

顔画像ディープフェイクを防御する ブラックボックス型敵対的ランドマーク搅乱

楊 力懿^{1,a)} 劉 正徳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。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。この問題に対して、ディープフェイクによる被害が広がる前に、生成プロセスそのものを妨害するために、顔画像における目鼻口などのランドマークに重みを設置し、ランドマーク抽出を妨害する敵対的サンプル手法が研究されている。しかし、本手法では、標的とする機械学習モデルのアルゴリズムを与えて学習重みの勾配を用いて摂動する、いわゆる、ホワイトボックスモデルを仮定しており、現実性に劣る。そこで、本研究では、学習アルゴリズムを既知とする仮定を弱めて、未知の学習アルゴリズムに対して、出力層のラベルだけで敵対的攻撃を行うブラックボックスモデルの上で敵対的サンプルを提案し、従来のホワイトボックス型の防御手法と比較してディープフェイク生成器の性能を低下させる安全性の向上を実験的に評価する。

キーワード：ディープフェイク、敵対的摂動

Black-Box Adversarial Perturbations for Disrupting Landmark Detection in Facial Deepfake

YANG LIYI^{1,a)} LIU CHENG-TE² KIKUCHI HIROAKI^{3,b)}

Abstract: In recent years, the malicious use of deepfake technology has raised serious concerns due to privacy violations and the spread of disinformation. While detection techniques for deepfakes have made significant progress, adversaries are concurrently developing new methods to evade such detection, resulting in a continuous technological arms race between defense and attack. To address this issue before deepfake damage becomes widespread, recent studies have explored adversarial sample techniques aimed at disrupting the generation process itself. These approaches interfere with landmark extraction (such as eyes, nose, and mouth positions) by assigning weights to facial landmarks to impair the alignment process critical for deepfake generation. However, these methods often rely on so-called white-box settings, which assume full knowledge of the target machine learning model's algorithm and gradients with respect to its learned weights, thereby limiting their applicability in real-world scenarios. In this study, we propose a black-box adversarial sample generation method that relaxes the assumption of known learning algorithms. Our method performs attacks using only output labels from the model, without requiring internal access to its architecture or gradients. We experimentally evaluate the proposed method by comparing it with conventional white-box-based defense techniques, and demonstrate its effectiveness in degrading the performance of deepfake generators, thereby enhancing the robustness of deepfake mitigation strategies.

Keywords: Deepfake, Adversarial Perturbation

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

1. はじめに

近年、生成AI技術の進展により、人間の顔や音声を高精度で合成することが可能となり、エンターテインメントや教育分野をはじめ幅広く応用されている。しかし一方で、特定の人物画像に対して表情や発話内容を改変するDeepfake[1]の悪用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[3]は機械学習を活用したフェイク画像検出手法を提示している。しかし、Hussainらは、攻撃者側も検出を回避する新たな生成技術[4]を取り入れており、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、深刻な被害を引き起こす可能性が高い。

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。これは標的モデルの内部アルゴリズムや勾配情報を攻撃者が事前に知っていることを前提とするホワイトボックス仮定に依存しており、現実性に劣るという点が課題であった。

そこで、本研究では、学習アルゴリズムを既知とする仮定を緩和し、出力ラベルのみを利用して攻撃を実現するブラックボックス仮定に基づく新たな敵対的サンプル手法Blackbox Landmark Breaker (BLB)を提案する。ホワイトボックス型防御手法との比較実験を通じて、本手法がディープフェイク生成器の性能を抑制し得ることを示し、防御技術としての有効性を評価する。

本研究では、ブラックボックス型防御手法の有効性を明らかにするため、以下の3つの研究課題(RQ)を設定する：

- RQ1: ブラックボックス型撮動は顔のランドマーク座標に対して、敵対的搅乱を効果的に付与するか。
- RQ2: ブラックボックス型防御手法はDeepfake攻撃に対する耐性があるか。
- RQ3: ブラックボックス型防御手法とホワイトボックス型撮動画像は、ランドマーク撮動性、ディープフェイク耐性、及び視覚的自然さの観点でどちらが優れて

いるか。

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[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、頬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Face Alignment Network (FAN)[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が敵対的撮動と呼ばれる意図的

b) kikn@meiji.ac.jp

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 Unbalanced Landmark Breaker

Yang 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、ホワイトボックスモデルを仮定している。

2.5 Zeroth-Order Optimization (ZOO)

Chen らは 2017 年に Zeroth-Order Optimization (ZOO) を提案した [15]。本手法は、モデルの入力と出力の確率分布のみを利用し、有限差分によって勾配を近似することで最適化を実行する。ZOO は、攻撃者が訓練済みモデルの内部パラメータにアクセスすることなく、ディープフェイクに対する敵対的攻撃を実現可能にした点に特長がある。

3. 提案方式

本研究では、ブラックボックス型の敵対的サンプル生成手法 Blackbox Landmark Breaker (BLB) を提案する。本手法は、ディープフェイク生成器における顔合成精度を効果的に低下させることを目的とする。

3.1 ランドマークベクトルのコサイン類似度

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 \mathbf{x} 、摂動を加えた敵対画像を \mathbf{x}^{pert} 、ランドマーク抽出器を \mathcal{F} とする。 $\mathcal{F}(\mathbf{x}) = (h_1, \dots, h_k)$ を元画像に対応する k 個のランドマークベクトル、 $\mathcal{F}(\mathbf{x}^{\text{pert}}) = (\hat{h}_1, \dots, \hat{h}_k)$ 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。 $\mathbf{w} = (w_1, \dots, w_k)$, $w_i \in [0, 1]$ 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、[5] のホワイトボックス ULB の損失関数は、

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

と定められていた。

本研究では、重みを全て 1 にし、これを

$$\text{Sim}(\mathcal{F}(\mathbf{x}), \mathcal{F}(\mathbf{x}^{\text{pert}})) = \sum_{i=1}^k \frac{h_i^\top \hat{h}_i}{\|h_i\| \|\hat{h}_i\|} \quad (2)$$

と定める。

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 \mathbf{x}^{pert} に対してノルム制約

$$\|\mathbf{x}^{\text{pert}} - \mathbf{x}\|_\infty \leq \epsilon \quad (3)$$

を定める。ここで、 ϵ は摂動の大きさを示す定数であり、 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

本研究で提案する Blackbox Landmark Breaker (BLB) は、ランドマーク検出器の内部ネットワーク層の勾配を利用せず、出力層のベクトルのみを参照して摂動を更新する。また、 \mathbf{u}_i は入力画像と同次元のランダム方向ベクトルであり、ガウス分布からサンプリングされ、有限差分による勾配近似に用いられる。

t 回目の反復では、まず N 本の方向ベクトルをサンプリングし、 $\mathbf{x}_t^{\text{pert}}$ に対して $\pm \eta$ (定数) の摂動を加えた画像 $\mathbf{x}^+ = \mathbf{x}_t^{\text{pert}} + \eta \mathbf{u}_i$ および $\mathbf{x}^- = \mathbf{x}_t^{\text{pert}} - \eta \mathbf{u}_i$ を生成する。それこれから得られるランドマークベクトル $\mathcal{F}(\mathbf{x}^+), \mathcal{F}(\mathbf{x}^-)$ と $\mathcal{F}(\mathbf{x})$ のコサイン類似度を比較することで、方向 \mathbf{u}_i における差分近似を次式で求める：

$$\delta_i = \frac{\text{Sim}(\mathcal{F}(\mathbf{x}^+), \mathcal{F}(\mathbf{x})) - \text{Sim}(\mathcal{F}(\mathbf{x}^-), \mathcal{F}(\mathbf{x}))}{2\eta}, \quad (4)$$

この結果を全方向で平均し、近似勾配を

$$\hat{\nabla} \mathcal{L} = \frac{1}{N} \sum_{i=1}^N \delta_i \mathbf{u}_i \quad (5)$$

として推定する。

この推定勾配を用いて、反復的に入力画像を更新する：

$$\mathbf{x}_{t+1}^{\text{pert}} = \text{clip} \left(\mathbf{x}_t^{\text{pert}} - \alpha \text{sign}(\hat{\nabla} \mathcal{L}) \right), \quad (6)$$

ここで α はステップサイズ、 ϵ はノルム制約に基づく最大摂動量、clip 関数は画素値が範囲外に出ないように制約を課す関数である。この更新を最大反復回数 T まで繰り返す。

Algorithm 1 Blackbox Landmark Breaker による敵対的画像生成

Require: ランドマーク抽出器 \mathcal{F} , 元画像 x , 最大反復回数 T , 方向数 N , 有限差分幅 η , ステップサイズ α , ノルム制約 ϵ

Ensure: 敵対的画像 x_T^{pert}

- 1: 初期化 : $x_0^{\text{pert}} \leftarrow x$
- 2: **for** $t \leftarrow 0, 1, \dots, T - 1$ **do**
- 3: N 本のランダム方向ベクトル u_i, \dots, u_N をサンプリング
- 4: **for** $i \leftarrow 1, \dots, N$ **do**
- 5: $x^+ \leftarrow x^{\text{pert}} t + \eta u_i$
- 6: $x^- \leftarrow x^{\text{pert}} t - \eta u_i$
- 7: $\delta_i \leftarrow \frac{\text{Sim}(\mathcal{F}(x^+), \mathcal{F}(x)) - \text{Sim}(\mathcal{F}(x^-), \mathcal{F}(x))}{2\eta}$
- 8: **end for**
- 9: $\hat{\nabla}\mathcal{L} \leftarrow \frac{1}{N} \sum_{i=1}^N \delta_i u_i$
- 10: $x^{\text{pert}} t + 1 \leftarrow \text{clip} \left(x_t^{\text{pert}} - \alpha \text{sign}(\hat{\nabla}\mathcal{L}) \right)$
- 11: **end for**
- 12: **return** x_T^{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[16] 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 [17] データセットを使用した。

FaceForensics++ [18] データセットから、男性 100 本、女性 100 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 100 枚、女性 100 枚のフェイク画像を生成した。本研究では、これら計 200 枚の生成画像を対象として耐性評価を行う。なお、RQ1 に関する実験は 300W データセット上で行い、RQ2 および RQ3 に関する実験は 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 $\alpha = 5$ 、繰り返し回数は $T = 50$ とした。有限差分による勾配近似に用いるパラメータは $\eta = 0.0005$ 、サンプリングする方向ベクトルの本数は $N = 30$ とした。

実験には、PyTorch[19] 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) [20]

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、

$$\text{NME}(P, \hat{P}) = \frac{1}{k} \sum_{i=1}^k \frac{\|p_i - \hat{p}_i\|_2}{d} \quad (7)$$

で定義される。ここで、 $p_i = \text{argmax}(h_i)$ 、 $\hat{p}_i = \text{argmax}(\hat{h}_i)$ 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) [21]

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、

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \quad (8)$$

で定義される。ここで、 MAX は画像画素値の最大値、 $\text{MAX} = 255$ 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) [22]

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、

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (9)$$

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

本節では、設定した 3 つの問い合わせ (RQ1~RQ3) に対して、実験結果に基づく検証を行う。

ランドマーク移動距離の比較 (RQ1)

異なる摂動方式 (random noise, blackbox, whitebox[5]) における摂動強度 ϵ の増加に伴う平均 NME の変化を図 1 に示す。横軸は摂動強度 ϵ 、縦軸は平均 NME である。

whitebox 方式 [5] では、 $\epsilon = 10$ から $\epsilon = 50$ にかけて平均 NME が大きく上昇し、最も高い値を示した。提案 blackbox 方式では、 ϵ の増加に伴い平均 NME が緩やかに上昇し、whitebox 方式より低い値を示したが、random

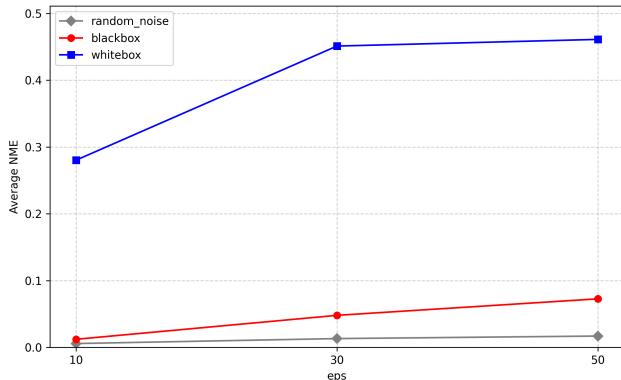


図 1: 異なる撮動方式における平均 NME の変化

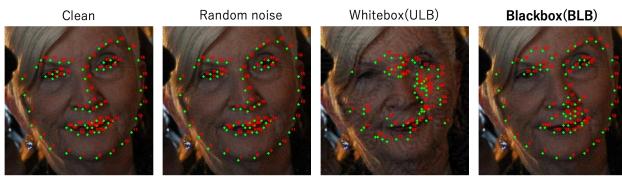


図 2: 異なる撮動方式におけるランドマーク搅乱の例

noise よりは高い。

撮動強度 $\epsilon = 30$ の条件下で生成された各方式における撮動画像の例を図 2 に示す。提案 blackbox 方式は whitebox 方式ほど搅乱が強くないが、random noise よりランドマークを搅乱していることが確認された。

Deepfake 攻撃に対する耐性 (RQ2)

図 3 および図 4 は、異なる撮動方式 (random noise, blackbox, whitebox) における撮動強度 ϵ の増加に伴う平均 SSIM の変化を、それぞれ男性および女性の顔画像について示したものである。

図 3 に示す男性データでは、whitebox 方式において SSIM が最も低下し、 $\epsilon = 10$ から $\epsilon = 50$ にかけて緩やかな減少傾向を示した。blackbox 方式でも SSIM の低下が観測されたが、whitebox 方式に比べて高い値を維持した。一方、random noise では全体を通じて SSIM が高く、ほとんど変化が見られなかった。

図 4 に示す女性データでも、同様の傾向が確認された。whitebox 方式は最も大きな SSIM の低下を示し、blackbox 方式は中間的な値を維持した。random noise では男女いずれの場合も SSIM がほぼ一定であった。

撮動強度 $\epsilon = 50$ 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 [8] の出力例を図 5 に示す。上段は男性、下段は女性の入力画像に対応する。

男性の例では、whitebox 方式において顔全体に強い歪みが生じ、著しい画質劣化が確認された。blackbox 方式では、whitebox 方式よりも搅乱は少ないものの、顔の各部位に明確な歪みが残り、ディープフェイク生成が不自然となっている。random noise では外観の変化が小さく、クリーン画像に近い結果となった。

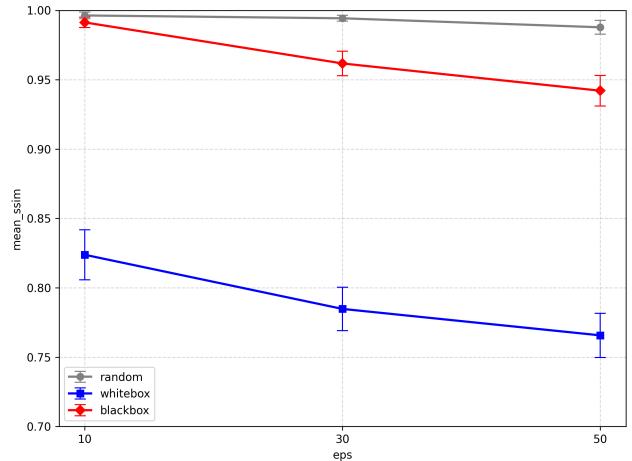


図 3: 摄動強度の増加に伴うディープフェイクの平均 SSIM の変化 (男性の場合)

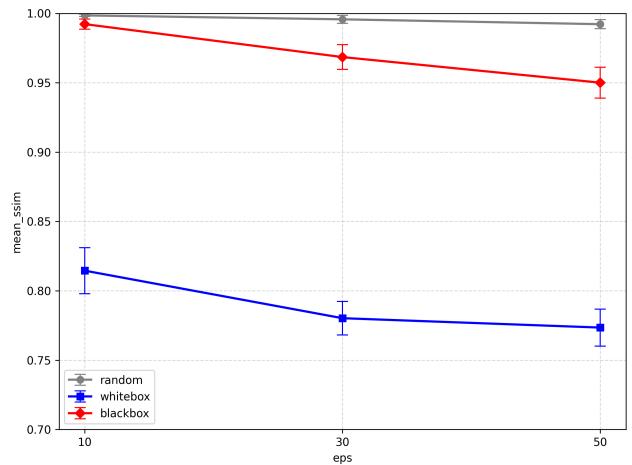


図 4: 摄動強度の増加に伴うディープフェイクの平均 SSIM の変化 (女性の場合)

女性の例においても同様の傾向が確認された。whitebox 方式では画質の劣化が顕著であり、blackbox 方式ではそれに比べて視覚的自然さをある程度保持しつつも、ディープフェイク結果に歪みが生じている。

撮動画像の視覚的品質の比較 (RQ3)

図 6 および図 7 は、男性および女性データにおける撮動画像の平均 PSNR を示している。

whitebox 方式は、 ϵ の増加に伴って PSNR が大きく低下し、最も画質劣化が顕著であった。一方、blackbox 方式と random noise は、全体を通じてほぼ同じ傾向を示しており、 ϵ の増加に伴って緩やかに低下した。これにより、blackbox 方式の PSNR は whitebox 方式よりも高く維持され、画質劣化が抑えられていることが確認できる。

図 8 は、撮動強度 $\epsilon = 50$ の条件下における各方式の撮動画像を示している。

whitebox 方式では、顔全体に強い撮動が付加され、視覚的に顕著な画質劣化が生じていることが確認できる。一



図 5: 各方式におけるディープフェイクの出力例

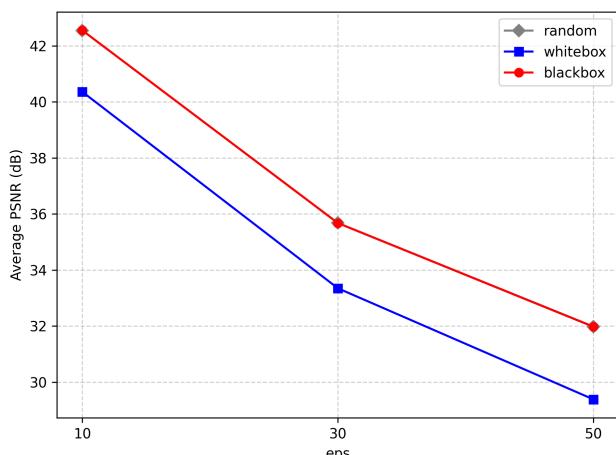


図 6: 摂動強度の増加に伴う摂動画像の平均 PSNR の変化（男性の場合）

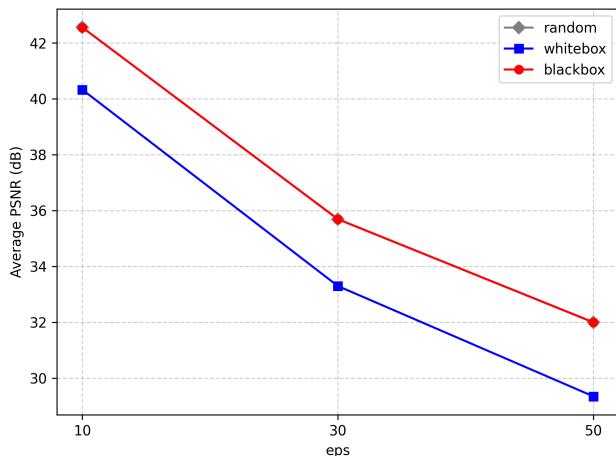


図 7: 摂動強度の増加に伴う摂動画像の平均 PSNR の変化（女性の場合）

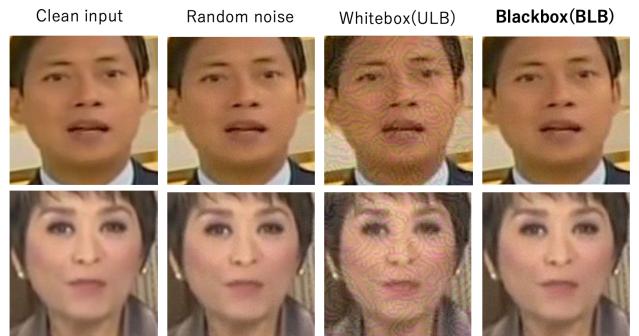


図 8: 各方式における撮動画像の例

方, blackbox 方式では, whitebox 方式に比べて撮動は抑えられており, より自然な外観を維持している. random noise では外観の変化は小さく, クリーン画像に近い結果が得られた.

4.4 考察

4.4.1 ブラックボックス方式の有用性

本研究で提案したブラックボックス方式は, ホワイトボックス方式に比べて撮動画像の画質を維持しつつ, ディープフェイクの質を低下させ得ることが確認された. 特に, NME の上昇や SSIM の低下といった指標から, ランダムノイズよりも一貫して高い攻撃効果を示した. これは, 勾配情報を直接利用できない環境においても, ランドマークベクトルのコサイン類似度を最適化目標とすることで, 有効な撮動が生成可能であることを示している. 一方で, 効果はホワイトボックス方式に比べて限定的であり, 特に ϵ が小さい場合には搅乱が不十分となる傾向が見られた.

4.4.2 Deepfake 耐性向上の現状と課題

Deepfake 実験の結果, BLB は ULB に比べて撮動画像の自然さをある程度維持しながらも, 生成結果に明確な歪

みを生じさせることができ、Deepfake 攻撃に対して一定の有効性を有することが確認された。しかし、依然として完全な防御には至っておらず、特に図 5 の女性の例においては、撮動を加えても比較的自然な合成が成立するケースが観察された。これは、Deepfake 生成器が入力撮動に対してある程度の頑健性を持つ可能性や、ランドマーク抽出の誤差が必ずしも最終合成に直結しない構造的特性に起因すると考えられる。改善のためには、ランドマーク搅乱のみならず、テクスチャ特徴や潜在表現に対する多面的な干渉を組み合わせる必要がある。また、現状の評価は SSIM や PSNR といった客観指標に依存しているが、Deepfake 防御の観点では「どの程度不自然に見えるか」という主観的評価の導入も不可欠である。

5. おわりに

本研究では、顔ランドマーク検出に対する新たな敵対的攻撃手法として、ブラックボックス環境における方向探索に基づく Blackbox Landmark Breaker を提案した。提案手法はランドマーク検出器の内部勾配を必要とせず、出力ヒートマップのみに基づいて撮動を更新することで、実際のシステムにおいても適用可能性が高いことを示した。実験結果から、ホワイトボックス手法と比較して精度は劣るもの、検出結果に明確な影響を与えられることが確認され、Deepfake 生成に対しても一部で耐性を低下させる効果が観察された。

一方で、本研究にはいくつかの課題が残されている。まず、近似勾配の推定には不確かさが伴うため、攻撃の一貫性や効率性に限界がある。また、撮動が画像全体に広がることで、視覚的自然さを損なうケースも見られた。さらに、本研究では単一のランドマーク検出モデルに対してのみ評価を行ったため、提案手法の一般性を論じるには十分ではない。

今後の展望としては、異なるランドマーク検出モデルや Deepfake 生成モデルを対象に評価実験を拡張し、提案手法の有効性と汎用性を検証する必要がある。加えて、ランドマーク間の依存関係を考慮した損失設計や、撮動を局所領域に制約する正則化の導入により、効果と自然さの両立を目指したい。さらに、人間の知覚実験や主観評価を取り入れることで、現実環境におけるリスク低減に資するより実践的な検証が可能になると考えられる。

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: 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/>
- [3] Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). DeepFake 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 Deep-Fake 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] 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).
- [16] HRNet-Facial-Landmark-Detection. <https://github.com/HRNet/HRNet-Facial-Landmark-Detection>.
- [17] 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.

- [18] 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.
- [19] Paszke, A., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.
- [20] 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.
- [21] Tanchenko, A. (2014). Visual-PSNR measure of image quality. *Journal of Visual Communication and Image Representation*, 25(5), 874-878.
- [22] 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.