Weighted Average Precision: Adversarial Example Detection for Visual Perception of Autonomous Vehicles

Weiheng Chai, Yantao Lu, Senem Velipasalar

Department of Electrical Engineering and Computer Science
Syracuse University

wchai01@syr.edu, ylu25@syr.edu, svelipas@syr.edu

The information, data, or work presented herein was funded in part by National Science Foundation under Grants 1739748 and 1816732, and by the Advanced Research Projects Agency-Energy, U.S. Department of Energy, under Award Number DE-AR0000940. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Outline

- Background and Motivation
- Proposed Method
- Experimental Results
- Conclusion

Background

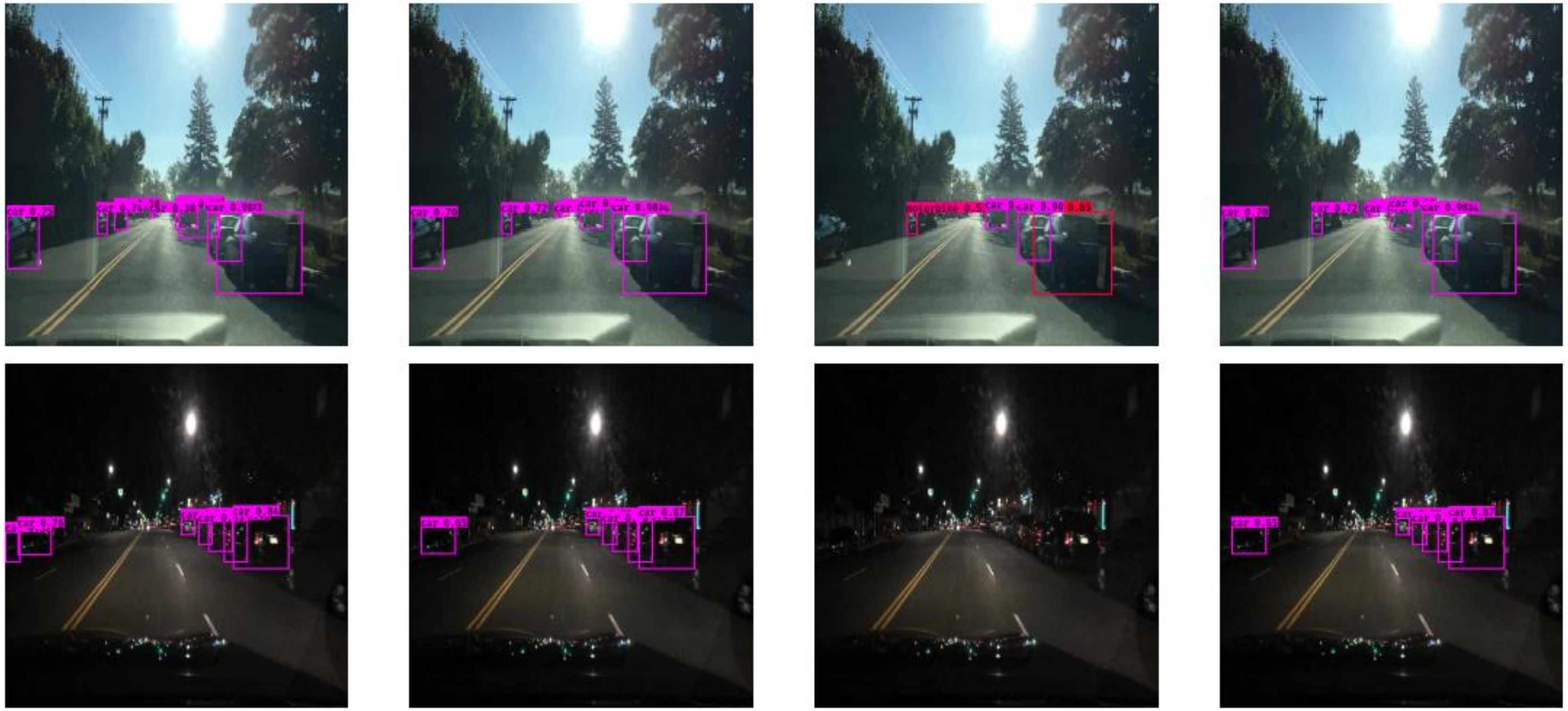
- Neural network-based object detection models are widely employed in autonomous driving perception systems.
- Control successors are highly dependent on the outputs of object detectors. Thus, object detection needs to be reliable for safe autonomous driving.
- Neural network-based object detectors have been shown to be vulnerable to adversarial examples (AEs).
- By adding small perturbations to original images, AEs can deceive victim models, and result in incorrect outputs.
- Most existing studies, focusing on the detection of AEs in autonomous driving applications, either use simplifying assumptions on the outputs of object detectors or ignore the tracking system in the perception pipeline.

Proposed Method

- We present a unified framework to bridge the gap between AE research and real-world autonomous driving systems.
- Our framework extends the AE detection mechanisms developed for image classifiers into the object detection domain by
  - designing a novel similarity distance metric to detect inconsistencies between dense object detection results of two video frames, and
  - adding temporal information into the detection pipeline to further improve accuracy

Proposed Method

It has been shown that the robustness of DNNs to local changes (e.g., squeezing, scale, position) does not generalize to the perturbations added by AEs [1,2,3].



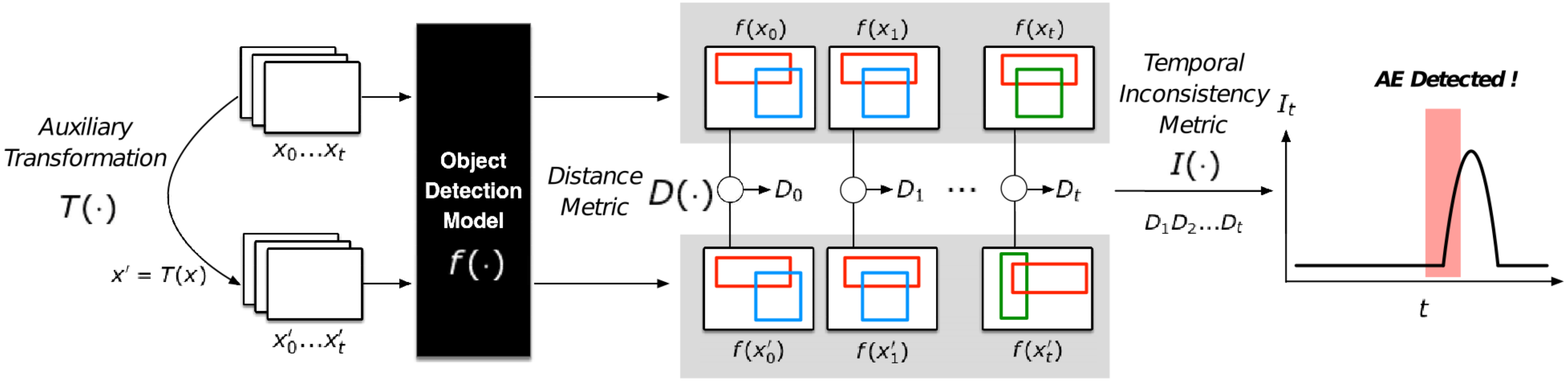
(a) Benign (b) Benign after squeeze (c) Adversarial Example (AE) (d) AE after squeeze

[1] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao, "Foveation-based mechanisms alleviate adversarial examples," arXiv preprint arXiv:1511.06292, 2015.

[2] Weilin Xu, David Evans, and Yanjun Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," arXiv preprint arXiv:1704.01155, 2017.

[3] Dongyu Meng and Hao Chen, "Magnet: a two-pronged defense against adversarial examples," in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017, pp. 135–147.

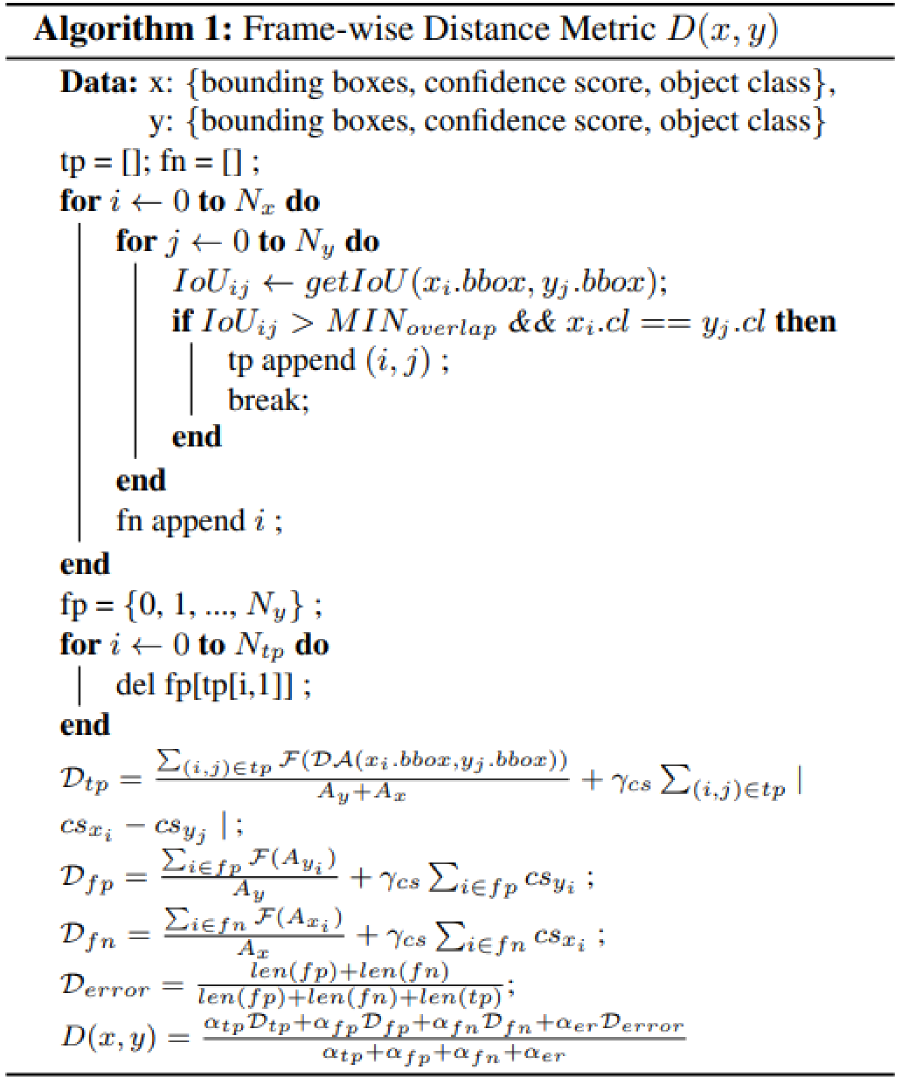
Proposed Method



- For each frame x_t from time t, a transformed image x'_t is generated using an auxiliary transformation $T(x_t)$. If x_t is an AE, it is highly likely that the object detection result $f(x_t)$ will be very different from that of $f(x'_t)$.
- Traditional mAP is not suitable for perturbation-based AE detection.
- We propose a novel distance metric, referred to as the weighted average precision (wAP), to compute D(.)

Proposed Frame-wise Distance Metric

- In the pseudocode, x and y refer to the object detection results from the original image and its transformation, respectively.
- x is considered as the ground truth.
- Based on the IoU value between the bounding boxes of x and y, and an overlap threshold ($\text{MIN}_{\text{overlap}}$), all bounding boxes are classified into true positive (TP), false positive (FP) and false negative (FN) sets.
- $\mathcal{F}(\text{m})=\frac{m}{m+a}$ is a weight function, which makes small bounding boxes contribute less.
- $\mathcal{DA}(\text{A,B})=(\text{A}-\text{B})\cup(\text{B}-\text{A})$ is the sum of the different areas.



Algorithm 1: Frame-wise Distance Metric $D(x, y)$

```
Data: x: {bounding boxes, confidence score, object class},
      y: {bounding boxes, confidence score, object class}
tp = []; fn = [] ;
for $i \leftarrow 0$ to $N_x$ do
    for $j \leftarrow 0$ to $N_y$ do
        $IoU_{ij} \leftarrow getIoU(x_i.bbox, y_j.bbox)$;
        if $IoU_{ij} > MIN_{overlap}$ && $x_i.cl == y_j.cl$ then
            tp append $(i, j)$ ;
            break;
        end
    end
    fn append $i$ ;
end
fp = {0, 1, ..., $N_y$} ;
for $i \leftarrow 0$ to $N_{tp}$ do
    del fp[tp[i,1]] ;
end
$\mathcal{D}_{tp} = \frac{\sum_{(i,j)\in tp} \mathcal{F}(\mathcal{DA}(x_i.bbox, y_j.bbox))}{A_y + A_x} + \gamma_{cs} \sum_{(i,j)\in tp} |$
$cs_{x_i} - cs_{y_j} |$;
$\mathcal{D}_{fp} = \frac{\sum_{i\in fp} \mathcal{F}(A_{y_i})}{A_y} + \gamma_{cs} \sum_{i\in fp} cs_{y_i}$ ;
$\mathcal{D}_{fn} = \frac{\sum_{i\in fn} \mathcal{F}(A_{x_i})}{A_x} + \gamma_{cs} \sum_{i\in fn} cs_{x_i}$ ;
$\mathcal{D}_{error} = \frac{len(fp)+len(fn)}{len(fp)+len(fn)+len(tp)}$;
$D(x, y) = \frac{\alpha_{tp}\mathcal{D}_{tp}+\alpha_{fp}\mathcal{D}_{fp}+\alpha_{fn}\mathcal{D}_{fn}+\alpha_{er}\mathcal{D}_{error}}{\alpha_{tp}+\alpha_{fp}+\alpha_{fn}+\alpha_{er}}$
```

Temporal Consistency

- In a perception system, a tracker will be deleted if it cannot be associated to an object for a duration of R frames.
- Thus, in a set of R consecutive frames, AEs that can successfully attack all R frames are regarded as an effective attack.
- Intermittent attacks will likely fail, since trackers save information from the benign samples.
- Thus, we propose a temporal detection approach, which can be expressed as:

$$I(D_0 \ldots D_t) = \prod_0^t (\mathbb{1}(Di - \mu))$$

where $\mathbb{1}(\cdot)$ is the indicator function returning 1 if $(D_i - \mu) > 0$, and 0 otherwise, and μ is the threshold for the distance metric $D(\cdot)$.

Only when distance metric values for all R single frame attacks are higher than the threshold μ, the temporal metric I(. . .) outputs a 1.

Experimental Setting

- Datasets: BDD10k [4] and Cityscapes [5]
- White-box attack: CW$_{\text{inf}}$ [6]
- Black-box attacks: Momentum Iterative Fast Gradient Sign Method (MIFGSM) [7], Translation Invariant Momentum Diverse Inputs (TI-DIM) [8], Momentum Diverse Inputs Fast Gradient Sign Method (DIM) [9] and Dispersion Reduction (DR) [10]
- Auxiliary image transformation T(.): Bit-wise squeeze[11]
- Object detector: Yolo-v3 [12]
- Baseline metrics: mAP, mIoU

[4] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell, “Bdd100k: A diverse driving video database with scalable annotation tooling,” arXiv preprint arXiv:1805.04687, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213– 3223.

[6] Nicholas Carlini and David Wagner, “Towards evaluating the robustness of neural networks,” in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 39–57.

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu, “Discovering adversarial examples with momentum,” CoRR, abs/1710.06081, 2017, 2017.

[8] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu, “Evading defenses to transferable adversarial examples by translation-invariant attacks,” arXiv preprint arXiv:1904.02884, 2019.

[9] Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan Yuille, “Improving transferability of adversarial examples with input diversity,” arXiv preprint arXiv:1803.06978, 2018.

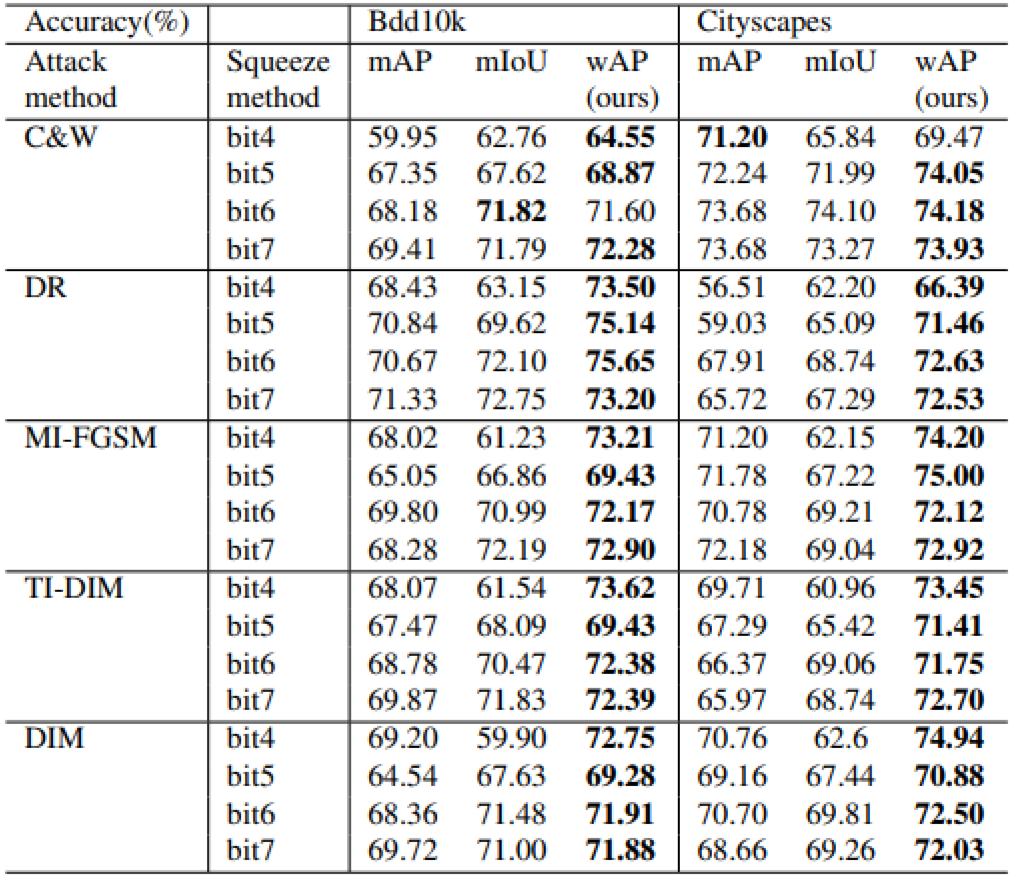
[10] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar, “Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 940–949, 2020.

[11] Weilin Xu, David Evans, and Yanjun Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” arXiv preprint arXiv:1704.01155, 2017.

[12] Joseph Redmon and Ali Farhadi, “Yolov3: An incremental improvement,” arXiv, 2018.

Single-frame AE Detection

- We compute the distance D() between an input image and its version transformed with T(), and use a threshold to decide whether the image is an AE.
- In 38 out of 40 experiment configurations, including black-box and white-box attacks, our approach achieves higher AE detection accuracy.

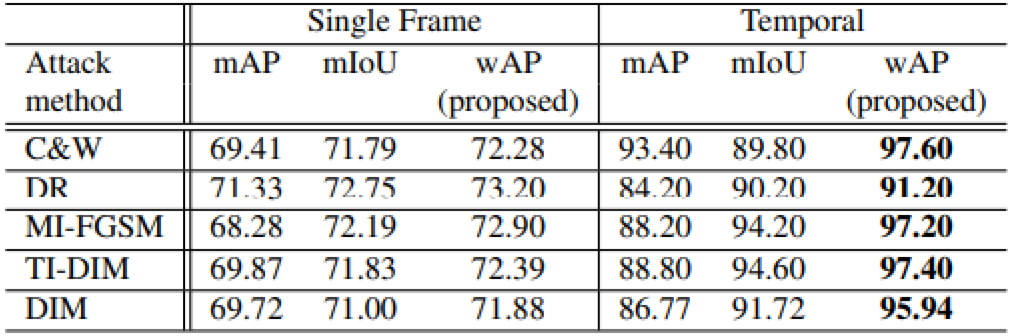


Accuracy(%)		Bdd10k			Cityscapes		
Attack method	Squeeze method	mAP	mIoU	wAP (ours)	mAP	mIoU	wAP (ours)
C&W	bit4	59.95	62.76	**64.55**	**71.20**	65.84	69.47
	bit5	67.35	67.62	**68.87**	72.24	71.99	**74.05**
	bit6	68.18	**71.82**	71.60	73.68	74.10	**74.18**
	bit7	69.41	71.79	**72.28**	73.68	73.27	**73.93**
DR	bit4	68.43	63.15	**73.50**	56.51	62.20	**66.39**
	bit5	70.84	69.62	**75.14**	59.03	65.09	**71.46**
	bit6	70.67	72.10	**75.65**	67.91	68.74	**72.63**
	bit7	71.33	72.75	**73.20**	65.72	67.29	**72.53**
MI-FGSM	bit4	68.02	61.23	**73.21**	71.20	62.15	**74.20**
	bit5	65.05	66.86	**69.43**	71.78	67.22	**75.00**
	bit6	69.80	70.99	**72.17**	70.78	69.21	**72.12**
	bit7	68.28	72.19	**72.90**	72.18	69.04	**72.92**
TI-DIM	bit4	68.07	61.54	**73.62**	69.71	60.96	**73.45**
	bit5	67.47	68.09	**69.43**	67.29	65.42	**71.41**
	bit6	68.78	70.47	**72.38**	66.37	69.06	**71.75**
	bit7	69.87	71.83	**72.39**	65.97	68.74	**72.70**
DIM	bit4	69.20	59.90	**72.75**	70.76	62.6	**74.94**
	bit5	64.54	67.63	**69.28**	69.16	67.44	**70.88**
	bit6	68.36	71.48	**71.91**	70.70	69.81	**72.50**
	bit7	69.72	71.00	**71.88**	68.66	69.26	**72.03**

Table 1

Temporal AE Detection

The state-of-the-art temporal attack can successfully attack a tracking system in 3 consecutive frames [13]. Thus, for this experiment, we set the number of frames for temporal consistency to i = 3. We randomly chose multiples of 3 consecutive frames from videos to insert AEs. Since these intervals can overlap or neighbor each other, the number of consecutive AEs can be greater than or equal to 3. As seen in Tab. 2, our proposed temporal wAP-based detection provides the highest AE detection accuracy for all the attack methods.



	Single Frame			Temporal		
Attack method	mAP	mIoU	wAP (proposed)	mAP	mIoU	wAP (proposed)
C&W	69.41	71.79	72.28	93.40	89.80	**97.60**
DR	71.33	72.75	73.20	84.20	90.20	**91.20**
MI-FGSM	68.28	72.19	72.90	88.20	94.20	**97.20**
TI-DIM	69.87	71.83	72.39	88.80	94.60	**97.40**
DIM	69.72	71.00	71.88	86.77	91.72	**95.94**

Table 2

[13] Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Zhenyu Zhong, and Tao Wei. Fooling detection alone is not enough: First adversarial attack against multiple object tracking. CoRR, abs/1905.11026, 2019.

Conclusion

- We have proposed a new weighted average precision (wAP) distance metric.
- We also proposed a temporal consistency method to improve the detection of AEs for object detection in autonomous driving perception systems.
- The proposed wAP metric focuses on bounding boxes in individual images and can be applied to a sequence of frames to fit into a tracking system.
- The evaluation on two different dataset with different attacks shows that the proposed pipeline greatly enhances the AE detection performance compared to the mAP-based and mIoU-based baselines.

Thank you!