画像識別器を用いた局所的な画像変換判定の性能評価

神城 潤光[1,a)]　関谷 勇司[1,b)]

概要：本研究では生成 AI による生成画像の識別のため，既存の識別器を用いた実験により，その識別性能の限界について検証する．GAN は生成器と識別器の並行した学習により高精度の生成画像や変換画像を作成する手法であり，CycleGAN などの様々な派生手法が提案されている．このような変換画像に対して，既存の識別器は Checkerboard Pattern と呼ばれる GAN の構造由来の特徴を学習することで高精度で汎化的な識別性能を実現しているが，画像の変換が局所的であるような画像については Checkerboard Pattern の痕跡が小さいために，依然として識別が困難である場合がある．本研究ではこうした局所的な画像変換を実際のデータセットと変換器の作成によって複数種類の画像データセットに対して実験を行い，既存の識別器の識別精度についてその現状と限界の検証を行った．本研究の成果は画像識別における既存の識別器の性能値を明らかにし，生成器や識別器の改良に関する今後の研究に貢献することが期待される．

キーワード：画像変換, GAN, Checkerboard Pattern, 性能評価

Performance evaluation of local image conversion judgment using image recognition

Junaki Kamishiro[1,a)]　Yuji Sekiya[1,b)]

Abstract: In this study, we verify the limitations of existing discriminators by conducting experiments using them to identify images generated by generative AI. GAN is a method that creates high-precision generated images and converted images through parallel learning of generators and discriminators, and various derivative methods such as CycleGAN have been proposed. For such transformed images, existing classifiers achieve high-precision and generalizable classification performance by learning a feature called the Checkerboard Pattern, which originates from the structure of GANs. However, for images where the transformation is localized, the traces of the Checkerboard Pattern are small, making classification difficult in some cases. In this study, we conducted experiments on multiple types of image datasets using actual datasets and transformers created for such localized image transformations, and verified the current status and limitations of the classification accuracy of existing classifiers. The results of this study clarify the performance values of existing classifiers in image classification and are expected to contribute to future research on improving generators and classifiers.

Keywords: image transform, GAN, Checkerboard Pattern, evaluation of discriminators

1. はじめに

1.1 背景

近年，生成 AI をはじめとする AI 技術の発展により，生成 AI を活用したテキスト，画像，動画の生成と，その社会的な活用が期待されつつある．一方でこうした生成 AI による出力は，フェイクニュースの作成[*1] やフィッシング等の詐欺[*2] に悪用される可能性もあり，こうした危険性へ

[1] 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology
[a)] kamishiro-junaki835@g.ecc.u-tokyo.ac.jp
[b)] sekiya@nc.u-tokyo.ac.jp

[*1] https://www3.nhk.or.jp/news/special/article/world20231107-01.html
[*2] https://www.theverge.com/news/624159/youtube-ai-generated-neal-mohan-phishing-scam

の対策が社会的に必要とされている．

このような生成 AI への技術的な対策として提案されているのが識別器による識別である．深層学習を用いた画像の識別や分類は，既に人手による分類性能を凌駕しており [1]，識別や分類タスクに応じた様々なモデルの構成や学習データセットが公開されている．特に画像の識別においては CNN モデルをベースとした識別器が主流になっており，ResNet[2] や Xception[3] などの様々なモデル構造の提案や，CNNDetection[4] などの既存のモデル構造を転移学習に用いた研究が存在し，GAN[5] や拡散モデル [6] といった生成手法に対する高精度な識別制度が実証されている．

しかし，こうした識別器に対して攻撃を試みる研究も存在し，例えば敵対的サンプルによる攻撃 [7] は，微小なノイズ画像を付与することにより識別器や分類器を誤認識させることができ，こうした既存の AI に対する攻撃手法は敵対的攻撃と呼ばれる．既存の識別器の社会的な実用のためには，こうした悪意のある画像データセットや画像生成器に対しても比較的高い識別制度を維持することを示す必要があるが，識別器の提案や改善に関する論文と比べて識別器に対する敵対的攻撃や耐性の評価を行う研究は少ない．そこで本研究では GAN による画像生成と識別に注目し，既存の識別器の構造を逆手に取った敵対的攻撃とその攻撃に対する既存の識別器の識別性能の評価と現行手法の限界に関する考察を行う．

1.2 本論文の構成

本論文の構成はまず第 2 節で既存の画像生成手法として GAN に注目し，その派生手法として CycleGAN を紹介する．またそれに対する既存の画像識別手法として CNNDetection や GANDetection，オンライン識別ツールなどを紹介し，こうした識別器の識別根拠に関する研究として Checkerboard Pattern についても説明する．第 3 節では局所的な画像変換と Checkerboard Pattern に関する説明から既存の識別器の限界について示唆し，第 4 節では実際にデータセットを用いて局所的な画像変換を行う生成器を作成し，識別実験を通じて，既存の識別器の現状と限界について確認する．最後に第 5 節で本研究のまとめについて述べる．

2. 関連研究

2.1 GAN に基づく画像生成手法

生成 AI における画像生成手法には GAN や拡散モデルなど様々なものが存在するが，特に GAN は派生手法の豊富さや先行研究の多さなどから，拡散モデルと並んでよく利用される画像生成手法である．

2.1.1 GAN

GAN（Generative Adversarial Network）は生成器と識別器の並行した学習により，高精度の生成画像や変換画像を作成する手法である．

GAN 生成器の一般的な構造としては，エンコーダとデコーダの 2 層の構造により，画像の次元削減（ダウンサンプリング）とアップサンプリングを行うことで画像を生成するが，ここでエンコーダは次元削減により，入力画像の概要的な情報の抽出と特徴ベクトルの生成を，デコーダはアップサンプリングにより，特徴ベクトルから最終的な出力画像の詳細を決定する役割を担うと解釈される．

GAN には様々な派生手法が知られ，例えば ProGAN[8] や CycleGAN[9] などが有名である．特に CycleGAN は高精度な画像変換を行うことができ，画像の変換タスクによく用いられている．

2.1.2 CycleGAN

CycleGAN は GAN から派生した画像生成手法の一種であり，主に入力画像のスタイル変換などの画像変換タスクに用いられることが多い．

CycleGAN の特徴としては，ペア化されていない画像群による学習が可能である点が挙げられる．例えばある二つのドメイン間の変換において，CycleGAN 以前の変換器は両ドメインの学習画像に物体や空間的な対応関係を必要としており，こうしたペア画像に関するデータセットを用意しなければならなかった．このような問題に対して CycleGAN はペア画像でなくとも各ドメインの全体的な特徴を学習し，高精度な画像変換を可能とした点が特徴的であり，例えば画像データの数や種類が制限される絵画データセットなどでも GAN による高精度な変換が可能となった．

2.2 CNN に基づく画像識別手法

画像の識別タスクにおいては主に CNN（Convolutional Neural Network）と呼ばれるモデル構造が利用されることが多く，CNN に基づく様々なモデルや学習が提案されている．中でも CNNDetection や GANDetection は様々なカテゴリに対する汎化的な識別性能や GAN 以外の生成手法に対する識別性能が期待されている手法である．

2.2.1 CNNDetection

CNNDetection は ResNet-50 を基にした GAN の識別モデルであり，幅広い GAN の派生手法に対して高精度な識別が可能であることが知られている．

CNNDetection の特徴としては，学習に一種類の GAN データセット (ProGAN) の学習のみを用いている点が挙げられる．一般にある生成手法 A による生成画像を識別するためには，その生成手法 A による訓練データセットで学習することにより高精度な識別が可能となり，学習したモデルは別の生成手法 B による生成画像も高精度に識別できるとは限らない．

しかし CNNDetection は損失関数やデータの前処理に独自の工夫を加え，また ProGAN の学習データセットにつ

いて，そのクラス数やサイズを巨大にすることで，GAN の一般的な特徴を学習し ProGAN 以外の派生手法によるテストセット (StyleGAN[10]，BigGAN[11]，CycleGAN[9]，StarGAN[12]，GauGAN[13]) についても高精度な識別が可能となっている．

また CNNDetection は GAN 以外の生成手法に対する識別性能も期待されており，CNNDetection の再学習によって，拡散モデル由来の生成画像の識別や顔画像の識別においても高い識別精度を示すことが報告されている [14]．

2.2.2 GANDetection

GANDetection[15] は複数の CNN モデルの学習によって GAN 生成画像の高精度な識別を可能とした識別モデルである．

CNNDetection が豊富なカテゴリ数の学習セットで学習しているのに対して，GANDetection は 5 つの異なる CNN モデルを並列させ，それぞれのモデルを顔画像，動物画像，絵画画像など異なるデータセットによって学習するように設計されている．これにより各カテゴリ特有の特徴が各モデルで学習され，また予測スコアはそれぞれのモデルの予測スコアの単純和をとることによって算出されるため，各モデルの総合的な評価によって入力画像が生成画像であるか否かを判断するような構造になっている．

実際に GANDetection は AFHQ2[16]，Metfaces[*3]，FFHQ[*4]などの様々なカテゴリのデータセットに対して高い識別精度を示しており，CNNDetection と同様に汎化的な識別性能を期待されている．

2.3 オンライン識別ツール

生成 AI がよりコモディティ化していく中で，生成画像の識別に対する需要の高まりからオンラインで利用可能な識別ツールも多く，GAN や拡散モデルなど様々な生成手法に対する高精度な識別が手軽に利用できる．特に Hive[*5]，Illuminarty[*6]，SightEngine[*7]はオンライン識別ツールの中でも著名である．

このうち Illuminarty と SightEngine は API を購入することにより完全な識別機能を利用することができるが，Hive は企業向けのみリリースされており大量の画像を識別させることはできない．しかし Hive には識別機能のデモ版が提供されており，一日に利用可能な回数は制限されているものの少量の画像であれば識別させることは可能である．

2.4 Checkerboard Pattern

GAN には様々な派生手法が存在することを紹介したが，GAN 生成画像には Checkerbord Pattern と呼ばれる GAN の構造由来の特徴が生じることが知られており，CNNDetection をはじめとする識別器の識別根拠に深く関係している．

GAN 生成器はエンコーダとデコーダの 2 層による一般的構造を持つことを説明したが，特にデコーダはエンコーダによって生成された特徴ベクトルから詳細な出力を決定する役割から，生成画像の精度や画質に大きく関係する層であると考えられる．デコーダにおけるアップサンプリングには主にフィルタを用いた次元拡張が行われるが，ここで用いられるフィルタは主に転置畳み込み（transposed convolution）と最近隣補完法（nearest neighbor interpolation）の二種類の補完アルゴリズムに基づくものが使われ，様々な GAN の派生手法についてこのアップサンプリングの構造は共通である [17]．

そのため GAN 生成画像にはこうした二種類のフィルタによる特徴的な画素パターンが生じることが知られており，これを Checkerboard Pattern と呼ぶ．Checkerboard Pattern は画像の周波数変換によって確認することができ，**図 2** のように特徴的な格子模様としてスペクトル画像に現れる．GAN に関する識別器はこうした Checkerboard Pattern を根拠に識別を行っていると考えられ，CNNDetection についてもその効率性の理由として，大規模な学習データセットによる Checkerboard Pattern の学習を挙げている [4]．

3. 既存研究の課題

3.1 局所的な画像変換

生成 AI を用いた画像変換タスクには様々なものがあるが，その一つに局所的な画像変換がある．例えば**図 1** は様々な画像変換器により画像修正を行った例であり，ツールによって修正の精度は異なるものの，比較的高精度な汚れやシミの修正が可能であることがわかる．特に CycleGAN は既存の消しゴムマジックツール (iPhone 16e[*8]，Google Pixel[*9]，フォト消しゴム 6[*10]) と比べて比較的自然な生成画像を生み出すことができることが確認できる．こうした画像修正は広告事業や創作活動等に利用可能な一方で，画像内に本来存在するものを消してしまうことから，フェイクニュースの作成や，オンラインショッピングにおける優良誤認につながる可能性がある．こうした危険性から CNNDetection をはじめとする既存の識別器は，このような局所的な画像変換による変換画像も高精度に識別することが求められる．

[*3] https://github.com/NVlabs/metfaces-dataset
[*4] https://github.com/NVlabs/ffhq-dataset
[*5] https://thehive.ai/
[*6] https://illuminarty.ai/en/
[*7] https://sightengine.com/
[*8] https://www.apple.com/jp/iphone-16e/
[*9] https://store.google.com/jp/category/phones?hl=ja
[*10] https://www.sourcenext.com/product/0000014506

図 1 様々な変換ツールによる変換サンプル
Fig. 1 Conversion samples using various conversion tools.

3.2 局所的な画像変換の Checkerboard Pattern

CNNDetection をはじめとする従来の識別器は Checkerboard Pattern がその識別根拠に深く関わっており，局所的な画像変換の識別のためにはこの Checkerboard Pattern の痕跡が一定程度顕著である必要があると考えられる．しかし図 2 と**図 3** を比較すればわかるように，従来の CycleGAN による画像変換は Checkerboard Pattern が顕著に表れるため，変換前後におけるスペクトルの変化は大きいのに対して，CycleGAN を局所的な画像変換に用いた時は従来画像と比べてそれほどスペクトルの変化が大きくないことが確認できる．

これらの結果から，Checkerboard Pattern を基に識別を行う既存の識別器が局所的な画像変換を識別できるかは不明である．そのため，局所的な画像変換に対する識別精度や性能に関する評価を行うことは，偽画像に騙されないという社会的な要求に応えるためにも必要な技術となる．仮に既存の識別器がこれを識別できないとき，悪意のある第三者がこの局所的な変換を利用した敵対的攻撃を行い，識別器の識別精度低下や分類の誤認識などを引き起こす可能性があるためである．

4. 実験内容

4.1 目的

今回の実験においては既存の識別器の限界について考察するために局所的な画像変換に着目し，実際のデータセットを用いた画像変換器の作成と，変換画像に対する既存の識別器や識別ツールの識別精度をいくつかの指標によって確認する．

4.2 実験手順

StainDoc Dataset[18] はテキストや図，画像などが様々な組み合わせで印刷された紙面について，清潔な状態の紙面（Clean）とシミや汚れなどが付着した紙面の 2 種類の画像データを集めたデータセットである．後者についてはマーカーによって紙面を塗ったもの（Marker），ウォーターマークを添付したもの（WaterMark），スタンプを押下したもの（Seal）の 3 つのカテゴリが存在し，**図 4** はそれぞれのサンプルを示したものである．

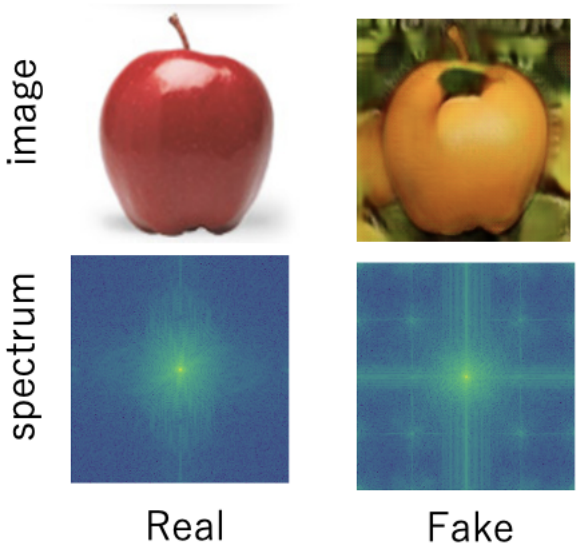

図 2 Apple2Orange による画像変換とスペクトルの変化
Fig. 2 Image conversion and spectral changes using Apple2Orange.

実験においては各カテゴリから 500 枚の画像を抽出し，Clean カテゴリと他 3 カテゴリが対になるようにして，紙面の汚れ除去訓練データセットを作成し（Marker2Clean, WaterMark2Clean, Seal2Clean），各データセットを用いて CycleGAN を 200 エポック分学習した．次に Marker, WaterMark, Seal の 3 カテゴリから訓練セットとは別に新たに 500 枚の画像を抽出し，各変換器によって偽 Clean 画像を作成した．そして変換前と変換後の画像をそれぞれ 0_real, 1_fake とラベル付けし，既存の識別器に識別させた．ただしオンラインツール（Hive, Illuminarty, SightEngine）に関しては API の制限のため，0_real, 1_fake からそれぞれ 50 枚の画像を抽出して識別させた．

4.3 実験結果

表 1 は各カテゴリのテストセットに対する既存の識別器

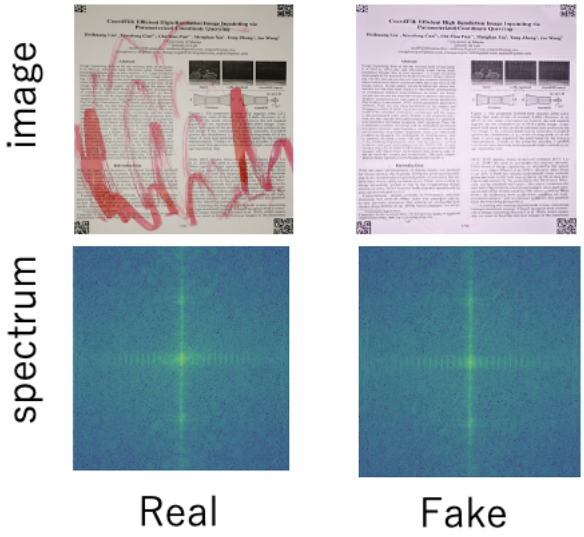

図 3 Marker2Clean による画像変換とスペクトルの変化
Fig. 3 Image conversion and spectral changes using Marker2Clean.

の ACC（Accuracy）と AP（Average Precision）である．一般に ACC はその識別器における識別精度を，AP は識別器の分類精度を表現しているものと考えられ，いずれの指標も 1 に近い程高精度であると解釈される．

表 1 を見ればわかるように CNN ベースの識別器である CNNDetection と GANDetection はいずれも ACC と AP が極端に低く，StainDoc Dataset による局所的な画像変換を認識できていないことがわかる．また Illuminarty をはじめとするオンライン識別ツールについても Hive を除きそれほど精度が高いとは言えない．Hive については ACC が 80%程度を示しており，局所的な画像変換を比較的識別できてはいるものの，依然として ACC の改善の余地を残していることがわかる．

以上の結果より既存の識別器は Hive を除き，識別精度や分類精度がそれほど大きくなく，局所的な画像変換に対する識別を苦手としていることが分かった．特に CNNDetection については既存の研究によって Checkerboard Pattern を基に識別を行っていることは明らかであったが，逆に Checkerboard Pattern が小さいような変換画像については識別を不得手としていることが明らかになった．また同じ CNN ベースの識別器である GANDetection についても Checkerboard Pattern との関連性は明らかではなかったが，CNNDetection と同様に識別精度が極端に低いことから Checkerboard Pattern を識別の根拠としており，CNNDetection と同様の課題を持つことが推測される．

5. おわりに

5.1 本研究の成果

本研究では局所的な画像変換に注目し，画像変換器の作成と既存の識別器についての性能評価を行った．作成した画像変換器は既存の消しゴムマジックツールと比べて，比較的自然な生成画像を生成できることが確認された．またその変換画像の識別においては，CNNDetection をはじめとする複数の CNN ベースの識別器と，いくつかのオンラインツールでは識別が困難であることが確認された．

本研究の成果を生成器の改良という視点で考えると，今回作成した画像変換器を紙面の変換以外のタスクへも適用可能な変換器へと改良することで，既存の消しゴムマジックツールよりも自然な画像変換が可能であると考えられる．また識別器の改良という視点では，局所的に修正された偽画像を見抜くことは現行の技術では難しいため，モデル構造や学習過程の改良により，このような敵対的攻撃に対しても識別精度を保つ識別手法の提案が必要である．

5.2 今後の展望

本研究では StainDoc Dataset を用いた実験を行ったが，別カテゴリによる画像変換や識別を行った場合，本研究の実験結果とは異なる結果が得られる可能性がある．例えば今回作成した画像変換器は StainDoc Dataset の画像変換タスクではより自然な画像を生成したが，別データセットによる学習や変換が常に消しゴムマジックツールよりも高精度ではない可能性もあり，より汎用的に利用可能な画像変換器を考える場合，そのような画像カテゴリによる変換精度のばらつきを抑えるような構造に改良する必要がある．

また識別器の改良においても，今回紹介した既存の識別器の全てがアルゴリズムを公開しているわけではないため，他の CNN ベースの識別器についても同様の実験を行ったり，別の DNN モデルを利用した識別実験を行う必要がある．これにより，局所的な画像変換に対する識別性能の向上を狙う．

参考文献

[1] Ha, A.Y.J., Passananti, J., Bhaskar, R., Shan, S., Southen, R., Zheng, H., Zhao, B.Y. (2024). *Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?*, CCS 2024, pp. 4822–4836.
[2] He, K., Zhang, X., Ren, S., Sun, J. (2016). *Deep Residual Learning for Image Recognition.*, CVPR 2016, pp. 770–778.
[3] Chollet, F. (2017). *Xception: Deep Learning with Depthwise Separable Convolutions.*, CVPR 2017, pp. 1800–1807.
[4] Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A. (2020). *CNN-Generated Images Are Surprisingly Easy to Spot... for Now.*, CVPR 2020, pp. 8692–8701.
[5] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y. (2014). *Generative Adversarial Nets.*, NIPS 2014, pp. 2672–2680.
[6] Ho, J., Jain, A., Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models.*, NIPS 2020. pp. 6840–6851.
[7] Goodfellow, I.J., Shlens, J., Szegedy, C. (2015). *Explain-*

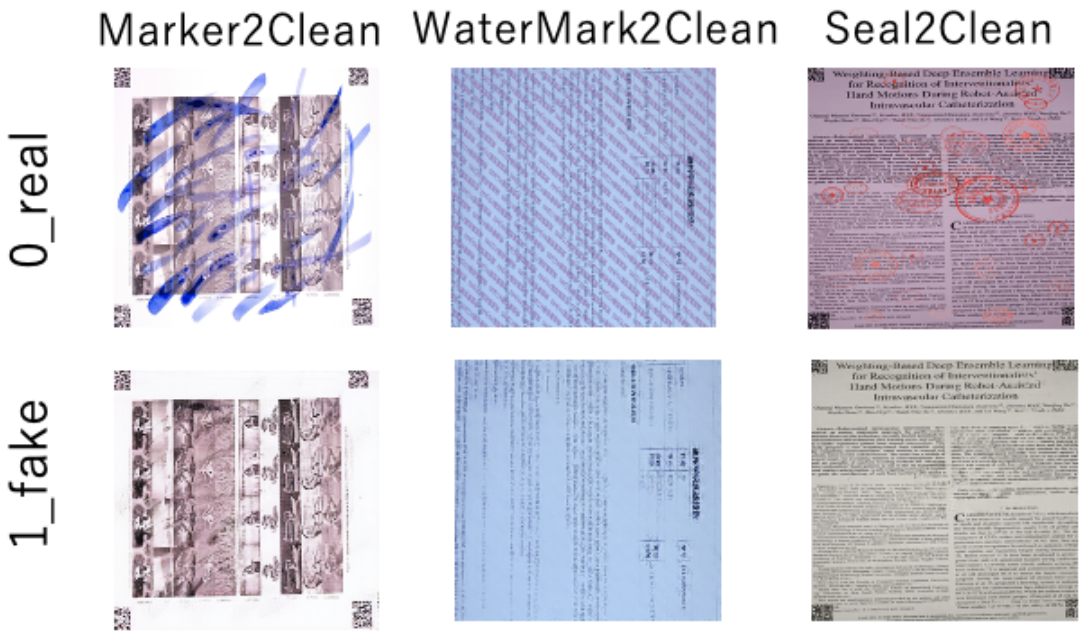

図 4 StainDoc Dataset から作成したテストセットのサンプル

Fig. 4 Sample test set created from the StainDoc Dataset.

表 1 局所的な変換画像に対する既存の識別器の識別精度 (ACC(%) / AP(%))

Table 1 Recognition accuracy of existing recognizers for locally transformed images.

	Marker2Clean	WaterMark2Clean	Seal2Clean
CNNDet	50.8 / 65.8	50.6 / 53.0	50.2 / 95.3
GANDet	51.1 / 69.2	50.0 / 40.8	49.4 / 46.9
Hive	80.8 / 97.3	78.7 / 97.3	77.7 / 97.3
Illuminarty	50.0 / 52.4	50.0 / 52.4	50.0 / 52.4
SightEngine	72.7 / 89.0	67.6 / 89.0	60.6 / 89.0

ing and Harnessing Adversarial Examples., ICLR 2015. WACV 2025. pp. 7614–7624.

[8] Karras, T., Aila, T., Laine, S., Lehtinen, J. (2018). *Progressive Growing of GANs for Improved Quality, Stability, and Variation.*, ICLR 2018.

[9] Zhu, J.Y., Park, T., Isola, P., Efros, A.A. (2017). *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.*, ICCV 2017. pp. 2242–2251.

[10] Karras, T., Laine, S., Aila, T. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks.*, CVPR 2019. pp. 4401–4410.

[11] Brock, A., Donahue, J., Simonyan, K. (2019). *Large Scale GAN Training for High Fidelity Natural Image Synthesis.*, ICLR 2019.

[12] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J. (2018). *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation.*, CVPR 2018. pp. 8789–8797.

[13] Park, T, Liu, M.Y., Wang, T.C., Zhu, J.Y. (2019). *Semantic Image Synthesis With Spatially-Adaptive Normalization.*, CVPR 2019. pp. 2337–2346.

[14] Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H. (2023). *DIRE for Diffusion-Generated Image Detection.*, ICCV 2023. pp. 22388–22398.

[15] Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S. (2022). *Detecting Gan-Generated Images by Orthogonal Training of Multiple CNNs.*, ICIP 2022. pp. 3091–3095.

[16] Choi, Y., Uh, Y., Yoo, J., Ha, J.W. (2020). *StarGAN v2: Diverse Image Synthesis for Multiple Domains.*, CVPR 2020. pp. 8185–8194.

[17] Zhang, X, Karaman, S., Chang, S.F. (2019). *Detecting and Simulating Artifacts in GAN Fake Images.*, WIFS 2019. pp. 1–6.

[18] Li, M., Sun, H., Lei, Y., Zhang, X., Dong, Y., Zhou, Y., Li, Z., Chen, X. (2025). *High-Fidelity Document Stain Removal via A Large-Scale Real-World Dataset and A Memory-Augmented Transformer.*,