画質変化に頑健なディープフェイク画像検出手法
－画質情報を導入したCLIPに基づく検出－

村上　尭良†　古垣 遵之‡　山名 早人§

† 早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
‡ 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
§早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: † orimaorima@akane.waseda.jp, ‡ waseda3979@akane.waseda.jp, §yamana@waseda.jp

あらまし　近年，生成AI技術の進歩により，高精細な画像や映像を生成することが可能となっている．しかし，同時に偽情報の拡散や，プライバシー・著作権侵害といった深刻な社会問題を引き起こしている．生成画像の悪用を防ぐため，数多くの検出器が存在するが，学習データのファイル形式の偏りに起因して，JPEG圧縮耐性が低いという課題がある．本研究では，その課題を解決するため，CLIPを活用した新たな検出手法を提案する．具体的には，CLIPの学習時に画像の「真贋」情報をキャプションに埋め込む先行研究の手法を拡張し，新たに「画質」に関する情報もキャプションに統合する．実験では，高・中・低の3段階の画質を持つJPEG画像を作成し，それぞれの画質状態に対応したキャプションを付与して学習を行った．評価実験はJPEGとPNGが混在した画像，品質係数96,90,80,70,60,50のJPEG画像に対して行い，ペア・ブートストラップ法（有意水準0.05）を用いた検定を行った．検定の結果，画質情報を付与しない場合と比較して，画質情報を付与した手法ではJPEG圧縮された画像に対してAccuracyが有意に低下したものの，AUCに関して，JPEGとPNGが混在したデータ，品質係数90,80,70,60,50のJPEG画像に対して有意な改善が認められた．

キーワード　画像認識，画像生成，偽情報の検出，ディープフェイク

1. はじめに

近年，生成AI技術の進歩により，真贋の判別が困難な画像や映像を生成することが可能となっている．高度な画像，映像生成技術は芸術・創作活動に応用される一方で，虚偽情報の拡散やプライバシー侵害，政治的操作への悪用が懸念されている．例えば，総務省の「令和6年版 情報通信白書」[1]によれば，2022年の台風15号による災害時，静岡県の水害被害を装った生成画像がSNS上で拡散され，多くの利用者が事実と誤認した事例が報告されている．誤情報の氾濫はデジタル空間における情報の信頼性を著しく損なう．したがって，安心・安全な情報社会を実現するために，AIによって生成された画像を高精度に識別する検出技術の確立が急務となっている．

しかし，既存のディープフェイク検出器はJPEG圧縮された画像に対して脆弱であることが指摘されている[2]．JPEG圧縮耐性の欠如をもたらす主な要因は，学習データセットに含まれる画像形式のバイアスであり，多くのデータセットにおいて，自然画像はJPEG形式であるのに対し，生成画像はPNG形式で構成されている．その結果，検出器が生成過程の痕跡ではなく，ファイル形式そのものを真贋判定の根拠として学習する弊害が生じる．結果としてJPEG圧縮を施した生成画像を自然画像と誤認し，検出精度が低下する実態が報告されている．

本研究では，JPEG圧縮された画像に対しても検出精度を維持する手法を提案する．まず，学習段階ですべての画像をJPEG形式に統一し，ファイル形式に起因するバイアスを排除する．加えて，形式の統一のみでは画質変化への適応が不十分であることを踏まえ，画質劣化の度合いに応じた柔軟な特徴抽出を促す目的から，画像の画質を言語情報としてモデルに提示する．

具体的な手順として，まず1枚の画像に対して品質係数（Quality Factor）96,70,50の3段階のJPEG圧縮を適用して多様な画質を持つ画像データセットを構築する.続いて，各圧縮率に対応した「high-quality」等の画質情報タグをテキストキャプションに統合する．画質情報を統合したキャプションを用いてCLIPベースのモデルを学習させることにより，画質劣化の影響を考慮した上での真贋特徴の識別，および真贋判定が可能な検出器の構築を図る．

以下2章で関連研究を紹介し，3章で使用するデータセットを説明する．4章で提案手法について説明し，５章で提案手法を用いた実験を行い，6章で評価する．7章でまとめる．

2.関連研究

本章では，まず画像生成手法について説明する．次に本研究の土台となる，大量の画像とテキストの対応関係を学習したモデルである CLIP(Contrastive Language-Image Pre-training）[3]について説明し，CLIP を用いたディープフェイク検出への応用例として Tan ら(2025)[4] が提案した C2P-CLIP を紹介する.そして Tan らのモデルをベースラインとして利用している，ディープフェイク検出器を紹介する.

2.1 画像生成手法

近年，画像生成技術は急速な進展を遂げており，アプローチは大きく分けて，敵対的生成ネットワーク（Generative Adversarial Networks: GANs）と，拡散モデル（Diffusion Models）の 2 つに分類される.

GANs は，Goodfellow ら(2014)[5]によって提案されたモデルであり，生成器と検出器を競合させることで学習を行う．GANs の学習の枠組みは時間とともに数多くの改良が加えられ，特に Brock ら(2019)[6]が提案した BigGAN は，モデルのパラメータ数とバッチサイズを極限まで大規模化することで，生成画像の品質を改善することに成功した.

GANs に対し，近年では，拡散モデルが台頭している．拡散モデルは，画像へ段階的にノイズを付加する拡散過程と，ノイズを段階的に除去する逆過程により構成される.

Dhariwal(2021)ら[7]は，従来の拡散モデルのアーキテクチャの最適化に加え，逆過程で外部分類器の勾配を用いて生成を制御する Classifier Guidance を導入した，Ablated Diffusion Model (ADM)を提案した．実験の結果，画像の忠実度と多様性において BigGAN より優れたスコアを達成し，拡散モデルが GANs の生成品質を凌駕し得ることを実証した.

また，Nichol ら(2022)[8]は GLIDE において，Ho ら(2021)[9]が提案した外部分類器を用いない Classifier-free Guidance を採用し，モデル自身の知識を活用することで，従来の CLIP[3]による誘導よりも高い写実性とキャプションへの忠実度が得られる事実を示した.

さらに，高解像度画像の生成に伴う計算コストの課題を解決するため, Rombach ら(2022)[10]は従来の画素空間ではなく，重要な特徴量を保持したまま次元圧縮された潜在空間上で拡散プロセスを行う Latent Diffusion Models (LDM)を提案した．LDM を基盤として Stability AI により公開された Stable Diffusion (SD) v1.4 および v1.5 は，一般消費者向けの GPU でも動作する軽量性と高い描画性能を両立している．またモデルのアーキテクチャや学習データにおける多様化も進んでおり，Gu ら(2022)[11]は画像を離散的なトークン列に変換し，離散的な潜在空間上で拡散プロセスを定義する VQ-Diffusion を提案した．また，Huawei の研究チーム(2022)[12]は大規模な中国語マルチモーダルデータセットを用いた Wukong を開発した.

商用サービスとしては Midjourney[13]があり，高度な芸術性と写実性を両立させた画像を生成可能である.上記の生成器を含む画像生成技術の発達により，生成画像と自然画像を識別することは困難である[14].

2.2 CLIP

CLIP (Contrastive Language-Image Pre-training)[3]は，4 億ペアの大規模な画像とテキストのペアを用いて学習されたマルチモーダルモデルである．CLIP は，画像エンコーダとテキストエンコーダで構成されており，対応する画像とテキストの特徴量ペアの類似度を最大化し，対応しないペアの類似度を最小化する対照学習(Contrastive Learning）によって訓練される．対照学習により，画像内の視覚的な特徴と言語的な意味や概念を共通の特徴空間上で結びつけることを可能とする.

2.3C2P-CLIP

Tan らは CLIP[3]を使ったディープフェイク検出器の一つとして，C2P-CLIP[4]を提案した．彼らは，CLIP の特徴量を解析し，モデルの判断が「真贋」の識別ではなく「類似した概念」のマッチングに基づいているという事実を明らかにした.その知見に基づき, Tan らは Clipcap[1][15]により生成された画像のキャプションに対して，生成画像には"Deepfake"，自然画像には"Camera" といった「カテゴリ共通プロンプト」を付与した．キャプションの作成と拡張を行い，モデルに対して明示的に真贋の概念注入(Concept Injection)を施すことで，GenImage(2023)[16]データセットにおいて，Mean Accuracy 95.8%を達成した.

2.4 ディープフェイク検出器

画像生成技術の発達とともに生成画像を検出する技術も多数提案されている．本節では，本研究が注目するデータセットのバイアス除去に着目した研究を紹介する.

Guillaro ら(2025)[17]は，自然画像と生成画像の内容を一致させるため，元の画像を生成のガイドとする自己条件付き再構成に加え，画像の一部を生成画像で補完するインペインティングを用いた学習方法を提案した．結果として，画像の内容に依存しない生成画像の検出を可能にしている．また，学習時に JPEG 圧縮やぼかしをランダムに適用することで，画像形式に起因するバイアスを低減している.

Chen ら(2025)[18]は，再構成した画像に，対応する

[1] https://github.com/rmokady/CLIP_prefix_caption

自然画像と同等の JPEG 圧縮を適用して周波数特性を一致させることで，データセットのバイアスを排除している．加えて，画像の一部をピクセルレベルで混合（Pixel Mixing）するデータ拡張を導入し，生成画像と自然画像の差異をさらに縮小させている．

上記の論文はいずれも Tan ら[4]のモデルをベースラインとして採用し，Tan らのモデルを上回ったと報告している．

2.5 まとめ

画像生成技術の発展に対し，C2P-CLIP[4]のように高い検出精度を達成する手法も提案されている．

しかしながら，C2P-CLIP は学習データセットのバイアスを保持した状態で学習が行われるため，JPEG 圧縮を施した画像に対する識別精度が低下する脆弱性を有している．

そこで，本研究では学習データセット全体に JPEG 圧縮を適用して画像のファイル形式バイアスの排除を行うと共に，C2P-CLIP の手法に着想を得て，画質の情報をテキストキャプションとして付与する手法を提案する．画質情報を明示的に学習プロセスへ組み込むことで画質劣化の度合いに応じた柔軟な特徴抽出を促し，先行研究とは異なるアプローチとして，画像圧縮に対してより頑健な検出器の構築を図る．

3.データセット

3.1GenImage

1 つ目に大規模ベンチマークデータセットである GenImage [2][16]を採用した．GenImage は，広範なカテゴリを網羅する ImageNet[19]の自然画像と，自然画像のカテゴリに対応する生成画像のペアによって構築されている．生成画像を作成する生成器は，ADM, GLIDE, Midjourney，Stable Diffusion v1.4，Stable Diffusion v1.5，VQDM，Wukong という 7 種類の拡散モデルに加え，GAN ベースのモデルである BigGAN を含めた計 8 種類が使用されている．多様な生成手法を網羅する構成は，特定の生成アルゴリズムに依存しない，汎用的な検出性能の評価を可能にする．

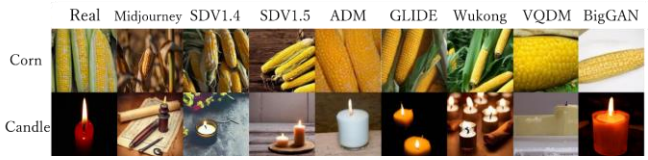

図 3.1 GenImage[16]の画像例
([16]の Fig.1 から一部改変・加工して引用)

3.2DDA-aligned MSCOCO

2 つ目に Chen ら[18]の手法（Dual Data Alignment; DDA）との比較検証を目的として，Chen らが作成した DDA-aligned MSCOCO[3]を学習データとして採用した．本データセットは，自然画像として Microsoft COCO[20]を用いており，VAE（Variational Autoencoder）を介して自然画像を再構成することで，画像内容が酷似した生成画像を作成している．

さらに，Chen らは自然画像と生成画像間の周波数特性を一致させるため，生成された画像に対して，生成の元となった自然画像と同等の JPEG 圧縮を適用した．加えて，ピクセルレベルでの自然画像と生成画像の混合（Pixel Mixing）によるデータ拡張を導入することで，自然画像と生成画像の差異をより縮小させている．

4.提案手法

4.1 概要

提案手法は，Tan らによって提案された C2P-CLIP [4]を基盤として用いる．C2P-CLIP ではモデルの判断が真贋の識別ではなく「類似した概念」のマッチングに基づいているというアイデアに基づき，明示的に真贋の概念注入を行い，モデルを学習させている．

提案手法では，本アイデアを画像の圧縮率の学習に用いることで JPEG 圧縮に頑健な検出器を構築することを考えた．はじめに，ファイル形式に基づいた学習を防ぐ目的で，複数の画質の JPEG に統一する．加えて，研究の核心となる二つの工程を実行する．まず，画質情報を付与した拡張キャプションを構築し，次いで画質損失を含む損失関数を用いた概念注入を行う．

4.2 キャプションの生成と拡張

4.2.1 初期キャプションの生成

キャプション生成に用いる学習データセット X を式 (1) の通り定義する．

$$X = \{x_j, y_j\}_{j=1}^{N},\ y \in \{0,1\} \quad (1)$$

ここで，x_jは入力画像，y_jは真贋ラベルを示し，$y_j = 1$は生成画像，$y_j = 0$は自然画像を意味する．画像の内容を記述する初期キャプションc_jの生成には，ClipCap [16] を用いる．生成された初期キャプションの集合Cを式(2) に示す．

$$C = \{c_j, y_j\}_{j=1}^{N},\ y \in \{0,1\} \quad (2)$$

4.2.2 画像形式の統一とデータ拡張

GenImage [16] および MSCOCO [20] データセットにおける画像形式のバイアスを排除するため，本研究では画像形式の統一と画質の多様化を実施した．

まず，Grommelt ら[2]による，GenImage の自然画像の多くは品質係数 96 の JPEG 形式で構成されているという指摘に基づき，本研究では自然画像の元画像を JPEG96 と定義した．また，MSCOCO データセット[20]についても自然画像の過半数が JPEG96 であったこと

[2] https://github.com/GenImage-Dataset/GenImage
[3] https://huggingface.co/datasets/Junwei-Xi/DDA-Training-Set

から，元画像を JPEG96 と定義した．元画像からそれぞれ再圧縮を行うことで，JPEG70 および JPEG50 の画像を作成し，画質のバリエーションを確保した．一方，PNG 形式である生成画像については，自然画像との条件を揃えるため，初めに JPEG96 で圧縮を施し，その後自然画像と同様に JPEG70, JPEG50 への再圧縮を行った．一連の流れにより,1 枚の画像x_jに対し，圧縮率$q \in \{96,70,50\}$に応じた 3 段階の画質バリエーション $x_{j,q}$を用意し，サンプル数を元の 3 倍に拡張した画像データセット$\tilde{X}$を式(3) の通り構築した．

$$\tilde{X} = \left\{ x_{j,q}, y_j \right\}_{j=1}^{N}, y \in \{0,1\}, q \in \{96,70,50\} \quad (3)$$

4.2.3 画質情報を導入したキャプションの拡張

モデルが圧縮による画質劣化の影響を考慮しつつ真贋特徴を学習できるよう，各画像$x_{j,q}$の圧縮率qに応じた画質情報タグQ_qをキャプションc_jに統合する手法を提案する．画質情報タグQ_qは，$q \in \{96,70,50\}$に対し，それぞれ "high-quality", "middle-quality", "low-quality" と定義した．画質情報タグQ_qと，C2P-CLIP で導入されたカテゴリ共通プロンプトである"Camera"（自然画像）および "Deepfake"（生成画像）を初期キャプションに組み合わせ，最終的な拡張キャプション$\tilde{c}_{j,q}$を式(4) の通りに構築した．

$$\tilde{c}_{j,q} = \begin{cases} (P_{real}, Q_q, c_j), if\ y_j = 0 \\ (P_{fake}, Q_q, c_j), if\ y_j = 1 \end{cases} \quad (4)$$

すべての拡張キャプションから構成される集合 $\tilde{C}$ を式 (5) に示す．

$$\tilde{C} = \left\{ \tilde{c}_{j,q} \right\}_{j=1}^{N} q \in \{96,70,50\} \quad (5)$$

構築された拡張キャプション集合$\tilde{C}$は，次節で述べる概念注入において，CLIP のテキストエンコーダの入力として使用される．

4.3 概念注入

先行研究である C2P-CLIP[4] に倣い，拡張キャプションに包含される画像の真贋情報および画質情報を画像エンコーダへ埋め込むため，学習プロセスに対照学習（Contrastive Learning）を導入する．学習時には，CLIP の画像エンコーダおよびテキストエンコーダのパラメータを固定した状態で，画像エンコーダに対して LoRA [22] を適用する．

学習の最適化には，先行研究で用いられた対照損失（Contrastive Loss）および分類損失（Classification Loss）に加え，本研究で新たに定義する画質損失（Quality Loss）を合わせた 3 種類の損失関数を用いる．3 種類の損失関数を統合して学習を行うことにより，画像エンコーダは画質劣化の影響を考慮しつつ真贋概念を効果的に獲得し，未知のデータに対する汎化性能が向上すると考えられる．

具体的な特徴量の算出において，テキスト特徴量u_jおよび画像特徴量v_jは以下の通り定義される．

$$u_j = \text{encoder}_{(text)}(\tilde{c}_{j,q}) \quad (6)$$

$$v_j = \text{encoder}_{img}^{(lora)}(x_{j,q}) \quad (7)$$

ここで，$\text{encoder}_{(text)}$はテキストエンコーダを$\text{encoder}_{img}^{(lora)}$は LoRA レイヤーを組み込んだ画像エンコーダを指す．

次に,対照損失$\mathcal{L}_{contrastive}$を以下のように計算する．

$$\mathcal{L}_{contrastive} = \frac{(\mathcal{L}_{u \to v} + \mathcal{L}_{v \to u})}{2} \quad (8)$$

$$\mathcal{L}_{v \to u} = -\frac{1}{N}\sum_{i=1}^{N} log\left(\frac{exp(v_i^T u_i)}{\sum_{j=1}^{N} exp(v_i^T u_j)}\right) \quad (9)$$

$$\mathcal{L}_{u \to v} = -\frac{1}{N}\sum_{i=1}^{N} log\left(\frac{exp(u_i^T v_i)}{\sum_{j=1}^{N} exp(u_i^T v_j)}\right) \quad (10)$$

さらに，画像特徴量v_jに対して線形識別機（Linear Classifier）を適用し，真贋の判定を行う．分類損失$\mathcal{L}_{classification}$および，画質損失$\mathcal{L}_{quality}$には,いずれもクロスエントロピー損失を採用する．

最終的な損失関数$\mathcal{L}$は，3 つの損失の加重和として定義する．

$$\mathcal{L} = \alpha * \mathcal{L}_{contrastive} + \beta * \mathcal{L}_{classification} + \gamma * \mathcal{L}_{quality} \quad (11)$$

上記のα，β，γは 3 つの損失のバランスを調整するためのハイパーパラメータである．

評価は，C2P-CLIP と同様に，LoRA 付きの画像エンコーダと線形識別機を用いて行われる．提案手法の概要を図 4.1 に示す．

5.評価実験

5.1 実験データセット

5.1.1GenImage

提案手法の有効性を検証するため，大規模ベンチマークデータセットである GenImage [16] を用いて評価実験を実施する．GenImage データセットは，モデルの訓練を行うための「トレーニングサブセット」と，未知の画像に対する性能を測定するための「テストサブセット」の 2 つにより構成される．GenImage を準拠し，自然画像と Stable Diffusion v1.4（SDv1.4）により生成された画像で構成されたトレーニングサブセットを採用した．また学習の際に，4.2.2 節で詳述したデータ拡張を行った．データ拡張後の生成画像および自然画像の枚数（いずれも学習に使用）の内訳を表 5.1 に示す．

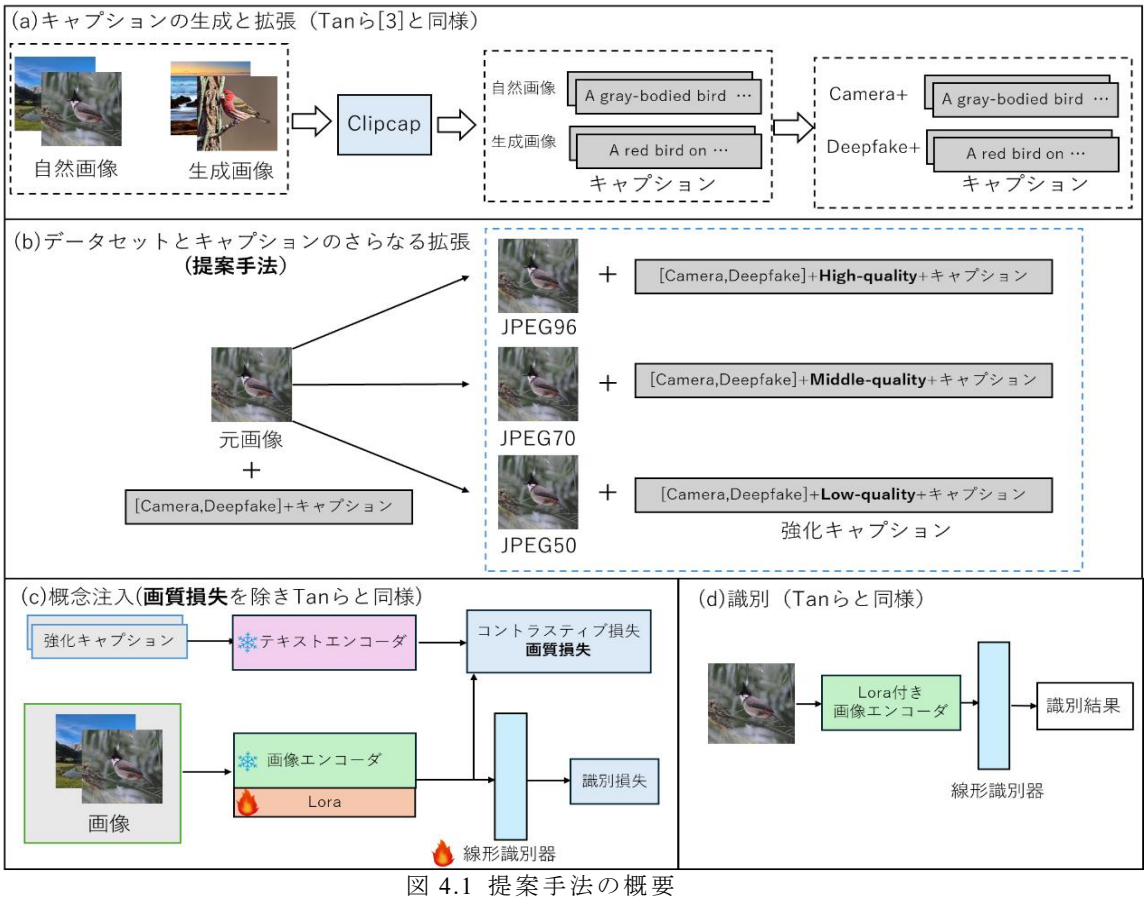

図 4.1 提案手法の概要

表 5.1 トレーニングサブセットの内訳

画質	生成画像	自然画像
JPEG 96	161,997	162,000
JPEG 70	161,997	162,000
JPEG 50	161,997	162,000
合計	485,991	486,000

4.2.2 節で述べた通り，単一の画像サンプルに対して 3 種類の品質係数を適用しているため,学習に使用した総画像枚数はオリジナルのデータセットの 3 倍に拡張されている．なお，トレーニングサブセットの生成画像において 3 枚のファイル破損が確認されたため，それらの画像は学習から除外した．

画像圧縮に対する頑健性を検証するため，複数の画質条件下で性能評価を実施する．評価においては，テストデータとして自然画像が JPEG 形式,生成画像が PNG 形式の未加工のデータに加え，PNG 形式の生成画像を JPEG 形式の自然画像にそろえる目的で JPEG 96 へ変換したデータを評価対象とする．加えて，自然画像および生成画像の両者に対し品質係数を 90,80,70,60,50 と段階的に設定して再圧縮を施し，未加工のデータを含む計 7 通りの条件下でモデルの性能を測定する．テストサブセットの詳細な内訳を，表 5.2 に示す．

表 5.2 テストサブセットの内訳

画質	生成画像	自然画像
未加工	50,000	50,000
JPEG 96	50,000	50,000
JPEG 90	50,000	50,000
JPEG 80	50,000	50,000
JPEG 70	50,000	50,000
JPEG 60	50,000	50,000
JPEG 50	50,000	50,000
合計	350,000	350,000

5.1.2DDA-aligned MSCOCO

Chen ら [18] が提案した手法 (Dual-Data-Alignment,DDA)との比較検証を行うため，同研究で構築された DDA-aligned MSCOCO を学習データとして採用した．GenImage[16]と同様に学習の際に，4.2.2 節で詳述したデータ拡張を行った．データ拡張後の生成画像および自然画像の枚数（いずれも学習に使用）の内訳は，表 5.3 に示す通りである．

表 5.3 トレーニングサブセットの内訳

画質	生成画像	自然画像
JPEG 96	118,287	118,287
JPEG 70	118,287	118,287
JPEG 50	118,287	118,287
合計	354,861	354,861

性能評価においては, 5.1.1 節で詳述した GenImage のテストサブセットを用いて評価を行った.

5.2 実験詳細

本実験におけるモデルの構築，および学習から評価に至る一連の手順は，先行研究である C2P-CLIP [4]の手法記述に基づき独自に実装したものである．論文内の記述に従うことで手法の再現を試みた.

提案手法の実装には PyTorch フレームワーク[4](version 2.2.2)を用い，バックボーンモデルには，Hugging Face Transformers ライブラリ[5](version 4.43.3)を通じて提供されている学習済み CLIP (ViT-L/14)を採用した．C2P-CLIP[4] の手法に準拠し，モデルの効率的な微調整のため，PEFT ライブラリ[6](version 0.11.1)[21]を用いて LoRA [22]を導入した．LoRA の設定に関しても，先行研究である C2P-CLIP を全面的に踏襲しており，画像エンコーダの Transformer ブロックにおける Query，Key，Value の各投射層に対して LoRA を適用した．ハイパーパラメータも C2P-CLIP の設定を参考に，ランク $r = 6$, スケーリング係数$\alpha = 6$, ドロップアウト率 0.8 に設定した．最適化アルゴリズムには Adam を使用し, 学習率は4.0×10^{-4}とした. また, CLIP の対照学習における温度パラメータτは，Tan らの論文に記載がなかったため，CLIP [3] の設定を参考にし 0.07 に固定した．学習におけるエポック数も C2P-CLIP と同様に 1 に設定した．再現性を担保するため，シード値は 123 に設定した．また，バッチサイズについては, ハードウェアの制約を考慮し, 先行研究の 128 から 64 へ変更して実施した.

また，本研究では識別閾値の最適化を目的として，検証サブセットを用いた検証（validation）を実施した．具体的には，学習に用いた SDv1.4 のトレーニングサブセットを除く 7 種類の生成器（Midjourney, SDv1.5, ADM, GLIDE, Wukong, VQDM, BigGAN）のトレーニングサブセットから各 1,000 枚(自然画像，生成画像共に 500 枚)の画像を抽出し，それらに対して品質係数$q \in \{96, 70, 50\}$の JPEG 圧縮を施したものを提案手法の検証サブセットとして用いた. そのため総計 21,000 枚の画像で検証サブセットは構成されている.

C2P-CLIP においては検証を用いた閾値の調整に関する具体的な記述が確認されなかったことから，本研究では実験条件の公平性を期すため，再現実験においても検証プロセスを導入した. C2P-CLIP の再現に際しては，同手法が学習データセットの拡張を行っていない点を考慮し，閾値調整には未加工の状態の画像を用いた．具体的には前述の 7 種類の生成器のトレーニングサブセットから各 3,000 枚(自然画像，生成画像共に 1,500 枚)の画像を抽出し，提案手法に用いた検証サブセットと同様に総計 21,000 枚の画像を用いた.

Chen ら[18]はベンチマーク評価において，閾値の調整を行わずに測定を実施していることから，検証サブセットによる閾値調整を省き，閾値を 0.5 に固定して測定を実施した.

本研究ではテキストキャプションの情報を強化したため，先行研究と比較して対照損失への重みを大きく設定した．一方で，本手法の主目的が画像の真贋検出であり，画質情報の学習は補助的な役割である点を考慮し，画質損失の重みは相対的に小さく設定した．先行研究である C2P-CLIP では対照損失の重みが 1.0, 分類損失の重みが 8.0 に設定されていたことを踏まえ, 式(11)に記載されている各係数について, α=2.0, β=8.0, γ=1.0 と設定して実験をおこなった.

5.3 評価方法

本研究では，提案手法の有効性を評価するため,Accuracy と AUC を評価指標として採用した．Accuracy の算出にあたって，識別閾値は 5.2 節で説明した検証サブセットを用いて調整を行った.評価は生成画像と自然画像の二値分類として行った．2 値分類時おける各要素を以下の図 5.4 の通り定義する.

実測値＼予測値	1	0
1	TP	FN
0	FP	TN

図 5.4 混合行列

TP，FN，FP，TN はそれぞれ True Positive，False Negative，False Positive，True Negative を意味する．本研究では，Positive は生成画像を指す．上記の 4 つの値を用いて，Accuracy，AUC を説明する．まず，Accuracy は式(14)で定義される.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{14}$$

[4] https://pytorch.org/
[5] https://github.com/huggingface/transformers
[6] https://github.com/huggingface/peft

真陽性率である TPR (True Positive Rate)は式(15)で定義される.

$$TPR = \frac{TP}{TP + FN} \tag{15}$$

偽陽性率である FPR(False Positive Rate)は式(16)で定義される.

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

AUC は，閾値を変化させた際の FPR を横軸，TPR を縦軸にプロットした ROC 曲線下の面積であり，0.0 から 1.0 の範囲でモデルの識別能力を表す.

また，モデル間の性能比較における統計的有意性を検証するため，ペア・ブートストラップ（Paired Bootstrap）法を採用した.

具体的には，GenImage[16]における 100,000 枚の画像を母集団とし，各画質条件下において，母集団と同数の画像（100,000 枚）を重複を許して抽出するリサンプリングを 2,000 回実施した．各試行において，抽出された同一の画像セットに対して両モデルを適用し，評価指標（Accuracy および AUC）の差分を算出することで，性能差の分布を推定した．分布から得られた 95% 信頼区間に基づき，信頼区間が 0 を跨がない場合に「統計的有意差がある」と判断した.

5.4 評価方法

本研究では，提案手法の有効性を詳細に分析するため，評価の目的に応じて対象モデルを以下の 2 つの枠組みで整理した.

5.4.1 既存手法との比較検証

提案手法を構成する各要素（データ拡張および画質情報の提示）が Accuracy,AUC に与える影響を明らかにするため，GenImage[16]のトレーニングサブセット(SDv1.4)を学習データとした 3 つの構成を比較する．各モデルの詳細は表 5.5 に示す通りである.

表 5.5 既存手法との比較検証における各モデルの詳細

モデル名	学習データ	キャプション構成
C2P-CLIP (再現)	GenImage オリジナル	真贋 + 内容
提案手法（画質なし）	GenImage JPEG 3 段階	真贋 + 内容
提案手法（画質あり）	GenImage JPEG 3 段階	真贋 + 画質 + 内容

5.4.2 提案手法における各要素の有効性検証

次に，先行研究である Chen ら[18]との比較を行うため，同研究で提案された DDA-aligned MSCOCO を用いた比較実験を設定した．Chen らのモデル(DDA)については，著者らにより公開されている学習済み重み を使用して性能を測定する.既存手法との公平な比較を目的として,提案手法(DDA 比較用)を設定した.各モデルの詳細は表 5.6 に示す.

表 5.6 DDA との比較検証における各モデルの詳細

モデル名称	学習データ
DDA	DDA-aligned MSCOCO オリジナル
提案手法 (画質キャプションあり)	DDA-aligned MSCOCO JPEG 3 段階

6.評価結果

6.1 提案手法の有効性評価

6.1.1 データ拡張の有効性調査

本研究では,学習データセットのファイル形式の差に起因するバイアスの影響を抑制するため,4.2.2 節で詳述したデータ拡張を実施した．バイアス低減に対する効果を調査するため，ベースライン（C2P-CLIP 再現モデル）と，JPEG 3 段階のデータ拡張を施した「提案手法（画質キャプションなし)」を比較した．ペア・ブートストラップ法による Accuracy の差の結果を表 6.1 に，AUC の差の結果を表 6.2 に示す.

表 6.2 より，画像に JPEG 圧縮を施すことで閾値の調整を行う場合でも Accuracy が低下する傾向が確認された．この結果は，先行研究の指摘通り，モデルが生成画像特有の痕跡ではなく，ファイル形式の差異を真贋判定の根拠として学習していることを強く示唆している.

6.1.2 画質キャプション付与の有効性調査

次に，本研究の主要な提案要素である「画質情報タグ（high/middle/low-quality）を用いたキャプション拡張」の有効性を検証するため，画質情報の有無による比較検証を行った．本節におけるベースラインである「画質キャプションなし」は，6.1.1 節で有効性が確認された「JPEG 3 段階のデータ拡張を施した提案手法」を指す．対する「画質キャプションあり」は，データ拡張に加え，キャプションへ画質情報を付与し，真贋概念と画質概念の同時注入を図った手法である．データ拡張によるバイアス排除効果を共通の前提とすることで，画質情報を言語情報としてモデルに提示することによる性能変化を評価する．5.1.1 節で説明した評価サブセットを用いて，ペア・ブートストラップ法による 7 つの画質条件の Accuracy の差の検定結果を表 6.3, AUC の差の検定結果を表 6.4 に示す.

表 6.4 より，JPEG 96 を除いたすべての画質条件下において，画質キャプションを導入したモデルは画質キャプションなしのモデルと比較して AUC が統計的に有意に改善した．AUC の向上は，画質情報を言語として明示的にモデルへ提示するアプローチが，圧縮に伴うノイズの影響と画像の本質的な真贋特徴を分離して学習する助けとなったことを示唆している.一方で,

表 6.1　C2P-CLIP と提案手法（画質キャプションなし）の Accuracy 差の検定結果

画質 (テストデータ)	提案手法の Accuracy (画質キャプションなし)	C2P-CLIP の Accuracy	ΔAccuracy [95% 信頼区間]	有意性
未加工	0.7454	0.8116	−0.0662 [−0.0682, −0.0641]	↓
JPEG 96	0.8691	0.7215	+0.1476 [+0.1449, +0.1501]	↑
JPEG 90	0.9066	0.7076	+0.1990 [+0.1965, +0.2016]	↑
JPEG 80	0.7698	0.6027	+0.1671 [+0.1647, +0.1696]	↑
JPEG 70	0.7921	0.5662	+0.2259 [+0.2231, +0.2287]	↑
JPEG 60	0.8251	0.5397	+0.2854 [+0.2825, +0.2881]	↑
JPEG 50	0.8263	0.5348	+0.2915 [+0.2869, +0.2942]	↑

※ Δ は「提案手法 − 再現手法」の差分を示す．
※ ↑ / ↓ は 95% 信頼区間が 0 を跨がず，正または負の方向に有意であることを示す．

表 6.2　C2P-CLIP と提案手法（画質キャプションなし）の AUC 差の検定結果

画質 (テストデータ)	提案手法の AUC (画質キャプションなし)	C2P-CLIP の AUC	ΔAUC [95% 信頼区間]	有意性
未加工	0.8934	0.9549	−0.0615 [−0.0632, −0.0597]	↓
JPEG 96	0.9171	0.8933	+0.0238 [+0.0215, +0.0260]	↑
JPEG 90	0.9325	0.8932	+0.0394 [+0.0371, +0.0418]	↑
JPEG 80	0.9281	0.9176	+0.0105 [+0.0085, +0.0123]	↑
JPEG 70	0.9361	0.9224	+0.0137 [+0.0117, +0.0156]	↑
JPEG 60	0.9507	0.9207	+0.0301 [+0.0281, +0.0320]	↑
JPEG 50	0.9727	0.9227	+0.0501 [+0.0486, +0.0516]	↑

※ Δ は「提案手法 − 再現手法」の差分を示す．
※ ↑ / ↓ は 95% 信頼区間が 0 を跨がず，正または負の方向に有意であることを示す．

表 6.3　画質キャプション導入有無による Accuracy 差の検定結果

画質 (テストデータ)	提案手法 (画質キャプションあり)	提案手法 (画質キャプションなし)	ΔAccuracy [95% 信頼区間]	有意性
未加工	0.7539	0.7454	+0.0085 [+0.0070, +0.0100]	↑
JPEG 96	0.8431	0.8691	−0.0260 [−0.0274, −0.0245]	↓
JPEG 90	0.8628	0.9065	−0.0437 [−0.0451, −0.0423]	↓
JPEG 80	0.7683	0.7698	−0.0015 [−0.0023, −0.0007]	↓
JPEG 70	0.7790	0.7921	−0.0130 [−0.0140, −0.0120]	↓
JPEG 60	0.8165	0.8252	−0.0087 [−0.0100, −0.0074]	↓
JPEG 50	0.7964	0.8263	−0.0299 [−0.0312, −0.0285]	↓

※ Δ は「画質キャプションあり − 画質キャプションなし」の差分を示す．
※ ↑ / ↓ は 95% 信頼区間が 0 を跨がず，正または負の方向に有意であることを示す．

表 6.4　画質キャプション導入有無による AUC 差の検定結果

画質 (テストデータ)	提案手法 (画質キャプションあり)	提案手法 (画質キャプションなし)	ΔAUC [95% 信頼区間]	有意性
未加工	0.8967	0.8933	+0.0033 [+0.0018, +0.0049]	↑
JPEG 96	0.9013	0.9171	−0.0158 [−0.0171, −0.0146]	↓
JPEG 90	0.9574	0.9325	+0.0249 [+0.0238, +0.0260]	↑
JPEG 80	0.9607	0.9281	+0.0326 [+0.0314, +0.0338]	↑
JPEG 70	0.9647	0.9361	+0.0286 [+0.0273, +0.0299]	↑
JPEG 60	0.9750	0.9507	+0.0243 [+0.0231, +0.0254]	↑
JPEG 50	0.9835	0.9728	+0.0107 [+0.0100, +0.0115]	↑

※ Δ は「画質キャプションあり − 画質キャプションなし」の差分を示す．
※ ↑ / ↓ は 95% 信頼区間が 0 を跨がず，正または負の方向に有意であることを示す．

表 6.5　既存手法(DDA)と提案手法（画質キャプションあり）による Accuracy 差の検定結果

画質 (テストデータ)	提案手法の Accuracy (画質キャプションあり)	DDA の Accuracy	ΔAccuracy [95% 信頼区間]	有意性
未加工	0.7435	0.9128	−0.1692 [−0.1719, −0.1665]	↓
JPEG 96	0.7639	0.8877	−0.1239 [−0.1264, −0.1213]	↓
JPEG 90	0.7200	0.8727	−0.1527 [−0.1554, −0.1500]	↓
JPEG 80	0.8087	0.8894	−0.0807 [−0.0835, −0.0779]	↓
JPEG 70	0.7995	0.8669	−0.0673 [−0.0702, −0.0645]	↓
JPEG 60	0.8107	0.8497	−0.0390 [−0.0419, −0.0358]	↓
JPEG 50	0.8036	0.8356	−0.0320 [−0.0350, −0.0291]	↓

※ Δ は「提案手法 − DDA」の差分を示す．
※ ↑ / ↓ は 95% 信頼区間が 0 を跨がず，正または負の方向に有意であることを示す．

表 6.6　既存手法(DDA)と提案手法（画質キャプションあり）による AUC 差の検定結果

画質 (テストデータ)	提案手法の AUC (画質キャプションあり)	DDA の AUC	ΔAUC [95% 信頼区間]	有意性
未加工	0.8604	0.9754	−0.1150 [−0.1172, −0.1129]	↓
JPEG 96	0.8905	0.9676	−0.0771 [−0.0789, −0.0753]	↓
JPEG 90	0.8446	0.9673	−0.1227 [−0.1250, −0.1205]	↓
JPEG 80	0.9225	0.9814	−0.0590 [−0.0606, −0.0574]	↓
JPEG 70	0.9248	0.9763	−0.0515 [−0.0530, −0.0499]	↓
JPEG 60	0.9213	0.9712	−0.0499 [−0.0516, −0.0483]	↓
JPEG 50	0.9117	0.9662	−0.0545 [−0.0563, −0.0528]	↓

※ Δ は「提案手法 − DDA」の差分を示す．
※ ↑ / ↓ は 95% 信頼区間が 0 を跨がず，正または負の方向に有意であることを示す．

表 6.3 より未加工のテストデータを除き，Accuracy は低下する結果となった．

6.2DDA をベースラインとした評価

次に本研究とは異なるアプローチでデータセット間のバイアス除去に取り組んだ，Chen ら[18]との比較を行った．比較の際には，6.1 節と同様にペア・ブートストラップ法による AUC と Accuracy の差の検定を行った．ペア・ブートストラップ法による 7 つの画質条件の Accuracy の差の検定結果を表 6.5，AUC の差の検定結果を表 6.6 に示す．表 6.5，表 6.6 より先行研究と比べて，閾値の調整を行ったにもかかわらず Accuracy，AUC 共に有意に低下した．

7.終わりに

従来のディープフェイク検出器は，学習データセットにおける自然画像（JPEG）と生成画像（PNG）の不整合に起因するバイアスにより，JPEG 圧縮された画像に対して識別精度が低下するという課題を抱えていた．

本研究では，学習データセットが持つバイアスを排除し，画質変化に頑健な検出器を構築するため，画像形式の統一と多様な圧縮強度の導入を目的としたデータ拡張，および画質情報を言語的に統合した新たな検出手法を提案した．

実験の結果，まず学習段階におけるデータ拡張，すなわち全画像の JPEG 形式への統一と 3 段階の圧縮適用が，ファイル形式バイアスの排除に有効であることが確認された．さらに，画質情報をキャプションに統合する提案手法により，JPEG 96 を除く全画質条件で Mean AUC が有意に向上した．これは，画質情報を明示的に提示することで，モデルが圧縮ノイズの影響を抑制し，画像の本質的な真贋特徴をより効果的に分離して学習できたためと考えられる．一方で，自然画像が JPEG 形式，生成画像が PNG 形式の未加工のテストサブセットを除き，Mean Accuracy は低下し，向上した分離能力を直接的な精度向上へ結び付けるには至らなかった．

既存手法である Chen ら[18](DDA）との比較においても，Accuracy および AUC 共に先行研究を上回る結果を得ることはできなかった．今後の課題として，今回調査できなかった損失の重みを含むパラメータの最適化による提案手法の性能改善が考えられる．

参　考　文　献

[1] 総務省, "令和 6 年版 情報通信白書," 2024, pp. 49–51.[Online].Available:https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/pdf/00zentai.pdf (Accessed:Jan.3, 2026).

[2] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8748–8763.

[3] C. Tan et al., "C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 7, 2025, pp. 7184–7192.

[4] P.Grommelt et al., "Fake or JPEG? Revealing common biases in generated image detection datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, vol. 15644, 2024, pp. 80–95

[5] I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.

[6] A. Brock et al., "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019

[7] P. Dhariwal et al., "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 8780–8794.

[8] A. Nichol et al., "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 162, 2022, pp. 16784–16804.

[9] J. Ho et al., "Classifier-free diffusion guidance," in *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, vol. 160, 2021.

[10] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.

[11] S. Gu et al., "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10696–10706.

[12] J. Gu et al., "Wukong: A 100 million large-scale Chinese cross-modal pre-training dataset and a benchmark," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 26418–26431.

[13] Midjourney, Inc., "Midjourney," 2025. [Online]. Available: https://www.midjourney.com/ (Accessed: Dec. 30, 2025).

[14] S. J. Nightingale et al., "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proc. Natl. Acad. Sci. USA (PNAS)*, vol. 119, no. 8, Art. no. e2120481119, 2022.

[15] R. Mokady et al., "ClipCap: CLIP prefix for image captioning," *arXiv preprint arXiv:2111.09794*, 2021.

[16] M. Zhu et al., "GenImage: A million-scale benchmark for detecting AI-generated image," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 77771–77782.

[17] F. Guillaro et al., "A bias-free training paradigm for more general AI-generated image detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[18] R. Chen et al., "Dual data alignment makes AI-generated image detector easier generalizable," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 38, 2025.

[19] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[20] T. Lin et al., "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comput. Vis. (ECCV), vol.8693,2014, pp. 740–755.

[21] S. Mangrulkar et al., "PEFT: Parameter-Efficient Fine-Tuning of Million-Scale Models on Consumer Hardware,"2022.[Online].Available: https://github.com/huggingface/peft(Accessed: Dec. 30, 2025).

[22] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*,2022