

# 敵対的再帰反射パッチ: 暗闇で有効化する敵対的攻撃

---

鶴岡 豪(早稲田大学)

野本一輝(早大), 小林 竜之輔(早大)

田中優奈(早大), 森 達哉(早大/NICT/理研AIP)



# 目次

---

1. 論文概要
2. 自動運転の仕組みと交通標識認識
3. 交通標識認識に対する攻撃と現状
4. 現状を踏まえた提案手法の概要
5. 評価
6. 今後の方針
7. まとめ

# 論文概要

## 敵対的再帰反射パッチ: 暗闇で有効化する敵対的攻撃

再帰反射パッチを用いた交通標識認識に対する検知回避攻撃

新規性: 夜のみで有効・昼間におけるステルス性

デジタル実験・物理実験で以下を検証

RQ1. 攻撃は実現可能であるか

RQ2. 角度や距離に対して頑健か

→攻撃実現性・頑健性が確認

→反射板は交通標識認識に対して脆弱



# 自動運転の仕組み

- Sensing...センサで周囲の把握
- Localization...位置推定(地図とセンサの情報)
- Perception...周囲の状況認識(Sensingに基づく状況把握)
- Planning...運転ルート等を決定する
- Control...決定に基づいた制御



# Perceptionについて

Perception : 障害物等の周囲の状況認識→安全に最も直結する部分

例: カメラ画像からの物体検出

- 障害物検知
- 交通標識認識

→現在運転支援などで利用

今回は交通標識認識に注目



# 交通標識認識の仕組み

## 交通標識認識モデル

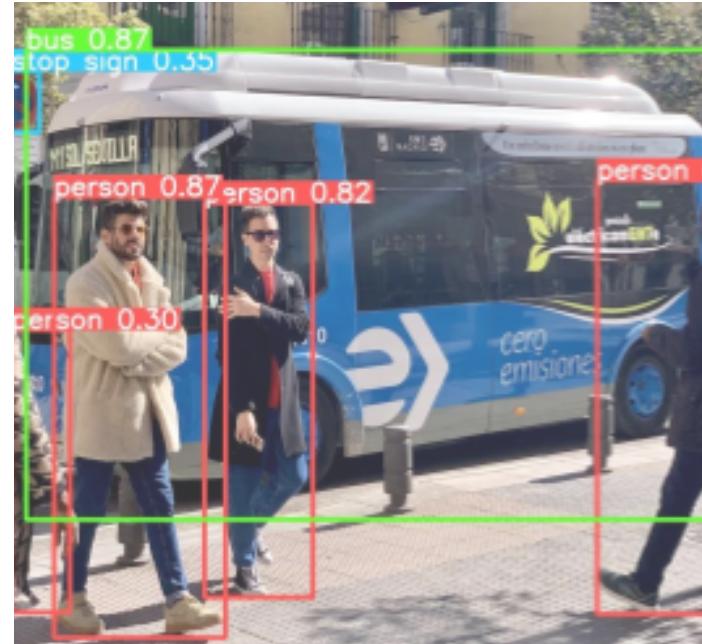
→物体検知モデル(YOLO等)を交通標識データセットで学習

YOLO : 以下を同時に予測

- 物体の位置
- 物体のクラス
- 信頼度: 物体が存在する尤度

信頼度は閾値がある

→下回ると存在しないと扱う

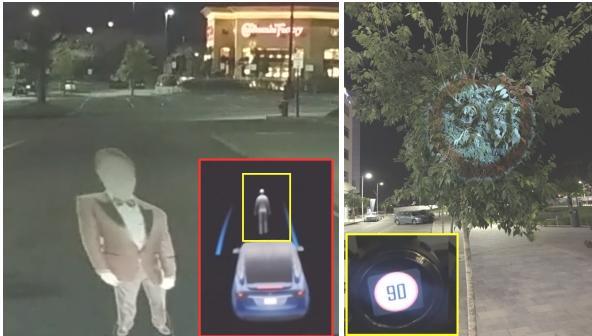


# 交通標識モデルに対する攻撃とその現状

交通標識等の機械学習モデルに対する攻撃: Adversarial Example 攻撃

夜における攻撃やパッチによる攻撃の既存研究

- Phantom Attack → プロジェクターで投影, 検知誤りを狙う
- SLAP → プロジェクターで投影, 検知回避を狙う
- AEパッチの研究(例が多い)→ 昼の状況下にパッチ貼り検知回避等



# 交通標識モデルに対する攻撃とその現状

---

## 現状の攻撃の問題点

夜での攻撃：プロジェクトが必要→ステルス性▲・予算・技術的ハードル

Adversarial Example パッチ：昼間での攻撃に特化，ステルス性▲

本論文：

- 昼間/ヘッドライトなしの夜間のステルス性(cf.夜，AEパッチ)

- 安価・簡単(cf.夜)

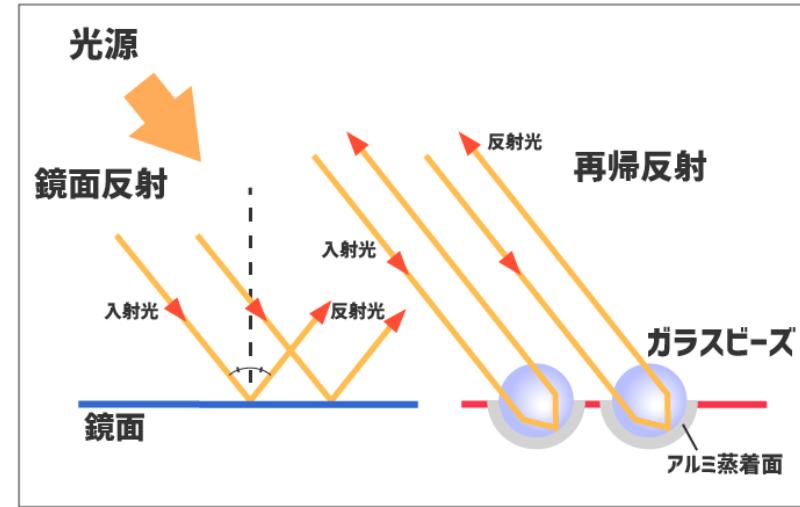
→暗闇における脅威性の高い攻撃

# 本攻撃の概要

ARPA攻撃 再帰反射パッチを使った夜に有効な敵対サンプル攻撃

- 再帰反射パッチがヘッドライト光を反射することで撮動を実現  
→YOLO v3-tinyのSTOPサインクラスの検知回避

再帰反射パッチ：光源の方向に反射する特殊素材



# 攻撃の手順

---

## Step1. パッチを貼る位置の決定

- 画像処理による最適化
  - 反射色は白と仮定
  - 信頼度を落とすような貼り付け位置を探索

## Step2. 実際に貼り付け

## Step3. 自動車のヘッドライト光が反射 → 摂動となり検知回避

- 脅威モデル
  - 交差点における標識無視
  - 速度制限無視による脱輪等の事故

# 最適化の概要

---

最適化の目的 : YOLOの検知回避

→信頼度が一定以下で物体が存在しないと扱う

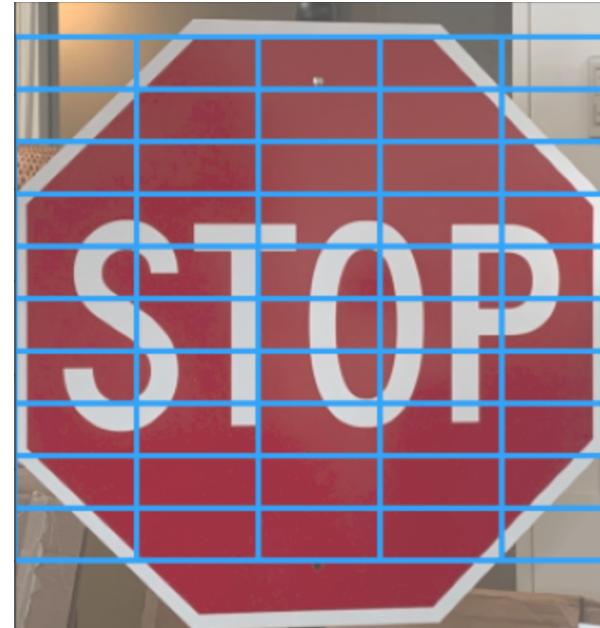
→信頼度が下がるようにパッチを貼り付ければOK

Step1. STOPサインの領域を特定

Step2. グリッドで区切る

Step3. 信頼度が下がるグリッドの探索

- 白で塗り信頼度評価
- ビームサーチを用いる
- 最大貼り付け枚数は5枚



# 評価内容

ディジタル実験、物理実験でそれぞれ画像20枚に対する攻撃成功率を評価

## 評価項目

1. 攻撃の実現性
2. 角度や距離に対するロバスト性

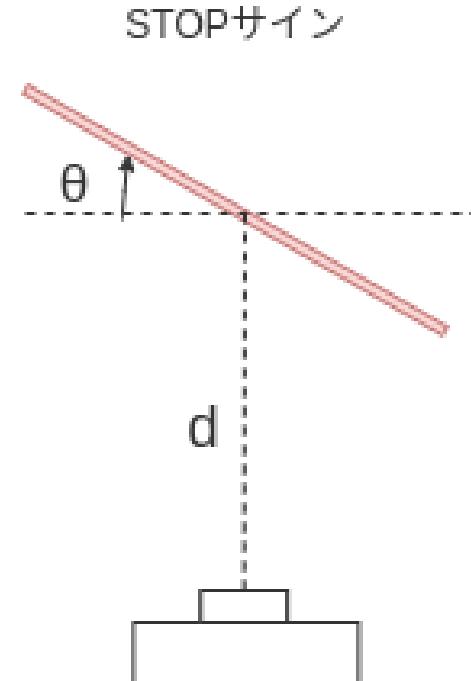
## 評価条件

### 攻撃の実現性

:5m 正面での攻撃成功率

### ロバスト性

3m,7m 正面、 $5m \pm 30\text{度}$ での攻撃成功率



# デジタル実験

## 評価手順

- 5m 正面での画像を1枚用意，貼り付け位置を最適化
  - 反射色を白と仮定して画像処理を利用
- 最適化した貼る位置を他の写真にも適用→攻撃成功率を測定



シミュレーション上でパッチを貼った場合の状況

# デジタル実験の結果

## 評価結果

攻撃の実現性：80%の攻撃成功率

口バスト性：一部を除き80～100%の攻撃成功率

→攻撃の成立・一定の口バスト性が確認

距離	攻撃成功率	角度	攻撃成功率
3 m	25%	30°	80%
5 m	80%	0°	80%
7 m	100%	-30°	10%

# 物理実験

---

## 評価手順

- 5m 正面での画像を1枚用意，貼り付け位置を最適化
- 実際に交通標識に貼る
- ヘッドライト光が当たる状況で撮影→撮影画像に対しYOLOで評価



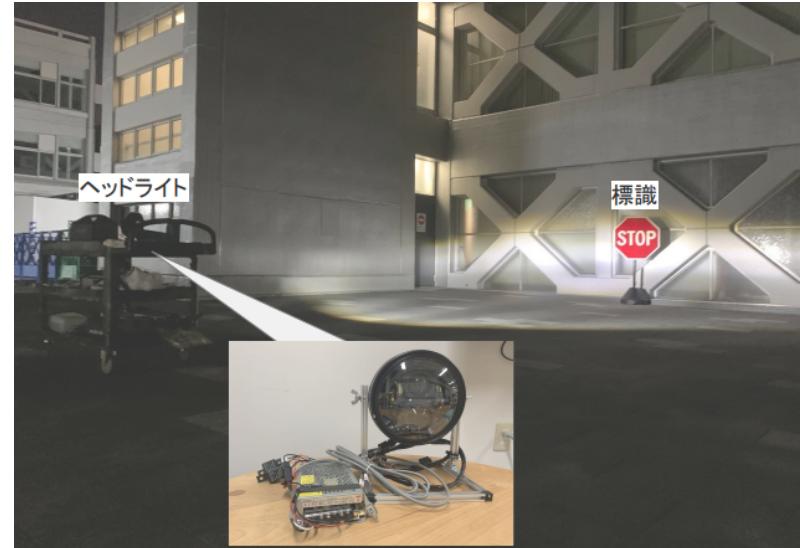
# 物理実験

## 実験設定

ヘッドライト：実際のヘッドライトを利用

カメラ：ミラーレスカメラを利用

カメラとヘッドライトの位置:同じ位置



# 物理実験の結果

攻撃実現性 : 5m 正面で 90% の攻撃成功率

口バスト性 : 35~90% の攻撃成功率を達成

→ 1. 攻撃は実際の環境で成功する 2. 攻撃には一定の口バスト性がある

距離	攻撃成功率	角度	攻撃成功率
3 m	50%	30°	65%
5 m	90%	0°	90%
7 m	50%	-30°	35%

# 評価結果まとめ・考察

---

## 攻撃実現性

- ディジタル・物理実験で80%以上の攻撃成功率  
距離・角度に対する口バスト性
- 一部を除き35~90%の攻撃成功率  
→攻撃実現性・一定の口バスト性が確認できた
- 攻撃成功率のゆらぎ:学習データセット等が原因

# 今後の方針

---

- 最適化方法の工夫
  - Optunaなどでの連続最適化の活用
  - 最適化時に環境変化についても考慮
- 攻撃ベクトルの増加
  - クラス誤り, 誤検知
- 交通標識認識専用モデルに対する評価(今回は一般の物体検出モデル)
- 実車を用いた実験等で現実性の検証
- 既存の攻撃手法に対する評価, 防御手法の提案

# まとめ

---

- 再帰反射パッチを使った夜に有効な敵対サンプル攻撃
- YOLO v3-tinyのSTOPサインクラスの検知回避を目的とした攻撃
- ディジタル実験・物理実験で攻撃の成立・一定のロバスト性を確認  
→交通標識は反射パッチ等を使った攻撃に対して脆弱
- 今後の方針
  - 最適化の工夫
  - 攻撃ベクトルの増加
  - 防御手法の提案等

