RUB

RUHR-UNIVERSITÄT BOCHUM

TARGETED PHYSICAL EVASION ATTACKS IN THE NEAR-INFRARED DOMAIN

Pascal Zimmer, Simon Lachnit, Alexander Jan Zielinski, Ghassan Karame

Network and Distributed System Security (NDSS) Symposium 2026

ML systems in the real-world

Forbes

INNOVATION > TRANSPORTATION

EDITORS' PICK

Waymo Targets 1 Million Robotaxi Rides A Week

Alphabet's fast-growing autonomous ride company expects to expand to over 20 cities, including London and Tokyo, by the end of 2026, up from its current six.

ML systems in the real-world (and its dangers)

Forbes

INNOVATION > TRANSPORTATION

EDITORS' PICK

Waymo Targets 1 Million Robotaxi Rides A Week

Alphabet's fast-growing autonomous ride company expects to expand to over 20 cities, including London and Tokyo, by the end of 2026, up from its current six.

PBS NEWS

U.S. opens Tesla probe after more crashes involving its so-called full self-driving technology

Traffic sign recognition systems

Camera input

Traffic sign recognition systems

Camera input **Detection**

Traffic sign recognition systems

Camera input **Detection** **Classification**

Traffic sign recognition systems

Brake

Camera input **Detection** **Classification** **Action**

Traffic sign recognition systems

Ignore stop

Camera input **Detection** **Classification** **Action**

Previous works

- ✗ Weak untargeted attacker (generic service disruption)
- ✗ Visible perturbations
- ✗ High cost

[1]

[2]

[1] Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification, **CVPR 2018.**
[2] Lovisotto et al., SLAP: Improving physical adversarial examples with Short-Lived adversarial perturbations, **USENIX Security 2021.**

Previous works

✗ Weak untargeted attacker (generic service disruption)
✗ Visible perturbations
✗ High cost

[1]

[2]

Can we mount invisible attacks that satisfy both strong attacker goals and cost constraints?

[1] Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification, **CVPR 2018.**
[2] Lovisotto et al., SLAP: Improving physical adversarial examples with Short-Lived adversarial perturbations, **USENIX Security 2021.**

Previous works

✗ Weak untargeted attacker (generic service disruption)
✗ Visible perturbations
✗ High cost

[1]

[2]

Our work

✓ Strong targeted attacker (specific actions: brake, accelerate,...)
✓ Invisible perturbations
✓ Low cost (US$ 50, 1 min. deployment)

Human vision

Camera vision

Can we mount invisible attacks that satisfy both strong attacker goals and cost constraints?

[1] Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification, **CVPR 2018.**
[2] Lovisotto et al., SLAP: Improving physical adversarial examples with Short-Lived adversarial perturbations, **USENIX Security 2021.**

Threat model

Knowledge:

- task of the ML vision system
- lack of infrared filters in cameras
- reliance on vision-based sensing

Threat model

Knowledge:

- task of the ML vision system
- lack of infrared filters in cameras
- reliance on vision-based sensing

Goal: integrity of model output (targeted, untargeted, hide)

Threat model

Knowledge:

- task of the ML vision system
- lack of infrared filters in cameras
- reliance on vision-based sensing

Goal: integrity of model output (targeted, untargeted, hide)

Capabilities: manipulation of inputs at inference-time, black-box access (score-based)

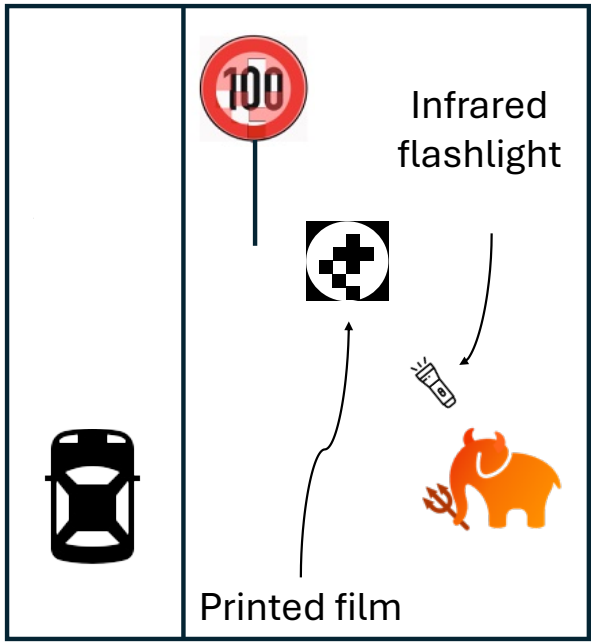
Threat model

Knowledge:

- task of the ML vision system
- lack of infrared filters in cameras
- reliance on vision-based sensing

Goal: integrity of model output (targeted, untargeted, hide)

Capabilities: manipulation of inputs at inference-time, black-box access (score-based)



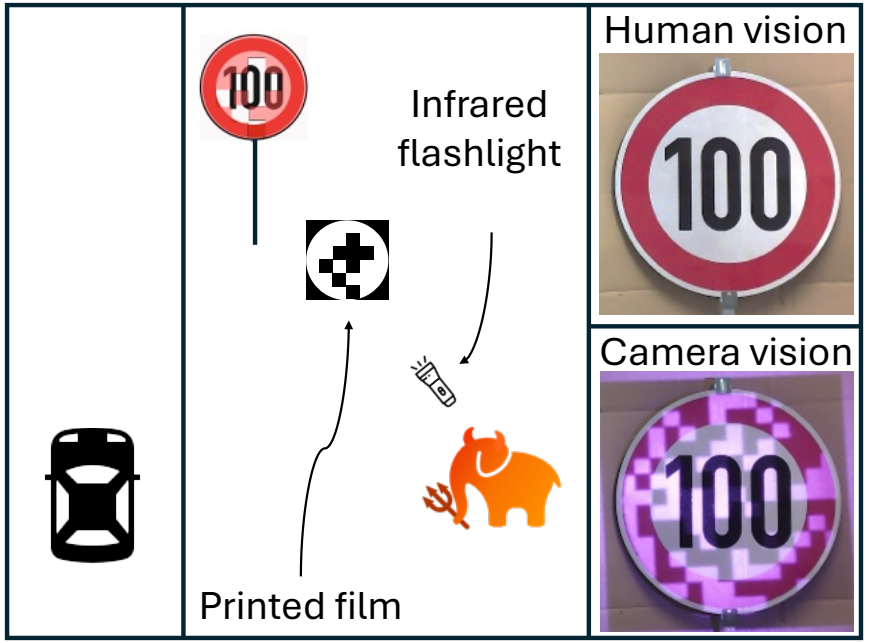
Threat model

Knowledge:

- task of the ML vision system
- lack of infrared filters in cameras
- reliance on vision-based sensing

Goal: integrity of model output (targeted, untargeted, hide)

Capabilities: manipulation of inputs at inference-time, black-box access (score-based)



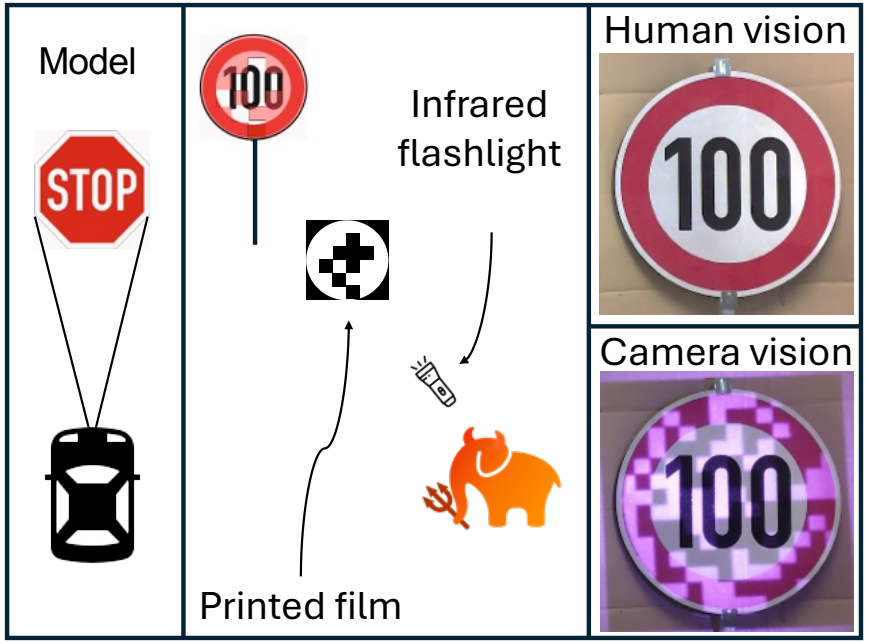
Threat model

Knowledge:

- task of the ML vision system
- lack of infrared filters in cameras
- reliance on vision-based sensing

Goal: integrity of model output (targeted, untargeted, hide)

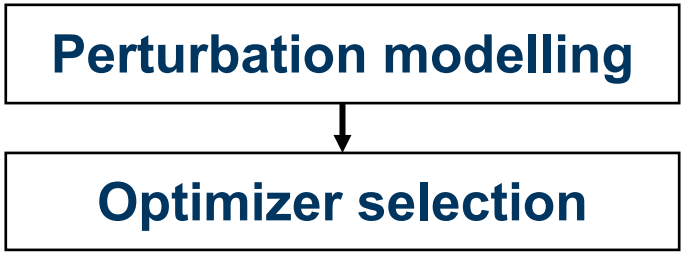
Capabilities: manipulation of inputs at inference-time, black-box access (score-based)



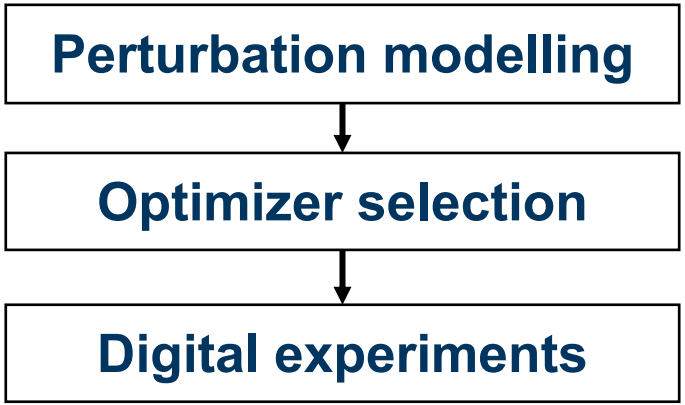
Methodology

Perturbation modelling

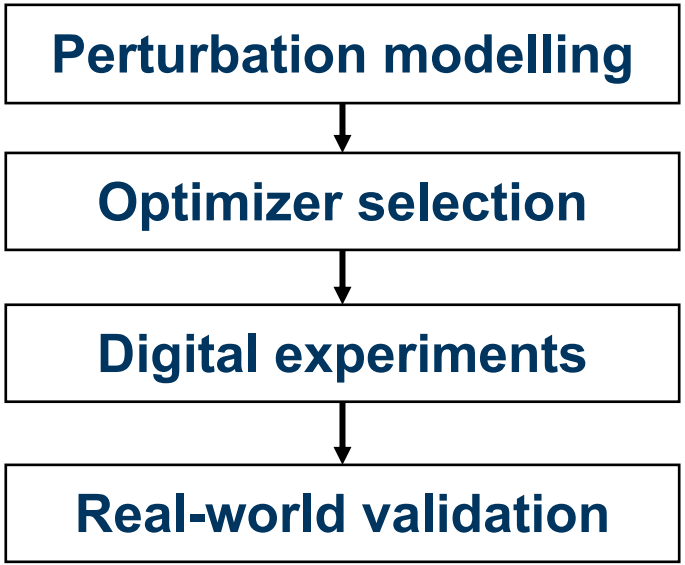
Methodology



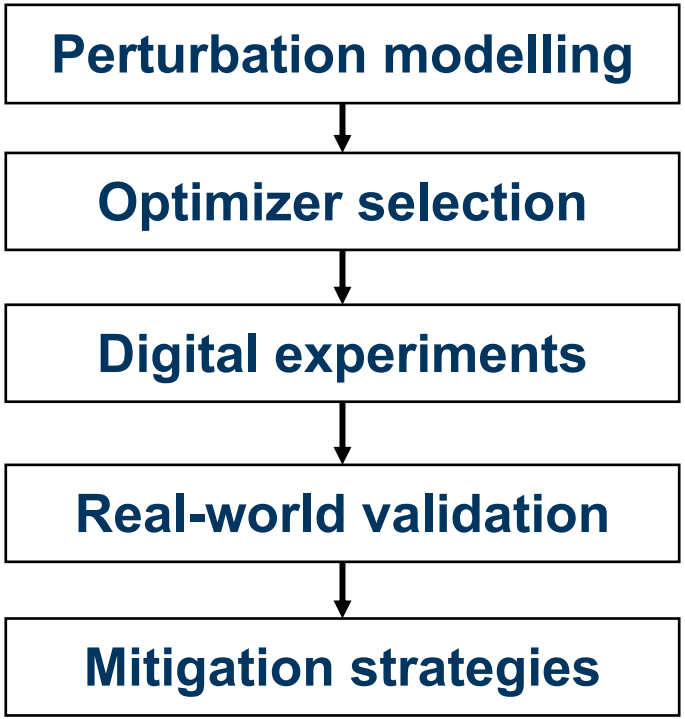
Methodology



Methodology

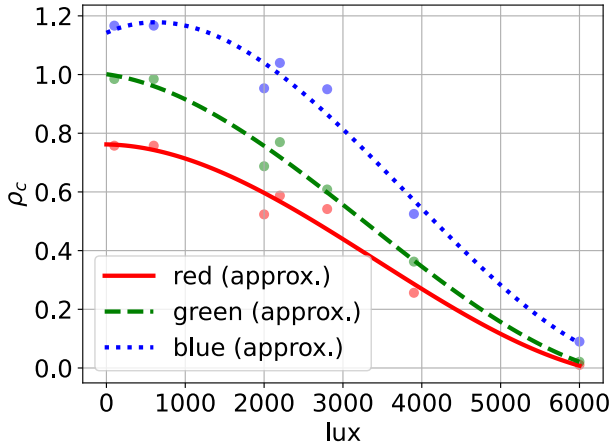


Methodology



Perturbation modelling

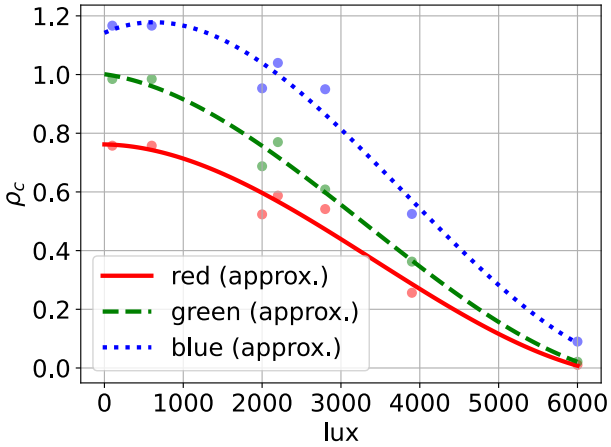
Color: simulation of an infrared light source under ambient light by fitting color-channel specific functions



Perturbation modelling

Color: simulation of an infrared light source under ambient light by fitting color-channel specific functions

Shape: grouping of pixels into square patches to facilitate optimization, manufacturing, and projection

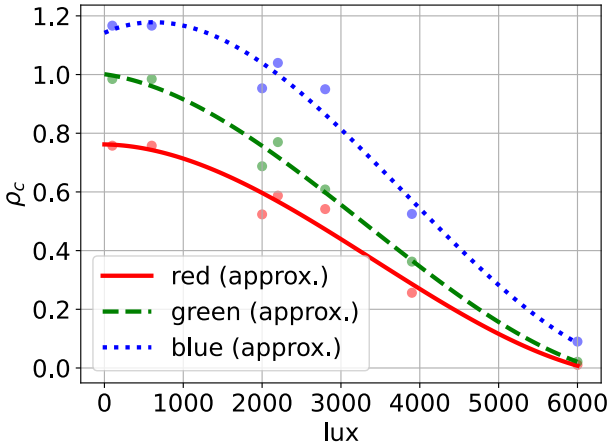


Perturbation modelling

Color: simulation of an infrared light source under ambient light by fitting color-channel specific functions

Shape: grouping of pixels into square patches to facilitate optimization, manufacturing, and projection

$$x_{\mathrm{IR}} = \texttt{ApplyIR}(x, \mathcal{P}) = x \odot \mathcal{P} + \mathbf{IR}(x) \odot (1 - \mathcal{P})$$



a) Perturbation mask $\mathcal{P}$

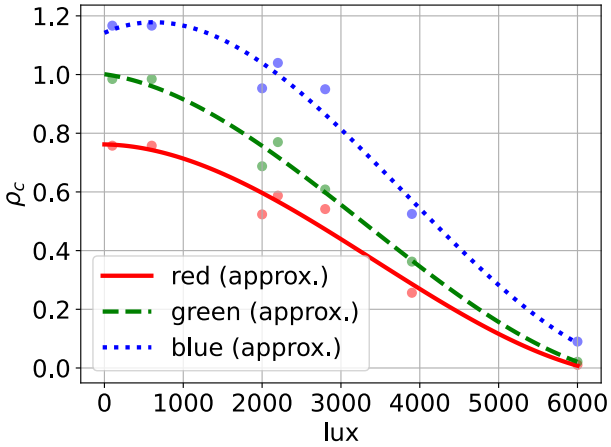
b) Perturbed sign x_{IR}

Perturbation modelling

Color: simulation of an infrared light source under ambient light by fitting color-channel specific functions

Shape: grouping of pixels into square patches to facilitate optimization, manufacturing, and projection

$$x_{\text{IR}} = \texttt{ApplyIR}(x, \mathcal{P}) = x \odot \mathcal{P} + \mathbf{IR}(x) \odot (1 - \mathcal{P})$$



Optimization: black-box optimization of loss with efficient *local random search* to generate perturbation mask:

$$\mathcal{L}_{adv}(x_{\text{IR}}) := \begin{cases} f_s(x_{\text{IR}}) - \max\limits_{i \neq s} f_i(x_{\text{IR}}) & \text{(untargeted attack)} \\ \max\limits_{i \neq t} f_i(x_{\text{IR}}) - f_t(x_{\text{IR}}) & \text{(targeted attack)} \end{cases}$$

a) Perturbation mask $\mathcal{P}$

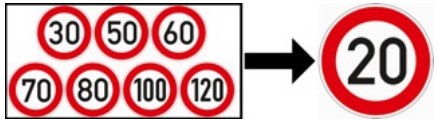
b) Perturbed sign x_{IR}

Digital experiments

Datasets/Models:

- GTSRB (*European traffic signs*)
- LISA (*North American traffic signs*)
- trained on lightweight CNN architectures

Digital experiments



Datasets/Models:

- GTSRB (*European traffic signs*)
- LISA (*North American traffic signs*)
- trained on lightweight CNN architectures

Scenarios (attacker goal):

- Targeted (*brake, accelerate, ignore stop)*
- Untargeted (*service disruption*)

Digital experiments

Datasets/Models:

- GTSRB (*European traffic signs*)
- LISA (*North American traffic signs*)
- trained on lightweight CNN architectures

Scenarios (attacker goal):

- Targeted (*brake, accelerate, ignore stop)*
- Untargeted (*service disruption*)

Metrics:

- Attack success rate (*ASR*)
- Average number of consumed queries (*Q*)

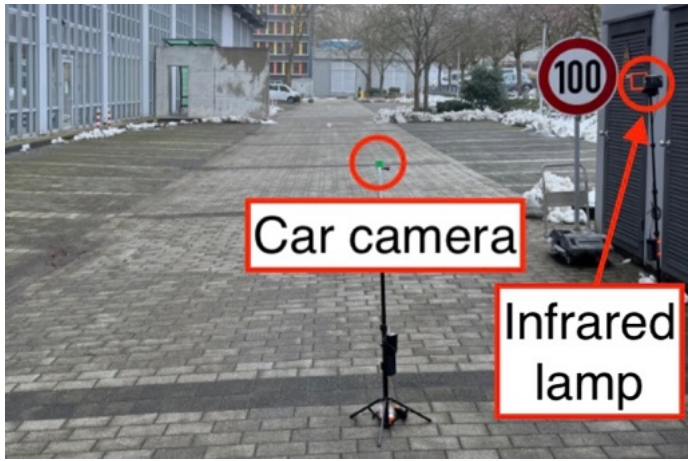
		GTSRB-CNN ASR	GTSRB-CNN Q
Lux	10	96.10	127.90
	1000	95.62	135.36
	2000	92.48	177.03
	3000	81.43	300.15
	4000	41.05	671.71
	5000	4.67	964.87
Patches (k)	16	57.14	559.79
	32	76.76	363.14
	64	88.67	218.16
	96	90.95	192.13
	128	92.57	172.00
	192	92.48	177.03

Real-world validation

Testbed:

- Raspberry Pi 4 with PiCam 3 (IMX 708)
- 808nm 10W IR LED chip
- Total projector cost: US$ 50

Environments: indoor, outdoor (at ~1000 lux)



Real-world validation

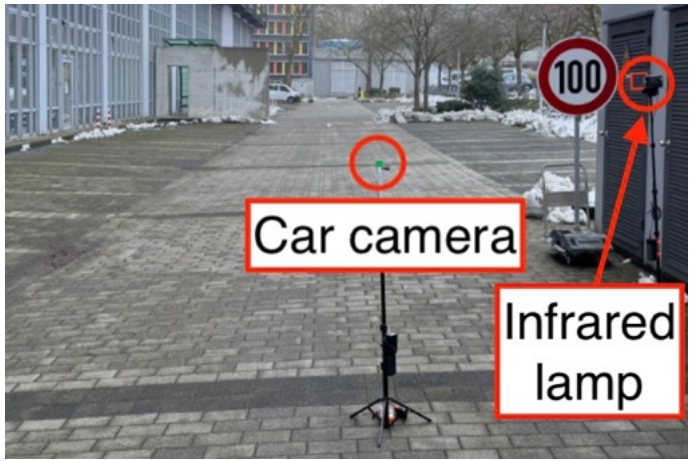
Testbed:

- Raspberry Pi 4 with PiCam 3 (IMX 708)
- 808nm 10W IR LED chip
- Total projector cost: US$ 50

Environments: indoor, outdoor (at ~1000 lux)

Robustness enhancement through expectation-over-transformation (EOT):

- optimization over a set of “real-world” transformations
- rotation, distance, perspective, exposure, alignment, motion blur, ...



Main results

Environment	**Scenario**			
	Targeted Brake	**Targeted** Accelerate	**Targeted** Ignore stop	Untargeted
Static (Indoor)	100.0	100.0	100.0	100.0
Static (Outdoor)	90.0	80.0	100.0	100.0

Main results

Environment	**Scenario**			
	Targeted Brake	**Targeted** Accelerate	**Targeted** Ignore stop	Untargeted
Static (Indoor)	100.0	100.0	100.0	100.0
Static (Outdoor)	90.0	80.0	100.0	100.0
Moving (10 km/h)	99.4	93.7	96.3	84.8
Moving (30 km/h)	98.0	90.0	84.5	84.4

Mitigation strategies

Off-the-shelf defenses (spatial smoothing, adversarial training) only decrease ASR to ~62%

	No defense	Spatial Smooth. (non-local)	Spatial Smooth. (local)	Adv. Training
CA ↑	98.76	95.35	96.56	98.67
ASR ↓	95.16	67.72	61.77	62.89

Mitigation strategies

Off-the-shelf defenses (spatial smoothing, adversarial training) only decrease ASR to ~62%

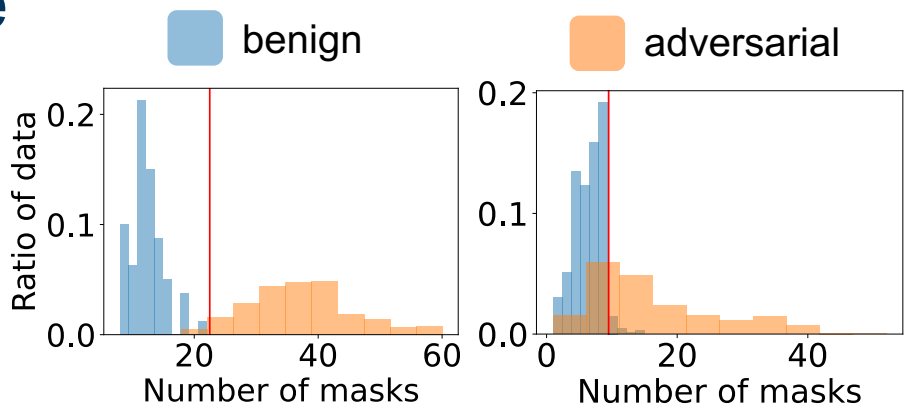
	No defense	Spatial Smooth. (non-local)	Spatial Smooth. (local)	Adv. Training
CA ↑	98.76	95.35	96.56	98.67
ASR ↓	95.16	67.72	61.77	62.89

Observation: perturbation increases the number of shapes on a traffic sign compared to a benign sign

Idea: use an empirical threshold on the number of shapes to determine presence of an adversary

Segmentation-based defense

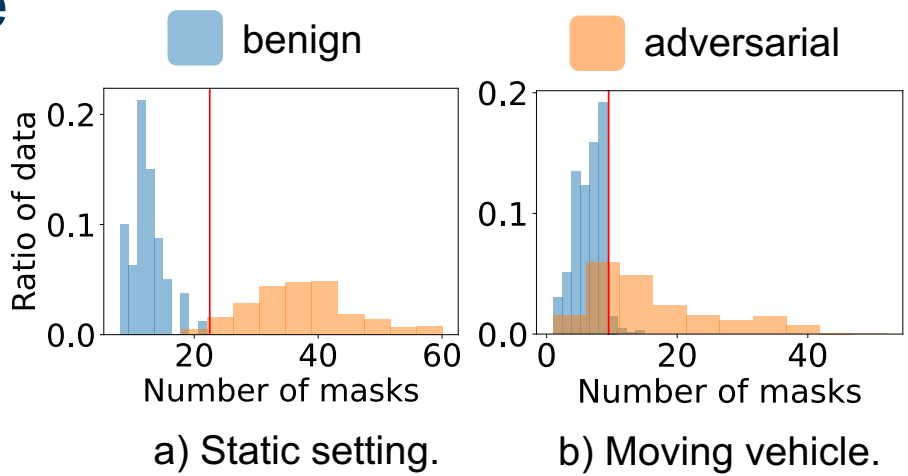
- Segmentation of all images captured during experiments
  - static: ~ 1000 images
  - moving: ~ 1700 images



a) Static setting. b) Moving vehicle.

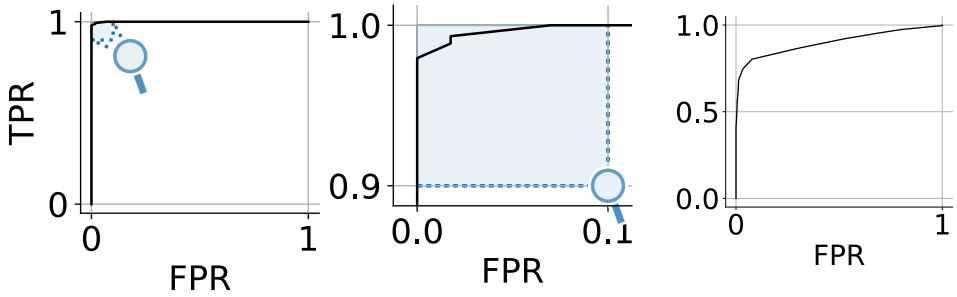
Segmentation-based defense

- Segmentation of all images captured during experiments
  - static: ~1000 images
  - moving: ~1700 images



a) Static setting. b) Moving vehicle.

- Detection performance @ ~0% FPR:
  - static: TPR 98%, ASR of ~2%
  - moving: TPR 75%, ASR of ~25%



a) Static setting. b) Moving vehicle.

- Segmentation-based strategy can effectively thwart the attack!

Takeaways

1. Attacks in the near-infrared domain are an underexplored opportunity for adversaries to mount invisible attacks.

Takeaways

1. Attacks in the near-infrared domain are an underexplored opportunity for adversaries to mount invisible attacks.

2. Invisible attacks alleviate constraints on the optimization and facilitate more powerful attacks.

Takeaways

1. Attacks in the near-infrared domain are an underexplored opportunity for adversaries to mount invisible attacks.

2. Invisible attacks alleviate constraints on the optimization and facilitate more powerful attacks.

3. Off-the-shelf defenses cannot reduce attack success significantly. Our segmentation-based defense can thwart the attack more effectively.

Takeaways

1. Attacks in the near-infrared domain are an underexplored opportunity for adversaries to mount invisible attacks.

2. Invisible attacks alleviate constraints on the optimization and facilitate more powerful attacks.

3. Off-the-shelf defenses cannot reduce attack success significantly. Our segmentation-based defense can thwart the attack more effectively.

Thank you for your attention!

Questions?

Paper →

Code & Dataset

Contact