Poster: Adversarial Retroreflective Patches: A Novel Stealthy Attack on Traffic Sign Recognition at Night[1]

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

***Abstract*—Reliable traffic sign detection is crucial in autonomous driving. Our research presents the Adversarial Retroreflective Patch (ARP) attack, a new threat that utilizes the optical properties of retroreflective materials. This method is particularly effective at night and uses retroreflective patches on traffic signs to interfere with camera-based recognition by reflecting light from vehicle headlights. Initial studies show a high success rate in digital simulations, slightly lower in real-world tests. This study is crucial for evaluating vulnerabilities in autonomous vehicles' sensors and AI, especially in low-light conditions, to maintain the safety and reliability of self-driving technologies.**

I. Introduction

Traffic signs are fundamental to the safety and efficiency of all road users. For autonomous vehicles, traffic sign recognition is important. For this task, the vision-based traffic sign recognition systems [1] are widely adopted in many vehicles.

However, many recent studies [2]–[5] have actively reported that vision-based traffic sign recognition systems could be vulnerable to adversarial attacks with malicious stickers [2]–[4] and IR laser projection attacks [5]. While these attacks are effective and stealthy in their targeted scenarios, each attack has complemental pros and cons. For example, patch attacks [2]–[4] are always visible to everyone, and pedestrians or road guards may notice and remove them.

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. 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.

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. We evaluate the effect of our proposed attack with the YOLOv3-tiny model [6] trained on the COCO dataset [7]. We achieved a success rate of up to 100% in digital simulations and up to 90 % in the real-world experiment. Furthermore, we discuss future research directions, particularly in developing defense methods against ARP and evaluating its broader implications for automated vehicle systems.

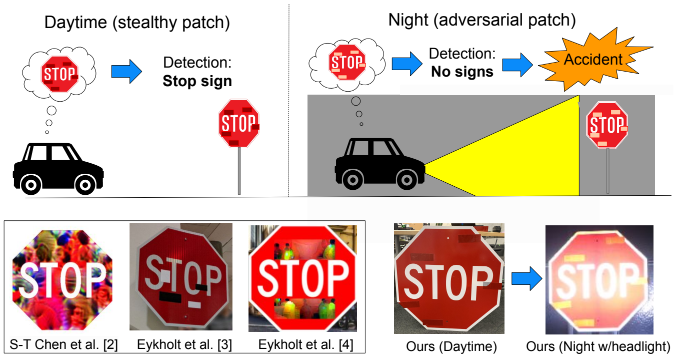


Fig. 1: Overview of the ARP Attack. The patch becomes visible only when the headlights shine on it, and the attack takes place. Our attacks are highly stealthy at daytime.

II. 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. Thus, we assume that the attacker 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: **Step 1.** Collect the images of the target sign. **Step 2.** Generate and deploy the ARP attack patches. **Step 3.**Wait for the victim vehicle to pass by the attack sign.

B. Optimization for Adversarial Reflective Patch Attacks

We present the methodology for optimizing the placement of 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. 2 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.

III. Evaluation

We evaluated the feasibility of our ARP attacks in the digital experiment with image processing and the real-world

[1] The extended version of this work will be presented at VehicleSec 2024 (WIP).

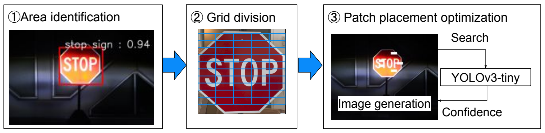


Fig. 2: Overview of optimization for patch placement of Adversarial Reflective Patch Attacks. We use image processing and beam search for search for the best patches' placement.

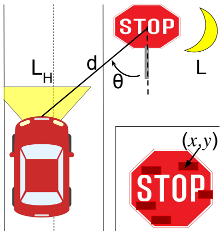


Fig. 3: Overview of variables and parameters of ARP attack.

TABLE I: 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

experiment. 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). We took a photo of the sign from a position 5 meters in front of it and optimize the patch placement with this photo. To measure the attack success rate, 20 images were taken under each condition, and the detection and evasion success rate with YOLOv3-tiny was measured.

In the digital experiment, we assumed that the reflected color was white and reproduced the reflection with image processing. Then, we investigated avoidance of detection using YOLOv3-tiny. In the physical experiment, we actually attached patches to a sign, took photos under headlight illumination, and investigated the success rate of attacks with photos. The results showed a 100% success rate in digital simulations, and less than or equal to 90% in real-world experiments.

IV. Discussions and Future Plans

In our study of ARP attacks in autonomous driving systems, we can use many types of the reflective materials. We will conduct a comprehensive survey of various retroreflective materials to assess their effectiveness in ARP attacks across different autonomous driving scenarios. It is also crucial to enhance the realism of our attack models. To accomplish this, we will use advanced shader techniques and ray-tracing in 3D software like Blender to model reflective patches more accurately. This will enable more precise attack simulations. Additionally, we will focus on improving attack optimization methodologies, particularly in the selection and placement of patches, by using heuristic-based optimization algorithms for a more nuanced approach. Expanding our focus, we plan to investigate additional attack targets, particularly those that involve deceiving traffic sign recognition models. We will conduct a thorough evaluation of the effectiveness of these new methods. Our research will also include an assessment of attack stealthiness, which will be evaluated through both metric-based assessments and user studies to provide a comprehensive evaluation of the subtlety of our attacks. To ensure practical application, it is imperative to test the attack's feasibility in real-world autonomous driving scenarios. This requires dynamic testing environments where the relative positions and angles of signs and cameras continuously change. Additionally, it is essential to develop and evaluate potential defense methods against ARP attacks. This involves scrutinizing the limitations of current defense strategies and innovating more robust countermeasures. We are currently considering both hardware and software countermeasures. The first method is to install a filter on the camera to prevent attacks by suppressing reflected light, and the second method is to detect attacks from the area where reflections are occurring. These approaches will lead our research to a deeper understanding of ARP attacks and contribute significantly to improving the security of autonomous driving systems.

V. Summary

In this study, we propose a novel attack vector, ARP attack, which is stealthy and effective. Our study evaluated the feasibility of the ARP attack and and recorded high attack success rates in both digital and real-world experiment. For future work, we plan to expand our material evaluation, conduct a large-scale attack test in 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, our study underscores the importance of sensor and AI security in nighttime conditions for 24/7 autonomous driving. Our future work will further enhance our understanding of ARP attacks and 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] 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 *CVPR 2018*, 2018.

[3] 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

[4] 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.

[5] 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.

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

[7] 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.

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

Go Tsuruoka[1], Takami Sato[2], Qi Alfred Chen[2], Kazuki Nomoto[1,3], Ryunosuke Kobayashi[1], Yuna Tanaka[1], Tatsuya Mori[1,4,5]

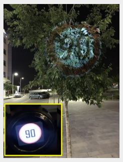
[1]Waseda University, [2]University of California, Irvine,[3]Deloitte Tohmatsu Cyber LLC,[4]RIKEN,[5]NICT,

Limitations of Prior Attack against Vision-based Traffic Sign Recognition (TSR)

- **Many prior attacks on vision-based TSR has been actively reported.**
- **But, many of them has stealthiness issues, particularly at daytime.**
  - **Patch attacks**:
    - **Less deployment effort** since adversary can just put patches at once
    - But, **attack patches are highly noticeable** and police or road guards might remove patches immediately
  - **Light projection attacks**:
    - Can be **stealth when projection is off**
    - But, adversary **must control attack on and off** and **attack device should be highly visible and suspicious.**

Chen et al. (2019)

Eykholt et al. (2018)

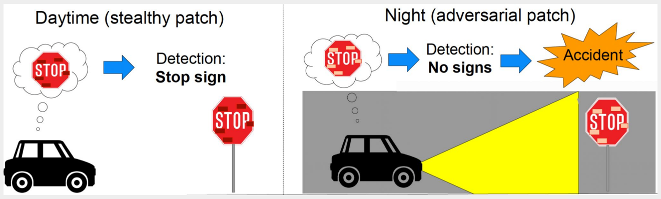

Nassi et al. (2020)

Lovisotto et al. (2021)

Can we design new attack just having only advantages of these attacks?

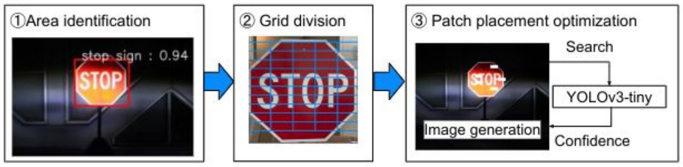
New Attack Vector: Adversarial Retroreflective Patch Attack (ARP Attack)

- **Key idea**: **Retroreflective patch**
  - Can be stealth at daytime but **effective at night with victim's headlight**
  - Retroreflective material returns light to its source
    - Can selectively targets the victim



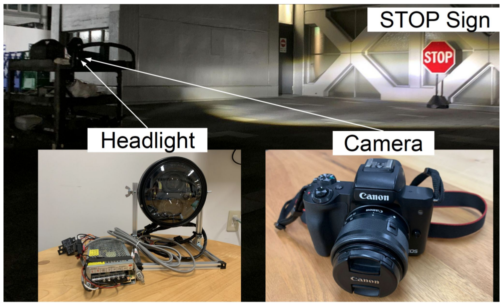
Current Methodology of Patch Location Optimization

- Size of each patch is now fixed for ease.
  - Each patch is 6 inch by 3 inch
- Explore min. score of stop sign with beam search
- Target model: YOLOv3-time

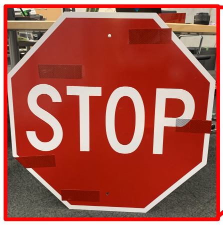


Feasibility Study in Real World

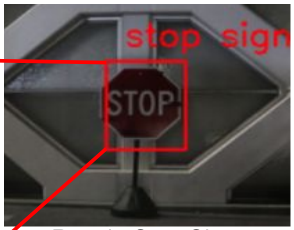
Experimental Setup



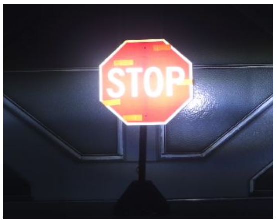
Optimized Attack



Significantly more stealthy than priors!



Result: Stop Sign



Result: No Sign

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

θ [°]	ASR [%]
30	65
0	90
-30	35

ASR: attack success rate
d: Distance: Car↔ Sign
θ: Angle: Car ↔ sign

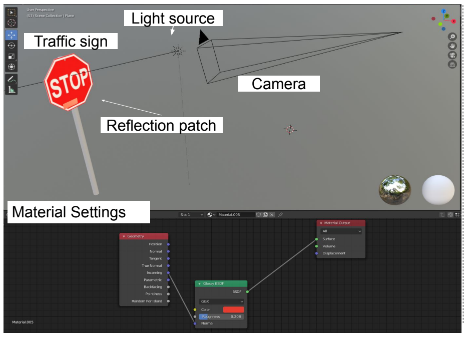
Conclusion & Future Plans

We confirmed feasibility of ARP attacks in real world

Future Plans:

- **Survey more materials** usable for ARP attacks
- **Explore more realistic attack modeling**
- Evaluate **other targets such speed limit sings**
- Perform **stealthiness evaluation with user study**
- **End-to-end autonomous driving evaluation**
- **Design effective defenses** against ARP attacks

Work-in-Progress Attack Modeling with Blender



Strong light reflection

Weak light reflection