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

楊 力懿[1,a)]　林 志訓[2]　菊池 浩明[3,b)]

概要：近年, ディープフェイク技術の悪用による偽画像の拡散などの脅威が深刻な問題となっている．ディープフェイクの検出技術が発展しているが, 攻撃者も新たな技術を用いて検出を回避しようとするため, 防御側と攻撃側の技術的な競争が続いている．そこで, Li らは, ディープフェイクによる被害が広がる前に，ランドマーク抽出を妨害する敵対的攻撃手法を提案した．しかし, 本手法ではすべてのランドマークに対して一様な強さのノイズを付加するため, 人間には知覚しやすい領域にも大きな誤差を生じさせ，画質を下げていた．そこで, 本研究では, 顔の部位による識別容易性の違いに基づいて, 最適な大きさのノイズを特定の領域に加える非均一敵対的攻撃によるディープフェイク生成防御法を提案する．提案するディープフェイク生成器の性能を低下させるセキュリティ強化機構を実験実装し, オープンデータを用いて Deepfake 耐性と画質低下を実験評価する．

Defense against Deepfake generation using Adversarial Example Disrupting Facial Landmark Extraction

1. はじめに

近年，生成 AI 技術の進化により，人物の顔や音声を高精度に合成することが可能となり，その応用はエンターテインメントや教育など多岐に広がっている．一方で，この技術を悪用して標的となる画像の表情を任意に変えたり, 任意の姿勢や発言をさせる Deepfake[1] が深刻な課題になっている．NHK が 2024 年 8 月から 12 月にかけて実施した調査 [2] によれば，俳優，歌手，アイドルなど 100 人以上の有名人を含む 1,000 件超の偽画像がフリマアプリ「メルカリ」で販売されていたことが判明しており，その深刻さが浮き彫りとなっている．

そこで，ディープフェイクの検出や対策技術に関する研究が盛んに行われている．Hsu ら [3] は，機械学習を用いたフェイク画像を検出する手法を提案した．攻撃者も検出を回避する新たな技術 [4] を次々と導入しており，ディープフェイクはますます検出が困難になってきている．しかも, 一度拡散された有害なディープフェイクコンテンツは，その影響を後から完全に取り消すことが難しく，極めて深刻な被害をもたらす恐れがある．

ディープフェイク生成においては，顔のランドマーク情報が重要な役割を果たしており，多くの手法が顔の特徴点を抽出し，それに基づいて顔画像の変換や合成を行っている．そのため，顔のランドマーク抽出過程を妨害することは，ディープフェイクの生成精度を低下させる効果的な手段の一つと考えられている．Li ら [5] は, ランドマーク抽出を妨害する敵対的攻撃手法として, LandmarkBreaker を提案した．しかし, 本手法ではすべてのランドマークに対して同じ強さのノイズを付加していた．しかし, 人は顔を知覚する時に, 目や口元などいくつかの重要な部位に着目して個人の識別や表情の認識をしており, 画一的なノイズ付加は, これらの重要なランドマークを著しく変動させる恐れがある．従って, 人間による主観的な画質評価はノイズの隠蔽性の結果を大きく左右すると考える．

そこで, 本研究では，顔の部位ごとの知覚的特徴に基づいて，特定のランドマークに対して重点的にノイズを加えることが可能な非均一敵対的攻撃 Unbalanced Landmark

1 明治大学大学院先端数理科学研究科
Nakano, Nakano-ku, Tokyo 164-8525, Japan
2 国立陽明交通大学
National Yang Ming Chiao Tung University
3 明治大学総合数理学部
Nakano, Nakano-ku, Tokyo 164-8525, Japan
a) cs242039@meiji.ac.jp
b) kikn@meiji.ac.jp

© 1959 Information Processing Society of Japan

Breaker (ULB) を提案する．これにより，画質の劣化を最小限に抑えつつ，ディープフェイク生成の妨害を効果的に実現することが期待できる．

本研究では，以下の 3 つの問い（RQ）に対する検証を通じて，顔ランドマーク抽出器に対する柔軟かつ安全な敵対的攻撃の可能性を探る：

- RQ1: 特定の顔部位のランドマーク座標だけを選択的に操作することは可能か？
- RQ2: 部分的に摂動した画像は，従来の一様な摂動画像よりも DeepFake 攻撃に対して高い耐性を持つか？
- RQ3: 部分的な摂動画像は，従来よりも人間にとって知覚されにくく，視覚的な自然さを保持できるか？

2. 基本定義

2.1 Deepfake

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 [8]．

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・定位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．

深層学習の進展に伴い, CNN に基づく手法 [11] が登場し，大幅な精度向上を実現した．現在広く使われている 2017 年に Bulat らによって提案された CNN ベースのランドマーク抽出器 [12] は，2 段階の処理を行う構造を持つ．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（各ランドマークごとの確率分布）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，精度と汎用性の観点から，CNN ベースのランドマーク抽出器を主な攻撃対象とする．中でも，High-Resolution Network（HRNet）[13] は，高解像度の特徴マップを維持しながら複数スケールの情報を統合できる点に優れており，顔ランドマーク検出タスクにおいても高い性能を示している．

2.3 敵対的攻撃

Goodfellow らは, 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 [14]．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．

敵対的攻撃には，モデルの構造やパラメータに完全にアクセスできるホワイトボックス攻撃とモデルの詳細が未知であることを前提とするブラックボックス攻撃がある．

2.4 Landmark Breaker

Li らは, 顔ランドマーク抽出に対する敵対的攻撃 LandmarkBreaker 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存在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．

LandmarkBreaker は，全てのランドマークに対して広範に影響を与えることができるという利点を持つ一方で，いくつかの課題も存在する．すべてのランドマークに対して一様な強度でノイズを加えるため，顔の目元や口元など，人間が視覚的に敏感な部位にも大きな変形を与えてしまい，画質が著しく劣化する傾向がある．

3. 提案方式

本手法は，顔の部位ごとの視覚的知覚の違いに基づいてランドマークごとに異なるノイズの重みを設定し，画質劣化を抑えつつディープフェイク生成器の精度を効果的に低下させることを目的とする．

3.1 損失関数

本研究の損失関数は, ランドマークベクトルのコサイン類似度をベースとする点は先行研究と共通であるが, 各ランドマークに異なる重みを付与することで, 特定ランドマークへの選択的な攪乱が可能となるよう次のように改良を加える．

入力画像を $\boldsymbol{x}$, 摂動を加えた敵対画像を $\boldsymbol{x}^{\mathrm{pert}}$, 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, $\mathcal{F}(\boldsymbol{x}^{\mathrm{pert}}) = (\hat{h}_1, \ldots, \hat{h}_k)$ を対応する敵対的画像のランドマークベクトルとする．このとき，先行手法の損失関数は,

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

と定められている．本研究では, これを

$$\mathcal{L}(\mathcal{F}(\boldsymbol{x}), \mathcal{F}(\boldsymbol{x}_{\mathrm{pert}})) = \sum_{i=1}^{k} w_i \frac{h_i^\top \hat{h}_i}{\|h_i\| \|\hat{h}_i\|} \tag{2}$$

と定める．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．この損失関数は，元画像と敵対的画像の対応するランドマークベクトル間のコサイン類似度を加重したものであり，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 $\boldsymbol{x}^{\mathrm{pert}}$ に対してノルム制約

$$\|\boldsymbol{x}^{\mathrm{pert}} - \boldsymbol{x}\|_p \leq \epsilon \tag{3}$$

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 敵対的画像の最適化

式 (2) 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 $\boldsymbol{x}^{\mathrm{pert}}$ 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,

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

と初期化する．

t 回目の反復において，勾配 $\nabla_{\boldsymbol{x}_t^{\mathrm{pert}}} \mathcal{L}$ に基づいてモメンタム $\boldsymbol{m}_t$ を

$$\boldsymbol{m}_{t+1} = \lambda \boldsymbol{m}_t + \frac{\nabla_{\boldsymbol{x}_t^{\mathrm{pert}}} \mathcal{L}(\mathcal{F}(\boldsymbol{x}_t^{\mathrm{pert}}), \mathcal{F}(\boldsymbol{x}_t))}{\|\nabla_{\boldsymbol{x}_t^{\mathrm{pert}}} \mathcal{L}(\mathcal{F}(\boldsymbol{x}_t^{\mathrm{pert}}), \mathcal{F}(\boldsymbol{x}_t))\|_1} \tag{4}$$

と更新する．摂動方向に沿って更新を適用し，範囲を (3) 式のノルム制約の元で

$$\boldsymbol{x}_{t+1}^{\mathrm{pert}} = \mathrm{clip}\left(\boldsymbol{x}_t^{\mathrm{pert}} - \alpha \mathrm{sign}(\boldsymbol{m}_{t+1})\right) \tag{5}$$

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．

Algorithm 1 Unbalanced Landmark Breaker 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 $\boldsymbol{x}_T^{\mathrm{pert}}$
1: 初期化：$\boldsymbol{x}_0^{\mathrm{pert}} \leftarrow \boldsymbol{x}, \quad \boldsymbol{m}_0 \leftarrow (0, \ldots, 0)$
2: **for** $t \leftarrow 0, 1, \ldots, T-1$ **do**
3: $\quad \boldsymbol{m}_{t+1} = \lambda \boldsymbol{m}_t + \frac{\nabla_{\boldsymbol{x}_t^{\mathrm{pert}}} \mathcal{L}(\mathcal{F}(\boldsymbol{x}_t^{\mathrm{pert}}), \mathcal{F}(\boldsymbol{x}_t))}{\|\nabla_{\boldsymbol{x}_t^{\mathrm{pert}}} \mathcal{L}(\mathcal{F}(\boldsymbol{x}_t^{\mathrm{pert}}), \mathcal{F}(\boldsymbol{x}_t))\|_1}$
4: $\quad \boldsymbol{x}_{t+1}^{\mathrm{pert}} \leftarrow \mathrm{clip}\left(\boldsymbol{x}_t^{\mathrm{pert}} - \alpha \mathrm{sign}(\boldsymbol{m}_{t+1}),\ \varepsilon\right)$
5: **end for**
6: **return** $\boldsymbol{x}_T^{\mathrm{pert}}$

4. 実験

4.1 実験設定

本研究では，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[15] に対する敵対的攻撃実験を行う．ランドマーク抽出モデルの学習および評価には，広く使用されている 68 個のランドマーク付きの 300W [16] データセットを使用した．

68 点ランドマークの部位対応関係は，表 1 の通りである（インデックスは 0 始まり）．

ランドマークごとの重み w_i は，攻撃したい部位に対応するインデックスの重みを $w_i = 1$ にし，それ以外のランドマークについては $w_i = 0$ に設定する．摂動部位と重み設定の対応関係は，表 1 の通りである．

FaceForensics++ [17] データセットの一部を用い，オートエンコーダをベースとしたフェイススワップモデルを実装してディープフェイク耐性を評価する．

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．摂動の最大許容範囲は $\epsilon = 15$，1 ステップあたりの更新量は $\alpha = 1$，繰り返し回数は 20 回とした．ノルム制約には ℓ_∞ ノルムを使用した．

実験は，PyTorch[18] を用い，Ubuntu 環境にて実行した．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

表 1: ランドマークインデックスと顔部位の対応

i	部位名	N_P	重み設定 w_i
0-16	顔の輪郭 (faceground)	17	$w_0 \sim w_{16} = 1$
17-21	左眉 (left eyebrow)	5	$w_{17} \sim w_{21} = 1$
22-26	右眉 (right eyebrow)	5	$w_{22} \sim w_{26} = 1$
27-35	鼻 (nose)	9	$w_{27} \sim w_{35} = 1$
36-41	左目 (left eye)	6	$w_{36} \sim w_{41} = 1$
42-47	右目 (right eye)	6	$w_{42} \sim w_{47} = 1$
48-67	口元 (mouth)	20	$w_{48} \sim w_{67} = 1$
	計	68	

© 1959 Information Processing Society of Japan

たワークステーションである．

4.2 実験方法

提案手法による表 1 の 5 つのランドマークの各々の w を 1 にし，他を 0 として，algorithm1 で以下の指標で評価する．

Normalized Mean Error（NME）[19]

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，顔のスケールを表す正規化係数で除したものであり，

$$\mathrm{NME}(P, \hat{P}) = \frac{1}{N_P} \sum_{i=1}^{N_P} \frac{\|p_i - \hat{p}_i\|_2}{d} \tag{6}$$

で定義される．ここで，$P = \{p_i\}_{i=1}^{N_P}$ は正解画像（クリーン画像）におけるランドマーク座標の集合，$\hat{P} = \{\hat{p}_i\}_{i=1}^{N_P}$ は敵対的画像におけるランドマーク座標の集合，N_P はランドマークの数を表す．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，攻撃効果は大きいことを意味する．

Peak Signal to Noise Ratio（PSNR）[20]

攻撃後の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，

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{MAX^2}{\mathrm{MSE}} \right) \tag{7}$$

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）[21]

攻撃後の摂動画像と元の入力画像との間の構造的な類似度を測定する指標であり，

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

で定義される．ここで，x と y は比較対象の画像における位置，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

本節では，設定した 3 つの問い（RQ1〜RQ3）に対して，実験結果に基づく検証を行う．

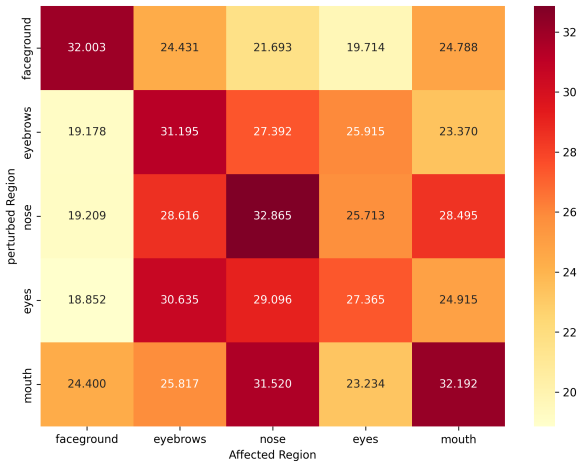

図 1: 部位別摂動におけるランドマーク移動距離の部位間平均（単位：ピクセル）

（実験 1）

ランドマーク移動距離の比較（RQ1）

提案手法による部位別攻撃が，各顔部位に与える撹乱の影響を可視化するため，表 1 の 5 つの部位のランドマーク移動距離の平均値部位間のヒートマップを図 1 に示す．

このヒートマップは，摂動対象の部位（行），移動量を測定した部位（列）について，ランドマークの平均移動距離を表す．

図より，摂動対象部位と同一部位における移動距離は他部位に比べて高い値を示す傾向が見られ，部位ごとの撹乱効果に差が存在することが確認された．たとえば，鼻に対する攻撃（3 行目）では，鼻部のランドマーク移動距離が最大であり，局所的な撹乱が達成されている．一方で，目（4 行目）への摂動においては，摂動対象でない眉（2 列目）のランドマークにも目の 27.36 よりも大きな撹乱（30.63）が生じている．

理想的部位別摂動では，ヒートマップ上の色の強い部分が対角線上に集中し，摂動対象部位に限定した撹乱が期待される．しかし，本結果は，対角線以外のセルにも比較的強い撹乱が生じていることを示している．

提案方式と先行研究によるランドマーク変動を示すため，顔の輪郭と目元にのみ摂動対象にした各部位のランドマーク平均移動距離を図 2, 3 のレーダーチャートで示す．図 2 より，摂動対象とした顔の輪郭のランドマークにおいて，既存方式と比べて高い移動距離が観測された．摂動対象以外のランドマークにおいて，既存方式と比べて移動距離が小さいことが示された．しかし，図 3 より，摂動対象以外のランドマークにおいても，既存方式と比べて高い移動距離が観測された．

眉（インデックス 17-26）のみを摂動対象とした場合における，各ランドマークの平均移動距離を図 4 に示す．

図より，摂動対象とした眉（オレンジ）のランドマーク

© 1959 Information Processing Society of Japan

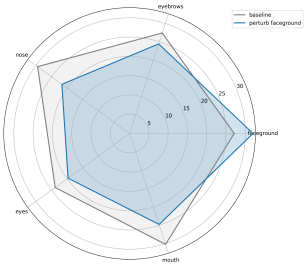

図 2: 顔の輪郭摂動と既存方式の比較

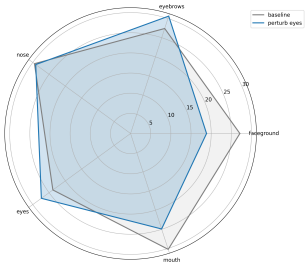

図 3: 目元摂動と既存方式の比較

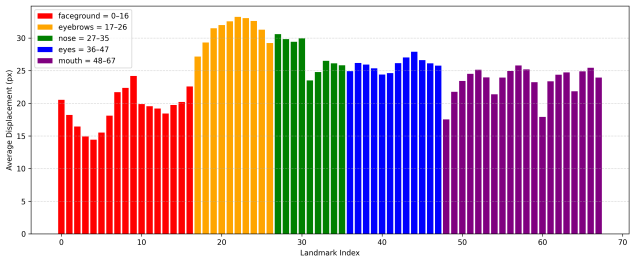

図 4: 眉摂動（17-26）における各ランドマークの平均移動距離

表 2: 部位別摂動における平均 NME の比較

摂動部位	ランドマーク数	平均 NME
鼻	9	0.2286
眉	10	0.2133
目元	12	0.2195
顔の輪郭	17	0.2217
口元	20	**0.2410**
全体 (Baseline)	68	0.2372

において，他部位と比べて高い移動距離が観測された．一方で，隣接する目（青）や鼻（緑）のランドマークにおいても大きく移動が生じており，摂動の影響が摂動対象部位以外にも波及している．顔輪郭（赤）や口（紫）など，眉から物理的に距離のある部位に対しても移動が発生している．

以上より，RQ1「特定の顔部位のランドマーク座標だけを選択的に操作することは可能か？」に対して，提案方式は不完全である．

表 2 は，摂動対象とするランドマーク数と平均 NME の関係を示す．

表より，摂動対象とするランドマーク数の増加に伴い，平均 NME が概ね高くなる傾向が見られる．一方で，鼻の摂動では，ランドマーク数が一番少ないにも関わらず，眉の摂動よりも平均 NME が高い結果となった．なお，口元の摂動においては，平均 NME が全体摂動よりも高い結果となった．

目元のみを摂動対象とした場合における，各ランドマークの変化を図 5 に示す．図より，提案手法（目元摂動）の方は，既存方式（Baseline）と比べてランドマーク全体の位置の乱れがより顕著に現れている．

表 3: 平均 PSNR / SSIM の比較

摂動部位	PSNR (dB)	SSIM
顔の輪郭	39.60	0.9496
眉	39.66	0.9503
鼻	**39.69**	0.9505
目元	39.67	0.9501
口元	**39.69**	**0.9508**
既存方式	39.64	0.9497

（実験 2）

Deepfake 耐性（RQ2）

元の画像（Clean），既存手法 (Baseline)，および提案手法（ULB）を入力した場合の DeepFake 生成の出力例を図 6 示す．各行は，左から順に摂動画像，フェイスアライメント結果，顔合成結果，DeepFake 生成結果を示している．

図の右列から，既存手法（baseline）と提案手法 ULB いずれも，DeepFake 生成画像に歪みや不自然なアーティファクトが生じていることが確認できる．特に，提案手法（右下）では，既存手法と比較して，顔全体の形状の乱れがより顕著に現れている．

視覚的品質の比較

表 3 に，各部位別摂動および既存手法における摂動画像の平均 PSNR および平均 SSIM を示す．

表より，提案手法による部位別摂動において，既存手法と比較して PSNR および SSIM は高いが，差が十分ではない．

特に，鼻および口を対象とした摂動では，PSNR および SSIM が比較的高く，局所的な摂動が視覚的自然さに与える影響が小さい．

（実験 3）

主観的評価（RQ3）

表 4 に，鼻摂動および既存手法による摂動画像の画質に対する主観的評価（画質の悪い方を選ぶ）の結果を示す．1 列目は提案方式と既存手法の PSNR の差分（いずれも提案方式の PSNR が高い）を表す．2 列目は既存方式を選択した人数を表す．3 列目は既存方式を選択した人数が全体人数（8 人）に占める割合を表す．9 回の選択の中で，既存方式を選択した人数の割合が高いのが 7 回あった．しかし，PSNR が高いにも関わらず，提案方式を選択した人数の割合が高いのが 2 回あった．

4.4 考察

4.4.1 部位選択的な操作の設計の不十分

実験結果より，提案手法は摂動部位に対して他部位より大きな摂動を加えることは可能であったが，摂動対象部位以外にも一定の影響が波及していることが確認された．これは，損失関数の設計上，全体の確率マップの勾配が完全に局所化されていないこと，および HRNet のネットワー

© 1959 Information Processing Society of Japan

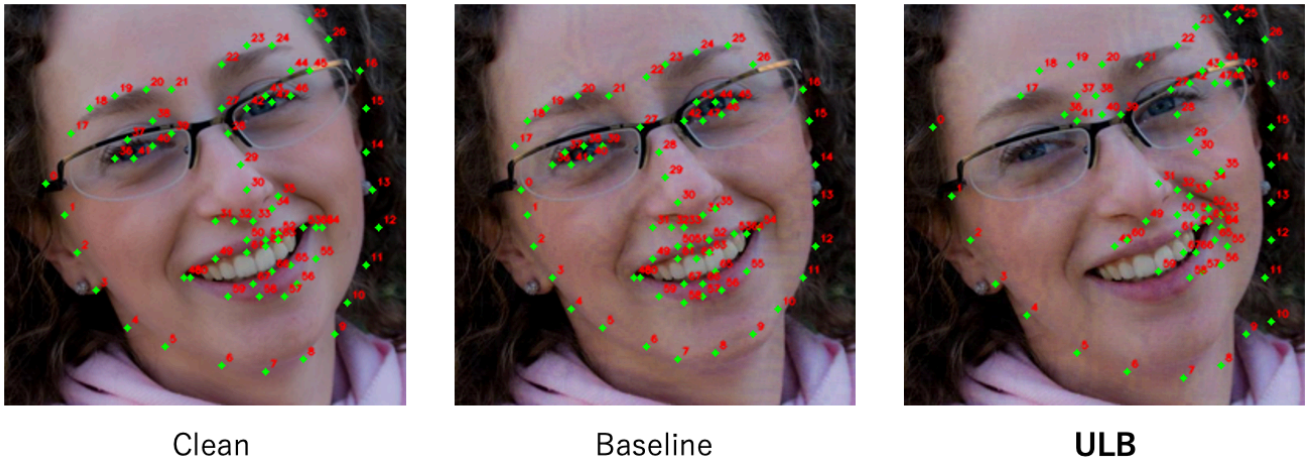

図 5: 目元摂動における各ランドマークの変化

表 4: 主観的評価の結果

PSNR 差分 (dB)	選択人数	割合
1.30	7	0.875
0.88	7	0.875
0.70	5	0.625
0.56	3	0.375
0.55	2	0.25
0.54	5	0.625
0.00	8	1
0.00	6	0.75
0.00	7	0.875

ク構造において各ランドマーク間に高い空間的依存性が存在することが一因と考えられる．

4.4.2 Deepfake 耐性向上の現状と課題

生成した Deepfake の比較結果から，提案手法は既存手法と比較して Deepfake 生成画像に対して一定の品質低下を引き起こす傾向が見られた．しかし，品質低下の程度は限定的であり，その一因として，顔アライメント工程において重要なランドマーク（例えば，目の位置や鼻先など）に対する撹乱が現状では十分でなく，顔の整列自体が完全には破綻していない可能性が考えられる．このため，Deepfake 生成器はある程度補正を行い，生成画像の品質を維持していると推察される．

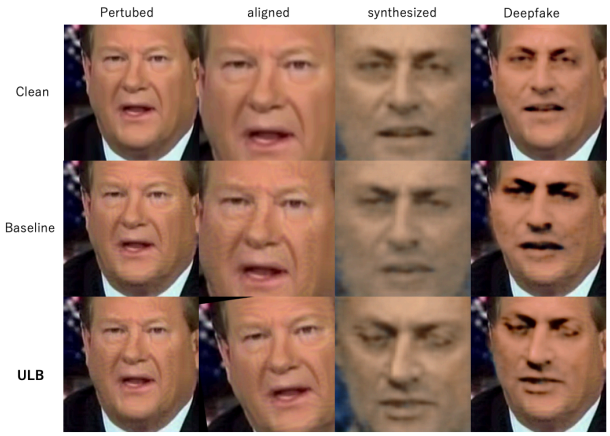

図 6: DeepFake 生成結果

4.4.3 視覚的自然さの維持と主観評価の必要性

PSNR および SSIM の結果から，提案手法は既存手法と比較して画質の劣化を抑制し，自然さを良好に保持できている傾向が見られた．差が十分ではないのは既存手法と同じく画像全体に摂動が加わっていると推測する．また，主観的評価においても，ある程度 PSNR や SSIM などの客観的評価指標との相関が見られたが，評価対象や評価人数が限定的であり，統計的に有意な関係性を示すには至っていない．

5. おわりに

本研究では，顔ランドマーク抽出器に対する新たな敵対的攻撃手法として，部位ごとに異なる重みを導入した損失関数を提案した．

実験では，輪郭，目元，口元など異なる部位を個別に摂動した際のランドマーク移動距離や NME を分析し，提案手法が狙った部位に対して集中的な撹乱を与えられる一方で，周辺部位への波及や全体的な影響範囲とのバランスに課題が残ることも確認した．

Deepfake モデルへの影響に関しては，一部の攻撃画像で生成品質の変化が確認されたが，従来手法と比べた一貫した耐性向上の有無については，今後より大規模かつ多様な評価が求められる．また，現状の手法では画像全体に摂動が加わるため，視覚的自然さの観点でもさらなる改善が必要である．

今後の課題としては，以下の点が挙げられる．まず，ランドマーク間の空間的相関や依存関係を考慮した損失関数や摂動最適化の導入により，より効果的かつ一貫性のある攻撃手法の構築を目指す．次に，モデルの勾配や応答を利用して，影響度の高いランドマークを自動的に特定し，重

© 1959 Information Processing Society of Japan

点的に摂動を加える適応的な攻撃戦略の設計を検討する．また，複数の Deepfake 生成モデルやデータセットを対象とした大規模な評価実験を通じて，提案手法の汎用性と安全性の向上を図りたい．加えて，局所領域に対する摂動範囲の制約やマスク適用などにより，視覚的自然さを高める攻撃設計も今後の重要な課題である．さらに，ユーザ調査や知覚実験を取り入れ，より定量的かつ実用的な評価の実施が求められる．

参考文献

[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: 芸能人の偽の性的画像 メルカリなどで大量販売 対策求める声, `https://www3.nhk.or.jp/news/html/20241214/k10014665821000.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] Li, Y., Sun, P., Qi, H., & Lyu, S. (2024). LandmarkBreaker: A proactive method to obstruct DeepFakes via disrupting facial landmark extraction. *Computer Vision and Image Understanding*, 204, 103935.

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

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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.

© 1959 Information Processing Society of Japan