社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

敵対的ランドマーク撹乱の転移性に関する研究

楊　　力懿†　菊池　浩明††

† 明治大学大学院先端数理科学研究科　〒164–0001 東京都中野区中野 4–21–1
†† 明治大学総合数理学部　〒164–0001 東京都中野区中野 4–21–1
E-mail: †{cs242039,kikn}@meiji.ac.jp

あらまし　近年，ディープフェイク技術の悪用によるプライバシー侵害や偽情報の拡散などのリスクが深刻な問題となっている．従来のディープフェイクの検出技術は，標的としている学習アルゴリズムに依存し，他の学習モデルに対して有効でなかった．そこで，本研究では，標的とする学習モデルに依存する敵対的ランドマーク撹乱を他のモデルでも有効性を維持する転移性を向上した防御方式の提案と評価を行う．
キーワード　ディープフェイク，敵対的摂動

Study on the Transferability of Adversarial Landmark Perturbations

Yang LIYI† and Kikuchi HIROAKI††

† Graduate School of Advanced Mathematical Sciences, Meiji University　4–21–1 Nakano, Nakano-ku, Tokyo, 164–0001 Japan
†† School of Interdisciplinary Mathematical Sciences, Meiji University　4–21–1 Nakano, Nakano-ku, Tokyo, 164–0001 Japan
E-mail: †{cs242039,kikn}@meiji.ac.jp

Abstract　In recent years, the malicious use of deepfake technologies has led to serious problems, including violations of privacy and the spread of misinformation. Conventional deepfake detection methods often depend on the specific learning algorithms they target and therefore fail to remain effective against other models. In this study, we propose and evaluate the transferability of adversarial landmark perturbations that are designed for a target model while maintaining their effectiveness across different models.
Key words　Deepfake，Adversarial Perturbation

1. はじめに

近年，生成 AI 技術の急速な発展により，人間の顔画像や音声を高精度に合成・編集することが可能となり，エンターテインメント，教育，コンテンツ制作など多様な分野での活用が進んでいる．特に，ディープラーニングを用いた画像生成技術の高度化により，実在する人物の顔画像を基に，表情，姿勢，発話内容を自在に改変することが可能となった [1]．一方で，このような技術の悪用により，特定の人物になりすました偽の画像や映像を生成するディープフェイク（Deepfake）が深刻な社会問題となっている．NHK が 2024 年 9 月に報じた調査 [2] によれば，アメリカや韓国，日本において，卒業アルバムや日常写真を基に生成された性的ディープフェイク画像が拡散し，未成年を含む多数の一般人が深刻な被害を受けている事実が明らかとなり，その被害の深刻さが浮き彫りになった．

こうした状況を背景に，ディープフェイクの対策技術に関する研究が活発化している．Hsu らは機械学習を活用したフェイク画像検出手法 [3] を提示している．しかし，攻撃側もその対策を行い，イタチごっこが続いている．例えば，Hussain らは，攻撃者側も検出を回避する新たな生成技術 [4] を取り入れている．こうして，ディープフェイクは年々検出が難しくなっている．さらに，一度拡散した有害コンテンツは後から完全に削除・無効化することが極めて困難であり，被害をより深刻にしている．

ディープフェイク生成の過程では，顔のランドマーク情報が重要な役割を担っており，多くの手法は抽出した特徴点を基に顔の変換や合成を実行している．したがって，ランドマーク抽出を妨害することは生成精度を低下させる有効な対策となり得る．我々は [5] にて，目鼻口などのランドマークに重みを付与し，その抽出を阻害する敵対的サンプル生成手法を提案した．これは単一モデルの内部アルゴリズムや勾配情報に依存しており，他のモデルへの転移性が足りないという点が課題であった．

そこで，本研究では，複数の顔ランドマーク抽出モデルの内

部構造や勾配情報にアクセス可能である状況を仮定し，特定モデルに対して求めた撹乱を他のモデルにおいても有効とする転移性を強化することを試みる．そのために，複数モデルに対して同時に有効な敵対的ランドマーク撹乱手法 Unbalanced Landmark Breaker Plus (ULB+) を提案する．本手法は単一モデルに最適化された撹乱に比べ，異なるモデル間でも効果的にディープフェイク生成の品質を低下させることを目的とする．

本研究では，以下の 3 つの研究課題（RQ）に基づいて，敵対的ランドマーク撹乱の転移性について検証を行う：

- RQ1: 単一のランドマーク抽出器に対して生成した敵対的ランドマーク撹乱は，他のランドマーク抽出器に対しても有効に作用するか？
- RQ2: 複数のランドマーク抽出器を考慮して設計した提案敵対的ランドマーク撹乱は，単一モデルに最適化された撹乱と比較して，ディープフェイク生成の妨害効果を維持できるか？
- RQ3: 転移性を考慮した提案敵対的ランドマーク撹乱は，視覚的自然さを保持しつつ，複数モデルに対して実用的な防御性能を発揮できるか？

2. 基本定義

2.1 ディープフェイク

ディープフェイクは，顔の入れ替えなど [6] を通じて合成された偽写真, 偽動画, 及び, それらを生成するための技術の総称である．

まず，入力となる動画から顔検出器によって対象人物の顔領域が検出され，続いて顔のランドマークを抽出する．次に，これらのランドマークを用いて顔領域を標準的な形状に整列（アライメント）し，整列後の顔画像を切り出してディープフェイクモデル [8] に入力する．このモデルはオートエンコーダ構造に基づいており，畳み込みニューラルネットワーク（CNN）[7] からなるエンコーダとデコーダで構成されている．エンコーダは表情や顔の向きなどを保持しつつ，個人識別に関わる特徴を除去し，デコーダはその特徴から提供者の顔を再構成する．

合成された顔は，ランドマークに基づいて元のフレームに再配置され，マスク処理によって自然に馴染ませることで完成する．このように，ディープフェイクにおいては，顔のランドマーク抽出が整列処理やマスク生成などの工程において中核的な役割を担っており，その精度が生成結果の品質に直結する．

2.2 ランドマーク抽出器

顔ランドマーク抽出器 [9] は，顔画像において目，鼻，眉，口，顎の輪郭などの特徴的な点（キーポイント）を高精度に検出するための技術である．初期のランドマーク抽出器は，機械学習に基づく比較的単純な手法が主流であり，たとえば Dlib [10] に実装されているような回帰木のアンサンブルなどが広く用いられていた．

近年，ディープラーニングの発展に伴い，畳み込みニューラルネットワークに基づく顔ランドマーク抽出手法が登場し，高精度かつロバストなランドマーク推定を実現した．現在主流となっている CNN ベースの顔ランドマーク抽出器の多くは，二段階の処理構造を採用している．第 1 段階では，各ランドマークの空間的な存在確率を示す 2 次元テンソルが生成される．第 2 段階では，これらの確率マップ内のピーク位置に基づいて，最終的なランドマーク座標が決定される．本研究では，このような高精度な CNN ベースの顔ランドマーク抽出器 High-Resolution Network（HRNet）[13] と Face Alignment Network（FAN）[12] を摂動対象とする．

2.3 敵対的摂動

CNN は，高い性能を示す一方で，敵対的摂動（Adversarial Perturbation）と呼ばれる意図的に設計された微小なノイズに対して脆弱であることが知られている．Goodfellow ら [14] は，CNN が敵対的摂動と呼ばれる意図的に設計された微小なノイズに対して脆弱であることを示した．敵対的摂動は人間には知覚できないほどの微細なノイズであるにもかかわらず，画像分類器や物体検出器など，さまざまな CNN ベースのモデルに対して誤分類や誤検出を引き起こすことが知られている．

敵対的摂動は，攻撃者が利用可能な情報の範囲に応じて，ホワイトボックス方式およびブラックボックス方式の二つの脅威モデルに分類される．ホワイトボックス方式とは，攻撃者が対象となるモデルの構造，学習済みパラメーター，および勾配情報に完全にアクセスできる状況を想定した攻撃方式である．ホワイトボックス方式に基づく代表的な敵対的攻撃手法としては，Fast Gradient Sign Method (FGSM) [14]，Basic Iterative Method (BIM) [15]，および Projected Gradient Descent (PGD) [16] などが知られている．FGSM は，損失関数の勾配に基づいて一度だけ摂動を付加する単一ステップ攻撃であり，計算コストが低い点が特徴である．一方，BIM や PGD は，FGSM を複数回反復的に適用することで摂動を逐次更新する反復型攻撃に分類される．ブラックボックス方式とは，攻撃者が対象モデルの内部構造やパラメータにアクセスできず，モデルの出力結果のみを利用可能とする状況を想定した攻撃方式である．代表的な手法としては，Zeroth-Order Optimization (ZOO) [17] のように，数値的に勾配を近似することで敵対的摂動を生成する手法が提案されている．

2.4 Unbalanced Landmark Breaker(ULB)

ディープフェイク生成においては，顔認識や顔ランドマーク抽出などの CNN ベースの解析モデルが前処理や中間処理として用いられる場合が多い．そのため，これらのモデルに対して敵対的摂動を付加することで，生成過程全体に影響を及ぼし，結果としてディープフェイク生成の品質を低下させることが可能であると考えられる．我々らは, 顔ランドマーク抽出に対する敵対的攻撃 Unbalanced Landmark Breaker [5] を提案した．本手法では，CNN ベースのランドマーク抽出器に対し，入力画像に敵対的ノイズを付加することで，ランドマークの推定位置に誤差を生じさせることを目的としている．予測された各ランドマークの座標確率分布と元画像に対応する分布のコサイン類似度の総和を損失関数として定義する．この損失を最小化する摂動を画像に加えることで，意図的に誤ったランドマーク出力を誘発する．

Unbalanced Landmark Breaker は，部位別のランドマークに対

して影響を与えることができるという利点を持つ一方で，単一モデルに依存している．

3. 提案方式

本研究では，ホワイトボックス型の敵対的サンプル生成手法 Unbalanced Landmark Breaker Plus(ULB+) を提案する．本手法は，複数のランドマーク抽出器に誤りを引き起こし，ディープフェイク生成器における顔合成を防止することを目的とする．

3.1 損失関数

本研究は敵対的画像の最適化するため，ランドマークベクトルのコサイン類似度を計算する．入力画像を $\boldsymbol{x}$，摂動を加えた敵対的画像を $\boldsymbol{x}^{\text{pert}}$，ランドマーク抽出器を $\mathcal{F}$ とする．$\mathcal{F}(\boldsymbol{x}) = (h_1, \ldots, h_k)$ を元画像に対応する k 個のランドマークベクトル，$\mathcal{F}(\boldsymbol{x}^{\text{pert}}) = (\hat{h}_1, \ldots, \hat{h}_k)$ を対応する敵対的画像のランドマークベクトルとする．ここで，各 $\hat{h}_i$ は，ランドマーク i に対応する 64×64 のヒートマップを平坦化したベクトルである．$\boldsymbol{w} = (w_1, \ldots, w_k)$, $w_i \in [0, 1]$ は第 i ランドマークに対応する重み係数である．このとき，ULB [5] の損失関数は，

$$\mathcal{L}(\mathcal{F}(\boldsymbol{x}), \mathcal{F}(\boldsymbol{x}_{\text{pert}})) = \sum_{i=1}^{k} w_i \frac{h_i^\top \hat{h}_i}{\|h_i\| \|\hat{h}_i\|} \tag{1}$$

と定められていた．ここで，$\boldsymbol{w} = (w_1, \ldots, w_k)$, w_i は第 i ランドマークに対応する重み係数であり，部位ごとの影響度を与える．w が大きいほどランドマークの出力に大きな攪乱を与える．

敵対的摂動が画像全体の画質に与える影響を制限するため，敵対画像 $\boldsymbol{x}^{\text{pert}}$ に対してノルム制約

$$\|\boldsymbol{x}^{\text{pert}} - \boldsymbol{x}\|_p \leq \epsilon \tag{2}$$

を定める．ここで，ϵ は摂動の大きさを示す定数であり，p はノルムを示す．本研究では主に ℓ_∞ ノルムに基づいてピクセル単位の変化量を制限することにより，人間の視覚で違和感が生じないレベルにノイズを抑制している．

3.2 HRNet と FAN の同時撹乱

本研究では，単一のランドマーク抽出器に対する攻撃だけでなく，異なるアーキテクチャを持つ 2 つの抽出器（HRNet と FAN）の内部にアクセスできることを仮定する．HRNet および FAN をそれぞれ $\mathcal{F}_{\text{HR}}$，$\mathcal{F}_{\text{FAN}}$ とし，$\mathcal{F}_{\text{HRNet}}(\boldsymbol{x}^{\text{pert}})$，$\mathcal{F}_{\text{FAN}}(\boldsymbol{x}^{\text{pert}})$ を各モデルの対応する敵対的画像のランドマークベクトルとする．

HRNet と FAN から得られる j 個の画素に対応する勾配を $\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{HRNet}} = (g_1^{\text{HRNet}}, \ldots, g_j^{\text{HRNet}})$, $\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{FAN}} = (g_1^{\text{FAN}}, \ldots, g_j^{\text{FAN}})$ とする．ここで，g_n^{HRNet} および g_n^{FAN} は，画素 n における RGB 空間における勾配ベクトルを表す．

しかし，与えられた $\boldsymbol{x}^{\text{pert}}$ に対する 2 つのモデルの勾配 $\nabla_{\boldsymbol{x}^{\text{pert}}}\mathcal{L}_{\text{HRNet}}$ と $\nabla_{\boldsymbol{x}^{\text{pert}}}\mathcal{L}_{\text{FAN}}$ が整合するとは限らない．

(i) g_n^{HRNet} と g_n^{FAN} は同一方向（内積が正）

(ii) g_n^{HRNet} と g_n^{FAN} は逆方向（内積が負）

がありえる．そこで，この問題に対して次の対策を提案する．

(1) 勾配の融合

HRNet と FAN から得られる勾配 $\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{HRNet}}$，$\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{FAN}}$ は，画素 n における RGB 空間における勾配ベクトルの内積が正となる場合は FAN の勾配ベクトル g_n^{FAN} と HRNet の勾配ベクトル g_n^{HRNet} を加算しても摂動効果が低下しない．そのため，$g_n^{\text{FAN}\prime} = g_n^{\text{FAN}}$ とする．

(2) 勾配衝突の直交化

HRNet と FAN から得られる勾配 $\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{HRNet}}$，$\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{FAN}}$ は，画素ごとに逆方向（内積が負）となるとき，単純に加算すると互いに打ち消し合い，更新効率が低下する．そこで，画素 n における RGB 空間における勾配ベクトルの内積が負となる場合は FAN の勾配ベクトルを HRNet の勾配ベクトルに直交化する：

$$g_n^{\text{FAN}\prime} = g_n^{\text{FAN}} - \frac{g_n^{\text{HRNet}\top} g_n^{\text{FAN}}}{\|g_n^{\text{HRNet}}\|} g_n^{\text{HRNet}}, \quad \text{if } g_n^{\text{HRNet}\top} g_n^{\text{FAN}} < 0, \tag{3}$$

したがって，直交化した FAN の勾配を $\nabla_{\boldsymbol{x}}\mathcal{L}_{\text{FAN}}{}' = (g_1^{\text{FAN}\prime}, \ldots, g_j^{\text{FAN}\prime})$ とする．

3.3 動的重み付け

すべてのランドマークを毎回同等に考慮すると，勾配が平均化されてしまい，摂動の効果が低下する恐れがある．そのため，本研究では，全ランドマークを毎回一様に撹乱するのではなく，反復回数に応じて重点的に撹乱するランドマーク領域を切り替える．反復 t における重み $\boldsymbol{w}$ は，所定の反復に対して $w_i = 1$，それ以外は $w_i = 0$ として構成する．反復回数と重み設定の対応関係は，表 1 の通りである．

表 1　重み切替スケジュール

反復 t	i	重み設定 w_i
0-9	0-16	$w_0 \sim w_{16} = 1$
10-19	17-26	$w_{17} \sim w_{26} = 1$
20-29	27-35	$w_{27} \sim w_{35} = 1$
30-39	36-47	$w_{36} \sim w_{47} = 1$
40-49	48-67	$w_{48} \sim w_{67} = 1$
50-59	0-67	$w_0 \sim w_{67} = 1$

3.4 敵対的画像の最適化

式 (1) で定義した損失関数 $\mathcal{L}$ を最大化することで，ランドマーク抽出器の出力を撹乱する敵対的画像 $\boldsymbol{x}^{\text{pert}}$ を生成する．この最適化には，モメンタム付きの更新アルゴリズムを採用する．$\boldsymbol{m}$ はモメンタムであり，λ はモメンタムパラメーターである．

最初の敵対画像は，

$$\boldsymbol{x}_0^{\text{pert}} = \boldsymbol{x}, \quad \boldsymbol{m}_0 = (0, \ldots, 0)$$

と初期化する．

t 回目の反復において，HRNet と FAN の勾配を

$$\nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}_{\text{HRNet}} = \nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}(\mathcal{F}_{\text{HRNet}}(\boldsymbol{x}_t^{\text{pert}}), \mathcal{F}_{\text{HRNet}}(\boldsymbol{x}_t)), \tag{4}$$

$$\nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}_{\text{FAN}} = \nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}(\mathcal{F}_{\text{FAN}}(\boldsymbol{x}_t^{\text{pert}}), \mathcal{F}_{\text{FAN}}(\boldsymbol{x}_t)) \tag{5}$$

と更新する．

勾配に基づいてモメンタム $\boldsymbol{m}_t$ を

$$\boldsymbol{m}_{t+1} = \lambda \boldsymbol{m}_t + \frac{\nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}_{\text{HRNet}} + \nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}_{\text{FAN}}{}'}{\|\nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}_{\text{HRNet}} + \nabla_{\boldsymbol{x}_t^{\text{pert}}}\mathcal{L}_{\text{FAN}}{}'\|_1} \tag{6}$$

Algorithm 1 Unbalanced Landmark Breaker Plus による敵対的画像の生成

Require: ランドマーク抽出器 $\mathcal{F}$, 元画像 $\boldsymbol{x}$, 最大反復回数 T, モメンタムパラメーター λ, ステップサイズ α, ノルム制限 ϵ

Ensure: 敵対的画像 $\boldsymbol{x}_T^{\text{pert}}$

1: 初期化：$\boldsymbol{x}_0^{\text{pert}} \leftarrow \boldsymbol{x}, \quad \boldsymbol{m}_0 \leftarrow (0, \ldots, 0)$

2: **for** $t \leftarrow 0, 1, \ldots, T-1$ **do**

3: $\quad \boldsymbol{m}_{t+1} = \lambda \boldsymbol{m}_t + \frac{\nabla_{\boldsymbol{x}_t^{\text{pert}}} \mathcal{L}_{\text{HRNet}} + \nabla_{\boldsymbol{x}_t^{\text{pert}}} \mathcal{L}_{\text{FAN}'}}{\| \nabla_{\boldsymbol{x}_t^{\text{pert}}} \mathcal{L}_{\text{HRNet}} + \nabla_{\boldsymbol{x}_t^{\text{pert}}} \mathcal{L}_{\text{FAN}'} \|_1}$

4: $\quad \boldsymbol{x}_{t+1}^{\text{pert}} \leftarrow \text{clip}\left(\boldsymbol{x}_t^{\text{pert}} - \alpha \text{sign}(\boldsymbol{m}_{t+1}),\ \epsilon\right)$

5: **end for**

6: **return** $\boldsymbol{x}_T^{\text{pert}}$

と更新する. 摂動方向に沿って更新を適用し，範囲を (2) 式のノルム制約の元で

$$\boldsymbol{x}_{t+1}^{\text{pert}} = \text{clip}\left(\boldsymbol{x}_t^{\text{pert}} - \alpha \text{sign}(\boldsymbol{m}_{t+1})\right) \tag{7}$$

と更新する. ここで，`clip` 関数は，得られた画像の各ピクセルの画素の最大変化量が所定の範囲内に収まるように切り捨て処理を行う関数である．`sign` 関数は，符号だけを取る処理を行う関数である．この処理は最大反復回数 T に達するまで繰り返される．以上をアルゴリズム 1 に示す.

4. 実　　験

4.1 実験設定

本章では，提案手法の有効性を検証するために，顔ランドマーク抽出モデル HRNet [19] と FAN [18] に対する敵対的摂動実験を行う. ランドマーク抽出モデルの評価には，広く使用されている 68 個のランドマーク付きの 300W [20] データセットを使用した.

FaceForensics++ [21] データセットから，男性 20 本，女性 20 本のビデオをランダムに選択した．これらのビデオを用いて，オートエンコーダベースのフェイススワップモデル [8] を実装し，男性 20 枚，女性 20 枚のフェイク画像を生成した．本研究では，これら計 40 枚の生成画像を対象として耐性評価を行う. なお，RQ1 に関する実験 1 は 300W データセット上で行い，RQ2 および RQ3 に関する実験 2 と実験 3 は FaceForensics++ データセットを用いて実施した.

敵対的摂動におけるパラメータ設定は以下の通りである．摂動の 1 ステップあたりの更新量は $\alpha = 1$, 繰り返し回数は $T = 60$ とした.

実験には，PyTorch [22] を用い，Ubuntu にて実行した．使用したハードウェア構成は，NVIDIA RTX A6000, および, AMD EPYC 7543P 32-Core Processor を搭載したワークステーションである.

4.2 評価方法

提案手法を以下の指標で評価する.

Normalized Mean Error（NME）[23]

予測されたランドマーク $\hat{P}$ と正解ランドマーク P のユークリッド距離の平均値を正規化したものであり，

$$\text{NME}(P, \hat{P}) = \frac{1}{k} \sum_{i=1}^{k} \frac{\|p_i - \hat{p}_i\|_2}{d} \tag{8}$$

で定義される．ここで，$p_i = \text{argmax}(h_i), \quad \hat{p}_i = \text{argmax}(\hat{h}_i)$ とする. $P = \{p_i\}_{i=1}^k$ はクリーン画像におけるランドマーク座標の集合，$\hat{P} = \{\hat{p}_i\}_{i=1}^k$ は摂動画像におけるランドマーク座標の集合を表す．d は顔のスケールを表す正規化項であり，この研究では目頭と目尻間の距離 $d = \|p_{36} - p_{45}\|_2$ を使用した．この定義により，顔の大きさに依存しない誤差の比較が可能となる．NME の値が大きいほどランドマークの予測精度が低く，摂動効果は大きいことを意味する.

Peak Signal to Noise Ratio（PSNR）[24]

摂動画像と元の入力画像との間の画質の劣化度合いを評価する指標であり，

$$\text{PSNR} = 10 \log_{10}\left(\frac{MAX^2}{\text{MSE}}\right) \tag{9}$$

で定義される. ここで, MAX は画像画素値の最大値, $MAX = 255$ とする. MSE（Mean Squared Error）は入力画素と摂動画素との間の平均二乗誤差である. PSNR の値が高いほど，摂動画像は元の画像に近く，視覚的な品質が高いことを意味する.

Structural Similarity（SSIM）[25]

摂動画像で生成されたフェイク画像とクリーンな画像で生成されたフェイク画像との間の構造的な類似度を測定する指標であり，

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$

で定義される．ここで，x はクリーンな画像で生成されたフェイク画像，y は摂動画像で生成されたフェイク画像を表す．μ_x, μ_y はそれぞれの平均値，σ_x^2, σ_y^2 は分散，σ_{xy} は共分散，C_1, C_2 は定数項である．SSIM の値は $[0, 1]$ の範囲を取り，1 に近いほど画像の構造的な類似性が高く，視覚的に自然な画像であることを意味する.

4.3 実験結果

4.3.1 実験 1：ランドマーク移動距離の比較（RQ1）

各モデルにおける各摂動方式の摂動強度 ϵ の増加に伴う平均 NME の変化を図 1 と図 2 に示す．横軸は摂動強度 ϵ，縦軸は平均 NME である.

FAN を評価モデルとした場合，FAN→FAN および ULB+→FAN では，$\epsilon = 1$ から $\epsilon = 7$ にかけて平均 NME が大きく上昇し，高い値を示した．一方，HRNet→FAN では，全ての ϵ において平均 NME が低い値にとどまった.

HRNet を評価モデルとした場合，HRNet→HRNet および ULB+→HRNet では，ϵ の増加に伴い平均 NME が上昇する傾向が確認された．特に ULB+→HRNet は，HRNet→HRNet と比較して，全ての ϵ においてやや高い値を示した．一方，FAN→HRNet では，平均 NME は低い値にとどまった.

摂動強度 $\epsilon = 3$ の条件下で生成された各方式における摂動画像の例を図 3 に示す．提案手法は従来の単一モデルに依存する方式よりランドマークを撹乱していることが確認された.

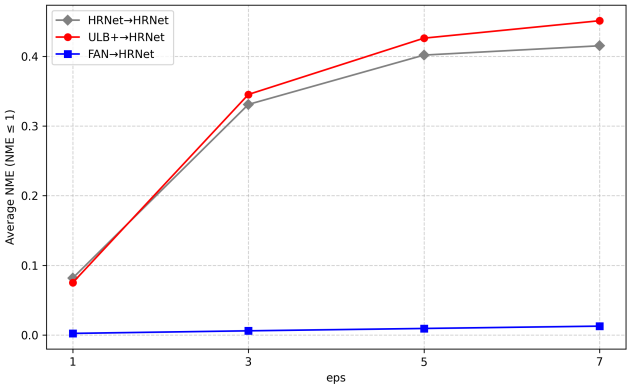

図 1　HRnet を対象とした場合の異なる摂動方式における平均 NME の変化

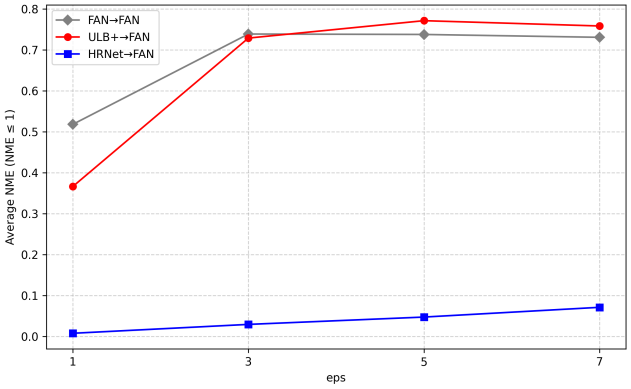

図 2　FAN を対象とした場合の異なる摂動方式における平均 NME の変化

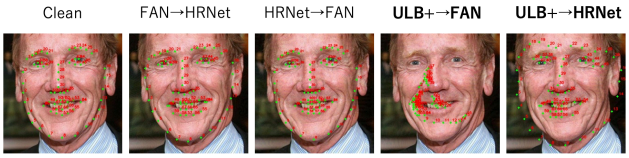

図 3　異なる摂動方式におけるランドマーク撹乱の例

4.3.2　実験 2：Deepfake 攻撃に対する耐性（RQ2）

図 4 および図 5 は，各モデルにおける各摂動方式の摂動強度 ϵ の増加に伴う平均 SSIM の変化を示している．図 4 は HRNet を評価モデルとした場合，図 5 は FAN を評価モデルとした場合の結果である．

HRNet を評価モデルとした場合，FAN→HRNet では，ϵ の増加に伴う平均 SSIM の変化は小さく，高い値を維持した．一方，HRNet→HRNet および ULB+→HRNet では，$\epsilon = 1$ から $\epsilon = 3$ にかけて平均 SSIM が大きく低下し，その後は緩やかな変化を示した．ULB+→HRNet は，HRNet→HRNet と比較して，全ての ϵ において低い値を示した．

FAN を評価モデルとした場合，HRNet→FAN では，$\epsilon = 1$ から $\epsilon = 5$ にかけて平均 SSIM が高い値を維持し，緩やかな減少にとどまった．一方，FAN→FAN および ULB+→FAN では，ϵ の増加に伴い平均 SSIM が低下する傾向が確認された．特に ULB+→FAN は，全ての ϵ において FAN→FAN よりも低い値

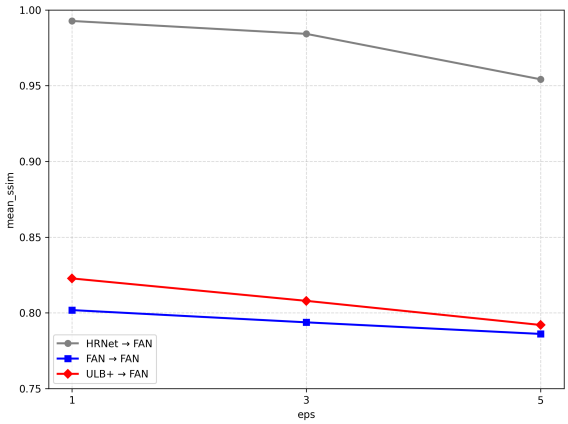

図 4　摂動強度 ϵ についてのディープフェイクの平均 SSIM の変化（FAN の場合）

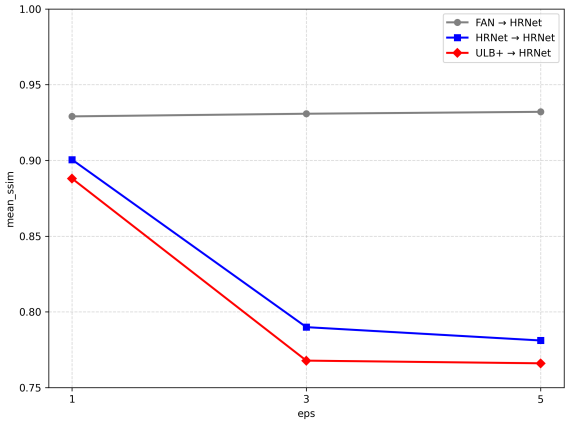

図 5　摂動強度 ϵ にていてのディープフェイクの平均 SSIM の変化（HRNet の場合）

を示した．

摂動強度 $\epsilon = 3$ の条件下で生成された各方式におけるディープフェイク出力例を図 6 に示す．上段は男性，下段は女性の入力画像に対応する．HRNet→FAN および FAN→HRNet では，外観の変化は比較的小さく，クリーン画像に近い結果が得られている．一方，ULB+→FAN および ULB+→HRNet では，顔全体に歪みが生じ，顕著な劣化が確認された．

4.3.3　実験 3：摂動画像の視覚的品質の比較（RQ3）

各モデルにおける各摂動方式の摂動強度 ϵ の増加に伴う平均 PSNR および平均 SSIM の変化を図 7 および図 8 に示す．

図 7 に示す PSNR の結果では，いずれの方式においても，ϵ の増加に伴い PSNR が低下していることが確認された．図 8 に示す SSIM の結果においても，PSNR と同様の傾向が確認された．提案手法はいずれの評価方式においても，HRNet における ULB(whitebox 方式) より高い値を示した．

摂動強度 $\epsilon = 3$ の条件下で生成された各方式の摂動画像を図 9 に示す．いずれの方式においても，顔全体に摂動が付加され，画質劣化が確認された．

4.3.4　Ablation Study

摂動強度 $\epsilon = 5$ の条件下で動的重み付け戦略および勾配衝突

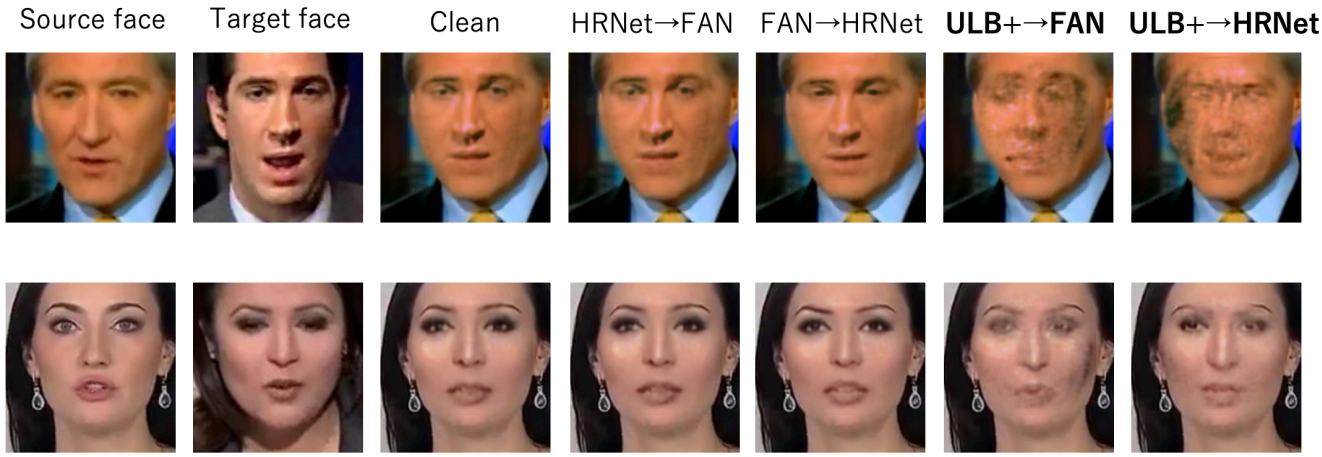

図 6　各方式における転移時のディープフェイクの出力例（提案方式 ULB ＋）

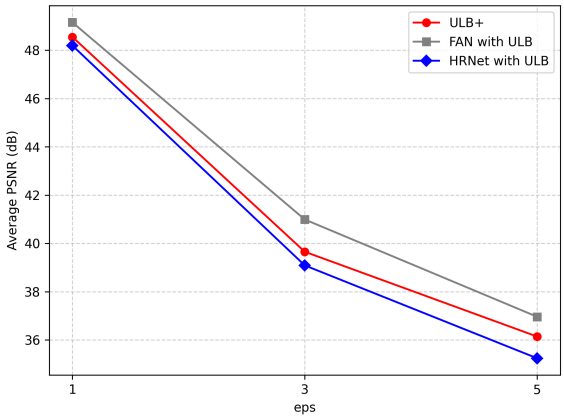

図 7　摂動強度 ϵ に関する摂動画像の平均 PSNR の変化

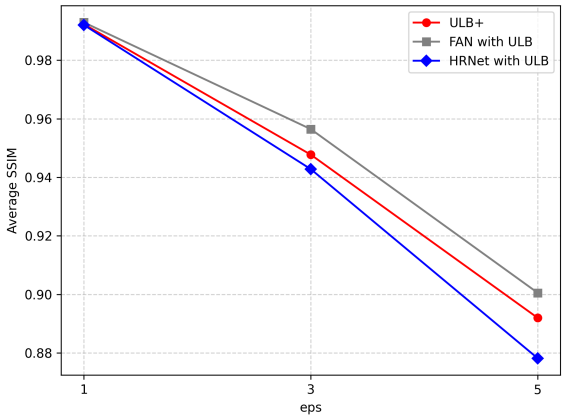

図 8　摂動強度 ϵ に関する摂動画像的平均 SSIM の変化

の緩和処理がモデルに対するランドマーク攪乱効果に与える影響を表 2 に示す．動的重み付けおよび勾配衝突の緩和は，それぞれ攪乱性能の向上に寄与しており，両者を併用することで高いモデル攪乱効果が得られることが確認された．

4.4　考　　察

4.4.1　提案手法の転移性

本章の実験結果より，単一のランドマーク抽出器に対して生成された従来の敵対的撹乱は，他のモデルに対する転移性が低

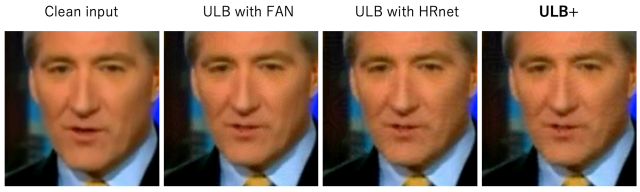

図 9　各方式における摂動画像の例

表 2　Ablation Study

ランドマーク抽出器	摂動方式	平均 NME
HRNet	ULB ＋ (動的重み付けなし)	0.392
HRNet	ULB ＋ (勾配衝突緩和なし)	0.420
HRNet	ULB ＋	**0.426**
FAN	ULB ＋ (動的重み付けなし)	0.714
FAN	ULB ＋ (勾配衝突緩和なし)	0.763
FAN	ULB ＋	**0.771**

いことが確認された．一方，複数のランドマーク抽出器の勾配情報を同時に考慮した提案方式 ULB+ による摂動生成では，評価モデルへの依存性が低く，より安定したランドマーク撹乱および Deepfake 出力の劣化（防御効果）が観測された．これは，複数モデルに共通する特徴空間に基づいて摂動が生成されていることを示唆しており，単一モデルに特化した摂動と比較して，より汎用的な撹乱が得られていると考えられる．ただし，ULB+ においても，ホワイトボックス条件下での直接摂動と同等の撹乱効果が常に得られるわけではなく，摂動強度が小さい場合やモデル差異が大きい場合には転移効果が低下する傾向が確認された．このことから，転移性を前提とした敵対的ランドマーク撹乱には，モデル間の構造的差異を考慮した設計が重要であることが示された．

4.4.2　Deepfake 防御の課題

転移性に関する実験結果から，顔ランドマーク抽出を標的とした敵対的撹乱が，Deepfake 生成結果の画質指標を低下させることは定量的に確認された．一方で，本研究では Deepfake 生成の成否を二値的に定義した成功率などの指標は導入しておらず，防御の成否を厳密に判定できていない．また，本章では

PSNR や SSIM といった客観的指標を用いて視覚的品質を評価したが，Deepfake 防御の実用性を議論する上では，人間による主観的な違和感や判別容易性を考慮した評価指標の導入も重要な課題である．これらの点を踏まえ，転移性を考慮したより実環境に即した防御手法の設計が，今後の課題として挙げられる．

5. おわりに

本章では，敵対的ランドマーク撹乱の転移性に着目し，異なる顔ランドマーク抽出器間における撹乱効果の挙動について検証を行った．単一モデルに対して生成された敵対的摂動は，他モデルに対して十分な転移性を示さない場合が多く，モデル構造の違いが転移効果に影響を与えることが確認された．一方，複数のランドマーク抽出器を同時に考慮した ULB+ による摂動生成では，単一モデルに依存せず，より一貫したランドマーク撹乱および Deepfake 生成結果の品質低下が観測された．これにより，複数モデルの情報を用いた摂動設計が，転移性の向上に寄与する可能性が示された．

ただし，本章で得られた転移効果は限定的であり，ホワイトボックス条件下での直接摂動と同等の撹乱が常に得られるわけではない．また，摂動強度が小さい場合やモデル間の差異が大きい場合には，転移効果が低下する傾向も確認された．これらの結果は，転移性を前提とした敵対的ランドマーク撹乱の設計において，モデル構造の多様性を考慮する必要性を示している．近年の Deepfake 生成器が，入力段階での幾何的誤差に対してある程度の頑健性を備えている可能性や，ランドマーク情報以外の特徴量を補完的に利用していることに起因すると考えられる．そのため，今後の Deepfake 防御においては，ランドマーク撹乱に加えて，テクスチャ特徴や潜在表現などに対する干渉を組み合わせた多層的な防御戦略の検討が必要だと考える．

文　献

[1] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2387–2395.

[2] NHK: もし、あなたの卒業アルバムが裸にされたら, https://www3.nhk.or.jp/news/html/20240914/k10014580201000.html

[3] Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, 10(1), 370.

[4] Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021). Adversarial DeepFakes: Evaluating vulnerability of DeepFake detectors to adversarial examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3348–3357.

[5] 楊力懿, 林志訓, 菊池浩明, 顔ランドマーク抽出妨害する敵対的攻撃によるディープフェイク生成防御, 第 110 回 CSEC 研究発表会, 2025.

[6] Chen, D., Chen, Q., Wu, J., Yu, X., & Jia, T. (2019). Face Swapping: Realistic Image Synthesis Based on Facial Landmarks Alignment. *Mathematical Problems in Engineering*.

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 1097-1105.

[8] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207-3216.

[9] Zhu, X., & Ramanan, D. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879-2886.

[10] Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1867-1874).

[11] Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3476-3483.

[12] Bulat, A., & Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1021-1030.

[13] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693-5703.

[14] Goodfellow, I.J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

[15] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

[16] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[17] Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017, November). ZOO: Zeroth Order Optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 15-26).

[18] HRNet-Facial-Landmark-Detection. https://github.com/HRNet/HRNet-Facial-Landmark-Detection.

[19] face-alignment. https://github.com/1adrianb/face-alignment.

[20] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 397-403.

[21] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1-11.

[22] Paszke, A., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

[23] Huang, Y., Yang, H., Li, C., Kim, J., & Wei, F. (2021). Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3080-3090.

[24] Tanchenko, A. (2014). Visual-PSNR measure of image quality. *Journal of Visual Communication and Image Representation*, 25(5), 874-878.

[25] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.