That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency

Yanmao Man[#], Raymond Muller[§], Ming Li[#], Z. Berkey Celik[§], Ryan Gerdes[‡]
[#]University of Arizona [§]Purdue University [‡]Virginia Tech

Autonomous Systems

Perception Security

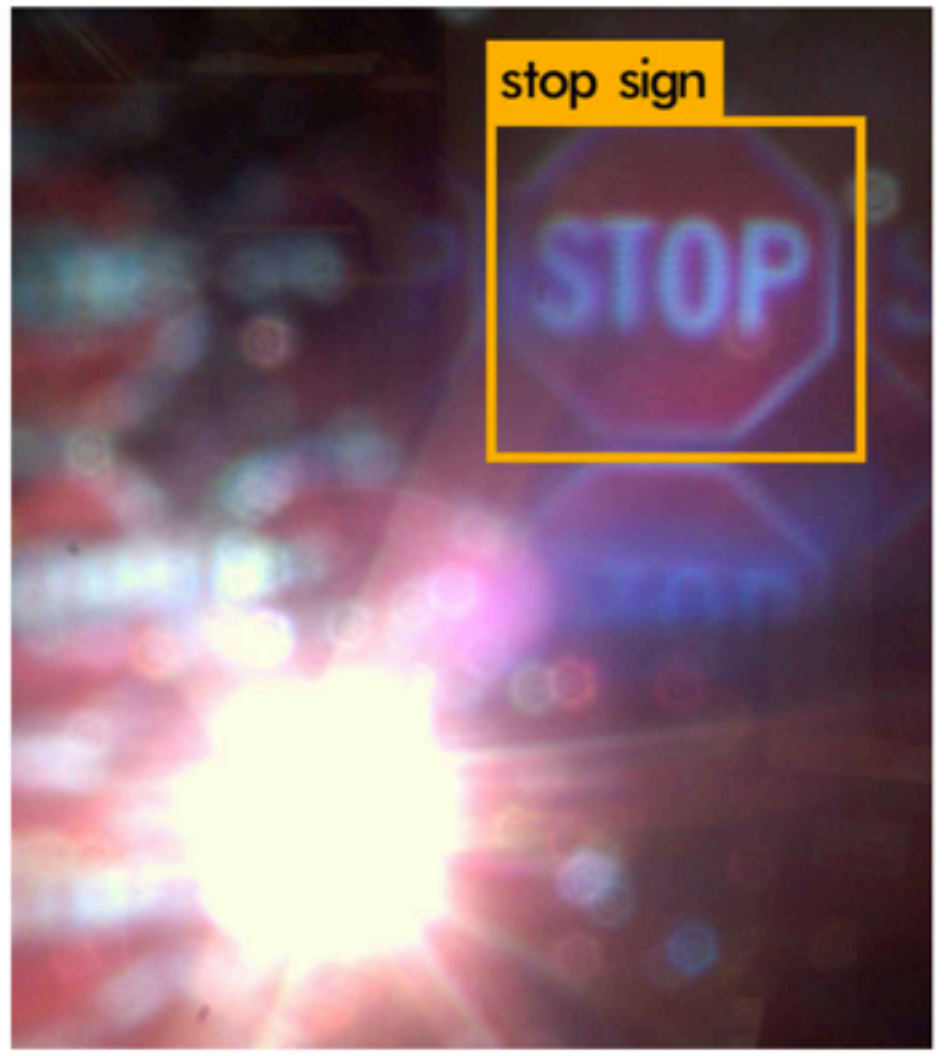
Environment → Perception —Data→ Prediction → Planning → Control —Command→ Actuation → Environment

Physically-realizable Perception Attacks

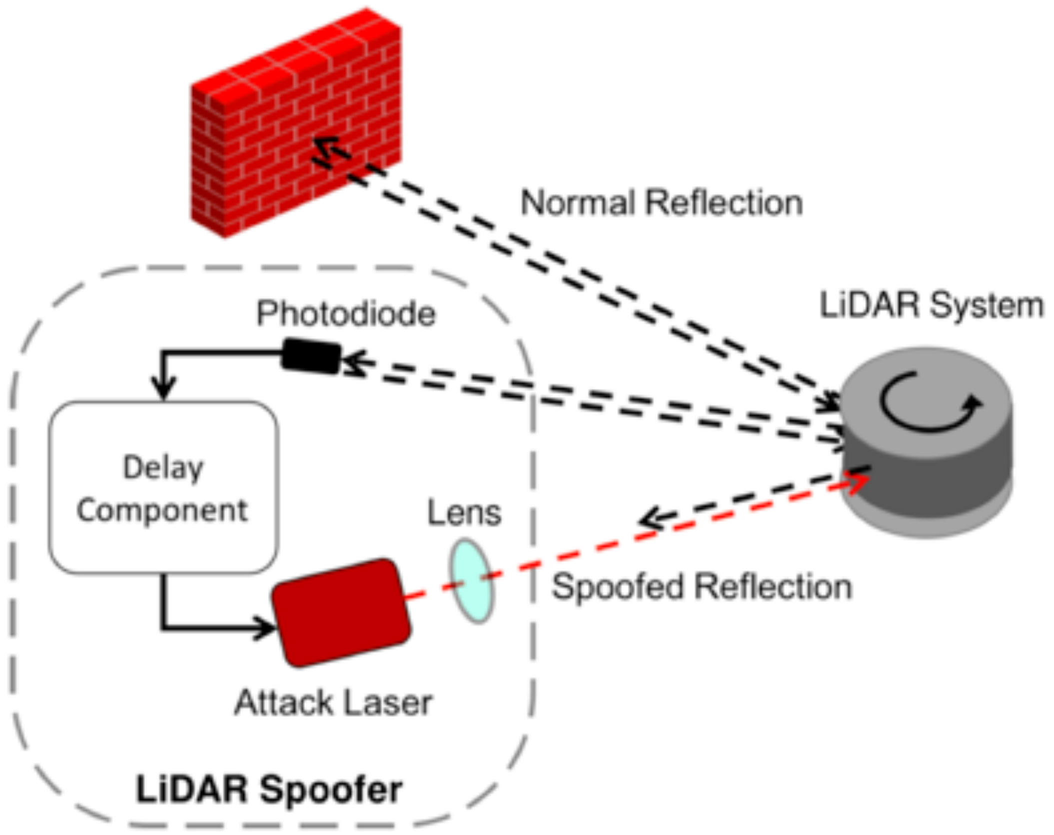
These perception attacks alter the data at the source, hence bypassing traditional digital defenses

Stop Sign Sticker

Phantom Attack



GhostImage Camera Attack



LiDAR Spoofing

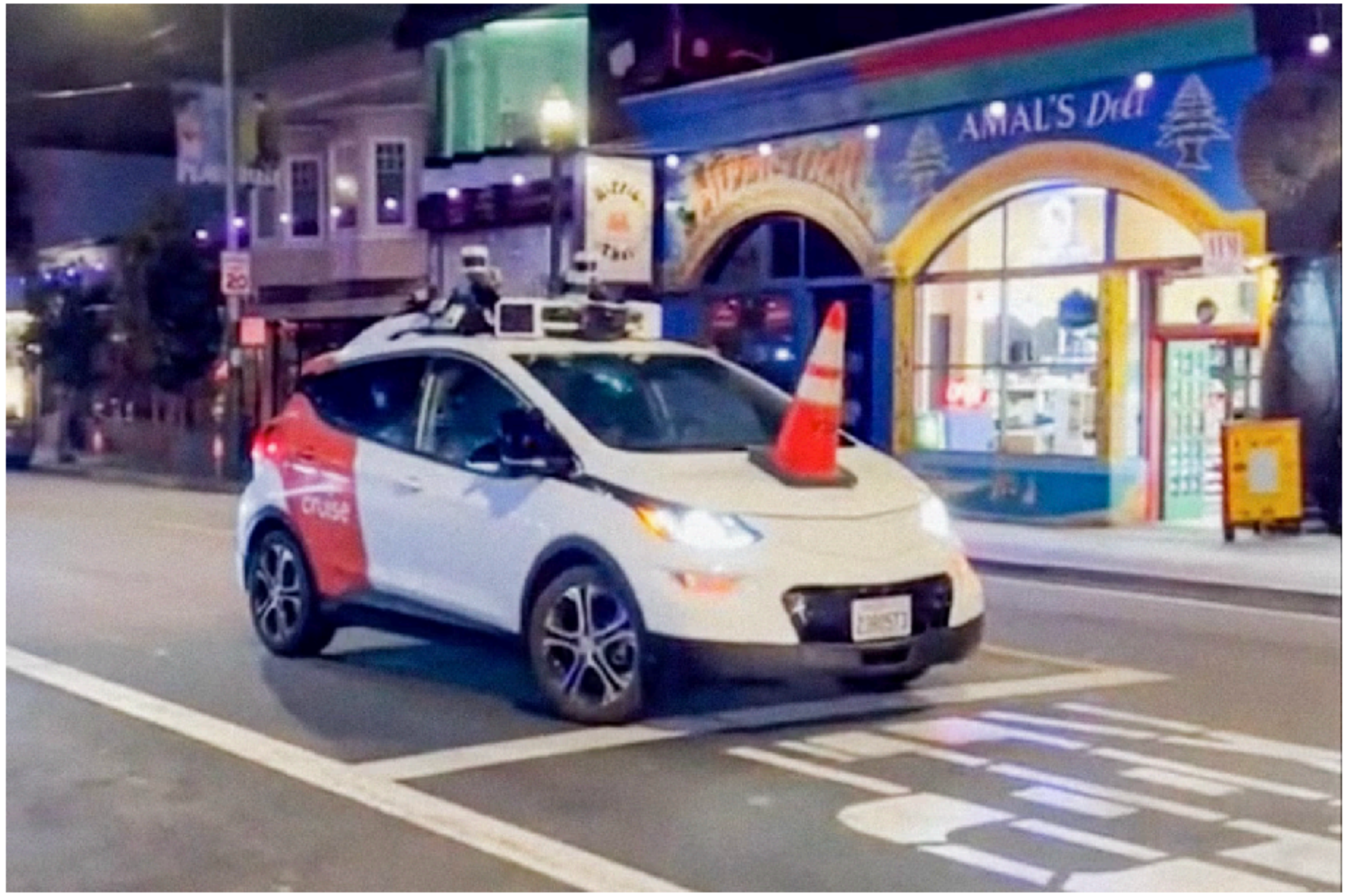
1. Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." *CVPR 2018*.
2. Nassi, Ben, et al. "Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks." CCS 2020
3. Man, Yanmao et al."GhostImage: Remote perception attacks against camera-based image classification systems." RAID 2020
4. Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." CCS 2019.

The Self-Driving Cars Wearing a Cone of Shame

There's a brilliant activist campaign to stop San Francisco's autonomous vehicles in their tracks.

BY ALISON GRISWOLD

JULY 11, 2023 • 10:45 AM

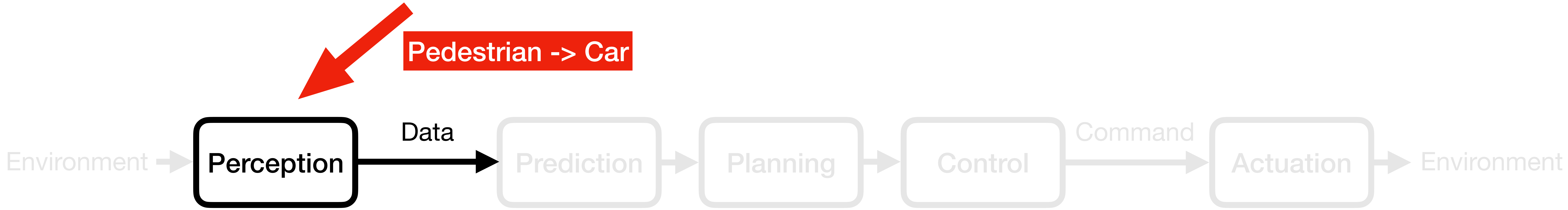


It looks like a sad unicorn (which, in a way, it is). Screengrab from TikTok/Safe Street Rebel

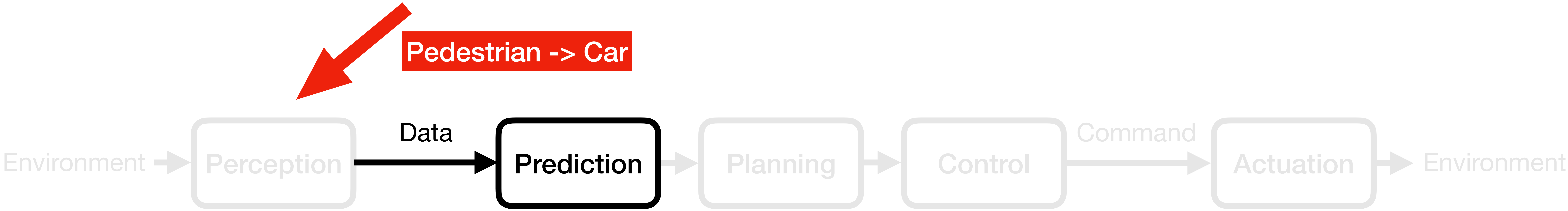
slate.com

Perception Security

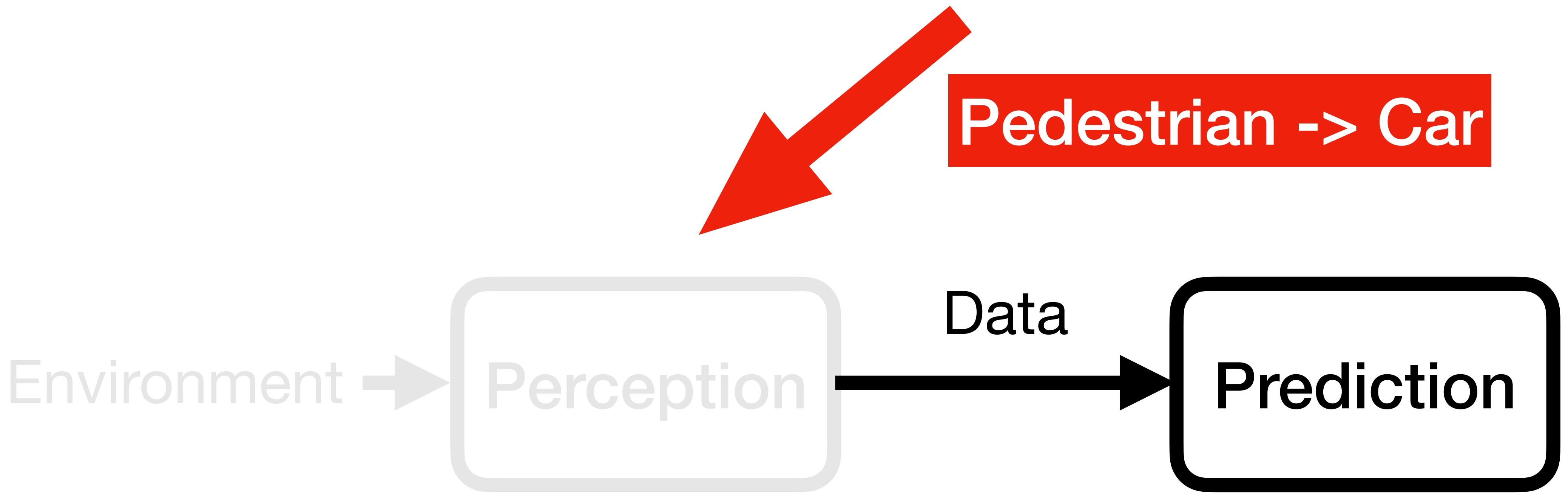
Misclassification Attacks



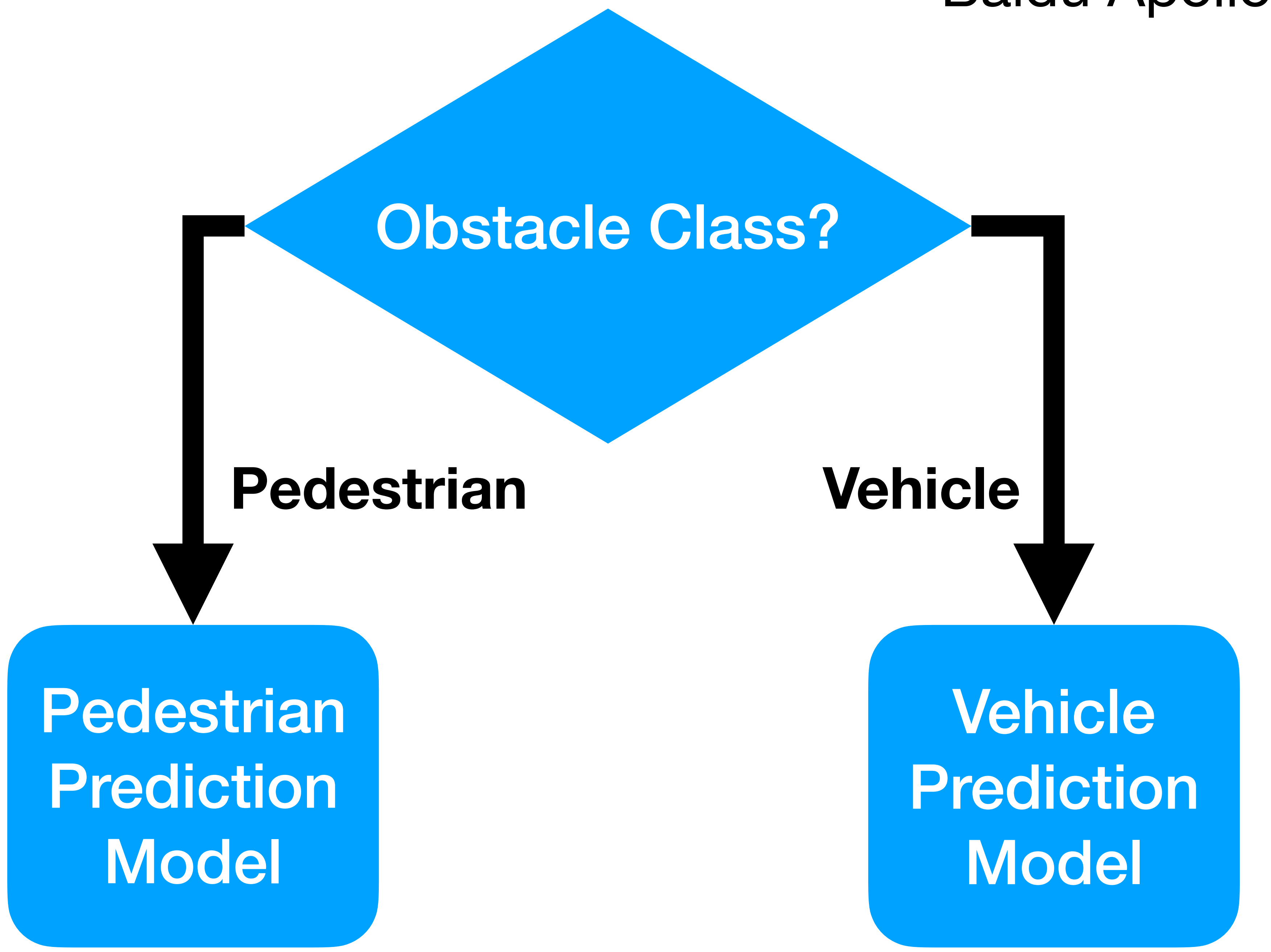
Misclassification Attacks



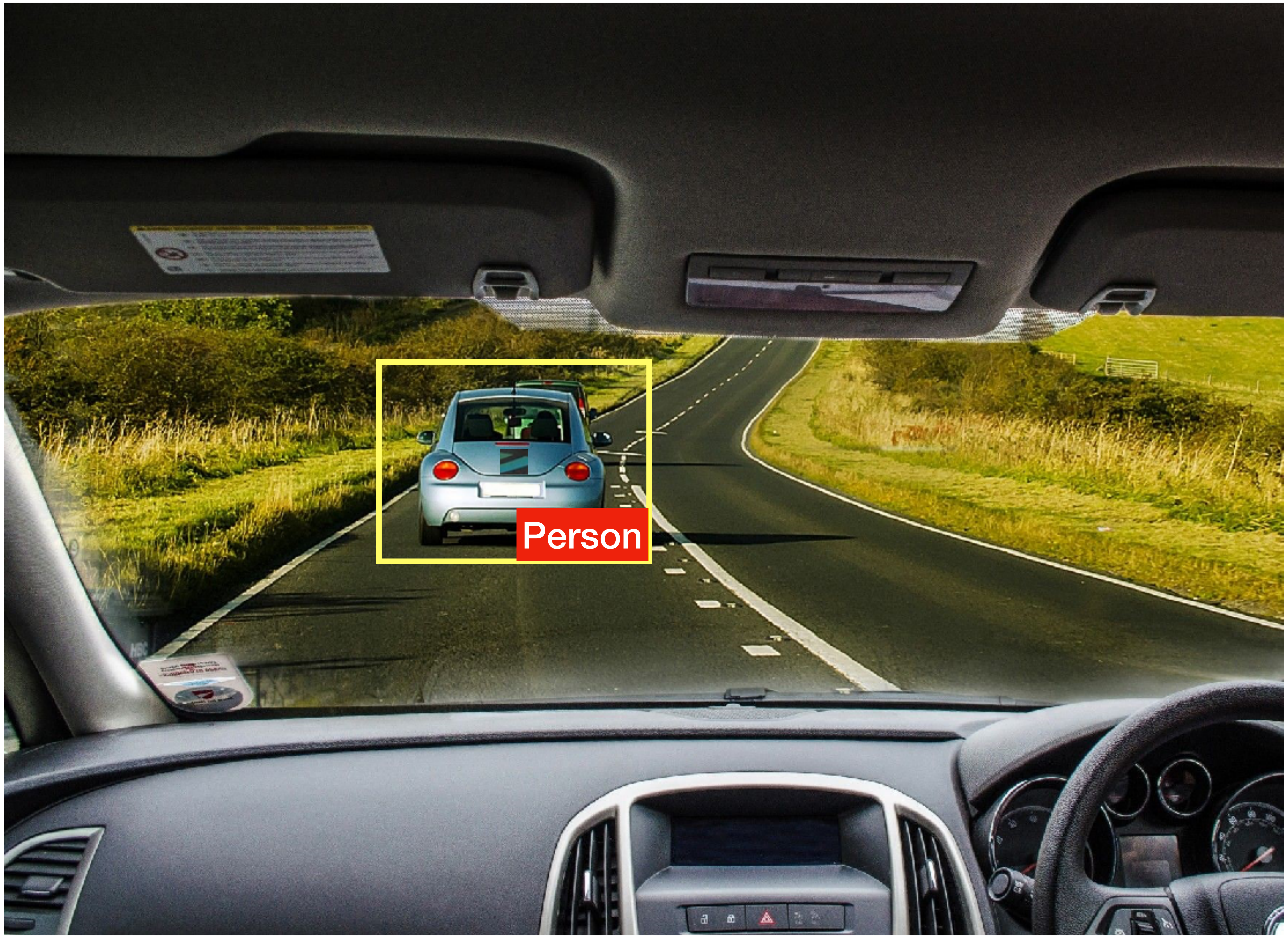
Misclassification Attacks



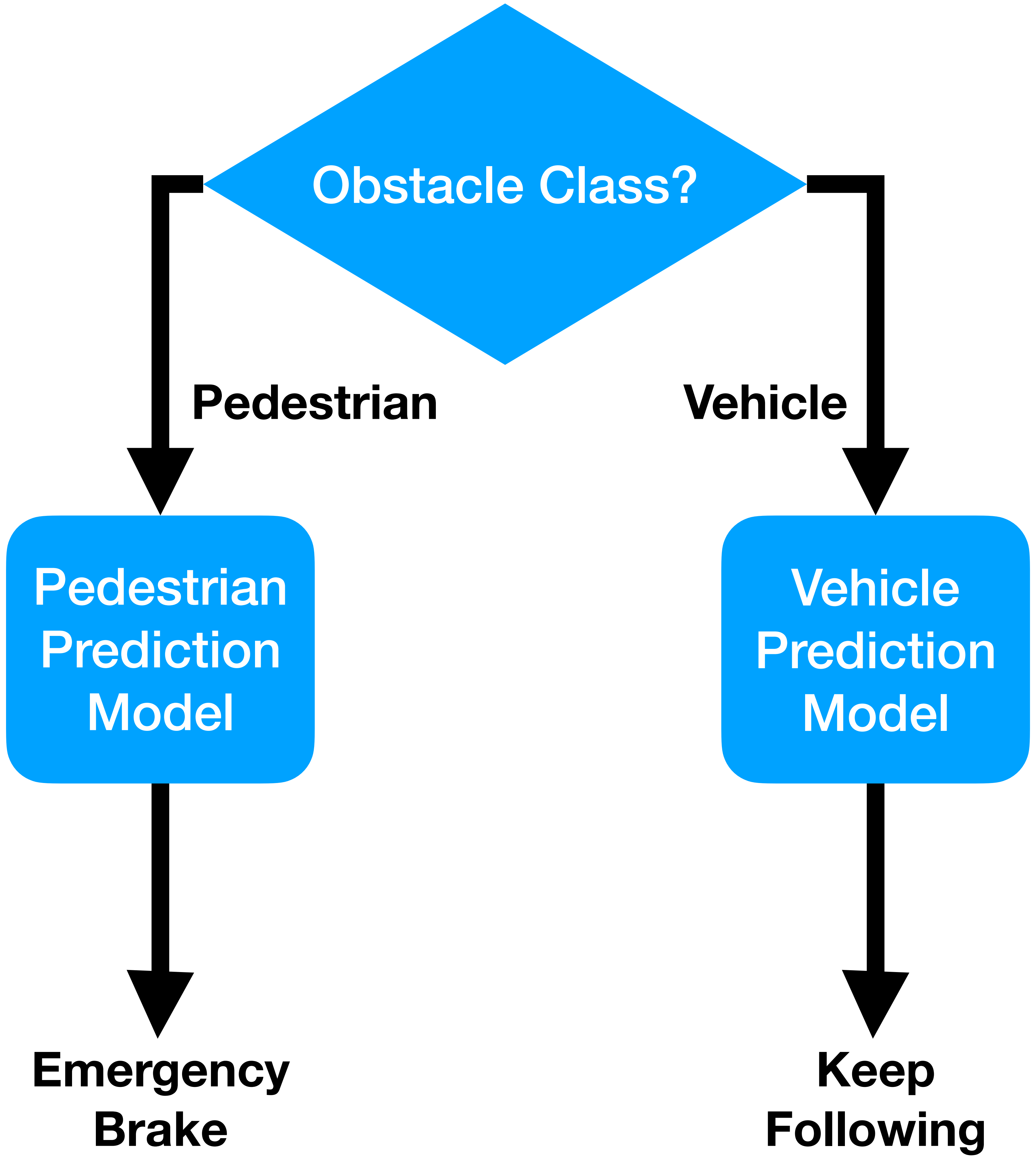
*Baidu Apollo



Misclassification Attacks



https://toocooltrafficschool.com/following-distance/



[5] Man, Yanmao, et al. "Evaluating perception attacks on prediction and planning of autonomous vehicles." *USENIX Security Symposium Poster Session*. 2022.

Misclassification Attacks

PercepGuard aims to detect misclassification attacks

Existing defenses against perception attacks are either

- Specific to some sensing modality
  - LiDARs
  - GPS
  - IMU
- Specific to some attack methodology
  - Adversarial Patch
  - Norm-bounded

PercepGuard aims to detect misclassification attacks

Agnostic to

- Attack methodologies
- Object detection and tracking algorithms

PercepGuard aims to detect misclassification attacks by verifying the spatiotemporal consistency of the perception result

Spatio-temporal Consistency

t=1

Spatio-temporal Consistency

t=2

Spatio-temporal Consistency

t=3

Spatio-temporal Consistency

t=3

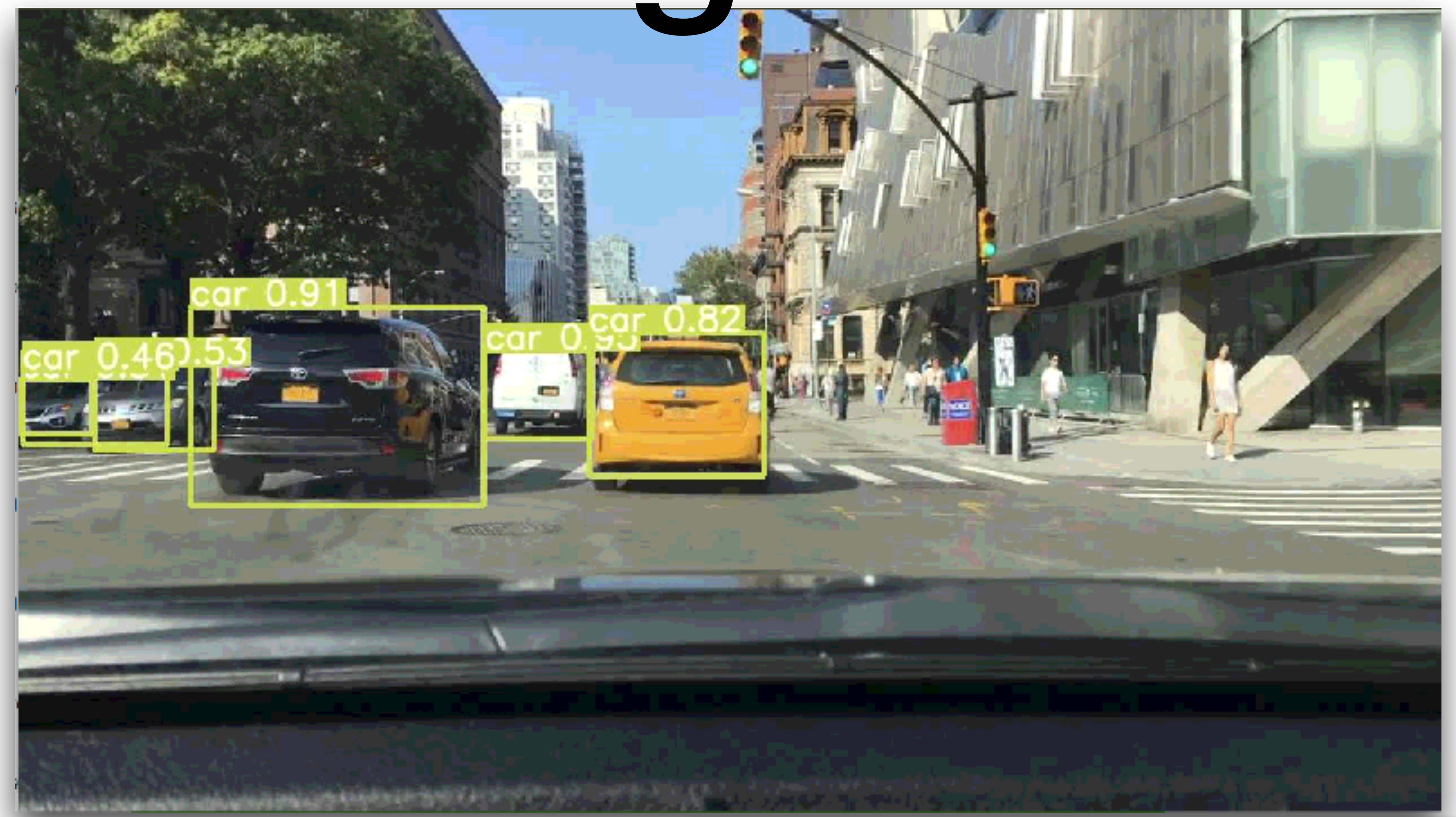
Spatio-temporal Consistency

Spatio-temporal Consistency

Research Questions:

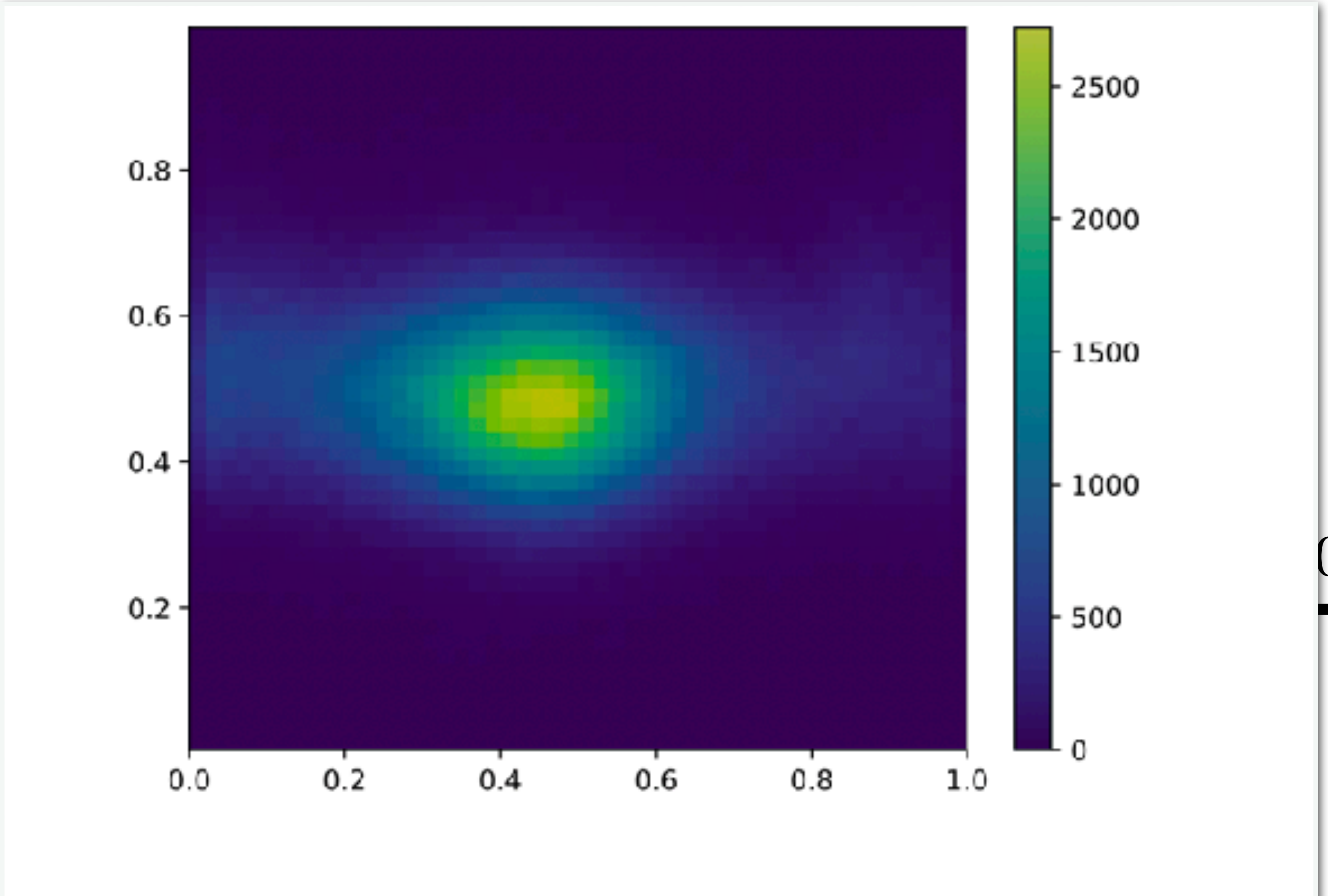
- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

PercepGuard Design

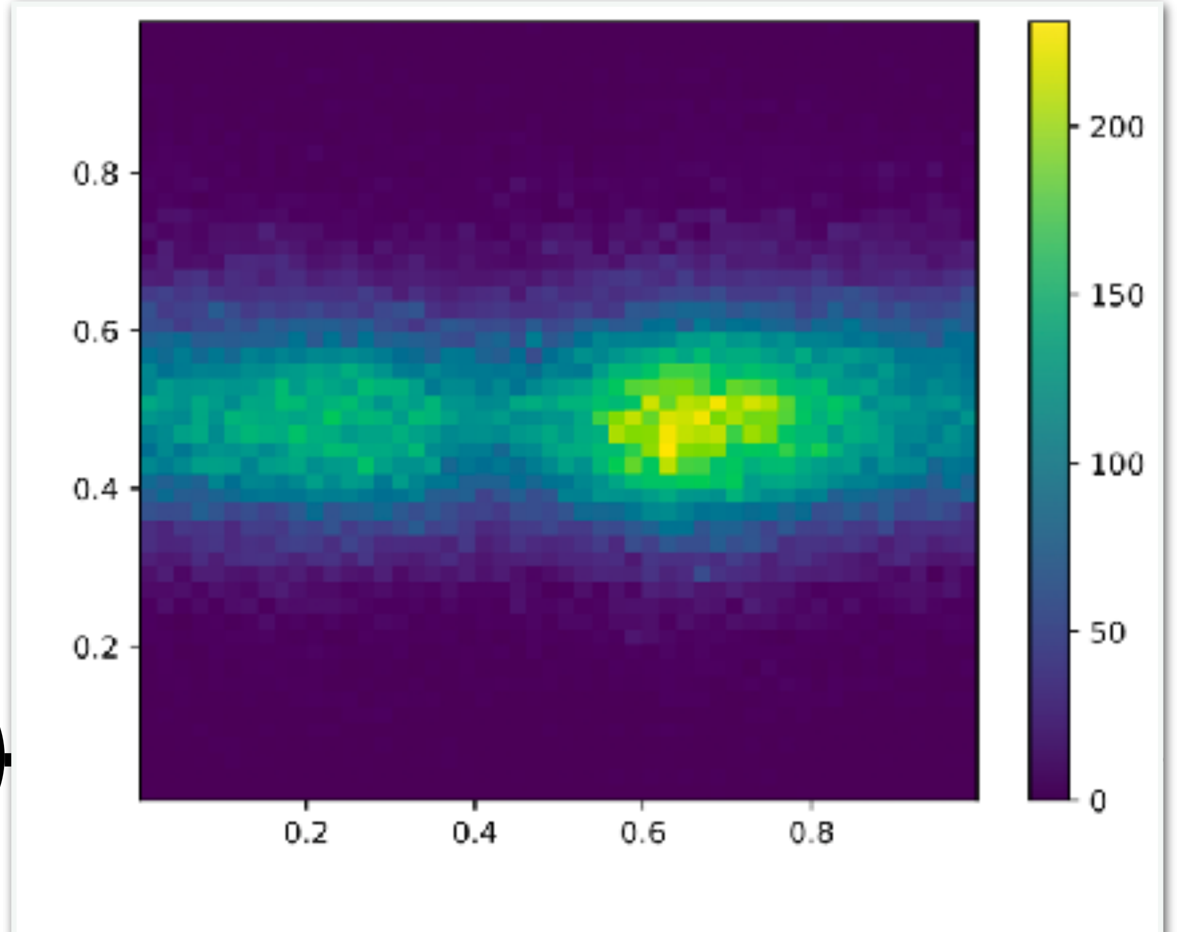


Research Questions:

- Do bounding boxes provide statistically-sufficient information?

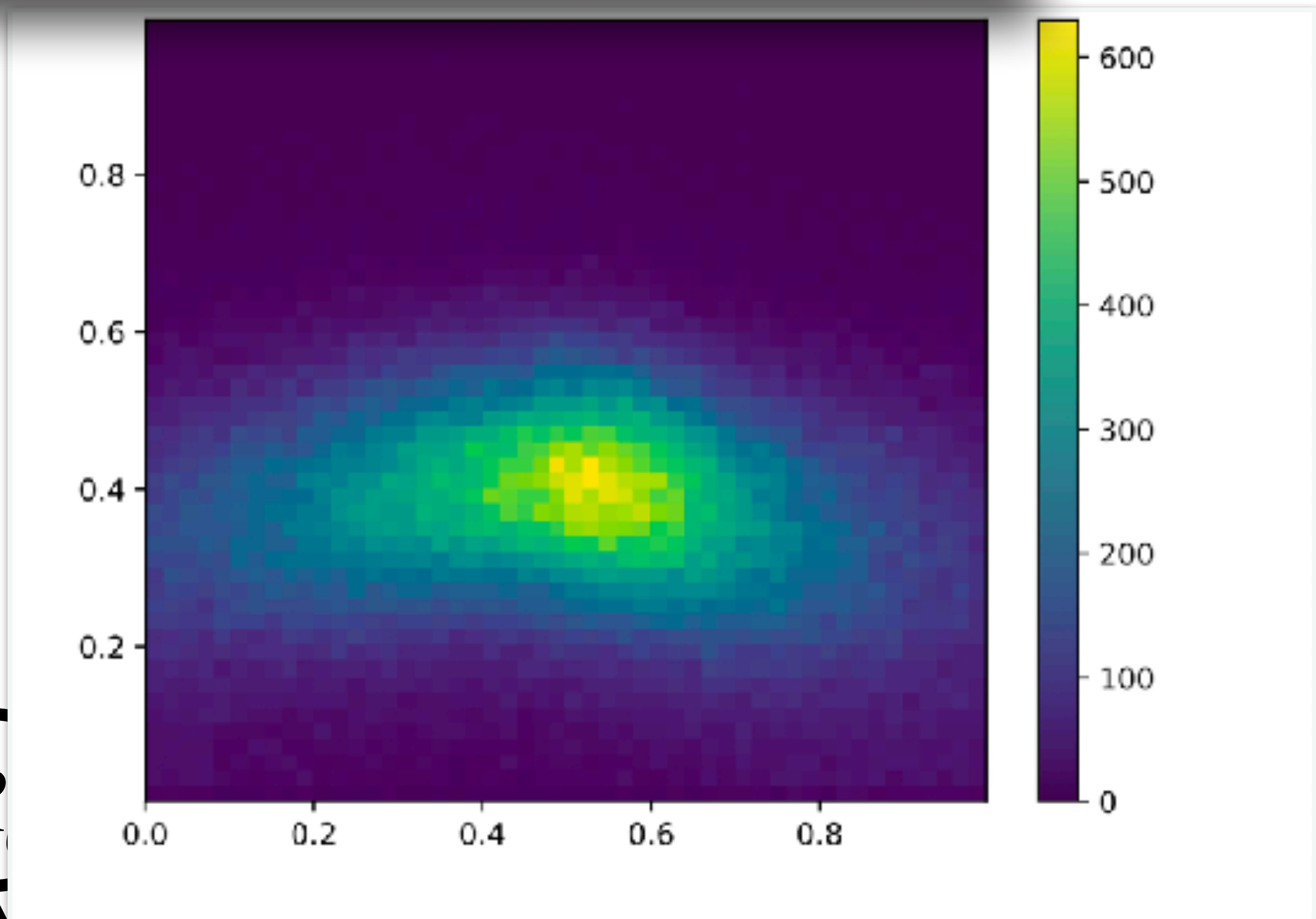


(a) Ground truth: Histogram of cars' locations.



(a) Ground truth: Histogram of locati

Figure 2: Persons' locations.



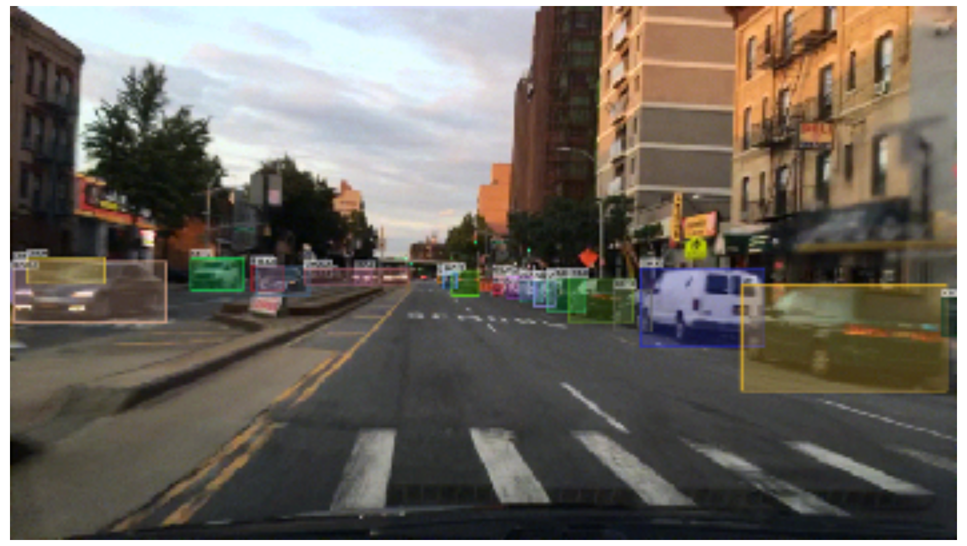
(a) Ground truth: Histogram of locations.

Figure 3: Traffic Signs' locations.

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

Evaluation



Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

Dataset:

- Berkeley Driving Dataset (BDD)
- Five object classes:
  - bike, bus, car, pedestrian, truck

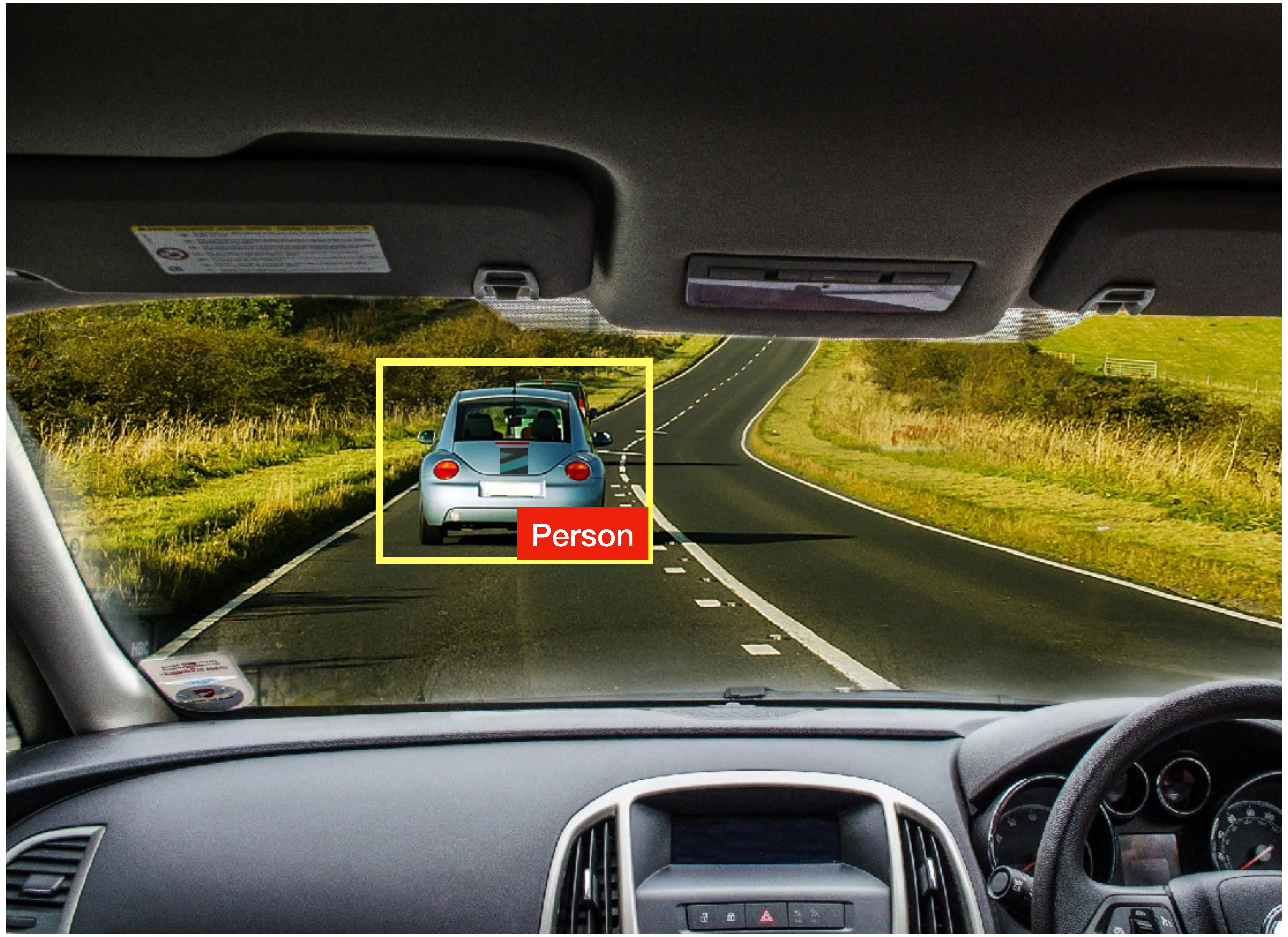
Classification Accuracy: 95%
False Negative Rate: 5%

Evaluation

$$\underset{\Delta}{\text{minimize}} \quad \|\Delta\| \qquad \text{such that} \quad \bar{c} = c''$$

"person"

YOLO's classification result

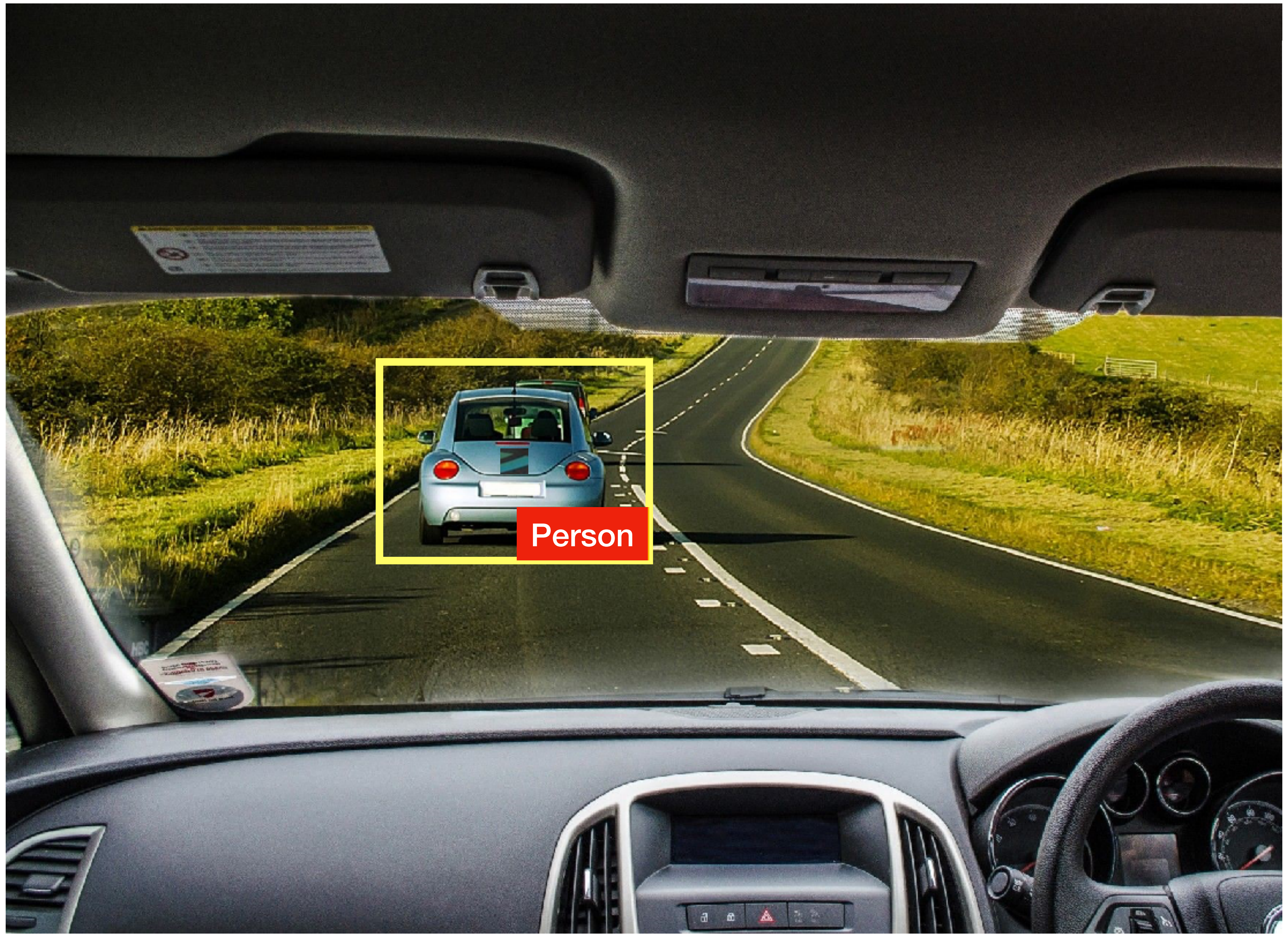


Attack Model:

- *Attack Goal*: Causing the rear car to recognize the front car as a person thus decides to stop (e.g., on a highway)
- *Attacker's Capability*: They utilize the adversarial machine learning to generate adversarial patches with white-box knowledge of the object detection algorithm (e.g., YOLO)

True Positive Rate: ??
False Negative Rate: 5%

Evaluation



$$\underset{\Delta}{\text{minimize}} \quad \|\Delta\| \qquad \text{such that} \quad \bar{c} = c''$$

"person" → $\bar{c} = c'$ ← Our classification result

Table I: Adversarial patch attacks with BDD100K

Attack Type	Patch Size	A.M.R.	T.P.R.	A.S.R.
Defense-unaware	20×20	83.47%	99.63%	0.3%
	40×40	89.41%	100%	0%
	60×60	92.94%	100%	0%

But, what about adaptive attackers, who are aware of our defense and try to evade it?

True Positive Rate: Above 99%

False Negative Rate: 5%

Evaluation

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

$$\underset{\Delta}{\text{minimize}} \quad ||\Delta|| \qquad \text{such that} \quad \bar{c} = c'' \\ \bar{c} = c'$$

Table I: Adversarial patch attacks with BDD100K

Attack Type	Patch Size	A.M.R.	T.P.R.	A.S.R.
Defense-unaware	20×20	83.47%	99.63%	0.3%
	40×40	89.41%	100%	0%
	60×60	92.94%	100%	0%

But, what about adaptive attackers, who are aware of our defense and try to evade it?

True Positive Rate: Above 99%

False Negative Rate: 5%

Evaluation

$$\underset{\Delta}{\text{minimize}} \quad \|\Delta\| \qquad \text{such that} \quad \bar{c} = c'' \\ \bar{c} = c'$$

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

Table I: Adversarial patch attacks with BDD100K

Attack Type	Patch Size	A.M.R.	T.P.R.	A.S.R.
Defense-unaware	20×20	83.47%	99.63%	0.3%
	40×40	89.41%	100%	0%
	60×60	92.94%	100%	0%

True Positive Rate: Above 99%

False Negative Rate: 5%

Evaluation

$$\underset{\Delta}{\text{minimize}} \quad \|\Delta\| \qquad \text{such that} \quad \bar{c} = c'' \\ \bar{c} = c'$$

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

Table I: Adversarial patch attacks with BDD100K

Attack Type	Patch Size	A.M.R.	T.P.R.	A.S.R.
Defense-unaware	20 × 20	83.47%	99.63%	0.3%
	40 × 40	89.41%	100%	0%
	60 × 60	92.94%	100%	0%
Defense-aware	20 × 20	73.25%	98.74%	0.92%
	40 × 40	80.49%	90.33%	7.78%
	60 × 60	87.6%	85.67%	12.55%

True Positive Rate: Above 99%

False Negative Rate: 5%

Evaluation

$$\underset{\Delta}{\text{minimize}} \quad \|\Delta\| \qquad \text{such that} \quad \bar{c} = c'' \\ \bar{c} = c'$$

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

Table I: Adversarial patch attacks with BDD100K

Attack Type	Patch Size	A.M.R.	T.P.R.	A.S.R.
Defense-unaware	20 × 20	83.47%	99.63%	0.3%
	40 × 40	89.41%	100%	0%
	60 × 60	92.94%	100%	0%
Defense-aware	20 × 20	73.25%	98.74%	0.92%
	40 × 40	80.49%	90.33%	7.78%
	60 × 60	87.6%	85.67%	12.55%

True Positive Rate: Above ~~99%~~ 85%

False Negative Rate: 5%

Evaluation

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

A sequence of bounding boxes $\xrightarrow{\textbf{RNN}}$ **Object Class**

+

Contextual Information

- Ego-vehicle velocity
- Relative velocity to the object

Defense-aware	20×20	73.25%	98.74%	0.92%
	40×40	80.49%	90.33%	7.78%
	60×60	87.6%	85.67%	12.55%
with contexts	60×60	88.76%	99.35%	0.6%

True Positive Rate: Above ~~99%~~ 85%

False Negative Rate: 5%

Evaluation

Research Questions:

- Do bounding boxes provide statistically-sufficient information?
- Does the detection algorithm produce low false positive and negative rates?
- Is it robust against adaptive attackers?

A sequence of bounding boxes $\xrightarrow{\textbf{RNN}}$ **Object Class**

+

Contextual Information

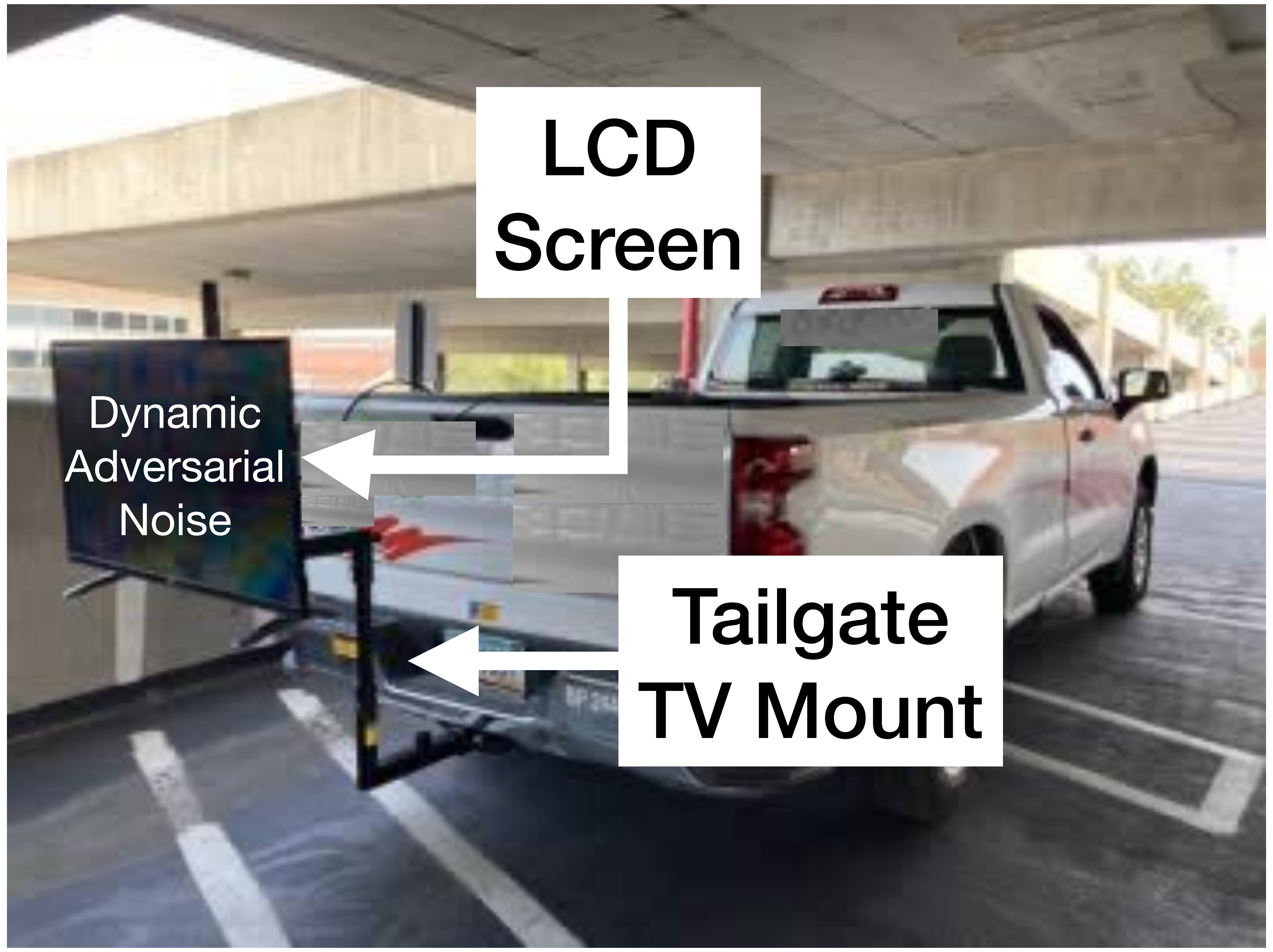
- Ego-vehicle velocity
- Relative velocity to the object

Defense-aware	20×20	73.25%	98.74%	0.92%
	40×40	80.49%	90.33%	7.78%
	60×60	87.6%	85.67%	12.55%
with contexts	60×60	88.76%	99.35%	0.6%

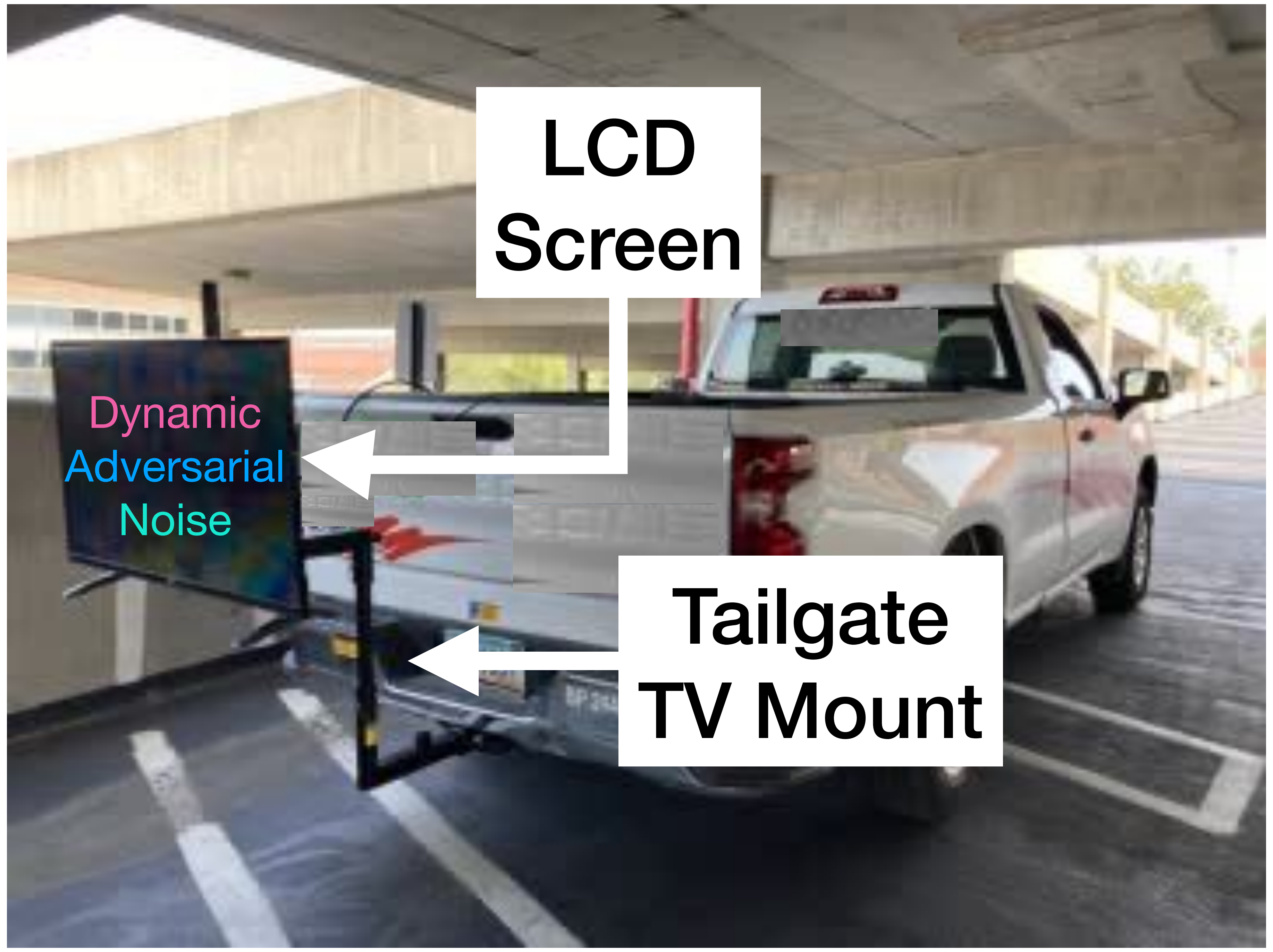
True Positive Rate: Above ~~99%~~ ~~85%~~

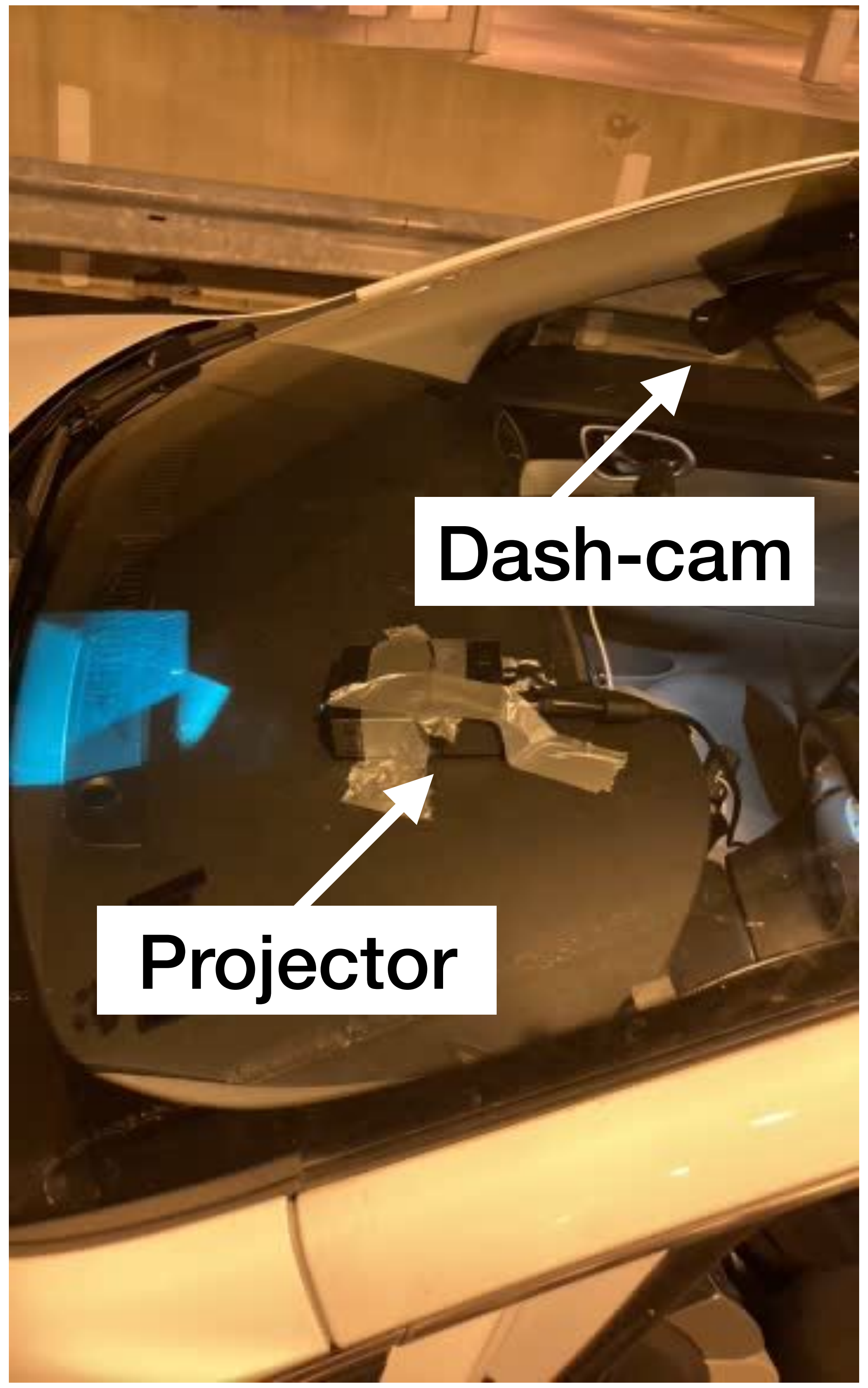
False Negative Rate: 5% 99%

Real-world Experiments

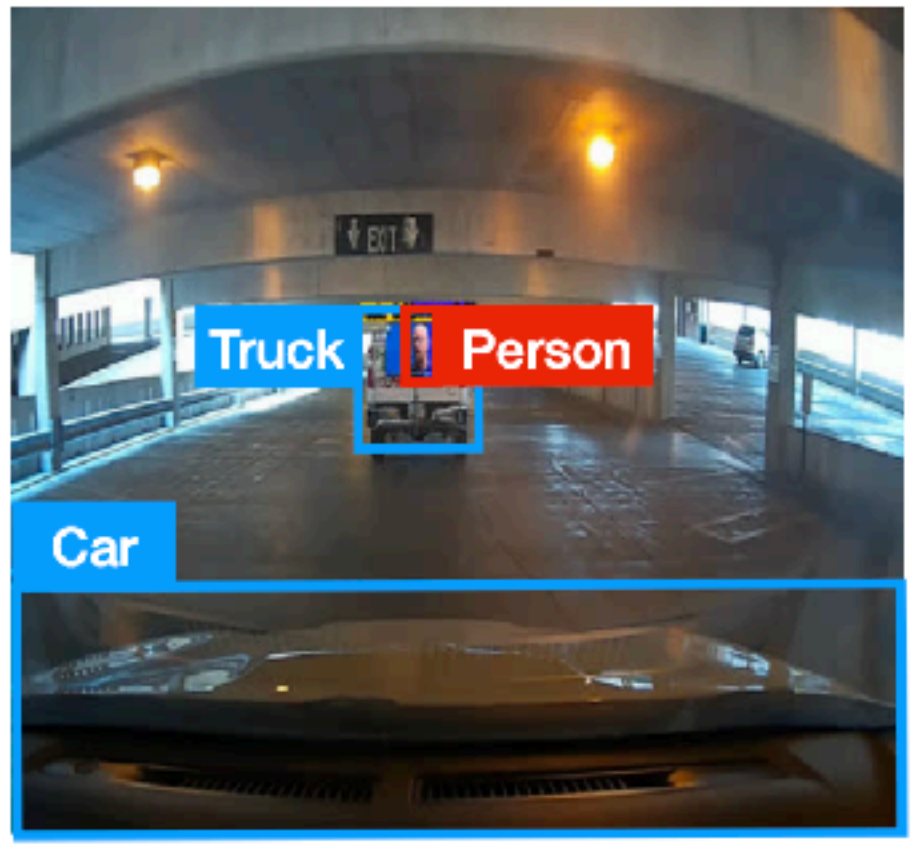


Real-world Experiments



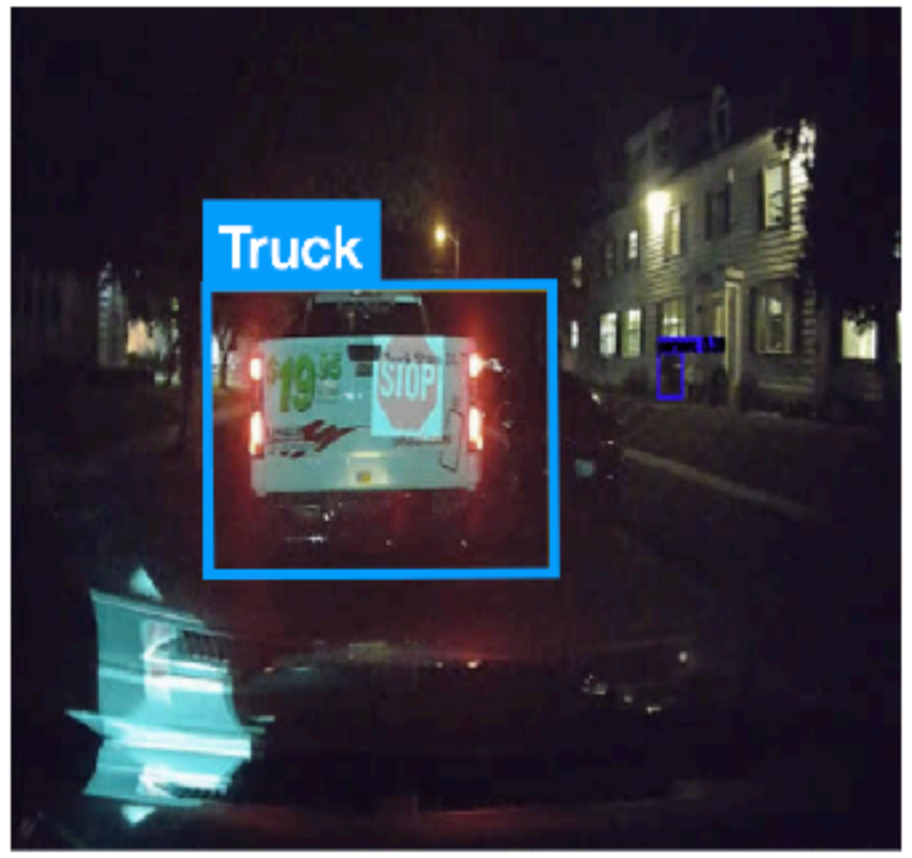


Real-world Experiments

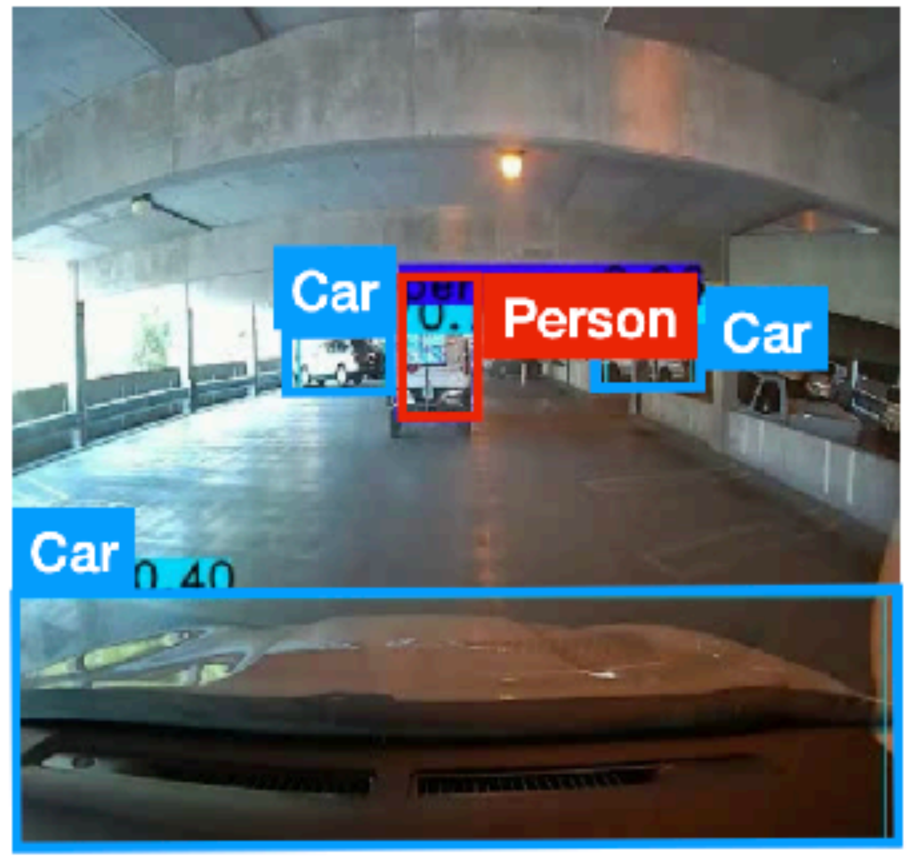


(a) Person

(b) Stop sign on monitor



(c) Projected stop sign



(d) Adversarial patch

Table 4: Real image attacks in the real-world

Real Images of	Device	ARR	TPR	ASR
People	Monitor	63.2%	83.3%	10.6%
	Projector	58.8%	100%	0%
Stop Signs	Monitor	40.0%	100%	0%
	Projector	20.0%	100%	0%

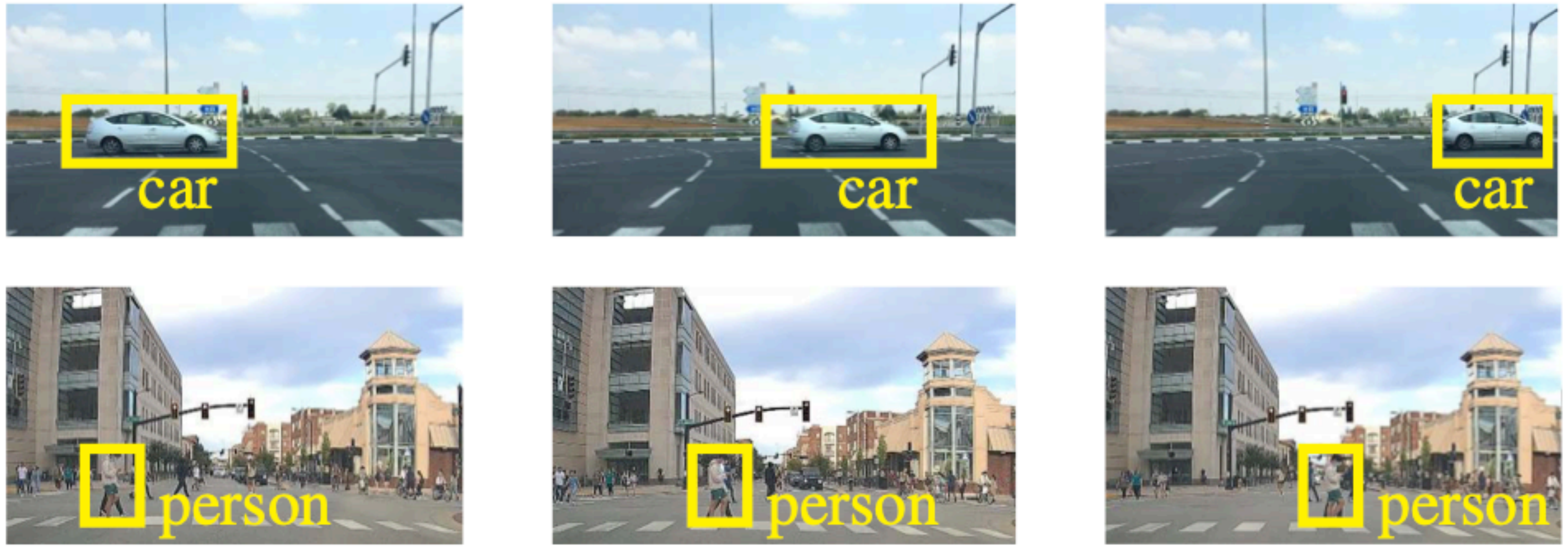
Table 5: Adversarial patch attacks in the real-world

Attack Type	Device	AMR	TPR	ASR
Defense-unaware	Monitor	52.2%	100%	0%
	Projector	27.3%	100%	0%
Defense-aware	Monitor	45.5%	100%	0%
	Projector	20.0%	100%	0%

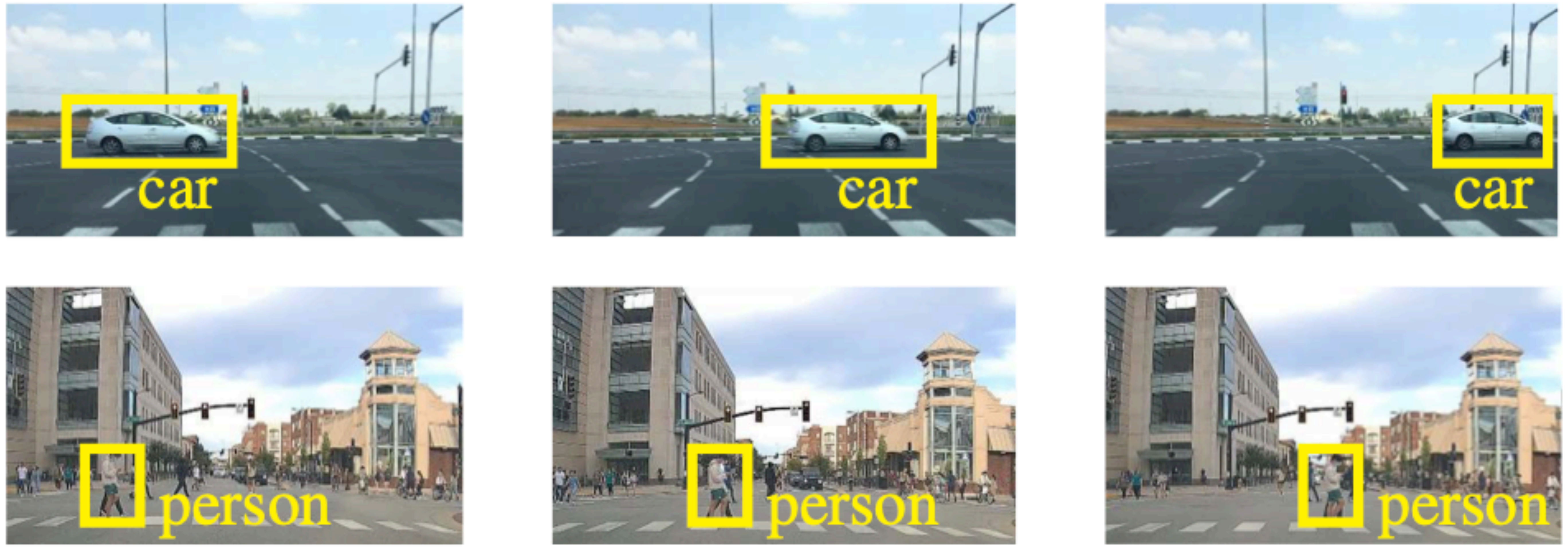
More Evaluation

- Baseline comparison
- Sensitivity analysis
- Alternative operating points
- Additional features

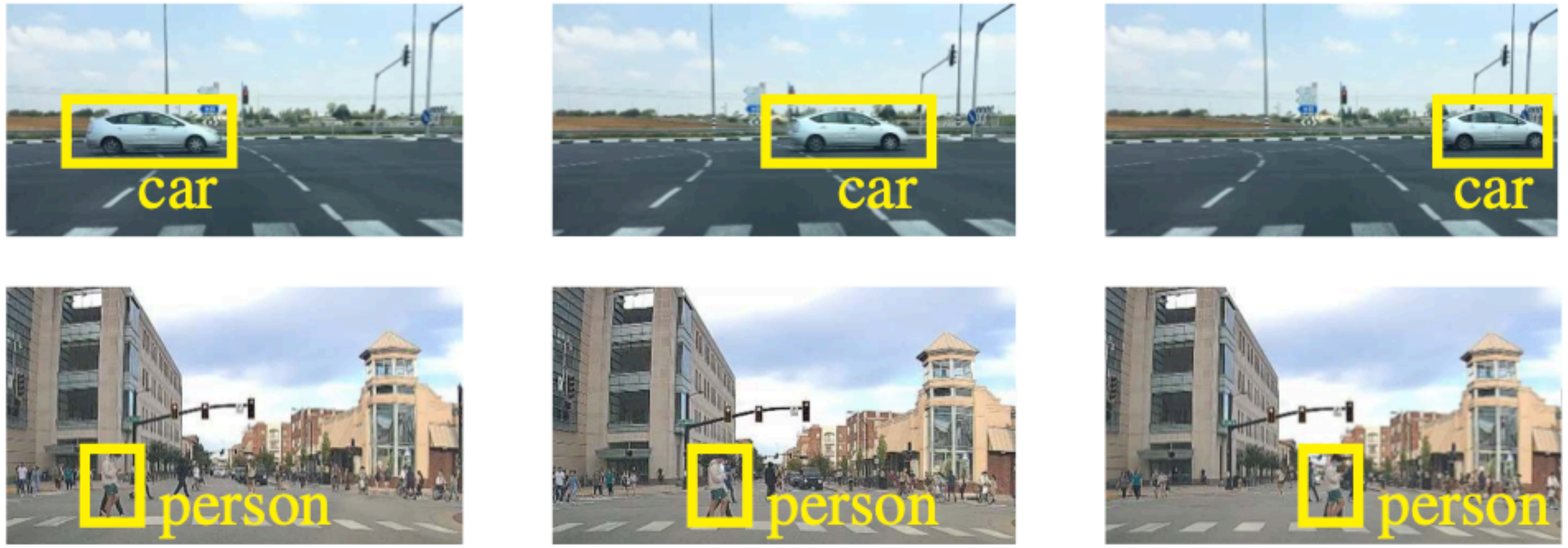
That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency



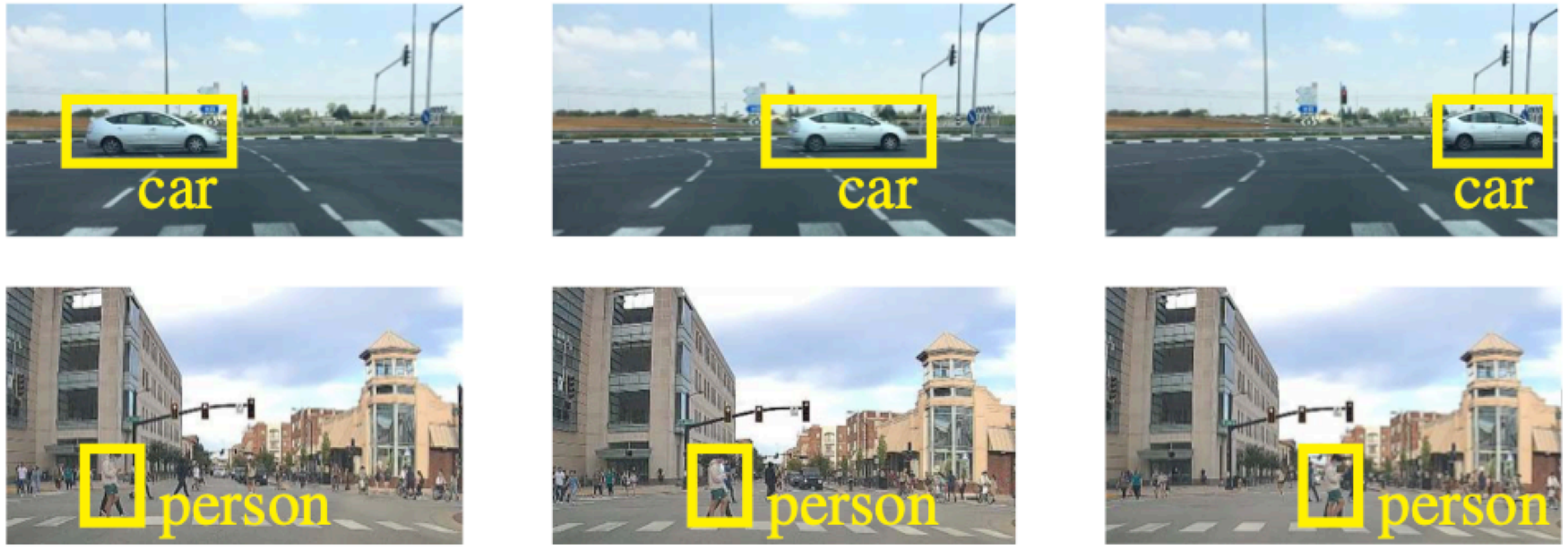
That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency



That Person Moves Like A Car:
Misclassification Attack Detection
for Autonomous Systems
Using Spatiotemporal Consistency

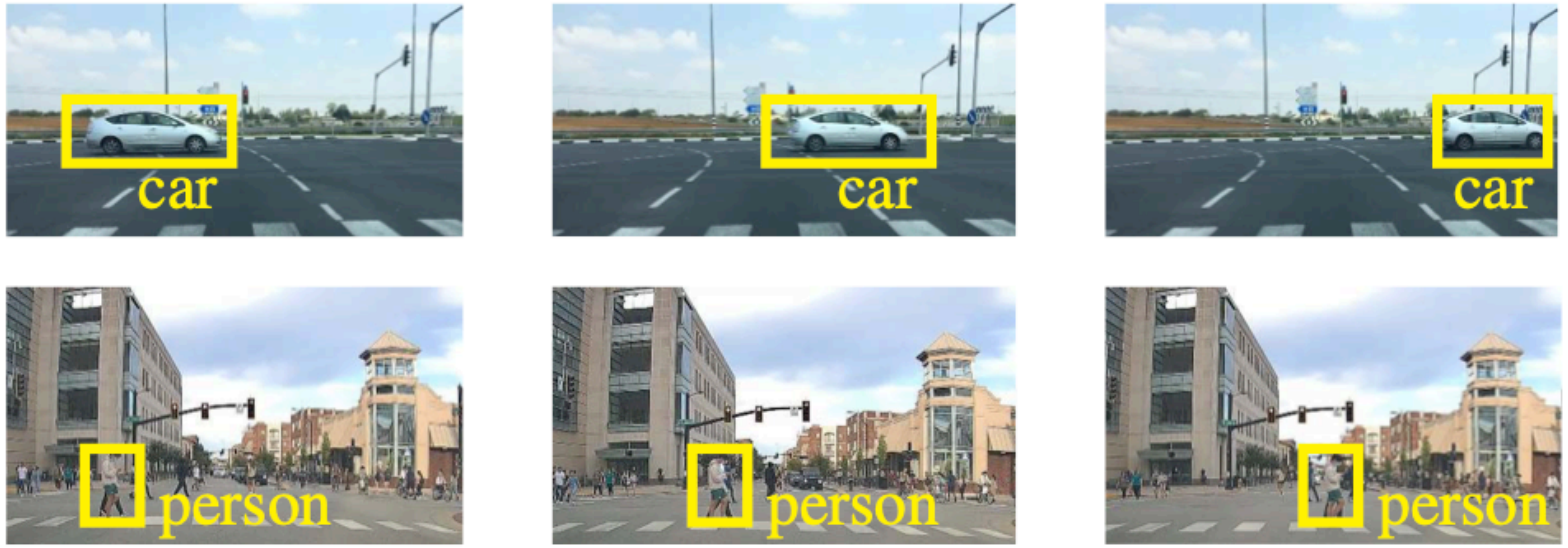


That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using **Spatiotemporal Consistency**



- Adaptive Attacks
- Contextual information
- True positive rate: above 99%
- False negative rate: 5%

Future Work



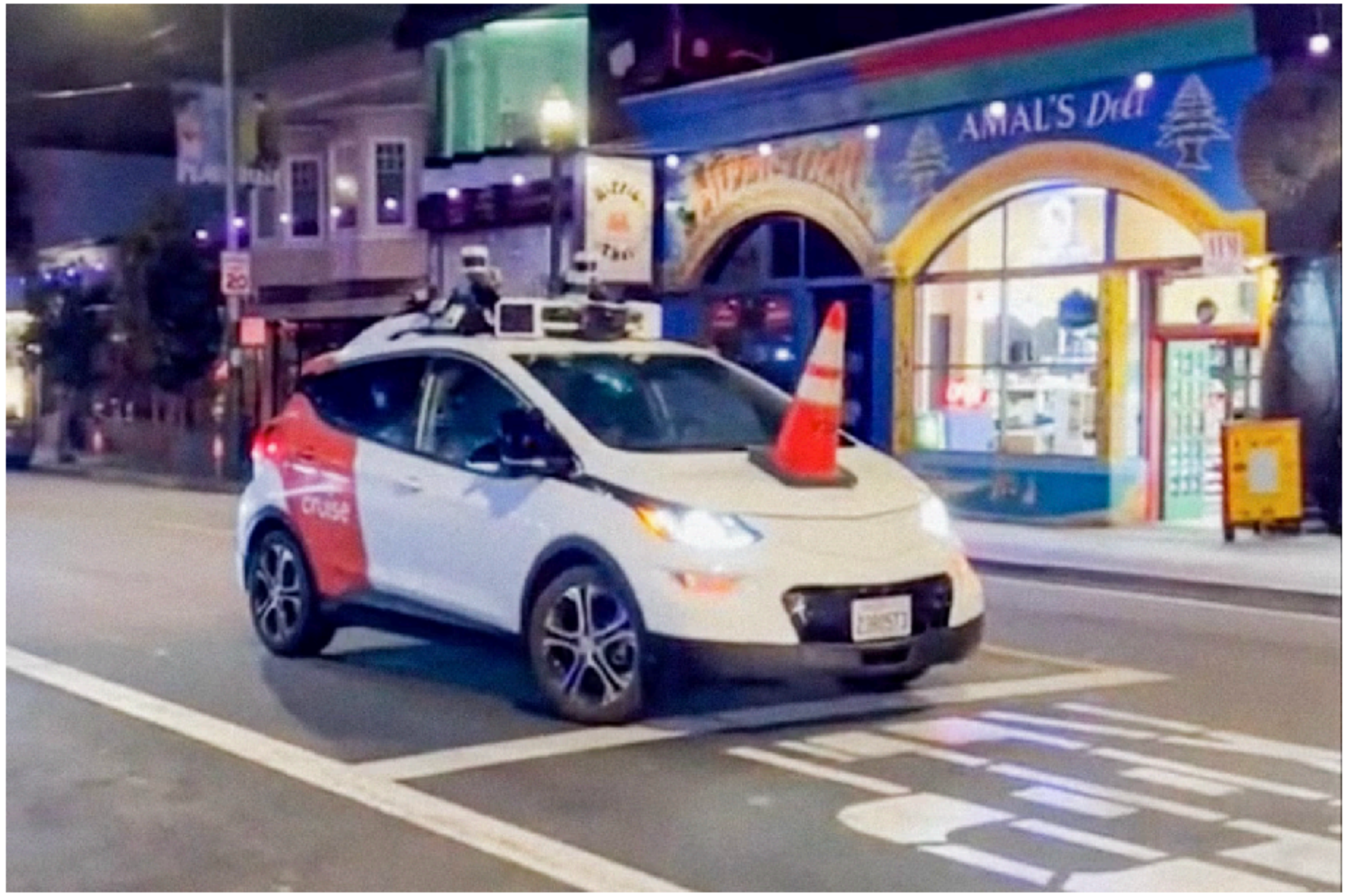
- More spatiotemporal features
  - Different Sensors
  - Semantic Segmentation
- Detecting object creation attacks
- Attention-based
- Sensor Configuration Randomization

The Self-Driving Cars Wearing a Cone of Shame

There's a brilliant activist campaign to stop San Francisco's autonomous vehicles in their tracks.

BY ALISON GRISWOLD

JULY 11, 2023 • 10:45 AM



It looks like a sad unicorn (which, in a way, it is). Screengrab from TikTok/Safe Street Rebel

slate.com

That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency

Yanmao Man[#], Raymond Muller[§], Ming Li[#], Z. Berkey Celik[§], Ryan Gerdes[‡]
[#]University of Arizona [§]Purdue University [‡]Virginia Tech

Q&A