Uncertainty-Based Detection of Adversarial Attacks in Semantic Segmentation

Kira Maag[1] and Asja Fischer[2]

[1]*Technical University of Berlin, Germany*

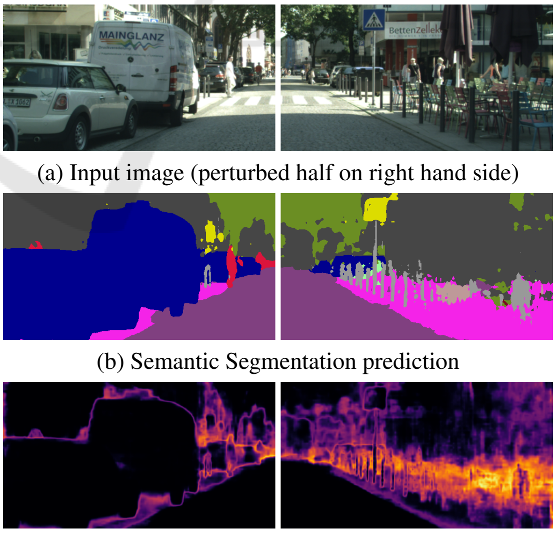
[2]*Ruhr University Bochum, Germany*

Keywords: Deep Learning, Semantic Segmentation, Adversarial Attacks, Detection.

Abstract: State-of-the-Art deep neural networks have proven to be highly powerful in a broad range of tasks, including semantic image segmentation. However, these networks are vulnerable against adversarial attacks, i.e., non-perceptible perturbations added to the input image causing incorrect predictions, which is hazardous in safety-critical applications like automated driving. Adversarial examples and defense strategies are well studied for the image classification task, while there has been limited research in the context of semantic segmentation. First works however show that the segmentation outcome can be severely distorted by adversarial attacks. In this work, we introduce an uncertainty-based approach for the detection of adversarial attacks in semantic segmentation. We observe that uncertainty as for example captured by the entropy of the output distribution behaves differently on clean and perturbed images and leverage this property to distinguish between the two cases. Our method works in a light-weight and post-processing manner, i.e., we do not modify the model or need knowledge of the process used for generating adversarial examples. In a thorough empirical analysis, we demonstrate the ability of our approach to detect perturbed images across multiple types of adversarial attacks.

1 INTRODUCTION

In recent years, deep neural networks (DNNs) have demonstrated outstanding performance and have proven to be highly expressive in a broad range of tasks, including semantic image segmentation (Chen et al., 2018; Pan et al., 2022). Semantic segmentation aims at segmenting objects in an image by assigning each pixel to a fixed and predefined set of semantic classes, providing comprehensive and precise information about the given scene. However, DNNs are vulnerable to *adversarial attacks* (Bar et al., 2021) which is very hazardous in safety-related applications like automated driving. Adversarial attacks are small perturbations added to the input image causing the DNN to perform incorrect predictions at test time. The perturbations are not perceptible to humans, making the detection of these examples very challenging, see for example Figure 1. This undesirable property of DNNs is a major security concern in real world applications. Hence, developing efficient strategies against adversarial attacks is of high importance. Such strategies can either increase the robustness of DNNs making it more difficult to generate adversarial examples (defense) or build on approaches to detect adversarial attacks (detection).



(a) Input image (perturbed half on right hand side)

(b) Semantic Segmentation prediction

(c) Entropy heatmap

Figure 1: Semantic segmentation prediction and entropy heatmap for clean (*left*) and perturbed image (*right*) generated by a dynamic target attack for hiding pedestrians.

Adversarial attacks have attracted much attention, and numerous attacks as well as detection strategies

Maag, K. and Fischer, A.
Uncertainty-Based Detection of Adversarial Attacks in Semantic Segmentation.
DOI: 10.5220/0012303500003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 37-46
ISBN: 978-989-758-679-8; ISSN: 2184-4321
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

have been proposed (Khamaiseh et al., 2022). However, adversarial examples have not been analyzed extensively beyond standard image classification models, often using small datasets such as MNIST (LeCun and Cortes, 2010) or CIFAR10 (Krizhevsky, 2009). The vulnerability of modern DNNs to attacks in more complex tasks like semantic segmentation in the context of real datasets from different domains has been rather poorly explored. The attacks which have been tested so far on semantic segmentation networks can be divided roughly into three categories. The first approaches transfer common attacks from image classification to semantic segmentation, i.e., pixel-wise classification (Agnihotri and Keuper, 2023; Gu et al., 2022; Rony et al., 2022). Other works (Cisse et al., 2017; Xie et al., 2017) have presented attacks specifically designed for the task of semantic segmentation, either attacking the input in a way that leads to the prediction of some predefined and image-unrelated segmentation mask or the complete deletion of one segmentation class (Metzen et al., 2017). Such attacks are more difficult to detect than attacks that perturb each pixel independently. The third group of attacks creates rectangular patches smaller than the input size, which leads to an erroneous prediction of the entire image (Nakka and Salzmann, 2020; Nesti et al., 2022). Defense methods aim to be robust against such attacks, i.e., to achieve high prediction accuracy even on perturbed images. For semantic segmentation tasks, defense approaches are often robust against only one type of attack (Arnab et al., 2018; Klingner et al., 2020; Yatsura et al., 2022). In contrast, detection methods aim at classifying an input as clean or perturbed based on the output of the segmentation model (Xiao et al., 2018).

In this paper, we present an uncertainty-based approach for detecting several kinds of adversarial attacks on semantic image segmentation models. Uncertainty information has already been exploited for the detection of adversarial attacks on classification DNNs, but has not been investigated in the context of segmentation so far. In (Feinman et al., 2017) an approximation to Bayesian inference (Monte-Carlo Dropout) is proposed, which is widely employed to estimate model uncertainty, for the detection of adversarial attacks. The gradient-based approach introduced in (Michel and Ewetz, 2022) generates salient features which are used to train a detector. While both methods require access to the inside of the model, our approach can be applied as a post-processing step using only information of the network output. We construct features per image based on uncertainty information provided by the DNN such as the entropy of the output distribution. In Figure 1 (c), the entropy heatmaps for a clean (left) and a perturbed image (right) are shown, indicating high uncertainties in the attacked regions, motivating the use of uncertainty information to separate clean and perturbed images. On the one hand, these features for clean and perturbed inputs are fed into an one-class support vector machine that performs unsupervised novelty detection (Weerasinghe et al., 2018). On the other hand, we train a logistic regression model with clean and perturbed data for classification. The perturbed data used during training is generated by only one kind of adversarial attack method, while the detector is applied to identify adversarial examples of other methods. Our approach neither modifies the semantic segmentation model nor requires knowledge of the process for generating adversarial examples. We only assume that our post-processing model is kept private while the attacker may have full access to the semantic segmentation model.

In our tests, we employ state-of-the-art semantic segmentation networks (Chen et al., 2018; Pan et al., 2022) applied to the Cityscapes (Cordts et al., 2016) as well as Pascal VOC2012 dataset (Everingham et al., 2012) demonstrating our adversarial attack detection performance. To this end, we consider different types of attackers, from pixel-level attacks designed for image classification (Goodfellow et al., 2015; Kurakin et al., 2017) to pixel-wise attacks developed for semantic segmentation (Metzen et al., 2017) and patch-based ones (Nesti et al., 2022). The source code of our method is publicly available at https://github.com/kmaag/Adversarial-Attack-Detection-Uncertainty. Our contributions are summarized as follows:

- We introduce a new uncertainty-based approach for the detection of adversarial attacks for the semantic image segmentation task. In a thorough empirical analysis, we demonstrate the capability of uncertainty measures to distinguish between clean and perturbed images. Our approach serves as a light-weight post-processing step, i.e., we do not modify the model or need knowledge of the process for generating adversarial examples.
- For the first time, we present a detection method that was not designed for a specific adversarial attack, rather has a high detection capability across multiple types. We achieve averaged detection accuracy values of up to 100% for different network architectures and datasets.

2 RELATED WORK

In this section, we discuss the related works on defense and detection methods for the semantic segmentation task. Defense methods aim to achieve high prediction accuracy even on perturbed images, while detection methods classify the model input as clean or attacked image. A dynamic divide-and-conquer strategy (Xu et al., 2021) and multi-task training (Klingner et al., 2020), which extends supervised semantic segmentation by a self-supervised monocular depth estimation using unlabeled videos, are considered as adversarial training approaches enhancing the robustness of the networks. Another kind of defense strategy is input denoising to remove the perturbation from the input without the necessity to re-train the model. In (Bar et al., 2021) image quilting and the non-local means algorithm are presented as input transformation techniques. To denoise the perturbation and restore the original image, a denoise autoencoder is used in (Cho et al., 2020). The demasked smoothing technique, introduced in (Yatsura et al., 2022), reconstructs masked regions of each image based on the available information with an inpainting model defending against patch attacks. Another possibility to increase the robustness of the model is during inference. In (Arnab et al., 2018) is shown how mean-field inference and multi-scale processing naturally form an adversarial defense. The non-local context encoder proposed in (He et al., 2019) models spatial dependencies and encodes global contexts for strengthening feature activations. From all pyramid features multi-scale information is fused to refine the prediction and create segmentation. The presented works up to now are defense methods improving the robustness of the model. To the best of our knowledge so far only one work focuses on detecting adversarial attacks on segmentation models, i.e., the patch-wise spatial consistency check which is introduced in (Xiao et al., 2018).

The described defense approaches are created for and tested only on a specific type of attack. The problem is that you assume a high model robustness, however, the defense method may perform poorly on new unseen attacks and does not provide a statement about this insecurity. Therefore, we present an uncertainty-based detection approach which shows strong results over several types of adversarial attacks. The presented detection approach (Xiao et al., 2018) is only tested on perturbed images attacked in such a way that a selected image is predicted. The spatial consistency check randomly selects overlapping patches to obtain pixel-wise confidence vectors. In contrast, we use only information of the network output from one inference and not from (computationally expensive) multiple runs of the network. For these reasons, the detection approach introduced in (Xiao et al., 2018) cannot be considered as a suitable baseline.

Post-processing classification models as well as simple output-based methods are used for false positive detection (Maag et al., 2020; Maag and Rottmann, 2023) and out-of-distribution segmentation (Maag et al., 2022; Hendrycks and Gimpel, 2016), but have not been investigated for adversarial examples.

3 ADVERSARIAL ATTACKS

For the generation of adversarial examples, we distinguish between *white* and *black box* attacks. White box attacks are created based on information of the victim model, i.e., the adversarial attacker has access to the full model, including its parameters, and knows the loss function used for training. In contrast, black box attackers have zero knowledge about the victim model. The idea behind these type of attacks is transferability, i.e., an adversarial example generated from another model works well with the victim one. The attacks described in the following belong to the white box setting and were proposed to attack semantic segmentation models.

Attacks on Pixel-Wise Classification. The attacks described in this paragraph were originally developed for image classification and were (in a modified version) applied to semantic segmentation. For semantic segmentation, given an image x a neural network with parameters w provides pixel-wise probability distributions $f(x;w)_{ij}$ over a label space $C = \{y_1, \ldots, y_c\}$ per spatial dimension (i, j). The single-step untargeted *fast gradient sign method* (FGSM, (Goodfellow et al., 2015)) creates adversarial examples by adding perturbations to the pixels of an image x with (one-hot encoded) label y that leads to an increase of the loss L (here cross entropy), that is

$$x_{ij}^{adv} = x_{ij} + \varepsilon \cdot \mathrm{sign}(\nabla_x L_{ij}(f(x;w)_{ij}, y_{ij})) \quad , \qquad (1)$$

where ε is the magnitude of perturbation. The single-step targeted attack with target label y_{ll} instead decreases the loss for the target label and is given by

$$x_{ij}^{adv} = x_{ij} - \varepsilon \cdot \mathrm{sign}(\nabla_x L_{ij}(f(x;w)_{ij}, y_{ij}^{ll})) \quad . \qquad (2)$$

Following the convention, the least likely class predicted by the model is chosen as target class. This attack is extended in (Kurakin et al., 2017) in an iterative manner (I-FGSM) to increase the perturbation strength

$$x^{adv}_{ij,t+1} = \operatorname{clip}_{x,\varepsilon}\big(x^{adv}_{ij,t} + \alpha \cdot \operatorname{sign}(\nabla_{x^{adv}_t} L_{ij}(f(x^{adv}_t; w)_{ij}, y_{ij}))\big) \tag{3}$$

with $x^{adv}_0 = x$, step size α, and a clip function ensuring that $x^{adv}_t \in [x-\varepsilon, x+\varepsilon]$. The targeted case (see eq. (2)) can be formulated analogously. Based on these attacks, further methods for pixel-wise perturbations in the classification context have been proposed such as projected gradient descent (Madry et al., 2018; Bryniarski et al., 2022) and DeepFool (Moosavi-Dezfooli et al., 2016). Some of these approaches have been further developed and adapted to semantic segmentation (Agnihotri and Keuper, 2023; Gu et al., 2022; Rony et al., 2022).

Stationary Segmentation Mask Attacks. Another type of attacks are so called *stationary segmentation mask methods* (SSMM) where the pixels of a whole image are iteratively attacked until most of the pixels have been mis-classified into the target class (Cisse et al., 2017; Xie et al., 2017). For each spatial dimension $(i, j) \in I$, the loss function per image x is given by

$$L(f(x;w), y) = \frac{1}{|I|} \sum_{(i,j)\in I} L_{ij}(f(x;w)_{ij}, y_{ij}) \,. \tag{4}$$

In (Metzen et al., 2017), the universal perturbation is introduced to achieve real-time performance for the attack at test time. To this end, training inputs $D^{\text{train}} = \{x^{(k)}, y^{(k),\text{target}}\}_{k=1}^{m}$ are generated where $y^{(k),\text{target}}$ defines a fixed target segmentation. The universal noise in iteration $t+1$ is computed by

$$\xi_{t+1} = \operatorname{clip}_{\varepsilon}\big(\xi_t - \alpha \cdot \operatorname{sign}(\frac{1}{m} \sum_{k=1}^{m} \nabla_x L(f(x^{(k)} + \xi_t; w), y^{(k),\text{target}})\big) \tag{5}$$

with $\xi_0 = 0$. The loss of pixels which are predicted as belonging to the desired target class with a confidence above a threshold τ are set to 0. At test time, this noise is added to the input image and does not require multiple calculations of the backward pass.

The *dynamic nearest neighbor method* (DNNM) presented in (Metzen et al., 2017) aims to keep the network's segmentation unchanged but to remove a desired target class. Let o be the object class being deleted and $\hat{y}(x)_{ij} = \arg\max_{y\in C} f(x;w)^{y}_{ij}$ the predicted class, where $f(x;w)^{y}_{ij}$ denotes the probability the model assigns for the pixel at position (i, j) to belong to class y, then $I_o = \{(i, j) | \hat{y}(x)_{ij} = o\}$ and $I_{\bar{o}} = I \setminus I_o$. The target label is chosen by $y^{\text{target}}_{ij} = \hat{y}(x)_{i'j'}$ with $\arg\min_{(i',j')\in I_{\bar{o}}} (i'-i)^2 + (j'-j)^2$ for all $(i, j) \in I_o$ and $y^{\text{target}}_{ij} = \hat{y}(x)_{ij}$ for all $(i, j) \in I_{\bar{o}}$. Since the loss function described in eq. (4) weights all pixels equally though both objectives, i.e., hiding a object class and being unobtrusive are not necessarily equally important, a modified version of the loss function with weighting parameter ω is given by

$$L^{\omega}(f(x;w), y) = \frac{1}{|I|}(\omega \sum_{(i,j)\in I_o} L_{ij}(f(x;w)_{ij}, y^{\text{target}}_{ij}) + (1-\omega) \sum_{(i,j)\in I_{\bar{o}}} L_{ij}(f(x;w)_{ij}, y^{\text{target}}_{ij})) \,. \tag{6}$$

Note, the universal perturbation can also be computed for the DNNM.

Patch-Based Attacks. The idea behind *patch attacks* is that perturbing a small region of the image causes prediction errors in a much larger region (Nakka and Salzmann, 2020). In (Nesti et al., 2022), the *expectation over transformation* (EOT)-based patch attack is introduced to create robust adversarial examples, i.e., individual adversarial examples that are at the same time adversarial over a range of transformations. Transformations occurring in the real world are for instance angle and viewpoint changes. These perturbations are modeled within the optimization procedure and an extension of the pixel-wise cross entropy loss is additionally presented in (Nesti et al., 2022) to enable crafting strong patches for the semantic segmentation setting.

4 DETECTION METHOD

Our method does not alter the semantic segmentation model, nor does it require knowledge of the adversarial example generation process. While the attacker may have full access to the semantic segmentation model, we only assume that our post-processing model is kept secret or not attacked. Our approach can be applied to any semantic segmentation network serving as a post-processing step using only information of the network output. In Figure 2 an overview of our approach is given.

The degree of uncertainty in a semantic segmentation prediction is quantified by pixel-wise dispersion measures like the entropy

$$E(x)_{ij} = -\sum_{y\in C} f(x;w)^{y}_{ij} \log f(x;w)^{y}_{ij} \;, \tag{7}$$

the variation ratio

$$V(x)_{ij} = 1 - \max_{y\in C} f(x;w)^{y}_{ij} \;, \tag{8}$$

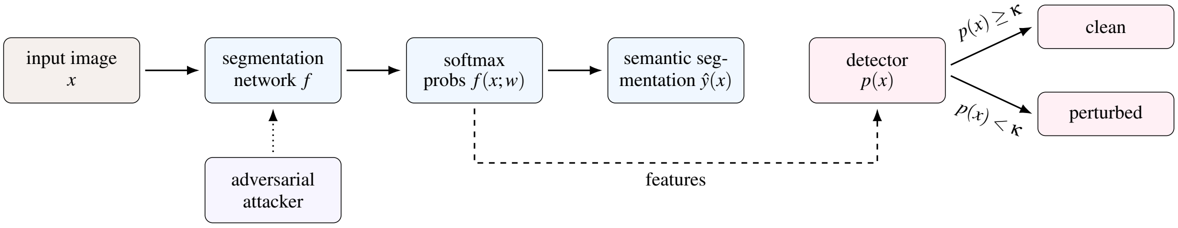


Figure 2: Schematic illustration of our detection method. The adversarial attacker can have full access to the semantic segmentation model. Information from the network output is extracted to construct the features which serve as input to the detector model classifying between clean and perturbed images.

or the probability margin

$$M(x)_{ij} = V(x)_{ij} + \max_{y \in \mathcal{C} \setminus \{\hat{y}(x)_{ij}\}} f(x;w)^y_{ij} \ . \qquad (9)$$

The entropy heatmaps for a clean (left) and a perturbed image (right) are shown in Figure 1 (c) indicating that higher uncertainties occur in the attacked regions which motivates the use of uncertainty information to separate clean and perturbed data. To obtain uncertainty features per image from these pixel-wise dispersion measures, we aggregate them over a whole image by calculating the averages $\bar{D} = 1/|I| \sum_{(i,j) \in I} D(x)_{ij}$ where $D \in \{E,V,M\}$. Moreover, we obtain mean class probabilities for each class $y \in \{1,\dots,C\}$

$$P(y|x) = \frac{1}{|I|} \sum_{(i,j) \in I} f(x;w)^y_{ij} \, . \qquad (10)$$

The concatenation of this $|C|+3$ features forms the feature vectors used in the following. We compute these image-wise features for a set of benign (and adversarially changed) images, which are then used to train classification models providing per image a probability $p(x)$ of being clean (and not perturbed). We classify x as perturbed if $p(x) < \kappa$ and as clean if $p(x) \geq \kappa$, where κ is a predefined detection threshold. We explore different ways to construct such a classifier.

First, we consider two basic outlier detection techniques which only require benign data, i.e., an one-class support vector machine (OCSVM, (Schölkopf et al., 1999)) and an approach for detecting outliers in a Gaussian distributed dataset learning an ellipse (Rousseeuw and Driessen, 1999). Second, we consider the supervised logistic regression (LASSO, (Tibshirani, 1996)) as classification model trained on the features extracted for clean and perturbed images. Importantly, we do not require knowledge of the adversarial example generation process used by the attacker, instead we use attacked data generated by any (other) adversarial attack (cross-attack). While the OCSVM and the ellipse approach are unsupervised and outlier detection is a difficult task, the supervised cross-attack method has the advantage of having already seen other types of perturbed data. Third, we threshold only on the mean entropy $\bar{E}$ (which requires only to choose the threshold value) proposing a very basic uncertainty-based detector. Note, applying our detection method is light-weight, i.e., the feature computation is inexpensive and classification models are trained in advance so that only one inference run is added after semantic segmentation inference.

5 EXPERIMENTS

First, we present the experimental setting and then evaluate our adversarial detection approach.

5.1 Experimental Setting

Datasets. We perform our tests on the Cityscapes (Cordts et al., 2016) dataset for semantic segmentation in street and on the Pascal VOC2012 (Everingham et al., 2012) (shorthand VOC) dataset of visual object classes in realistic scenes. The Cityscapes dataset consists of 2,975 training and 500 validation images of dense urban traffic in 18 and 3 different German towns, respectively. The VOC dataset contains 1,464 training and 1,449 validation images with annotations for the various objects of categories person, animal, vehicle and indoor.

Segmentation Networks. We consider the state-of-the-art DeepLabv3+ network (Chen et al., 2018) with Xception65 backbone (Chollet, 2017). Trained on Cityscapes, we achieve a mean intersection over union (mIoU) value of 78.93 on the validation set and trained on VOC, a validation mIoU value of 88.39. Moreover, we use the BiSeNet (Yu et al., 2018) trained on Cityscapes obtaining a validation mIoU of 70.32. We consider also two real-time models for the Cityscapes dataset, the DDRNet (Pan et al., 2022) achieving 77.8 mIoU and the HRNet (Wang et al., 2021) with 70.5 mIoU.

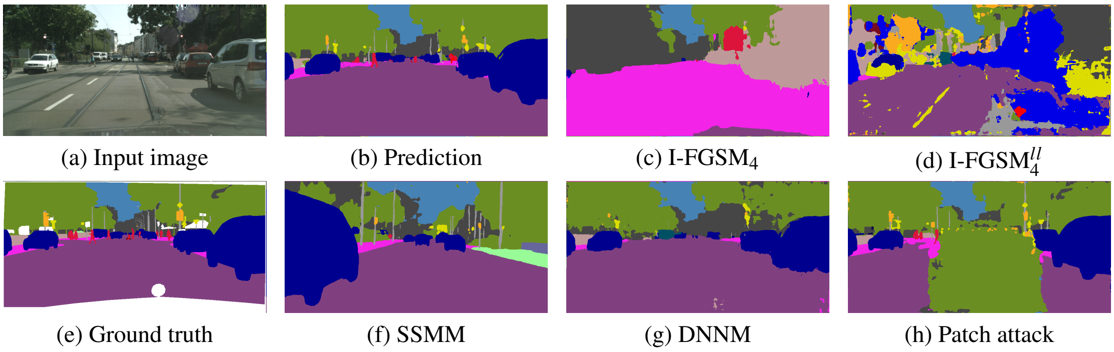


Figure 3: Input image (a) with corresponding ground truth (e). Semantic segmentation prediction for clean (b) and perturbed image generated by an untargeted (c) and a targeted FGSM attack (d) as well as by SSMM (f), DNNM (g) and patch attack (h).

Adversarial Attacks. In many defense methods in semantic segmentation, the adapted FGSM and I-FGSM attack are employed (Arnab et al., 2018; Bar et al., 2021; He et al., 2019; Klingner et al., 2020; Xu et al., 2021). Thus, we study both attacks in our experiments with the parameter setting presented in (Kurakin et al., 2017). The magnitude of perturbation is given by $\varepsilon = \{4, 8, 16\}$, the step size by $\alpha = 1$ and the number of iterations is computed as $n = \min\{\varepsilon + 4, \lfloor 1.25\varepsilon \rfloor\}$. We denote the attack by FGSM$_\varepsilon^\#$ and the iterative one by I-FGSM$_\varepsilon^\#$, $\# \in \{\_, ll\}$, where the superscript discriminates between untargeted and targeted (here ll refers to "least likely"). For the re-implementation of SSMM and DNNM (Metzen et al., 2017), we use the parameters $\varepsilon = 0.1 \cdot 255$, $\alpha = 0.01 \cdot 255$, $n = 60$ and $\tau = 0.75$. For SSMM, the target image is chosen randomly for both datasets and for DNNM, the class person is to be deleted for the Cityscapes dataset. For the VOC dataset, the DNNM attack makes no sense, since on the input images often only one object or several objects of the same class are contained. For our experiments, we use a model zoo[1] where we add the implementations of the adversarial attacks FGSM, I-FGSM, SSMM and DNNM. As we also use the pre-trained networks provided in the repository, we run experiments for the Cityscapes dataset on both models, DeepLabv3+ and HRNet, and for VOC on the DeepLabv3+ network.

For the patch attack introduced in (Nesti et al., 2022), we use the provided code with default parameters and consider two different segmentation models, BiSeNet and DDRNet, applied to the one tested real world dataset (Cityscapes). Since we use the cross-attack procedure (logistic regression) as detection model, i.e., we train the classifier on clean and perturbed data attacked by an attack other than the patch, we use the data obtained from the DeepLabv3+ to train the detector and test on the DDRNet. For the HRNet and the BiSeNet we proceed analogously, since in each case the prediction performance (in terms of mIoU) of both networks is similar.

As the Cityscapes dataset provides high-resolution images (1024×2048 pixels) which require a great amount of memory to run a full backward pass for the computation of adversarial samples, we re-scale image size for this dataset to 512×1024 when evaluating. In Figure 3, a selection of these attacks is shown for the Cityscapes dataset and the DeepLabv3+ network (or DDRNet for the patch attack).

Evaluation Metrics. Our detection models provide per image a probability $p(x)$ of being clean (and not perturbed). The image is then classified as attacked if the probability exceeds a threshold κ for which we tested 40 different values equally spaced in $[0, 1]$. The first evaluation metric we use is the averaged detection accuracy (ADA) which is defined as the proportion of images that are classified correctly. As this metric depends on a threshold κ, we report the optimal ADA score obtained by $\text{ADA}^* = \max_{\kappa \in [0,1]} \text{ADA}(\kappa)$. Secondly, we compute the area under the receiver operating characteristic curve (AUROC) to obtain a threshold independent metric. Lastly, we consider the true positive rate while fixing the false positive rate on clean images to 5% (TPR$_{5\%}$).

5.2 Numerical Results

In the following, we study the attack success performance and evaluate our adversarial attack detection performance.

[1] https://github.com/LikeLy-Journey/SegmenTron

Table 1: APSR results for the semantic segmentation predictions on clean data.

	DeepLabv3+	DDRNet	HRNet	BiSeNet
Cityscapes	6.84	4.00	5.48	5.26
VOC	2.92	-	-	-

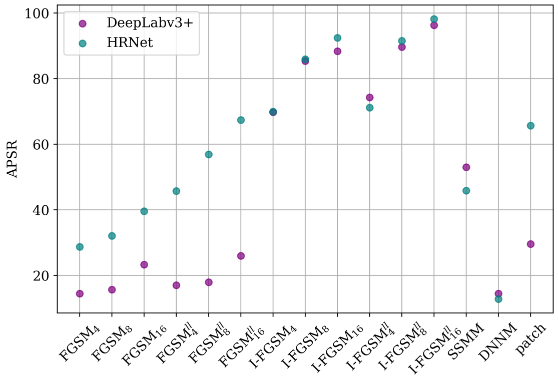


Figure 4: APSR results for the Cityscapes dataset and both networks perturbed by different attacks.

Attack Performance. In order to access the performance, i.e., the strength, of the attack generating methods, we consider the attack pixel success rate (APSR) (Rony et al., 2022) defined by

$$\mathrm{APSR} = \frac{1}{|I|} \sum_{(i,j)\in I} \operatorname*{arg\,max}_{y\in C} f(x^{adv};w)^{y}_{ij} \neq y^{GT}_{ij} \quad , \quad (11)$$

where y^{GT}_{ij} denotes the true class of the pixel at location (i,j). Note, this metric is the opposite of the accuracy, as it focuses on falsely (and not correct) predicted pixels. If we replace x^{adv} in eq. (11) by the input image x, we obtain a measure how well the semantic segmentation model performs on clean data. The values for this measure for the different networks and both datasets are given in Table 1. As expected, we observe small APSR scores: all values are below 6.84%.

The APSR results for various attacks on the Cityscapes dataset are shown in Figure 4. For all variations of the FGSM attack (untargeted vs. targeted, non-iterative vs. iterative) the APSR increases with larger magnitude of perturbation strength. Moreover, targeted attacks lead to larger ASPR values than their untargeted counterpart. The I-FGSM outperforms the FGSM due to the iterative procedure of individual perturbation steps. For the SSMM attack, the target is a randomly chosen image from the dataset. The examples shown in Figure 3 (a) and (f) indicate, that the correct and target classes of the clean and the perturbed image coincide in several areas, such as the street or the sky reflecting the nature of street scenes. This may explain the relatively low ASPR values around 50%. For the DNNM attack, the APSR scores are comparatively small since most parts of the images are not perturbed but only one class is to be deleted. We observe that the performance of the patch attack has more or less impact depending on the model. Comparing the two models, we find that DeepLabv3+ is more robust against adversarial attacks, as the APSR values are mostly smaller than those of the HRNet network.

Table 2: APSR results for the VOC dataset and the DeepLabv3+ network perturbed by different attacks.

FGSM$_4$	FGSM$^{ll}_4$	I-FGSM$_4$	I-FGSM$^{ll}_4$	SSMM
17.81	19.68	56.73	64.30	62.99
FGSM$_8$	FGSM$^{ll}_8$	I-FGSM$_8$	I-FGSM$^{ll}_8$	
17.93	19.66	74.94	82.51	
FGSM$_{16}$	FGSM$^{ll}_{16}$	I-FGSM$_{16}$	I-FGSM$^{ll}_{16}$	
16.67	17.65	80.98	91.68	

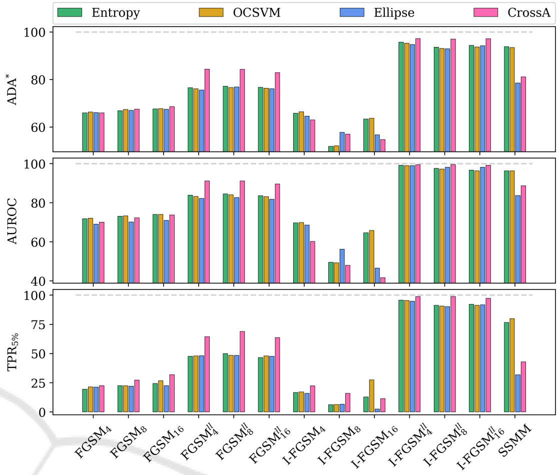


Figure 5: Detection performance results for the VOC dataset and the DeepLabv3+ network.

The results for the VOC dataset given in Table 2 are qualitatively similar to the findings for the Cityscapes dataset. However, the outcome for the (targeted as well as untargeted) non-iterative FGSM attack differs, i.e., the APSR scores are not increasing with higher noise but stay at similar values. In summary, for both datasets and the different network architectures, most attacks achieve high APSR values and greatly alter the prediction. Thus, the detection of such attacks is extremely important.

Evaluation of Our Detection Method. The defense approaches described above are created for and tested only on a specific type of attacks. The sole presented detection approach (Xiao et al., 2018) is only tested on stationary segmentation mask methods and is computationally expensive due to the requirement of multiple runs of the network. For this reason, neither this detection approach nor the defense methods

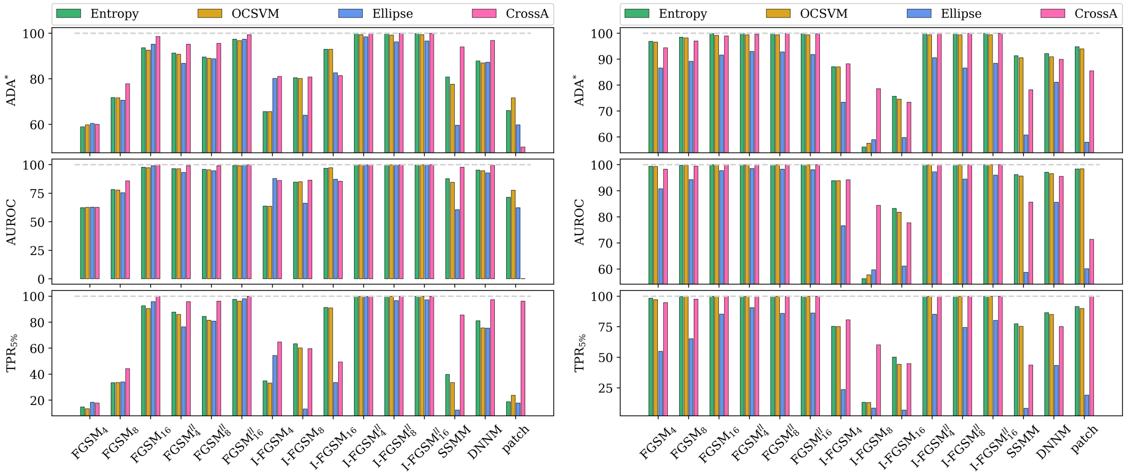


Figure 6: Detection performance results for DeepLabv3+ (*left*) and the HRNet (*right*) trained on the Cityscapes dataset.

can be considered as suitable baselines.

In the following, we denote the single feature, mean entropy based classification by *Entropy*, the ordinary one-class support vector machine by *OCSVM*, the outlier detection method proposed in (Rousseeuw and Driessen, 1999) by *Ellipse* and the logistic regression model by *CrossA*. For training the regression model, we chose data resulting from the iterative targeted FGSM attack with a magnitude of noise of 2 as perturbed data (assuming that it might be advantageous to have malicious data stemming from an attack method with small perturbation strength, which makes the attack harder to detect). The resulting classifier is then evaluated against all other attacks. Note, the training of the detection models is light-weight and no knowledge of the process for generating adversarial examples is needed. For evaluating our detection models, we use cross-validation with 5 runs.

The detection results for the VOC dataset are shown in Figure 5 and for the Cityscapes dataset in Figure 6. We observe comparatively lower detection performance results for smaller perturbation magnitudes for the untargeted FGSM attacks which may be explained by the fact that weaker attacks lead to a change of prediction for a lower number of pixels and are thus more difficult to detect. Targeted attacks are better detected than untargeted attacks (with ADA* values over 80% for all models). This could be due to the procedure of picking the most unlikely class as target which results in larger changes of the uncertainty measures used as features. The detectors perform not that well on the adversarial examples resulting from the untargeted I-FGSM despite the strength of the attack. An inspection of these examples shows that during segmentation only a few classes are predicted (see Figure 3 (c)) with often low uncertainty for large connected components which complicates the distinction between clean and perturbed data. Interesting are the high detection results of up to 96.89% ADA* values (obtained by CrossA for the Cityscapes dataset and the DeepLabv3+ network) for the DNNM attack as the perturbation targets only a few pixels, (APSR values around 15%) and is therefore difficult to detect. For the patch attack, it is noticeable that the detection performance for the DeepLabv3+ network on the Cityscapes dataset is low compared to the HRNet which is explained by the higher disturbance power of this attack on the HRNet, reflected in 36.11 percentage points higher APSR values. In general, the detection capability for attacks on the HRNet is stronger than for the DeepLabv3+, since the HRNet is more easily to attack (see higher APSR values in Figure 4). Generally, our experiments highlight the high potential of investigating uncertainty information for successfully detecting adversarial segmentation attacks. Already the basic method *Entropy* leads to high accuracies often outperforming OCSVM and Ellipse. However, across different attacks and datasets, CrossA achieves ADA* values of up to 100%. Thus, our light-weight and uncertainty-based detection approach should be considered as baseline for future methods.

6 CONCLUSION

In this work, we introduced a new uncertainty-based approach for the detection of adversarial attacks on semantic image segmentation tasks. We observed that uncertainty information as given by the entropy behaves differently on clean and perturbed images and used this property to distinguish between the two cases with very basic classification models. Our approach works in a light-weight and post-processing manner, i.e., we do not modify the model nor need knowledge of the process used by the attacker for generating adversarial examples. We achieve averaged detection accuracy values of up to 100% for different network architectures and datasets. Moreover, it has to be pointed out, that our proposed detection approach is the first that was not designed for a specific adversarial attack, but has a high detection capability across multiple types. Given the high detection accuracy and the simplicity of the proposed approach, we are convinced, that it should serve as simple baseline for more elaborated but computationally more expensive approaches developed in future.

ACKNOWLEDGEMENTS

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

Agnihotri, S. and Keuper, M. (2023). Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks. 2, 4

Arnab, A., Miksik, O., and Torr, P. (2018). On the robustness of semantic segmentation models to adversarial attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3, 6

Bar, A., Lohdefink, J., Kapoor, N., Varghese, S., Huger, F., Schlicht, P., and Fingscheidt, T. (2021). The vulnerability of semantic segmentation networks to adversarial attacks in autonomous driving: Enhancing extensive environment sensing. *IEEE Signal Processing Magazine*. 1, 3, 6

Bryniarski, O., Hingun, N., Pachuca, P., Wang, V., and Carlini, N. (2022). Evading adversarial example detection defenses with orthogonal projected gradient descent. In *International Conference on Learning Representations (ICLR)*. 4

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*. 1, 2, 5

Cho, S., Jun, T. J., Oh, B., and Kim, D. (2020). Dapas : Denoising autoencoder to prevent adversarial attack in semantic segmentation. In *International Joint Conference on Neural Network (IJCNN)*. 3

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5

Cisse, M., Adi, Y., Neverova, N., and Keshet, J. (2017). Houdini: Fooling deep structured prediction models. In *Conference on Neural Information Processing Systems (NeurIPS)*. 2, 4

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 5

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2, 5

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. (2017). Detecting adversarial samples from artifacts. 2

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y., editors, *International Conference on Learning Representations (ICLR)*. 2, 3

Gu, J., Zhao, H., Tresp, V., and Torr, P. (2022). Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision (ECCV)*. 2, 4

He, X., Yang, S., Li, G., Li, H., Chang, H., and Yu, Y. (2019). Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In *AAAI Conference on Artificial Intelligence*. 3, 6

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. 3

Khamaiseh, S. Y., Bagagem, D., Al-Alaj, A., Mancino, M., and Alomari, H. W. (2022). Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification. *IEEE Access*. 2

Klingner, M., Bär, A., and Fingscheidt, T. (2020). 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)*. 2, 3, 6

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. 2

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2017). Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*. 2, 3, 6

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database. 2

Maag, K., Chan, R., Uhlemeyer, S., Kowol, K., and Gottschalk, H. (2022). Two video data sets for tracking and retrieval of out of distribution objects. In *Asian Conference on Computer Vision (ACCV)*, pages 3776–3794. 3

Maag, K. and Rottmann, M. (2023). False negative reduction in semantic segmentation under domain shift using depth estimation. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications. 3

Maag, K., Rottmann, M., and Gottschalk, H. (2020). Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. 3

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations, (ICLR)*. 4

Metzen, J. H., Kumar, M. C., Brox, T., and Fischer, V. (2017). Universal adversarial perturbations against semantic image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*. 2, 4, 6

Michel, A. and Ewetz, R. (2022). Gradient-based adversarial attack detection via deep feature extraction. In *SoutheastCon*. 2

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4

Nakka, K. K. and Salzmann, M. (2020). Indirect local attacks for context-aware semantic segmentation networks. In *European Conference on Computer Vision (ECCV)*. 2, 4

Nesti, F., Rossolini, G., Nair, S., Biondi, A., and Buttazzo, G. C. (2022). 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)*. 2, 4, 6

Pan, H., Hong, Y., Sun, W., and Jia, Y. (2022). Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*. 1, 2, 5

Rony, J., Pesquet, J.-C., and Ayed, I. B. (2022). Proximal splitting adversarial attacks for semantic segmentation. 2, 4, 7

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. 5, 8

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., and Platt, J. (1999). Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*. MIT Press. 5

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*. 5

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B. (2021). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 5

Weerasinghe, P. S., Erfani, S. M., Alpcan, T., Leckie, C., and Kuijper, M. (2018). Unsupervised adversarial anomaly detection using one-class support vector machines. *International Symposium on Mathematical Theory of Networks and Systems*. 2

Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., and Song, D. (2018). Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *European Conference on Computer Vision (ECCV)*. 2, 3, 7

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., and Yuille, A. L. (2017). Adversarial examples for semantic segmentation and object detection. *IEEE International Conference on Computer Vision (ICCV)*. 2, 4

Xu, X., Zhao, H., and Jia, J. (2021). Dynamic divide-and-conquer adversarial training for robust semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*. 3, 6

Yatsura, M., Sakmann, K., Hua, N. G., Hein, M., and Metzen, J. H. (2022). Certified defences against adversarial patch attacks on semantic segmentation. 2, 3

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision*. 5