Ensemble-based Blackbox Attacks on Dense Prediction

Zikui Cai*, Yaoteng Tan*, and M. Salman Asif
University of California Riverside
{zcai032,ytan073,sasif}@ucr.edu

Abstract

We propose an approach for adversarial attacks on dense prediction models (such as object detectors and segmentation). It is well known that the attacks generated by a single surrogate model do not transfer to arbitrary (blackbox) victim models. Furthermore, targeted attacks are often more challenging than the untargeted attacks. In this paper, we show that a carefully designed ensemble can create effective attacks for a number of victim models. In particular, we show that normalization of the weights for individual models plays a critical role in the success of the attacks. We then demonstrate that by adjusting the weights of the ensemble according to the victim model can further improve the performance of the attacks. We performed a number of experiments for object detectors and segmentation to highlight the significance of the our proposed methods. Our proposed ensemble-based method outperforms existing blackbox attack methods for object detection and segmentation. Finally we show that our proposed method can also generate a single perturbation that can fool multiple blackbox detection and segmentation models simultaneously. Code is available at https://github.com/CSIPlab/EBAD.

1. Introduction

Computer vision models (e.g., classification, object detection, segmentation, and depth estimation) are known to be vulnerable to carefully crafted adversarial examples [4, 11, 16, 17, 46]. Creating such adversarial attacks is easy for whitebox models, where the victim model is completely known [14,16,24,37,55]. In contrast, creating adversarial attacks for blackbox models, where the victim model is unknown, remains a challenging task [1, 33, 54]. Most of the existing blackbox attack methods have been developed for classification models [10, 21, 35, 47]. Blackbox attacks for dense prediction models such as object detection and segmentation are relatively less studied [4, 17, 27], and most of the existing ones mainly focus on untargeted attacks [17]. Furthermore, a vast majority of these methods are based on transfer attacks, in which a surrogate (whitebox) model is used to generate the adversarial example that is tested on the victim model. However, the success rate of such transfer-based attacks is often low, especially for targeted attacks [10, 21, 47].

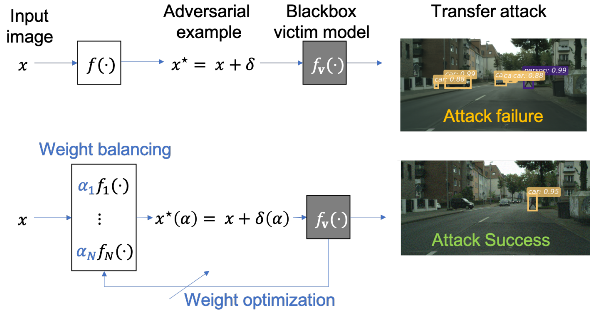


Figure 1. Illustration of the targeted ensemble-based blackbox attack. (Top) Attack generated by a single surrogate model does not transfer on the victim blackbox model (person does not map to car). (Bottom) Attack generated by weight balancing and optimization can transfer on a variety of victim models (person is mapped to car).

In this paper, we propose and evaluate an ensemble-based blackbox attack method for objection detection and segmentation. Our method is inspired by three key observations: 1) targeted attacks generated by a single surrogate model are rarely successful; 2) attacks generated by an ensemble of surrogate models are highly successful if the contribution from all the models is properly normalized; and 3) attacks generated by an ensemble for a specific victim model can be further improved by adjusting the contributions of different surrogate models. The overall idea of the proposed work is illustrated in Fig. 1. Our proposed method can be viewed as a combination of transfer- and query-based attacks, where we can adjust the contribution based on the feedback from the victim model using a small number of queries (5–20 in our experiments). In contrast, conventional query-based attacks require hundreds or thousands of queries from the victim model [9, 19, 22, 49].

We conduct comprehensive experiments to validate our proposed method and achieve state-of-the-art performance for both targeted and untargeted blackbox attacks on ob-

[*] Equal contribution

ject detection. Specifically, our proposed method attains 29–53% success rate using only 5 queries for targeted attacks on object detectors, whereas the current state-of-the-art method [4] achieves 20–39% success rate with the same number of queries. Furthermore, we extend our evaluation to untargeted and targeted attacks on blackbox semantic segmentation models. Our method achieves 0.9–1.55% mIoU for untargeted and 69–95% pixel-wise success for targeted attacks. By comparison, the current state-of-the-art method [17] obtains 0.6–7.97% mIoU for untargeted attacks and does not report results for targeted attacks. To the best of our knowledge, our work is the first approach for targeted and query-based attacks for semantic segmentation.

Below we summarize main contributions of this work.

- We design a novel framework that can effectively attack blackbox dense prediction models based on an ensemble of surrogate models.
- We propose two simple yet highly effective ideas, namely weight balancing and weight optimization, with which we can achieve significantly better attack performance compared to existing methods.
- We extensively evaluate our method for targeted and untargeted attacks on object detection and semantic segmentation models and achieve state-of-the-art results.
- We demonstrate that our proposed method can generate a single perturbation that can fool multiple blackbox detection and segmentation models simultaneously.

2. Related work

Blackbox adversarial attacks. In the context of blackbox attacks, the attacker cannot access the model parameters or compute the gradient via backpropagation. Blackbox attack methods can be broadly divided into two groups: transfer-based [25, 33, 39, 40] and query-based attacks [9, 22, 49]. Transfer-based attacks rely on the assumption that surrogate models share similarities with the victim model, such that an adversarial example generated for the surrogate model can also fool the victim model. Query-based methods generate attacks by searching the adversarial examples space based on the feedback obtained from the victim model through queries. They can often achieve higher success rate but may require a large number of queries.

Ensemble-based attacks. Ensemble-based attacks leverage the idea of transfer attack and assume that if an adversarial example can fool multiple models simultaneously, the chances of fooling an unseen model are higher [14, 33, 56]. Recently, some methods have combined ensemble-based transfer attacks with limited feedback from the victim models to improve the overall success rate [19, 21, 26, 35, 45, 47]. These methods have mainly focused on classification models, and ensemble attacks on dense prediction tasks such as object detection and semantic segmentation are relatively less studied, especially for targeted attacks [52].

Attacks against object detectors and segmentation. Dense (pixel-level) prediction tasks such as object detection and semantic segmentation have higher task complexities [50] compared to classification tasks. Existing attacks on object detectors mainly focus on whitebox setting, although there are a few exceptions [3, 51]. A recent study [4] generates blackbox attacks on object detectors by using a surrogate ensemble and context-aware attack-based queries. Another approach [51] trains a generative model to generate transferable attacks. While some patch-based attacks [32, 44] are effective, the patches are easily noticeable. Recent works [17, 18] have investigated adversarial robustness for semantic segmentation and proposed a transferable untargeted attack using a single surrogate model. While most existing methods are based on a single surrogate model, we demonstrate that using multiple surrogates with weight balancing/search in the attack generation process, we can generate more effective adversarial examples for both untargeted and targeted scenarios, as well as for various types of dense prediction tasks.

3. Method

3.1. Preliminaries

We consider a per-instance attack scenario in which we generate adversarial perturbation δ for a given image x. To keep the perturbation imperceptible, we bound its ℓ_p norm as $\|\delta\|_p \leq \varepsilon$. In our experiments, we mainly use ℓ_∞ or max norm that limits the maximum level of perturbation. Our goal is to find δ such that the perturbed image, $x^\star = x + \delta$, can disrupt a victim image recognition system $f_\mathbf{v}$ to make wrong predictions. Suppose the original prediction for the clean image x is $y = f_\mathbf{v}(x)$. The attack goal is $f(x^\star) \neq y$ for untargeted attack, and $f(x^\star) = y^\star$ for targeted attack, where $y^\star$ is the desired output (e.g., label or bounding box or segmentation map).

For classification models, the label $y \in \mathbb{R}$ is a scalar. However, dense prediction models can have more complex output space. For object detection, the variable-length output $y \in \mathbb{R}^{K\times 6}$, where K is the number of detected objects, and each object label and position are encoded in a vector of length 6 that include the object category, bounding box coordinates, and confidence score. Some other tasks like keypoint detection and OCR are similar to object detection. For semantic segmentation, the prediction $y \in \mathbb{R}^{H\times W}$ is per-pixel classification, where H and W are the height and width of the input image, respectively. Depth and optical flow estimation tasks have similar output structure.

The adversarial loss functions for object detection and semantic segmentation can be defined using their respective training or prediction loss functions. Let us consider a whitebox model f and an input image x with output

$y = f(x)$. For untargeted attack, we can search for the adversarial example $x^\star$ by solving the following maximization problem:

$$x^\star = \operatorname*{\mathbf{arg\,max}}_x \; \mathcal{L}(f(x), y), \tag{1}$$

where $\mathcal{L}(f(x), y)$ represents the training loss of the model with input x and output y. For targeted attacks, with a target output $y^\star$, we solve the following minimization problem:

$$x^\star = \operatorname*{\mathbf{arg\,min}}_x \; \mathcal{L}(f(x), y^\star). \tag{2}$$

Different from classification, which mostly use cross-entropy loss across different models, dense predictions have different loss functions for different models due to the complexity of the output space and diversity of the architectures. For example, two-stage object detector, including Faster RCNN [43], has losses for object classification, bounding box regression, and losses on the region proposal network (RPN). But for one-stage object detectors like YOLO [41,42], they do not have losses corresponding to RPN. Due to the large variability of the loss functions used in different dense prediction models, we use the corresponding training loss $\mathcal{L}$ for each model as the optimization loss to guide the backpropagation.

We employ PGD [37] to optimize the perturbation as

$$\delta^{t+1} = \Pi_\varepsilon \left(\delta^t - \lambda \,\mathbf{sign}(\nabla_\delta \mathcal{L}(f(x+\delta^t), y^\star))\right), \tag{3}$$

for targeted attack and

$$\delta^{t+1} = \Pi_\varepsilon \left(\delta^t + \lambda \,\mathbf{sign}(\nabla_\delta \mathcal{L}(f(x+\delta^t), y))\right), \tag{4}$$

for untargeted attack. Here t indicates the attack step, λ is the step size, and Π_ε projects the perturbation into a ℓ_p norm ball with radius ε. In the rest of the paper, we focus on targeted attacks without loss of generalization.

3.2. Ensemble-based attacks

In an ensemble-based transfer attack, we use an ensemble of N surrogate (whitebox) models: $\mathcal{F} = \{f_1, \dots, f_N\}$ to generate perturbations to attack the victim model $f_\mathbf{v}$. Note that if the ensemble has a single model, then such an attack becomes a simple transfer attack with a single surrogate model. Let us denote the training loss function for ith model as $\mathcal{L}_i(f_i(x), y^*)$. A natural approach to combine the loss functions of all surrogate models is to compute an average or weighted average of the individual loss functions. For instance, we can generate the adversarial image by solving the following optimization problem:

$$x^\star(\boldsymbol{\alpha}) = \operatorname*{\mathbf{arg\,min}}_x \sum_{i=1}^{N} \alpha_i \mathcal{L}_i(f_i(x), y^\star), \tag{5}$$

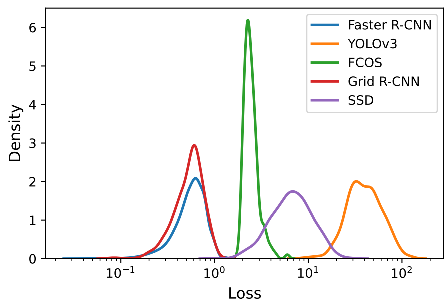


Figure 2. Distribution of losses for different object detection models. $\mathbb{P}(\mathcal{L}_i(f_i(x), y^\star))$. Calculated on 500 images from VOC dataset.

where $x^\star(\boldsymbol{\alpha})$ is a function of the weights of the ensemble $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_N\}$. One of our key observations is that the choice of weights plays a critical role in the transfer attack success rate of the ensemble models.

Weight balancing (victim model agnostic). In ensemble-based transfer attacks, we build on the intuition that if an adversarial example can fool all models simultaneously, it would potentially be more transferable to any unseen victim model. This concept has been empirically corroborated by numerous works [14, 33]. However, most attack methods have only been verified on classification models, all of which use the same cross-entropy loss and yield similar loss values. In contrast, the loss functions for object detectors in an ensemble can differ significantly and cover a large range of values (as shown in Fig. 2). In such cases, models with large loss terms heavily influence the optimization procedure, reducing the attack success rate for models with small losses (see Tab. 1). To overcome this issue, we propose a simple yet effective solution to balance the weights assigned to each model in the ensemble as follows. For each input image x and target output $y^\star$, we adjust the weight for ith surrogate model loss as

$$\alpha_i = \frac{\sum_{i=1}^{N} \mathcal{L}_i(f_i(x), y^\star)}{N \mathcal{L}_i(f_i(x), y^\star)}. \tag{6}$$

The weights are adjusted in a whitebox setting as it allows us to measure the loss of each whitebox model accurately. The purpose of weight balancing is to ensure that all surrogate models can be successfully attacked, making the generated example more adversarial for blackbox victim models.

Weight optimization (victim model specific). Note that the weight normalization, as discussed above, is agnostic to the victim model. We further observe that such transfer-based attacks can be further improved by optimizing the weights of the ensemble according to the victim model, input image, and target output. In particular, we can change the individual α_i to create the perturbations that reduce the victim model loss $\mathcal{L}_\mathbf{v}$. To achieve this goal, we need to solve the following optimization problem with respect to $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^\star = \operatorname*{\mathbf{arg\,min}}_{\boldsymbol{\alpha}} \; \mathcal{L}_\mathbf{v}(f_\mathbf{v}(x^\star(\boldsymbol{\alpha})), y^\star). \tag{7}$$

The optimization problem in (7) is a nested optimization that we can solve as an alternating minimization routine.

Step 0. Given input x, output $y^\star$, and surrogate ensemble $\mathcal{F}$, we initialize $\boldsymbol{\alpha}$ using (6).

Step 1. Solve (5) to generate an adversarial example $x^\star(\boldsymbol{\alpha})$.

Step 2. Test the victim model. Stop if attack is successful; otherwise, change one of the α_i and repeat Step 1.

In our experiments, we update the α_i in a cyclic manner (one coordinate at a time) as $\alpha_i \pm \gamma$ in **Step 2**, where γ denotes a step size. In every round, we select the value of α_i that provides smallest value of the victim loss. We count the number of queries as the number of times we test the generated adversarial example on the victim model and denote it as Q in our experiments.

4. Experiments

To evaluate the effectiveness of our method, we performed extensive experiments on attacking various object detection and semantic segmentation models. We first show that the attacks generated by a single surrogate model fail to transfer to arbitrary victim models. Then we show that the attack transfer rate can be increased by using an ensemble with weight balancing. Additional optimization of the weights surrogates for each victim model can further improve the attack performance. Finally, we show that we can generate single perturbations to fool object detectors and semantic segmentation models simultaneously.

4.1. Experiment setup

4.1.1 Object detection

Models and datasets. We utilize `MMDetection` [7] toolbox to select various model architectures and weights pre-trained on COCO 2017 dataset [30]. To construct the surrogate ensemble, we start with two widely used models, `Faster R-CNN` [43] and `YOLO` [41, 42], and expand the ensemble by appending models with different architectures, including {`FCOS` [48], `Grid R-CNN` [36], `SSD` [31]}. We select different victim models, including {`RetinaNet` [29], `Libra R-CNN` [38], `FoveaBox` [23], `FreeAnchor` [59], `DETR` [6]}. We evaluate attack performance on COCO 2017 [30] and Pascal VOC 2007 [15] datasets. Since the models from this repository are trained on COCO, which contains 80 object categories (a superset of VOC dataset's 20 categories), while testing on VOC dataset, we only return the objects that exist in VOC. We follow the setup in [4] and randomly select 500 images containing multiple (2–6) objects from VOC 2007 test and COCO 2017 validation sets.

Evaluation metrics. We mainly focus on targeted attacks for object detection since they are more challenging than untargeted or vanishing attacks. We measure the performance of the attack using attack success rate (ASR), which equals the number of successfully attacked images over the total number of attacks. We follow the setting in [4], where if the target label is detected within the victim object region with IOU > 0.3, the attack is determined a success.

Perturbation and query budget. We tested different perturbation levels with $\ell_\infty = \{10, 20, 30\}$ out of 255. We use at most 10 queries for attacking object detectors, and we show the trends of how ASR increases with the number of queries. To align with [4] that uses 5 attack plans, we set the maximum query budget to $Q = 5$ in Tab. 1.

Comparing methods. We compare with [4], which is a state-of-the-art transfer-based approach that leverages context information to design attack plans to iteratively attack the victim object. The method generates different perturbations by iterating over a set of predefined attacks, and the total number of queries is the number of attempted attacks. BASES [3] is a recent work on ensemble-based blackbox attacks, which mainly focused on classification tasks and did not consider the loss distributions of different surrogate models. In our experiments, the ensemble with weight optimization and without balancing is equivalent to BASES [3].

4.1.2 Semantic segmentation

Models and datasets. We use `MMSegmentation` [12] toolbox to select different model architectures and weights pre-trained on Cityscapes [13] ($x \in \mathbb{R}^{512\times1024\times3}$) and Pascal VOC ($x \in \mathbb{R}^{512\times512\times3}$) datasets. We select `PSPNet` [60] and `DeepLabV3` [8] with `ResNet50` and `ResNet101` [20] backbones as our blackbox victim models. For the surrogate ensemble, we start with the primary semantic segmentation model `FCN` [34], and expand the ensemble with {`UPerNet` [53], `PSANet` [61], `GCNet` [5], `ANN` [62], `EncNet` [57]}. All models are built on `ResNet50` [20] backbone trained with the cross-entropy loss. The loss values across all surrogate models have similar range; therefore, the effect of weight balancing for semantic segmentation is not as significant as it is for object detection. We use validation datasets from Cityscapes [13] and Pascal VOC 2012, which contains 500 and 1499 images with 19 and 21 classes, respectively.

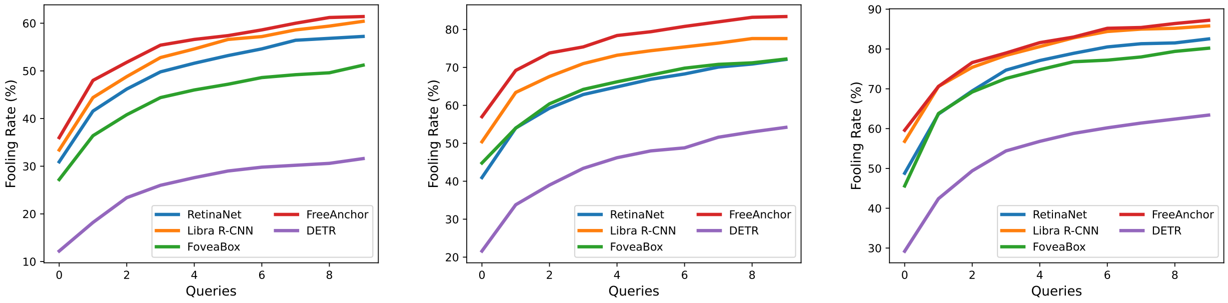
Evaluation metrics. We use different metrics for untargeted and targeted attack performance evaluation. In untargeted experiments, the attack performance is evaluated using the mIoU score (in percentage %), the lower mIoU score the better attack performance. For targeted experiments, we report the pixel success ratio (PSR), which indicates the percentage of pixels successfully assigned the desired label in the target region, the higher the better attack performance.

Perturbation and query budget. We use the perturbation budget $\ell_\infty \leq 8$ out of 255 and query budget $Q = 20$.

Comparing methods. We compare with dynamic scale (DS) attack [17] which is the most recent method that

Table 1. Targeted attack success rate (%) of different methods at different perturbation budgets on VOC dataset. For each perturbation level, the first 4 rows correspond to different settings of our attacks, *i.e.* with (✓) or without (✗) weight balancing and weight optimization. We show comparison with context-aware attack [4], the state-of-the-art method for query-based blackbox attacks.

Perturbation	Weight	Weight	**Surrogate Ensemble**		**Blackbox Victim Models** (ASR ↑)				
Budget	Balancing	Optimization	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**
	Context-aware Attack [4]		55.8	75.6	22.6	20.4	33.6	39.2	20.2
$\ell_\infty = 20$	✗	✗	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**
	Context-aware Attack [4]		78.6	87.2	35.2	38.4	51.6	56.6	34.0
$\ell_\infty = 30$	✗	✗	43.4	91.1	17.1	22.6	17.4	27.2	11.4
	✗	✓	82.7	**99.6**	47.2	54.8	47.0	57.4	33.4
	✓	✗	85.3	90.2	48.8	56.8	45.6	59.6	29.2
	✓	✓	**96.0**	98.1	**78.9**	**82.8**	**76.8**	**83.0**	**58.8**
	Context-aware Attack [4]		80.6	88.0	42.0	44.2	56.8	63.6	40.2



(a) $\ell_\infty \leq 10$ **(b)** $\ell_\infty \leq 20$ **(c)** $\ell_\infty \leq 30$

Figure 3. Attack success rate (or fooling rate) vs number of queries (Q). The maximum value of Q is set to 10 for these results.

achieves the highest attack transfer rate on semantic segmentation untargeted attacks.

4.2. Attacks against object detection

Following settings in [4], we randomly select one object from the output of victim model as the victim object and perturb it into a target object that does not exist in the original detection. This approach rules out the possibility of mis-counting existing objects as the target object.

We report our main results in Tab. 1. The baseline method uses a surrogate ensemble without weight balancing and models are assigned weight of 1. Such a baseline method is same a transfer-based method and results in highly imbalanced success rate for different surrogate models. For instance, at $\ell_\infty \leq 10$, the success rate for `YOLOv3` is above 90% while the success rate for `Faster R-CNN` is less than 30%. Low success rate on surrogate side translates to low success rate on blackbox victim side. The main reason for such imbalance is that the loss of different object detectors can be highly unbalanced (e.g., the loss value for `YOLOv3` is nearly 60× larger than the loss of `Faster RCNN` for targeted attacks, *cf*. Fig. 2). With weight balancing, the success rate increases for surrogate and blackbox victim models. The success rate is further increased on surrogate and victim blackbox models if we optimize the weights, same as BASES [3]. Our method (with weight balancing and optimization) achieves a significantly higher ASR compared to context-aware attack across different datasets and different perturbation budgets. On average, our ASR on blackbox victim models is over 4× better than baseline method and over 1.5× better than context-aware attack. On whitebox surrogate models, weight balancing and optimization also achieves the highest ASR. Context-aware attack fixes weight ratio for surrogate models, $\alpha_{\text{FRCNN}}/\alpha_{\text{YOLO}} = 4$, which is sub-optimal according to our analyses. Even though it achieves much higher performance than baseline, it still largely under-performs our method. Similar trend is observed for COCO dataset (see Tab. S1).

Table 2. Targeted ASR (%) for blackbox victim models and whitebox surrogate models with different ensemble sizes (N). On VOC dataset, $\ell_\infty \leq 10$.

N	**Surrogate Ensemble**					**Blackbox Victim Models** (ASR ↑)				
	FRCNN	YOLOv3	FCOS	Grid R-CNN	SSD	Retina	Libra	Fovea	Free	DETR
1	74.7	-	-	-	-	31.3	31.2	29.8	40.4	10.6
2	86.0	96.9	-	-	-	53.2	56.6	47.2	57.4	29.0
3	87.9	96.1	74.2	-	-	63.1	62.0	57.3	66.6	38.0
4	89.6	94.7	75.2	87.9	-	68.7	**71.0**	67.6	74.4	49.6
5	89.7	91.8	73.5	86.1	82.4	**68.9**	70.2	**68.4**	**77.6**	**53.2**

Table 3. Untargeted attack mIoU scores (%) of ensemble sizes $N = 2, 4, 6$ on Cityscapes dataset. We compare $Q = 0$ (i.e. direct transfer attack) with $Q = 20$ ensemble attack performance. DS uses DeepLabV3-Res50 (DL3-50) as the surrogate model for attack generation; thus the DS on DL3-50 is a whitebox attack. While our method used an ensemble that does not include any victim models for attack generation, we still achieved comparable mIoU scores to DS on DL3-50. Blue numbers represent whitebox attacks.

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
DS [17]	PSPNet-Res50	0.82	8.04	1.36	9.00
	DeepLabV3-Res50	1.23	7.97	0.61	7.11
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**

Table 4. Targeted attack performance on Cityscapes as pixel success rate (higher the better). The attack performance increases as we increase ensemble size (N) and number of queries for weight optimization (Q). $N = 1$ has zero query. We note PSPNet-Res50 as PSP-r50, and DeepLabV3-Res50 as DL3-r50, similar abbreviations apply to Res101.

Q	N	**Blackbox Victim Models** (PSR ↑)			
		PSP-r50	PSP-r101	DL3-r50	DL3-r101
0	1	39.15	10.21	35.02	7.58
	2	52.15	12.28	47.99	10.59
	3	43.17	11.34	42.10	9.87
	4	51.44	26.13	49.14	17.42
	5	52.24	23.88	51.75	16.08
20	2	83.97	51.80	82.70	46.95
	3	88.88	64.63	85.55	60.88
	4	91.51	64.28	87.19	63.88
	5	**92.91**	**69.09**	**88.95**	**69.65**

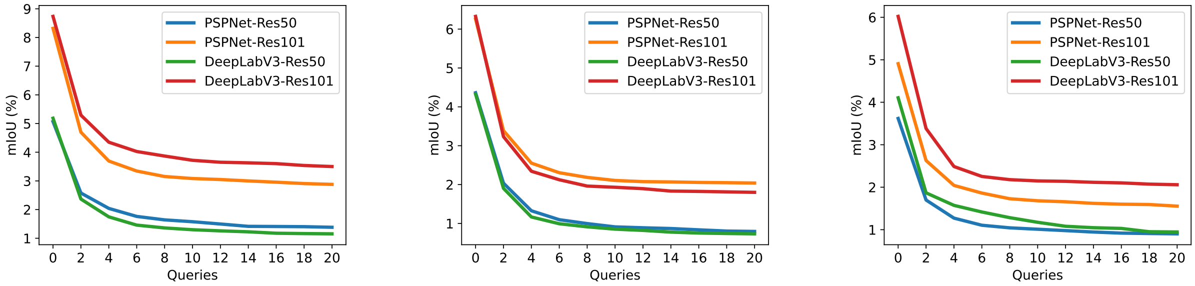
Fig. 3 shows the effect of the number of queries on the ASR that gradually improves as we optimize the weights. We observe the largest increase in the first two steps and then the improvement plateaus as $Q \to 10$.

We also conducted an experiment to test our method with varying ensemble sizes. The results for $\ell_\infty \leq 10$, $Q = 5$ are presented in Tab. 2. As we increase the number of models in the ensemble from $N = 1$ to $N = 5$, we observe an increased ASR on all blackbox victim models.

4.3. Attacks against semantic segmentation

We evaluate the effectiveness of our attack on semantic segmentation in both untargeted and targeted settings. For the sake of consistency and a fair comparison, we adopt adversarial attack settings in DS attack [17].

Untargeted attacks. We generate adversarial attacks using different ensemble sizes and report mIoU scores on Cityscapes in Tab. 3 and Fig. 4 (and Pascal VOC in supplementary material). In the untargeted setting, semantic segmentation models are attacked to maximize the loss between clean and modified annotation; hence, the lower mIoU implies better attack performance. All of the victim models achieve high performance on clean images. The baseline method (direct transfer attack with one surrogate model using PGD) performs well in the whitebox setting but suffers when the victim uses another backbone. For example, the attacks generated on `PSPNet-Res50` achieves 3.43% mIoU on `PSPNet-Res50` but only attains 24.18% mIoU on `PSPNet-Res101`. DS attack achieves better results than the baseline method but still suffers from cross-backbone transfers. On the other hand, our method, without weight optimization (*i.e.*, $Q = 0$) and using a surrogate ensemble of $N = 2$ models, can achieve results comparable to DS attack, particularly for attacks on Res101 models. As we increase the number of surrogate models to 4 or 6, our



(a) $N = 2$ **(b)** $N = 4$ **(c)** $N = 6$

Figure 4. mIoU vs number of queries (Q) for different ensemble sizes (N).

attack performance further improves. Furthermore, when we apply weight optimization (*e.g.*, $Q = 20$), the attack improves by updating the weights of the surrogate models, allowing us to outperform DS attack for all victim models. Fig. 4 shows how the mIoU changes with the number of queries. We observe that the mIoU gradually reduces as we query the victim model and optimize the weights. The largest decrease happens in the first 3–4 steps and then the reduction plateaus as $Q \to 20$.

Targeted attack. To evaluate our method in a more challenging setting, we consider a targeted attack scenario, where instead of changing every pixel in the segmentation to some arbitrary label, we focus on attacking a dominant class (*i.e.*, the class occupying the largest area) in the scene to its least likely class $y^\star$. For each clean image, we first select a region with the dominant class y (*e.g.*, "road" or "building" for most of the Cityscapes images. See Fig. S3 as an example). Then based on the least-likely class of each pixel in that region, we select the class that appears most frequently as the target label $y^\star$ of the entire region. We use PSR as our evaluation metric, which represents the percentage of pixels in the selected region that are successfully assigned to $y^\star$. The higher percentage indicates more pixels are successfully attacked to the desired class, which indicates better attack performance. Our targeted attack results are reported in Tab. 4. Results show that as we increase the number of surrogate models (N), the ASR improves for most instances without any weight optimization step (i.e., $Q = 0$). If we perform weight optimization for $Q = 20$ steps, then the success rate increases for all the models. For instance, with $N = 4$, the ASR for Res101 models increases from 17–26% to 63-64%.

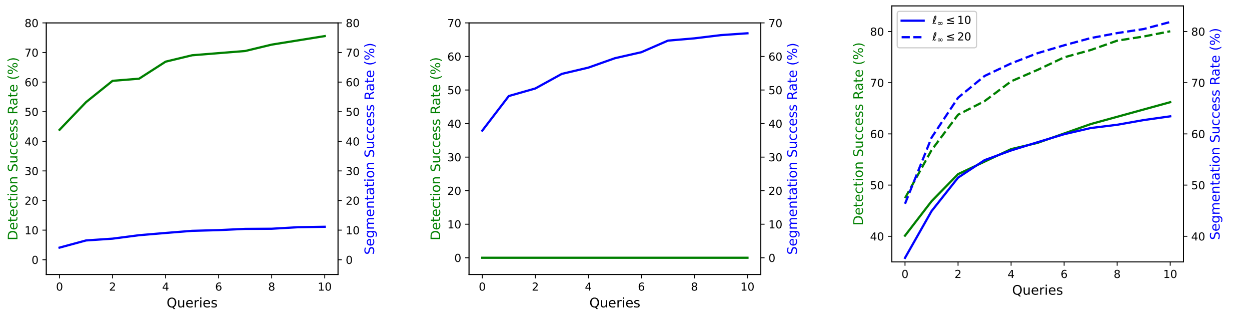
4.4. Joint attack for multiple models and tasks

We first show that generally adversarial examples generated for object detection do not transfer to semantic segmentation, and vice versa. Then we show that we can generate single perturbations to fool object detectors and semantic segmentation models simultaneously, by using a surrogate ensemble including both detection and segmentation models. We choose targeted attacks in our experiments because they are more challenging than untargeted attacks.

Experiment setup. On the blackbox (victim) side, we tested `RetinaNet` as the victim object detector and `PSPNet-Res50` as the victim semantic segmentation model. On the whitebox (surrogate) side, we used `Faster RCNN`, `YOLOv3` as the surrogate object detectors and `FCN`, `UPerNet` as the surrogate semantic segmentation models. We performed targeted attacks on 500 test images selected from the validation set of CityScapes dataset.

Results. We present the ASRs for task-specific and joint attacks in Fig. 5. Green curves denote ASR for object detectors, and blue curves denote PSR for semantic segmentation. Fig. 5a presents the results when we generate attacks using an object detector surrogate ensemble. Note that success rate for victim object detector (`RetinaNet`) increases as we optimize the weights but the success rate for the semantic segmentation model (`PSPNet`) remains small. Similarly, Fig. 5b presents the results when we generate attacks using a segmentation surrogate ensemble. The success rate for the victim semantic segmentation model increases, but the success rate for the object dector remains close to zero. Fig. 5c presents the results when we perform a joint attack using an ensemble that consists of both object detectors and segmentation models. The blackbox ASR is high on both detection and segmentation Fig. 5c, and the attack performance improves as we update the weights of the surrogate models. In Fig. 5c, we show the results for different perturbation budgets, with $\ell_\infty \leq 10$, the success rates on detection and segmentation are between $60\% - 70\%$, which are close to in-domain detection attacks in Fig. 5a and in-domain segmentation attacks Fig. 5b. When we increase the perturbation to $\ell_\infty \leq 20$, the success rate for both detection and segmentation can surpass 80%.

Visualization of adversarial examples. In this example, our goal is to perturb the car in the middle to a traffic light. We assign the target label for car region to traffic light. Fig. 6a shows the results where a single adversarial image generated by the surrogate model can successfully fool the blackbox models `RetinaNet` and `PSPNet`.

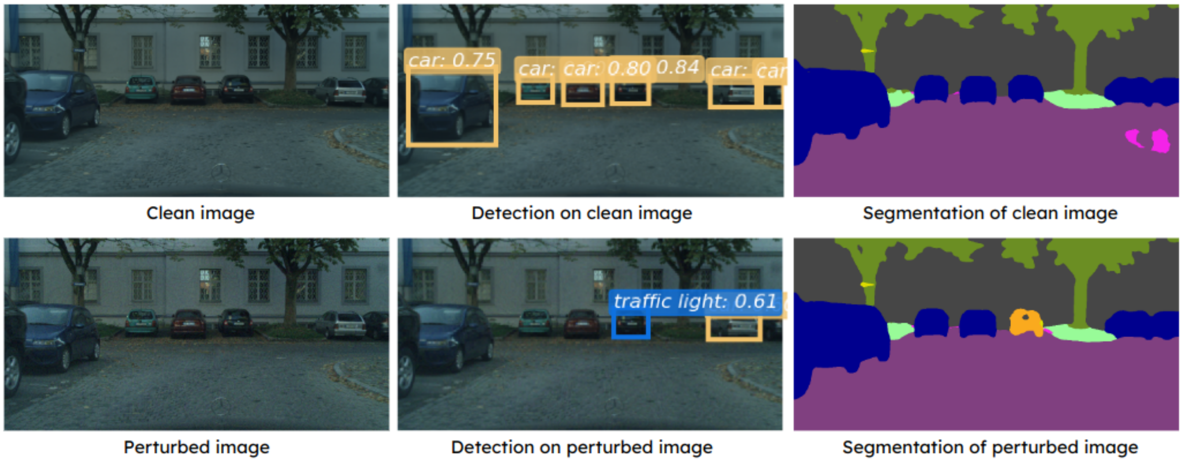


(a) Attacks generated for object detection only, $\ell_\infty \leq 10$

(b) Attacks generated for semantic segmentation only, $\ell_\infty \leq 10$

(c) Attacks generated jointly for object detection and semantic segmentation, $\ell_\infty \leq \{10, 20\}$

Figure 5. Comparison between task-specific attacks and joint attack performance on blackbox object detector (`RetinaNet`) and segmentation model (`PSPNet`). Green curves denote attack success rate for object detectors, and blue curves denote pixel success rate for semantic segmentation. (a) Attacks generated with an object detector surrogate do not transfer for semantic segmentation. (b) Attacks generated with semantic segmentation models surrogate do not transfer for object detectors. (c) Attacks generated by a surrogate of object detectors and semantic segmentation (along with weight balancing and optimization) provide successful attacks for blackbox object detectors and semantic segmentation models.



(a) Our method can generate attacks to fool multiple blackbox object detector and blackbox semantic segmentation models jointly. First row shows a clean image from Cityscapes dataset, detection with `RetinaNet` and segmentation with `PSPNet`. Second row shows the perturbed image using ensemble surrogates {`Faster RCNN`, `YOLOv3`, `FCN`, `UPerNet`}, and detection and segmentation results on the perturbed image. We generate perturbation to map the Car in the middle to Traffic Light. Image id: `lindau_000026_000019`

Void	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic Light	Traffic Sign	Vegetation
Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle

(b) Color encoding for segmentation maps in CityScapes dataset

Figure 6. Visual adversarial examples of our method that generates successful attacks to fool a blackbox object detector and a blackbox semantic segmentation model using a single perturbed image.

5. Conclusion

We presented a new method to generate targeted attacks for dense prediction task (e.g., object detectors and semantic segmentation) using an ensemble of surrogate models. We demonstrate that (victim model-agnostic) weight balancing and (victim model-specific) weight optimization play a critical role in the success of attacks. We present an extensive set of experiments to demonstrate the performance of our method with different models and datasets. Finally, we show that our approach can create adversarial examples to fool multiple blackbox models and tasks jointly.

Limitations. Our method employs an ensemble of surrogate models to generate attacks, which inevitably incurs higher memory and computational overhead. Moreover, the success of our method hinges on the availability of a diverse set of surrogate models, which could potentially limit its efficacy if such models are not readily obtainable.

Acknowledgments. This material is based upon work supported by AFOSR award FA9550-21-1-0330, NSF award 2046293, UC Regents Faculty Development grant. We acknowledge the computing support from Nautilus PRP.

References

[1] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018. 1

[2] Zikui Cai, Shantanu Rane, Alejandro E Brito, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. Zero-query transfer attacks on context-aware object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15024–15034, 2022. 12, 14

[3] Zikui Cai, Chengyu Song, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and M. Salman Asif. Blackbox attacks via surrogate ensemble search. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5

[4] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 149–157, 2022. 1, 2, 4, 5, 12, 14

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 4

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020. 4

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[9] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. 1, 2

[10] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32, 2019. 1

[11] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 514–532. Springer, 2022. 1

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 4

[13] 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*, pages 3213–3223, 2016. 4

[14] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 1, 2, 3

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1

[17] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip Torr. Adversarial examples on segmentation models can be easy to transfer. *arXiv preprint arXiv:2111.11368*, 2021. 1, 2, 4, 6

[18] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022. 2

[19] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019. 1, 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[21] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2019. 1, 2

[22] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018. 1, 2

[23] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 4

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017. 1

[25] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 641–649, 2020. 2

[26] Shasha Li, Abhishek Aich, Shitong Zhu, M. Salman Asif, Chengyu Song, Amit Roy-Chowdhury, and Srikanth Krishnamurthy. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34:2085–2096, 2021. 2

[27] Siyuan Liang, Longkang Li, Yanbo Fan, Xiaojun Jia, Jingzhi Li, Baoyuan Wu, and Xiaochun Cao. A large-scale multiple-objective method for black-box attack against object detection. In *European Conference on Computer Vision*, pages 619–636. Springer, 2022. 1, 12, 14

[28] Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, and Xiaochun Cao. Parallel rectangle flip attack: A query-based black-box attack against object detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7677–7687. IEEE, 2021. 12, 14

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, pages 2980–2988, 2017. 4

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 4

[32] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *AAAI Workshop on Artificial Intelligence Safety*, 2019. 2

[33] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. 1, 2, 3

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[35] Nicholas A. Lord, Romain Mueller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. In *International Conference on Learning Representations*, 2022. 1, 2

[36] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2019. 4

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 3

[38] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019. 4

[39] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 2

[40] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 2

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 3, 4

[42] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3, 4

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 3, 4

[44] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. Role of spatial context in adversarial robustness for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 784–785, 2020. 2

[45] Fnu Suya, Jianfeng Chi, David Evans, and Yuan Tian. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. *USENIX Security Symposium*, 2019. 2

[46] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1

[47] Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white-and black-box attacks. *Advances in Neural Information Processing Systems*, 33:4536–4548, 2020. 1, 2

[48] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 4

[49] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019. 1, 2

[50] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2

[51] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object

detection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 954–960, 2019. 2

[52] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[53] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 4

[54] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 1

[55] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 1

[56] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan. Meta gradient adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7748–7757, 2021. 2

[57] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[58] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 12, 14

[59] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. FreeAnchor: Learning to match anchors for visual object detection. In *Neural Information Processing Systems*, pages 147–155, 2019. 4

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 4

[61] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 4

[62] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019. 4