Ensemble-based Blackbox Attacks on Dense Prediction

Zikui Cai, Yaoteng Tan, and M. Salman Asif

TUE-AM-386

University of California, Riverside

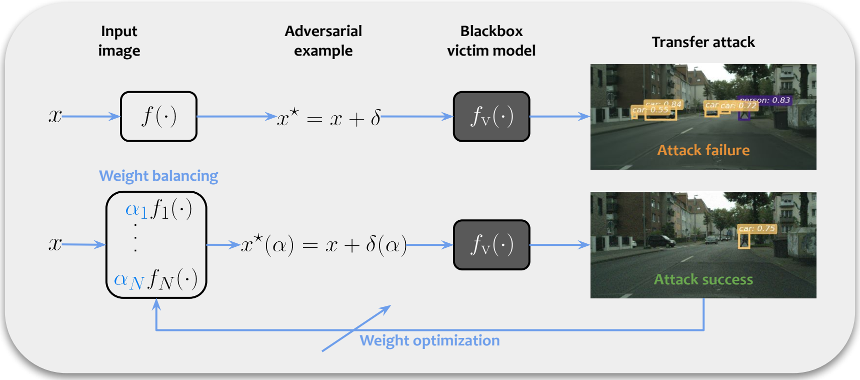
Introduction

- Investigate vulnerabilities of dense prediction models
- Performance
  - Achieves SOTA blackbox attacks on object detection and segmentation
  - Attack multiple tasks at the same time
- Blackbox attacks
  - Transfer-based
  - Query-based
  - Dense prediction is less studied
- Motivation
  - Combine advantages of Transfer- and Query-based attacks
  - Balance the ensemble weights for better whitebox attacks
  - Optimize the ensemble weights according to blackbox feedback

Framework

$$x^\star(\boldsymbol{\alpha}) = \underset{x}{\operatorname{argmax}} \sum_{i=1}^{N} \alpha_i \mathcal{L}_i(f_i(x), y)$$

$$x^\star(\boldsymbol{\alpha}) = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{N} \alpha_i \mathcal{L}_i(f_i(x), y^\star)$$



Weight balancing

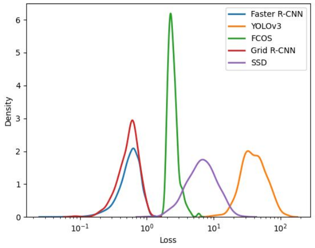
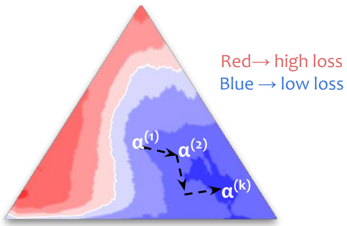


Fig 1. Distribution of loss values for different object detection models

$$\alpha_i = \frac{\sum_{i=1}^{N} \mathcal{L}_i(f_i(x), y^\star)}{N \mathcal{L}_i(f_i(x), y^\star)}$$

Weight optimization [1]



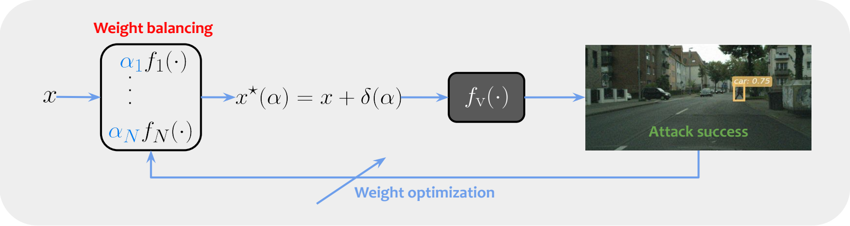
$$\boldsymbol{\alpha}^\star = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\ \mathcal{L}_V(f_V(x^\star(\boldsymbol{\alpha})), y^\star)$$

$$\boldsymbol{\alpha}^\star = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\ \mathcal{L}_{\mathbf{v}}(f_{\mathbf{v}}(x^\star(\boldsymbol{\alpha})), y)$$

Fig 2. Illustration of weight optimization

[1] Cai, Zikui, et al. "Blackbox attacks via surrogate ensemble search." Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

Weight balancing



- Motivation
    - To balance variances in the architectures and loss functions of different dense prediction models

Weight balancing

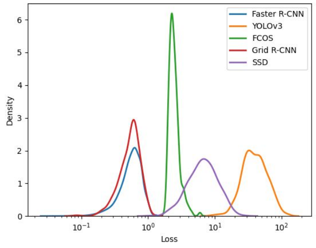
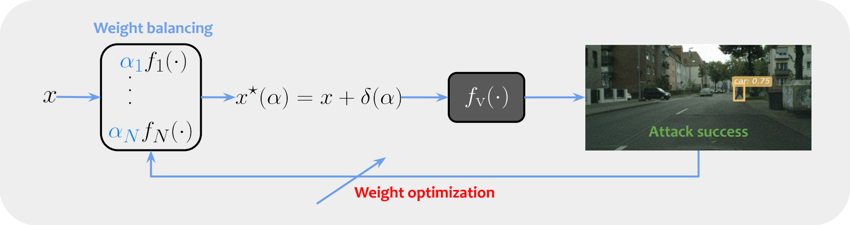


Fig 1. Distribution of loss values for different object detection models

Model Architecture	Training Loss function
Faster R-CNN	$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*)$
FCOS	$L(\{p_{x,y}\},\{t_{x,y}\}) = \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{x,y}, c^*_{x,y}) + \frac{\lambda}{N_{pos}} \sum_{x,y} \mathbb{I}_{\{c^*_{x,y}>0\}} L_{reg}(t_{x,y}, t^*_{x,y})$
SSD	$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g))$
Grid R-CNN	Binary cross-entropy loss
YOLOv3	...

$$\alpha_i = \frac{\sum_{i=1}^{N} \mathcal{L}_i(f_i(x), y^\star)}{N \mathcal{L}_i(f_i(x), y^\star)}$$

Weight optimization



- Motivation
  - Combine transfer-based attacks and query-based attacks
  - BASES [1]

Weight optimization [1]

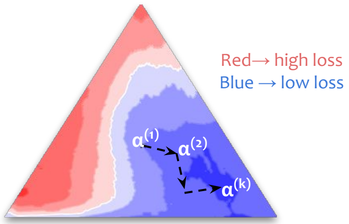


Fig 2. Illustration of weight optimization

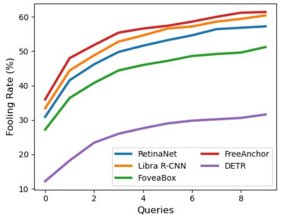
$$\boldsymbol{\alpha}^\star = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\ \mathcal{L}_{\mathbf{v}}(f_{\mathbf{v}}(x^\star(\boldsymbol{\alpha})), y)$$

$$\boldsymbol{\alpha}^\star = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\ \mathcal{L}_{\mathrm{V}}(f_{\mathrm{V}}(x^\star(\boldsymbol{\alpha})), y^\star)$$

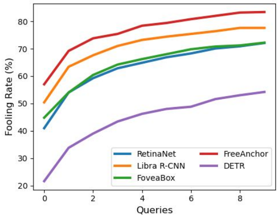
[1] Cai, Zikui, et al. "Blackbox attacks via surrogate ensemble search." Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

Attack on Object Detection

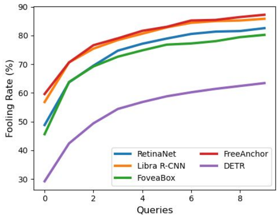
- Fooling rates v.s. Numbers of queries for targeted attack on perturbation budgets {10, 20, 30}.



$\ell_\infty \leq 10$



$\ell_\infty \leq 20$

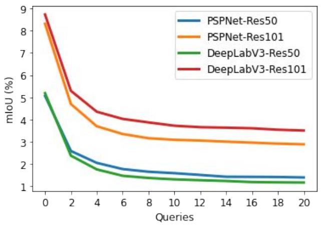


$\ell_\infty \leq 30$

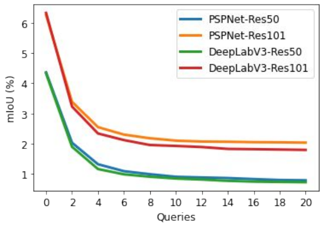
Perturbation Budget	Weight Balancing	Weight Optimization	**Surrogate Ensemble**		**Blackbox Victim Models** (ASR ↑)				
			FRCNN	YOLOv3	Retina	Libra	Fovea	Free	DETR
$\ell_\infty = 10$	✗	✗	27.9	91.5	11.6	9.2	9.0	13.4	5.6
	✗	✓	61.4	**99.4**	24.3	28.0	22.4	31.0	15.4
	✓	✗	71.1	85.7	30.9	33.4	27.2	36.0	12.2
	✓	✓	**86.0**	96.9	**53.2**	**56.6**	**47.2**	**57.4**	**29.0**
$\ell_\infty = 20$	✗	✗	40.1	92.2	16.9	20.4	15.4	23.2	9.7
	✗	✓	77.7	**99.8**	41.0	45.4	37.8	47.0	22.5
	✓	✗	82.7	89.8	41.0	50.4	44.8	57.0	21.6
	✓	✓	**94.6**	98.0	**66.9**	**74.4**	**68.0**	**79.4**	**48.0**

Attack on Semantic Segmentation

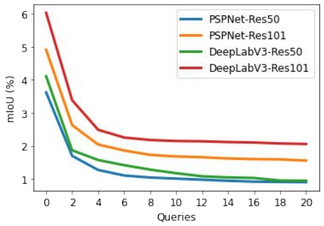
- mIoU v.s. Numbers of queries for untargeted attack on ensemble sizes {2, 4, 6} . Perturbation budgets is fixed to 8/255



$N = 2$



$N = 4$



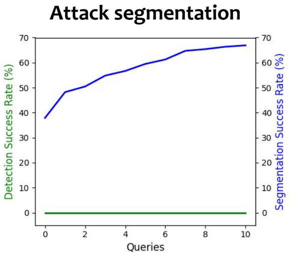
$N = 6$

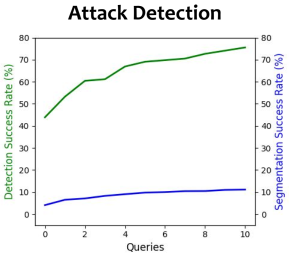
Method	**Whitebox Surrogate**	**Blackbox Victim Models** (mIoU ↓)			
		PSPNet-Res50	PSPNet-Res101	DeepLabV3-Res50	DeepLabV3-Res101
Clean Images	-	77.92	78.28	79.12	77.12
Baseline	PSPNet-Res50	3.43	24.18	5.05	25.74
	DeepLabV3-Res50	4.76	21.72	3.92	22.23
Ours ($Q = 0$)	N=2	5.07	8.32	5.19	8.74
	N=4	4.33	6.26	4.32	6.33
	N=6	3.62	4.91	4.02	4.84
Ours ($Q = 20$)	N=2	1.38	2.88	1.15	3.50
	N=4	**0.79**	2.04	**0.73**	1.80
	N=6	0.90	**1.55**	0.94	**1.09**

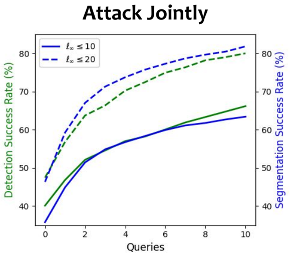
Blue numbers are whitebox attacks

Joint attack on detection and segmentation

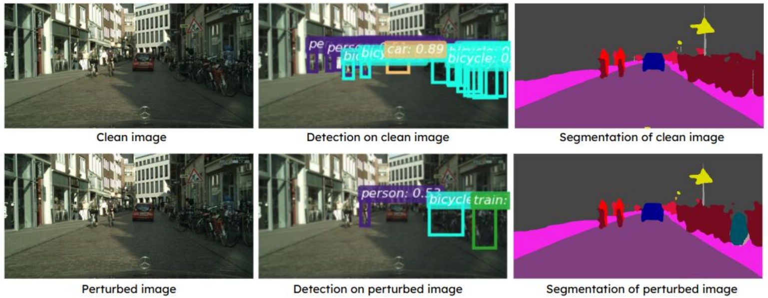
- Some results and comparison

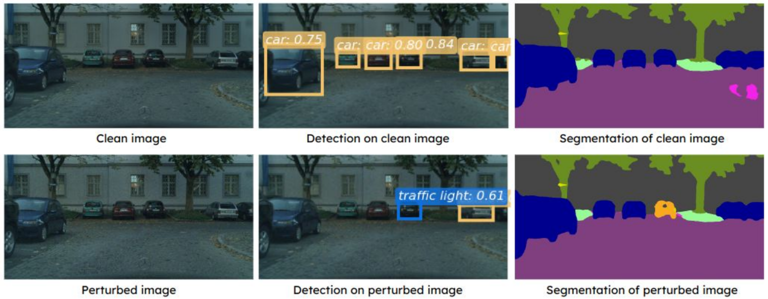






- Some visualizations

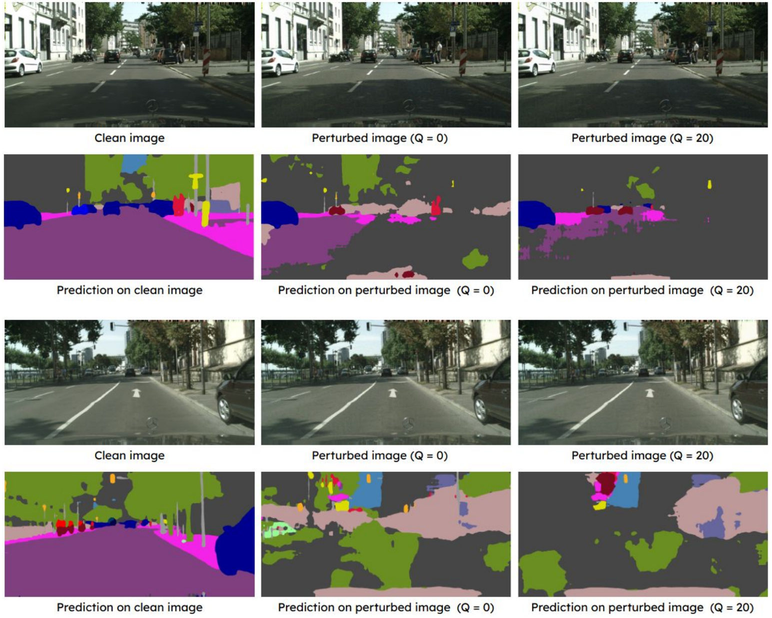




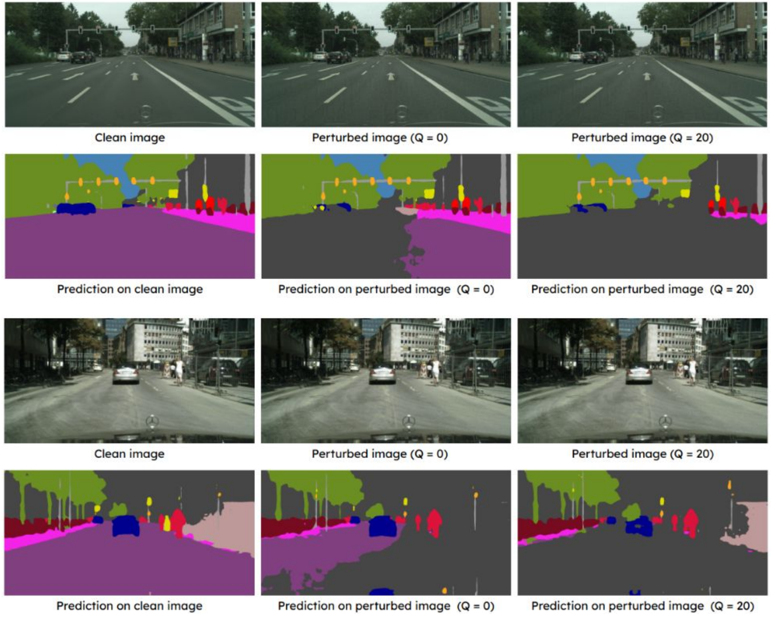
More visualizations

- Visualizations of attacking segmentation

Untargeted attacks



Targeted attacks



Conclusion

- Summary:
  - We propose a new method to generate targeted attacks for dense predictions using an ensemble of surrogate models.
  - We demonstrate that (victim model-agnostic) weight balancing and (victim model-specific) weight optimization can play a critical role in the success of attacks.
- Poster: TUE-AM-386
- Paper: https://arxiv.org/abs/2303.14304
- Code: https://github.com/CSIPlab/EBAD

Acknowledgment: AFOSR award FA9550-21-1-0330, NSF award 2046293, UC Regents Faculty Development grant. Computing support from Nautilus PRP.
More information. Zikui cai (zcai032@ucr.edu), Yaoteng Tan (ytan082@ucr.edu), M. Salman Asif (sasif@ucr.edu)