Improving Adversarial Robustness of Detector via Objectness Regularization

Jiayu Bao[1,2], Jiansheng Chen[1,2,3](✉), Hongbing Ma[1], Huimin Ma[2], Cheng Yu[1], and Yiqing Huang[1]

[1] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
jschen@ustb.edu.cn
[3] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

Abstract. Great efforts have been made by researchers for achieving robustness against adversarial examples. However, most of them are confined to image classifiers and only focus on the tiny global adversarial perturbation across the image. In this paper, we are the first to study the robustness of detectors against vanishing adversarial patch, a physically realizable attack method that performs vanishing attacks on detectors. Based on the principle that vanishing patches destroy the objectness feature of attacked images, we propose objectness regularization (OR) to defend against them. By enhancing the objectness of the whole image as well as increasing the objectness discrepancy between the foreground object and the background, our method dramatically improves the detectors' robustness against vanishing adversarial patches. Compared with other defense strategies, our method is more efficient but robust to adaptive attacks. Another benefit brought by our method is the improvement of recall on hard samples. Experimental results demonstrate that our method can generalize to adversarial patches of different strengths. We reduce the vanishing rate (VR) on YOLOv3 and YOLOv4 under the vanishing attack by 49% and 41% respectively, which is state-of-the-art.

Keywords: Adversarial defense · Vanishing patch · Object detection · Objectness regularization

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success on object detection [10,19,23] and fueled the development of many applications. However, DNNs are found to be easily fooled by adversarial examples [21]. In the computer vision field, adversarial examples are maliciously crafted images that aim at misleading the predictions of DNNs. Typically, for an input image x and a model $F(.)$, the goal of adversarial attacks is to find the adversarial example x' which satisfies Eq. (1), where Δ is a metric to measure the difference between x and x'. In most

H. Ma et al. (Eds.): PRCV 2021, LNCS 13022, pp. 252–262, 2021.
https://doi.org/10.1007/978-3-030-88013-2_21

of studies [17,24,29], Δ is the p-norm ($p = 2$ or ∞) metric, and the perturbation is restricted to be within the p-norm ball of radius ϵ, which means the adversarial perturbation is imperceptible to humans.

$$F(x) \neq F(x') \wedge \Delta(x, x') < \epsilon \tag{1}$$

Object detection is wildly used in many security critical areas like autonomous driving and medical diagnosis. Hence the robustness of deep detectors has attracted great attention. Many attack methods [20,22,30] have been proposed to study the robustness of detectors. DAG [26] misleads the classification of all the bounding boxes, while RAP [14] also includes regression loss in attacks. TOG [8] designs different objectness loss functions to conduct three types of adversarial attacks. However, detectors are wildly used in the real world while attacks with global tiny perturbations are not realizable in the physical world. Among those physically realizable attacks, adversarial patches [4,6,12,16] are the most common attacks which misleading networks with localized image patches. Like what shows in Fig. 1, localized adversarial patches can mislead the object detector to detect nothing. Attackers can implement those patches in the real world and poses real threats to detectors [6,13,30]. So we pay attention to the physically realizable adversarial patch in this work.

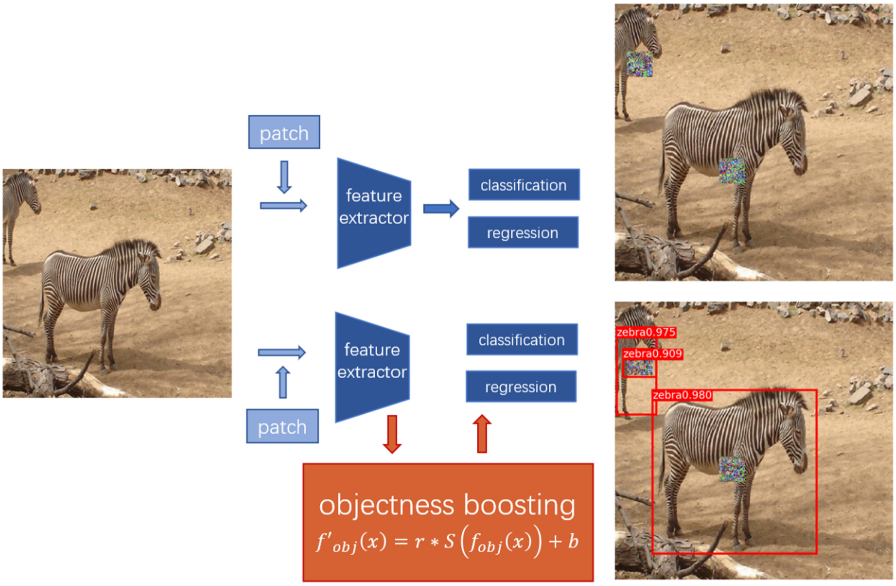

Fig. 1. Illustration of our objectness regularization method. The objectness regularization module is inserted into the detector. We regenerate adversarial patch for the model with objectness regularization.

Defenses against adversarial patches, such as LGS [18] and DW [11], are designed originally for classifiers. Besides, those strategies are proved to be easily circumvented [7] by exploiting BPDA [1] to approximate gradients. By contrast, much less work has been done to improve the adversarial robustness of detectors. It has proven to be an effective way for classifiers to achieve robustness by adversarial training (AT) [17]. AT continuously generates adversarial examples and optimizes the model on these samples in the training phase. By such min-max optimizations, AT can achieve a classifier with a very smooth decision boundary, thus it's more difficult for a benign example to cross the decision boundary with only an imperceptible perturbation. Therefore, AT is extremely time-consuming and usually serves as a trade-off between clean accuracy and adversarial robustness [29]. However, when applying to object detectors, AT can even cause a 25.8 mAP drop on clean images [28], which is far from satisfactory. AT methods are robust to adaptive attacks [1] but only effective to adversarial perturbations of small l_2 or l_∞ norm, which is inappropriate for defending against physically realizable perturbations (often have large l_∞ norm and small l_0 norm).

The vanishing attack [2,8] is the most frequently used adversarial attack on detectors that allowing the object to evade detection. This type of attack is harmful for it is natural-style [25,27] but can be catastrophic to security-critical systems like automatic driving [6,30]. And the adversarial patch that performing vanishing attacks poses security concerns for the employment of detectors in the real world. Hence we shine a light on the robustness of detectors against vanishing adversarial patches in this paper.

Our contributions are as follows:

- To the best of our knowledge, we are the first to bridge the gap of detection robustness against adversarial patches. We develop vanishing adversarial patch to judge the robustness.
- We propose objectness regularization, a simple yet effective method for achieving robustness against vanishing adversarial patches, with a proper trade-off between clean performance and adversarial robustness.
- Our method is efficient and robust to adaptive attacks. We reduce the vanishing rate (VR) on YOLOv3 [10] and YOLOv4 [3] under the adaptive attack by 49% and 41% respectively.

2 Method

In this section, we first revisit the vanishing adversarial patch method in Sect. 2.1. Then we introduce our objectness regularization method in Sect. 2.2.

2.1 Vanishing Adversarial Patch

For an input image x, the object detector outputs bounding box predictions $b(x)$ and classification predictions $c(x)$. The objectness scores of bounding boxes are denoted as $o(x)$, representing the confidence of containing objects. In some detectors [3,10], objectness is directly defined. While in detectors of other structures [19,23], objectness can be represented by the sum of all foreground object probabilities. Objectness plays an import role in distinguishing the foreground objects and the background. So it is often attacked by vanishing attacks [8].

The most import aspects of the vanishing patch are the positions of patches and the vanishing loss function. We choose the positions of patches in an empirical way for attack efficiency. The center of objects are proven to be good choices for vanishing attacks in TOG [8] and SAA [2], so we add vanishing patches at the centers of corresponding objects. Specifically, for an input image x, we get the detection result $F(x)$ and choose patch locations using bounding box positions in $F(x)$. Since it is difficult to optimize the l_0 perturbation directly, we exploit a predefined location mask m with bool values to restrict the adversarial perturbation positions. The adversarial example x' is then denoted as Eq. (2). Here x' denotes the image with perturbation, $\odot$ is the element-wise multiplication operator and p is the vanishing patch we want to get.

$$x' = x \odot (1 - m) + p \odot m \tag{2}$$

We optimize patch p using stochastic gradient descent (SGD) with properly designed loss function and step length. The loss function we employ here is as what exactly used in SAA [2] as Eq. (3). Here $o(x')$ is the objectness predictions of x' and the loss function erases all the predictions with a high objectness. The step length of adversarial attack is fixed in most of cases [5,9,17]. However, we find exponential decay step length will create more powerful adversarial examples under the same attack budgets (e.g. iterations, time). So we use both fixed step length and decay step length attacks to evaluate the adversarial robustness.

$$loss = max(o(x')) \tag{3}$$

Like many other works [17,24,29], we consider the attack budgets in evaluating the robustness of detectors. In this work, we use three indicators to indicate the attack strength, the total pixels of patches (l_0 norm), the maximum iteration numbers (resource constraints), and the step length (optimization strategy).

2.2 Objectness Regularization

The vanishing adversarial patch in the previous Sect. 2.1 is an objectness gradient based attack. To erase objects in detections, vanishing attacks have to fool the objectness of the victim object. In a clean image, foreground objects often have much higher objectness than the background. However, the vanishing patch changes this situation. Compared to clean images, adversarial images with vanishing patches often have much lower objectness. Like what shows in Fig. 2(b), the highest value of objectness feature map is far less than 0, and is far less than 0.1 even after the *sigmoid* activation. We conclude the impacts of the vanishing patch as follows. First, the vanishing patch reduces the whole objectness of the image. Second, the vanishing patch compresses the objectness to a

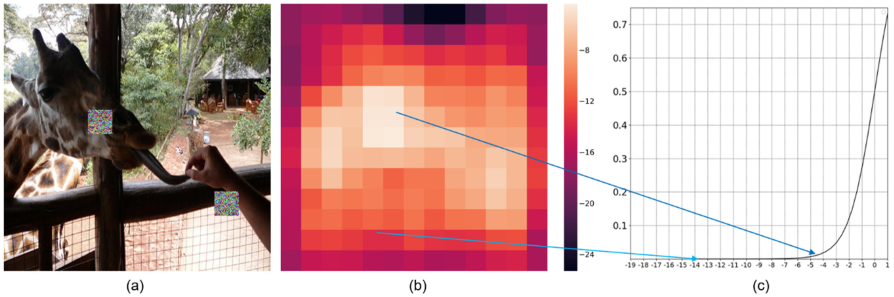

Fig. 2. The objectness distribution of an adversarial image. (a) is an image with vanishing patches and (b) is the objectness distribution of the image. (c) is the sigmoid function. We mark the objectness of the object and the background with blue arrows in (c).

small range, where the discrepancy between the foreground and the background becomes smaller.

To defend against such vanishing attacks, we have to boost the objectness of the image. However, there are two problems we have to solve. First, we should consider the robustness to adaptive attacks. That is, when attackers know our defense strategy and regenerate adversarial examples for attacks, our defense method should be still effective. Second, since the discrepancy between the foreground and the background is smaller, it is of vital importance to avoid too many false predictions when boosting the objectness of the whole image. To solve these problems, we design an objectness boosting function with the post-process parameter of the detector included (to defend adaptive attacks). Before the final classification and regression, we apply our objectness boosting function to correct the objectness feature, like what shows in Fig. 1. We describe the function design in the following.

The basic idea of our method is to make it more difficult for attackers to reduce the objectness of foreground objects. We observe that the foreground object still remains a slightly higher objectness than the background due to the feature discrimination, though the vanishing patch reduces the objectness of the whole victim image to a large scale. So we boost the objectness of the whole image as well as increase the discrepancy between the foreground and the background. We denote the objectness feature of an image x as $f_{obj}(x)$, and the objectness feature after regularization can be formulated by Eq. (4).

$$f'_{obj}(x) = r * S(f_{obj}(x)) + b \tag{4}$$

Here r serves as a range control, b is a necessary balancing bias and $S(.)$ denotes the boosting function which has the form of Eq. (5).

$$S(t) = \frac{1}{1+e^{-t}} \tag{5}$$

We choose the boosting function $S(.)$ as the *sigmoid* function. The *sigmoid* function can map the objectness feature to the range of 0 and 1 while maintaining the internal relative numerical size relationship in it. The object attacked by the vanishing patch has low objectness but is often slightly higher than that of the background. We choose the *sigmoid* function for it also increases the objectness discrepancy between the attacked object and the background when boosting the objectness. As can be seen from Fig. 2, the objectness of the attacked object and that of the background are in different segments of the *sigmoid* function (indicates with blue arrows). In our method, higher objectness results in a greater gain due to the attacked image typically has all values of objectness feature lower than 0. Therefore, we increase the objectness of the foreground object and avoiding too high objectness of the background. That is, our method can enhance object detection under vanishing adversarial attacks without generating too many unreasonable bounding boxes.

The regularization parameters r and b are closely related to the objectness threshold τ. The τ is used in the post process of the detector to erase redundant bounding boxes. We argue that a larger tangent slope in the *sigmoid* function corresponding to τ is supposed to have a smaller range r. So we define r as the reciprocal of the *sigmoid* derivative. Due to the convenience of *sigmoid* derivative calculation, we can easily get r as Eq. (6). We introduce a constant product factor 1/4 into Eq. (6) to avoid too high objectness.

$$r = \frac{1}{4\tau(1-\tau)} \tag{6}$$

The b is an essential bias that control the lowest prediction objectness. We must ensure that the lowest objectness after regularization is high enough for achieving robustness against adaptive adversarial attacks. However, considering the time consumption of post-process, the lowest objectness after regularization should not be higher than $S^{-1}(\tau)$ (otherwise there will be too many redundant bounding boxes with prediction objectness higher than threshold τ). Therefore we design the b as Eq. (7), where $S^{-1}(.)$ is the inverse function of the *sigmoid* function and ϵ is a small constant to filter redundant bounding boxes with relative low objectness. We choose the value of ϵ quite empirically and will study the effect of it in Sect. 3.

$$b = S^{-1}(\tau) - \epsilon = \ln(\frac{\tau}{1-\tau}) - \epsilon \tag{7}$$

3 Experiment

In this section, we evaluate the proposed objectness regularization (OR) method. Both standard object detection and adversarial object detection are investigated.

3.1 Experimental Setup

We introduce our experimental setup in this section. For standard object detection and adversarial object detection, we use different settings accordingly.

Datasets. For standard object detection, we evaluate the performance of detectors on COCO2014 [15] validation set. However, many tiny objects are contained in COCO dataset and will be covered by patches directly. For a fair judgment, we generate adversarial examples on 1000 images chosen from the COCO dataset. All the objects in the chosen 1000 images are large enough to not be covered by patches directly.

Models. We evaluate our method on YOLOv3 [10] and YOLOv4 [3] that are both pre-trained on COCO2014 [15]. The input size of both detectors is 416 * 416 pixels in our experiments. The performances of detectors on clean images are evaluated on COCO validation set with the objectness threshold of 0.001 for non-max suppression (NMS). While the objectness threshold τ of detectors is set to be the most frequently used 0.5 in all our robustness evaluations.

Patches. The method in Sect. 2.1 is exploited to generate adversarial patches. We evaluate our defense method under patches with different strengths. The attacks are strong enough that the iteration number of patches is set to be at least 100. We generate adversarial patches for every single image independently. And we evaluate our method under defense aware adversarial patches. That is, we regenerate vanishing patches for models equipped with objectness regularization.

Metrics. The mAP is chosen to demonstrate the performance on clean images. For convenience, we use mAP-50 in experiments. While the vanishing rate (VR) is introduced to demonstrate the performance under vanishing attacks. The lower the VR, the better robustness of detectors against vanishing patches. The VR is calculated as Eq. (8), where $B(x)$ and $B(x')$ denote the prediction of the clean image x and the adversarial image x' severally. The $IOU(.)$ is a function that calculates the reasonable detections in $B(x')$ where $B(x)$ serves as the ground truth result.

$$VR = 1 - \frac{IOU(B(x), B(x'))}{B(x)} \tag{8}$$

3.2 Experimental Result

The performance of detectors on clean images and adversarial images is reported in this section.

Resilience to Adaptive Attacks. We investigate the VR of detectors under attacks with different strengths. We exploit adversarial patches of various sizes (2500, 3000, and 3500 total pixels respectively) to attack detectors. For each size of the patch, we design three iteration number budgets (100, 200, and 255 respectively) for robustness evaluation. A constant update step of 1/255 is used in all attacks of 100 and 200 iterations. While a decaying update step with initial update step 8/255, decay rate 1/2 and decay point at 30, 90, 190 is employed in all attacks of 255 iterations.

The VR on YOLOv3 and YOLOv4 under attacks with different budgets are presented in Table 1. The hyper-parameter ϵ in experiments of Table 1 is 0.01 for YOLOv3 and 0.05 for YOLOv4. From Table 1, adversarial patches using

Table 1. The VR on YOLOv3 and YOLOv4 (with and without OR defense) under attacks of different strengths. Attacks to defense models are all adaptive attacks in this table.

Patch pixels	Attack iters	YOLOv3	YOLOv3_OR	YOLOv4	YOLOv4_OR
2500	100	34.5%	20.4% (↓ 14.1%)	28.3%	25.4% (↓ 2.9%)
	200	48.1%	22.1% (↓ 26.0%)	37.5%	27.6% (↓ 9.9%)
	255	79.7%	31.5% (↓ 48.2%)	59.5%	32.0% (↓ 27.5%)
3000	100	41.3%	21.9% (↓ 19.4%)	34.4%	26.5% (↓ 7.9%)
	200	56.2%	23.9% (↓ 32.3%)	45.7%	29.0% (↓ 16.7%)
	255	85.2%	37.4% (↓ 47.8%)	70.1%	34.5% (↓ 35.6%)
3500	100	48.8%	23.1% (↓ 25.7%)	38.3%	27.3% (↓ 11.0%)
	200	64.4%	26.0% (↓ 38.4%)	52.5%	30.0% (↓ 22.5%)
	255	91.0%	42.2% (↓ **48.8%**)	78.3%	37.0% (↓ **41.3%**)

strategies of SAA greatly increase the VR on the two detectors, with even 91% and 78% of objects evade detection. Our method reduces the VR on YOLOv3 and YOLOv4 by 48.8% and 41.3% respectively under the strongest attack in Table 1. As also can be seen from Table 1, the OR method is particularly effective against strong adversarial attacks, which is more representative of the real robustness of models.

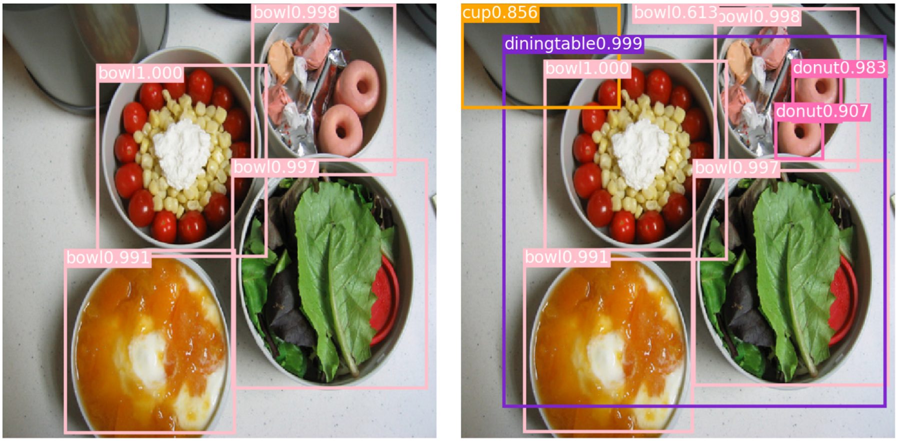

Fig. 3. Detection results on a clean image. *left*: detection results of YOLOv3, *right*: detection results of YOLOv3_OR.

Effects on Clean Images. Our method changes the original objectness for achieving detection robustness. However, the method only causes a decrease of 2.67 mAP on clean images in YOLOv3, as demonstrated in Table 2. The performance of YOLOv4 on clean images only has a slight drop from 56.74 mAP to 56.45 mAP. Despite the comprehensive performance drop on clean images, our method improves the recall of weak-feature objects significantly. It can be observed from Table 2 that the recall of tiny objects like spoon and baseball glove has an increase of over 5% when using OR. The recall of the refrigerator even reaches 96.3% with our method. As demonstrated in Fig. 3, our method is helpful to detect small objects like donuts and hard samples like the dining-table and the cup.

Table 2. The recall of some COCO dataset categories in YOLOv3 with and without OR.

Category	YOLOv3 (54.30 mAP)	YOLOv3_OR (51.63 mAP)
Stop sign	88.1%	90.5%(↑ 2.4%)
Baseball glove	67.6%	72.7%(↑ 5.1%)
Spoon	62.5%	71.1%(↑ **8.6%**)
Banana	70.9%	76.6%(↑ 5.7%)
Refrigerator	90.7%	**96.3%**(↑ 5.6%)

Ablation Study. The effects of the hyper-parameter ϵ are presented in Table 3. The adversarial attack used in Table 3 has a strength of pixel 3000 and iteration number 255. It's obvious that a smaller ϵ typically results in a better performance on adversarial images and a slightly worse performance on clean images, at the expense of inference speed. The negative values of ϵ are abandoned by us in experiments due to the explosion of inference time. We choose ϵ as 0.01 for balance in most of the experiments.

Table 3. The effects of ϵ on clean mAP (mAP on clean images of COCO validation set), VR, and inference time per image (test on GTX 2080 Ti) in YOLOv3_OR.

ϵ	Clean mAP	VR (%)	Time (ms)
0.3	50.32	76.2	23.6
0.1	51.69	58.3	23.9
0.01	51.63	37.4	27.0
0.001	51.63	28.9	41.8
0.0001	51.63	27.4	96.0

4 Conclusion

In this paper, we propose a defense method called objectness regularization (OR) against vanishing adversarial patch attacks. The *sigmoid* function is chosen to enhance the image objectness as well as increase the objectness discrepancy between the foreground object and the background. Our method is efficient (compared to adversarial training methods) but effective against adaptive adversarial attacks. The experimental results on YOLOv3 and YOLOv4 demonstrate that OR can generalize to attacks of different strengths. This method significantly improves the recall of hard samples on clean images with only little mAP degradation. We will further generalize our method to detectors of other structures and adjust it to resist different types of adversarial attacks.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 61673234, No. U20B2062), and Beijing Science and Technology Planning Project (No. Z191100007419001).

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: International Conference on Machine Learning, pp. 274–283 (2018)
2. Bao, J.: Sparse adversarial attack to object detection. arXiv preprint arXiv:2012.13692 (2020)
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
4. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
6. Chen, S.-T., Cornelius, C., Martin, J., Chau, D.H.P.: ShapeShifter: robust physical adversarial attack on faster R-CNN object detector. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11051, pp. 52–68. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10925-7_4
7. Chiang, P.y., Ni, R., Abdelkader, A., Zhu, C., Studor, C., Goldstein, T.: Certified defenses for adversarial patches. In: International Conference on Learning Representations (2019)
8. Chow, K.H., et al.: Adversarial objectness gradient attacks in real-time object detection systems. In: 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pp. 263–272. IEEE (2020)
9. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193 (2018)
10. Farhadi, A., Redmon, J.: YOLOv3: an incremental improvement. Computer Vision and Pattern Recognition, cite as (2018)
11. Hayes, J.: On visible adversarial perturbations & digital watermarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1597–1604 (2018)
12. Karmon, D., Zoran, D., Goldberg, Y.: LaVAN: localized and visible adversarial noise. In: International Conference on Machine Learning, pp. 2507–2515 (2018)

13. Komkov, S., Petiushko, A.: AdvHat: real-world adversarial attack on ArcFace face ID system. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 819–826. IEEE (2021)
14. Li, Y., Tian, D., Bian, X., Lyu, S.: Robust adversarial perturbation on deep proposal-based models
15. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
16. Liu, X., Yang, H., Liu, Z., Song, L., Chen, Y., Li, H.: DPatch: an adversarial patch attack on object detectors. In: SafeAI@ AAAI (2019)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
18. Naseer, M., Khan, S., Porikli, F.: Local gradients smoothing: defense against localized adversarial attacks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1300–1307 (2019)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, pp. 91–99 (2015)
20. Song, D., et al.: Physical adversarial examples for object detectors. In: 12th USENIX Workshop on Offensive Technologies (WOOT 2018) (2018)
21. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
22. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
23. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636 (2019)
24. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems 33 (2020)
25. Wu, Z., Lim, S.-N., Davis, L.S., Goldstein, T.: Making an invisibility cloak: real world adversarial attacks on object detectors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 1–17. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_1
26. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1369–1378 (2017)
27. Xu, K., et al.: Adversarial T-shirt! Evading person detectors in a physical world. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 665–681. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_39
28. Zhang, H., Wang, J.: Towards adversarially robust object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 421–430 (2019)
29. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning, pp. 7472–7482. PMLR (2019)
30. Zhao, Y., Zhu, H., Liang, R., Shen, Q., Zhang, S., Chen, K.: Seeing isn't believing: towards more robust adversarial attack against real world object detectors. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 1989–2004 (2019)