Preventing Adversarial Attacks on Autonomous Driving Models

Junaid Sajid[1], Bareera Anam[1], Hasan Ali Khattak[1(✉)], Asad Waqar Malik[1], Assad Abbas[2], and Samee U. Khan[3]

[1] National University of Sciences and Technology (NUST), Islamabad, Pakistan
{jsajid.msit20seecs,banam.msit20seecs,hasan.alikhattak}@seecs.edu.pk
[2] Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan
assadabbas@comsats.edu.pk
[3] Electrical and Computer Engineering, Mississippi State University, Starkville, USA
skhan@ece.msstate.edu

Abstract. Autonomous driving systems are among the exceptional technological developments of recent times. Such systems gather live information about the vehicle and respond with skilled human drivers' skills. The pervasiveness of computing technologies has also resulted in serious threats to the security and safety of autonomous driving systems. Adversarial attacks are among one the most serious threats to autonomous driving models (ADMs). The purpose of the paper is to determine the behavior of the driving models when confronted with a physical adversarial attack against end-to-end ADMs. We analyze some adversarial attacks and their defense mechanisms for certain autonomous driving models. Five adversarial attacks were applied to three ADMs, and subsequently analyzed the functionality and the effects of these attacks on those ADMs. Afterward, we propose four defense strategies against five adversarial attacks and identify the most resilient defense mechanism against all types of attacks. Support Vector Machine and neural regression were the two machine learning models that were utilized to categorize the challenges for the model's training. The results show that we have achieved 95% accuracy.

Keywords: Autonomous driving models · Autonomous driving system · Adversarial attacks · Support vector machine

1 Introduction

Applications of artificial intelligence (AI) and deep neural networks (DNN) are growing in the field of autonomous driving systems (ADS) as they provide intelligent transportation systems. The field of pervasive computing is developing quickly, especially with ADS. Various prediction and categorization issues are used in several ubiquitous computing applications of deep learning. CNN-based

Z. J. Haas et al. (Eds.): WiCON 2022, LNICST 464, pp. 3–13, 2023.
https://doi.org/10.1007/978-3-031-27041-3_1

regression models can be used to forecast when the car will collide and prevent it from being involved in accidents [1]. Response time and throughput are the main parameters that must be considered in the design of ADMs. The globe presently has a large number of autonomous vehicles. By placing sensors such as camera and LiDAR, autonomous vehicles may gather data that is then fed into systems to execute judgments [2].

Autonomous automobiles also help to enhance the driving skills of a driver [3]. However, they must be tested several times to ensure safety. It even now requires a few critical actions to increase their precision and drive to avoid dangerous collisions. Most of the world is yet to work to install autonomous driving systems on autonomous unmanned aerial vehicles to enable them to take off unassisted by pilots as generic model of autonomous vehicle is shown in Fig. 1. However, to prevent any threat, humans should make sure it is fully established [4]. Deep neural networks have become well-known for their use in autonomous driving functions in intelligent transportation systems. However, safety regulations, as well as comfort and functionality, are essential [5]. Despite their significant contribution to autonomous vehicle visual perception, DNNs are easily fooled by adversarial attacks [6]. These attacks can alter the prediction of autonomous cars by bringing minor changes in the pixels of the images. In this way, neural networks fail by making incorrect predictions [7].

Many adversarial attacks have been studied, and their defense methods to improve adversarial robustness [8]. Some adversarial strategies have already been put forth and shown to be successful against image classifications [9,10] and several mechanisms to strengthen neural networks have already been suggested to protect against adversarial threats [11]. But prior studies have mostly concentrated on image classifications. The effectiveness of such adversarial attacks and responses against regression models is unknown, such as ADMs. These ambiguities highlight possible privacy threats and expand the field of study. Intruders may potentially trigger road collisions and endanger human protection if adversarial assaults on ADS are effective. It is essential to propose a defensive technique appropriate for ADMs if present defensive strategies cannot be modified to protect against attacks on regression models.

In this paper, some adversarial attacks are analyzed given in Sect. 3 in this investigation. Those adversarial attacks were defended using four strategies. These tests employ Nvidia DAVE-2, Epoch, and VGG16 as the ADMs. The Nvidia DAVE-2 ADMs for autonomous cars are well-known in the community, Additionally, in the Udacity dataset, Epoch is seen to behave effectively, and VGG16 employs the extremely reliable framework of neural networks that are widely utilized in transfer learning for image categorization. Neither of the four defense strategies can successfully stop all five types of attacks. According to adversarial training and defensive distillation, only the IT-FGSM and an optimization-based strategy are capable of lowering the effectiveness of the attacks. Those two attacks, as well as AdvGan, can be successfully detected using a mechanism that recognizes unusual device status. Considering appropriate conditions, feature squeezing can identify all five attacks with significantly

higher than 78% accuracy, although it has a significant false positive rate. This suggests that to develop a strong defensive mechanism against different threats, a system or middleware must employ many security techniques simultaneously.

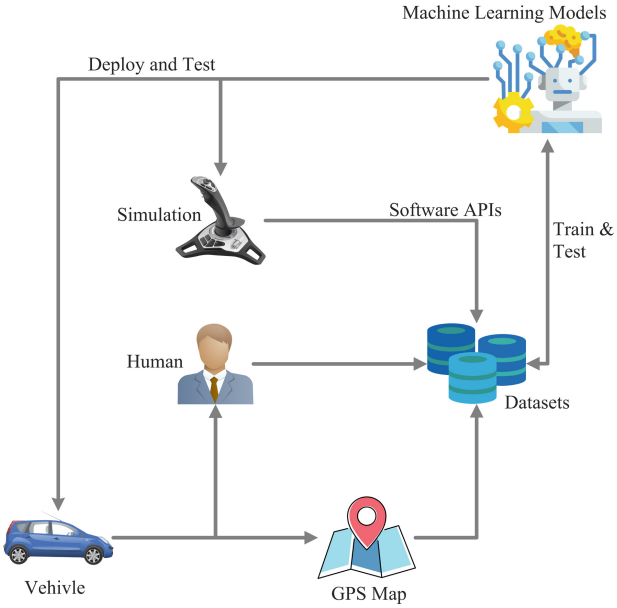


Fig. 1. Generic autonomous driving model based on machine learning

Structure of the Paper: The introduction of this paper is presented in Sect. 1. The rest of the paper is organized as follows: Literature Review is covered in Sect. 2. The methodology is represented in Sect. 3. The Sect. 4 provides the results, and the paper is concluded in Sect. 5.

2 Literature Review

2.1 Autonomous Driving Systems

Autonomous vehicles, which are developed on various ADMs, represent one of the finest competent and fascinating examples of innovation that has been rapidly evolving around the globe. The very first autonomous driving competition was organized in 2005, which goes by the name of DARPA [12] While following, the topic has experienced tremendous expansion, and several ADS proposals have been made. One of the essential elements that support this reason is DNN-based LiDAR-based item identification [13,14]. Additionally, TESLA uses a variety of deep neural networks and learning models for its AutoPilot capabilities. It goes without saying that every technique that incorporates a learning model is going to provide various flaws and be vulnerable to adversarial assaults [11]. This typically happens whenever the initial forecast differs from reality and the actual prediction that exceeds the adversarial threshold.

2.2 Adversarial Attacks on ADS

In [15], an examination of common autonomous vehicle weaknesses is provided. The study in [16] states that adversarial attacks might potentially be successful against every legitimate input or successful against a certain parameter alone. To execute these assaults, there are primarily two types of approaches: those depending upon optimizing methods and those dependent upon the fast gradient step method. Both types of attacks discussed above may be conducted using optimization-based techniques. L-BFGS, which was suggested by the authors of [17], is one illustration. The Fast gradient step method and its expansions, such as iterative-based-FGSM, and momentum-based iterative-FGSM, are those techniques [18].

The patch attack only impacts a limited percentage of image dots. The variations in image pixels are considerably bigger than those caused by the perturbations assault, which often impacts most of the pixels inside of the source picture. These assaults examined some operational components required for vision-based ADMs. For instance, the patched attack has been investigated concerning lane changing in [19], and 2D item identification in [20]. The other attacks were studied, considering symbol categorization in [21] and 2D item identification in [22]. Neither of the aforementioned research specifically examines how the assaults affect ADS safety and driving patterns.

2.3 Adversarial Defences

In general, preemptive defense and responsive defense are two categories of the various strategies that have been put forth to counter adversarial assaults. The goal of preemptive defenses is to increase the neural network's durability versus adversarial examples. The typical approach is to add batch normalization elements into the baseline models [23] or train the system employing datasets containing adversarial cases [11]. Defensive distillation was also presented by the authors in [24]. This hardens a neural network by increasing the size of signals to the soft max by altering a factor termed temperature. Conversely, responsive approaches look for adversarial cases. By examining the characteristics of the incoming pictures [25] and monitoring the state of the soft max [26], DNN might be utilized to ascertain if the network is being attacked. In this paper, the goal is to analyze some possible adversarial attacks versus some different CNN-based ADMs, present four countermeasures to counteract such attacks, and then present the best countermeasure versus these five attacks.

3 Proposed Methodology

3.1 Adversarial Attack

In image classification models, adversarial assaults are extremely effective. While this differs beyond the underlying model's forecast, perhaps we may conclude

that adversarial attacks are effective and productive in undermining the models. There are three different forms of adversarial attacks that are based on the perturbation generation approach. The Fast Gradient Sign-Based technique (FGSM) executes assaults by applying the signature of the loss gradient to every pixel inside this baseline picture. Confrontational instances are produced using an optimization-based strategy. Using compositional models' capabilities, generative model-based methodology creates adversarial instances. There are some distinct adversarial attacks known as universal attacks. This produces a solitary adversarial case to screw up the whole database or every sample of the data.

ITFGSM. IT-FGSM [27] is an adversarial attack, and a special kind of FGSM. In this type of attack, every pixel of the source picture may have an additional signature added to it to create adversarial attacks, and as the term implies, the adversarial attack is strengthened by applying the targeted fast gradient sign method more than once.

Opt. This type of attack [28] is yet an additional technique for producing adversarial assaults. By using the equation underneath to solve the optimization issue, it determines the adversarial perturbation ϵ.

$$argmin_{\epsilon}||\epsilon||_2 \, s.t. \, f(x+\epsilon) = b \; , x + \epsilon \, \epsilon[0,1]^m \tag{1}$$

In this analysis of, we modify the b to $f(x) + \delta$ and both the aforementioned equation and the following equation.

$$argmin_{\epsilon}||\epsilon||_2 \; + Z_{\theta} \left(Clips(x+\epsilon), f(x) + \delta\right) \tag{2}$$

AdvGAN. AdvGAN [9] provides the adversarial instances by implementing other objective $J_y = Z_{\theta} \left(G(x), f(x) + \delta\right)$ through initial picture into the objective function $J_{AdvGAN} = J_y + \alpha J_{GAN}$, where α is the thing which determines how important each target seems to be. Once learning is finished, the produced G can provide new adversarial attack examples that resemble our initial picture however can also update the forecast or produce forecasting which deviates by *delta* from $f(x)$.

Opt-uni. An Opt-uni [29] attack is a type of optimization-based approach.For that kind of attack, we disrupt a picture initially, then each picture in the database individually at a time. By converting v to $v + \delta v$, we determine that the least difference is δv. The general disturbance was obtained by repeatedly performing this assault across the entire data set.

AdvGAN-uni. This is the variant of the Advance GAN method [9]. Rather than creating the distinct perturbations $G(x)$ per each picture within the data as is the case in that sort of attack, we suggested utilizing GAN to produce the general perturbations. An overall disturbance is produced by the generators.

Table 1. Methodology and their capabilities

Method	Comments
Regression model	This model is weak to adversarial attacks
Classification model	This model is highly successful against adversarial attacks
Adversarial training	In contrast to other sorts of attacks, adversarial training is particularly successful against ITFGSM and opt
Defensive distillation	This defensive technique works well versus ITFGSM and certain similar types of attacks, although not versus others
Feature squeezing	This method is effective in detecting an adversarial attack. However, it is not well enough in defense
Anomaly detection	Although this defensive technique can identify adversarial attacks, and also it has several drawbacks by itself

3.2 Adversarial Defenses

As we know, each model has a defense strategy for all kinds of attacks. To defend against adversarial attacks on ADMs, we incorporated four defenses in this investigation as given in Table 1. All four techniques are given below:

Adversarial Training. In order to defend against this, we first create adversarial instances and attempt to relearn the existing system using hostile instances. By using this strategy, the models would understand the characteristics of adversarial cases, resulting in resilient and universally applicable models.

Defensive Distillation. The defensive distillation [24] type of approach uses the original model's prediction of class probabilities to train a new model.

Anomaly Detection. Continually checking the conditions of the automobile and ADMs is a live supervision mechanism found in autonomous driving systems. In order to identify possible additional deaths brought on by adversarial attacks, we initially evaluate the model's forecasting after they are applied. We next utilize NvidiaSMI to analyze the processor and memory usage. This anomaly detection method analyzes GPU utilization and forecasting times for each picture, either using adversarial instances or without them.

Feature Squeezing. This methods have two types of techniques [30] in order to facilitate defense versus adversarial attacks. A picture's native 24 bit color is reduced in the first technique to 1 bit or 8 bit color. While the bit quality declines, one could more easily foresee an adversarial attack. By our next technique, median temporal flattening, we move the strainer to change the main pixel number to the average of the pixel number. Thus, one may state that there is a chance of an adversarial attack if there is a gap between the forecast of the source picture and the estimation of the compressed picture that surpasses the cutoff value.

3.3 Neural Regression

The Keras library supports a neural network regression model to provide the solution to the regression problems. We needed to determine or forecast the accuracy of the trained models using a few dependent variables.

3.4 SVM Classification

A support vector machine classification is just the model's discrete observations coordinated together. It belongs to the class of supervised machine learning, which is utilized for issues involving the categorization of distinct categories. It has been employed in this study since we needed to ensure whether the methodology backed training models.

4 Results

A thorough examination of adversarial attacks and countermeasures against ADMs To achieve this, we applied four defensive techniques and five adversarial attacks to three convolutional neural network-based driving models. According to the study's findings, all three driving models aside from IT-FGSM are not resistant to these adversarial attacks, whereas neither of four defensive strategies can counteract all five of them. Following a thorough analyzation of every attack, we discovered that CNN based training is particularly susceptible towards the attack we mentioned in the article because they have a strong achievement percentage when used against regressors. As a result, they are not appropriate for use when training against such attacks.

	precision	recall	f1-score	support
0	0.97	0.93	0.95	1074
1	0.82	0.93	0.87	354
accuracy			0.93	1428
macro avg	0.90	0.93	0.91	1428
weighted avg	0.94	0.93	0.93	1428

Fig. 2. Accuracy of the trained model

Models for classifying attacks defend well against adversarial attacks. The outcomes also demonstrate that defensive distillation and adversarial training are only somewhat successful against IT-FGSM and opt attacks, however, hardly versus alternative forms of attacks. Feature squeezing and defensive distillation, additional major defense strategies, are also capable of spotting certain attacks, although each has its particular drawbacks and restrictions.

4.1 Accuracy and Prediction Error

In the Fig. 3, two images have been shown. On the left image the results of the training model are shown when the models are being trained. It is observed that at start the error rate was very high, as the model continue training, the error rate are reduced down linearly. In the right image, the total performance of the 1428 samples have been discussed. It is observed that, more that 95% samples did not deviates from the attacks when defensive techniques are being used as shown in the Fig. 3.

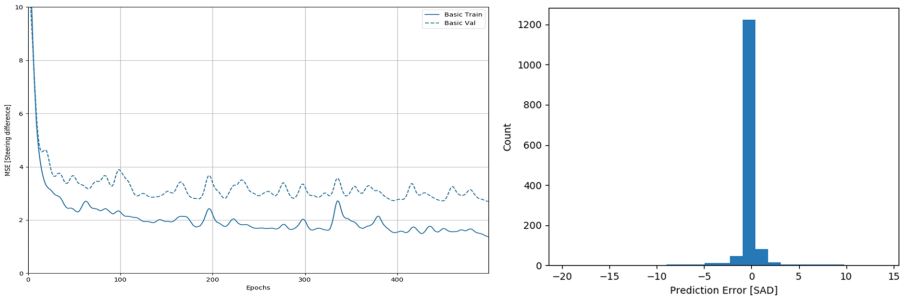


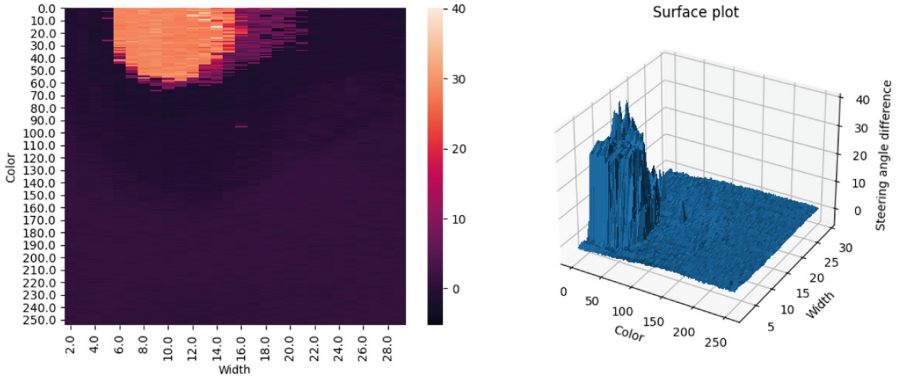
Fig. 3. The results of the simulation with 1428 samples

4.2 SVM Classification

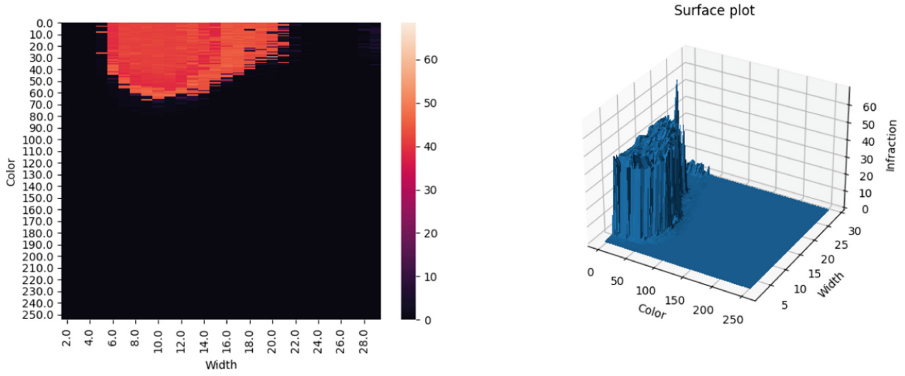
The results of the SVM classification on the testing samples are given in Fig. 2. There are two classifications in this diagram; '0' and '1'. The normal behavior of the ADMs is categorized as '0', and the other behaviors are categorized as '1' where the infraction of the vehicles is greater than 0. Precision, recall, f1-score, and support are used as measurement metrics. The support for 0 is 1074 and the support for 1 is 354. This means that we have 75% support for autonomous models. The precision score was 0.97, the recall of our model was 0.93, and the f1-score was 0.95. That means that, collectively, the performance of the classification model was 95.7%.

4.3 Perturbations and Decision Boundary

In Fig. 4 two results have been shown: steering angle difference and infraction of the vehicle. We used two types of graphs to find the difference and infraction in the form of decision boundary and perturbation. Color is given on the Y-axis, width is given on the X-axis, steering angle difference, and infraction is on the Z-axis in the three-dimensional graph. On the other side of the image, few color combinations show different types of results. The dark color shows that there is no attack on the surface, and the lighter color shows the region where adversarial attacks happen. The color that is in the middle acts as the decision boundary between the two regions where adversarial attacks happen and the safe region.



(a) Steering angle difference



(b) Infraction of the vehicle

Fig. 4. Difference of Steering Angle and Infraction of the Vehicle

5 Conclusion

In this research, we examine several of the adversarial attacks against ADMs as well as defense methods to identify and thwart these assaults. We employed neural regression and support vector machines as our two models for training samples. As we can observe, classifiers are far better adapted to adversarial assaults than regressors. Neither of the four defense strategies can successfully stop all five types of attacks. According to the adversarial training and defensive distillation, only IT-FGSM and optimization-based strategy are capable of lowering the effectiveness of the attacks. Those two attacks, as well as AdvGan, can be successfully detected using a mechanism that recognizes unusual device status. This suggests that in order to develop a defensive mechanism that is strong against different threats, a system or middleware must employ many security techniques simultaneously. The findings reveal that, out of 1428 testing examples, 95.7% of the specimens in SVM demonstrate that there is no divergence beyond +0.5 and −0.5.

References

1. Mozaffari, S., Al-Jarrah, O.Y., Dianati, M., Jennings, P., Mouzakitis, A.: Deep learning-based vehicle behavior prediction for autonomous driving applications: a review. IEEE Trans. Intell. Transp. Syst. **23**(1), 33–47 (2020)
2. Singh, S., Saini, B.S.: Autonomous cars: recent developments, challenges, and possible solutions. In: IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012028. IOP Publishing (2021)
3. Raja, F.Z., Khattak, H.A., Aloqaily, M., Hussain, R.: Carpooling in connected and autonomous vehicles: current solutions and future directions. ACM Comput. Surv. **1**(1), 1–35 (2021)
4. Ni, R., Leung, J.: Safety and liability of autonomous vehicle technologies. Massachusetts Institute (2014)
5. Yuan, T., da Neto, W.R., Rothenberg, C.E., Obraczka, K., Barakat, C., Turletti, T.: Machine learning for next-generation intelligent transportation systems: a survey. Trans. Emerg. Telecommun. Technol. **33**(4), e4427 (2022)
6. Malik, S., Khattak, H.A., Ameer, Z., Shoaib, U., Rauf, H.T., Song, H.: Proactive scheduling and resource management for connected autonomous vehicles: a data science perspective. IEEE Sens. J. **21**(22), 25151–25160 (2021)
7. Kyrkou, C., et al.: Towards artificial-intelligence-based cybersecurity for robustifying automated driving systems against camera sensor attacks. In: 2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 476–481. IEEE (2020)
8. Deng, Y., Zheng, X., Zhang, T., Chen, C., Lou, G., Kim, M.: An analysis of adversarial attacks and defenses on autonomous driving models. In: 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 1–10. IEEE (2020)
9. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4422–4431 (2018)
10. Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610 (2018)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
12. Buehler, M., Iagnemma, K., Singh, S.: The 2005 DARPA Grand Challenge: The Great Robot Race, vol. 36. Springer, Cham (2007)
13. Cao, Y., et al.: Adversarial objects against lidar-based autonomous driving systems. arXiv preprint arXiv:1907.05418 (2019)
14. Arnold, E., Al-Jarrah, O.Y., Dianati, M., Fallah, S., Oxtoby, D., Mouzakitis, A.: A survey on 3D object detection methods for autonomous driving applications. IEEE Trans. Intell. Transp. Syst. **20**(10), 3782–3795 (2019)
15. Ren, K., Wang, Q., Wang, C., Qin, Z., Lin, X.: The security of autonomous driving: threats, defenses, and future directions. Proc. IEEE **108**(2), 357–372 (2019)
16. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: attacks and defenses for deep learning. IEEE Trans. Neural Netw. Learn. Syst. **30**(9), 2805–2824 (2019)
17. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 23296–23308 (2021)

18. Ren, H., Huang, T., Yan, H.: Adversarial examples: attacks and defenses in the physical world. Int. J. Mach. Learn. Cybern. **12**(11), 3325–3336 (2021). https://doi.org/10.1007/s13042-020-01242-z
19. Zhou, H., et al.: Deepbillboard: systematic physical-world testing of autonomous driving systems. In: 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE), pp. 347–358. IEEE (2020)
20. Song, D.,et al.: Physical adversarial examples for object detectors. In: 12th USENIX Workshop on Offensive Technologies (WOOT 18) (2018)
21. Feng, R., Chen, J., Fernandes, E., Jha, S., Prakash, A.: Robust physical hard-label attacks on deep learning visual classification. arXiv preprint arXiv:2002.07088 (2020)
22. Lu, J., Sibai, H., Fabry, E.: Adversarial examples that fool detectors. arXiv preprint arXiv:1712.02494 (2017)
23. Yan, Z., Guo, Y., Zhang, C.: Deep defense: Training DNNs with improved adversarial robustness. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
24. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy (SP), pp. 582–597. IEEE (2016)
25. Zheng, Z., Hong, P.: Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
26. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
27. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
28. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
29. Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1765–1773 (2017)
30. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)