Uncertainty-weighted Loss Functions for Improved Adversarial Attacks on Semantic Segmentation

Kira Maag
Technical University of Berlin, Germany
maag@tu-berlin.de

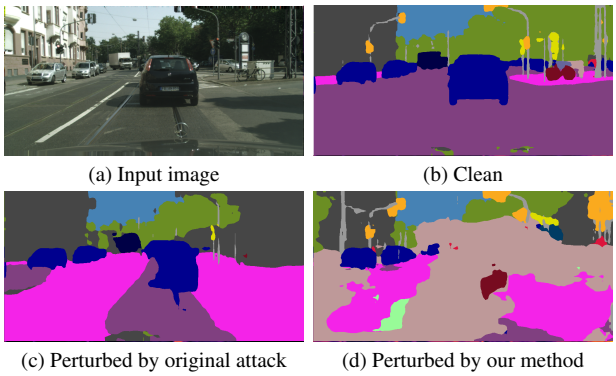
Asja Fischer
Ruhr University Bochum, Germany
asja.fischer@rub.de

Abstract

State-of-the-art deep neural networks have been shown to be extremely powerful in a variety of perceptual tasks like semantic segmentation. However, these networks are vulnerable to adversarial perturbations of the input which are imperceptible for humans but lead to incorrect predictions. Treating image segmentation as a sum of pixel-wise classifications, adversarial attacks developed for classification models were shown to be applicable to segmentation models as well. In this work, we present simple uncertainty-based weighting schemes for the loss functions of such attacks that (i) put higher weights on pixel classifications which can more easily perturbed and (ii) zero-out the pixel-wise losses corresponding to those pixels that are already confidently misclassified. The weighting schemes can be easily integrated into the loss function of a range of well-known adversarial attackers with minimal additional computational overhead, but lead to significant improved perturbation performance, as we demonstrate in our empirical analysis on several datasets and models.

1. Introduction

Deep neural networks (DNNs) have been shown to be extremely powerful in a wide range of perceptual tasks, such as semantic image segmentation [5, 25] for which they demonstrate an outstanding prediction performance. Semantic segmentation provides comprehensive and precise information about the given image by assigning each pixel to a predefined and fixed set of semantic classes resulting in segmented objects. However, many studies have found that DNNs are vulnerable to *adversarial attacks* [2, 3]. Adversarial attacks generate slightly perturbed versions of the input images which fool the DNN, i.e., change the network predictions at test time, see for example Fig. 1. These small perturbations are not perceptible to humans making adversarial examples very hazardous in safety-related applications like automated driving. Thus, the development of efficient defense strategies against adversarial attacks is of highest interest. These strategies either enhance the robustness of DNNs rendering it more challenging to generate adversarial examples, or rely on approaches to detect adversarial attacks. In general, there are three common approaches to increase the robustness of DNNs. The first class of approaches aims to enhance the robustness of the network by modifying the training process (e.g. [15, 31]). Second, input denoising procedures, like autoencoder based reconstruction [7] or inpainting [32], are considered to remove the perturbation from the input. The third class of approaches increases the robustness during inference, e.g. by multi-scale processing [2]. For the detection of adversarial examples, the patch-wise spatial consistency check [29] and an uncertainty-based method [17] have been proposed. Regardless of strategy, it is a continuous loop between the development of defense/detection strategies and adversarial attackers. This also means that the development of new faster and stronger attacks is important in order to strengthen the models against them and thus, enhancing general model robustness.



(a) Input image (b) Clean (c) Perturbed by original attack (d) Perturbed by our method

Figure 1. Semantic segmentation prediction for a clean image, a perturbed image generated by the original attack proposed in [16] and a perturbed image created by the same attack using our uncertainty-weighted loss.

Prior works on adversarial attacks focus on the image

classification task and some of them are transferred to the semantic segmentation task by treating each pixel labeling independently as a separate classification task [12, 16, 20]. Moreover, there were adversarial examples specifically developed for the semantic segmentation task where all pixels of an image are attacked until selected pixels have been misclassified into the target class, i.e., pixels of a selected class appear or disappear, or even the entire image changes [8, 21]. In comparison to these methods, the patch-wise attack [23, 24] perturbs a small rectangular region of the image aiming to cause prediction errors in the whole image. Recently, the certified radius-guided attack framework for segmentation models [26] was introduced. The idea is to disturb pixels with comparatively smaller certified radii since a smaller theoretically certified radius should relate to lower robustness to adversarial perturbations.

In this paper, we present an uncertainty-based weighting scheme which can be incorporated into the loss function of any untargeted attack on semantic segmentation models that is composed out of pixel-wise attacks. Uncertainty information, such as Monte-Carlo Dropout [11] or maximum softmax [14], is considered for prediction error [19, 28] and out-of-distribution detection [18]. These works demonstrate the correlation between uncertainty measures and erroneous predictions. The idea behind our method is to include uncertainty information into the loss function of an adversarial attack to degrade the performance of the network even more. To this end, we consider different *white box* attacks, i.e., the attacker has full access to the model including parameters and loss function used during training. In contrast, *black box* methods gain zero knowledge about the model to attack. The expectations for any attack are low runtimes and computational effort while at the same time having powerful perturbation effects. We modify the loss function of well-known adversarial methods by introducing an uncertainty-weighted loss. On the one hand, we put higher weights on pixel classifications which can more easily perturbed and on the other hand, we zero-out the pixel-wise losses corresponding to those pixels that are already confidently misclassified. Thus, our approach can be incorporated into any attack with minimal additional computational overhead, but improved perturbation performance. First, we apply our loss function to pixel-wise attacks for semantic segmentation. Second, we replace the certified radius-guided loss function introduced in [26] by our uncertainty-based weighting scheme. Last, we introduce an alternative approach for the patch-based attack where only a few pixels are attacked. To this end, we choose randomly a subset of pixels to attack and apply our loss function in combination with the pixel-wise iterative *fast gradient sign method* [16].

In our tests, we employ state-of-the-art semantic segmentation networks [5, 25, 33, 34] applied to the Cityscapes [9] as well as the Pascal VOC2012 dataset [10] demonstrating our adversarial attack performance. We apply our approach to different types of attacks, such as pixel-level attackers designed for image classification [12, 20] and pixel-wise attacks developed for semantic segmentation [26, 27]. The source code of our method is publicly available at `https://github.com/kmaag/Uncertainty-weighted-Loss`. Our contributions are summarized as follows:

- For the first time, we present an uncertainty-based weighting scheme which can be incorporated into the loss function for white box adversarial attacks which has low computationally overhead compared to the original attack.
- Our method is not designed for a specific adversarial attack, rather we enhance different types of attackers in a light-weight manner. We achieve attack pixel success rate values of up to 99.82% across different network architectures and datasets.
- We propose an approach which attacks only a subset of the image pixels, similar to the patch attack, but also leading to erroneous prediction of the entire image.

The paper is structured as follows. In Sec. 2, we present various adversarial attacks for the semantic segmentation task which serve as baselines in our work. We introduce our method in Sec. 3. In Sec. 4, the numerical results are shown, followed by a comparison with related work in Sec. 5 and a conclusion in Sec. 6.

2. Background

In this section, we recall the semantic segmentation task as well as different previously proposed adversarial attackers to semantic segmentation models. Note, that all described attacks belong to the white box setting.

Semantic segmentation To obtain a semantic segmentation, i.e., pixel-wise classification of image content, each pixel z of an input image x gets assigned a label $\tilde{y}_z$ from a prescribed label space $C = \{y_1, \dots, y_c\}$. A neural network given learned weights w provides for the z-th pixel a probability distribution $f(x;w)_z \in \mathbb{R}^{|C|}$ specifying the probability for each class $y \in C$ denoted by $p(\cdot|x)_z \in \mathbb{R}$. The predicted class is then computed by $\hat{y}_z^x = \arg\max_{y \in C} p(y|x)_z$. To train the semantic segmentation network, a pixel-wise loss function (generally the cross entropy) is simultaneously minimized for all pixels $z \in Z$ of an image x. The complete loss function is then given by

$$L(f(x;w),y) = \frac{1}{|Z|} \sum_{z \in Z} L_z(f(x;w)_z, y_z) \ , \qquad (1)$$

where y_z denotes the one-hot vector of the label.

Adversarial attacks A well-known adversarial attack developed for image classification but also applied to semantic segmentation is the *fast gradient sign method* (FGSM, [12]). This (untargeted) single-step attack adds small perturbations to the image x leading to an increase of the loss of

$$x^{adv} = x + \varepsilon \cdot \text{sign}(\nabla_x L(f(x; w), y)) \ , \tag{2}$$

where ε describes the magnitude of perturbation, i.e., the ℓ_∞-norm of the perturbation is bounded to be (at most) ε. This attack is extended to the *iterative FGSM* (I-FGSM, [16]) increasing the perturbation strength by

$$x_{t+1}^{adv} = \text{clip}_{x,\varepsilon}(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} L(f(x_t^{adv}; w), y))) \ , \tag{3}$$

where $x_0^{adv} = x$, α defines the step size, and a clip function ensures that $x_t^{adv} \in [x - \varepsilon, x + \varepsilon]$. The *projected gradient descent* (PGD, [20]) attack is similar to the iterative FGSM. The difference between the two methods is that PGD choose the starting point randomly within the ℓ_∞ ball of interest (and does random restarts), while I-FGSM initializes to the original point. These approaches serve as basis for further elaborated attacks like the *orthogonal PGD* [4] or *DeepFool* [22].

Furthermore, there have been developed adversarial attacks especially for the semantic segmentation task such as adaptions of the PGD attack [1, 13]. The introduced *ALMA prox* attack [27] is based on a proximal splitting to produce adversarial perturbations with much smaller ℓ_∞-norm in comparison to FGSM and PGD. Recently, a *certified radius-guided* (CR) attack framework for segmentation models was proposed [26]. The certified radius specifies the size of an ℓ_p ball around a pixel in which a perturbation is guaranteed to not change the class predicted for the pixel. Thus, a larger certified radius indicates more robustness to adversarial perturbations. The idea of the framework is to focus on disrupting pixels with relatively smaller certified radii.

In contrast to attacks adding perturbations to all pixels, *patch attacks* [23] disrupt a small rectangular region of the image aiming at prediction errors in a much larger region, i.e., the whole image. In [24], an individual patch attack is introduced, the *expectation over transformation-based attack*, creating robust adversarial examples to perturb a range of transformations at the same time. In the real world scenario, transformations consist of angle and viewpoint changes for instance. To generate these strong perturbing patches for the semantic segmentation task within the optimization procedure an extension of the pixel-wise cross entropy loss is introduced.

There also exist targeted attacks specifically developed for semantic segmentation. For the *stationary segmentation mask method* [8,21,30] the pixels of an image are iteratively perturbed until most of the pixels have been misclassified as belonging to the target class given by an arbitrary segmentation defined by the attacker. The *dynamic nearest neighbor method* [6,21] is intended to remove one desired target class (like pedestrians from street scene images) but keep for all other classes the network's segmentation unchanged.

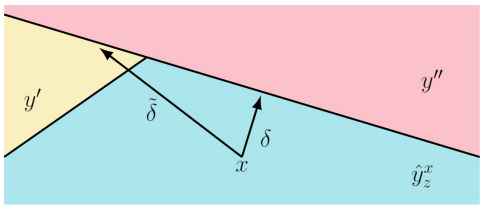


Figure 2. An illustration of the decision boundaries between three different classes where $\hat{y}_z^x$ is the predicted class of input x, y'' the class with second highest probability and y' another class. The arrows δ and $\tilde{\delta}$ represent distances from x to the decision boundaries.

3. Uncertainty-weighted loss functions

In the following, we describe the two weighting schemes which follow different motivations: on the one hand focusing on pixels whose classification is easily perturbable, and on the other paying no attention to those pixels which are already misclassified with high confidence.

3.1. Focusing on easily to perturb pixels

Recall, that for creating an adversarial attack we want to add a perturbation (of a small predefined magnitude) to the image that changes as many pixel-wise classifications as possible. Intuitively, due to the restriction of the magnitude of the perturbation, it makes sense to focus on those pixel classifications that are easily to perturb. Geometrically, a small shift of the input is more likely to lead to a different classification result as closer the input is to the decision boundary between the current and other classes. Let $\hat{y}_z^x$ be the predicted class for pixel z and lets define y' to be another class. In a linear model the minimal distance to the decision boundary separating class $\hat{y}_z^x$ from class y' is then defined by

$$\delta = \frac{p(\hat{y}_z^x|x)_z - p(y'|x)_z}{||\nabla_x (p(\hat{y}_z^x|x)_z - p(y'|x)_z)||_2} \ . \tag{4}$$

An illustration is given in Fig. 2. If an attack would focus on the misclassification of only a single pixel it would make sense to focus at the one with the smallest distance to the second most likely class. In this case the numerator of Eq. (4) is given by the probability margin

$$M(x)_z = p(\hat{y}_z^x|x)_z - \arg\max_{y \in C \setminus \{\hat{y}_z^x\}} p(y|x)_z \ . \tag{5}$$

However, neural networks are not linear models and the attacks aim at the misclassification of all pixels. Therefore,

the adversarial perturbation of the input does not necessarily points into the direction of shortest distance to a decision boundary. To take this into account a better indication of which pixel classification is easiest to attack could be given by the difference between the highest and lowest class probability

$$D(x)_z = p(\hat{y}_z^x|x)_z - \min_{y\in\mathcal{C}} p(y|x)_z \ , \tag{6}$$

which serves as a proxy of maximal distance to a decision boundary, i.e., the distance to the least likely class. If the prediction is highly certain, the probability of the least likely class is equal or close to zero.

Instead, one can also estimate the mean of the margins to all other classes

$$\bar{M}(x)_z = \frac{1}{|C|-1} \sum_{y\in\mathcal{C}\setminus\{\hat{y}_z^x\}} p(\hat{y}_z^x|x)_z - p(y|x)_z \ . \tag{7}$$

Moreover, the entropy

$$E(x)_z = -\sum_{y\in\mathcal{C}} p(y|x)_z \cdot \log p(y|x)_z \tag{8}$$

could be a good indicator, since it is often used as uncertainty measure for the semantic segmentation task and it is related to a weighted mean margin (where each margin is weighted by $log(p(y|x)_z)$, i.e., smaller margins get a higher weight).

The proposed indicators for the closeness to decision boundaries can than be used as weighting factors for the pixel-wise loss functions when calculating the adversarial attacks. Such weighted versions of the FGSM and I-FGSM method would for example be obtained by replacing $L(f(x;w),y)$ in Eq. (2) and Eq. (3), respectively, by

$$L^U(f(x;w),y) = \frac{1}{|Z|} \sum_{z\in Z} e^{U(x)_z} \cdot L_z(f(x;w)_z, y_z) \tag{9}$$

where $U(x)_z \in \{1 - M(x)_z, 1 - D(x)_z, 1 - \bar{M}(x)_z, E(x)_z\}$. The pixels with high uncertainty, i.e., larger values of $U(x)_z$, are weighted stronger in the loss during the adversarial example generation in order to lead as many pixels as possible to a wrong prediction. We re-scale the uncertainty values $U(x)$ by the exponential function to further emphasize high uncertainties.

3.2. Ignoring confidently wrongly classified pixels

If pixels are already predicted incorrectly, it makes no sense to continue to give them a high weighting. Thus, it is important to focus on pixels, when applying the loss function, that are still correctly classified. That is, the loss of pixels which are already misclassified with sufficiently high confidence can be neglected. Therefore, during the attack generation process, we set the loss of all pixels which are misclassified with a probability of at least 75% to zero. This is achieved by the following weighting scheme

$$L^0(f(x;w),y) = \frac{1}{|Z|} \sum_{z\in Z} \mathbb{1}_{(\hat{y}_z^x = \tilde{y}_z \vee p(\hat{y}_z^x|x) < 0.75)} \cdot L_z(f(x;w)_z, y_z) \tag{10}$$

with ground truth class $\tilde{y}_z$. The confidence of the misclassification (as measured by $p(\hat{y}_z^x|x)$) has to be taken into account, since the corresponding pixel is still perturbed based on the gradients of the other pixel-wise loss functions. If the uncertainty is high, i.e., the confidence is low, this could shift the pixel accidentally back into the correct class. In general, our uncertainty-based weighting scheme, L^U and L^0, can be inserted into the loss function used in any untargeted and pixel-wise adversarial attack.

4. Experiments

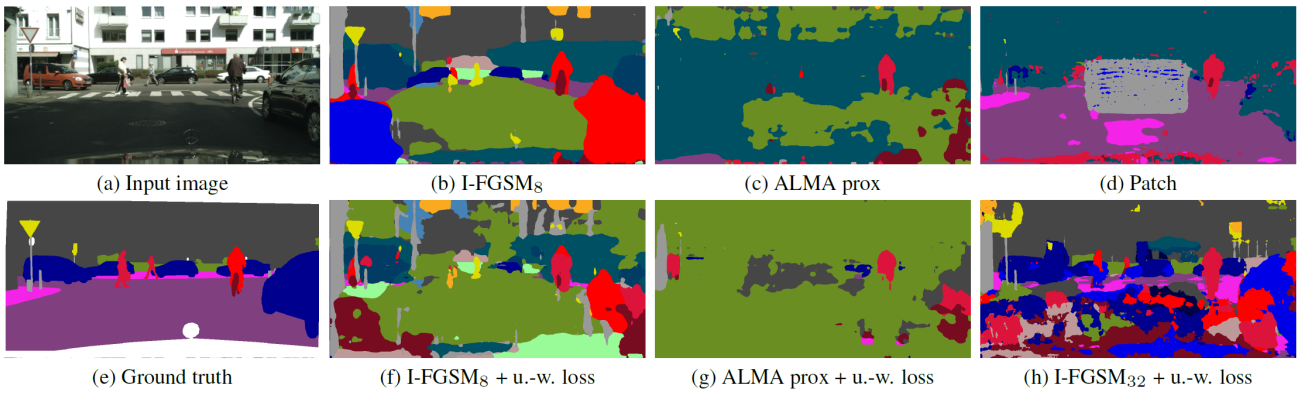
In this section, we describe the experimental setting first and then evaluate our adversarial attack performance.

4.1. Experimental setting

Datasets The experiments are conducted on two datasets, Cityscapes [9] and Pascal VOC2012 [10] (shorthand VOC). The latter dataset, for visual object classes in realistic scenes, consists of $1{,}464$ training and $1{,}449$ validation images with annotations for different objects of categories person, animal, vehicle and indoor. The Cityscapes dataset, for semantic segmentation in street scenes, contains $2{,}975$ training and 500 validation images of dense urban traffic in 18 and 3 different German towns, respectively.

Segmentation networks In our tests, we consider four different pre-trained state-of-the-art networks. Trained on the Cityscapes dataset, the BiSeNet [33] achieves a mean intersection over union (mIoU) of 74.37% on the validation set and the DDRNet [25] of 77.80%. Moreover, we employ the DeepLabv3+ network [5] trained on Cityscapes obtaining a validation mIoU of 79.61% and on VOC of 76.81%. The PSPNet [34] achieves a mIoU value of 76.78% on the VOC validation set.

Adversarial attacks We consider the well-known and in defense approaches often considered [2,3,7,15] FGSM and I-FGSM attacks in our tests with parameter setting proposed in [16]. The step size is given by $\alpha = 1$ and the perturbation magnitude by $\varepsilon = \{4, 8, 16\}$ resulting in a number of iterations of $n = \min\{\varepsilon + 4, \lfloor 1.25\varepsilon \rfloor\}$. The corresponding (iterative) FGSM attack is denoted by FGSM_ε and $\text{I-FGSM}_\varepsilon$. As another pixel-wise attack originally developed for image classification, we use the PGD attack with parameters $\alpha = 1/30$, $\varepsilon = 1$, $n = 40$ and one restart. In addition, we



(a) Input image (b) I-FGSM$_8$ (c) ALMA prox (d) Patch

(e) Ground truth (f) I-FGSM$_8$ + u.-w. loss (g) ALMA prox + u.-w. loss (h) I-FGSM$_{32}$ + u.-w. loss

Figure 3. Segmentations for weighted and un-weighted attacks. (a) Input image from the Cityscapes dataset and (e) corresponding ground truth. Semantic segmentation prediction for perturbed images generated by (b) iterative FGSM, (c) ALMA prox, (d) patch attack, (f) FGSM with uncertainty-weighted (u.-w.) loss, (g) ALMA prox with u.-w. loss and (h) iterative FGSM with u.-w. loss applied to only a subset of pixels.

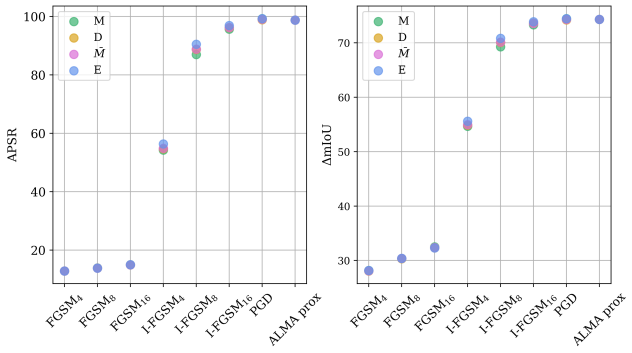


Figure 4. Comparison of uncertainty measures M, D, $\bar{M}$ and E used for weighting loss functions of different adversarial attacks for the DeepLabv3+ network applied to the Cityscapes dataset.

employ various adversarial example generation techniques specifically designed for the semantic segmentation task. Firstly, we consider the ALMA prox attack with default parameters using the implementation of [27]. For the described attacks above, we employ the model zoo[1] including the pre-trained models. Secondly, for the certified radius-guided approach [26], we use the provided code with two parameter settings, i.e., ℓ_2-norm with $\varepsilon = 1$ and ℓ_∞-norm with $\varepsilon = 0.004$ (hyperameters all under default setting). Lastly, we consider the patch attack [24] using the available repository with default parameters applied to the BiSeNet and the DDRNet tested on the real world Cityscapes dataset.

The model used to compute the certified radius-guided method re-scales the VOC images to 473×473 which we keep also for the other attacks. Since the Cityscapes dataset provides high-resolution images of size 1024×2048, we re-scale the image size to 512×1024 for the computation of the adversarial examples to reduce the amount of memory to run a full backward pass. Figure 3 (top row) shows semantic segmentation predictions for a few attacks applied to the Cityscapes dataset and the BiSeNet network.

[1] https://github.com/open-mmlab/mmsegmentation

Evaluation metrics To access the performance of the adversarial attackers, we use the attack pixel success rate (APSR) [27] which is defined by

$$\text{APSR} = \frac{1}{|Z|} \sum_{z \in Z} \arg\max_{y \in \mathcal{C}} p(y|x)_z \neq \tilde{y}_z \qquad (11)$$

with ground truth class $\tilde{y}_z$. This metric measures the number of falsely predicted pixels and thus, successfully attacked pixels. Furthermore, we consider the difference of the mIoU obtained on clean images and the mIoU obtained on perturbed images as performance metric, denoted by ΔmIoU. Note, this metric is bounded by the mIoU value on clean images.

4.2. Comparison of different weighting schemes

In Sec. 3.1, we introduced four different measures for estimating which pixel classification can easily be disturbed: the probability margin M, the difference between the highest and lowest probability value D, the mean of the margins $\bar{M}$, and the entropy E. The first experiment aimed at analysing which of these measures performs best. For that, we have computed the different weighted attacks on the DeepLabv3+ network applied to the Cityscapes dataset. Figure 4 shows a comparison of the performance in terms of APSR and ΔmIoU. The scores are fairly close and for the weaker (FGSM) as well as stronger attacks (PGD and ALMA prox) all measures lead to almost the same results. For all attacks, the values resulting from weighting with the difference between the highest and lowest probability D and

the mean of the margins $\bar{M}$ are almost equal. Overall, the probability margin performs the worst, while the entropy shows slight improvements for the iterative FGSM attack. These results indicate, that it is advantageous to consider the margin for more than one class and that higher weighting margins to more likely classes (as approximately done by the entropy) is also beneficial. Thus, for the following experiments, we consider only the entropy as uncertainty measure.

4.3. Evaluation of uncertainty-weighted attacks

A comparison between the attack performance of the original proposed attacks and the attacks resulting from replacing the original loss function with our incorporated uncertainty-weighted loss function, i.e., Eq. (9) with entropy as uncertainty measure and Eq. (10), is given in Fig. 5 for the Cityscapes dataset and in Fig. 6 for the VOC dataset. We observe increased APSR and ΔmIoU performance for larger magnitudes of perturbation for the non-iterative as well as the iterative FGSM attacks. Our approach clearly outperforms the original FGSM attack, as well in its simple as in its iterative version. Examples of segmentations stemming from the original I-FGSM as well as the uncertainty-weighed counterpart (L^0) are shown in Fig. 3 (b) and (f), respectively. We obtain the largest performance boost for the PGD attack where the incorporation of the weighting leads up to 62.5 percentage points (pp) higher APSR values. A closer inspection shows, that the original PGD attack performs poorly for the Cityscapes dataset. The reason for this is that using this attack, the same wrong class is often predicted for all pixels of an image. However, if the perturbation fails to do that, only a few pixels are predicted incorrectly. But with the uncertainty-weighted loss (L^0), almost all images are predicted incorrectly. Alma prox is comparatively the strongest attack and achieves APSR values of over 99%. It is therefore difficult to improve the results any further, see for example Fig. 3 (c) and (g). The APSR values for the original attack and our uncertainty-based weighting scheme are very similar, although we can enhance the ΔmIoU values for the Cityscapes dataset. In general, the more extreme weighted loss function L^0 outperforms the entropy-weighted loss function L^E for both datasets and investigated networks. This behavior is in the nature of the weighting manner, i.e., in L^0 pixels that are certainly incorrectly predicted are set to zero in the loss function and are therefore no longer considered, while in L^E only a weaker weighting is applied. We also experimented with a combination of both loss functions which however did not increase the attack performance over using L^0 alone.

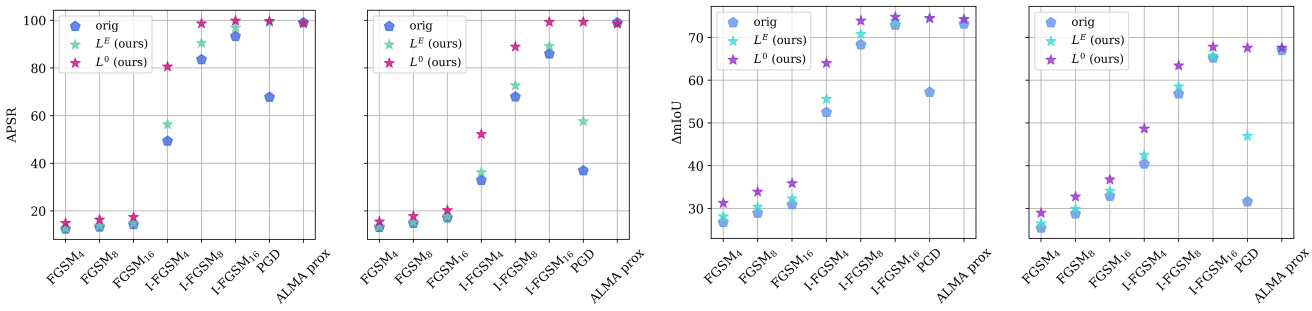
The runtimes are given in Tab. 1 for the different attackers. The results are averaged over the number of validation images and measured on a NVIDIA A40 GPU. The runtimes for the weighted loss functions are quite comparable and we quote the highest value here. Incorporating our uncertainty-based weighting scheme into existing adversarial attack generation models increases the runtimes negligible but improves the performance greatly. Using our proposed weighting scheme, the I-FGSM$_{16}$ attack as well as the PGD attack attain similar performance values as the strong ALMA prox attack, but at lower runtimes. Only needing about 3 to 5% of the runtime of ALMA prox, the advantage is especially drastic for the iterative FGSM. Note, the original attacks only achieve weaker performance.

4.4. Comparison with CR attack

The certified radius-guided approach is similar to our uncertainty-based weighting scheme as both methods weight the pixel-wise losses of the adversarial example generator to achieve a high attack success rate. We use the framework provided in the original paper [26] and replace their CR weighting procedure with our uncertainty-based one to have a fair comparison of both methods. As shown in Sec. 4.3, the L^0 loss function performs best and is used in the following experiments. In Tab. 2 (left), the numerical results are shown for the VOC dataset and the PSPNet. Note, the code for more datasets and models is not released by the authors up to now. With up to 26.62 pp higher APSR and 8.86 pp higher ΔmIoU values, our approach clearly outperforms the CR method for both parameter settings and evaluation metrics. Moreover, in Tab. 2 (right) the runtimes for the adversarial attackers are given where "clear" means that the underlying attack is considered without any weighting scheme in the loss function. Our approach shows only a minimally extended runtime, while the CR method shows a runtime 1.4 times larger. Thus, the proposed weighting scheme substantially improves over the CR based weighting in terms of performance and runtime.

4.5. Perturbation of a reduced number of pixels

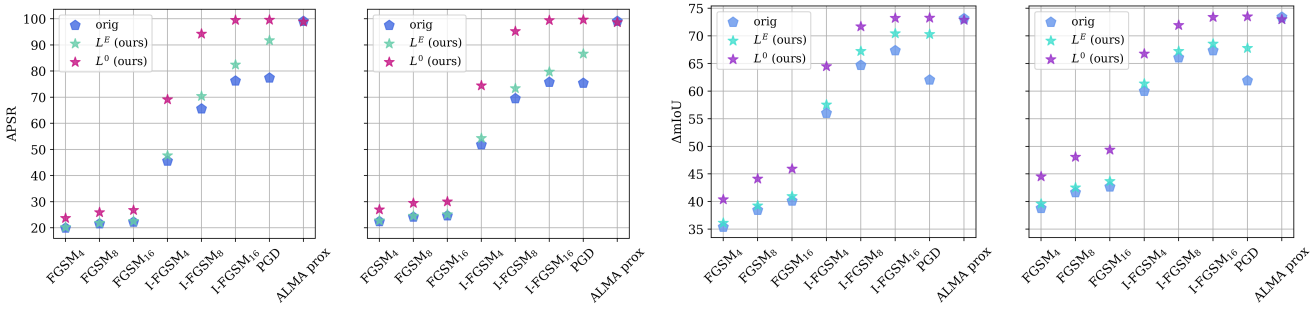
Patch attacks disrupt a small rectangular region of the image aiming at prediction errors in a much larger region, i.e., the whole image. We propose an alternative to the patch attack where also only a few pixels are perturbed. To this end, we choose randomly a subset of pixels to attack and apply our uncertainty-based loss function in combination with the iterative FGSM. Note, we disturb the same number of pixels like the considered patch method. Both methods are difficult to compare since the expectation over transformation-based attacking patch trains the attacker to successfully perturb the image over a range of transformations while our method perturbs only random pixels with the I-FGSM attack. Our aim is to propose another way of attacking only a few pixels and achieving at the same time a high prediction damage. In Fig. 7, the performance results of the patch attack for the Cityscapes dataset are given in comparison to our approach using various magnitudes



(a) DeepLabv3+ (left) and BiSeNet (right)

(b) DeepLabv3+ (left) and BiSeNet (right)

Figure 5. APSR (a) and ΔmIoU (b) results for different attacks on two networks trained on the Cityscapes dataset.



(a) DeepLabv3+ (left) and PSPNet (right)

(b) DeepLabv3+ (left) and PSPNet (right)

Figure 6. APSR (a) and ΔmIoU (b) results for different attacks on two networks trained on the VOC dataset.

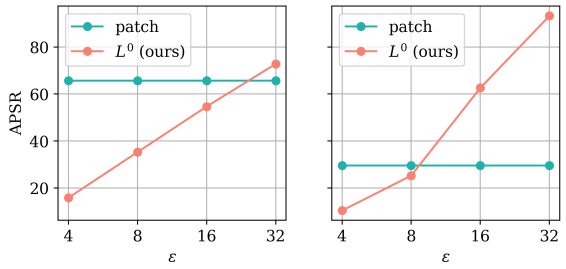


Figure 7. APSR results for the BiSeNet (left) and the DDRNet (right) applied to the Cityscapes dataset for a comparison between the patch attack and our approach (iterative FGSM attack with uncertainty-weighted loss applied to only a subset of pixels).

of perturbation for the I-FGSM attack. For the BiSeNet, we need a perturbation magnitude of 32 to outperform the patch attack in terms of APSR, while for the DDRNet a magnitude of 16 is sufficient. In Fig. 3 (d) and (h), a qualitative result of the patch attack and our approach is shown. Both attacks perturb the prediction in different ways, i.e., the patch attack targets the upper part of the image while our approach focuses on the lower part. This observation is not specific to the shown example but holds more generally for segmentations of the discussed adversarial examples.

5. Related work

The only similar work to our uncertainty-based weighting scheme is the certified radius-guided approach proposed in [26] focusing on the attack of pixels with relatively smaller certified radii. While we use different uncertainty measures that are simple to compute or set values in the loss function to zero, the calculation of the certified radius for each pixel produces high computational overhead which is reflected in the runtimes of the method. In addition to the more expensive calculation and longer runtimes, the attack performance is also worse compared to our method.

The patch attack [24] perturbs only a few pixels of an image, i.e., a rectangular region of fixed size. The expectation over transformation-based method trains the attacker to successfully perturb the image over a range of transformations. In contrast, our approach also perturbs only a small number of pixels (exactly the same number as the patch attack) but uses the I-FGSM procedure with uncertainty-weighted loss and targets only a specific image. Therefore, both methods are not comparable with each other, rather, they demonstrate two different ways to create a high degree of attack damage to the image while perturbing only a few pixels. Figure 3 shows that both attacks focus on different regions, so combining both attacks could be interesting.

			FGSM$_4$	FGSM$_8$	FGSM$_{16}$	I-FGSM$_4$	I-FGSM$_8$	I-FGSM$_{16}$	PGD	ALMA prox
Cityscapes	Deep-Labv3+	orig	0.16	0.16	0.16	0.74	1.50	3.01	60.15	64.75
		ours	0.16	0.16	0.16	0.77	1.53	3.07	61.72	66.80
	BiSe-Net	orig	0.14	0.14	0.14	0.24	0.36	0.56	9.81	16.81
		ours	0.15	0.15	0.15	0.24	0.36	0.56	10.05	17.74
VOC	Deep-Labv3+	orig	0.08	0.08	0.08	0.34	0.69	1.39	27.41	30.76
		ours	0.08	0.08	0.08	0.35	0.71	1.43	28.16	31.55
	PSP-Net	orig	0.08	0.08	0.08	0.32	0.64	1.28	26.21	29.91
		ours	0.09	0.09	0.09	0.33	0.65	1.30	26.96	30.21

Table 1. Runtimes (sec. per frame) for different adversarial attacks with original loss function in comparison to our uncertainty-weighted loss for both datasets and different networks.

	APSR		ΔmIoU	
	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞
orig	59.92	82.50	61.36	67.89
L^0 (ours)	86.54	99.69	70.22	73.46

clean	3.64
orig	5.18
ours	3.70

Table 2. APSR and ΔmIoU results for the PSPNet applied to the VOC dataset comparing the CR framework with our approach (left). Corresponding runtimes in seconds per frame for the underlying attack without weighting scheme, the CR loss function as well as our approach (right).

Potential negative societal impact

Adversarial attacks are generally considered to be malicious, as they can compromise the security of neural networks and rapidly degrade performance. However, the development and especially the free availability are of highest interest to develop detection and defense methods to prevent attacks.

6. Conclusion and outlook

In this work, we proposed an uncertainty-based weighting scheme which can be incorporated into the loss function of any untargeted attack on semantic segmentation models that is composed out of pixel-wise attacks. We exploited the correlation between uncertainty measures and erroneous predictions to strongly degrade the prediction performance of neural networks. The expectations for any attack are low runtimes and computational effort while at the same time having powerful perturbation effects. Our approach can be applied to any attack with minimal computational overhead compared to the original attack, but results in significantly enhanced perturbation performance. We achieved attack pixel success rate values of up to 99.82% across different network architectures and datasets. Moreover, we presented a method which attacks only a subset of the image pixels, similar to the patch attack, but also leading to an erroneous prediction of big parts of the image.

As a further improvement, we plan to develop an uncertainty-weighted loss function for targeted adversarial attacks as our approach is limited to untargeted attacks.

Acknowledgement

This work is supported by the Ministry of Culture and Science of the German state of North Rhine-Westphalia as part of the KI-Starter research funding program and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2092 CASA – 390781972.

References

[1] Shashank Agnihotri and Margret Keuper. Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks, 2023. 3

[2] Anurag Arnab, Ondrej Miksik, and Philip Torr. On the robustness of semantic segmentation models to adversarial attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 4

[3] Andreas Bar, Jonas Lohdefink, Nikhil Kapoor, Serin Varghese, Fabian Huger, Peter Schlicht, and Tim Fingscheidt. The vulnerability of semantic segmentation networks to adversarial attacks in autonomous driving: Enhancing extensive environment sensing. *IEEE Signal Processing Magazine*, 2021. 1, 4

[4] Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, and Nicholas Carlini. Evading adversarial example detection defenses with orthogonal projected gradient descent. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 4

[6] Zhenhua Chen, Chuhua Wang, and David Crandall. Semantically stealthy adversarial attacks against segmentation models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 3

[7] Seungju Cho, Tae Joon Jun, Byungsoo Oh, and Daeyoung Kim. Dapas : Denoising autoencoder to prevent adversar-

ial attack in semantic segmentation. In *International Joint Conference on Neural Network (IJCNN)*, 2020. 1, 4

[8] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4

[10] Mark Everingham, Luc Van Gool, Chris K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 2, 4

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on International Conference on Machine Learning*, 2016. 2

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 2, 3

[13] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[14] Katharina Viktoria Hoebel, Vincent Andrearczyk, Andrew L Beers, Jay B. Patel, Ken Chang, Adrien Depeursinge, Henning Mueller, and Jayashree Kalpathy-Cramer. An exploration of uncertainty information for segmentation quality assessment. In *Medical Imaging: Image Processing*, 2020. 2

[15] Marvin Klingner, Andreas Bär, and Tim Fingscheidt. Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2020. 1, 4

[16] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 3, 4

[17] Kira Maag and Asja Fischer. Uncertainty-based detection of adversarial attacks in semantic segmentation. *ArXiv*, 2023. 1

[18] Kira Maag and Tobias Riedlinger. Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation. *ArXiv*, 2023. 2

[19] Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020. 2

[20] 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, (ICLR)*, 2018. 2, 3

[21] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[23] Krishna Kanth Nakka and Mathieu Salzmann. Indirect local attacks for context-aware semantic segmentation networks. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[24] Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio C. Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2, 3, 5, 7

[25] Huihui Pan, Yuanduo Hong, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 1, 2, 4

[26] Wenjie Qu, Youqi Li, and Binghui Wang. A certified radius-guided attack framework to image segmentation models. *IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 200–220, 2023. 2, 3, 5, 6, 7

[27] Jérôme Rony, Jean-Christophe Pesquet, and Ismail Ben Ayed. Proximal splitting adversarial attacks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5

[28] Kristoffer Wickstrøm, Michael Kampffmeyer, and Robert Jenssen. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60:101619, 2019. 2

[29] 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 *European Conference on Computer Vision (ECCV)*, 2018. 1

[30] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Loddon Yuille. Adversarial examples for semantic segmentation and object detection. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[31] Xiaogang Xu, Hengshuang Zhao, and Jiaya Jia. Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[32] Maksym Yatsura, Kaspar Sakmann, N. Grace Hua, Matthias Hein, and Jan Hendrik Metzen. Certified defences against adversarial patch attacks on semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[33] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision*, 2018. 2, 4

[34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 4