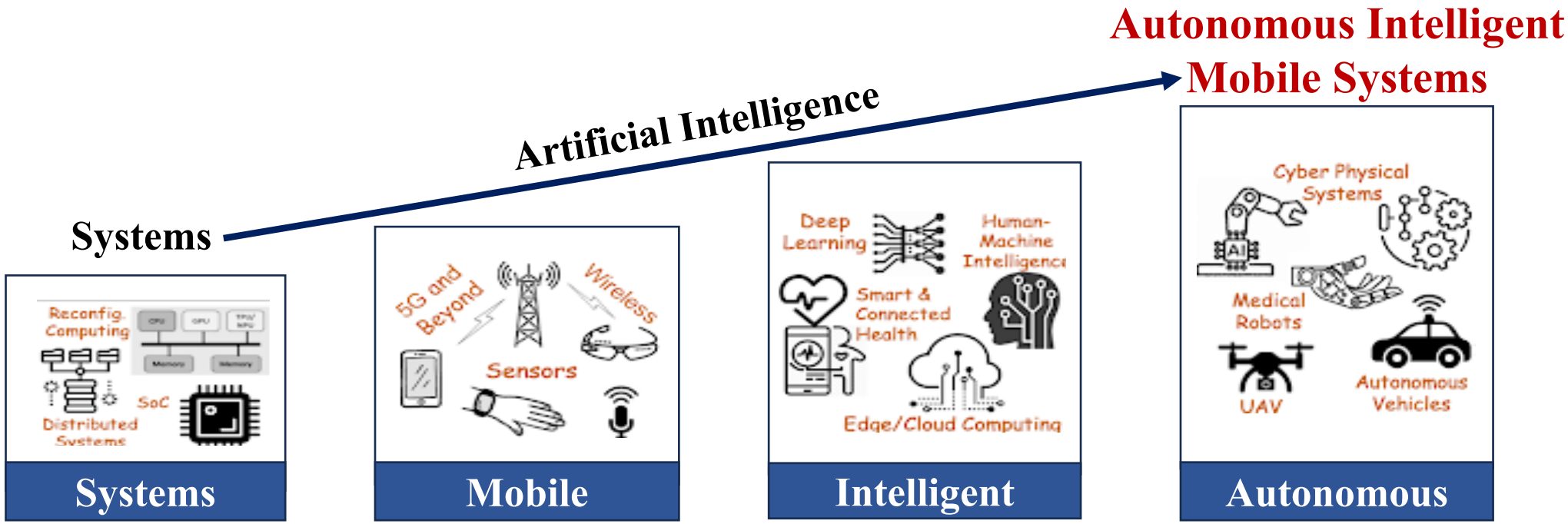
# L-HAWK: A Controllable Physical Adversarial Patch Against a Long-Distance Target

**Taifeng Liu**[1], Yang Liu[1], Zhuo Ma[1], Tong Yang[2], Xinjing Liu[1], Teng Li[1], Jianfeng Ma[1]

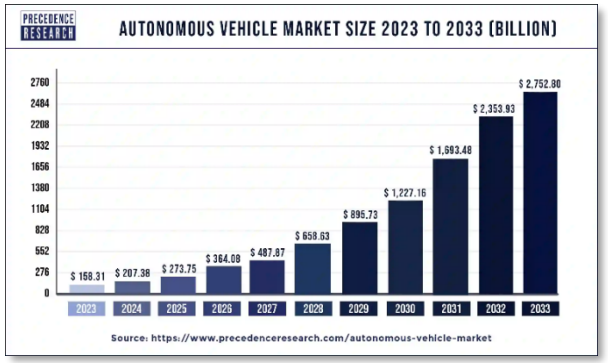
[1]AI-Ants Lab, Xidian University

[2]Peking University

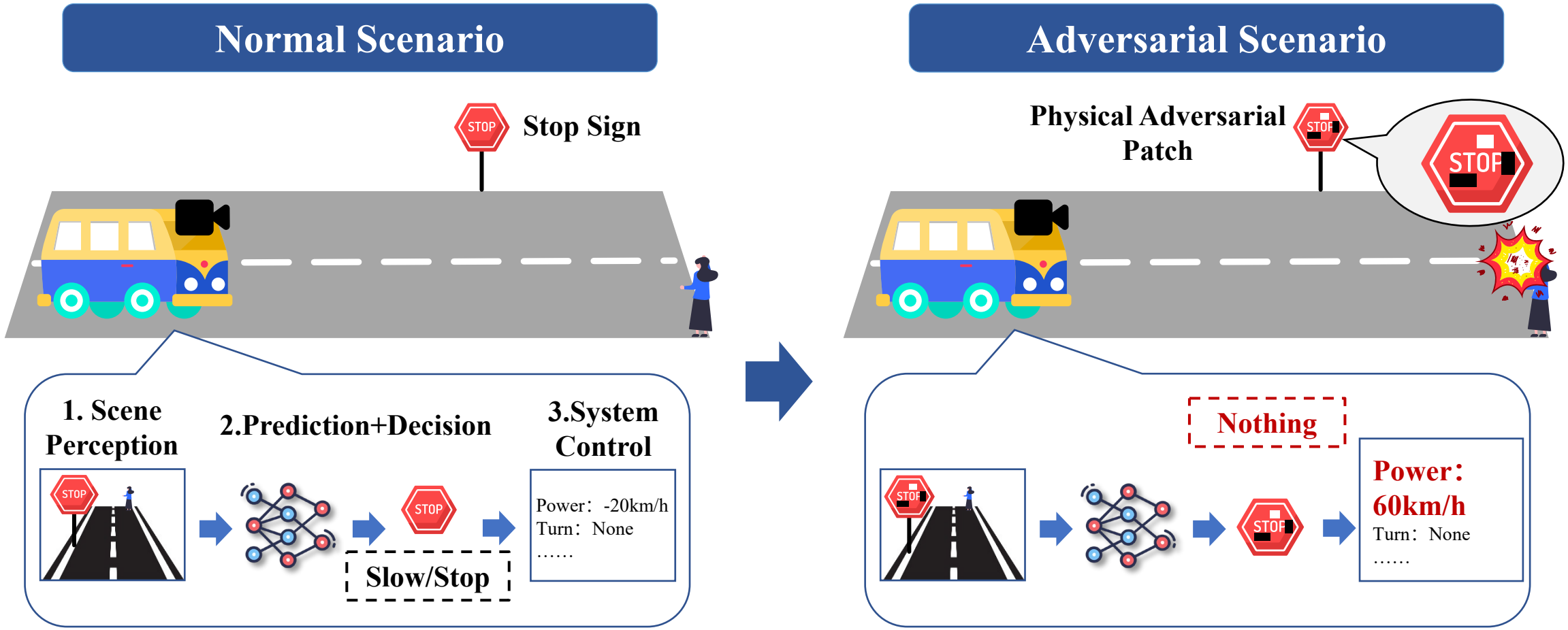
# Autonomous Intelligent Mobile Systems





# Visual Recognition in Autonomous Driving

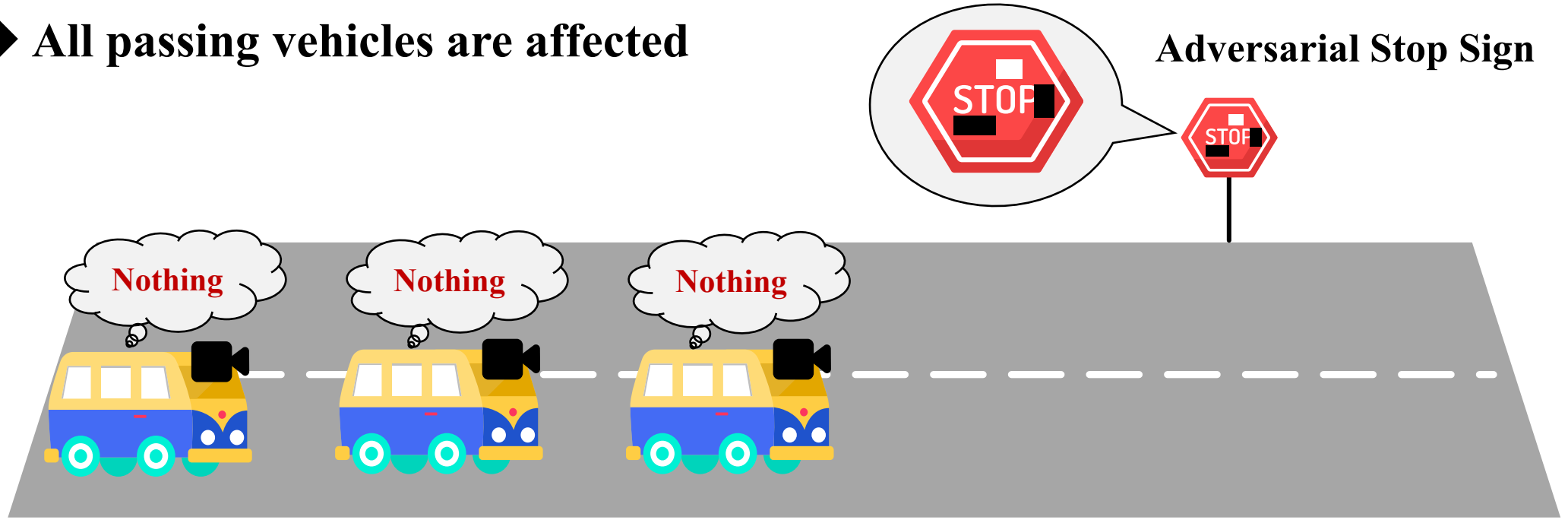



◆ AD is vulnerable and susceptible to physical adversarial patches.

# Research on Adversarial Patch Attacks

## Traditional Adversarial Attack is Not Controllable

◆ All passing vehicles are affected

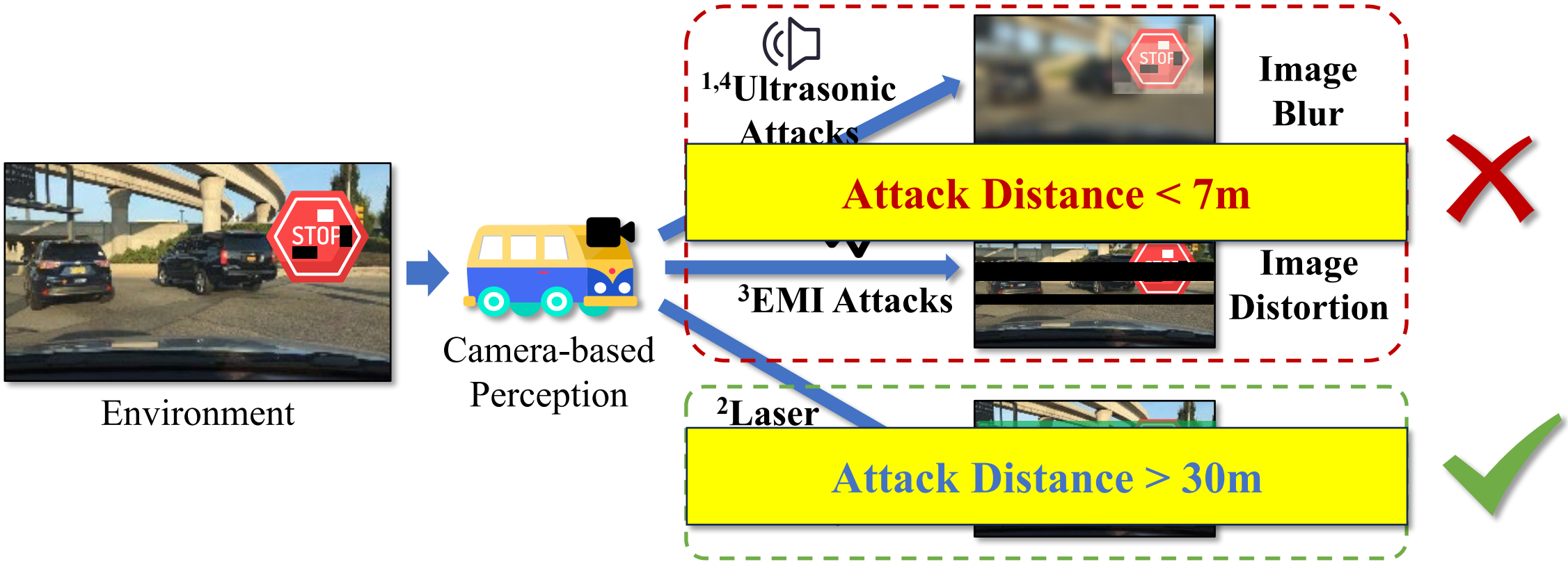



**The risk of adversarial patches being discovered increases.**

*Can we **control** an adversarial patch to affect **only one** specific target, rather than all passing vehicles?*

# Research on Adversarial Patch Attacks

**Discovery：** The camera sensor is susceptible to **special signals** such as ultrasonic waves, electromagnetic pulses, lasers, etc.

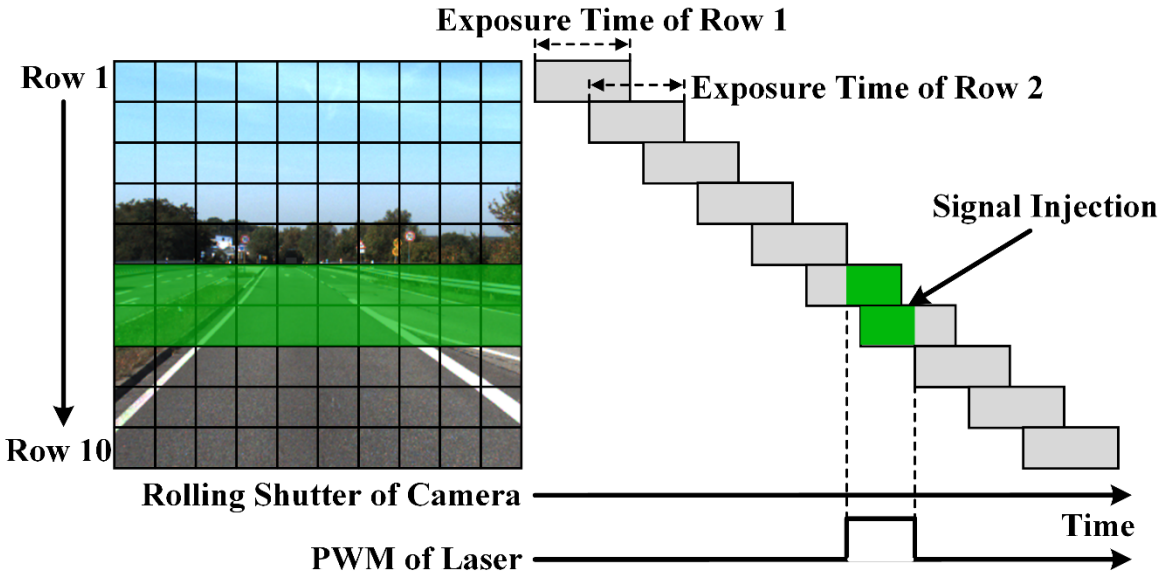




[1] Poltergeist: Acoustic adversarial machine learning against cameras and computer vision (S&P 2021)
[2] Rolling colors: Adversarial laser exploits against traffic light recognition (USENIX Security 2022)
[3] GlitchHiker: Uncovering Vulnerabilities of Image Signal Transmission with IEMI (USENIX Security 2023)
[4] TPatch: A Triggered Physical Adversarial Patch (USENIX Security 2023)

# L-HAWK: Controllable Adversarial Attacks

◆ Why **laser signal** can be used to **control** the adversarial patch?





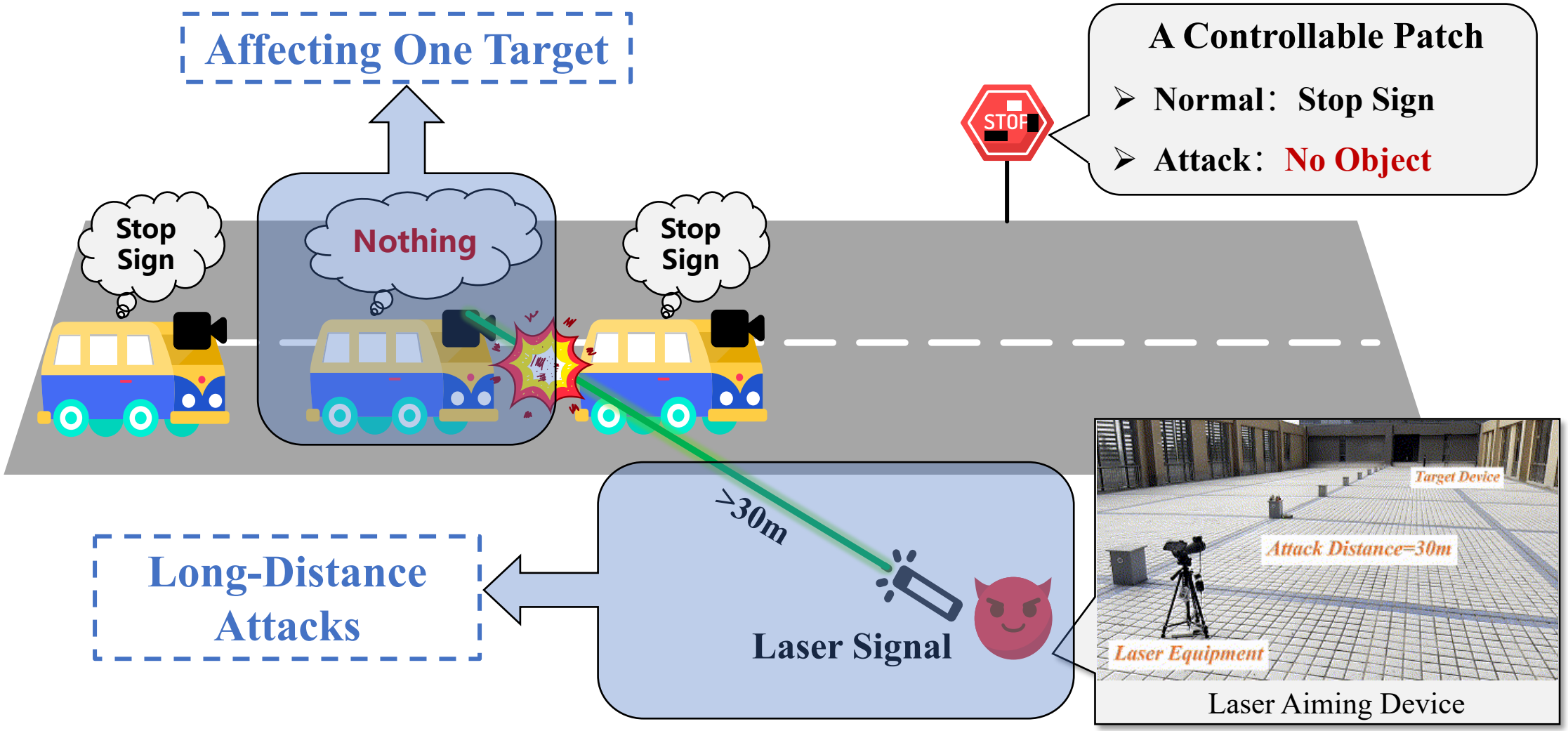
**Without** laser attacks


**With** laser attacks

➢ The **color stripe** caused by laser signal attack.

# L-HAWK: Controllable Adversarial Attacks

◆ We propose a controllable adversarial attack based on long-distance laser triggering



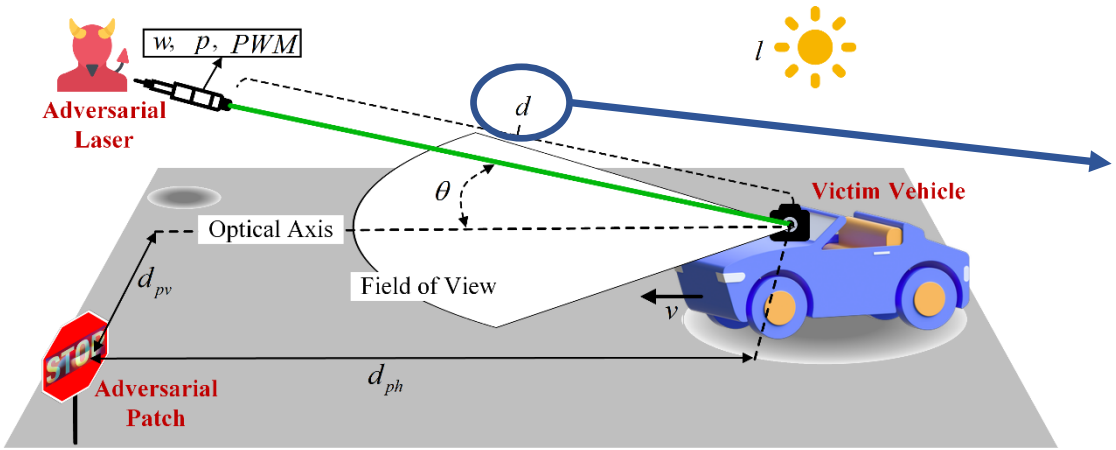
# Challenges

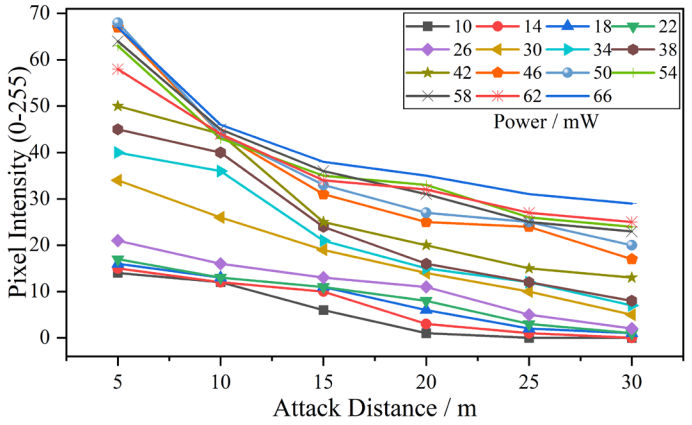
- **Challenge 1**: *How to make the adversarial patch can be **controlled** by the laser signal?*

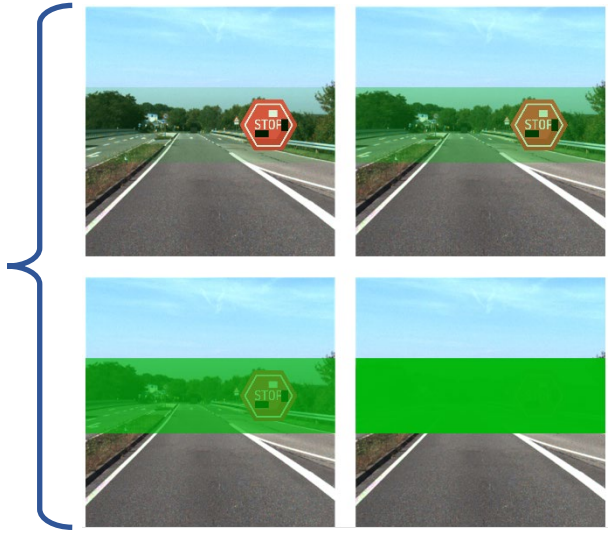
- **Challenge 2**: *How to increase the **attack robustness** of in the physical world?*

# Challenges

- **Challenge 1**: *How to ensure that the adversarial patch can be **controlled** by the laser signal?*



- The intensity of color stripe is affected by **multiple factors**, e.g., laser power, attack distance, ambient light intensity.

# Challenges

- **Challenge 1**: *How to ensure that the adversarial patch can be **controlled** by the laser signal?*
  - The previous patch optimization method[1]: (**only optimize the patch**)

$$arg\min_{\delta} \mathbb{E}_{x,t}[\mathcal{L}(x, \delta, t)]$$

**Some color stripes can not control the patch due to *strong or weak intensity*.**

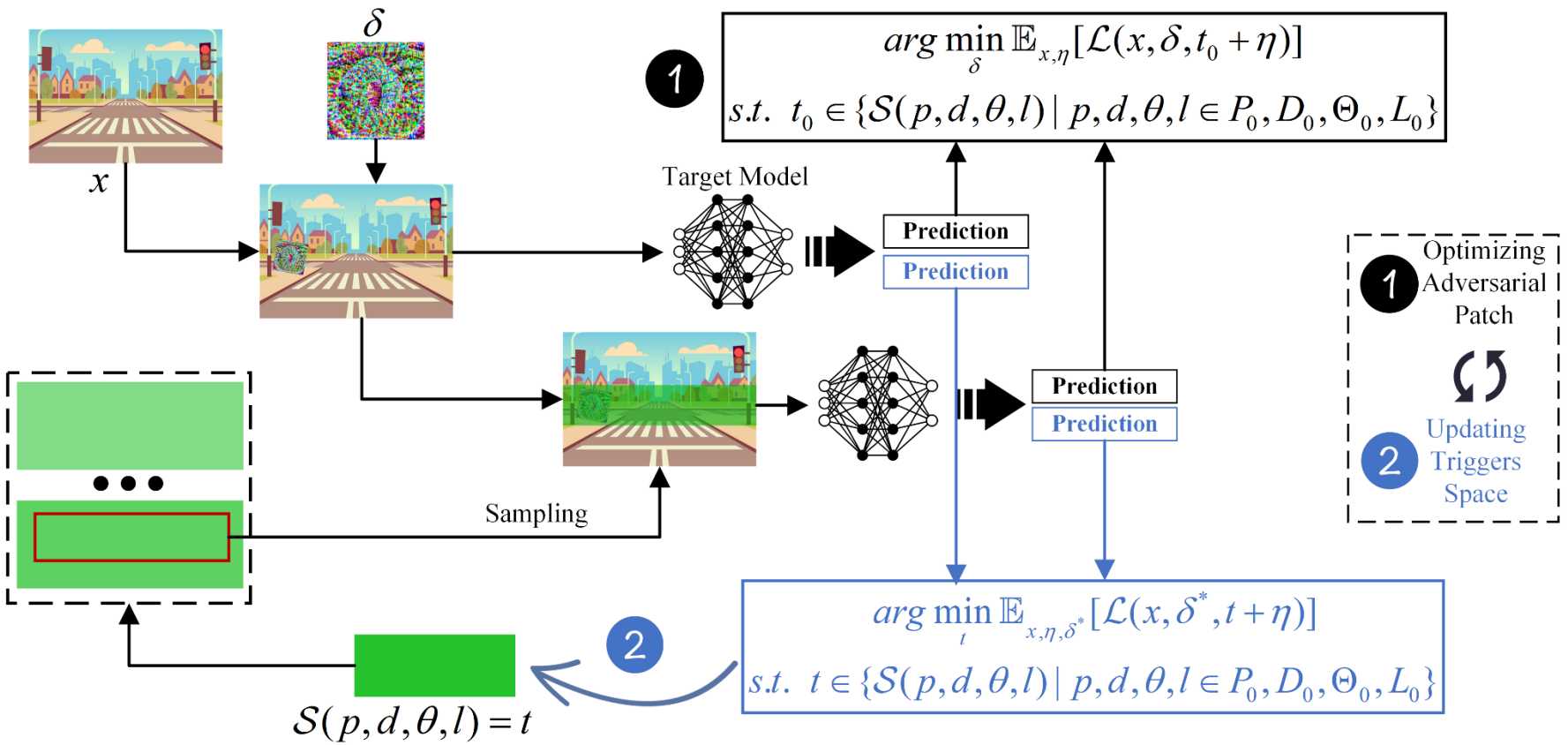
  - Thus, we optimize not only the patch but also the color stripe:

$$arg\min_{\delta,t} \mathbb{E}_{x}[\mathcal{L}_{*}(x, \delta, t)],$$
$$s.t.\ t \in \{\mathcal{S}(p, d, \theta, l) \mid p, d, \theta, l \in P, D, \Theta, L\}.$$

[1] TPatch: A Triggered Physical Adversarial Patch (USENIX Security 2023)

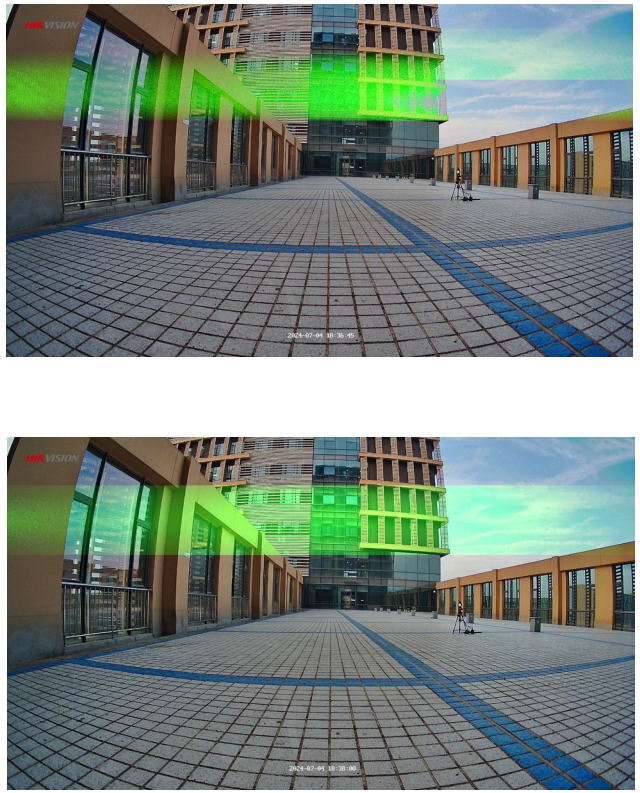
# Asynchronous Learning For Optimizing

- We propose an asynchronous learning method that facilitates multi-objective adversarial patch and color stripe optimization.



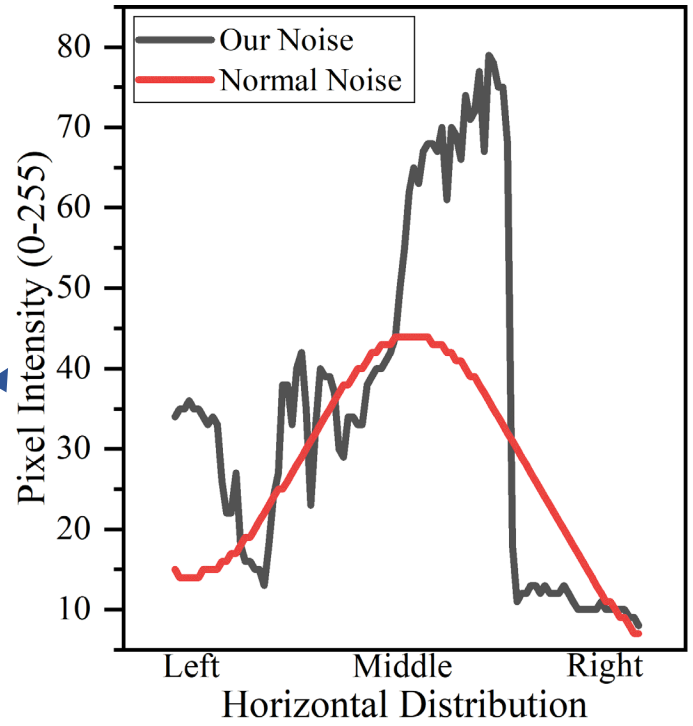
# Challenge

- **Challenge 2**: *How to increase the **attack robustness** of laser signal in the physical world?*



The real color stripe

The simulated color stripe[1]

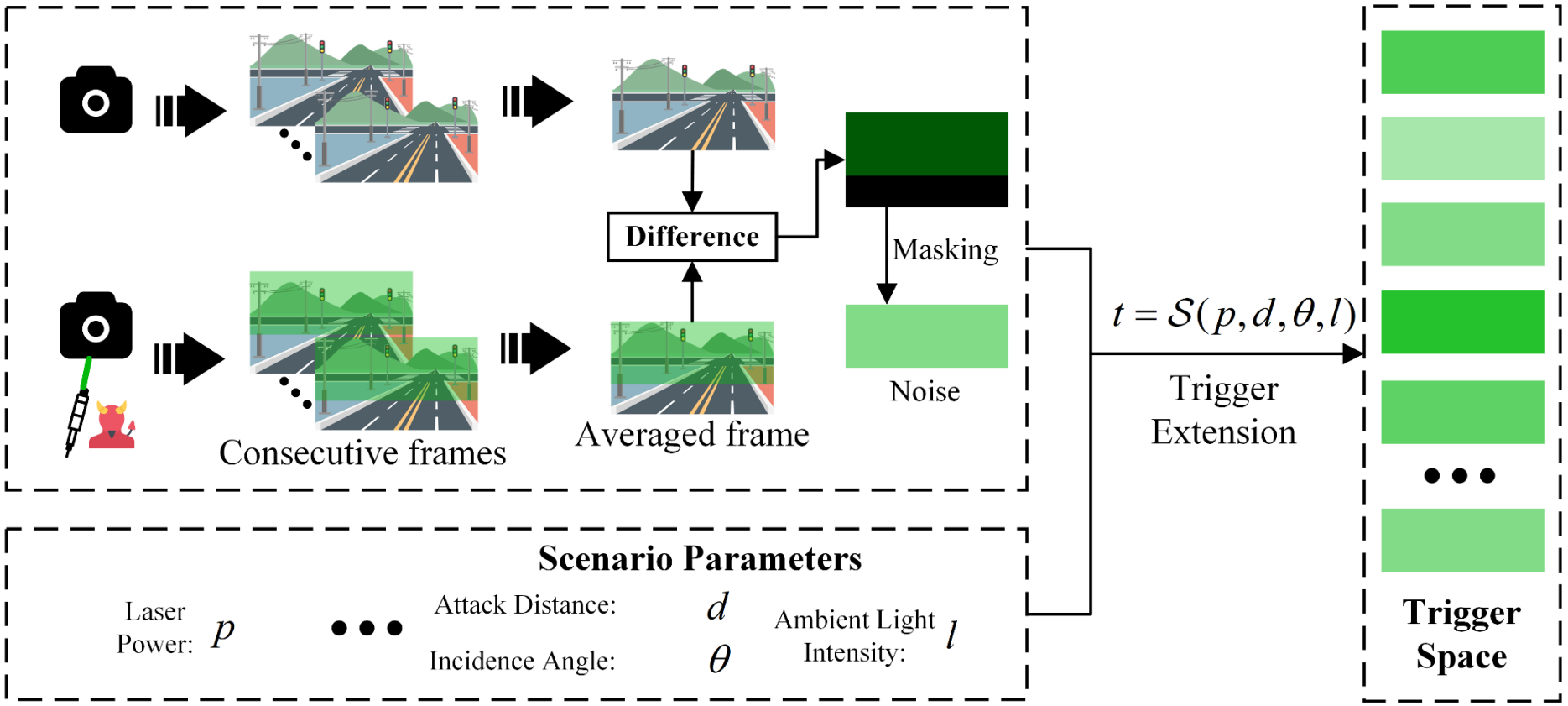



***Fail to control adversarial patches***

[1] Rolling colors: Adversarial laser exploits against traffic light recognition (USENIX Security 2022)

# Trigger Modeling

- We approximate real-world noise by evaluating differences between continuous camera frames.



# Evaluation

## ◆ Digital Evaluation

- ➢ 3 object detectors and 8 image classifiers
- ➢ **94.4%** average attack success rate of four attacks

## ◆ Physical World Evaluation in Stationary Setups

- ➢ **92.8%** average attack success rate against 4 cameras

## ◆ Physical World Evaluation in Moving Setups

- ➢ **56%** average attack success rate across all attacks at **50 km/h**
- ➢ **91.9%** average attack success rate at **50m**

# Digital Evaluation

- Compared to **TPatch (Usenix'23)**, our average ASRs has improved by more than **5 times**.

| Method | HA | CA | TA-D | TA-C |
|---|---|---|---|---|
| Patch optimization in TPatch [10] | 10.7% | 0.5% | 14.4% | 35.1% |
| Our joint optimization | 36.9% | 25.1% | 42.8% | 62.3% |
| Our joint optimization & trigger modeling | 97.6% | 95.5% | 99.6% | 85% |

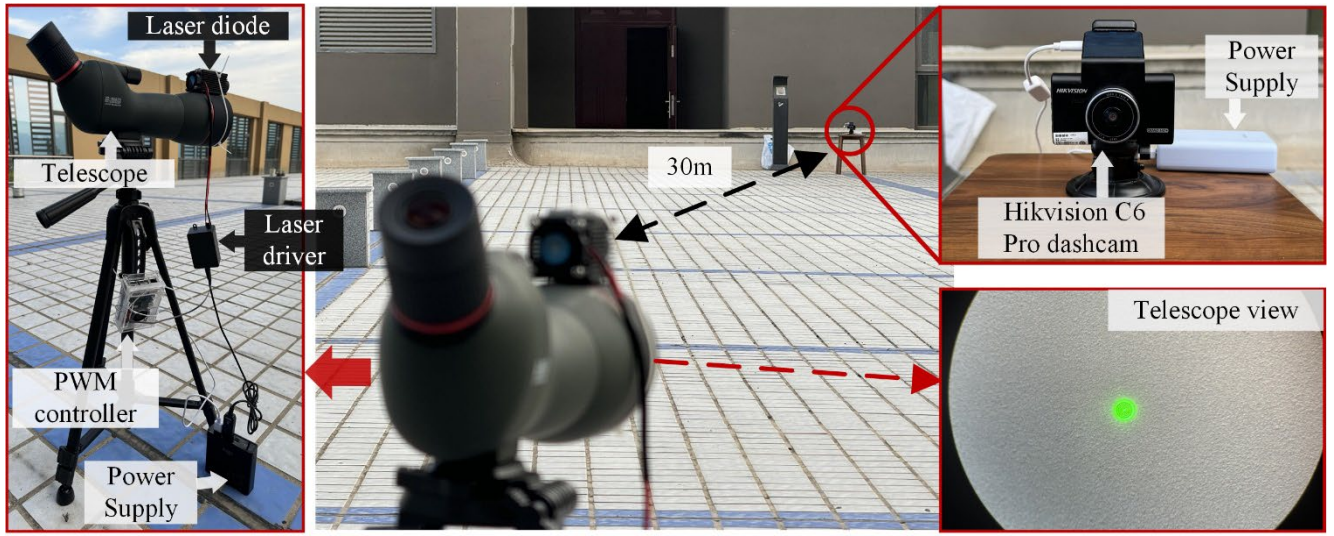
- **Transfer Attacks**: we achieve an average ASR above **44%**.

| Atack Type | HA | | | CA | | | TA-D | | |
|---|---|---|---|---|---|---|---|---|---|
| Black Detector / White Detector | Faster R-CNN | YOLO v3 | YOLO v5 | Faster R-CNN | YOLO v3 | YOLO v5 | Faster R-CNN | YOLO v3 | YOLO v5 |
| Faster R-CNN | 96.6% | 59.3% | 27.4% | 97.0% | 93.4% | 68.8% | 100.0% | 94.0% | 45.6% |
| YOLO v3 | 56.3% | 98.6% | 97.6% | 6.1% | 99.3% | 72.4% | 0.12% | 100.0% | 21.9% |
| YOLO v5 | 89.0% | 63.6% | 99.9% | 10.4% | 99.6% | 98.4% | 48.0% | 99.4% | 99.6% |

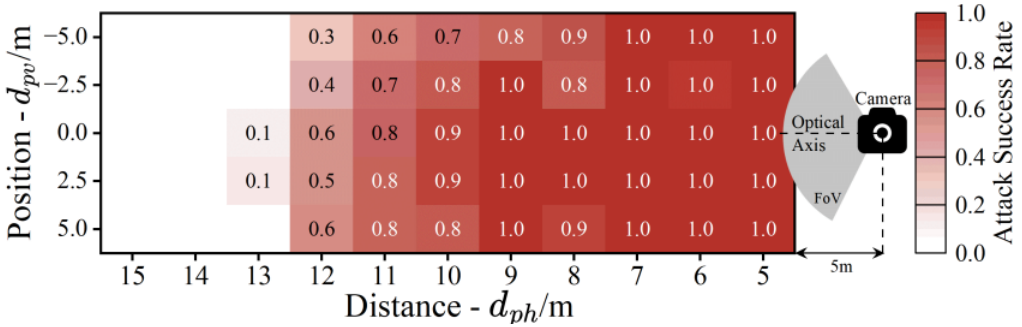
| Black Classifier / White Classifier | VGG-13 | VGG-16 | VGG-19 | ResNet-50 | ResNet-101 | ResNet-152 | Inception-v3 | MobileNet-v2 |
|---|---|---|---|---|---|---|---|---|
| VGG-ens | 94.9% | 97.1% | 99.9% | 58.0% | 44.0% | 62.9% | 35.5% | 40.3% |
| ResNet-ens | 14.9% | 50.1% | 57.4% | 95.8% | 96.6% | 93.6% | 22.4% | 56.8% |

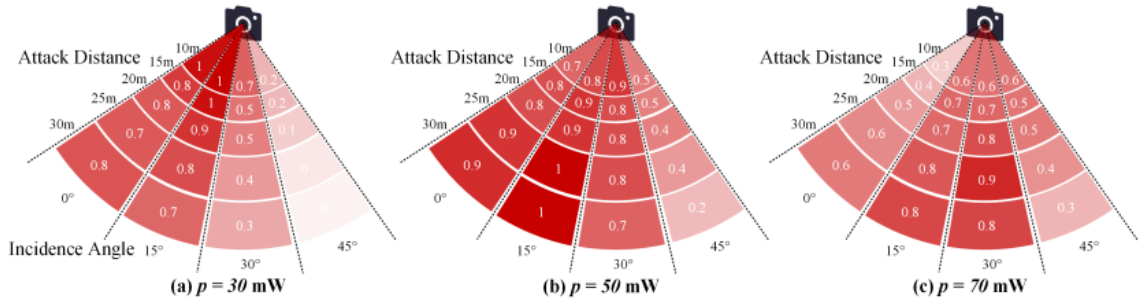
# Physical World Evaluation in Stationary Setups

- **Laser Aiming Equipment**: attack distance **>50m**.



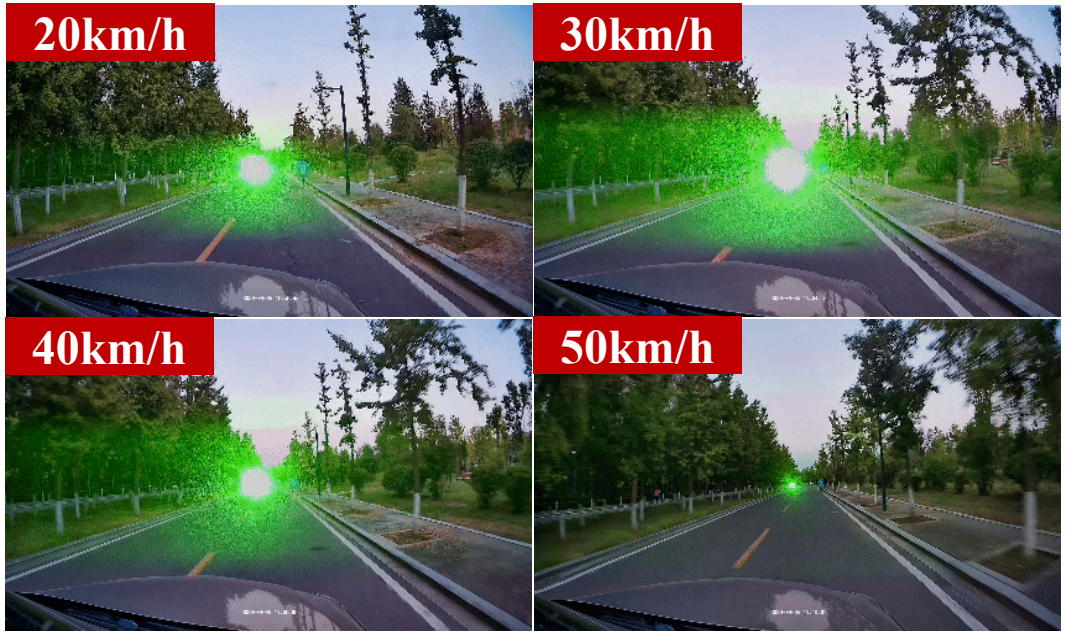

- Evaluation in complex physical environments.



# Physical World Evaluation in Moving Setups

- **Various Speeds Evaluation**: the attack is still effective at all.
- **End-to-End Evaluation: >80%** average ASRs.




# Countermeasures

## ◆ Algorithm Level

- ➢ Adversarial Training and Input Transformation-Based Method
- ➢ Adversarial Patch Detection

## ◆ Sensor Level

- ➢ Multi-Sensor Fusion
- ➢ Random Rolling Shutter Mechanism

# Conclusion

- A controllable physical adversarial patch attack based on laser signal attacks.
- A comprehensive study of laser attacks.
- An asynchronous learning method for optimizing laser parameters and physical adversarial patches and a progressive sampling-based method are proposed to improve the attack robustness in the real world.
- Validated attacks in the physical world.

# L-HAWK: A Controllable Physical Adversarial Patch Against a Long-Distance Target

---

**Thanks for listening!**