単眼深度推定により顕在化する幾何学的不整合に着目した生成画像検出の基礎検討

A Preliminary Study on AI-Generated Image Detection Focusing on Geometric Inconsistencies Revealed by Monocular Depth Estimation

坂拓実 [1]　　山内悠嗣 [1]
Takumi Ban　　Yuji Yamauchi

中部大学 [1]
Chubu University

1　はじめに

近年，画像生成 AI の急速な発展に伴い，生成画像の高品質化が進み，生成画像であるか否かを判別することが困難になっている．生成画像の悪用，虚偽情報やフェイクニュースの拡散などの問題に繋がり得るため，生成画像検出技術の確立が求められる．生成画像は局所的なテクスチャが高品質である一方で，物体間の前後関係や大域的構造に幾何学的不整合を含む場合がある．しかし，その差異は RGB 画像だけでは視覚的に微細であり，正確な判別は難しい．そこで，単眼深度により画像の 3 次元構造を擬似的に復元し，幾何学的矛盾を深度マップ上の歪みや不連続性として顕在化させることで，生成画像検出の性能向上が期待できる．本稿では，単眼深度推定モデルを用いた生成画像検出の有効性を基礎的に検討する．

2　提案手法

提案手法の流れを図 1 に示す．まず，単眼深度推定モデルにより入力画像 (実画像/生成画像) から深度マップを生成する．次に，得られた深度マップを検出モデルに入力し，実画像/生成画像の 2 値分類を行うことで生成画像を検出する．

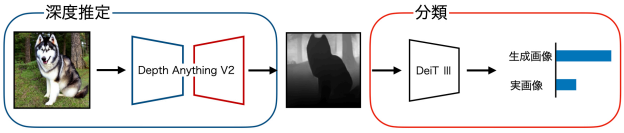

図 1　提案手法の流れ

単眼深度推定モデルには，Depth Anything V2[1] を採用する．Depth Anything V2 は，画像から各画素の距離を推定可能であり，生成画像に対して適用することで，RGB 画像上では差異がわずかである幾何学的不整合が深度マップ上の不自然さとして現れることが期待できる．検出モデルには，Data-Efficient Image Transformers III(DeiT III)[2] を採用する．Self-Attention 機構が大域的な相関関係を捉えることで，背景と被写体の距離感の不整合といった，画面全体に及ぶ幾何学的矛盾の検出に有効であると考えられる．

3　評価実験

提案手法の有効性を検証するため生成画像検出のベンチマークである GenImage[3] を用いて評価した．GenImage は，実画像 1,331,167 枚と生成画像 1,350,000 枚から構成されている．実画像は ImageNet 由来，生成画像は BigGAN や Stable Diffusion(V1.4/V1.5)，Midjourney V5(Midjour) などの生成モデルにより作成されている．

本研究では，「Stable Diffusion V1.4 で学習し，全ての生成モデルのテストデータで評価をする」クロス生成器評価を採用した．評価指標は検出精度とし，深度マップを入力とする提案手法と，RGB 画像を入力とする既存検出器 (GenImage 論文で報告された手法) を比較する．

表 1　各手法の検出精度の比較 [%]

Detectors	Midjour	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Ave
ResNet-50	54.9	**99.9**	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S	55.6	**99.9**	**99.8**	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T	62.1	**99.9**	**99.8**	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
GramNet	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
Ours	**71.8**	99.4	99.4	**97.5**	**98.6**	**99.5**	**98.8**	**83.8**	**93.6**

表 1 に各手法の 8 つのデータセットに対する検出精度を示す．提案手法は，評価に用いた 8 つのデータセットの平均で 93.6%を達成し，既存の RGB ベースの手法 (Swin-T 等) を大きく上回った．特に，ADM や GLIDE に対し，既存手法が 40〜60%台に留まる一方で，提案手法は 97%以上の精度を達成した．これは，単眼深度推定により，RGB 画像上では顕在化しにくい幾何学的不整合が深度マップ上の歪みや境界の不連続として表れ，検出に寄与した可能性がある．一方で，Midjour に対する精度は 71.8%に留まった．高品質な拡散モデルでは幾何学的一貫性も改善され，深度マップ上の破綻が相対的に目立ちにくい可能性がある．このため，幾何学的情報に加えて照明整合性やエッジ整合性など，複合的特徴の導入が今後の課題である．

4　おわりに

本稿では，単眼深度推定モデルで生成した深度マップを用いる生成画像検出について検討し，GenImage における有効性を確認した．今後は，あまり精度が向上しなかった生成モデルに対する要因分析に加え，深度推定と生成画像検出を同時に学習するマルチタスク学習を導入し，特徴表現の汎化を促すことで検出性能の向上を図る．

参考文献

[1] L. Yang *et al.*: “Depth Anything V2”, NeurIPS, 2024.
[2] H. Touvron *et al.*: “DeiT III: Revenge of the ViT”, ECCV, 2022.
[3] M. Zhu *et al.*: “GenImage: A million-scale benchmark for detecting AI-generated image”, NeurIPS, 2024.