NEURAL INFORMATION PROCESSING SYSTEMS

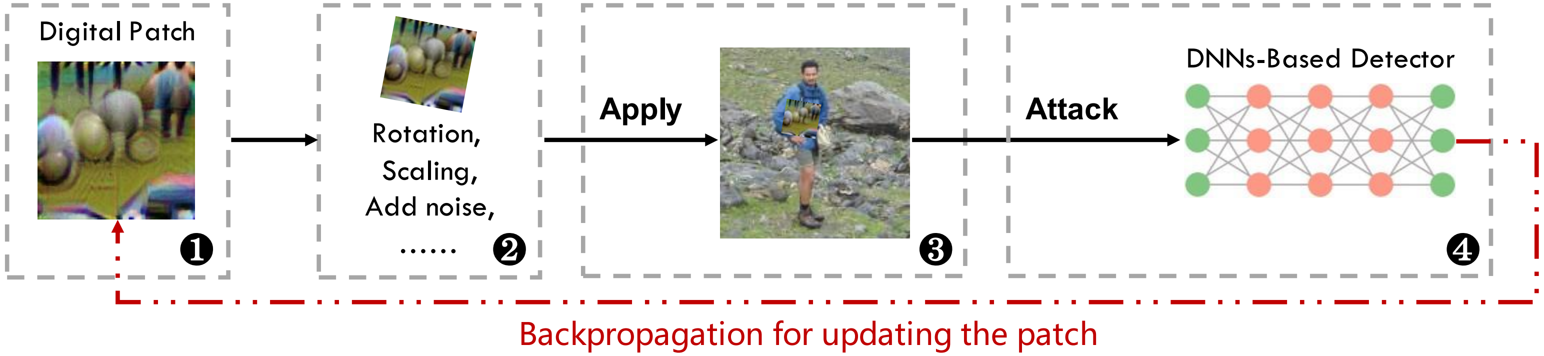
Revisiting Adversarial Patches for Designing **Camera-Agnostic Attacks** against Person Detection

Hui Wei*, Zhixiang Wang*, Kewei Zhang*,
Jiaqi Hou, Yuanwei Liu, Hao Tang, Zheng Wang†
Oct. 10th, 2024

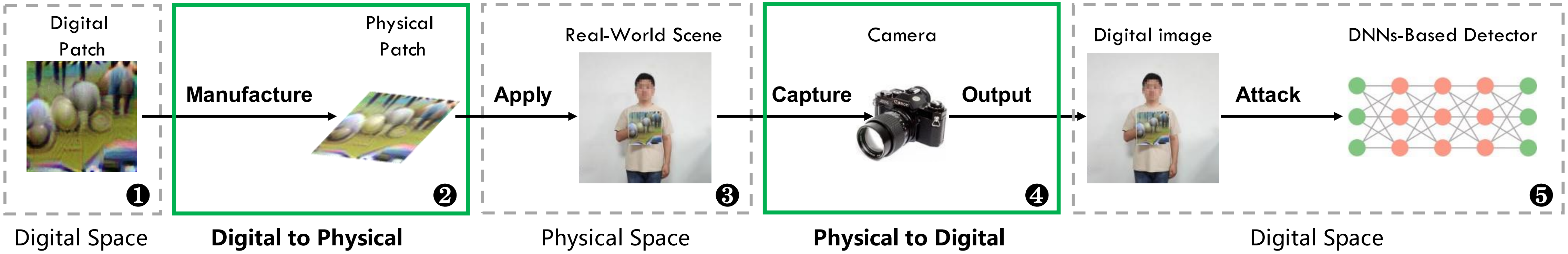
Adversarial Patch

Existing Patch-based Methods

Digital-space

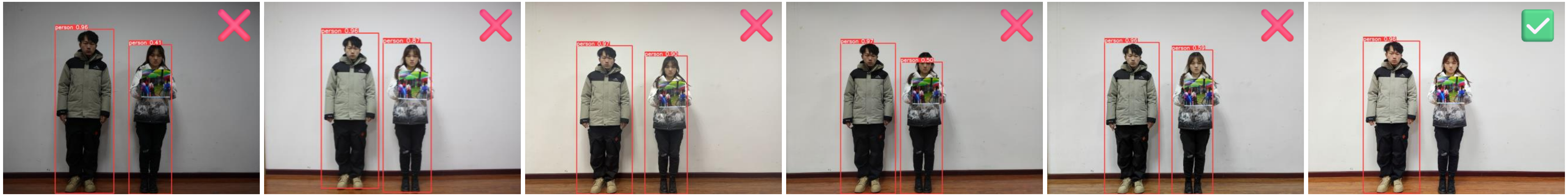


Physical-space



Limitations of Existing Methods

[Thys et al., AdvPatch, 2019]



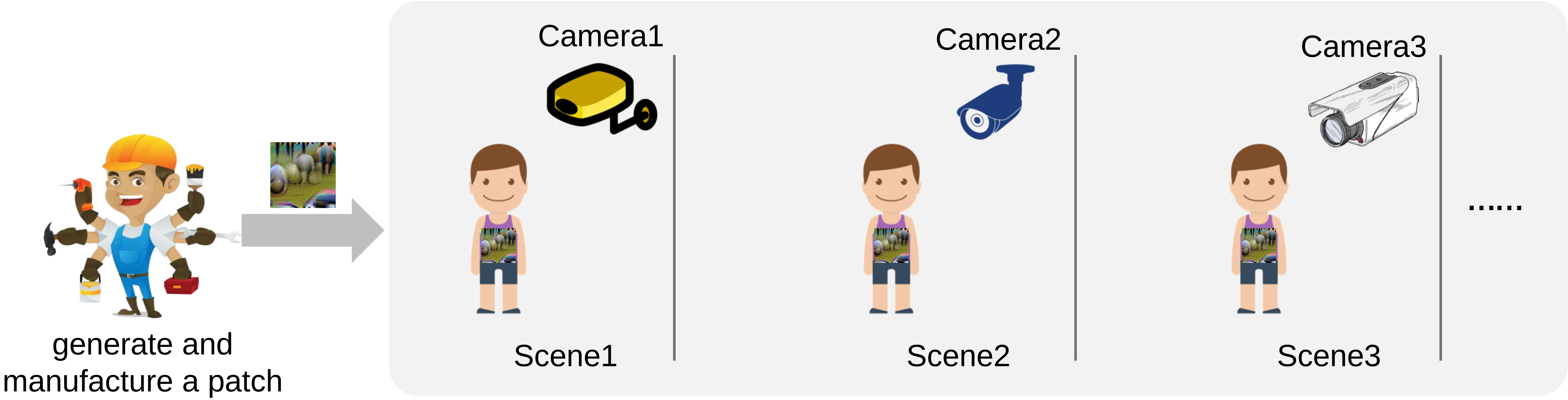
Sony Canon iPhone Redmi Huawei Samsung

[Huang et al., T-SEA, 2023]

Sony Canon iPhone Redmi Huawei Samsung

Deployment in the real world

Diverse camera device



Method

Designing Camera-Agnostic Attacks

Problem Definition

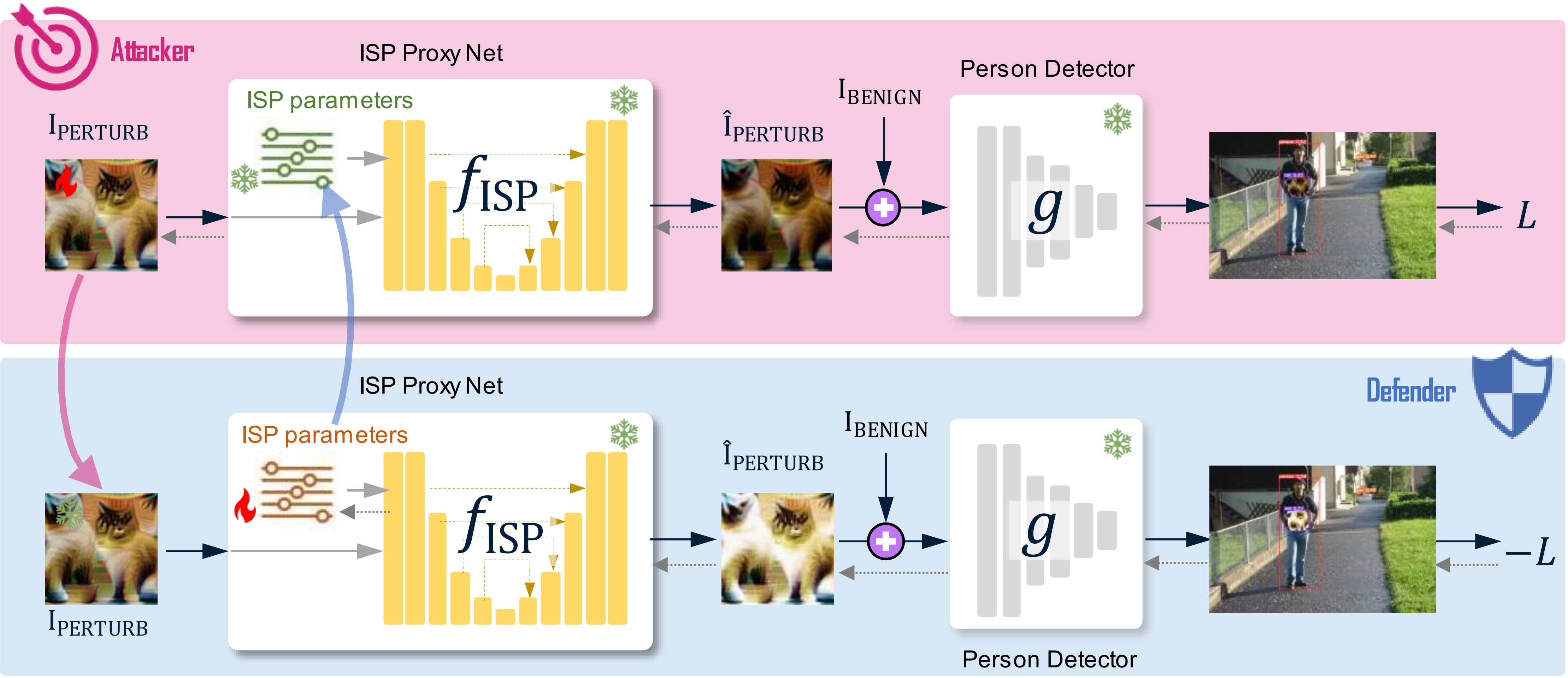
Attack module

$$P^* = \arg\max_i L(f(I^i_{\texttt{SCENE}}, f_{\texttt{ISP}}(P; \boldsymbol{\Theta})), GT),$$

Defense module

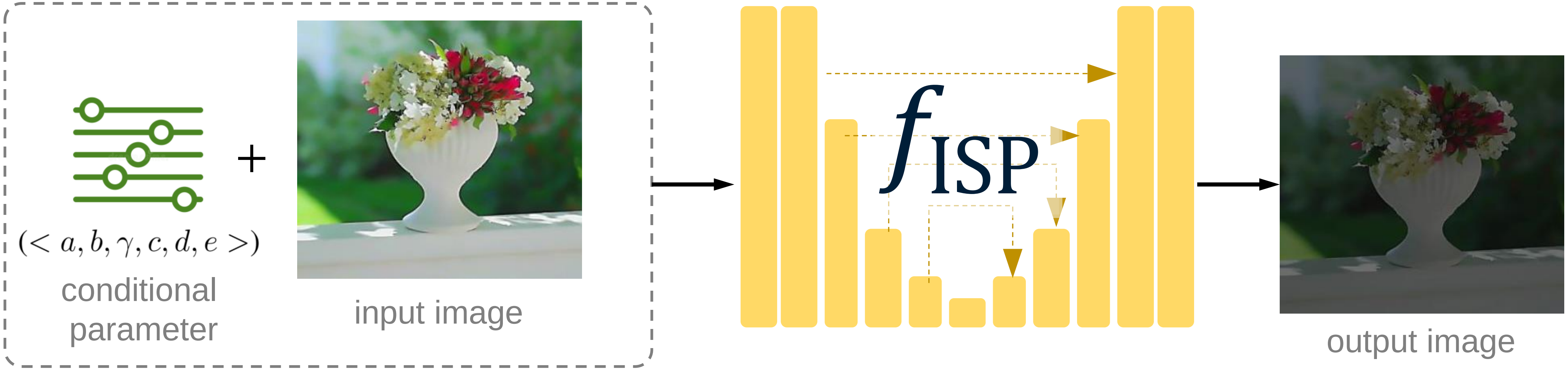
$$\boldsymbol{\Theta}^* = \arg\min_i L(f(I^i_{\texttt{SCENE}}, f_{\texttt{ISP}}(P; \boldsymbol{\Theta})), GT),$$

Overall Framework



Camera ISP Proxy Net

U-Net



Camera ISP Proxy Net

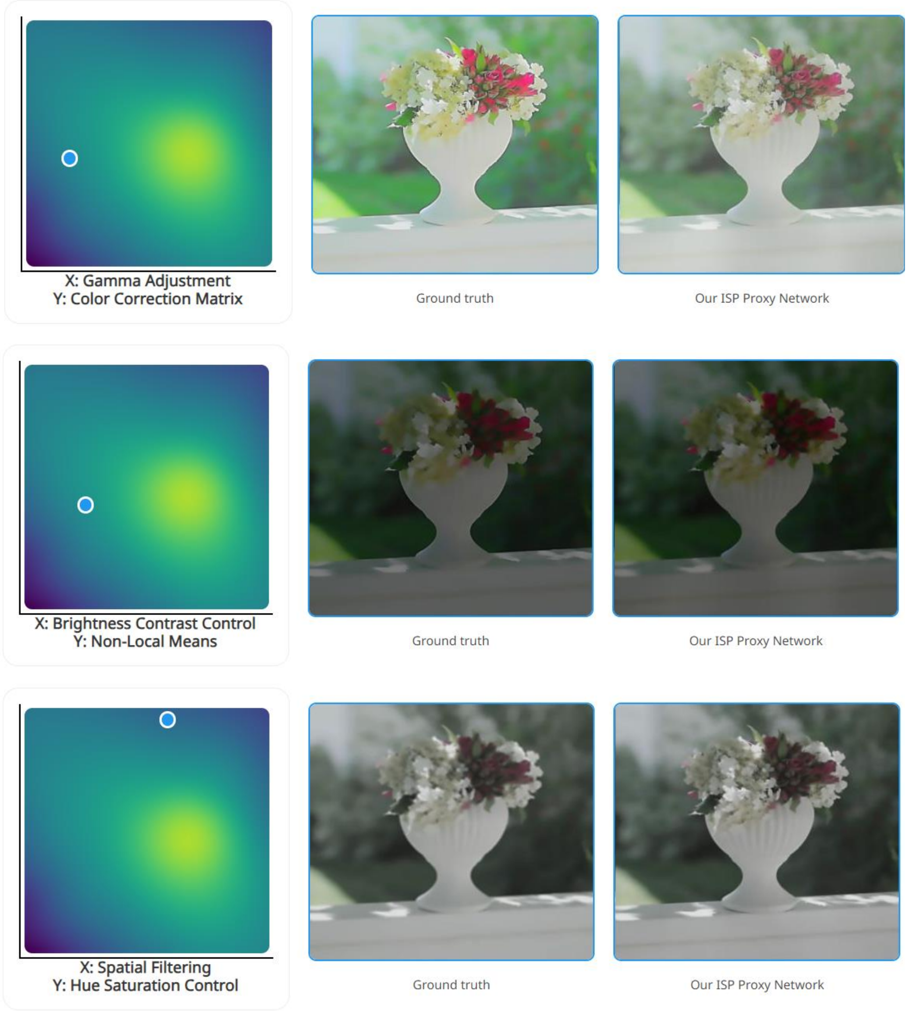
U-Net

(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



Optimization

Algorithm 1 The proposed adversarial optimizaiton (Attacker and Defender)

1: Given source image $x \in X$, targeted person detector f, and the trained camera ISP network f_{ISP};
2: Initialize the adversarial patch P and input hyperparameters $\mathbf{\Theta}$ of f_{ISP};
3: **for** $t = 1, 2, \ldots, T$ **do**
4:     // `Optimize the adversarial patch` P `to maximize attack effectiveness`
5:     **for** $batch = 1, 2, \ldots, M$ **do**
6:         Sample a batch of data x_b from X;
7:         $x_{adv} \leftarrow apply(x_b, P_{\text{ISP}})$, $P_{\text{ISP}} = f_{\text{ISP}}(P, \mathbf{\Theta})$;
8:         x_{adv} are fed into the person detector f to obtain predictions and compute the loss;
9:         Update the adversarial patch P via Eq. 1;
10:     **end for**
11:     // `Optimize input hyperparameters` $\mathbf{\Theta}$ `to minimize the attack effectiveness`
12:     **for** $batch = 1, 2, \ldots, M$ **do**
13:         Sample a batch of data x_b from X;
14:         $x_{adv} \leftarrow apply(x_b, P_{\text{ISP}})$, $P_{\text{ISP}} = f_{\text{ISP}}(P, \mathbf{\Theta})$;
15:         x_{adv} are fed into the person detector f to obtain predictions and compute the loss;
16:         Update the input hyperparameters $\mathbf{\Theta}$ via Eq. 2;
17:     **end for**
18: **end for**

```python
total_iterations = 0
for epoch in range(opt.epochs):
    for i, (imgs, targets, paths, _) in pbar:
        if total_iterations % 40 < 20:
            optimizer_isp.zero_grad()
            ...
            ISP_loss.backward()
            optimizer_isp.step()
        else:
            optimizer_patch.zero_grad()
            ...
            patch_loss.backward()
            optimizer_patch.step()

    total_iterations += 1
...
```

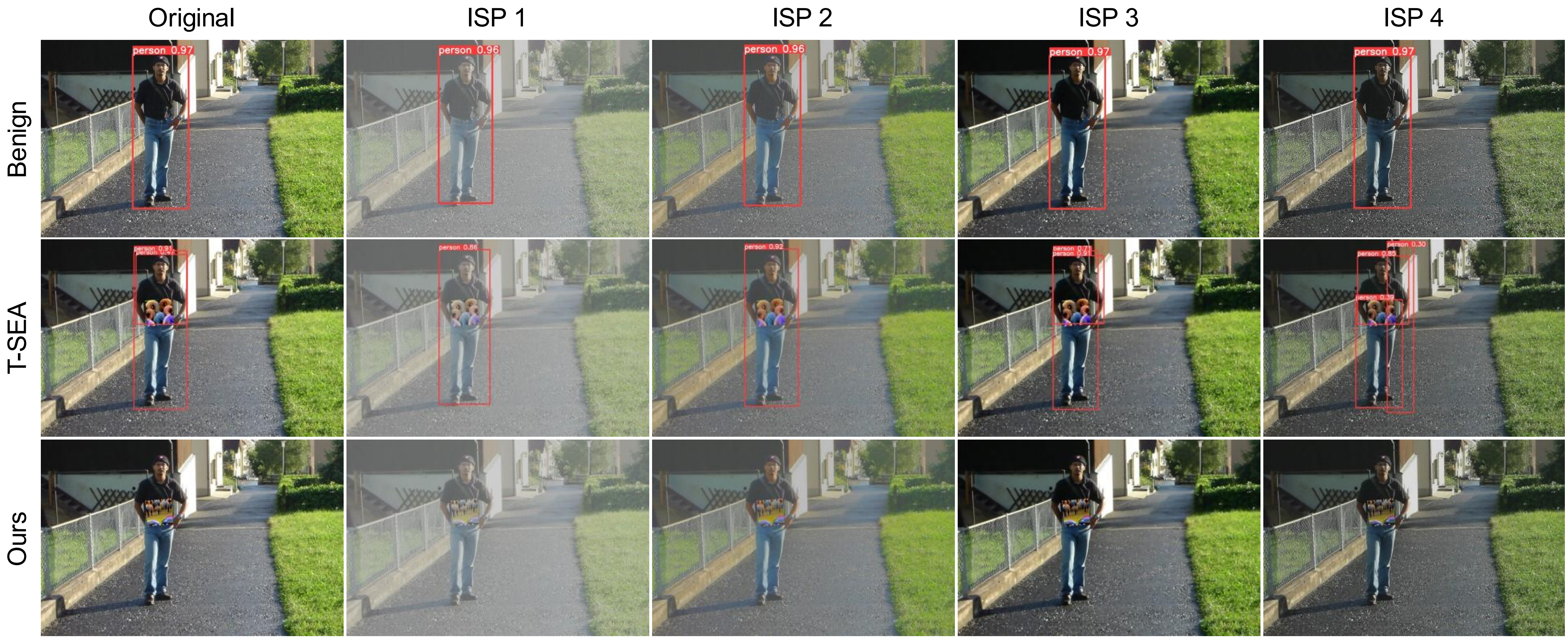
Experiments

Attacking under multiple cameras

Experimental Setup

- **Dataset**: INRIAPerson
- **Compared Methods**: AdvPatch, AdvT-shirt, AdvCloak, NAP, LAP, TC-EGA, and T-SEA.
- **Metrics**: Average Precision (AP%), Attack Success Rate (ASR%)
- **Implementation Details**: two NVIDIA GeForce RTX 3090 GPUs

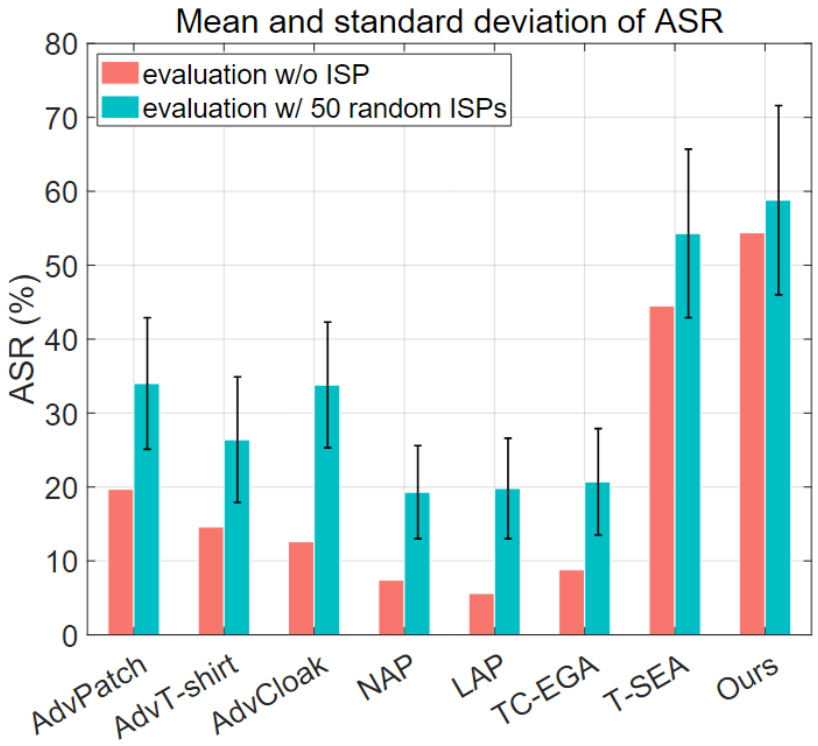
Digital-space Attacks



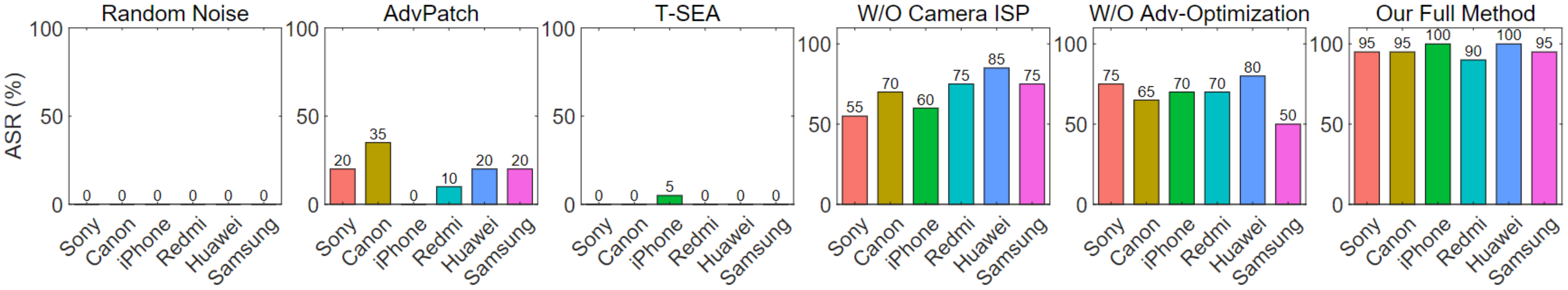
Digital-space Attacks

Quantitative results

Method	Original		ISP 1		ISP 2		ISP 3		ISP 4	
	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑
	Confidence threshold = 0.001, IoU threshold = 0.6									
Random Noise	81.7	7.3	79.3	14.9	80.2	11.0	79.8	10.9	80.1	8.5
AdvPatch [29]	67.7	19.7	60.4	38.3	65.8	30.4	64.5	28.2	68.6	22.9
AdvT-shirt [35]	76.6	14.6	73.0	21.9	76.1	18.8	71.7	21.2	76.5	14.1
AdvCloak [34]	70.5	12.6	65.3	30.4	68.9	23.7	64.3	25.0	68.6	15.8
NAP [11]	81.3	7.4	76.8	16.9	79.1	12.9	76.5	13.8	80.2	8.8
LAP [28]	81.0	5.6	76.3	17.2	78.6	11.6	77.8	12.1	79.4	10.1
TC-EGA [14]	79.9	8.8	71.3	20.3	76.4	14.4	75.6	17.1	76.8	13.3
T-SEA [15]	21.2	44.5	27.0	53.0	22.8	52.7	26.3	44.7	24.7	47.4
CAP (Ours)	37.7	54.4	24.3	64.5	25.7	73.8	37.8	57.4	31.8	68.2



Physical-space Attacks



Ablation Study

Defense Discussion

Digital-space

Defense strategy \ Attack method	Non-attack	CAP*	CAP†	CAP
Non-defense	85.0	52.8	45.5	37.7
JPEG compression [6]	84.7	52.7	45.8	36.8
SAC [20]	85.0	56.2	52.2	46.0
Adversarial training-CAP*	84.1	95.7	91.7	94.3
Adversarial training-CAP†	84.0	92.6	95.4	92.8
Adversarial training-CAP	84.6	94.2	91.6	96.3

Physical-space

Defense strategy \ Attack method	CAP*	CAP†	CAP
Non-defense	70.0	68.3	95.8
JPEG compression [6]	90.8	89.2	93.3
SAC [20]	70.0	68.3	95.8
Adversarial training-CAP*	0.0	4.2	1.7
Adversarial training-CAP†	5.8	0.0	10.8
Adversarial training-CAP	0.0	4.0	0.0

Future Possibilities

- **Better camera simulation.**
  - Design camera simulator containing lens.

- **More camera devices.**
  - Industrial camera: Sony IMX415, Hikvision DS-2CD2043G1-I …

- **Extension to black-box models.**
  - YOLOv8, YOLOv10, DETR, …

NEURAL INFORMATION PROCESSING SYSTEMS

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

Hui Wei*, Zhixiang Wang*, Kewei Zhang*,
Jiaqi Hou, Yuanwei Liu, Hao Tang, Zheng Wang†
Oct. 10th, 2024