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*, 林志訓**, 菊池浩明*

*明治大学大学院

**国立陽明交通大学

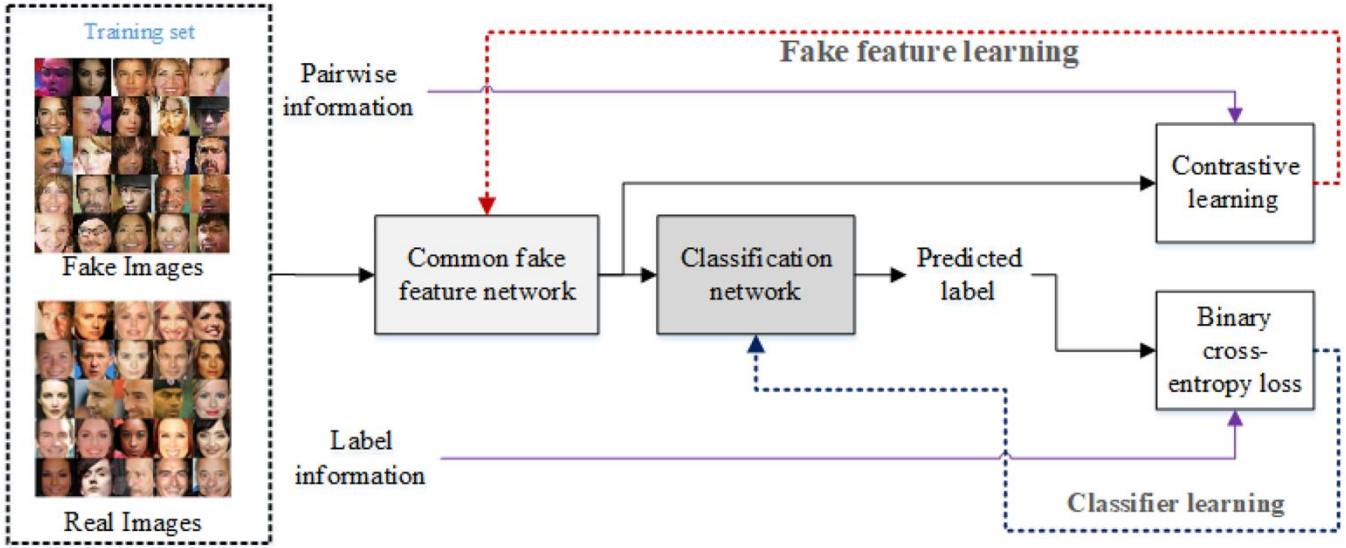
Deepfake問題

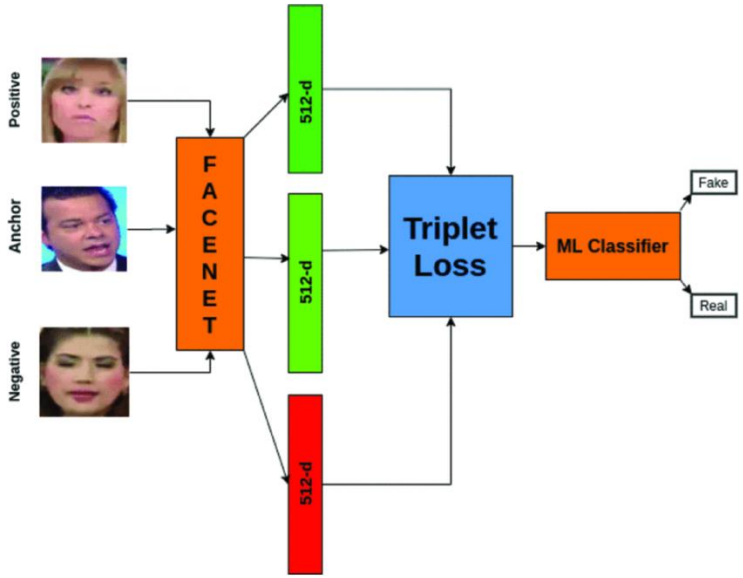
- BBCの報道によると，トランプの支持者がDeepfakeを利用し，黒人の支持者を増やそうとしていた

Reference: https://www.bbc.com/news/world-us-canada-68440150

従来の対策（検出）

- Hsu et.al. 2020 “DeepFake Image Detection Based on Pairwise Learning”：画像ベースの特徴学習を用いたフェイク画像の検出
- Akash et.al. 2020 “Detecting Deepfakes with Metric Learning”：ビデオベースのメトリック学習を用いたフェイクビデオの検出





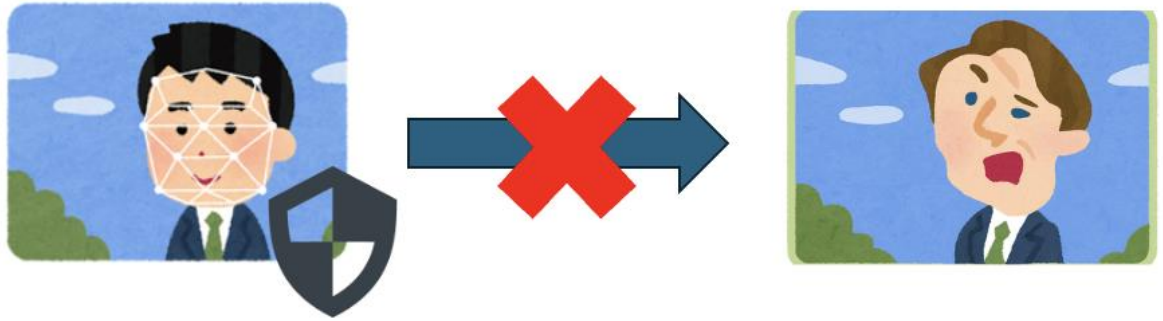
課題：**有害なフェイク画像が流出すると, 取り返しがつかない**

本研究の目的

- Deepfakeの**被害を未然に防ぐ**

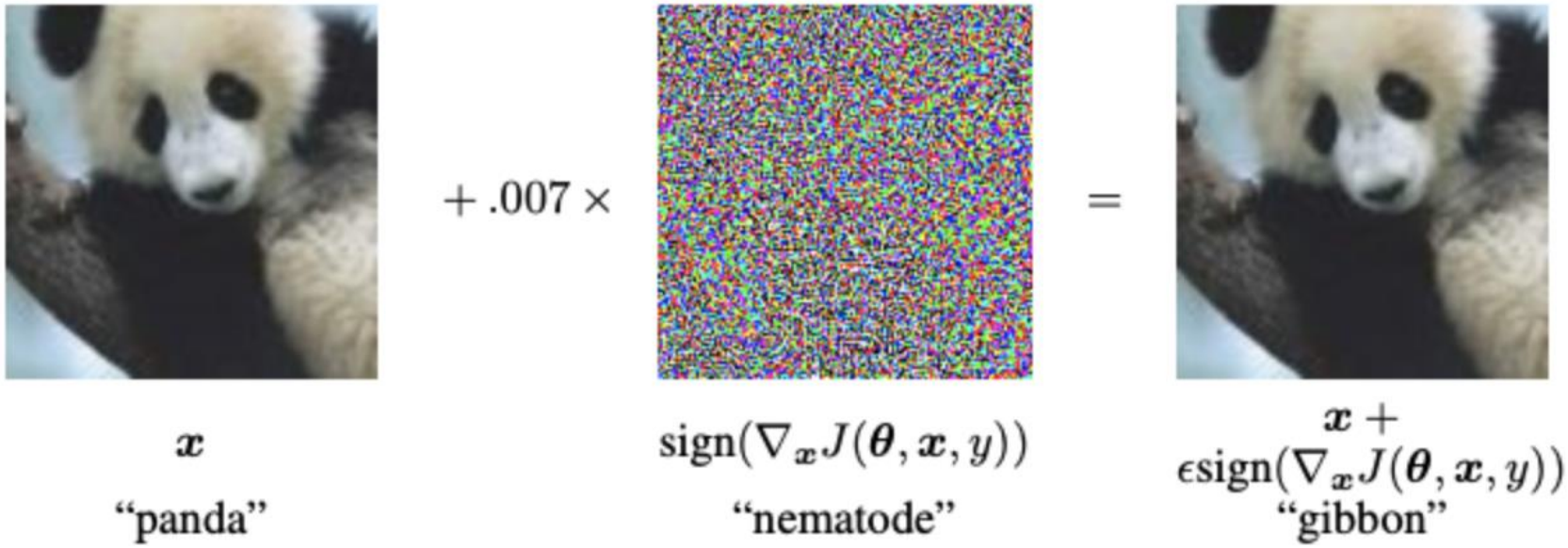
検出（事後）

本研究（事前）



本研究のアプローチ

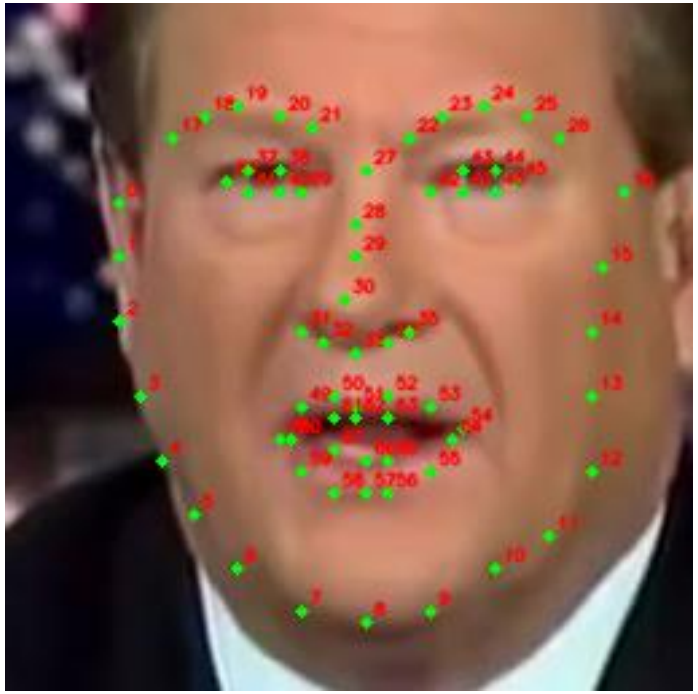
- 敵対的摂動



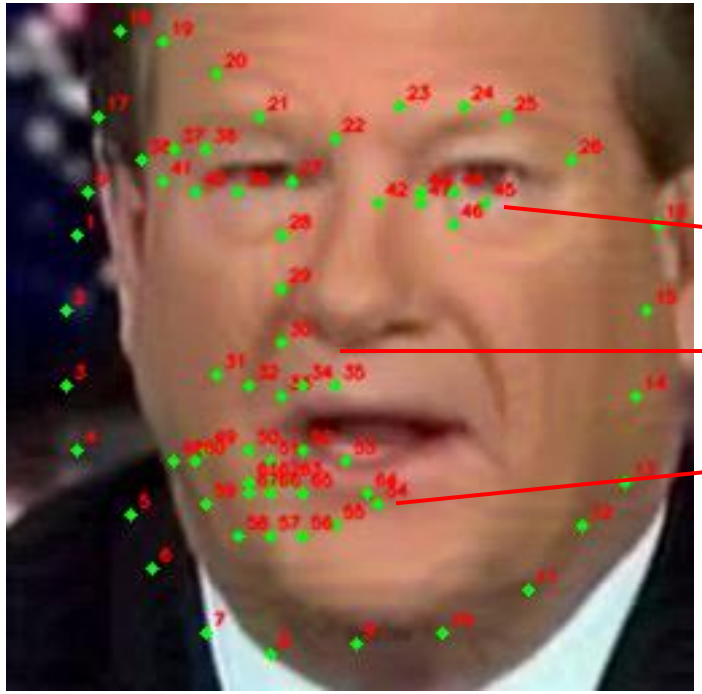
Reference: Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations. ICLR.

敵対的摂動を用いた防御

- Landmark Breaker[Li 2024]
- ランドマーク座標を狂わせる



通常ランドマーク



Landmark Breakerの効果

全ての
部位に
一様に
摂動

Landmark Breakerの課題

- 課題
  - 画質の劣化
  - 主観評価を行なっていない
- 仮説
  - **均一な摂動**を使ったから, 画質の劣化がひどい. **特定な部位にだけ**摂動を強めた方が知覚されにくい（画質を維持できる）
  - 全てのランドマークを撹乱しなくても効果が出る. **不均一な摂動**を使った方がランドマークに対する撹乱効果が増し, Deepfakeにされにくくなる

Landmark Breaker

本研究

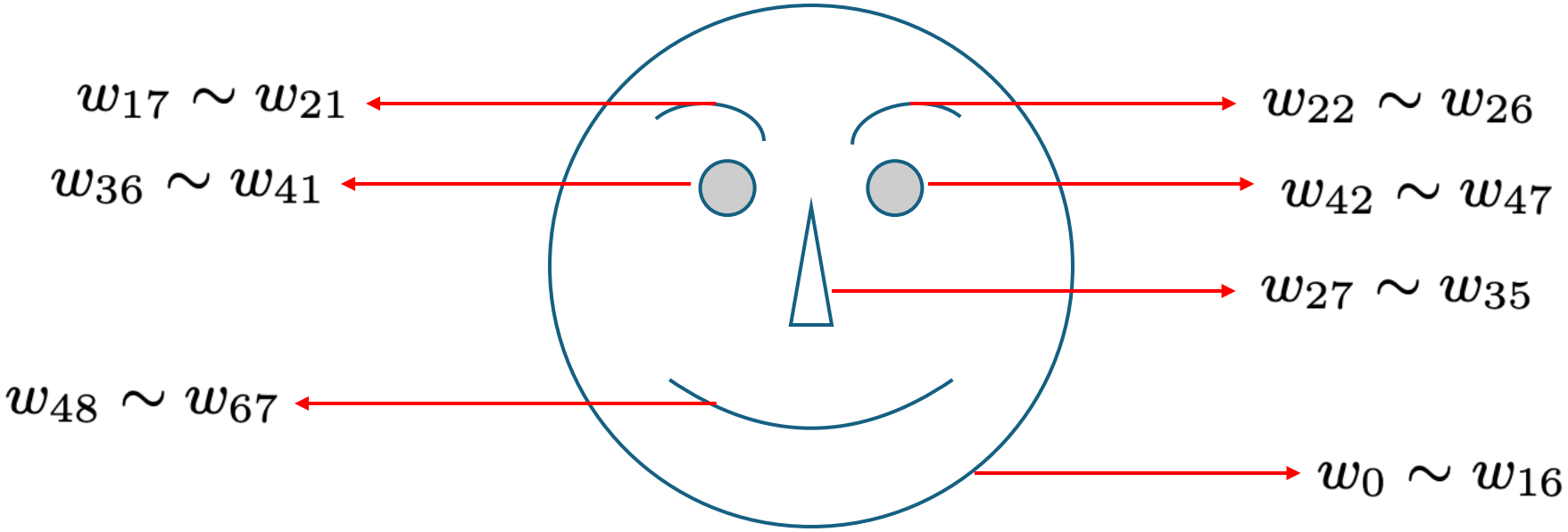
損失関数の工夫

Landmark Breaker

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

提案方式（Unbalanced Landmark Breaker, 略してULB）

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



例：
目元のランドマークだけを撹乱したい時は
w36〜w47＝1にして、
他は全部０にする

リサーチクエスチョン

- 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？

実験

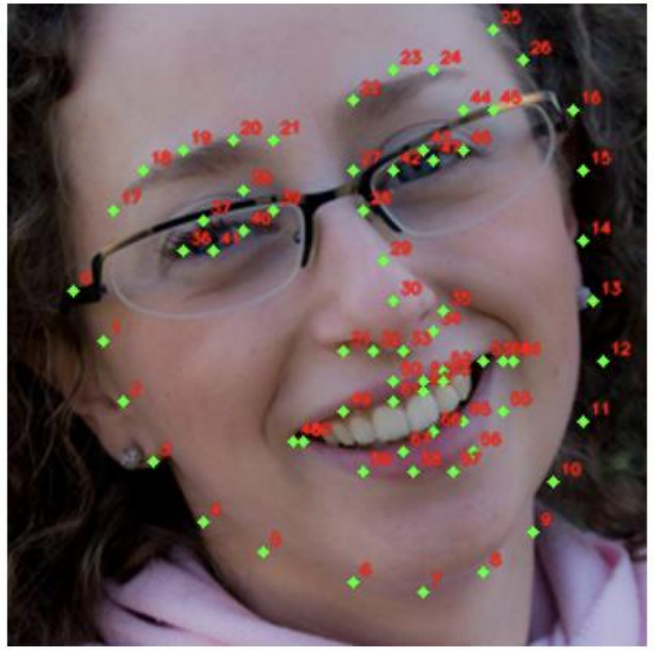
- 実験環境：
  - Ubuntu
  - AMD EPYCTM (2.8GHz, 256MB)
  - NVIDIA RTX A6000
- 顔画像データセット：
  - 300W [16]
  - FaceForensics++ [17]
- ランドマーク抽出モデル
  - **HRnet**(High-Resolution Network) [13]

結果1-1

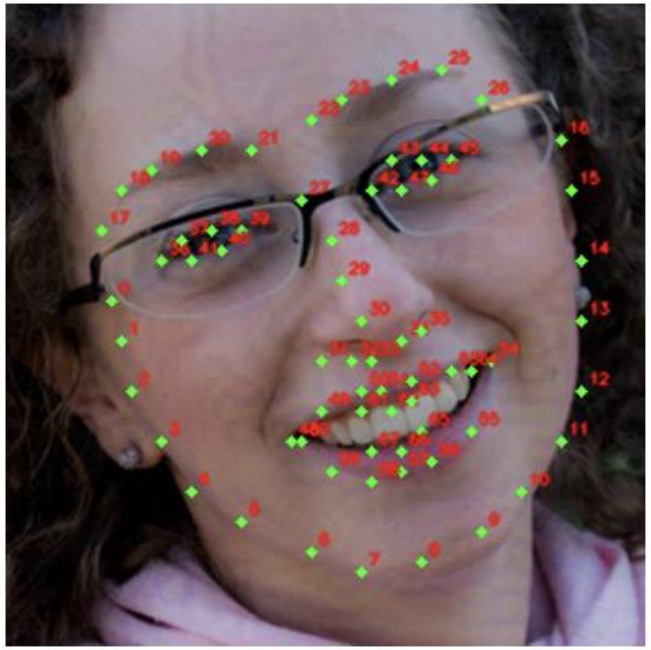
- 部分的な摂動は全体的な摂動よりも撹乱効果が強い

全てのランドマークに一様

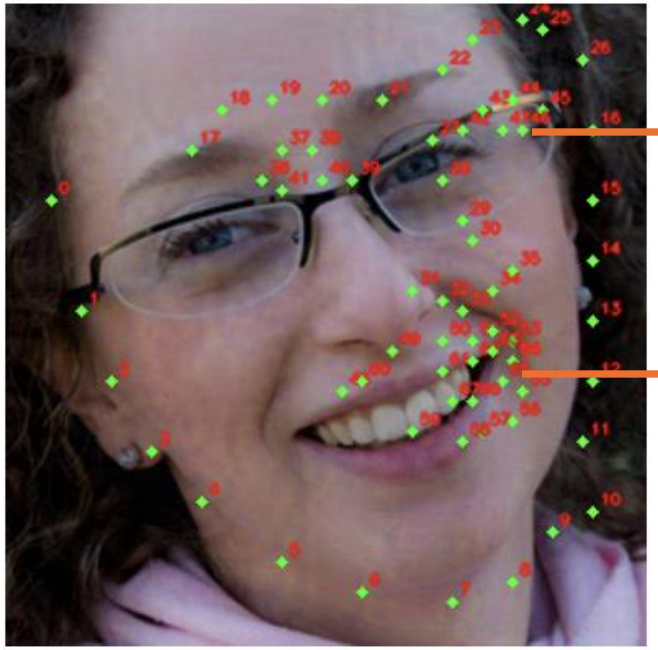
目元のランドマークにだけ



元画像



Landmark Breaker



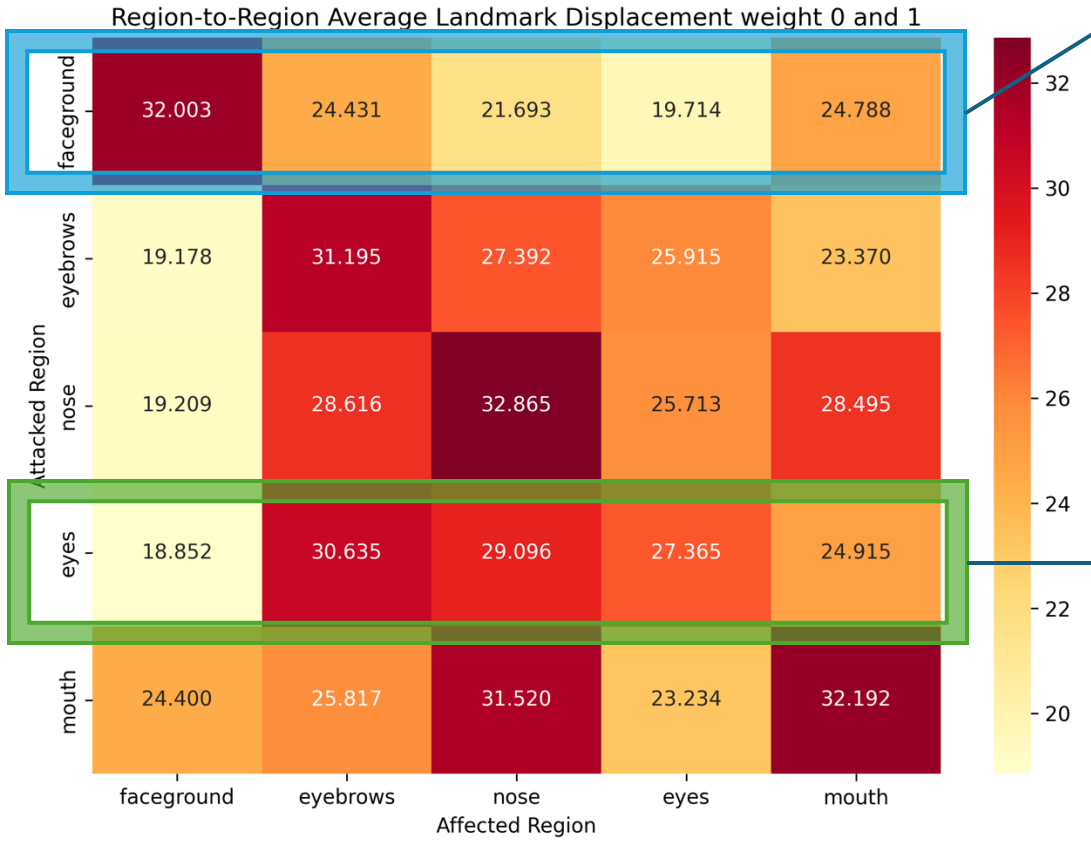
提案方式（ULB）

目元のランドマークの移動が激しい（good）

他の部位のランドマークにも影響している（bad）

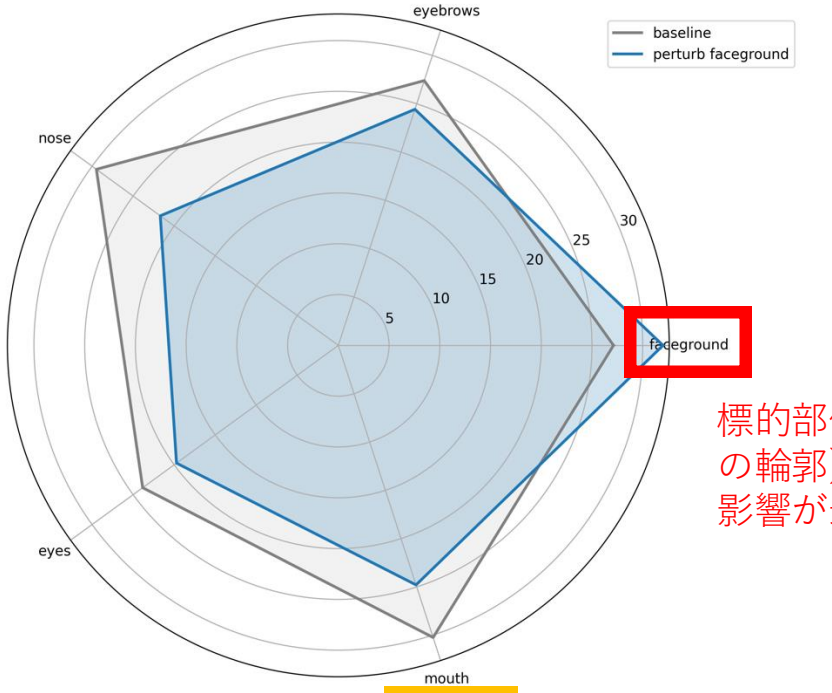
結果1-2

- 標的部位以外への影響



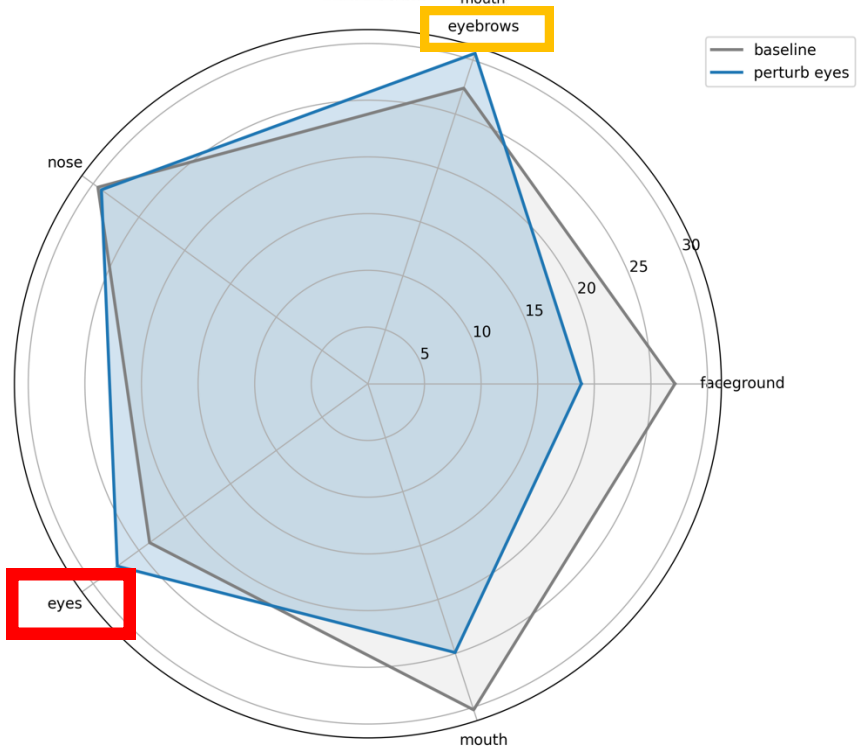
各部位のランドマークの平均移動距離（ピクセル）

Good



標的部位（顔の輪郭）への影響が最大

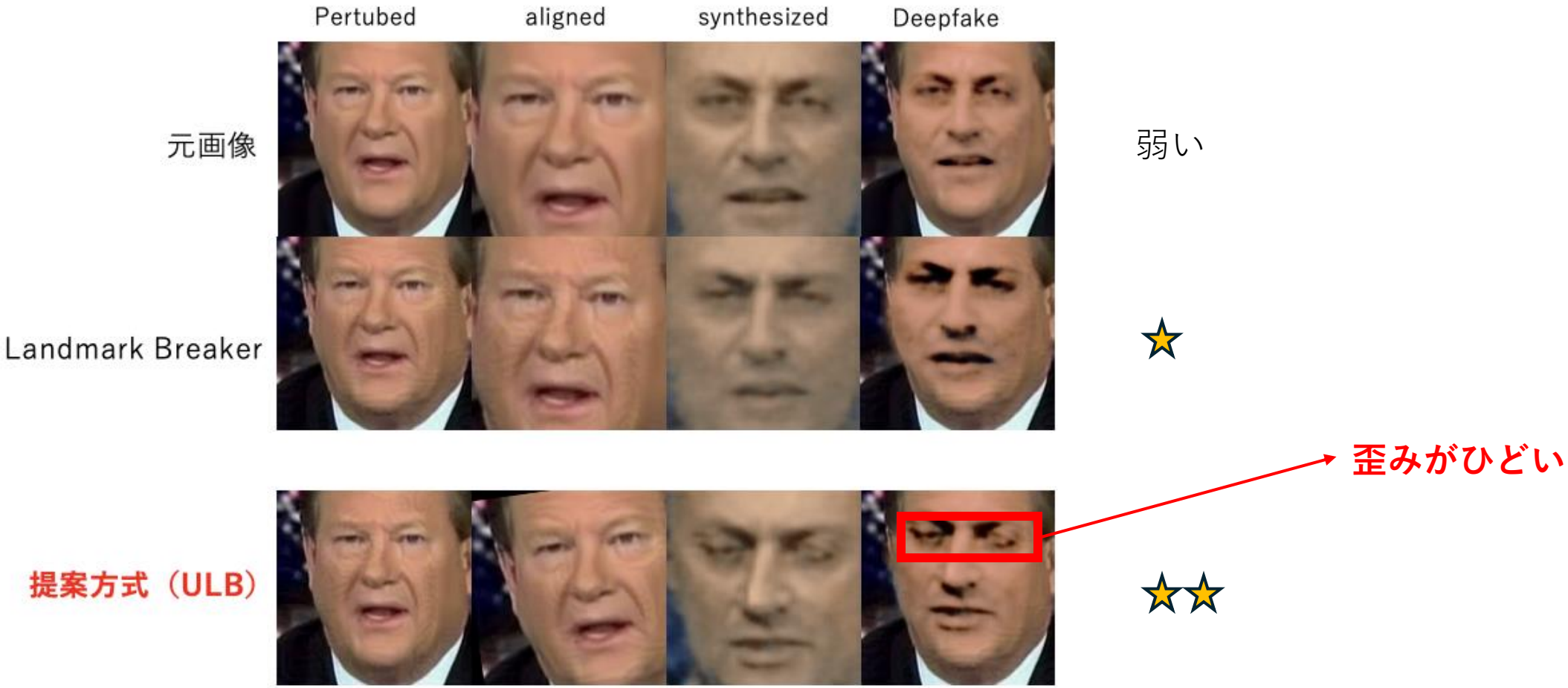
Bad



標的部位（目元）への影響が最大ではない

結果2

- Deepfake耐性



結果3-1（客観的評価）

- **PSNR**（Peak Signal to Noise Ratio）：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
- **SSIM**（Structural Similarity）：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

高いほど望ましい

提案方式

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

Landmark Breaker

提案方式の方が高いが,
差が小さい

結果3-2（主観的評価）

- 被験者は９回の実験中の７回（77%）で**提案方式の方が画質がいい**と答えた
- PSNRとの**相関関係がない**

質問：どちらの画像が画質悪いか

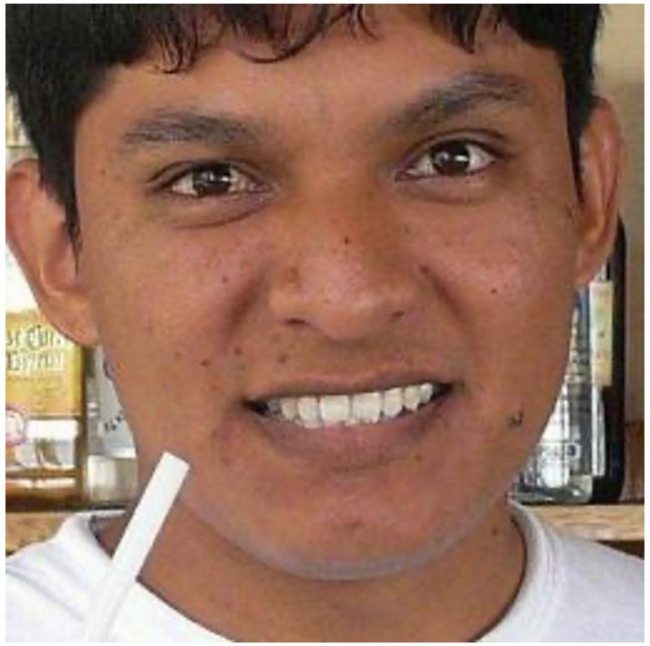


表 4: 主観的評価の結果

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

Landmark Breakerの方が画質が悪いと選択した人

結論

- 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Deepfake対策方式を提案した
- 部分的にランドマークを撹乱できることを確認したが, 標的ランドマーク以外にも影響を生じさせることが明らかになった
- 部分的にランドマークを摂動した画像は，従来の一様摂動画像よりもDeepfake耐性が高い
- 提案方式は，Landmark Breaker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とは言えない
- これからは**Deepfake耐性の評価**と**攻撃効果の改善**