Future Digital Technologies and Artificial Intelligence

Open Access Journal

ISSN 3070-0965

November 2025| Volume 01 | Issue 02 | Pages 08-18

https://doi.org/10.55670/fpll.fdtai.1.2.2

Journal homepage: https://fupubco.com/fdtai

Future Publishing LLC

Review

A study on adversarial attacks in Deep Learning-based traffic signal recognition for autonomous vehicles

Sheik Murad Hassan Anik[1], Yolguly Allaberdiyev[1], Sharmin Afrose[2], Tahsin Mullick[3], Fatih Karabiber[1]*

[1]Department of Computer Science and Computer Information Systems, Auburn University at Montgomery, Montgomery, AL 36117, USA
[2]Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA
[3]Department of Systems and Information Engineering, University of Virginia, VA 22903, USA

ARTICLE INFO

Article history:
Received 02 June 2025
Received in revised form
12 July 2025
Accepted 25 July 2025

Keywords:
Adversarial attack, Autonomous vehicle, Traffic light detection, Image classification, ImageNet, Inception-v3

*Corresponding author
Email address:
fkarabib@aum.edu

DOI: 10.55670/fpll.fdtai.1.2.2

ABSTRACT

Autonomous vehicles are gradually occupying the streets and are expected to become ubiquitous in the near future. However, recent incidents involving these vehicles have raised serious concerns about their safety, particularly regarding the reliability of their onboard machine learning systems. In this paper, we expose a critical yet underexplored vulnerability—misclassifying street signs as traffic lights—by conducting a targeted white-box adversarial attack. To the best of our knowledge, this specific vulnerability has not been addressed in the existing literature. We craft adversarial examples using the Fast Gradient Sign Method (FGSM) to generate minimal perturbations that can deceive a state-of-the-art image classification model, Inception-V3, trained on the ImageNet dataset. We also introduce a custom dataset consisting of real-world street sign and traffic light images to test the attack under more domain-specific conditions. Our evaluation metrics include attack success rate, Structural Similarity Index (SSIM), and L2 distance, with our method achieving a 100% success rate in misclassification. These results highlight the pressing need to design robust defenses against adversarial attacks in safety-critical systems. We further discuss technical challenges, potential defenses such as adversarial training and obfuscated gradients, and directions for future research to enhance the resilience of deep learning systems in autonomous vehicles.

1. Introduction

Autonomous vehicles have gained considerable attention in recent years due to their potential to enhance transportation safety and efficiency. Several Autonomous vehicles, including Tesla Autopilot, Volkswagen and Audi's Traffic Jam Pilot, Ford Argo AI, Daimler Intelligent Drive, etc., are increasingly visible on public roads. These driverless vehicles are designed to provide numerous options to the drivers with regard to safety as well as comfort. Autonomous vehicles can sense their surrounding environment and classify objects (pedestrians, traffic lights, street signs, etc.) to make decisions and take appropriate actions. It is highly essential for the manufacturers to integrate a model architecture that can classify objects accurately. Otherwise, misclassification can lead to erroneous decisions, potentially resulting in accidents that jeopardize human safety. Consequently, the development of robust and accurate object classification models is essential for the safe deployment of autonomous driving systems. It is challenging to build an accurate classifier model. On the initiation of an adversarial attack, the task of accurate classification becomes more challenging. Adversarial attacks involve the intentional manipulation of input data, typically through subtle modifications and perturbations, that lead a classifier to produce incorrect outputs. For example, a minor alteration to an image of a traffic sign can cause a model to misclassify it as a different sign, despite the changes being imperceptible to the human eye. Recent studies [1, 2] have demonstrated the effectiveness of such adversarial perturbations in deceiving traffic sign

recognition systems. The challenges of creating adversarial images are that a minimal amount of perturbation has to be done so that the changes are not visible to the human eye, yet the classifier model of the autonomous car misclassifies the images.

In this study, the adversarial attack is implemented on street signs to be misclassified as traffic lights and deceive the autonomous car. To accomplish the task, the Inception-v3 model [3] is utilized. Specifically, the input image is fed into a convolutional neural network (CNN), and the classification loss is computed at the softmax layer. The gradient descent optimizer calculates the perturbation from the loss function. The perturbation is added in the target image and re-evaluated by the CNN to assess the effect of the adversarial modification. We evaluate the performance of the generated image using L2 distance and the Structural Similarity Index (SSIM). Figure 1 presents the classifier's correct prediction on the unaltered image, while Figure 2 illustrates the misclassification result after applying the crafted adversarial perturbation using the same model. Visually, there may be no discernible difference between the original and adversarial images. The main contributions of our work are summarized as follows:

- We designed a novel application case of generating adversarial examples of a street sign that is misclassified as a traffic light.
- Our generated perturbations are so small that they are imperceptible to the human eye. The evaluation result of SSIM and L2 distance shows the quality of adversarial images to confirm the high visual similarity between the original and adversarial images.
- We further perform transfer learning to develop a custom image classifier. This model is trained and tested on the customized dataset.

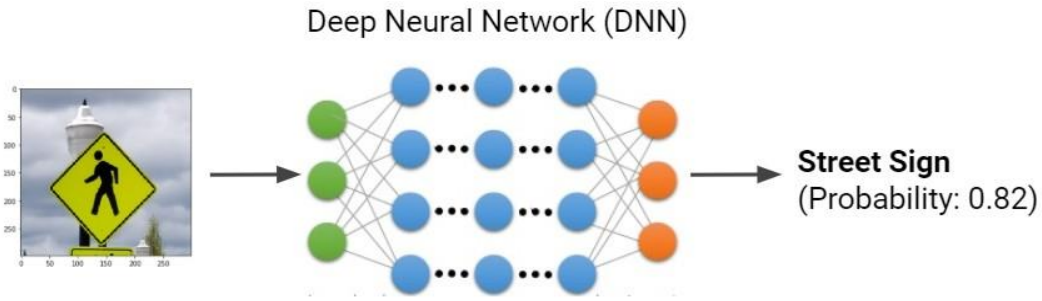


Figure 1. Regular image classification

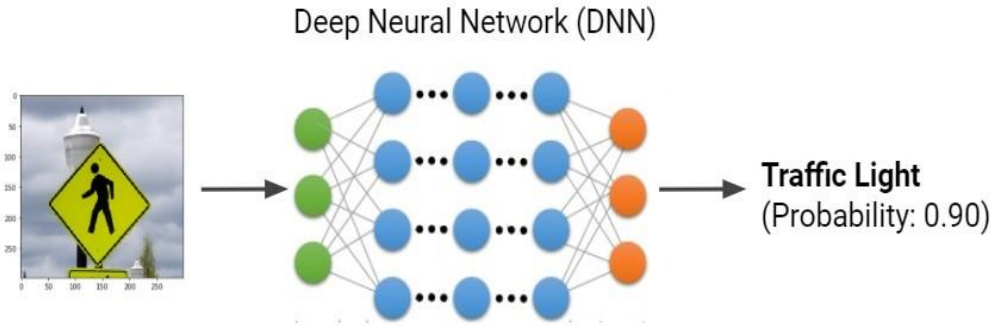


Figure 2. Misclassification due to an adversarial sample with no visible perturbation

2. Motivation

While autonomous vehicles are an imminent future, they are also open to a wide array of security threats. These systems are equipped with multiple sensors to perceive and interpret their surroundings. A failure in any component—particularly within the learning algorithms responsible for perception and decision-making—can result in critical, potentially life-threatening consequences. This project is motivated to expose these threats, particularly with respect to generating and utilizing adversarial perturbations that can severely affect the learning algorithms. The main aim of this study is to encourage the development of more robust and secure machine learning models for autonomous systems by demonstrating the effectiveness of these attacks.

3. Related works

Generation of adversarial examples relies on two important components: first, the accurate detection of traffic lights and street signs, and second, the generation of targeted perturbation. The problem of detecting traffic signs and lights has been extensively studied in the literature,

with numerous robust algorithms developed to leverage recent advancements in computer vision. These detection systems serve as the foundation upon which adversarial attacks can be constructed, underscoring the importance of both accurate perception and security in autonomous driving applications. Detection methods for traffic signs and lights can broadly be categorized into two approaches: image processing-based methods and machine learning-based methods. The image processing methods primarily rely on visual cues such as shape and color for detection. A work done by Linderet al. [4] and Franke et al. [5] on traffic lights focused on the classification of the color of each pixel and followed it with the usage of component analysis for segmentation to locate regions of interest. In the context of traffic sign detection, Athrey et al. [6] employed thresholding and blob detection combined with template matching. Omachi et al. [7] grouped pixels that exceeded a specific threshold by normalizing the color space of images. Image processing techniques offer advantages based on the fact that they are not data dependent, and thus do not suffer from overfitting and work well in specific scenarios. They are, however, prone to face challenges under slight variability, which can limit their robustness in real-world scenarios.

The lack of robustness in image processing approaches to variable scenarios can be effectively addressed through learning-based models. The advantage of learning-based models is that they can be trained on a broad set of traffic lights or signs under different lighting conditions in various environments. For instance, Aggregate Channel Feature (ACF) based detectors presented by Morten et al. [8] and Philipsen et al. [9] outperformed image processing detectors on the LISA dataset. One notable deep learning model, YOLO (You Only Look Once), developed by Redmon et al. [10], is an algorithm that is used to detect traffic lights, which includes a separate CNN to classify states of traffic lights. Similarly, traffic sign detection research makes use of a variety of learning based models, such as the multi-scale CNN proposed by Sermanet and LeCun [11]. Support Vector Machines (SVMs) coupled with CNNs to develop finer classification as presented in Yang et al. [12]. Pon et al. [13] presented a way to enable the detection of both traffic lights and traffic signs using a combined data set. Their deep hierarchical architecture, one of the first networks to perform joint detection on traffic light and traffic sign, incorporates a mini-batch proposal selection mechanism to solve overlapping issues of the dataset in their proposed method.

Once the challenge of detecting traffic lights and signs is addressed, the next step in creating adversarial examples is the generation of perturbations. One of the seminal works in this field was by Szegedy et al. [14], where they discovered several models, including state-of-the-art neural networks, being vulnerable to adversarial examples and misclassification. Their method of generating perturbations involved taking advantage of discontinuities in the input-output mappings of deep neural networks. The perturbations in their work were designed to maximize the network's prediction error. They were also able to show that the same perturbation could be applicable to multiple different networks, which were trained on different subsets of the dataset, misclassifying the same input. This paper led to many other studies where researchers have applied adversarial examples to different applications to expose security vulnerabilities. Amongst the work done, Eykholt et al. [2] propose a general attack algorithm, Robust Physical Perturbation (RP2), which was robust under diverse physical conditions. Their algorithm was tested on real-world road sign classification. In another study, Chawin et al. [15] showed that traffic sign recognition could be compromised when physically printed perturbations were overlaid on traffic signs.

4. The proposed approach

This section outlines the approach taken by the project and serves to present a brief overview of the experimentation section, which subsequently delves into extensive details. Additionally, we present the threat model for the project, highlighting the assumed constraints and the intended targets of attack.

4.1 Methodology

Developing a model capable of successfully misclassifying street signs requires a robust learning-based approach. In this work, we employ the Inception-v3 architecture, which has been pre-trained on the ImageNet dataset. Inception-v3 was selected for its strong performance and deep architecture, which enables it to extract fine-grained features from images. The ImageNet dataset, comprising a wide range of classes, includes images captured under various lighting conditions and orientations. These characteristics make it well-suited for our purposes in generating adversarial examples. As shown in Figure 3, in Step 1, an image from our custom dataset is passed to the Inception-v3 model. After being processed through the layers of the convolutional network, the softmax layer produces class probabilities in Step 2. These probabilities represent the model's confidence in the input image belonging to each class. In Step 3, this softmax output is also used to extract values relevant for generating perturbations, forming the basis for crafting adversarial examples. As shown in Figure 3 in Step 1, an image from our own data set is passed on to the Inception-v3. After being processed through the layers of the convolutional network, the softmax layer produces class labels with probabilities in Step 2. These probabilities represent the model's confidence in the input image belonging to each class. In Step 3, the softmax layer output is also used to extract values relevant for generating perturbations, forming the basis for crafting adversarial examples. If the loss is below a predefined threshold, the perturbation is considered complete. Otherwise, the values are passed to the Fast Gradient Descent optimizer, as illustrated in Step 4. In Step 5, the optimizer generates perturbations based on the specified parameters. Finally, in Step 6, these perturbations are superimposed onto the original image in an attempt to deceive the convolutional neural network (CNN) into misclassifying it as a different class. An amplified visualization of the perturbation is shown in Figure 4.

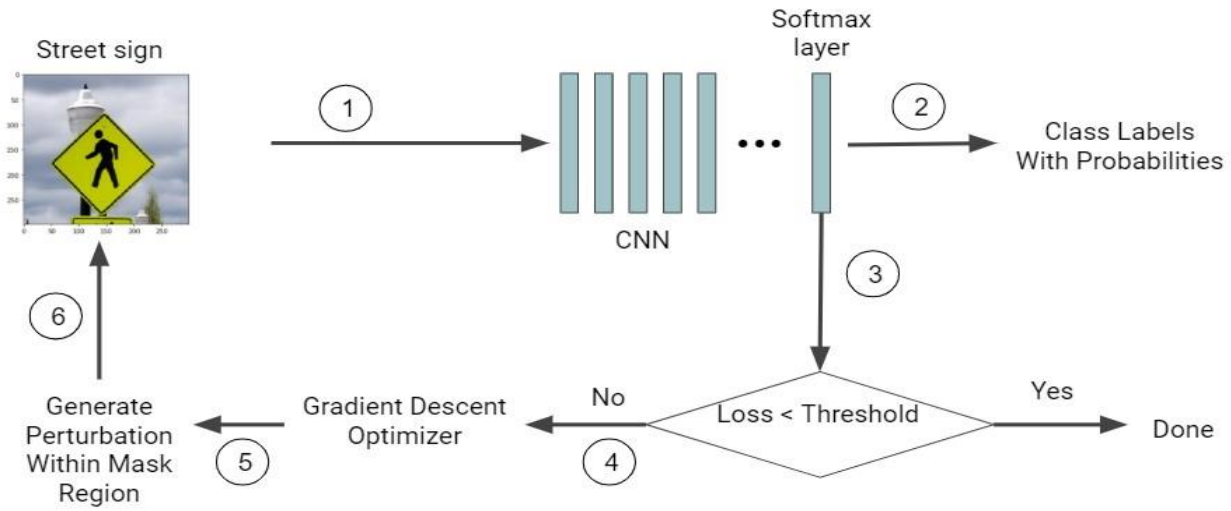


Figure 3. System architecture overview

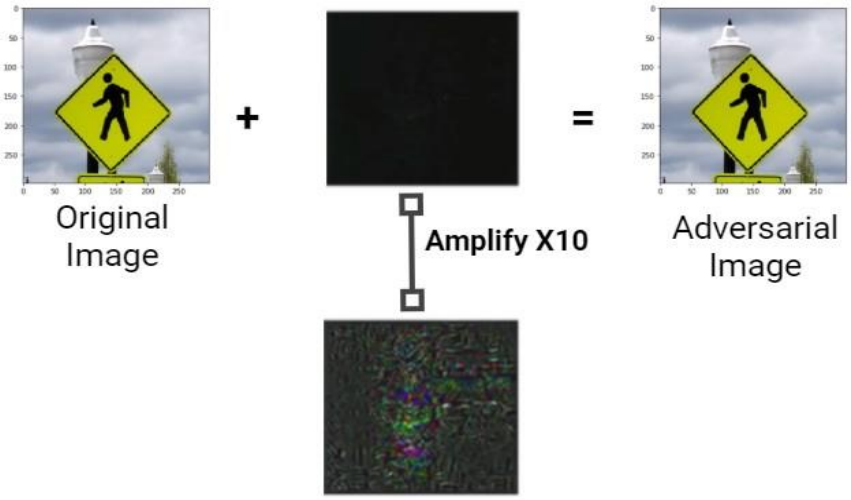


Figure 4. Perturbation added to the street sign

4.2 Threat model

This work, presented in this study, aims to highlight potential security vulnerabilities in image classification systems by crafting adversarial examples. Below, we define our threat model, outlining the type of attack conducted, the constraints imposed on the adversary, and the specific objectives of the attack. For our adversarial attack, the attacker is bound by some constraints. The threat model discussion is as follows:

- White-box attack scenario: We assume a white-box scenario where the attacker has full knowledge of the detection model's architecture.
- Modification restriction: Even though the adversary has white-box access to the detection system, we will maintain the model modification restriction, meaning the attacker will not be able to change the detection model or any of its parameters.
- Targeted attack: We only consider the targeted attack, where the attacker classifies street sign images as traffic lights.
- Training data poisoning: The adversary cannot poison the image classification system by altering training data.

5. Technical challenges

In this section, we outline several technical challenges encountered during the project and how we addressed them in the process of crafting adversarial samples for an image classification model.

- Computational Capabilities: Deep neural networks, which work as the backbone of image classification systems, typically require a large amount of data and significant processing power for training the model. They require multiple Graphics Processing Units (GPU) for parallel computation. However, due to hardware constraints, we conducted this project in a CPU-only environment. Since the training requires months to complete, we used a pre-trained model for our approach.
- Pre-trained Model: We employed the pre-trained Inception-v3 [3] model for classification. Although this model is not specifically trained for traffic lights and street signs, it performed considerably well on our test images. We initiated designing our custom model because we were unable to find a dedicated pre-trained model for traffic lights and street signs.
- Transfer Learning: For our custom classifier, we adopted Inception-v3 as our teacher model. During development, we encountered difficulties in saving the student model as a checkpoint. We addressed this issue by generating the student model in graph format, which preserved its ability to accurately classify various street signs and traffic lights.

6. Experiment

In this section, we discuss details of our approach for crafting adversarial samples aimed at fooling an image classifier for detecting street signs. We begin by describing the dataset used in our experiments, followed by an overview of the classification model employed. We then explain the data pre-processing steps necessary to prepare the input for training and testing. After that, we introduce the custom dataset we created and the model we built for more targeted classification. We also describe the test data used to evaluate the model's performance under adversarial conditions. Next, we present the attack method used to generate adversarial samples. We then define the evaluation metrics applied to measure the effectiveness of our approach. Finally, we discuss the results obtained from our experiments.

6.1 Data description

In this project, we used the ImageNet [16] dataset, which contains over 14 million labeled images spanning 1,000 different object categories. Our classifier was trained on all 1000 classes of ImageNet. The images in the dataset were collected from various sources and come in different resolutions and formats, contributing to the diversity and robustness of the training data. Among the 1000 classes, classes #919 and #920 denote street signs and traffic lights, respectively. Although ImageNet is a large and general-purpose dataset, we assume that it may be used in a specific manner for the navigation of autonomous vehicles. Autonomous vehicle manufacturers do not reveal their training and testing datasets to minimize security threats, but there are only a limited number of large datasets and classifiers available. So, it is an easy assumption that the vehicle manufacturers extend the work of these datasets and classifiers for their own cause. ImageNet is one of the largest available public datasets, and for this reason, we decided to work with this dataset in our project. Table 1 indicates the specifications of the dataset and classifier used.

Table 1. Parameter specifications of the experiment

Parameters	Specification	
Architecture	Inception-v3	Custom Classifier
Dataset	Google ImageNet	Custom Dataset (Blacksburg)
Dataset Size	15 million	1,250
Image Size	Variable	300 X 300

6.2 Classification model

In this project, we utilized the Inception-v3 [3] image classification model developed by Google and trained on the ImageNet dataset [16]. It has 1000 classes including street signs and traffic lights which are of our primary focus of our study. The model is a 42-layer deep convolutional neural network where the input layer takes the images as input and the final softmax layer produces the probability distribution of the input image across the 1000 classes. Table 1 summarizes the key specifications of the dataset and the classification model used. Given the size and complexity of the ImageNet dataset, training a deep network from scratch would be computationally intensive and time-consuming—potentially requiring months without access to high-performance hardware. To address this, we employed a pre-trained version of Inception-v3, which has already been trained on the full ImageNet dataset. Figure 5 and Figure 6 illustrate the classification performance of the model. As shown, the model is capable of classifying both traffic lights and street signs with high confidence. These results confirm the model's appropriateness for use in our adversarial attack experiments.

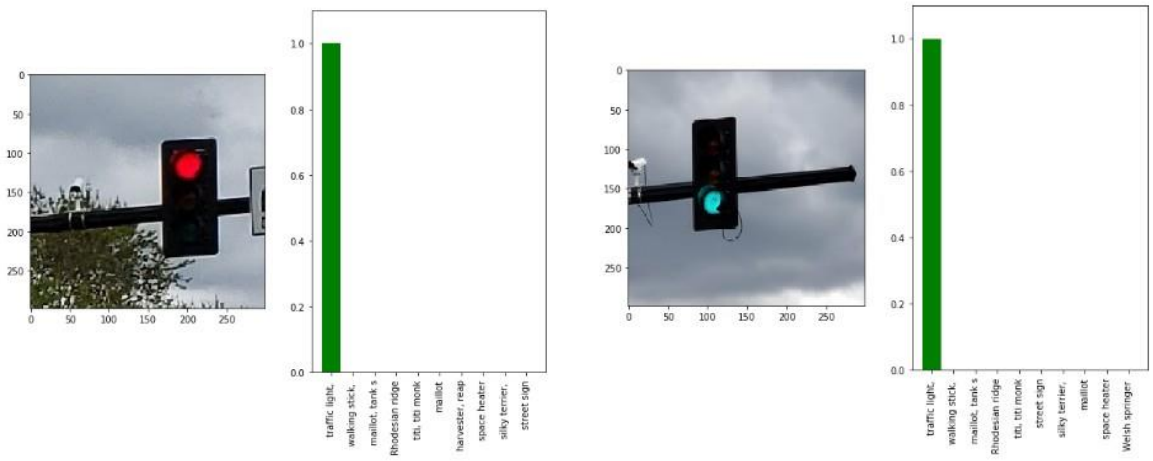


Figure 5. Correct classification of traffic light using original image

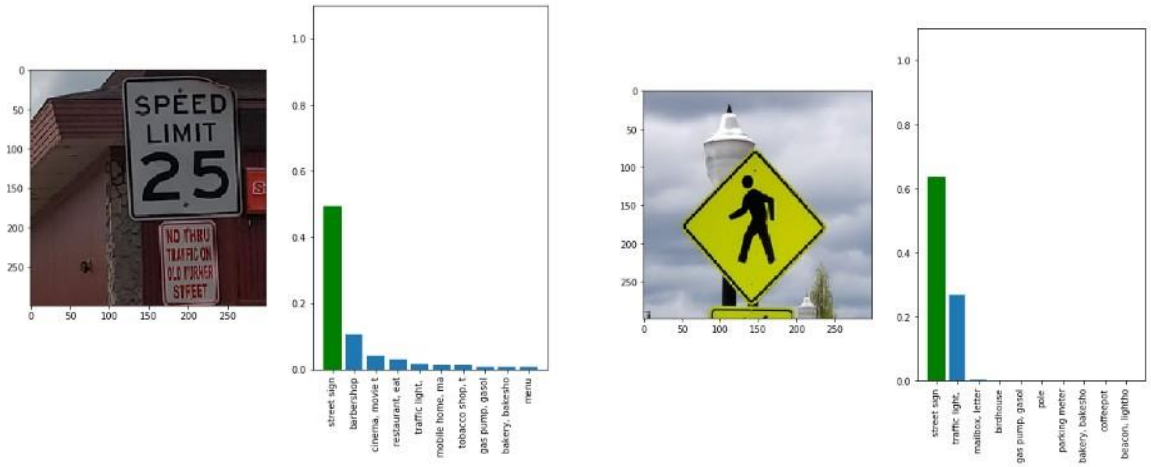


Figure 6. Correct classification of the street sign using the original image

6.3 Data pre-processing

As mentioned in the previous section, the ImageNet dataset contains over 14 million images, and our classifier is trained on these images with 1,000 classes. Since these images originate from various sources, they have different resolutions and formats. To make these images workable with our model and to ensure compatibility, all input images were resized to a standard resolution of 300×300 pixels**.** We implemented this resizing step directly in our code. Although this approach alters the original aspect ratio of the images, it was necessary to conform to the input requirements of the classification model. Despite this limitation, resizing was the only practical method for ensuring that all images could be processed consistently by the Inception-v3 architecture.

6.4 Custom dataset & model

As explained earlier, manufacturers of autonomous vehicles tend to use their proprietary datasets and classifiers for their vehicles to minimize security risk. For this reason, we also decided to build up our own dataset and classifier specifically for the task of classifying street signs and traffic lights. The processing of building the custom dataset and classifier is still at the initial phase. We collected over 1200 images from the city of Blacksburg. These images span five distinct classes: red light, green light, yellow light, 25 MPH sign, and pedestrian walk sign. Each class contains more than 200 images. We plan to expand the dataset by increasing both the number of images per class and the total number of classes in the future. To support the dedicated classification of street signs and traffic lights, we designed a custom classifier using transfer learning. The Inception-v3 model was employed as the teacher model of transfer learning. As our dataset is still growing, we

are in the process of fine-tuning the classifier to improve its accuracy and robustness. All classification and adversarial attack results presented in this paper are based on our custom-collected dataset. The source images shown in the classification results in Figures 5 and 6 are taken directly from our custom dataset.

6.5 Test data

To evaluate the effectiveness of our adversarial attack method, we tested the classification model using our custom-built dataset. The test images were collected from real-world environments in the city of Blacksburg, ensuring a variety of natural lighting conditions and backgrounds. The evaluation focused specifically on images of traffic lights and street signs, consistent with the classes defined in our dataset. This real-world testing approach allowed us to assess the model's performance under practical and diverse conditions.

6.6 Attack method

For crafting adversarial samples of traffic lights from street signs to mislead image classification systems, we attempted several well-known attack methods of adversarial image generation, including Gradient Attack [17], L2 Iterative Attack, and C&W Attack [18]. However, due to limited computational resources—specifically, the lack of access to GPU acceleration—these methods proved too slow and computationally intensive on our system. So, we adopted the Fast Gradient Sign Descent Method (FGSM) [19] to generate adversarial samples for our attack. FGSM is a single-step, gradient-based attack that perturbs the input data in the direction of the gradient of the loss function with respect to the input. This makes it both computationally efficient and effective at generating adversarial examples that can fool deep learning models with minimal perturbation. The algorithm uses an optimization-based approach to generate adversarial perturbation. The single-image optimization problem searches for a perturbation δ to be added to the input x, such that the perturbed instance $x = x+\delta$ is misclassified by the target classifier $f_\theta(.)$:

$$min\ H(x+\delta,x),\ s.t.\ f_\theta(x+\delta) = y^* \tag{1}$$

where H denotes the $lp - norm$ of the distance function, and y^* is the target class. We only consider targeted attacks in this project. The above optimization problem is reformulated in the Lagrangian-relaxed form as:

$$\text{argmin}_\delta\ \lambda||\delta||_p + \quad J(f_\theta(x+\delta) = y^*) \tag{2}$$

Here, $J(...)$ is the Jacobian-based loss function, which measures the difference between the target label y^* and the model's prediction. λ controls the regularization and distance function H is specified here as $||\delta||_p$, denoting the l_p norm of δ.
For our attack, we include two stopping criteria, (i) a loss value threshold of 0.0001 and (ii) 3500 as maximum number of iterations. So, if the adversarial image gets classified as the original class even after 3500 iterations, then it is called as a failure.

6.7 Evaluation matrices

In this section, we discuss the evaluation metrics used to assess the effectiveness of our proposed adversarial attack on image classification systems. The attack success rate will be evaluated using the following equation [14]:

$$successRate = \frac{imagesMisclassifiedDueToAttack}{imagesClassifiedCorrectlyBeforeAttack} \tag{3}$$

Our second evaluation is based on the Structural SIMilar measure (SSIM) [20], which measures the similarity between the original and the adversarial image. The SSIM is the similarity metric for the perturbed image and the actual image. DSSIM (Structural Dissimilarity) is a distance metric derived from SSIM (Structural SIMilarity). The basic form of SSIM compares three aspects of the two image samples, luminance (l), contrast (c), and structure (s). The SSIM score is then given in the following equation:

$$SSIM(x,y) = l(x,y) * c(x,y) * s(x,y) \tag{4}$$

Here, x and y represent the original and adversarial images, respectively. We will also use the L2 distance [21] for our evaluation. That can be expressed as:

$$D_{L_2}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5}$$

It is the Euclidean distance between the two images. In our case, we pass the original and the adversarial image in the function. In this scenario, achieving a lower L2 distance is desirable, as it indicates that the adversarial example remains closer to the original input. A smaller L2 distance typically results in more subtle perturbations, making the adversarial image harder to detect while still being effective.

7. Results

To evaluate the effectiveness of our approach, we generated 10 adversarial examples of street signs of class #919 to be classified as traffic lights of class #920. All crafted adversarial images were successful in fooling the classification model. As a result, the attack success rate, as defined by equation (3), is 100%**,** indicating complete success in fooling the model. As mentioned earlier, we used both the Structural Similarity Index (SSIM) [20] and the L2 distance [21] to evaluate the similarity between the original source image and the corresponding adversarial image. The results of these evaluations are summarized in Table 2. Additionally, Figure 7 illustrates the variation of our SSIM

scores over different source images and their corresponding adversarial images. Figure 8 presents the L2 distance measures for various original-adversarial image pairs. These distances help quantify the difference between the source image and the perturbed adversarial image.

Figure 7. Evaluation result of 10 adversarial images in terms of SSIM

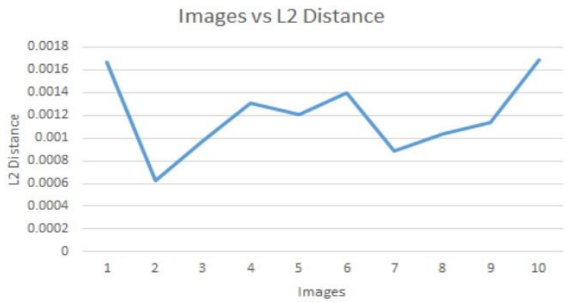


Figure 8. Evaluation Result of 10 Adversarial Images in Terms of L2 Distance

Table 2. Result summary

Image #	SSIM	L2 Distance
1	0.999983	0.00167
2	0.999997	0.000624
3	0.999993	0.000978
4	0.999988	0.001306
5	0.999991	0.001203
6	0.999988	0.001401
7	0.999995	0.000882
8	0.999993	0.001033
9	0.999993	0.001134
10	0.999981	0.001688

We also include some examples of classification and attack results in this report. Figure 5 and Figure 6 show the classification results before the attack. As seen, the model can accurately classify traffic lights and street signs with high confidence. Figure 9 shows the convergence of loss against the number of iterations for our attack. It is clearly visible that the loss approaches near zero around the 2500th iteration. We set the attack to stop when the loss reached a threshold of 0.0001 or after 3500 iterations, whichever came first. Figure 10 (a) shows the attack results of our method. Here we can see, a pedestrian walk sign is being classified as a traffic light with nearly 100% confidence. Figure 10 (b) displays the generated perturbation for the attack. The leftmost image of 10 (b) is the original perturbation. Since the perturbation is not visible at this scale, we first amplified it 10 times (middle image) and then 100 times for better comprehensibility.

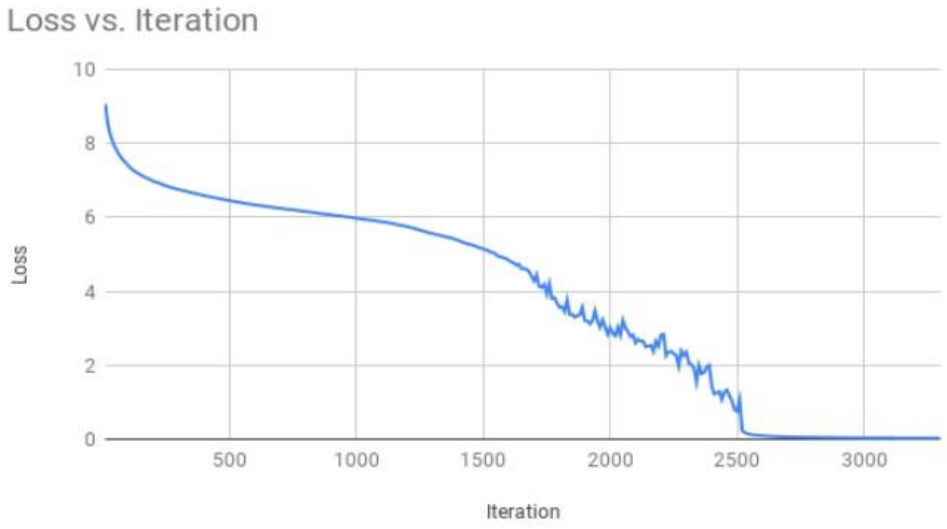
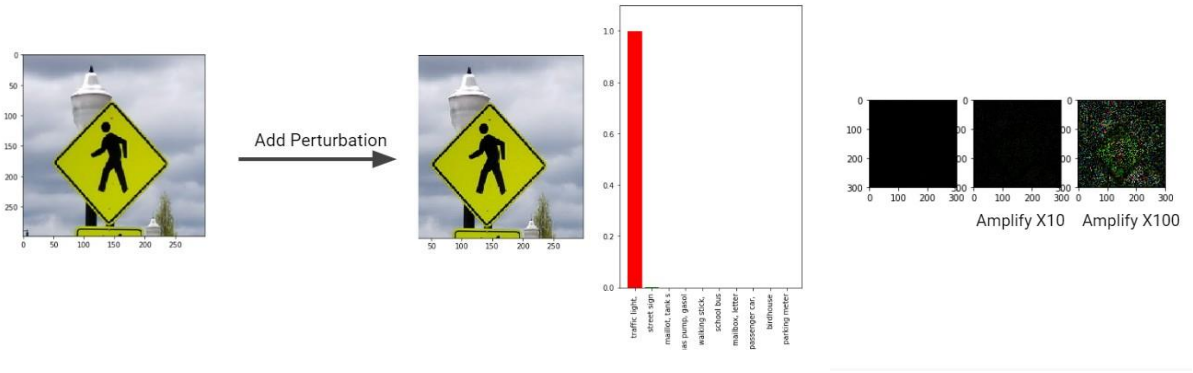


Figure 9. Convergence of adversarial attack



(a) Original Image Transform to Adversarial Image (b) Perturbation

Figure 10. Incorrect classification of a street sign using an adversarial image

8. Discussion

The results presented earlier clearly show that our crafted adversarial samples were consistently misclassified into the targeted class with high confidence, while maintaining a high Structural Similarity Index (SSIM) and low L2 distance from the original images. This highlights the alarming effectiveness of such attacks, even when executed with minimal computational resources and within a short time frame. The implications of this are significant, especially for real-time, safety-critical applications like autonomous driving. In this section, we explore potential defense strategies that could mitigate the impact of such adversarial attacks and discuss the limitations of our current approach.

8.1 Possible defenses

To address adversarial threats similar to the one proposed in our study, we suggest several defense mechanisms that can be implemented at various stages of the model pipeline:

- Adversarial Training: One effective method is to augment the training dataset with synthetically generated adversarial examples. This approach enables the classifier or a preceding detector module to learn and recognize adversarial patterns, enhancing its resilience. A two-step system, where a detector flags adversarial inputs before classification, can add an additional layer of protection.
- Shape and Color Constraints: Traffic lights and road signs have distinct shapes (e.g., circular, octagonal) and standardized colors (e.g., red, green, yellow). Enforcing constraints based on these physical characteristics can help validate whether an input image conforms to expected real-world features before passing it to the classifier. This validation step can effectively narrow the decision space and reduce susceptibility to manipulated inputs.

- Gradient Obfuscation: Introducing gradient obfuscation techniques can obscure the loss landscape seen by attackers, thereby making gradient-based attacks—like FGSM—less effective. By disrupting the clarity of the gradient information, the attacker's optimization process becomes less predictable and less efficient.
- Noise Filtering and Image Preprocessing: Employing classical image processing techniques such as denoising filters, Gaussian blurring, or compression-based methods can help eliminate subtle perturbations introduced by adversarial attacks. These preprocessing steps act as a form of input sanitization that can neutralize minor manipulations without significantly degrading image quality.

8.2 Limitations

Although our attack is 100% successful for this test dataset, one important thing is that our dataset is small, and some hyperparameters needed to be tuned a lot for some of the input images to converge. This particular attack is not robust, which means it may or may not work in different environmental situations, like rain, snow, or different angles and distances. We plan to address these issues in future work to improve the attack's robustness. The attack was also performed on the entire image, which is not practical in a real-world scenario. So, for future work, we will include masks on source images to constrain the spatial region of perturbation. In addition, the notion of inconspicuousness is subjective, and the only way to quantify it adequately requires incorporating human-subject studies. We did SSIM (Structure Similarity Index) and L2 distance for that measure. In future work, we plan to work on making this attack more robust and more compatible with the real world. So that we can perturb different angles of an image and generate a uniform perturbation for that image.

9. Conclusion

In this project, we crafted adversarial samples to attack the traffic light detection and recognition systems likely to be employed in self-driving or autonomous vehicles. The industry of autonomous vehicles is rapidly advancing, and the accurate detection of traffic lights is critical not only for the safety of the vehicle and its passengers but also for the protection of pedestrians and surrounding infrastructure. By targeting such a vital component, our work sheds light on a key vulnerability that needs urgent attention from both researchers and industry stakeholders. Our results demonstrate that such adversarial attacks are feasible and, if successfully executed in real-world settings, could lead to catastrophic consequences, including accidents and loss of life. This highlights the pressing need to anticipate potential threats in the AI pipeline of autonomous systems. The broader vision of this project is to raise awareness and encourage proactive development of robust security measures before such threats manifest on public roads. We also discussed the potential to extend our attack model to generate more resilient and transferable adversarial examples, which could be even more difficult to detect and defend against. We shed light upon some promising defense mechanisms that may be effective against our proposed approach, although no single method can safeguard against all types of adversarial attacks. Hence, we strongly advocate for ongoing collaborative research efforts to build comprehensive and adaptive defense strategies tailored to real-world autonomous driving applications.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically concerning authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere in any language.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the corresponding author.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: deceiving autonomous cars with toxic signs," CoRR, vol. abs/1802.06430, 2018. [Online]. Available: http://arxiv.org/abs/1802.06430

[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. X. Song, "Robust physicalworld attacks on deep learning visual classification," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1625–1634, 2018.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[4] F. Lindner, U. Kressel, and S. Kaelberer, "Robust recognition of traffic signals," in IEEE Intelligent Vehicles Symposium, 2004, June 2004, pp. 49–53.

[5] U. Franke, D. Gavrila, S. Gorzig, F. Lindner, F. Puetzold, and C. Wohler, "Autonomous driving goes downtown," IEEE Intelligent Systems and their Applications, vol. 13, no. 6, pp. 40–48, Nov 1998.

[6] K. S. Athrey, B. M. Kambalur, and K. K. Kumar, "Traffic sign recognition using blob analysis and template matching," in Proceedings of the Sixth International Conference on Computer and Communication Technology 2015, ser. ICCCT '15. New York, NY, USA: ACM, 2015, pp. 219–222. [Online]. Available: http://doi.acm.org.ezproxy.lib.vt.edu/10.1145/2818567.2818609

[7] M. Omachi and S. Omachi, "Traffic light detection with color and edge information," in 2009 2nd IEEE International Conference on Computer Science and Information Technology. IEEE, 2009, pp. 284–287.

[8] M. B. Jensen, M. P. Philipsen, C. Bahnsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Traffic light detection at night: Comparison of a learning-based detector and three model-based detectors," in Advances in Visual Computing, G. Bebis, R. Boyle, B. Parvin, D. Koracin, I. Pavlidis, R. Feris, T. McGraw, M. Elendt, R. Kopper, E. Ragan, Z. Ye, and G. Weber, Eds. Cham: Springer International Publishing, 2015, pp. 774–783.

[9] M. P. Philipsen, M. B. Jensen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Traffic light detection: A learning algorithm and evaluations on challenging dataset," in 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Sep. 2015, pp. 2341–2345.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 779–788.
[11] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in The 2011 International Joint Conference on Neural Networks, July 2011, pp. 2809–2813.
[12] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 7, pp. 2022–2031, July 2016.
[13] A. Pon, O. Adrienko, A. Harakeh, and S. L. Waslander, "A hierarchical deep architecture and mini-batch selection method for joint traffic sign and light detection," in 2018 15th Conference on Computer and Robot Vision (CRV), May 2018, pp. 102–109.
[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," CoRR, vol. abs/1312.6199, 2013. [Online]. Available: http://arxiv.org/abs/1312.6199
[15] C. Sitawarin, A. N. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang, "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos," CoRR, vol. abs/1801.02780, 2018. [Online]. Available: http://arxiv.org/abs/1801.02780
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
[17] K.-H. Chow, L. Liu, M. Loper, J. Bae, M. E. Gursoy, S. Truex, W. Wei, and Y. Wu, "Adversarial objectness gradient attacks in real-time object detection systems," in 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). IEEE, 2020, pp. 263–272.
[18] B. Aksoy and A. Temizel, "Attack type agnostic perceptual enhancement of adversarial images," arXiv preprint arXiv:1903.03029, 2019
[19] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 39–57.
[20] Z. Wang, "The ssim index for image quality assessment," https://ece. uwaterloo. ca/˜ z70wang/research/ssim, 2003.
[21] L. Baccour and R. I. John, "Experimental analysis of crisp similarity and distance measures," in 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR). IEEE, 2014, pp. 96–100.