WIP: Adversarial Retroreflective Patches: A Novel Stealthy Attack on Traffic Sign Recognition at Night

Go Tsuruoka*, Takami Sato†, Qi Alfred Chen †,
Kazuki Nomoto*‡, Yuna Tanaka*, Ryunosuke Kobayashi*, Tatsuya Mori*§¶,
*Waseda University † University of California, Irvine
‡Deloitte Tohmatsu Cyber LLC §RIKEN AIP ¶NICT

***Abstract*—Traffic signs, essential for communicating critical rules to ensure safe and efficient traffic for entities such as pedestrians and motor vehicles, must be reliably recognized, especially in the realm of autonomous driving. However, recent studies have revealed vulnerabilities in vision-based traffic sign recognition systems to adversarial attacks, typically involving small stickers or laser projections. Our work advances this frontier by exploring a novel attack vector, the Adversarial Retroreflective Patch (ARP) attack. This method is stealthy and particularly effective at night by exploiting the optical properties of retroreflective materials, which reflect light back to its source. By applying retroreflective patches to traffic signs, the reflected light from the vehicle's headlights interferes with the camera, causing perturbations that hinder the traffic sign recognition model's ability to correctly detect the signs. In our preliminary study, we conducted a feasibility study of ARP attacks and observed that while a 100% attack success rate is achievable in digital simulations, it decreases to less than or equal to 90% in physical experiments. Finally, we discuss the current challenges and outline our future plans. This research gains significance in the context of autonomous vehicles' 24/7 operation, emphasizing the critical need to assess sensor and AI vulnerabilities, especially in low-light nighttime environments, to ensure the continued safety and reliability of self-driving technologies.**

I. Introduction

Traffic signs, essential for communicating critical traffic rules, are fundamental to the safety and efficiency of all road users, including pedestrians and motor vehicles. Their compliance is particularly critical in autonomous driving to prevent significant physical and human damages. To recognize traffic signs, the vision-based traffic sign recognition systems [1], [2] are widely adopted in many vehicles. However, many recent studies [3]–[7] have actively reported that vision-based traffic sign recognition systems could be vulnerable to against adversarial attacks with malicious stickers [2]–[4], visible light projection [5], [6], [8] or IR laser projection attacks [7]. While these attacks are effective and stealthy in their targeted scenarios, each attack has complemental pros and cons. Patch attacks [3], [4] are always visible to everyone, and pedestrians or road guards may notice and remove them. Light or laser projection attacks [5]–[8] can be effective only for the target victim by selectively turning on and off the light or laser. However, this implies that the attacker, or at the very least the equipment utilized in the attack, must be present in close proximity to the target traffic sign.

Symposium on Vehicles Security and Privacy (VehicleSec) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-7-6
https://dx.doi.org/10.14722/vehiclesec.2024.23025
www.ndss-symposium.org

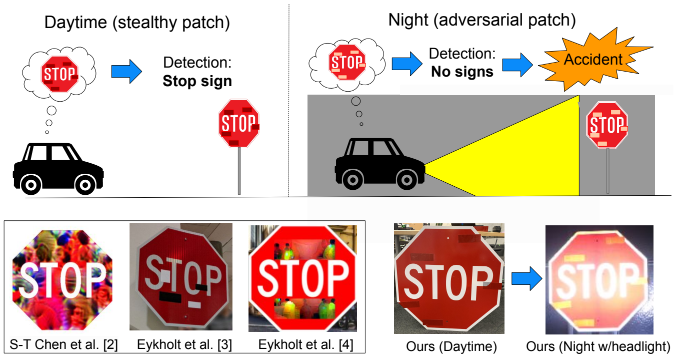


Fig. 1: Overview of the ARP Attack. The patch becomes visible only when the headlights shine on it, and the attack takes place. Our attacks are highly stealty at daytime.

To advance the frontier and address limitations in existing adversarial attack research, we introduce the Adversarial Reflective Patch (ARP) as a novel attack vector. ARP employs stealthy reflective patches, transparent or matching the background color, that are imperceptible during the day but become visible when exposed to vehicle headlights at night. The ARP uses retroreflective materials that are specially designed to reflect light back to its source. This property enables the ARP to create adversarial patterns that are only visible under nighttime illumination, effectively misleading traffic sign detection or classification systems while maintaining a high degree of stealth in daylight conditions. The covert nature of these patches, which only become visible under strong illumination in the dark, greatly enhances their stealthiness, making it a versatile and easy-to-deploy tool for challenging autonomous vehicle security.

In our study, we thoroughly examined the effectiveness of ARP, using both digital and real-world experiments. We first focused on outlining ARP's threat model and its optimization strategies. The impact of ARP on traffic sign recognition was then carefully evaluated using the YOLOv3-tiny model [9] trained on the COCO dataset [10], where we achieved a success rate of up to 100% in digital simulations. The study was subsequently extended to real-world scenarios, applying ARP to actual traffic signs and observing a success rate of up to 90%. Furthermore, the paper discusses future research directions, particularly in developing defense methods against

ARP and evaluating its broader implications for automated vehicle systems.

Our work-in-progress paper explores the feasibility of ARP attack during nighttime, its primary operational scenario due to its reliance on light reflection. The relevance of our research is underscored by an important development in the field: starting in August 2023, the California Public Utilities Commission granted GM Cruise [11] and Alphabet's Waymo [12] authorization to operate unmanned self-driving taxi cabs 24/7 in San Francisco [13]. In an era where 24/7 operation of autonomous vehicles is now commonplace, the critical need to assess sensor and AI vulnerabilities, especially in nighttime environments, is even more pronounced. This shift emphasizes the importance of our research in ensuring the continued safety and reliability of self-driving technologies.

II. Background and Related work

A. *Vision-Based Traffic Sign Recognition*

In the field of autonomous driving, Traffic Sign Recognition (TSR) is a critical safety feature that involves the real-time identification of traffic signs from images captured by vehicle-mounted cameras. This is primarily achieved through advanced object detection models based on Deep Neural Networks (DNNs), which are an integral part of Level 4 autonomous driving technologies and Advanced Driver Assistance Systems (ADAS) found in brands such as Tesla [14] and Toyota [15]. These systems enhance vehicle safety by ensuring that traffic signs are accurately detected and acted upon.

DNN-based TSR models are designed to quickly and accurately detect specific traffic sign in images or video feeds. They take input from the vehicle's cameras and output the classification of the identified object along with its location in the form of bounding boxes. These bounding boxes include the object's position and a confidence score that represents the likelihood of the object's presence in the scene. An object is confirmed to be present if its confidence score exceeds a certain threshold, resulting in the display of a bounding box. Conversely, objects with confidence scores below this threshold are ignored and no bounding box is reported, effectively ignoring objects that are considered unlikely to be present.

B. *Reflective Materials*

Reflection can be categorized into three distinct types: 1) diffuse reflection, 2) specular reflection, and 3) retroreflection. [16]–[18] Table I shows a comparison of three reflection properties. "Single Direction" indicates whether the reflection is limited to a single direction, and "Direction" indicates the specific direction of the reflection. Diffuse reflection is characterized by its ability to emit light of equal intensity in multiple directions, regardless of the angle of incidence. In contrast, specular reflection occurs when light is reflected in a manner that the angle of incidence and the angle of reflection are equal. Finally, retroreflection is characterized by the property of strongly reflecting incident light back in its original direction as shown in Fig. 2. This type of reflective material is mainly used for traffic signs. In this study, we attach patches with reflective properties to traffic signs and the color of the patch is the same as that of the traffic sign. Therefore, regardless of the type of reflection, when the patch is not illuminated, it appears natural to humans. On the other hand, when the headlight hits the reflective patch, light is reflected according to the reflective characteristics of the reflective patch. As a result of this reflection, the reflected light enters the camera of the automatic vehicle, resulting in a perturbation. This perturbation leads to problems in traffic sign recognition, specifically causing phenomena such as detection avoidance and misclassification.

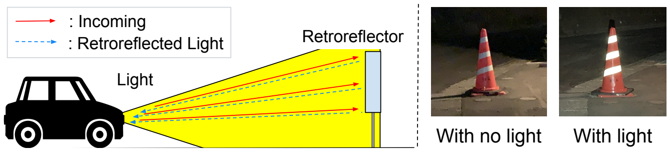


Fig. 2: Retroreflective light path (left) and actual retroreflective material when illuminated (right). The reflected light has the property of returning to the light source.

TABLE I: Comparison of three reflection properties.

	Single Direction	Direction	Intensity
Diffuse reflection	No	Random	Low
Specular reflection	Yes	Directional	High
Retroreflection	Yes	Backward	High

C. *Physical-World Adversarial Attacks against TSR systems*

In this section, we analyze current attack strategies and explore the effectiveness of our proposed approach. Existing attack methods towards object detection models include applying stickers and projecting visible patterns. Nevertheless, these methods have problems with nighttime effectiveness and stealth. In attacks using patches, their color differs from that of traffic signs, which makes them less stealthy. In addition to this, these attacks mainly focus on the daytime and the effectivity of attacks in night time is unclear. Although projector-based attacks such as Phantom Attack [8] and SLAP [5] focus on nighttime attacks, they rely on the use of projectors. This dependence increases the risk of police detection and restricts their effectiveness. Our study proposes a new method that uses reflective patches to address these issues. This method does not require projectors or other equipment and, unlike existing adversarial patch attacks, uses reflective patches of the same color as the traffic signs. This approach achieves effective adversarial patch attacks on DNNs at night while maintaining high stealthiness during the day.

III. Methodology

A. *Threat Model and Attack Goals*

Fig. 1 shows an overview of our ARP attack. We generally follow the similar threat models adopted in prior patch attacks [2]–[4]. The major difference is that the ARP attack patch is enabled by the headlights of the victim vehicles at night so that the patch can be very stealthy at daytime and stealthy to other vehicles or pedestrians even at night since the retroreflective materials reflect the majority of lights back to its light source. Thus, we assume that the adversary can know the camera and headlights used in the victim vehicle, but we do not assume white-box access to the traffic sign recognition model, i.e., the ARP attack is a black-box attack. The ARP attack consists of the following three steps:

Collect the images of the target sign. The attacker selects target traffic signs and takes photos of the target signs at night, with the vehicle's headlights illuminating them. These photos are used in later steps to optimize patch placement.

Generate and deploy the ARP attack patches. Using the obtained images, the attacker optimizes the placement of the patches on the traffic signs. The details of this optimization process are described in the following sections. Following the optimization results, the attacker applies reflective patches to the targeted traffic signs.

Wait for the victim vehicle to pass by the attack sign. Finally, the deployed ARP attack will be automatically enabled by the illumination from the victim's headlights We note that the attacker does not need to be directly present at the location while they may want to stay nearby to observe the consequence of the attack.

The main goal of the ARP attack is to cause misidentification of traffic sign recognition tasks at nighttime, which could lead to accidents.

B. Optimization for Adversarial Reflective Patch Attacks

We present the methodology for optimizing the placement of adversarial reflective patches used in ARP attacks. The optimization process consists of three steps: **Step 1:** Locating traffic signs in images, **Step 2:** dividing the identified signs into a grid, and **Step 3:** searching for optimal grid positions to place patches. Fig. 3 presents the overview of the procedure. While patch design can vary in placement, size, and shape, this paper focuses primarily on optimizing placement for simplicity. The size of each patch is set to one-tenth the height and one-fifth the width of the traffic sign.

Step 1: To locate traffic signs in images, object detection is performed using YOLOv3-tiny, resulting in bounding boxes around the identified signs.

Step 2: The bounding box of each traffic sign identified in step 1 is divided into a grid of 10 vertical by 5 horizontal sections (see Fig. 3). This grid provides potential placement areas for the reflective patches.

Step 3: Assuming the patches reflect white color, image processing is used to simulate reflection and optimize placement. Beam search, a widely-used heuristic search method, is adopted due to the large number of possible patch positions. This approach efficiently balances accuracy and computational complexity. Due to the space limitation, we omit the details of the algorithm. The beam width, b, is an important parameter that determines the breadth of the search. A larger b value expands the search area, potentially improving the results, but also increasing the computational complexity. In this study, we empirically set the beamwidth to $b = 3$.

We note that the approaches described here are only one of many possible strategies. The grid size used in this study was determined empirically. However, there is room for optimization in this area as well, which will be further discussed in future work (§VI).

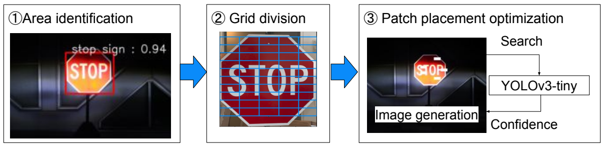


Fig. 3: Overview of optimization for patch placement of Adversarial Reflective Patch Attacks. We use image processing and beam search for search for the best paches' placement.

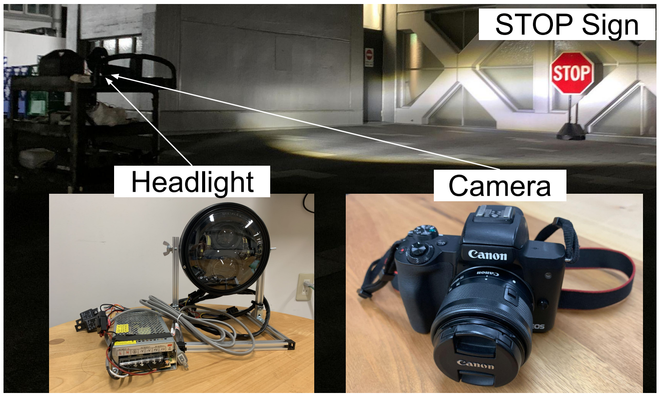


Fig. 4: Experimental setup for taking photos at night.

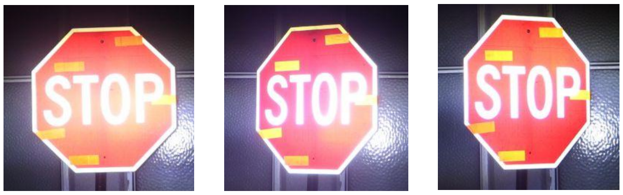


Fig. 5: Physical-World Demo of ARP attacks illuminated by headlights at night.

IV. Evaluation in Digital Space

A. Experimental Setup

As described in §III, our proposed attack begins with capturing images of traffic signs illuminated by headlights at night. In the following, we describe the detail of the equipment we use in the experiment especially in taking photos of the traffic sign. Fig. 4 presents the appearance of the experimental setup in taking the images of traffic signs and examples of photographs taken at this setup. Fig. 5 presents the example of photos of traffic signs with adversarial patches illuminated by headlights. As the targeted traffic sign in this evaluation, we use a STOP sign which is the US standard Manual on Uniform Traffic Control Devices (MUTCD). The size of the sign is 30 inches, and it is constructed of reflective aluminum. For car's headlight, we utilize a 7-inch headlight unit, typically used on motorcycles. This headlight's color temperature is pure white with a range of 6000 K to 6500 K. The light mode offers two options: high beam (3400 LM) and low beam (2800 LM) and we use a low beam. A mirrorless camera is used to collect the STOP sign images. Photographs are taken in auto mode with the focus centered on the STOP sign. The reflective tape to create the adversarial patch described below is red, soft, high-intensity prismatic reflective tape. The tape measures 0.26 mm in thickness and is 50 mm wide.

The headlight's height matches that of the center of the STOP sign, and its optical axis aligns with the center of the STOP sign. Therefore, the distance from the STOP sign to the headlight is the same as that from the STOP sign to the camera. We utilize YOLOv3-tiny trained on the COCO dataset as the object detection model. YOLOv3-tiny is lightweight and has better real-time performance than other models. Therefore, we used YOLOv3-tiny as a potential model for object detection and traffic sign detection used in automated driving systems. The reason we use the model trained on the COCO dataset is this dataset has a "STOP Sign" class. The default confidence threshold for determining the presence of an object is 0.4. In other words, if the confidence level is less than 0.4, the object is assumed not to exist. In this experimental setup, YOLOv3-tiny is able to detect a sign taken from the front at a distance of 5 m from the STOP sign with a high probability of 85%.

Evaluation Procedure We outline the evaluation procedure, which primarily consists of four steps.

Step 1: In this step, we take pictures of STOP signs according to the experimental specifications. This involves varying the distance d and angle (θ) as shown in Fig. 6. We take 20 photographs for each set of d and θ.

Step 2: Here, we optimize the placement of patches. To optimize placement, we conduct this process using a photograph taken under specific conditions where d = 5 and θ = 0.

Step 3: In this step, we apply the patch to the traffic sign at the paste position optimized in Step 2. Please note that the optimized patch position for one photo at d = 5, θ = 0 also applies to photos taken at other distances and angles. The process involves using YOLOv3-tiny to identify the regions of the STOP sign, dividing them into grids, and digitally applying the patch according to the optimized placement from Step 2.

Step 4 With modified image in Step 3 , analyze to investigate whether it can be detected by YOLOv3-tiny and calculate the attack success rate.

B. Results and Observations

Table III (left) shows the observed success rates of ARP attacks at various distances. The data shows that ARP attacks tend to be more successful at the 5-meter and 7-meter marks. In contrast, a lower success rate at the 3-meter distance suggests that the larger image size resulting from being closer to the target may improve the security system's initial detection capabilities, leading to higher confidence in identifying attacks. This result underscores the importance of further research to improve the effectiveness of ARP attacks at different distances.

Table III (right) presents the observed success rates at various angles between the autonomous vehicle and the traffic sign, highlighting how the angle of approach affects the effectiveness. Success rates at an angle of $-30°$ appear to be lower compared to a direct frontal approach or an angle of $30°$. Given that only 20 trials were conducted, this preliminary finding, along with potentially high initial detection confidence, may have played a role in the results. We postulate that the non-symmetric design of the traffic sign may have an influence on the detection process, especially when the angle of attack deviates from frontal. To elucidate these effects, future research could focus on increasing the number of trials and evaluating steeper angles, which we expect would provide a more thorough understanding of how the non-symmetrical patterns of traffic signs affect the angle sensitivity of ARP attacks.

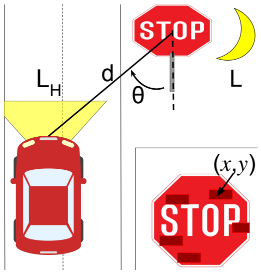


Fig. 6: Overview of variables and parameters of APR attack.

TABLE II: Definition of Variables.

d	Distance: Car↔ sign
θ	Angle: Car ↔ sign
N	Number of Patches
(x, y)	Coordinates of a patch.
L	Intensity of ambient light
L_H	Intensity of Headlight

TABLE III: Distance vs. attack success rate (ASR) with angle fixed at 0 ° (left), and angle vs. ASR with distance fixed at 5 [m] (right).

d [m]	ASR [%]
3	25
5	80
7	100

θ [deg]	ASR [%]
30	80
0	80
-30	10

V. Evaluation in Physical Space

A. Experimental Setup

The procedure for the physical experiments is mostly similar to the digital experiments described in §IV. The difference is in Step 3 and Step 4. In the digital experiment, we reproduced the effects of reflections through image processing. We then tested the generated images for attack feasibility using Yolov3-tiny. In the real-world experiment, we applied the reflective patch to a real traffic sign instead of using reproducing reflection with image processing. Then, we take images of the sign illuminated by headlights and evaluate the feasibility of attack with Yolov3-tiny.

B. Results and Observations

In the empirical results presented in Table IV, we observe a deviation from the results of the digital experiment. The peak attack success rate occurs at a distance of 5 meters and exactly at an angle of 0°, while the success rate decreases as the distance increases to 7 meters. Contrary to the trends observed in the digital domain, the drop in success rate at an angle of -30° is not as drastic. It's important to emphasize that the overall success rates of the attacks are not negligible, suggesting a reasonable level of effectiveness under varying conditions. However, the need to develop more robust attack methods that can withstand variations in both distance and angle is a clear takeaway for future research efforts.

Table V presents the results of the YOLOv3-tiny model's detection capabilities in a non-attack context, serving as a

TABLE IV: Distance vs. ASR (left), and angle vs. ASR (right).

d [m]	ASR [%]
3	50
5	90
7	50

θ [deg]	ASR [%]
30	65
0	90
-30	35

TABLE V: Benchmark detection rates of YOLOv3-tiny without attacks for comparative control.

d [m]	DFR [%]
3	0%
5	15%
7	45%

θ [deg]	DFR [%]
30	0
0	15
-30	0

control benchmark within our experimental parameters and physical experimental setup and the results also show that there is little effect on the detection results when the headlight light strikes the sign and is reflected. There is a noticeable trend: the detection rate decreases as the distance increases. This pattern likely contributes to the declining attack success rates observed at the longer distance of 7 meters. At the closer distance of 3 meters, even though the detection mechanism proves effective, the attack success rate reaches the lower 50% threshold. These findings could be attributed to robustness issues inherent in the detection model's performance, rather than attack design alone, that reduce success rates at different distances and angles. Notably, robust detection is achieved at most angles when not under attack, with the exception of 0°. This highlights the impact that environmental variables such as distance and angle have on the effectiveness of our attack model, and reinforces the need for further investigation into adaptive attack strategies.

VI. Discussions and Future Plans

In this section, we discuss the intricate aspects of ARP attacks using retroreflective materials, with a focus on their implications for autonomous driving systems. The discussion revolves around the evaluation of different reflective materials for their effectiveness in ARP attacks, the realism of their modeling using 3D software, and the optimization of attack strategies. We also address the stealth of these attacks, both through mathematical evaluations and user studies. Finally, we explore the feasibility of these attacks in real-world autonomous driving scenarios and the potential for developing effective software-based defenses.

1) Survey on Reflective Patch Materials: We evaluated the feasibility of ARP attack with a retroreflective material in this WIP paper. We are currently surveying various retroreflective materials and measure their attack capability in the ARP attack. The United States, Japan, and other countries have set standards for retroreflective materials [19], [20]. In the United States, for example, the American Society for Testing and Materials (ASTM) provides guidelines defining the performance characteristics of retroreflective materials used in traffic control devices [19]. Meanwhile, in Japan, the Japanese Industrial Standards (JIS) provide guidelines and standards that regulate the use and specifications of retroreflective materials in a variety of applications [20]. We will conduct a comprehensive survey by utilizing these standards.

Furthermore, we also explore other reflective materials since the original definition of the ARP attack does not limit the attack vector only to retroreflective materials. We plan to perform a large-scale measurement study to evaluate the attack capability of a wide variety of reflective materials in different practical autonomous driving scenarios. We particularly plan to evaluate scenarios under different lighting conditions, which should largely affect the effectiveness of the ARP attack because of the nature of ARP attack enabled by he headlights.

2) Further Realistic Modeling of Reflective Patches: In this WIP paper, we evaluate an naive patch modeling assuming the patch reflection color is aways fixed – white color to confirm the feasibility of ARP attack vector. However, real reflection colors should not be just white and thus we are working on further realistic modeling of the reflective patches. The simulation of the light effects are known as a shader technique. Recent ray-tracing technology enables photo-realistic simulations of lights and its shadow [21]. To leverage the advanced technology in the computer graphics area, we are building a simulation environment on a 3D modeling software such as Blender [22]. Some 3D modeling software programs offer physically based rendering, which enables the reproduction of reflections that are closer to realistic scenarios. This approach is expected to allow the optimization process to be performed under conditions that more closely resemble actual situations as shown in Fig. 7. As discussed in §II, the ARP attack requires to model many types of reflective materials.

To address this issue, we are considering two methods for reproducing reflective patches. The first method involves reproducing the reflective performance based on datasheets. This approach designs reflective patches in 3D simulations based on the information about the reflective characteristics obtained from datasheets related to the surface materials or strength of the light reflected. Datasheets contain detailed information on how specific materials reflect light, and by directly applying this information to the simulation's material parameters, it is possible to reproduce reflections that are close to real scenarios. The advantage of this approach is the ability to conduct highly accurate simulations based on the physical properties of actual materials, assuming that the necessary datasheets are available and that the simulation environment supports these detailed physical characteristics.

The second method aims to approximate the RGB values through optimization. This approach involves optimizing parameters that represent the physical characteristics of the reflective patch, such as roughness, to reproduce realistic reflection characteristics. By using Optuna, a black-box optimization tool, it is possible to efficiently find the optimal combination of these parameters, thereby reproducing the reflection of light on objects in 3D simulations more accurately. This method allows for a more flexible adaptation to various reflective conditions by fine-tuning the parameters to match the observed reflection characteristics closely.

Both approaches offer promising paths to enhance the realism of reflective patch modeling in the context of ARP attacks. By incorporating detailed material characteristics from datasheets or optimizing reflection parameters to match observed phenomena, we can develop more sophisticated and realistic simulations. These improved simulations will not only aid in the understanding and analysis of ARP attacks but also contribute to the development of more effective defense mechanisms against such attacks.

3) Attack Optimization Improvements: We plan to further improve the optimization methodology to explore the most effective patch's location, size, and material for each scenario.

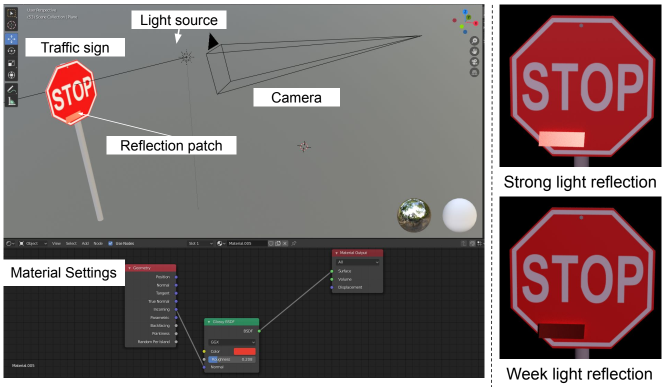


Fig. 7: Simulation of ARP attack in Blender: The left panel illustrates the setup of the attack, while the right panel displays the variations in light reflection intensity and the resulting rendering outcomes.

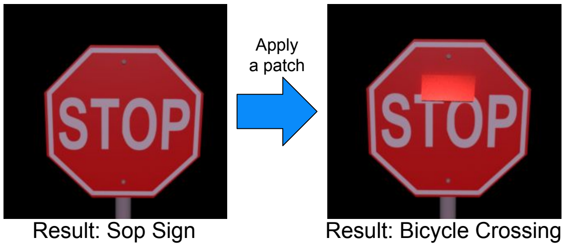


Fig. 8: 3-D simulation of an adversarial patch attack on a traffic sign classifier using Blender: The left image shows a stop sign without a patch, which is correctly classified as a stop sign, while the right image, with a patch applied, misleadingly leads to its classification as a bicycle crossing.

In this work, we determined the size and number of patches empirically and just explore the discrete binary values whether patch is located or not for each grid. To improve this, we are working on more heuristic-based optimization algorithms with more detailed parameterization of the problem setting, e.g., directly optimizing patch coordinates and their sizes.

4) Exploration on Other Attack Targets: In this WIP paper, we focused on an attack of detection avoidance of object recognition models. In addition to this approach, we plan to develop and evaluate attack methods to deceive the class recognized by traffic sign recognition models. We are working on an attack against traffic sign classifier as shown in Fig. 8. We will apply this attack technique to traffic sign recognition models and evaluate attack effectiveness.

5) Stealthiness Evaluation: We consider that the ARP attack has a high advantage in the attack stealthiness compared to existing adversarial patch attacks. We plan to conduct two methods to systematically verify the stealthiness of our attack: metric-based evaluation and user study. Mathematical verification methods include methods for measuring the differences between traffic signs with and without patches. Specifically, it is conceivable to acquire photos or rendering results with and without patches in the actual traffic sign environment or simulation environment, and then calculate the differences. By comparing the difference in the measured L2 norm to that of the adversarial sample patch from previous research, we compare the stealthiness. The second user study approach involves presenting real driving videos. We measure the probability of being aware of the difference between the adversarial sample patch from existing research and the patch from the proposed method when attached to traffic signs that appear while driving to compare the two.

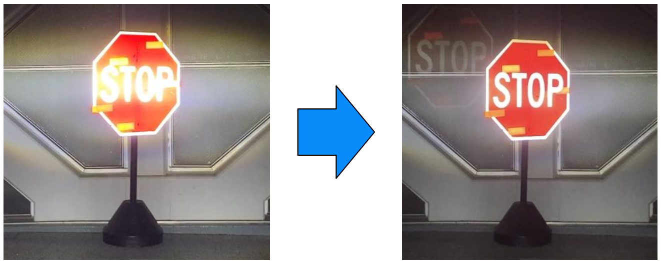
6) Evaluation in Realistic Autonomous Driving Scenarios : We also plan to evaluate the capabilities of our attack in real autonomous driving scenario. In this paper, we evaluated the feasibility of the proposed method only in a fixed position setting, where the camera and the STOP sign are fixed. However, in real-world autonomous driving scenarios, the distance and angle between the traffic sign and camera change continuously, and the attack must be continuously successful. Therefore, we will assess the feasibility of continuous attacks in situation with constantly changing distance and angle, and the attacks against traffic sign recognition systems of real autonomous vehicles.

In this study, the experiment was conducted under the condition that the height of the traffic sign and the headlight were the same. However, in an actual driving environment, a height difference exists between the sign and the headlight. The effect of this height difference on the success rate of attacks should be considered as an issue for future research. In addition, although only ambient light under nighttime conditions was considered in this study, the brightness of ambient light and headlights may in fact affect the success rate of attacks. A detailed investigation of the impact of these factors is another issue that should be addressed in future research.

7) Potential Defences: We plan to evaluate and discuss defense methods against the proposed method. Our attack is difficult to see with the naked eye under conditions such as daytime. However, when the attack is triggered, some parts of the sign change color significantly due to the reflecting patch. Therefore, existing defense methods against adversarial patch attacks can be applied. There are two types of defense methods against adversarial patch attacks: empirical defense methods such as the detection of attack patch patterns, and theoretically guaranteed defense methods. The former is known to be vulnerable to adaptive attacks. The latter has problems such as low accuracy and high computational complexity and is not suitable for real-time environments that require autonomous driving. Therefore, it is essential to explore defense methods against the proposed method. We will evaluate and discuss countermeasures to ARP attack.

Specifically, we are considering two methods: attack detection and defense by suppressing reflections. Regarding the detection of the first attack, some research has already been conducted on detecting reflection parts [23]–[25]. We are planning to utilize this research to detect reflection areas in real time and determine whether there is an attack based on their size. We also plan to suppress reflected light and prevent attacks by attaching a polarizer to the camera. It is known that reflected light is polarized in a specific direction [26]. Therefore, we believe that reflection can be suppressed by using a polarizing filter to allow only light from a specific direction to enter the camera, as shown in Fig. 9, and as a result, it will be possible to defend against ARP attacks.

When discussing defensive methods, it is important to



without a polarizer

with a polarizer

Fig. 9: Images with and without polarizer under ARP attack. The polarizer suppresses the reflected light from the patches and this will lead to defence against ARP attack.

consider whether the defensive method can be implemented. Traffic signs recognition systems using cameras are currently widely used. Therefore, even if an effective defensive method is found, it is necessary to apply the method to many autonomous vehicles. We plan to evaluate the defense methods from the perspectives of both flexibility and accuracy.

VII. Summary

In this study, we identify a novel attack vector, ARP attack, which is stealthy and effective. Its feasibility was evaluated using reflective patches with high success rates in both digital simulations and real-world experiments. For future work, we plan to expand our material evaluation, conduct a large-scale attack test in different driving scenarios, and improve 3D simulations for patch reflection. We also aim to optimize attack strategies for traffic sign recognition systems, improve the stealthiness of the attack, and develop potential countermeasures. Importantly, the value of our study lies in highlighting the need to focus on sensor and AI security in nighttime conditions for 24/7 autonomous driving operations. Our future efforts will allow us to deepen our understanding of ARP attacks and develop effective countermeasures.

Acknowledgment

We thank Mr. Tatsuya Takehisa of Office A.H.O. for building the set of headlights used in this study. A part of this work was supported by JSPS KAKENHI 22S0604, JST CREST JPMJCR23M4, NSF CNS-2145493, CNS-1929771,CNS-1932464, and USDOT UTC Grant 69A3552348327.

References

[1] S. B. Wali, M. A. Abdullah, M. A. Hannan, A. Hussain, S. A. Samad, P. J. Ker, and M. B. Mansor, "Vision-Based Traffic Sign Detection and Recognition Systems: Current Trends and Challenges," *Sensors*, vol. 19, no. 9, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/9/2093

[2] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Machine Learning and Knowledge Discovery in Databases*, M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Eds. Cham: Springer International Publishing, 2019, pp. 52–68.

[3] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *Computer Vision and Pattern Recognition*, 2018.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, T. Kohno, and D. Song, "Physical Adversarial Examples for Object Detectors," *CoRR*, vol. abs/1807.07769, 2018. [Online]. Available: http://arxiv.org/abs/1807.07769

[5] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, "SLAP: Improving Physical Adversarial Examples with Short-Lived Adversarial Perturbations," in *USENIX Security*, 2021.

[6] D. Wang, W. Yao, T. Jiang, C. Li, and X. Chen, "RFLA: A Stealthy Reflected Light Adversarial Attack in the Physical World," 2023.

[7] T. Sato, S. H. V. Bhupathiraju, M. Clifford, T. Sugawara, Q. A. Chen, and S. Rampazzi, "WIP: Infrared Laser Reflection Attack Against Traffic Sign Recognition Systems," in *VehicleSec*, 2023.

[8] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, "Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks," in *ACM Conference on Computer and Communications Security*, 2020.

[9] darknet, "YOLO: Real-Time Object Detection," https://pjreddie.com/darknet/yolo/.

[10] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 2015.

[11] "Cruise," https://www.getcruise.com/.

[12] "Ride-Hailing App - Make the Most of Your Drive - Waymo One," https://waymo.com/waymo-one/.

[13] "California Issues Permits to Cruise, Waymo for Autonomous Vehicle Service," https://www.reuters.com/technology/california-issues-permits-cruise-waymo-autonomous-vehicle-service-2022-02-28/, 2023.

[14] "Tesla autopilot," https://www.tesla.com/autopilot.

[15] "Road sign assist — toyota safety sense — toyota au," https://www.toyota.com.au/toyota-safety-sense/road-sign-assist.

[16] M. Folsom, P. Hausladen, J. Hayward, J. Nattress, and K. Ziock, "Characterization of retroreflective tape optical properties for use with position-sensitive scintillator detectors," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1005, p. 165365, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900221003491

[17] P. D. of Transportation, "Conducting sign retroreflectivity inspections," https://higherlogicdownload.s3.amazonaws.com/PSATS/b571db8e-24cb-4351-9048-c6ee55c076c8/UploadedImages/LTAP/00_Workbook_2020-10-27_ConductingSignRetroreflectivity_Inspections__37.pdf.

[18] F. Sun, Y. Liu, Y. Yang, Z. Chen, and S. He, "Arbitrarily shaped retro-reflector by optics surface transformation," *Chin. Opt. Lett.*, vol. 18, no. 10, p. 102201, Oct 2020. [Online]. Available: https://opg.optica.org/col/abstract.cfm?URI=col-18-10-102201

[19] A. International, *Standard Specification for Retroreflective Sheeting for Traffic Control*, ASTM International standard D4956-19, 2019.

[20] "Japanese Industrial Standards (JIS)." [Online]. Available: https://ndlsearch.ndl.go.jp/books/R100000002-I000003327815

[21] NVIDIA, "Ray tracing — nvidia developer," https://developer.nvidia.com/discover/ray-tracing.

[22] T. B. Foundation, "blender.org - home of the blender project - free and open 3d creation software," https://www.blender.org/.

[23] T. Yoshida, I. Funahashi, N. Yamashita, and M. Ikehara, "Saturated Reflection Detection for Reflection Removal Based on Convolutional Neural Network," *IEEE Access*, vol. 10, pp. 39 800–39 809, 2022.

[24] M. A. Ahmed, F. Pitie, and A. Kokaram, "Reflection Detection in Image Sequences," in *Computer Vision and Pattern Recognition*, 2011, pp. 705–712.

[25] A. Morgand and M. Tamaazousti, "Generic and real-time detection of specular reflections in images," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 1, 2014, pp. 274–282.

[26] T. T. Corporation, "Reflection & polarization of light in machine vision," 2023. [Online]. Available: https://www.toshiba-teli.com/en/technology/reflection-polarization-light-machine-vision