Characterizing Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation

Chaowei Xiao[1], Ruizhi Deng[2], Bo Li[3], Fisher Yu[3], Jinfeng Yi[4], Mingyan Liu[1], and Dawn Song[3]

[1]University of Michigan [2]Simon Fraser University [3]UC Berkeley
[4]JD AI Research

Abstract. Deep Neural Networks (DNNs) have been widely applied in various recognition tasks. However, recently DNNs have been shown to be vulnerable against adversarial examples, which can mislead DNNs to make arbitrary incorrect predictions. While adversarial examples are well studied in classification tasks, other learning problems may have different properties. For instance, semantic segmentation requires additional components such as dilated convolutions and multiscale processing. In this paper, we aim to characterize adversarial examples based on spatial context information in semantic segmentation. We observe that spatial consistency information can be potentially leveraged to detect adversarial examples robustly even when a strong adaptive attacker has access to the model and detection strategies. We also show that adversarial examples based on attacks considered within the paper barely transfer among models, even though transferability is common in classification. Our observations shed new light on developing adversarial attacks and defenses to better understand the vulnerabilities of DNNs.

Keywords: Semantic segmentation, adversarial example, spatial consistency

1 Introduction

Deep Neural Networks (DNNs) have been shown to be highly expressive and have achieved state-of-the-art performance on a wide range of tasks, such as speech recognition [20], image classification [24], natural language understanding [54], and robotics [32]. However, recent studies have found that DNNs are vulnerable to *adversarial examples* [38,17,31,47,45,40,9,8,7]. Such examples are intentionally perturbed inputs with small magnitude adversarial perturbation added, which can induce the network to make arbitrary incorrect predictions at test time, even when the examples are generated against different models [27,5,33,46]. The fact that the adversarial perturbation required to fool a model is often small and (in the case of images) imperceptible to human observers makes detecting such examples very challenging. This undesirable property of deep networks has

become a major security concern in real-world applications of DNNs, such as self-driving cars and identity recognition systems [16,37]. Furthermore, both white-box and black-box attacks have been performed against DNNs successfully when an attacker is given full or zero knowledge about the target systems [2,17,45]. Among black-box attacks, transferability is widely used for generating attacks against real-world systems which do not allow white-box access. Transferability refers to the property of adversarial examples in classification tasks where one adversarial example generated against a local model can mislead another unseen model without any modification [33].

Given these intriguing properties of adversarial examples, various analyses for understanding adversarial examples have been proposed [29,30,43,42], and potential defense/detection techniques have also been discussed mainly for the image classification problem [13,21,30]. For instance, image pre-processing [14], adding another type of random noise to the inputs [48], and adversarial retraining [17] have been proposed for defending/detecting adversarial examples when classifying images. However, researchers [4,19] have shown that these defense or detection methods are easily attacked again by attackers with or even without knowledge of the defender's strategy. Such observations bring up concerns about safety problems within diverse machine learning based systems.

In order to better understand adversarial examples against different tasks, in this paper we aim to analyze adversarial examples in the semantic segmentation task instead of classification. We hypothesize that adversarial examples in different tasks may contain unique properties that provide in-depth understanding for such examples and encourage potential defensive mechanisms. Different from image classification, in semantic segmentation, each pixel will be given a prediction label which is based on its surrounding information [12]. Such spatial context information plays a more important role for segmentation algorithms, such as [50,55,26,23]. Whether adversarial perturbation would break such spatial context is unknown to the community. In this paper we propose and conduct image spatial consistency analysis, which randomly selects overlapping patches from a given image and checks how consistent the segmentation results are for the overlapping regions. Our pipeline of spatial consistency analysis for adversarial/benign instances is shown in Figure 1. We find that in segmentation task, adversarial perturbation can be weakened for separately selected patches, and therefore adversarial and benign images will show very different behaviors in terms of the spatial consistency information. Moreover, since such spatial consistency is highly random, it is hard for adversaries to take such constraints into account when performing adaptive attacks. This renders the system less brittle even facing the sophisticated adversaries, who have full knowledge about the model as well as the detection/defense method applied..

We use image scale transformation to perform detection of adversarial examples as a baseline, which has been used for detection in classification tasks [39]. We show that by randomly scaling the images, adversarial perturbation can be destroyed and therefore adversarial examples can be detected. However, when the attacker knows the detection strategy (adaptive attacker), even without the

exact knowledge about the scaling rate, attacker can still perform adaptive attacks against the detection mechanism, which is similar with the findings in classification tasks [4]. On the other hand, we show that by incorporating spatial consistency check, existing semantic segmentation networks can detect adversarial examples (average AUC 100%), which are generated by the state-of-the-art attacks considered in this paper, regardless of whether the adversary knows the detection method. Here, we allow the adversaries to have full access to the model and any detection method applied to analyze the robustness of the model against adaptive attacks. We additionally analyze the defense in a black-box setting, which is more practical in real-world systems.

In this paper, our goal is to further understand adversarial attacks by conducting spatial consistency analysis in the semantic segmentation task, and we make the following contributions:

1. We propose the spatial consistency analysis for benign/adversarial images and conduct large scale experiments on two state-of-the-art attack strategies against both DRN and DLA segmentation models with diverse adversarial targets on different dataset, including Cityscapes and real-world autonomous driving video dataset.
2. We are the first to analyze spatial information for adversarial examples in segmentation models. We show that spatial consistency information can be potentially leveraged to distinguish adversarial examples. We also show that spatial consistency check mechanism induce a high degree of randomness and therefore is robust against adaptive adversaries. We evaluate image scaling and spatial consistency, and show that spatial consistency outperform standard scaling based method.
3. In addition, we empirically show that adversarial examples generated by the attack methods considered in our studies barely transfer among models, even when these models are of the same architecture with different initialization, different from the transferability phenomena in classification tasks.

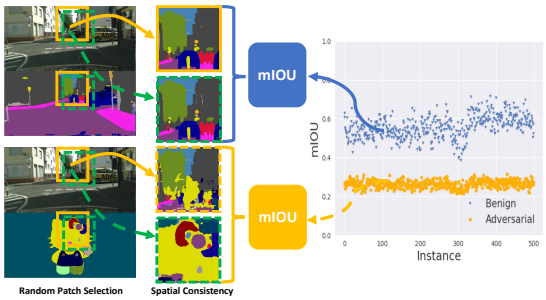


Fig. 1: Spatial consistency analysis for adversarial and benign instances in semantic segmentation.

2 Related work

Semantic Segmentation has received long lasting attention in the computer vision community [25]. Recent advances in deep learning [24] also show that deep convolutional networks can achieve much better results than traditional methods [28]. Yu et al. [50] proposed using dilated convolutions to build high-resolution feature maps for semantic segmentation. They can improve the performance significantly compared to upsampling approaches [28,34,1]. Most of the recent state-of-the-art approaches are based on dilated convolutions [51,55,44] and residual networks [18]. Therefore, in this work, we choose dilated residual networks (DRN) [51] and deep layer aggregation (DLA) [52] as our target models for attacking and defense.

Adversarial Examples for Semantic Segmentation have been studied recently in addition to adversarial examples in image classification. Xie et al. proposed a gradient based algorithm to attack pixels within the whole image iteratively until most of the pixels have been misclassified into the target class [49], which is called dense adversary generation (DAG). Later an optimization based attack algorithm has been studied by introducing a surrogate loss function called Houdini in the objective function [10]. The Houdini loss function is made up of two parts. The first part represents the stochastic margin between the score of actual and predicted targets, which reflects the confidence of model prediction. The second part is the task loss, which is independent with the model and corresponds to the actual task. The task loss enables Houdini algorithm to generate adversarial examples in different tasks, including image segmentation, human pose estimation, and speech recognition.

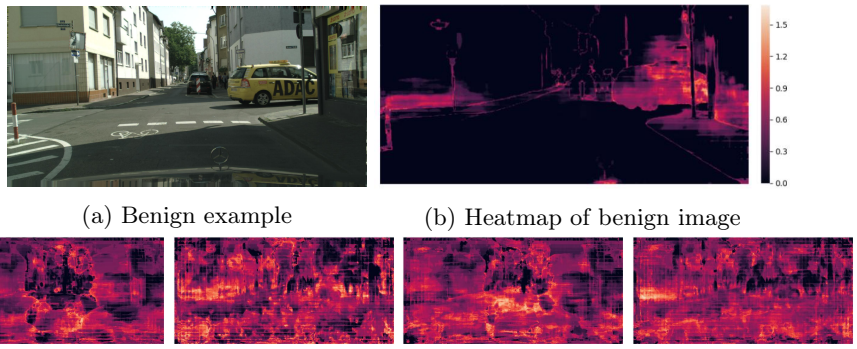
Various detection and defense methods have also been studied against adversarial examples in image classification. For instance, adversarial training [17] and its variations [41,30] have been proposed and demonstrated to be effective in classification task, which is hard to adapt for the segmentation task. Currently no defense or detection methods have been studied in image segmentation.

(a) Cityscapes (b) BDD

Fig. 2: Samples of benign and adversarial examples generated by Houdini on Cityscapes [11] (targeting on Kitty/Pure) and BDD100K [53] (targeting on Kitty/Scene). We select DRN as our target model here. Within each subfigure, the first column shows benign images and corresponding segmentation results, and the second and third columns show adversarial examples with different adversarial targets.

3 Spatial Consistency Based Method

In this section, we will explore the effects that spatial context information has on benign and adversarial examples in segmentation models. We conduct different experiments based on various models and datasets, and due to the space limitation, we will use a small set of examples to demonstrate our discoveries and relegate other examples to the supplementary materials. Figure 2 shows the benign and adversarial examples targeting diverse adversarial targets: "Hello Kitty" (Kitty) and random pure color (Pure) on Cityscapes; and "Hello Kitty" (Kitty) and a real scene without any cars (Scene) on BDD video dataset, respectively. In the rest of the paper, we will use the format "attack method | target" to label each adversarial example. Here we consider both DAG [49] and Houdini [10] attack methods.



(a) Benign example (b) Heatmap of benign image

(c) DAG | Kitty (d) DAG | Pure (e) Houdini | Kitty (f) Houdini | Pure

Fig. 3: Heatmap of per-pixel self-entropy on Cityscapes dataset against DRN model. (a) and (b) show a benign image and its corresponding per-pixel self-entropy heatmap. (c)-(f) show the heatmaps of the adversarial examples generated by DAG and Houdini attacks targeting "Hello Kitty" (Kitty) and random pure color (Pure).

3.1 Spatial Context Analysis

To quantitatively analyze the contribution of spatial context information to the segmentation task, we first evaluate the entropy of prediction based on different spatial context. For each pixel m within an image, we randomly select K patches $\{P_1, P_2, ..., P_K\}$ which contain m. Afterwards, within each patch P_i, the pixel m will be assigned with a confidence vector based on Softmax prediction, so pixel m will correspond to K vectors in total. We discretize each vector to a one-hot vector and sum up these K one-hot vectors to obtain vector $\mathcal{V}_m$. Each component $\mathcal{V}_m[j]$ of the vector represents the number of times pixel m is predicted to be

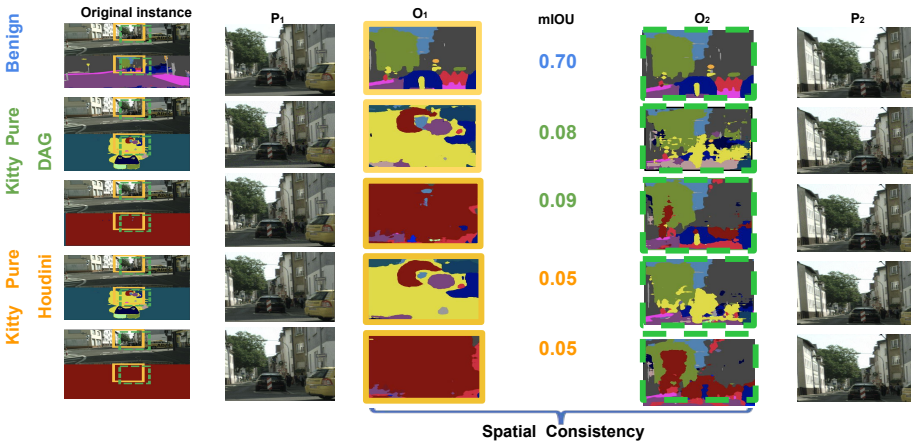


Fig. 4: Examples of spatial consistency based method on adversarial examples generated by DAG and Houdini attacks targeting on Kitty and Pure. First column shows the original image and corresponding segmentation results. Column P_1 and P_2 show two randomly selected patches, while column O_1 and O_2 represent the segmentation results of the overlapping regions from these two patches, respectively. The mIOU between O_1 and O_2 are reported. It is clear that the segmentation results of the overlapping regions from two random patches are very different for adversarial images (low mIOU), but relatively consistent for benign instance (high mIOU).

class j. We then normalize $\mathcal{V}_m$ by dividing K. Finally, for each pixel m, we calculate its self-entropy

$$\mathcal{H}(m) = -\sum_j \mathcal{V}_m[j] \log \mathcal{V}_m[j]$$

and therefore calculate the self entropy for each vector. We utilize such entropy information of each pixel to convey the consistency of different surrounding patches and plot this information in the heatmaps in Figure 3. It is clear that for benign instances, the boundaries of original objects have higher entropy, indicating that these are places harder to predict and can gain more information by considering different surrounding spatial context information.

3.2 Patch Based Spatial Consistency

The fact that surrounding spatial context information shows different spatial consistency behaviors for benign and adversarial examples motivates us to perform the spatial consistency check hoping to potentially tell these two data distributions apart.

First, we introduce how to generate overlapping spatial contexts by selecting random patches and then validate the spatial consistency information. Let

s be the patch size and w, h be the width and height of an image $\mathbf{X}$. We define the first and second patch based on the coordinates of their top-left and bottom-right vertices $(u_1, u_2, u_3, u_4), (v_1, v_2, v_3, v_4)$, where Let $(d_{u_1,v_1}, d_{u_2,v_2})$ be displacement between the top-left coordinate of the first and second patch: $d_{u_1,v_1} = v_1 - u_1, d_{u_2,v_2} = v_2 - u_2$. To guarantee that there is enough overlap, we require (d_{u_1,v_1} and d_{u_2,v_2}) to be in the range $(b_{\mathbf{low}}, b_{\mathbf{upper}})$. Here we randomly select the two patches, aiming to capture diverse enough surrounding spatial context, including information both near and far from the target pixel. The **patch selection algorithm (getOverlapPatches)** is shown in supplementary materials.

Next we show how to apply the spatial consistency based method to a given input and therefore recognize adversarial examples. The detailed algorithm is shown in Algorithm 1. Here K denotes the number of overlapping regions for which we will check the spatial consistency. We use the mean Intersection Over Union (mIOU) between the overlapping regions O_1, O_2 from two patches P_1, P_2 to measure their spatial consistency. The mIOU is defined as $\frac{1}{n_{cls}} \sum_i n_{ii} / (\sum_j n_{ij} + \sum_j n_{ji} - n_{ii})$, where n_{ij} denotes the number of pixels predicted to be class i in O_1 and class j in O_2, and n_{cls} is the number of the unique classes appearing in both O_1 and O_2. **getmIOU** is a function that computes the mIOU given patches P_1, P_2 along with their overlapping regions O_1 and O_2 shown in supplementary materials.

Algorithm 1: Spatial Consistency Check Algorithm

input: Input image $\mathbf{X}$;
number of overlapping regions K;
patch size s;
segmentation model f;
bound $b_{\mathbf{low}}, b_{\mathbf{upper}}$;
output: Spatial consistency threshold c;

Initialization : **cs** $\leftarrow$[], $w \leftarrow x.width, h \leftarrow x.height$;
for $k \leftarrow 0$ **to** K **do**
  $(u_1, u_2, u_3, u_4), (v_1, v_2, v_3, v_4) \leftarrow$ **getOverlapPatches**$(s, w, h, b_{\mathbf{low}}, b_{\mathbf{upper}})$;
  $P_1 = X[u1 : u3, u2 : u4], P_2 = X[v_1 : v_3, v_3 : v_4]$;
  /* get prediction result of two random patches from f */;
  $pred^1 \leftarrow \text{argmax}_c\, f_c(P_1), pred^2 \leftarrow \text{argmax}_c\, f_c(P_2)$;
  /* get prediction of the overlap area between two patches */;
  $p_1 \leftarrow \{pred^1_{i,j} | \forall (i,j) \in pred^1, i > v_1 - u_1, j > v_2 - u_2\}$;
  $p_2 \leftarrow \{pred^2_{i,j} | \forall (i,j) \in pred^2, i < s - (v_1 - u_1), j < s - (v_2 - u_2)\}$;
  /* get consistency value (mIOU) from two patches */;
  cs $\overset{+}{\leftarrow}$ **getmIOU**$(p1, p2)$;
end
$c \leftarrow$ **Mean**(**cs**);
Return: c

4 Scale Consistency Analysis

We have discussed how spatial consistency can be utilized to potentially characterize adversarial examples in segmentation task. In this section, we will discuss another baseline method: image scale transformation, which is another natural factor considered in semantic segmentation [22,28]. Here we focus on image blur operation by applying Gaussian blur to given images [6], which is studied for detecting adversarial examples in image classification [39]. Similarly, we will analyze the effects of image scaling on benign/adversarial samples. Since spatial context information is important for segmentation task, scaling or performing segmentation on small patches may damage the global information and therefore affect the final prediction. Here we aim to provide quantitative results to understand and explore how image scale transformation would affect adversarial perturbation.

4.1 Scale Consistency Property

Scale theory is commonly applied in image segmentation task [35], and therefore we train scale resilient models to obtain robust ones, which we perform attacks against. On these scale resilient models, we first analyze how image scaling affect segmentation results for benign/adversarial samples. We applied the DAG [49] and Houdili [10] attacks against the DRN and DLA models with different adversarial targets. The images and corresponding segmentation results before and after scaling are shown in Figure 5. We apply Gaussian kernel with different standard deviations (std) to scale both benign and adversarial instances. It is clear that when we apply Gaussian blurring with higher std (3 and 5), adversarial perturbation is harmed and the segmentation results are not longer adversarial targets for scale transformed adversarial examples as shown in Figure 5 (a)-(e).

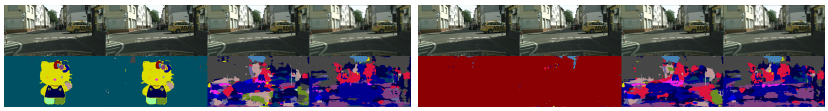
5 Experimental Results

In this section, we conduct comprehensive large scale experiments to evaluate the image spatial and scale consistency information for benign and adversarial examples generated by different attack methods. We will also show that the spatial consistency based detection method is robust against sophisticated adversaries with knowledge about defenders, while scale transformation method is not.

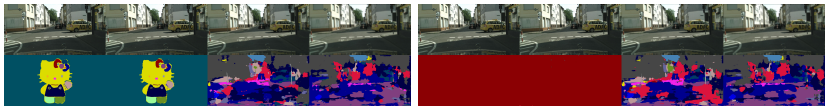
5.1 Implementation Details

Datasets. We apply both Cityscapes [11] and BDD100K [53] in our evaluation. We show results on the validation set of both datasets, which contains 500 high resolution images with a combined 19 categories of segmentation labels. These two datasets are both outdoor datasets containing instance-level annotations, which would raise real-wold safety concerns if they were attacked. Comparing with other datasets such as Pascal VOC [15] and CamVid [3], these two dataset are more challenging due to the relatively high resolution and diverse scenes within each image.

(a) Benign example



(b) DAG | Kitty (c) DAG | Pure



(d) Houdini | Kitty (e) Houdini | Pure

Fig. 5: Examples of images and corresponding segmentation results before/after image scaling on Cityscapes against DRN model. For each subfigure, the first column shows benign/adversarial image, while the later columns represent images after scaling by applying Gaussian kernel with std as 0.5, 3, and 5, respectively. (a) shows benign images before/after image scaling and the corresponding segmentation results; (b)-(e) present similar results for adversarial images generated by DAG and Houdini attacks targeting on Kitty and Pure.

Semantic Segmentation Models. We apply Dilated residual networks (DRN) [51] and Deep Layer Aggregation (DLA) [52] as our target models. More specifically, we select DRN-D-22 and DLA-34. For both models, we use 512 crop size and 2 random scale during training to obtain scale resilient models for both the BDD and Cityscapes datasets. The mIOU of these two models on pristine training data are shown in Table 1. More result on different models can be found in supplementary materials.

Adversarial Examples We generate adversarial examples based on two state-of-the-art attack methods: DAG [49] and Houdini [10] using our own implementation of the methods. We select a complex image, Hello Kitty (Kitty), with different background colors and a random pure color (Pure) as our targets on Cityscapes dataset. Furthermore, in order to increase the diversity, we also select a real-world driving scene (Scene) without any cars from the BDD training dataset as another malicious target on BDD. Such attacks potentially show that every image taken in the real world can be attacked to the same scene without any car showing on the road, which raises great security concerns for future autonomous driving systems. Furthermore, we also add three additional adversarial

targets, including "ECCV 2018", "Remapping", and "Color strip" in supplementary materials to increase the diversity of adversarial targets.

We generate 500 adversarial examples for Cityscapes and BDD100K datasets against both DRN and DLA segmentation models targeting on various malicious targets (More results can be found in supplementary materials).

5.2 Spatial Consistency Analysis

To evaluate the spatial consistency analysis quantitatively for segmentation task, we leverage it to build up a simple detector to demonstrate its property. Here we perform patch based spatial consistency analysis, and we select patch size and region bound as $s = 512$, $b_{low} = 32, b_{upper} = 64$. We select the number of overlapping regions as $K \in \{1, 5, 10, 50\}$. Here we first select some benign instances, and calculate the normalize mIOU of overlapping regions from two random patches. We record the lower bound of theses mIOU as the threshold of the detection method. Note that when reporting detection rate in the rest of the paper, we will use the threshold learned from a set of benign training data; while we also report Area Under Curve (AUC) of Receiver Operating Characteristic Curve (ROC) curve of a detection method to evaluate its overall performance. Therefore, given an image, for each overlapping region of two random patches, we will calculate the normalize mIOU and compare with the threshold calculated before. If it is larger, the image is recognized as benign; vice versa. This process is illustrated in Algorithm 1. We report the detection results in terms of AUC in Table 1 for adversarial examples generated in various settings as mentioned above. We observed that such simple detection method based on spatial consistency information can achieve AUC as nearly 100% for adversarial examples that we studied here. In addition, we also select s with a random number between 384 to 512 (too small patch size will affect the segmentation accuracy even on benign instances, so we tend not to choose small patches on the purpose of control variable) and show the result in supplementary materials. We observe that random patch sizes achieve similar detection result.

5.3 Image Scale Analysis

As a baseline, we also utilize image scale information to perform as a simple detection method and compare it with the spatial consistency based method. We apply Gaussian kernel to perform the image scaling based detection, and select $\mathbf{std}_{detect} \in \{0.5, 3, 5\}$ as the standard deviation of Gaussian kernel. We compute the normalize mIOU between the original and scalled images. Similarly, the detection results of corresponding AUC are shown in Table 1. It is demonstrated that detection method based on image scale information can achieve similarly high AUC compared with spatial consistency based method.

5.4 Adaptive Attack Evaluation

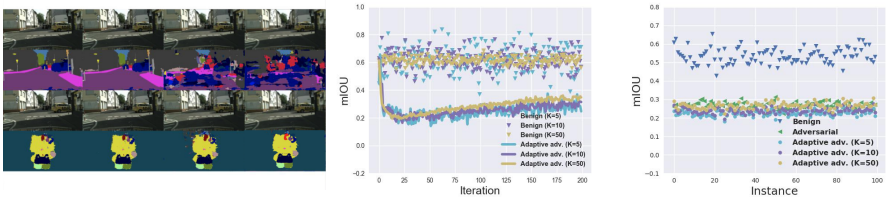
Regarding the above detection analysis, it is important to evaluate *adaptive attacks*, where adversaries have knowledge of the detection strategy.

Method		Model	mIOU	Detection DAG Pure	Detection DAG Kitty	Detection Houdini Pure	Detection Houdini Kitty	Detection Adap DAG Pure	Detection Adap DAG Kitty	Detection Adap Houdini Pure	Detection Adap Houdini Kitty
Scale (std)	0.5	DRN (16.4M)	66.7	100%	95%	100%	99%	100%	67%	100%	78%
	3.0			100%	100%	100%	100%	100%	0%	97%	0%
	5.0			100%	100%	100%	100%	100%	0%	71%	0%
	0.5	DLA (18.1M)	74.5	100%	98%	100%	100%	100%	75%	100%	81%
	3.0			100%	100%	100%	100%	100%	24%	100%	34%
	5.0			100%	100%	100%	100%	97%	0%	95%	0%
Spatial (K)	1	DRN (16.4M)	66.7	91%	91%	94%	92%	98%	94%	92%	94%
	5			100%	100%	100%	100%	100%	100%	100%	100%
	10			100%	100%	100%	100%	100%	100%	100%	100%
	50			100%	100%	100%	100%	100%	100%	100%	100%
	1	DLA (18.1M)	74.5	96%	98%	97%	97%	99%	99%	100%	100%
	5			100%	100%	100%	100%	100%	100%	100%	100%
	10			100%	100%	100%	100%	100%	100%	100%	100%
	50			100%	100%	100%	100%	100%	100%	100%	100%

Table 1: Detection results (AUC) of image spatial (Spatial) and scale consistency (Scale) based methods on Cityscapes dataset. The number in parentheses of the Model shows the number of parameters for the target mode, and mIOU shows the performance of segmentation model on pristine data. We color all the AUC less than 80% with red.

As Carlini & Wagner suggest [4], we conduct attacks with full access to the detection model to evaluate the adaptive adversary based on Kerckhoffs principle [36]. To perform adaptive attack against the image scaling detection mechanism, instead of attacking the original model, we add another convolutional layer after the input layer of the target model similarly with [4]. We select std $\in \{0.5, 3, 5\}$ to apply adaptive attack, which is the same with the detection model. To guarantee that the attack methods will converge, when performing the adaptive attacks, we select 0.06 for the upper bound for adversarial perturbation, in terms of L_2 distance (pixel values are in range [0,1]), since larger than that the perturbation is already very visible. The detection results against such adaptive attacks are shown in Table 1 on Cityscapes (We omit the results on BDD to supplementary materials). Results on adaptive attack show that the image scale based detection method is easily to be attacked (AUC of detection drops dramatically), which draws similar conclusions as in classification task [4]. We show the qualitative results in Figure 6 (a), and it is obvious that even under large std of Gaussian kernel, the adversarial example can still be fooled into the malicious target (Kitty).

Next, we will apply adaptive attack against the spatial consistency based method. Due to the randomness of the approach, we propose to develop a strong



(a) Image scaling (b) Convergence analysis (c) spatial consistency

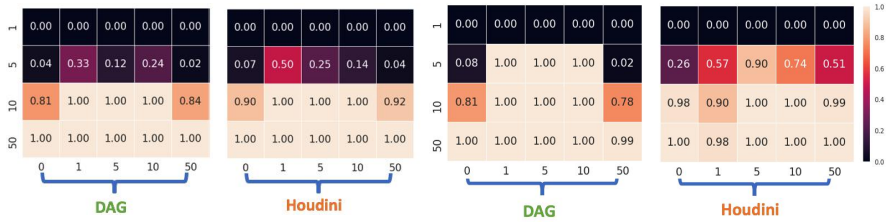
Fig. 6: Performance of adaptive attack. (a) shows adversarial image and corresponding segmentation result for adaptive attack against image scaling. The first two rows show benign images and the corresponding segmentation results; the last two rows show the adaptive adversarial images and corresponding segmentation results under different std of Gaussian kernel (0.5, 3, 5 for column 2-4). (b) and (c) show the performance of adaptive attack against spatial consistency based method with different K. (b) presents mIOU of overlapping regions for benign and adversarial images during along different iterations. (c) shows mIOU for overlapping regions of benign and adversarial instances at iteration 200.

adaptive adversary that we can think of by randomly select K patches (the same value of K used by defender). Then the adversary will try to attack both the whole image and the selected K patches to the corresponding part of malicious target. The detailed attack algorithm is shown in the supplementry materials. The corresponding detection results of the spatial consistency based method against such adaptive attacks on Cityscapes are shown in Table 1. It is interesting to see that even against such strong adaptive attacks, the spatial consistency based method can still achieve nearly 100% detection results. We hypothesize that it is because of the high dimension randomness induced by the spatial consistency based method since the search space for patches and the overlapping regions is pretty high. Figure 6 (b) analyzes the convergence of such adaptive attack against spatial consistency based method. From figure 6 (b) and (c), we can see that with different K, the selected overlapping regions still remain inconsistent with high probability.

Since the spatial consistency based method can induce large randomness, we generate a confusion matrix of detection results for adversaries and detection method choosing various K as shown in Figure 7. It is clear that for different malicious targets and attack methods, choosing $K = 50$ is already sufficient to detect sophisticated attacks. In addition, based on our empirical observation, attacking with higher K increases the computation complexity of adversaries dramatically.

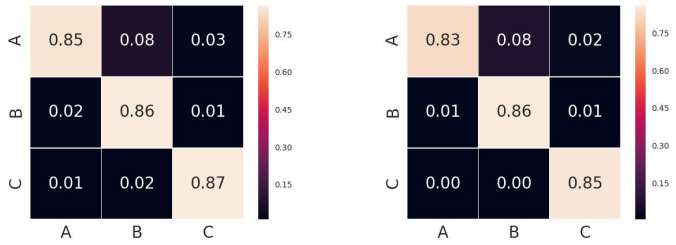
5.5 Transferability Analysis

Given the common properties of adversarial examples for both classifier and segmentation tasks, next we will analyze whether transferability of adversarial



(a) Kitty (b) Pure

Fig. 7: Detection performance of spatial consistency based method against adaptive attack with different K on Cityscapes with DRN model. X-axis indicates the number of patches selected to perform the adaptive attack (0 means regular attack). Y-axis indicates the number of overlapping regions selected for during detection.



(a) DAG (b) Houdini

Fig. 8: Transferability analysis: cell (i, j) shows the normalized mIoU value or pixel-wise attack success rate of adversarial examples generated against model j and evaluate on model i. Model A,B,C are DRN (DRN-D-22) with different initialization. We select "Hello Kitty" as target

examples exists in segmentation models considering they are particularly sensitive to spatial and scale information. *Transferability* is demonstrated to be one of the most interesting properties of adversarial examples in classification task, where adversarial examples generated against one model is able to mislead the other model, even if the two models are of different architectures. Given this property, transferability has become the foundation of a lot of black-box attacks in classification task. Here we aim to analyze whether adversarial examples in segmentation task still retain high transferability. First, we train three DRN models with the same architecture (DRN-D-22) but different initialization and generate adversarial images with the same target.

Each adversarial image has at least 96% pixel-wise attack success rate against the original model. We evaluate both the DAG and Houdini attacks and evaluate the transferability using normalized mIoU excluding pixels with the same

label for the ground truth adversarial target. We show the transferability evaluation among different models in the confusion matrices in Figure 8[1]. We observe that the transferability rarely appears in the segmentation task. More results on different network architectures and data sets are in the supplementary materials.

As comparison with classification task, for each network architecture we train a classifier on it and evaluate the transferability results as shown in supplementary materials. As a control experiments, we observe that classifiers with the same architecture still have high transferability aligned with existing findings, which shows that the low transferability is indeed due to the natural of segmentation instead of certain network architectures.

This observation here is quite interesting, which indicates that black-box attacks against segmentation models may be more challenging. Furthermore, the reason for such low transferability in segmentation is possibly because adversarial perturbation added to one image could have focused on a certain region, while such spatial context information is captured differently among different models. We plan to analyze the actual reason for low transferability in segmentation in the future work.

6 Conclusions

Adversarial examples have been heavily studied recently, pointing out vulnerabilities of deep neural networks and raising a lot of security concerns. However, most of such studies are focusing on image classification problems, and in this paper we aim to explore the spatial context information used in semantic segmentation task to better understand adversarial examples in segmentation scenarios. We propose to apply spatial consistency information analysis to recognize adversarial examples in segmentation, which has not been considered in either image classification or segmentation as a potential detection mechanism. We show that such spatial consistency information is different for adversarial and benign instances and can be potentially leveraged to detect adversarial examples even when facing strong adaptive attackers. These observations open a wide door for future research to explore diverse properties of adversarial examples under various scenarios and develop new attacks to understand the vulnerabilities of DNNs.

Acknowledgments We thank Warren He, George Philipp, Ziwei Liu, Zhirong Wu, Shizhan Zhu and Xiaoxiao Li for their valuable discussions on this work. This work was supported in part by Berkeley DeepDrive, Compute Canada, NSERC and National Science Foundation under grants CNS-1422211, CNS-1616575, CNS-1739517, JD Grapevine plan, and by the DHS via contract number FA8750-18-2-0011.

[1] Since the prediction of certain classes presents low IoU value due to imperfect segmentation, we eliminate K classes with the lowest IoU values to avoid side effects. In our experiments, we set K to be 13.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Bhagoji, A.N., He, W., Li, B., Song, D.: Exploring the space of black-box attacks on deep neural networks. arXiv preprint arXiv:1712.09491 (2017)
3. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: ECCV '08 Proceedings of the 10th European Conference on Computer Vision: Part I. pp. 44–57 (2008)
4. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. pp. 3–14. ACM (2017)
5. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49, `https://doi.org/10.1109/SP.2017.49`
6. Chan, T.F., Wong, C.K.: Total variation blind deconvolution. IEEE transactions on Image Processing **7**(3), 370–375 (1998)
7. Chen, H., Zhang, H., Chen, P.Y., Yi, J., Hsieh, C.J.: Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 2587–2597 (2018)
8. Chen, P.Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.J.: Ead: elastic-net attacks to deep neural networks via adversarial examples. arXiv preprint arXiv:1709.04114 (2017)
9. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. pp. 15–26. ACM (2017)
10. Cisse, M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured prediction models. arXiv preprint arXiv:1707.05373 (2017)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
12. Cui, W., Wang, Y., Fan, Y., Feng, Y., Lei, T.: Localized fcm clustering with spatial information for medical image segmentation and bias field estimation. Journal of Biomedical Imaging **2013**, 13 (2013)
13. Das, N., Shanbhogue, M., Chen, S.T., Hohman, F., Chen, L., Kounavis, M.E., Chau, D.H.: Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. arXiv preprint arXiv:1705.02900 (2017)
14. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853 (2016)
15. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (Jan 2015)
16. Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., Song, D.: Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945 (2017)

17. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. He, W., Wei, J., Chen, X., Carlini, N., Song, D.: Adversarial example defense: Ensembles of weak defenses are not strong. In: 11th USENIX Workshop on Offensive Technologies (WOOT 17). USENIX Association, Vancouver, BC (2017), https://www.usenix.org/conference/woot17/workshop-program/presentation/he
20. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine **29**(6), 82–97 (2012)
21. Hosseini, H., Chen, Y., Kannan, S., Zhang, B., Poovendran, R.: Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318 (2017)
22. Johnson, B., Xie, Z.: Unsupervised image segmentation evaluation and refinement using a multi-scale approach. ISPRS Journal of Photogrammetry and Remote Sensing **66**(4), 473–483 (2011)
23. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. pp. 109–117 (2011)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
25. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International journal of computer vision **43**(1), 29–44 (2001)
26. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2016)
27. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 (2016)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
29. Ma, X., Li, B., Wang, Y., Erfani, S.M., Wijewickrema, S., Houle, M.E., Schoenebeck, G., Song, D., Bailey, J.: Characterizing adversarial subspaces using local intrinsic dimensionality. arXiv preprint arXiv:1801.02613 (2018)
30. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
31. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 427–436. IEEE (2015)
32. Noda, K., Arie, H., Suga, Y., Ogata, T.: Multimodal integration learning of robot behavior using deep neural networks. Robotics and Autonomous Systems **62**(6), 721–736 (2014)
33. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)

34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
35. Saha, P.K., Udupa, J.K., Odhner, D.: Scale-based fuzzy connected image segmentation: theory, algorithms, and validation. Computer Vision and Image Understanding **77**(2), 145–174 (2000)
36. Shannon, C.E.: Communication theory of secrecy systems. Bell Labs Technical Journal **28**(4), 656–715 (1949)
37. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 1528–1540. ACM (2016)
38. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
39. Tabacof, P., Valle, E.: Exploring the space of adversarial images. In: Neural Networks (IJCNN), 2016 International Joint Conference on. pp. 426–433. IEEE (2016)
40. Tong, L., Li, B., Hajaj, C., Xiao, C., Vorobeychik, Y.: Hardening classifiers against evasion: the good, the bad, and the ugly. CoRR, abs/1708.08327 (2017)
41. Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)
42. Weng, T.W., Zhang, H., Chen, H., Song, Z., Hsieh, C.J., Boning, D., Dhillon, I.S., Daniel, L.: Towards fast computation of certified robustness for relu networks. arXiv preprint arXiv:1804.09699 (2018)
43. Weng, T.W., Zhang, H., Chen, P.Y., Yi, J., Su, D., Gao, Y., Hsieh, C.J., Daniel, L.: Evaluating the robustness of neural networks: An extreme value theory approach. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=BkUHlMZ0b
44. Wu, Z., Shen, C., Hengel, A.v.d.: Wider or deeper: Revisiting the resnet model for visual recognition. arXiv preprint arXiv:1611.10080 (2016)
45. Xiao, C., Li, B., yan Zhu, J., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 3905–3911. International Joint Conferences on Artificial Intelligence Organization (7 2018). https://doi.org/10.24963/ijcai.2018/543, https://doi.org/10.24963/ijcai.2018/543
46. Xiao, C., Sarabi, A., Liu, Y., Li, B., Liu, M., Dumitras, T.: From patching delays to infection symptoms: Using risk profiles for an early discovery of vulnerabilities exploited in the wild. In: 27th USENIX Security Symposium (USENIX Security 18). USENIX Association, Baltimore, MD (2018), https://www.usenix.org/conference/usenixsecurity18/presentation/xiao
47. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=HyydRMZC-
48. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (2018)
49. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: International Conference on Computer Vision. IEEE (2017)

50. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (ICLR) (2016)
51. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Computer Vision and Pattern Recognition (CVPR) (2017)
52. Yu, F., Wang, D., Darrell, T.: Deep layer aggregation. arXiv preprint arXiv:1707.06484 (2017)
53. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 (2018)
54. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2335–2344 (2014)
55. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)