H-014

実環境評価型最適化を用いた単眼深度推定器に対する投光型敵対的攻撃
Optical Adversarial Attack on Monocular Depth Estimators Using Physics-in-the-loop Optimization

日下部 尊[1)]　向田 眞志保[1)]　小野 智司[1)]
Takeru Kusakabe　Mashiho Mukaida　Satoshi Ono

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を基に 3 次元情報を推定する技術である [1]．近年，深層ニューラルネットワーク（Deep Neural Network: DNN）の発展により深度推定の精度は飛躍的に向上した．一方で，DNN は特殊な摂動を加えられた敵対的事例（Adversarial Example: AE）に対して，誤認識を引き起こす脆弱性が明らかにされている [2]．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，誤推定が事故の要因となる可能性がある．このため，単眼深度推定器を含む DNN ベースのコンピュータビジョンシステムの頑健性の強化が急務となっている．一般に，脆弱性を発見する敵対的攻撃は，参照可能な情報によってホワイトボックス攻撃とブラックボックス攻撃に大別される．ホワイトボックス攻撃は，DNN モデルのパラメータ，勾配情報などを含む内部情報を参照する手法であり，ブラックボックス攻撃は内部情報を用いない手法である．AI を用いたサービスや商用システムでは，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が多いため，内部情報を利用しないブラックボックス攻撃は，商用サービス等の脆弱性を外部から検証する際に有用である．

画像処理を行う DNN を対象とした敵対的攻撃の研究は当初，入力となる画像に対して摂動やパッチ等を加えるデジタル攻撃が主であった [3]．一方で近年，実世界での物体やシーンに何らかの変化を加える物理攻撃の研究も行われている [4]．物理攻撃は対象となる装置やシステムに侵入することなく攻撃が可能であるため，より現実的な攻撃シナリオとなる．また，物理的な AE を発見することは悪意を持った攻撃に対する脆弱性を検証することに加えて，DNN の認識に影響を与える最悪の外乱を発見する観点でも重要である．

本研究では，対象シーンにプロジェクタを用いて摂動光を投影する投光型敵対的攻撃手法を提案する．本手法は，摂動の設計において，解候補の評価を実環境下で実際に投光および撮影を行う実環境評価型（Physics-In-The-Loop: PITL）最適化 [5] を利用する．これにより，実環境に含まれる複雑な条件や外乱を考慮した物理攻撃が可能となる．実験を行い，提案手法により設計された AE が単眼深度推定器の誤推定を引き起こすことを確認した．

2　関連研究

物理攻撃手法は，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 [3]．また，パッチベース，カモフラージュベースは侵略的攻撃，投光ベースは非侵略的攻撃に分類される．非侵略的攻撃は，対象物体に対して物理的な接触を必要としない点，実世界の自然環境から発生する影響に似ている点から脅威性が高い [6]．

Daimo らはブラックボックス条件下で単眼深度推定器に対する AE を生成する手法を提案した [7]．評価実験により，CG シミュレーションを用いて敵対的投光パターンを生成し，実環境で同様のパターンを投影したが CG シミュレーション同等の結果を得られなかった．これは，物体の反射特性や環境光，カメラノイズなどの複雑な実環境の再現が困難であったため，攻撃性能が低下したと考えられる．

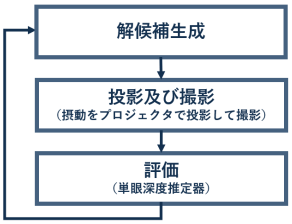

図 1: 提案手法の処理手順

3　提案手法

本研究では，ブラックボックス条件かつ実環境下で，単眼深度推定器の誤推定を引き起こす投光型敵対的攻撃方式を提案する．摂動の設計を行う最適化において，実環境を用いて解候補の評価を行う実環境評価型進化計算を利用する．これにより，環境光の外乱やカメラノイズ等の複雑な実環境条件を考慮した物理攻撃が可能となる．提案手法では図1の手続きによって，最適な摂動光を求める．まず，生成した摂動光（解候補生成）をプロジェクタによって対象物体の表面へ投影し，撮影を行う．その撮影した画像を単眼深度推定器へ入力し，深度マップを推定する．その後，推定された深度マップから目的関数の計算を行う．この処理を繰り返し，目的関数を最小化することで実環境における最適な摂動光を求めることができる．

単眼深度推定器は画像認識と同様に，画像内のエッジによる交点や物体間の境界といった画像の重要な特徴を根拠に，深度推定を行っている [8]．また，画像内の重要な特徴を持つ領域に摂動が加わることによって，敵対的攻撃の影響が大きくなることが知られている [9]．そのため，画像内の重要な特徴を持つ物体の表面に摂動を加えることで，物体のエッジやテクスチャの視覚的手がかりを失い，誤推定を引き起こす可能性が高まる．したがって，提案手法では摂動の投影領域を対象物体の表面とし，投影された領域の深度を誤認識させる．

提案手法では，深度推定に関する目的関数 f_1 と摂動量に関する目的関数 f_2 を同時に最適化する進化型多目的最適化（Evolutionary Multiobjective Optimization: EMO）を用いる．EMO では，最適解の集合であるパレートフロントを探索し，各目的関数のバランスを取った解を求

1）　鹿児島大学，Kagoshima University

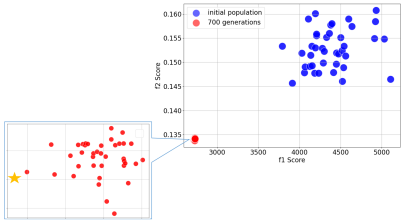

図 2: $f_1(\chi)$ スコア及び $f_2(\chi)$ の推移

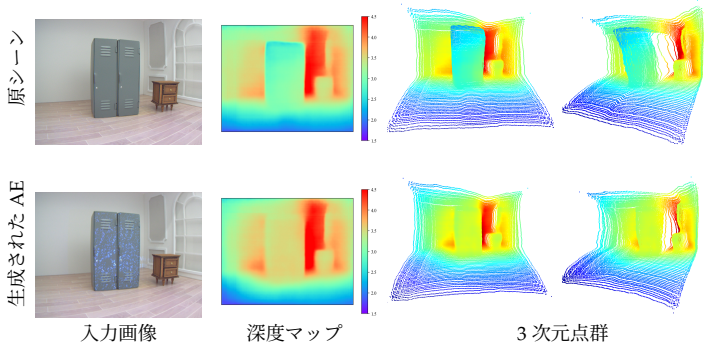

図 3: 原画像および生成された AE の深度推定結果

めることができる．f_1 は以下のように定義する．

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

ここで，R は対象物体を含む周辺領域（$w \times h$ 画素）の画素位置の集合である．$d^{est}_{(w,h)}(\chi)$ は摂動光 χ を投影したときの画素位置 (w,h) における深度マップの深度値である．$d^{(target)}_{w,h}$ は画素位置 (w,h) におけるターゲット深度マップの深度値である．ターゲット深度マップの詳細は 4 章の評価実験で説明する．f_2 は以下のように定義する．

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

ここで，$||\cdot||_2$ はユークリッドノルムである．$\rho(\chi)$ は摂動光 χ を投影したときの摂動量である．

4 評価実験

提案手法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 [10] の単眼深度推定器を攻撃対象とした．ターゲット深度マップは，対象シーンにおいて対象物体を取り除いた際の撮影画像とし，対象物体がシーンから消えるような誤推定を目標とした AE の生成を試みた．本実験では，最適化アルゴリズムとして MOEA/D [11] を用いた．スカラー化関数として Chebyshev 法を選択し，近傍サイズ $N_n = 10$，$\delta = 0.8$，$n_r = 1$，個体数 N_p を 40，世代数を 700 として最適化を行った．また，ブロック単位の摂動パターン数 N_{AP} を 15，サイズ N_{pat} を 8×8 に設定した．

本実験で使用した機材を以下に示す．カメラは Basler acA1300-30gc，レンズは DV3.4X3.8SA-1，プロジェクタは EPSON EB-E01，フィルタは可変 ND フィルタを使用した．また，プロジェクタの映像における設定を以下に示す．カラーモードの種類は，sRGB モード（明るさ 42，コントラスト 65，色の濃さ 95，色合い 43，シャープネス 20，ホワイトバランス 6,000K，オートアイリス標準）に設定した．また，ND フィルタ（ND2）を使用し，プロジェクタの光量を 1/2 に抑えた．

図2に初期集団における非劣解集合と 700 世代目の非劣解集合の分布を示す．最適化が進むにつれて，ターゲット深度マップとの誤差と摂動量の双方が改善され，両者のトレードオフ関係が明確になるような非劣解集合が得られたことがわかる．

図2において星印で表される解候補を投影したシーンと，摂動が加えられていないシーンとにおける深度マップおよび点群を図3に示す．提案手法により生成された AE では，元画像の深度マップと比較して，ロッカーがより後方に存在するように誤推定が生じていることがわかる．また，ロッカー上部は壁との境が曖昧になっており，ロッカー上部と壁が同化しているような結果になった．

5 結論

本研究では，単眼深度推定器を対象とした実環境評価型最適化を用いた投光型敵対的攻撃手法を提案した．実験により，屋内シーンにおいて対象物体を本来の位置より後ろへ誤推定させる投光パターンを生成できることを確認した．今後，最適化アルゴリズムの改善や，他の深度推定モデルにおける有効性の検証を行う．

参考文献

[1] A. Bhoi, “Monocular depth estimation: A survey,” *arXiv preprint arXiv:1901.09402*, 2019.
[2] I. J. Goodfellow *et al.*, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
[3] D. Wang *et al.*, “A survey on physical adversarial attack in computer vision,” *arXiv preprint arXiv:2209.14262*, 2022.
[4] A. Guesmi *et al.*, “Physical adversarial attacks for camera-based smart systems: Current trends, categorization, applications, research challenges, and future outlook,” *IEEE Access*, vol. 11, pp. 109617–109668, 2023.
[5] T. Minamata *et al.*, “A coded aperture as a key for information hiding designed by physics-in-the-loop optimization,” *IEICE Trans. Inf. Sys.*, vol. 107, no. 1, pp. 29–38, 2024.
[6] T. Sato *et al.*, “Invisible reflections: Leveraging infrared laser reflections to target traffic sign perception,” *arXiv preprint arXiv:2401.03582*, 2024.
[7] R. Daimo and S. Ono, “Projection-based physical adversarial attack for monocular depth estimation,” *IEICE Trans. Inf. Sys.*, vol. E106.D, no. 1, pp. 31–35, 2023.
[8] T. v. Dijk and G. d. Croon, “How do neural networks see depth in single images?,” in *Proc. of the IEEE/CVF Int. Conf. Computer Vision (ICCV)*, October 2019.
[9] J. Hu and T. Okatani, “Analysis of deep networks for monocular depth estimation through adversarial attacks with proposal of a defense method,” *arXiv preprint arXiv:1911.08790*, 2019.
[10] I. Laina *et al.*, “Deeper depth prediction with fully convolutional residual networks,” in *Proc. 2016 Fourth Int. Conf. 3D Vis. (3DV)*, pp. 239–248, IEEE, 2016.
[11] Q. Zhang and H. Li, “Moea/d: A multiobjective evolutionary algorithm based on decomposition,” *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, 2007.