

Defending Against Physical Adversarial Patch Attacks on Infrared Human Detection

Lukas Strack^{1*}, Futa Waseda^{2,3*}, Huy H. Nguyen³, Yinqiang Zheng², and Isao Echizen^{2,3}.

(* Equal contribution)

¹ University of Freiburg, Germany

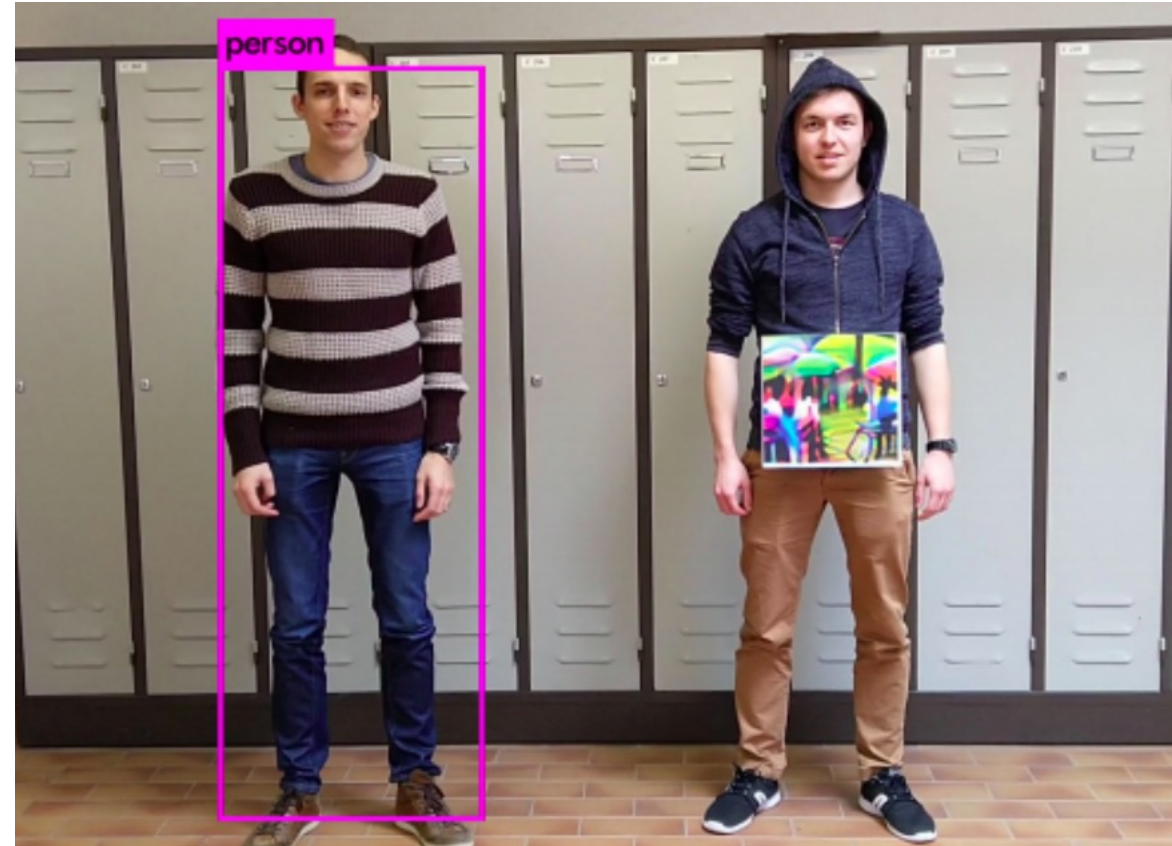
² The University of Tokyo, Japan

³ National Institute of Informatics, Japan

Correspondence to: Lukas Strack <christian.tomani@tum.de>, Futa Waseda <futa-waseda@nii.ac.jp>



- **Adversarial Patch**
 - Localized
 - Visible
 - Physically implementable
- **Realistic attack scenario!**



[Simen et al., "Fooling automated surveillance cameras: adversarial patches to attack person detection." CVPRW 2019.]

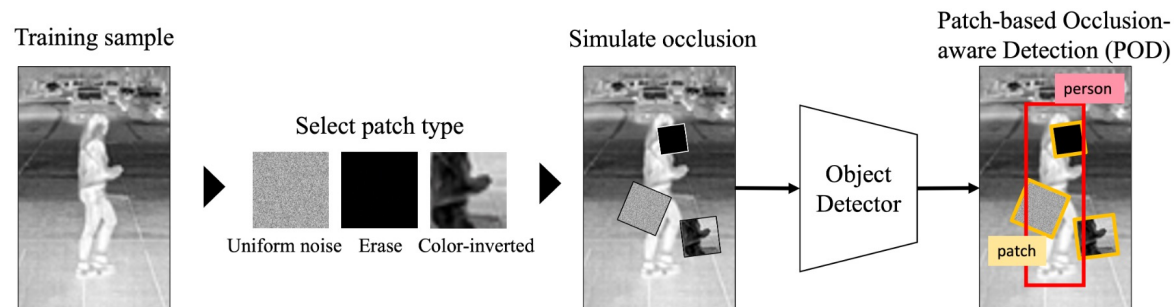
Motivation: Adversarial Patch × Infrared Detection

- **Infrared detectors:**
 - Used in safety-critical tasks, **but** vulnerable to **adversarial patches**
- **Infrared Adversarial Patch:**
 - Manipulate temperature patterns, using thermal objects
- **Challenge:**
 - Defending against **physically realizable** infrared attacks is **underexplored**



e.g., HotCold Block Attack [Wei+ 2023]

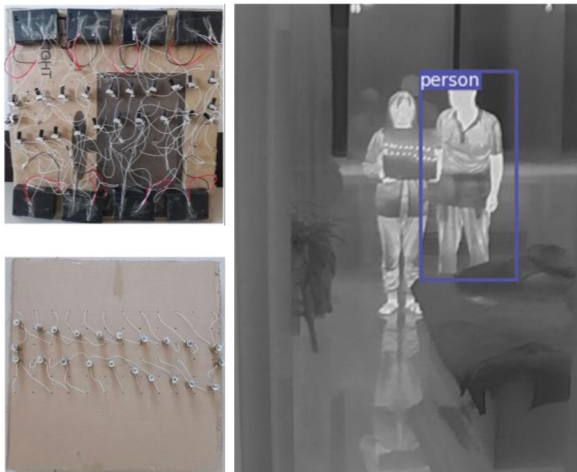
- **The first defense** study against **infrared adversarial patch**
- **We propose Patch-based Occlusion-aware Detection (POD)**, a simple robust training method that is
 - **Efficient**: little-to-no training overhead
 - **Generalizable**: robust to unseen patch attacks



	Defense Capability			Training Efficiency
	Adv.	HC	Shape	
Std. Training				✓
POD _{noDet}	△	✓	✓	✓
POD	✓	✓	✓	✓
Adv-POD _{noDet}	✓	✓	✓	
Adv-POD	✓	✓	✓	

We extend robust-feature learning beyond digital perturbations to physical-world infrared patch attacks

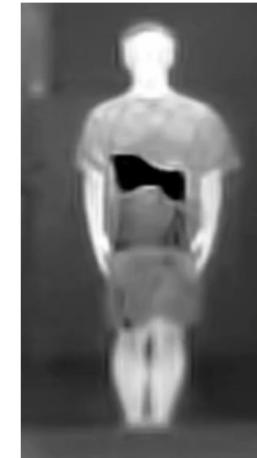
- **Attackers' Goal:**
 - Universal patch to fool infrared detector (viewpoint agnostic)
- **Attackers' Constraints:**
 - Infrared images have **less texture** than RGB images
 - Physical limitations: **low resolution, binary heat maps**



Bulb attack [Zhu+ 2021]



HotCold Block [Wei+ 2023]



Shape-Loc. Attack [Xingxing+ 2023]

- **Attackers' Goal:**

- Universal patch to fool infrared detector (viewpoint agnostic)

- **Hypothesis:**

Infrared patch attacks are largely **occlusion-like**, not fine-grained feature manipulation

→ **Occlusion-robust training** may improve robustness against infrared patch attacks!



Bulb attack [Zhu+ 2021]



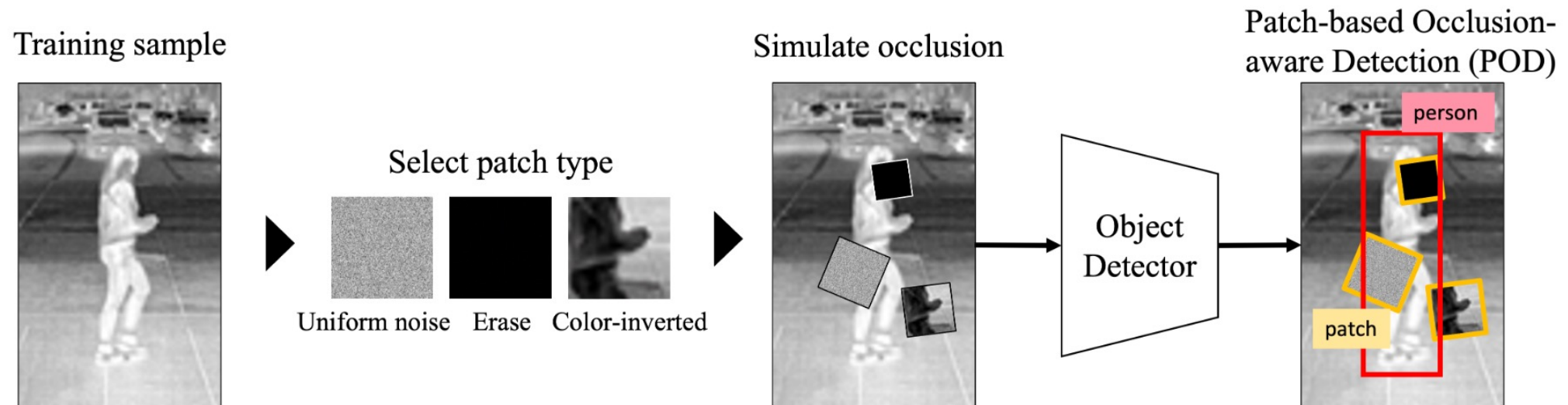
HotCold Block [Wei+ 2023]



Shape-Loc. Attack [Xingxing+ 2023]

Patch-based Occlusion-aware Detection (POD)

- Train with **random occlusion patches** to simulate physical attacks



➤ Robust feature learning against occlusions

Task: single-class **human** detection

Training:

- Model: RGB-pretrained YOLO-v5
- Fine-tuned on FLIR-ADAS (infrared)
 - 50 epochs, less than 1 day with NVIDIA A100 GPU

Test: CVC-14 dataset (infrared)

Attacks:

- **RGB patch** (upper bound baseline)
- **Infrared patches:** HotColdBlock, Shape-loc attack

POD improves robustness with ~no training-time overhead

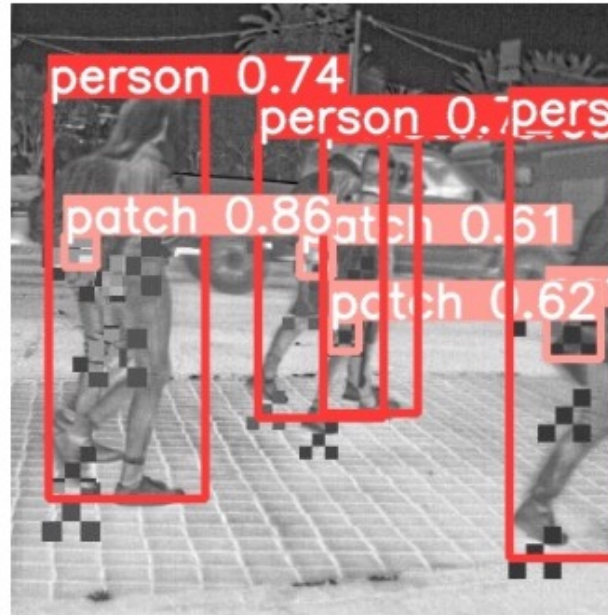
		Average Precision $\uparrow$			Time (rel.)
		Clean	AdvPatch (RGB)	HCB (infrared)	
Ours {	Std. Finetune	92.0	20.7	36.1	$\times 1$
	POD	93.6	79.0	76.4	$\times 1.0$
	Adv-POD (w/ adv. patch)	94.0	81.6	74.3	$\times 5.0$

Adversarial Training version of POD takes long

Qualitative results: POD generalizes well to unseen attacks



AdvPatch [Thys+ 2019]



HCB [Wei+ 2023]



Shape-Loc [Xingxing+ 2023]

➤ **Occlusion-robust feature learning** generalizes to diverse infrared patch attacks

- **First defense study against physical infrared adversarial patch**
- **Propose Patch-based Occlusion-aware Detection (POD):**
 - **Efficient:** 5× faster training than AT
 - **Generalizes** to unseen attacks



Inspired by the insight that **robust-feature learning is key**, we propose a **practical and efficient** method to defend against **physical adversarial attacks**