Supplementary Materials for Revisiting Adversarial Patches for Designing Camera-Agnostic Attacks against Person Detection

Hui Wei[1]* **Zhixiang Wang**[2]* **Kewei Zhang**[1]*
Jiaqi Hou[1] **Yuanwei Liu**[1] **Hao Tang**[3] **Zheng Wang**[1]†
[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University
[2]The University of Tokyo [3]School of Computer Science, Peking University
https://camera-agnostic.github.io/

A Workflow

Adversarial patches are commonly used to attack person detection models [1, 2, 3, 8, 9]. As shown in Figure A , we summarize the workflow of these methods. Typically, designing a patch-based physical adversarial attack involves five general steps: ❶ Adversarial patch generation, ❷ Adversarial patch manufacturing, ❸ Attack deployment, ❹ Threat image capturing, and ❺ Attack launching. In these steps, we can observe two domain transitions. Specifically, the first transition takes place in Step ❷, where digital patches are translated into tangible real-world patch material. The second transition occurs in Step ❹, where the physical scene undergoes the transformation into digital images, typically achieved by hardware camera devices.

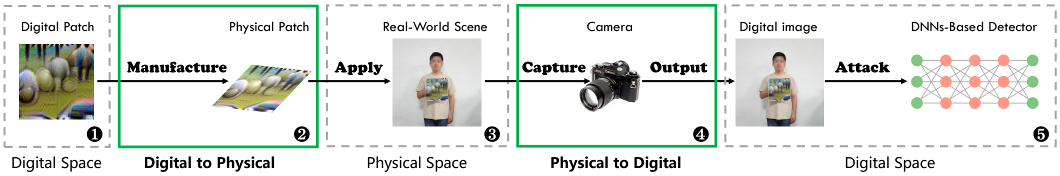


Figure A: **Workflow of patch-based physical adversarial attacks against person detection.** Generally, designing a patch-based physical adversarial attack involves five steps: adversarial patch generation, adversarial patch manufacturing, attack deployment, threat image capturing, and attack launching.

B Attacks under Multiple Detectors

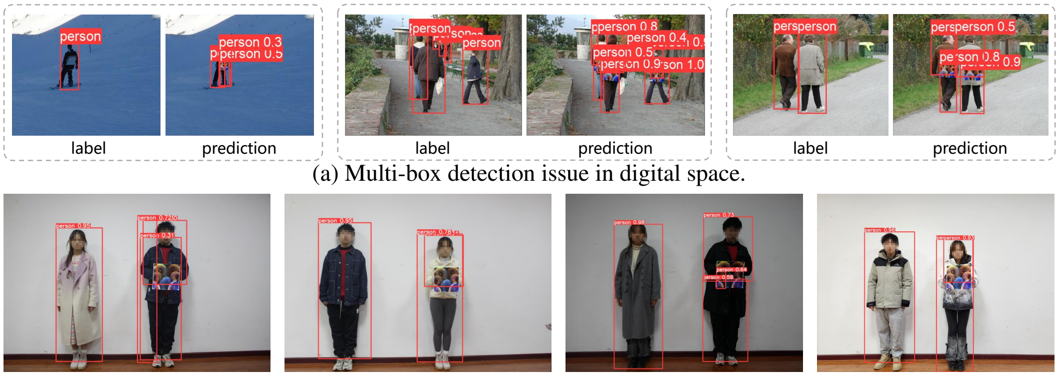
Our primary investigation focuses on maintaining physical adversarial attack effectiveness across multiple cameras, and we initially conducted experiments on the YOLOv5 [4] detector. To comprehensively evaluate the cross-model generalization capability of our proposed method, we further assess its performance across different object detection models, specifically YOLOv3 [7] and YOLOv8 [5]. Table A presents the comparative results. The experimental results demonstrate that our proposed CAP method consistently outperforms existing approaches across all tested architectures in terms of Average Precision (AP) and Attack Success Rate (ASR). Specifically, CAP achieves ASRs of 43.3%, 54.4%, and 14.7% on YOLOv3, YOLOv5, and YOLOv8, respectively. More recent approaches like NAP [1] and LAP [8], while focusing on naturalistic perturbations, demonstrate limited cross-model attack capability, with ASRs below 15% across all tested models.

*Equal contribution †Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table A: Open-source resources utilized in this paper.

Method	YOLOv3		YOLOv5		YOLOv8	
	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑
Random Noise	71.3	11.3	81.7	7.3	77.9	4.0
AdvPatch [9]	48.1	33.3	67.7	19.7	75.6	8.8
AdvT-shirt [13]	55.8	24.4	76.6	14.6	77.2	6.2
AdvCloak [12]	52.7	30.5	70.5	12.6	73.7	10.1
NAP [1]	66.2	14.0	81.3	7.4	78.1	5.0
LAP [8]	65.2	14.6	81.0	5.6	78.6	4.6
TC-EGA [2]	56.9	24.7	79.9	8.8	77.3	6.7
CAP (Ours)	41.5	43.3	37.7	54.4	60.5	14.7



(a) Multi-box detection issue in digital space.

(b) Multi-box detection issue in physical space.

Figure B: **Illustration of the multi-box detection issue in the T-SEA attack [3].** We observe that applying the T-SEA patch results in multiple bounding boxes for the same person instance. This issue exists in both digital and physical spaces.

C Multi-Box Detection Issue

We discovered a multi-box detection issue in the T-SEA attack [3], which occurs in both digital and physical spaces. This issue explains why the attack method significantly reduces Average Precision (AP) but does not achieve a high Attack Success Rate (ASR).

Figure Ba presents some typical examples in digital space. We observe that the detector predicts multiple bounding boxes for a single person instance compared to the ground truth labels. Figure Bb illustrates the attack in physical space. The multi-box detection issue does not occur when the person is without an adversarial patch, but it appears when the person is equipped with the T-SEA patch.

In object detection tasks, the AP value is related to the crucial metric of Precision [6]. The formula for calculating Precision is as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

where TP denotes the true positives and FP represents the false positives. Precision quantifies the ratio of correctly predicted positive instances to the total predicted positive instances. Clearly, the multi-box detection issue results in an increase in FP, consequently reducing Precision and AP. Therefore, although the AP decreases, the attack is not truly successful, as the person is not hidden from the detector.

D Hyperparameters of Camera ISP Proxy Net

Camera ISPs involve multiple processing stages, summarized by Tseng *et al.* [10] as follows: (1) Optics, (2) White Balance & Gain, (3) Demosaicking, (4) Denoising, (5) Color & Tone Correction, and (6) Color Space Conversion & Compression. While the first three stages apply to RAW data, Table B: **Hyperparameters we select from the software camera ISP for building a differentiable camera ISP proxy network.** We select six parameters, with four belonging to the Color & Tone Correction module and two to the Denoising module.

(a) Color & Tone Correction

Parameter	Symbol	Value interval	Max
Brightness Contrast Control	a	(64, 256)	2^8
Hue Saturation Control	b	(64, 256)	2^8
Gamma Adjustment	γ	(0.4, 2.0)	2^1
Color Correction Matrix	c	(512, 1024)	2^{10}

(b) Denoising

Parameter	Symbol	Value interval	Max
Spatial Filtering	d	(0.1, 2.0)	2^1
Non-Local Means	e	(1.0, 32.0)	2^5

the latter three operate on RGB values. As our task centers on RGB images, we selected conditional parameters from the final three stages, focusing on six critical factors, such as Brightness Contrast Control and Gamma Adjustment, shown in Table B. Experimental results reveal that these parameters significantly affect attack outcomes.

Certain parameter combinations may result in complete information loss. To ensure image quality and diversity from the ISP proxy network, we defined ranges as specified in Table B. For instance, values for parameter a below 64 yield overly dark images, while values for b under 64 lead to desaturated colors. Adjustments to γ settings above the range introduce noise in dark regions, while lower values diminish contrast. Deviation in parameter c can degrade image quality, insufficient values for d introduce excessive noise, and e values below 1 inadequately suppress noise.

E Additional Results

E.1 Qualitative Analysis of Digital-Space Attacks

Figure C demonstrate the digital-space attacks of seven different patch configurations under five different camera ISP settings. We notice that changes in ISP affect image attributes like brightness and contrast. This, in turn, impacts the attack performance. Specifically, we find that varying camera ISPs have minimal impact on benign images, as the detector consistently identifies person instances across all four ISP settings with marginal confidence variation. A similar phenomenon occurs for adversarial patches with low attack effectiveness, like Random Noise. These results indicate that the person detector is inherently robust, having camera-agnostic detection capabilities. AdvPatch [9] and T-SEA [3], the two comparative methods, did not successfully attack all 4 camera ISPs. However, the confidence of person instances exhibited noticeable fluctuations. For example, T-SEA decreased from the highest score of 0.92 (ISP 2) to 0.71 (ISP 3). Ours without camera ISP has demonstrated improved attack effectiveness, yet it is influenced by the camera ISP. Successful attacks are observed under ISP 1 and ISP 4, while attacks fail under ISP 2 and ISP 3. Ours without adversarial optimization is similarly affected by variations in camera ISPs, achieving success only under ISP 1. In contrast, our full method maintains stable attacks across 5 settings, successfully concealing the person. This results illustrate that our CAP mitigates the instability of cross-camera attacks and enhances the attack efficacy of adversarial patches.

E.2 50 Random Camera ISPs Used for Evaluation.

To validate the camera-agnostic attack capability of our method, we selected 50 random camera ISP settings to evaluate CAP. Given that the samples are six-dimensional data, we employed t-SNE [11] for visualization. As shown in Figure D, we observe that the samples tested cover the entire sample space, reflecting the performance of our adversarial patches across diverse camera ISP settings and enhancing the credibility of our experimental results.

Furthermore, compared to ours without o the camera ISP, Figure D illustrates that our full method exhibits superior attack performance (lower AP) at each sample. Additionally, we computed the standard deviations of the AP for both baseline and our method across 50 samples, which are 3.89 and 1.82, respectively. This indicates that our method demonstrates better robustness to changes in camera ISP settings.

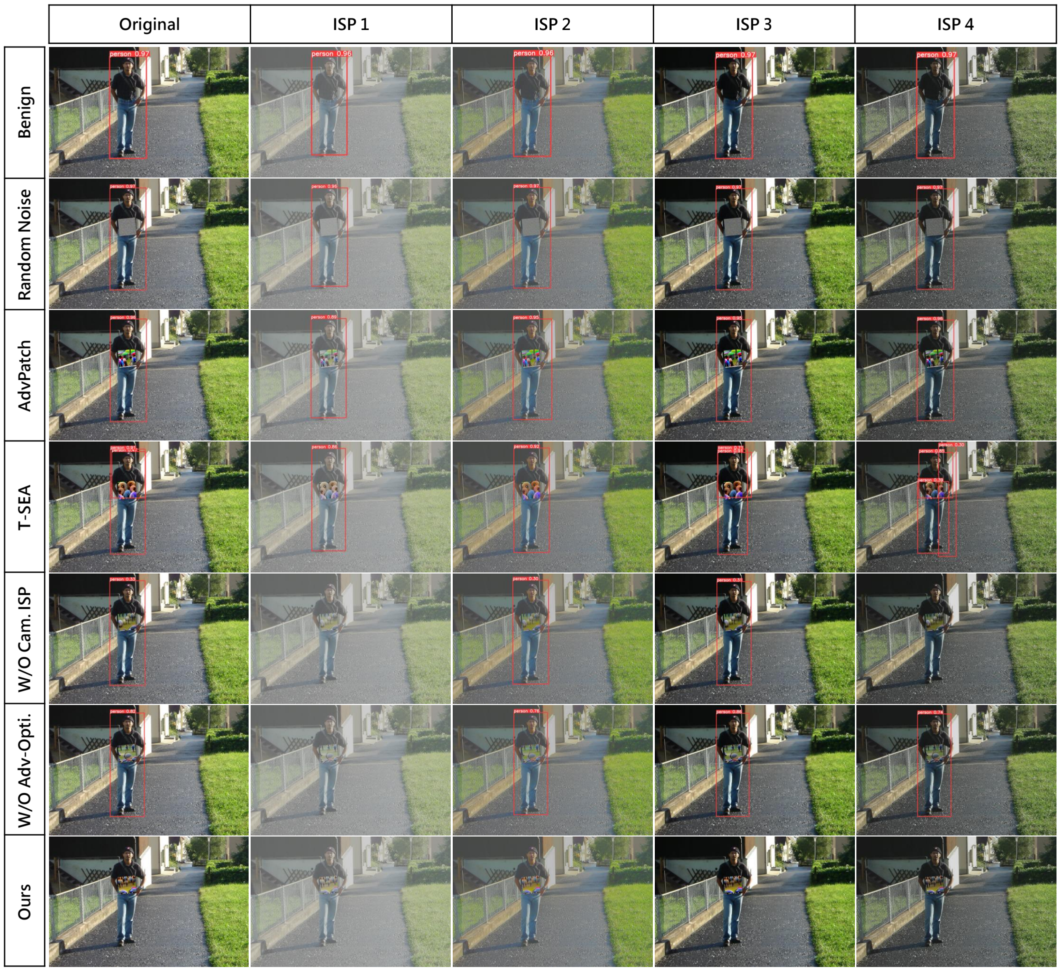
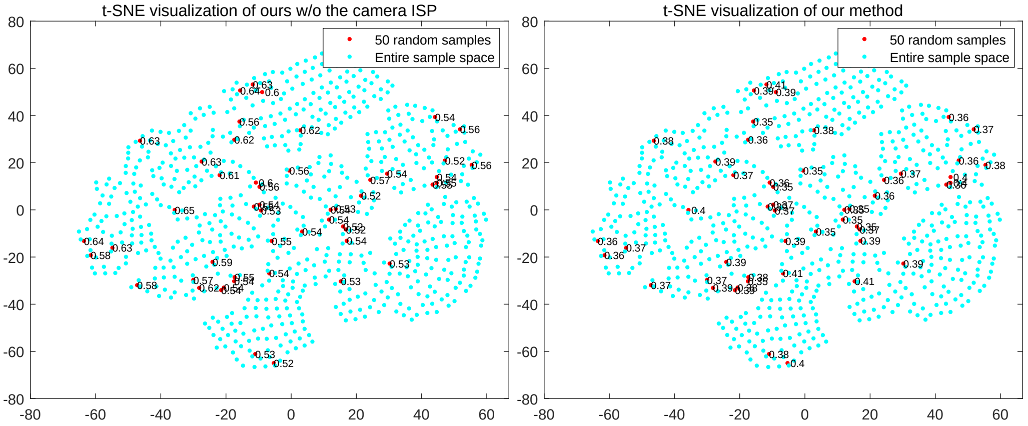


Figure C: **Display of digital-space attacks.** We demonstrate the attacks of seven different patch configurations under five different camera ISP settings. The bounding boxes indicate the YOLOv5 [4] successfully detects the person instances, i.e., the attack fails.

E.3 Qualitative Analysis of Physical-Space Attacks

In Figure E, we display two comparative methods (AdvPatch [9] and T-SEA [3]) alongside our method. We can observe significant imaging variations when the same scene is captured by different cameras. These disparities manifest in features such as brightness and saturation in the images. For instance, images captured by the selected Sony device exhibit the lowest brightness among the 6 devices. These imaging differences have an impact on the attack effectiveness of adversarial patches. For AdvPatch [9] and T-SEA [3], we observe that participants without carrying adversarial patches exhibit stable recognition by detectors across different cameras, maintaining a confidence level of around 0.96. However, for participants carrying adversarial patches, there is a significant fluctuation in the confidence of person instances. For instance, in the case of AdvPatch, the confidence of the attacker is below 0.25 on the Samsung camera (0.25 being the confidence threshold set by the detector; instances below this threshold are discarded), while on the iPhone camera, the confidence reaches as high as 0.90. T-SEA induces incomplete detection (half-body) or multiple detections in detectors, yet consistently fails to conceal the presence of a person. The consistent performance improvement of our adversarial patch, resulting in successful attacks across all six cameras, underscores the capability of our CAP for real-world cross-camera attacks.



(a) AP of ours w/o the camera ISP.

(b) AP of our full method.

Figure D: **The entire sample space of Camera ISP input hyperparameters and our sampled points**. For digital-space attack evaluation, we randomly sampled 50 sets of camera ISP input hyperparameters to assess the cross-camera attack capability. Here, we employ t-SNE visualization to illustrate them, with AP values annotated at corresponding locations.

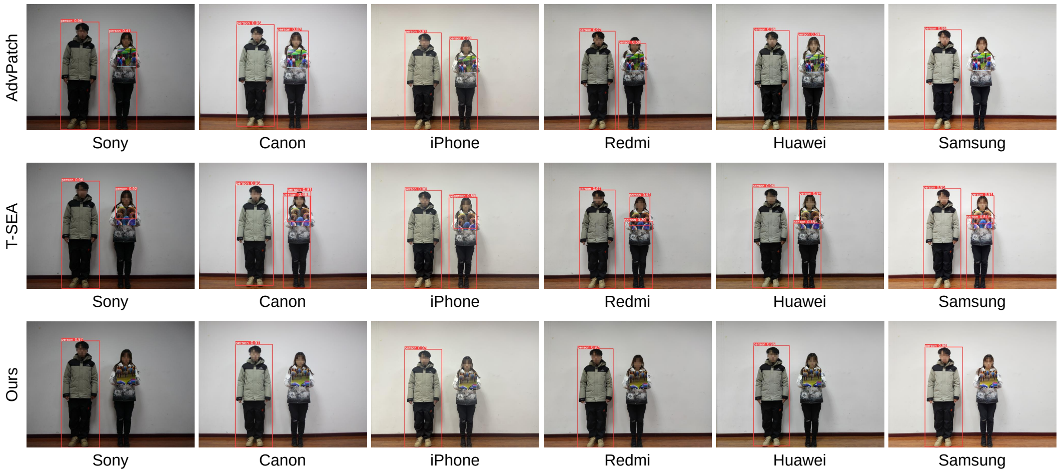


Figure E: **Display of physical-space attacks.** We showcase the attack outcomes of three distinct patch configurations across six different cameras. The bounding boxes indicate the detector successfully detects the person instances, i.e., the attack fails.

F Licenses

Table C provides a list of the resources that have been used in this research paper and their associated licenses.

Table C: Open-source resources utilized in this paper.

Name	License	URL
INRIAPERSON Dataset	CC BY 4.0	link
YOLOv5	AGPL-3.0, Enterprise	link
COCO Dataset	Creative Commons Attribution 4.0	link
fast-openISP	MIT	link
Pytorch	BSD-style	link

References

[1] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[2] Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13307–13316, 2022.

[3] Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20514–20523, 2023.

[4] Glenn Jocher. Ultralytics yolov5, 2020.

[5] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[7] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[8] Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *ACM International Conference on Multimedia*, 2021.

[9] 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.

[10] Ethan Tseng, Felix Yu, Yuting Yang, Fahim Mannan, Karl ST Arnaud, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Hyperparameter optimization in black-box image processing using differentiable proxies. *ACM Trans. Graph.*, 38(4):27–1, 2019.

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[12] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, 2020.

[13] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision*, 2020.