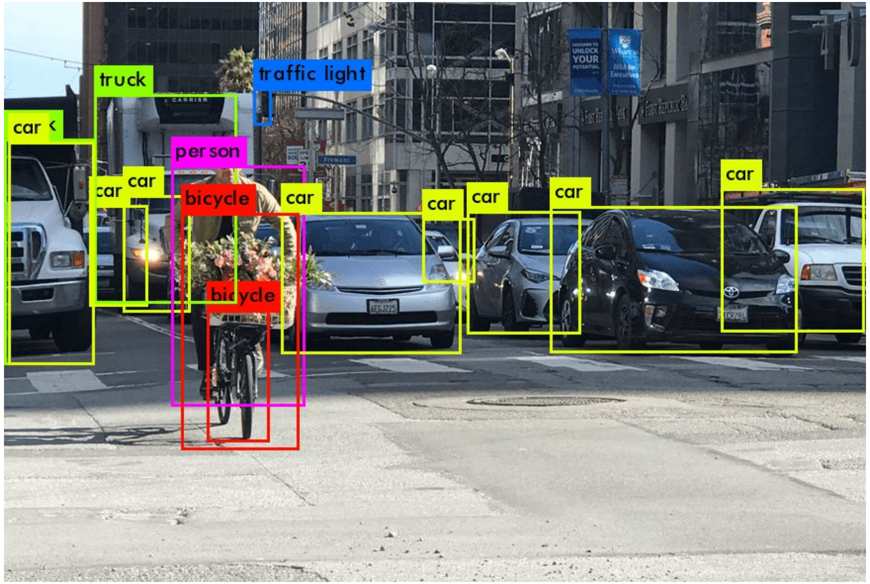
DATE 25
DESIGN, AUTOMATION AND TEST IN EUROPE
THE EUROPEAN EVENT FOR ELECTRONIC SYSTEM DESIGN & TEST
31 MARCH – 2 APRIL 2025
LYON, FRANCE
CENTRE DE CONGRÈS DE LYON

Generating and Predicting Output Perturbations in Image Segmenters

Matthew Bozoukov, Anh Vu Doan, **Bryan Donyanavard**

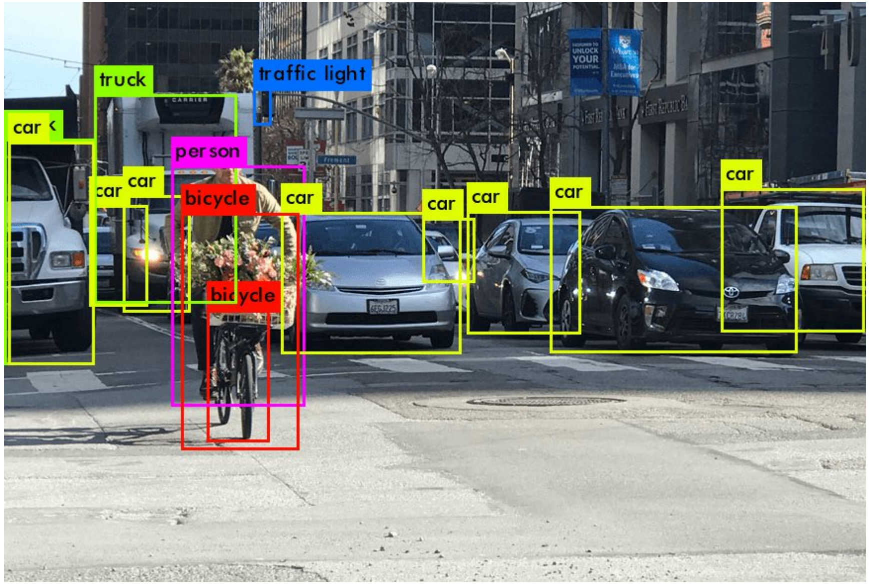
SDSU | San Diego State University

Background: AI in (safety critical) autonomous systems



[The Complete Guide to Object Detection: An Introduction to Detection in 2024 — visionplatform]

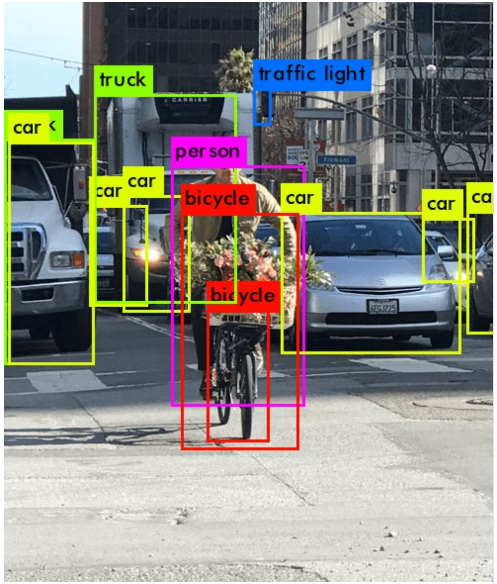
Background: AI in (safety critical) autonomous systems



[The Complete Guide to Object Detection: An Introduction to Detection in 2024 — visionplatform]

[How are Satellites Bringing Low-Latency Internet to Autonomous Vehicles? - Zuken US]

Background: AI in (safety critical) autonomous systems



[The Complete Guide to Object Detection: An Intro visionplatform]

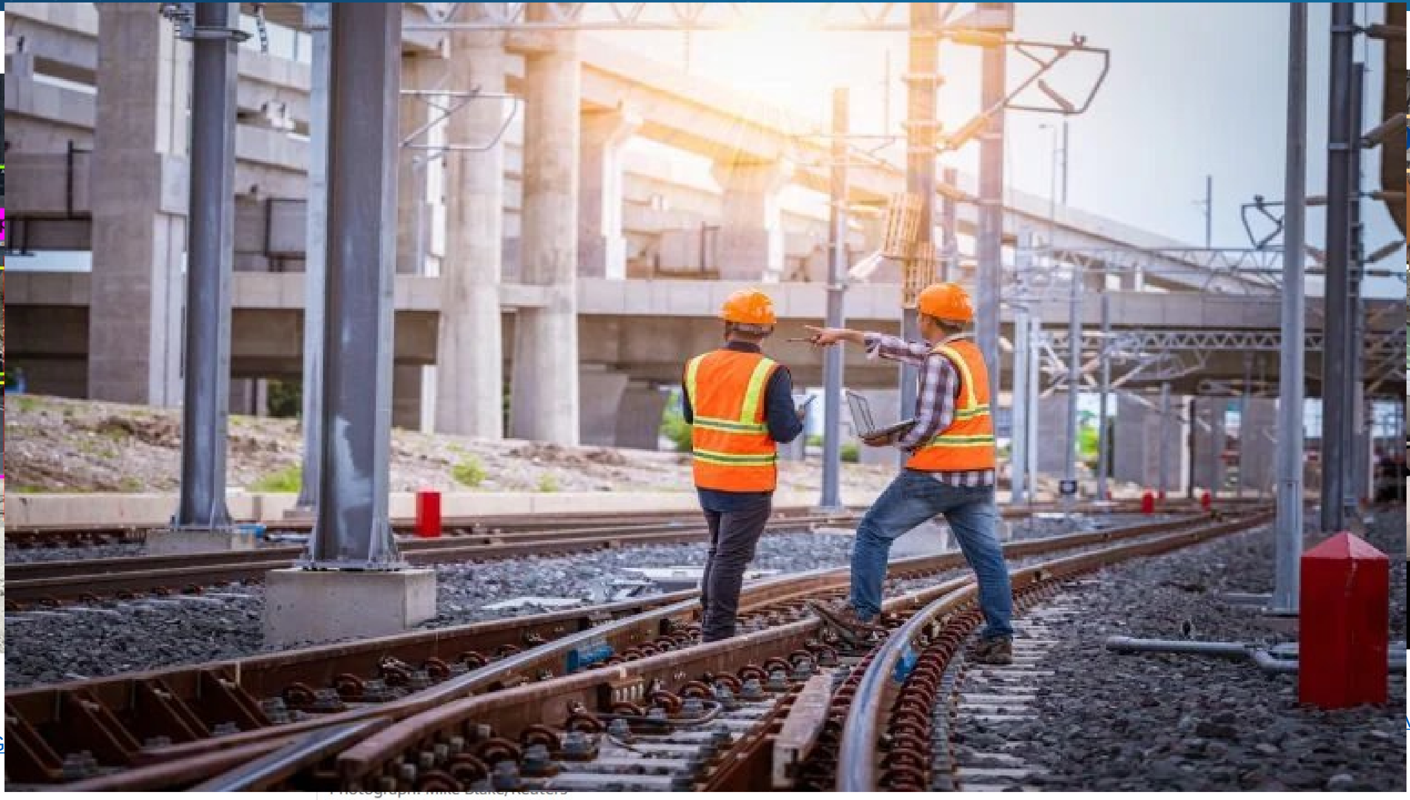
Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says

Federal transportation agency finds Tesla's claims about feature don't match their findings and opens second investigation

A Tesla model 3 drives on autopilot along the 405 highway in Westminster, California, in 2022. Photograph: Mike Blake/Reuters

[Latency Internet to Autonomous Vehicles? - Zuken US]

Background: AI in (safety critical) autonomous systems

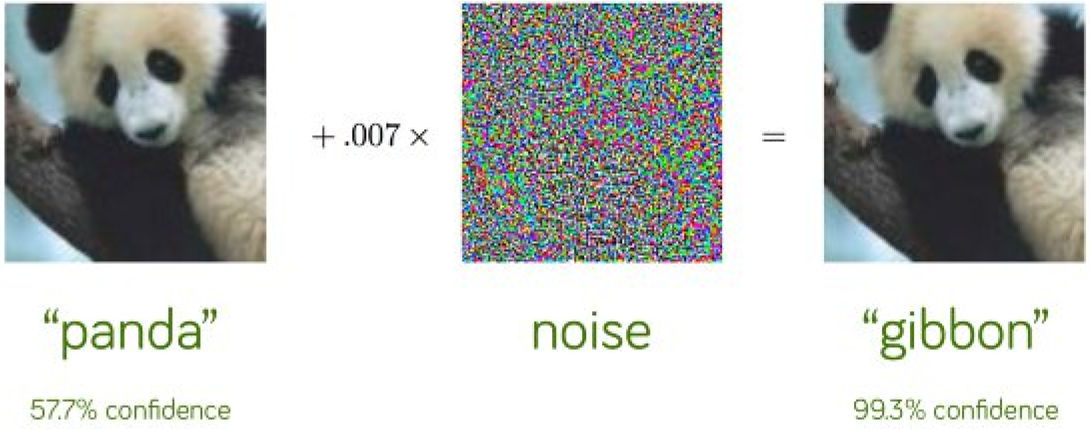


[The Complete G](#) [visionplatform](#)

[Vehicles? - Zuken US](#)

[TrackEi enables real-time defect detection and predictive maintenance using NVIDIA Jetson edge AI. Credit: APChanel/Shutterstock.](#)

Prior Work: Perturbations in object detection



Source: Goodfellow et al., "Explaining and Harnessing Adversarial Examples", *ICLR 2015*

Prior Work: Perturbations in object detection

Stop sign classified as a 45 mph speed limit!

Source: Eykholt et al., “Robust Physical-World Attacks on Deep Learning Visual Classification”, *CVPR 2018*

Prior Work: Butterfly effect attack

Small perturbations in object detection

Butterfly Effect Attack: Tiny and Seemingly Unrelated Perturbations for Object Detection Doan et al, DATE 2023

Prior Work: Butterfly effect attack

Small perturbations in object detection

Butterfly Effect Attack: Tiny and Seemingly Unrelated Perturbations for Object Detection Doan et al, DATE 2023

Our work: Generating and detecting perturbation

Part 1: Generate problematic perturbation in image segmenters

Part 2: Detect problematic perturbation in image segmenters

Adversarial perturbation based on multi-objective optimization

Multi-objective optimization-based exploration with NSGA-II and AGE-MOEA

⇒explicit encoding of the filter mask applied to the image

⇒mutation emulates sensor degradation

Adversarial perturbation based on multi-objective optimization

Multi-objective optimization-based exploration with NSGA-II and AGE-MOEA

⇒explicit encoding of the filter mask applied to the image

⇒mutation emulates sensor degradation

Objective functions:

1. **Maximize performance degradation**

   ⇒bounding-box based

Adversarial perturbation based on multi-objective optimization

Multi-objective optimization-based exploration with NSGA-II and AGE-MOEA

⇒explicit encoding of the filter mask applied to the image

⇒mutation emulates sensor degradation

Objective functions:

1. **Maximize performance degradation**

   ⇒bounding-box based

1. **Minimize perturbation**

   ⇒L2 norm between images

Adversarial perturbation based on multi-objective optimization

Multi-objective optimization-based exploration with NSGA-II and AGE-MOEA

⇒explicit encoding of the filter mask applied to the image

⇒mutation emulates sensor degradation

Objective functions:

1. **Maximize performance degradation**

   ⇒bounding-box based

1. **Minimize perturbation**

   ⇒L2 norm between images

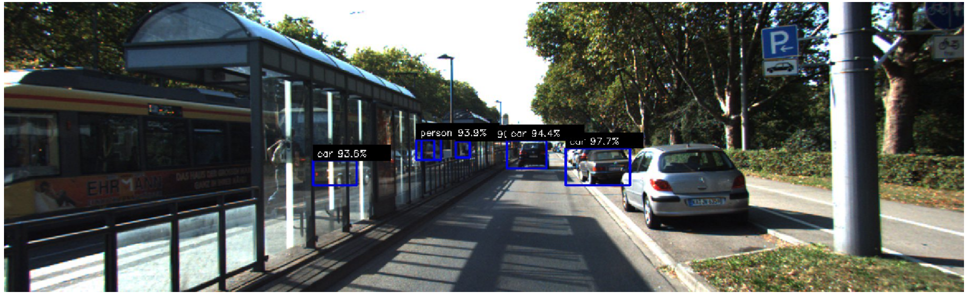
1. **Maximize unrelatedness**

   ⇒distance from perturbation to object

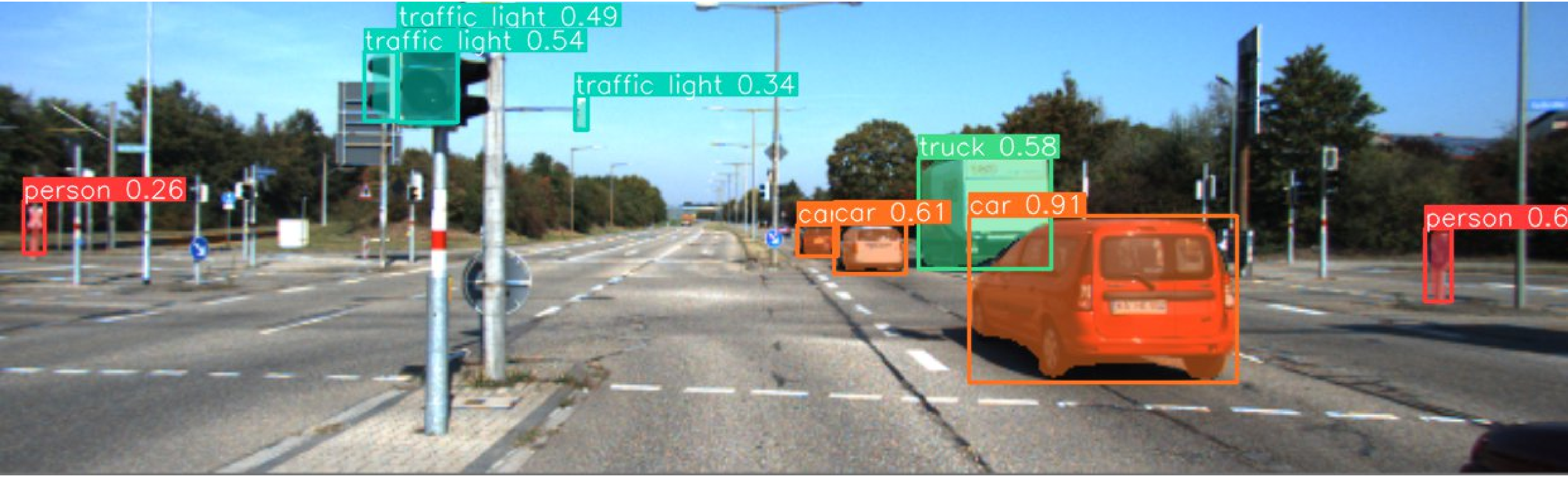
Adversarial perturbation based on multi-objective optimization

Experimental setup:

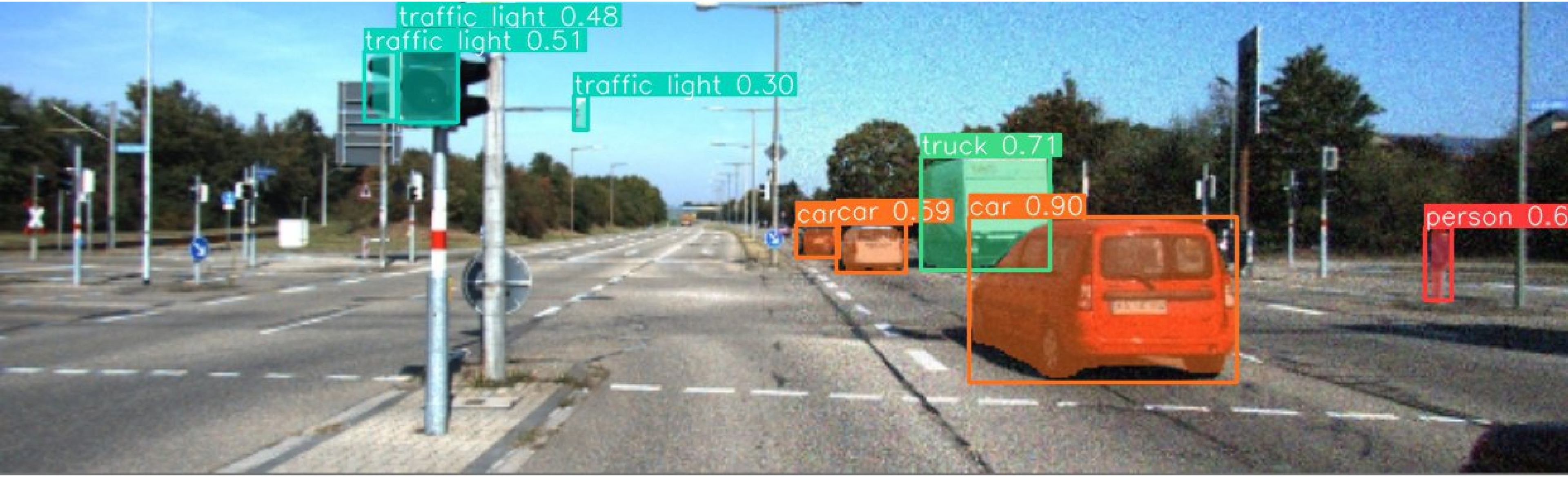
- KITTI dataset
- Transformer-based (DETR) and CNN-based (YOLOv5) object segmentation
- Perturbation injection on opposite half of image



YOLO segmentation <u>without</u> perturbation added

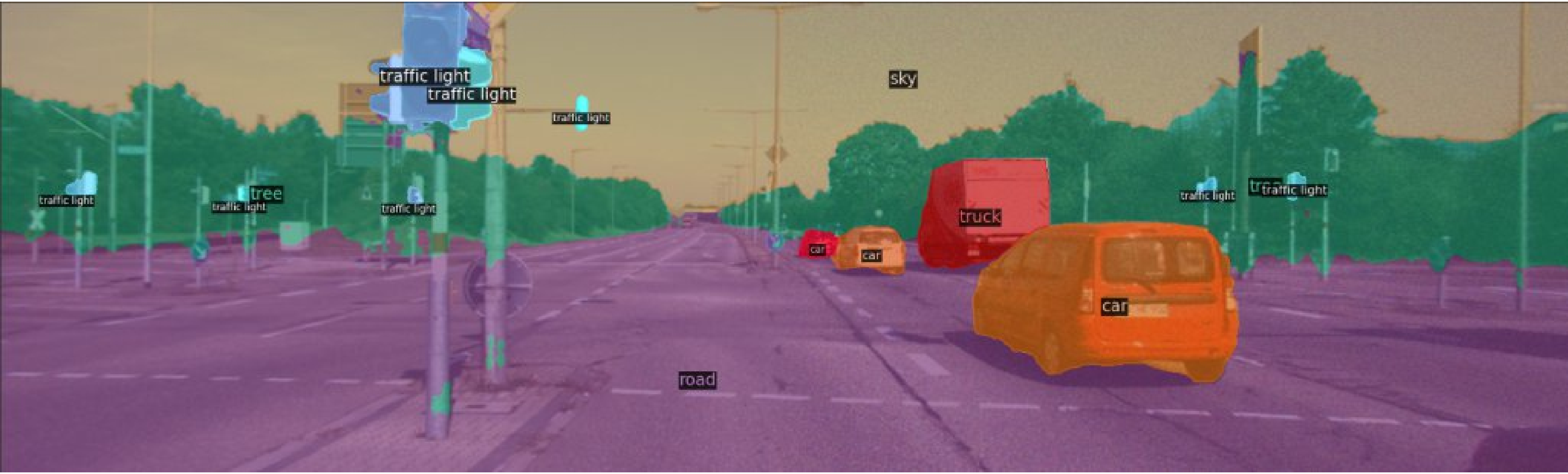


YOLO segmentation with perturbation added



DETR segmentation without perturbation added

DETR segmentation with perturbation added



Our work: Generating and detecting perturbation

Part 1: Generate problematic perturbation in image segmenters

Part 2: Detect problematic perturbation in image segmenters

Our work: Generating and detecting perturbation

Part 1: Generate problematic perturbation in image segmenters

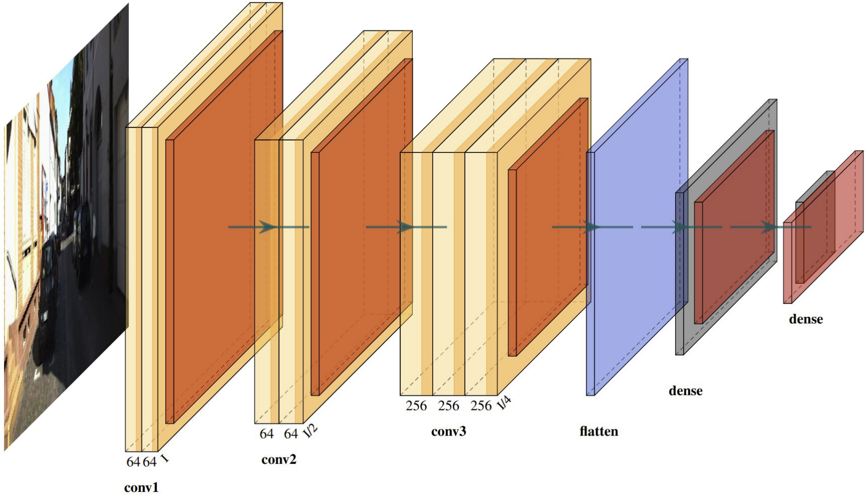
Part 2: Detect problematic perturbation in image segmenters

Can we predict the degradation these perturbations will cause?

Evaluation: Detecting perturbation

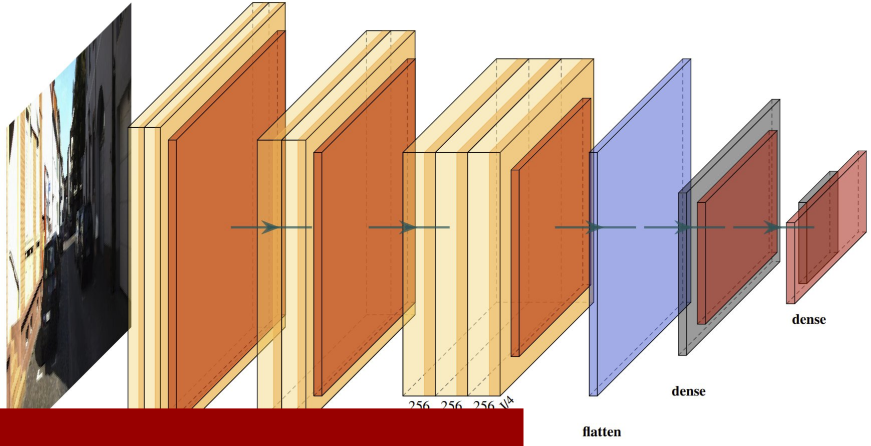
CNN with 3 conv layers and 2 dense layers

Trained with segmentation output of optimal perturbed images



CNN with 3 conv layers and 2 dense layers

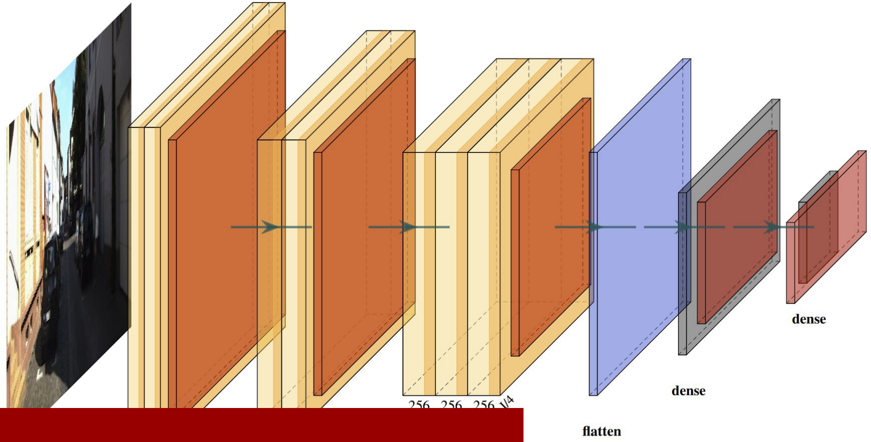
Trained with segmentation output of optimal perturbed images



90% precision, 100% recall

CNN with 3 conv layers and 2 dense layers

Trained with segmentation output of optimal perturbed images



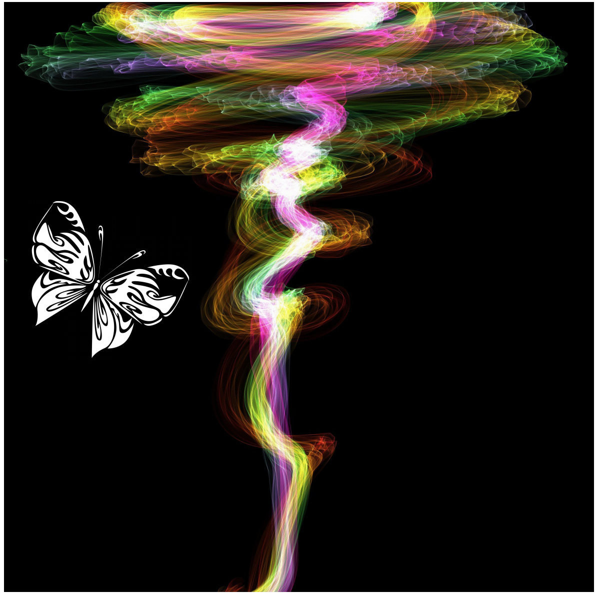
Summary

Tiny and seemingly unrelated perturbations can cause mis-identification and -segmentation of objects

⇒true positives become false negatives

⇒true negatives become false positives

⇒segmentation mask degradation



Summary

Tiny and seemingly unrelated perturbations can cause mis-identification and -segmentation of objects

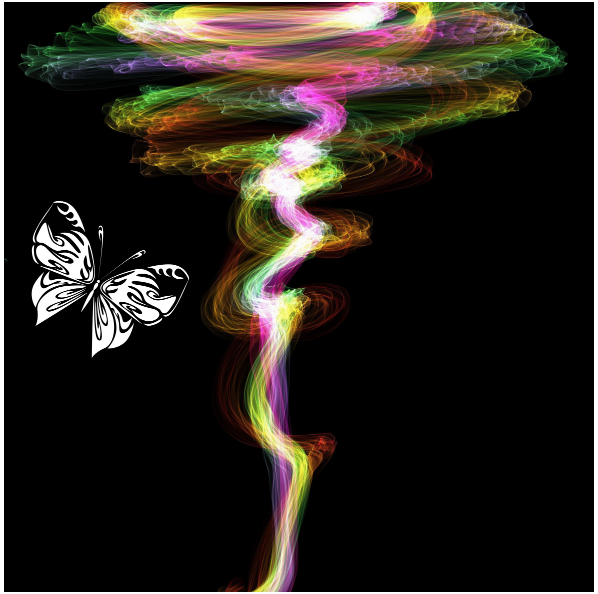
⇒true positives become false negatives

⇒true negatives become false positives

⇒segmentation mask degradation

Errors due to perturbation can be predicted

⇒environmentally sensitive

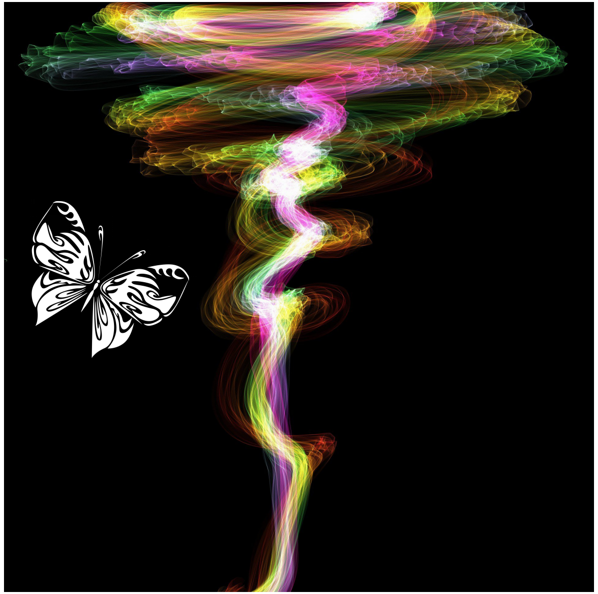


Future work

Can we root cause the errors in the network architecture?

How does this generalize to broader computer vision applications?

Can we generate perturbations in realtime?



Thank You!