Is Your Autonomous Vehicle Safe? Understanding the Threat of Electromagnetic Signal Injection Attacks on Traffic Scene Perception

Wenhao Liao[1*], Sineng Yan[1*], Youqian Zhang[2†], Xinwei Zhai[1], Yuanyuan Wang[1], Eugene Fu[3]

[1]Shenzhen University, Shenzhen, China
[2]The Hong Kong Polytechnic University, Hong Kong S.A.R., China
[3]The Education University of Hong Kong, Hong Kong S.A.R., China
liaowenhao23@email.szu.edu.cn, yansineng2023@email.szu.edu.cn, you-qian.zhang@polyu.edu.hk, zhaixinwei2022@email.szu.edu.cn, yywang@szu.edu.cn, eugenefu@eduhk.hk

Abstract

Autonomous vehicles rely on camera-based perception systems to comprehend their driving environment and make crucial decisions, thereby ensuring vehicles to steer safely. However, a significant threat known as Electromagnetic Signal Injection Attacks (ESIA) can distort the images captured by these cameras, leading to incorrect AI decisions and potentially compromising the safety of autonomous vehicles. Despite the serious implications of ESIA, there is limited understanding of its impacts on the robustness of AI models across various and complex driving scenarios. To address this gap, our research analyzes the performance of different models under ESIA, revealing their vulnerabilities to these attacks. Moreover, due to the challenges in obtaining real-world attack data, we develop a novel ESIA simulation method and generate a simulated attack dataset for different driving scenarios. Our research provides a comprehensive simulation and evaluation framework, aiming to enhance the development of more robust AI models and secure intelligent systems, ultimately contributing to the advancement of safer and more reliable technology across various fields.

Introduction

Autonomous driving is advancing at a rapid pace, and becoming a reality with real products now being used in everyday life. Notable examples, including Tesla, with its Autopilot, Google's Waymo, Baidu's Apollo Go, etc., exemplify the cutting-edge autonomous driving capabilities in practice (Badue et al. 2021). However, safety concerns still hinder the public trust and constrain the widespread adoption of autonomous driving, particularly with the potential threat of adversarial attacks (Liu, Nikitas, and Parkinson 2020; Ljubi and Groznik 2023; Nastjuk et al. 2020). Unlike human drivers, autonomous vehicles can be potentially affected by adversarial attacks, leading to inaccurate perception of driving environment, and eventually, unsafe operation (Zhang et al. 2021b).

Figure 1 illustrates a general process of autonomous driving systems, composed of three modules: "Perception", "Planning", and "Control". The perception module gathers input data through sensors such as cameras, LiDARs, and radars, then employs artificial intelligence (AI) to accurately interpret road markings, vehicles, pedestrians, and other environmental elements for further processing. The planning module uses this environmental understanding to generate driving strategies, such as motion trajectories, while the control module converts these strategies into control commands to steer wheels so as to achieve autonomous driving.

The perception module is arguably the most critical, which serves as the fundamental component and directly influences the decisions made by the other modules. It is thus the primary target of adversarial attacks (Guesmi et al. 2023). Of specific interest is the camera and AI models within the perception module due to their versatility and critical role, including functions such as obstacle detection, lane recognition, and traffic sign identification. If such a perception module is attacked, it may result in traffic accidents involving the autonomous car. For example, failing to detect obstacles ahead, the car may not stop, leading to a catastrophic car crash. It is essential to point out that such attacks are not imagined, and many studies have been demonstrated to manipulate images captured by cameras through different attack vectors, including adversarial patches/stickers, lasers, ultrasound, and electromagnetic interference (detailed discussion is presented in the Related Work).

Among these attacks, one emerging and concerning threat is Electromagnetic Signal Injection Attack (ESIA) (Köhler, Baker, and Martinovic 2022; Jiang et al. 2023; Zhang et al. 2024a,b; Kang et al. 2024), a real-world attack that injects adversarial electromagnetic signals into the camera circuits, causing malicious image distortion, e.g., color strips as shown in Figure 1. Jiang et al. (2023), and Zhang et al. (2024b) provided a detailed workflow and error modeling of ESIA, demonstrating and confirming that ESIA can cause color strips. Jiang et al. (2023) also showed that it is feasible for attackers to deploy compact attack devices, as small as a credit card, close to car cameras and trigger the attack while the car is in operation. ESIA could mislead AI models within the autonomous driving system, leading to erroneous decisions with potentially catastrophic consequences for road safety. Understanding the robustness and reliability

[*]These authors contributed equally.
[†]Corresponding author.

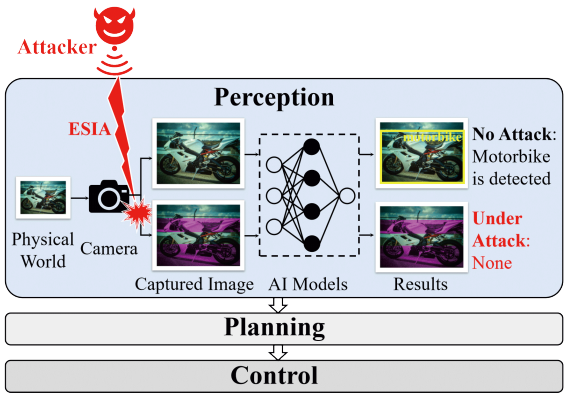


Figure 1: A general process of autonomous driving systems consists of three modules: Perception, Planning, and Control. Electromagnetic signal injection attacks (ESIA) can manipulate the output image of the autonomous vehicle's cameras, hiding the motorbike from being detected, potentially leading to an accident.

of the existing AI models under such attacks is crucial for ensuring safety, building public trust, and enhancing future models to better handle under-attack scenarios. Although related work has explored various attacks on cameras, the focus of this paper lies on the severe implications of ESIA on the perception system and the inherent risks it poses to autonomous driving operations, which have not been investigated yet.

A critical research gap exists in understanding the varying severity of ESIA attacks across different driving scenarios. Uncertainties persist regarding the specific risks posed by such attacks in diverse environments, necessitating a comprehensive investigation to gauge their potential impact accurately. However, conducting practical experiments to gather extensive under-attack data is hindered by the complexity and cost associated with setting up professional-grade attack systems and camera systems that save images in raw format (Jiang et al. 2023; Zhang et al. 2024a). This limits the research on understanding the threat posed by ESIA on real traffic images. Additionally, comprehending why AI models falter under attack conditions presents a significant challenge, complicating efforts to bolster system defenses against adversarial threats.

To bridge these research gaps, we made the following contributions:

- We develop a novel simulation method that mimics ESIA's adversarial patterns, making the attack scalable and facilitating extensive research without the need for costly practical setups.
- We introduce an ESIA simulation dataset focused on traffic scenarios, categorized by the attack severity, which will aid in the development of robust and reliable autonomous driving systems in the future.
- We gain a deep understanding of how different models perform under ESIA attacks through systematic experiments in diverse traffic scenarios, highlighting the threat of ESIA and potential safety risks.

Related Work

Adversarial attacks on cameras manifest in two primary forms: digital attacks, and physical attacks.

Digital Attacks

In digital attacks, an attacker can arbitrarily alter the image at a granular pixel level. In an early work, Szegedy et al. (2014) demonstrated how even minor perturbations, imperceptible to the human eye, could significantly disrupt the performance of AI models. Following this, numerous algorithms, for example, Goodfellow, Shlens, and Szegedy (2015); Carlini and Wagner (2017); Chen, Jordan, and Wainwright (2020), have been developed to craft adversarial perturbations that exploit the vulnerabilities in the model's learning process, causing it to make incorrect predictions or classifications.

Physical Attacks

Physical attacks involve manipulating inputs from the physical world to deceive the camera system. Several techniques have been explored in this domain.

Adversarial patches or stickers, strategically positioned on physical surfaces of objects, aim to deceive recognition systems, leading to misclassification or false interpretations. For instance, a carefully designed sticker on a stop sign could cause an autonomous vehicle's camera to misinterpret it (Eykholt et al. 2018; Song et al. 2018; Duan et al. 2020), or an adversarial marking on the ground can cause an autonomous vehicle to drive in the wrong direction (Jing et al. 2021).

Beyond object manipulation, other attack modalities aim to interfere with the cameras directly. Light-based attacks aim to alter the lighting conditions around the camera, potentially causing crucial visual information to be overlooked or misinterpreted. For example, directing a strong light source at the camera can result in temporary blindness (Petit et al. 2015; Yan, Xu, and Liu 2016; Fu et al. 2021), or fine-grained laser pointed at the camera can achieve a malicious control of traffic light color (Yan et al. 2022). Acoustic-based manipulation involves broadcasting ultrasound to cause perturbations in the car camera's stabilizer, leading to blurred images, so that the detection is severely degraded (Ji et al. 2021; Zhu et al. 2023). Furthermore, electromagnetic manipulation, although a significant concern, has been previously discussed and is not repeated here.

Simulation of ESIA

This section introduces a novel simulation method for generating adversarial patterns. The similarity between these simulated patterns and those produced by real-world attacks is then verified. Finally, the method is employed to create a dataset of simulated adversarial patterns.

Simulation Method

Normally, a camera transmits its captured images to subsequent processing stages as illustrated in Figure 1. Each im-

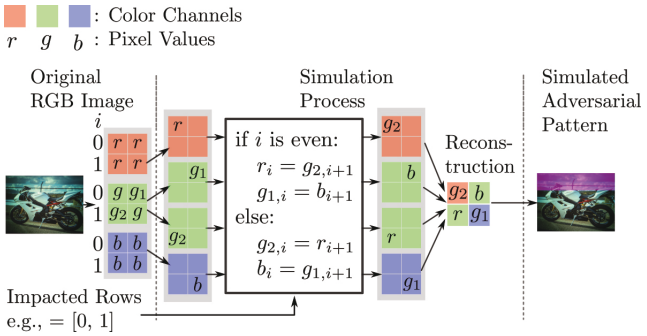


Figure 2: The simulation process generates arbitrary adversarial patterns, i.e., color strips, in an RGB image.

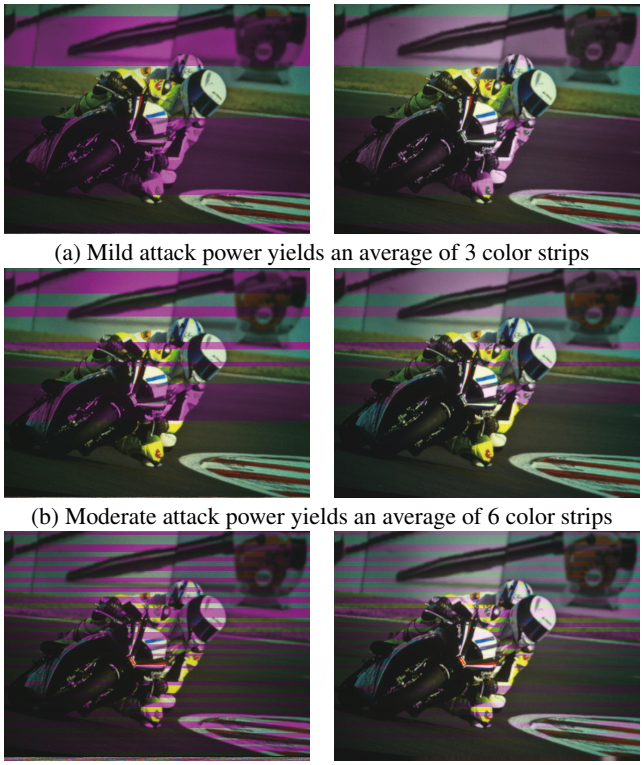
age row is encapsulated as a packet, and these packets are reassembled to reconstruct the image. Previous research (Jiang et al. 2023) has demonstrated that injected malicious signals can induce packet loss, leading to incorrect color interpretation, and consequently, color strips in the reconstructed image.

By carefully analyzing the attack mechanism explained in previous work, we observe that incorrect color interpretation can be characterized by a swap of red/blue and green channel pixel values. We then propose a simulation method, outlined in Figure 2, to mimic this behavior.

The input of our method is an RGB image, with pixel values represented as r, g, and b for red (R), green (G), and blue (B) channels, respectively. To simplify, we visualize the simulation process with two specific rows (two columns for each), termed "Impacted Rows" (row number, $i = 0, 1$). In particular, we simulate the computation based on the values of the upper-left corner in the r channel (denoted as r), the bottom-right corner of the blue channel (denoted as b), and the upper-right corner (g_1) and bottom-left corner (g_2) for green channel respectively.

Assuming neighboring pixel values are similar, we approximated color information using these four values, i.e., r, g_1, g_2, b. For each impacted row with index i, if the row index i is even, the red pixel value (r) is replaced with the green pixel value from the next row (g_2), and the green pixel value (g_1) is replaced with the blue pixel value from the next row (b). If the row index i is odd, the green pixel value (g_2) is replaced with the red pixel value from the next row (r), and the blue pixel value (b) is replaced with the green pixel value from the next row (g_1).

After applying the above transformations to the color channels, the pixels are reassembled to form a new image. The reconstruction step, which is essentially demosaicing that reconstructs color information for all three channels, brings together the modified red, green, and blue channel values to form a new image. The resulting image exhibits an adversarial pattern, as exemplified in the Figure 2 where rows at the top (e.g., 0 and 1) are impacted. The position of the color strip can be controlled by adjusting the selected impacted rows.



(a) Mild attack power yields an average of 3 color strips

(b) Moderate attack power yields an average of 6 color strips

(c) Severe attack power yields an average of 15 color strips

Figure 3: Example real (left) and simulated (right) attack images with different levels of attack severity.

Evaluation of Similarity between Simulated Attack and Real Attack

Zhang et al. (2024a) generated a set of real attack images by re-taking 100 randomly selected images from the COCO2017 (Lin et al. 2014) testing dataset using a physical setup to inject electromagnetic signals during the re-taking process. By adjusting the power of the injected signals, they conducted attacks at different severity levels, producing images under mild, moderate, and severe attacks. This process yields 300 under-attack images in total in their dataset, with three per original image, respective to the three severity levels. According to this real attack image set, the counts of color strips induced by mild, moderate, and severe attacks, fall in the ranges of *[1, 6]*, *[7, 12]*, and *[13, 20]*, respectively.

To validate the efficiency of our proposed ESIA simulation method. We apply our simulation method to the same 100 images, generating another set of 300 under-attack images, again, three per original image, one for each severity level. Specifically, we use the simulated attack method to produce the same number and location of color strips as the real ones (for each image at each severity level) as shown in Figure 3. We then evaluate the consistency of the impact of attacks – the performance degradation they caused, between the images generated by our simulated method and those from real attacks in object detection task, one of the critical traffic-related applications (i.e., traffic object detection).

Severity	ΔmAP50	ΔmAP75	ΔmAP50:95
Mild	0.876	0.803	0.950
Moderate	0.493	0.675	0.704
Severe	0.491	0.373	0.896

Table 1: P-values of t-tests for different levels of ΔmAP

A diverse range of object detection models are selected for the evaluation, to ensure the comprehensiveness and reliability of the results. They include *CNN-based single-stage detectors* (e.g., YOLOv3 (Redmon and Farhadi 2018), YOLOx (Ge et al. 2021), Retinanet (Lin et al. 2017), VFNet (Zhang et al. 2021a), and EfficientDet (Tan, Pang, and Le 2020)), *CNN-based multi-stage detectors* (e.g., Mask R-CNN (He et al. 2017) and Cascade Mask R-CNN (Cai and Vasconcelos 2019)), and *Transformer-based detectors* (e.g., DETR (Carion et al. 2020), DINO (Zhang et al. 2022), Co-DETR (Zong, Song, and Liu 2023), ViT-DET (Li et al. 2022), and SWIN (Liu et al. 2021)).

We calculated the mean Average Precision (mAP) for each model based on the aforementioned data to illustrate the performance differences between simulated attack images and real attack images. Specifically, mAP50 is computed as the average of the Average Precision (AP) across all classes at a single Intersection over Union (IoU) threshold of 0.5. To visually demonstrate the performance degradation across models, we first calculated the mAP values (mAP50, mAP75, and mAP50:95) for each model on unattacked images, real attack images, and simulated attack images. Subsequently, we computed the differences in mAP values (ΔmAP50, ΔmAP75, and ΔmAP50:95) between the attacked and unattacked states for both real and simulated attacks. The results for ΔmAP50 are shown in Figure 4, where (a) represents performance degradation under real attacks and (b) illustrates degradation under simulated attacks. By examining the trend lines in the two image sets, it is evident that as the severity of the attack increases, the ΔmAP values rise, indicating a decrease in mAP values. Moreover, the trend of ΔmAP values under both real and simulated attacks is consistent across different models.

Furthermore, to evaluate whether there is a significant difference between the impacts of simulated and real attacks on AI models, we performed a t-test at a standard significance level of 5% based on the calculated and summarized ΔmAP results for all selected models under both attack types. The p-values for ΔmAP50, ΔmAP75, and ΔmAP50:95 are shown in Table 1. As indicated by the results, all p-values from the t-tests for ΔmAP exceed the 5% threshold, indicating that there is insufficient evidence to suggest significant differences between the two datasets. These tests and analyses demonstrate that images generated by simulated attacks do not differ significantly from those generated by real attacks, and the impact on AI models follows a highly consistent trend.

Traffic Images with Simulated Adversarial Pattern

We then apply the simulation method on a set of traffic images, to simulate attacks that autonomous vehicles might encounter in real world driving with complex environmental conditions. Specifically, we randomly select 10,000 images from the validation set of the BDD100k dataset (Yu et al. 2020), which are well-annotated for traffic object detection and drivable area segmentation tasks.

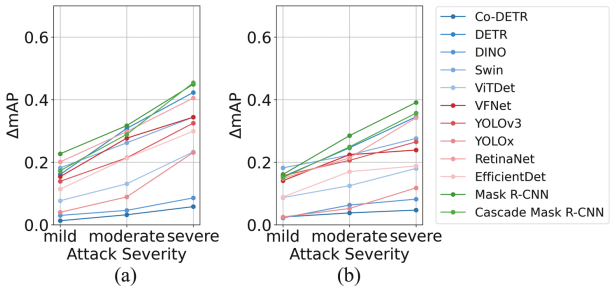


Figure 4: Similar performance (ΔmAP50) between (a) real attack images and (b) simulated attack images across different object detection models.

Weather	Time of Day	Scene
Overcast (1239)		
Clear (5346)	Daytime (5258)	City Street (6112)
Rainy (738)	Night (3929)	Highway (2499)
Snowy (769)	Dawn (778)	Residential (1253)
Partly Cloudy (738)		
Total (8830)	Total (9965)	Total (9864)

Table 2: Distribution of images for three driving environmental groups (The numbers in parentheses indicate the number of images in each category).

We divide the images into three environmental groups with several subcategories based on driving conditions, specifically focusing on weather, scene, and time of day, – three key driving environment variables. To explore the impact of attacks in specific scenarios and ensure the reliability of the experimental results with a sufficient amount of data, we filter out undefined subcategories and some subcategories with fewer images. The number of images for each subcategory after filtering is shown in Table 2.

To investigate the threat of ESIA at different severity levels, we randomly divide the images in each subcategory into four equal attack severity groups: unattacked, mild, moderate, and severe. Following the observations in the real attack dataset (Zhang et al. 2024a), for the mild, moderate, and severe groups, we perform simulated attacks with color strip numbers ranging from *[1, 6]*, *[7, 12]*, and *[13, 20]*, respectively.

Impacts of ESIA in Different Driving Scenarios

This section will explore and discuss the impacts of ESIA on perception tasks across various driving scenarios.

Group	Sub-Category	Model	Traffic object detection (mAP50)				Drivable area segmentation (mIoU)			
			No attack	Mild	Moderate	Severe	No attack	Mild	Moderate	Severe
Weather	Clear	HybridNets	0.755	0.647	0.439	0.250	0.908	0.889	0.842	0.790
		A-YOLOM	0.796	0.696	0.474	0.291	0.914	0.892	0.843	0.791
		YOLOP	0.755	0.648	0.438	0.255	0.913	0.898	0.856	0.807
	Rainy	HybridNets	0.800	0.695	0.501	0.341	0.866	0.864	0.827	0.787
		A-YOLOM	0.835	0.746	0.538	0.378	0.874	0.860	0.824	0.758
		YOLOP	0.782	0.680	0.486	0.337	0.881	0.870	0.827	0.789
	Snowy	HybridNets	0.777	0.680	0.575	0.284	0.901	0.870	0.822	0.785
		A-YOLOM	0.808	0.733	0.517	0.322	0.886	0.863	0.812	0.759
		YOLOP	0.761	0.667	0.467	0.284	0.907	0.878	0.841	0.789
	Partly Cloudy	HybridNets	0.782	0.665	0.414	0.245	0.906	0.880	0.842	0.790
		A-YOLOM	0.813	0.706	0.456	0.291	0.919	0.887	0.827	0.781
		YOLOP	0.768	0.662	0.421	0.257	0.911	0.891	0.858	0.809
	Overcast	HybridNets	0.785	0.702	0.465	0.268	0.904	0.884	0.842	0.790
		A-YOLOM	0.821	0.742	0.498	0.310	0.909	0.895	0.846	0.800
		YOLOP	0.780	0.697	0.458	0.277	0.916	0.902	0.863	0.811
Scene	Highway	HybridNets	0.741	0.576	0.347	0.208	0.915	0.894	0.846	0.790
		A-YOLOM	0.782	0.622	0.379	0.240	0.928	0.897	0.848	0.791
		YOLOP	0.742	0.575	0.347	0.212	0.917	0.895	0.854	0.800
	City Street	HybridNets	0.775	0.692	0.495	0.305	0.897	0.875	0.834	0.789
		A-YOLOM	0.812	0.739	0.530	0.348	0.902	0.880	0.834	0.780
		YOLOP	0.769	0.683	0.489	0.307	0.915	0.896	0.852	0.805
	Residential	HybridNets	0.804	0.701	0.492	0.302	0.907	0.895	0.840	0.795
		A-YOLOM	0.837	0.748	0.541	0.349	0.919	0.901	0.844	0.791
		YOLOP	0.798	0.703	0.486	0.303	0.913	0.893	0.849	0.796
Time of Day	Night	HybridNets	0.740	0.630	0.453	0.262	0.903	0.884	0.839	0.796
		A-YOLOM	0.787	0.676	0.487	0.299	0.910	0.888	0.842	0.794
		YOLOP	0.745	0.632	0.445	0.264	0.908	0.894	0.855	0.811
	Dawn	HybridNets	0.787	0.716	0.440	0.263	0.897	0.883	0.835	0.792
		A-YOLOM	0.824	0.756	0.469	0.304	0.894	0.888	0.844	0.780
		YOLOP	0.777	0.718	0.428	0.267	0.907	0.896	0.843	0.800
	Daytime	HybridNets	0.796	0.686	0.473	0.283	0.906	0.884	0.836	0.790
		A-YOLOM	0.829	0.727	0.513	0.327	0.907	0.884	0.834	0.782
		YOLOP	0.783	0.680	0.465	0.289	0.917	0.896	0.852	0.801

Table 3: Vulnerability of models to different environmental conditions in traffic object detection and drivable area segmentation.

Tasks and Models

We select two critical tasks, which are object detection and drivable area segmentation (Feng et al. 2020). Object detection is an important task to identify a specific object in an image and determine its location. The detection not only needs to identify the categories of objects in the image but also needs to mark the exact position of each object in the image. Drivable area segmentation is one of the key tasks in autonomous driving designed to identify and segment drivable areas in road scenarios. The goal of this task is to detect areas where the vehicle is safe to travel in order to aid in path planning, obstacle avoidance, and decision-making for autonomous vehicles.

Considering that multi-task models offer better resource efficiency compared to single-task models and are more practical for resource-limited devices, we select three state-of-the-art multi-task models: HybridNets (Vu, Ngo, and Phan 2022), A-YOLOM (Wang, Wu, and Zhang 2024), and YOLOP (Wu et al. 2022). Specifically, HybridNets is based on EfficientNet (Tan and Le 2019), a CNN architecture known for delivering comparable performance to other CNN architectures with fewer resources. A-YOLOM, based on YOLOv8, excels in detection and segmentation within the real-time-oriented YOLO family. YOLOP utilizes a Spatial Pyramid Pooling (SPP) (He et al. 2015) module for feature generation and fusion. Previous studies (Zhang et al. 2021b; Ranjan et al. 2019) have demonstrated that the SPP module is robust against adversarial attacks, and we wonder whether it also demonstrates robustness in the context of ESIA.

For evaluation metrics, we follow standard practices: we measure mAP50 for traffic object detection and mIoU for drivable area segmentation. Average Precision is computed as the area under the precision-recall curve. The Intersection over Union (IoU) metric is used to evaluate the drivable area, and mIoU is the average IoU for each class. To ensure a fair comparison between different models, we follow the evaluation setup specified in the work of three selected models, which means a confidence threshold of 0.001 and a Non-Maximum Suppression (NMS) threshold of 0.6 for all three models. We use the best-performing pretrained models of HybridNets, A-YOLOM, and YOLOP on the BDD100k dataset. It is worth noting that all three pre-trained models focus exclusively on four classes – car, bus, truck, and train during training, and combine these classes into a single "vehicle" classification for the detection task.

Group	Sub-Category	Threat on traffic object detection			Threat on drivable area segmentation		
		D_S_{mild}	D_S_{moderate}	D_S_{severe}	D_S_{mild}	D_S_{moderate}	D_S_{severe}
Weather	Clear	-13.68%	-41.43%	-65.52%	-2.05%	-7.09%	-12.69%
	Rainy	-12.28%	-36.93%	-56.34%	-1.03%	-5.45%	-10.95%
	Snowy	-11.37%	-33.55%	-62.09%	-3.08%	-8.13%	-13.41%
	Partly Cloudy	-13.97%	-45.38%	-66.47%	-2.85%	-7.63%	-13.01%
	Overcast	-10.28%	-40.46%	-64.20%	-1.76%	-6.53%	-12.02%
Scene	Highway	-21.74%	-52.65%	-70.89%	-2.68%	-7.68%	-13.73%
	City Street	-10.29%	-35.76%	-59.29%	-2.32%	-7.15%	-12.53%
	Residential	-11.78%	-37.76%	-60.92%	-1.82%	-7.52%	-13.03%
Time of Day	Night	-14.71%	-39.06%	-63.72%	-2.02%	-6.80%	-11.76%
	Dawn	-8.29%	-44.03%	-65.11%	-1.15%	-6.52%	-12.08%
	Daytime	-13.09%	-39.77%	-62.70%	-2.42%	-7.62%	-13.08%

Table 4: The threat of simulated attack on traffic object detection (i.e., degradation in detection accuracy in mAP) and drivable area segmentation (i.e., degradation of segmentation accuracy in mIoU) across different driving environmental conditions.

Model	Group	Threat on traffic object detection			Threat on drivable area segmentation		
		D_M_{mild}	D_M_{moderate}	D_M_{severe}	D_M_{mild}	D_M_{moderate}	D_M_{severe}
HybridNets	Weather	-13.09%	-38.61%	-64.45%	-2.17%	-6.89%	-12.08%
	Scene	-15.26%	-42.70%	-65.00%	-2.02%	-7.32%	-12.68%
	Time of Day	-12.57%	-41.15%	-65.21%	-2.03%	-7.24%	-12.12%
A-YOLOM	Weather	-11.06%	-39.06%	-60.95%	-2.33%	-7.76%	-13.61%
	Scene	-13.36%	-40.54%	-61.59%	-2.58%	-8.11%	-14.07%
	Time of Day	-11.55%	-39.77%	-61.89%	-1.87%	-7.04%	-13.09%
YOLOP	Weather	-12.80%	-40.99%	-63.37%	-1.96%	-6.25%	-11.54%
	Scene	-15.20%	-42.91%	-64.51%	-2.22%	-6.92%	-12.53%
	Time of Day	-11.97%	-41.93%	-64.43%	-1.68%	-6.66%	-11.71%

Table 5: The threat of simulated attack on traffic object detection (i.e., degradation in detection accuracy in mAP) and drivable area segmentation (i.e., degradation of segmentation accuracy in mIoU) across different models.

Experimental Results and Analysis

In the experiments, driving scenarios are defined by three categories: weather, scene, and time of day, with subcategories such as residential, city street, and highway for the scene, as detailed in Table 2. The results for each category are presented in Table 3. It can be observed that model performance significantly deteriorates with increasing attack intensity across all categories. For example, in the HybridNets model under the clear subcategory, the mAP50 for detection decreases by 10.80%, 31.60%, and 50.50% under mild, moderate, and severe intensity attacks, respectively, compared to the unattacked condition. Correspondingly, the mIoU for segmentation drops by 1.9%, 6.6%, and 11.8% under the same conditions.

To clearly illustrate the impact of attack across different driving scenarios and models, we further calculate the percentage decrease in performance compared to the "No Attack" condition, referred to as Degradation. Furthermore, We calculate D_S_i by averaging the Degradation across the three models for the same attack intensity within the same subcategory and D_M_i by averaging the Degradation across all subcategories for the same attack intensity within the same model, where $_i$ represents attack intensity, i.e., mild, moderate, severe. The results are presented in Table 4 and Table 5, respectively.

It is observed that: (1) For the weather category, in the traffic object detection task, all three models are less affected by simulated attacks in the snowy subcategory compared to other subcategories. Conversely, in the drivable area segmentation task, the models experience less impact from attacks in the rainy subcategory. (2) For scene types, regardless of whether the task is traffic object detection or drivable area segmentation, all three models tend to experience the least impact from attacks in the city street subcategory. This may be attributed to the more complex and cluttered nature of city streets, which compels the models to focus on finer details, thereby making them less susceptible to attacks. In contrast, highway scenarios, which typically feature simpler and more predictable environments, are most affected by attacks. This observation highlights how the simplicity of an environment can exacerbate the impact of ESIA on model performance. (3) For model robustness, all three models are significantly affected. In the comparison among models, A-YOLOM experiences the least impact in traffic object detection, while YOLOP is the least affected in drivable area segmentation.

Potential Driving Risks

The above experiments focus on changes in model metrics. We know that not every erroneous prediction by a model will lead to driving risks, though it may cause a decline in model metrics. Does a decline in model metrics pose risks in real-world driving scenarios? In the following parts, we discuss cases that could lead to driving risks.

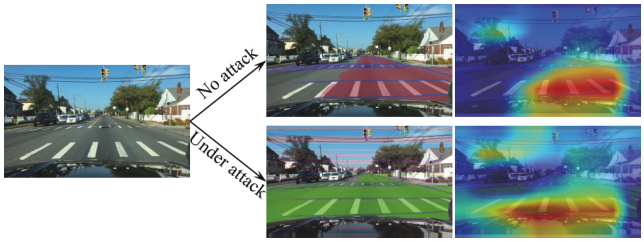


Figure 5: Model attention variations across attack intensities in the case of "driving against traffic".

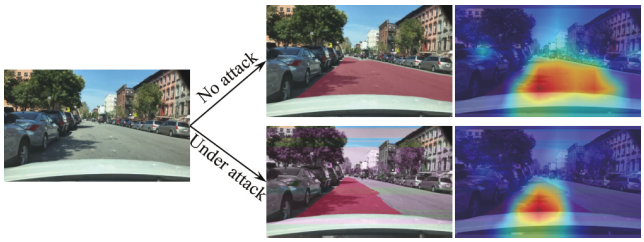


Figure 6: Model attention variations across attack intensities in the case of "drivable area reduction".

We visualize the model prediction results for images in the "no attack" and "under attack" conditions, with prediction parameters consistent with evaluation parameters. For traffic object detection, previous studies have demonstrated issues like hiding and creating objects in the presence of attacks (Jiang et al. 2023; Zhang et al. 2024b), and we do not repeat them here. In this part, we focus on cases related to drivable area segmentation. We identify two potential driving risk scenarios: (1) Driving against traffic: After the attack, the detection of reverse lanes changes from non-drivable to drivable. (2) Drivable area reduction: After the attack, the model detects a significant reduction in the drivable area.

To further investigate the changes in model attention related to the two potential driving risk scenarios caused by simulated attacks, we apply Grad-CAM (Selvaraju et al. 2017) to the feature extraction layer (e.g., bifpn[5].conv6-down) close to the segmentation head. In Figure 5, it can be observed that after the attack, the model tends to focus more on erroneous areas, which indicates a shift in the model's attention from relevant features to misleading ones, thereby compromising its performance and reliability. However, as shown in Figure 6, the model's focus on the correct regions decreases, indicating that the attack causes a significant reduction in attention towards relevant features, even at mild attack intensity. Interestingly, as the attack intensity increases, the model tends to revert to its pre-attack state. We speculate that the higher attack intensity affects more modules within the model simultaneously, leading to a counteracting effect that diminishes the overall impact of the attack. We will explore this phenomenon further in our future work.

Future Work

In future work, we plan to extend our methodology to evaluate the impact of integrating data streams from multiple sensors (e.g., cameras, LiDARs, radars) on end-to-end models under electromagnetic signal injection attacks (ESIA). A key focus will be examining whether the compromise of camera systems introduces security vulnerabilities when multiple sensor modalities are present. Additionally, we aim to broaden our evaluations to include fully autonomous systems, assessing their resilience in real-world scenarios. Lastly, we will investigate mitigation strategies for ESIA, designing more robust and resilient systems.

Conclusion

In this paper, we systematically explore the previously unaddressed vulnerabilities of the perception system in autonomous vehicles to electromagnetic signal injection attacks (EISA) in various driving scenarios. An efficient method is developed to simulate adversarial patterns, generating a set of under-attack traffic images for the analytics in this and future work. Our extensive experiments reveal that model performance degrades to varying extents under attacks, with our observations highlighting severe safety risks such as contraflow driving. This further underscores the significance of evaluating AI models' robustness and reliability in the face of potential adversarial attacks.

This study establishes a foundation for future investigation of techniques to enhance the robustness of AI models (in autonomous driving systems) against such attacks, thereby strengthening security and providing passengers with reliable and safe services. It also showcases a simulation and evaluation framework for other security- or safety-critical applications under potential attacks, shedding light on future research on AI robustness.

Acknowledgements

This work was, in part, supported by the Hong Kong Polytechnic University under grant P0048514, and Shenzhen University.

References

Badue, C.; Guidolini, R.; Carneiro, R. V.; Azevedo, P.; Cardoso, V. B.; Forechi, A.; Jesus, L.; Berriel, R.; Paixao, T. M.; Mutz, F.; et al. 2021. Self-driving cars: A survey. *Expert systems with applications*, 165: 113816.

Cai, Z.; and Vasconcelos, N. 2019. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1483–1498.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

Chen, J.; Jordan, M. I.; and Wainwright, M. J. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, 1277–1294. IEEE.

Duan, R.; Ma, X.; Wang, Y.; Bailey, J.; Qin, A. K.; and Yang, Y. 2020. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1000–1008.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1625–1634.

Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; and Dietmayer, K. 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3): 1341–1360.

Fu, Z.; Zhi, Y.; Ji, S.; and Sun, X. 2021. Remote Attacks on Drones Vision Sensors: An Empirical Study. *IEEE Transactions on Dependable and Secure Computing*, 19(5): 3125–3135.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guesmi, A.; Hanif, M. A.; Ouni, B.; and Shafique, M. 2023. Physical adversarial attacks for camera-based smart systems: Current trends, categorization, applications, research challenges, and future outlook. *IEEE Access*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9): 1904–1916.

Ji, X.; Cheng, Y.; Zhang, Y.; Wang, K.; Yan, C.; Xu, W.; and Fu, K. 2021. Poltergeist: Acoustic Adversarial Machine Learning against Cameras and Computer Vision. In *2021 IEEE Symposium on Security and Privacy (SP)*, 160–175. IEEE.

Jiang, Q.; Ji, X.; Yan, C.; Xie, Z.; Lou, H.; and Xu, W. 2023. GlitchHiker: Uncovering Vulnerabilities of Image Signal Transmission with IEMI. In *32nd USENIX Security Symposium (USENIX Security 23)*, 7249–7266.

Jing, P.; Tang, Q.; Du, Y.; Xue, L.; Luo, X.; Wang, T.; Nie, S.; and Wu, S. 2021. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, 3237–3254.

Kang, D.; Zhang, Y.; Tam, W. C.; and Fu, E. Y. 2024. Anti-ESIA: Analyzing and Mitigating Impacts of Electromagnetic Signal Injection Attacks. *arXiv preprint arXiv:2409.10922*.

Köhler, S.; Baker, R.; and Martinovic, I. 2022. Signal Injection Attacks against CCD Image Sensors. In *Proc. 2022 ACM ASIA Conference on Computer and Communications Security (ACM ASIACCS 2022)*. ACM.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, 280–296. Springer.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, N.; Nikitas, A.; and Parkinson, S. 2020. Exploring expert perceptions about the cyber security and privacy of Connected and Autonomous Vehicles: A thematic analysis approach. *Transportation research part F: traffic psychology and behaviour*, 75: 66–86.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ljubi, K.; and Groznik, A. 2023. Role played by social factors and privacy concerns in autonomous vehicle adoption. *Transport policy*, 132: 1–15.

Nastjuk, I.; Herrenkind, B.; Marrone, M.; Brendel, A. B.; and Kolbe, L. M. 2020. What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. *Technological Forecasting and Social Change*, 161: 120319.

Petit, J.; Stottelaar, B.; Feiri, M.; and Kargl, F. 2015. Remote attacks on automated vehicles sensors: Experiments on camera and lidar. *Black Hat Europe*, 11(2015): 995.

Ranjan, A.; Janai, J.; Geiger, A.; and Black, M. J. 2019. Attacking optical flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2404–2413.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Song, D.; Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Tramer, F.; Prakash, A.; and Kohno, T. 2018. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790.

Vu, D.; Ngo, B.; and Phan, H. 2022. Hybridnets: End-to-end perception network. *arXiv preprint arXiv:2203.09035*.

Wang, J.; Wu, Q. J.; and Zhang, N. 2024. You only look at once for real-time and generic multi-task. *IEEE Transactions on Vehicular Technology*.

Wu, D.; Liao, M.-W.; Zhang, W.-T.; Wang, X.-G.; Bai, X.; Cheng, W.-Q.; and Liu, W.-Y. 2022. Yolop: You only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6): 550–562.

Yan, C.; Xu, W.; and Liu, J. 2016. Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle. *Def Con*, 24(8): 109.

Yan, C.; Xu, Z.; Yin, Z.; Mangard, S.; Ji, X.; Xu, W.; Zhao, K.; Zhou, Y.; Wang, T.; Gu, G.; et al. 2022. Rolling Colors: Adversarial Laser Exploits against Traffic Light Recognition. In *31st USENIX Security Symposium (USENIX Security 22)*, 1957–1974.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021a. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8514–8523.

Zhang, J.; Lou, Y.; Wang, J.; Wu, K.; Lu, K.; and Jia, X. 2021b. Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. *IEEE Internet of Things Journal*, 9(5): 3443–3456.

Zhang, Y.; Cheung, M.; Yang, C.; Zhai, X.; Shen, Z.; Ji, X.; Fu, E. Y.; Chau, S.-Y.; and Luo, X. 2024a. Modeling Electromagnetic Signal Injection Attacks on Camera-based Smart Systems: Applications and Mitigation. *arXiv preprint arXiv:2408.05124*.

Zhang, Y.; Yang, C.; Fu, E. Y.; Jiang, Q.; Yan, C.; Chau, S.-Y.; Ngai, G.; Leong, H.-V.; Luo, X.; and Xu, W. 2024b. Understanding Impacts of Electromagnetic Signal Injection Attacks on Object Detection. *2024 IEEE International Conference on Multimedia and Expo (ICME)*.

Zhu, W.; Ji, X.; Cheng, Y.; Zhang, S.; and Xu, W. 2023. TPatch: A Triggered Physical Adversarial Patch. In *32nd USENIX Security Symposium (USENIX Security 23)*, 661–678.

Zong, Z.; Song, G.; and Liu, Y. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6748–6758.