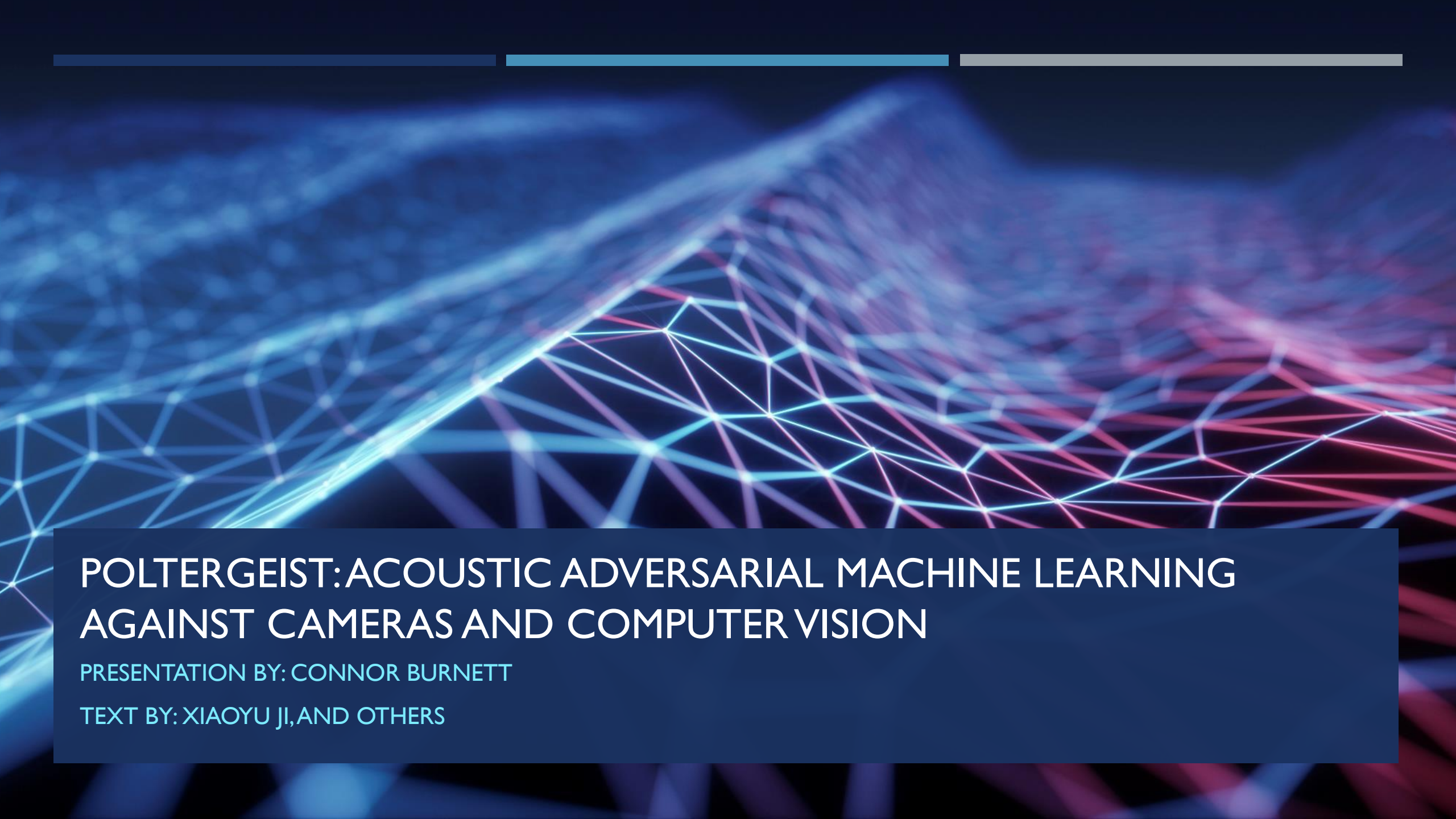
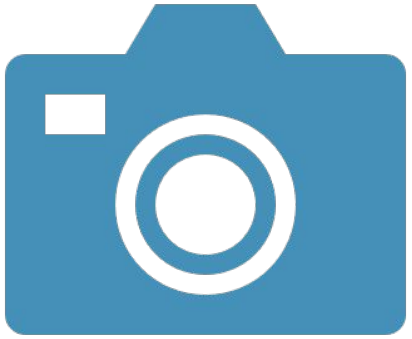


POLTERGEIST: ACOUSTIC ADVERSARIAL MACHINE LEARNING AGAINST CAMERAS AND COMPUTER VISION

PRESENTATION BY: CONNOR BURNETT

TEXT BY: XIAOYU JI, AND OTHERS



Onboard Cameras

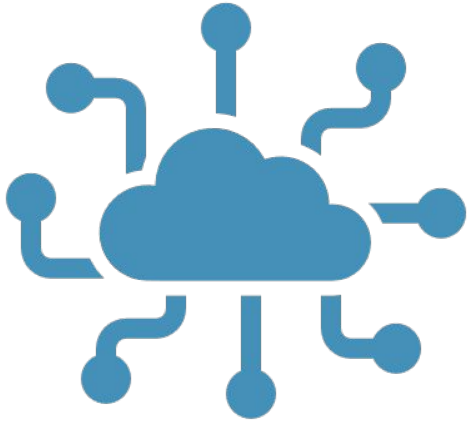
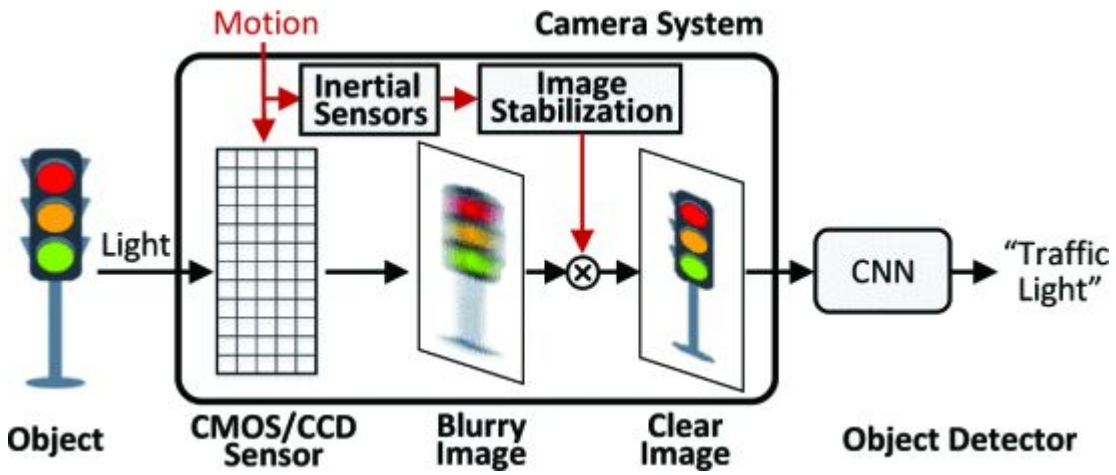


Image Classification and Detection

Car Decision and Movement

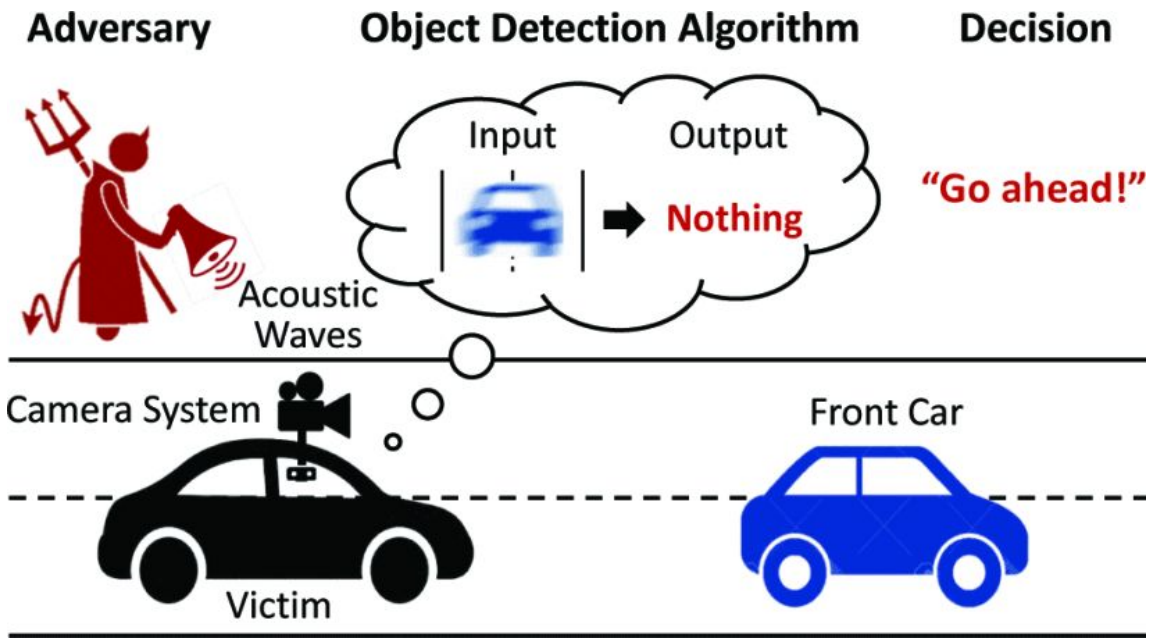
AUTONOMOUS VEHICLE FLOW

IMAGE STABILIZATION



- For smooth video, cameras rely on stabilizers that cancel out sudden movements.
- Stabilizers use Inertial Measurement Units (IMU) which contain accelerometers and gyroscopes to detect movement.
- **However, inertial sensors are susceptible to acoustic attacks, which cause stabilizers to overcorrect, resulting in blurry images.**
- **This is the basis for a poltergeist attack (PG attack)**

POLTERGEIST ATTACK



- Acoustic waves are sent from an adversary, directed at the camera system
- The camera system overcorrects, causing the image of car to be blurry
- The object detection algorithm does not detect a vehicle
- The vehicle determines it is safe to move forward

ATTACK TYPES

- Hiding Attacks
- Creating Attacks
- Altering Attacks

HIDING ATTACKS (HA)

(a) Car detected without any motion blur (confidence score 0.997)

(b) Car detected (0.919) after linear motion blur (slight, horizontal)

(c) Nothing detected after linear motion blur (medium, horizontal)

(d) Nothing detected after linear motion blur (heavy, horizontal)

- The goal is to produce an image where the object detector fails to identify an object of interest.
  - The greater the blur, the more the object detector struggles to detect the SUV

CREATING ATTACKS (CA)

(a) Nothing detected for the original image without any motion blur

(b) Person detected (0.902) after linear motion blur (slight, horizontal)

(c) Boat detected (0.894) after linear motion blur (heavy, inclined)

(d) Car detected (0.851) after linear motion blur (heavy, horizontal)

- The goal is to produce an image where the object detector detects a non-existent object.
  - Despite no object being present, the blur makes the object detector believe a person, boat, or car exists.

ALTERING ATTACKS (AA)

(a) Car detected without any motion blur (confidence score 0.979)

(b) Car is misclassified as bus (0.99) after linear motion blur (slight, vertical)

(c) Car is misclassified as bottle (0.439) after rotational motion blur (slight, anticlockwise)

(d) Car is misclassified as person (0.969) after rotational motion blur (heavy, anticlockwise)

- The goal is to produce an image where an existent object is incorrectly detected as a different object.
  - Based on the blur, the car is detected as a bus, bottle, or person.

DESIGNING AN EFFECTIVE ATTACK

CREATING A BLUR MODEL

ADVERSARY ASSUMPTIONS

- Black-box Object Detector
  - The adversary has no prior knowledge of the object detector
  - The adversary can obtain the classification results and confidence scores
- Camera and Sensor Awareness
  - The adversary can acquire and analyze a camera of the same model used in the target system
- Attack Capability
  - The adversary can set up an ultrasonic speaker along the roadside, attach speakers inside the vehicle, or control a compromised onboard speaker system in the target vehicle

BLUR PATTERN MODELING

- Camera stabilization motion has up to six DOFs (degree-of-freedom)
  - $\vec{M} = \{\overrightarrow{a_x}, \overrightarrow{a_y}, \overrightarrow{a_z}, \overrightarrow{\omega_r}, \overrightarrow{\omega_p}, \overrightarrow{\omega_y}\ \}$
- Camera motions can be converted into pixel motions
  - Three types of pixel motion (pitch and yaw are excluded since they require additional pixel information)
    - $\{\ \overrightarrow{a_x}, \overrightarrow{a_y}\ \}$
    - $\{\ \overrightarrow{a_z}\ \}$
    - $\{\ \overrightarrow{\omega_r}\}$
- Pixel motions can be converted into Blur Patterns
  - Each pixel motion has an equivalent blur pattern
    - Linear
    - Radial
    - Rotational

MOTION BLURS PART I

Linear Motion Blur

- Blur pattern caused by linear pixel motions (x,y)

(a) The *truck* in the clear image (left) is hidden after blurring (right).

$\{\vec{a}_x, \vec{a}_y\}$	Linear	$\vec{L}_{xy} = \frac{f}{2u}(\vec{a}_x + \vec{a}_y)T^2$ $\alpha = \arccos(\frac{\vec{a}_x \cdot \vec{a}_y}{\lvert\vec{a}_x\rvert \lvert\vec{a}_y\rvert})$

Radial Motion Blurs

- Blur pattern caused by radial pixel motions towards or away from center image (z)

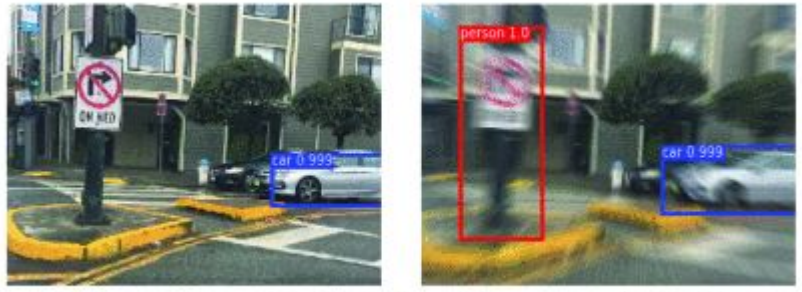
(b) The *person* and *bicycle* in the clear image (left) is hidden after blurring (right).

$\vec{a}_z$	Radial	$p = \frac{\vec{a}_z T^2}{2u}$

MOTION BLURS PART 2

Rotational Motion Blur

- Blur pattern causes by rotational motions along an arc



(c) A *person* is created with a high confidence score after blurring.

$\vec{\omega}_r$	Rotational	$\beta = \omega_r T$

Heterogenous Motion Blur

- Blur pattern that combines the linear, radial, and rotational blur. Can simulate any combination of each motion blur
  - Equation returns the entire blurred image

$$[u(k), v(k)]^T = \begin{bmatrix} \cos\alpha & \cos\left(\frac{k}{n}\beta+\gamma\right) & \cos\delta \\ \sin\alpha & \sin\left(\frac{k}{n}\beta+\gamma\right) & \sin\delta \end{bmatrix} \begin{bmatrix} \frac{kf|\vec{a}_x+\vec{a}_y|T^2}{2nu} \\ r_c \\ \frac{k|\vec{a}_z|T^2 r_o}{2nu} \end{bmatrix}$$

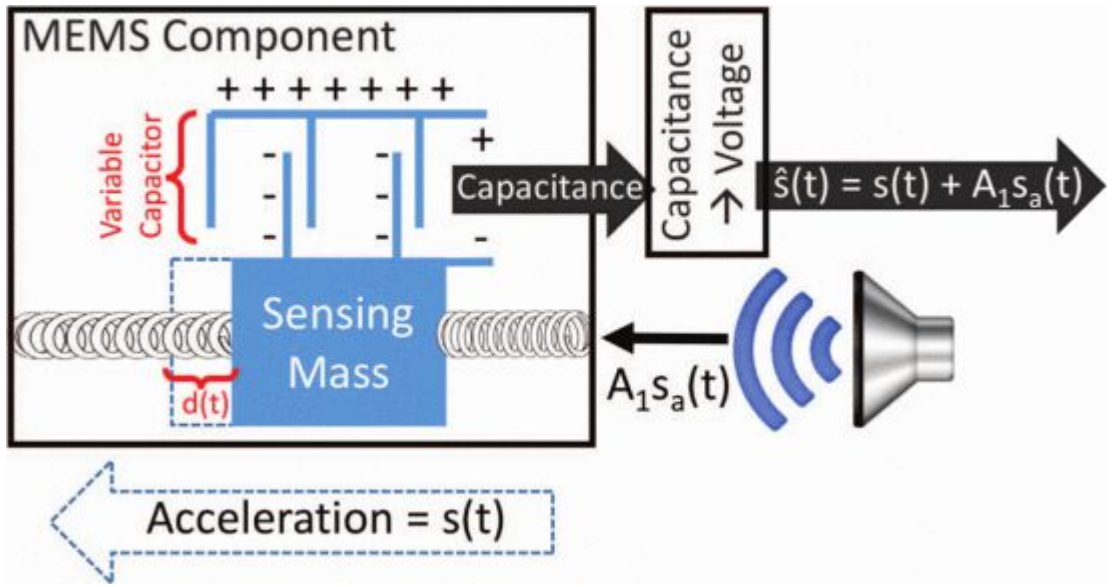
$$\gamma = \arctan\left(\frac{j-c_1}{i-c_0}\right), \quad r_c = \|(i,j),(c_0,c_1)\|_2$$

$$\delta = \arctan\left(\frac{j-o_1}{i-o_0}\right), \quad r_o = \|(i,j),(o_0,o_1)\|_2$$

OBJECTIVE FUNCTIONS

- Assuming a black-box object detector, we can represent the predictions as:
  - $Y_i = (B_i, S_i^B, C_i, S_i^C)$
    - B_i, C_i are the bounding box and class of prediction
    - S_i^B, S_i^C are the corresponding confidence scores
    - Y_i is the prediction
- Hiding Attack (HA)
  - The product of S_i^B, S_i^C should be less than the threshold that determines whether an object exists
- Creating Attack (CA)
  - The product of S_o^B, S_o^C should be larger than the threshold that determines whether an object exists
  - Bounding box intersection needs to be minimized to guarantee we are not altering an object
- Altering Attacks (AA)
  - The product of $S_i^{B\prime}, S_i^{C\prime}$ should be larger than the threshold
  - Bounding box intersection needs to be maximized to make sure we are altering an object

LAUNCHING THE SENSOR ATTACK



- Attack utilizes the sampling deficiencies at the analog-to-digital converter (ADC)
- Find the acoustic resonant frequency
  - Perform a frequency sweep until output measurements deviate from normal
- Shift the acoustic resonant frequency to induce a direct current alias at the ADC
- Control the desired output signal by transmitting arbitrary information signals over another carrier signal.
  - Amplitude Modulation: Varying the amplitude of the carrier signal overtime.
  - Phase Modulation: Varying the phase of the carrier signal overtime

EVALUATION

COMPARING PG ATTACKS AGAINST OBJECT-DETECTION SYSTEMS

SIMULATION EVALUATION

- Used the BDD100K and KITTI driving datasets
  - Both datasets are large and diverse datasets for computer vision evaluation
  - The images were blurred and tested against commercial and academic object detectors
- Found that hiding attacks (HA) have a 100% success rate against black-box object detectors
- Creating attacks (CA) and Alerting attacks (AA) had high success for untargeted attacks, but much less success for targeted attacks

SIMULATION EVALUATION

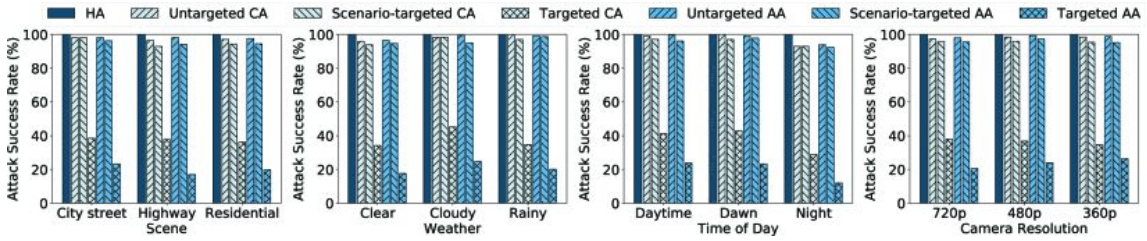
Creating Attacks

Black-box Detector	Creating Attack	Overall Attack Success Rate			
		BDD100K		KITTI	
YOLO V3	Untargeted†	69.5%		80.0%	
	Scenario-targeted‡	68.5%		77.0%	
	Targeted§	16.6% (Avg.)	person (12.0%), car (57.5%), truck (8.5%), bus (7.0%), traffic light (13.5%), stop sign (1.0%)	19.7% (Avg.)	person (31.0%), car (58.0%), truck (8.5%), bus (7.0%), traffic light (10.5%), stop sign (3.0%)
YOLO V4	Untargeted	93.0%		91.5%	
	Scenario-targeted	88.5%		85.0%	
	Targeted	34.3% (Avg.)	person (42.5%), car (83.5%), truck (30.0%), bus (12.5%), traffic light (34.5%), stop sign (2.5%)	31.6% (Avg.)	person (52.5%), car (72.5%), truck (31.5%), bus (10.0%), traffic light (22.5%), stop sign (0.5%)
YOLO V5	Untargeted	97.5%		96.5%	
	Scenario-targeted	96.0%		95.0%	
	Targeted	37.7% (Avg.)	person (57.5%), car (90.5%), truck (23.5%), bus (14.0%), traffic light (37.5%), stop sign (3.0%)	39.8% (Avg.)	person (71.0%), car (87.0%), truck (25.5%), bus (9.5%), traffic light (40.5%), stop sign (5.5%)
Faster R-CNN	Untargeted	97.4%		97.9%	
	Scenario-targeted	95.9%		96.9%	
	Targeted	37.9% (Avg.)	person (65.0%), car (88.7%), truck (19.6%), bus (30.9%), traffic light (20.1%), stop sign (3.1%)	40.9% (Avg.)	person (88.7%), car (80.4%), truck (12.4%), bus (31.4%), traffic light (16.0%), stop sign (16.5%)
Apollo	Untargeted	91.2%		96.0%	
	Targeted	40.2% (Avg.)	person (47.4%), car (79.9%), truck (18.0%), bus (15.5%)	46.2% (Avg.)	person (67.7%), car (83.8%), truck (15.2%), bus (18.2%)

Altering Attacks

Black-box Detector	Altering Attack	Overall Attack Success Rate			
		BDD100K		KITTI	
YOLO V3	Untargeted†	91.8%		98.7%	
	Scenario-targeted‡	82.2% (Avg.)	OOI* → OOU** (82.1%), OOU → OOI (75%)	96.9% (Avg.)	OOI → OOU (96.8%), OOU → OOI (100%)
	Targeted§	23.7% (Avg.)	Top 5: bus → car (100%), stop sign → car (100%), truck → car (96.7%), bus → truck (88.9%), traffic light → car (77.8%)	19.8% (Avg.)	Top 5: bus → car (100%), truck → car (92.9%), traffic light → car (84.2%), bus → person (83.3%), bus → truck (66.7%)
YOLO V4	Untargeted	98.1%		97.2%	
	Scenario-targeted	97.9% (Avg.)	OOI → OOU (97.8%), OOU → OOI (100%)	95.6% (Avg.)	OOI → OOU (95.5%), OOU → OOI (97.3%)
	Targeted	32.3% (Avg.)	Top 5: bus → car (100%), truck → car (97.8%), car → person (95.6%), person → car (90.1%), car → truck (73.2%)	28.3% (Avg.)	Top 5: bus → person (100%), truck → car (96.5%), bus → car (95.9%), car → person (82.3%), car → truck (77.9%),
YOLO V5	Untargeted	99.6%		99.3%	
	Scenario-targeted	98.2% (Avg.)	OOI → OOU (98.1%), OOU → OOI (100%)	97.1% (Avg.)	OOI → OOU (96.9%), OOU → OOI (99.6%)
	Targeted	34.1% (Avg.)	Top 5: truck → car (97.8%), bus → car (97.2%), traffic light → car (90.3%), person → car (89.2%), person → truck (76.2%)	32.4% (Avg.)	Top 5: bus → person (100%), bus → car (100%), truck → car (92.1%) bus → truck (85.2%) person → car (81.1%)
Faster R-CNN	Untargeted	98.0%		99.4%	
	Scenario-targeted	95.5% (Avg.)	OOI → OOU (95.3%), OOU → OOI (100%)	97.2% (Avg.)	OOI → OOU (96.9%), OOU → OOI (100%)
	Targeted	20.5% (Avg.)	Top 5: truck → car (94.2%), bus → car (92.9%), person → car (75.9%) stop sign → person (75.0%), person → bus (70.1%)	30.6% (Avg.)	Top 5: bus → person (100%), car → person (97.6%), truck → car (97.4%) stop sign → person (95.7%), truck → person (92.3%)
Apollo	Untargeted	67.0%		73.1%	
	Targeted	16.6% (Avg.)	Top 5: truck → car (76.0%), person → car (75.0%), bus → car (68.4%), bus → truck (26.3%), person → truck (25.8%)	18.3% (Avg.)	Top 5: truck → car (75.0%), person → car (70.2%), bus → car (66.7%), truck → bus (25.0%), bus → truck (25.0%)

ATTACK ROBUSTNESS



- Scenes
  - Attack performance across different scenes (city street, highway, residential street) showed no performance loss.
- Weather
  - Different weather conditions (clear, cloudy, rainy) displayed minimal performance loss
- Times of Day
  - Night proved to decrease the performance of CA and AA since darkness has more similar colored pixels
- Camera Resolution
  - Found no performance loss using different quality cameras

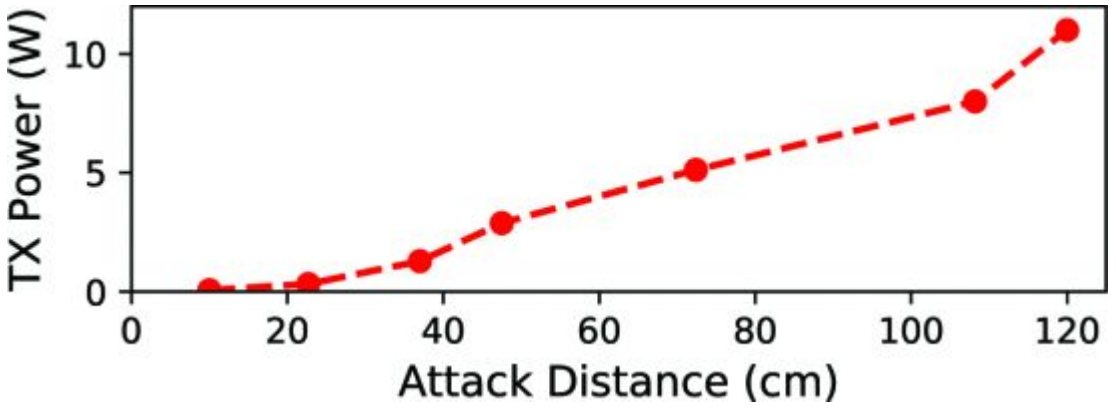
REAL WORLD ATTACK

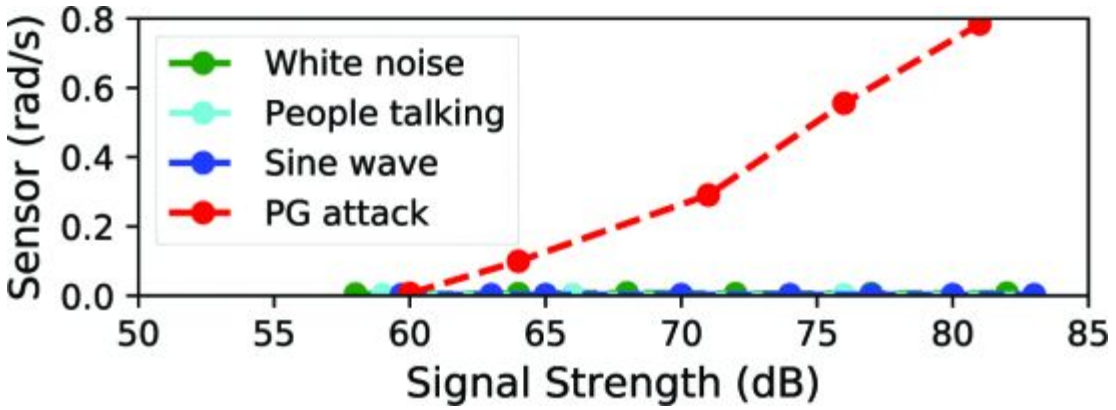
- Used Faster R-CNN

Attacks	Scenes							
	City Lane		City Crossroad		Tunnel		Campus Road	
	Goal	SR	Goal	SR	Goal	SR	Goal	SR
Hiding	hide a "person"	98.1%	hide a "car"	100%	hide a "car"	100%	hide a "car"	95.2%
Creating	create a "truck"	17.1%	create a "bus"	75.7%	create a "truck"	43.9%	create a "person"	37.9%
Altering	alter a "car" into a "bus"	81.4%	alter a "car" into a "boat"	54.4%	alter a "traffic light" into a "person"	15.0%	alter a "car" into a "person"	21.7%

REAL WORLD LIMITATIONS

- A more powerful audio device is needed to conduct the attack from larger distances
  - 10 W is needed to conduct an attack from 1.2 m away
- Other noises could disturb the effectiveness of the attack
  - Minimal interference was found





COUNTERMEASURES

Physical Safeguards

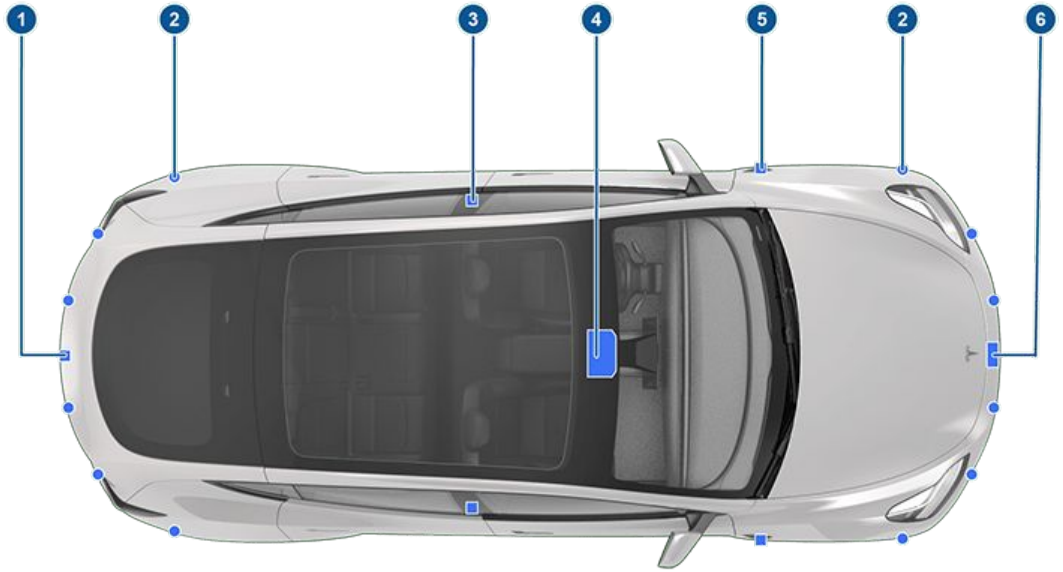
- Surround the inertial sensor with MEMS fabricated acoustic metamaterial
  - Reduces susceptibility of the inertial system to resonant acoustic signals
- Secure a low-pass filter to eliminate out-of-band analog signals
  - Reduces adversaries' ability of controlling the sensor output via signal aliasing
- Attach a microphone to the sensor which can detect acoustic injection, alerting the system to a potential attack

Software Safeguards

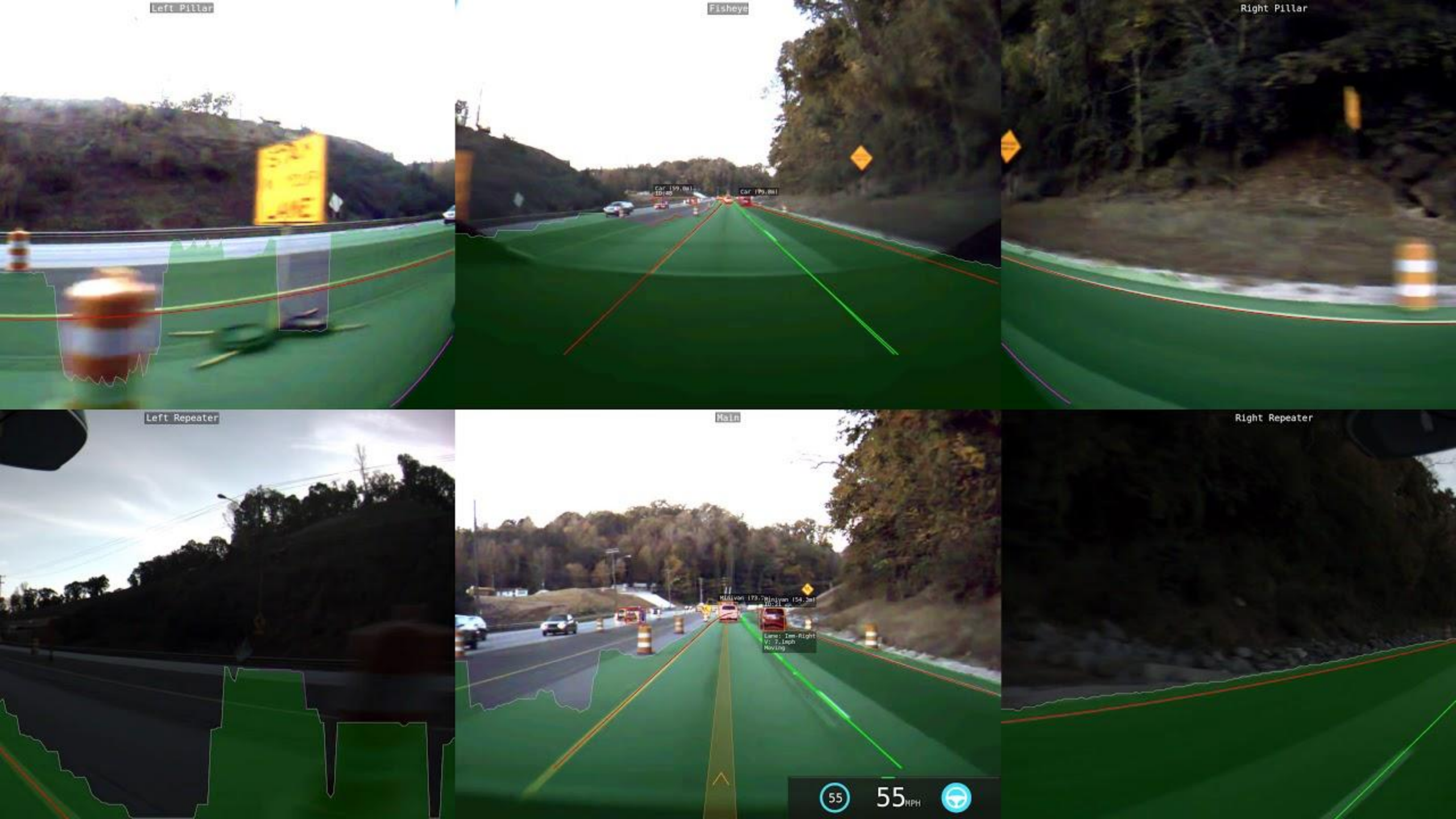
- Digital stabilization by de-blurring images
- Sensor Fusion
  - More cameras, LiDAR, radars
- Object Detection Algorithms
  - Remove adversarial blur patterns via a guided de-noiser
  - Improve detection models by increasing detection criteria

QUESTIONS??

DISCUSSION



- How would you conduct a poltergeist attack against this Tesla Model Y? Is it feasible at all?
  - Front facing Autopilot cameras are located at 3 (one on each B pillar) and 4 (3 cameras on the rear-view mirror)
  - Rear racing cameras are located at 5 (one on each side fender) and 1 (above license plate)


Left Pillar
Fisheye
Right Pillar
Left Repeater
Main
Right Repeater
55
55 MPH

THANK YOU