7Q-05

実環境評価型進化計算を用いた単眼深度推定器への敵対的攻撃についての基礎検討

日下部　尊　河野　竜士　水俣　友希　大毛　廉也　小野　智司 †
鹿児島大学 †

概要

深層ニューラルネットワーク（Deep Neural Network: DNN）の進歩により性能が飛躍的に向上した単眼深度推定器において，誤推定を誘発させるような微小な摂動が存在することが明らかにされており，単眼深度推定器の頑健性の向上を目的とした敵対的攻撃技術の開発が望まれている．本研究は，最適化における解候補の評価を実環境で行うことで，外乱を考慮した敵対的攻撃を行う方式を提案する．

1　はじめに

単眼深度推定は, 単眼カメラで撮影されたシーンの 3 次元情報，すなわち深度を推定する技術である．近年，DNN の発展により深度推定の精度は飛躍的に向上し，工場や倉庫における物資の自動搬送装置，自動車の自動運転などへの活用が期待されている．一方で，入力画像に微小な摂動を加えることで画像分類器モデルの誤認識を誘発させる敵対的攻撃の危険性が明らかにされており，単眼深度推定用の DNN にも同様の危険性が懸念される．単眼深度推定を自律移動ロボット等の自動運転に用いる場合，誤推定が事故に繋がる可能性がある．このため単眼深度推定用の DNN の脆弱性を敵対的攻撃により検証する研究が行われている．

近年の人工知能技術はソースコードのオープン化により，飛躍的な発展を遂げている．しかし，商用システムでは内部構造及びパラメータへのアクセスが禁止されていることも多い．このため，内部情報を利用しないブラックボックス攻撃による DNN の脆弱性を検証する技術の重要性が高まっている．

本研究では，単眼深度推定用 DNN を対象とした物理的なブラックボックス敵対的攻撃方式を提案する．本方式は，最適化のなかで実環境における物理攻撃を行うことで解候補の評価を行う．実験により，提案手法が実環境における多様な外乱を考慮して物理攻撃を行えることを示す．

A Preliminary Study on Adversarial Attack to Monocular Depth Estimator Using Evolutionary Computation with Real Environment Evaluation
† Takeru Kusakabe, Ryuji Kawano, Tomoki Minamata, Renya Daimo, Satoshi Ono, Kagoshima University

2　関連研究

深度推定器など画像を入力とする DNN に対する物理攻撃は，パッチベース，カモフラージュベース，投光ベースに大別される [1]．Xu らは，生成した敵対的事例を T シャツの表面に張り付け，物体検出器を回避する手法を提案した [2]．一方，投影機を用いたパターン光の投射による敵対的攻撃の研究は，顔認証モデルを対象とした研究 [3] が行われている程度である．著者らの先行研究 [4] を除くと，単眼深度推定用の DNN を投光ベースで物理的に攻撃する研究は著者らが確認した限りでは行われていない．

3　提案手法

本研究では，単眼深度推定器の脆弱性となる敵対的事例を，実環境評価を用いた最適化により発見する手法を提案する．提案手法はブラックボックス条件下で攻撃を行うため，商用のロボットに搭載された装置やシステムの脆弱性を検証することが可能である．実環境評価を行うことで，先行研究 [4] では考慮が困難であった物体の反射特性や環境光の影響などの外乱を考慮することが可能となる．

提案手法は，進化型多目的最適化（Evolu-tionary-Multi-objective　Optimization: EMO）を用いて深度推定誤差と摂動量の 2 つの目的関数を同時に最適化する．EMO は，解候補の生成と評価を繰り返しながら解候補の更新を行うため，最適化におけるすべての解に対して実環境評価を行うと膨大な計算時間を要する．このため，提案手法は図 1 に示す，multi-fidelity 最適化, すなわち物理的な事象を表現する忠実度が異なる複数の評価モデルを順次用いて計算時間の短縮を行う．忠実度が高く，計算コスト

Algorithm 1 Multi-fidelity 最適化による AE 生成.

解候補を初期化，世代数 $gen \leftarrow 0$
while $gen \leq N_{gen}^{(\mathrm{LF})}$ do
    LF モデルを用いた解候補の評価
    解候補の更新
    $gen \leftarrow gen + 1$
end while
while $gen \leq N_{gen}^{(\mathrm{LF})} + N_{gen}^{(\mathrm{HF})}$ do
    HF モデルを用いた解候補の評価
    解候補の更新
    $gen \leftarrow gen + 1$
end while

図 1: 提案手法のアルゴリズム

の高い High-fidelity（HF）モデルはプロジェクタによる投影で対象物体の表面に摂動を加え，計算コストの低い Low-fidelity（LF）モデルでは，実際の投光を行わずに画像処理により投光結果を疑似撮影して評価を行う．

投影する摂動パターン χ を解候補とする．解候補である投影画像 χ を実環境にてプロジェクタで投影した画像 $F_H(\chi)$ を LF モデルで近似した式は $F_L(\chi)$ は以下のように示す．

$$\hat{F_L}(\chi) = I_R + (a^{(adj)} H\chi + b^{(adj)})$$

ここで，I_R は対象物体に摂動を加えずに撮影した画像である．また，解候補 χ を，ホモグラフィ行列 H を用いて，対象物体の摂動を乗せる領域部分のテクスチャに射影した後，最小二乗法 f_{LSM} によって近似計算を行った調整パラメータ $a^{(adj)}, b^{(adj)}$ を用いて色味の調整をする．

第 1 の目的関数 f_1 は，深度マップの全画素（W × H画素）における，対象物体が存在しない室内の画像の深度値 $d_{w,h}^{(target)}$ をターゲットとし解の深度値 $d_{w,h}^{(est)}$ との差の絶対値の総和を最小化する．

$$\text{minimize } f_1(\chi) = \sum_{(w,h)} \left| d_{w,h}^{(est)}(\chi) - d_{w,h}^{(target)} \right| \quad (1)$$

第 2 の目的関数 f_2 は，生成された摂動 $\rho(\chi)$ の L2 ノルムを最小化する．

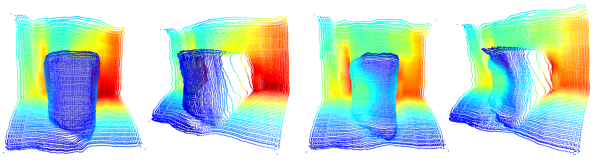
$$\text{minimize } f_2(\chi) = ||\rho(\chi)||_2 \quad (2)$$

4 評価実験

提案方式の有効性を検証するため，屋内シーンを実際の 1/12 のスケールで再現した模型を利用して実験を行った．本実験では，屋内シーンのデータセット NYU Depth v2 を訓練した Laina ら [5] の単眼深度推定器を攻撃対象とした．

(a) 屋内シーン（原画像）　(b) 提案手法で生成した AE

図 2: 生成された AE の例



(a) 屋内シーン（原画像）　(b) 生成された AE

図 3: 原画像および AE の深度推定結果

提案手法により設計された敵対的事例を図 2(b) に示し，単眼深度推定器による推定された深度マップを図 3 に，推定された 3 次元点群を示す．原画像の深度マップと比較し，AE の深度を推定した結果は対象物体の左中央部分が本来の位置より奥に移動したような誤推定が生じた．

5 結論

パターン光の投影により単眼深度推定器の誤認識が生じる脆弱性を検証するために，実環境で解候補を評価する最適化による手法を提案した．実験により，屋内シーンにおいて誤推定を生じさせる敵対的事例を生成できることを確認した．今後，他のシーンや他の深度推定モデルを対象とした検証を行う．

参考文献

[1] D. Wang, et al., “A survey on physical adversarial attack in computer vision,” arXiv:2209.14262, 2022.
[2] K. Xu, et al., “Evading real-time person detectors by adversarial t-shirt,” CoRR, vol. abs/1910.11099, 2019.
[3] D.-L. Nguyen, et al., “Adversarial light projection attacks on face recognition systems: A feasibility study,” CVPR workshops, pp. 814–815, 2020.
[4] R. DAIMO and S. ONO, “Projection-based physical adversarial attack for monocular depth estimation,” IEICE Trans. Information and Systems, vol. E106.D, no. 1, pp. 31–35, 2023.
[5] I. Laina, et al, “Deeper depth prediction with fully convolutional residual networks,” 3DV, pp. 239–248, IEEE, 2016.