

ヘッドライト光の反射を悪用した標識認識への攻撃： 商用車両を用いた影響評価と対策の提案 *

鶴岡 豪^{1,2,a)} 佐藤 貴海³ Qi Alfred Chen⁴ 野本 一輝^{1,7} 小林 竜之輔¹ 田中 優奈^{1,7}
森 達哉^{1,5,6}

概要：標識認識システム (TSR) は自動運転車の安全において重要であるが、近年の研究では、TSR が敵対的攻撃に対して脆弱であることが知られ、様々な攻撃が提案されてきた。しかし既存の攻撃はステルス性と実現可能性に制約があり、それらを解消する新しい攻撃ベクトルとして、我々は再帰反射素材を利用した夜間のヘッドライト照射時のみ動作する敵対的再帰反射パッチ (ARP) 攻撃を提案した。これまで我々は研究用の TSR に対する評価を行ってきたが、現実的な脅威評価及び有効な防御の開発が課題として残されていた。本研究では、(1) 商用 TSR への脅威評価、(2) 様々な速度での攻撃有効性評価、(3) ARP 攻撃への防御手法の開発という課題に取り組んだ。その結果として、実車両を用いた動的環境評価では、速度 5-25 km/h の範囲で 93.4%以上の攻撃成功率を達成し、速度変化に対する頑健性を実証した。また、2024 年製の商用車両に搭載された TSR に対して評価を行い、SpeedLimit サインへの攻撃で 60-75%の成功率を達成し、ARP 攻撃が商用の TSR に対しても脅威であることを明らかにした。また ARP 攻撃の防御として、反射の物理を利用した防御手法である DPR Shield を設計し、75%以上の防御成功率を達成した。

キーワード：自動運転、物体検知、交通標識認識、敵対的サンプル、再帰反射

Adversarial Retroreflective Patch Attacks against Traffic Sign Recognition Systems: Impact for Commercial Vehicles and Defenses

GO TSURUOKA^{1,2,a)} TAKAMI SATO³ QI ALFRED CHEN⁴ KAZUKI NOMOTO^{1,7}
RYUNOSUKE KOBAYASHI¹ YUNA TANAKA^{1,7} TATSUYA MORI^{1,5,6}

Abstract: Traffic sign recognition systems are crucial for autonomous vehicle safety but are vulnerable to adversarial attacks. Due to constraints in stealth and feasibility of existing attacks, we proposed the Adversarial Reflective Patch (ARP) attack, which utilizes retroreflective materials and become effective only when illuminated by headlights during nighttime. In this work, we addressed three challenges: (1) evaluation against commercial TSR, (2) attack effectiveness evaluation at various speeds, and (3) development of defense methods. We demonstrated that ARP attack could achieve over 93.4% attack success rate at speeds of 5-25 km/h, and 60-75% success rate against commercial vehicles manufactured in 2024, which demonstrate that ARP attacks could be effective even against commercial TSR. We also designed the DPR Shield defense method using two polarizing plates, achieving over 75% defense success rate.

Keywords: Autonomous driving, Object detection, Traffic sign recognition, Adversarial Example, Retroreflective material

¹ 早稲田大学/Waseda University

² 産総研/AIST

³ 慶應義塾大学/Keio University

⁴ カリフォルニア大学アーバイン校/UC Irvine

⁵ 情報通信研究機構/NICT

⁶ 理研 AIP/RIKEN

⁷ デロイトトーマツサイバー合同会社/Deloitte Tohmatsu Cyber

LLC

a) go@nsl.cs.waseda.ac.jp

*1 本研究は著者らの過去の発表 [18] (CSS2024) を発展させ、自動車走行時における攻撃有効性評価、商用 TSR を用いた攻撃有効性評価、再帰反射の物理に基づく効果的な防御手法の提案・評価を新たに実施し、その結果を報告するものである。

1. はじめに

自動運転車を含めたすべての道路交通者は交通標識を遵守しなければならない。自動運転車両が標識を認識するための標識認識システム（以下 TSR）には深層学習ベースの手法が用いられており、高い性能を誇っている。しかしながら、深層学習モデルは敵対的攻撃に対して脆弱であることが知られており、主に 2 つの攻撃ベクトルが提案されている。一つは悪意のあるパッチを標識に貼り付けるパッチ攻撃でもう一つは、プロジェクターなどで悪意のあるパターンを投影する光・レーザー投影攻撃である。それぞれの攻撃は想定された環境において高い有効性を示すが、それぞれ制約がある。パッチ攻撃は攻撃の実現が容易な一方で、一度攻撃が有効化すると常に目立つため、ステルス性にかける。その一方で、光・レーザー投影攻撃は攻撃時のみ投影すれば良いためステルス性に優れるが、大型の機材が必要な上に、標識のそばに攻撃者がいる必要があるため、攻撃実現性にかける。これらの制約を解決するために、図 3 に示すようなヘッドライトによって有効化する、敵対的再帰反射パッチ攻撃 (ARP) 攻撃を提案し、評価していた [18]。しかし、以前の研究では、実際に TSR が使われる環境での影響評価及び、より有効な防御手法の探索が課題となっていた。特に前者については、5 km/h に制約されていた上に、商用の TSR に対する脅威の有効性が未知であった。そのため、本研究では、様々な速度における実車両を用いた ARP 攻撃の有効性評価及び、実際の TSR に対する攻撃の有効性の評価そして、より効果的な防御の提案と評価を行った。

本研究は、敵対的再帰反射パッチ攻撃の有効性をより現実的な状況で評価し、実社会における脅威を評価すること及び ARP 攻撃に対するより有効な防御を探査することを目的とする。§ 3においては、脅威モデルや ARP 攻撃の生成方法の概要について示す。§ 4においては、実際の商用車にカメラを搭載し、オープンソースの TSR への影響を様々な速度での走行環境下で評価した。§ 5においては、実際の商用車に搭載されている TSR 2 つを対象に攻撃の影響を評価し、オープンソースの TSR の攻撃との比較を行った。そして、§ 6においては、ARP 攻撃の防御として、反射の物理特性を利用した防御方法である DPR Shield を提案し、その有効性を評価した。最後に § 7 では、本研究の将来的な課題について議論した。

本研究の貢献は以下のとおりである。

- 様々な速度での走行下において ARP 攻撃に対して評価を行い、速度によらず安定して平均 93.4%以上の攻撃成功率 (ASR) を達成することを確認した。
- 最新の商用 TSR モデルに対して SpeedLimit サインに対する ARP 攻撃の有効性を検証し、60-75%の ASR を確認し、既存の SpeedLimit サインへの商用モデル

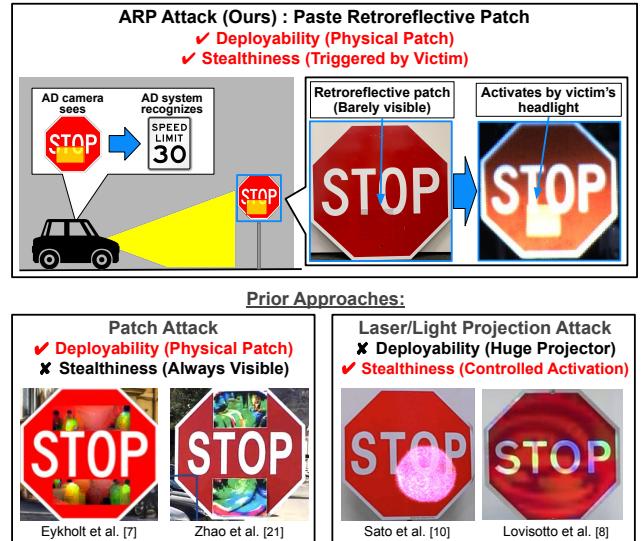


図 1: ARP 攻撃の概要図。攻撃に用いる再帰反射パッチはヘッドライト光に照射された時のみ有効化し攻撃として機能する。ARP 攻撃はパッチ攻撃の攻撃実現性と、光・レーザー照射攻撃のステルス性を兼ね備えている。

への攻撃の成功率 (0%) よりも高いことを確認し、脅威の実社会への影響を比較検討した。

- ARP 攻撃に対する防御として、反射の物理を利用した防御である DPR Shield を提案し、評価した。その結果として single-stage, two-stage ともに 75%以上の防御成功率を達成した。

2. 背景と関連研究

2.1 カメラベースの交通標識認識 (TSR) システム

交通標識認識 (TSR) はカメラから取得した画像からリアルタイムで交通標識を識別するものである。自動運転の安全にとって重要な機能であり、すでに Toyota の safetysense [13] などの ADAS にも使われている。カメラベースの TSR は主に 2 つのタイプに分類され、single-stage と、two-stage と呼ばれている [5]。Single-stage TSR では、DNN ベースの物体検出器を利用し、画像内の交通標識の位置とクラスを同時に認識する。Two-stage TSR は、物体検出器と分類器の 2 つの DNN モデルを利用する。第一段階の物体検出器は、画像中の交通標識の領域を決定し、画像を切り取る。第二段階の分類器では、物体検出器で切り取られた画像のクラスを分類する。本研究では両方のアーキテクチャの TSR に対する防御の影響を評価した。

2.2 TSR に対する物理世界での敵対的攻撃

TSR は悪意のある入力、敵対的攻撃に対して脆弱であることが知られている。主に敵対的攻撃は 2 つの攻撃ベクトルに分類され、パッチ攻撃 [6, 16] とレーザー・光投影攻撃 [10, 12] に分けられる。前者は標識に悪意のあるパッチを貼り付ける攻撃であり、攻撃を設置するのは容易な一

方、常に見えるためステルス性には劣る。後者は攻撃時のみレーザーや光を照射すればいいため、ステルス性には優れるが、プロジェクター等の大型の機材が必要な上、攻撃者が標識の近くにいる必要があり、ステルス性にかける。我々が提案する ARP 攻撃は、これら両方の手法の利点を併せ持つ。再帰反射パッチを用いることで、攻撃の実行が容易でありながら、ターゲット車両のヘッドライトが当たる瞬間以外はステルス性を保つことができる。つまり、攻撃実行時以外は標識の外観が変わらない、高いステルス性と実現可能性を両立させる新たな攻撃手法である。

2.3 商用 TSR への攻撃評価

TSR に対する敵対的攻撃の実世界での影響を評価するため、近年では学術的な TSR モデル (YOLO や Faster R-CNN など) に加えて、商用車に搭載された実際の TSR システムに対する攻撃評価が行われている。Wang et al. [14] の研究では、商用 TSR システムに対する大規模な測定実験が初めて実施された。その結果、学術モデルでは高い成功率を示す既存の攻撃手法が、商用システムに対しては大幅に成功率が低下することが明らかになった。この成功率低下の主要因として、商用 TSR システムに実装されている空間的記憶 (spatial memorization) 機能が特定された。この機能は、一度検知された標識を車両がその位置を通過するまで記憶し続けるというもので、従来のオブジェクトトラッキング (数秒程度) とは異なり、時間に依存せず空間的条件に基づいて動作する。この空間的記憶の影響により、攻撃者が標識検知を妨害しようとしても、車両が標識に到達するまでの間に一度でも検知に成功すれば、システムレベルでは攻撃は失敗となる。論文では、この設計により特に hiding attack (標識を隠す攻撃) の成功が数学的に困難になることが証明されており、TSR システムのセキュリティ評価において空間的記憶の考慮が不可欠であることが示されている。本研究では、商用の TSR モデルを 2 つ選び、ARP 攻撃の有効性を検証した。

2.4 再帰反射パッチの物理と偏光

再帰反射: 再帰反射は、光を光源の方向に返す反射であり、ガラスピーズやプリズムなどの特殊に設計された物質によって実現される。再帰反射素材は、交通標識に広く応用されており、自動車のヘッドライト光を効果的に反射することで、夜間でもドライバーにとって高い視認性を確保する。

偏光フィルタと反射の防御: 光は電磁波の一種であり、自然光では電磁波が様々な方向に振動している。偏光フィルタ（偏光板）は特定方向の振動成分のみを通過させる光学素子であり、自然光が偏光フィルタを通過すると、特定の一方向のみに振動する偏光に変換される。重要な点として、偏光状態は反射においても維持される性質がある。そのため、再帰反射パッチからの反射光は、入射光の偏光特

表 1: パラメータの定義。 i はパッチのインデックスを示す

シナリオ パラメータ	説明	攻撃 パラメータ	説明
d_{lon}	距離: 車 \leftrightarrow 標識	(x_p^i, y_p^i)	i の座標
d_{lat}	距離: 車 \leftrightarrow 標識	(W^i, H^i)	i の幅・高さ
h_s	標識の高さ	$C^i \in \mathbb{R}^3$	i の色
h_l	ヘッドライト高さ	R^i	i の再帰反射係数
L	環境光強度		
L_H	ヘッドライト強度		

性をそのまま保持して返される。この物理的性質を利用すると、既に偏光状態にある光が、その振動方向と垂直に配置された偏光板に入射した場合、理論的に完全に遮断される。本手法はこの偏光の原理を応用したものである。具体的には、車両のヘッドライトに偏光フィルタを配置して偏光化された光を生成し、カメラ前には垂直方向の偏光板を設置する。これにより、再帰反射パッチからの攻撃光を選択的に遮断しながら、正常な標識からの拡散反射光は部分的に透過させる ARP 攻撃に対する防御手法、DPR Shield を実現する。

3. ARP 攻撃の脅威モデルと攻撃生成方法

ARP 攻撃についての脅威モデルと攻撃生成方法の概要について説明する。詳細は CSS2024 における ARP 攻撃の提案 [18] に譲る。

3.1 脅威モデル

攻撃モデル: 図 2 に ARP 攻撃の攻撃モデルの概要を示す。基本的に従来のパッチ攻撃と同じ脅威モデルに従っているが、相違点として攻撃を引き起こすための前提条件にターゲット車両のヘッドライトが挙げられる。攻撃とシナリオパラメータの定義は表 1 に示す。攻撃者は、予め定めた交通標識上に、悪意ある再帰反射パッチを (x_p^i, y_p^i) の位置に (W^i, H^i) のサイズで貼り付ける。再帰反射パッチの基本色 C_i は、ステルス性を保つために、貼り付ける標識の表面と同じ色とする。この脅威モデルでは、標識の高さ h_s 、ヘッドライトの高さ h_l 、環境光 L 、ヘッドライトの強度 L_H を定義している。攻撃者は、ターゲットとなる車両及び周囲の環境を測定することで、これらのシナリオパラメータを推定及び測定することができると仮定する。攻撃者の目的は、自動運転車両の TSR を騙し、事故に繋がりうる交通規則違反を引き起こすことである。例えば、STOP サインの検知失敗は交差点において、他の車両との衝突を引き起こす可能性がある。また、速度制限標識の誤分類は、予期せぬ加速や減速を引き起こす可能性がある。

パッチの選択: 以前の ARP 攻撃の研究 [18] においては、ARP 攻撃の能力を網羅的かつ体系的に調査するために、表 2 に示す 4 種類の代表的な再帰反射素材を選択した。これらの素材は、再帰反射材料のグレードと構造を網羅する

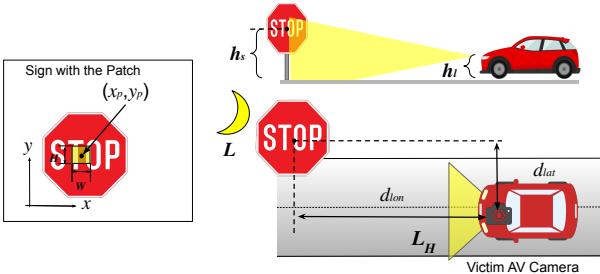


図 2: ARP 攻撃のパラメータ. 本攻撃はヘッドライトによって有効化される攻撃であるため、既存のパッチ攻撃でのパラメータに加え、ヘッドライトの明るさ等を考慮する.

表 2: 選択した再帰反射素材. ASTM D4956, AASHTO M268 に定められたグレードを網羅する.

名前	NittoL [19]	HIP3930 [2]	Nikkalite [11]	DG4090 [1]
メーカー	Nitto	3M	Nikkalite	3M
ブランド	Engineer Gr.	HIP	Crystal Gr.	Diamond Gr.
シリーズ	HT	3930	CRG 92000	4000
再帰反射材	GlassBeads	MicroPrism	MicroPrism	MicroPrism
1m ² の価格	\$80	\$277	\$479	\$724
ASTM	I	III, IV	VIII	XI
AASHTO	N/A	B	B	D

*Nitto: 日東エルマテリアル

ように選択した. 今回はその研究の結果に基づき、最も攻撃性能が高いパッチを攻撃研究では採用する. また、防御評価においては、構造が網羅的になるようにパッチを選択し、評価した. 実際に用いたパッチについては、各実験の実験設定に記載する.

3.2 攻撃生成手法の概要

ARP 攻撃の生成パイプラインの概要について述べる. 効果的な ARP 攻撃を実現するために、再帰反射素材の特性を考慮した攻撃生成パイプラインを構築した. このパイプラインの目的は、攻撃パッチの最適な貼付位置およびサイズを探査し、攻撃の効果を最大化することである. 図 3 に、ARP 攻撃生成パイプラインの全体像を示す. ARP 攻撃生成パイプラインは、主に 3 つの過程から構成される: (1) データ収集、(2) パラメータ推定、(3) パッチ最適化. データ収集のステップにおいては、再帰反射素材の光学特性を正確に把握するため、対象となる自動運転車両で使用されるカメラとヘッドライトを用いて実測データを取得する. 具体的には、ASTM E810 基準に基づいて 15 m 離れた距離から再帰反射パッチにヘッドライト光を照射し、カメラ画像を収集する. また、再帰反射素材とヘッドライトのデータシートから仕様値も併用することで、より信頼性の高いパラメータ推定を可能にする.

次にパラメータ推定のステップでは、収集したデータに

基づいて Blender 上で再帰反射挙動を再現するための物理パラメータを算出する. 標準的な 3D シェーディングツールには再帰反射機能が実装されていないため、垂直反射面を仮想的に配置することで再帰反射を模擬する手法を開発した. IoR Level (反射光強度), Specular Tint (反射色), Roughness (表面粗さ) の 3 つの主要パラメータを、光学の理論に基づいて計算し、決定する.

最後にパッチ最適化では、昼間のステルス性と夜間の攻撃効果を両立させる最適なパッチ配置を探索する. 以下の式 1 に示す目的関数を用いることにより、昼間でのステルス性と夜間における攻撃性能の両立を図り、Tree-structured Parzen Estimator [3] によるブラックボックス最適化を適用する. この過程により、視覚的に目立たず、かつ高い攻撃効果を持つ ARP 攻撃の自動生成が可能となる.

$$\begin{aligned} \min_{x_p, y_p, W, H} & \mathbb{E}_{t \sim T} [L_{\text{attack}}(X_{\text{night}}) + \alpha L_{\text{stealth}}(X_{\text{day}})] \quad (1) \\ \text{s.t. } & X_{\text{day}} = \text{SimulateARP}_{\text{day}}(x_p, y_p, W, H) \\ & X_{\text{night}} = \text{SimulateARP}_{\text{night}}(x_p, y_p, W, H) \\ & W \times H \leq MPR \cdot W_{\text{bbox}} \cdot H_{\text{bbox}} \end{aligned}$$

4. 自動車走行環境での攻撃性能評価

ARP 攻撃のより現実的な脅威を調査するために、自動車実験走行環境における攻撃成功率の評価を行った.

4.1 実験設計

§ 3.2 で述べた攻撃パイプラインを用いて、攻撃のために位置やサイズを最適化した再帰反射パッチを実際の標識に適用し、設置する. そして、自動車にカメラを乗せて録画を行い、動画の画像フレームを TSR に入力し、出力結果を見ることで攻撃の有効性を評価する. 攻撃対象の TSR は single-stage と two-stage の両方を用いた. 前者は ARTS データセットで訓練した YOLOv5 を用いた. このモデルは、 $mAP_{0.5}$ が 0.83 であった. 本研究では、先行研究 [12] に従い、信頼度のしきい値を 0.3 に設定した. Two-stage TSR には、二段階目の分類に着目し、ARTS データセットで訓練された独自の分類器を使用した. この分類器は、畳み込み層、マックスプーリング層、パッチ正規化層、全結合層、ドロップアウト層を含む計 14 層からなる CNN モデルアーキテクチャ (以下 simple CNN) を採用する. このモデルの分類精度は 83% であった.

攻撃対象の標識としては、アメリカの STOP サイン及び、SpeedLimit 65 mph サインを対象とした. 攻撃の生成においては、標識とカメラ及びヘッドライトが 15 m 離れた環境で攻撃の最適化を行った. 標識の高さ (h_s) は、先行研究 [8] に従い 1.5 m とし、ヘッドライトの高さ (h_l) は 0.75 m に設定した. カメラは、オープンソース自動運転システムの Autoware [9] で参照されている、FLIR Machine

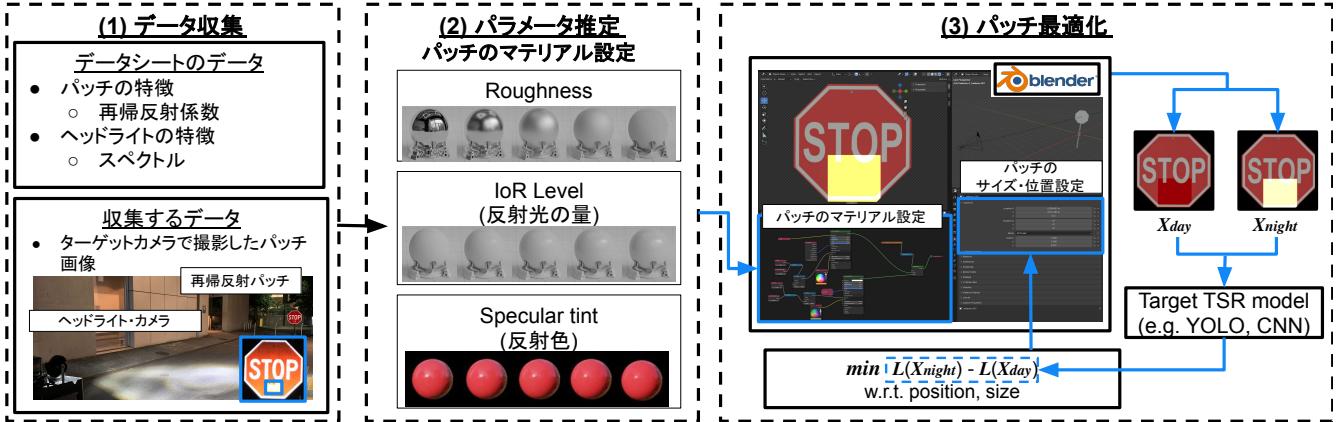


図 3: ARP 攻撃生成パイプライン。以下の 3 つのフェーズからなる: 1. パラメータ推定のためのデータ収集 2. 再帰反射のパラメータ推定と再現 3. 最適化ベースの ARP 攻撃生成

Vision Camera BlackflyS (BFS-PGE-16S2C-CS) [7] の特性を再現した。攻撃に使用したパッチは、最も効果的な素材及びサイズの条件で最適化したものを利用した。STOP サインに対しては DG4090 素材を用い、single-stage TSR では最大パッチサイズ 18.75%, two-stage TSR では 12.5% とした。SL65 サインに対しては NittoL 素材を用い、両アーキテクチャで最大パッチサイズ 6.25% とした。

自動車走行環境については、一般的な乗用車のダッシュボードの中央に FLIR BlackflyS (BFS-PGE-16S2C-CS) を設置し、標識が設置された 50 m の直線コースを一定速度で走行することにより、ARP 攻撃の攻撃性能を測定した。なお、速度は 5, 15, 25 km/h で評価を行った。

評価指標: 既存研究に従って攻撃成功率 (Attack Success Rate, ASR) を採用した。これは全体のフレーム数のうち、攻撃に成功した、つまり誤認識を引き起こしたフレーム数の割合である。今回は 5 m の区間に区切り、5 回の走行データを統合して攻撃成功率を算出した。その際、single-stage TSR については撮影した動画のフレーム画像をそのまま用い、two-stage TSR については、第二段階の分類に焦点を当て、標識部分を手動でクロップして分類器に入力して評価する。

4.2 結果

攻撃の有効性については標識の種類によって、異なる距離依存の有効性を示した。SL65 標識に対する攻撃は、TSR アーキテクチャや車両速度によらず一貫して 90% 以上の高い攻撃成功率を示した。これは、SpeedLimit 標識が他の速度制限値 (30 mph, 45 mph 等) と視覚的に類似しており、数字部分の微小な変化でも誤分類を誘発しやすいためと考えられる。一方、STOP サインに対する攻撃では距離依存性が観察され、35m より遠い距離では速度・アーキテクチャによらず 90% 以上の攻撃成功率を達成したが、近距離では性能が低下した。この距離依存性は再帰反射の物理



図 4: 運転環境における実験のセットアップ図。カメラはダッシュボードの中央に設置し、一定の速度で標識の 50 m 手前から走行し標識を通過する様子をカメラで録画した。

的特性で説明できる。遠距離では、ヘッドライトと標識、標識とカメラが成す角度が小さくなり、再帰反射素材の光源方向への反射するという特性が最も効果的に発揮されるためであると考えられる。逆に近距離では角度が大きくなることで反射効率が低下し、攻撃効果が減衰したと考えられる。両標識に共通する重要な特徴として、攻撃効果が車両速度 (5-25 km/h) に依存しない高いロバスト性が確認された。特に 35 m よりも遠い位置においては、多くの場合において 95% 以上の攻撃成功率となっている。これは、高速道路に SL65 標識が実際に適用される環境や都市部の STOP 標識のような多様な交通環境において本攻撃が有効であることを示しており、APR 攻撃の脅威が現実世界においても高いことを示唆していると考えられる。

5. 商用の TSR に対する攻撃有効性評価

ARP 攻撃の実社会での影響を調査するため、商用の自動車に搭載された TSR を対象に攻撃の有効性を調査した。

5.1 実験設計

X 社のモデル A (2024 年モデル) 及び、Y 社のモデル B (2024 年モデル) に対して搭載された TSR を対象に攻撃の有効性を検証した。攻撃の対象標識として、日本の速度制限 30 km/h 標識を用いた。これは、今回対象にしたモデルに共通して認識が可能な標識であったためである。攻撃

表 3: ASR of the ARP Attack in driving scenarios with actual vehicles at different speeds.

Sign	Model	Speed	ASR at different distances. Bold numbers indicate ASR above 80%.						
			15-20 m	20-25 m	25-30 m	30-35 m	35-40 m	40-45 m	45-50 m
STOP	YOLOv5	5 km/h	61.7%	72.2%	76.4%	99.8%	94.7%	97.3%	100%
		15 km/h	73.5%	51.4%	64.5%	73.0%	99.8%	100%	100%
		25 km/h	44.5%	91.3%	100%	100%	100%	100%	100%
	SimpleCNN	5 km/h	55.6%	60.9%	70.1%	67.1%	93.4%	98.6%	94.9%
		15 km/h	78.2%	75.9%	98.3%	100%	98.1%	96.4%	94.7%
		25 km/h	28.0%	60.8%	95.6%	90.2%	93.6%	87.2%	89.0%
SL65	YOLOv5	5 km/h	100%	100%	100%	100%	100%	100%	100%
		15 km/h	100%	100%	100%	100%	100%	100%	100%
		25 km/h	100%	100%	100%	100%	100%	100%	100%
	SimpleCNN	5 km/h	99.7%	96.2%	98.9%	99.4%	99.5%	100%	100%
		15 km/h	99.6%	100%	99.8%	100%	100%	100%	100%
		25 km/h	100%	99.7%	100%	98.6%	99.6%	100%	

生成においては、GTSRB で訓練した YOLOv5 を対象にしたモデルに対して攻撃を生成し、それを実際の速度制限 30km/h 標識に対して適用した。なお、再帰反射パッチについては、最大パッチサイズ 0.0375, 再帰反射パッチの種類は NittoL を用いた。

実験指標: 商用の TSR は § 2 でも述べた通り、標識通過して初めて認識結果が明らかになる。そのため、標識通過後に標識を正しく認識しなかった場合を攻撃成功と定義し、総試行回数 (20 回) に対する、攻撃成功の割合を攻撃成功率として定義した。

5.2 実験結果

実験結果から、モデル A では 60% の ASR、モデル B では 75% の ASR となり、学術用 TSR と比較して低い ASR を示した。これは商用 TSR が空間記憶機能を実装しており、一度でも SpeedLimit サインを正常に認識した場合、その後のフレームで仮に認識をしなかったとしても標識が認識するものとして判定する機能によるものであると考えられる。しかし従来の攻撃である SIB [16] や FTE [8] は、SpeedLimit サインに対する攻撃成功率が 0% であったことを踏まえると本攻撃は高い攻撃成功率を達成しており、実社会への脅威が大きい攻撃と言える。

これは、再帰反射パッチが光源の位置によらず、光源の方向へ強く光を返すという特性が、自動車走行環境において継続的に攻撃が有効かすることにつながり検知回避が可能となった結果、空間記憶を持った商用 TSR に対しても、有効な攻撃となったと考えられる。

6. 防御評価

本研究はパッチ攻撃の特性を受け継ぐものである。そのため、理論的にはパッチ攻撃の防御が適用可能である。そこでまず、既存のパッチ攻撃に対する防御の議論をした後に、ARP 攻撃に対するより効果的な防御手法として、偏光

板フィルタを二枚用いる防御手法 DPR Shield を開発し、その有効性を評価した。

6.1 既存のパッチ攻撃に対する防御

既存のパッチ攻撃に対する防御は主に 2 つ存在し、経験的防御と証明可能な防御に分けられる。前者は一般的に Adaptive Attack に脆弱であることが知られている [4]。そのため、本研究では最先端の証明可能防御である PatchCleanser [15] が適用された環境における ARP 攻撃の防御性能について評価した。防御の評価においては STOP サイン及び SL65 サインを対象に single-stage, two-stage の両方のモデルに対する有効性を評価した。パッチや攻撃対象のモデルについては、§ 4 と同様のものを用いた。

防御性能の評価のために、カメラから 15 m 離れた環境に標識を設置し、ARP 攻撃が有効化された環境の画像を撮影し、PatchCleanser による防御性能を記録した。防御性能は PatchCleanser によって正しく分類できた画像数の割合である Clean Acc. と PatchCleanser が理論的な保証を持って正しく分類できた画像の割合を Certified Acc. として評価した。結果は表 4 に示す結果となった。STOP サインに対してはベースライン条件でも認識率が 0% となり、SL65 サインでは 100% の認識率を保つものの、攻撃条件下では完全に防御に失敗した。この結果は、PatchCleanser がランダムに画像の一部を隠して分類する戦略をとっており、これが TSR タスクにおいては、分類に必要な要素を隠してしまうことにつながった結果引き起こされた副作用であると考えられる。

6.2 DPR Shield の提案と評価

ARP 攻撃が再帰反射の物理的特性を利用することに着目し、偏光フィルタを用いた物理的な防御手法である DPR Shield を提案する。本手法の概要を図 5 に、実際のセットアップ図を図 6 に示す。本手法は、ヘッドライトに偏光フィルタを配置して偏光化された光を生成し、カメラ前に

表 4: ARP 攻撃に対する PatchCleanser による防御性能の評価. Clean Acc. PatchCleanser が正しく分類をした割合である. Certified Acc. は Patch Cleanser が理論的な保証の下で正しく分類した割合である.

	Benign		Attack			
	STOP		SL65	STOP		SL65
	No Defense	100%	100%	0%	0%	
Clean Acc	0%	100%	0%	0%		
Certified Acc	0%	0%	0%	0%		

表 5: 提案した DPRShield の攻撃なし (Benign) 及び攻撃あり (Attack) 状況下における分類精度の比較. DPRShield は攻撃条件下における認識精度を向上させつつ、攻撃なしの状況における認識精度も維持していることがわかる.

	STOP				SL65			
	Benign		Attack		Benign		Attack	
	Single	Two	Single	Two	Single	Two	Single	Two
No Defense	100%	100%	0%	0%	100%	100%	0%	0%
With Defense	100%	100%	100%	100%	100%	100%	100%	75%

垂直な偏光フィルタを配置することで、再帰反射パッチからの攻撃光を選択的に遮断する. 従来の単一偏光フィルタとは異なり、DPR Shield は光源と受光部の両方に偏光制御を適用する. これにより、正常な標識からの拡散反射光は部分的に透過する一方、再帰反射パッチからの偏光を保持した強い反射光は効果的に遮断される.

物理実験による評価の結果を表 5 に示す. また、実際の様子を図 7 に示す. DPR Shield は高い防御効果を示した. Single-stage TSR に対しては 100% の防御成功率を達成し、two-stage TSR に対しても 75% の防御効果を確認した. なお、攻撃がないベースライン条件下での認識率に影響を与えたかった. Two-stage TSR に対する SL65 サインへの攻撃の防御においては、75% の認識精度にとどまったのは、用いた再帰反射パッチの素材がグラスビーズタイプのパッチであり、DPR Shield では再帰反射光を吸収しそぎてしまい、黒色になってしまったためだと考えられる. 次の節において、グラスビーズタイプのパッチを用いた攻撃の防御の有効性について議論する.

6.3 グラスビーズタイプのパッチに対する防御評価

前の節において、グラスビーズタイプのパッチを用いた ARP 攻撃についての有効性について議論した. ここでは、STOP サインに対してグラスビーズタイプのパッチを用いた場合の DPR Shield の防御有効性について議論する.

実験では、NittoL グラスビーズパッチに対して DPR Shield を適用した結果、single-stage, two-stage TSR の両方で 100% の防御成功率を達成した. しかし、防御過程においてパッチ部分が黒くなる現象が観察された. これは、

DPR Shield がグラスビーズからの再帰反射光を過度に吸収したためである. しかし、分類精度に影響を与えたかったのは、STOP サインの赤色背景においては、この黒化現象による視覚的コントラストは最小限に抑えられ、実用上の問題とはならないことが確認された.

以上の結果より、DPR Shield は SpeedLimit サインを対象に、グラスビーズタイプのパッチで攻撃した場合において、一定の防御性能の低下がみられるが、その他の場合においては、有効な防御となりうると言える.

7. 議論

End-to-End レベルでの評価: 本研究では、TSR システムに対する ARP 攻撃の有効性を包括的に評価したが、これらの攻撃が経路計画や制御に与える連鎖的影響については評価を行っていない. ARP 攻撃の真の脅威を理解し、実用的な対策を講じるためには、攻撃が自動運転車の意思決定プロセス全体に与える影響の定量的評価が不可欠である. 特に、誤認識された標識情報が経路計画アルゴリズムや車両制御システムにどのように伝播し、最終的な車両挙動にどの程度の変化をもたらすかの分析は、攻撃のリスク評価において重要な要素となる. 今後の研究では、Overpass [17] のような自動運転システムを使ったエンドツーエンド評価や、制御された環境下での実車両テストにより、TSR への攻撃が自動運転システム全体に与える包括的な影響を評価する.

Adaptive Attack: 今回提案した DPR Shield には、グラスビーズタイプの再帰反射パッチについては、光を過度に抑制してしまうという問題があった. この特性を利用して、DPR Shield を破るような攻撃が実現できる可能性がある. このような攻撃の影響の評価と防御の提案については今後の課題とする.

倫理的考慮: 本研究では、倫理的考慮および実験時の安全の確保を行った. ARP 攻撃の自動車を用いた実験では、大学の敷地内部で実験を行い、進行方向上に人がいないよう配慮した上で実験を行った. また、公道を走行する TSR を搭載した自動車から見えないよう配慮をし、他の道路利用者の安全を損うことがないようにした. さらに、本研究の結果については自動車企業等に対して responsible disclosure を進めている.

8. 結論

本論文では再帰反射パッチによりステルスかつ攻撃実現性の高い攻撃をより実社会の状況に近い形で評価した. 具体的には、自動車が走行する環境や、実際の商用車に対する影響を調査した. その結果として、ARP 攻撃は自動車の速度によらず 35m より遠い距離においては 90% 以上の高い攻撃性能を持つことが明らかになった. また商用車への攻撃性能について、既存の攻撃では ASR が 0% であっ

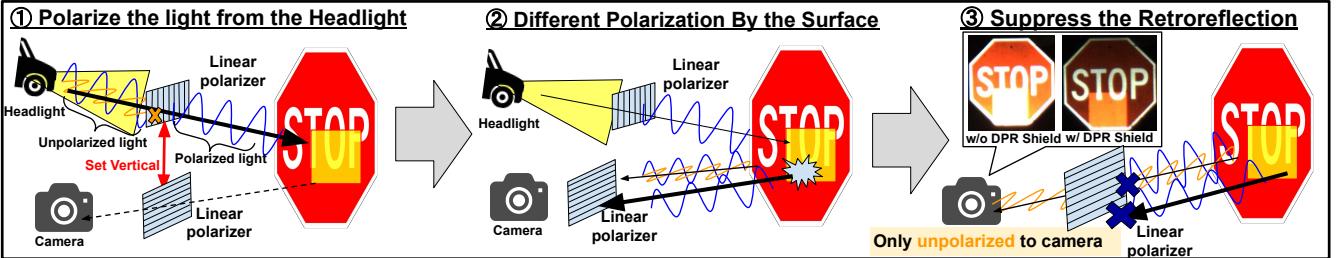


図 5: 偏光版を用いた ARP 攻撃への防御 DPR Shield の概要. DPRShield は 3 つのステップからなる. (1) ヘッドライト光を変更状態にする. (2) 反射時に異なる変更状態となり, 再帰反射光のみ偏光状態が維持される (3) カメラに取り付けられた偏光板によって ARP 攻撃による再帰反射光のみを除く.

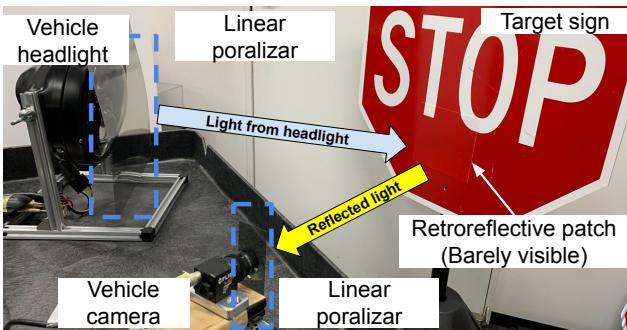


図 6: DPR Shield のセットアップ図. 偏光板はカメラとヘッドライトの両方に取り付ける. これらの偏光板はフィルムとしても入手可能でカメラのレンズ及びヘッドライトの表面に取り付けることも可能である.



図 7: STOP サイン及び SL65 サインに対する, 防御手法別の ARP 攻撃条件下における図.

たのに対して, ARP 攻撃では最大 75% の攻撃成功率が確認され, 現実世界における脅威性が高い攻撃であることが明らかになった. また, これまでの再帰反射パッチの防御方法の制約や既存の防御の制約を踏まえ, 再帰反射の物理を利用した DPR Shield を評価し結果として, 一部のパッチ・標識の条件以外では 100% の防御成功率を示し, 有効性を明らかにした.

謝辞 本研究の一部は JSPS 科研費 22H00519, JST CREST JPMJCR23M4, NEDO JPNP25006 の結果得られたものです. また, 本研究は JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援）JPMJBS2429 の支援を受けたものです.

参考文献

- [1] 3M Company: 3M™ Diamond Grade™ DG3 Reflective Sheeting Series 4000 (2023).
- [2] 3M Company: 3M™ High Intensity Prismatic Reflective Sheeting Series 3930 (2023).
- [3] Akiba, T. et al.: Optuna: A Next-generation Hyperparameter Optimization Framework, *ACM SIGKDD*, pp. 2623–2631 (2019).
- [4] Chiang, P.-Y. et al.: Certified Defenses for Adversarial Patches, *ICLR* (2020).
- [5] Ertler, C., Mislej, J., Ollmann, T., Porzi, L. and Kuang, Y.: Traffic Sign Detection and Classification around the World, *CoRR*, Vol. abs/1909.04422 (online), available from <<http://arxiv.org/abs/1909.04422>> (2019).
- [6] Eykholt, K. et al.: Physical Adversarial Examples for Object Detectors, *CoRR*, Vol. abs/1807.07769 (2018).
- [7] FLIR Systems, Inc.: Blackfly S GigE (2024).
- [8] Jia, W. et al.: Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems (2022).
- [9] Kato, S. et al.: Autoware On Board: Enabling Autonomous Vehicles with Embedded Systems, *ICCP'S18*, IEEE Press, pp. 287–296 (2018).
- [10] Lovisotto, G. et al.: SLAP: Improving Physical Adversarial Examples with Short-Lived Adversarial Perturbations, *USENIX Security* (2021).
- [11] Nippon Carbide Industries (USA) Inc.: Nikkalite Brand 92000 Series White Crystal Grade Super High Intensity Microprismatic Reflective Sheeting (2023).
- [12] Sato, T. et al.: Invisible Reflections: Leveraging Infrared Laser Reflections to Target Traffic Sign Perception, *NDSS* (2024).
- [13] US, T.: Road Sign Assist, Toyota Safety Sense, <https://www.toyota.com.au/toyota-safety-sense/road-sign-assist>.
- [14] Wang, N. et al.: Revisiting Physical-World Adversarial Attack on Traffic Sign Recognition: A Commercial Systems Perspective, *NDSS* (2025).
- [15] Xiang, C. et al.: PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier, *USENIX Security 22*, pp. 2065–2082 (2022).
- [16] Zhao, Y. et al.: Seeing isn't Believing: Practical Adversarial Attack Against Object Detectors (2019).
- [17] 野本一耀ほか: Overpass: セキュリティ評価のための自動運転プラットフォームの提案と評価, 情報処理学会研究報告, pp. 199– (2024).
- [18] 鶴岡 豪ほか: ヘッドライトの反射光を悪用する敵対的パッチ攻撃の提案と評価, CSS2024 論文集, pp. 401–408 (2024).
- [19] 日東エルマテリアル: 広角反射テープ (2023).