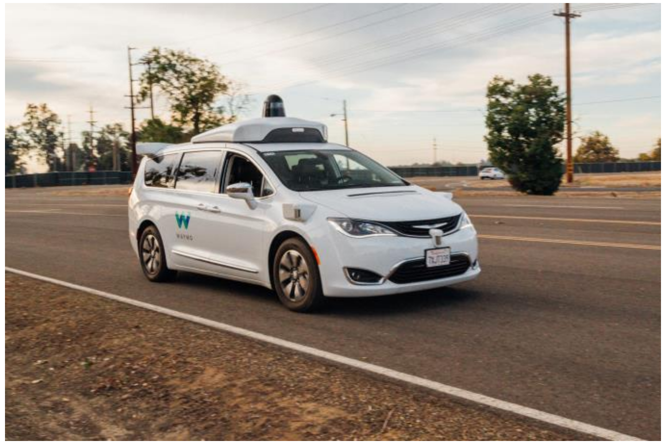
WIP: Towards the Practicality of the Adversarial Attack on Object Tracking in Autonomous Driving

Chen Ma[1], Ningfei Wang[2], Qi Alfred Chen[2], Chao Shen[1]

1 西安交通大學 XI'AN JIAOTONG UNIVERSITY

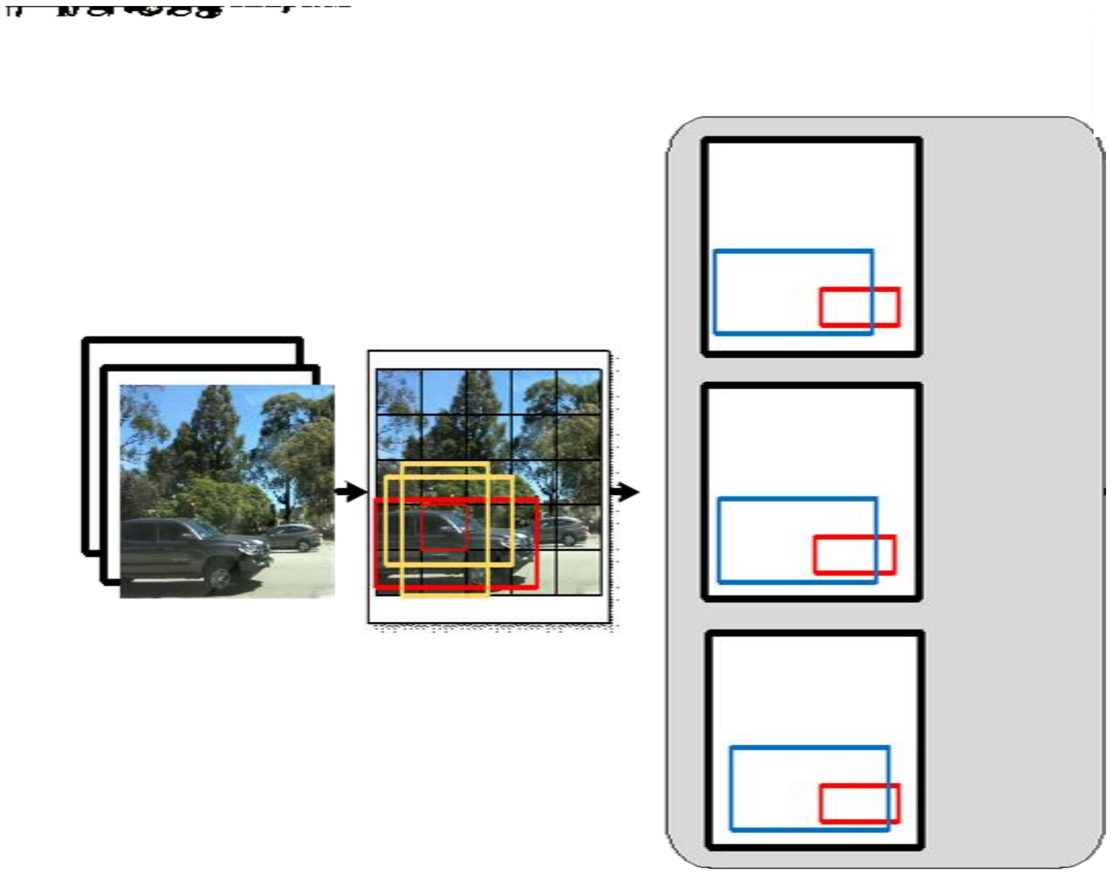
2 UCI

Autonomous Driving (AD) Vehicles are Increasingly Deployed



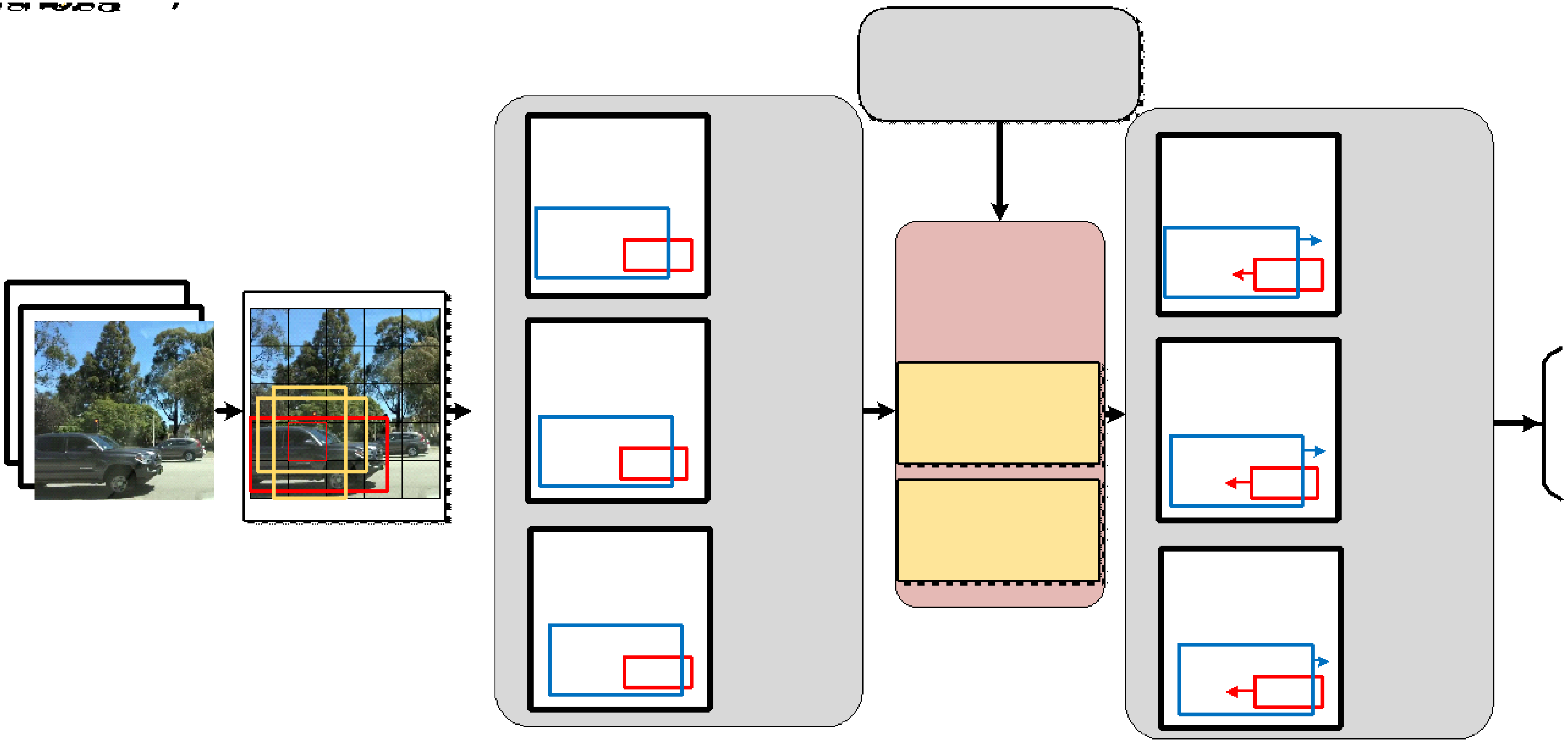
Autonomous Driving (AD) Visual Perception

- Autonomous Driving visual perception consists of object **detection** and object **tracking**.



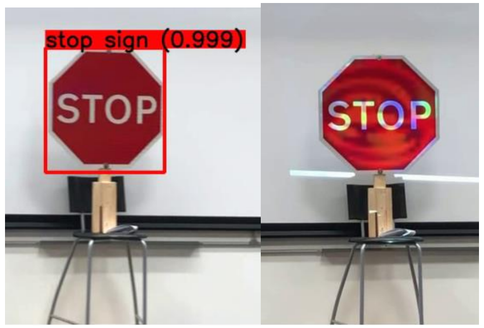
Autonomous Driving (AD) Visual Perception

- Autonomous Driving visual perception consists of object **detection** and object **tracking**.



Prior Attacks on AD Object Detection

- Object detection attack is well studied.
  - ➢ Various forms of adversarial attacks successfully in the physical world.



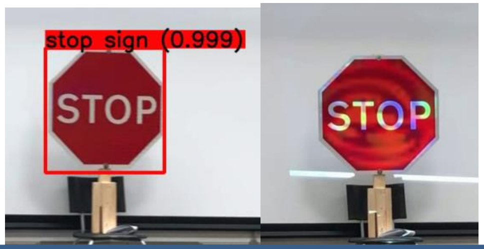
[Lovisotto et al., USENIX Security'21]

[Zhao et al., CCS'19; Eykholt et al. Woot'18]

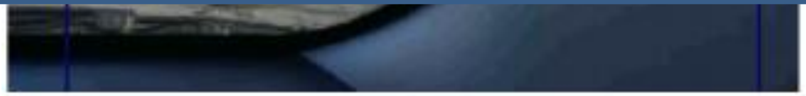
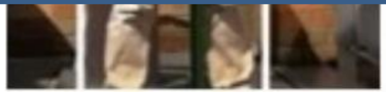
[Huang et al., CVPR'20]

Prior Attacks on AD Object Detection

- Object detection attack is well studied.
  - ➢ Various forms of adversarial attacks successfully in the physical world.



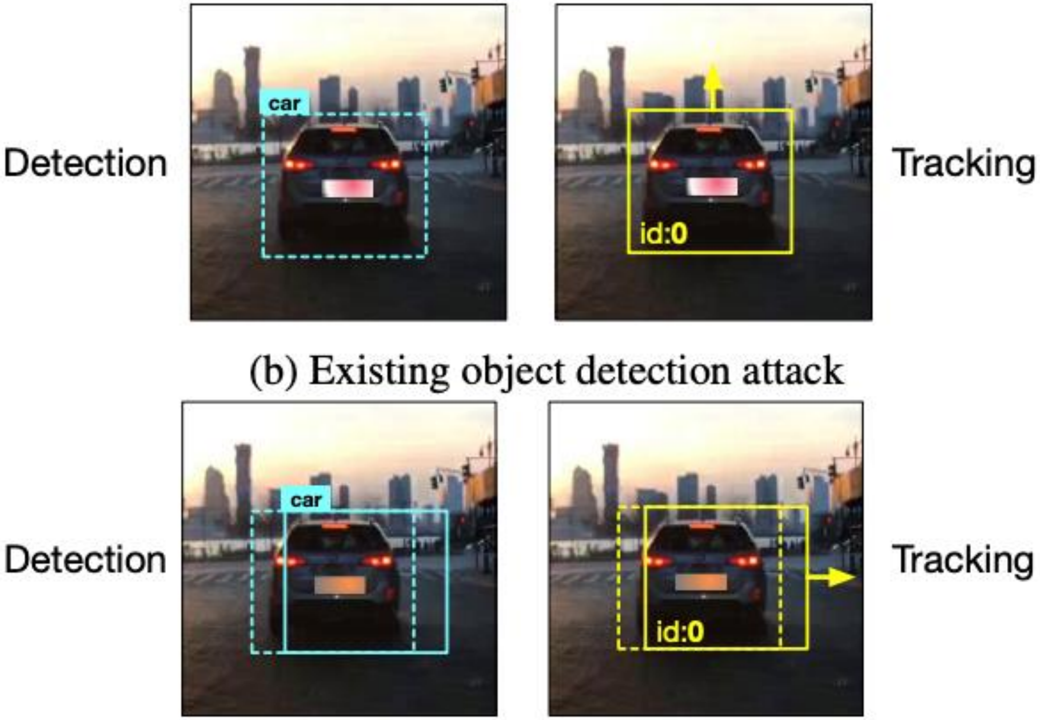
None of them consider the object tracking, which thus does not necessarily lead to end-to-end attack effects in practical AD settings

[Zhao et al., CCS'19; Eykholt et al. Woot'18]

[Huang et al., CVPR'20]

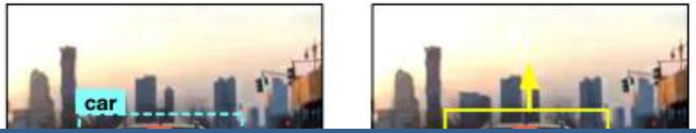
Prior Attacks on AD Object Tracking



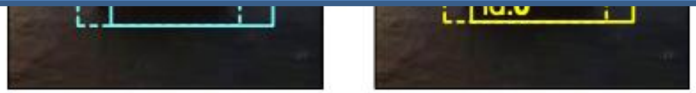
[Jia et al., ICLR’20]: digital attack

[Muller et al., CCS’22]: single-object tracker

Prior Attacks on AD Object Tracking



None of them consider attacking Multiple-Object Tracking (MOT) in the physical world, which is a more representative setup
in the real world



[Muller et al., CCS’22]: single-object tracker

[Jia et al., ICLR’20]: digital attack

Threat Model & Attack Goal

- Threat Model
  - White-box access to the perception pipeline of target AD vehicle
  - Dynamic adversarial patches using the monitors or projectors
- Attack Goal
  - Fool AD vehicles to have tracking errors of a front object to cause crashes or emergency stop

Generating Adversarial Patch

- Prior work simply select all bounding boxes (bbox) around the object

Generating Adversarial Patch

- Prior work simply select all bounding boxes (bbox) around the object
- Prior work simply optimize the shape and the position of the bbox, which is less effective using the standard Lagrangian relaxation method

Generating Adversarial Patch

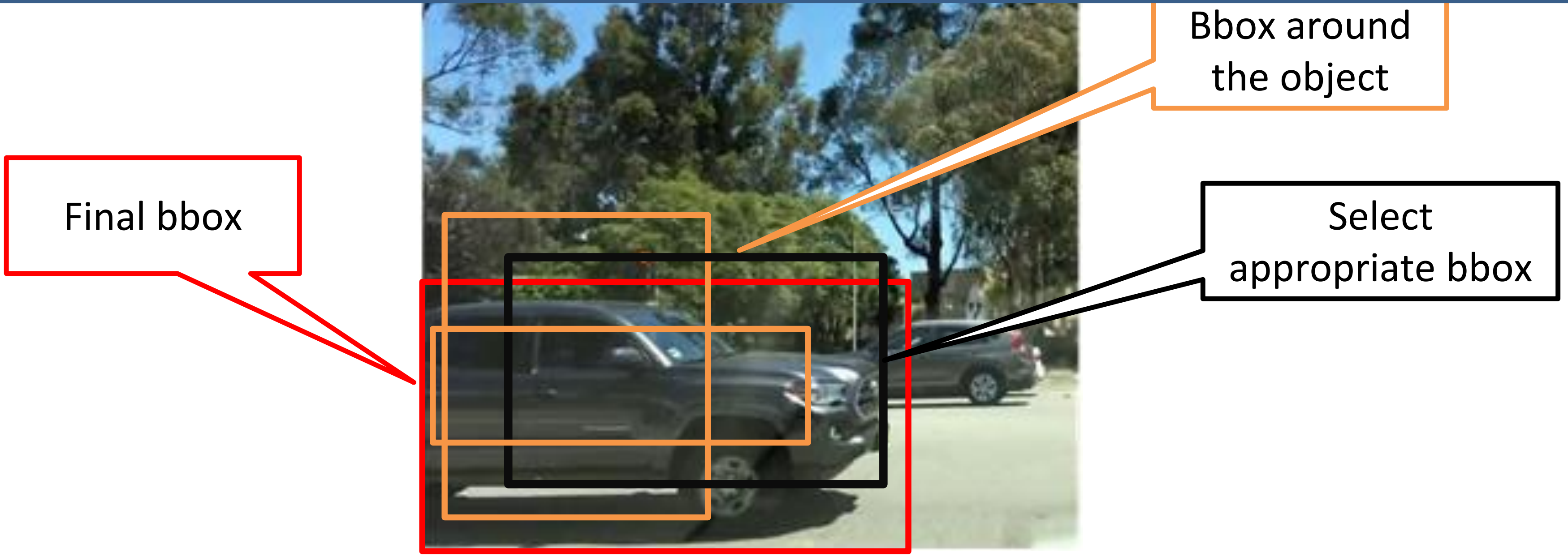
- Strategically select one bounding box as optimization goal
- Optimize the score to keep this box after NMS (Non-Maximum Suppression)
- Optimize the position to satisfy the condition of bbox

Generating Adversarial Patch

Therefore, we need to optimize both the shape and position loss L_r , and the score loss L_s

$$\arg\min_{\Delta} L_r(x+\Delta, b_t, b_s, D) \text{ such that } b_s \in B' \quad (1)$$

$$L_s = \lambda \cdot L_c(x+\Delta, b_s, D) - \sum_{i=0}^{B} \mathbb{1}_i^{obj} \cdot L_c(x+\Delta, b_i, D) \quad (2)$$



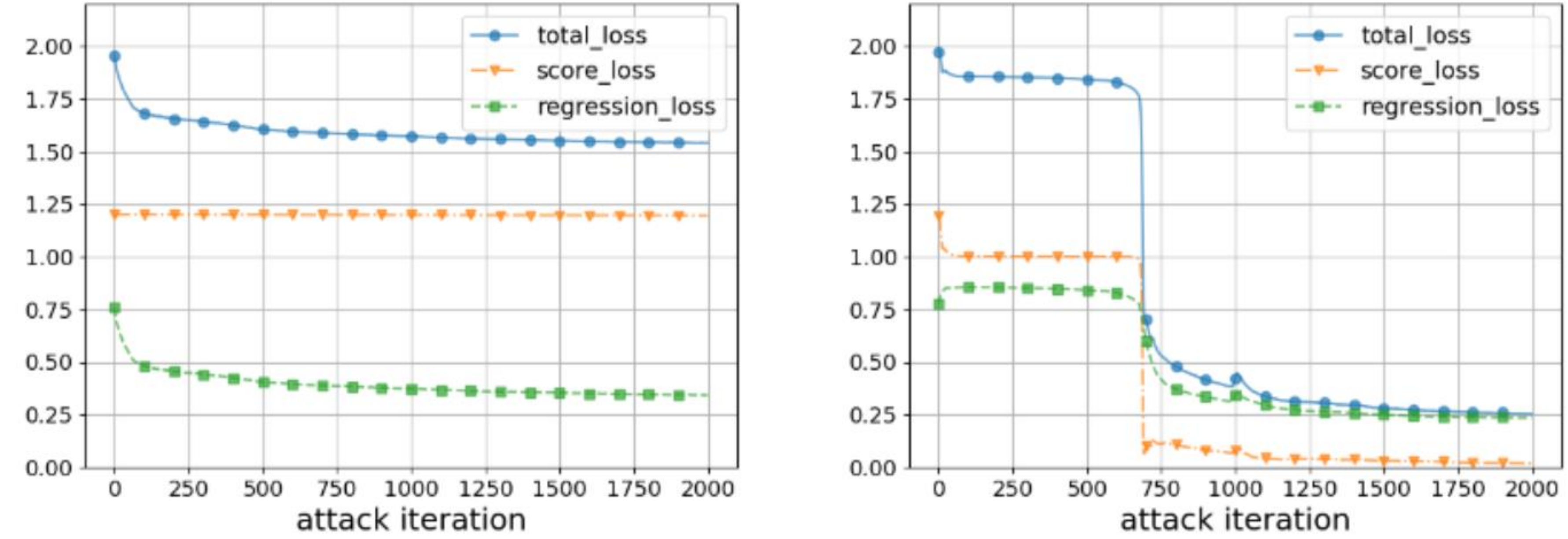
Generating Adversarial Patch

- To solve the optimization problem
  - Standard Lagrangian relaxation method can not work well
  - Score loss is not the lower the better, only need to keep selected bbox after NMS
  - There is conflict between the two losses

$$\arg\min_{\Delta} \mathbb{1}[b_s \in B'] \cdot L_r(x+\Delta, b_t, b_s, D) + \mathbb{1}[b_s \notin B'] \cdot L_s(x+\Delta, b_s, D) \quad (3)$$

Generating Adversarial Patch

- To solve the optimization problem
  - Standard Lagrangian relaxation method can not work well
  - Score loss is not the lower the better, only need to satisfy the co
  - T



(a) Lagrangian relaxation method used by previous works

(b) our optimization method

Preliminary Evaluation

- Evaluate on 2 anchor-based detectors included in YOLO v3 (adopted in Autoware.AI) & camera-based object detection model in Baidu Apollo
  - Select 10 video clips from the Berkeley Deep Driving Dataset
  - Capture video data in the real world and stick cardboard on the back of the car to mark the patch location

Preliminary Evaluation

- Evaluate on 2 anchor-based detectors included in YOLO v3 (adopted in Autoware.AI) & camera-based object detection model in Baidu Apollo
  - Select 10 video clips from the Berkeley Deep Driving Dataset
  - Capture video data in the real world and stick cardboard on the back of the car to mark the patch location
- Effectiveness
  - 90% success rate on YOLO v3 and 80% success rate on the Apollo model

Preliminary Evaluation

- Evaluate on 2 anchor-based detectors included in YOLO v3 (adopted in Autoware.AI) & camera-based object detection model in Baidu Apollo
  - Select 10 video clips from the Berkeley Deep Driving Dataset

Detection results

Tracking results

Conclusion & Future Work

- Conclusion
  - Achieve an adversarial attack against the complete visual pipeline of real-world AD systems
  - Adopt an optimization-based approach with novel designs to solve adversarial patch generation problem
  - Evaluate our attack on complete visual perception of real-world AD systems
- Future work
  - Comprehensive evaluation: evaluate our attack in a **large-scale** dataset, evaluate the **generality**, and **compare** our work to the state-of-the-art practical tracking attack.
  - Practicality: improve the **practicality** and **robustness** of the adversarial patch to make our adversarial patch work successfully in the physical world

Thank you for listening.