WIP: Infrared Laser Reflection Attack Against Traffic Sign Recognition Systems

Takami Sato*†, Sri Hrushikesh Varma Bhupathiraju*‡, Michael Clifford§, Takeshi Sugawara¶,
Qi Alfred Chen†, and Sara Rampazzi‡
†University of California, Irvine; ‡University of Florida; §Toyota InfoTech Labs; ¶The University of Electro-Communications

***Abstract*— All vehicles must follow the rules that govern traffic behavior, regardless of whether the vehicles are human-driven or Connected, Autonomous Vehicles (CAVs). Road signs indicate locally active rules, such as speed limits and requirements to yield or stop. Recent research has demonstrated attacks, such as adding stickers or dark patches to signs, that cause CAV sign misinterpretation, resulting in potential safety issues. Humans can see and potentially defend against these attacks. But humans can not detect what they can not observe. We have developed the first physical-world attack against CAV traffic sign recognition systems that is invisible to humans. Utilizing Infrared Laser Reflection (ILR), we implement an attack that affects CAV cameras, but humans can not perceive. In this work, we formulate the threat model and requirements for an ILR-based sign perception attack. Next, we evaluate attack effectiveness against popular, CNN-based traffic sign recognition systems. We demonstrate a 100% success rate against stop and speed limit signs in our laboratory evaluation. Finally, we discuss the next steps in our research.**

I. Introduction

All vehicles must obey traffic signs, whether those vehicles are human-driven, connected, autonomous vehicles (CAVs), or semi-autonomous vehicles. Failure to do so can have severe legal and safety consequences. Recent studies [1]–[3] focused on the security of traffic sign recognition systems. In these studies, physical adversarial attacks degrade the accuracy of traffic sign recognition systems. Examples of these attacks include modifying signs using stickers [2], visible light projection [1], or shadow projection [3] in order to cause autonomy stacks to classify traffic signs incorrectly. For example, these attacks may cause a stop sign to be misclassified as a yield sign or speed limit sign, resulting in potential accidents.

These attack studies have several important limitations. First, the attacks are detectable by human drivers. A human may notice the presence of stickers or illumination inconsistent with the surrounding environment on a sign. This limits applicability to semi-autonomous vehicles, such as Tesla's, whose drivers may see the attack's effects and assume complete vehicle control. Additionally, the studies do not address the scenarios where attacks are invisible to humans, such as those using Infrared (IR) light.

*co-first authors

Symposium on Vehicles Security and Privacy (VehicleSec) 2023
27 February 2023, San Diego, CA, USA
ISBN 1-891562-88-6
https://dx.doi.org/10.14722/vehiclesec.2023.23055
www.ndss-symposium.org

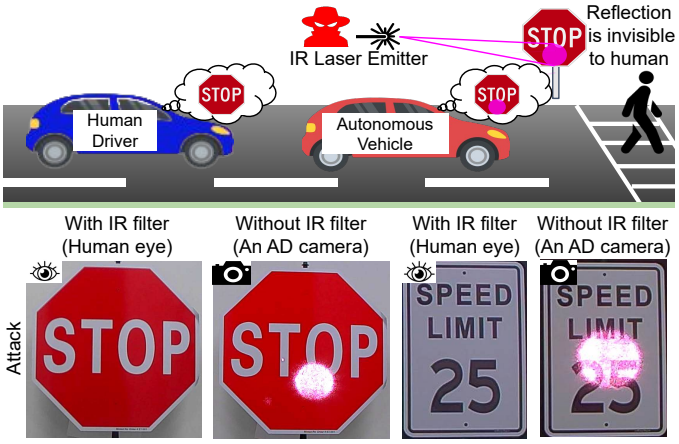

Fig. 1: Overview of our Infrared Laser Reflection (ILR) attack. Infrared laser light is invisible to humans but is visible to CAV cameras that lack infrared filters. The resulting altered images may be misclassified by CAV autonomy stacks such as [4] as the wrong type of sign, or as entirely different objects.

We develop and implement a novel adversarial attack using IR laser light. While invisible to humans, this light can be seen by CAV cameras that lack an IR filter (demonstrated in Fig. 1). The output from a victim camera is an image altered at the pixel level, which may be crafted to cause misclassification by the CAV's perception stack.

Camera sensors count the number of photons that hit each sensor photosite. These sensors are typically sensitive to both visible and infrared spectrum light. To reduce image noise, an infrared filter can be placed between the lens and the sensor. However, some CAV sensors, such as those used in the front cameras of the Tesla Model 3, lack these filters. While improving nighttime sensitivity, this leaves those sensors vulnerable to IR attacks such as the I-Can-See-the-Light (ICSL) attack [5], which projects an IR laser directly onto camera sensors. ICSL has not been demonstrated against traffic sign recognition systems. Moreover, ISCL requires that the laser beam be aimed continuously and precisely at a moving target camera. These limitations motivate our novel IR Laser Reflection (ILR) attack.

Our attack causes CAV perception stacks to misclassify images of traffic signs. Our attack alters the sign appearance for IR-sensitive cameras by illuminating parts of the signs with an IR laser. Unlike ICSL, our attack does not require the attacker to aim at a camera on a moving vehicle. In addition, by projecting onto signs, ILR can fool perception

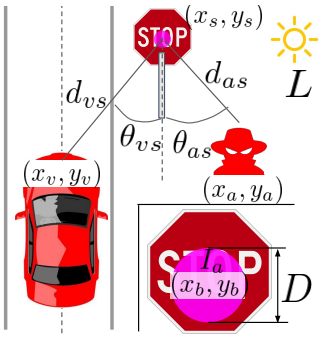

Fig. 2: Overview of variables and parameters of ILR attack

TABLE I: Definition of Variables

Symbol	Description	Type
d_{vs}	Distance: victim ↔ sign	V
θ_{vs}	Angle: victim ↔ sign	V
(x_v, y_v)	Position of victim	V
d_{as}	Distance: attacker ↔ sign	A
θ_{as}	Angle: attacker ↔ sign	A
D	Diameter of attack beam	A
(x_a, y_a)	Position of laser emitter	A
(x_b, y_b)	Position of attack beam	A
I_a	Laser beam power	A,S
L	Intensity of ambient light	S
(x_s, y_s)	Position of traffic sign	S

V: Victim's variable, A: Attacker's parameter, S: Scenario parameter

module classifiers using significantly less effort than attacks that target camera sensors directly, making implementation easier in real-world scenarios. Additionally, we designed a methodology that generates optimized, simulated traces of IR laser reflections. This optimization allows us to create effective, real-world adversarial examples that cause target models to misclassify traffic signs.

In this work, we present the initial results of our ILR attack design and evaluation. We evaluate the attack using a CNN-based classification model trained on US and European traffic sign data sets: GTSRB and LISA [6], [7].

In §III, we formulate our ILR threat model variables and parameters and introduce our work to design the most effective ILR attack through the simulation and optimization of both the IR laser reflection itself and the position of the reflection on the sign. In §IV, we demonstrate our ILR attack under indoor conditions, achieving a 100% attack success rate against both stop sign and speed limit sign targets. We consider the attack successful when the perception stack's classification is changed from a correct sign to an incorrect sign.

We compare our results against the baseline random attack in which an IR laser beam hits random portions of the target signs. The ILR attack's success rate on the stop sign is approximately 15 times higher than the baseline. Against speed limit signs, the attack performs 3 times better. We also perform preliminary evaluations of attack robustness versus the resulting first-state object detection inaccuracy and differences in sign background. Finally, we characterize ILR attacks and discuss our next steps in §V.

II. Background and Related Work

Vision-Based Traffic Sign Recognition. Vision-based traffic sign recognition detects signs in real-time using inexpensive camera sensors, combined with fast neural networks, to provide object recognition and classification [8]. These cost and capability advantages have led to wide-spread adoption in commercially available Level 2 autonomy systems such as Tesla and OpenPilot. Unfortunately, real-time sign perception is vulnerable to attacks that affect what the camera sees. Our novel work addresses perception attacks because these systems are both vulnerable, and available, in production vehicles.

IR Laser Light and Human Perception. Humans see only visible light, but some IR camera sensors used in vehicle vision systems can see both visible and IR light. Their lack of IR-cut filters allows IR light to reach the sensor, improving nighttime vision and object detection [9]. Unfortunately, this also creates a gap between what humans and sensors can see, as shown in Fig 1. Our attack projects a spot onto a target sign using IR light. While invisible to humans, the IR projection is seen by the unfiltered sensor, becoming part of the sensor output image.[1] We demonstrate that adversaries can use this effect to create invisible (to a human) adversarial shapes on the target sign that induce traffic sign misclassification.

Adversarial Attacks Against Traffic Sign Recognition. Prior research showed that Deep Neural Network (DNN) models are generally vulnerable to adversarial examples and adversarial attacks [10]. These attacks have been explored in the physical world [1]–[3]. For traffic sign recognition, DNNs are known to be vulnerable to attaching small stickers [2], projecting visible patterns [1], and shading shadows [3]. However, none of these prior efforts consider invisible attacks as the one proposed in this analysis.

Wang et al. [5] utilized IR LEDs to attack cameras in the Tesla Model 3. Their attack fools Tesla's traffic light and front vehicle recognition systems, as well as its vision-based localization system. An IR light is projected directly onto the vehicle's camera, causing the perception of a fake traffic signal. Their solution distinguishes between IR light sources, which do not reflect off roadways, and active street lights, which do. They also distinguish the difference in color between the reflected, valid light and the non-reflected attack light. Their solution cannot be applied to detect our attack because street signs do not use active illumination, and because our IR laser reflects off street signs rather than hitting the camera directly. They also do not evaluate the attack against traffic sign recognition systems nor validate their attack for realistic scenarios involving moving vehicles.

III. Methodology: IR Laser Reflection Attack

A. Threat Model and Attack Goal

Fig. 1 shows an overview of the ILR attack. The attacker selects a traffic sign to attack and places the IR laser emitter 3–10 meters away from the traffic sign, as in prior work [1]. The attacker's goal is to cause the victim's perception stack to misclassify the target sign (such as a stop sign to a yield sign). We assume the attacker can use public information such as manuals and teardown documents [11] to obtain camera specifications. The attacker can (pre-) compute attack IR traces for each potential target vehicle. We also assume that the attacker has a basic understanding of IR light and optics. The attack is remote and does not require any firmware access or information about the images captured by the camera.

B. ILR Attack Model

We formulate the ILR attack as in Fig. 2, with variables shown in Table I. The victim's parameters change dynamically as the victim CAV moves. These include: (1) the position of the victim's camera, d_{vs}, and (2) its orientation, θ_{vs} – both with respect to the stop sign. The attacker's parameters include: (1) the position of the attacker's laser, d_{as}, and (2)

[1]The IR light appears white and purple in the output image, but can not be seen by humans. The sensor is monocolor, and uses color filters to filter the light that will hit each photosite, allowing through red, green, or blue wavelengths. However, the sensor can only sense intensity (photon count) at each photosite. Because the filters are not perfect, a false color representation of the photon count where IR light hits each photosite is generated.

its orientation, θ_{as} – both with respect to the stop sign; (3) the resulting laser beam's power (in mW), I_a, (4) the beam diameter, *D*, and (5) the beam spot position on the traffic sign surface, (x_b, y_b). The attacker can select parameters to maximize attack effectiveness. We determine viable attack capabilities by mapping the feasible range of attacker parameters given the IR laser's capabilities. Scenario parameters represent environmental factors, including ambient light intensity and traffic sign position.

Our model accounts for *temporal image noise*, meaning random fluctuations in the output image pixel intensity (photon count) values [12], as stray photons hit different sensor photosites across consecutive image frames. We thus evaluate the attack effects on CAV sign classification over multiple, consecutive camera image frames. Note that our attack formulation has not been finalized in this work – we continue to improve both our attack formulation and model. We present our preliminary attacker capability model in §III-C.

CAV cameras generate image streams as output. We refer to the set of output image pixels altered by ILR attacks as *attack traces* (see Fig. 3). Each parameter above independently affects the camera's perception, resulting in different attack traces.

C. Optimization Problem Formulation for ILR Attack

Attackers can use attack formulation (§III-B) to determine the best parameters to achieve traffic sign misclassification. While testing all possible attack scenarios in the physical world is infeasible, an offline digital-space simulation allows the evaluation of a large number of attack parameter combinations. There are two technical challenges to doing so: (1) Generating an IR laser reflection model that accurately simulates attack effects; (2) Creating an effective black-box optimization method to explore a large set of parameter variations. For (1), we present our initial laser attack trace model in §III-D. For (2), we explore the best attacker position (x_b, y_b) under static indoor scenarios. The detailed setup is discussed in §IV-A. Our future work plan is discussed in §V.

In our initial work, we used the Tree-structured Parzen Estimator algorithm [13] (a black-box optimization method) to minimize the target sign classification accuracy. In other words, we consider the ILR implementation a black-box misclassification attack. To determine attacker capability, we assume the attacker can change the position of the attack trace on the target traffic sign, (x_b, y_b). We measure attack success by whether the traffic sign recognition module's second-stage classifier generates different classifications for benign and attack images. The results of this analysis are shown in Section §IV-B.

D. Image Difference-based IR Trace Modeling

Optimizing ILR for a given target sign requires finding the best position for the IR beam on the sign to maximize its effects. We optimized the attack beam position (x_b, y_b), while holding all other attacker parameters constant. We first modeled the attack trace by calculating the color differences between corresponding RGB pixel values of the trace in the benign and attack output images. These pixel differences are then applied to the benign traffic sign image to simulate IR beams targeted at different sign locations. See Fig. 4.

Experiment Configuration. We generated attack traces using a laser power, I_a, of 51.4 mW. This was the power level where we observed average trace pixel saturation when using a 15 cm spot diameter, *D*. The distance from the laser to the sign, d_{as}, was set to 3m, and ambient light was set to 100 lux using the attack model in §III-B. The angles θ_{vs} and θ_{as} were set to 0° by placing the victim camera and the laser diode next to each other, in front of the traffic sign. To reduce temporal image noise in the target camera, we modeled and averaged 10 consecutive frames for each benign and attack case.

Results and Observations. We observed that the RGB pixel values in the trace difference image depend on the camera's perceived surface color for the traffic sign. This required trace re-modeling for each color on the sign surface. As each traffic sign has two surface colors (red and white for the stop sign and white and black for the speed limit sign), we collected four total attack traces. We then created individual attack traces for each surface color by interpolating and adjusting RGB values. Fig. 4 in the Appendix shows an overview of the modeling and simulation process. This modeling was used to optimize the attack on the sign classification model. We evaluate the consistency between our model and real-world attacks in §IV.

IV. Evaluation

A. Evaluation Setup

As in our modeled attack, we evaluated our attack on real-world aluminum stop and 25 mph speed limit signs in a controlled indoor environment, as shown in Fig 3. We placed the victim camera and IR laser diode at $d_{vs} = d_{as} = 3$ m in front of the target sign with angles $\theta_{vs} = \theta_{as} = 0^\circ$. We configured the attack beam to project a $D = 15$ cm spot using 51.4 mW laser power. Ambient light, L was stable at 100 Lux. To increase scenario diversity, we evaluated 4 different background colors (room wall, green, black, and white). We observe that changing the background changes the camera's measured scene brightness, which causes the camera to automatically change the exposure.[2]

Evaluation Models. We evaluated two classification models, one trained on European traffic signs and the other on US signs. For the European traffic signs, we trained a CNN classification model on the GTSRB [6] dataset using a 30×30-pixel image size. For the US traffic signs, we train a CNN model on the LISA [7] dataset using a 60×60-pixel image size. We doubled the input image size of the US signs because the US sign classifier requires a higher resolution to correctly classify each sign, even in benign cases. To focus on the analysis of the second-stage classification model, we manually generated a square crop for each sign, as in Fig. 3, and resized it to match the model's input size.

Evaluation Metrics. We selected two evaluation metrics based on our threat model (§III-A) and attack design: the attack success rate (ASR), and the simulation consistency rate (SCR). ASR measures the percentage of cases in which a sign is misclassified, thus satisfying our attack goal, as shown in Fig. 2. Conversely, SCR, which is defined as the percentage of cases in which the classification is the same as in our simulated attack process, is used to evaluate attack beam modeling (§III-C) quality. While ASR is our primary metric, SCR is necessary to evaluate the validity of our attack design and select the most effective attack beam model in future work.

[2]Most auto-exposing cameras use average brightness to set exposure.

TABLE II: Comparison between success rates for our ILR attack and random attacks. Note that random attacks are tested physically, so SCR does not apply to the random attack.

	Stop Sign		25 mph Speed Limit	
	Random Attack	ILR	Random Attack	ILR
ASR	6.75%	100%	34%	100%
SCR	N/A	100%	N/A	80%

ASR: Attack Success Rate, SCR: Simulation Consistency Rate

B. Attack Effectiveness Evaluation

We observed a 100% traffic sign classification accuracy for both traffic sign types in the benign case. We first evaluated the effectiveness of the ILR attack through comparison with a simple baseline attack named *random attack*. In random attack, we manually aim the laser beam at the sign and collect the resulting camera images without any optimization process. We then conducted a preliminary robustness evaluation of the ILR attack to measure the effects of first-stage object detection and differences in sign backgrounds on our attack.

Results and Observations. Table II lists the evaluation results of the random (baseline) and ILR attacks. We evaluated the attacks against the stop sign using a simple CNN model trained on the GTSRB dataset, and against the 25 mph speed limit sign using the CNN model trained on the LISA dataset. To evaluate the ILR attack, we collected 40 images with 4 different backgrounds and 10 image frames each for each traffic sign. For random attacks, we collected 10 random beam positions across the same 4 backgrounds and traffic signs with 10 image frames for each. The ILR attack has a significantly higher attack success rate (100%) than the random attack. Moreover, the observed class labels are generally consistent with the labels simulated when SCR is $\geq$80%. These results indicate that the second-stage classification model is vulnerable to manipulation with circular IR laser reflections when the attacker has control over IR beam spot position on the sign.

Preliminary evaluation of attack robustness. We evaluated attack robustness to differences in sign background and first-stage object detection inaccuracies. We initially evaluated the attack results on three backgrounds in an indoor setting and observed that the ASR and SCR were always 100%. We also observed that regardless of the background tested, the optimal IR beam spot position on the sign remained constant. This indicates robustness to sign background differences.

We also evaluated how inaccuracies in first-stage object boundary detection can affect the input of the second-stage classification model and, consequently, the classification results. To do this, we manually apply random translations to the annotated bounding boxes. We evaluated 100 random cases for each attack and observed that ASR and SCR for stop signs decreased with the noise level. For speed limit signs, on the other hand, we observed that ASR is always 100% and that SCR remains constant for low noise levels while gradually decreasing for higher noise levels. Appendix A provides a preliminary analysis of attack robustness.

V. Concluding Remarks and Future Plans

In this work in progress, we present our ILR attack and provide initial results on the vulnerability of traffic sign recognition systems to ILR attacks. We show that the ILR attack has a significantly higher attack success rate (100%) than the random attacks in our indoor test environment and that ILR attacks are robust against first-stage object detection inaccuracy and traffic sign background.

Fig. 3: Comparison of the simulated trace with our modeling and the real ILR attack results on the stop and speed limit sign. The classification results of the simulated and real attacks are consistent in $\geq$80% cases and always different from the correct class. Note that the size of the beam is always 15 cm diameter, i.e., the stop sign is larger than the speed limit sign.

Laser Safety. All the experiments in this work were conducted in controlled environments, using appropriate eye and skin protection. Note that we set our laser output power to 51 mW (class 3B laser), and used lenses and an iris to create the diverging beam suitable for the experiments (an optical power reduction of more than 9%).

Future Work. Next, we will investigate the attack capabilities of the ILR attack in real-world autonomous driving scenarios in a variety of environmental conditions. The ILR attack is highly effective in lab conditions, as shown in §IV, but real-world environments are more challenging. Environmental factors, such as dynamic lighting and weather, as well as differences in victim vehicle cameras and software, affect how traffic signs appear in camera data.

To address these limitations, we plan to improve attack optimization and trace modeling accuracy. For trace modeling improvements, we are considering two strategies: *(1) Function-fitting modeling:* This technique consists of using function fitting (e.g., $f(I_a, L, x_a, y_a, \theta_{vs}, \theta_{as}, ...)$) to model the RGB pixel differences caused by the IR laser. While this is robust against environmental changes, function fitting requires a careful design of the functions to reflect complex dependencies among parameters. *(2) Ray-tracing-based modeling:* Ray-tracing-based shading [14] can simulate the reflection of visible light very accurately and is typically used in rendering engines such as Blender [15]. In our case, the model needs to be modified to match our IR laser reflection parameters.

We also plan to systematically evaluate and discuss countermeasures to ILR. While ILR attacks are invisible to humans, attack traces are visible in camera images. Thus, the existing defense methods against adversarial patch attacks are theoretically applicable. Generally, two types of defenses against adversarial patch attacks have been proposed: empirical defenses, such the as detection of anomalous patterns in attack patches [3], [16]), and certified defenses that can provide theoretical guarantees [17], [18]. Although theoretically applicable to the ILR attack vector in this paper, the former is known to be vulnerable to adaptive attacks [17], and the latter suffers from low accuracy and high computational overhead, which is especially critical in autonomous driving settings. Further,

since IR laser traces are not the same as human-visible adversarial patches studied in these prior works, the effectiveness of these defenses against ILR attacks are unknown. Thus, an exploration of mitigating defenses against ILR are both novel and necessary.

ACKNOWLEDGEMENTS

This research was supported in part by the NSF CNS-1932464, CNS-1929771, CNS-2145493, USDOT UTC Grant 69A3552047138, JST SPRING JPMJSP2123, JSPS KAKENHI 22H00519, and unrestricted research funds from Toyota InfoTech Labs.

REFERENCES

[1] 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.

[2] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, "Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems," in *NDSS*, 2022.

[3] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be Dangerous: Stealthy and Effective Physical-world Adversarial Attack by Natural Phenomenon," in *CVPR*, 2022.

[4] "Baidu Apollo," https://github.com/ApolloAuto/apollo.

[5] W. Wang, Y. Yao, X. Liu, X. Li, P. Hao, and T. Zhu, "I Can See the Light: Attacks on Autonomous Vehicles Using Invisible Lights," in *ACM CCS*, 2021.

[6] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark," in *IJCNN*, 2013.

[7] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey," *IEEE TITS*, 2012.

[8] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, and Y. Kuang, "The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale," in *ECCV*. Springer, 2020.

[9] R. Thakur, "Infrared Sensors for Autonomous Vehicles," in *Recent Development in Optoelectronic Devices*, 2017.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," in *ICLR*, 2014.

[11] "Teardown: Lessons Learned From Audi A8," https://www.eetasia.com/teardown-lessons-learned-from-audi-a8/, 2020.

[12] H. Tian, B. Fowler, and A. Gamal, "Analysis of Temporal Noise in CMOS Photodiode Active Pixel Sensor," *Solid-State Circuits*, 2001.

[13] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in *NeurIPS*, 2011.

[14] "Ray Tracing," https://developer.nvidia.com/discover/ray-tracing.

[15] "blender.org - Home of the Blender project," https://www.blender.org/.

[16] C. Yu, J. Chen, Y. Xue, Y. Liu, W. Wan, J. Bao, and H. Ma, "Defending against Universal Adversarial Patches by Clipping Feature Norms," in *ICCV*, 2021.

[17] P. yeh Chiang*, R. Ni*, A. Abdelkader, C. Zhu, C. Studor, and T. Goldstein, "Certified Defenses for Adversarial Patches," in *ICLR*, 2020.

[18] C. Xiang, S. Mahloujifar, and P. Mittal, "PatchCleanser: Certifiably Robust Defense against Adversarial Patches for Any Image Classifier," in *WOOT*, 2022.

APPENDIX A
ATTACK ROBUSTNESS EVALUATION

1) Robustness Against Inaccuracy in First-Stage Object Detection: While we focused on attacking the second-stage classification model of traffic sign recognition systems, as discussed in §II, we also observed how inaccuracies in the first stage can change the automatic bounding cropping and consequently alter the input of the second-stage classification model. To evaluate the impact of the inaccuracy on the classification results, we apply vertical and horizontal translation noise to our manually annotated bounding boxes. Fig. 5 shows the ASR for random vertical and horizontal displacement. Since the bounding box sizes are different for each image, we use the percentage over the width and height of the bounding box as a displacement level, δ, instead of the corresponding sizes in pixels. Given δ, we generate a random number under the uniform distribution $\mathcal{U}(-\delta, \delta)$ and displace the bounding box based on the result. For example, a 10 pixel displacement will be applied on a bounding box with 100 pixel height and width if the random number is 10%. As shown, bounding box inaccuracy has a greater impact for the stop sign than the speed limit sign. The ASR and SCR for the stop sign decrease with increasing noise levels. In contrast, the ASR for the speed limit sign is always 100%, while the SCR eventually starts to decrease around a noise level of 8%.

We hypothesize that these results are due to the shape, and the resulting pixel RGB values, of the attack beam traces. For example, the majority of attacks against the speed limit signs are classified as stop signs as shown in Fig 3. This suggests that the beam and stop sign may have similar features, which result in similar classifications. Thus, small translations of the IR spot can negate adversarial attacks, resulting in the correct stop sign classification. We are not able to precisely control the RGB pixel value generated by the IR laser because this depends on the CMOS image sensor and camera lens used. However, in the future, we will evaluate the impact of various IR spot shape changes on the effectiveness of the attack.

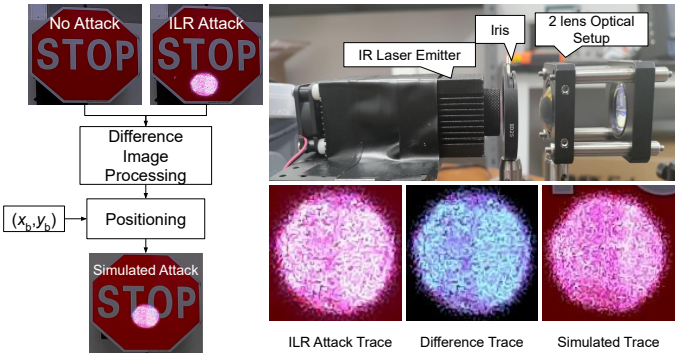

Fig. 4: Overview of Image Difference-based IR Trace Modeling (Left). The IR Laser module with two lens optical setup (right-top). The ILR Attack trace, the calculated Difference trace, and the corresponding Simulated trace (right-bottom).

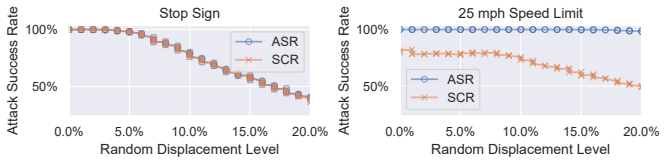

Fig. 5: Evaluation results of bounding box position noise detected in the first stage. Given noise level δ, the resulting perturbation, $\mathcal{U}$, obeys: $\mathcal{U}(-\delta, \delta)$ = percentage of bounding box width or height.