On the Robustness of Panoptic Segmentation Models to Adversarial Attacks

Wen-Fu Lee
UW-Madison
wlee256@wisc.edu

Huawei Wang
UW-Madison
hwang665@wisc.edu

Shuoxuan Dong
UW-Madison
sdong34@wisc.edu

Abstract

Panoptic segmentation unifies the typically distinct tasks of semantic segmentation and instance segmentation. With the growing research interest in panoptic segmentation models, the impacts of adversarial attacks on these panoptic segmentation models still remains unknown. In our study, we are working on a rigorous evaluation of adversarial attacks on panoptic segmentation model. In addition, some defense methods are applied to improve the robustness of the original model.

1. Introduction

Panoptic segmentation draws attentions from researchers since 2018. While currently most of the researchers focus on improving the performance of the models themselves, the robustness of those popular models still remain unknown to our best knowledge. While these models have potential usage in safety-crucial applications such as self-driving cars[8], they could be vulnerable to adversarial samples. So it is crucial to evaluate the robustness of these models under adversarial attacks and try to defend the models.

The problem we are focusing on is how to generate adversarial input image examples to fool the panoptic segmentation models, so that the models gives different outputs thus decrease the performance. Moreover, how to make these models more robust under adversarial attacks would be extensively studied.

In this study, we choose UPSNet[18], a cutting-edge panoptic segmentation model that permits back-propagation in an end-to-end manner, as the target network to implement the attack methodologies and also try some defense method.

2. Related Work

Panoptic segmentation is proposed by FAIR and Heidelberg University first time in [10]. The paper recalls this pre-deeplearning tasks to revive the interest of the community in a more unified view of image segmentation. This paper gives the definition of panoptic segmentation and compares the differences between this task with semantic/instance segmentation. Also, it proposes a new comprehensive panoptic quality(PQ) metric to treat all classes in a uniform way and to better describe the task. In [9], the proposed Panoptic FPN solves semantic and instance segmentation efficiently and makes it memory and computationally efficient. It is a modified version of Mask-RCNN with FPN and tries to add a new semantic segmentation branch to FPN to acquire better performance. [13] claims to propose the first end-to-end network incorporating 2 segmentation tasks which shares the backbone features but applies different head branches for the two tasks. It also introduces the unique spatial ranking module to solve the overlapping issue and becomes the state-of-art solution for this challenge on COCO dataset. [12] proposes AUNet that can leverage proposal and mask level attention and get better background results. In [18], a single unified panoptic head is used to combine the outputs from instance segmentation head and semantic segmentation head to speed up the complex task inference. [2] summarizes some attack methods commonly used for deep neural networks including FGSM and iterative FGSM. It also provides a initial report on the robustness of ordinary segmentation network.

It has been shown that machine learning models are often vulnerable to adversarial manipulation of their input intended to cause incorrect classification[4]. In particular, neural networks and many other categories of machine learning models are highly vulnerable to attacks based on small modifications of the input to the model at test time. [3] [6] [16]

3. Method

3.1. Attack Method

We use the popular attack methods on UPSNet to test the robustness of the original model. The attack methods are introduced as following:

- Fast gradient sign method(FGSM): find the gradient of loss with respect to the correct labels, take element wise sign and update the input image samples in resulting directions. This method is easy to apply as long as

you acquire the gradient of the input after doing backward propogation.

$$x \leftarrow x + \epsilon * sgn\left(\frac{\partial L(x, y^*)}{\partial x}\right) \quad (1)$$

- Projected Gradient Descent: Take multiple small steps using either FGSM or least likely label method, each step clip the result to the ϵ neighborhood of the original image. We tune both the learning rate, ϵ value and iteration numbers during the experiments.
- Least likely label method: this method is pretty similar to fast gradient sign method except that we use the least likely label in the network output as the direction to misclassify the input images.

$$x \leftarrow x - \epsilon * sgn\left(\frac{\partial L(x, \hat{y})}{\partial x}\right) \quad (2)$$

3.2. Defense Method

Several methods are also proposed to defense the adversarial attack phenomena. [15] claims that the first order method can efficiently solve the saddle point formulation and motivate projected gradient descent (PGD) as a universal first-order adversary. It also shows that network capacity can play an important role in correctly classifying adversarial samples. [17] introduces Ensemble Adversarial Training, a technique that augments training data with perturbations transferred from other models. [14] proposed a SafetyNet architecture. It consists of the original classifier, and an adversary detector which looks at the internal state of the later layers in the original classifier. If the adversary detector declares that an example is adversarial, then the sample is rejected. [7] applying image transformations such as bit-depth reduction, JPEG compression, total variance minimization, and image quilting before feeding the image to a convolutional network classifier and shows that total variance minimization and image quilting are very effective defenses which can eliminate most of the gray-box and black-box attacks. We use the similar input preprocess method to defense our model. Specifically, we design a noise removal filter using OpenCV's function $cv2.fastNlMeansDenoisingColored()$. This function converts an image to CIELAB colorspace and then separately denoises L and AB components.

4. Experiments and Results

We describe the dataset, evaluation metrics, evaluations of UPSNet, adversarial attacks, and defenses in this section. Exhaustive details are included in the supplementary. We have also attached our codes with this report to Canvas.

Dataset Cityscapes has 5000 images of ego-centric driving scenarios in urban settings which are split into 2975, 500 and 1525 for training, validation and testing respectively. Its image size is 2048 x 1024, and it consists of 8 and 11 classes for *thing* and *stuff*.

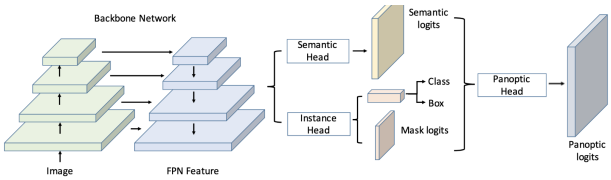

Figure 1. Overall Architecture of UPSNet

Figure 2. Panoptic segmentation result of UPSNet

Table 1. Evaluation metric comparison

	PQ	SQ	RQ
Paper	59.3	79.7	73.0
Reproduce	57.9	79.1	71.7

Panoptic Quality Metric. We plan to use a new panoptic quality (PQ) proposed in the first panoptic segmentation paper [10]. PQ will be calculated for each class independently and averaged over classes. This makes PQ insensitive to class imbalance. Given true positives (TP), false positives (FP), and false negatives (FN), PQ is defined as

$$PQ = \frac{\Sigma_{(p,g)\in TP} IoU(p, g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (3)$$

, where *p* is the predicted segment and *g* is the ground truth segment.

PQ can be seen as the multiplication of a segmentation quality (SQ) term and a recognition quality (RQ) term. This decomposition provides insight for analysis:

$$PQ = SQ * RQ \quad (4)$$

$$SQ = \frac{\Sigma_{(p,g)\in TP} IoU(p, g)}{|TP|} \quad (5)$$

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (6)$$

4.1. UPSNet Implementation

UPSNet is designed as a unified panoptic segmentation network to approach the panoptic segmentation problem. Unlike previous methods which have two separate branches designed for semantic and instance segmentation

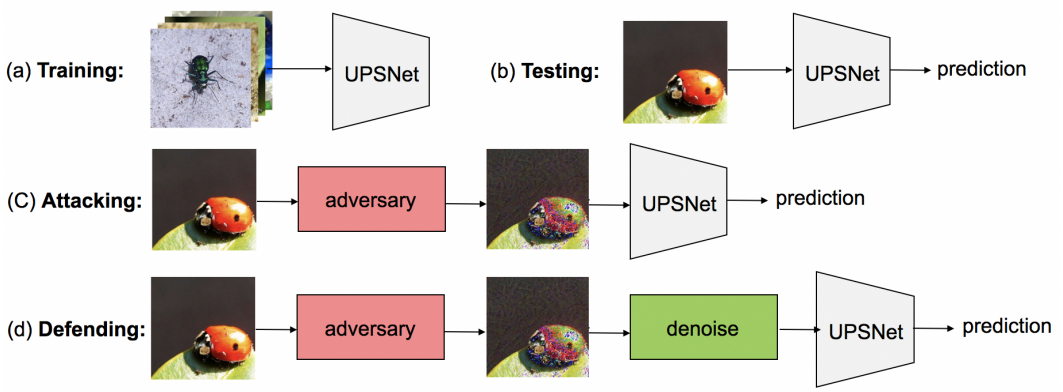

Figure 3. The pipelines of training, testing, attacking, and defending, respectively. (a) and (b) are for 4.1 UPSNet Implementation; (a) and (c) are for 4.2 Adversarial Attacks; (a) and (d) are for 4.3 Defenses.

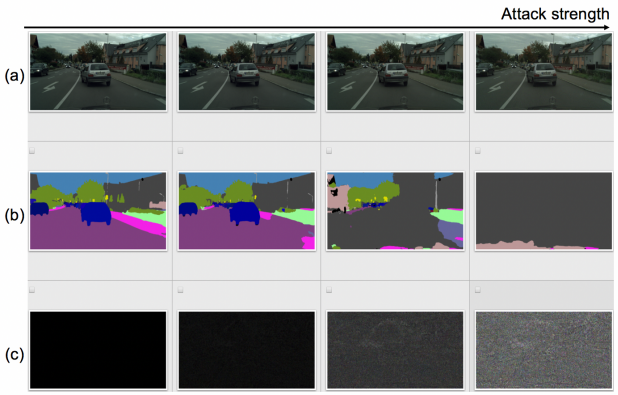

Figure 4. Example 1 of adversarial attacks. (a) are adversarial images; (b) are panoptic segmentations; (c) are difference images, the absolute difference between original images and (a). Difference images are multiplied by a constant scaling factor to increase visibility.

individually, this model exploits a single network as backbone to provide shared representations and stands as a unified end-to-end model. Besides, two heads on top of the backbone are used for solving these tasks simultaneously. The semantic head builds upon deformable convolution and leverages multi-scale information from feature pyramid networks (FPN). The instance head follows the Mask RCNN design and outputs mask segmentation, bounding box and its associated class. In addition, a light weight panoptic head is designed to provides a better way of resolving the conflicts between semantic and instance segmentation. We reproduce the original UPSNet model using the link [1] provided by the author. After solving the environment issue and fixing some model configuration, we can achieve almost the same evaluation metrics as the author claimed in the paper. The segmentation results is shown in Fig.2.

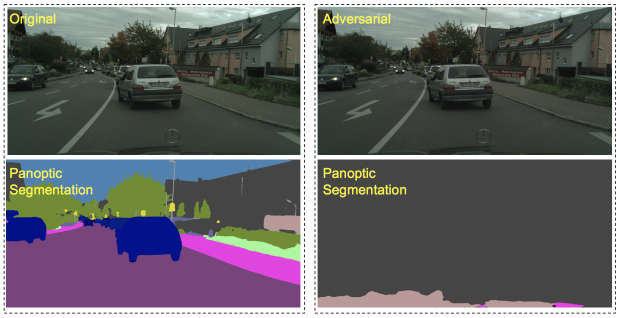

Figure 5. Comparison between an original image and an adversarial image. The left column shows the original image and its panoptic segmentation; the right column shows the adversarial results.

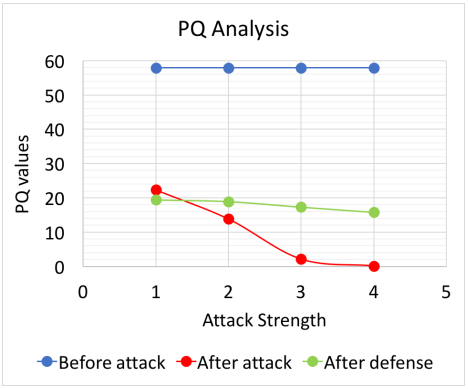

Figure 6. PQ analysis. The blue, red, and green lines show the PQ values before attack, after attack, and after defense, respectively.

4.2. Adversarial Attacks

The attacking pipeline is shown in Fig. 3 (c). We use the Project Gradient Descent (PGD) attack described in Sec. 3. Following [11], we set the number of iterations of iterative attacks to $\min(\epsilon + 4, 1.25\epsilon)$ and step-size = 1 meaning that the value of each pixel is changed by 1 every iteration.

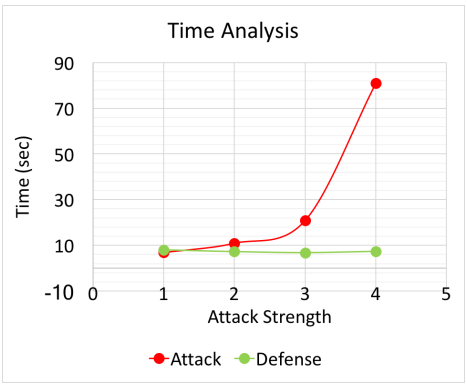

Figure 7. Time analysis. The red and green lines show the time spent to generate attack and defense samples, respectively.

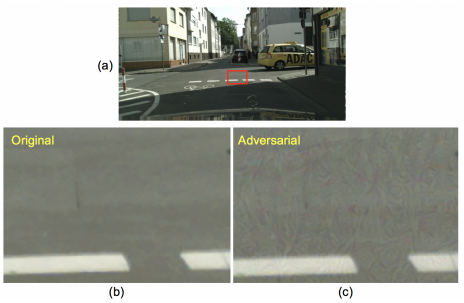

Figure 8. The noise pattern brought by the adversarial generation can be observed.

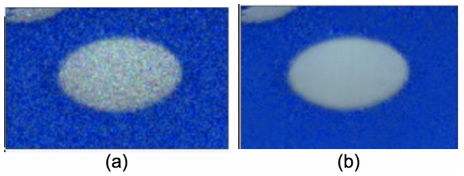

Figure 9. The denoising filter can be applied to remove noises. (a) and (b) are the images before and after noise removal, respectively.

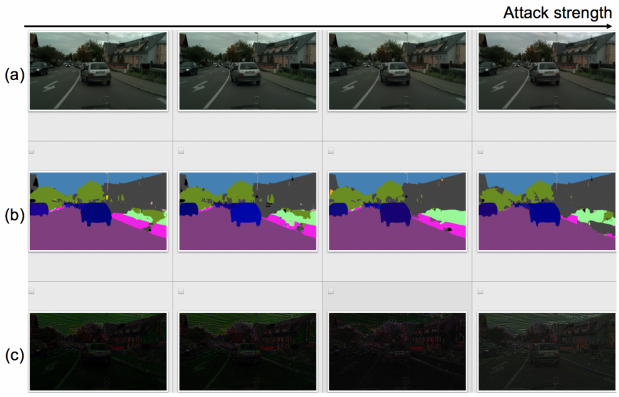

Figure 10. Example 1 of defenses. (a) are denoised images; (b) are panoptic segmentations; (c) are difference images, the absolute difference between adversarial images and (a). Difference images are multiplied by a constant scaling factor to increase visibility.

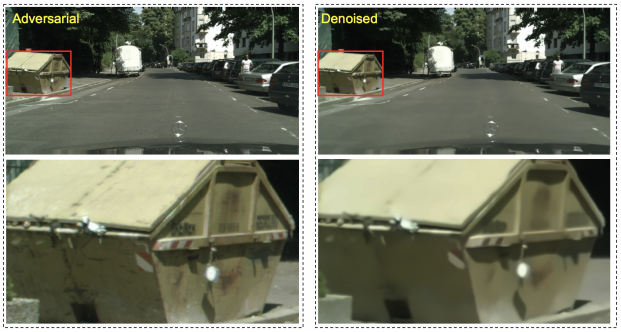

Figure 11. Comparison between an adversarial image and an denoised image. The left column shows the adversarial image and its cropped and resampled region; the right column shows the denoised results.

We evaluate these attacks when setting the l_∞ norm of the perturbations ϵ to each value from $\{0.25, 1, 4, 16\}$. The larger the value of ϵ, the stronger the attack strength. Even small values such as $\epsilon = 0.25$ introduce errors in the UPSNet model re-trained in Sec. 4.1, as shown in Fig. 3 (a). The maximum value of ϵ is chosen as 16 since the perturbation is quite conspicuous at this point.

One qualitative example of these attacks is shown in the Fig. 4. As expected, when the attack strength is stronger, the results of panoptic segmentation get worse. The interesting is that as shown in Fig. 5, even if the attack is so strong that the panoptic segmentation result is not meaningful anymore, we can still hardly tell the difference between the original image and the adversarial one. More results can be found in the appendix.

The attack method is also evaluated in a quantitative way, PQ analysis and time analysis shown in the Fig. 6 and 7, respectively. Fig. 6 shows that even the weakest attack can downgrade the original PQ value from 57.9 to 22.3. The PQ value gets even worse, which is 0.1, at the strongest attack. However, there is no free lunch to generate a strong attack. Fig. 7 shows that the stronger the attack is, the more time we need to generate it.

4.3. Defenses

There are multiple ways to defend a learning model. One way is to re-train the model with the adversarial samples. However, training a regular UPSNet alone already takes around 3 days. It will take another couple of weeks to re-train the model since the generation of adversarial samples involves back propagation of gradients and iterative processing, all of which are time-consuming steps. So, a different approach is adopted here. First, we observe the irregular pattern, shown in Fig. 8, brought by the adversarial generation. Obviously, it is unlike natural textures, which can be found in real objects, but instead more like noises, which can be removed using a denoising filter. Thus,

in our implementation, we adopt the fastNlMeansDenoising function in OpenCV for noise removal. This function converts an image to CIE*LAB* colorspace and then separately denoises *L* and *AB* components. For example, its effect is shown in Fig. 9. The setting adopted here is cv2.fastNlMeansDenoisingColored(image, None, 10, 10, 7, 21). The details of each parameter can be found in [5]. As for the defending pipeline, it is shown in Fig. 3 (d).

Compared to Fig. 4, the qualitative example in Fig. 10 shows that the results of panoptic segmentation after defenses are still well-structured for each attack. This proves that the denosing filter adopted here can defend the adversarial attacks and keep the panoptic segmentation quality. More specifically, Fig. 6 shows that the PQ values can be improved for attack strengths 2-4 although the original PQ values cannot be recovered. On the other hand, Fig. 7 shows that it takes around 7 seconds to finish a defense which is much faster than an attack generation. One more thing to note is that in Fig. 6, we see that the PQ value downgrades a little at the attack strength 1 after the defense. This downgrade can be explained in Fig. 11. Compared to the adversarial image, the denoised image is less noisy but also loses some details and textures. However, they are sometimes the keys for a learning model to be able to do the inference well. This issue can be improved by dynamically adjusting the parameters of the denoising filter according to the image quality, but this is not addressed in this report.

5. Challenge & Difficulty

We meet up with several difficulties listed as below:

- We meet some difficulties in re-implementing the state-of-art solutions. Since there are still not any open-source solutions for the state-of-art methods, we choose UPS-Net which is a relatively complete solution for panoptic segmentation task. However, many environment issues arise as the original implementation uses some python & C combined programming, the CUDA, GCC and Pytorch version have to be in consistent with the author's environment. We discussed with the author and fix several issues during the model setup process.

- In order to attack the model, we want to apply general attack methods such as fast gradient method and project gradient descent method. Those method all required that we acquired the gradient with respect to input tensor such that we can generate adversarial samples. However, at the beginning the input's gradient is always a None object. Through tracing the gradient for several layers after backward on the whole network, we found that the author basically uses detach method to prevent the loss propagation to previous layers.

- We try to use the least likely methods to attack the model but with no success at the end. The reason is that the UPSNet has several head for different metrics output. The segmentation head has a label while the head using for output the bounding box doesn't. Thus it is hard to define a combined label for the whole networks' output.

- We explore different methods to defense the model. We pick up the relatively simple methods which preprocess the adversarial samples and then feed in the network. The experiments proved that this simple method is an efficient way to make the model more robust.

6. Conclusion & Future Work

1. Traditional adversarial attack can have great impact on the deep learning models that are specially designed for the panoptic segmentation task. For example, as our experiment shows, the PGD attack reduce the PQ metric to a really low level.

2. Simple defensive method works well. For example, in our case, perform image denoising using Nonlocal Means Denoising algorithm with several computational optimizations improves PQ when the model is under strong attack.

3. However, when the model is attacked by weak adversarial examples, the defensive method can downgrade the PQ value.

4. For future work, it would be good the try the work on other different network of implementations.

5. Different attack methods can be tried in the future so that the we can better understand the whole picture.

6. It would be great if more robust defense method can be tried on the panoptic segmentation networks. Due to time limit of this project, we only tried some preprocessing filter that denoise the original figure. Although this method work in our case, it might not be generic.

7. Contributions of Team Members

- Wen-Fu Lee

    1. Set up the UPSNet's environment on GCP
    2. Trained and tested the UPSNet model
    3. Implemented PGD to attack the UPSNet model
    4. Implemented the denoising filter for defenses
    5. Prepared the experimental results in both presentation slides and this report

- Huawei Wang

    1. Setting up UPSNet and another model for semantic segmentation using feature pyramid network as backbone.
    2. Conduct research on attacking the panoptic segmentation models.
    3. Organize experiments data and finish project reports.

- Shuoxuan Dong

    1. Tried Several implementations of Panoptic Segmentation other than UPSNet.
    2. Tried the adversarial attack method mentioned in [2] on our models.
    3. Literature review on the related work.

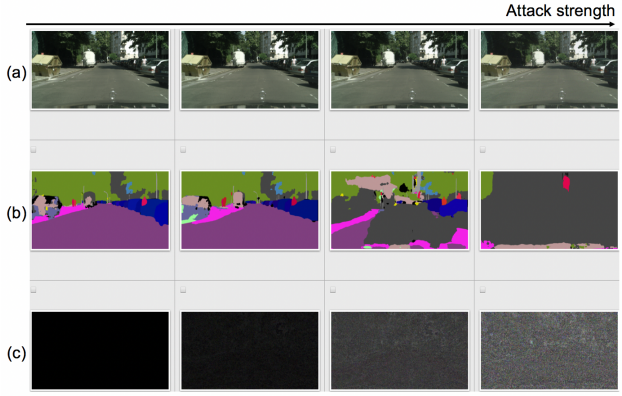

Figure 12. Example 2 of adversarial attacks. (a) are adversarial images; (b) are panoptic segmentations; (c) are difference images, the absolute difference between original images and (a). Difference images are multiplied by a constant scaling factor to increase visibility.

References

[1] https://github.com/uber-research/upsnet.

[2] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. *CoRR*, abs/1711.09856, 2017.

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[4] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004.

[5] fastNlMeansDenoisingColored. https://docs.opencv.org/3.0-alpha/modules/photo/doc/denoising.html. *OpenCV*.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

[8] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *arXiv preprint arXiv:1704.05519*, 2017.

[9] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *CoRR*, abs/1901.02446, 2019.

[10] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CoRR*, abs/1801.00868, 2018.

[11] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ICLR*, 2017.

[12] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided Unified Network for Panoptic Segmentation. *arXiv e-prints*, page arXiv:1812.03904, Dec 2018.

[13] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An End-to-End Network for Panoptic Segmentation. *arXiv e-prints*, page arXiv:1903.05027, Mar 2019.

[14] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. 03 2017.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[16] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[17] Florian Tramr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[18] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. UPSNet: A Unified Panoptic Segmentation Network. *arXiv e-prints*, page arXiv:1901.03784, Jan 2019.

Appendices

A. More Attack Results

B. More Defense Results

Figure 13. Example 3 of adversarial attacks. (a) are adversarial images; (b) are panoptic segmentations; (c) are difference images, the absolute difference between original images and (a). Difference images are multiplied by a constant scaling factor to increase visibility.

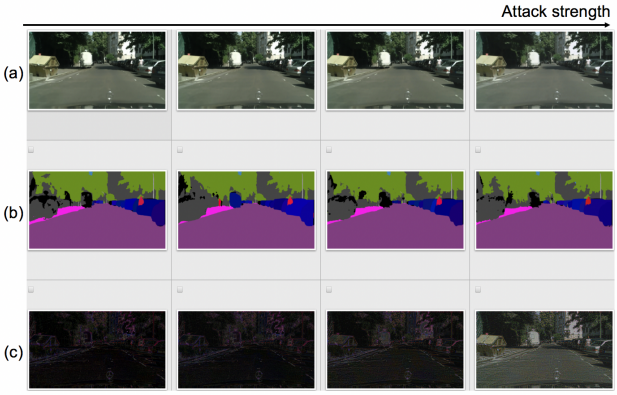

Figure 14. Example 2 of defenses. (a) are denoised images; (b) are panoptic segmentations; (c) are difference images, the absolute difference between adversarial images and (a). Difference images are multiplied by a constant scaling factor to increase visibility.

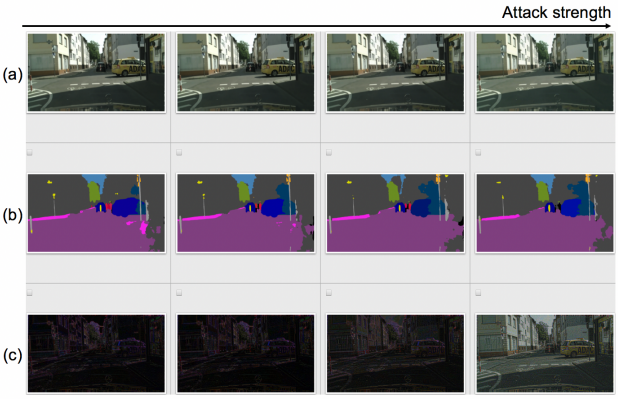

Figure 15. Example 3 of defenses. (a) are denoised images; (b) are panoptic segmentations; (c) are difference images, the absolute difference between adversarial images and (a). Difference images are multiplied by a constant scaling factor to increase visibility.