Evaluating the Adversarial Robustness of Semantic Segmentation: Trying Harder Pays Off

Levente Halmosi[1], Bálint Mohos[1], and Márk Jelasity[1,2]

[1] University of Szeged, Hungary
[2] HUN-REN-SZTE Research Group on AI, Szeged, Hungary

Abstract. Machine learning models are vulnerable to tiny adversarial input perturbations optimized to cause a very large output error. To measure this vulnerability, we need reliable methods that can find such adversarial perturbations. For image classification models, evaluation methodologies have emerged that have stood the test of time. However, we argue that in the area of semantic segmentation, a good approximation of the sensitivity to adversarial perturbations requires *significantly more effort* than what is currently considered satisfactory. To support this claim, we re-evaluate a number of well-known robust segmentation models in an extensive empirical study. We propose new attacks and combine them with the strongest attacks available in the literature. We also analyze the sensitivity of the models in fine detail. The results indicate that most of the state-of-the-art models have a *dramatically larger sensitivity* to adversarial perturbations than previously reported. We also demonstrate a size-bias: small objects are often more easily attacked, even if the large objects are robust, a phenomenon not revealed by current evaluation metrics. Our results also demonstrate that a diverse set of strong attacks is necessary, because different models are often vulnerable to different attacks. Our implementation is available at `https://github.com/szegedai/Robust-Segmentation-Evaluation`.

1 Introduction

It has long been known that deep neural networks (and, in fact, most other machine learning models as well) are sensitive to adversarial perturbation [37,39]. In the case of image processing tasks, this means that—given a network and an input image—an adversary can compute a specific input perturbation that is invisible to the human eye yet changes the output arbitrarily. This is not only a security problem but, more importantly, also a clue that the models trained on image processing tasks have fundamental flaws regarding the feature representations they evolve [18, 24]. In the context of image classification, this problem has received a lot of attention, leading to a large number of attacks under various assumptions (just to mention a few, [7–9, 15, 34]) and defenses (for example, [13, 30]).

In image segmentation, the vulnerability to adversarial perturbation attacks has also been demonstrated many times, for example, [2, 12, 20, 23, 36, 38, 41].

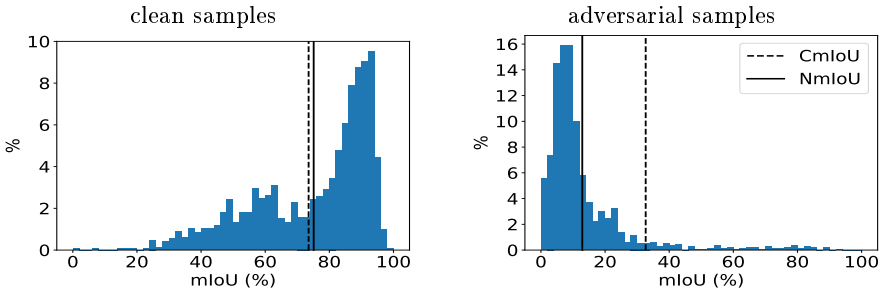

Fig. 1: Single image mIoU distributions of the 'small' adversarially trained model of Croce et al. [17] over the PASCAL VOC 2012 validation set. The image-wise (NmIoU) and class-wise (CmIoU) aggregated mIoU metrics are shown using vertical lines (see Sec. 5). The lack of robustness is apparent and the gap between the two aggregated mIoU metrics for adversarial input indicates a size-bias (see Sec. 5).

However, interestingly, the problem of training models that are robust to adversarial perturbation has not received a lot of attention until recently. The first work that focuses on this problem in depth is by Xu et al. [42] where the DDC-AT method was proposed, followed by the improved SegPGD-AT method by Gu et al. [22]. More recently, Croce et al. [17] also experimented with several configurations for adversarial training.

These methods all use adversarial training [21, 30], a method that has stood the test of time in image classification. On clean input samples, all of the resulting models perform similarly to normally trained models. At the same time, they are reported to have a non-trivial robustness. This is rather surprising, because in image classification, it is well-known that there is a tradeoff between accuracy and robustness [43]. *This motivates our hypothesis that these models are in fact not robust and the current methodology for evaluating robustness is insufficient.*

To test this hypothesis, we perform a thorough empirical evaluation addressing two shortcomings of current practice. First, we apply the combination of the strongest attacks available in the literature, along with our own attacks, to get state-of-the-art upper bounds on the robust performance. The recently proposed attacks in our set include the ALMAProx attack [36] and the Segmentation Ensemble Attack (SEA) [17]. We apply the attack set in an ensemble fashion, attacking each input with all the attacks. We then select the most successful attack for each input, according to a given metric.

Second, we demonstrate that the usual practice of using only pixel accuracy and class-wise aggregated mIoU over the test set hides an important robustness problem, because these metrics are relatively insensitive to the misclassification of small objects. We therefore propose to examine the image-wise average mIoU as well when evaluating robust models, because it balances between smaller and larger objects better, especially over datasets where some classes have only a few instances in most images.

1.1 Contributions

In stark contrast with the results reported previously, using a thorough evaluation methodology we demonstrate that the best-known models proposed in the literature *cannot be considered significantly robust*. We focus on the state-of-the-art adversarial training methods including DDC-AT [42] and SegPGD-AT [22], as well as the approach of Croce et al. [17]. We show that

- using 50% adversarial and 50% clean samples during training—the setup used by DDC-AT and SegPGD-AT—results in a *complete lack of robustness* regardless of which performance metric is considered
- using 100% adversarial samples might result in some robustness according to some of the metrics, but *in terms of the image-wise average mIoU metric these models are also vulnerable*
- the models of Croce et al.—that have the highest robust pixel accuracy among the models we examine—are vulnerable in terms of image-wise average mIoU, which suggests that these models sacrifice the small objects and focus on protecting the large ones (see Fig. 1)
- the *diversity of our attack set is essential*, because different scenarios and models might require different attacks, no attack dominates all the others

Our results indicate that, in the case of semantic segmentation models, a very thorough robustness evaluation is necessary using a diverse set of attacks, and the distribution of the mIoU values over the dataset should also be examined.

1.2 Related work

Here, we overview related work specifically in the area of adversarial robustness in semantic segmentation.

Attacks. Adaptations of the gradient-based adversarial attacks using the segmentation loss function have been proposed relatively early [2, 20, 41]. The Houdini attack by Cissé et al. uses a novel surrogate function more tailored to adversarial example generation [12]. Agnihotri et al. propose the cosine similarity as a surrogate [1]. The ALMAProx attack proposed by Rony et al. [36] defines a constrained optimization problem to find the minimum perturbation to change the prediction over a given proportion of pixels. This is a fairly expensive, yet accurate baseline to evaluate defenses. The dynamic divide-and-conquer (DDC) attack by Xu et al. [42] is based on grouping pixels dynamically during the attack. The segmentation PGD (SegPGD) attack by by Gu et al. [22] is an efficient adaptation of the PGD algorithm for the segmentation task. Finally, the Segmentation Ensemble Attack (SEA) by Croce et al [17] is a collection of four different adaptive attacks with the goal of serving as a reliable ensemble for evaluating robust models.

Special purpose attacks. Some works introduce different versions of the adversarial perturbation problem with specific practical applications in mind. Metzen et al. study universal (input independent) perturbations [33]. Cai et

al. [6] study semantically consistent (context sensitive) attacks that can fool defenses that are based on verifying the semantic consistency of the predicted scene. Chen et al [11] also consider semantic attacks where the predicted scene is still meaningful, only some elements are deleted, for example.

Detection as a defense. Klingner et al. proposed an approach to detect adversarial inputs based on the consistency on different tasks [26]. Another detection approach was suggested by Bär et al. [4, 5] where a specifically designed ensemble of models is applied that involves a dynamic student network that explicitly attempts to become different from another model of the ensemble. While detection approaches are useful in practice, they are not hard defenses because adversarial inputs can be constructed to mislead multiple tasks or multiple models as well simultaneously, as the authors also note.

Multi-tasking as a defense. Another approach involves multi-task networks, with the underlying idea that if a network is trained on several different tasks then it will naturally become more robust [25, 31]. While this is certainly a very promising direction for defense, many uncertainties are involved such as generalizability to different tasks and datasets, and the critical number of tasks.

Adversarial training. In image classification, the most successful approach is adversarial training [30]. In semantic segmentation, a notable application of adversarial training is the DDC-AT algorithm by Xu et al. [42]. More recently, the SegPGD-AT algorithm was proposed by Gu et al. [22], where SegPGD was used to implement adversarial training. Croce et al. [17] have also proposed techniques for adversarial training such as using a robust backbone. These approaches are the subject of our study.

2 Background

Here, we briefly summarize the basic notions of adversarial attacks and adversarial training. We focus on the white-box setting, and we assume that the model to be attacked is differentiable and deterministic. Let the pre-trained model be $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ and the loss function $\mathcal{L}(\theta, x, y) \in \mathbb{R}$ that characterizes the error of the prediction $f_\theta(x)$ given the ground truth output $y \in \mathcal{Y}$.

2.1 Adversarial Attacks

Intuitively, the goal of an adversarial attack is to find a very small perturbation of a given input x in such a way that the prediction of the model $f_\theta(x)$ is completely wrong. Clearly, for an input $x \in \mathcal{X}$ and a perturbation δ we require that $x + \delta \in \mathcal{X}$ as well. For simplicity, we will omit this constraint from the discussion below.

The most common type of attack is based on solving the constrained maximization problem

$$\delta^* = \arg\max_{\delta \in \Delta} \mathcal{L}(\theta, x + \delta, y), \tag{1}$$

which gives us the perturbed input $x + \delta^*$ that *causes the most damage* in terms of the loss function *within a small domain* Δ. Here, the set Δ captures the idea

of "very small perturbation". Throughout the paper, we adopt the widely used definition $\Delta_\epsilon = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ that defines the neighborhood of an input x in terms of the maximum absolute difference in any tensor value.

Another possible type of attack is defined by the constrained minimization problem

$$\delta^* = \arg\min \|\delta\|_\infty \quad \text{s.t.} \quad \mathcal{C}(f_\theta(x+\delta), f_\theta(x)) > 0, \tag{2}$$

where the function $\mathcal{C}$ expresses the amount of damage. This allows the definition of constraints that, for example, require that the predicted label is wrong (in classification) or that 99% of the predicted pixel labels are wrong (in segmentation). Many early approaches adopted this formalism, for example, [7, 34, 39].

Note that Eq. (2) does not guarantee the perturbation to stay within a certain small domain Δ. Since here, we are interested in perturbations from a given Δ_ϵ, we will use the clipped perturbation $\max(\delta^*, \epsilon\frac{\delta^*}{\|\delta^*\|})$ in the case of the attacks that solve the minimization problem in Eq. (2).

2.2 Adversarial Training

Adversarial training has proven to be a reliable heuristic solution to achieve robustness [21, 30]. The idea behind adversarial training is to use adversarial examples during training as a form of augmentation. The adversarial examples are always created based on the current model in the given update step. Formally, we wish to solve the following learning task:

$$\theta^* = \arg\min_\theta \mathbb{E}_{p(x,y)}[\max_{\delta\in\Delta} \mathcal{L}(\theta, x+\delta, y)], \tag{3}$$

where $p(x, y)$ is the distribution of the data and we assumed the more usual maximum damage attack formalism.

In the outer minimization (learning) task one can use an arbitrary learning method, and in the inner maximization task one can select any suitable attack to perturb the samples used by the learning algorithm.

3 Our Battery of Segmentation Attacks

Every attack is constrained to the perturbation set $\Delta_\epsilon = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ with $\epsilon = 8/255$, in line with general practice.

We build on the notions in Sec. 2, noting that in semantic segmentation, $\mathcal{X} = [0,1]^{H\times W\times 3}$, that is, the inputs are 3-channel color images of width W and height H, with all the values normalized into the interval $[0, 1]$. The output space $\mathcal{Y}$ is $[0,1]^{H\times W\times C}$, where C is the number of possible categories for each pixel. Note that $\mathcal{Y}$ is the softmax output, which defines a probability distribution for each pixel over the categories that can be used to compute the final segmentation mask by taking the maximum probability category.

3.1 PAdam attacks

We include two attacks of our own: PAdam-CE and PAdam-Cos. PAdam stands for Projected Adam. It is an algorithm similar to projected gradient descent (PGD) [27,30], but it uses the Adam optimizer [3] instead of the vanilla gradient descent used by PGD. PAdam can be thought of as an alternative to APGD [16] for adaptive step-size control. Our two PAdam attacks both use the AMSGrad variant [35] of Adam with 200 iterations and a step size of 2/255, projected onto the feasible solution set Δ after each update like in PGD.

PAdam-CE solves the problem in Eq. (1) using PAdam assuming the cross entropy loss

$$\mathcal{L}_{CE}(\theta, x, y) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} -y_{h,w,c} \log f_\theta(x)_{h,w,c}, \tag{4}$$

while PAdam-Cos solves the problem

$$\delta^* = \arg\min_{\delta \in \Delta} \text{CosSim}(\text{OneHot}(y), F_\theta(x + \delta)) \tag{5}$$

using PAdam, where F_θ computes the logit layer of f_θ, $\text{CosSim}(x, y) = \frac{x \cdot y}{\|x\|_2 \|y\|_2}$, and $\text{OneHot}(y)$ computes the one-hot encoded ground truth label so that the dimensions of the label match the logit tensor dimensions.

PAdam-Cos is not to be confused with CosPGD [1] where cosine similarity is also used but in a different way, combined with cross entropy. Croce et al. [17] found that SEA dominates CosPGD. However, PAdam-Cos is clearly not dominated as we demonstrate later.

3.2 The SEA attack set

We include the four attacks in SEA introduced by Croce et al. [17]. The authors propose an improved version of APGD [16] that uses progressive radius reduction and apply it for 300 iterations. They include four attacks defined by four loss functions: *balanced cross-entropy (SEA-BCE)* as used in SegPGD, *masked cross-entropy (SEA-MCE)*, where pixels that are incorrectly classified are excluded, *Jensen-Shannon divergence (SEA-JSD)* between the softmax output and the one-hot encoded label, and *masked spherical loss (SEA-MSL)*, where the logit of the correct class is minimized, but first the logit is projected on the unit sphere. The parameter settings we used were identical to those in [17].

3.3 Clipped Minimum Perturbation Attacks

We include a number of attacks that solve the problem in Eq. (2), also clipping the result into the bounded perturbation space Δ, as described in Sec. 2.1. The attacks we include are ALMAProx [36], DAG [41], and PDPGD [32].

Proposed by Rony et al. [36], the ALMAProx attack is a proximal gradient method for solving the perturbation minimization problem in Eq. (2), where

the constraint requires a 99% pixel label error. The problem is transformed into an unconstrained problem by moving the constraints into the objective using Lagrangian penalty functions. The parameter settings we used were identical to those in [36].

Xie et al. proposed the dense adversary generation (DAG) algorithm [41], which attempts to attack each pixel using a gradient method that takes into consideration only the correctly predicted pixels in each gradient step. The algorithm terminates when reaching the maximum iteration number, or when all the targeted pixels are successfully attacked. We can use this algorithm in our evaluation framework by recording the perturbation size at termination, along with the ratio of the successfully attacked pixels. The maximum iteration number was set to 200, and we used two step-sizes: 0.001 and 0.003.

Matyasko and Chau proposed the primal-dual proximal gradient descent adversarial attack (PDPGD) [32] that, like ALMAProx, also uses proximal splitting but uses a different optimization method. PDPGD was adapted to the semantic segmentation task in [36] via introducing a constraint on each pixel. We applied the same parameter settings as [36].

3.4 Aggregating the Attacks

We use our set of ten different attacks by running each attack on each input and taking the most successful result. Depending on the metric in question, the most successful attack is the one with the smallest pixel accuracy, or the smallest mean IoU, respectively. Note that the aggregated performance of the set of attacks can in principle be much better than any of the individual attacks.

4 Investigated Models

We investigate the best known state-of-the-art robust models [17, 22, 42]. We also include a number of our own models to illustrate the effect of some design choices. Our model-set includes 22 robust models as described below.

DDC-AT. Xu et al. [42] propose the DDC attack and design an adversarial training algorithm based on DDC. We include their model checkpoints in our set. These checkpoints are trained over the Cityscapes dataset [14] and the PASCAL VOC 2012 [19] dataset with the PSPNet [45] and DeepLabv3 [10] architectures, using a ResNet-50 backbone pretrained on ImageNet. Thus, we include four DDC-AT model instances altogether. DDC-AT uses a sophisticated method to create the training batches with 50% adversarial samples and we do not include any modified versions of the published method.

PGD-AT. Xu et al. [42] include a simple baseline adversarial training algorithm that uses the 3-step PGD attack. The implementation uses training batches with 50% adversarial and 50% clean samples. Here, we also include the checkpoints the authors shared in all the four settings similar to those of DDC-AT. In addition,

we train our own robust models as well in all the four settings with an identical configuration, except using 100% adversarial batches during training. Thus, we include eight PGD-AT models altogether.

SegPGD-AT. Gu et al. [22] improve DDC-AT with the help of the SegPGD attack. They publish measurements with models trained using the 3-step and 7-step SegPGD, using batches with 50% adversarial and 50% clean samples. Since the checkpoints used in the paper are not available publicly, we trained our own models using the implementation provided by the authors. We created eight models altogether using configurations that are identical to those of the PGD-AT models.

SEA-AT. Croce et al. [17] test their SEA ensemble attack using their own adversarially trained models. Note that instead of the SEA attack, they use PGD for adversarial training. The dataset is an extended PASCAL VOC 2011 dataset, and the architecture of the model is UPerNet [40] with a ConvNeXt [28] pretrained ImageNet backbone. All the training batches are 100% adversarial. We include in our set two checkpoints provided by the authors that were both trained for 50 epochs with a 5-step PGD and a robust backbone initialization. The two models use the tiny and the small version of the ConvNeXt architecture, respectively.

Normal. We also include models trained on clean samples for all the possible combinations of databases and architectures mentioned above. The normal PSPNet and DeepLabv3 models are identical to the ones used in [42]. The normal UPerNet models were trained by us using the implementation of [17] for 50 epochs.

4.1 General Notes on Training

The internal attacks applied in adversarial training used the ℓ_∞-norm neighborhood $\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ with $\epsilon = 0.03 \approx 8/255$, the same perturbation set the attacks use. The input channels are scaled to the range $[0, 1]$.

We note that on the PASCAL VOC dataset adversarial training was implemented assuming that the background class is a regular class, both in terms of training and attack. However, the Cityscapes models were all trained with masking the 'void' class out. We adopt this slightly inconsistent methodology from related work in order to obtain comparable results. For more details on the training methodology, please refer to the supplementary material (Sec. S.7 and the implementation).

5 Evaluation

Our models, datasets and attacks allow for 280 possible combinations. We evaluated all these combinations. Here, we present a representative sample of our

Table 1: Clean / robust metrics (%). PN: PSPNet, DL: DeepLabv3, P: PASCAL VOC 2012, CS: Cityscapes, $\not\mathcal{B}$: pixels with background ground truth label ignored.

		Normal	DDC-AT	PGD-AT	SegPGD-AT	PGD-AT-100	SegPGD-AT-100
Accuracy	PN+P	91.85 / 0.00	91.42 / 0.00	91.23 / 0.51	88.19 / 0.19	79.26 / 39.25	79.12 / 46.15
	DL+P	91.81 / 0.00	91.45 / 0.01	89.81 / 0.00	88.23 / 0.01	80.08 / 46.16	79.37 / 48.04
	PN+P $\not\mathcal{B}$	87.63 / 0.00	86.98 / 0.00	86.71 / 0.04	67.34 / 0.00	40.59 / 8.15	38.46 / 9.06
	DL+P $\not\mathcal{B}$	87.84 / 0.00	86.71 / 0.00	85.58 / 0.00	67.87 / 0.00	39.81 / 8.20	39.68 / 9.46
	PN+CS	93.09 / 0.00	92.80 / 0.01	92.52 / 0.03	91.77 / 0.01	83.83 / 51.45	83.32 / 69.84
	DL+CS	93.08 / 0.00	92.78 / 0.04	92.62 / 0.01	91.83 / 0.00	84.69 / 48.38	84.04 / 68.76
CmIoU	PN+P	68.87 / 0.00	67.53 / 0.00	66.73 / 0.06	51.59 / 0.02	24.31 / 5.37	23.27 / 5.88
	DL+P	68.80 / 0.00	67.30 / 0.00	63.24 / 0.00	52.50 / 0.01	24.70 / 5.70	24.15 / 6.20
	PN+P $\not\mathcal{B}$	78.84 / 0.00	78.12 / 0.00	77.28 / 0.02	53.82 / 0.00	25.10 / 4.50	23.59 / 4.98
	DL+P $\not\mathcal{B}$	79.08 / 0.00	77.46 / 0.00	75.28 / 0.00	55.26 / 0.00	24.70 / 4.60	24.51 / 5.02
	PN+CS	66.28 / 0.00	63.90 / 0.01	61.85 / 0.04	59.71 / 0.01	32.74 / 16.37	31.19 / 20.29
	DL+CS	66.96 / 0.00	64.01 / 0.04	63.29 / 0.01	59.87 / 0.00	34.79 / 15.73	33.32 / 20.72
NmIoU	PN+P	70.70 / 0.00	69.32 / 0.00	68.46 / 0.12	55.76 / 0.05	37.01 / 11.35	36.10 / 13.31
	DL+P	69.00 / 0.00	67.92 / 0.00	64.01 / 0.00	54.94 / 0.01	35.23 / 11.61	33.85 / 12.21
	PN+P $\not\mathcal{B}$	43.41 / 0.00	42.25 / 0.00	41.86 / 0.02	28.00 / 0.01	14.50 / 3.14	13.55 / 3.59
	DL+P $\not\mathcal{B}$	42.89 / 0.00	41.59 / 0.00	39.26 / 0.00	27.89 / 0.01	13.73 / 3.06	13.41 / 3.52
	PN+CS	52.57 / 0.00	50.97 / 0.01	49.14 / 0.02	47.27 / 0.01	33.13 / 17.64	32.41 / 22.67
	DL+CS	51.18 / 0.00	49.63 / 0.02	48.28 / 0.01	45.91 / 0.00	33.37 / 15.86	32.68 / 21.82

Table 2: Clean / robust metrics of SEA-AT (%). $\not\mathcal{B}$: pixels with background ground truth label ignored.

	Normal-Tiny	SEA-AT-Tiny	Normal-Small	SEA-AT-Small
Accuracy	93.10 / 0.00	92.75 / 71.84	93.78 / 0.00	93.10 / 70.57
Accuracy $\not\mathcal{B}$	88.21 / 0.00	86.39 / 52.99	90.12 / 0.00	88.53 / 55.50
CmIoU	72.49 / 0.00	72.09 / 32.79	75.40 / 0.00	73.53 / 32.66
CmIoU $\not\mathcal{B}$	80.55 / 0.00	79.68 / 43.14	83.20 / 0.00	81.76 / 43.76
NmIoU	73.33 / 0.00	74.07 / 13.20	77.08 / 0.00	75.11 / 12.89
NmIoU $\not\mathcal{B}$	43.10 / 0.00	42.29 / 7.31	45.54 / 0.00	43.48 / 7.51

results to support our main findings, the complete set of results is presented in the supplementary material.

We apply the same evaluation methodology for every model we study, as we discuss below. This methodology differs from the ones used in the original evaluations of these models. We present the details of the original methodologies in the supplementary material in Sec. S.8, noting here that some of these details are not documented and had to be learned directly from the implementation.

The evaluation procedure. Following general practice, we evaluate over the validation sets of the PASCAL VOC 2012 and Cityscapes datasets. On PASCAL VOC, before evaluation, the longer dimension of each image is scaled to 512 pixels, but no further rescaling or cropping is applied. On Cityscapes, the images are scaled down to 1024x512. After computing the prediction mask, no augmentation method is applied to improve the mask.

Here, we deviate from related work in that Croce et al. [17] resize to 512x512 and then crop to 473x473, and Gu et al. and Xu et al. [22, 42] use a tiled evalu-

ation with overlapping tiles, and in addition, over the clean inputs further augmentations are applied based on mirrored inputs. Due to these differences, our measurements are not identical to those published in the original works. However, when applying these techniques, we are able to reproduce the published values.

How to aggregate IoU? We use *Pixel Accuracy* (or simply accuracy) and *mean intersection over union (mIoU)* as our performance metrics. Regarding mIoU, the aggregation procedure to compute the global mIoU value based on the images in the evaluation set is often overlooked. We will demonstrate that in the case of evaluating robustness, this aggregation procedure plays an important role. The usual approach, also used in [17, 22, 42], performs an aggregation over the images to compute the global IoU of each class, and then computes the average:

$$\mathrm{CmIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{n=1}^{N}\mathrm{TP}_{cn}}{\sum_{n=1}^{N}\mathrm{FP}_{cn}+\mathrm{TP}_{cn}+\mathrm{FN}_{cn}}, \tag{6}$$

where N is the number of images and C is the number of classes.

However, we will also apply a different aggregation to emphasize the errors made on individual images, which is the main goal of adversarial attacks. Here, we simply average the image-wise IoU values:

$$\mathrm{NmIoU} = \frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C}\frac{\mathrm{TP}_{cn}}{\mathrm{FP}_{cn}+\mathrm{TP}_{cn}+\mathrm{FN}_{cn}}. \tag{7}$$

This metric captures the image-wise performance much better, especially over datasets where there are only a few objects in most images, some of which are large and some are small. This is the case, for example, in the PASCAL VOC dataset.

5.1 Results

Tab. 1 contains our results with the DDC-AT and SegPGD-AT models, along with baselines (normal model and PGD-AT models). As described in Sec. 4, the two models that use 100% adversarial batches during training (PGD-AT-100 and SegPGD-AT-100) were trained by us, just like SegPGD-AT. The remaining models are checkpoints from [42].

Tab. 2 shows our results with the SEA-AT models. These models are all checkpoints taken from [17]. The normal models were trained by ourselves.

The tables contain clean and robust metrics. The robust metrics were computed over the adversarially perturbed inputs that were generated using the aggregated attack of our set of ten attacks.

No robustness with 50% adversarial batches. The most striking observation is that our aggregated attack achieves a value of near zero according to all the three metrics for all the models that use only 50% adversarial samples. This means that the models published in [22, 42] show *no sign of robustness*. (See Sec. S.10.2 in the supplementary material for further evidence.)

Using 100% adversarial samples is not sufficient. While we can achieve non-trivial robustness in terms of pixel accuracy, our attacks significantly reduce both the CmIoU and NmIoU metrics in the case of PGD-AT-100 and SegPGD-AT-100, indicating a very low robustness. Also, the performance of these models over the clean samples is much worse than that of the normal model.

In the case of the SEA-AT models (Tab. 2) the clean performance is very close to that of the normal model according to all the metrics. However, robust NmIoU drops to a fraction of the clean NmIoU, again, indicating a very low level of robustness.

Size-bias in the SEA-AT models. In the case of the SEA-AT models the robust CmIoU is significantly higher than the robust NmIoU, while the rest of the models show the opposite behavior. To understand this better, we take a closer look at the distribution of the single image mIoU over the evaluation set. Figs. 1 and 3 show a representative sample of such distributions, with CmIoU and NmIoU indicated. (For more, please refer to Sec. S.9).

It is striking that in the case of the PASCAL VOC dataset the robust single image mIoU distributions of SegPGD-AT-100 (Fig. 3) and SEA-AT-S (Fig. 1) are very similar but the aggregated mIoU measures dramatically differ. Since the NmIoU metric is more sensitive to errors in the segmentation of small objects, a possible explanation is that the SEA-AT model learns to protect the largest objects while it sacrifices the small ones. In other words, the attacks essentially remove smaller objects from the image, causing very low mIoU values on many images. This problem is not captured well by the CmIoU or the accuracy metrics. Looking at the predicted masks confirms this hypothesis. Fig. 2 shows an example for this phenomenon.

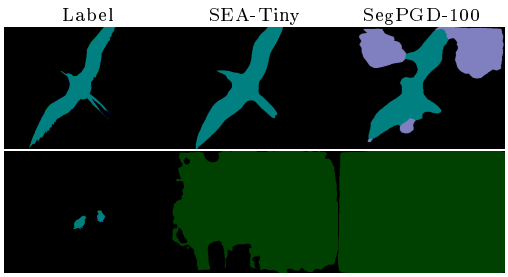

Fig. 2: Illustration of the size-bias on two bird samples. The predicted masks on adversarial input are shown for two models. The small birds are deleted completely by our adversarial attacks, while the large bird is better preserved, especially by SEA-AT-Tiny. The NmIoU metric captures this important problem better than CmIoU or accuracy do.

The foreground is more vulnerable. In the PASCAL VOC dataset, the images contain lots of pixels in the background class that might dominate our measurements. In fact, the background pixels form 74% of all the pixels in the validation set.

To examine the effect of the background class, in all the tables, we also show the measurement results with the background pixels removed. That is, the metrics are computed only on the foreground pixels, and the IoU of the background class is not taken into account.

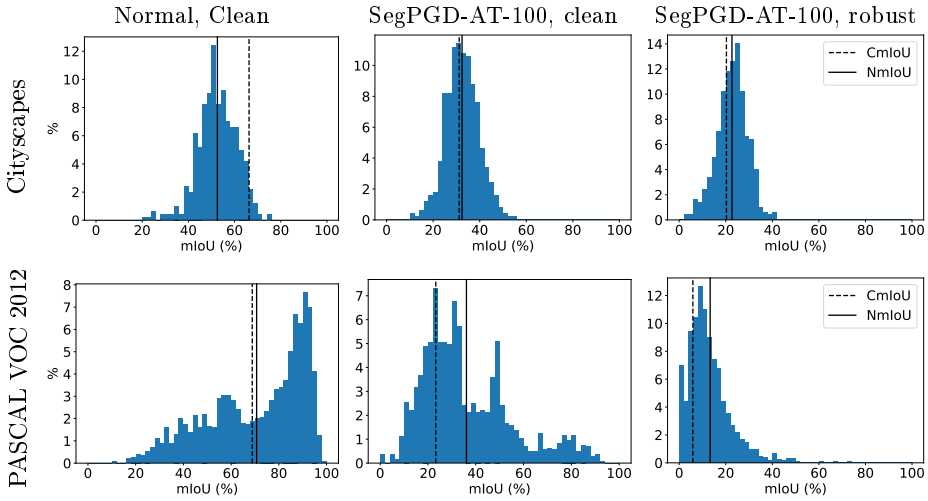

Fig. 3: Distributions of single image mIoU with PSPNET for the Normal model and for SegPGD-AT-100, for clean and adversarial inputs. CmIoU and NmIoU are shown using vertical lines.

We can see that the accuracy and the NmIoU over the foreground are much worse than on the complete image. This problem is more severe in the case of SegPGD-AT-100 and PGD-AT-100. These models essentially focus on predicting and protecting the background, which is not the intended behavior. The SEA-AT models also show this effect to some extent.

Robustness-accuracy tradeoff. Our results confirm that the apparent lack of robustness-accuracy tradeoff can be explained by the insufficient evaluation methodology both in terms of using only weak attacks, or not using the right performance metrics. Indeed, those models that have a good clean performance, close to that of the normally trained models, turn out not to be very robust, when measured appropriately. Clearly, the best models we examined are the SEA-AT models, but even those models show a weak performance in the NmIoU metric due to their mIoU distribution (Fig. 1).

5.2 A Closer Look at the Individual Attacks

Let us now examine how the individual attacks in our attack-set contribute to the aggregated robust metrics. Tab. 3 shows the results of every attack separately in one scenario (every scenario is similar in this regard, please refer to the supplementary material in Sec. S.10.1). We can observe that the aggregated attack is often much stronger than any of the individual attacks. Also, different attacks are effective against different kinds of models. For example, for the normal models PAdam-Cos tends to be the best individual attack, while for the more robust models the SEA attacks perform better.

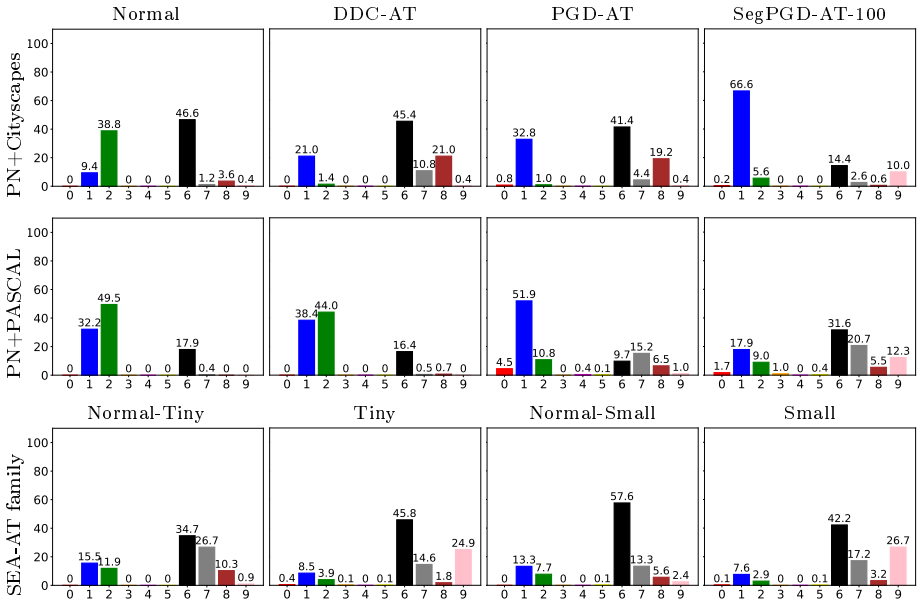

Fig. 4: Best attack distribution over the validation set according to mIoU (PN: PSPNet). The attacks are 0: ALMAProx, 1: PAdam-CE, 2: PAdam-Cos, 3: DAG-0.001, 4: DAG-0.003, 5: PDPGD, 6: SEA-JSD, 7: SEA-MCE, 8: SEA-MSL, 9: SEA-BCE.

Fig. 4 shows the distribution of the most successful attack over the validation set in a number of scenarios. In other words, for each input example, we determine which attack was the most successful and show the resulting distribution. The supplementary material covers all the scenarios in Sec. S.10.1.

Table 3: Attacks on Cityscapes with DeepLabv3 (NmIoU %).

	Normal	DDC-AT	PGD-AT	SegPGD -AT	PGD-AT -100	SegPGD -AT-100
clean	51.18	49.63	48.28	45.91	33.37	32.68
PAdam-CE	1.39	1.09	0.31	0.85	20.66	23.11
PAdam-Cos	0.00	0.64	0.30	0.43	22.58	25.40
SEA-JSD	0.67	0.74	0.58	1.12	17.15	25.14
SEA-MCE	1.51	1.03	0.67	0.79	18.13	26.23
SEA-MSL	1.53	0.73	0.62	0.83	20.95	28.41
SEA-BCE	1.73	2.81	1.20	2.43	17.74	25.48
ALMAProx	1.87	1.35	1.73	1.37	30.86	30.53
DAG-0.001	5.48	45.29	35.25	27.67	33.51	32.80
DAG-0.003	1.04	20.80	13.24	9.31	33.51	32.80
PDPGD	1.48	11.22	9.88	1.75	32.34	31.87
aggregated	0.00	0.02	0.01	0.00	15.86	21.82

It is striking how different these distributions are in the various scenarios. The two attacks proposed in this work, PAdam-CE and PAdam-Cos, dominate many scenarios including normal and some robust models as well. Interestingly, the SEA-AT family of models is most sensitive to the SEA attack set. Nevertheless, it is clear that every attack has its contribution, and different models require different attacks.

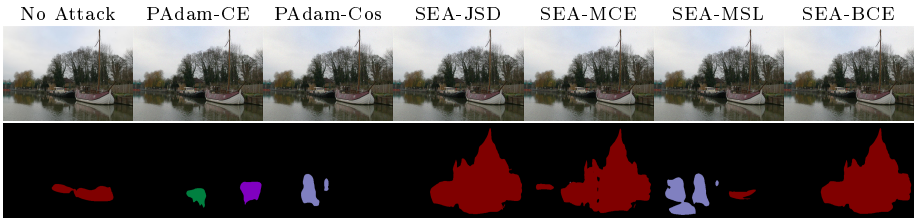

Fig. 5: Result of some attacks on an example PASCAL-VOC image on SEA-AT-Tiny. Top row: perturbed images; bottom row: predicted mask on the perturbed image.

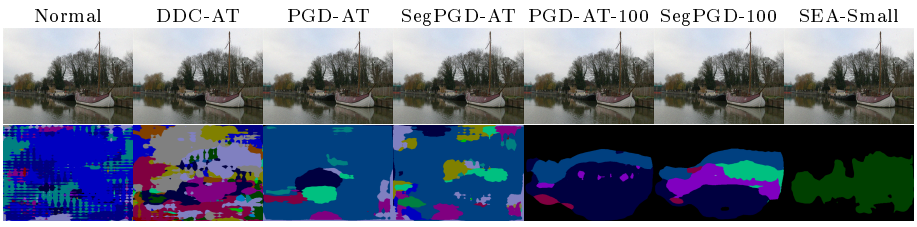

Fig. 6: Result of PAdam-Cos attack on an example PASCAL-VOC image for various PSPNet models and SEA-AT-Small. Top row: perturbed images; bottom row: predicted mask on the perturbed image.

Figs. 5 and 6 illustrate the diversity of the attacks and the models through an example image. The perturbations remain invisible in all the cases. For this particular image, PAdam-Cos can completely alter the output of all the models. Further examples are shown in the supplementary material in Sec. S.11.

6 Conclusions and Limitations

Our contribution was of a methodological nature. We empirically proved that using a strong set of attacks can dramatically reduce the known upper bounds of the robustness metrics of the state-of-the-art models, in many cases completely diminishing them. We also pointed out that the choice of mIoU aggregation method matters, because robust models tend to have a strong size-bias that is not revealed by class-wise aggregation, only by image-wise aggregation (NmIoU).

We demonstrated our methodology mostly on public model checkpoints from related work and we did only a limited exploration of variants. It would be very informative to also study, for example, the effect of the proportion of adversarial samples in the batches, or combinations of backbone models, architectures and training hyperparameters. These limitations are mostly due to the prohibitive cost of such a study.

We do believe that our thorough analysis of the highly cited models we selected still provides a useful contribution to the community in terms of how to move forward with the analysis of robust semantic segmentation models.

Acknowledgements

This research was supported by the AI Competence Centre of the Cluster of Science and Mathematics of the Centre of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged. We also thank the support by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory, and by project TKP2021-NVA-09 that has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme. We thank András Balogh for his comments on earlier versions that helped us improve the presentation.

References

1. Agnihotri, S., Keuper, M.: CosPGD: a unified white-box adversarial attack for pixel-wise prediction tasks (2023). https://doi.org/10.48550/ARXIV.2302.02213, https://arxiv.org/abs/2302.02213
2. Arnab, A., Miksik, O., Torr, P.H.S.: On the robustness of semantic segmentation models to adversarial attacks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 888–897. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00099, http://openaccess.thecvf.com/content_cvpr_2018/html/Arnab_On_the_Robustness_CVPR_2018_paper.html
3. Ba, J., Kingma, D.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR) (2015), http://arxiv.org/abs/1412.6980
4. Bär, A., Hüger, F., Schlicht, P., Fingscheidt, T.: On the robustness of redundant teacher-student frameworks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1380–1388. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPRW.2019.00178, http://openaccess.thecvf.com/content_CVPRW_2019/html/SAIAD/Bar_On_the_Robustness_of_Redundant_Teacher-Student_Frameworks_for_Semantic_Segmentation_CVPRW_2019_paper.html
5. Bär, A., Klingner, M., Varghese, S., Hüger, F., Schlicht, P., Fingscheidt, T.: Robust semantic segmentation by redundant networks with a layer-specific loss contribution and majority vote. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020. pp. 1348–1358. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPRW50498.2020.00174, https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Bar_Robust_Semantic_Segmentation_by_Redundant_Networks_With_a_Layer-Specific_Loss_CVPRW_2020_paper.html
6. Cai, Z., Rane, S., Brito, A.E., Song, C., Krishnamurthy, S.V., Roy-Chowdhury, A.K., Asif, M.S.: Zero-query transfer attacks on context-aware object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15024–15034 (June 2022)

7. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49, https://arxiv.org/abs/1608.04644
8. Chen, J., Jordan, M.I., Wainwright, M.J.: HopSkipJumpAttack: A query-efficient decision-based attack. In: 2020 IEEE Symposium on Security and Privacy (SP). pp. 1277–1294. IEEE Computer Society, Los Alamitos, CA, USA (May 2020). https://doi.org/10.1109/SP40000.2020.00045, https://doi.ieeecomputersociety.org/10.1109/SP40000.2020.00045
9. Chen, J., Gu, Q.: Rays: A ray searching method for hard-label adversarial attack. In: Gupta, R., Liu, Y., Tang, J., Prakash, B.A. (eds.) KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020. pp. 1739–1747. ACM (2020). https://doi.org/10.1145/3394486.3403225, https://doi.org/10.1145/3394486.3403225
10. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017), http://arxiv.org/abs/1706.05587
11. Chen, Z., Wang, C., Crandall, D.J.: Semantically stealthy adversarial attacks against segmentation models. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022. pp. 2846–2855. IEEE (2022). https://doi.org/10.1109/WACV51458.2022.00290, https://doi.org/10.1109/WACV51458.2022.00290
12. Cissé, M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 6977–6987 (2017), https://proceedings.neurips.cc/paper/2017/hash/d494020ff8ec181ef98ed97ac3f25453-Abstract.html
13. Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. pp. 1310–1320 (2019), http://proceedings.mlr.press/v97/cohen19c.html
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 3213–3223. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.350, https://doi.org/10.1109/CVPR.2016.350
15. Croce, F., Gowal, S., Brunner, T., Shelhamer, E., Hein, M., Cemgil, A.T.: Evaluating the adversarial robustness of adaptive test-time defenses. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 4421–4435. PMLR (2022), https://proceedings.mlr.press/v162/croce22a.html
16. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 2206–2216. PMLR (2020), http://proceedings.mlr.press/v119/croce20b.html

17. Croce, F., Singh, N.D., Hein, M.: Robust semantic segmentation: Strong adversarial attacks and fast training of robust models. CoRR **abs/2306.12941** (2023). https://doi.org/10.48550/arXiv.2306.12941, https://doi.org/10.48550/arXiv.2306.12941
18. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., Madry, A.: Adversarial robustness as a prior for learned representations. Tech. Rep. 1906.00945, arXiv.org (2019), https://arxiv.org/abs/1906.00945
19. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010). https://doi.org/10.1007/S11263-009-0275-4, https://doi.org/10.1007/s11263-009-0275-4
20. Fischer, V., Kumar, M.C., Metzen, J.H., Brox, T.: Adversarial examples for semantic image segmentation. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=S1SED1MYe
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations (ICLR) (2015), https://arxiv.org/abs/1412.6572
22. Gu, J., Zhao, H., Tresp, V., Torr, P.H.S.: Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIX. Lecture Notes in Computer Science, vol. 13689, pp. 308–325. Springer (2022). https://doi.org/10.1007/978-3-031-19818-2_18, https://doi.org/10.1007/978-3-031-19818-2_18
23. Gupta, P., Rahtu, E.: MLAttack: Fooling semantic segmentation networks by multi-layer attacks. In: Fink, G.A., Frintrop, S., Jiang, X. (eds.) Pattern Recognition - 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10-13, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11824, pp. 401–413. Springer (2019). https://doi.org/10.1007/978-3-030-33676-9_28, https://doi.org/10.1007/978-3-030-33676-9_28
24. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32. pp. 125–136. Curran Associates, Inc. (2019), http://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.pdf
25. Klingner, M., Bär, A., Fingscheidt, T.: Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020. pp. 1299–1309. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPRW50498.2020.00168, https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Klingner_Improved_Noise_and_Attack_Robustness_for_Semantic_Segmentation_by_Using_CVPRW_2020_paper.html
26. Klingner, M., Kumar, V.R., Yogamani, S.K., Bär, A., Fingscheidt, T.: Detecting adversarial perturbations in multi-task perception. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, Kyoto, Japan, October 23-27, 2022. pp. 13050–13057. IEEE (2022). https://doi.org/10.1109/IROS47612.2022.9981559, https://doi.org/10.1109/IROS47612.2022.9981559

27. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=HJGU3Rodl
28. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 11966–11976. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01167, https://doi.org/10.1109/CVPR52688.2022.01167
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
30. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb
31. Mao, C., Gupta, A., Nitin, V., Ray, B., Song, S., Yang, J., Vondrick, C.: Multitask learning strengthens adversarial robustness. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12347, pp. 158–174. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_10, https://doi.org/10.1007/978-3-030-58536-5_10
32. Matyasko, A., Chau, L.: PDPGD: primal-dual proximal gradient descent adversarial attack. CoRR **abs/2106.01538** (2021), https://arxiv.org/abs/2106.01538
33. Metzen, J.H., Kumar, M.C., Brox, T., Fischer, V.: Universal adversarial perturbations against semantic image segmentation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2774–2783. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.300, https://doi.org/10.1109/ICCV.2017.300
34. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2574–2582 (June 2016), https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.pdf
35. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of Adam and beyond. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=ryQu7f-RZ
36. Rony, J., Pesquet, J.C., Ayed, I.B.: Proximal splitting adversarial attacks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20524–20533 (2023), https://arxiv.org/abs/2206.07179
37. Sturm, B.L.: A simple method to determine if a music information retrieval system is a "horse". IEEE Trans. Multim. **16**(6), 1636–1644 (2014). https://doi.org/10.1109/TMM.2014.2330697, https://doi.org/10.1109/TMM.2014.2330697
38. Sun, S., Song, B., Cai, X., Du, X., Guizani, M.: CAMA: class activation mapping disruptive attack for deep neural networks. Neurocomputing **500**, 989–1002 (2022). https://doi.org/10.1016/j.neucom.2022.05.065, https://doi.org/10.1016/j.neucom.2022.05.065

39. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations (ICLR) (2014), http://arxiv.org/abs/1312.6199
40. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V. Lecture Notes in Computer Science, vol. 11209, pp. 432–448. Springer (2018). https://doi.org/10.1007/978-3-030-01228-1_26, https://doi.org/10.1007/978-3-030-01228-1_26
41. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.L.: Adversarial examples for semantic segmentation and object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 1378–1387. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.153, https://doi.org/10.1109/ICCV.2017.153
42. Xu, X., Zhao, H., Jia, J.: Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 7466–7475. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00739, https://doi.org/10.1109/ICCV48922.2021.00739
43. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 7472–7482. PMLR (2019), http://proceedings.mlr.press/v97/zhang19p.html
44. Zhao, H.: semseg. https://github.com/hszhao/semseg (2019)
45. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6230–6239. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.660, https://doi.org/10.1109/CVPR.2017.660