black hat EUROPE 2018

DECEMBER 3-6, 2018

EXCEL LONDON / UNITED KINGDOM

Perception Deception: Physical Adversarial Attack Challenges and Tactics for DNN-based Object Detection

Zhenyu (Edward) Zhong, Yunhan Jia, Weilin Xu, Tao Wei

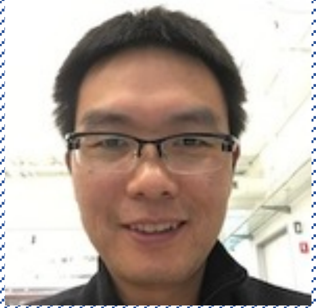
black hat EUROPE 2018

Scan Me

Our Team X-Lab

Chief Security Scientist
Dr. Tao Wei

AI Security Research



WE'RE HIRING!

Dr. Zhenyu(Edward) Zhong
edwardzhong [at] baidu DOT com

Dr. Yunhan Jia

Weilin Xu

System Security Research

MesaTEE
https://www.mesatee.org

RUST · SGX
https://github.com/baidu/rust-sgx-sdk

AdvBox
https://github.com/baidu/AdvBox

MesaLock
https://github.com/mesalock-linux

MesaLink
https://github.com/mesalock-linux/mesalink

MesaPy
https://github.com/mesalock-linux/mesapy

Disclaimer

- This talk doesn’t target any commercial autonomous driving systems.
- We don’t provide any comments to the vulnerabilities of the perceptions of existing autonomous driving systems.
- We focus on state-of-the-art object detection methods, all the results/techniques are proof-of-concept.

black hat EUROPE 2018

Car Safety – Unintended Acceleration

AP / May 25, 2010, 7:08 PM

https://www.cbsnews.com/news/toyota-unintended-acceleration-has-killed-89/

Toyota "Unintended Acceleration" Has Killed 89

Unintended acceleration in Toyota vehicles may have been involved in the deaths of 89 people ov… deaths possibly linked to the massive …

NASA Engineering a… Technical Assess…

Title:

National Highway Traffic Safet… Toyota Unintended Accelerati…

Proof for the hypothesis that the ETCS-i caused t… submitted VOQs … a single failure mode found that, combined with d… degrees in certain conditions and may not generat… releasing the accelerator pedal or overridden by th… openings, the NESC team found single failure mo… openings less than 5 degrees. These failures may … as described in submitted VOQs and may not gen… releasing the accelerator pedal or overridden by the braking system.

…the large throttle ope…

submitted VOQs could …

Because proof that the ETCS-i caused the reported UAs was not found does not mean it could not occur. However, the testing and analysis described in this report did not find that TMC ETCS-i electronics are a likely cause of large throttle openings as described in the VOQs.

…does not mean it could not occur…

The New York Times

Toyota Will Pay $1.6 Billion Over Faulty Accelerator Suit

By Jaclyn Trop

July 19, 2013

…ingle Bit Flip That Killed

…13 108

…tly did not have time to do," Barr said. For one thing, by looking …

"unprotected critical variables."…

…b-CPU," and they "uncovered gaps and defects in the throttle fail…

Green Hills Simulator. "This confirmed tasks can die without the … roup also independently checked worst-case stack depth. "We found … s that NASA relied on."

The e… testing… Barr said. "We also obtained and reviewed the source code for the black box and found that it can record false information about the driver's actions in the final seconds before a crash."

…the defects we found were linked to unintended Acceleration through vehicle testing, …

https://www.eetimes.com/document.asp?doc_id=1319903&page_number=2

Car Safety – Autonomous Driving

THE WALL STREET JOURNAL.

TECH

Uber Self-Driving Car That Struck, Killed Pedestrian Wasn't Set to Stop in an Emergency

Pedestrian tested positive for methamphetamine and marijuana

National Transportation Safety Board investigators inspected the self-driving Uber vehicle after the fatal crash in Tempe, Ariz. **PHOTO:** NATIONAL TRANSPORTATION SAFETY BOARD/REUTERS

Car Safety – Rear Ended Into Fire Truck

CBS/AP / *May 15, 2018, 3:25 AM*

Tesla driver says she slammed into fire truck on Autopilot

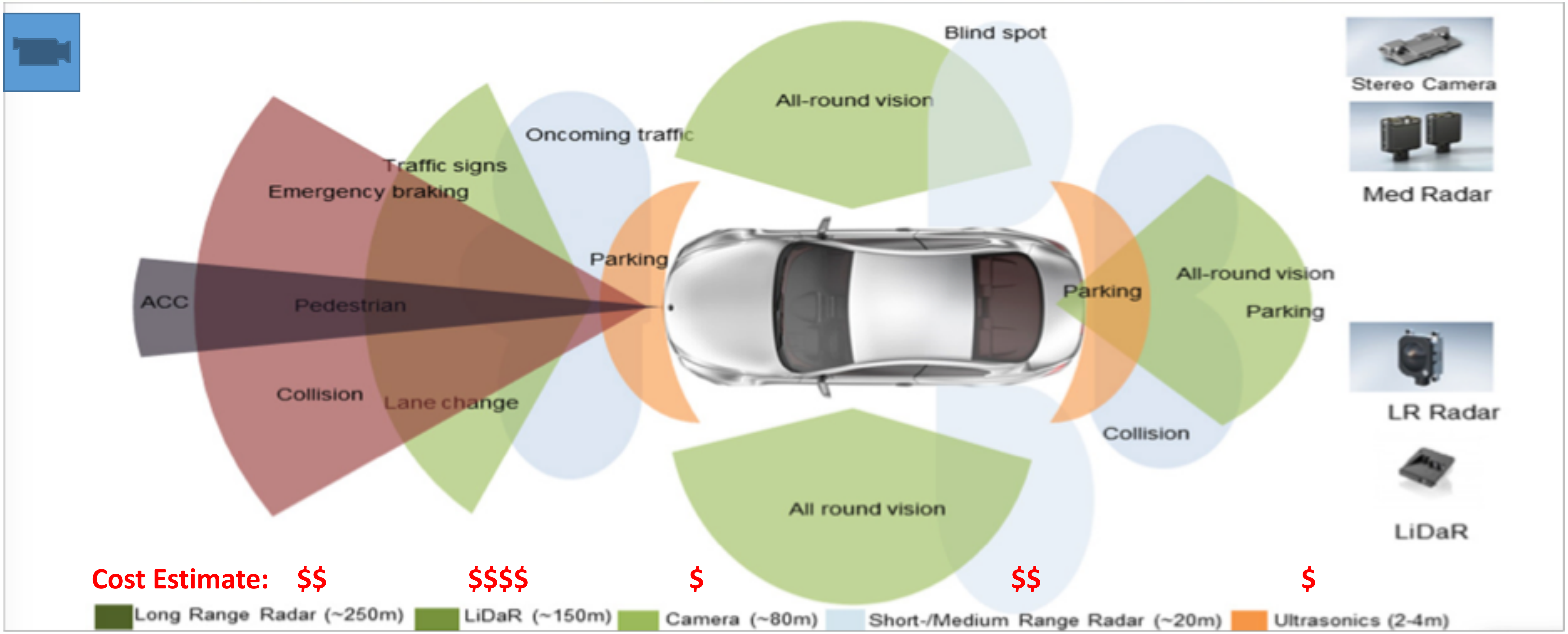
A photo released by the South Jordan Police Department shows a traffic collision involving a Tesla Model S sedan with a Fire Department mechanic truck stopped at a red light in South Jordan, Utah, May 11, 2018.

SALT LAKE CITY -- The driver of a Tesla electric car had the vehicle's semi-autonomous Autopilot mode engaged when she slammed into the back of a Utah fire truck over the weekend, in the latest crash involving a car with self-driving features. The 28-year-old driver of the car told police in suburban Salt Lake City that the system was switched on and that she had been looking at her phone before the Friday evening crash.

Tesla's Autopilot system uses radar, cameras with 360-degree visibility and sensors to detect nearby cars and objects. It's built so cars can automatically change lanes, steer, park and brake to help avoid collisions.

The auto company markets the system as the "future of driving" but warns drivers to remain alert while using Autopilot and not to rely on it to entirely avoid accidents. Police reiterated that warning Monday.

Car Perception While Driving



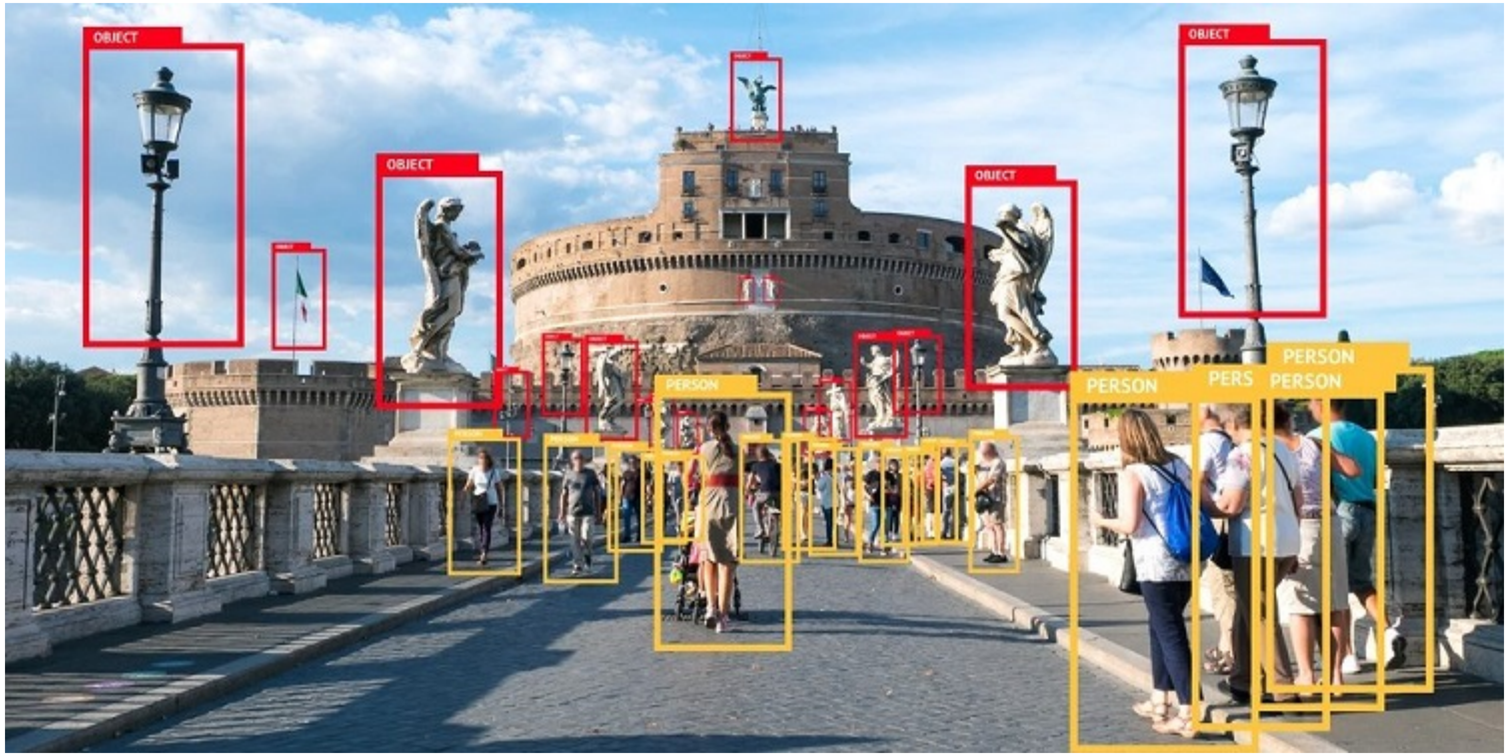
Cost Estimate: $$ $$$$ $ $$ $

Imgsrc: http://www.crosslinksolutions.co.uk/service-listing/adas-calibration-equipment/

Behind Perception: End2End Object Detection

Object Detection:

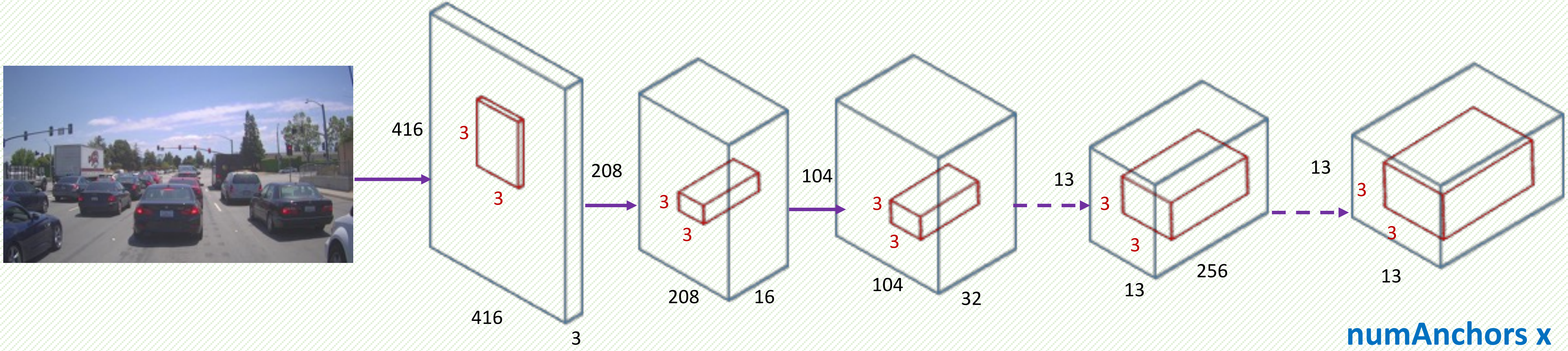
a technology related to computer vision and image processing that deals with instances of semantic objects of certain class in digital images and videos.



https://software.intel.com/en-us/articles/a-closer-look-at-object-detection-recognition-and-tracking

State-of-the-Art Vision-based Object Detection

YOLO (You Only Look Once)



numAnchors x
(5 + numClasses)

Pick Our Target – YOLOv3

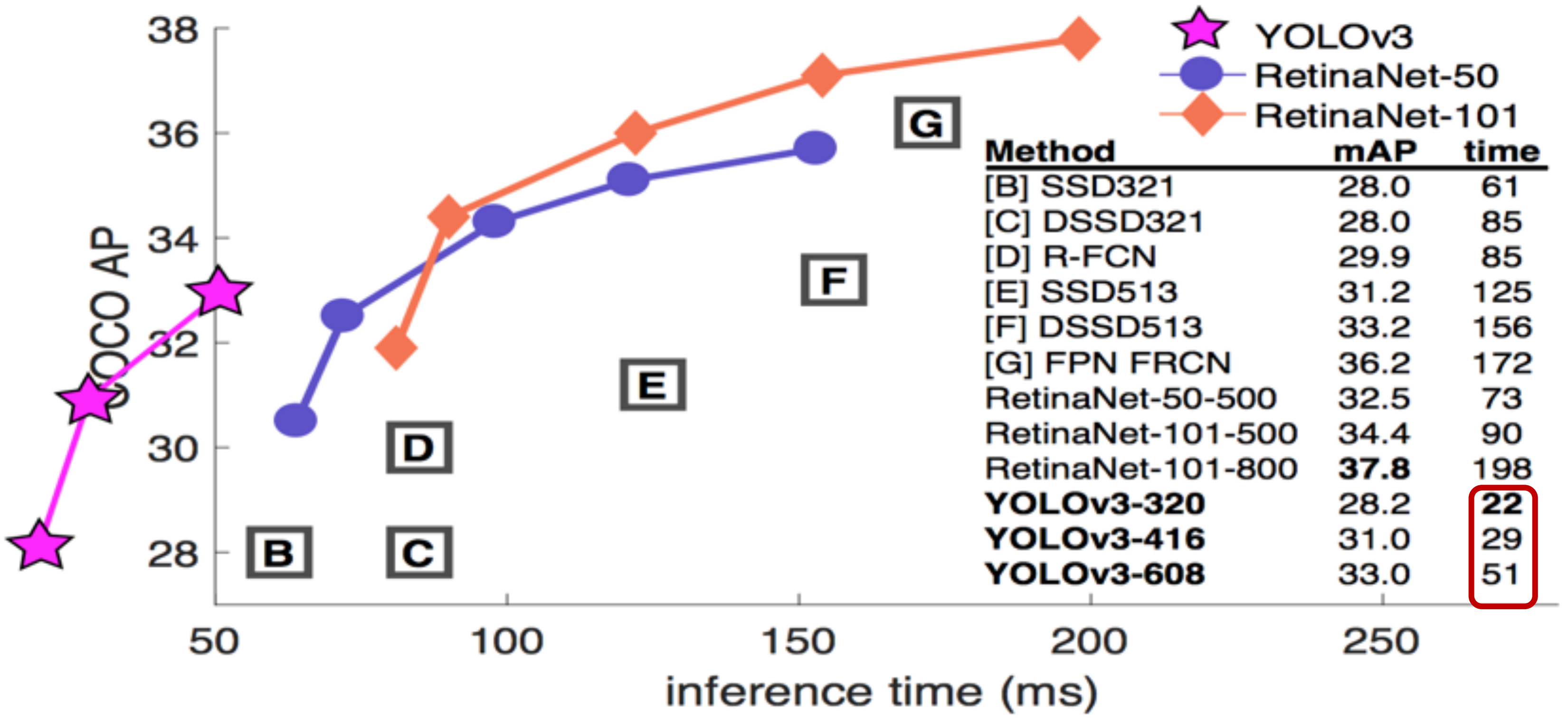
Accuracy on MS COCO

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [3]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [6]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [4]	Inception-ResNet-v2 [19]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [18]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [13]	DarkNet-19 [13]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [9, 2]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [2]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [7]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [7]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

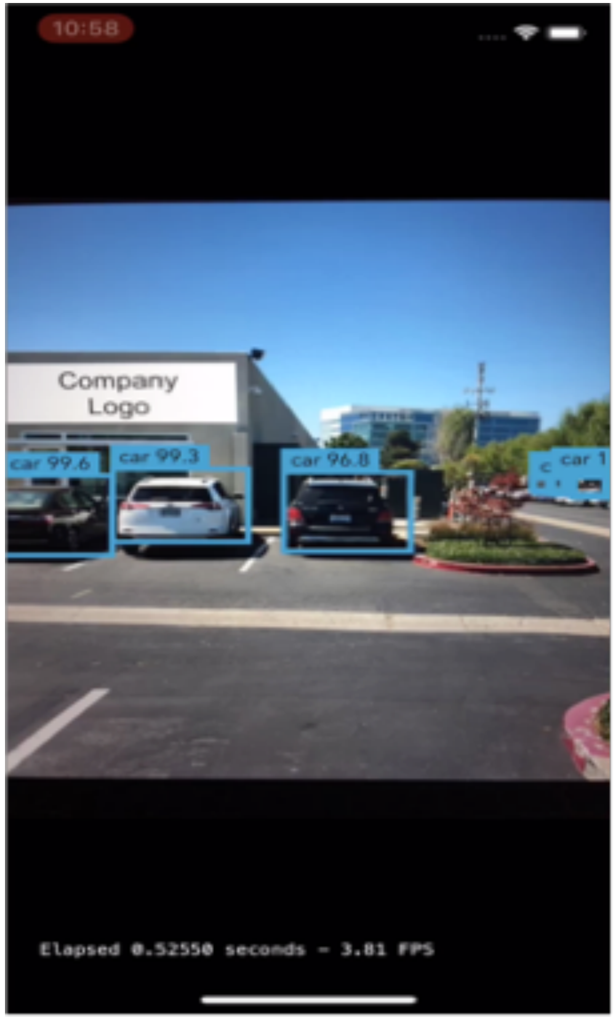
YOLOv3: An Incremental Improvement. Joseph Redmon, Ali Farhadi
https://arxiv.org/pdf/1804.02767.pdf

Pick Our Target – YOLOv3

Performance



Method	mAP	time
[B] SSD321	28.0	61
[C] DSSD321	28.0	85
[D] R-FCN	29.9	85
[E] SSD513	31.2	125
[F] DSSD513	33.2	156
[G] FPN FRCN	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	**37.8**	198
YOLOv3-320	28.2	**22**
YOLOv3-416	31.0	**29**
YOLOv3-608	33.0	51



YOLOv3: An Incremental Improvement. Joseph Redmon, Ali Farhadi
https://arxiv.org/pdf/1804.02767.pdf

Current Status of Adversarial Example

Definition:

For an input image x,

minimize $\mathcal{D}(x, x+\delta)$, $s.t.\ C(x+\delta) = t, x+\delta \in [0,1]^n$

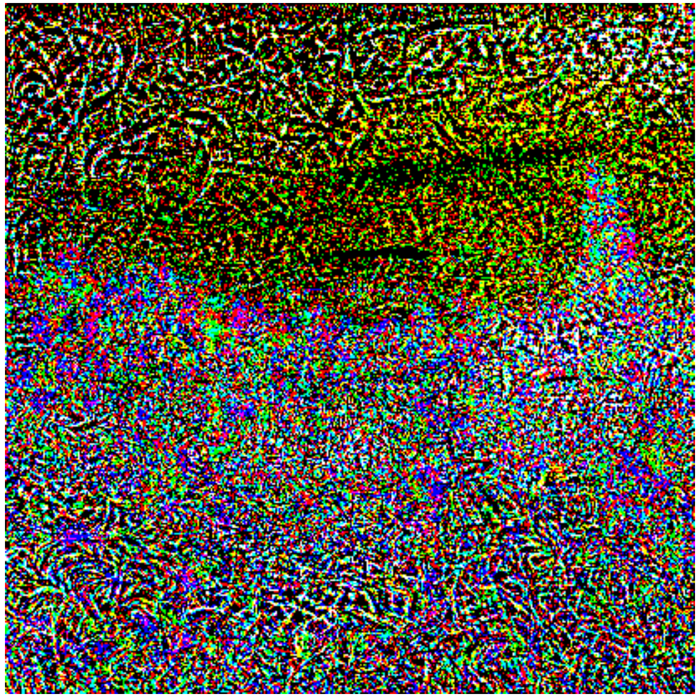
The most well-studied distance metric: $\boldsymbol{L_p}$ $\boldsymbol{Norm}$ Perturbations

- L_∞ -- each pixel is allowed to be changed by up to a limit
- $L_\mathbf{0}$ -- number of pixels altered that matter most
- $L_\mathbf{2}$ -- many small changes to many pixels

Adversarial Examples & L_∞ Norm Perturbations Impact to DNN

Source Image

Perturbations



Perturbed Images

FGSM L_∞ **based Perturbation Method**

Intuition: each pixel is allowed to change by up to a limit

$$x' = x - \epsilon \cdot sign(\nabla Loss_{F,t}(x))$$

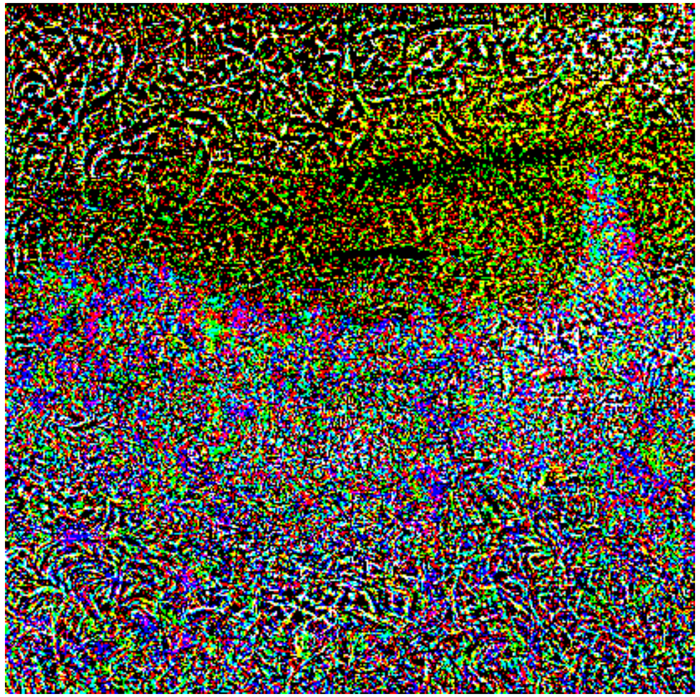
Still in Digital Context

Adversarial Examples & L_∞ Norm Perturbations Impact to DNN

Source Image

$+$

Perturbations



$\rightarrow$

YOLOv3 Detection

FGSM L_∞ **based Perturbation Method**

Intuition: each pixel is allowed to change by up to a limit

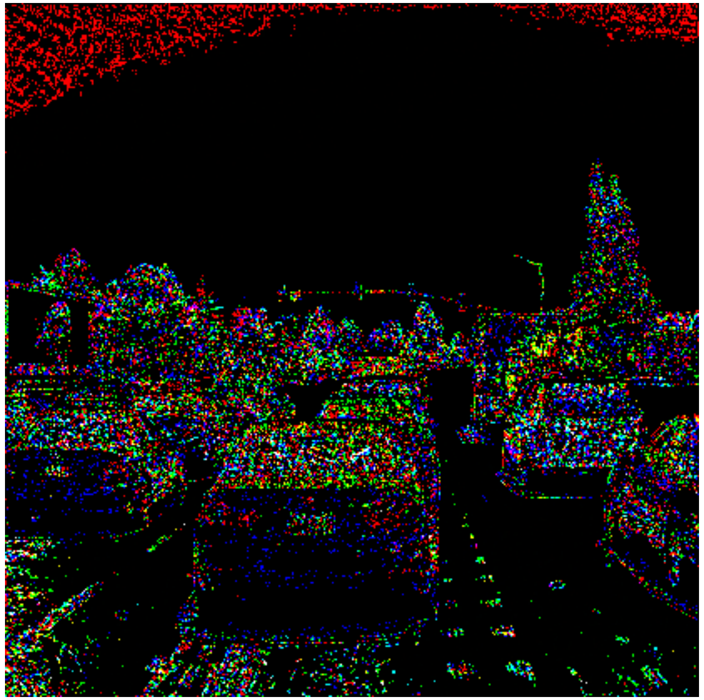
$$x' = x - \epsilon \cdot sign(\nabla Loss_{F,t}(x))$$

Still in Digital Context

Adversarial Examples & L_0 Norm Perturbations Impact to DNN

Source Image

Perturbations



Perturbed Image

$JSMA$ L_0 based Perturbation Method

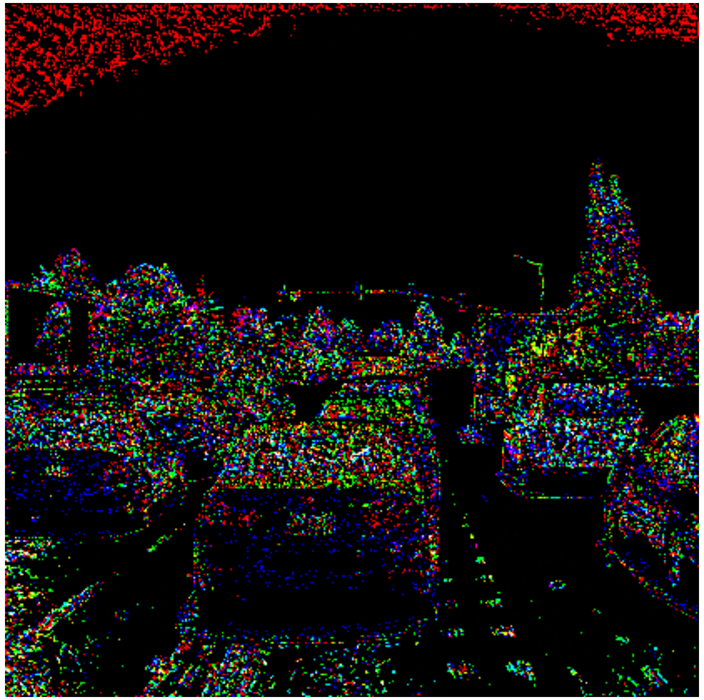
Intuition: # of pixels altered that matter the most

Still in Digital Context

Adversarial Examples & L_0 Norm Perturbations Impact to DNN

Source Image

Perturbations



YOLOv3 Detection

$JSMA$ L_0 based Perturbation Method

Intuition: # of pixels altered that matter the most

Still in Digital Context

Adversarial Examples & L_2 Norm Perturbations Impact to DNN

Source Image

Perturbations

Perturbed Image

$CW2$ L_2 based Perturbation Method

Intuition: many small changes to many pixels

$$minimize \ \| x - x' \|_2^2 + c \cdot f(x')$$

Still in Digital Context

Adversarial Examples & L_2 Norm Perturbations Impact to DNN

Source Image | **Perturbations** | **Perturbed Image**

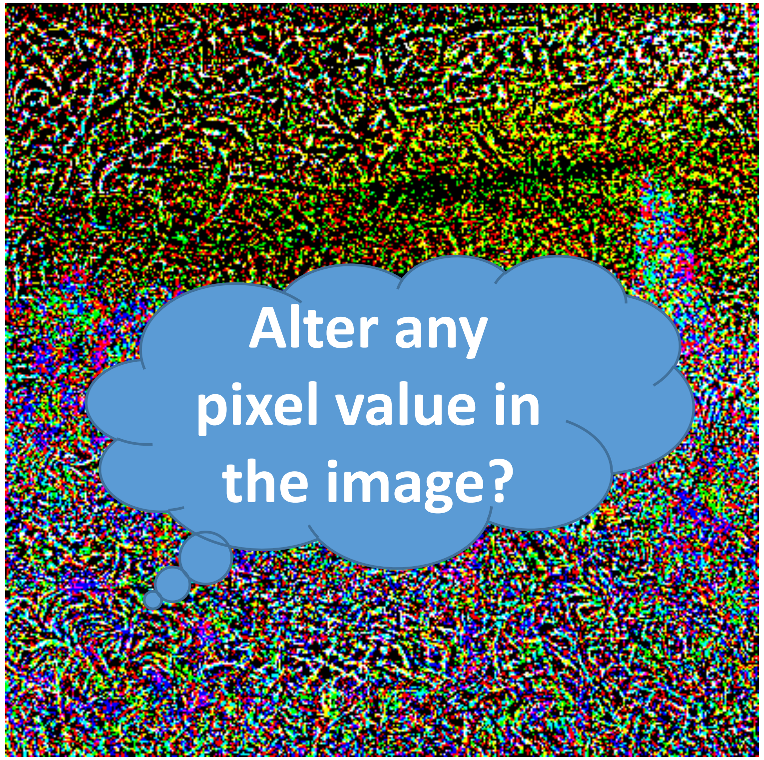
+

→

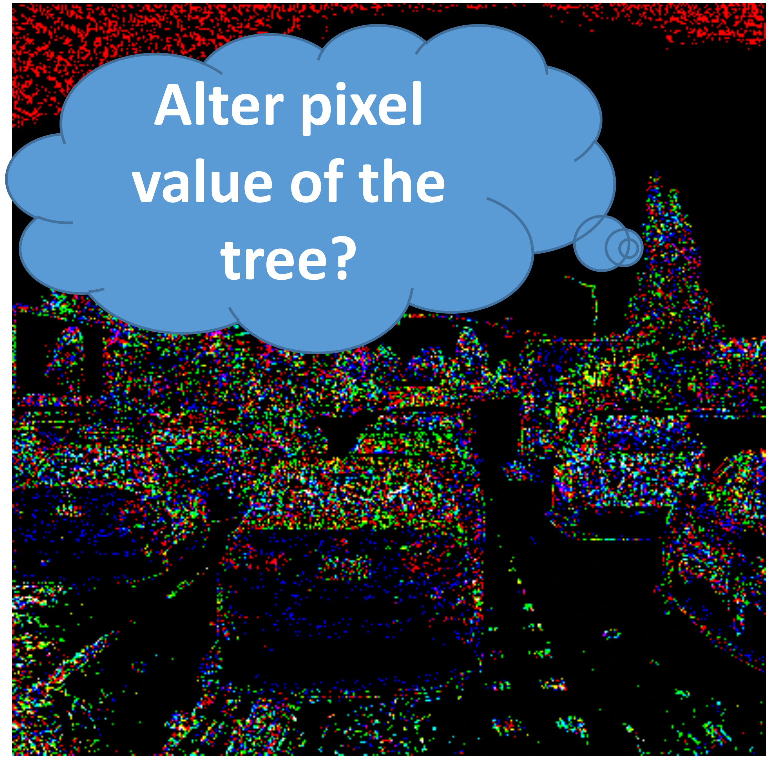
$CW2$ L_2 **based Perturbation Method** **Intuition: many small changes to many pixels** $minimize \; \| x - x' \|_2^2 + c \cdot f(x')$

Still in Digital Context

Digital Perturbations Realistic Enough?



FGSM



JSMA

CW2

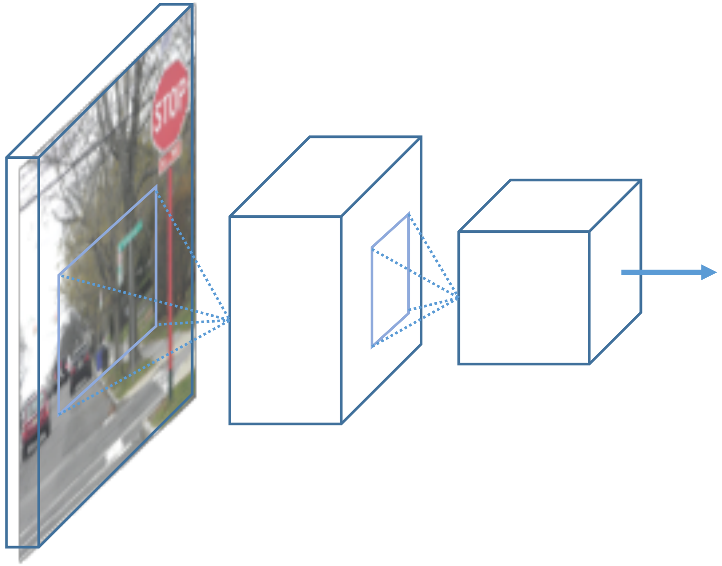
Feasible

Explore Chances of Physical White Box Attack against YOLOv3

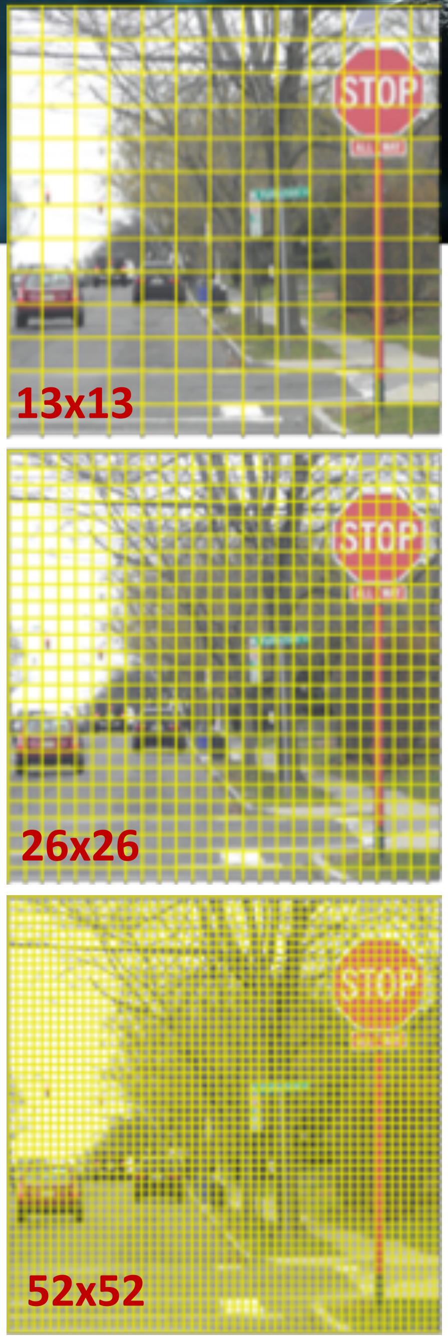
Identify Opportunities by Completely Understanding YOLOv3 Inference Mechanism

Deep Dive into YOLOv3

Input
[416x416x3]



YOLO v3
Object Detection Model
[147 Layers, 62M Parameters]



Output
[10,647
Bounding
Boxes]

Image: http://media.nj.com/traffic_impact/photo/all-way-stop-sign-that-flashes-in-montclairjpg-30576ab330660eff.jpg

Training Dataset – MS COCO Dataset

- Common Objects in Context
- 80 Classes: person, [car, truck, bus], [bicycle, motorcycle], [stop sign, traffic light], etc.

black hat EUROPE 2018

YOLOv3 Prediction

Anchor Boxes

p_w

116x 90 p_h

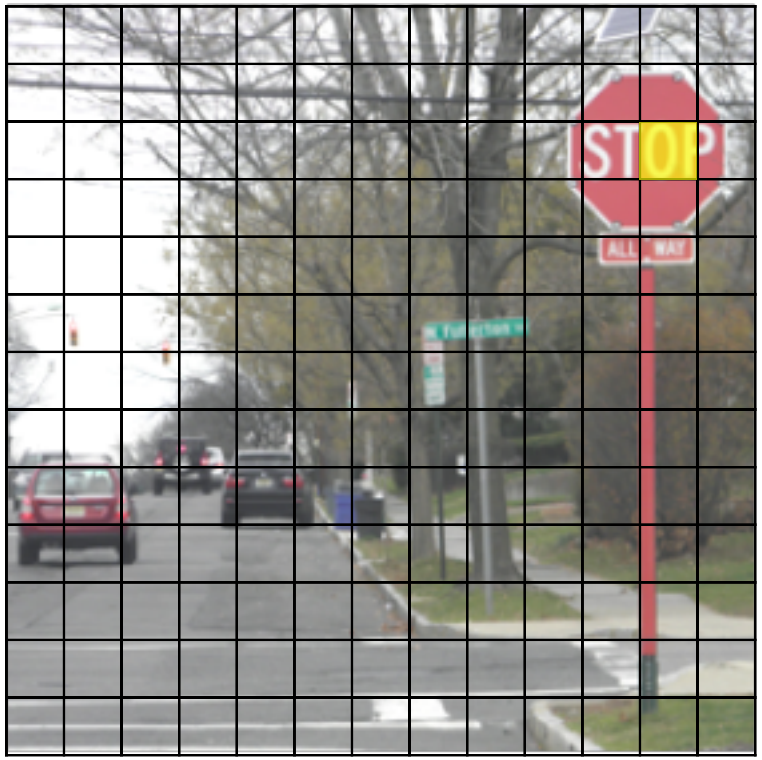
156x 198

373x326

Prediction Vector

t_x	t_y	t_w	t_h	p_{obj}	c_1	c_2	…	…	c_{79}	c_{80}

Bounding Box Objectness 80 Class Confidence



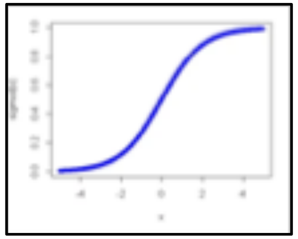
13 x 13 Grid

Center Point

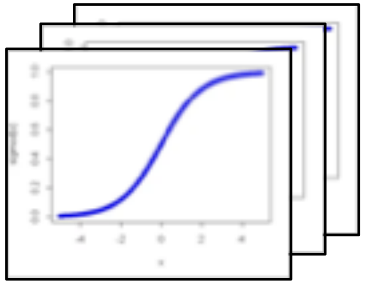
$$(c_x, c_y) = (11,2)$$
$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y$$

Object Size

$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$



×



stop sign 99%

car 0.01%

Threat Model : Physical Image Patch Attack

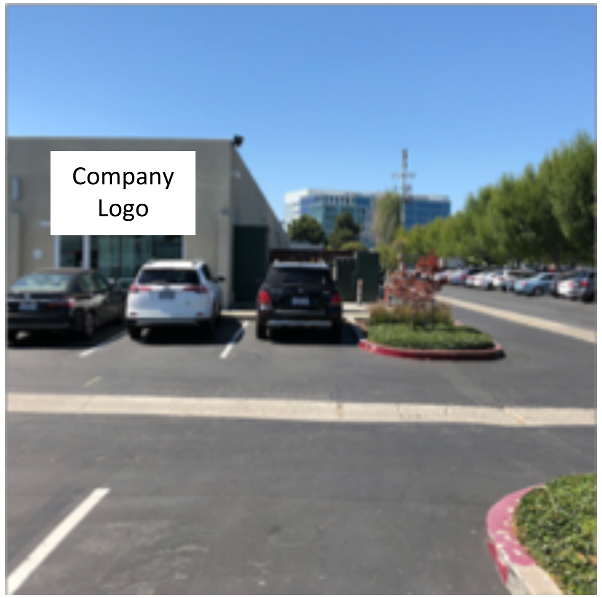
Image Patches

Our Physical Attack Approach & Objectives

- Input Patch Construction
  - Differentiable to craft adversarial examples
- Attack Objectives
  - Make YOLOv3 detect fake object
  - Make object disappear in front of YOLOv3

Differentiable Input Patch Construction

Resize → Perspective Transformation



Our Physical Attack Approach & Objectives

- Input Patch Construction
  - Differentiable to craft adversarial examples

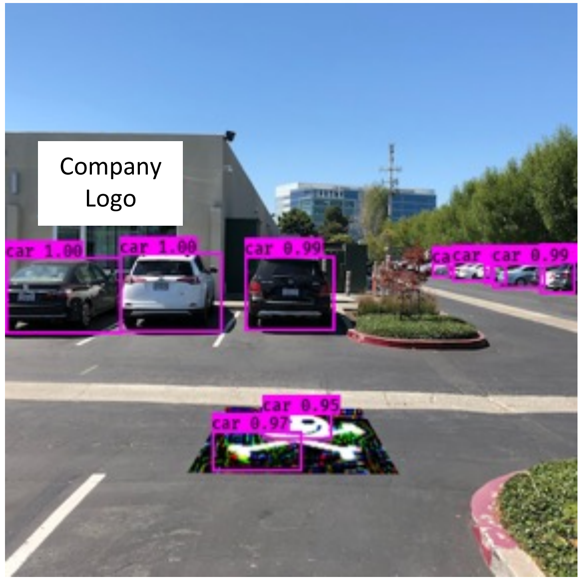
- Attack Objectives
  - Object Fabrication: make YOLOv3 detect fake object
  - Object Vanishing: make object disappear in front of YOLOv3

Attack Objective 1 – Object Fabrication

A. Naive Fabrication

- Push more detections towards a certain object

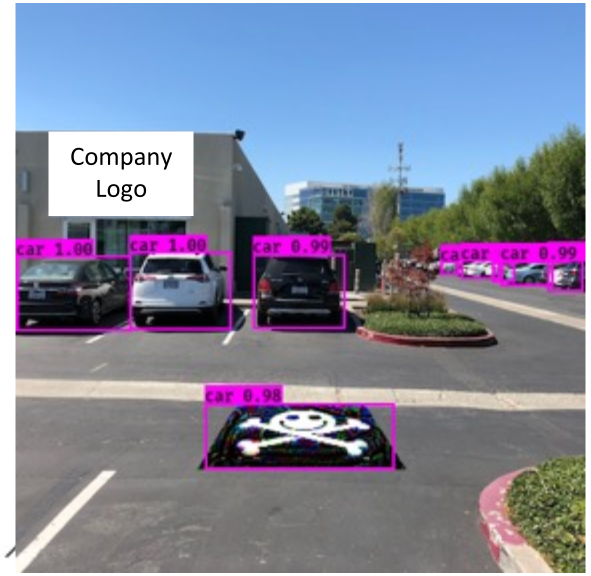
```
tgt_cls_id = self.model.class_names.index("car")
loss_box_class_conf = -tf.reduce_mean(y_box_class_probs[:, tgt_cls_id])
loss_box_conf = -tf.reduce_mean(y_box_confidence)
loss_final = loss_box_class_conf + loss_box_conf
```



B. Precise Fabrication

- Produce fake object at specific location

```
loss_boxes = 0
idx_pred_dict = self.yolo3_calc.calculate_box_preds(x1_y1_x2_y2)
for idx, pred in idx_pred_dict.items():
    loss_boxes += tf.losses.mean_squared_error(pred, y_box_preds[idx])
```

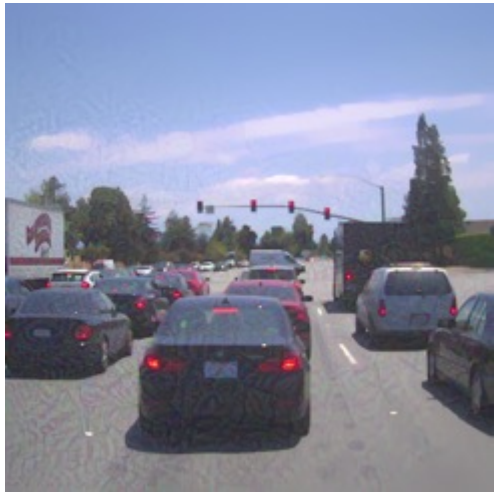


Attack Objective 2 – Object Vanishing

Make a certain object class disappear in the whole image.

```
tgt_cls_id = self.model.class_names.index("car")
loss_box_class_conf = tf.reduce_mean(y_box_class_probs[:, tgt_cls_id])
loss_box_conf = tf.reduce_mean(y_box_confidence)
loss_final = loss_box_class_conf + loss_box_conf
```

Challenges to the Success of Physical Attack

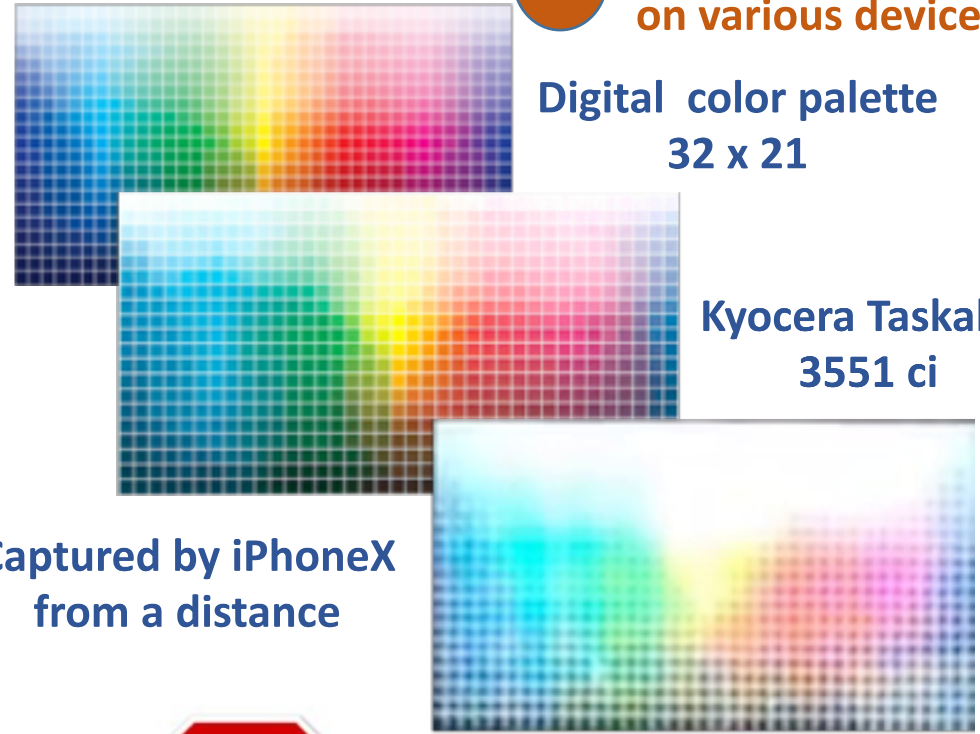


1 Controlled Perturbation Area

2 Object appearance changes at various distances, angles

3 Various Light conditions: e.g. glaring, dimming

4 Color Distortion on various devices



Digital color palette
32 x 21

Kyocera Taskalfa
3551 ci

Captured by iPhoneX from a distance

5 Inaccurate Patch Location

Tactics to the Challenges:

- [**Controlled Perturbation Area**] Image-patch based Attack
- [**Color Distortion**] Color Management with the **N**on-**P**rintability **L**oss (NPS)
- [**Inaccurate Patch**] **R**andom **T**ransformation (RT) during optimization iterations
- [**Various Distances & Angles**] RT + Total Variation regularization instead of Expectation-Over-Transformation
- [**Various Light Condition**] Get a stable environment
- **More** …

Color Management with Non-Printability Loss

Given $P \subset [0,1]^3$, a set of printable RGB triplets. $NPS(\hat{p}) = \prod_{p \in P} |\hat{p} - p|$

For the perturbated δ, NPS(δ) $= \sum_{\hat{p} \in \delta} NPS(\hat{p})$. NPS(δ) ↓, color reproducibility ↑

No NPS

Printed&Captured by iPhone

With NPS

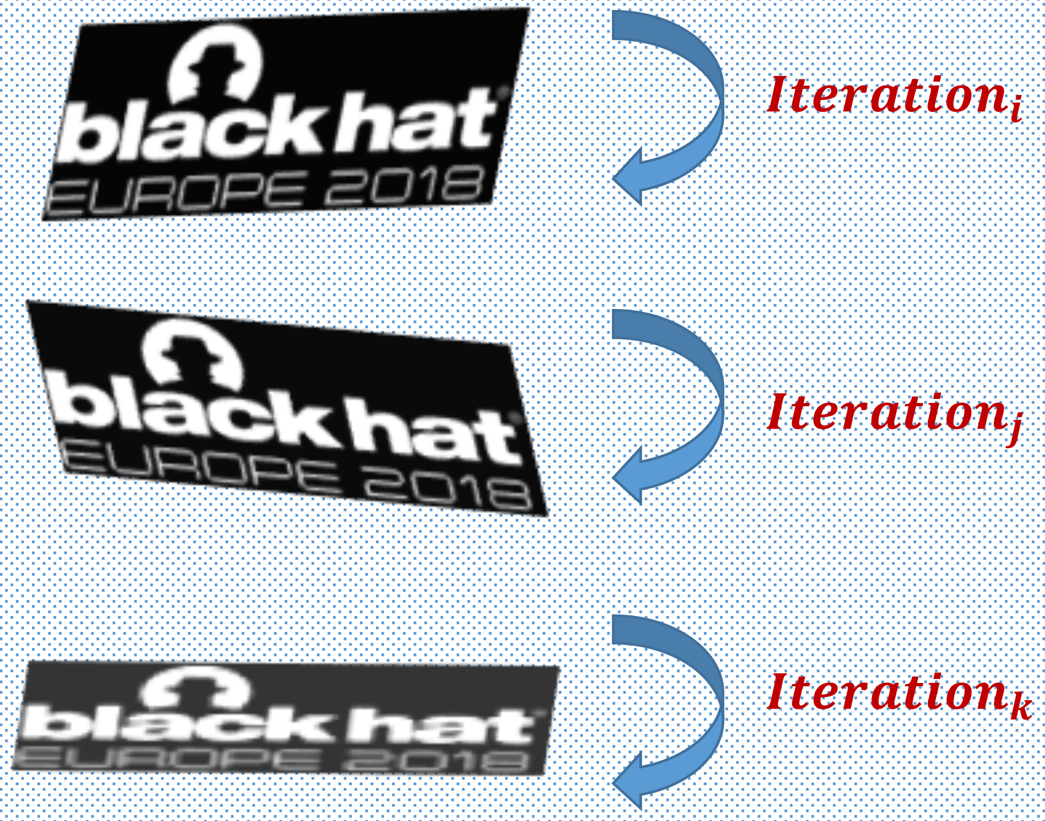
Accessorize to a Crime: Real and Stealthy Attacks on State-Of-The-Art Face Recognition.
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael ReiterIn. In Proceedings of CCS 2016

Random Transformation During Optimization Iterations

Perfectly positioned?

Generated Perturbation Patch

troduce Random Perspective Transformation

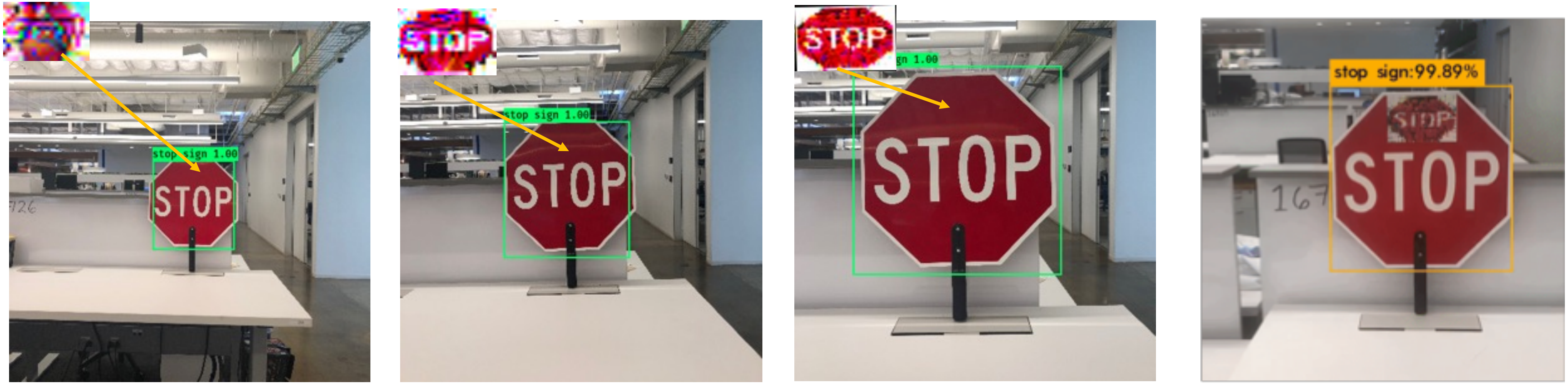


RT + TV for Various Distances & Angles

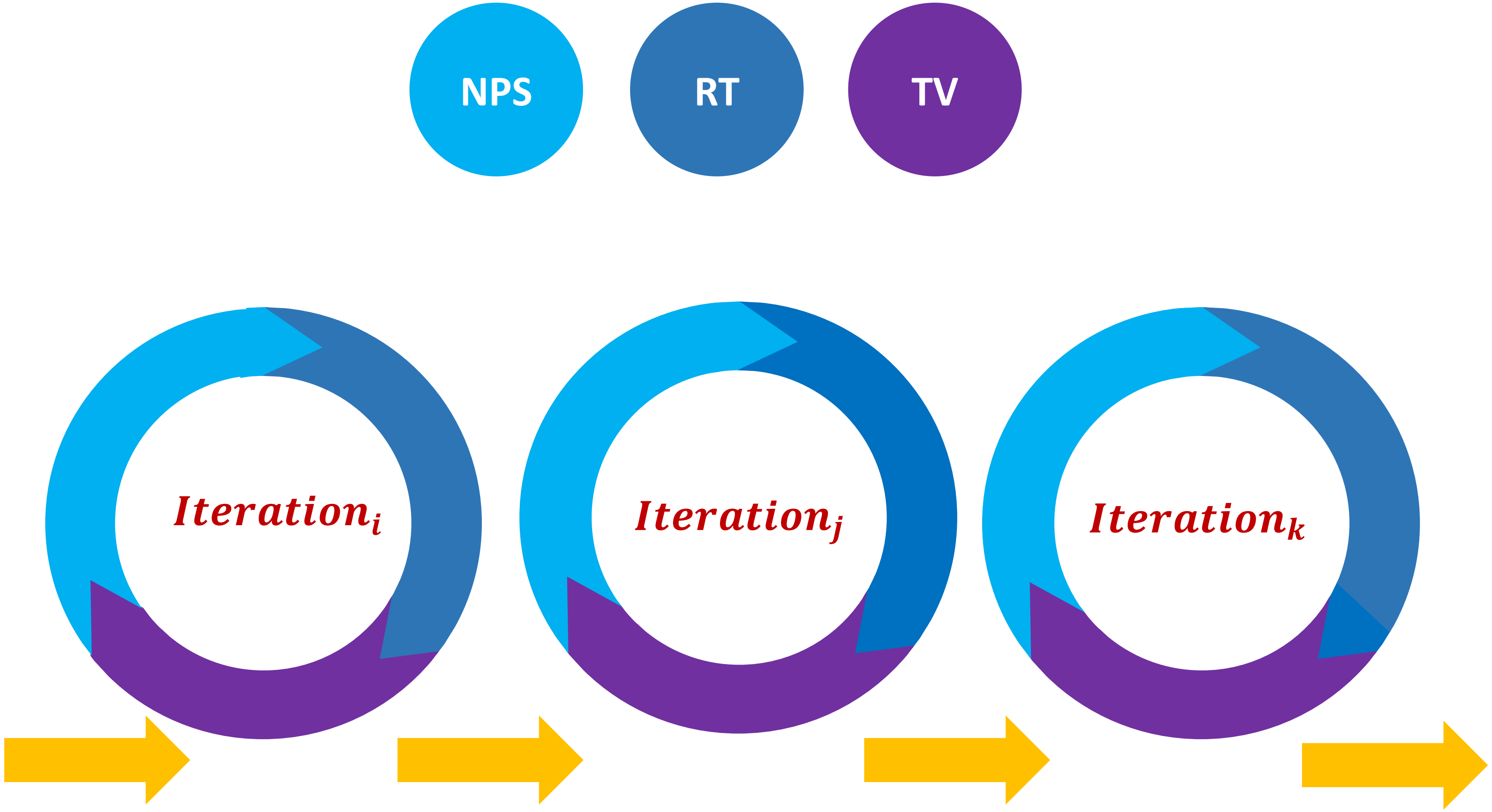
- **Random Transformation + Total Variance Regulation ：**

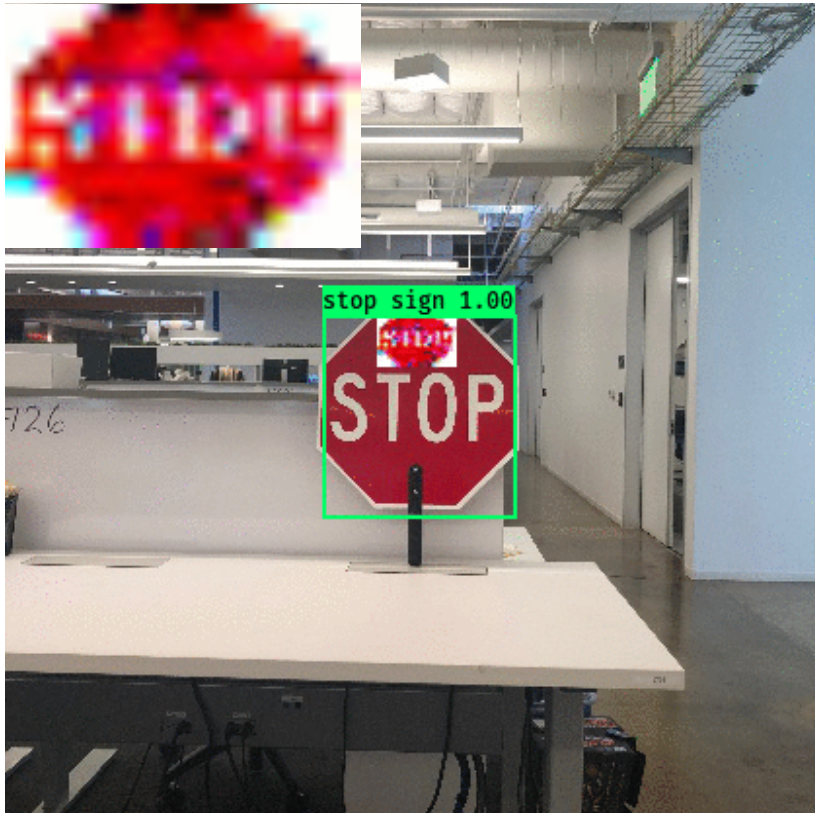
  a different approach from EOT

  Simulate the transformations using RT + TV for various distances & angles instead of drawing from a distribution



Put Everything Together: An Iterative Optimization






DEMO

Conclusion & Takeaway

- **With careful setup, physical attacks are achievable against DNN-based object detection methods in a white box setting**

- **Defense is hard, a good safety and security metric has to be explored**

- **We call out efforts for a robust, adversarial example resistant model that is required in safety critical system like autonomous driving system**

Scan Me