Adversarial Attacks and Mitigations on Scene Segmentation of Autonomous Vehicles

Yuqing Zhu[1], Sridhar Adepu[1,2(✉)], Kushagra Dixit[1,2], Ying Yang[3], and Xin Lou[3,4]

[1] University of Bristol, Bristol, UK
sridhar.adepu@bristol.ac.uk
[2] Reperion, Singapore, Singapore
kd@reperion.io
[3] Advanced Digital Sciences Center, Singapore, Singapore
[4] Singapore Institute of Technology, Singapore, Singapore
https://reperion.io

Abstract. In this study, we focus on the effectiveness of adversarial attacks on the scene segmentation function of autonomous driving systems (ADS). We explore both offensive as well as defensive aspects of the attacks in order to gain a comprehensive understanding of the effectiveness of adversarial attacks with respect to semantic segmentation. More specifically, in the offensive aspect, we improved the existing adversarial attack methodology with the idea of momentum. The adversarial examples generated by the improved method show higher transferability in both targeted as well as untargeted attacks. In the defensive aspect, we implemented and analyzed five different mitigation techniques proven to be effective in defending against adversarial attacks in image classification tasks. The image transformation methods such as JPEG compression and low pass filtering showed good performance when used against adversarial attacks in a white box setting.

Keywords: Security · Autonomous vehicles · Deep learning · Adversarial attacks · Semantic segmentation

1 Introduction

With the rapid development of deep learning, fully autonomous driving is gradually becoming a reality. Deep Neural Networks (DNNs) show incredible performance in solving computer vision tasks such as classification, detection, and segmentation, and provide efficient solutions to Autonomous Driving Systems (ADS) for the same. ADS use a wide range of sensors including cameras, RADAR's and LIDAR's to monitor the environment around them and collect visual, positioning and mapping data. This data is then used by the ADS to

Sridhar Adepu: Primary affiliation is University of Bristol.

S. Katsikas et al. (Eds.): ESORICS 2022 Workshops, LNCS 13785, pp. 46–66, 2023.
https://doi.org/10.1007/978-3-031-25460-4_3

have a comprehensive understanding of the surrounding environment with the help of DNNs where the techniques which are used for sensor fusion and scene segmentation are fairly mature. However, ADS is extremely security critical, and any safety and reliability issues can lead to severe and irreversible consequences. In [14], authors divided the attacks on autonomous driving vehicles into three categories. These are attacks on the physical sensors, control systems and connection mechanisms. In this work, we focus on a specific attack technique called the adversarial attack which is a technique that utilizes the vulnerabilities of the DNN to mislead control systems into making wrong decisions.

In [8], authors demonstrated that DNNs are vulnerable to adversarial attacks. These adversarial attacks can cause machine learning models to give an incorrect output with a high level of confidence by adding subtle perturbations to the input samples. Therefore, adversarial attacks become a potential security threat to autonomous vehicles that use DNN. In [11], authors mitigated the adversarial attack by utilizing JPEG compression. However, most of these studies about adversarial attacks and defence against them focus on image classification tasks, which require less computational complexity compared to semantic segmentation. Semantic segmentation plays a key role in autonomous vehicles since it helps the ADS to differentiate between various important regions in visual data. In [1], authors evaluated the robustness of semantic segmentation models to adversarial attacks. In [20], authors created a dense attack generation approach to generate adversarial instances that challenge DNN-based scene segmentation and object detection models at the same time. However, the required computational intensity demands harder optimization for training segmentation models and thus adversarial attacks require much more effort.

In this paper, we propose a momentum based adversarial attack that specifically addresses the semantic segmentation tasks in autonomous vehicles. The proposed method utilizes momentum which is a technique used in deep learning to achieve an efficient black-box attack, i.e., the attack can work well against various segmentation models. Moreover, our methodology can launch effective attacks in either targeted or untargeted scenarios, which gives flexibility for the attacker's objectives. We also implemented and analyzed five mitigation methods based on image transformation. In summary, following are our contributions:

- We analyze the robustness of DNN based semantic segmentation models against adversarial attacks in an autonomous vehicles scenario. To address the computationally demanding nature of semantic segmentation models, we propose to leverage the idea of momentum to the Iterative Fast Gradient Sign Method (I-FGSM) adversarial attack algorithm which can reduce the required computational effort and significantly increase the transferability.
- We validate adversarial attack methodology by attacking state-of-the-art semantic segmentation models on a common real-world segmentation dataset i.e. "Cityscapes". Our experiments show that momentum based I-FGSM performs significantly better than the original I-FGSM in a targeted setting.
- We verified the viability of using image transformations as a mitigation technique against adversarial attack in the context of semantic segmentation models. We add another preprocessing layer before sending data into the semantic

segmentation model that can remove the effect of the adversarial perturbation in the input image without modify the architecture of the model or the training process. The results show that image transformation functions such as low pass filtering and JPEG compression can mitigate adversarial attacks in a white box setting against semantic scene segmentation models.

The remaining article organisation: Sect. 2 reviews prior work in adversarial attack. Section 3 elaborates on the momentum based I-FGSM attack. Section 4 shows the experimental settings together with the results including both attack and defence scenarios. Section 5 concludes the article.

2 Background

This section aims to provide an introduction to semantic segmentation (Sect. 2.1) and adversarial attacks (Sect. 2.2).

2.1 Semantic Segmentation

Semantic segmentation is a pixel-level classification task. The semantic segmentation model needs to assign each pixel of the input image to a class. It is an important task in autonomous vehicles that is used to help the ADS understand the input image and solve vision tasks such as discovering drivable/undrivable areas. DNN-based semantic segmentation models have been widely employed by ADS to help autonomous vehicles when it comes to performing tasks such as scene perception. However, the safety of DNN when it comes to such tasks is questionable at best, for example, DNN shows low reliability while facing malicious attacks that use adversarial attack methodologies [8,18].

2.2 Adversarial Attack

The adversarial attack is a technique that can cause a malfunction in a DNN. It can cause the DNN to give an incorrect output with a high level of confidence by adding subtle disturbances to the input samples. In [18], authors showed that adversarial examples have strange transferability. That is, the neural network is statistically vulnerable to the adversarial examples generated by another neural network. There are two types of adversarial attacks - white box attacks and black box attacks. For white box attacks, the attacker has information about the architecture of the target neural network. For a black box attack, the architecture of the target neural network is not available to the attacker. There are various ways of generating adversarial examples. The Fast Gradient Sign Method (FGSM) [8] utilizes the gradient of the loss function to generate adversarial examples. Carlini & Wagner's attack [4] utilizes optimization-based methods to launch an adversarial attack. Jacobian-based Saliency Map attack [13] exploits saliency maps and increases high-saliency pixels to lead to a misclassification by the deep neural network. In the following sub-section, we briefly introduce FGSM and its variants.

Fast Gradient Sign Method (FGSM). FGSM was one of the first effective adversarial attacks introduced in [8]. FGSM generates adversarial perturbations by maximizing the gradient of the loss for the input. Equation (1) shows the detail of untargeted FGSM:

$$x_{adv} = x + \varepsilon \cdot \text{sign}\left(\nabla_x L\left(x, y\right)\right), \tag{1}$$

where x_{adv} is the adversarial example, x is the input without perturbation, y is the label of the input, L is the loss function of the model, ∇_x is the gradient function and sign $(\nabla_x L\,(x, y))$ is the direction that will maximise the loss. The constant value ε is the magnitude of the perturbation. The attack calculates $(\nabla_x L\,(x, y))$ by back-propagating the gradient. Then it adjusts the input in the direction of sign $(\nabla_x L\,(x, y))$.

Iterative Fast Gradient Sign Method (I - FGSM). In [12], the author raised an iterative version of FGSM (I-FGSM) that applied FGSM in a recurring fashion with a smaller step size to increase the efficiency of the attack. Equation (2) shows the detail of untargeted I-FGSM:

$$x_{adv}^{t+1} = x_{adv}^{t} + \varepsilon \cdot \alpha \cdot \text{sign}\left(\nabla_x L\left(x_{adv}^{t}, y\right)\right), \tag{2}$$

where Eq. (2) is inherited from Eq. (1), α is the step size of I-FGSM and set to ε/T to restrict the adversarial example in a bounded L2 norm where T is the number of iterations. For a targeted attack, the aim is to minimize the loss between the adversarial example and the target label $y*$ such that the adversarial example will be predicted as target label $y*$. Equation (3) shows the detail of targeted I-FGSM:

$$x_{adv}^{t+1} = x_{adv}^{t} + \varepsilon \cdot \alpha \cdot \text{sign}\left(\nabla_x L\left(x_{adv}^{t}, y^*\right)\right). \tag{3}$$

The I-FGSM can generate finer adversarial examples that do not spoil the visual content even with a greater attack magnitude [12].

3 Work Execution

Section 3.1 list the two main drawbacks of the original I-FGSM. Section 3.2 elaborates our Momentum-based I-FGSM attack method. Section 3.3 shows the structure of mitigation methodologies.

3.1 Drawbacks of I-FGSM

As introduced in Sect. 2, I-FGSM can successfully cause an incorrect prediction during image classification tasks. However, this method shows two drawbacks when attacking semantic segmentation models. The first drawback being that adversarial examples generated by I-FGSM show poor transferability, which leads to deficient performance in a black box setting. The transferability of adversarial examples occur because multiple machine learning models learn comparable decision boundaries around a data point [7], making adversarial examples

designed for one model effective against others. However, the I-FGSM is prone to falling into a suboptimal local optimum which greatly reduces the transferability of the adversarial examples. The second drawback is that it is hard to achieve convergence with I-FGSM. This disadvantage becomes more apparent in the case of segmentation networks as they are more complex in nature than classification models, with higher computational complexity, rendering the process of searching for minute perturbations difficult. To improve I-FGSM and to overcome these drawbacks, we integrate the idea of momentum into the original I-FGSM algorithm.

3.2 Momentum-Based I-FGSM

The momentum based I-FGSM attack is inspired by the momentum technique which is used to optimize the Stochastic Gradient Descent (SGD) algorithm in DNN [15]. Figure 1 shows the details of the progression of our momentum based adversarial attack algorithm.

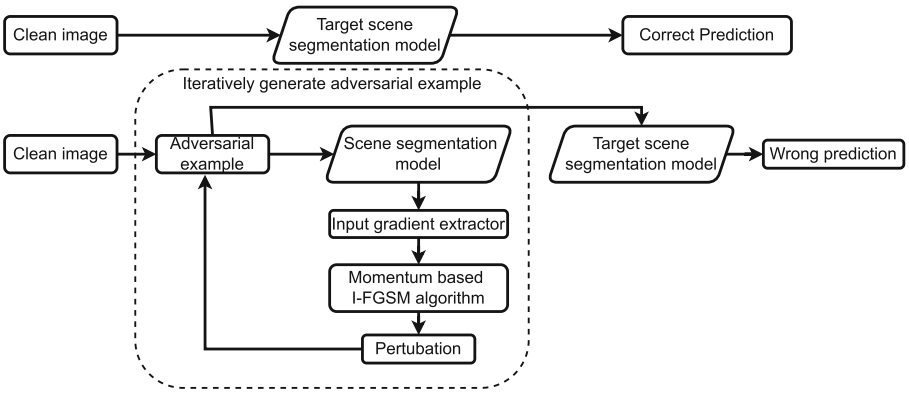


Fig. 1. Workflow of momentum based I-FGSM adversarial attack. The segmentation model used to generate adversarial examples is same with the target scene segmentation model in white box settings and is different in black box settings

In DNN, SGD is widely used to modify the network parameters in minimizing the difference between the prediction by the network and the real data. For each iteration, the weights are updated, and the weight vector is moved towards the direction of the negative gradient at the current position. However, there is a certain probability that SGD is stuck in a local minimum or saddle instead of the global minimum. Momentum is used to mitigate and optimize the SGD algorithm in this aspect. In Gradient Descent with Momentum, the change in the weight vector depends on both the current gradient and the previous sequence of gradients. The Eqs. (4) and (5) show the Gradient Descent with Momentum:

$$V_t = \beta V_{t-1} + \alpha \nabla_w L(W, X, y), \tag{4}$$

where:

$$W = W - V_t, \tag{5}$$

Here, L is the loss function, α is the learning rate, β is a hyperparameter that is used to adjust the influence of the earlier gradients. V_t stands for the "current descent velocity" which is based on the metaphor of velocity from physics. V_t is updated depending on the current gradient and the previous velocity.

The idea of momentum can level out the variations and lead to faster convergence when the direction of the gradient keeps changing. When in a ravine, it is difficult to find the global minimum utilising pure SGD because the direction of the gradient is almost perpendicular to the direction of the global minimum hence the algorithm will oscillate in the ravine and make small actual progress in the direction towards the global minimum. Momentum can be used to mitigate this oscillatory behavior and accelerate the SGD. This is because the optimization direction depends on both the current as well as the previous gradient directions which leads to the oscillations being counteracted between them. The technical background of the I-FGSM is introduced in Sect. 2.2. Here we try to integrate momentum into the I-FGSM. In following section, the I-FGSM with momentum for both targeted attack and untargeted attack will be introduced.

Reviewing the equations for I-FGSM:

$$x_{adv}^{t+1} = x_{adv}^{t} + \varepsilon \cdot \alpha \cdot \operatorname{sign}\left(\nabla_x L\left(x_{adv}^{t}, y\right)\right). \tag{6}$$

In I-FGSM for each iteration the adversarial example is updated along the direction of the current gradient. In I-FGSM with momentum, for each iteration the adversarial example is updated along the direction of momentum where the momentum accumulates the direction vector for gradients in previous steps. Equations (7) and (8) are the equations for I-FGSM with momentum:

$$x_{adv}^{t+1} = x_{adv}^{t} + \alpha \cdot \operatorname{sign}\left(g_{t+1}\right), \tag{7}$$

where:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x L\left(x_{adv'}^{t}, y\right)}{||\nabla_x L\left(x_{adv'}^{t}, y\right)||_p}, \tag{8}$$

where μ is a hyperparameter called decay factor that is used to adjust the influence of the earlier gradients, α is the learning rate, p denotes the order of the norm which is normally set as 1 or 2 to represent L1 norm and L2 norm. Algorithm 1 shows the algorithm for momentum based I-FGSM untargeted attack bounded by L2 norm.

Algorithm 1: Momentum based I-FGSM

Input :
A semantic segmentation network f with loss function L;
Input image x;
Ground-truth label y;
The size of perturbation ϵ;
Iteration number T;
Decay factor μ;

Output:
An adversarial example x^* with $\|x^* - x\|_2 < \varepsilon$

$\alpha = \epsilon/T$;
$\boldsymbol{g}_0 = 0$;
$\boldsymbol{x}_0^* = \boldsymbol{x}$;
for $t = 0$ *to* $T - 1$ **do**
  Input $\boldsymbol{x}_t^*$ to f and calculate the gradient $\nabla_x J\left(\boldsymbol{x}_t^*, y\right)$;
  Update $\boldsymbol{g}_{t+1}$ by accumulating the velocity vector in the gradient direction as:
  $\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_x J\left(\boldsymbol{x}_t^*, y\right)}{\left\|\nabla_x J\left(\boldsymbol{x}_t^*, y\right)\right\|_2}$;
  Update $\boldsymbol{x}_{t+1}^*$ by applying the sign gradient as
  $\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* + \alpha \cdot \operatorname{sign}\left(\boldsymbol{g}_{t+1}\right)$;
end
return $\boldsymbol{x}^* = \boldsymbol{x}_T^*$;

In a targeted attack, the goal is to make the model misclassify with a specific target in mind, i.e., the predicted class of input x to be a targeted class y^* in y. This kind of attack is a source target misclassification. When it comes to the result, the predicted class of the input x will be changed from the original class label to y^*.

Equations (9) and (10) are the details to targeted I-FGSM with momentum:

$$x_{adv}^{t+1} = x_{adv}^t - \alpha \cdot \operatorname{sign}\left(g_{t+1}^*\right), \tag{9}$$

where:

$$g_{t+1}^* = \mu \cdot g_t - \frac{\nabla_x L\left(x_{adv}^t, y^*\right)}{\left|\left|\nabla_x L\left(x_{adv}^t, y^*\right)\right|\right|_p}, \tag{10}$$

Here Y^* is the target label. Unlike a nontargeted attack, a targeted attack tries to drive the output towards a target classification. Hence it needs to minimize the loss function. Algorithm 2 shows the algorithm for targeted I-FGSM with momentum. Compared with algorithm 1, the ground truth label y is replaced by the target label Y^* so that the adversarial example x will lead the DNN to make the prediction as target label Y^*.

Algorithm 2: Momentum based I-FGSM for targeted attack

Input :
A semantic segmentation network f with loss function L;
Input image x;
Target label y^*;
The size of perturbation ϵ;
Iteration number T;
Decay factor μ;

Output:
An adversarial example x^* with $\|x^* - x\|_2 < \varepsilon$

$\alpha = \epsilon/T$;
$\boldsymbol{g}_0 = 0$;
$\boldsymbol{x}_0^* = \boldsymbol{x}$;
for $t = 0$ *to* $T - 1$ **do**
  Input $\boldsymbol{x}_t^*$ to f and calculate the gradient $\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_t^*, y^*\right)$;
  Update $\boldsymbol{g}_{t+1}$ by accumulating the velocity vector in the gradient direction as:
  $g_{t+1}^* = \mu \cdot g_t - \frac{\nabla_x L\left(x_{adv}^t, y^*\right)}{\left|\nabla_x L\left(x_{adv}^t, y^*\right)\right|_2}$;
  Update $\boldsymbol{x}_{t+1}^*$ by applying the sign gradient as
  $x_{adv}^{t+1} = x_{adv}^t - \alpha \cdot \operatorname{sign}\left(g_{t+1}^*\right)$;
end
return $\boldsymbol{x}^* = \boldsymbol{x}_T^*$;

3.3 Mitigation

We studied five different image pre-processing mitigation techniques In this experiment, we added another preprocessing layer before sending the data into the semantic segmentation model. Using this layer, we evaluated five different image transformation functions, these are: JPEG compression [6,11], bit-depth reduction [21], total variance minimization [9], low pass filtering [16] and PCA denoising [3]. Figure 2 shows the workflow of the mitigation methodology.

Fig. 2. Workflow of the preprocessing defence. We tested 5 different image transformation functions in the preprocessing layer including JPEG Compression, Bit-depth Reduction, Total Variance Minimization, Low Pass Filter and PCA Denoising.

4 Experiments and Results

This section presents the setting and results for the experiments that demonstrate the performance of I-FGSM with momentum.

4.1 Experiment Settings

Dataset. In this study, we evaluate the proposed adversarial attack methodology using the Cityscapes segmentation dataset. Cityscapes [5] is a widely used segmentation dataset and entire dataset consists of street scenes from 50 different cities. In this study, we generate adversarial examples against and evaluate the performance of this validation data set.

Target Models. Dual Graph Convolutional Network (Dual-GCN) [22] uses a graph neural network to capture object correlation and improve semantic linkages. On Cityscapes, Dual-GCN achieves SOTA performance of 76% mIoU. In this exercise, Dual-GCN is used to generate adversarial examples against the target model to evaluate the performance of the adversarial attacks in white box setting. Image Cascade Network (ICNet) [23] is a real-time lightweight semantic segmentation model that guarantees speed and accuracy. It uses a cascade of image inputs to employ a cascade of feature fusion units and uses cascade label guidance during training, which can refine semantic predictions with relatively low computational cost. On Cityscapes, ICNet achieves 74% mIoU. In this study, ICNet is used as the target model in a black box setting to test the transferability of adversarial examples.

Experimental Setups. This work has been implemented using Pytorch on Python 3.7. The experiments were run using Google Colab and The Bristol Blue Crystal 4 supercomputer. Both platforms provided a single Nvidia Tesla P100 GPU with 16GB of memory as the main AI accelerator. In the experiments mean Intersection over Union per class (mIoU) is utilised as a metric to assess the performance of untargeted attacks while the class wise Intersection over Union (IoU) utilized for the assessment of targeted attacks. The adversarial examples in the experiments are generated by attacking the Dual-GCN. In a white box attack setting the adversarial examples are tested against the same model that they are generated from i.e. the Dual-GCN. In a black box attack setting the adversarial examples are evaluated against ICNet.

Hyperparameter Configuration. We evaluate the relationship between decay factor and the effects of momentum based I-FGSM. The decay factor controls the size of impact of the earlier gradients as mentioned in Sect. 3.2. With a larger decay factor the past gradients have a greater impact on the direction of change of the weight vector. When the decay factor is equal to 0 the past gradients have no impact on the update direction and the momentum based I-FGSM reverts to a normal I-FGSM. In this experiment the attack strength is set to 40 and the iteration number is set to 10. The decay factor is evaluated from 0 to 2 with an interval of 0.2. Both white box as well as black box attack scenarios are evaluated.

Number of Iterations. We compare the effect of the number of iterations between I-FGSM and momentum based I-FGSM. In this experiment the attack

strength is set to 40 for both I-FGSM and momentum based I-FGSM. The decay factor for momentum based I-FGSM is set to 1.0, based on the results detailed in Sect. 4.2. Iterations ranging from 1 to 10 times are tested for both attack methodologies. White box as well as black box attack testing is done for both I-FGSM and momentum based I-FGSM to check the relationship between the number of iterations and the transferability of adversarial examples.

Attack Strength. We evaluate the effectiveness of adversarial attacks with differing attack strengths for I-FGSM as well as momentum based I-FGSM. The number of iterations for both the attacks are set to 10. Same as the previous experiment, the decay factor of the momentum based I-FGSM is set to 1.0. The attack strength is evaluated from 5 to 40 in increments of 5. Both white box as well as black box attacks are evaluated.

Target Attack. The study of targeted attacks is important for autonomous vehicles since the detection accuracy of certain classes such as "person" and "car" largely affects the safety of such ADS. Therefore, in this experiment we evaluate the performance of a targeted attack using I-FGSM and momentum based I-FGSM with varying attack strengths. Configurations for the target labels are inspired by [10]. Two sets of targeted labels are generated by modifying the original labels from the Cityscapes dataset. The details of the targeted sets are as follows:

- Set 1: The labels of classes "person", "rider", "motorcycle" and "bicycle" are replaced by the label "vegetation".
- Set 2: The labels of classes "car", "truck", "bus" and "train" are changed to "road".

Defence. We also implement and evaluate five mitigation methodologies based on different image transformation functions. All of the five mitigation methodologies are introduced in Sect. 3.3. Specifically, JPEG compression is performed with a 75% quality ratio. For bit-depth reduction input image bit depth is reduced to 5 bits. The scikit-image package [19] is used to implement the total variance minimization and low pass filter. For total variance minimization the strength is set to 2.5 and we have applied the low pass filter to each color channel with a 20% frequency cut of ratio. PCA was performed on each input image by selecting the 150 largest principal components. All the defence methods are tested against both I-FGSM based as well as momentum based I-FGSM adversarial attacks with different attack strengths in the range of 5 to 40. Testing is done in both a white box as well as a black box setting for a comprehensive analysis.

4.2 Impact of Parameters

This sub-section shows the results of the experiments detailed in Sect. 4.1.

Decay Factor. As mentioned in Sect. 3.2, the decay factor is an important hyperparameter for momentum based I-FGSM. Figure 3 shows the mIoU of the ICNet and Dual-GCN with the momentum based I-FGSM adversarial examples generated from Dual-GCN. The Y axis shows the mIoU of the model. In this experiment, smaller mIoU means better performance of the adversarial attack. X axis shows the size of the decay factor. Larger the decay factor greater the effect the previous gradients have on the updated direction of the adversarial example. For a white box attack, the performance of the attack decreases with an increase of the decay factor starting from 0.2.

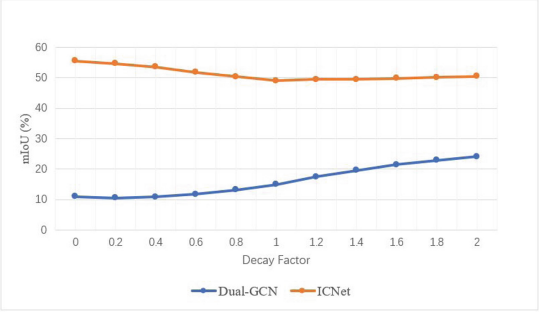
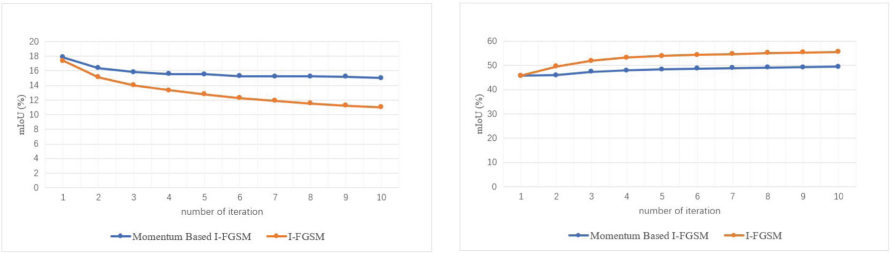


Fig. 3. The mIoU (%) of the adversarial examples generated for Dual-GCN against Dual-GCN (whitebox) and ICNet (blackbox), with a decay factor ranging from 0 to 2

These results show that the momentum based I-FGSM has the best performance when the decay factor is equal to 0.2 for a white box attack setting. However, for a black box attack, the performance of momentum based I-FGSM increases with an increase of the decay factor and archives best performance when the decay factor is equal to 1.0. Subsequently the performance slowly decreases with an increase in the decay factor. When the decay factor is equal to 1.0, the weight update for each iteration is simply represented by the sum of all prior gradients.

Number of Iteration. The number of iterations influences the performance of iterative adversarial attacks. Here are the results of the experiments that study the effect of the number of iterations against momentum based I-FGSM and I-FGSM. Figure 4(a) shows the result for a white box attack while Fig. 4(b) shows the result for a black box attack. In these two figures the Y axis shows the mIoU of the model and X axis shows the number of iterations.

Momentum based I-FGSM converged at around 4 iterations while I-FGSM shows no evidence of convergence even at 10 iterations. When the number of iterations is 10, the I-FGSM has a 4% reduction in the mIoU when compared with the momentum based I-FGSM. The I-FGSM has a constant learning rate. The reason I-FGSM has difficulty converging may be due to the direction of the update being completely dependent on the current gradient, but the gradient



(a) Dual-GCN (white box) (b) ICNet (black box)

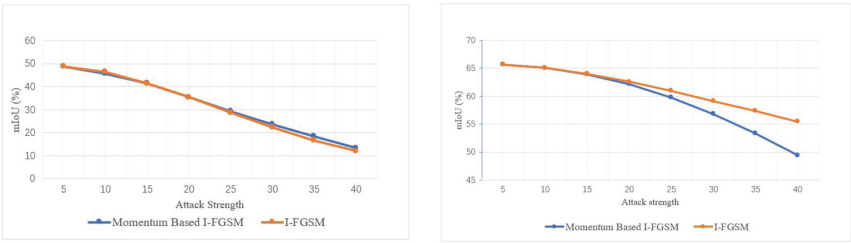
Fig. 4. The mIoU (%) of the adversarial examples generated for Dual-GCN with momentum based I-FGSM and I-FGSM against (a): Dual-GCN (white box) (b): ICNet (black box), with the number of iterations ranging from 1 to 10.

will become exceedingly small when approaching the optimal value and because of the constant learning rate, the I-FGSM will slow down, and might even fall into a local optimum. From the result it is obvious that I-FGSM shows better performance in a white box attack setting which proves that I-FGSM can very easily overfit a specific model. Figure 4(b) shows that the momentum based I-FGSM outperformed the I-FGSM in a black box attack setting. The momentum based I-FGSM reduces the mIoU of the model by 5% compared to the I-FGSM when the number of iterations equals 10. This also proves that the adversarial examples generated from I-FGSM can easily overfit with the white box model and have poor transferability.

Attack Strength. We then study the relationship between the attack strength of adversarial examples and the accuracy of the semantic segmentation models. Figure 5(a) shows the results of attacking Dual-GCN by Momentum based I-FGSM and I-FGSM. Here the Y axis is the mIoU of the model and the X axis is the attack strength. The lines for momentum based I-FGSM and I-FGSM almost overlap when the attack strength is small in a white box setting and the I-FGSM is shown to have an exceedingly small advantage when the attack strength is larger than 35. Both the momentum based I-FGSM and I-FGSM show good performance in a black box setting and the mIoU of the semantic segmentation model decreases linearly with the strength of the attack.

Figure 5(b) shows the results for a black box setting. With an increase in the attack strength, the momentum based I-FGSM leads to a faster decrease of the mIoU of the semantic segmentation model compared with I-FGSM. When the attack strength is 40 the momentum based I-FGSM leads to a 6% greater decrease in mIoU compared with the original I-FGSM. In a black box attack, the momentum based I-FGSM can reach the required effect with a smaller attack strength which means it would be more difficult to detect such an attack manually.

Targeted Attack. In this section, we demonstrate the results of the targeted adversarial attack. As introduced in Sect. 4.1. We designed two sets of target labels. Figure 6 shows two examples of momentum based I-FGSM targeted



(a) Dual-GCN (white box) (b) ICNet (black box)

Fig. 5. The mIoU (%) of the adversarial examples generated for Dual-GCN with momentum based I-FGSM and I-FGSM against (a): Dual-GCN (white box) (b): ICNet (black box), with the attack strength ranging from 5 to 40.

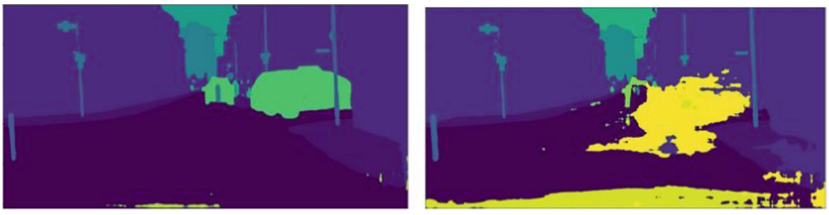


Fig. 6. Two examples for targeted momentum based I-FGSM: The left image shows an example from target label set 1 and where right image is from target label set 2.

attack. the left image shows an example from target label set 1 and the right image shows an example from target label set 2. In the image on the left the model cannot classify the pixels belonging to the class "person" correctly and in the right image the network cannot correctly classify the pixels in proximity to the car.

I: Targeted Attack with White-Box Setting. Table 1 shows the results of a targeted adversarial attack with label set 1 and Table 2 shows the results of label set 2. The details of the label set 1 and 2 are shown in Sect. 4.1. The results shows that both momentum based I-FGSM and I-FGSM show good performance in a white box attack. Both adversarial attack methods can reduce the targeted categories IoU to 0 with a small attack magnitude. It is worth noting that the mIoU increases as the attack magnitude increases. A similar observation is made with the untargeted attack, I-FGSM shows better performance with the same attack magnitude. Adversarial examples generated from I-FGSM lead to a higher mIoU while keeping the IoU of targeted classes at 0. In other words, the adversarial examples generated by I-FGSM cause less damage to the other classes while maintaining 100% attack success rate for targeted classes.

Table 1. Black box targeted attacks set 1: misclassified person, rider, motorcycle, and bicycle into the label of vegetation

Attack method	Attack strength	mIoU	Categories IoU			
			Person	Rider	Motorcycle	Bicycle
Momentum based I-FGSM	40	70.770	0	0	0	0
Momentum based I-FGSM	5	62.677	0	0	0	0
I-FGSM	40	74.028	0	0	0	0
I-FGSM	5	62.864	0	0	0	0
No attack	0	76.113	80.952	60.235	62.777	76.125

Table 2. White box targeted attacks set 2: misclassified car, truck, bus, and train into the label of road

Attack method	Attack strength	mIoU	Categories IoU			
			Car	Truck	Bus	Train
Momentum based I-FGSM	40	73.600	0	0	0	0
Momentum based I-FGSM	5	49.882	0	0	0	0
I-FGSM	40	71.520	0	0	0	0
I-FGSM	5	59.203	0	0	0	0
No attack	0	76.113	94.178	74.254	83.002	67.480

II: Targeted Attack with Black-box Setting. Table 3 shows the results of a targeted adversarial attack with label set 1 and Table 4 shows the results of label set 2 in a black box setting. The momentum based I-FGSM performs better than I-FGSM in a black box attack. The momentum based I-FGSM significantly reduces the IoU for all the targeted classes with the same attack strength when compared with I-FGSM. This proves that the addition of momentum helps to increase the transferability of adversarial examples. From Table 3, it can be observed that the effect of the attack varies for the various categories. The experimental results do not clearly show a reason for such a difference.

Table 3. Black box targeted attacks set 1: misclassified person, rider, motorcycle, and bicycle into the label of vegetation

Attack method	Attack strength	mIoU	Categories IoU			
			Person	Rider	Motorcycle	Bicycle
Momentum based I-FGSM	40	73.520	69.369	41.148	36.341	65.658
Momentum based I-FGSM	5	65.966	74.501	53.059	47.163	71.101
I-FGSM	40	68.970	73.332	49.496	43.461	70.015
I-FGSM	5	65.959	74.500	53.061	47.165	71.010
No attack	0	74.068	78.707	57.704	58.407	74.274

Table 4. Black box targeted attacks set 2: misclassified car, truck, bus, and train into the label of road

Attack method	Attack strength	mIoU	Categories IoU			
			Car	Truck	Bus	Train
Momentum based I-FGSM	40	67.916	88.322	36.989	53.580	49.906
Momentum based I-FGSM	5	65.934	93.037	49.303	62.108	56.909
I-FGSM	40	68.537	92.095	43.962	58.154	51.463
I-FGSM	5	65.924	93.037	49.260	62.064	56.909
No attack	0	74.068	94.159	76.607	81.329	60.075

4.3 Defence

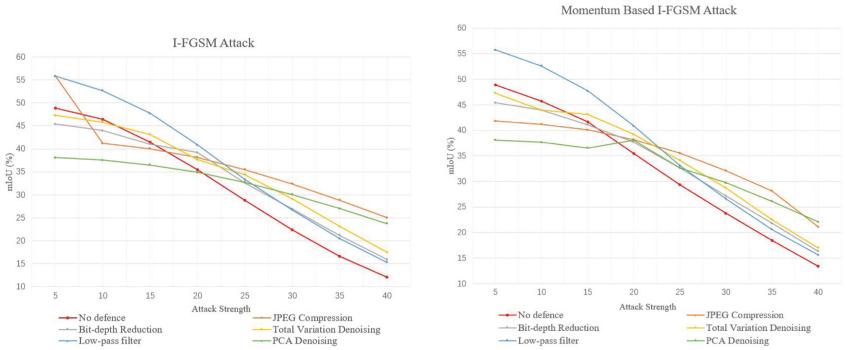
This sub-section details the results for the various mitigation techniques when used against adversarial attacks. The results are separated into white box and black box scenarios. For both white box and black box setting the effect of the five defence methodologies against the momentum based I-FGSM and I-FGSM are tested at attack strengths ranging from 5 to 40. The impact of the defence methodologies on the model with a clean input are shown in Table 5.

Table 5. mIoU of defence methods on Dual-GCN model with clean input

Defence	No Defence	JPEG	Low Pass Filter	Bit-depth Reduction	TVM	PCA
mIoU (%)	76.113	65.911	70.990	50.095	59.487	65.989

It is important to note that all of the defensive methodologies have a negative impact on the accuracy of the model. As it can be observed, the Bit-depth Reduction and PCA Denoising yield the smallest decrease in the mIoU on clean inputs, followed by JPEG. Low-pass filtering and Total Variance Minimization lead to a more significant decrease in the performance of the semantic segmentation model. The negative impact of these defensive methodologies may have a worse effect on the network accuracy compared to adversarial attacks of a smaller intensity.

I: Defence against White-box Attack Fig. 7 shows the performance of the various defence methodologies performance in a white box setting. Figure 7(a) shows the effectiveness of the 5 defence methodologies against an I-FGSM attack and fig.7(b) shows the results of the defence methodologies against a momentum based I-FGSM attack. Here the Y axis is the mIoU of the model and the X axis is the attack strength.



(a) (b)

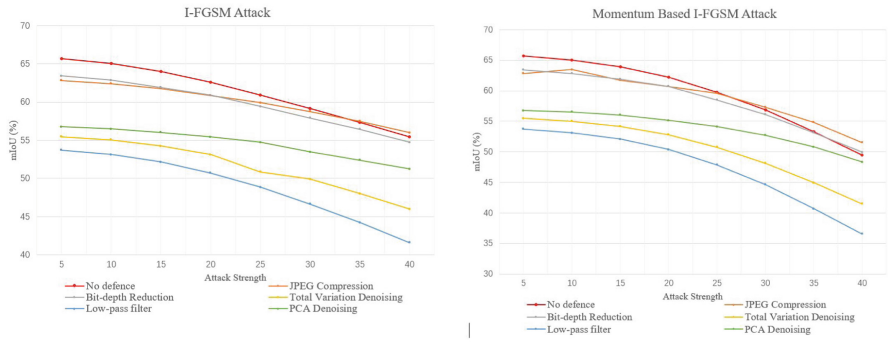
Fig. 7. The mIoU (%) of Dual-GCN with defence methods against (a): I-FGSM attack and (b): momentum based I-FGSM attack (white box setting).

In a white box setting the five defence methodologies show a similar effect when utilised against momentum based I-FGSM and I-FGSM. The low pass filter is effective at all attack magnitudes and increases the accuracy of the model in both the adversarial attack scenarios. The remaining four defence methodologies do not perform very well when the intensity of the attack is low. However, JPEG compression and PCA denoising show better performance compared to the other three defence methods as the strength of the attack increases.

As mentioned above, JPEG compression shows the best defensive performance under high intensity adversarial attacks and low pass filtering shows the best performance among the five defences against adversarial attacks of a low intensity. Both the JPEG compression and the low pass filtering remove the high frequency information from the input image. Therefore, it is reasonable to believe that the perturbation produced by the adversarial attack contain high frequency components.

In [16] and [9], the results show that these basis transformation functions are more effective when used against adversarial attacks in a classification task. This may be due to the classification networks being more sensitive to adversarial attacks and having a larger tolerance for image transformations. For the dataset used in [16] and [9], each image contains only one object so the image transformation functions such as blurring have a lesser effect on the accuracy of the classification models compared to this experiment.

II: Defence Against Black-Box Attack. Figure 8 shows the performance of the various defence methodologies in a black box setting. Figure 8(a) is the results of five defence methods against I-FGSM attack and Fig. 8(b) shows the results of defence methodologies against a momentum based I-FGSM attack.



(a) (b)

Fig. 8. The mIoU (%) of ICNet with defence methods against (a): I-FGSM attack and (b): momentum based I-FGSM attack (black box setting).

Among the five defence methodologies, JPEG compression becomes effective only when the attack strength is greater than 35 for I-FGSM and greater than 30 for momentum based I-FGSM. The bit depth reduction is effective for momentum based I-FGSM when the attack strength reaches 40. PCA denoising, total variation denoising and PCA denoising do not perform well against any of the attacks at any given strength.

None of the tested defence methodologies are effective when used against small perturbations in this setting. This is consistent with the result of basis transformation defences when used against classification in [16]. Since the adversarial examples with a small attack strength have a limited impact on the performance of the network in a black box setting, the various image pre-processing methodologies lead to negative impacts on the accuracy of the network which are comparable to adversarial perturbations. Table 6 Shows the impact of the 5 image pre-processing methodologies on the ICNet network with a clean input. The lowpass filter denoising and total variance denoising lead to a significant reduction in the accuracy of the network. This largely impacts the effectiveness of these two defensive methodologies to eliminate the adversarial perturbation and mitigate the adversarial attack.

Table 6. mIoU of defence methods on ICNet model with clean input

Defence	No defence	JPEG	Low pass filter	Bit-depth reduction	TVM	PCA
mIoU (%)	74.068	69.499	71.232	54.318	54.879	67.170

4.4 Discussions

The Effectiveness of Adversarial Attacks on Scene Segmentation. The experiment results show that adversarial attacks can significantly decrease the

performance of semantic segmentation models based on DNN. In a white box setting, the mIoU of SOTA semantic segmentation models can easily drop from around 75% to about 12% with the adversarial examples generated from early adversarial attack methods. The result from a targeted attack is also not optimistic. Adversarial attacks can greatly decrease the models' accuracy for certain classes while maintaining a high mIoU. This means such targeted attacks are more difficult to detect since the model can still make a correct prediction for the rest of the classes and function normally. In autonomous vehicles, this deserves more attention since the accuracy for certain classes such as pedestrians and traffic lights are naturally more important than the accuracy for some other classes and hence average accuracy is not a good indicator of reliability. In a black box setting, although the semantic segmentation model shows better resistance to adversarial attacks, the adversarial examples still lead to about 25% decrease in the mIoU. This shows that the adversarial examples are effective against different models.

Momentum Based I-FGSM Shows Better Transferability. The results show that the idea of momentum improves the transferability of I-FGSM in both targeted as well as untargeted attacks which compensates for the drawbacks of I-FGSM. The momentum based I-FGSM outperforms the original I-FGSM in a black box setting and has similar performance in a white-box settings. The transferability of adversarial examples is based on the fact that different DNNs learn through similar decision boundaries [7]. The original I-FGSM adjusts the adversarial examples by relying solely on the current gradient of the iteration which has a high probability of falling into suboptimal local maximas. As a result, the adversarial noise only interferes with a local decision boundary and the adversarial examples have extremely poor transferability. The momentum based I-FGSM solves this problem by changing the direction of adversarial examples utilizing past gradients as well as the current gradient into a single gradient which can level out the variations in the weight change direction and hence help the adversarial examples in finding the global maxima. Therefore, the adversarial examples generated by momentum based I-FGSM have better transferability.

The Performance of Defence Methods. In a black box setting all of these image transformation methods show relatively poor effectiveness when used against adversarial attacks with a small attack strength in semantic segmentation tasks. This is in comparison to their performance in image classification tasks detailed in prior research. Transformations such as low pass filter denoising and total variance denoising lead to a large decrease in the accuracy of the semantic segmentation model itself. The image transformation functions tested in this study all have negative impacts on the quality of the images to some extent. For example, the JPEG compression is a lossy compression that discards some of the high frequency components of the image. This quality loss caused by the transformation function leads to a greater impact on the model accuracy compared to the impact of adversarial examples in the context of black box testing.

Related Works: In [25], the authors evaluated the adversarial attack against semantic segmentation models. Unlike this work that uses the visual data collected from camera, [25] the focus was on the data from LiDAR (Light Detection and Ranging) sensors. They showed that the LiDAR semantic segmentation models used in ADS are also vulnerable to adversarial attacks. Combined with this work, the adversarial attack could still be a security threat to ADS that utilise different sensors since semantic segmentation models utilising various types of data are vulnerable to adversarial attacks. In [24] authors designed a pre-processing model that exploits the invariant features. The pre-processing model can disentangle the invariant features that represent sematic classification informations from adversarial noise and then restore the examples without adversarial perturbation by utilizing these invariant features. The authors declared that this defence methodology presents superior effectiveness when used against previously unseen adversarial attacks so it is effective when used against a black box attack. However, like most of the prior studies, this study focused on the image classification models so the effectiveness of this mitigating method in semantic segmentation scenario is uncertain.

Recently many researches are also focusing on using model-specific strategies to mitigate adversarial attacks. These strategies usually change the architecture or the training procedures of the DNN and utilise the learning algorithms or regularization method to enforce features such as invariance and smoothness [17]. [2] focused on utilizing adversarial training and defensive distillation to increase the robustness of traffic sign classification models. The results showed that the combination of these two defence techniques can achieve higher accuracy when used against different kinds of adversarial attacks in traffic sign classification tasks.

5 Conclusion

In this study, we focused on the adversarial attack and its mitigations in semantic segmentation tasks. We first applied the I-FGSM adversarial attack methodology to the task of semantic segmentation. Next, in order to enhance the transferability of the adversarial examples, we integrated the idea of momentum into the original I-FGSM algorithm. Extensive experiments were conducted to verify the efficacy of this momentum based I-FGSM technique. The results showed that momentum based I-FGSM has similar performance when compared to the original I-FGSM in both targeted as well as untargeted attack in a white box settings and outperformed the same in black box settings. From a mitigation standpoint, we focused on image pre-processing, and applied and tested five different image transformation functions. The results of the experiments showed that Low pass filtering and JPEG compression have superior performance when used against adversarial attacks in a white box setting. However, all five transformation methods showed limited performance when used against adversarial attacks in a black box setting. In future work, we want to investigate the feasibility of combining pre-processing defence methodologies with adversarial training to improve the robustness of AV systems.

Acknowledgment. This project is supported by the National Research Foundation, Singapore and National University of Singapore through its National Satellite of Excellence in Trustworthy Software Systems (NSOE-TSS) office under the Trustworthy Computing for Secure Smart Nation Grant (TCSSNG) award no. NSOE-TSS2020-01. This research was supported by grants from NVIDIA and utilised NVIDIA Quadro RTX 6000 GPUs.

References

1. Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 888–897 (2018)
2. Aung, A.M., Fadila, Y., Gondokaryono, R., Gonzalez, L.: Building robust deep neural networks for road sign detection. arXiv preprint arXiv:1712.09327 (2017)
3. Bhagoji, A.N., Cullina, D., Sitawarin, C., Mittal, P.: Enhancing robustness of machine learning systems via data transformations. In: 2018 52nd Annual Conference on Information Sciences and Systems (CISS). pp. 1–5. IEEE (2018)
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
5. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
6. Das, N., et al.: Keeping the bad guys out: protecting and vaccinating deep learning with jpeg compression. arXiv preprint arXiv:1705.02900 (2017)
7. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193 (2018)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
10. Kang, X., Song, B., Du, X., Guizani, M.: Adversarial attacks for image segmentation on multiple lightweight models. IEEE Access **8**, 31359–31370 (2020)
11. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
12. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
13. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P), pp. 372–387. IEEE (2016)
14. Pham, M., Xiong, K.: A survey on security attacks and defense techniques for connected and autonomous vehicles. Comput. Secur. **109**, 102269 (2021)
15. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural Netw. **12**(1), 145–151 (1999)
16. Shaham, U., et al.: Defending against adversarial images using basis functions transformations. arXiv preprint arXiv:1803.10840 (2018)
17. Shaham, U., Yamada, Y., Negahban, S.: Understanding adversarial training: Increasing local stability of neural nets through robust optimization. arXiv (2015)
18. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv (2013)

19. Van der Walt, S., et al.: scikit-image: image processing in python. PeerJ **2**, e453 (2014)
20. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1369–1378 (2017)
21. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
22. Zhang, L., Li, X., Arnab, A., Yang, K., Tong, Y., Torr, P.H.: Dual graph convolutional network for semantic segmentation. arXiv preprint arXiv:1909.06121 (2019)
23. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: ICNet for real-time semantic segmentation on high-resolution images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 418–434. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_25
24. Zhou, D., Liu, T., Han, B., Wang, N., Peng, C., Gao, X.: Towards defending against adversarial examples via attack-invariant features. In: International Conference on Machine Learning, pp. 12835–12845. PMLR (2021)
25. Zhu, Y., Miao, C., Hajiaghajani, F., Huai, M., Su, L., Qiao, C.: Adversarial attacks against lidar semantic segmentation in autonomous driving. In: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, pp. 329–342 (2021)