自己教師あり学習を導入したWavelet Vision Transformerによる Deepfake検出の高精度化

○高瀬俊希 †，山内悠嗣 †（講演者に○を付けてください）
†：中部大学大学院工学研究科ロボット理工学専攻
tr22008-4099@sti.chubu.ac.jp

概要：深層学習により画像上の顔の交換や属性・表情を変更する Deepfake が問題となり，Deepfake により生成された画像を検出する研究が活発に行われている．近年，画像認識分野において Vision Transformer ベースの手法が優れた性能を発揮することが報告されているが，Deepfake 検出問題においては畳み込みニューラルネットワークベースの手法と比較して性能が低下することが分かっている．そこで，本研究では自己教師あり学習を導入した Wavelet Vision Transformer による高精度な Deepfake 検出法を提案する．Wavelet Vision Transformer は，Deepfake 検出において画像中の重要な高周波成分を捉えることができるモデルである．表現学習の一種である自己教師あり学習と組み合わせることで，Deepfake 画像における加工の痕跡を正確に検知することが可能となり，高精度な検出が期待できる．
＜キーワード＞ Deepfake 検出，Vision Transformer，自己教師あり学習

1. はじめに

深層学習を用いた画像生成技術の発展に伴い，Deepfake と呼ばれる画像上の顔の交換や属性・表情変更が容易に操作できるようになった．Deepfake によって生成される画像は，非常に精巧であり一見すると真の顔画像[1]と区別がつかない画像が生成される．生成された画像や動画は，ニュースサイトや Social Networking Service(SNS)，動画配信サービスなどで悪意のある目的で利用されることが社会的な問題となっている．そのため，Deepfake を自動的に検出する研究 [1][2] が盛んに取り組まれている．

Deepfake を検出する手法としては，畳み込みニューラルネットワーク (CNN) に基づく多くの手法が提案されている．一方，画像認識の分野においては Vision Transformer(ViT)[3] に基づく手法が優れた性能を発揮することが報告されている．しかしながら，Deepfake 検出問題においては ViT に基づく手法 [4][5] は CNN に基づく手法に比べて性能が低いことが報告されている．その理由は，Deepfake 画像における偽造痕跡の特性にある．Deepfake により画像を加工すると，偽造痕跡として高周波成分のノイズが画像に加わる [6]．そのため，モデルにより高周波成分を抽出することが重要となるが，ViT は高周波成分よりも低周波成分の画像特徴を抽出する傾向にある [7]．ViT ベースの手法により高精度な Deepfake 検出を実現するためには，画像中の高周波成分をよく捉えるようなアプローチが必要である．

そこで，本研究では高周波成分の特徴を抽出できる Wavelet Vision Transformer[8] に基づく Deepfake 検出法を提案する．また，自己教師あり学習を導入することで，Data Augmentation により画像に加わる摂動と Deepfake によって加工した際に残る偽造痕跡の違いを区別するように特徴抽出器を学習する．

[1]これ以降，真の顔画像を Real 画像，何らかの画像生成手法で生成した顔画像を Deepfake 画像と呼ぶ．

2. 関連研究

2.1. Deepfake 検出

Deepfake の生成には深層学習を用いた画像生成手法が用いられる．画像生成手法として，画像を生成する Generator と入力画像の真偽を見分ける Discriminator で構成される Generative Adversarial Network(GAN)[9] や，入力画像を圧縮して潜在変数を抽出する Encoder と潜在変数を基に画像を復元する Decoder で構成される AutoEncoder(AE)[10] などが挙げられる．これらの手法は，コンピュータビジョンやコンピュータグラフィックスの分野においてデザイン作成や映画，アバターの作成などに利用されている．一方で，これらの技術を利用した画像上の顔の交換や属性・表情変更 [11][12] などの加工がなりすまし等へ悪用され問題となっている．

Deepfake 検出は，一般的に Real 画像と Deepfake 画像を分類する 2 値分類問題として扱われる．FaceForensics++(FF++)[13] では，Deepfake 検出における一般的なデータセットとシンプルな CNN を用いた Deepfake 検出手法を提案している．また，Face X-Ray[14]，Pair-wise self-Consistency Learning-Inconsistency Image Generator(PCL-I2G)[15]，Self-Blended Images(SBIs)[16] では生成した Deepfake 画像を学習に利用している．中でも SBIs は，1 枚の顔画像に異なる Data Augmentation を適用して疑似的な Deepfake 画像を生成するアプローチを提案している．微細な偽造痕跡が含まれる Deepfake 画像を用いて学習することで，高精度かつ汎化性能の高い Deepfake 検出を可能にしている．

2.2. Vision Transformer

Vision Transformer(ViT)[3] は，自然言語処理の分野において注目を集めている Transformer[17] を画像処理に応用した手法である．ViT は高精度な画像認識が可能である一方で，帰納バイアスが弱いため CNN を超える性能を発揮するためには大規模なデータセットを用いた事前学習が必要となる．

この問題に対し，Data-efficient image Transformer (DeiT)[18] は，知識蒸留を導入することで大量の教師付き画像を用いた事前学習がなくても ViT を超える高精度な画像認識を実現した．また，DeiT では Data Augmentation や正則化，学習率などの ViT ベースの研究における基本となる設定を提示しており，以降の研究においても DeiT で提示されたパラメータが踏襲されている．

Class Attention in Image Transformers (CaiT)[19] は，Class Token をモデルの冒頭にて入力せず，Encoder の後半で加える処理に変更した．これにより，画像認識に不要な情報や悪影響を与える情報が蓄積されることを防止することが可能となった．Class Token の追加後は，Class Attention と Class Token を用いて Attention の計算と推論を行う．本研究においても，Class Attention を導入する．

近年では，CNN と ViT の違いを明らかにするために画像に対して特定のノイズを付与し，認識精度に与える影響を調査する研究もある [7]．この調査の結果によると，CNN は高周波ノイズを付与した際に精度が低下し，ViT は低周波ノイズを付与した際に精度が低下することが明らかとなった．この調査結果は，CNN は画像に含まれる高周波成分に高い反応を示し，ViT は低周波成分に高い反応を示すことを示唆している．Deepfake 検出においては，画像を加工した境界線付近に発生する偽造痕跡となる情報を抽出することが重要であるため，高周波成分の情報を如何に捉えるかが重要であると考えられている．しかしながら，先の調査結果からも明らかなように ViT は CNN よりも高周波成分の特徴をよく抽出できないため，Deepfake 検出問題においては ViT は検出精度が劣ることが報告されている [4][5]．一方，高精度な画像認識を実現するために周波数成分の情報を利用した Wavelet Vision Transformer(Wave-ViT)[8] が提案されている．Wave-ViT は，ウェーブレット変換と ViT のアテンション機構を組み合わせた手法である．ViT が苦手とする画像の高周波成分を捉える機能を強化することで，高性能な画像認識を実現している．

2.3. 自己教師あり学習

自己教師あり学習は，学習用画像に対して Data Augmentation により生成した画像に元の画像と同じ教師情報を付与して学習する手法である．教師データ作成の人的コストの削減や，異なる見た目の画像の類似度を最大化するように学習するため頑健な特徴抽出が可能である．

自己教師あり学習の手法として Momentum Contrast(MoCo)[20] や SimCLR[21] が提案されている．例えば SimCLR は，図 1 に示すようにミニバッチ内の画像を拡張する Data Augmentation と

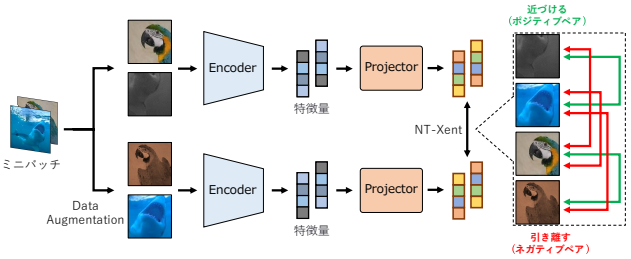

図 1 SimCLR の流れ．

特徴量を抽出する Encoder，特徴量を投影変換する Projector で構成される．同じ画像から得られた特徴量をポジティブペア，異なる画像から得られた特徴量をネガティブペアと呼ぶ．ポジティブペアの特徴量は類似度が大きくなるように，ネガティブペアの特徴量は類似度が小さくなるように学習することで Encoder の性能を向上させている．ネガティブペアを使用しない手法として Bootstrap Your Own Latent(BYOL)[22] や SimSiam[23] が提案されている．SimCLR はミニバッチ内の全データ間の類似度を計算するが，BYOL や SimSiam ではポジティブペアの類似度のみを最大化するように学習をするため計算量を削減できる．

2.4. 提案手法の概要

本研究では，ViT ベースのモデルに基づく Deepfake 検出の高精度化のために下記の 2 点について改善を施す．

- 周波数変換を導入した ViT の採用
  Deepfake 検出の精度向上のためには，偽造痕跡として表現される高周波成分の特徴を抽出することが重要である．しかし，ViT は CNN に比べて高周波成分の特徴を抽出する機能が弱いことが分かっている．そこで，本研究では画像に含まれる高周波成分の情報を抽出するためにウェーブレット変換を ViT に導入した Wavelet Vision Transformer(Wave-ViT)[8] を採用する．ウェーブレット変換を導入することで高周波成分の特徴を抽出する機能の向上が期待できる．
- 自己教師あり学習の導入
  画像の変動と偽造痕跡の違いを明確に区別するモデルを学習するために自己教師あり学習を導入する．Real 画像及び Deepfake 画像に対して Data Augmentaion により摂動を与え，自己教師あり学習によりモデルをファインチューニングする．Deepfake 画像としては，Self-Blended Images(SBIs)[16] により生成した擬似的な Deepfake 画像を用いる．

3. 提案手法

ウェーブレット変換を導入した Wavelet-ViT と自己教師あり学習を用いた Deepfake 検出法について述べる．

3.1. 提案手法の流れ

図 2 に提案手法による学習の流れを示す．提案手法は，図 2(a) に示す Deepfake 画像を生成する

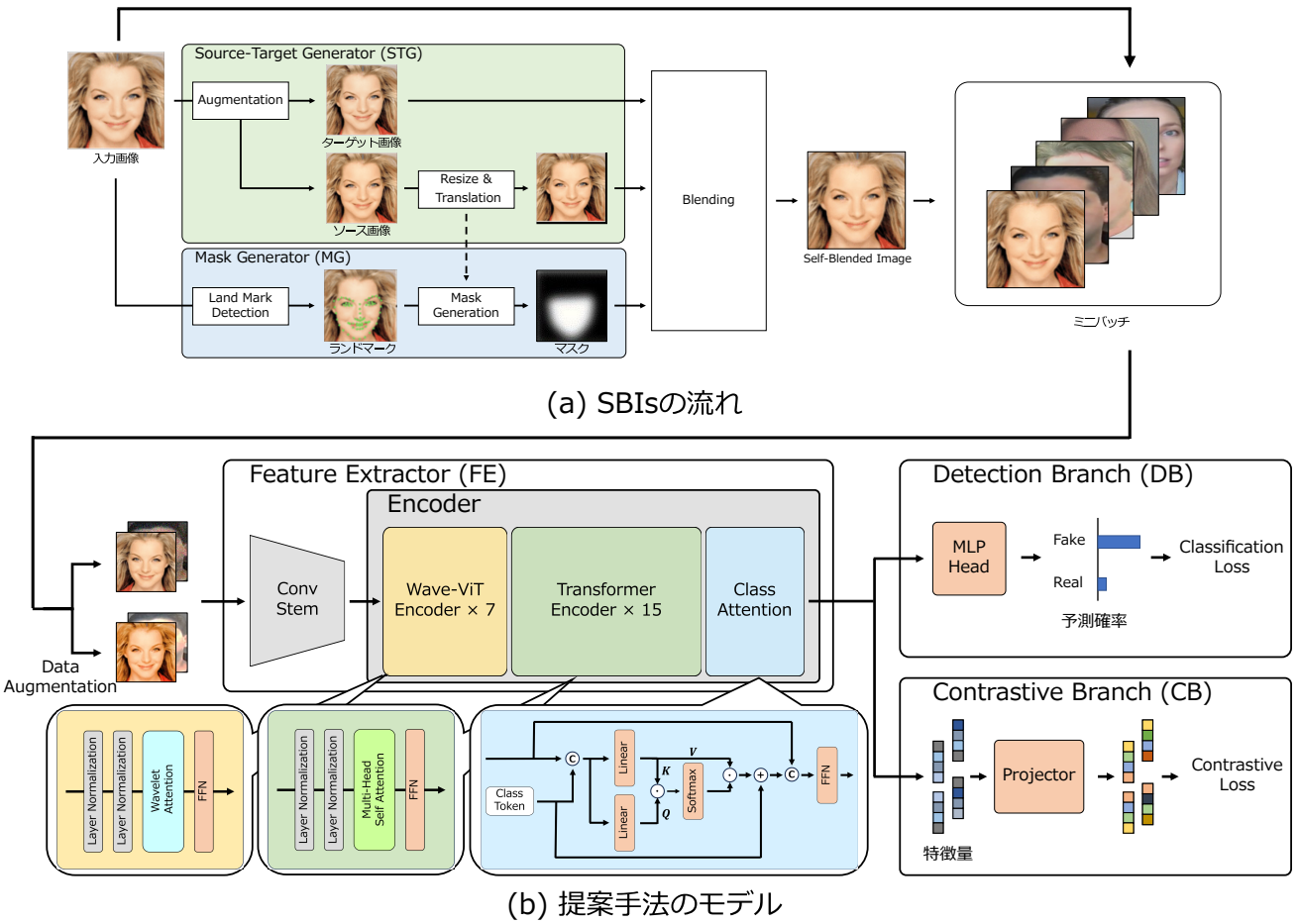

図 2　提案手法による学習の流れ．(a)SBIs による Deepfake 画像の生成の流れ，(b) 提案手法のモデルの構成．

Self-Blended Images(SBIs)[16] と図 2(b) に示す特徴抽出及び分類を行うモデルによって構成される．

SBIs は入力画像を 2 つに拡張する Source-Target Generator と拡張した画像を合成するためのマスクを生成する Mask Generator，2 つに拡張した画像をマスクを用いて疑似的な Deepfake 画像を合成する Blending の 3 つの処理で構成される．提案手法のモデルは SBIs で生成した画像を入力し，特徴抽出を行う Feature Extractor と Feature Extractor で得られた特徴量に基づいてクラス予測を行う Detection Branch，Data Augmentation により拡張した画像を用いた自己教師あり学習により Feature Extractor の性能を向上させる Contrastive Branch の 3 つの処理で構成される．クラス分類と自己教師あり学習を統一的に行うことで分類を意識した特徴抽出が可能となる．

3.2.　SBIs によるデータ生成

SBIs[16] により生成した擬似的な Deepfake 画像を学習に用いる．SBIs では，図 2(a) に示すように 3 つのステップで疑似的な Deepfake 画像を生成する．まず，入力画像を Source-Target Generator(STG) と Mask Generator(MG) に入力する．STG では，入力画像に Data Augmentation を施してソース画像とターゲット画像に拡張し，ソース画像に対して平行移動とサイズ変更を行う．MG では，入力画像からランドマークとなる顔の器官点を推定し，この結果を用いてソース画像とターゲット画像を合成するためのマスクを生成する．生成するマスクは，ターゲット画像に対してソース画像の合成領域を指定するものである．最後に，STG で得られたソース画像とターゲット画像，MG で生成したマスクを用いて疑似的な Deepfake 画像を生成する．

3.3.　モデル

特徴抽出を行う Feature Extractor(FE) と FE で得られた特徴量に基づいて分類を行う Detection Branch(DB)，FE の性能を向上させる Contrastive Branch(CB) の 3 つの処理で構成される．FE から得られた特徴量を DB と CB に入力し，それぞれから得られた損失によりモデルを最適化する．提案手法のモデルは，ViT の派生手法の 1 つである Wavelet Vision Transformer(Wave-ViT)[8] と特徴抽出の性能を向上させるために SimCLR[21] を組み合わせている．

3.3.1.　Wavelet Vision Transformer

Wave-ViT はウェーブレット変換と ViT のアテンション機構を組み合わせた手法である．Wave-ViT は入力画像をパッチに分割する ViT と異なり，入力画像を畳み込むことで埋め込みを行う Convolutional Stem(Conv Stem)[24] を採用している．Conv Stem を利用することで学習の安定化を図っ

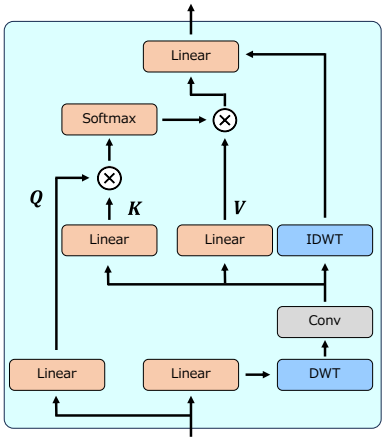

図 3 Wavelet Attention の流れ.

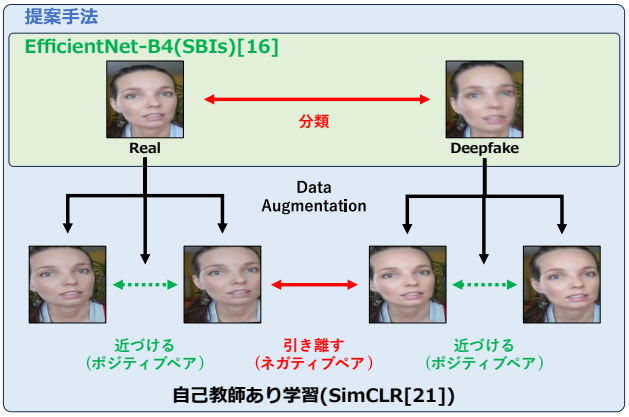

図 4 提案手法におけるペアの扱い方.

ている．そして，埋め込まれた特徴マップを平坦化して Encoder に入力する．その後，平坦化した特徴マップに対して離散ウェーブレット変換 (DWT：Dictate Wavelet Transform) を適用しアテンションを計算する．

図 3 に Wavelet Attention の流れを示す．Wavelet Attention では，ViT と同様に特徴量 Q(Query)，K(Key)，V(Value) を用いてアテンションを計算する．この時，元の画像から得られた特徴量である Q と DWT を施した特徴量である K，V を用いてアテンションを計算することで，元の画像から得られた特徴量と DWT を施した特徴量との対応関係を学習することが可能である．また，DWT を適用した特徴量に対して逆離散ウェーブレット変換 (IDWT：Inverse Dictate Wavelet Transform) を適用することで元の特徴量を再構成することが可能である．さらに，Wave-ViT では Class Attention[19] を採用し，画像認識に不要な情報や悪影響を与える情報が蓄積することを防止する．

FE では SBIs で生成した疑似的な Deepfake 画像を入力し，計算した特徴量を DB と CB に入力する．DB は FE で抽出した特徴量を基にクラスを予測する分類器である．CB は後述する自己教師あり学習により FE の特徴抽出機能を強化する役割を持つ．

3.3.2. SimCLR

画像の変動と偽造痕跡の違いを明確に区別するモデルを学習するために自己教師あり学習を導入する．Real 画像及び Deepfake 画像を Data Augmentation により摂動を与えてデータ拡張し，これらを区別するように学習する．CB には自己教師あり学習の 1 手法である SimCLR を使用する．図 4 に従来法及び SimCLR を適用した提案手法のイメージ図を示す．SBIs[16] では，Real 画像と Deepfake 画像を判別するためにモデルを学習する．一方，提案手法では Real 画像と Deepfake 画像に対して Data Augmentation を施し，拡張した画像を用いて自己教師あり学習によりモデルを学習する．これにより，画像の変動と偽造痕跡を分けやすいモデルを学習することが可能となる．

SimCLR は，ミニバッチ内の画像を拡張する Data Augmentation と特徴量を抽出する Encoder，特徴量を投影変換する Projector で構成される．まず，Real 画像と SBIs で生成した疑似的な Deepfake 画像で構成されているミニバッチ内の画像に対して，Data Augmentation を施して 2 つの見えの異なる画像に拡張する．次に，拡張した画像を Encoder を介して特徴量を抽出し，Projector により投影変換する．同じ画像から得られた特徴量をポジティブペア，異なる画像から得られた特徴量をネガティブペアと呼ぶ．ポジティブペアの場合は類似度が大きく，ネガティブペアの類似度が小さくなるように互いの特徴量を予測する pre-text タスクを解くことで Encoder の性能を強化する．

3.3.3. 損失関数

提案手法で用いる損失 $\mathcal{L}$ を式 (1) に示す．

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{con} \quad (1)$$

$$\mathcal{L}_{cls} = -(y \log(p) + (1-y)\log(1-p)) \quad (2)$$

$$\mathcal{L}_{con} = -\log \frac{\exp\left(\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)} \quad (3)$$

ここで，$\mathcal{L}_{cls}$ は DB におけるクラス分類の損失，$\mathcal{L}_{con}$ は CB における特徴量の類似度から求めた損失を表す．p は MLP Head から出力される予測確率，y は正解ラベル，$\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)$ はポジティブペアの類似度，$\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)$ はネガティブペアの類似度，τ は温度パラメータ，λ は各ブランチの損失のバランスをとる重みパラメータである．なお，特徴量の類似度はコサイン類似度により算出する．

4. 評価実験

提案手法の有効性を確認するために 2 つの評価実験を行う．1 つ目は，Deepfake 検出において一般的なデータセットである FaceForensics++(FF++)[13] を用いて評価する．本データセットは，顔の交換を行う Deepfakes[2]，FaceSwap[3]，表情を変更する Face2Face[12]，Neural Textures[11] の 4 つの Deepfake 手法で生成された画像で構成される．本実験により，加工の種類に対して有効的な手法を確認する．2 つ目は，提案手法の汎化性能を評価する．本実験では，

[2]https://github.com/deepfakes/faceswap
[3]https://github.com/MarekKowalski/FaceSwap

Deepfake 検出の評価において利用される FF++, Celeb-DF(CDF)[25], DeepFake Detection Challenge Preview(DFDCP)[26], FaceForensics in the Wild(FFIW)[27] の 4 種類のデータセットに対して評価する．

評価実験では EfficientNet-B4[28], ResNet-50[29], ViT[3], Wave-ViT[8] 及び提案手法の Deepfake 検出精度を Area Under Curve(AUC) により比較する．なお，Deepfake 検出問題において SBIs を提案した文献 [16] では幾つかのモデルに対して比較実験しており，EfficientNet-B4 が最も高い性能であることが報告されている．

4.1. 実験条件

提案手法は ImageNet で学習済みの Wave-ViT を使用し，FF++の Real 画像と FF++の Real 画像から SBIs によって生成した疑似的な Deepfake 画像を用いてモデルをファインチューニングする．ミニバッチ内の画像を拡張する Data Augmentation としては，色の変換やコントラストの調整，移動や回転，画像の圧縮，ノイズ付与やぼかしなど 13 種類を用いる．

また，入力画像のサイズは 224 × 224 画素，エポックは 100，バッチサイズは 32，最適化手法は AdamW，Weight Decay は 0.05，学習率を 5e-5 として Cosinedecay によりスケジューリングする．この際，WarmUp Epoch を 5 とする．

4.2. FF++に対する AUC の比較

表 1 に FF++に対する各手法の AUC を示す．表 1 より，EfficietnNet-B4 や ResNet-50 に比べて ViT は AUC の平均が 10%以上低下していることが分かる．Deepfake 検出問題において，ViT は CNN よりも性能が低いと報告している文献 [4][5] と同じ傾向であった．

一方で，EfficietnNet-B4，Wave-ViT，提案手法は同等の AUC であることが分かる．この結果より，ViT に周波数変換を導入した Wave-ViT を用いることで性能が改善したと考えられる．また，4 つの Deepfake 手法のうち，Neural Textures に対しては ViT が大きく性能を損ねたが，他の手法に関しては大きな傾向がないことを確認した．

4.3. 各手法の平均 AUC の比較

表 2 に 4 つのデータセットに対する各手法の AUC とその平均を示す．4.2 の実験結果と同様に，ViT は CNN ベースである EfficientNet-B4 や ResNet-50 よりも低い性能であることが確認できる．

一方で，Wave-ViT は EfficientNet-B4 とよりも平均 AUC が 2.31%低下したが，Wave-ViT に自己教師あり学習を導入した提案手法は EfficietnNet-B4 よりも平均 AUC が 1.70%向上した．これは，ウェーブレット変換を導入したことで特徴抽出の性能が向上したことに加え，自己教師あり学習を導入したことで画像上の変動と Deepfake の偽造痕跡を区別できるようになったためと考えられる．

4.4. アテンションマップの比較

Real 画像と Deepfake 画像を判別するためには，2 つの画像を合成した際の境界付近に表れる偽造痕跡に注目する必要がある．そこで，ViT と提案手法においてアテンションマップを可視化し，偽造痕跡を注視しているか確認する．ただし，提案手法は Class Attention を採用しているため 1 枚のアテンションマップで表現されるが，ViT は一般的に複数のアテンションマップで表現される．比較が困難であるため，ViT は Attention Rollout[30] で可視化した 1 枚のアテンションマップを示す．

図 5 に SBIs により生成した擬似的な Deepfake 画像及びマスク画像，そして ViT と提案手法において可視化したアテンションマップを示す．ViT のアテンションマップは，顔領域を注視しているもの多かったが，偽造痕跡が残っていると考えられるマスク画像の境界付近を注視しているものは少なかった．一方，提案手法は偽造痕跡が現れやすいマスク画像の境界線付近を注視しているものが多い．

5. おわりに

本研究では，SimCLR を導入した Wavelet Vision Transformer による Deepfake 検出手法を提案した．提案手法では，高周波成分の特徴を抽出できる Wavelet Vision Transformer を特徴抽出器に採用し，画像の変動と Deepfake の偽造痕跡の違いを明確に区別するために自己教師あり学習手法である SimCLR を導入した．評価実験より，提案手法は Deepfake 検出の精度において従来手法である EfficientNet-B4 と比較して平均 AUC が 1.70%向上したことを確認した．今後は，クラス情報を考慮した自己教師あり学習に拡張し，Deepfake 検出の精度向上を図る予定である．

参考文献

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection”, *Information Fusion*, vol. 64, pp. 131–148, 2020.

[2] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, “Robust deepfake on unrestricted media: Generation and detection”, *Frontiers in Fake Media Generation and Detection*, pp. 81–107, 2022.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale”, *International Conference on Learning Representations*, 2021.

[4] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, “M2tr: Multi-modal multi-scale transformers for deepfake detection”, *International Conference on Multimedia Retrieval*, pp. 615–623, 2022.

[5] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, “Deepfake detection algorithm based on improved vision transformer”, *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527, 2023.

[6] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues”, *European Conference on Computer Vision*, pp. 86–103, 2020.

表 1 FF++に対する各手法の AUC[%]．太字は各評価データにおける最高性能を示す．

手法	Deepfakes	Face2Face	FaceSwap	NeuralTextures	平均
EfficietnNet-B4[28]	99.99	99.88	**99.91**	**98.79**	**99.64**
ResNet-50[29]	99.92	98.69	97.42	94.68	97.68
ViT[3]	96.23	89.34	90.89	71.15	86.90
Wave-ViT[8]	**100.00**	**100.00**	99.40	98.52	99.48
提案手法	99.99	99.97	99.60	98.39	99.49

表 2 各評価データに対する各手法の AUC[%]．太字は各評価データにおける最高性能を示す．

手法	FF++	CDF	DFDCP	FFIW	平均
EfficientNet-B4[28]	**99.64**	93.18	86.15	84.83	90.95
ResNet-50[29]	97.68	90.09	84.71	80.35	88.20
ViT[3]	86.90	82.32	**88.50**	75.60	83.33
Wave-ViT[8]	99.48	87.66	79.56	**86.65**	88.34
提案手法	99.49	**97.49**	87.54	86.07	**92.65**

[7] N. Park, and S. Kim, "How do vision transformers work?" *International Conference on Learning Representations*, 2022.

[8] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-vit: Unifying wavelet and transformers for visual representation learning", *European Conference on Computer Vision*, pp. 328–345, 2022.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[10] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures", *Acm Transactions on Graphics*, vol. 38, no. 4, pp. 1–12, 2019.

[12] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos", *Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395, 2016.

[13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images", *International Conference on Computer Vision*, pp. 1–11, 2019.

[14] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection", *Conference on Computer Vision and Pattern Recognition*, pp. 5001–5010, 2020.

[15] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection", *International Conference on Computer Vision*, pp. 15 023–15 033, 2021.

[16] K. Shiohara, and T. Yamasaki, "Detecting deepfakes with self-blended images", *Conference on Computer Vision and Pattern Recognition*, pp. 18 720–18 729, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention", *International Conference on Machine Learning*, vol. 139, pp. 10 347–10 357, 2021.

[19] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers", *International Conference on Computer Vision*, pp. 32–42, 2021.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning", *Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", *International Conference on Machine Learning*, pp. 1597–1607, 2020.

[22] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning", *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[23] X. Chen, and K. He, "Exploring simple siamese representation learning", *Conference on Computer Vision and Pattern Recognition*, pp. 15 750–15 758, 2021.

[24] P. Wang, X. Wang, H. Luo, J. Zhou, Z. Zhou, F. Wang, H. Li, and R. Jin, "Scaled relu matters for training vision transformers", *Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2495–2503, 2022.

[25] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics", *Conference on Computer Vision and Pattern Recognition*, pp. 3207–3216, 2020.

[26] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset", *CoRR, abs/1910.08854*, 2020.

[27] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild", *Conference on Computer Vision and Pattern Recognition*, pp. 5778–5788, 2021.

[28] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks", *International conference on machine learning*, pp. 6105–6114, 2019.

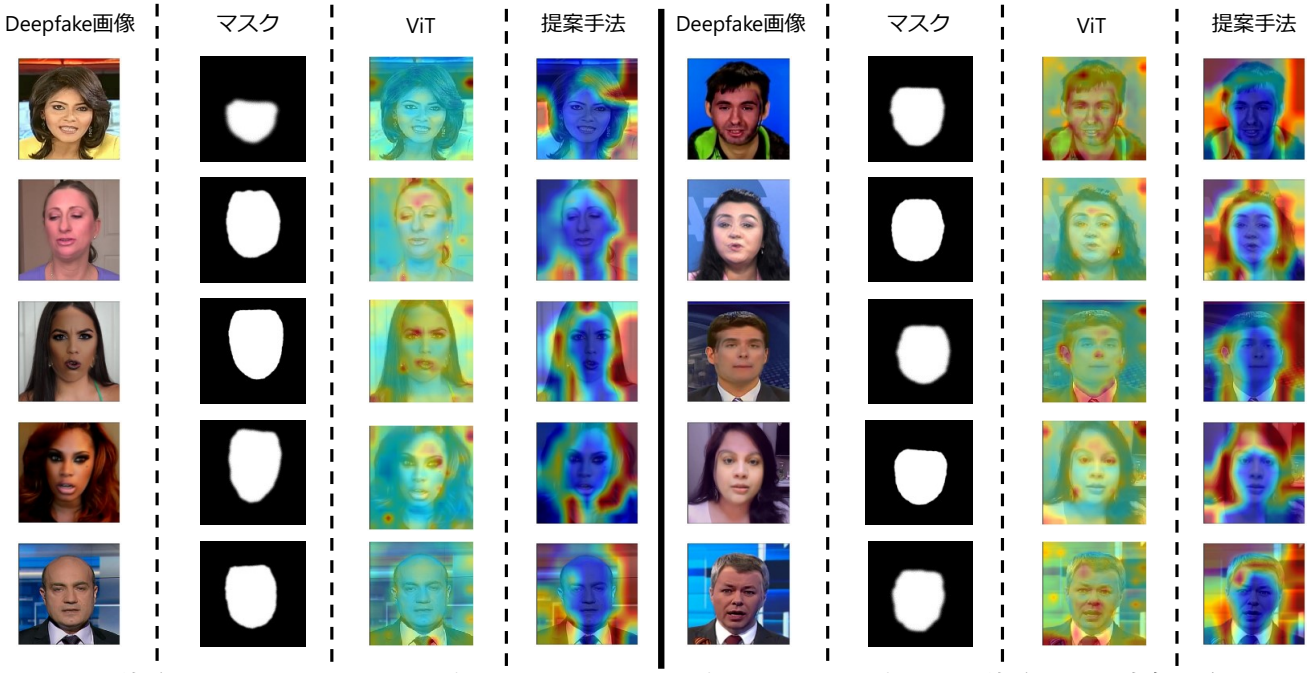

図 5 可視化したアテンションマップ．アテンションマップはヒートマップとして可視化され，赤色が強いほど注目度が高いことを表す．

[29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[30] S. Abnar, and W. Zuidema, “Quantifying attention flow in transformers”, *Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

高瀬 俊希：中部大学大学院工学研究科ロボット理工学専攻在学．現在，エッジコンピューティングのための圧縮画像認識に関する研究と深層学習を用いた Deepfake の検出の研究を進めている．

山内 悠嗣：2012 年中部大学大学院博士後期課程修了．同大学助手を経て 2018 年より講師．2010 年独立行政法人日本学術振興会特別研究員．画像認識，機械学習，知能ロボティクスの研究に従事. 博士 (工学)