have low contextual effects. In training settings, *Saha et al.* [13] shows interesting contextual effects. Its attack can push the probability of the "person" class toward zero, producing no detections of persons in almost a quarter of the image. However, changing network parameters or moving from the learning position suppresses almost the entire effect of the patch.

5. Conclusion

In this paper, we define various categories of criteria: namely geometric, radiometric, and transferability, for the evaluation of Adversarial Patch Attacks and, using these criteria, we propose an evaluation framework able to rank them. The framework has been applied on three state-of-the-art patch based adversarial attacks.

Typically, we noticed that the patches trained to be top left have a little perturbation impact when placed bottom right. The same resilience to patch attack is true when changing learning conditions (other initial starting weights, other architecture). What this first study reveals is that the actual threat caused by the presence of state-of-the-art adversarial patch attacks is low when deployed in a realistic context. This first analysis does not claim, however, that all possible patch attacks have low impact on detection performance: the idea was rather to propose an evaluation framework able to assess their potential threat in a more physically realistic way, a framework which, we hope, future patch attacks will use.

Future works will follow two directions: complement the categories of evaluation criteria, typically with other types of transferability features, and design patch attacks resilient to a larger set of viewing and learning conditions.

Acknowledgements

This work has been supported by the French government under the France 2030 program, as part of the SystemX Technological Research Institute.

References

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012).

[2] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).

[6] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: Joint European conference on machine learning and knowledge discovery in databases, Springer, 2013, pp. 387–402.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

[8] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via randomized smoothing, in: International Conference on Machine Learning, PMLR, 2019, pp. 1310–1320.

[9] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial machine learning at scale, arXiv preprint arXiv:1611.01236 (2016).

[10] F. Tramer, N. Carlini, W. Brendel, A. Madry, On adaptive attacks to adversarial example defenses, Advances in Neural Information Processing Systems 33 (2020) 1633–1645.

[11] T. B. Brown, D. Mané, A. Roy, M. Abadi, J. Gilmer, Adversarial patch, arXiv preprint arXiv:1712.09665 (2017).

[12] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, T. Kohno, Physical adversarial examples for object detectors, in: 12th USENIX workshop on offensive technologies (WOOT 18), 2018.

[13] A. Saha, A. Subramanya, K. Patil, H. Pirsiavash, Role of spatial context in adversarial robustness for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 784–785.

[14] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, Y. Chen, Dpatch: An adversarial patch attack on object detectors, SafeAI 2019 (AAAI Workshop on Artificial Intelligence Safety) (2018).

[15] M. Lee, Z. Kolter, On physical adversarial

patches for object detection, arXiv preprint arXiv:1906.11897 (2019).

[16] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.

[17] M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, et al., The pascal visual object classes challenge 2007 (voc2007) results (2008).

[18] S. T. Chen, C. Cornelius, J. Martin, D. H. P. Chau, Shape shifter: Robust physical adversarial attack on faster r-cnn object detector: Recognizing outstanding, D. Research (2019).

[19] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), IEEE, 2017, pp. 39–57.

[20] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, in: International conference on machine learning, PMLR, 2018, pp. 284–293.

[21] S. Thys, W. Van Ranst, T. Goedemé, Fooling automated surveillance cameras: adversarial patches to attack person detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0–0.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.