SENSOR ADVERSARIAL TRAITS: ANALYZING ROBUSTNESS OF 3D OBJECT DETECTION SENSOR FUSION MODELS

Won Park[⋆] *Nan Liu*[⋆] *Qi Alfred Chen*[†] *Z. Morley Mao*[⋆]

[⋆] University of Michigan
[†] UC Irvine

ABSTRACT

A critical aspect of autonomous vehicles (AVs) is the object detection stage, which is increasingly being performed with *sensor fusion models*: multimodal 3D object detection models which utilize both 2D RGB image data and 3D data from a LIDAR sensor as inputs. In this work, we perform the first study to analyze the robustness of a high-performance, open source sensor fusion model architecture towards adversarial attacks and challenge the popular belief that the use of additional sensors automatically mitigate the risk of adversarial attacks. We find that despite the use of a LIDAR sensor, the model is vulnerable to our purposefully crafted image-based adversarial attacks including disappearance, universal patch, and spoofing. After identifying the underlying reason, we explore some potential defenses and provide some recommendations for improved sensor fusion models.

***Index Terms*—** Adversarial examples, multimodal, 3D object detection, sensor fusion

1. INTRODUCTION

Autonomous vehicle (AV) manufacturers often use *sensor fusion* models to help vehicles detect the environment around them. These types of models are multimodal 3D object detection models that take in two types of inputs: a 2D image from a camera and 3D depth data usually from a LIDAR sensor. With the growing proliferation of autonomous vehicles, their security is becoming more paramount, especially against adversarial examples [1, 2].

It has long been known in the community that machine learning models are vulnerable to adversarial examples, maliciously crafted inputs designed to intentionally fool the model into outputting an erroneous result. These range from attacks in the raw pixel space [3] to launching these attacks in the physical world [4, 1, 5, 6].

There is a belief that the use of additional inputs can mitigate the effect of adversarial examples. While recent work [7] has shown theoretically that models that take in multiple inputs are vulnerable to potential perturbations in a single input, no one has actively explored the robustness and crafted adversarial examples against sensor fusion models. Our work is the first to demonstrate the insecurity of sensor fusion models to several realistic adversarial attacks for 3D object detection.

Though there are many multimodal 3D detection architectures available, we focus this study on models that take in the two types of inputs simultaneously. We purposefully choose to ignore model architectures such as Frustum-Pointnet [8] that utilize a "pipeline" structure in which image data is taken in first followed by LIDAR data because these models are trivially vulnerable to image-based attacks — any existing attack algorithm to fool 2D image object detectors will be able to fool the entire model. Instead, for this work, we choose AVOD [9], an open-source 3D object detection model, because of its near-top performances among open-source models in the KITTI benchmark. Note that this model differs from a prior work [10] that does not evaluate attacks on a properly created sensor fusion model whose architecture is conducive to 3D multimodal object detection. Instead, they simply combine existing architectures - a LIDAR featurizer with a YOLO model, for example. The difference between the architectures is made apparent in the AP scores reported: AVOD, studied here, reports 71.88 while the paper's architecture has a score of 60.3. In short, our attacks are more meaningful because we evaluate and attack a model with a higher baseline accuracy. Secondly, whereas the aforementioned work only explores one attack, we are able to develop a wide variety of attacks and delve deeper into the nature of sensor fusion models.

We are interested in understanding if the use of an additional input prevents adversarial attacks on the other input. Though we could utilize adversarial attacks on the LIDAR input like previous work [11, 2, 12, 13], we instead choose to focus on modifying the image input. This is because we find that the model relies more heavily on LIDAR data and successful attacks using modification of just the LIDAR is more trivial. Furthermore, physical attacks on the camera detector are more realistic and potent than the ones on the LIDAR sensor. Thus, by restricting our attacks to just images, we are assuming a less powerful and more realistic attacker. Because this is the first foray into this field, we assume that the adversary is a white-box attacker, having full access to the model. Despite this, in order to guide future research works, we aim to be as realistic as possible; this includes restrictions that the adversary will not be able to modify the model arbitrarily, in-

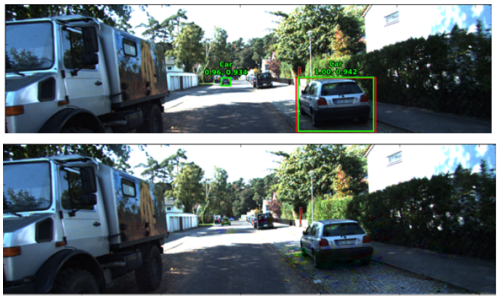

Fig. 1. Results of some of our disappearance attacks. Top is benign images and bottom is adversarial images. The 1st value corresponds to the classification confidence and the 2nd value corresponds to the IOU with the ground truth bounding box. The red boxes are ground truth and the green boxes are bounding boxes outputted by the model.

cluding any post-processing steps.

Our key contributions are as follows:

- We perform the first study of adversarial examples on proper sensor fusion models for 3D object detection. We modify existing techniques to show that sensor fusion models are vulnerable to adversarial attacks that modify just the image input. These attacks include the *raw pixel disappearance attack* (94.17% success rate) and a spoofing attack (89.1%). We then analyze the model architecture to show that despite the symmetric architecture, the model frequently leans heavily on the LIDAR input to detect obstacles.
- Building upon the raw pixel disappearance attack, we develop a new methodology of constructing generalized adversarial examples in which one single noise can fool many samples.
- We explore some basic defenses, including robust training and a novel fusion layer [7]. We comment on their effectiveness and put forth suggestions for future directions.

2. CRAFTING ATTACKS

In this section, we explore attacks that are able to fool the model into not detecting an object it had previously detected (*raw-pixel disappearance attacks*) and those that fool the model into detecting an object that is not actually present (*spoofing attack*).

The *raw-pixel disappearance attack* is motivated by a desire to create a disappearance attack that results in an adversarial example that is less noticeable to the human eye. We explore a different kind of attack - patch attacks - in Section 3. To cause a desired object to disappear, we aim to force the output softmax probabilities of all potential bounding boxes around the said object below the detection threshold. We will call this set of all potential boxes that we need to attack B.

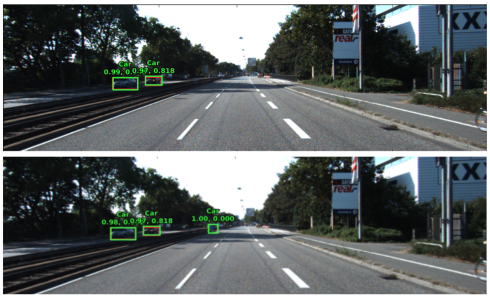

Fig. 2. Results of some of our spoofing attacks. Top is benign and bottom is adversarial. bounding boxes with an IOU of 0 (second value) are the spoofed objects.

For ease of notation, suppose $C(w, b) \in R^c$ denotes the output classification softmax of bounding box b on image w and $C(w)$ outputs all the potential bounding boxes of image w in decreasing order according to softmax score. We attempt to find an adversarial sample δ that minimizes the following function:

$$L(w + \delta, B) = \sum_{b \in B} [C(w + \delta, b)] + \epsilon * D(w + \delta, w) \quad (1)$$

The second element $D()$, which measures the L_2 norm between the adversarial image and the regular image, is added to make the perturbation to the image as small as possible, as suggested by the CW attack [14]. The optimal value of the weighting coefficient ϵ is found through binary search. Unlike previous work, we choose to attack the second-stage detector (instead of the first stage RPN) as it results in an attack with less distortion.

However, due to NMS and the restrictions we set on the adversary, not all of the elements of B will be visible and we are unable to know the set B a priori. For some related previous work, this was not a huge limiting factor [15] while others went around this by trying to attack the NMS algorithm itself or to modify the NMS threshold to obtain all bounding boxes [3].

Since neither solution is allowed under our threat model, we instead modify the algorithm to greedily attack the top confidence bounding box that is visible - as we keep trying to lower the confidence of the bounding box with the highest score, one of two outcomes will happen. In one, the object in question will no longer be detected, in which case our attack goal is accomplished. In the other case, the bounding box in question will be removed via NMS and the next top-score bounding box will appear and the process can be repeated. This process will remove all objects present in an image, but an adversary can selectively remove certain objects by applying a mask. In this case, the objective function needs to be modified to attack the top k bounding boxes simultaneously.

For the spoofing attack, in addition to switching the sign of the loss function, we add another element to our adversarial objective function that will help ensure the desired bound-

ing box is outputted by the RPN. The final loss function is a weighting of the two losses where the RPN loss function is weighted much more heavily:

$$L_{spoof} = L_{RPN} + \alpha * L_{Stage2} \quad (2)$$

It is important to note that a simple defense against spoofing attacks is to remove any anchor boxes that do not have any LIDAR data as a pre-processing step before inputting into the model. Acknowledging this, we run all of our experiments under this assumption to increase the likelihood that our results cannot be defeated by this simple defense.

2.1. Evaluation and Results

When training AVOD, we follow the methodology followed in the paper proposed by Chen et al [16] on the KITTI dataset [17], We split the trainval set into a training set with 3712 samples and a validation set with 3769 samples. We train all models to closely match the results stated in the original paper.

For our experiment, we choose 3000 random samples containing a total of 10,920 detected vehicles and 7,585 pedestrians / cyclists and try constructing adversarial examples using the method stated above. We are able to achieve a 94.17% success rate on vehicles and 97.11% on pedestrians and cyclists. Some of the attacks are shown in Figure 1. For the spoofing attack, we are able to achieve a 89.1% and 91.33%, respectively, upon evaluating on these samples 2.

3. TOWARDS GENERALIZABILITY

In this section, we attempt to create a single adversarial patch that, despite being more noticeable to the human eye, would be able to be universally applied to any vehicleand cause them to escape detection from the model. This is a key step in determining the feasibility of attacks in the physical world.

To start, we draw inspiration from the expectation over transformation (EOT) algorithm [18]. Due to the difficulty of applying any transformation to an image and also properly modify the corresponding LIDAR data, however, we use different object samples available in KITTI instead as input. For each image, we identify an area on the vehicle we wish to apply a patch. Note that if an image has multiple objects, we may have to apply the patch separately to different areas. Let $P(w, \delta)$ be the operation that applies adversarial patch δ to image w, appropriately resizing the patch as necessary. If we have a set of images T (along with their corresponding bounding box set), we would be able to create a universal patch by solving the following objective function:

$$\text{argmin}_{\delta} \mathop{\mathbb{E}}_{w \in T}[L(P(w, \delta))] \quad (3)$$

Normally, this would be done simultaneously via batching. However, since AVOD and many other sensor fusion models do not support batching, we alter the algorithm: instead of operating over all the images simultaneously, we perform the objective function on one image at a time, keeping the noise in between images and iterate over all the images trying to ensure convergence. The number of times to iterate is a hyperparameter that must be tuned but for our experiments we iterated 25 times. This is a similar approach as suggested by [19], however, we do not project all perturbations onto a p-norm since we find it slows our algorithm due to the nature of our loss function trying to target every potential bounding box. The resizing and the nature of multiple bounding boxes are also reasons why the GD-UAP[20] is less than ideal for this work. Therefore, we take advantage of updating ϵ to control the distortion. We find that gradually decreasing the value until a certain floor value works well. The algorithm can be viewed in 1.

Algorithm 1: Modified EOT

input : Set of images T, k, n, ϵ
output: Adversarial noise δ
begin
  $\delta \longleftarrow$ RandomInit;
  for $i = 0$ **to** n **do**
    $w \leftarrow$ NextImage(T, i);
    $B' \leftarrow C(P(w, \delta))[0...k]$;
    $\delta \leftarrow argmin_{\delta} L(P(w, \delta), B', \epsilon)$
    $\epsilon \leftarrow$ UpdateEpsilon(i)
  end
  return δ
end

3.1. Results

We run this algorithm on the validation set and are able to achieve a success rate of 64.03%. To establish a baseline comparison, we apply a random noise patch to the same vehicles. These random noise patches, when applied to the vehicle in the same location, achieve a 0% success rate. The results of this case study suggests a worrisome fact that sensor fusion models are still vulnerable to universal physical adversarial examples, similar to what is shown in Huang et al [5].

4. ANALYSIS OF SENSOR INPUT

Motivated by the results of our experiments on various attacks, we suspect that the model architecture, while symmetrical, heavily utilizes the LIDAR sensor input over the image. To test this, we run an experiment in which we use the LIDAR from one sample and the image for another to understand how the model performs when the image and the LIDAR are at odds with each other. This was done for 600 random samples, swapping the image of one and the LIDAR of another, resulting in 360,000 combinations. For the sake of simplicity,

Type	Disappearance	Spoof
Baseline	0.94	0.89
Distorted Inputs	0.92	0.85
MaxSSN [7]	0.87	0.81
MaxSSN + LEL [7]	0.80	0.39
Adversarial Training	0.63	0.51

Table 1. Table showing the success rate (in %) of our attacks against various defenses.

we consider an object as "correctly identified" if the bounding box was correctly drawn according to the LIDAR sensor [1]. Amongst all the potential bounding boxes, 91% were correctly identified, despite having conflicting image data. Furthermore, only 19% of all bounding boxes detected did not correspond to any ground truth bounding box.

This experiment strongly suggests that the model favors LIDAR data when detecting objects, which helps explain the difficulty in the spoofing attacks. For instance, when we compare the L2 norm per pixel, we find that the spoofing attack requires much more distortion: while the disappearance attack required a median per-pixel distortion of roughly 0.28 in L2 space, the spoofing attack required a L2 distortion of over 1.3 in L2 space. This is line with a previous work that found another sensor fusion model, MV3D, also favors LIDAR [2]. While this is contrary to what was found in Wang et al [10], we believe this is because their lack of a true sensor fusion model built from the ground up. This also suggests that the use of image in this architecture proves to be an "Achilles' heel": while most of the detection of an object is done using the LIDAR input, it is not sufficient, as the image provides a way for adversaries to override this and attack the model.

5. EXPLORING DEFENSES

In this section we analyze some potential defenses against adversarial examples on sensor fusion models. For each defense, we test our raw-pixel disappearance attack and our spoofing attack on the same vehicle samples as in the evaluation for our attacks. All results can be seen in Table 1.

We first attempt to address a belief that training on a wider range of inputs (like fog and snow) will help mitigate adversarial examples. We train an instance of AVOD on an augmented version of the dataset in which we apply all distortions as suggested by Hendryks and Dietterich [21] to every training image. While training on various input distortions may be important for safe performance of AVs, we find that it does not eliminate the threat against adversarial examples.

We next analyze methodologies proposed by Kim and Ghosh [7] that provide robustness against single-source distortion in sensor fusion models. We refer readers to the paper for details, but in short, the authors propose novel loss functions and a new fusion layer called LEL to protect against noisy distortion. We test the two designs proposed that achieved the best performance under noise: using the new loss function called MaxSSN with and without LEL. We train both models according to the specifications shown in the paper and achieve a similarly stated accuracy. We find that the models do mitigate against our disappearance attacks, but only to a success rate of around 80%. However, the addition of the LEL does better in defending against spoofing attacks. Unfortunately, it is worth noting that there exists a trade-off as both models suffer in a drop of AP score compared to the original model when run on benign inputs.

A popular methodology to protect against adversarial examples is adversarial training. We utilize a preliminary technique proposed by Zhang and Wang [22]. We find that the AP score drops after adversarial training, but so does the success rate of the adversarial attacks drop as well; the raw-pixel disappearance attack drops to a 63% effectiveness while the spoofing attack drops to a 51% success rate. While these results are not ideal, they do not necessarily eliminate adversarial training as a viable option to provide robustness. On the contrary, they demonstrate that a better adversarial training algorithm may be able to provide robustness. We leave this exploration to future work. We also believe another avenue to explore is incorporating defenses *outside* the model. In a larger system, it could be feasible to utilize the different sensors, for example, to validate one another before feeding into the final detection model. We leave these possible defenses for future work to explore.

6. CONCLUSION

In this paper we explore a fusion model's security against adversarial examples. We discover that the use of a secondary input provides limited defense against a myriad of different adversarial attacks. Though we only evaluate on one model due to the lack of availability of open-source models, our evaluation on alternative fusion layers and training loss functions suggest that other models may also be vulnerable to single image attacks. We urge future works on sensor fusion models to help increase robustness.

7. ACKNOWLEDGEMENTS

We thank our reviewers for their comments. We also wish to thank Fahad Kamran, Jiachen Sun, Yulong Cao for their help and insights. This project is partially supported by Mcity and NSF under the grants CMMI-2038215, CNS-1930041, CCF-1628991, CNS-1544678, CNS-1850533, CNS-1929771, and CNS-1932464.

[1]Note that we could have easily swapped and used the image input as "ground truth" but that would not make any change to the final result

8. REFERENCES

[1] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, 2019, CCS '19, p. 1989–2004, Association for Computing Machinery.

[2] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z. Morley Mao, "Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures," in *29th USENIX Security Symposium (USENIX Security 20)*. Aug. 2020, pp. 877–894, USENIX Association.

[3] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille, "Adversarial examples for semantic segmentation and object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, Tadayoshi Kohno, and Dawn Song, "Physical adversarial examples for object detectors," in *Proceedings of the 12th USENIX Conference on Offensive Technologies*, USA, 2018, WOOT'18, p. 1, USENIX Association.

[5] Lifeng Huang, Chengying Gao, Yuyin Zhou, Changqing Zou, Cihang Xie, Alan Yuille, and Ning Liu, "Upc: Learning universal physical camouflage attacks on object detectors," 2019.

[6] S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 49–55.

[7] Taewan Kim and Joydeep Ghosh, "On single source robustness in deep fusion models," in *NeurIPS*, 2019.

[8] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Lake Waslander, "Joint 3d proposal generation and object detection from view aggregation," *CoRR*, vol. abs/1712.02294, 2017.

[10] Shaojie Wang, Tong Wu, and Yevgeniy Vorobeychik, "Towards robust sensor fusion in visual perception," 2020.

[11] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Zhuoqing Morley Mao, "Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving," in *Proceedings of the 26th ACM Conference on Computer and Communications Security (CCS'19)*, London, UK, November 2019.

[12] Yuxin Wen, Jiehong Lin, Ke Chen, and Kui Jia, "Geometry-aware generation of adversarial and cooperative point clouds," *CoRR*, vol. abs/1912.11171, 2019.

[13] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun, "Physically realizable adversarial examples for lidar object detection," in *IEEE CVPR*, June 2020.

[14] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017.

[15] Yi Huang, Adams Wai Kin Kong, and Kwok-Yan Lam, "Adversarial signboard against object detector," in *British Machine Vision Conference*, 2019.

[16] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia, "Multi-view 3d object detection network for autonomous driving," in *IEEE CVPR*, 2017.

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[18] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, "Synthesizing robust adversarial examples," 2017.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and P. Frossard, "Universal adversarial perturbations," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 86–94, 2017.

[20] Konda Reddy Mopuri, Aditya Ganeshan, and R. Babu, "Generalizable data-free objective for crafting universal adversarial perturbations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, 01 2018.

[21] Dan Hendrycks and Thomas Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[22] H. Zhang and J. Wang, "Towards adversarially robust object detection," in *ICCV*, 2019, pp. 421–430.