Adversarial Patch: Physical Patch in Carla Simulator

Anonymous CVPR submission

Paper ID *****

Abstract

We plan to generate adversarial patches against object detection models in a simulated environment by changing the material or texture of an object, which is a special kind of physical patch.

1. Introduction

Adding a small and intentional drift to the input distribution, also known as adversarial perturbations, can substantially decrease the deep neural network's performance. Adversaries can exploit optimization methods, sensitive features, geometric transformations, and generative models to generate adversarial examples [15].

In 2017, Brown et al. brought adversarial examples to the physical world by printing human-noticeable patches on stickers [1], posing threats against real-world applications. Then, research interests gradually shifted from digital attacks (modifying image pixel values) to physical attacks.

Adversarial examples in the physical world must be able to be captured by sensors, such as cameras. As a result, physical attacks require a substantial intensity of perturbations, making them visually perceptible by human eyes. Among physical attacks, adversarial patches [1] are the most widely studied methods. In a survey on visually adversarial attacks, Wei et al. categorized adversarial patches into meaningful patches (e.g., QR Code) and meaningless patches that do not correspond to real-world objects [23].

Though most physical attacks are visually perceptible by human eyes, some optical attacks can only be captured by sensors (e.g., rolling shutter attacks). In [8], Li et al. summarized optical adversarial attacks, including attacks that use high-frequency light, laser, and projector.

Our research will focus on adversarial patches that are noticeable by human eyes. In [16], Sharma et al. surveyed adversarial patch attacks in vision-based tasks that involve three mainstream models: classification, detection, and re-identification [22], and we will focus on adversarial attacks against object detection models.

2. Digital Patch

Digital patches generate adversarial examples by directly modifying image pixel values. Some digital patches can be applied to the physical world by adding extra constraints during the optimization. For example, Lee et al. extended digital DPatch [9] to the physical world [7].

However, some research generates asteroid and grid-shaped patches [24] or small patches [6] to reduce the number of perturbed pixels, making it infeasible to be printed out on physical objects.

Besides, the effectiveness of adversarial patches should be position invariant. Digital patches can fool detection models without overlapping with objects [14].

3. Physical Patch

Physical patches pose a great threat against autonomous driving vehicles as they are invariant to input images and thus can inherently achieve real-time attacks [18].

Prior research generates stop signs that cannot be recognized by detection models [17] [2], and adversarial posters to vanish pedestrian [19] [21]. Chindaudom et al. produce meaningful patches by combining patches and QR Codes [3] [4]. Though most patches are static, Hoory et al. generate and display dynamic patches on a monitor attached to a vehicle [5]. However, physical patches can only attack object detection models when the camera is close enough (see the demo video [20] [10]). Besides, testing physical patches could waste a lot of printing materials.

For autonomous driving, it is more popular to test safety-critical edge cases in simulation. Testing physical attacks in simulators is more efficient than in the physical world since we can easily vary weather and lighting conditions. In simulators, adversarial patches can be applied by changing the material or texture of an object, but prior research only applies patches by modifying pixel values [11] [12] [13].

4. Summary

In summary, most prior research tests physical attacks in the real world or applies adversarial patches by modifying pixel values in simulators.

References

[1] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 1

[2] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 52–68. Springer, 2019. 1

[3] Aran Chindaudom, Prarinya Siritanawan, Karin Sumongkayothin, and Kazunori Kotani. Adversarialqr: An adversarial patch in qr code format. In *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 1–6. IEEE, 2020. 1

[4] Aran Chindaudom, Prarinya Siritanawan, Karin Sumongkayothin, and Kazunori Kotani. Surreptitious adversarial examples through functioning qr code. *Journal of Imaging*, 8(5):122, 2022. 1

[5] Shahar Hoory, Tzvika Shapira, Asaf Shabtai, and Yuval Elovici. Dynamic adversarial patch for evading object detection models. *arXiv preprint arXiv:2010.13070*, 2020. 1

[6] Hao Huang, Yongtao Wang, Zhaoyu Chen, Zhi Tang, Wenqiang Zhang, and Kai-Kuang Ma. Rpattack: Refined patch attack on general object detectors. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1

[7] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*, 2019. 1

[8] Haiyan Li, Bing Bai, and Geyang Xiao. A survey on optical adversarial examples against dnns. In *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 1–6. IEEE, 2022. 1

[9] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018. 1

[10] Mingming Lu, Qi Li, Li Chen, and Haifeng Li. Scale-adaptive adversarial patch attack for remote sensing image aircraft detection. *Remote Sensing*, 13(20):4078, 2021. 1

[11] Yael Mathov, Lior Rokach, and Yuval Elovici. Enhancing real-world adversarial patches through 3d modeling of complex target scenes, 2021. 1

[12] Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2280–2289, 2022. 1

[13] Giulio Rossolini, Federico Nesti, Gianluca D'Amico, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving, 2022. 1

[14] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. Adversarial patches exploiting contextual reasoning in object detection. *arXiv preprint arXiv:1910.00068*, 2019. 1

[15] Alexandru Constantin Serban, Erik Poll, and Joost Visser. Adversarial examples - a complete characterisation of the phenomenon, 2019. 1

[16] Abhijith Sharma, Yijun Bian, Phil Munz, and Apurva Narayan. Adversarial patch attacks and defences in vision-based tasks: A survey. *arXiv preprint arXiv:2206.08304*, 2022. 1

[17] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018. 1

[18] Michael Threet, Colin Busho, Josh Harguess, Melanie Jutras, Nicole Lape, Sara Leary, Keith Manville, Mike Tan, and Chris Ward. Physical adversarial attacks in simulated environments. In *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–5. IEEE, 2021. 1

[19] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[20] Derui Wang, Chaoran Li, Sheng Wen, Qing-Long Han, Surya Nepal, Xiangyu Zhang, and Yang Xiang. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples, 2021. 1

[21] Yajie Wang, Haoran Lv, Xiaohui Kuang, Gang Zhao, Yu-an Tan, Quanxin Zhang, and Jingjing Hu. Towards a physical-world adversarial patch for blinding object detection models. *Information Sciences*, 556:459–471, 2021. 1

[22] Hui Wei, Hao Tang, Xuemei Jia, Hanxun Yu, Zhubo Li, Zhixiang Wang, Shin'ichi Satoh, and Zheng Wang. Physical adversarial attack meets computer vision: A decade survey. *arXiv preprint arXiv:2209.15179*, 2022. 1

[23] Xingxing Wei, Bangzheng Pu, Jiefan Lu, and Baoyuan Wu. Visually adversarial attacks and defenses in the physical world: A survey, 2023. 1

[24] Shudeng Wu, Tao Dai, and Shu-Tao Xia. Dpattack: Diffused patch attacks against universal object detection. *arXiv preprint arXiv:2010.11679*, 2020. 1