Adversarial Signal Generation for Fused-Sensor Autonomous Driving

Cleveland James Crossling
May 2026
Machine Learning CSCI-5270

East Tennessee State University

1 Introduction

Autonomous Vehicles (AVs) are integral to modern transportation infrastructure, relying on multi-modal sensor arrays (LiDAR, Radar, and Cameras) for environmental perception. While AV perception models have increased in robustness, they remain statistical systems and therefore are susceptible to adversarial poisoning. This project investigates the vulnerability of Tiny-VAD, a vectorized autonomous driving model, to malicious perturbations that can induce dangerous vehicle actions. Current research seeks to bridge the gap between digital perturbations and real world consequence. Subsequently, this work employs the Image Joint Embedded Predictive Architecture (I-JEPA) as a novel baseline for generating adversarial data for the purpose of exploring semantically aided visual perturbations.

2 Motivation

As systems continue to adopt autonomous capabilities precise decision making becomes fundamentally critical for safe operation. Uniquely, AV models go beyond the requirement of a need for precise decision making from the perspective of quality, and instead is a necessity for feasible usage on the road. State-of-the-art architectures driving AVs are typically statistical machine learning models, and therefore are vulnerable to Out-of-Distribution adversarial perturbations where sufficient safeguards are not deployed.

3 Research Questions

This investigation seeks to answer the following:

1. Is it possible to bypass the mathematical disconnect of black-box architectures using a surrogate-guided adversarial generator?
2. To what extent can semantically-rich features (I-JEPA) induce measurable trajectory drift in robust, vectorized perception models?

4 Literature Review

Common attacks on sensors include physical objects designed to misrepresent raw sensor data within the AV model or exploit its statistical reasoning architecture. This literature review is not comprehensive in exploring all domains of attack as it excludes training oriented attacks (training poisoning, context drift, etc.) which are beyond the scope of relevancy in this work.

AVs rely on a semantic understanding of the environment around them gathered through raw data collected by their sensors. More recent systems, including state-of-the-art systems, rely on object detection and classification to represent the environment surrounding the vehicle in a semantically meaningful way. By utilizing this environmental representation, subsequent decisions can be determined and executed by the AV.

Physical attacks specifically targeting the object detection systems needed to evaluate surroundings can be particularly impactful in creating false decisions with a model. More blatant attacks include physically obscuring road signs or features but are not necessarily limited to autonomous vehicles and can affect human drivers. Consequently, subtle attacks such as sensor obfuscation through IR blocking foam or applied artifacts, such as shapes or coloured plastics, can cause object detection to misclassify or totally ignore objects.

Conversely, non-physical sensor attacks specifically targeting the underlying model of the autonomous vehicle can serve as a precursor to more sophisticated physical attacks. These adversarial approaches are referred to as black box as the internal architecture of the model is hidden despite controlled input. Model decision making is captured based on output with regards to variable inputs and the distilled knowledge is transferred to a malicious, surrogate model. The surrogate model, once sufficiently trained on the AV model's behaviour, can be used to generate more powerful adversarial attacks despite structure of the model not being explicitly known.

5 Dataset

The experiment utilizes the nuScenes-mini dataset, a subset of the larger nuScenes collection. The dataset consists of multi-modal data necessary for effective road navigation including LiDAR, radar and camera data. The mini subset contains a prepared eight training and two validation observation split with a total of ten samples. For this work, the goal is to create perturbations on known input, therefore a small set of samples is permissible as there is no direct differentiation between the driving model and generator.

6 Experiment Setup

6.1 Data Collection and Cleaning

NuScenes-mini serves as the dataset for the experiment; to ensure computational feasibility, the validation dataset is utilized. The focus is restricted to the

forward-facing camera image sequences to evaluate the efficacy of visual perturbations on frontal perception. Ground truth trajectory values are Z-score scaled prior to training to ensure stability during the surrogate learning phase.

6.2 Feature Extraction

Tiny-VAD, alongside other state-of-the-art AV models, are highly specialized to environmental perception, a prerequisite for effective autonomous driving. Consequently, neural networks with no world experience are insufficient for generating perturbations that a specialized model will not recognize or ignore as noise. To circumvent this challenge, a pre-trained model specialized to image data or world semantics is preferable as a baseline feature extractor.

For this work, I-JEPA is selected for feature extraction as it is specifically trained for semantically relevant feature representation. Models such as ResNet50 or DINOv3 are also suitable for feature extraction but rely on pixel-level feature extraction; I-JEPA is specialized to extracting the intrinsic underlying real-world semantics of image data.

6.3 Model Building and Implementation

The experimental pipeline consists of two distinct components specialized to distinct tasks in effectively generating adversarial perturbations. These components are the surrogate, which is trained on output from the Tiny-VAD model, and the generator, which attempts to confuse the surrogate by creating perturbations.

Surrogate As a result of backpropagation being impossible between Tiny-VAD and generator because of the black-box nature of AV models, a surrogate is necessary to bridge the gap between the two models. The surrogate receives both clean input, and output perturbations aggregated to input within two distinct training cycles. Surrogate performance is evaluated on a validation subset of clean and perturbed data (respective to the current training cycle). The metrics gathered to determine convergence is the Mean Squared Error between the Tiny-VAD trajectory outputs and the surrogate corresponding output.

The surrogate architecture consists of two convolutional layers with incorporated dropout regularization which are pooled and flattened. Subsequently, a feed-forward network corresponding to trajectory coordinate structure is appended for model output. After training, the best model collected through checkpointing on validation metrics is saved.

Generator The generator consists of a distinct convolutional network with two separate inputs in the forward pass. The model is similar structure to the surrogate but instead of image data receives I-JEPA features and a noise value. The noise value ensures that perturbations are generated from image data that influenced by a degree of entropy, ensuring adversarial convergence is not a result of surrogate mode collapse. Output consists of values passed through a

hyperbolic tangent function ([-1,1]) which allows generated perturbations to both darken and lighten the input image pixels.

To ensure convergence on a feasible adversarial perturbation, loss is designed to maximize deviation between a clean trajectory prediction of an image and a perturbed example. This adversarially maximizing loss, defined as $\mathcal{L}_{total}$ is computed as follows:

$$\mathcal{L}_{total} = -\lambda_{adv}\mathcal{L}_{mse}(\hat{y}_c, \hat{y}_a) + \lambda_{ssim}(1 - SSIM(x, x_a)) \tag{1}$$

Where the variables are defined as follows:

- $\mathcal{L}_{total}$: The loss value minimized during the backpropagation of the generator.
- $-\lambda_{adv}\mathcal{L}_{mse}$: The adversarial loss component, which calculates the Mean Squared Error between the clean trajectory ($\hat{y}_c$) and the adversarial trajectory ($\hat{y}_a$). Computed as negative as it must be minimized.
- $\lambda_{ssim}(1 - SSIM)$: The structural dissimilarity loss, ensuring the perturbed image (x_a) remains visually similar to the original image (x).

To prevent immediate modal collapse and to promote subtle impactful changes, perturbations generated by the model are limited to a 5% change from the previous iteration. This approach ensures model stability during the training process by preventing large changes that can lead to full black or white images which beat the surrogate, but are functionally useless.

7 Results and Insights

The implementation successfully produced adversarial signals that induced a measurable semantic shift in the target model's output.

The generated perturbations resulted in a 2cm average deviation from the ground truth trajectory predictions. While seemingly minimal, a consistent 2cm drift in an autonomous system is non-negligible, as it represents a significant reduction in decision precision that could result in lane-drifting or decision planning failures in dense traffic environments. The generation process exemplified the robustness of the Tiny-VAD architecture, as powerful perturbations were exceptionally difficult to create.

The lack of modal collapse, specifically avoiding white or black-out effects, confirms that the configuration effectively balanced adversarial strength while preserving visual structure. Within safety-critical transport, any successful induction of a predictable error in a vectorized model serves as a proof of concept for more sophisticated physical attacks.

8 Future Research

Future research should expand the scope of the adversarial generation process to target the full multi-modal sensor array of AVs. While this work focused on visual perturbations, investigation into fused-sensor attacks, which incorporate

LiDAR and Radar data, is necessary to determine if cross-modal consistency can be exploited. Additionally, an objective for future studies is the transmissibility of these signals to physical applications. This involves the creation of real-world adversarial objects, such as specialized road objects or surfaces, to evaluate their efficacy in physical driving environments. Finally, future work must explore potential defensive strategies to prevent discovered adversarial attacks. This could involve the implementation of stronger data augmentation techniques during the training phase to improve model robustness against semantic shifts.

9 Answers to Research Questions

Regarding Question 1, the experiment confirms that a surrogate model can bridge the gradient gap between a generator and a decoupled target. By mimicking Tiny-VAD's output behavior, the surrogate allowed for effective backpropagation. Regarding Question 2, the I-JEPA features enabled a somewhat-successful semantic attack, resulting in a consistent 2cm trajectory deviation.

10 Things Learned

Throughout this project, the main takeaway was the complexity involved in fine-tuning the surrogate against a high-precision model. Additionally, I learned the structure behind black-box attacks and the generative adversarial network applications to circumvent the introduced limitations.

11 Conclusion

This project demonstrated a decoupled adversarial sensor attack on the Tiny-VAD architecture. By utilizing an I-JEPA guided generator and a surrogate bridge, the lack of direct weight connectivity to influence the model's trajectory planning was bypassed. This serves as a critical reminder that even robust, vectorized models are susceptible to precision drift under semantic pressure.