Poltergeist: Acoustic Adversarial Machine Learning against Cameras and Computer Vision

Xiaoyu Ji[1], Yushi Cheng[1], Yuepeng Zhang[1], Kai Wang[1],
Chen Yan[1], Wenyuan Xu[1], Kevin Fu[2]

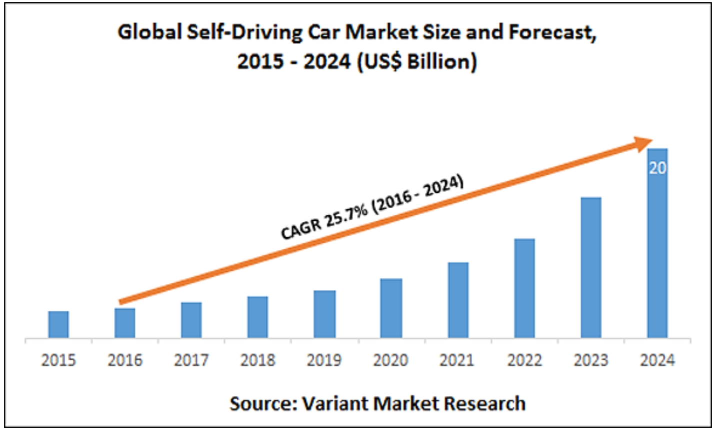
[1] Ubiquitous System Security Lab (USSLAB), Zhejiang University
[2] Security and Privacy Research Group (SPQR), University of Michigan

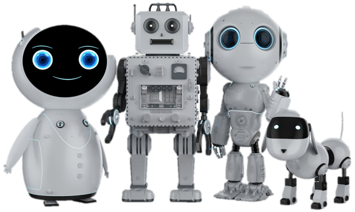
Autonomous unmanned systems are booming！

Unmanned Ground Vehicles

Unmanned Arial Vehicles

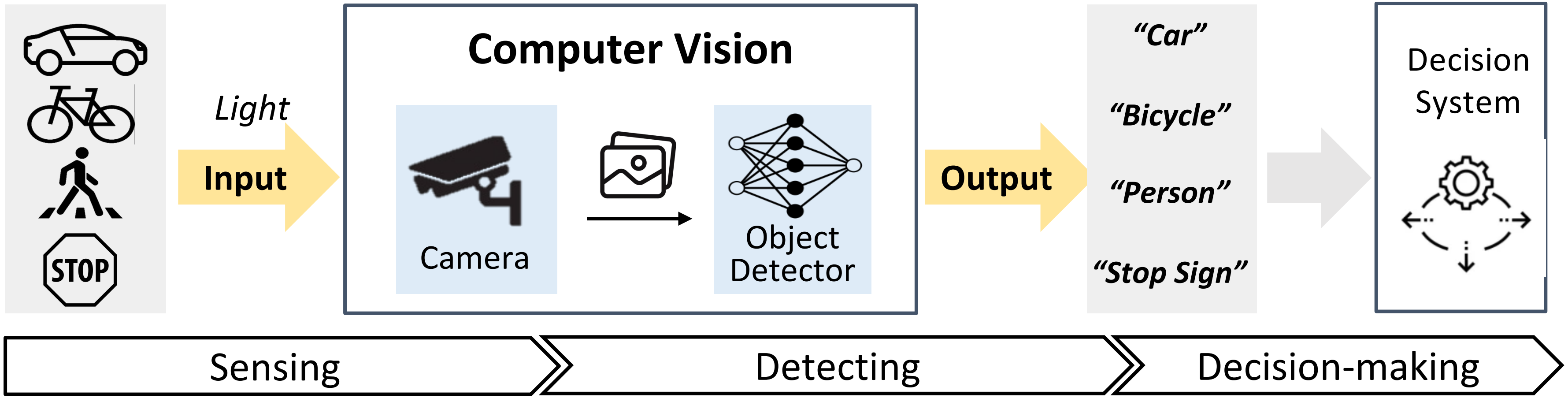


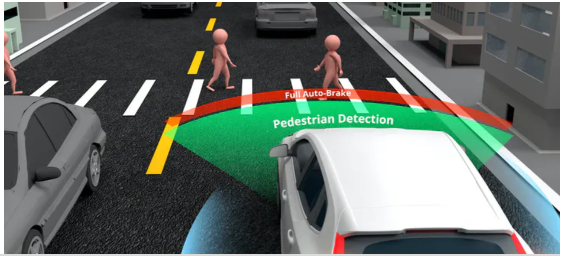
Unmanned Ships



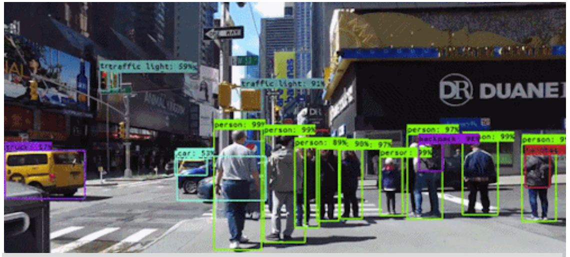
Robots

Computer Version in Autonomous Vehicles

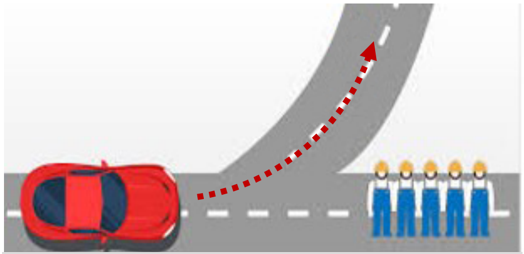




① *Camera sensing*



② *Pedestrian detected*



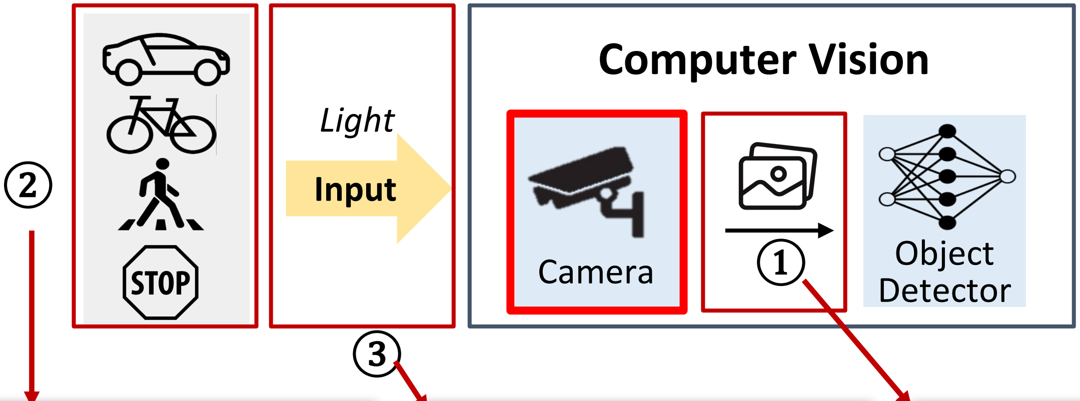
③ *“Take a turn”*

Adversarial Attacks against Computer Vision

Manipulating computer vision may result in tragic decisions

Existing Work

Focus on altering the images, objects and lights



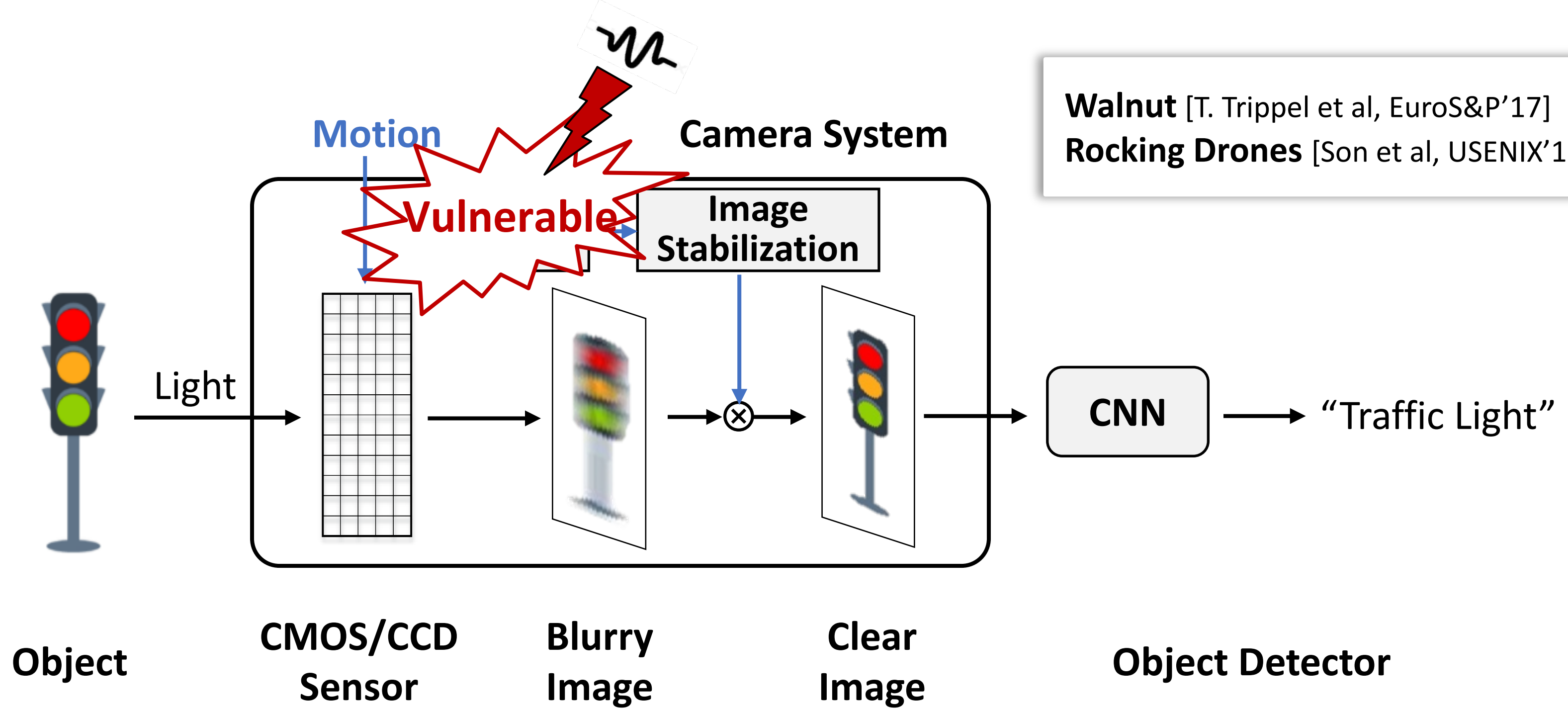
Physical stickers or patterns on the objects
[K. Eykholt et al, CVPR'18]
[Y. Zhao et al, CCS'19]

Projecting lights into cameras
[Y. Man et al, RAOD'20]

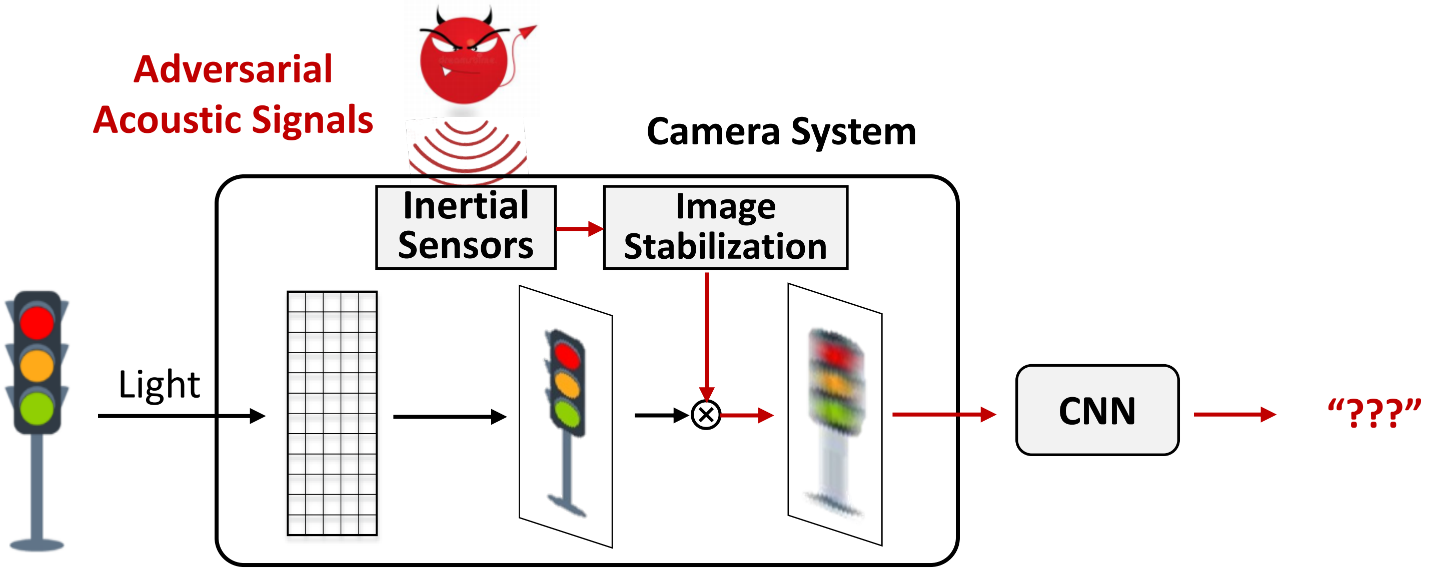
Digital pixel perturbation on the images
[S.-M. Moosavi-Dezfooli et al, CVPR'16]
[N. Carlini and D. Wagner, S&P'17]
[J. Su et al, TEC'19]

Poltergeist Attacks- Utilizing auxiliary sensors

Walnut [T. Trippel et al, EuroS&P'17]
Rocking Drones [Son et al, USENIX'15]



Poltergeist Attacks- Utilizing auxiliary sensors



Object

CMOS/CCD Sensor

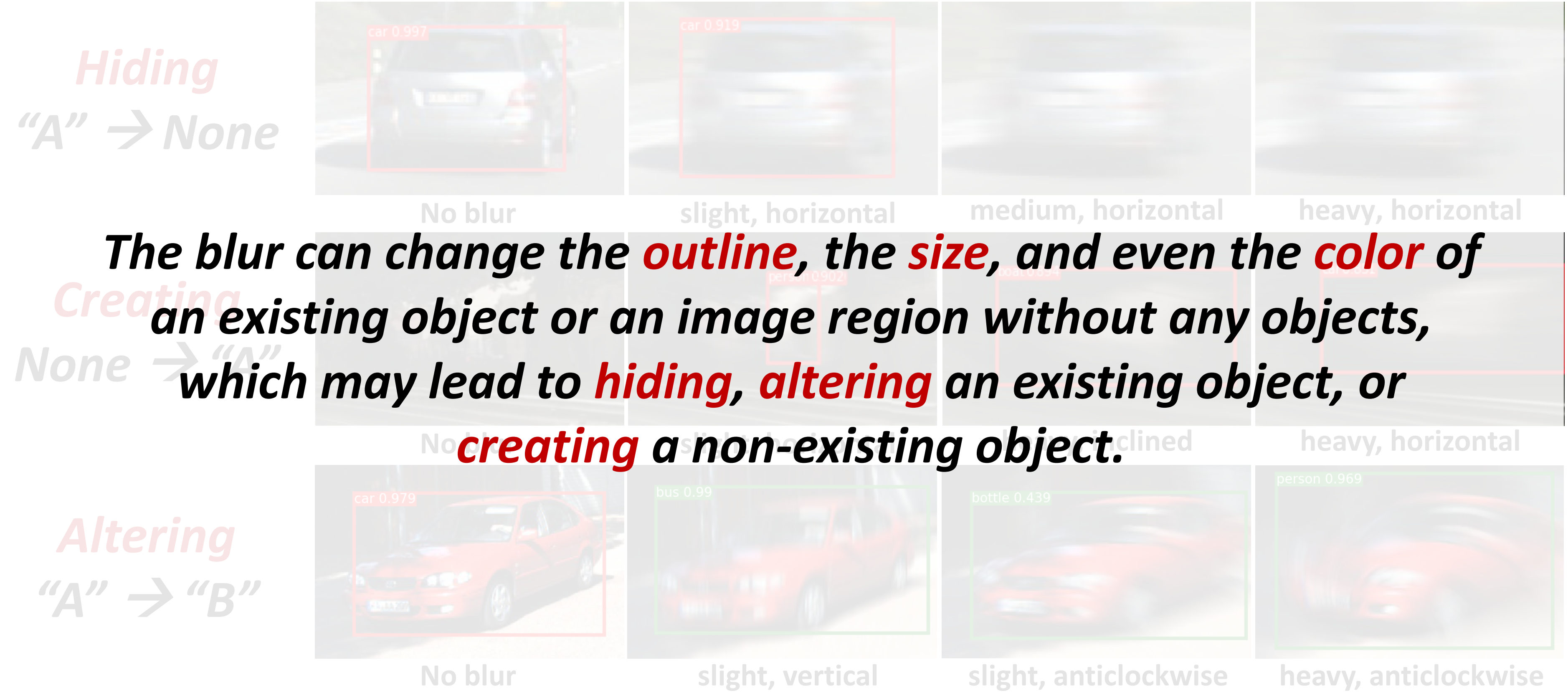
Clear Image

Blurry Image

Object Detector

Can we inject acoustic signals into adversarial examples?

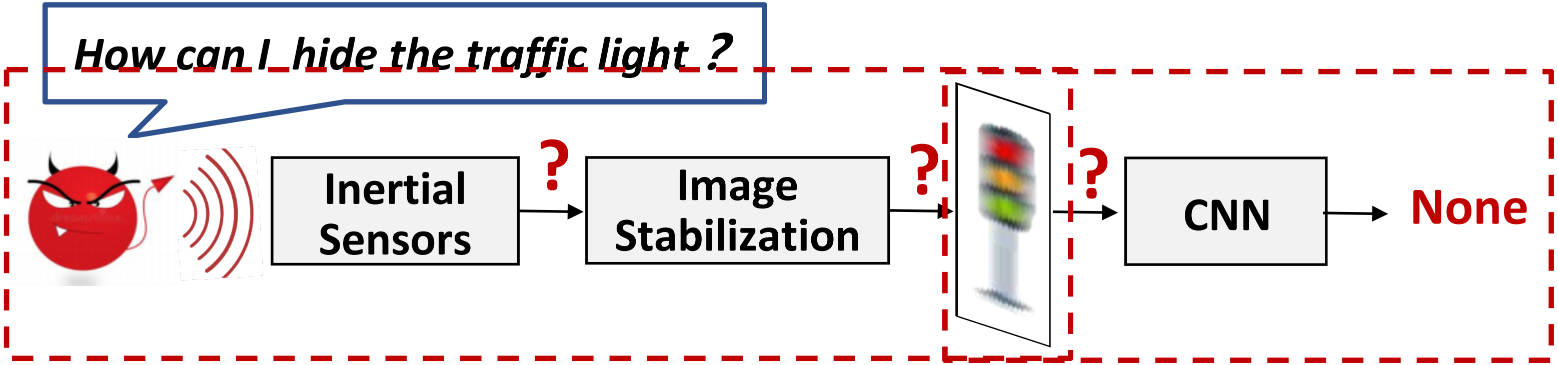
Preliminary Analysis- Stimulation



The blur can change the outline, the size, and even the color of an existing object or an image region without any objects, which may lead to hiding, altering an existing object, or creating a non-existing object.

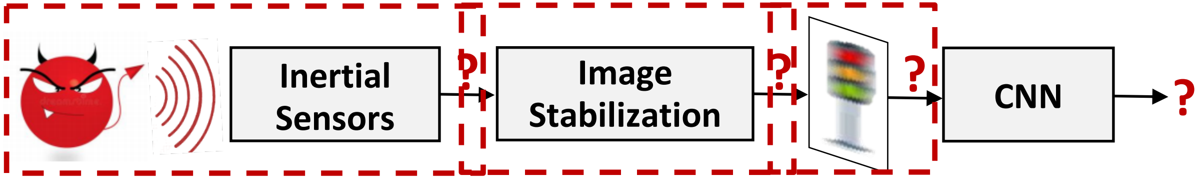
Challenges

- How to quantify the impact of acoustic signals on the level and patterns of the image blur?
- How to optimize the blur patterns for an effective and efficient attack against black-box object detectors?



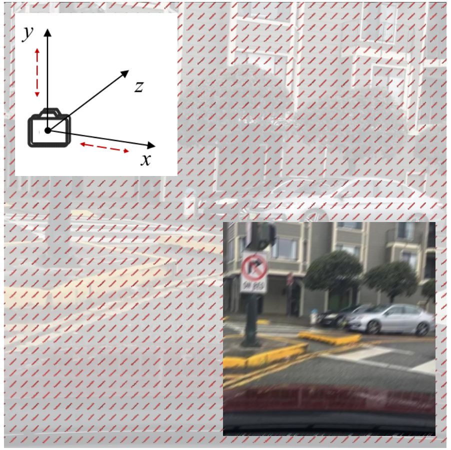
Challenge 1: Acoustic signals → Image blur patterns

- Acoustic signals → Sensor readings
  - Walnut [T. Trippel et al, EuroS&P'17], Rocking Drones [Yunmok Son et al, USENIX'15]
  - Accelerometer readings: $\{\vec{a}_x, \vec{a}_y, \vec{a}_z\}$
  - Gyroscope readings: $\{\vec{\omega}_r, \vec{\omega}_p, \vec{\omega}_y\}$
- Sensor readings → Compensatory camera motions → Pixel motions
  - $\{\vec{a}_x, \vec{a}_y\}$ → $\{-\vec{a}_x, -\vec{a}_y\}$ → linear motion: $\vec{L}_{xy} = \frac{f}{2u}(\vec{a}_x + \vec{a}_y)T^2$, $\alpha = \arccos\left(\frac{\vec{a}_x \cdot \vec{a}_y}{|\vec{a}_x||\vec{a}_y|}\right)$
  - $\vec{a}_z$ → $-\vec{a}_z$ → radial motion: $p = \frac{\vec{a}_z T^2}{2u}$
  - $\vec{\omega}_r$ → $-\vec{\omega}_r$ → rotational motion: $\beta = \omega_r T$

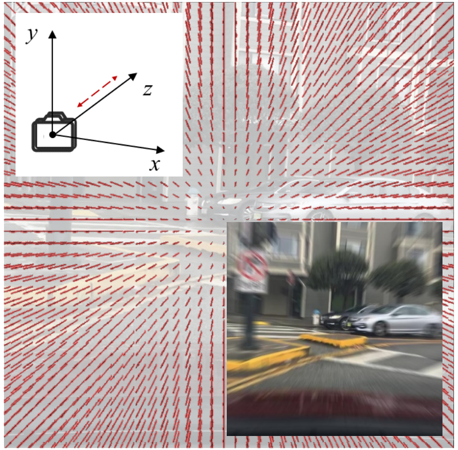


Challenge 1: Acoustic signals → image blur patterns

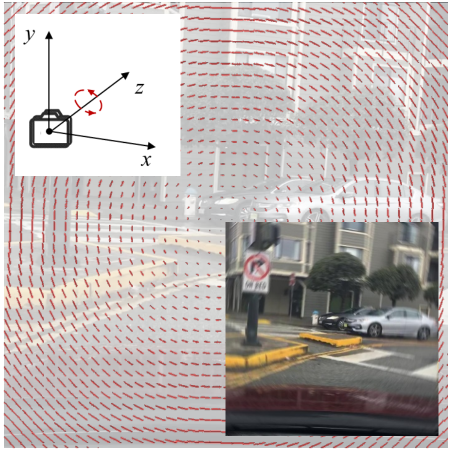
- Pixel motions → Four types of adversarial blur patterns



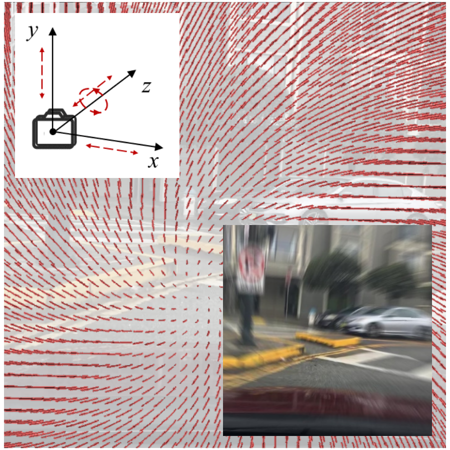
Linear Motion Blur



Radial Motion Blur



Rotational Motion Blur



Heterogeneous Motion Blur

Challenge 2: Blurry images→Object misclassification

- Large parameter space
  - Four degrees-of-freedom
  - Four kinds of motion blur patterns
- Black-box object detector
  - Unknown architecture, parameters
  - No gradient
- Physical Constraints
  - Attack distance
  - Attack power

Challenge 2: Blurry images→Object misclassification

Objective functions

- Attack effectiveness, Attack cost, Physical attack capability restriction

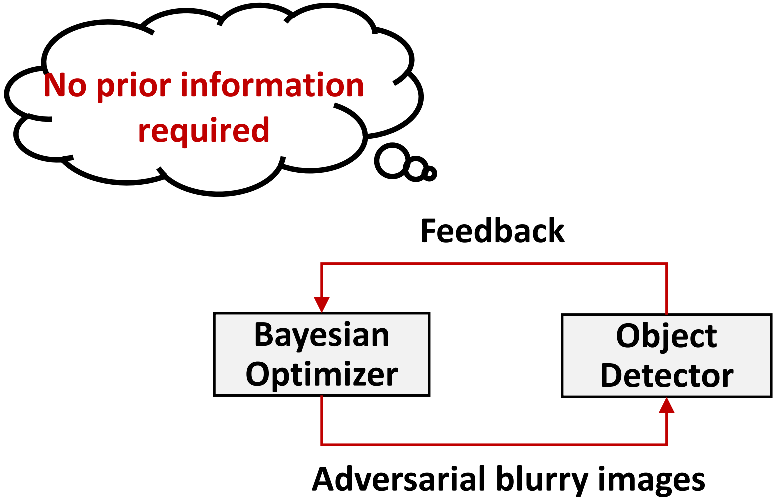
$$\textbf{HA:}\quad \underset{\vec{a}_x,\vec{a}_y,\vec{a}_z,\vec{\omega}_r}{\operatorname{argmin}} \quad w_1 S_i^B S_i^C \; + \; w_2||\Delta||_p \qquad \text{s.t.} \quad |\vec{a}_x+\vec{a}_y+\vec{a}_z| < \xi_1, \quad |\vec{\omega}_r| < \xi_2$$

$$\textbf{CA:}\quad \underset{\vec{a}_x,\vec{a}_y,\vec{a}_z,\vec{\omega}_r}{\operatorname{argmin}} \quad -w_3 \frac{S_o^B S_o^C|_{C_o=T}}{\sum_{i=1}^{m} U_{oi}} \; + \; w_4||\Delta||_p \qquad \text{s.t.} \quad |\vec{a}_x+\vec{a}_y+\vec{a}_z| < \xi_1, \quad |\vec{\omega}_r| < \xi_2$$

$$\textbf{AA:}\quad \underset{\vec{a}_x,\vec{a}_y,\vec{a}_z,\vec{\omega}_r}{\operatorname{argmin}} \quad -w_5 U_{ii'} S_o^B S_o^C|_{C_o=T} + w_6||\Delta||_p \qquad \text{s.t.} \quad |\vec{a}_x+\vec{a}_y+\vec{a}_z| < \xi_1, \quad |\vec{\omega}_r| < \xi_2$$

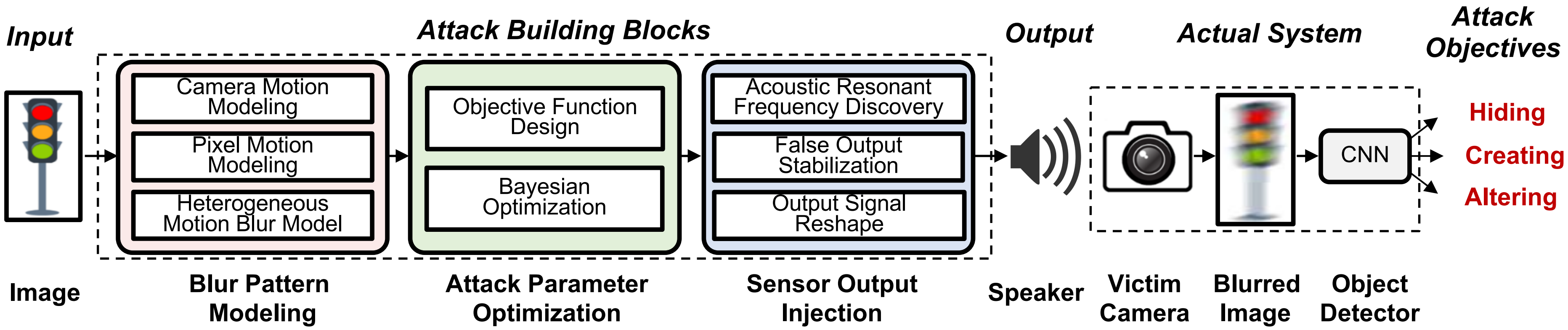
Bayesian Optimizer

- Gradient-free strategy
- Global optimization for black-box functions



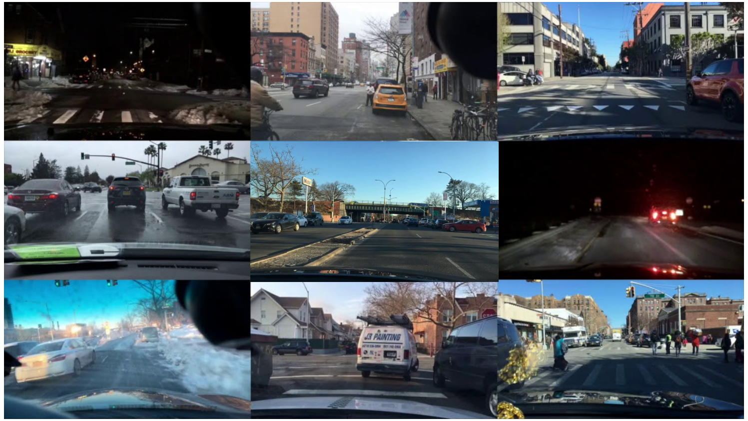
System Design

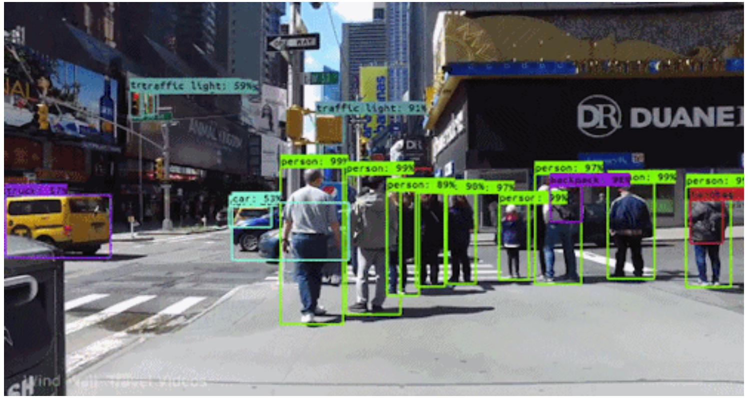
- Three key **attack building blocks**
  - Blur Pattern Modeling
  - Attack parameter Optimization
  - Sensor Output Injection



Evaluation-Simulation

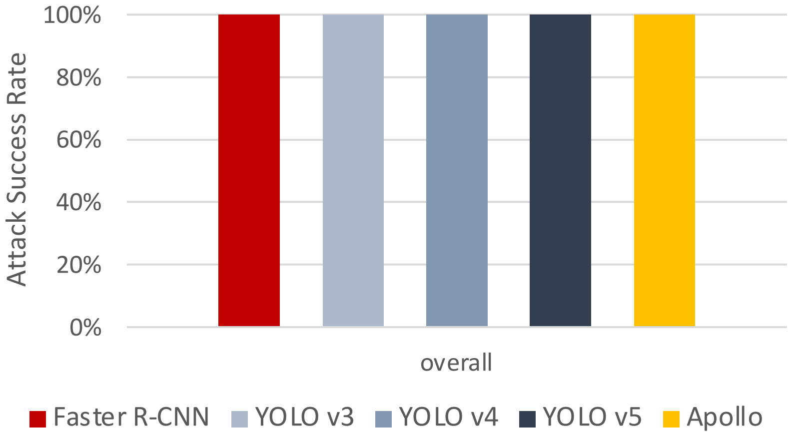
- **Datasets:**
  - 2 popular self-driving datasets
  - BDD100K, KITTI
- **Object Detectors:**
  - 5 state-of-the-art object detectors
  - Academic: Faster R-CNN, YOLO v3/v4/v5
  - Commercial: Apollo
- **Object of Interest (OOI):**
  - person, car, truck, bus, traffic light, stop sign

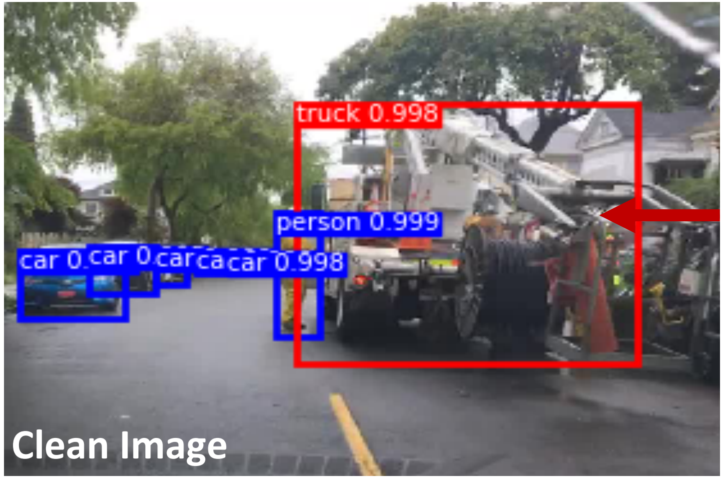




Attack Effectiveness

- **Hiding Attack (HA)**
  - Targeted: One → None



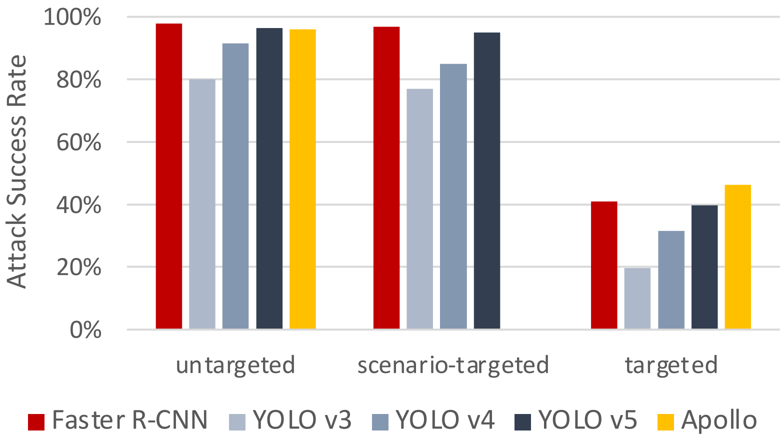


Truck

None

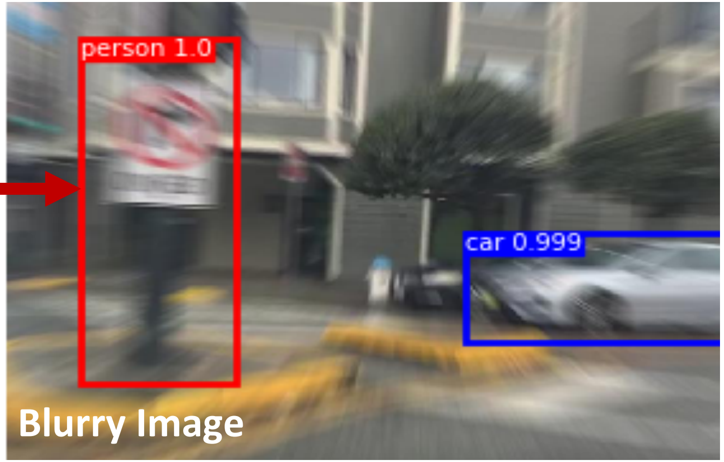
Attack Effectiveness

- **Creating Attack (CA)**
  - Untargeted: None → Any
  - Scenario-targeted: None → A Set
  - Targeted: None → One



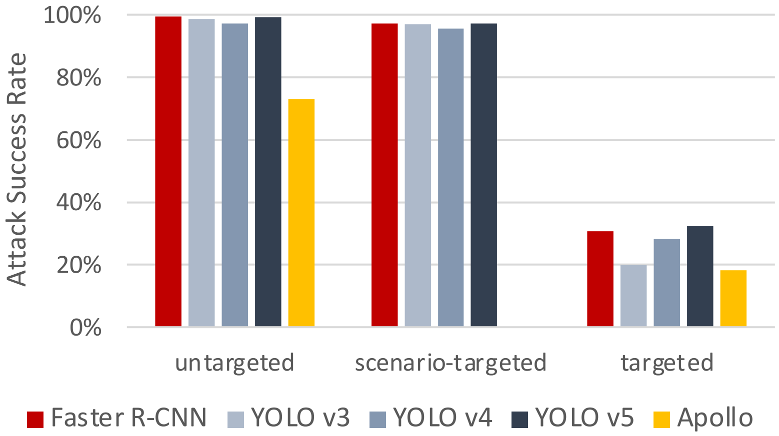
None

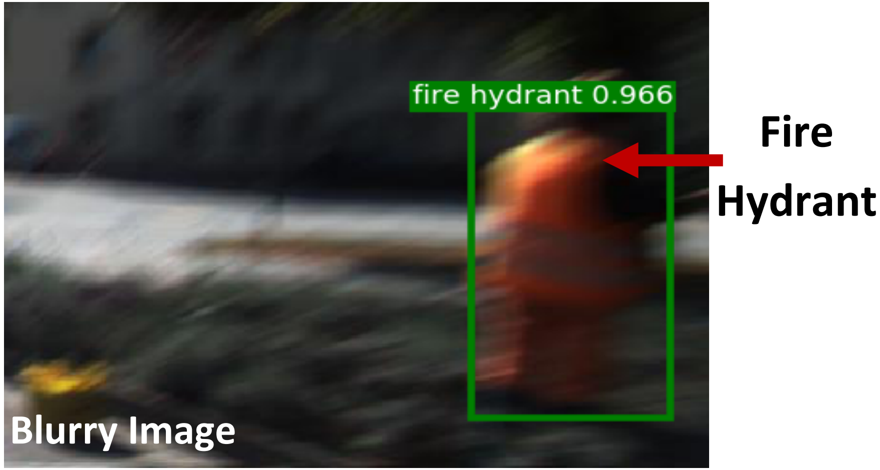
Person



Attack Effectiveness

- **Altering Attack (AA)**
  - Untargeted: One → Any
  - Scenario-targeted: One → A Set
  - Targeted: One → One

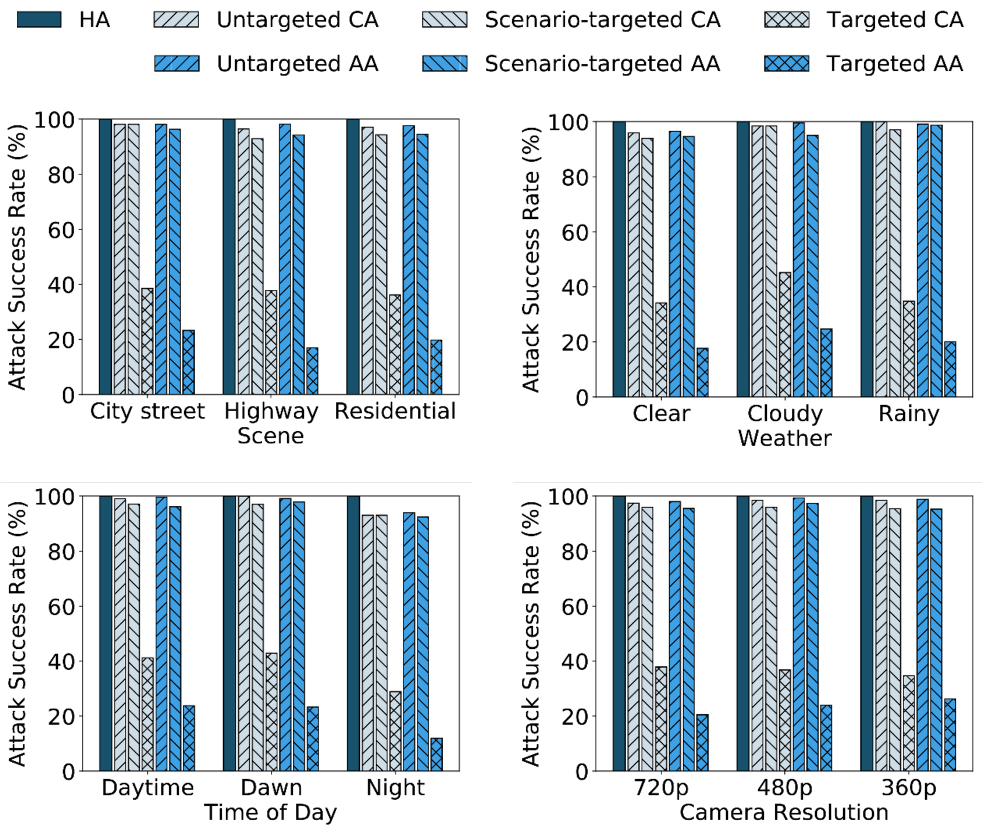




Attack Robustness

- ☐ **Scene**
- ☐ **Weather**
- ☐ **Time of Day**
- ☐ **Camera Resolution**

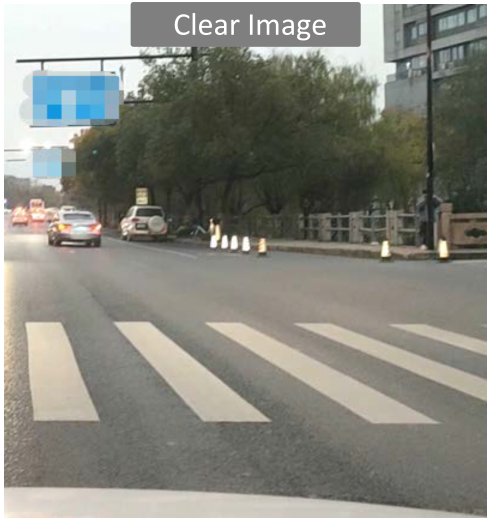
PG attacks are robust across various scenes, weathers, time periods of a day, and camera resolutions.

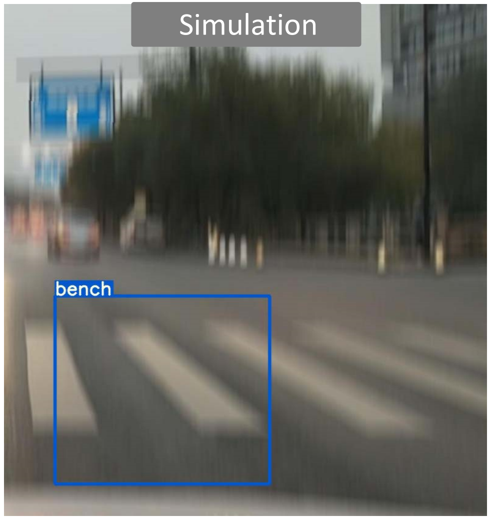


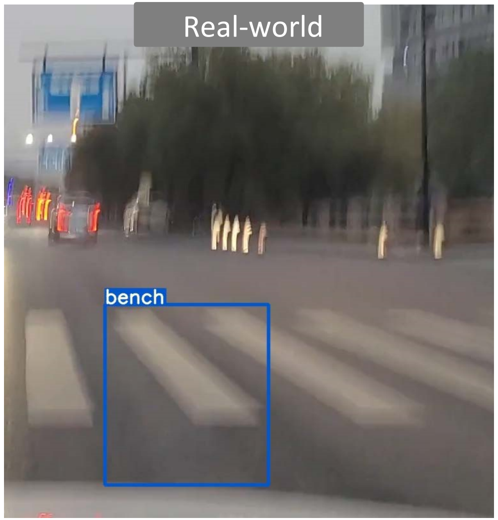
Evaluation-Real World

- **Target:** Samsung S20 smartphone in a moving vehicle
- **Attack device:** Ultrasonic Speaker
- **Scenes:**
  - City Lane
  - City Crossroad
  - Tunnel
  - Campus Road

Simulation vs. Real-world



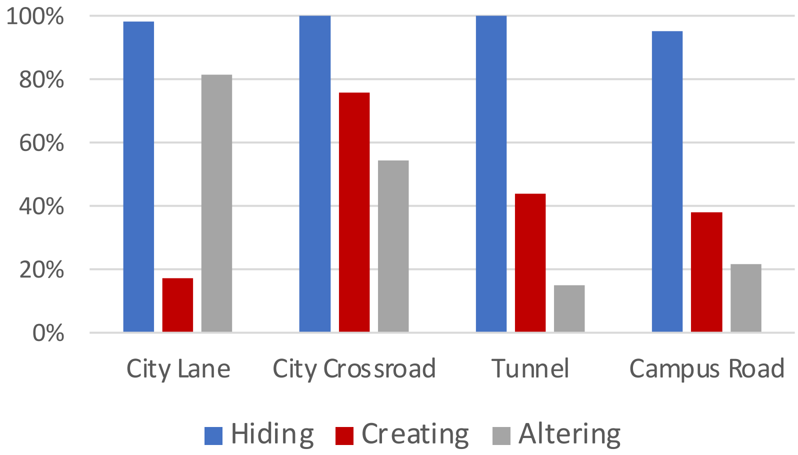




The simulated images are representative of the ones created in the presence of real attacks.

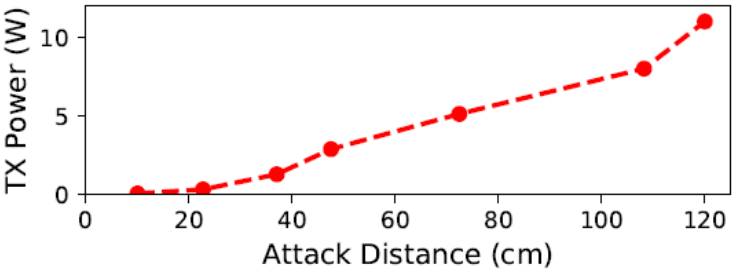
Attack Effectiveness

☐ Overall Performance



HA shows a good performance in any scenes
CA and RA works well in special enviroments

☐ Impact of Attack Distances



An attack powerof 10 W suffices to launch an attack from 1.1 m away

Real-world Attack Videos

Altering car into person

Creating a truck

Hiding the car

https://github.com/USSLab/PoltergeistAttack

Countermeasures

- **MEMS Inertial Sensors Safeguarding**
  - Acoustic Isolation
  - Secure Low-pass Filter
- **Image Stabilization Techniques**
  - Additional Digital Image Stabilization
- **Object Detection Algorithms**
  - Input Image De-blur
  - Detection Model Improvement
- **Sensor Fusion Techniques**
  - LiDARs, radars combined with cameras

Conclusion

- Discovered a new class of system-level vulnerabilities, **AMpLe attacks**, injecting physics into Adversarial Machine Learning
- Proposed **Poltergeist attacks**, acoustic adversarial machine learning against cameras and computer vision
- Evaluation showed high performance against 4 academic and 1 commercial object detectors

- Future work
  - Leveraging signal transmission via ultrasound, visible light, infrared, lasers, radio, magnetic fields, heat, fluid, etc. for AMpLe attacks

Poltergeist: Acoustic Adversarial Machine Learning against Cameras and Computer Vision

Thank you !

Contact the authors at:

xji@zju.edu.cn

yushicheng@zju.edu.cn

wyxu@zju.edu.cn

kevinfu@umich.edu

Lab websites:

usslab.org

spqr.eecs.umich.edu

Paper websites:

https://github.com/USSLab/PoltergeistAttack