Adversarial Robustness of Deep Sensor Fusion Models

Shaojie Wang, Tong Wu, Ayan Chakrabarti, Yevgeniy Vorobeychik
Washington University in St. Louis
{joss, tongwu}@wustl.edu, ayan.chakrabarti@gmail.com, yvorobeychik@wustl.edu

Abstract

We experimentally study the robustness of deep camera-LiDAR fusion architectures for 2D object detection in autonomous driving. First, we find that the fusion model is usually both more accurate, and more robust against single-source attacks than single-sensor deep neural networks. Furthermore, we show that without adversarial training, early fusion is more robust than late fusion, whereas the two perform similarly after adversarial training. However, we note that single-channel adversarial training of deep fusion is often detrimental even to robustness. Moreover, we observe cross-channel externalities, where single-channel adversarial training reduces robustness to attacks on the other channel. Additionally, we observe that the choice of adversarial model in adversarial training is critical: using attacks restricted to cars' bounding boxes is more effective in adversarial training and exhibits less significant cross-channel externalities. Finally, we find that joint-channel adversarial training helps mitigate many of the issues above, but does not significantly boost adversarial robustness.

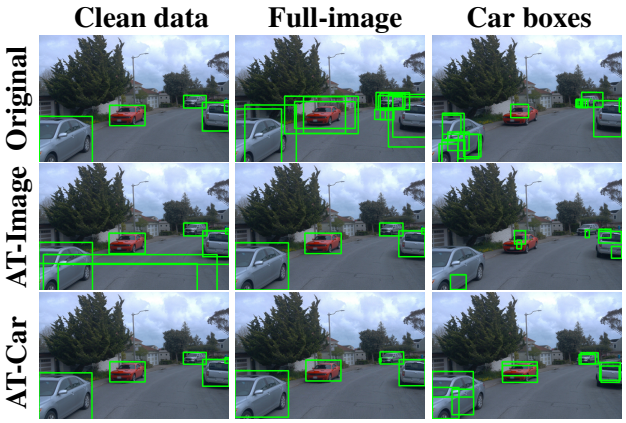


Figure 1: Visualization of late fusion 2D detection results. Top row: original late fusion model. Middle row: late fusion after adversarial training with full-image attacks. Bottom row: late fusion after adversarial training with attacks restricted to cars' bounding boxes. Left column: clean data. Middle column: full-image attacks. Right column: car bounding box attacks. Note the image in middle row, left column, which demonstrates "adversarial overfitting", that is, far more conservative bounding box generation after full-image adversarial training.

1. Introduction

Autonomous driving (AD) depends heavily on visual perception, such as camera, radar, and LiDAR. While deep neural network architectures have been transformative in improving automated perceptual efficacy in tasks such as classification and object detection, a series of demonstrations have shown that these perceptual frameworks are also vulnerable to small adversarial perturbations. While much of the attention has been devoted specifically to attacks on image-channel perception [5, 8, 9, 12, 19, 31], there is now a great deal of evidence that the LiDAR channel is also vulnerable [3, 22, 26].

The fusion of information from the myriad of available sensors, however, appears to be an opportunity for creating both more effective and more robust perceptual systems by integrating complementary information. For example, a LiDAR point cloud can provide additional depth information for RGB images. However, sensor fusion models are also not without limitations. For example, suboptimal fusion can lead to performance degradation even compared to single-channel models [17]. Additionally, several efforts have

shown that fusion models can exhibit single-channel and multi-channel vulnerabilities [32, 19, 14].

However, research so far has largely focused on identifying vulnerabilities in specific fusion architectures. We take a broader perspective and consider robustness of sensor fusion along four dimensions: 1) comparing robustness of deep sensor fusion to single-input-channel neural networks in the context of single-channel attacks, 2) the impact of fusion architecture on robustness, 3) the impact of the nature of the threat model, and 4) the efficacy of adversarial training (AT) in improving robustness to single-channel attacks. Regarding AT specifically, recent work by Kim and Ghosh [13] showed that single-channel adversarial training may be ineffective, whereas adversarial training with joint-channel attacks yields robustness to *single-channel* attacks. However, their evaluation only considered Gaussian noise with and used the AVOD architecture. Our investigation, in contrast, is focused on adversarial examples, and is far broader in scope.

We consider deep sensor fusion models with two perceptual input channels: camera (RGB image) and LiDAR (point cloud). Our analysis is in the context of 2D object detection. Our focus on 2D rather than 3D detection enables a direct comparison in robustness between single-channel models such as camera-only deep neural networks and sensor fusion models (dimension (1) above). Furthermore, we construct fusion architectures using YOLOv4 2D detection architecture as the anchor. Specifically, we compare two general fusion paradigms: early fusion (YOLO-Early), in which the two inputs are fused immediately at the input layer of the model architecture, and late fusion (YOLO-Late), in which we fuse (concatenate) the *features* extracted from each channel input (followed by additional neural network layers), where feature extraction paths are independent in the neural network architectures (dimension (2) above). For either early or late fusion, as well as a LiDAR-based single-channel model, we use a *depth* map extracted from the point cloud as the neural network input. This, again, enables direct comparison across architectures. We also study AVOD model, which uses a two-stage detection framework, in contrast of single-stage YOLO architectures.

We consider three threat models for each channel: the first one mirrors typical digital attacks where the entire input can be perturbed (for example, the typical l_∞-bounded adversarial example attacks [16]); the second one is meant to be more alike physically realizable attacks by masking adversarial inputs to be only within car bounding boxes (analogous to adversarial patch attacks [2]); and the third one, black-box version of the first threat model, where attackers has no access to parameters of the detection network. This allows us to explore dimension (3) above. Finally, we study two classes of adversarial training: first, single-channel adversarial training, where adversarial examples are only introduced for a single channel in the adversarial training loop, and second, joint-channel training, in which attacks on both channels are used in training. This addresses dimension (4) above.

Our main findings are:

1. Sensor fusion models are both more effective *and* more robust against single-source attacks than single-channel models.
2. Prior to adversarial training, early fusion is typically more robust than late fusion. This advantage largely disappears after adversarial training.
3. Single-channel adversarial training often *decreases* robustness to single-channel attacks *on the same channel*.
4. Single-channel adversarial training exhibits negative cross-channel externalities, decreasing robustness to attacks on the other channel.
5. Joint-channel adversarial training boosts robustness, but only slightly.
6. Conventional digital attacks used in adversarial training yield models that are robust neither to these attacks, nor to the more physically realizable attacks restricted to a car's bounding box. This appears to be a form of adversarial overfitting (the model becomes too conservative), and is illustrated in Figure 1. Adversarial training with restricted attacks is more effective and has fewer deleterious side-effects.

2. Related Work

Robust Sensor Fusion Data collected from multiple sensors are fused together to deal with complex tasks, e.g., audio-visual fusion in speech recognition (e.g. [7, 18]) and video captioning (e.g. [27, 20]). We refer the readers to [10] for a detailed literature review on AD sensor fusion. As for robust sensor fusion, [17] studies the robustness of audio-visual fusion

against catastrophic fusion problem in speech recognition. [34] proposes a multimodal model which is robust to the absence of some sensors. [21] focuses on the uncertainty of input data from different sensors, including the differences in physical units of measurement, sampling resolutions, and spatio-temporal alignment. [13] are the first to mention adversarial robustness of sensor fusion models, but only provide a theoretical proof for robustness of logistic regression against single-source adversarial noise. [25] experimentally study AT to achieve robust sensor fusion, but use highly visible attacks.

Adversarial Perturbations Adversarial perturbations add small noise to sensor inputs [16, 12]. Recently, physically realizable attacks have been proposed in the context of AD. For example, a number of attacks add stickers or adversarial patches to objects, such as stop signs and road pavement, or otherwise perturb these, in order to cause AD mistakes [8, 9, 5, 15, 6, 2, 24]. Similarly, attacks on point cloud input modalities range from perturbing all points [30] to physically realizable attacks that spoof the LiDAR sensor [3] and introducing real objects that cannot be detected by a LiDAR [4]. However, relatively few approaches study attacks specifically against fusion architectures. The few that do only consider the image inputs [32, 19], or consider a single target model, such as AVOD [25].

3. Attacking and Defending Sensor Fusion for Object Detection

We begin by describing the object detection framework, and the use and several forms of sensor fusion in this context. We then present a generalization of single-sensor attacks to apply to both multiple input modalities in fused models, as well as to model digital and physically realizable attacks. Finally, we describe natural variations of adversarial training as a means to defend object detection models that utilize sensor fusion from single-channel and multi-channel attacks.

3.1. Sensor Fusion for Object Detection

In addition to studying the AVOD [14] sensor fusion architecture specifically, we are investigate the effect of different fusion paradigms on adversarial robustness. For this, we select YOLOv4 [1], a single-stage detector, as our base model, and design two variations of YOLOv4 - YOLO-Early and YOLO-Late, where YOLO-Early fuses the two sensor inputs immediately at the beginning of the model, and YOLO-Late delayes the fusion operation until the feature output layer.

For consistency, we convert the input point cloud representation into a depth map, which is then used as the LiDAR sensor input into deep neural networks. Additionally, we project each point of the depth map onto the RGB image, and use nearest interpolation to generate a *dense depth map*, which serves as one LiDAR input channel. In addition, we have a second LiDAR channel, the *distance map*, which records the distance between the interpolated point and the original point. Since the two LiDAR channels are pixelwise consistent with the RGB images, we can either add them as a fourth and fifth channel, resulting in our RGBD representation, or use the two LiDAR channels as independent LiDAR inputs. We call the RGBD input representation *early fusion* (and the corresponding YOLO variant *YOLO-Early*), since both image and LiDAR inputs are combined early in the deep neural network architecture. Our *late fusion* variants (referred to as *YOLO-Late*) use separate feature extraction routes for RGB and two-channel LiDAR-based depth and distance maps, with the extracted features being fused later in the architectural structure of the neural network. Finally, we can also use single-sensor variants, such as RGB-input-only and LiDAR (depth and distance)-input-only models (referred to as *YOLO-RGB* and *YOLO-Depth* respectively). We implement these architectures using YOLOv4. Specifically, *YOLO-Early* directly feeds the 5-channel RGBD image into the original YOLOv4 model. *YOLO-Late* uses two separate feature extraction routes for the two channels, each has its individual Darknet backbone and feature pyramid network (FPN) neck. Then, we concatenate both features and subsequently feed them into the shared bounding box head for final prediction.

All models are trained to detect three classes of objects: cars, pedestrians, and cyclists. We use mean average precision (mAP) to evaluate the effectiveness of object detection, both with and without attacks, where significant reduction in mAP after an attack entails a successful attack.

3.2. Attacks on Sensor Fusion for Object Detection

We consider white-box decision-time (adversarial perturbation) attacks on inputs into the neural net-

works that combine data from a collection of n sensors. To formalize, let $f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta})$ be a neural network model for object detection that takes inputs $\boldsymbol{x}_i$ from the n sensors. While our specific focus below will be on only two sensor modalities, RGB image and LiDAR (represented by a distance and depth maps, as discussed above), we describe the adversarial framework more generally. The learner's goal is to learn model parameters $\boldsymbol{\theta}$ that minimize the prediction loss $L(f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}), y)$, and we model the adversary's goal as maximizing this loss *for a fixed vector of model parameters* θ. Let $\mathcal{I}$ denote the the set of indeces of sensors which are attacked, and $\mathcal{I}^c$ the indeces of sensors which are not (with $\mathcal{I} \cup \mathcal{I}^c = [1..n]$, where $[1..n]$ is the set of integers between 1 and n, and $\mathcal{I} \cap \mathcal{I}^c = \emptyset$). A natural generalization of adversarial example attacks to the sensor fusion setting is then

$$\begin{aligned} &\arg\max_{\boldsymbol{\delta}} L(f(\ldots, \boldsymbol{x}_i + \boldsymbol{\delta}_i * \boldsymbol{m}_i, \ldots; \boldsymbol{\theta}), y) \qquad (1) \\ &s.t. \quad \|\boldsymbol{\delta}_i\|_p \leq \epsilon_i, \forall i \in \mathcal{I},\ \boldsymbol{\delta}_i = 0, \forall i \in \mathcal{I}^c, \end{aligned}$$

where $\boldsymbol{\delta}_i$ is the perturbation on the i-th sensor input, ϵ_i is the bound on the l_p norm perturbation of the ith sensor attack, and $\boldsymbol{m}_i$ is a mask which constrains the attack on sensor i to an exogenously specified contiguous region of the input (e.g., image or LiDAR). In our setting, perturbations to images take the form of modifying pixel values for the three color channels. For LiDAR, in contrast, we modify the input *point cloud* (depth and distance maps are thereby indirectly affected) by shifting each point in 3D space.

The idea behind introducing masks $\boldsymbol{m}_i$ is that it allows us to capture a common distinction between *digital* and *physically realizable* attacks [29]. In particular, we will consider digital attacks, in which the mask is simply the entire input (i.e., no mask), as well as physically realizable attacks in which the mask is restricted to the bounding boxes of cars in scene. As our focus is on image and LiDAR input modalities, and since perturbations to these inputs are not directly comparable, we emphasize this by using ϵ to refer to the bound on the RGB inputs (denominated in pixel color value intensities) and γ to denote the bound on LiDAR point cloud perturbations (denominated in meters). Throughout, we focus discussion on l_∞-norm perturbations. We use the l_∞ variant of projected gradient descent (PGD) for all attacks, on both the image and LiDAR sensor modalities.

In object detection one can have several possible objectives in adversarial perturbations, such as confidence scores [31] and regression loss [33]. We focus on the latter, which is highly effective.

3.3. Defense through Adversarial Training

A common and usually most effective way to obtain increased robustness to adversarial perturbation attacks is *adversarial training*. A common mathematical framework for adversarial training is robust optimization of the following form:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[\max_{\boldsymbol{\delta}} L(f(\ldots, \boldsymbol{x}_i + \boldsymbol{\delta}_i * \boldsymbol{m}_i, \ldots; \boldsymbol{\theta}), y)] \qquad (2)$$

Learning is then a form of gradient descent with respect to $\boldsymbol{\theta}$, evaluated at the value of $\boldsymbol{\delta}$ which approximately maximizes the inner optimization of Problem (1). While PGD has commonly been used to obtain the approximate solution to the inner adversarial optimization problem [16], we utilize a recently proposed fast alternative that uses FGSM with random initialization and cyclic learning rate [28].

We utilize two forms of adversarial training. The first form only adds adversarial perturbations for a single input channel, with the aim of making that channel only more robust to attack. The second form adds perturbations to both input channels (image and LiDAR) concurrently in the adversarial training loop. Below we study the relative efficacy of these two variations. In addition, we consider two variations of single-channel threat models in adversarial training: digital attacks (i.e., adversarial training using attacks without a mask) and physically realizable attacks (in which adversarial perturbations are restricted to a small contiguous region of the input).

4. Experiments

4.1. Experiment Setup

Data We perform experiments on Waymo [23] and KITTI [11] datasets. Since the results on both datasets are highly consistent, we only present the results on Waymo in the main text, with KITTI results deferred to the supplement. Waymo has 5 cameras in total. In this work, we only use the front camera for images.

We select front images and LiDAR point clouds evenly from $1/10$ of all the frames for training. For testing, we select the first frame from each sequence in the validation set, which gives us 202 samples in total. The resemblance among different frames within the same sequence helps reduce the effect of sampling on the evaluation performance to the minimum.

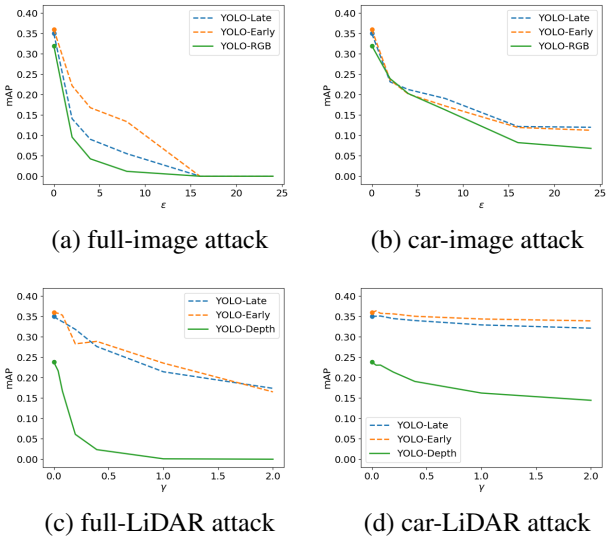
Training When training YOLOv4 models, we start everything from scratch, i.e. no pretrained weights are used. This is to avoid the problem where models pretrained on other image datasets might make the model biased to images and ignore the LiDAR sensor in 2D detection. More details can be found in the supplementary. For AVOD model, we follow the instructions in the original paper to train it on KITTI.

Attacks and Defenses We consider three attacks on both image and LiDAR channels - *digital* attacks that perturb the entire input, *physically realizable* attacks, and *black-box* digital attacks, which are described in Section 3.2. For defenses, we consider variants of adversarial training (AT), described in Section 3.3, based on each of the four single-sensor attacks: *AT-image*, which uses the full-image l_∞ attack, *AT-LiDAR*, which attacks the full point cloud, as well as *AT-Car* and *AT-LiDAR-Car*, which use attacks within the car bounding boxes only. Additionally, we consider adversarial training using attacks on *both* sensor modalities jointly, referred to as *AT-Joint*. For adversarial training on the image channel, $\epsilon = 2$, while adversarial training on the LiDAR channel uses $\gamma = 0.3$.

4.2. Adversarial Robustness of Fusion Models

Our first question involves relative performance of LiDAR-image fusion models in comparison with single-sensor models. First, we compare the effectiveness of early and late fusion models on *clean*, that is, original unperturbed, test data. It is, of course, not surprising that both early fusion (0.361)[1] and late fusion (0.350) models significantly outperform both the image-only (0.322) and LiDAR (depth)-only (0.238) single-channel models.

Figure 2 offers a more complete comparison in terms of adversarial robustness to our four attacks: the top row presents robustness to image attacks, whereas the bottom row shows robustness to LiDAR attacks. In all cases, we can see that not merely accuracy is considerably improved by fusion, but also single-channel robustness. Most significant improvements of fusion models in comparison to single-channel variants are in the context of full-image attacks (where early fusion is much more robust than the image-only model), and in the context of LiDAR attacks, where either fusion variant is far more effective both in clean data accuracy and robustness than the LiDAR-only counterpart. We can also observe from Figure 2 that early fusion is typically better than late fusion. In the case of full-image attacks, this advantage is substantial. However, in the case of the other three attack variants, the difference between early and late fusion is nearly negligible.



(a) full-image attack (b) car-image attack

(c) full-LiDAR attack (d) car-LiDAR attack

Figure 2: Robustness of deep sensor fusion, compared to single-channel neural network object detection models, to adversarial attacks on image and LiDAR modalities.

4.3. Effects of Adversarial Training

Adversarial Training and Image Channel Attacks Above, we observed that early fusion appears to be somewhat more robust to *full-image attacks* than late fusion, and slightly better even on clean data. At the same time, we saw that the difference between these approaches is minimal in the case of other attacks we consider. Furthermore, in fusion models are also more robust than single-sensor counterparts. We now investigate the impact of adversarial training on these observations.

[1]Numbers in parenthesis following the model name denotes its mean average precision (mAP) performance.

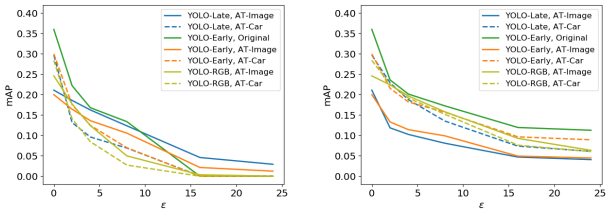


Figure 3: Early and late fusion after adversarial training, under attacks on the image channel. Left: full-image attack. Right: car bounding box attack.

In Figure 3, we consider both attacks (car bounding boxes only, and full-image), and adversarial training, with respect to the *image* channel. First, we can note that after adversarial training, the difference between early and late fusion (blue vs. orange lines) becomes small, whichever adversarial training method is used. However, if we compare AT-image early fusion between (solid orange lines) and original early fusion (solid green lines), we can see something rather striking: adversarial training actually makes it *less robust*, except for the exceptionally strong attacks (full-image, large ϵ), in which case all models are very bad in any case. In contrast, adversarial training does boost the robustness of late fusion, although only on the full-image attacks. Indeed, even if we consider AT-Car models, that is, adversarial training using the weaker attacks inside cars' bounding boxes (dashed lines in Figure 3), there seems little improvement in robustness to either type of attack.

A related observation pertains to a comparison between AT-Image, that is, adversarial training using attacks on the full image, and AT-Car, where only attacks within the car bounding boxes are used in training. We can note that AT-Image (solid lines in Figure 3) models are not much more robust than AT-Car models when facing full-image attacks (left plot), but are *much less robust* when faced with car-image attacks. AT-Image is also significantly worse in performance on clean data, e.g., mAP of early fusion with AT-Image is only 0.200 compared to its original performance 0.361. Furthermore, in the case of car-image attacks, any advantage of sensor fusion over the image-only model largely dissipates after adversarial training.

Interestingly, the observations above appear to be specific to sensor fusion models. In particular, if we consider image-only models (YOLO-RGB in the plots), adversarial training has both a comparable or lesser impact on clean accuracy, but also less value for robustness to full-image attacks with a large ϵ. On the other hand, these image-only models remain comparable to AT-Car variants of sensor fusion, even when trained using full-image attacks.

We can conclude that *the choice of threat model in adversarial training is critical, and conventional digital attacks in which every pixel in an image is suspect are perhaps a particularly poor choice*. More generally, if the adversarial model is too strong, and likely not a realistic representation of the attacker's capabilities, as is the case with full-image attacks, using it for adversarial training leads one to become too conservative (we can think of this as a form of overfitting to the adversarial model). This is visualized in Figure 1, where AT-Image late fusion model yields a very large bounding box for a car: indeed, it is safe in the sense that the bounding box includes the object, but liveness (ability to precisely identify boundaries of an obstacle) is critically compromised.

Adversarial Training and LiDAR Channel Attacks Next, we explore the effect of full-LiDAR attacks and LiDAR attacks restricted to cars' bounding boxes used both for adversarial training and evaluation (in other words, the threat model is a LiDAR-channel attack, consistent for both attack and defense).

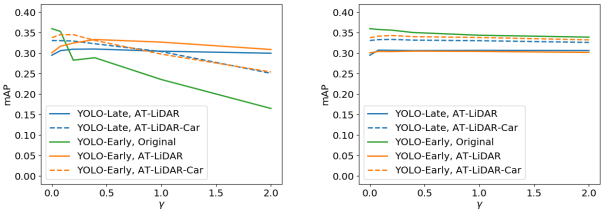


Figure 4: AT with LiDAR attacks, and LiDAR Channel Attacks. Left: full-LiDAR attack. Right: car-LiDAR attack.

Figure 4 presents the results of LiDAR attacks on AT-LiDAR variants. AT for YOLO-Depth is not presented because of its inferior performance. One interesting difference from our results where attacks are on the image channel is that here we do see a clear improvement in robustness compared to the original early fusion model in the case of full-LiDAR attacks (left plot). This is true whether adversarial training used the car-LiDAR or full-LiDAR attack model. Moreover,

the difference between AT-LiDAR and AT-LiDAR-Car appears less important than the analogous difference between AT-Image and AT-Car earlier, although AT using LiDAR-Car attacks does yield somewhat better results when evaluated against the same attack. We similarly observe that AT-LiDAR-Car (early fusion: 0.338; late fusion: 0.331) has superior clean data mAP to AT-LiDAR (early fusion: 0.301; late fusion: 0.295). Nevertheless, we do broadly still observe that the choice of the threat model is important. Specifically, if we view attacks that are restricted to the cars' bounding boxes (which are more likely to be physically realizable), even the baseline (pre-AT) model is quite robust—in fact, more robust than any of the adversarially trained variants. Moreover, in this case, adversarial training with full-LiDAR attacks decreases performance the most.

Overall, so far we observe that *single-channel adversarial training appears to be surprisingly ineffective in sensor fusion settings in the context of same-channel attacks*: not only does accuracy on clean data decrease, but even robustness to attack often does. This appears to be in contrast with single-channel adversarial training, which does increase robustness at least to the threat model used in AT (compare YOLO-RGB between Figures 2(a) and 3(left) for $\epsilon \leq 5$).

Cross-Channel Externalities of Single-Channel Adversarial Training Thus far, we considered only attacks on the same single channel for both adversarial training and robustness evaluation. We now study a uniquely fusion phenomenon: cross-channel externalities of adversarial training. Specifically in the context of image and LiDAR channels, we have in mind the following experiments. On the one hand, we can perform adversarial training with image-channel attacks, but evaluate the resulting model using LiDAR-channel attacks. On the other hand, we can reverse this, evaluating LiDAR-based AT using image-channel attacks. The question we ask is: what impact does adversarial training with respect to one channel has on robustness to attacks on the other channel?

The results of this evaluation are in Figure 5 when we use image attacks for training and LiDAR for evaluation, and in Figure 6, where adversarial training uses LiDAR attacks and evaluation uses image attacks. First, consider image-based adversarial training in Figure 5. AT with full-image attacks, in particular, causes

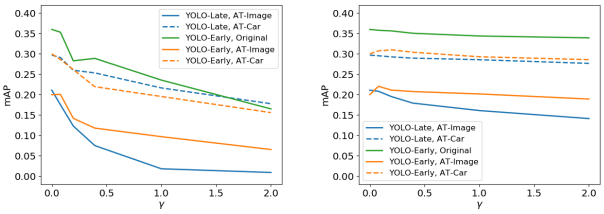


Figure 5: Early and late fusion after adversarial training using image attacks, but subject to LiDAR attacks. Left: full-LiDAR attack. Right: car-LiDAR attack.

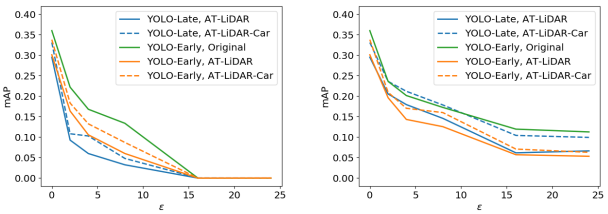


Figure 6: Early and late fusion after adversarial training using LiDAR attacks, but subject to image attacks. Left: full-image attack. Right: car-image attack.

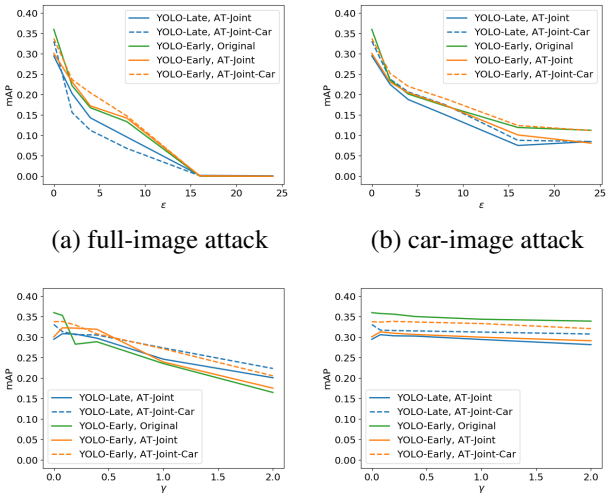
dramatic degradation in both performance on clean data and in robustness to LiDAR-based attacks of either variety. Interestingly, here we also see that early fusion (solid orange lines) generally does better than late fusion (solid blue lines). While adversarial training using only the car bounding box adversarial examples exhibits less loss in robustness compared to the model before no adversarial training, there is still a tangible reduction in performance against adversarial examples. Overall, we see a clear negative externality of image-based adversarial training on robustness to the LiDAR-channel attacks.

Turning to the reverse setting in Figure 6, the results are broadly consistent, but the differences less dramatic than above. In particular, we still typically see that single-channel adversarial training results in lower performance on the other channel compared to no adversarial training at all. The one interesting exception is in the performance of AT-LiDAR-Car late fusion model, which is essentially identical (i.e., no cross-channel negative externality) to the original late fusion model (i.e., before AT).

Overall, we therefore find that there are negative cross-channel externalities for both channels. However, here, too, we see that there is a great deal of importance in the choice of the threat model. For ex-

ample, to the extent that we see externalities, they are much stronger when we train with adversarial examples generated on the entire image or point cloud, than AT using only attacks within cars' bounding boxes.

Joint-Channel Adversarial Training Given the broad failure of AT that we had observed above in inducing much robustness in deep image-LiDAR fusion models, it is natural to consider this failure a product of training with only single-source attacks. One can therefore hypothesize that all will be well once we adversarially train with attacks on *both* channels simultaneously. Indeed, precisely this remedy has been proposed in a largely theoretical framework by Kim and Ghosh [13] with the goal of inducing single-source robustness to *random noise* in classification problems. Our goal now is to study this hypothesis experimentally in 2D object detection domain.



(a) full-image attack (b) car-image attack

(c) full-LiDAR attack (d) car-LiDAR attack

Figure 7: Early fusion v.s. late fusion with Joint-Channel Adversarial Training.

Figure 7 presents the results on the efficacy of joint-channel adversarial training (Joint-AT). Indeed, we see here a much improved picture compared to our results with single-channel adversarial training above. For example, when facing full-image and full-LiDAR attacks, early fusion AT-Joint models now indeed outperform the pre-AT baseline in terms of robustness, although one has to acknowledge that this improvement isn't especially striking. However, improvement is also not universal even here. For example, late fusion still under-performs the original fusion model on image attacks (Figure 7(a) and (b)). While all AT fusion models appear more robust in full-LiDAR attack evaluation (Figure 7(c)), they are all *less* robust compared to original fusion model when the LiDAR attack is restricted to cars' bounding boxes (Figure 7(d)). Finally, and somewhat surprisingly, now joint-channel training using attacks restricted only within cars' bounding boxes almost always outperform alternative Joint-AT approaches, *even in robustness to full-image and full-LiDAR attacks*. This last observation is very surprising, but overall serves to bolster our general observation that unrestricted conventional digital adversarial example models are not necessarily useful in developing robust classifiers.

5. Conclusion

We presented a systematic analysis of comparative robustness of deep fusion architectures involving RGB image and LiDAR channels. Our first finding is that sensor fusion models are more robust against single-channel adversarial perturbation attacks than single-channel models before and after adversarial training, highlighting the role of fusion in improving robustness. Second, adversarial training on a single channel is problematic for sensor fusion models - it not only overfits with the background perturbation, but also makes the other sensor less robust. Third, jointly training on both sensors can mitigate the problem of single-sensor AT. Nevertheless, even joint-channel training fails to significantly improve robustness to single-channel attacks. Fourth, we find throughout that the nature of the attack really matters when it comes to adversarial training. Indeed, it is typically the case that adversarial training using perturbations to the entire input (say, every pixel in the input image) often overfits to the attack, and usually performs worse than fusion models before adversarial training. Since such abstractions of attacks are not particularly well motivated in autonomous driving security in any case, our results suggest that threat modeling *specifically for the purpose of defense* must be a critical area for future research in robust sensor fusion.

Acknowledgement

This research was partially supported by the NSF (IIS-1905558, ECCS-2020289) and ARO (W911NF1910241).

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020.

[2] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017.

[3] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z. Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz, editors, *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 2267–2281. ACM, 2019.

[4] Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. *CoRR*, abs/1907.05418, 2019.

[5] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng (Polo) Chau. Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part I*, volume 11051 of *Lecture Notes in Computer Science*, pages 52–68. Springer, 2018.

[6] Alesia Chernikova, Alina Oprea, Cristina Nita-Rotaru, and BaekGyu Kim. Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction. In *2019 IEEE Security and Privacy Workshops, SP Workshops 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 132–137. IEEE, 2019.

[7] Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. Multimedia*, 2(3):141–151, 2000.

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. IEEE Computer Society, 2018.

[9] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, and Florian Tramèr. Note on attacking object detectors with adversarial stickers. *CoRR*, abs/1712.08062, 2017.

[10] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Fabian Timm, Claudius Glaeser, Werner Wiesbeck, and Klaus Dietmayer. Deep multimodal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *arXiv:1902.07830*, 2019.

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

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. In *Neural Information Processing Systems*, pages 4815–4826, 2019.

[14] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1-5, 2018*, pages 1–8. IEEE, 2018.

[15] Jiajun Lu, Hussein Sibai, Evan Fabry, and David A. Forsyth. NO need to worry about adversarial examples in object detection in autonomous vehicles. *CoRR*, abs/1707.03501, 2017.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[17] Javier R. Movellan and Paul Mineiro. Robust sensor fusion: Analysis and application to audio visual speech recognition. *Mach. Learn.*, 32(2):85–100, 1998.

[18] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 2130–2134. IEEE, 2015.

[19] Won Park, Qi Alfred Chen, and Z. Morley Mao. Crafting adversarial examples on 3D object detection sen-

sor fusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4321–4325. IEEE, 2020.

[20] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In Alan Hanjalic, Cees Snoek, Marcel Worring, Dick C. A. Bulterman, Benoit Huet, Aisling Kelliher, Yiannis Kompatsiaris, and Jin Li, editors, *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1092–1096. ACM, 2016.

[21] Varuna De Silva, Jamie Roche, and Ahmet M. Kondoz. Robust fusion of lidar and wide-angle camera data for autonomous mobile robots. *Sensors*, 18(8):2730, 2018.

[22] 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 *USENIX Security Symposium*, pages 877–894, 2020.

[23] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[24] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 49–55. Computer Vision Foundation / IEEE, 2019.

[25] James Tu, Huichen Li, Xinchen Yan, Mengye Ren, Yun Chen, Ming Liang, Eilyan Bitar, Ersin Yumer, and Raquel Urtasun. Exploring adversarial robustness of multi-sensor perception systems in self driving, 2021.

[26] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 13713–13722, 2020.

[27] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 795–801. Association for Computational Linguistics, 2018.

[28] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training, 2020.

[29] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *International Conference on Learning Representations*, 2020.

[30] Chong Xiang, Charles R. Qi, and Bo Li. Generating 3d adversarial point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9136–9144. Computer Vision Foundation / IEEE, 2019.

[31] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision*, pages 1378–1387. IEEE Computer Society, 2017.

[32] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. *CoRR*, abs/2005.10987, 2020.

[33] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 421–430. IEEE, 2019.

[34] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2365–2374. IEEE, 2019.