ウェーブレット変換とABTのスタッキングによるDeepFake検出

猪ノ口 拓夢[1,a)]　青木 茂樹[1,b)]　宮本 貴朗[1]

概要：近年 Deepfake と呼ばれる，画像や動画の顔領域を操作し，加工する技術が急速に進歩しており，Deepfake を見つけ出すことは，人間にも難しくなりつつある．プライバシーの侵害やフェイクニュースの拡散といった Deepfake による脅威の拡大に伴い，検出技術の研究が注目されている．従来の Deepfake の検出に関する研究では，学習した画像と同一の Deepfake 生成手法で生成された画像は高精度に検出できるものの，異なる Deepfake 生成手法で生成された画像の検出精度が大幅に低下することが問題となっている．そこで本稿では，Attention Branch Transformer (ABT) と呼ばれる Vision Transformer に Attention Branch Network を導入したモデルとスタッキングを組み合わせることにより Deepfake を検出する手法を提案する．まず，疑似フェイク画像とウェーブレット変換適用後の画像をそれぞれ ABT で学習する．次に，それぞれの予測値を特徴量としてメタモデルで再学習することで最終的な予測結果を出力する．実験では，Celeb-DF データセット及び MesoNet データセットを用いて，提案手法の有効性を確認した．

キーワード：Deepfake 検出，Attention Branch Transformer，スタッキング，ウェーブレット変換

Deepfake Detection via Stacked Attention Branch Transformer with Wavelet-Transformed Images

Takumu Inoguchi[1,a)]　Shigeki Aoki[1,b)]　Takao Miyamoto[1]

Abstract: In recent years, technologies known as Deepfake, which manipulate and synthesize facial regions in images and videos, have rapidly advanced, making detection increasingly difficult even for humans. As threats such as privacy violations and the spread of fake news continue to grow, research into Deepfake detection has attracted significant attention. In previous studies, Deepfake images generated using the same method as the training data can be detected with high accuracy, but detection performance significantly degrades when the generation method differs.To address this issue, we propose a Deepfake detection method that combines the Attention Branch Transformer (ABT)—a Vision Transformer enhanced with an Attention Branch Network—with a stacking ensemble approach. First, pseudo-fake images and wavelet-transformed images are individually trained using ABT. Then, the prediction outputs from each model are used as features for a meta-model, which performs retraining to produce the final prediction.Experiments using the Celeb-DF and MesoNet datasets confirmed the effectiveness of the proposed method.

Keywords: Deepfake Detection, Attention Branch Transformer, Staking Ensemble, Wavelet Transform

1. はじめに

近年，Deepfake と呼ばれる画像や動画の顔領域を操作し，加工する技術が急速に進歩しており，DeepFake を見つけ出すことは，人間にも難しくなりつつある．また，Deepfake

[1] 大阪公立大学院情報学研究科
Graduate School of Informatics, Osaka Metropolitan University
[a)] sp25497r@st.omu.ac.jp
[b)] aoki@omu.ac.jp

が誰にでも作成可能になったことで，プライバシーの侵害やフェイクニュースの拡散などの被害件数が急増している．このような被害を低減するために，Deepfake を検出する必要性が増加している．

文献 [1] では，XceptionNet[2] を用いて, Deepfake を検出する手法を提案しており, 90 %以上の検出精度を達成している．文献 [3] では，CNN (Convolutional Neural Network) に Attention Branch を導入して出力される Attention Map を重み付けして利用する Attention Branch Network[4] により Deepfake を検出する手法を提案している．この手法では，Attention Map で重み付けすることによる認識精度の向上や，Deep Learning による認識結果の可視性の向上が図られている．また文献 [5] では，複数のモデルを用いて重み付きアンサンブル学習を行う手法を提案している．実験では，複数のモデルの予測値を統合することで全体的な予測精度が向上することを確認している．これらの Deepfake 検出に関する手法では，学習した画像と同一の Deepfake 生成手法で生成された画像は高精度に検出できるものの，異なる Deepfake 生成手法で生成された画像の検出精度が大幅に低下することが問題となっている [6].

この問題に対処するため文献 [7] では，1 枚の顔画像を 2 枚に複製し，2 枚の顔画像からフェイク画像を擬似的に作成する手法を提案している．そして，作成した疑似フェイク画像を識別器に学習させることで，未学習の Deepfake 画像の検出精度を向上させることに成功している．

近年注目されている Deepfake 検出手法として，PUDD (Prototype-based Unified Framework for Deepfake Detection) [8] が提案されている．この手法では，プロトタイプベースの類似度学習により，既知の特徴空間との距離を計測することで未知のフェイク画像を検出する手法であり，未学習の生成手法に対しても高い汎化性能を示している．また，再学習にかかる時間が非常に短く，環境負荷も低い点が特徴である．そして，実験では既存の多くの手法を上回る性能を示している．

本研究では，高精度な DeepFake 検出モデルを構築するために，Vision Transformer に Attention Branch Network を導入したモデルである Attention Branch Transformer (ABT) とスタッキング手法を組み合わせた手法を提案する．まず，学習データとして文献 [7] の手法に基づいて作成した疑似フェイク画像を生成する．また，学習データセットをウェーブレット変換する．次に，学習データと学習データをウェーブレット変換した結果のそれぞれについて ABT で学習してベースモデルとする．その後，ベースモデルの予測値を統合して最終的な分類を行うメタモデルで学習し，DeepFake を検出する．

以下，2 節で関連研究について述べ，3 節では提案手法について説明する．4 節では実験と考察について述べ，5 節でまとめと今後の課題を示す．

2. 関連研究

本研究に関する従来研究として，Attention Branch Network を利用して Deepfake を検知する手法 [4] と重み付きアンサンブル学習手法により Deepfake を検知する手法 [5], 疑似フェイク画像を利用し，Deepfake を検知する手法 [7] の概要について述べる．

文献 [4] では，Attention Branch Network を使用して，Deepfake 動画像を検出する手法を提案している．Attention Branch Network では, CNN を途中で分割し, 前半を特徴抽出部, 後半を予測結果を出力する Perception Branch とし，特徴抽出部の出力に Attention Map を出力する Attention Branch を追加している．Attention Branch から出力される Attention Map は画像中の重要な部分を強調する．そして，特徴抽出部の出力を Attention Map で重み付けして Perception Branch に入力して識別する．画像分類など様々な問題に使用される CNN では，ネットワークからの出力が何を根拠に決定されたのかが分からないという問題がある．Attention Branch Network を用いることにより，Attention Branch Network が重要だと判断した箇所を可視化するとともに，Attention Map で重み付けすることにより認識精度を向上させている．Attention Branch Network により，Deepfake の合成部分を重要な領域として重み付けすることで，CNN モデルを単体で用いるよりも，高い精度で Deepfake を検出することに成功している．

文献 [5] では，3 つの異なるニューラルネットワークアーキテクチャを利用した重み付きアンサンブル学習手法を提案している．3 種類の異なる深層学習モデル (SE-ResNet，Data-efficient Image Transformer (DeiT)，XceptionNet with Wavelet) をベースモデルとし，各モデルの予測を重み付けして合成している．複数のモデルの予測値を統合することで，個々のモデルを単独で動作させる場合より高い精度を実現している．SE-ResNet，DeiT，XceptionNet with Wavelet がそれぞれ Deepfake 画像の異なる特徴に焦点を当てることにより，各モデルの弱点を補い合い，Deepfake の検出精度の向上に成功している．

これらの手法では学習した Deepfake 生成手法で生成された画像は高精度に検出できるものの，異なる生成手法で生成された Deepfake 画像は検出できない問題があった．そこで文献 [7] では，同一人物の画像から擬似的な Fake 画像を作成する Self Blended Images (SBIs) を提案している．疑似フェイク画像に注目した従来手法では，ランドマークが類似する画像を 2 枚検索し，それらの画像を合成して Fake 画像を作成する手法が採用されているが，個人特徴の不一致に注目するように学習するために，最終的な検出精度が低下することが問題となっていた．文献 [7] の手法では，同一人物の 1 枚の画像を複製して合成することで，画像中の合成された領域のみを注目するように学習させるこ

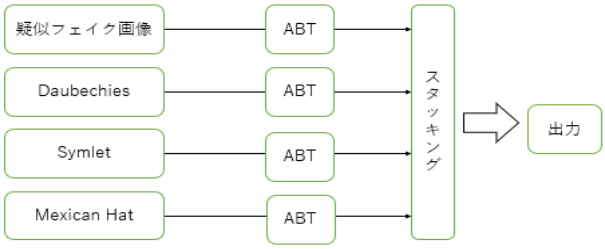

図 1 提案手法の概要
Fig. 1 Overview of the proposed Deepfake detection method.

とができ，未学習の Deepfake 画像に対する検出精度を向上させている．

3. 提案手法

提案手法の概要を図 1 に示す．まず，文献 [7] に基づいて疑似フェイク画像のデータセットを作成し，さらに作成した疑似フェイク画像に対して 3 種類のウェーブレット変換 (Daubechies Wavelet，Symlet Wavelet，Mexican Hat Wavelet) を適用したデータセットを用意する．次に，疑似フェイク画像データセット，3 種類のウェーブレット変換ごとのデータセットをそれぞれ ABT で学習し，学習した 4 種類の ABT をベースモデルとする．そして，各ベースモデルの予測結果を統合してメタモデルに入力し，最終的な予測結果を出力する．スタッキング手法では，個々のベースモデルの長所を組み合わせたメタモデルで予測することで，多様なデータ特性に対応し，汎化性能を向上させることができる．

3.1 疑似フェイク画像データセットの作成

疑似フェイク画像の生成手順の概要を図 2 に示す．まず，1 枚の顔画像を Source 画像 I_s, Target 画像 I_t に複製する．また，顔のランドマークの凸包であるマスク画像 Att_{MA} を作成する．ここで，顔のランドマークは，画像処理ライブラリの Dlib[9] で取得した 68 個の特徴点を基に抽出している．次に，確率 0.5 で I_s と I_t のどちらかを選択し，色変換と周波数変換を行う．そして, I_s に対してランダムなリサイズと平行移動処理を適用する．また, Att_{MA} に対しても同じパラメータでのリサイズと平行移動処理を適用する．更に，マスク領域全体に [0.25, 0.5, 0.75, 1, 1, 1] の中からランダムに選択した定数を掛けることでブレンド比率を決定する．その後，次式に従い，疑似フェイク画像 I_{SBI} を作成する．

$$I_{SBI} = I_t \odot Att_{MA} + I_s \odot (1 - Att_{MA})$$

また，疑似フェイク画像の負例として与えるリアル画像としては，疑似フェイク画像作成に使用した元画像を使用する．図 3 に元画像と元画像から生成した疑似フェイク画像の例を示す．

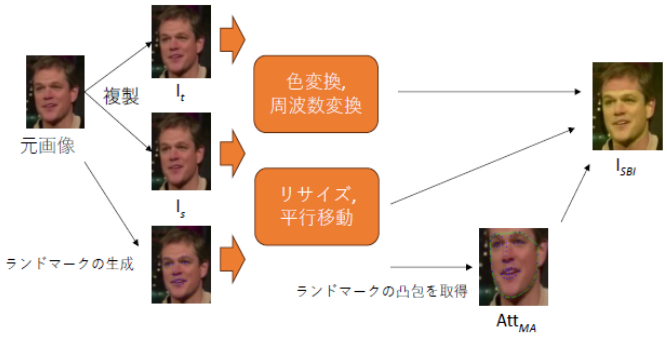

図 2 疑似フェイク画像の生成手順の概要
Fig. 2 Generation process of Self-Blended Images.

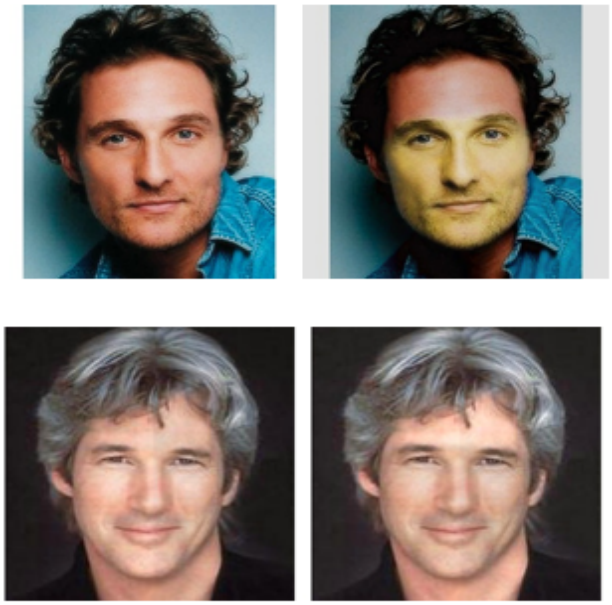

図 3 元画像 [10] と疑似フェイク画像の例
Fig. 3 Example of a Self-Blended Image (left: original, right: pseudo-fake).

3.2 データ拡張

オーバーフッティングを緩和し，モデルの汎化性能を向上させるために，学習データに対して以下の 3 つの手法を適用してデータを拡張する．用いたデータ拡張手法は，ランダム水平反転，ランダム回転，カラージッターである．ランダム水平反転は，50 %の確率で画像を垂直軸に沿って反転させる．ランダム回転は，画像を-10 度 ~10 度の範囲でランダムに回転させる．カラージッターでは，指定した範囲内で画像の明るさ，コントラスト，彩度，色相をランダムに変化させる．明るさ，コントラスト，彩度は-20 % ~20 %，色相は-10 % ~10 %の範囲で変化するように指定している．

3.3 ウェーブレット変換

ウェーブレット変換は，周波数解析手法の一つで，画像のエッジや，テクスチャ，微細なパターンを特定する際に有効な手法である．本研究では，異なる周波数解析特徴をもつ以下の 3 つのウェーブレット変換を使用する．ここで，データ拡張した画像をウェーブレット変換するとデータ拡張手法の影響を ABT が学習する可能性が考えられるため，ウェーブレット変換を適用するデータにはデータ拡張した画像を含んでいない．

3.3.1 Daubechies Wavelet

Daubechies ウェーブレット変換は，エッジ検出能力に優れており，高周波のノイズ解析に用いられる．画像における境界や特徴点の不整合の検出に適している．本研究では，合成領域の境界の効果的な検出を想定している．

3.3.2 Symlet Wavelet

Symlet ウェーブレット変換は，顔の特徴点間における位相関係の解析に優れており，微細な変化の検出に用いられる．画像における境界の不自然さの検出に適している．本研究では，Deepfake の生成プロセスで発生する微細な不整合の検出を想定している．

3.3.3 Mexican Hat Wavelet

Mexican Hat ウェーブレット変換は，画像における特徴の境界異常の強調に用いられる．本研究では，局所的な異常や急激な変化の強調による顔画像の境界異常の検出を想定している．

3.4 Attention Branch Transformer

3.2 でデータ拡張した疑似フェイク画像と，疑似フェイク画像を 3.3 で述べた 3 種類のウェーブレット変換で変換した結果をそれぞれ ABT で学習して Deepfake を検出する．図 4 に ABT の概略図を示す．本研究では，ViT (Vision Transformer) を改良したモデルである DeiT (Data-efficient Image Transformer) に Attention Branch Network を導入したモデルを使用する．DeiT は，事前学習済みの CNN モデルからの学習を使用することで迅速かつ効率的な学習を実現する．蒸留トークンと呼ばれる教師モデル (CNN モデル) からの情報を学習に活用し，最終的な分類結果に教師モデルの知識が反映されるように訓練している．DeiT はデータと計算資源が限られている環境でも Transformer を画像分類に使えることが大きな利点であり，少ないデータ量と学習量であっても学習できる．

ImageNet で事前学習済みの DeiT をファインチューニングしたモデルを特徴抽出に使用し，特徴抽出した出力を Attention Branch と Perception Branch に分岐する．Attention Branch では，画像内のどの領域が重要であるかを特定する Attention Map を生成する．Attention Branch で出力した Attention Map の重みを，DeiT で特徴抽出した出力に掛け合わせて特徴マップを強調し，Perception Branch に入力する．損失関数として，以下に示す各 Branch の学習損失の和 L を用いる．ここで，L_{att} は Attention Branch の学習損失，L_{per} は Perception Branch の学習損失を表す．

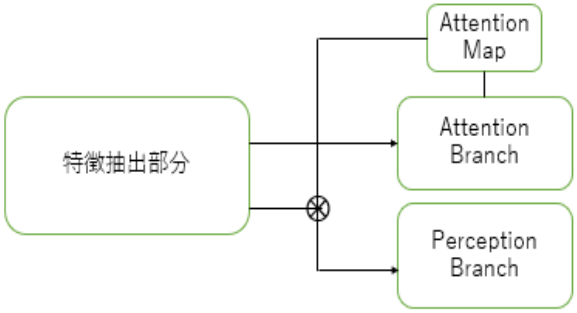

図 4 Attention Branch Transformer の概略図
Fig. 4 Structure of the Attention Branch Transformer.

$$L = L_{att} + L_{per}$$

3.5 スタッキング

スタッキング手法は，複数のモデルの予測結果を組み合わせてより高精度な予測を行うアンサンブル学習手法の一種である．本手法では，疑似フェイク画像，Daubechies ウェーブレット変換, Symlet ウェーブレット変換, Mexican Hat ウェーブレット変換の 4 つの ABT の予測値をメタモデルに統合する構成としている．メタモデルにはニューラルネットワークを使用し，ニューラルネットワークは，入力層，2 つの隠れ層，出力層という構造で構成しており，2 つの隠れ層では過学習を抑制するため，Dropout を使用する．入力層では，ABT からの出力である各ベースモデルからの予測値を特徴として使用する．出力層の活性化関数にはシグモイド関数を使用し，0.0〜1.0 の範囲で Deepfake である確率を出力する．

4. 実験

4.1 実験条件

4.1.1 実験データセット

本手法の有効性を確認するためにリアル画像データセットと 2 つの Deepfake データセットを用いて実験した．疑似フェイク画像の作成には，SCUT-FBP V2[10] データセットを使用した．SCUT-FBP V2 データセットはリアル画像のデータセットであり，5500 件の画像を収録している．5500 件の内，顔のランドマークを抽出できた 5484 件を疑似フェイク画像の作成に使用している．

Celeb-DF[11] データセットは 1 種類の Deepfake 生成手法で構成された Deepfake 動画のデータセットであり，不自然な特徴の少ないリアルなフェイク動画を収録している．Celeb-DF データセット内の 5939 本のフェイク動画と

表 1　実験結果

Table 1　Detection accuracy for each model on Celeb-DF and MesoNet.

手法	Celeb-DF	MesoNet
提案手法	87.03	**62.04**
ABT のみ	**90.62**	46.53
Daubechies+ABT	89.53	47.25
Symlet+ABT	74.28	53.66
Mexican Hat+ABT	**90.62**	46.34

590 本のリアル動画からランダムに 1 フレーム抽出した画像データを使用した．MesoNet[6] データセットは 2 種類の Deepfake 生成手法で構成された Deepfake 画像のデータセットであり，5,103 件のフェイク画像と 4,259 件のリアル画像を使用した．

4.1.2　学習と評価指標

各ベースモデルは，エポック数 10，バッチサイズ 24，学習率 0.0005，損失関数はクロスエントロピー誤差，最適化アルゴリズムは Adam を使用して学習した．また，メタモデルは，エポック数 30，バッチサイズ 24，学習率 0.0005，損失関数はクロスエントロピー誤差，最適化アルゴリズムは Adam を使用して学習した．評価指標には Accuracy を使用した．

4.2　実験結果

表 1 に，各ベースモデルおよび提案手法の検出精度 (Accuracy) を示す．Celeb-DF データセットでは，ABT 単体及び Mexican Hat + ABT が最も高い精度 (90.62 %) の結果が得られたが，スタッキングによる統合結果は 87.03 % とやや低下した. 一方，MesoNet データセットにおいては，スタッキングによって 62.04 %の精度を達成し，全てのベースモデル単体よりも高い結果となった．

図 5，図 6 に各ベースモデルにおける Attention Map の例を示す．赤色に示された領域が Attention Map の注目領域となっている．図 5 に示す Celeb-DF データセットの結果では，ウェーブレット変換の種類によって注目領域が異なっており，顔の輪郭や目元，口元などに注目が集まっていることを確認できる. 図 6 に示す MesoNet データセットの結果では，より明瞭に合成領域や境界に注目している様子が見られ，ウェーブレット変換による特徴強調が有効に働いていることを確認できる．

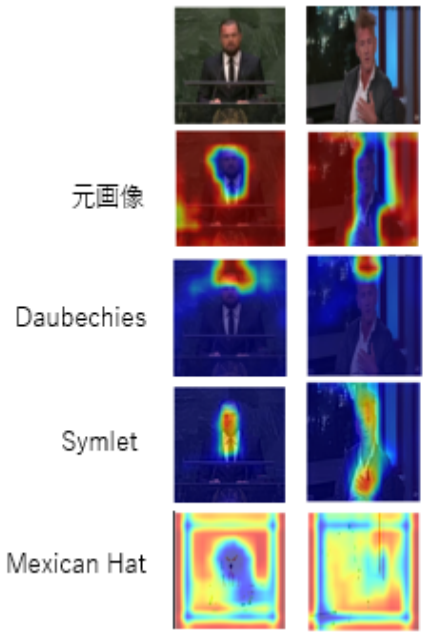

図 5　各ベースモデルごとの Celeb-DF の Attention Map

Fig. 5　Attention maps for each base model on Celeb-DF.

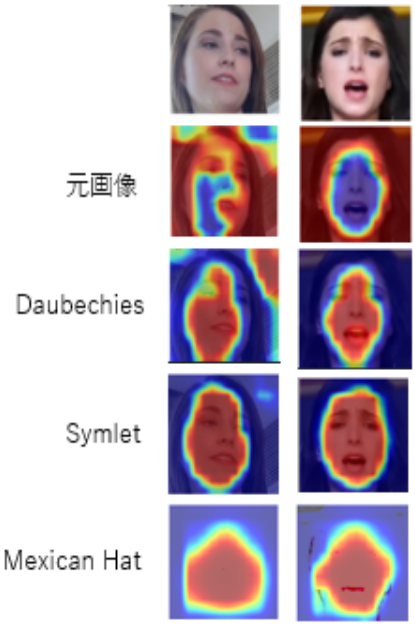

図 6　各ベースモデルごとの MesoNet の Attention Map

Fig. 6　Attention maps for each base model on MesoNet.

4.3　考察

MesoNet データセットにおいては，複数のウェーブレット変換を適用したデータセットを ABT で学習し，その出力をスタッキングすることで，Deepfake の検出精度が向上した. これは，MesoNet データセットのフェイク画像が比較的単純な合成手法で生成されており，境界の不整合や微細な異常が明瞭に現れるため，ウェーブレット変換による特徴強調が有効に機能したためであると考えられる．図 6 から確認できるように，各ベースモデルが異なる領域に注目しており，スタッキングによってそれらの特徴が補完的に統合されたことが精度向上に寄与したと推察される．

一方，Celeb-DF データセットにおいては，ABT 単体の方がスタッキングよりも高い精度を示した．これは，Celeb-DF データセットのフェイク動画が高品質であり，

合成による不自然さが少ないため，ウェーブレット変換によって強調されるべき特徴が十分に抽出されなかった可能性が考えられる．また，Symlet + ABT のように精度の低いベースモデルがスタッキングに含まれることで，全体の性能が希釈されたことや，図 5 に示すように，注目領域が分散しており，統合による効果が限定的であったことも要因であると考えられる．

文献 [8] で提案されている PUDD は，プロトタイプベースの類似度学習により，既知の特徴空間との距離を計測することで未知のフェイク画像を検出する手法であり，Celeb-DF データセットにおいて 95.1 %の高い検出精度を達成している．また，PUDD は再学習にかかる時間が非常に短く，環境負荷も低いという利点を持つ．本手法は Attention Map による視覚的な解釈性や，ウェーブレット変換による周波数領域の特徴強調により，既知の特徴を多角的に捉えることが可能であるが，現在用いているベースモデルだけでは未知の生成手法に十分に対応できているとは言い難い．特に，Self-Blended Images (SBIs) を用いた学習だけでは，実際の Deepfake 画像との分布の違い (ドメインギャップ) を生じさせている可能性があるために，汎化性能で PUDD に劣る結果となったと考えられる．

以上の結果から，スタッキングに用いるベースモデルの選定を見直し，精度の低いモデルを除外することで統合結果の精度低下を防ぐこと，メタモデルの構造を最適化し，ニューラルネットワーク以外の識別器の導入や，Attention Map に基づく特徴選択を行うことで統合性能の向上を図ること，さらに学習データとしてより実際の Deepfake に近い擬似フェイク画像を生成する手法を検討し，ドメインギャップの緩和を図ることなどが本手法の改善策として考えられる．さらに今後，PUDD のようなプロトタイプベースの類似度学習手法を本手法に部分的に導入する手法を検討することで，未知の Deepfake 生成手法への対応能力を補完することなどを検討したいと考えている．

5. まとめ

本稿では，Attention Branch Transformer (ABT) とウェーブレット変換を組み合わせたベースモデルを構築し，それらをスタッキングすることで Deepfake 検出精度の向上を図る手法を提案した．実験では，MesoNet データセット及び Celeb-DF データセットの 2 つのデータセットを用いて検証を行った．MesoNet データセットにおいては，スタッキングによって各ベースモデルの特徴が補完され，検出精度の向上が確認できた．一方，Celeb-DF データセットにおいては，ABT 単体の方が高い精度を示し，スタッキングによる効果を十分に確認できなかった．Attention Map の分析から，各ウェーブレット変換が異なる領域に注目していることが確認され，スタッキングによる多様な特徴の統合が有効であることを確認した．しかしながら，高品質なフェイク画像に対しては，既存の特徴抽出手法では十分な精度での検出が難しかった．

今後の課題としては，スタッキングに用いるベースモデルの選定方法やメタモデルの構造の最適化，さらに，実データに近い学習画像の生成手法の導入，およびプロトタイプベース手法との融合方法の検討などが挙げられる．

参考文献

[1] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images, Proc. International Conference on Computer Vision(ICCV), (2019).

[2] Chollet, F.: Xception: Deep learning with depthwise separable convolutions, Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference(CVPR), pp.1800–1807 (2017).

[3] 小野 尚記, 史 又華: Attention Mask によるディープフェイク動画像の検出, 人工知能学会全国大会論文集, (2023).

[4] Fukui, H., Hirakawa, T., Yamashita, T. and Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation, Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference(CVPR), pp.10705–10714 (2019).

[5] Kafi, A., Anindya, B., Ashir, I., Kaidul, I., Abrar, A.F., Utsab, S., Hafiz, I.:Hybrid Deepfake Image Detection: A Comprehensive Dataset-Driven Approach Integrating Convolutional and Attention Mechanisms with Frequency Domain Features, arXiv preprint (2025).

[6] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I.: MesoNet: a Compact Facial Video Forgery Detection Network, In IEEE Workshop on Information Forensics and Security(WIFS), (2018).

[7] Shiohara, K. and Yamasaki, T.: Detecting Deepfakes with Self-Blended Images, Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference(CVPR), pp.18720–18729 (2022).

[8] Pellicer, A. L., Li, Y., and Angelov, P.: PUDD: Towards Robust Multi-modal Prototype-based Deepfake Detection, Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference(CVPR), pp.3809-3817 (2024).

[9] Kazemi, V. and Sulivan, J.: One Millisecond Face Alignment with an Ensemble of Regression Trees, Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference(CVPR), pp. 3460–3467 (2014)

[10] Liang, L., Lin, L., Jin, L., Xie, D., and Li, M. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction, International Conference on Pattern Recognition (ICPR), pp. 1598-1603 (2018).

[11] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S.: Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics, Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference(CVPR), pp. 5888–5897 (2020).