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

What is the Poltergeist attack?

Poltergeist attack is a new kind of attack that exploits the **cameras' auxiliary sensor vulnerabilities via acoustic manipulation to create misclassification in object detection systems.** By emitting deliberately designed acoustic signals, Poltergeist attacks can control the output of an inertial sensor, which triggers unnecessary motion compensation and results in a blurred image, even if the camera is stable. The blurred images can then induce object misclassification affecting safety-critical decision making, as shown in Fig. 1.

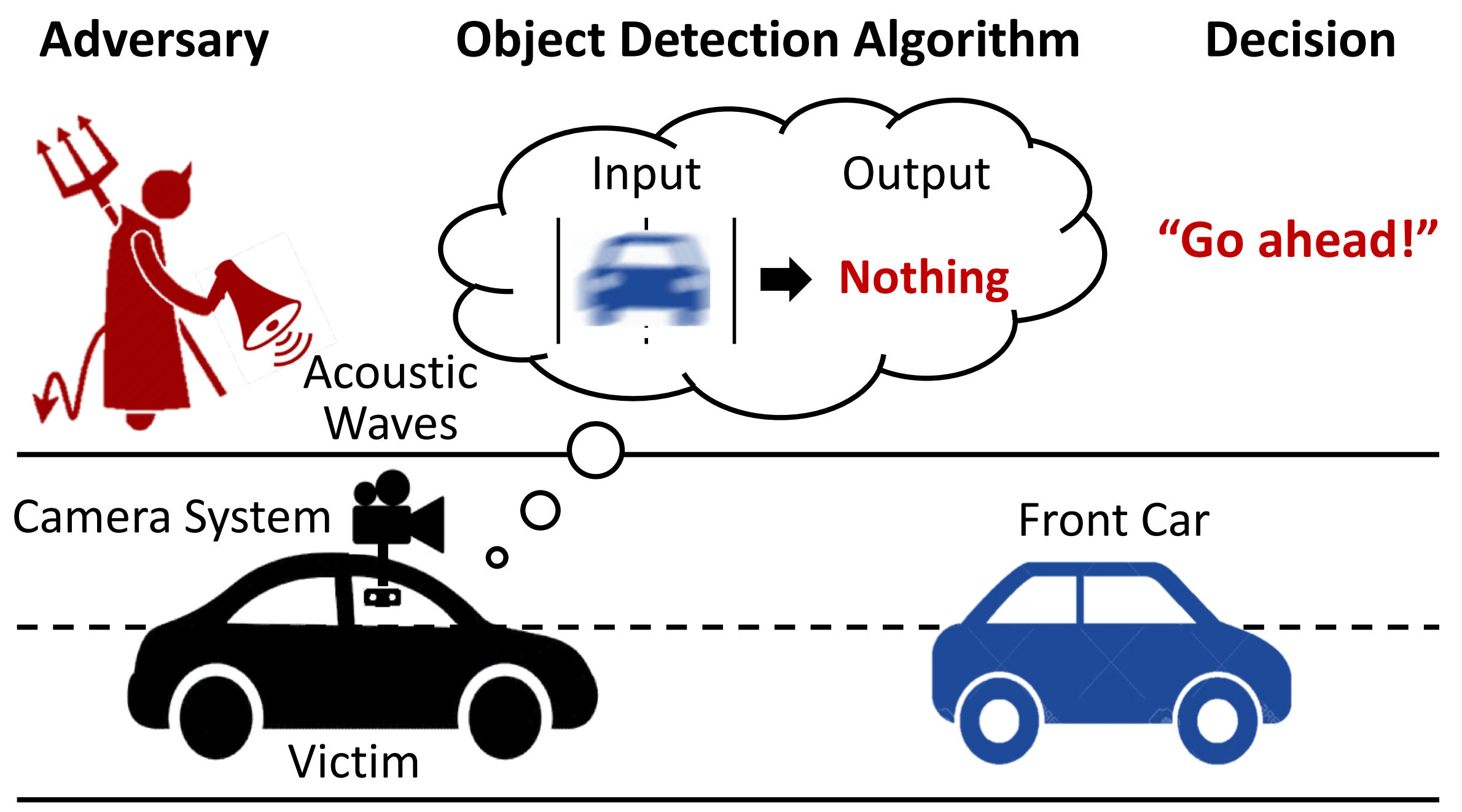


Figure 1. By injecting acoustic signals into the inertial sensors of object-detection systems in autonomous vehicles, an adversary can fool decision making.

What can the Poltergeist attack do?

Poltergeist attacks can achieve **three types of attack objectives** against modern object detectors:

- **Hiding attacks (HA)**, which cause an object to become undetected, e.g., make a front car "disappear".
- **Creating attacks (CA)**, which induce a non-existent object, e.g., create a car or a person in the driveway.
- **Altering attacks (AA)**, which cause an object to be misclassified, e.g., render a person detected as a fire hydrant.

Why is the Poltergeist attack feasible?

To increase the quality of captured images and thus the detection accuracy, object-detection systems utilize inertial sensors and image stabilization to reduce the blur effect caused by camera motions, as shown in Fig. 2. However, inertial sensors haven been proved to be **vulnerable to resonant acoustic injection attacks**, and it is feasible to have finer control over their outputs using acoustic signals [1].

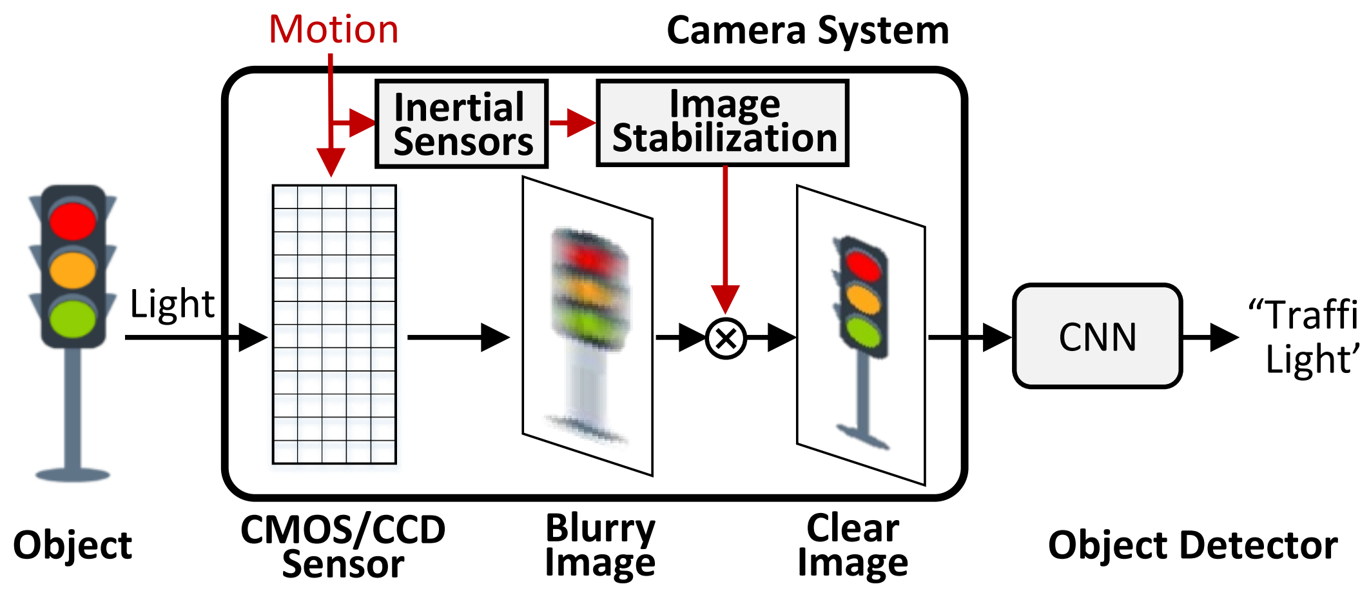


Figure 2. Object-detection systems

As a result, an attacker can manipulate the sensor outputs and the motion compensation process using acoustic signals. The blur caused by unnecessary motion compensation then 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.

How does the Poltergeist attack work?

When launching Poltergeist attacks, the adversary first uses an image of the target object to generate feasible attack parameters with **blur pattern modeling** and **attack parameter optimization**. Then, the adversary manipulates the sensor outputs according to the calculated parameters via acoustic signals to deceive the object detector, which may lead to hide, create, or alter objects.

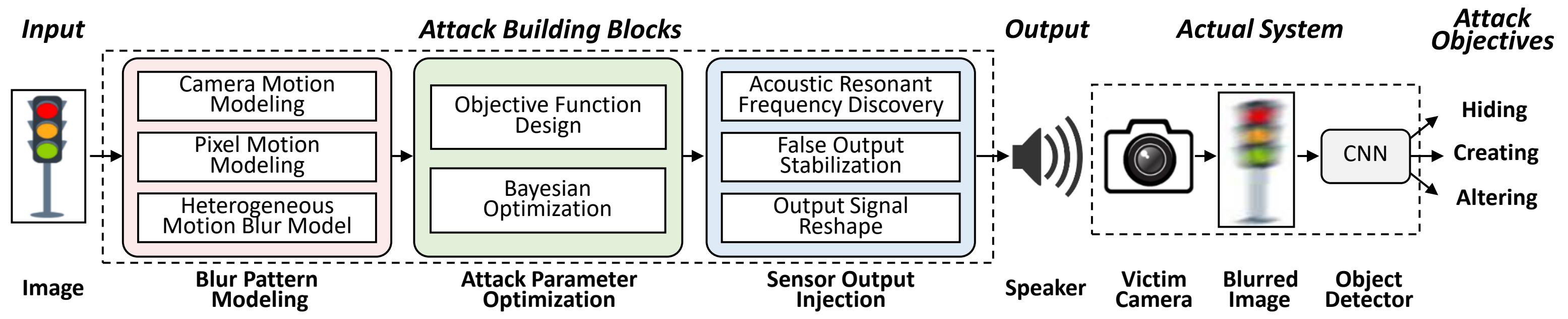
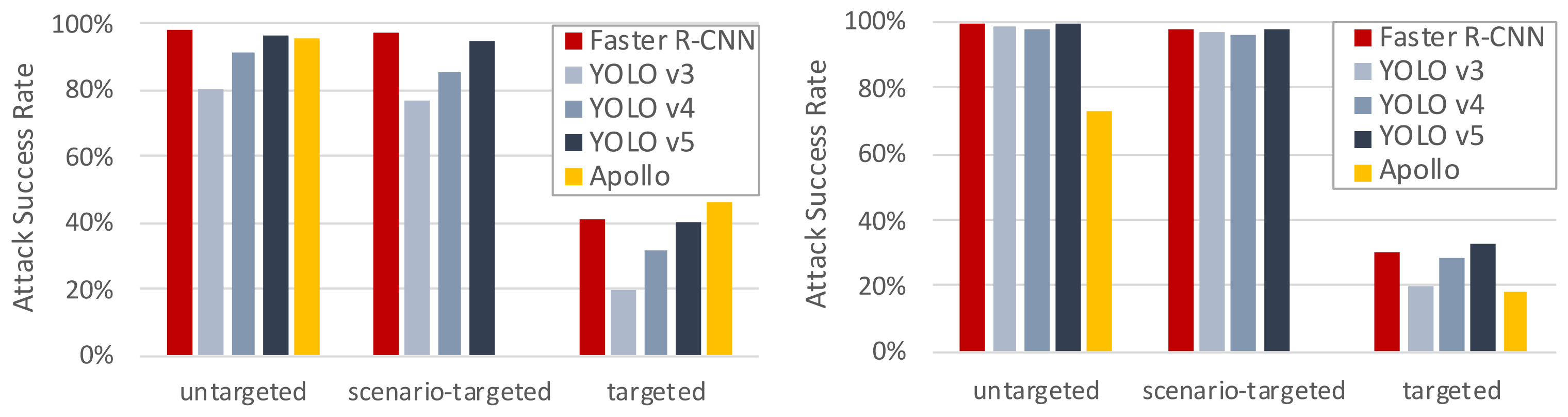


Figure 3. Overview of Poltergeist attacks

Simulation Evaluation

- **Datasets:** BDD100K, KITTI
- **Object Detectors:** Faster R-CNN, YOLO v3/v4/v5, Apollo
- **Attack Forms:** (1) Untargeted, (2) Scenario-targeted, (3) Targeted

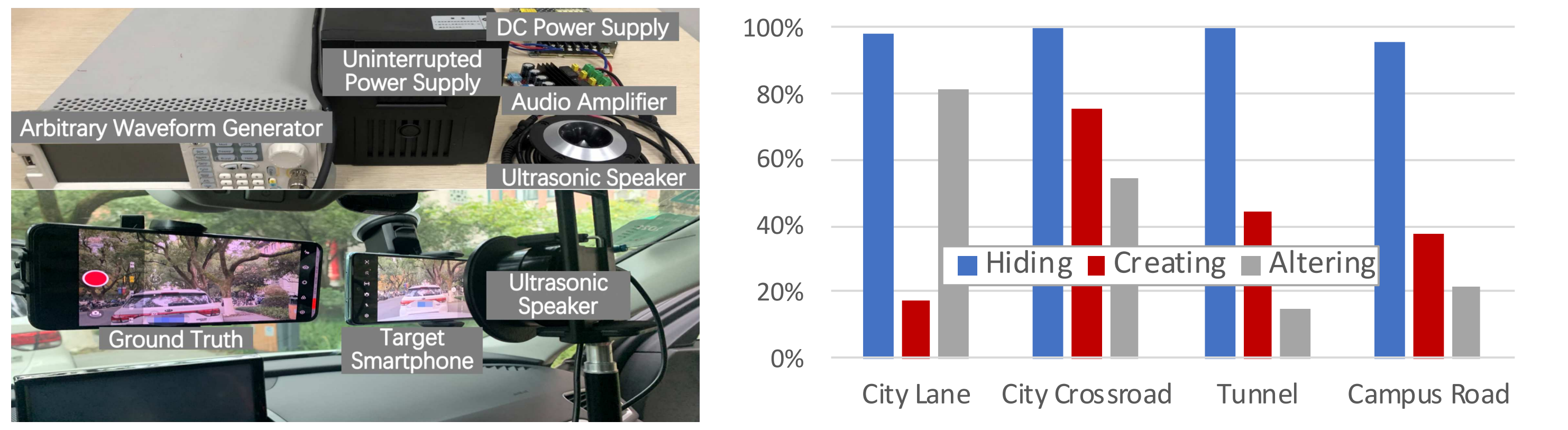


(a) Creating attacks (b) Altering attacks

Figure 4. Simulation evaluation results

Real-world Evaluation

- **Target Device:** Samsung S20 smartphone in a moving vehicle
- **Attack Device:** Ultrasonic Speaker
- **Scenes:** (1) City Lane, (2) City Crossroad, (3) Tunnel, (4) Campus Road



(a) Real-world attack setups (b) Real-world attack success rates

Figure 5. Real-world evaluation

References

[1] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks. In *Proceedings of the 2017 IEEE European Symposium on Security and Privacy (EuroS&P'17)*, pages 3–18. IEEE, 2017.