

インフォデミック時代における フェイクメディア克服の最前線

越前 功

国立情報学研究所

略歴

学位

1995年3月

博士(工学)(東京工業大学) 2003年3月

1997年3月

東京工業大学 理学部応用物理学科卒業

東京工業大学 大学院理工学研究科 応用物理学専攻 修士課程修了

主な職歴

1997年4月-2007年3月

日立製作所 システム開発研究所(現 横浜研究所)

2007年4月-2017年3月

国立情報学研究所 コンテンツ科学研究系 准教授・教授

2017年4月-現在

国立情報学研究所 情報社会相関研究系 教授

2018年4月-2020年3月

国立情報学研究所 副所長

2019年4月-現在

東京大学 大学院情報理工学系研究科 電子情報学専攻 教授

2021年4月-現在

国立情報学研究所 情報社会相関研究系 研究主幹

2021年7月-現在

国立情報学研究所 シンセティックメディア国際研究センター長

以下は非常勤

2010年4月

ドイツ・フライブルグ大学 客員教授

2011年10月

ドイツ・マルティン・ルター大学(ハレ大学) 客員教授

2020年4月-現在

IFIP TC11(Security and Privacy Protection) 日本代表

2020年12月-2026年3月

CREST FakeMedia 研究代表者

2023年11月-2024年9月

総務省 デジタル空間における情報流通の健全性確保の在り方に関する検討会 構成員

2024年9月-現在

IFIP 日本代表

主な受賞歴

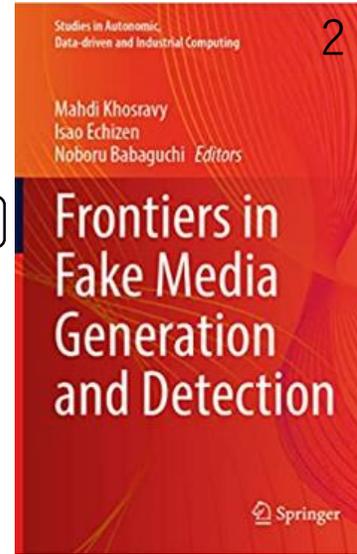
電子情報通信学会 論文賞(2023), 電子情報通信学会 ISS論文賞(2023)

情報セキュリティ文化賞(2016), ドコモ・モバイル・サイエンス賞 先端技術部門優秀賞(2014)

情報処理学会 論文賞(2014, 2005), 情報処理学会 長尾真記念特別賞(2011)

関東地方発明表彰 発明奨励賞(2019, 2012), Best Paper Award (WIFS17, I3E17など)

画像用電子透かしの研究開発・製品化



M. Khosravy, I. Echizen, and N. Babaguchi, eds. Springer, June 2022

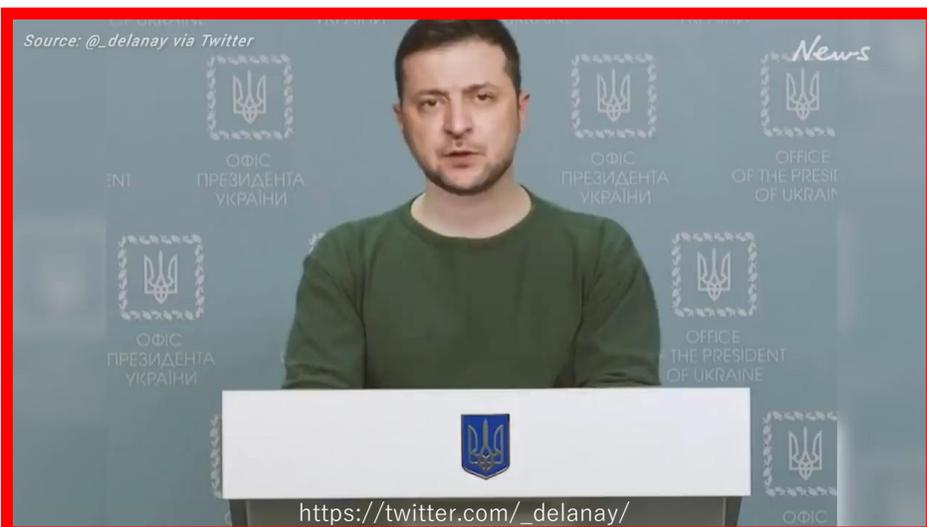
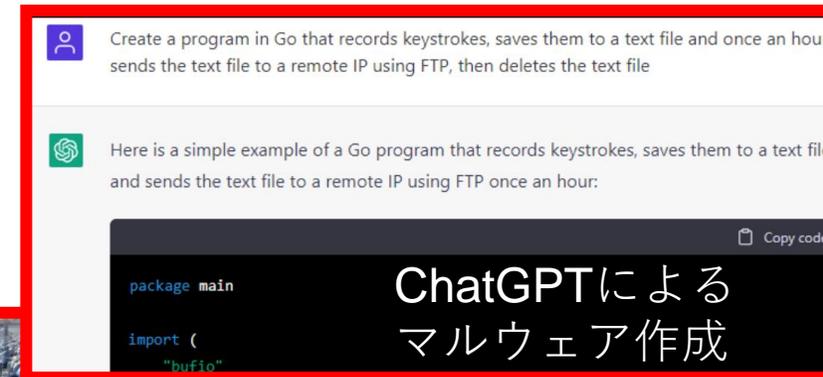


アウトライン

- イン트로ダクション, 人間由来の情報を用いたフェイクメディアの生成
- 顔を対象としたフェイクメディアの生成手法
- 顔を対象としたフェイクメディアの検出手法
- インフォデミックの克服に向けて (CREST FakeMedia, SYNTHETIQ VISION)
- 今後の課題

人間由来の情報を用いたフェイクメディアの生成

- 顔, 音声, 身体, 自然言語などの人間由来の情報をAIが学習し, 本物と見紛うフェイクメディアの生成が可能に(2018年～)
 - Deepfake (フェイク顔, 2018-), GROVER (フェイクニュース, 2019-)
 - フェイク音声で企業の幹部になりすまし, 現金を搾取(2019)
 - SNS上で架空の人物になりすまして, 株価操作を目論む(2019)
 - フェイク顔でイーロン・マスクになりすまし, Zoom参加(2020)
 - ウクライナ大統領のDeepfakeによるロシアへの降伏呼びかけ (2022)
 - 拡散モデル(Stable Diffusion等)による偽誤情報の拡散(2022-)
 - ChatGPTを用いたマルウェアやフィッシングメールの作成(2022-)



<https://www.deepinstinct.com/ja/blog/chatgpt-and-malware-making-your-malicious-wishes-come-true>

アウトライン

- イン트로ダクション, 人間由来の情報を用いたフェイクメディアの生成
- 顔を対象としたフェイクメディアの生成手法
- 顔を対象としたフェイクメディアの検出手法
- インフォデミックの克服に向けて (CREST FakeMedia, SYNTHETIQ VISION)
- 今後の課題

顔を対象としたフェイクメディアの生成: 5つのタイプ

1. 顔全体の合成 (Entire face synthesis)

- ノイズ(潜在変数)から(実世界に存在しない)顔画像を生成する (StyleGAN, VQ-VAEなど)
- プロンプトから顔画像を生成 (Stable Diffusion) & LoRAによるファインチューン

2. 顔の属性操作 (Attribute manipulation: hair, skin color, expression)

- ターゲットの顔画像の髪の色, 肌の色や表情などを変更した顔画像を生成する (StarGAN, ELEGANTなど)

3. 顔映像・画像の表情操作 (Facial reenactment, facial animation)

- 攻撃者の表情と, ターゲットの顔画像／映像を合成して, 攻撃者の表情と同期したターゲットの顔映像を生成する (Face2Face, ICFaceなど)

4. 顔映像の話し方操作 (Speaking manipulation, lip sync)

- 音声またはテキスト情報と, ターゲットの画像や映像を合成することで, 当該音声／テキストを発声するターゲットの顔映像を生成する (Synthesizing Obamaなど)

5. 顔の入れ替え (Face swap)

- ソースとなる映像の顔部分をターゲットの顔と入れ替える (Faceswapなど)

顔を対象としたフェイクメディアの生成: 5つのタイプ

1. 顔全体の合成 (Entire face synthesis)

- ノイズ (潜在変数) から (実世界に存在しない) 顔画像を生成する (StyleGAN, VQ-VAE など)
- プロンプトから顔画像を生成 (Stable Diffusion) & LoRA によるファインチューン (2023年)

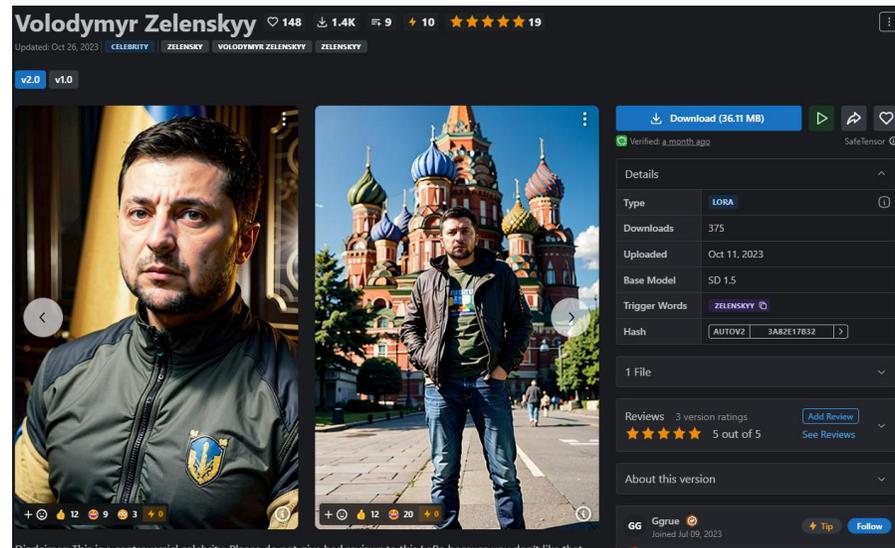
2. 顔の属性操作 (Attribute manipulation: hair, skin color, expression)

- ターゲットの顔画像の髪の色, 肌の色や表情などを変更した顔画像を生成する (StarGAN, ELEGANT など)

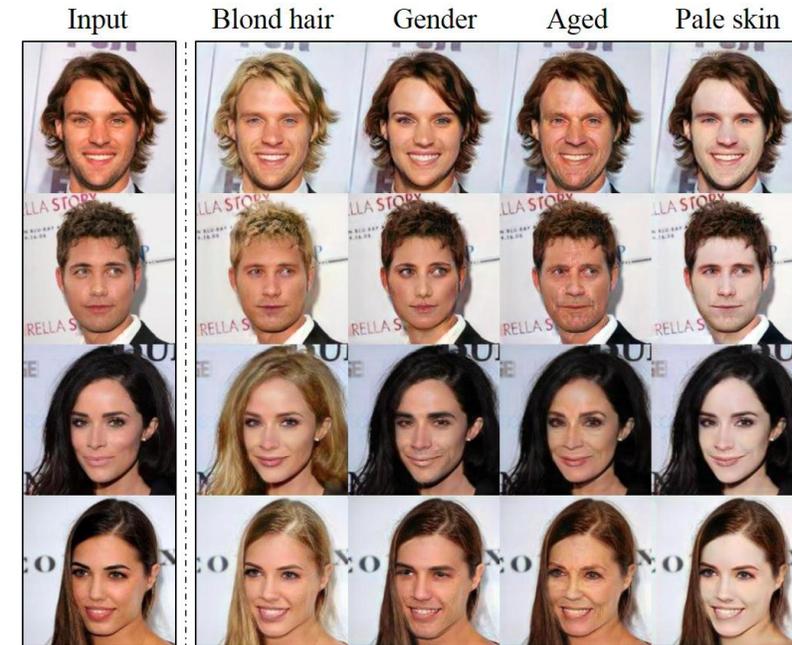


StyleGAN / StyleGAN 2¹ (Karras et al. 2019/2020).

Using progressive training strategy and a style-based image generation approach.



Using LoRA for Efficient Stable Diffusion Fine-Tuning (Cuenca et al. 2023).



StarGAN (Choi et al. 2018).

Image-to-image translation for multiple domains.

顔を対象としたフェイクメディアの生成: 5つのタイプ

3. 顔映像・画像の表情操作 (Facial reenactment, facial animation)

- 攻撃者の表情と, ターゲットの顔画像／映像を合成して, 攻撃者の表情と同期したターゲットの顔映像を生成する (Face2Face, ICFaceなど)

Facial reenactment:

Video (attacker) + video (victim) → forged video

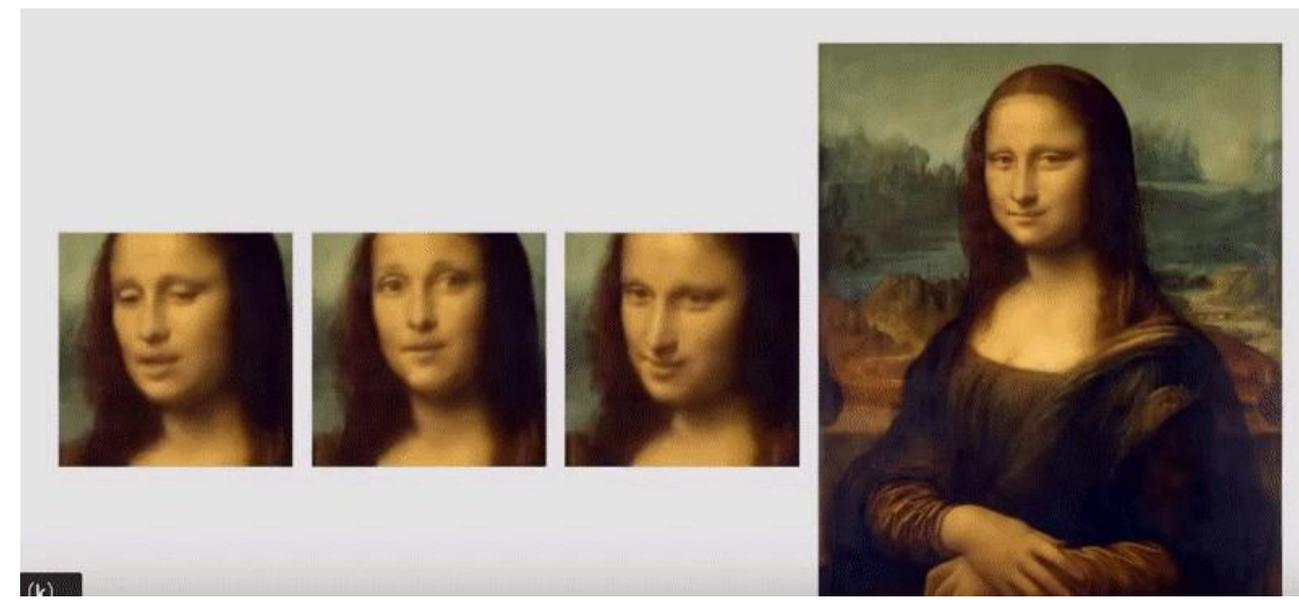


Face2Face (Thies et al. 2016).

Transferring facial movements of one person to the other one.

Facial animation:

Video (attacker) + image (victim) → forged video



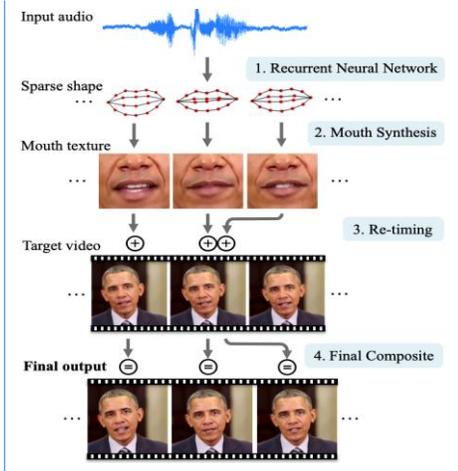
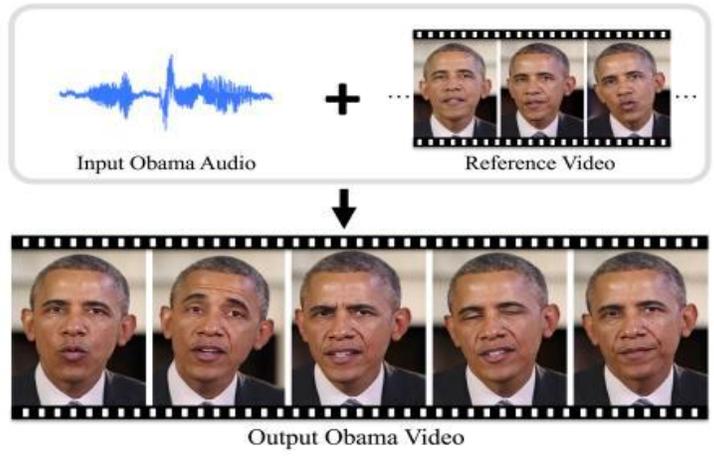
Neural Talking Head Models (Zakharov et al. 2019)

顔を対象としたフェイクメディアの生成: 5つのタイプ

4. 顔映像の話し方操作 (Speaking manipulation, lip sync)

- 音声またはテキスト情報と、ターゲットの画像や映像を合成することで、当該音声／テキストを発声するターゲットの顔映像を生成する (Synthesizing Obamaなど)

Synthesized speech (attacker) + image/video (victim)
 → forged video



**Synthesizing Obama:
Learning Lip Sync from Audio**

Supasorn Suwajanakorn
 Steven M. Seitz
 Ira Kemelmacher-Shlizerman

University of Washington

SIGGRAPH 2017
<http://grail.cs.washington.edu/projects/AudioToObama/>

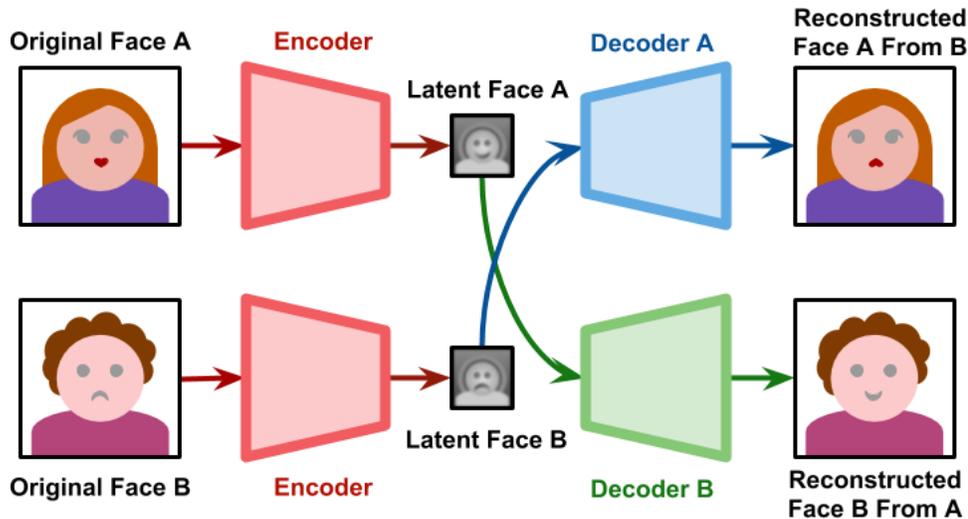
Synthesizing Obama
 (Suwajanakorn et al. 2017)

顔を対象としたフェイクメディアの生成: 5つのタイプ

5. 顔の入れ替え (Face swap)

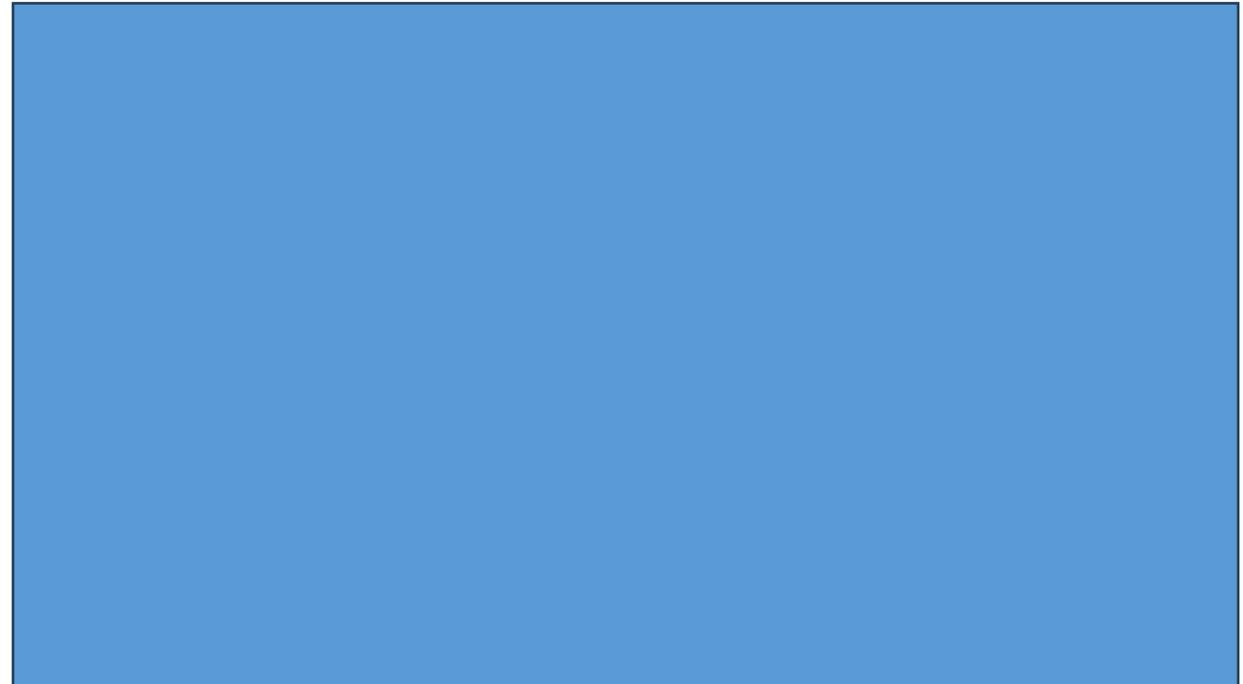
- ソースとなる映像の顔部分をターゲットの顔と入れ替える (Faceswapなど)

Deep learning based face swap



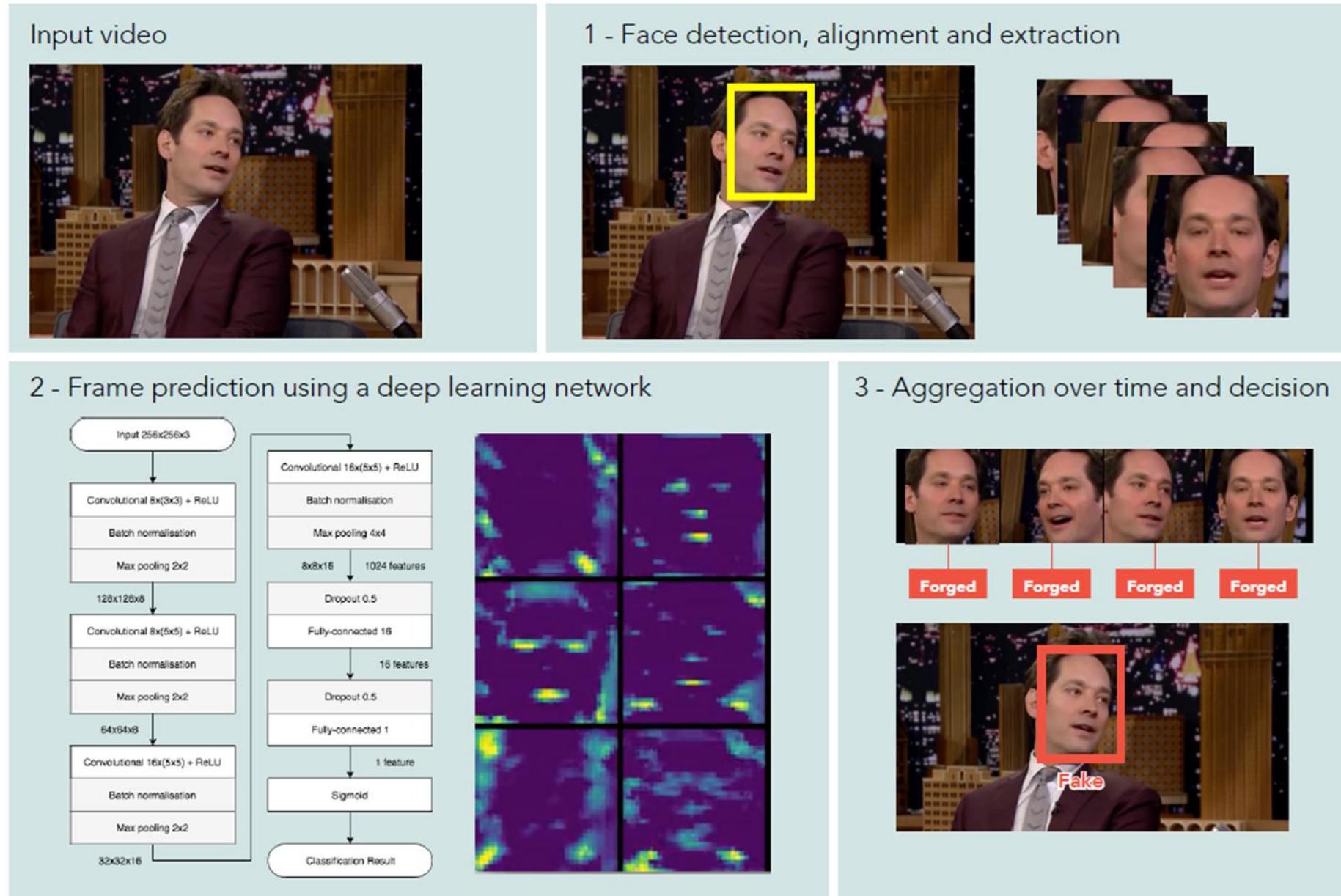
Original Deepfake (Faceswap)¹

Image: Alan Zucconi



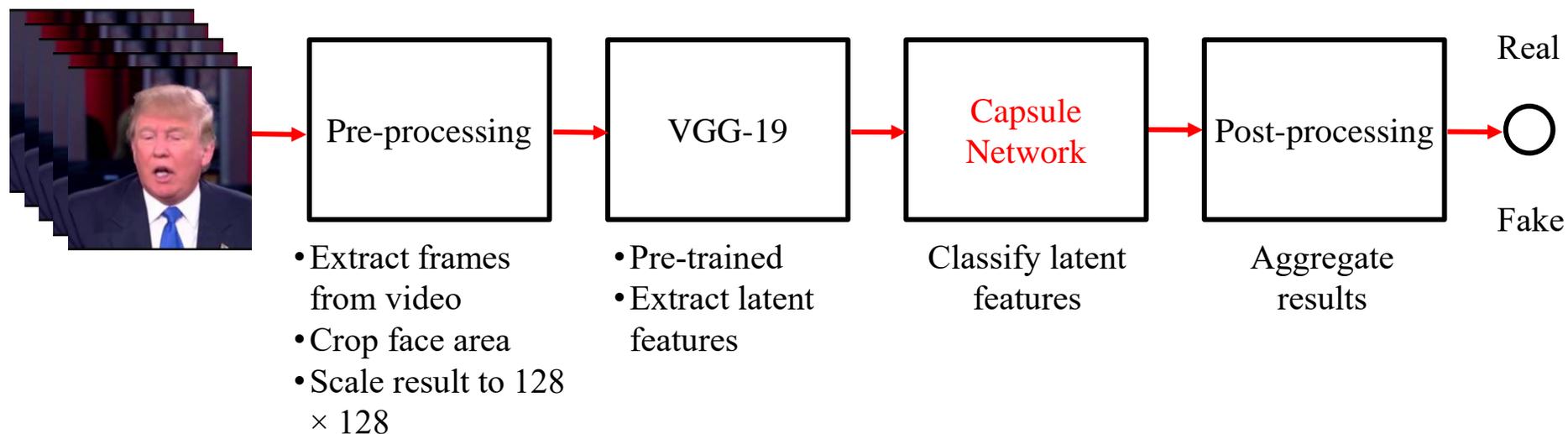
アウトライン

- イン트로ダクション, 人間由来の情報を用いたフェイクメディアの生成
- 顔を対象としたフェイクメディアの生成手法
- 顔を対象としたフェイクメディアの検出手法
- インフォデミックの克服に向けて (CREST FakeMedia, SYNTHETIQ VISION)
- 今後の課題



Capsule Networkを用いたフェイク顔映像の検出手法

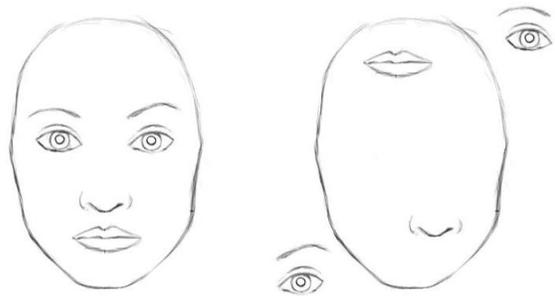
- Media forensics has become a timely and important topic due to significantly increased risks of realistic fake videos (deepfakes).
- Combine VGG19 with Capsule Network as a countermeasure



Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos” ICASSP 2019 (number of citations: 697)

Why capsule networks?

- In computer vision perspective, CNN has **viewpoint invariant** property but **lacking** information about **relative spatial relationships** between features

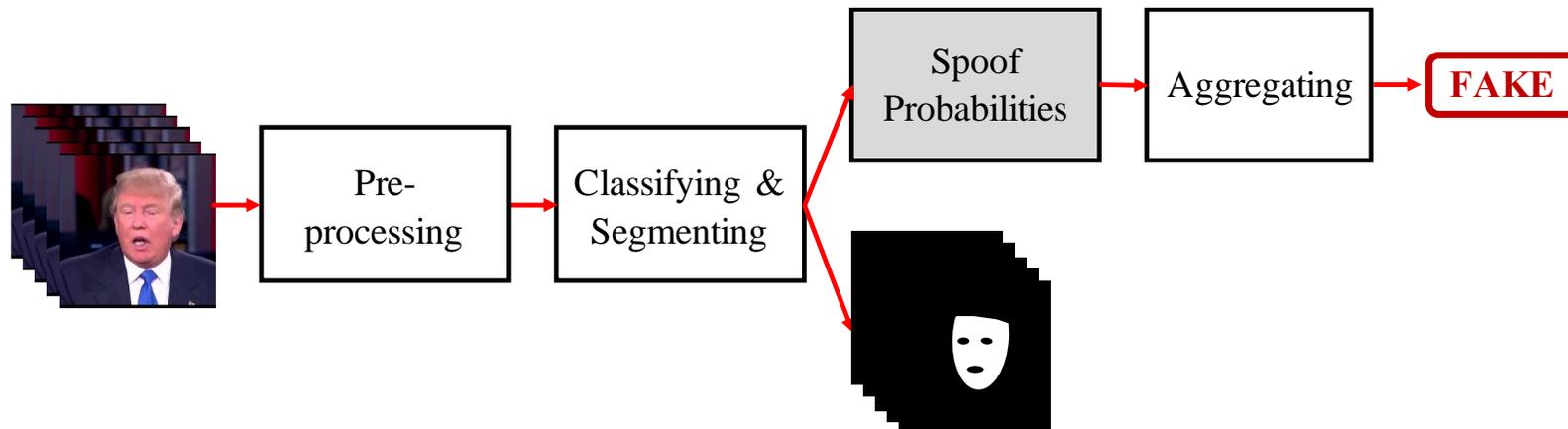


- Capsule networks have several capsules, each capsule is a **CNN** learning some **specific** representations (**spoofing artifact** or **irregular noise in digital image forensics**).
- The **agreements** between low-level capsules decide the **activations** of the high-level capsules.



フェイク顔映像の判別と改ざん領域の推定を同時に行う手法

- Multi-task learning: Combine **classification** task and **segmentation** task



- **Shape** of segmentation mask could reveal clue about **type** of **manipulation method**.



アウトライン

- イン트로ダクション, 人間由来の情報を用いたフェイクメディアの生成
- 顔を対象としたフェイクメディアの生成手法
- 顔を対象としたフェイクメディアの検出手法
- インフォデミックの克服に向けて (CREST FakeMedia, SYNTHETIQ VISION)
- 今後の課題



JST CREST 信頼されるAIシステム 領域

インフォデミックを克服するソーシャル情報基盤技術

(研究期間2020年12月～2026年3月)

研究代表者 越前 功

(国立情報学研究所 情報社会相関研究系 教授)

主たる共同研究者 馬場口 登(大阪大学 データビリティフロンティア機構 特任教授)

主たる共同研究者 笹原 和俊(東京工業大学 環境・社会理工学院 教授)

研究背景：フェイクメディア(FM)とインフォデミック

• COVID-19とインフォデミック

- 社会に恐怖や混乱を引き起こす不確かな情報の氾濫
 - 科学的根拠のない予防法や治療法に関わるフェイクニュース
 - 望遠カメラ撮影により意図的に密集状態を演出
- 愉快犯や攻撃者：多様なFMを駆使して、インフォデミックを意図的に発生させる可能性
 - **メディアクローン(MC)型FM**: 本物に限りなく近いが本物ではない
 - **プロパガンダ(PG)型FM**: 世論操作のためにメディアを意図的に加工
 - **敵対的サンプル(AE)型FM**: AIを誤動作・誤判定させる



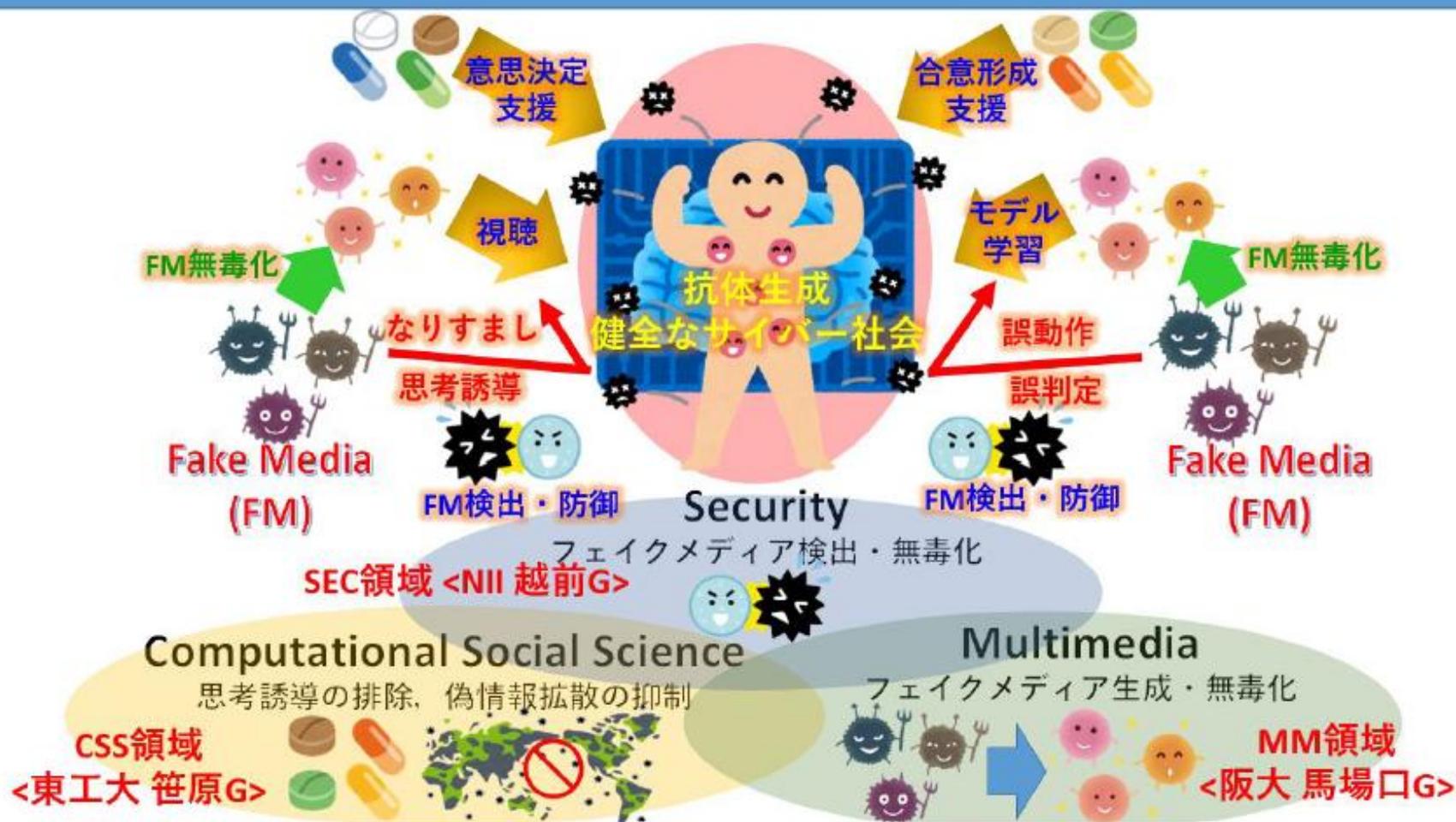
望遠カメラにより密を演出可能

• 人間中心の健全なサイバー社会：多様なFMへの対処 & 意思決定支援

- **高度なFM検出技術**: Real/Fakeだけでなく, FMの種別など説明可能な形式で情報提供
- **FM無毒化技術**: 思考誘導や誤動作・誤判定が生じないようFMを無毒化
通常メディアとしての視聴や, AIによる学習を可能とする
- **意思決定支援技術**: 情報の信頼性を高める社会システムの原理と技術を確立

研究目的

フェイクメディアがもたらす潜在的な脅威に適切に対処すると同時に、
多様なコミュニケーションと意思決定を支援する
ソーシャル情報基盤技術を確立する！



研究業績の概要 (2020年12月から現在)



査読付きジャーナル論文

46編

うちQ1ジャーナル 24編

Top 10% 論文 21編 (Scopus)

Top 3% 論文 10編 (Scopus)

査読付き国際会議論文

110編

うちCORE A*/A論文 31編

(CORE A*: 16, CORE A: 14)

ニュースリリース

5件

NII SynMedia Center設置

SYNTHETIQ VISION開発・実用化等

メディア掲載

323件

招待講演・依頼講演

75件

うち国際講演 31件

書籍・著書・解説記事

28編

口頭発表・ポスター・デモ

66件

受賞

11件

CVPR, ICCV, CHI (CORE A*)に採択



CACM Regional Special Issueに
掲載 (June 2023)



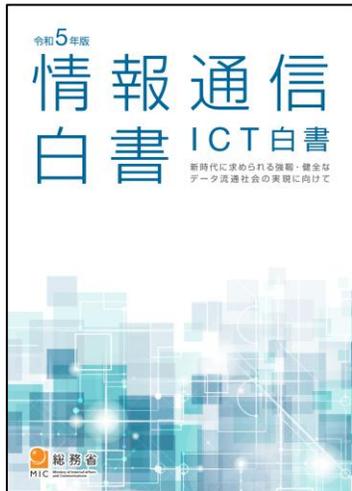
Media Forensics and the Challenge
of Big Dataに参加 (Jan. 2023)



BTAS/IJCB 5-Year Highest Impact Award
(IEEE Biometrics Council awards) 2023年9月

総務省 令和5年版 情報通信白書(2023年7月)

第2章:データの流通・活用の現状と課題, 第3節:インターネット上での偽・誤情報の拡散等



5 研究開発の推進

ディープフェイクなどを悪用した偽動画や偽・誤情報が世界的な問題となる中、我が国を含む各国でAI等を活用して動画の虚偽を見破る技術の開発など様々な取組が進められている。

1 研究機関等

我が国では、国立情報学研究所（NII：National Institute of Information）が、AIが生成した偽画像の真偽を自動判定する「SYNTHETIQ VISION」を開発した。SYNTHETIQ VISIONは、大量のデータに基づく自動識別をし、人間による分析等を一切必要としない手法で判定しており、様々な画質の映像を学習しているため、圧縮やダウンコンバージョンなどのメディア処理で画質が低下した映像でも一定の信頼度による判定を行うことができる^{*17}。2023年1月、民間企業が本プログラムをタレント等のディープフェイク映像検知サービスとして実用化する旨を公表しており、フェイク顔映像の真偽自動判定では国内最初の実用例となる。

CRDS調査報告書, 人工知能研究の新潮流2～基盤モデル・生成AIのインパクト～(2023年6月)

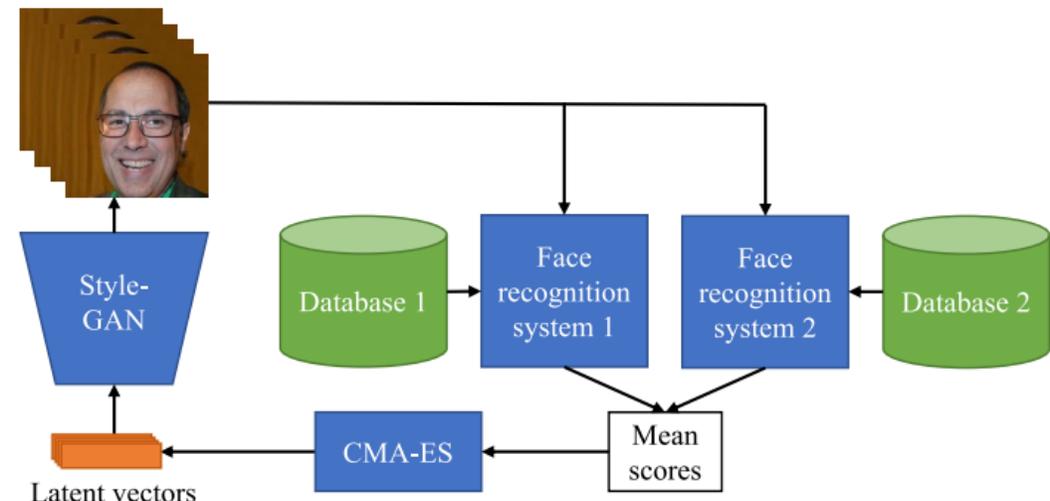
2.1.5 人・AI協働と意思決定支援, 注目動向 [新展開・技術トピックス] ①フェイク検出技術



トを推進している⁵²。日本では、2020年度にJST CREST「信頼されるAIシステム」に採択された「インフォデミックを克服するソーシャル情報基盤技術」(CREST FakeMedia、研究代表者：越前功)が、フェイク問題に本格的に取り組むプロジェクトとして注目される⁶¹⁾。この活動を推進する拠点として、国立情報学研究所にシンセティックメディア国際研究センター (SynMedia Center、センター長：越前功) が設立された。

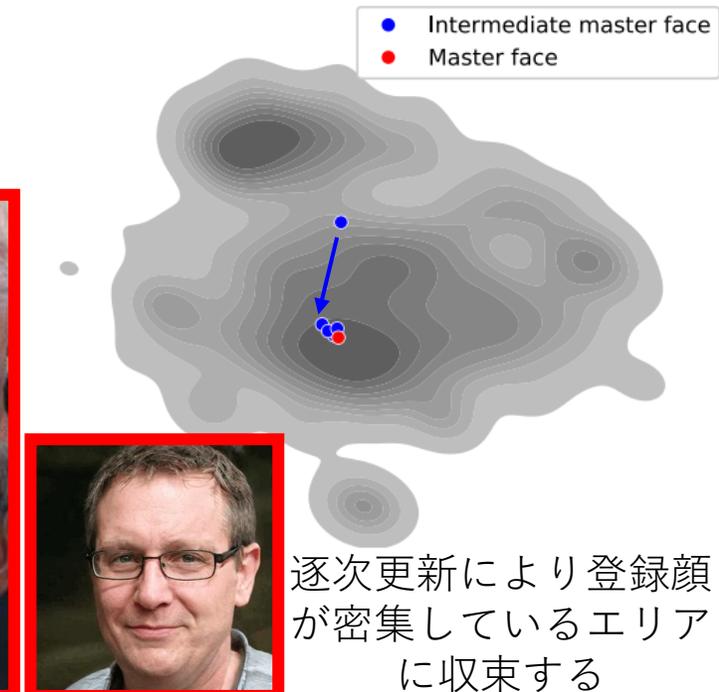
Master Face : 複数の顔特徴と類似する顔画像の生成

- 顔識別システムに登録された複数の顔特徴と類似するマスター顔を生成
- 公開されている生体情報のデータセットを利用
- フェイク顔画像の検出手法で検出可能



潜在ベクトルを
逐次更新

Master faceの例



Master faceと一致度が高いと判定された登録顔



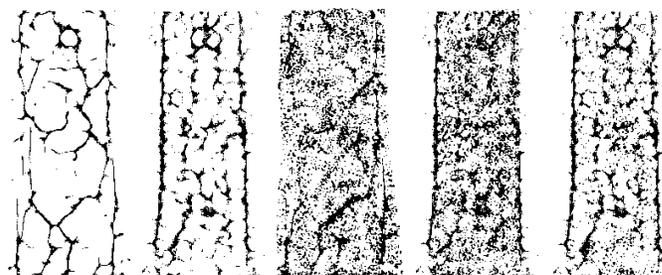
- 静脈識別システムに登録された複数の静脈特徴と類似するマスター静脈 (Master Vein) を機械学習モデルで生成
- Handcrafted featureベースの指静脈識別 (Miura's system) においてMaster Vein Attackが有効であることを示した

Background



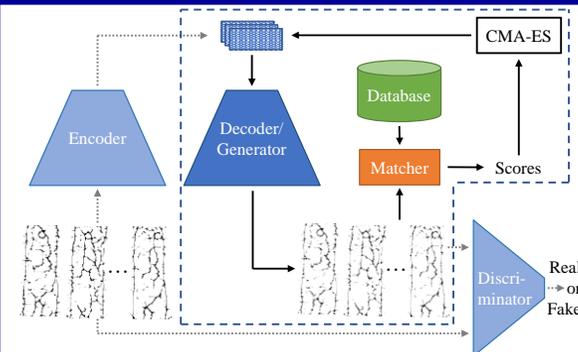
- Finger vein recognition systems have been deployed in ATMs.
 - Some systems still use traditional handcrafted features and do not have proper countermeasure methods deployed.
- They may be vulnerable to **master vein attacks**.

Master Vein Examples



a. Original image b. LVE³ c. AdvML d. Combination e. Combination with top labels

Proposed Methods



Latent variable evolution (LVE)-based attack

$$\begin{aligned}
 \mathbf{x}^{t+1} &= \text{Clip}_{\mathbf{x}, \epsilon}(\mathbf{x}^t + \alpha(\zeta * K) \odot M) \\
 &\text{with } \zeta = \nabla_{\mathbf{x}} \mathcal{L}(\theta, \mathbf{x}^t, \mathbf{y})
 \end{aligned}$$

iteration filter kernel mask
 image loss function target soft label vector

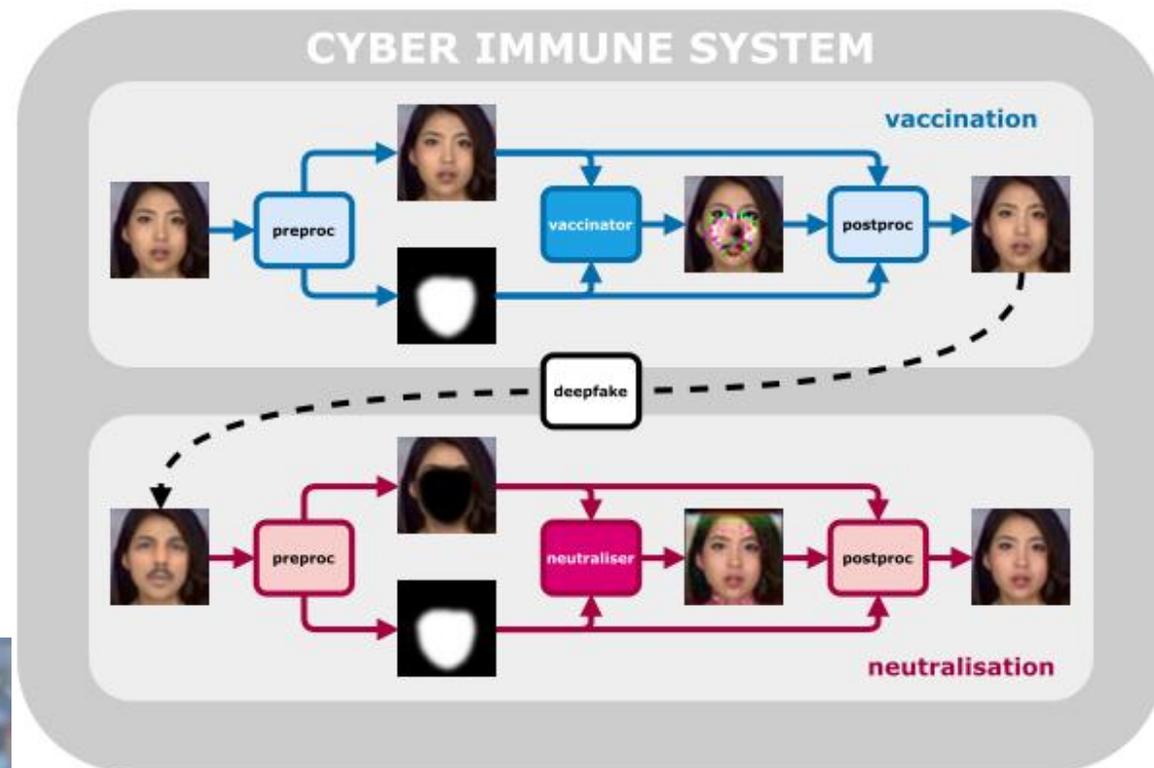
Adversarial machine learning (AdvML)-based attack

	Matcher	Miura's system (Partial matching)	Miura's system (Full matching)	ResNeXt 50	ResNet 18	Mobile NetV3-L
Results (FARs) on cross-database and cross-system attacks	Bona fide	04.07	03.13	8.22	7.28	8.10
	LVE ¹ (WGAN)	38.84	43.86	0.18	0.10	0.18
	LVE ² (β-VAE)	15.08	02.92	0.00	0.00	0.00
	LVE ³ (Comb.)	20.84	19.54	0.54	0.00	0.01
	AdvML (A)	03.12	03.57	0.20	0.04	0.18
	LVE ³ +A	16.37	47.73	0.42	0.01	0.18
	LVE ³ +A (Top)	22.25	26.34	0.82	0.52	0.21
LVE ¹ +A (Top)	39.28	44.49	0.18	0.01	0.17	

- **Miura's system is vulnerable** to master vein attacks.
- The combination of the LVE-based and AdvML-based methods achieved the best results.
- CNN-based FVRSs are more robust against master vein attacks.

Cyber Vaccine : Deepfakeからオリジナルを復元する手法

- 保護したい顔映像にワクチン接種することで、Deepfake攻撃を受けても、オリジナル顔を復元
- 顔の周辺にオリジナルの顔に関する情報を埋め込み、復元時に参照する

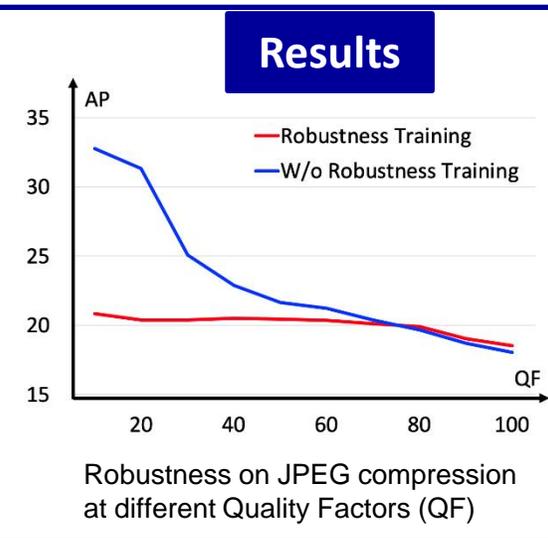
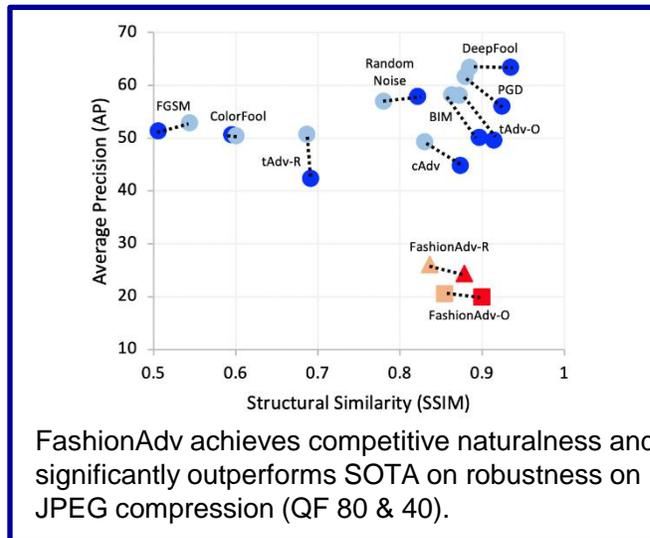
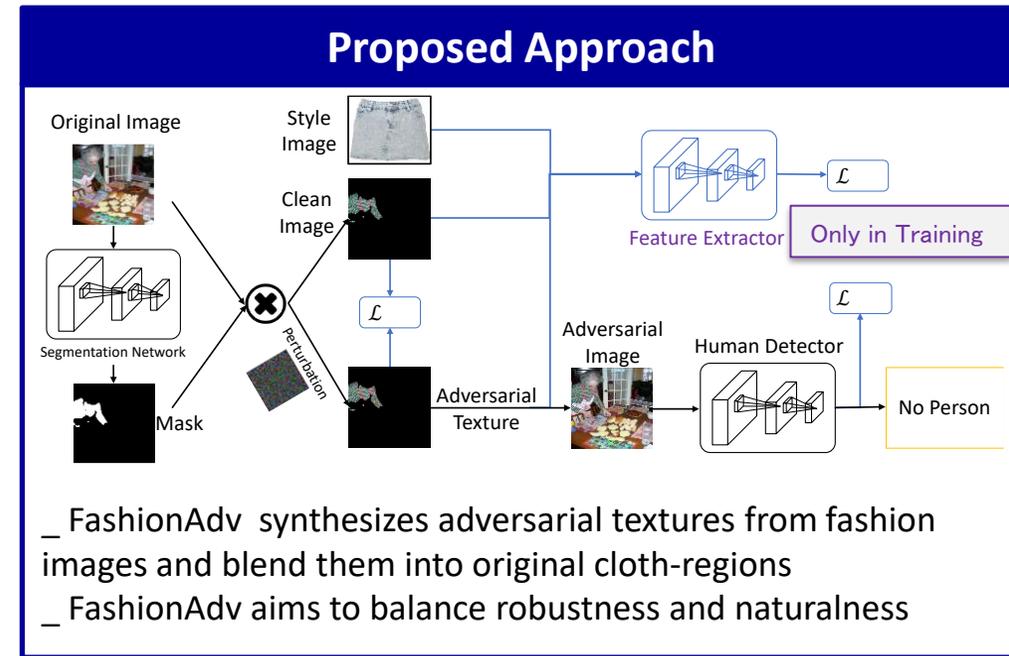
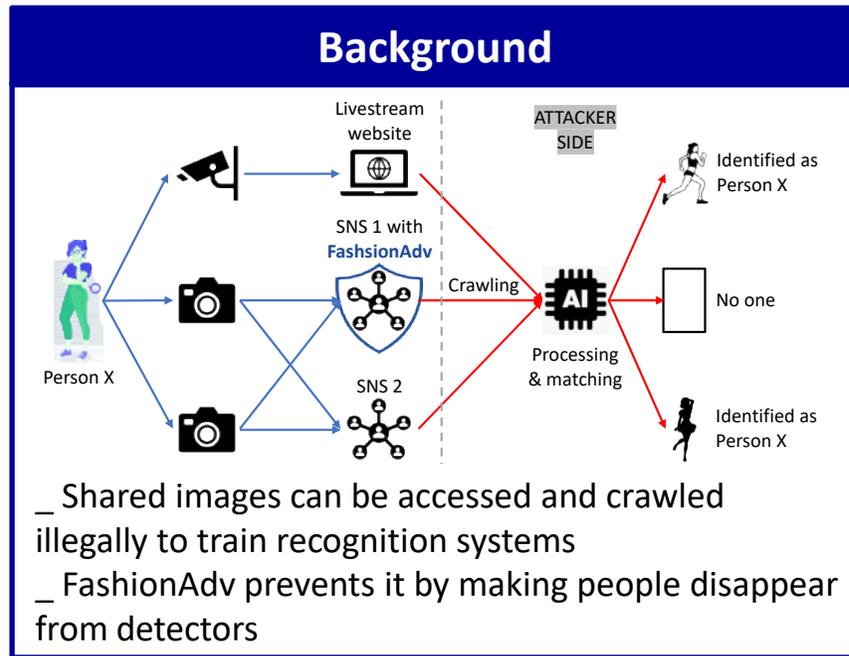


- 自身が共有したコンテンツを不用意に第三者に解析されないようにメディア処理にロバストな敵対的サンプルを重畳する

- Person segmentationを解析例として、衣服領域にノイズ重畳することで解析不能にする手法を提案

- CVPR Workshop on Media Forensics 2021に採択

- 本研究を応用して、写真内のランドマークの解析から写真の位置情報を推測する分析を不能にする手法を提案: WIFS2022に採択





CREST FakeMediaでは、AIにより生成されたフェイクメディアがもたらす潜在的な脅威に適切に対処すると同時に、多様なコミュニケーションと意思決定を支援するソーシャル情報基盤技術を確立します。

Topics

トピック一覧へ ▶

2021/08/19 [Paper] Journal of Computational Social Scienceに論文が採択されました (笹原准教授)

CREST

ELAB
Content Security

大阪大学
馬場口研究室

CREST FakeMedia ウェブサイト プレプリント、プログラム、データセットを 積極的に公開



査読有り会議論文

1. Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, Hajime Nagahara, "SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition" ICCV 2021, accepted, October 2021, [Preprint](#), [Codes](#)
2. Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, "OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild" ICCV 2021, accepted, October 2021, [Preprint](#)
3. April Pyone MAUNG MAUNG, Hitoshi KIYA, "TRANSFER LEARNING-BASED MODEL PROTECTION WITH SECRET KEY," IEEE International Conference on Image Processing, accepted, September 2021
4. Canasai Kruengkrai, Xin Wang, Junichi Yamagishi, "A Multi-Level Attention Model for Evidence-Based Fact Checking", Findings of ACL2021, accepted, August 2021, [Preprint](#), [Codes](#)
5. M. Kuribayashi, T. Tanaka, S. Suzuki, T. Yasui, Nobuo Funabiki, "White-box watermarking scheme for fully-connected layers in fine-tuning model," 9th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'21), accepted, June 2021.
6. April Pyone MAUNG MAUNG, Hitoshi KIYA, "Piracy-Resistant DNN Watermarking by Block-Wise Image Transformation with Secret Key," ACM Workshop on Information Hiding and Multimedia Security 22th, accepted, June 2021.
7. Marc Treu, Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, "Fashion-Guided Adversarial Attack on Person Segmentation", Computer Vision and Pattern Recognition WORKSHOP ON MEDIA FORENSICS 2021, accepted, June 2021, [Preprint](#), [Presentation Video](#)

フェイク顔映像検出AIaaSの開発

-SYNTHETIQ VISION: Synthetic video detectorの概要-

- 判定対象となる映像のアップロードから、判定結果を示した映像をダウンロードするまでの全てのプロセスをWeb APIとして利用可能
- ウェブAPIの活用により、AIを活用したウェブサービス「AI as a service」を容易に実現



NEWS RELEASE

NII 大学共同利用機関法人 情報・システム研究機構
 国立情報学研究所
 National Institute of Informatics

2021年(令和3年)9月22日

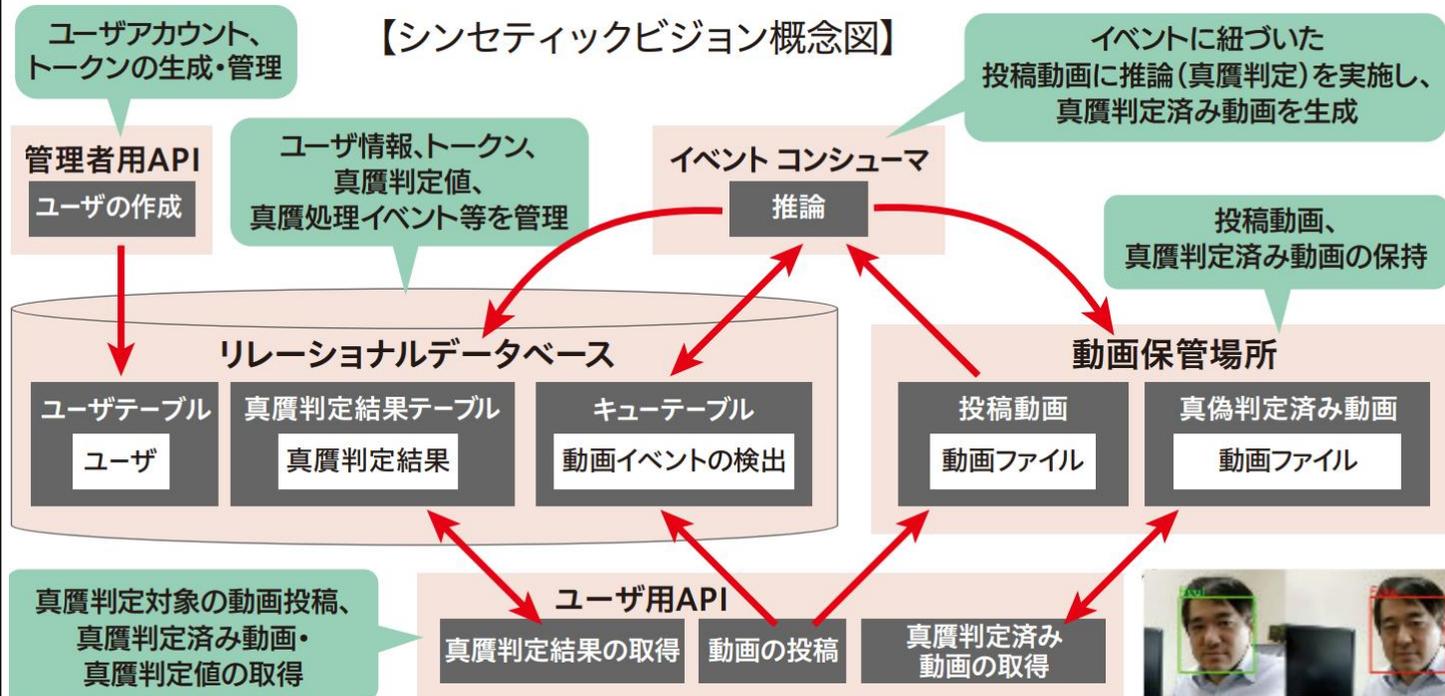
AIにより生成されたフェイク顔映像を自動判定するプログラム

SYNTHETIQ: Synthetic video detectorを開発

～AI動画の生成、フェイクメディアの検知、メディアの信頼性確保の研究を推進～

大学共同利用機関法人情報・システム研究機構 国立情報学研究所 (NII, 所長: 喜連川 優、東京都千代田区) のシンセティックメディア国際研究センター長の越前 功と副センター長の山岸 順一の研究チームはDeepfakeに代表される AIにより生成されたフェイク顔映像を自動判定するプログラム「SYNTHETIQ: Synthetic video detector」を開発しました。本プログラムは、判定対象となる映像のアップロードから、判定結果を示した映像をダウンロードするまでの全てのプロセスをウェブAPIとして利用可能なものです。このウェブAPIの活用により、AIを活用したウェブサービス「AI as a service, AIaaS」を容易に実現できると期待されます。

本研究成果は、科学技術振興機構 (JST, 理事長: 濱口 道成、東京都千代田区) の戦略的創造研究推進事業の「CREST VoicePersonae: 声のアイデンティティクローニングと保護 (研究代表者 山岸 順一)」、 「CREST インフォデミックを克服するソーシャル情報基盤技術 (研究代表者 越前 功)」、およびJST 研究成果最適展開支援プログラム A-STEP (トライアウト) の「AIにより生成された顔映像フェイクメディアを検出する技術の確立 (研究代表者 越前 功)」のもとで開発されました。



SYNTHETIQ VISIONライセンス状況

- サイバーエージェント (有償ソフトウェアライセンス 2022年)
 - デジタルツインレーベルに登録した著名人のdeepfake検知
 - **【研究成果の実用化】事業利用開始のためライセンス開始(2023.1.13 NII NR)**
- 国内企業 (有償ソフトウェアライセンス 2022年)
 - eKYC向けdeepfake検知
- その他, 複数の企業に有償ソフトウェアライセンス実施中

プレスリリース

▶ プレスリリース一覧

印刷用画面

著名人のデジタルツインをキャストするサービス「デジタルツインレーベル」を、芸能事務所および著名人向けに開始
「分身」となる公式3DCGモデルを制作・管理し、デジタル空間でのタレント活動を促進へ



<https://www.cyberagent.co.jp/news/detail/id=26503>

AI作成の偽動画で顔認証突破 成り済まし口座開設可能

2021/7/3(土) 10:00 配信 1



偽動画で顔認証を突破するイメージ

人工知能 (AI) を使って偽動画を作る「ディープフェイク」の技術を使うと、銀行口座を開設する際のオンライン認証で、本人に成り済まして顔認証を突破できる恐れがあることが3日分かった。日立製作所の研究グループが論文を人工知能学会で発表した。犯罪に悪用される恐れもあり、偽動画を見破るための対策が必要と提案している。

オンラインの本人確認システムは「eKYC」と呼ばれ、利用者がスマートフォンで撮影した本人の免許証などと一
緒に、カメラの前で本人の動画や顔写真を提示する仕組み。両方が同じなら同一人物と認識する。非対面で本人確認が可能で、金融機関などで導入が進んでいる。

<https://jp.reuters.com/article/idJP2021070301000235>

2023年（令和5年）1月13日

AIが生成したフェイク顔映像を自動判定するプログラム 「SYNTHETIQ VISION」をタレントのDeepfake映像検知に採用 ～フェイク顔映像の真偽自動判定では国内最初の実用例～

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（エヌアイアイN I I、所長：喜連川 優、東京都千代田区）のシンセティックメディア国際研究センター長えちげん いさおの越前 功と副センター長やまざしの山岸 順一じゅんいちの研究チームが開発した、AIが生成したフェイク顔映像の真偽を自動判定するプログラム「シンセティック ビジョンSYNTHETIQ VISION: Synthetic video detector」を株式会社サイバーエージェント（サイバーエージェント、代表取締役：藤田 晋、東京都渋谷区）が採用し、タレント等の著名人のディープフェイクDeepfake映像検知で実用化することになりました。NIIは情報学分野における研究成果を社会問題解決のために応用、展開する社会実装に取り組んでおり、今回の「SYNTHETIQ VISION」の実用化は、流通する多様なメディアの信頼性確保に寄与するものです。

本研究成果は、国立研究開発法人 科学技術振興機構（ジェイエスティJ S T、理事長：橋本 和仁、東京都千代田区）の戦略的創造研究推進事業クレストCREST「VoicePersonae: 声のアイデンティティクローニングと保護（研究代表者：山岸順一）」、「インフォデミックを克服するソーシャル情報基盤技術（研究代表者：越前 功）」、およびJST 研究成果最適展開支援プログラムエースステップA-STEP（トライアウト）の「AIにより生成された顔映像フェイクメディアを検出する技術の確立（研究代表者：越前 功）」により開発されました。



[サイバーエージェント様](#)
 デジタルツインレーベルに登録した著名人のdeepfake検知
【研究成果の実用化】事業利用開始のためライセンス開始
（2023.1.13 NIIニュースリリース）

2023年（令和5年）5月24日

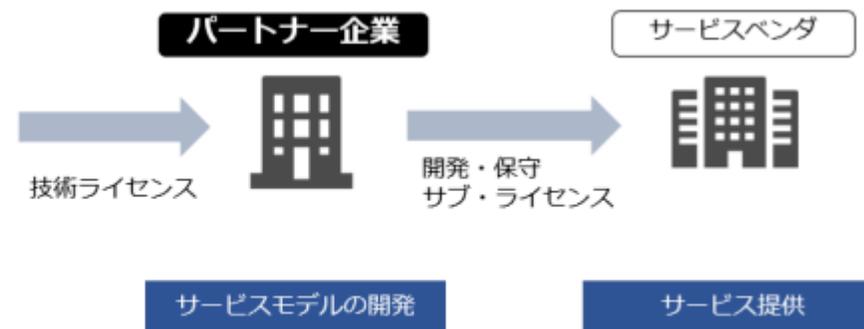
国立情報学研究所によるフェイク顔映像の真贋自動判定プログラム「SYNTHETIQ VISION」のライセンス事業者を募集 ～NIIの最新AI研究成果を社会に広めるパートナー企業を求む～



大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（^{エヌアイアイ}NII、所長：黒橋 禎夫、東京都千代田区）は、NIIが開発したフェイク顔映像の自動判定プログラム「^{シンセティックビジョン}SYNTHETIQ VISION」のライセンス事業を行う事業者の提案」を5月24日から募集します。

e-KYC（electronic Know Your Customer：オンライン本人確認）の普及やフェイク映像の増加などで、シンセティックメディア（AIで作られた映像）の真贋判定に対する社会的な期待やニーズが急速に増大してきました。これらに対応して企業がビジネスを持続的に実施できる環境が求められてきており、最新の研究成果を希望する企業が効率的かつ速やかに活用しています。この課題を解決するため、NIIの研究成果であるフェイク顔映像の真贋判定プログラム「SYNTHETIQ VISION」の技術移転を円滑に進めることを目的とします。NIIと当該企業が連携し研究成果を幅広く社会に広めることを目指し

SYNTHETIQ VISION
のライセンス事業者募集
(2023.5.24 NIIニュースリリース)



<図 2> SYNTHETIQ VISION の技術移転を円滑に進めるパートナー企業を募集

SYNTHETIQ VISION: 広がる潜在的なユースケース

これまでのユーザ企業とのやり取りから読み取った
B2B事業における想定ユーザ企業とユースケース

分野・業界	想定ユーザ企業	リスク	ユースケース
エンターテインメント業界	映画制作会社, 芸能事務所, ITサービス事業者	著作権侵害, 名誉棄損	真贋判定によるフェイクメディア削除
報道機関	テレビ局, 新聞社, オンラインメディア, ITサービス事業者	誤報道, 報道機関の信用低下	報道前の真贋判定による誤報道の防止
ソーシャルメディア	プラットフォーム, ITサービス事業者	偽・誤情報の拡散, プラットフォームの信頼性低下	真贋判定結果(確信度)のユーザへの提示による偽・誤情報の拡散防止
金融機関	銀行, 証券会社, クレジットカード会社, 金融ITサービス事業者	リモート本人確認(eKYCなど)に対するなりすまし	真贋判定によるなりすまし検知
法執行機関	警察庁, 弁護士事務所, フォレンジクス事業者	誤った証拠の使用, 特定の遅延	真贋判定による証拠等の真正性確認
ビデオコミュニケーション	ビデオコミュニケーションサービス事業者	身元詐称, オンライン面接・試験の不正	真贋判定によるなりすまし検知・不正監視

アウトライン

- イン트로ダクション, 人間由来の情報を用いたフェイクメディアの生成
- 顔を対象としたフェイクメディアの生成手法
- 顔を対象としたフェイクメディアの検出手法
- インフォデミックの克服に向けて (CREST FakeMedia, SYNTHETIQ VISION)
- 今後の課題

今後の課題

AIを活用したコンテンツモデレーション(AIを用いた自動検知や自動ファクトチェック)の必要性

- ✓効果と効率性の観点から必須技術となりうる
- ✓透明性やアカウントビリティの確保が重要だが、課題もある
 - ◆AIによる推論は原則ブラックボックス
 - ◆AIの学習データやベンチマークを公開すると、それを逆手にとって、自動検知を迂回する偽・誤情報の生成手法が出現する可能性

多種多様な偽情報の生成手法が出現

- ✓定期的なデータセット更新や自動検知モデルの追加学習が必要
 - ◆既知の手法で生成された偽情報の検知精度を確保しながらの追加学習はコスト大
 - ◆極めて多種多様の生成手法を安定的に自動検知できるか？
- ✓AI製を示す情報をコンテンツに不可分に埋め込む電子透かしの活用に期待
 - ◆ただし、電子透かしにも様々な課題がある

今後の課題

自動検知モデル, データセット, ベンチマークにおける課題

- ✓ 研究レベルでは様々なものが提案されているが, 現実の環境を反映していないもの多数
 - ◆ 課題解決のために産学連携による開発・実証が極めて重要
 - ◆ プラットフォーム事業者やAI関連事業者から研究者へのデータ提供も重要

自動ファクトチェックの課題

- ✓ 自動検出と相補的な活用が期待される
- ✓ 「信頼できる情報源」を誰がどのように収集し, メンテナンスしていくのか

ご清聴ありがとうございました