Robust Semantic Segmentation: Strong Adversarial Attacks and Fast Training of Robust Models

Francesco Croce [*1 2] **Naman D. Singh** [*1 2] **Matthias Hein** [1 2]

Abstract

While a large amount of work has focused on designing adversarial attacks against image classifiers, only a few methods exist to attack semantic segmentation models. We show that attacking segmentation models presents task-specific challenges, for which we propose novel solutions. Our final evaluation protocol outperforms existing methods, and shows that those can overestimate the robustness of the models. Additionally, so far adversarial training, the most successful way for obtaining robust image classifiers, could not be successfully applied to semantic segmentation. We argue that this is because the task to be learned is more challenging, and requires significantly higher computational effort than for image classification. As a remedy, we show that by taking advantage of recent advances in robust IMAGENET classifiers, one can train adversarially robust segmentation models at limited computational cost by fine-tuning robust backbones.

1. Introduction

The vulnerability of systems based on neural networks to adversarial perturbations, that is small changes in the input can drastically modify the output of the models, is now well-known (Biggio et al., 2013; Szegedy et al., 2014; Grosse et al., 2016; Jin et al., 2019). A large amount of work has been dedicated to developing adversarial attacks in several threat models for image classification, including ℓ_p-bounded perturbations (Carlini and Wagner, 2017; Chen et al., 2018; Rony et al., 2019), sparse attacks (Brown et al., 2017; Croce et al., 2022), and those defined by perceptual metrics (Wong et al., 2019; Laidlaw et al., 2021). At the same time, evaluating the adversarial robustness in semantic segmentation, arguably a very relevant vision domain, has received significantly less attention. While a few early works (Xie et al., 2017; Hendrik Metzen et al., 2017; Arnab et al., 2018) have proposed methods to generate adversarial attacks in different threat models, Gu et al. (2022); Agnihotri and Keuper (2023) have recently shown that even for the most popular ℓ_∞-bounded attacks significant improvements are possible. In particular, they suggest that the PGD attack (Madry et al., 2018), which is commonly used against image classifiers, with the sum of pixelwise cross-entropy losses as objective function might not be suitable for the case of semantic segmentation: in fact, the key difference to image classification is that for semantic segmentation we have to flip the predictions of all pixels instead of just the prediction for the image.

In this work, we make significant progress towards a better adversarial robustness evaluation in semantic segmentation: first, we propose novel loss functions and optimization schemes for this domain which are better suited to the task of flipping *all* pixelwise predictions; second, observing that these losses have complementary properties and thus are successful on different images, we assemble them for a more reliable robustness evaluation, similar to AutoAttack (Croce and Hein, 2020) for image classification, into Segmentation Ensemble Attack (SEA) and use the worst-case across attacks. With our SEA we show that clean and robust semantic segmentation models can be more than 10% less robust in average pixel accuracy and up to 6% lower in mIOU than suggested by existing attacks (Gu et al., 2022; Agnihotri and Keuper, 2023). For detailed overview of literature on the robustness of semantic segmentation see App. D.

We are also interested in advancing the state-of-the-art in robust semantic segmentation. For image classification the most successful methods (Rebuffi et al., 2021; Wang et al., 2023) are based on adversarial training (Madry et al., 2018). However, for semantic segmentation Gu et al. (2022) could find only limited improvement in robustness using adversarial training compared to standard models. We show that obtaining robust segmentation models with adversarial training is indeed possible but requires larger computational effort: in fact, more epochs and attacks steps are needed. However, we show that the training effort can be signifi-

[*]Equal contribution [1]University of Tübingen [2]Tübingen AI Center. Correspondence to: Francesco Croce <francesco.croce@uni-tuebingen.de>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

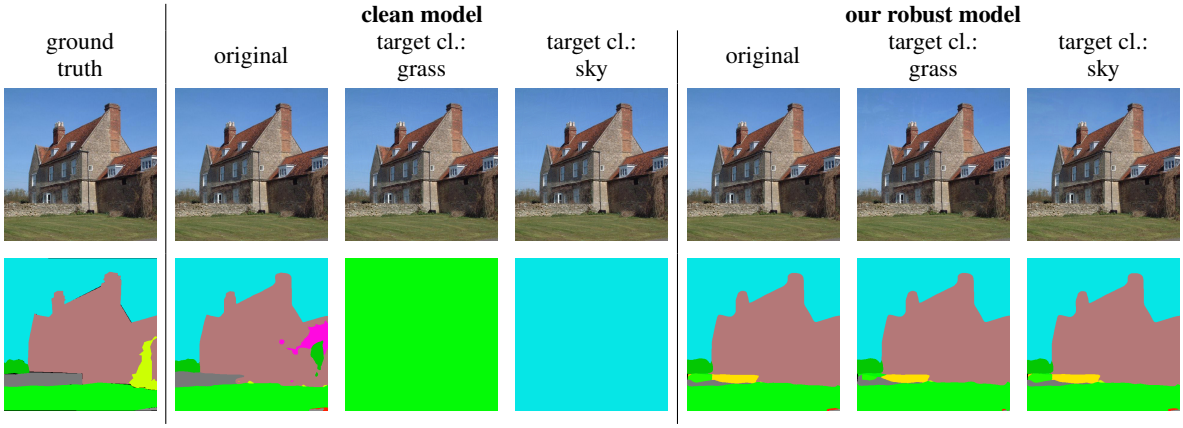

Figure 1: **Effect of adversarial attacks on semantic segmentation models.** For an image from the ADE20K validation set (first column, original image and ground truth mask), we show the image perturbed by targeted ℓ_∞-attacks with size $\epsilon_\infty = 2/255$ and target classes "grass" and "sky", and the resulting predicted segmentation maps. For a standard (clean) model, the attack is highly successful, whereas for our adversarially trained robust model the attack leads only minimal changes of the the segmentation. We use targeted attacks for illustration but untargeted attacks in the rest of the paper.

cantly reduced by leveraging recent advances in training robust models on IMAGENET (Debenedetti, 2022; Singh et al., 2023; Liu et al., 2023). By initializing the backbone of our segmentation model with a robust ConvNeXt, adversarially pre-trained on IMAGENET, we achieve similar or better adversarial robustness at up to 6 times lower computational cost than models trained with clean initialization.

Contributions. As a summary (see also an illustration in Fig. 1), in this work

- we propose novel loss functions to generate adversarial attacks against semantic segmentation, show how to adapt the optimization algorithms to significantly improve their efficiency, and validate the findings in extensive experiments.
- we propose Segmentation Ensemble Attack (SEA), an ensemble of attacks based on complementary losses for the ℓ_∞ threat model, which improves significantly over each individual attack.
- we show how to leverage existing robust image classifiers to achieve adversarially robust segmentation models at reduced training time. To our knowledge, this provides the SOTA robust models on PASCAL-VOC (Everingham et al., 2010) and the first ones on ADE20K (Zhou et al., 2019).

2. Adversarial Robustness of Semantic Segmentation Models

In the following we discuss details about the experimental setup focused on the ℓ_∞-threat model. We postpone the discussion on threat model and metric selection to App. E.

Setup. The goal of semantic segmentation consists in classifying each pixel of a given image into the available classes (corresponding to different objects or background). We denote a segmentation model $f : \mathbb{R}^{w \times h \times c} \longrightarrow \mathbb{R}^{w \times h \times K}$, which for an image $\boldsymbol{x}$ of size $w \times h$ (and c color channels) returns $\boldsymbol{z} = f(\boldsymbol{x})$, where $\boldsymbol{z}_{ij} \in \mathbb{R}^K$ contains the score of each of the K classes for the pixel $\boldsymbol{x}_{ij}$. Then, similar to image classification, the class predicted by f for $\boldsymbol{x}_{ij}$ is given by $\boldsymbol{m}_{ij} = \underset{k=1,\dots,K}{\arg\max}\, \boldsymbol{z}_{ijk}$, and $\boldsymbol{m} \in \mathbb{R}^{w \times h}$ is the segmentation map of $\boldsymbol{x}$. Assuming access to ground truth map $\boldsymbol{y} \in \mathbb{R}^{w \times h}$, one can compute the average pixel accuracy of f for $\boldsymbol{x}$ as $\frac{1}{w \cdot h} \sum_{i,j} \mathbb{I}(\boldsymbol{m}_{ij} = \boldsymbol{y}_{ij})$. Then, the goal of an adversarial attack on f is to reduce its segmentation performance. This can be formalized as solving

$$\min_{\boldsymbol{\delta}} \frac{1}{w \cdot h} \sum_{i,j} \mathbb{I}(\arg\max_k f(\boldsymbol{x} + \boldsymbol{\delta})_{ijk} = \boldsymbol{y}_{ij}) \qquad (1)$$
$$\text{s. th.} \quad \|\boldsymbol{\delta}\|_p \leq \epsilon, \quad \boldsymbol{x} + \boldsymbol{\delta} \in [0,1]^{w \times h \times c}$$

where one wants to minimize the number of correctly classified pixels with perturbations of bounded ℓ_p-norm and remaining in the image domain. Since the objective function in Eq. (1) is non-differentiable, it is common to rephrase the problem as

$$\max_{\boldsymbol{\delta}} \frac{1}{w \cdot h} \sum_{i,j} \mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta})_{ij}, \boldsymbol{y}_{ij}) \qquad (2)$$
$$\text{s. th.} \quad \|\boldsymbol{\delta}\|_p \leq \epsilon, \quad \boldsymbol{x} + \boldsymbol{\delta} \in [0,1]^{w \times h \times c}$$

where $\mathcal{L} : \mathbb{R}^K \times \mathbb{R} \longrightarrow \mathbb{R}$ is a (almost everywhere) smooth function whose maximization induces misclassification: this can then be (approximately) solved by standard techniques

for constrained optimization such as projected gradient descent (PGD). Designing surrogate losses specific for segmentation models is one of the key challenges to obtain effective attacks. In this work we focus on the ℓ_∞-threat model, which means that every pixel of an input image can be modified independently.

3. Adversarial Attacks on Segmentation Models

Before developing methods to obtain adversarially robust models, it is necessary to have effective attacks to test their robustness. Projected gradient descent (PGD) (Madry et al., 2018), together with its variants, is the most popular choice to solve the optimization problem in Eq. (2). In fact, PGD is also the basis of the existing attacks for semantic segmentation of Gu et al. (2022); Agnihotri and Keuper (2023), which we show (Table 4 in App. C.1) yields weaker attacks as opposed to APGD (Croce and Hein, 2020).

3.1. Loss functions

In the case of semantic segmentation an attacker wants to get as many pixels as possible to be misclassified. This is exemplified by the objective function in Eq. (1) consisting in the sum of pixelwise losses. However, this sum of losses can give rise to conflicting descent directions for different pixels, which can hinder the optimization. In the following we give a short overview of the loss functions which have been used in the literature and the ones which we propose as new alternatives. We denote $\boldsymbol{u} \in \mathbb{R}^K$ the logits of each pixel, and $y \in \mathbb{N}$ its correct label. Moreover, given $\boldsymbol{u}$, we indicate as $\boldsymbol{p} \in \mathbb{R}^K$ the predicted probability distribution via the softmax function: $\boldsymbol{p}_r = e^{\boldsymbol{u}_r} / \sum_{l=1}^K e^{\boldsymbol{u}_l}$, $r = 1, \dots, K$.

Losses used in literature for segmentation attacks

Cross-entropy (CE): the most common choice as objective function in PGD based attacks is the cross-entropy between the one-hot encoding of the ground truth label and the softmax of the logits, i.e. $\mathcal{L}_{\text{CE}}(\boldsymbol{u}, y) = -\log \boldsymbol{p}_y = -\boldsymbol{u}_y + \log\left(\sum_{j=1}^K e^{\boldsymbol{u}_j}\right)$. The cross-entropy loss is unbounded, which is problematic for semantic segmentation as already misclassified pixels will still be optimized instead of focusing on still correctly classified pixels (see discussion for the Jensen-Shannon-divergence).

Balanced cross-entropy: Gu et al. (2022) propose to balance the importance of the cross-entropy loss of correctly and wrongly classified pixels over iterations. In particular, at iteration $t = 1, \dots, T$, they use, with $\lambda(t) = (t-1)/(2T)$. Let $j^* = \underset{j=1,\dots,K}{\arg\max}\, \boldsymbol{u}_j$, then the loss can be defined as

$$\mathcal{L}_{\text{Bal-CE}}(\boldsymbol{u}, y) = \left((1-\lambda(t)) \cdot \mathbb{I}_{j^*=y} + \lambda(t) \cdot \mathbb{I}_{j^*\neq y}\right) \cdot \mathcal{L}_{\text{CE}}(\boldsymbol{u}, y)$$

In this way the algorithm first focuses only on the correctly classified pixels and then progressively balances the attention on the two subset of pixels: this has the goal of avoiding to make updates which find new misclassified pixels but leads to correct decisions for already misclassified pixels.

Weighted cross-entropy: Agnihotri and Keuper (2023) propose to weigh the importance of the pixels via cosine similarity between the prediction vector (post-applying the sigmoid function $\sigma(t) = 1/(1+e^{-t})$) and one-hot encoding $\boldsymbol{e}_y$ of the ground truth class. This can be written as

$$\begin{aligned}\mathcal{L}_{\text{CosSim-CE}}(\boldsymbol{u}, y) &= \frac{\langle \sigma(\boldsymbol{u}), \boldsymbol{e}_y \rangle}{\|\sigma(\boldsymbol{u})\|_2 \, \|\boldsymbol{e}_y\|_2} \cdot \mathcal{L}_{\text{CE}}(\boldsymbol{u}, y) \\ &= \sigma(\boldsymbol{u}_y) / \|\sigma(\boldsymbol{u})\|_2 \cdot \mathcal{L}_{\text{CE}}(\boldsymbol{u}, y)\end{aligned}$$

and again has the effect of reducing the importance of the pixels which are confidently misclassified.

Novel losses for attacks on semantic segmentation

Masked cross-entropy: in order to avoid over-optimizing misclassified pixels one can apply a mask which excludes such pixels from the loss computation, that is

$$\mathcal{L}_{\text{Mask-CE}}(\boldsymbol{u}, y) = \mathbb{I}(\underset{j=1,\dots,K}{\arg\max}\, \boldsymbol{u}_j = y) \cdot \mathcal{L}_{\text{CE}}(\boldsymbol{u}, y).$$

The downside of using such a mask is that the loss becomes discontinuous and ignoring misclassified pixels might lead to changes which revert back wrongly classified pixels into correctly classified ones with the danger of creating a situation where one starts oscillating. We note that Hendrik Metzen et al. (2017) proposed, for targeted attacks, to not optimize the loss for pixels already classified into the target class with confidence higher than a fixed threshold. Similarly Xie et al. (2017) did not include the pixels already belonging to the target class in the loss computation for unconstrained attacks. However, the masked CE-loss has not been thoroughly explored for ℓ_∞-bounded untargeted attacks.

Jensen-Shannon (JS) divergence: an intermediate behavior between losses which do not consider whether the attack is successful on a certain pixel and the classification mask used above would adjust the importance in the updates of each pixel depending on the confidence in the correct class. Given two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, the Jensen-Shannon divergence is defined as

$$\begin{aligned}D_{\text{JS}}(\boldsymbol{p} \,\|\, \boldsymbol{q}) &= \left(D_{\text{KL}}(\boldsymbol{p} \,\|\, \boldsymbol{m}) + D_{\text{KL}}(\boldsymbol{q} \,\|\, \boldsymbol{m})\right)/2 \\ \text{with} \quad \boldsymbol{m} &= (\boldsymbol{p}+\boldsymbol{q})/2\end{aligned}$$

where D_{KL} indicates the Kullback–Leibler divergence. If we assume $\boldsymbol{p}$ to be the softmax output of the logits $\boldsymbol{u}$ and $\boldsymbol{q}$ the one-hot encoding of the target y, we get $\mathcal{L}_{\text{JS}}(\boldsymbol{u}, y) = D_{\text{JS}}(\boldsymbol{p} \,\|\, \boldsymbol{q})$. Since D_{JS} measures the similarity between the two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, maximizing $\mathcal{L}_{\text{JS}}$ drives the

prediction of the model away from the ground truth. Unlike the KL divergence or the CE loss, the JS divergence is bounded, which means that the influence of every pixel is limited. In particular, it has the following property (see App. A for a proof)

$$\lim_{p_y \to 0} \frac{\partial \mathcal{L}_{\text{JS}}(\boldsymbol{u}, y)}{\partial u_t} = 0, \quad \text{for} \quad t = 1, \dots, K$$

In contrast, the CE loss has a non-zero gradient if $p_y \to 0$: thus, even clearly misclassified pixels still influence the optimization of the loss, hence one has to use masking. For the JS-divergence this is not necessary as misclassified pixels with $\boldsymbol{p}_y$ being small do not significantly influence the gradient, and thus the attack can focus on pixels which are not successfully perturbed yet without any mask.

Masked spherical loss: using the softmax output of a classifier can make, in some circumstances, the attack (partially) fail or weaker (Croce and Hein, 2020). A more direct approach is to directly minimize the logit of the correct class. However, we found it to work better when projecting the logits on the unit sphere: this recovers the structure of the spherical scoring rule (Bickel, 2007) which is a proper multi-class loss. We hypothesize that the projection first brings the logits of different pixels on the same scale, which balances the gradients deriving from each of them, and, second, involves the logits of all classes in the loss as part of the denominator. Since this loss is directly targeted to misclassification, we use it in combination with the mask for misclassified pixels:

$$\mathcal{L}_{\text{Mask-Sph.}} = -\mathbb{I}(\arg\max_{j=1,\dots,K} \boldsymbol{u}_j = y) \cdot \boldsymbol{u}_y / \left\|\boldsymbol{u}\right\|_2 .$$

3.2. Complementary performance of different losses for varying attack radii

In Table 1 we compare the effectiveness of the attacks (all using APGD) based on existing or proposed losses for a standard and an adversarially trained semantic segmentation model on PASCAL-VOC (details in Sec. 4). We are mainly interested in the attack performance across different radii ϵ for the ℓ_∞-threat model. We note that the best attack/loss depends on the radius, and is almost always achieved by one of our novel proposed losses. In particular, existing attacks have problems when using large ϵ values, as already observed in (Agnihotri and Keuper, 2023). When considering the image-wise worst attack regarding accuracy, we see that there is quite a gap between the worst-case over all attacks and the best single attack. This motivates our ensemble of attacks discussed next.

3.3. Segmentation Ensemble Attack (SEA)

Progressive radius reduction. In Table 1 we see that the worst case over losses is significantly lower than each individual attack. Besides the complementarity of the losses, this suggests that the optimization algorithm faces some issue regardless of the objective function, and can get stuck in suboptimal local minima. At the same time, increasing the perturbation set, i.e. larger ϵ_∞, reduces robust accuracy, which means that the gradient information provided by the models are still valid (no gradient masking is occurring). Thus, we take inspiration from Croce and Hein (2021), who adapted APGD to the ℓ_1-threat model to tackle the difficulties in optimizing the perturbations. In the case of semantic segmentation, we hypothesize that jointly optimizing the loss of thousands of pixels raises similar issues, and thus we adapt this technique to our task: we split the budget of iterations into three slots (with ratio $3:3:4$) where we run the attack with $2 \cdot \epsilon_\infty$, $1.5 \cdot \epsilon_\infty$ and ϵ_∞ respectively. The best adversarial attack found during each stage is then projected onto the smaller ℓ_∞-ball to start the algorithm in the next stage.

Radius reduction vs more iterations. To assess the effectiveness of the scheme with progressive reduction of the radius ϵ (red-ϵ) described above, we compare it with 300 iterations to the original APGD scheme (const-ϵ) of either 300 iterations or 100 iterations and 3 random restarts, so that all schemes have the same computational budget. We show in Fig. 2 (in the Appendix) the robust accuracy achieved by the three attacks with different objective functions, for $\epsilon_\infty \in \{8/255, 12/255\}$, on the adversarially trained model on PASCAL-VOC. One can observe that the red-ϵ APGD yields the best results (lowest accuracy) for almost every case, with large improvements especially at $\epsilon_\infty = 12/255$. This suggests that this scheme is better suited for generating stronger attacks on semantic segmentation models than common options used in image classification like more iterations or random restarts.

Final scheme. Distilling the findings of the complementary nature of the various losses at different robustness levels and the improvement in the optimization algorithm provided by the scheme with progressive radius shrinkage, we propose Segmentation Ensemble Attack, or SEA, as an evaluation protocol for segmentation models. It includes four runs of 300 iterations with red-ϵ APGD optimizing each of the four best losses found above, namely $\mathcal{L}_{\text{Mask-CE}}$, $\mathcal{L}_{\text{Bal-CE}}$, $\mathcal{L}_{\text{JS}}$ and $\mathcal{L}_{\text{Mask-Sph}}$. The motivation for this choice comes from the fact that the worst-case over these four losses leads to maximum 0.1% higher robust average accuracy or mIOU than using all six losses, and thus the two left-out losses, $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{CosSim-CE}}$, do not add much further value (see App. C.2). For more analysis see App. C.2.

4. Adversarially Robust Segmentation Models

In the following we discuss methods for robust segmentation models presented by prior work, and propose to take advantage of pre-trained robust classifiers to obtain robust

Table 1: **Comparison of losses at different robustness levels.** We use each of the losses discussed in Sec. 3.1 as objective in APGD with 100 iterations on one clean and one robust model. We report average pixel accuracy and mIOU (clean performance is indicated next to the model name). Depending on the radius ϵ_∞, the best results are realized by different losses. The worst-case over all runs is often significantly lower than each individual one.

ϵ_∞	losses used in prior works						proposed losses						Worst case over losses	
	$\mathcal{L}_{\text{CE}}$		$\mathcal{L}_{\text{Bal-CE}}$		$\mathcal{L}_{\text{CosSim-CE}}$		$\mathcal{L}_{\text{JS}}$		$\mathcal{L}_{\text{Mask-CE}}$		$\mathcal{L}_{\text{Mask-Sph.}}$			
Clean model (93.4 77.2)														
0.5/255	49.3	25.1	42.3	18.5	46.9	24.0	39.4	18.3	**36.9**	14.9	37.5	**14.5**	32.5	12.1
2/255	7.4	4.0	2.9	1.5	3.4	2.3	0.5	0.4	0.3	0.2	**0.1**	**0.1**	0.1	0.0
Adversarially trained classifier (92.7 75.9)														
4/255	88.9	65.7	88.7	**64.8**	88.9	65.4	**88.4**	**64.8**	88.9	65.6	90.4	69.7	88.3	64.4
8/255	78.9	48.9	**74.2**	**41.3**	77.8	47.3	75.3	43.5	74.6	41.8	80.3	49.6	72.3	38.4
12/255	59.9	28.9	43.3	14.9	56.6	26.4	45.1	18.6	**38.8**	13.2	38.9	**12.1**	31.9	8.4
16/255	41.5	18.1	20.7	5.7	34.0	15.3	19.1	7.4	12.9	3.4	**8.4**	**2.0**	6.4	1.1

segmentation models. Standard adversarial training is done and all evaluations are carried out with our SEA on the entire validation set, see App. B for details.

4.1. Existing work on adversarially robust semantic segmentation models

As mentioned above, unlike for image classification, only a few works have applied adversarial training to obtain robust segmentation models. Gu et al. (2022) do adversarial training with 3 or 7 steps of their SegPGD at $\epsilon_\infty = 8/255$, using a ResNet-50 as backbone in a PSPNet (Zhao et al., 2017) architecture. However, the obtained adversarial robustness, see upper part of Table 2, is relatively low. Since the models are not available, we can only show the robustness values reported in their paper which are based on the original SegPGD attack and not evaluated using SEA, but one can see that the improvement in mIOU provided by AT^7 is smaller than 14% compared to the clean model. Moreover, under the evaluation with SEA our clean model has no robustness already at $\epsilon = 4/255$, suggesting that the reported robustness for the models from Gu et al. (2022) might be overestimated and would significantly decrease with SEA.

4.2. Robust models via robust initialization

Since Liu et al. (2022) showed that their ConvNeXt, one of the currently most popular architectures for vision tasks, is very effective also for semantic segmentation, we use it as backbone in UPerNet (Xiao et al., 2018b). Moreover, Singh et al. (2023) have recently shown substantial improvements in adversarial training on IMAGENET using the ConvNeXt architecture. We use here ConvNeXt-T (similar size as ResNet-50), and present the results for larger ConvNeXt backbones in App. C.3.

PASCAL-VOC. Table 2 reports the statistics about the robustness of the various models trained on PASCAL-VOC. First, we use a ConvNeXt given by clean pre-training on IMAGENET as initialization for the backbone (the decoder is randomly initialized). When using 2 steps of PGD for generating the perturbations for training (denoted as AT^2), 50 epochs of adversarial training are not sufficient to achieve non-trivial robustness. However, increasing the length of training to 300 epochs makes the model significantly more robust, suggesting that learning robust segmentation models is a challenging problem. Then, in order to give a warm start to the training algorithm, we initialize the backbone with a robust image classifiers, adversarially trained on IMAGENET at $\epsilon_\infty = 4/255$. 50 epochs of AT^2 from robust initialization leads to 86.7% of robust accuracy at $\epsilon_\infty = 4/255$, and above 50% at $\epsilon_\infty = 8/255$ (while having only 0.5% lower clean accuracy than the clean model which has 0% robustness). This is significantly better than using 300 epochs from clean initialization. Moreover, it already exceeds, even compared to their original evaluation, the results reported by Gu et al. (2022). We further test the effect of increasing the number of steps for training from 2 to 5 (AT^5), which doubles the computational cost per epoch. In this case, even 50 epochs from clean initialization give a model with good robustness, which again improves with 300 epochs. Even with AT^5, using the robust initialization allows us to match (or outperform at large ϵ values) with 50 epochs the robustness of the models with clean initialization and 300 epochs. This shows that robust classifiers, commonly available, can significantly help in reducing the cost of getting robust segmentation models. Finally, our models show more than 2x higher robustness than reported in Gu et al. (2022). Interestingly, the large gains in robustness do not degrade much the clean performance, which is a typical drawback of adversarial training.

Table 2: **Comparison of training schemes for PASCAL-VOC.** For each model and ϵ_∞ value we report robust average pixel accuracy (white background columns) and mIoU (grey background columns) given by our SEA. * indicates that the result is taken from the original paper (since the model is not available) and obtained using a weaker attack.

Training scheme	0		4/255		8/255		12/255		16/255	
PSPNet with ResNet-50 backbone (previous works)										
clean (Gu et al., 2022)	–	76.6	–	–	–	3.4*	–	–	–	–
AT^3 (Gu et al., 2022)	–	75.4	–	–	–	10.3*	–	–	–	–
AT^7 (Gu et al., 2022)	–	74.4	–	–	–	17.0*	–	–	–	–
UPerNet with ConvNeXt-T backbone (ours)										
clean	93.4	77.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AT^2 clean init. 50 ep.	93.4	77.4	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
AT^2 clean init. 300 ep.	93.1	76.3	86.5	59.6	44.1	16.6	4.6	0.1	0.0	0.0
AT^2 robust init. 50 ep.	92.9	75.9	86.7	60.8	50.2	21.0	9.3	2.4	0.8	0.3
AT^5 clean init. 50 ep.	91.9	73.1	86.2	59.2	64.6	28.3	20.7	4.9	2.0	0.4
AT^5 clean init. 300 ep.	92.8	75.5	88.6	64.4	71.4	37.7	23.4	6.6	2.7	0.6
AT^5 robust init. 50 ep.	92.7	75.2	88.3	63.8	71.2	37.0	27.4	8.1	4.2	0.9

Table 3: **Comparison of training schemes for ADE20K.** We repeat the evaluation from Table 2 on ADE20K, for which no previous work presented robust models. Robust ACC and mIOU evaluated with SEA are shown for UPerNet with ConvNeXt-T backbone.

Training scheme	0		4/255		8/255		12/255	
clean 128 ep.	75.5	41.1	0.0	0.0	0.0	0.0	0.0	0.0
AT^5 clean init. 128 ep.	68.0	26.1	52.4	14.0	24.7	4.7	2.4	0.3
AT^5 robust init. 32 ep.	68.8	25.2	55.4	15.6	28.3	5.9	3.8	0.7
AT^5 robust init. 128 ep.	70.5	31.7	55.6	18.6	26.4	6.7	3.3	0.8

ADE20K. We further test the effectiveness of our scheme for obtaining robust models on the more challenging ADE20K dataset, with 150 object classes compared to 20 of PASCAL-VOC. We remark that, we train our models to predict also a background class, and similarly the attacks can use it to induce misclassification. Similar to PASCAL-VOC, Table 8 in App. C.3 one can see that 128 epochs (used following Liu et al. (2022)) of AT^2 from clean initialization are not sufficient to obtain a robust model, while they are with robust initialization. For AT^5 in Table 3, the model initialized with the robust backbone has higher clean and robust performance than that with standard backbone.

To test whether the robust initialization allows us to save training time, we additionally report (in Table 3) a model trained for only 32 epochs: it outperforms the one from clean initialization with 4x lower computational cost. Moreover, it has similar performance to the model with 128 epochs and robust initialization in the target threat model ($\epsilon = 4/255$), while it trades-off some clean performance for robustness at higher radii (longer training can fit better the training data, improving clean accuracy). We highlight that ours are the first adversarially trained models reported for ADE20K, which explains the lack of baselines.

5. Conclusion

We have shown that adversarial attacks on semantic segmentation models can be improved by adapting the optimization algorithms and objective functions, developing SEA, an ensemble of attacks which outperforms existing methods. This may open new research directions, for example for losses which take into account the interaction of neighboring pixels or directly target mIOU to achieve stronger attacks. Moreover, we could train segmentation models with SOTA robustness, even at limited computational cost by taking advantage of adversarially pre-trained image classifiers. It will be interesting to test the effect of applying our method to other architectures.

Limitations. We consider SEA an important step towards strong evaluation of robustness for semantic segmentation models. However, as shown for image classification (Croce and Hein, 2020), PGD-based attacks should be complemented by white-box attacks of different type and especially black-box methods.

References

Shashank Agnihotri and Margret Keuper. Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*, 2023.

Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than CNNs? In *NeurIPS*, 2021.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

J Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49–65, 2007.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *ECML/PKKD*, 2013.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *NeurIPS 2017 Workshop on Machine Learning and Computer Security*, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In *AAAI*, 2018.

Seungju Cho, Tae Joon Jun, Byungsoo Oh, and Daeyoung Kim. Dapas: Denoising autoencoder to prevent adversarial attack in semantic segmentation. In *IJCNN*. IEEE, 2020.

Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Francesco Croce and Matthias Hein. Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers. In *ICML*, 2021.

Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In *AAAI*, 2022.

Edoardo Debenedetti. Adversarially robust vision transformers. Master's thesis, Swiss Federal Institute of Technology, Lausanne (EPFL), 2022.

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: 303–338, 2010.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*, 2016.

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *ECCV*, 2022.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? natural language attack on text classification and entailment. In *AAAI*, 2019.

Xu Kang, Bin Song, Xiaojiang Du, and Mohsen Guizani. Adversarial attacks for image segmentation on multiple lightweight models. *IEEE Access*, 8:31359–31370, 2020.

Nikhil Kapoor, Andreas Bär, Serin Varghese, Jan David Schneider, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. From a fourier-domain perspective on adversarial examples to a wiener filter defense for semantic segmentation. In *IJCNN*, 2021.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*, 2021.

Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *arXiv preprint, arXiv:2302.14301*, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018.

Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2280–2289, 2022.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *CVPR*, 2019.

Guangyu Shen, Chengzhi Mao, Junfeng Yang, and Baishakhi Ray. Advspade: Realistic unrestricted attacks for semantic segmentation. *arXiv preprint arXiv:1910.02354*, 2019.

Naman D Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *arXiv preprint arXiv:2303.01870*, 2023.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *NeurIPS*, 2020.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*, 2023.

Eric Wong, Frank R Schmidt, and J Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *ICML*, 2019.

Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *ECCV*, 2018a.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018b.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.

Xiaogang Xu, Hengshuang Zhao, and Jiaya Jia. Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In *ICCV*, 2021.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.