Poster: Cloaking Signals with Signs: Traffic-Sign-Camouflaged Adversarial Patches against Traffic Light Detection

Kyo Suetsugu
Waseda University

Go Tsuruoka
Waseda University

Yuna Tanaka
Deloitte Tohmatsu Cyber LLC

Kazuki Nomoto
Deloitte Tohmatsu Cyber LLC

Hiroto Onoda
Waseda University

Tatsuya Mori
Waseda University, RIKEN AIP, NICT

1 Introduction

Traffic light recognition is a safety-critical component of autonomous driving systems. A typical pipeline first *detects* traffic lights in the scene and then *classifies* their signal colors, with deep neural networks (DNNs) widely used in both stages. Since DNNs are known to be vulnerable to adversarial perturbations, the security of traffic light recognition is an important concern. Prior work, however, has focused almost exclusively on the classification stage; Pavlitska et al. [1], for example, generated patches that flip the predicted signal color. This leaves two issues open: the detection stage itself has received little attention, even though suppressing detection denies the downstream system any access to traffic light state and is arguably the more consequential failure mode; and, as shown in Figure 1, existing patches tend to look visibly artificial, undermining their real-world plausibility.

In this paper, we take a different route. Rather than perturbing what a traffic light *is*, we make the traffic light *disappear* from the detector's view, and we do so behind something that already belongs on the roadside. Our key observation is that traffic signs are frequently co-located with traffic lights: in 30 randomly sampled images from the DTLD traffic light dataset [2], 18 contained a traffic sign immediately adjacent to a traffic light. We exploit this fact by designing an adversarial patch that is explicitly shaped to resemble a real traffic sign, so that an observer sees nothing unusual about the roadside scene while the detector silently loses the traffic light. We refer to this strategy as *cloaking signals with signs*, and illustrate the overall idea in Figure 2.

Realizing this idea requires two design choices that, together, distinguish our work from prior patch attacks. (i) We target the detection stage directly, using a loss that suppresses objectness and class confidence within the target region rather than steering a downstream classifier. (ii) We introduce a *camouflage loss* that penalizes the L_2 distance between the patch and a chosen traffic sign template, so that visual plausibility is treated as a first-class optimization objective rather than an afterthought. To make the resulting patch deployable in the physical world, we additionally adopt standard techniques



Figure 1: Comparison between prior patches and ours. Our patch blends into the roadside as an ordinary sign instead of looking like an adversarial artifact.

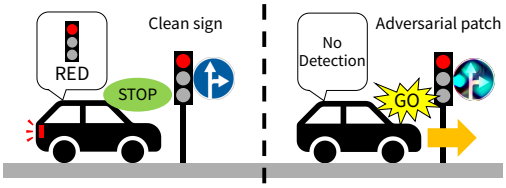


Figure 2: Overview of our attack. A patch disguised as a traffic sign hides the adjacent traffic light from the detector.

from prior patch attacks, namely Expectation over Transformation (EoT) and a Total Variation (TV) smoothness term, so that the patch is robust to lighting, viewing angle, and scale variations and remains printable.

We evaluate the resulting attack in both digital and physical-world settings. The strongest patch reaches an 85% attack success rate (ASR) in digital experiments, and printed patches mounted in a real outdoor scene achieve up to 100% ASR in recorded video, making the traffic light effectively invisible to a YOLOv5-based detector while remaining unremarkable to a human passer-by.

2 Traffic-Sign-Camouflaged Patches

We design adversarial patches for attacking traffic light detection that are visually similar to natural traffic signs placed near traffic lights, achieving stealthiness against human observers.

Threat Model. Following prior work [1], we assume a white-box setting in which the attacker has knowledge of the architecture and parameters of the target traffic light detection model. The attacker can generate an adversarial patch offline and place it near a target traffic light or overlay it onto an

existing traffic sign.

Patch Generation. Our goal is to generate adversarial patches that look visually similar to traffic signs placed near traffic lights, achieving stealthiness against human observers. To this end, we optimize the patch using a loss function consisting of three terms.

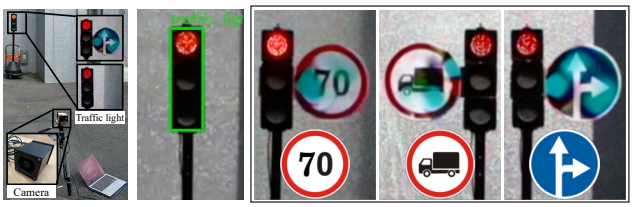
First, to suppress traffic light detection, we adopt a detection suppression loss that minimizes the product of the objectness score and the class probability for the traffic light class within an expanded region around the target bounding box. Second, to improve the patch's stealthiness, we introduce the camouflage loss, which is defined as the L_2 distance between the patch and the target traffic sign image, to encourage visual similarity. Third, we employ TV loss as a smoothness loss to reduce fine details and improve printability. To further improve robustness under physical-world conditions, we apply EoT, which simulates brightness changes, rotation, scaling, and additive noise during optimization.

3 Evaluation

Experimental Setup. We used a YOLOv5-based traffic light detector trained on 6,657 DTLD images for both patch generation and evaluation. As camouflage targets, we used the 43 reference images, one per class, from the GTSRB traffic sign dataset [3]. The patch width was fixed to 2.4 times the traffic light width in both digital and physical-world experiments. We report attack success rate (ASR): in the digital setting, the proportion of traffic lights detected in clean images but no longer detected after patch application; in the physical setting, the proportion of recorded video frames in which no detection matched the ground-truth traffic light with IoU ≥ 0.5.

Digital-World Evaluation. We generated patches using all 43 GTSRB classes as camouflage targets, placing each at four positions around the target traffic light (above, below, left, right). For every condition, a separate patch was optimized on the same 1,041 DTLD images and then evaluated on a held-out set of 1,388 DTLD images. We additionally used a GTSRB-trained ResNet-50 to verify that each generated patch was classified as its intended traffic sign class. The patch shown in Figure 2 achieved the highest ASR of 85% when placed to the right of the traffic light. Left and right placements were consistently more effective than above and below, with average ASRs of 62%, 58%, 33%, and 23%, respectively. Circular patches also outperformed non-circular ones (average ASR 54% versus 29%).

Physical-World Evaluation. For physical-world evaluation, we tested three high-ASR patches from the digital experiments: "Speed limit 70" (SL70, right), "Go straight or right" (GoSR, right) and "No vehicles over 3.5 t permitted" (NV3.5, left). We confirmed that these patches were classified as their target sign class. We used a traffic light mock-up [4] and a representative automotive-grade C1-085 camera [5], and recorded 10-second videos under six recording conditions: 2.5 m, 5.0 m (at 0°, left 30°, and right 30° viewing angles), 7.5 m, and 10 m. Figure 3 and Table 1 show representative examples and quantitative results, respectively. The baseline failure rate was near zero, while 11 of 18 patched conditions achieved ASR $\geq$80%. The lower ASR at 2.5 m may reflect the underrepresentation of close-range traffic lights in DTLD. For GoSR, false positives appeared near the patch at 2.5 m and in the left 30° condition; a ResNet-50 color classifier labeled the false positive areas as green despite the red traffic light, suggesting possible downstream effects.



(a) Physical-world setup (b) Traffic light w/o patch (c) Traffic light w/ patch

Figure 3: Physical-world examples of the proposed attack.

Table 1: Baseline detection failure rate and ASR under physical-world conditions. Baseline reports the failure rate without a patch; the remaining columns report ASR.

Condition	Baseline	SL70	NV3.5	GoSR
2.5 m	0.000	0.000	0.000	0.690
5.0 m	0.000	0.043	**1.000**	**1.000**
7.5 m	0.000	**0.887**	**1.000**	**0.970**
10 m	0.003	0.723	0.000	**1.000**
Left 30°	0.000	**1.000**	**1.000**	**1.000**
Right 30°	0.000	**0.983**	**1.000**	0.353

4 Future Work

We plan to explore GAN-based patch generation to make adversarial patches resemble realistic stains, such as dirt or rust. We also plan to evaluate the attack's impact on end-to-end autonomous driving behavior.

Acknowledgments. This work was partially supported by JST CREST JPMJCR23M4.

Ethical Considerations. Our attack targets a general class of DNN-based detectors rather than any specific commercial product. We disclose it to motivate safer autonomous driving, not to enable misuse; all experiments were conducted on controlled private property.

Use of Generative AI. Generative AI tools were used for editorial purposes in this manuscript, and all outputs were inspected by the authors to ensure accuracy and originality.

References

[1] Svetlana Pavlitska et al. Fool the Stoplight: Realistic Adversarial Patch Attacks on Traffic Light Detectors. In *IEEE IV*, pages 734–741, 2025.

[2] Andreas Fregin et al. The DriveU Traffic Light Dataset: Introduction and Comparison with Existing Datasets. In *2018 IEEE ICRA*, pages 3376–3383, 2018.

[3] J. Stallkamp et al. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. Selected Papers from IJCNN 2011.

[4] AliExpress. Full-color LED traffic light, red/yellow/green, 100 mm. https://ja.aliexpress.com/item/1005006386087304.html.

[5] TIER IV. C1 Camera Technical Reference Manual. https://tier4.github.io/edge-auto-docs/reference_manual/C1_technical_reference_manual.html.