ディープフェイクの検出で世界最高性能を達成
～SBIs でディープフェイク動画の高精度判定を可能に～

【発表概要】

ディープラーニングを使用し、画像を入れ替えて作られるディープフェイク動画は、政治家のような著名人の偽動画生成などに悪用され、世界中で問題となっています。
今回、東京大学大学院情報理工学系研究科の塩原 楓 大学院生と山崎 俊彦 准教授は、ディープフェイク動画を世界最高性能で真贋判定できる技術の実現に成功しました。本研究成果は、2022 年 6 月 19 日-24 日に米国で開催されるコンピュータビジョンの分野で最も著名な国際会議 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)で発表されます。

【発表のポイント】

◆与えられた動画像内の人物の顔がディープフェイク（注 1)かどうかを判定するタスクにおいて、既存研究を大きく上回る性能を達成するディープフェイク検出 AI を開発しました。
◆検出が難しい疑似フェイク画像を生成する新しい方法 SBIs (Self-Blended Images)を提案しました。SBIs で生成した画像をディープフェイク検出 AI にフェイク画像として学習させることで、実際のフェイク画像に対しても高い汎用性と頑健性で検出を行うことができます。
◆ディープフェイクは架空の事実を作り上げることで、デジタルメディアの信頼性を損ね、その価値を大きく低下させます。今回の提案は、高い精度でディープフェイク動画を検出することで、その悪用の根絶に貢献することが期待できます。

【発表内容】

＜研究の背景＞

ディープフェイクは、偽動画がフェイクニュースとして流布される、また犯罪に使われるなど悪用されるケースも多く、日本でも逮捕者を出すなど社会的な問題となっています。そのため多くの機関がディープフェイク検出の研究に取り組んでいます。
ディープフェイク検出研究では、深層学習（AI）ベースの検出器が高い性能を達成していますが、ディープフェイクの作り方は複数あり、多くの深層学習ベースの検出手法は訓練時に学習したタイプのフェイク画像・映像しか検出することができないことが知られています。それゆえ、検証時に入力されるディープフェイクの作成プロセスが訓練時に見たものと少し異なるだけでも、検出性能が大きく低下します。このような未知のディープフェイク画像への脆弱性の問題に対して、マイクロソフト社の研究者が 2020 年の CVPR で発表し、研究分野で代表的とされている手法では、ディープフェイク作成時に顔や背景を含む元画像と生成モデルによって生成された顔領域だけの合成画像をブレンドする際に生じる画像の不整合を再現した疑似フェイク画像を生成し、検出 AI にフェイクとして学習させることで汎用的な検出 AI の訓練を可能にしました。しかし、圧縮率が高く画像が潰れているものや高/低露光下のフェイク画像に対しては検出精度が低下する問題がありました。
今回研究グループは、これらの問題を解決するために、不整合が少ない疑似フェイク画像を生成する方法を提案しました。提案された疑似フェイク画像で訓練された検出 AI は、高圧縮率

や高/低露光下のフェイク画像のわずかな偽造痕跡を見逃さずに高い確信度を持って検出することができます（図 1）。

＜研究内容＞
前述のマイクロソフト社の手法では、似た顔の特徴点をもつ 2 枚の異なる人物画像をブレンドすることによって疑似的なフェイク画像を生成し、これらをフェイクとして検出 AI に学習させる際に、画像の色情報や画質を考慮しないで 2 枚の画像を合成するため、非常に検出が容易な疑似フェイク画像を生成していました。これに対して本研究では、1 枚の人物画像の色や周波数成分、画像サイズをわずかに変更した 2 枚の画像をブレンドすることによって生成される疑似フェイク画像、Self-Blended Images (SBIs)を提案しました（図 2）。SBIs はほぼ同一の画像をブレンドしているため、既存の疑似フェイク画像に比べて画像内の不整合が非常に少なく、SBIs を用いて訓練された検出 AI は実際のフェイク画像に対してもわずかな不整合があるだけでフェイクと判定することが可能になります。5 種類の評価用のデータセットで実施した既存手法との比較評価においても、今回の提案手法は 4 種類で世界最高性能を達成しました（表 1）。

＜社会的意義・今後の展開＞
ディープフェイクの技術は映画産業など新たな価値を生み出している一方で、スマートフォンのカメラアプリなどを使用することで誰でも簡単にフェイク動画を作ることが可能になるため、政治やポルノなど多くの偽動画で悪用され、社会問題に発展しています。本研究では、ディープフェイク動画の精度の高い検出を可能とし、喫緊の課題となっているディープフェイクの根絶に向けた有望な方法を提案しています。
また今後ディープフェイクの脅威が更に高まることが予想される中、研究グループはディープフェイク画像における時間的な不整合にも着目し、更に検出の精度を高めることを目指します。

【研究支援】
本研究の一部は、科研費「ディープ・フェイク撲滅に向けた検出技術基盤（課題番号：22H03640）」の支援により実施されました。

【発表学会】
学会名：IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022
学会公式 URL：https://cvpr2022.thecvf.com/
論文タイトル：Detecting Deepfakes with Self-Blended Images
著者：Kaede Shiohara and Toshihiko Yamasaki

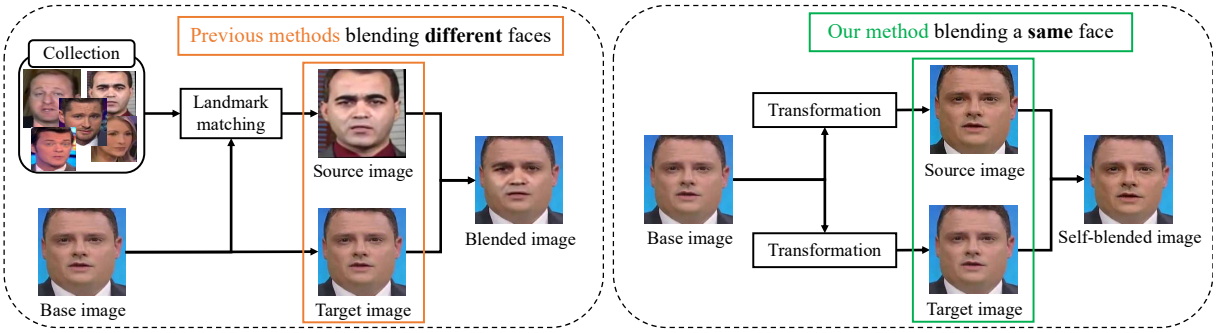
【発表者】
塩原 楓　　（シオハラ　カエデ）
（東京大学　大学院情報理工学系研究科電子情報学専攻　博士課程 1 年）
山崎 俊彦　（ヤマサキ　トシヒコ）
（東京大学　大学院情報理工学系研究科電子情報学専攻　准教授）

【用語解説】

(注 1) ディープラーニング（深層学習)などの人工知能技術を用いて、顔を他人と入れ替えたり偽の口元や表情を作ったりする技術、もしくはその技術を用いて作られた画像や動画。フェイクニュースを生成することが可能になってしまう他、有名人の顔を使った不適切な画像・映像が作られてしまうことも社会問題化している。なお、他に音声や文章を対象にしたフェイク技術もある。

(注 2) 合成や編集が行われていない生の画像

【添付資料】



(a) 既存手法 （マイクロソフト社提案手法 CVPR2020）　　(b) 提案手法 SBIs

図 1 ：既存手法と提案手法の違い。既存手法(a)は異なる 2 枚の自然画像をブレンドするが、提案手法(b)は同一の画像をわずかに変更した 2 枚の画像をブレンドする。

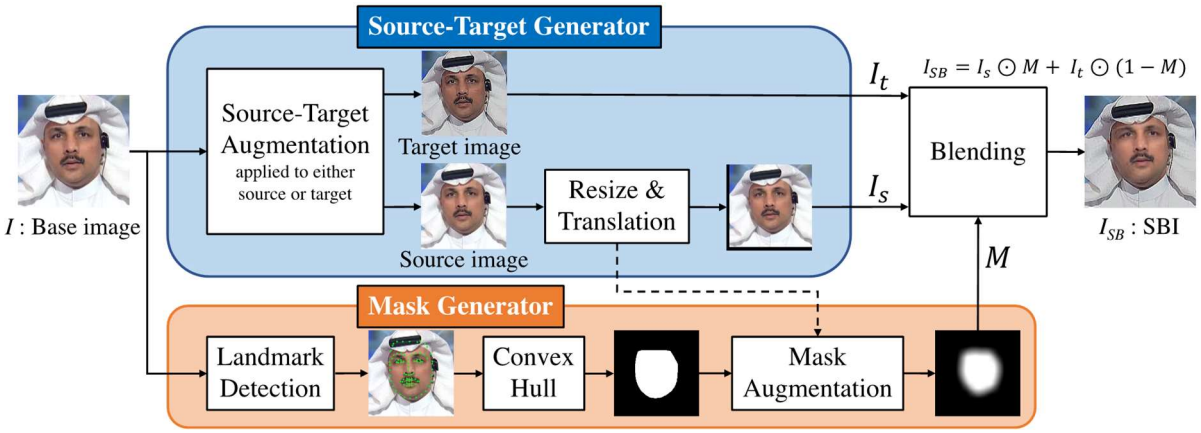


図 2： Self-Blended Images の生成フロー。ベース画像は Source-Target Generator (STG)と Mask Generator (MG)に入力される。STG では入力されたベース画像にいくつかの画像処理を用いて画像を変換して疑似的なソース画像とターゲット画像を生成する。ソース画像はその後リサイズと平行移動が施される。MG はベース画像の顔の特徴点からブレンドのためのマスク画像を生成する。ソース画像とターゲット画像をマスク画像でブレンドすることで Self-Blended Images が生成される。

Method	Input Type	Training Set		Test Set AUC (%)				
		Real	Fake	CDF	DFD	DFDC	DFDCP	FFIW
DSP-FWA	Frame	✓	✓	69.30	-	-	-	-
Face X-ray + BI	Frame	✓		-	93.47	-	71.15	-
Face X-ray + BI	Frame	✓	✓	-	95.40	-	<u>80.92</u>	-
LRL	Frame	✓	✓	78.26	89.24	-	76.53	-
FRDM	Frame	✓	✓	79.4	91.9	-	79.7	-
PCL + I2G	Frame	✓		<u>90.03</u>	**99.07**	67.52	74.37	-
Two-branch	Video	✓	✓	76.65	-	-	-	-
DAM	Video	✓	✓	75.3	-	-	72.8	-
LipForensics	Video	✓	✓	82.4	-	-	-	-
FTCN	Video	✓	✓	86.9	94.40*	<u>71.00</u>*	74.0	<u>74.47</u>*
EFNB4 + SBIs (Ours)	Frame	✓		**93.18**	<u>97.56</u>	**72.42**	**86.15**	**84.83**

表１：５種類のテストセットで実施した評価の結果。既存の手法と比較するとSBIsで生成した画像を使った今回の提案手法が（表内最下行：EFNB4＋SBIs）、DFDを除く4種のテストセットで最高性能を示していることがわかる。