単眼深度推定器に対する敵対的事例の進化的生成に関する基礎検討

大毛 廉也[1,a)]　鈴木 崇大[1,b)]　小野 智司[1,c)]

概要：近年の深層ニューラルネットワーク（DNN）の急速な進展により，1 枚の画像からシーンの深度情報を推定する単眼深度推定にも応用され，実用に向けての期待が高まっている．一方，DNN には特有の脆弱性が存在し，微小な摂動を入力画像に加えることにより誤認識が生じることが指摘されている．単眼深度推定を対象とする DNN においても同様の脆弱性が懸念され，その実用化に向けて解析を行うことは重要である．このため，本研究では，進化計算を用いて単眼深度推定器の誤りを引き起こす敵対的事例を生成する方式を提案する．提案手法は，物体表面のテクスチャに摂動を加えることで，推定された深度に誤りが生じる敵対的事例の生成を試みる．提案手法は特に，ブラックボックス的に敵対的攻撃を行う点に特徴があり，商用システムやサービスの解析も可能である．実験により，DNN を利用した単眼深度推定器は，敵対的事例により精度低下を起こしうる事を示す．

A Preliminary Study on Evolutionary Design of Adversarial Examples for One-Shot Depth Estimation

Abstract: The performance of monocular depth estimation has been improved by combining it with Deep Neural Networks (DNNs), and expectations for its practical use are increasing. However, recent studies have revealed DNNs are vulnerable to adversarial attacks, which are carefully designed perturbations to make DNN misclassify it. To apply DNN-based one-shot depth estimation to real-world applications, it is essential to investigate such vulnerabilities. In this study, we propose an evolutionary computation-based method to generate adversarial examples that cause incorrect measurements in one-shot depth estimator. The proposed method assumes a target attack and attempts to generate an adversarial example so that a target object would not be detected.
In addition, the proposed method also assumes black-box condition so that it can analyze commercial systems and services. Experimental results demonstrated that the proposed method successfully generated AEs that misleaded a DNN-based one-shot depth estimation method.

1. はじめに

1 枚の画像からシーンの深度を推定する単眼深度推定は，ロボット工学，シーン理解，3D 再構成などの用途で有用であり，広く研究されている [1]．近年，深層ニューラルネットワーク（Deep Neural Network: DNN）との組み合わせにより単眼深度推定の性能が向上する事が明らかにされ [2]，DNN に基づいた単眼深度推定器に関する多くの研究が行われており，実用に向けての期待が高まっている [3–6]．

一方，近年の研究により，DNN には特有の脆弱性があり，攻撃者によって意図的にモデルが誤認識するよう設計された敵対的事例（Adversarial Example: AE）の影響を受けやすく，正しく物体認識を行なえないことが明らかにされている [7]．このため，単眼深度推定器を対象とした NN においても同様の脆弱性が懸念され，これを明らかにすることは実用化に向けて重要な課題である．

本研究では，進化計算を用いて分類器に対する AE を生成する方式 [8] を応用し，単眼深度推定器の誤りを引き起こす AE の生成方式を提案する．提案手法は進化型多目的最適化を用いることで，目視で判断した奥行と深度推定器

[1] 鹿児島大学
Kagoshima University, 1-21-40, Korimoto, Kagoshima 890-0065, Japan
[a)] sc116091@ibe.kagoshima-u.ac.jp
[b)] sc115029@ibe.kagoshima-u.ac.jp
[c)] ono@ibe.kagoshima-u.ac.jp

が出力した奥行に相違が生じるような AE 生成を可能とする．特に，DNN 内部の情報および信頼度スコアを必要とせず，ブラックボックス的に敵対的事例を生成できる点に特徴があり，商用システムの解析にも有効である．評価実験により，DNN を利用した単眼深度推定器は，ブラックボックス的に生成された敵対的事例により，深度推定の精度低下を起こしうる事を確認した．

2. 関連研究

2.1 単眼深度推定

単一画像からの深度推定は，長年にわたって研究が行われており，近年は DNN における著しい進歩により，単眼深度推定に CNN を使用する研究が行われている [3–6]．Eigen 等は，疎な深度推定をする層および密な深度推定をする層の二つの深いネットワーク層によって深度を推定する手法を提案しており [2]，この研究が単眼深度推定に初めて CNN を用いた研究であると考えられる．また，Laina らが提案した手法は，完全畳み込み NN を用いることでパラメータの数を抑えつつより高解像度の出力を可能とした [3]．Hu 等によってより物体の境界線が際立った深度マップを得られる単眼深度推定の手法も提案されている [6]．

一方，単眼深度推定の発展に付随してそれらの分析に関する研究も活発に行われ始めている．Dijk らの研究では，車載カメラの映像をデータセットとした単眼深度推定器は，障害物の見かけのサイズを無視して画像内の垂直位置を優先すること，カメラのピッチとロール角の変化を部分的にしか認識しないことが明らかになった [9]．Hu 等の研究では，CNN はエッジの強度ではなく、シーン形状の推論に重要なエッジを参照すること，境界だけでなく個々のオブジェクトの内側の領域を参照する傾向が確認された．また，屋外シーンにおいては，消失点周辺の画像領域が重要であることが明らかになった [10]..

2.2 DNN に対する敵対的攻撃

Goodfellow らにより，DNN を含む機械学習モデルの誤りを引き起こす AE の生成が容易に行えることが明らかにされて以来，DNN に対する敵対攻撃の研究が広く行われている [7]．当初は対象となる DNN の内部情報を参照するホワイトボックス条件下の AE 生成が主流であった．しかし，分類器の内部情報を必要とする方法は市販のソフトウェアやサービスなど，NN 分類器内部の情報を利用できない際には適用が困難である．近年は，対象モデル内部の情報を参照しないブラックボックス条件下で AE を生成する手法の需要が高まっている．Suzuki らは進化型多目的最適化を用いた AE 生成手法を提案している [8]．この手法は，NN の内部情報および，信頼度スコアを必要としないため，ブラックボックス的に AE を生成できる．また，制約条件や目的関数の自由度が高い定式化が可能である．

DNN に対する敵対的攻撃の研究の多くは，物体認識を行う DNN を対象とした研究が多いが，単眼深度推定を行う DNN を対象とした研究も行われはじめている．Yamanaka 等が提案する AE 生成手法は，生成したパッチ画像にて入力画像の局所領域を上書きすることで，その領域内の深度を任意に誤推定させる事を可能とした．しかし，yamanaka 等の手法は損失関数の勾配情報を用い繰り返し演算を行う必要があり，推定器の内部情報が明らかでない商用サービス等には適用が難しい [11]．また，Hu 等はホワイトボックス的な AE 生成と，その AE に対する防御手法を研究することで，単眼深度推定の計算メカニズムを理解するための手がかりとしている [12]．

3. 提案手法

3.1 基本アイデア

本研究では進化型多目的最適化（Evolutionary Multi-criterion. Optimization: EMO ）[13] を用いて単眼深度推定器に対する AE をブラックボックス的に生成する手法を提案する．本方式の基本アイデアを以下に示す．

進化型多目的最適化 (EMO) アルゴリズムの適用: AE を生成する際，推定精度と摂動強度のように複数の目的関数はトレードオフの関係にある．多目的最適化はこれらの目的関数を単一の目的関数に統合せずにパレート解集合を発見しやすく，有効なアプローチであると考える．提案手法は目的関数を統合するためのパラメータを必要としない，微分不可能かつ多峰性の目的関数を最適化できる，といった利点がある．また，制約条件や目的関数の自由度が高い定式化が可能であり，未知である単眼深度推定器の脆弱性を調査するに当たって妥当なアプローチと言える．

テクスチャへの摂動の付加: 単眼深度推定器は，対象画像において，シーン形状の推論に重要なエッジと，個々の物体の内側の領域を認識した上で深度の推定を行うと考えられている [10]．したがって，単眼深度推定において物体表面のテクスチャは重要な役割を果たすと考え，提案手法ではテクスチャに摂動を付加する．

3.2 定式化

3.2.1 設計変数

高解像度の画像に対して摂動を直接的に設計すると，画像解像度の高さが設計変数の次元数の増大を招き，最適化が困難となる．このため，提案手法では，複数の局所的な摂動パターンを設計するとともに，画像のどの部分にどの摂動パターンを加えるかを表すパターンマップを作成することとした．これにより，設計変数の次元を削減する．

図 1 に摂動付加の手順を示す．提案手法では，入力画像 I の対応する画像ブロックごとに摂動を付加するために，N_{AP} 種類の摂動パターンを持つ．摂動パターンは，

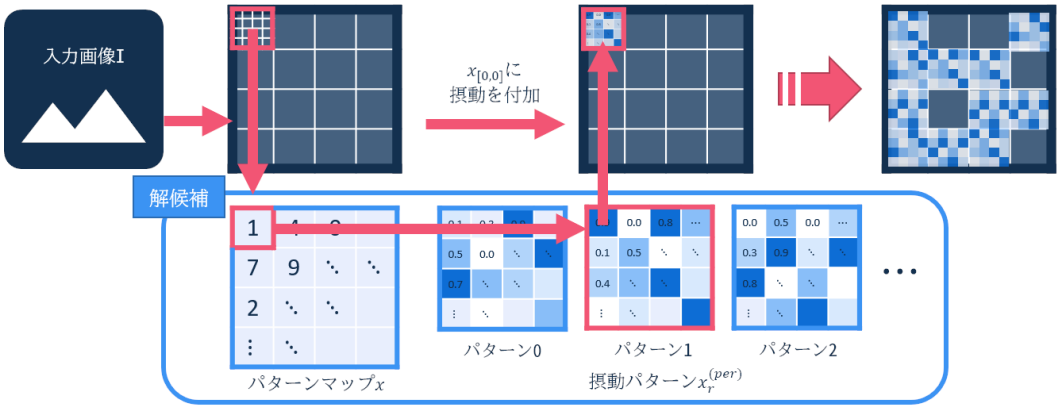


図 1　摂動の加え方

図 2　処理手順

$n \times n$ 画素の正方形領域における各画素値の増減幅である．$x_{p,q,r}^{(pat)}$ は画像ブロック (p,q) に対する変更画素値を表し，r はパターンの番号を表す．また，パターンマップ $x_{u,v}^{(PS)}$ は，入力画像 I を摂動パターンのサイズで分割した際に，各領域 (u,v) 内でどの摂動パターンを適用するかを定める $(x_{u,v}^{(PS)} = 0, 1, ...N_{AP})$．$x_{u,v}^{(PS)} > 0$ の場合，対応する摂動パターンを画像ブロック (u,v) に適用し，$x_{u,v}^{(PS)} = 0$ の場合は，当該画像ブロックの画素値は変化しない．

$$\chi = x \cup x^{(pat)} \tag{1}$$

$$x = \{x_{u,v}^{(PS)}\}_{(u,v)\in I} \tag{2}$$

$$x^{(pat)} = \{x_r^{(pat)}\}_{1\le r\le N_{AP}} \tag{3}$$

$$x_r^{(pat)} = \{x_{p,q,r}^{(pat)}\}_{1\le p\le N_{pat}, 1\le q\le N_{pat}} \tag{4}$$

3.3　処理手順

提案手法では，図 2 に示すように，解候補の生成と評価の繰り返しにより，良い AE を生成する．すなわち，画像に加える摂動の強度，摂動パターンの設定の決定，および，生成された AE の深度推定器への入力とその結果に基づく目的関数の計算を繰り返す．

[Step 1] 画像を入力
　　AE 生成の対象となる画像を入力する．

[Step 2] 解候補群を生成
　　画像に加える摂動パターンを複数生成する．

[Step 3] 解候補群を評価
　　Step2 で生成した摂動パターンを画像に加え AE を生成する．この AE を単眼深度推定器に入力し評価を行う．

[Step 4] 最良解を出力
　　最適化によって得られた最も良い摂動パターンを出力する．

4.　評価実験

提案手法が単眼深度推定器の誤りを引き起こす AE を生成できることを示すために，以下の評価実験を行った．多様な撮影シーンを対象に AE を生成することで，単眼深度推定器が AE の影響を受けるシーンを模索した．また，本実験では，ターゲットとなるシーンに誤認識するように AE の生成を試みた．

4.1　実験設定

解析対象の DNN として，典型的な構造を持ち，また，推定精度が高い Laina らの手法 [3] を選択した．

本モデルは，残差学習を含む完全畳み込みニューラルネットワークであり，ReLU 関数を活性化関数ととして用い，また，逆フィーバー損失を用いて訓練を行う．室内画像のデータセット[*1]より抽出，拡張したデータセット約 95,000 枚の RGB-D 画像のペアを使用して，上記のネットワークの学習を行った学習済みモデルが公開されており，本実験ではこの学習済モデルを用いた．

4.1.1　入力画像

深度推定器への入力は統合型 3DCG 作成ソフト[*2] を用いて作成したレンダリング画像を用いる．事後の考察を容易に行えるように，また，データセットを模倣するように，図 5(a) に示すような，テクスチャを付与する物体 1 個と最大 3 個の光源，床，壁，天井にて構成されるシーンを設計した．レンダリング画像のサイズは深度推定器の入力サイズに合わせ 228 × 304 画素とした．

[*1] NYU Depth v2:Microsoft Kinect の RGB カメラと Depth カメラの両方で記録された，屋内シーンのビデオシーケンスで構成されている

[*2] blender:3DCG アニメーションを作成するための統合環境アプリケーション

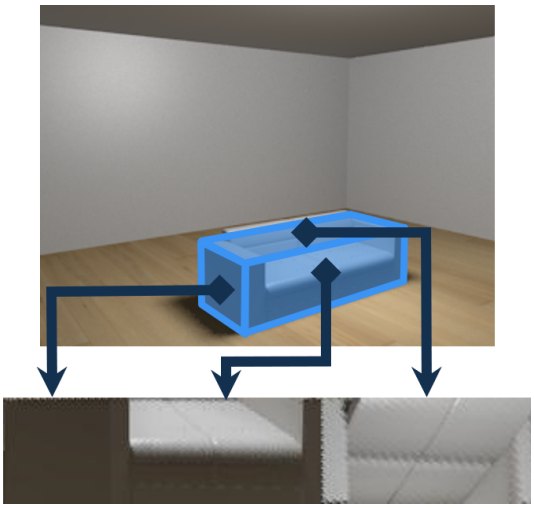


図 3　テクスチャの展開

また，摂動は物体のテクスチャに付与することとしたが，計算量を削減するために，物体を 1 つの直方体とみなし，射影変換することで得られる展開図に付加することとした．図 3 に示す例では，物体の側面の解像度を 64×64 画素，全面および上面を 64×128 画素とした．本実験における設計変数の総数は，摂動パターンを決定する変数 $8 \times 8 \times 10$ 次元と展開図における領域数 $8 \times 8 + 8 \times 16 \times 2$ 次元をあわせて 960 次元となる．

4.1.2　目的関数

本実験では，2 目的を同時に最小化するように目的関数を設定した．目的関数 f_1 は，推定された深度マップの深度値 $d^{est}_{(i,j)}$ と目標とする深度マップの深度値 $d^{target}_{(i,j)}$ との差の絶対値の，深度マップ全体（$W \times H$ 画素）における総和である．

$$\text{minimize } f_1(\boldsymbol{x}) = \sum_{(w,h)\in(W,H)} \left| d^{(est)}_{w,h}(\boldsymbol{x}) - D^{(target)}_{w,h} \right| \tag{5}$$

目的関数 f_2 は摂動量の L2 ノルムとした．

本実験では，入力画像において視認できる物体が，深度推定によって見落とされるように AE の生成を試みた．したがって，detail は物体が何も置かれていないシーンとし，3DCG 作成ソフトにて深度マップを作成し，目標とする深度マップ $D^{(target)}$ とした．ここで，摂動が付加されていない画像を入力したときの深度マップをもとに，$D^{(target)}$ の深度方向のスケーリングを行った．

4.1.3　最適化アルゴリズム

最適化アルゴリズムは，多目的最適アルゴリズムである MOEA/D を用いた．MOEA/D では，多目的最適化問題をスカラー最適化問題の集合に変換するために，チェビシェフ法を選択した．近傍サイズ $N_n = 10$，$\delta = 0.8$，$n_r = 1$ と設定した．最適化は 100 個体 5000 世代で行った．

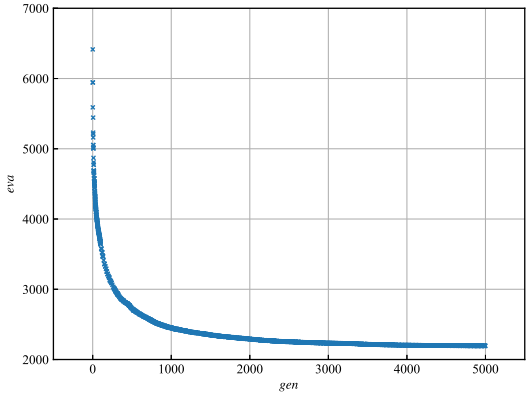
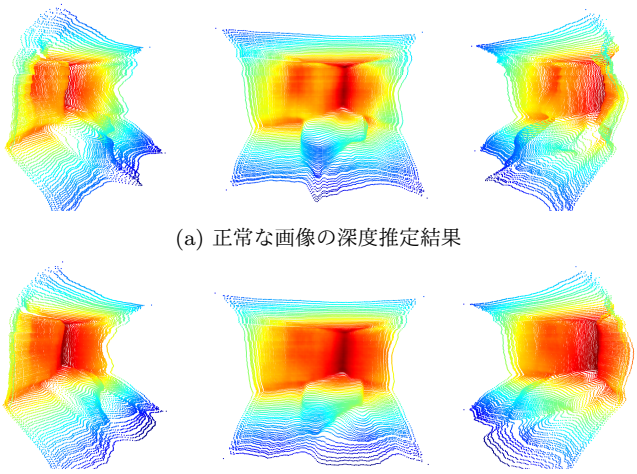


図 4　最良解の f_1（式 (5)）の推移

(a) 入力画像　　(b) 生成された AE

図 5　生成された AE の一例



(a) 正常な画像の深度推定結果

(b) AE の深度推定結果

図 6　深度推定結果の比較

4.2　実験結果

提案手法により AE を生成する際の最良解の f_1 の推移を図 4 に示す．また，本実験で得られた非劣解集合のうち，f_1（5）が最良であった解を図 5(b) に示し，深度推定の結果を 3 次元点群の形式で図 6 に示す．

図 6 より，正常な画像の深度推定結果ではソファーの手前側のひじ掛け部分が顕著に現れているが，摂動が加えられた画像における深度推定結果では，同じ部分がなだらかに推定されており，実際の深度よりも奥側に，かつ，高さが低く推定されていることがわかる．

本稿では，深度推定における AE による誤認識の影響度を定量化するために，テクスチャを付与した物体の体積 (ただし，物体裏から壁面までの範囲を含む) を計測した．図

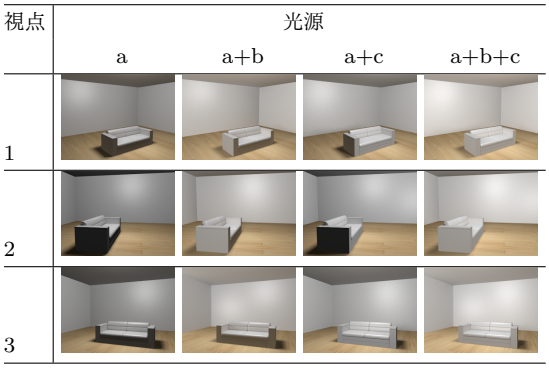


図 7 入力画像

表 1 対象物体の仮想的な体積

視点	画像	光源			
		a	a+b	a+c	a+b+c
1	原画像	1202.8	930.3	2161.1	2095.9
	AE（体積）	214.0	188.6	127.4	117.2
	AE（減少率）	82.2%	79.7%	94.1%	94.4%
2	原画像	2912.6	2308.3	3355.3	2342.9
	AE	1407.8	1126.1	1491.5	954.6
	AE（減少率）	51.7%	51.2%	55.5%	59.3%
3	原画像	2315.6	2287.9	2460.7	2607.5
	AE	720.7	366.0	583.8	193.9
	AE（減少率）	68.9%	84.0%	76.3%	72.5%

7 に示す 3 方向の視点，および，3 種類の照明を変化させた場合の，深度推定結果への影響を表 1 に示す．視点 1 は図 5 の AE を生成した際と同じ視点であり，物体の体積が大幅に減少していることから，AE が深度推定の結果に大きく影響を与えていることがわかる．視点や光源を変えた場合，視点 2 のように AE の影響を受けにくい場合もあるものの，多くの場合で視点 1 と同様に体積が減少することが確認でき，生成した AE が視点変化等に頑健であることがわかる．

5. おわりに

本研究では，単眼深度推定器の誤りを引き起こす AE を進化型多目的最適化に生成する手法を提案した．提案手法は，ブラックボックス条件下で AE を生成できる点，制約条件や，目的関数の自由度が高い定式化が可能な点に特徴がある．実験により，単眼深度推定器は敵対的事例によって誤りを起こし得る事を確認した．今後，より実環境に近いシーンでの AE 生成手法や，ロボットなどの行動決定により影響を与えるような AE 生成手法について検討する．

参考文献

[1] H. Laga, “A survey on deep learning architectures for image-based depth reconstruction,” *arXiv preprint arXiv:1906.06113*, 2019.

[2] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[3] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[4] F. Mal and S. Karaman, “Sparse-to-dense: Depth prediction from sparse depth samples and a single image,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[5] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8001–8008.

[6] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, “Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1043–1051.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[8] T. Suzuki, S. Takeshita, and S. Ono, “Adversarial example generation using evolutionary multi-objective optimization,” in *2019 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2019, pp. 2136–2144.

[9] T. v. Dijk and G. d. Croon, “How do neural networks see depth in single images?” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2183–2191.

[10] J. Hu, Y. Zhang, and T. Okatani, “Visualization of convolutional neural networks for monocular depth estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3869–3878.

[11] 幸. 山中, 隆. 松本, 桂. 高橋, and 俊. 藤井, “単眼深度推定 cnn における敵対的画像生成,” 名古屋大学大学院工学研究科藤井研究室, 名古屋大学大学院工学研究科藤井研究室, 名古屋大学大学院工学研究科藤井研究室, 名古屋大学大学院工学研究科藤井研究室, Tech. Rep. 17, nov 2019.

[12] 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.

[13] C. M. Fonseca, P. J. Fleming, E. Zitzler, K. Deb, and L. Thiele, “Evolutionary multi-criterion optimization,” in *Second International Conference, EMO 2003*. Springer, 2003.