大域的な空間構造の破綻に着目した単眼深度推定によるロバストな生成画像検出

坂 拓実[1,a)]　山内 悠嗣[1,b)]

概要

従来の生成画像検出は微細な生成痕跡に依存するため，JPEG 圧縮等の画像劣化により検出精度が著しく低下する．本稿では，単眼深度推定により大域的な幾何学的矛盾を顕在化させる検出手法を提案する．GenImage を用いた評価実験により，画像劣化に対する高いロバスト性と汎化性能を実証し，RISE 分析により幾何学的アプローチの有効な適用条件を明らかにした．

1. はじめに

近年，敵対的生成ネットワーク（Generative Adversarial Networks: GAN）[7] や拡散モデル [10] を用いた画像生成技術は飛躍的な進歩を遂げ，実画像と遜色ない高品質な画像の生成が可能となった．一方で，生成画像の悪用によるフェイクニュースが SNS 等で加工を伴いながら拡散されることが社会問題化しており，高精度かつロバストな生成画像検出技術の確立が急務である．

既存の生成画像検出手法の多くは，RGB 画素値の空間的パターン [24] やアップサンプリング処理に起因する周波数成分の周期的ノイズ [6] など，微細な生成痕跡の検出に依存している．しかし，これらの微細な痕跡は，画像が SNS 等で共有・拡散される過程で生じる劣化処理により容易に欠落してしまう．その結果，画像圧縮や平滑化などの劣化を受けた画像に対して，検出精度が低下するという課題に直面している．

一方で，人間が生成画像に抱く違和感の多くは，ピクセルレベルのノイズではなく，遠近法の狂いや物理的にあり得ない空間構造といった大域的な幾何学的不整合に起因する．現在の生成モデルは，局所的なテクスチャにおいて高い描画能力を保有する一方で，主に 2 次元画像の見え方から特徴を学習しているため，幾何学的一貫性を完全に保つことは依然として困難である．しかし，これらの不整合は 2 次元の RGB 画像上ではテクスチャ情報に埋もれてしまうため，直接的な検出は難しい．上記の問題を解決するためには，画像に潜む幾何学的矛盾を，深度の急激な歪みや不連続性として顕在化させるアプローチが有効であると考えられる．

そこで，本稿では単眼深度推定モデル Depth Anything V2[27] を用いて深度マップを生成し，大域的な相関関係の捕捉に長けた Vision Transformer（DeiT III[23]）により深度マップの幾何学的不整合を検出することで，生成画像を検出する手法を提案する．

本稿の貢献は以下の 3 点である．

(1) 深度情報における大域的幾何特徴に着目した新しい生成画像検出フレームワークの提案．
(2) 画像劣化に対するロバスト性の実証，および未知の生成モデルに対する幾何学的アプローチの有効な適用条件の提示．
(3) 顕著性マップを用いた定性・定量分析による，分類器が深度マップ上の幾何学的破綻を判定の根拠としていることの可視化．

2. 関連研究

2.1 深層学習を用いた生成画像検出手法

既存の生成画像検出手法は主に二値分類問題として扱われており，これらの手法は空間領域と周波数領域のアプローチに大別される．空間領域の手法である CNNSpot[24] は，微細なテクスチャパターンを抽出することで生成画像か実画像かの判定を行う．しかし，これらのアプローチは画素レベルの微細な痕跡に依存するため，拡散モデルのような全く異なる生成過程や，画像圧縮・平滑化などの劣化に対して脆弱であるという課題を抱えている．また，GramNet[13] は Gram 行列を用いて大域的なテクスチャ統計量を捉えることで改善を試みたが，本質的には RGB 空間の統計的特徴に依存しており，同様の限界を有する．

一方，周波数領域の手法である F3Net[18] や Spec[28] は，フーリエ変換を介して画像をスペクトル情報に変換し，生成モデルのアップサンプリング処理に起因する周期的アーティファクトを捉える．これらの手法は空間領域では視認困難な周波数的な痕跡を検出することで，一定の識別能力

1 中部大学
a) tr26010-4465@sti.chubu.ac.jp
b) yuu@fsc.chubu.ac.jp

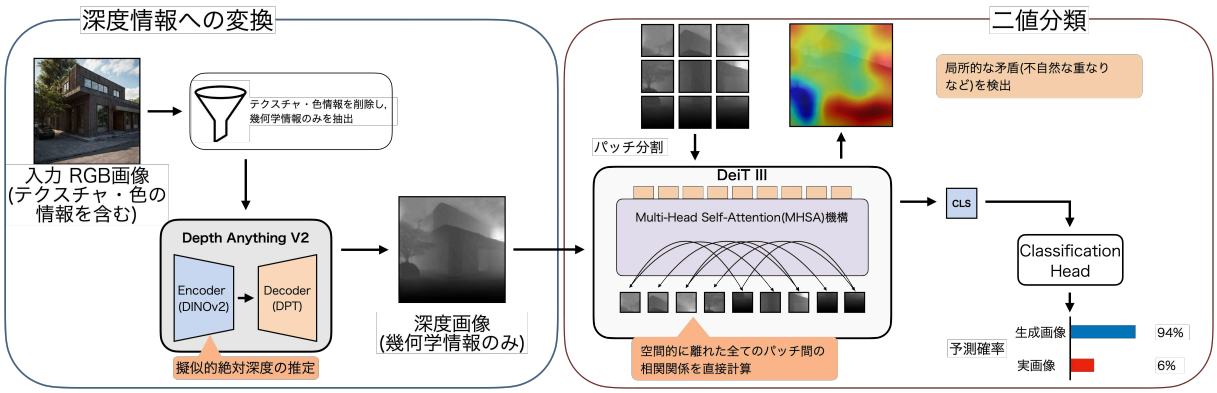

図 1 提案手法の流れ．

を獲得している．しかし，拡散モデルのように根本的に生成プロセスが異なる手法に対しては，従来手法が前提とする周波数的な痕跡が生じないため，検出精度が低下する．

既存手法の限界は，Zhu ら [29] が構築した 100 万枚規模の GenImage ベンチマーク実験においても実証されている．実際，Stable Diffusion V1.4[21] の生成画像で学習した ResNet-50[9] や Swin-T[12] などの検出器が，ADM[4] や GLIDE[15] といった異なるアーキテクチャの生成モデルに対して，検出精度が大幅に低下することが報告されている．この結果は，RGB 画像を入力する手法が学習に使用した生成モデルに過剰適合することによる検出精度の低下を示している．

2.2 単眼深度推定

単眼深度推定（Monocular Depth Estimation: MDE）は，一枚の RGB 画像から各画素の奥行き情報を推定する技術である．近年，大規模データセットで学習された MiDaS[20] や Dense Prediction Transformer (DPT)[19] などの深度推定モデルにより，未知のシーンに対する相対深度推定能力は大きく向上した．さらに，膨大なラベルなし画像を活用して汎用的な特徴表現を獲得する Depth Anything V1[26] が登場し，MDE の精度は飛躍的な進歩を遂げた．

その後継モデルとして提案された Depth Anything V2[27] は，DPT アーキテクチャをベースとしたエンコーダ・デコーダ構造を採用している．特に特徴抽出を担うエンコーダ部分には，強力な自己教師あり学習モデルである DINOv2[16] を組み込んでおり，画像内の微細な構造的関係性を深く理解する能力を持つ．さらに，約 6200 万枚の膨大な疑似ラベルデータと強力なデータ拡張を用いた学習パイプラインを経ることで，劣化画像や複雑なシーンに対しても破綻のない深度マップを出力できる能力を備えている．本研究では，これを空間解析の基盤として採用している．

3. 提案手法

提案手法の概要を図 1 に示す．提案手法は，入力画像から深度を推定するプロセスと，推定した深度マップから生成画像を判定する二値分類プロセスの 2 段階で構成される．

3.1 深度情報への変換

本手法の第一段階では，単眼深度推定モデルである Depth Anything V2[27] に実画像と生成画像を入力し，各画素の奥行きを表す深度マップへと変換する．特徴抽出を担う DINOv2 が画像から物体の形や配置といった本質的な構造を読み取る．次に，デコーダである DPT がその構造情報を手がかりとして，画素ごとの奥行きを計算する．

一般的な相対深度は入力画像間でスケールが変動するため，後段の分類プロセスにおける幾何学的矛盾の正確な学習を阻害する恐れがある．そこで本手法では，正確な物理的距離を正解データとして持つ Virtual KITTI 2 [2] で学習された重みを採用する．画像間で統一された物理スケールを持つ絶対深度を推定することで，安定した矛盾パターンの抽出を実現する．

3.2 二値分類プロセス

本手法の第二段階では，推定した深度マップから生成画像か実画像かを判定する二値分類を行う．分類器には，Vision Transformer (ViT)[5] ベースである Data-Efficient Image Transformers III (DeiT III)[23] を採用する．第二段階では，入力された深度マップを細かいパッチに分割し，自己注意機構によって全パッチ間の相関関係を計算する．次に，評価された画像全体の空間的な整合性を分類用の特徴量（Class Token）に集約し，それに基づいて生成画像か実画像かの判定を行う．

局所的な特徴抽出を行う畳み込みニューラルネットワーク [11] に対し，ViT は画像全体を同時に考慮して遠近法の崩れなど，離れた領域間における幾何学的不整合を直接比

較できる．そのため，大域的な矛盾の検出において有効な構造である．

4. 評価実験

提案手法の有効性を検証するために 3 つの評価実験を行う．1 つ目は，未知の生成モデルにより生成された画像に対する汎化性能の定量評価である．2 つ目は，画像劣化に対するロバスト性の評価である．3 つ目は，顕著性マップを用いた分類根拠の定性分析である．

4.1 実験条件

評価には，大規模な生成画像検出ベンチマークである GenImage[29] を使用する．GenImage は，ImageNet[3] 由来の実画像約 1,331,000 枚と，BigGAN[1]，Stable Diffusion（V1.4/V1.5）[21]，Midjourney V5（Midjour）[14]，ADM[4]，GLIDE[15]，Wukong[25]，VQDM[8] の計 8 種類の生成モデルによる生成画像約 1,350,000 枚で構成される．学習には Stable Diffusion V1.4 の生成画像を使用し，評価時には学習に含まれていない全ての生成モデルに対して個別に評価するクロス生成器評価を採用する．評価指標には検出精度として正解率を用いる．比較手法には，GenImage[29] においてベースラインとして報告されている空間領域ベースの検出手法 ResNet-50[9]，DeiT-S[22]，Swin-T[12]，CNNSpot[24]，GramNet[13]，および周波数領域ベースの検出手法 Spec[28]，F3Net[18] を用いる．なお，これらの比較手法はすべて RGB 画像を入力とする．

4.2 実験 1：未知の生成モデルに対する汎化性能

表 1 にクロス生成器評価の結果を示す．提案手法は 8 つの生成モデルの平均検出精度が 95.4%であり，比較手法において最も検出精度が高い Swin-T の 74.8%と比較して 20.6%上回る結果となった．また，比較手法が 40〜60%台の検出精度に留まる ADM や GLIDE に対し，提案手法は 99%以上の検出精度を達成した．RGB 画像を入力とする比較手法は，学習に用いた特定の生成モデルに過剰に適合してしまうため，未知のモデルに対しての汎化性能が低下する．これに対し，提案手法は大域的な空間構造の不整合性を捉えることで，未知の生成モデルに対しても高い検出精度を実現したと考えられる．一方，提案手法は Midjourney に対する検出精度が 70.7%に留まった．この結果は，極めて高い描画性能を持つ最新の生成モデルである Midjourney が 3 次元的整合性を獲得しており，深度マップ上の破綻が微細であるためと考えられる．

4.3 実験 2：画像劣化に対するロバスト性

画像が SNS などで共有される際，データ容量や通信量の削減のための圧縮処理により画質が劣化することがある．このような実運用上のシナリオを想定し，画像劣化に対するロバスト性を検証する．Stable Diffusion V1.4 のテストデータに対して，品質係数 $q = 65, 30$ の JPEG 圧縮および標準偏差 $\sigma = 3, 5$ の Gaussian blur といった劣化を適用した生成画像を用いて，検出精度を評価する．

表 1 クロス生成器評価における検出精度の比較 [%].

Datasets	ResNet-50	DeiT-S	Swin-T	CNNSpot	Spec	F3Net	GramNet	Ours
Midjour	54.9	55.6	62.1	52.8	52.0	50.1	54.2	**70.7**
SD V1.4	**99.9**	**99.9**	**99.9**	96.3	99.4	**99.9**	99.2	99.7
SD V1.5	99.7	99.8	99.8	95.9	99.2	**99.9**	99.1	99.6
ADM	53.5	49.8	49.8	50.1	49.7	49.9	50.3	**99.5**
GLIDE	61.9	58.1	67.6	39.8	49.8	50.0	54.6	**99.6**
Wukong	98.2	98.9	99.1	78.6	94.8	99.9	98.9	**99.6**
VQDM	56.6	56.9	62.3	53.4	55.6	49.9	50.8	**99.5**
BigGAN	52.0	53.5	57.6	46.8	49.8	49.9	51.7	**95.0**
Ave	72.1	71.6	74.8	64.2	68.8	68.7	69.9	**95.4**

表 2 に画像劣化に対する検出精度の比較結果を示す．比較手法は特定の劣化に対して大きく検出精度が低下していることがわかる．F3Net は $\sigma = 5$ の Gaussian blur を適用した生成画像において 51.7%まで検出精度が低下した．これは，周波数領域ベースの検出器が依存する高周波アーティファクトが平滑化されたためである．また，ResNet-50 や Swin-T も品質係数 $q = 30$ の JPEG 圧縮によって 50%台前半まで検出精度が低下した．このことは，空間領域ベースの検出器が依存する微細なテクスチャが，JPEG 圧縮特有のブロックの歪みにより破綻したことを示している．これに対し，提案手法は全劣化条件下で平均検出精度 99.4%という極めて高い水準を維持した．比較手法が依存する微細ノイズは画像劣化で容易に欠落する．しかし，本手法が着目する大域的な 3 次元構造は空間的により低周波な情報であり，激しい劣化を受けてもその構造的特徴が反転・喪失しにくい．そのため，Depth Anything V2 が劣化画像からでも一貫した深度マップを推定し，幾何学的破綻を正確に捕捉し続けたことがロバストな検出を可能とした要因であると考えられる．

表 2 画像劣化に対する検出精度の比較 [%].

Detectors	JPEG		Blur		Ave
	$q = 65$	$q = 30$	$\sigma = 3$	$\sigma = 5$	
ResNet-50	51.9	51.2	97.9	69.4	70.6
DeiT-S	55.6	50.5	94.4	67.2	69.8
Swin-T	52.5	50.9	94.5	52.5	67.0
CNNSpot	97.3	97.3	97.4	77.9	92.5
Spec	50.8	50.4	49.9	49.9	50.1
F3Net	89.0	74.4	57.9	51.7	62.1
GramNet	68.8	53.4	95.9	81.6	82.2
Ours	**99.4**	**99.3**	**99.4**	**99.5**	**99.4**

4.4 実験 3：顕著性マップを用いた分類根拠の分析

提案手法の分類器が深度マップのどの領域を根拠として判定しているかを検証するため，ブラックボックス可視化手法である Randomized Input Sampling for Explanation (RISE)[17] による分類根拠の分析をする．RISE は，部分的にマスクを適用した画像をモデルに入力し，その際の分類確信度の低下量を観測する．分類確信度の低下が最も大

表 3　各生成モデルにおける局所的破綻捕捉スコアの比較．

Datasets	局所的破綻捕捉スコアの平均値	提案手法の検出精度 [%]
ImageNet (実画像)	0.295	-
Midjourney	0.353	70.7
SD V1.4	0.736	99.7
SD V1.5	0.726	99.6
ADM	0.674	99.5
GLIDE	0.735	99.6
Wukong	0.701	99.6
VQDM	0.735	99.5
BigGAN	0.692	95.0

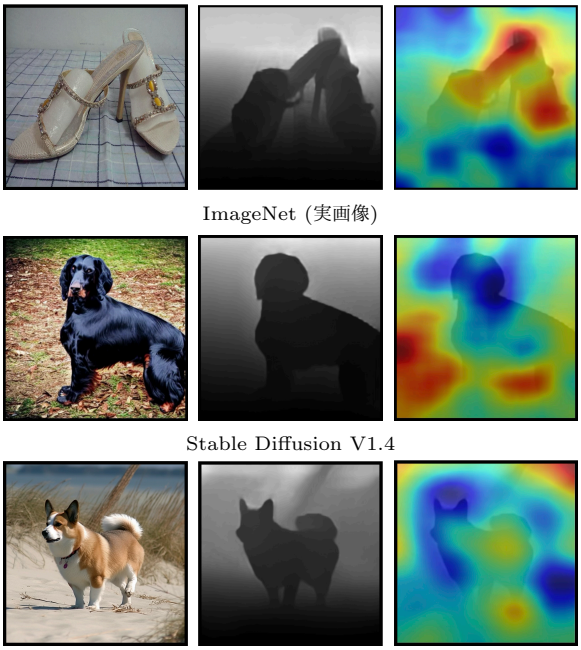

図 2　実画像と各生成モデルの注目領域比較（左から RGB 画像，深度マップ，RISE ヒートマップ）．

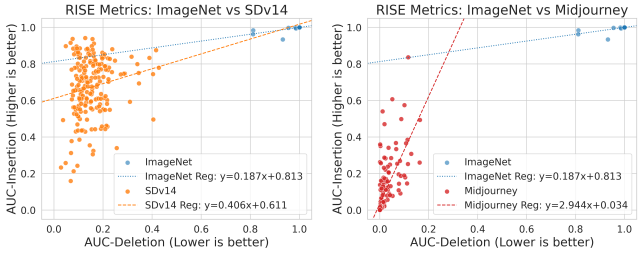

図 3　AUC-Deletion と AUC-Insertion の分布．

きいマスク領域が判定において重要であると推定し，注目領域として可視化する．

図 2 に提案手法の各画像に対する注目領域を可視化した結果の例を示す．実画像では被写体全体に注目しているのに対し，Stable Diffusion V1.4 では物体間の境界や奥行きが急激に変化する箇所など，空間構造の不整合が生じている領域に注目が集中している．一方，Midjourney では明確な破綻箇所がなく，注目が全体に分散していることがわかる．これは，最新の生成モデルが高度な 3 次元的整合性を獲得し，決定的な幾何学的破綻が消失しているためであり，表 1 の検出精度低下と整合性がある．

RISE では，判断根拠を表すヒートマップがモデルの判断根拠を正確に捉えているかを，モデルの出力変化から客観的に評価する指標として AUC-Deletion（d_{del}）と AUC-Insertion（d_{ins}）を提案している．AUC-Deletion は重要な画素を消去した際に低下した分類確信度を測定し，AUC-Insertion は重要な画素を追加した際に上昇した分類確信度を測定する指標である．図 3 に実画像で構成される ImageNet と各生成モデルの AUC-Deletion と AUC-Insertion の分布比較結果を示す．なお，本実験にはそれぞれランダムに選択した 200 サンプルを用いた．ImageNet は散布図の右上にプロットが密集していることがわかる．画像のどの領域をマスクしても分類確信度が低下せず，画像全体が均等に分類根拠となっていることがわかる．一方，Stable Diffusion V1.4 は左上から左下へ広く分散しており，特定の破綻箇所をマスクすれば分類確信度が低下し，提示すれば上昇するという局所的破綻に基づく挙動を示している．これに対し Midjourney は原点付近に密集し，モデルの注目領域を提示しても分類確信度が上昇せず，明確な分類根拠が存在しないことが確認できる．

次に，全テストデータを用いた統計的な指標により，分類器が局所的な幾何学的破綻をどの程度正確に捕捉できているかを定量評価するため，理想座標 $(0,1)$ からのユークリッド距離に基づく局所的破綻捕捉スコア S を式 (1)　で定義する．

$$S = 1 - \frac{\sqrt{d_{\mathrm{del}}^2 + (1 - d_{\mathrm{ins}})^2}}{\sqrt{2}} \qquad (1)$$

表 3 に各生成モデルの局所的破綻捕捉スコアの平均値を示す．表 3 から明らかなように，提案手法の検出精度が高い生成モデルほど局所的破綻捕捉スコアも高い値を示す傾向があり，定義したスコアがモデルの判定根拠の確からしさを評価する指標として妥当であることが確認できる．Stable Diffusion V1.4 をはじめとする 7 つの生成モデルのスコアは 0.674~0.736 と高い値を示す一方，Midjourney は 0.353 と低く，幾何学的破綻が少ないため分類器が明確な判定根拠を見出せていないことを裏付ける．

5.　おわりに

本稿では，RGB 画像の微細なノイズに依存する従来の生成画像検出手法の限界に対し，単眼深度推定による深度情報への変換と DeiT III を組み合わせた生成画像検出手法を提案した．GenImage を用いた評価実験により，未知の生成モデルに対する汎化性能と画像劣化に対するロバスト性を実証した．さらに，RISE を用いた定性・定量分析により，提案手法が物体間の境界や空間の歪みを的確に捉えていることを確認した．今後は，幾何学的一貫性の高い生成モデルに対応するため，深度情報の幾何学的特徴と RGB 画像の微細特徴を統合するマルチモーダル学習の導入を目指す．

参考文献

[1] Brock, A., Donahue, J. and Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis, *International Conference on Learning Representations* (2019).

[2] Cabon, Y., Murray, N. and Humenberger, M.: Virtual KITTI 2, *CoRR*, Vol. abs/2001.10773 (2020).

[3] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Li, F.-F.: ImageNet: A Large-Scale Hierarchical Image Database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (2009).

[4] Dhariwal, P. and Nichol, A.: Diffusion Models Beat GANs on Image Synthesis, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34, pp. 8780–8794 (2021).

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale, *International Conference on Learning Representations (ICLR)* (2021).

[6] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D. and Holz, T.: Leveraging frequency analysis for deep fake image recognition, *International conference on machine learning*, pp. 3247–3258 (2020).

[7] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in neural information processing systems*, Vol. 27 (2014).

[8] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L. and Guo, B.: Vector Quantized Diffusion Model for Text-to-Image Synthesis, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10696–10706 (2022).

[9] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).

[10] Ho, J., Jain, A. and Abbeel, P.: Denoising diffusion probabilistic models, *Advances in neural information processing systems*, Vol. 33, pp. 6840–6851 (2020).

[11] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).

[12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022 (2021).

[13] Liu, Z., Qi, X. and Torr, P. H. S.: Global Texture Enhancement for Fake Face Detection in the Wild, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8060–8069 (2020).

[14] Midjourney: Midjourney, `https://www.midjourney.com/home/` (2022). Accessed: 2026-02-03.

[15] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I. and Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, *Proceedings of the 39th International Conference on Machine Learning* (Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. and Sabato, S., eds.), Proceedings of Machine Learning Research, Vol. 162, PMLR, pp. 16784–16804 (2022).

[16] Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al.: DINOv2: Learning Robust Visual Features without Supervision, *Transactions on Machine Learning Research* (2024).

[17] Petsiuk, V., Das, A. and Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models, *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, BMVA Press, p. 151 (2018).

[18] Qian, Y., Yin, G., Sheng, L., Chen, Z. and Shao, J.: Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues, *European Conference on Computer Vision (ECCV)*, pp. 86–103 (2020).

[19] Ranftl, R., Bochkovskiy, A. and Koltun, V.: Vision transformers for dense prediction, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188 (2021).

[20] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V.: Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 44, No. 3, pp. 1623–1637 (2022).

[21] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695 (2022).

[22] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H.: Training data-efficient image transformers & distillation through attention, *International Conference on Machine Learning (ICML)*, pp. 10347–10357 (2021).

[23] Touvron, H., Cord, M. and Jégou, H.: Deit iii: Revenge of the vit, *European Conference on Computer Vision (ECCV)*, pp. 516–533 (2022).

[24] Wang, S.-Y., Wang, O., Zhang, R., Owens, A. and Efros, A. A.: CNN-generated images are surprisingly easy to spot... for now, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704 (2020).

[25] Wukong: Wukong, `https://xihe.mindspore.cn/modelzoo/wukong` (2022).

[26] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. and Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[27] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H.: Depth anything v2, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37, pp. 21875–21911 (2024).

[28] Zhang, X., Karaman, S. and Chang, S.-F.: Detecting and Simulating Artifacts in GAN Fake Images, *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6 (2019).

[29] Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J. and Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image, *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36, pp. 77771–77782 (2023).