国立情報学研究所 市民講座
2024年3月4日

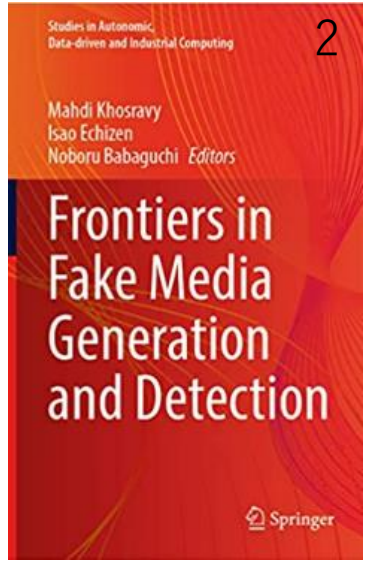
フェイクから身を守るには？
－ 創るAI vs 守るAI －

越前 功
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

略歴

学位
1995年3月　東京工業大学 理学部応用物理学科卒業
1997年3月　東京工業大学 大学院理工学研究科修士課程修了（応用物理学専攻）
博士（工学）（東京工業大学）2003年3月

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-現在　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



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

SynMedia Center

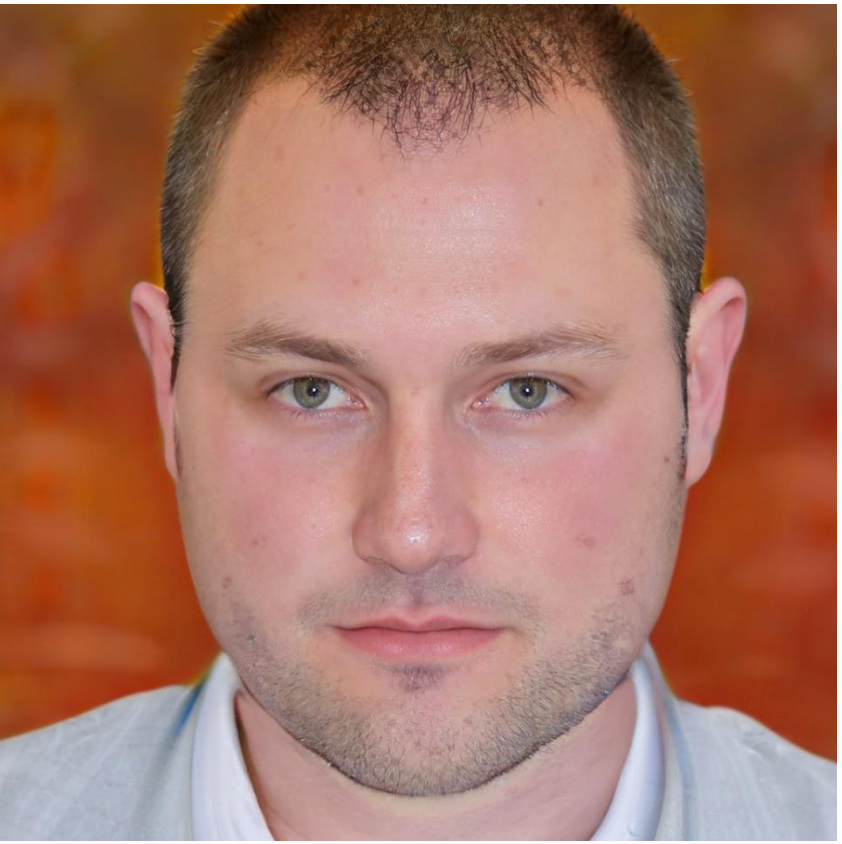
主な受賞歴
電子情報通信学会 論文賞(2023)，電子情報通信学会 ISS論文賞(2023)
情報セキュリティ文化賞(2016)，ドコモ・モバイル・サイエンス賞 先端技術部門優秀賞(2014)
情報処理学会 論文賞(2014, 2005)，情報処理学会 長尾真記念特別賞(2011)
関東地方発明表彰 発明奨励賞(2019, 2012)，Best Paper Award (WIFS17, I3E17など)

CREST FakeMedia

アウトライン

- イントロダクション，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）

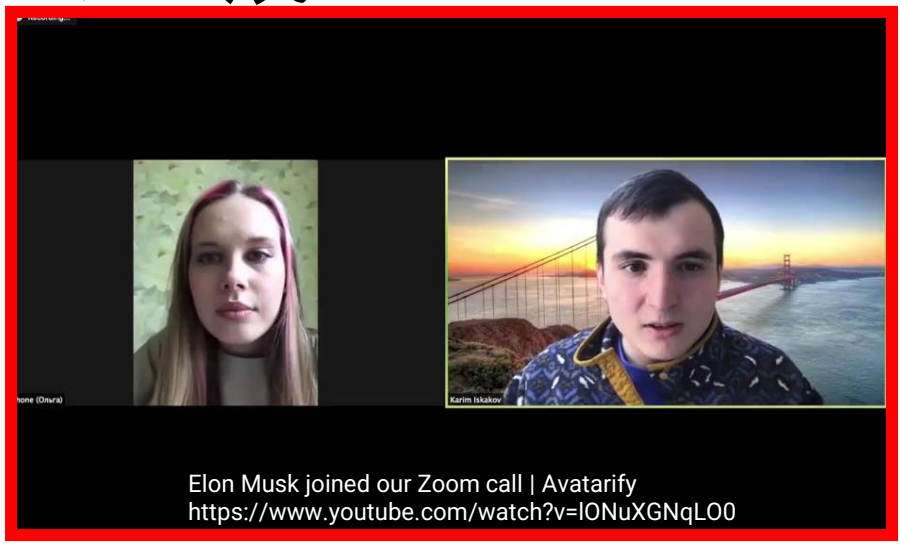
Fake or Real?

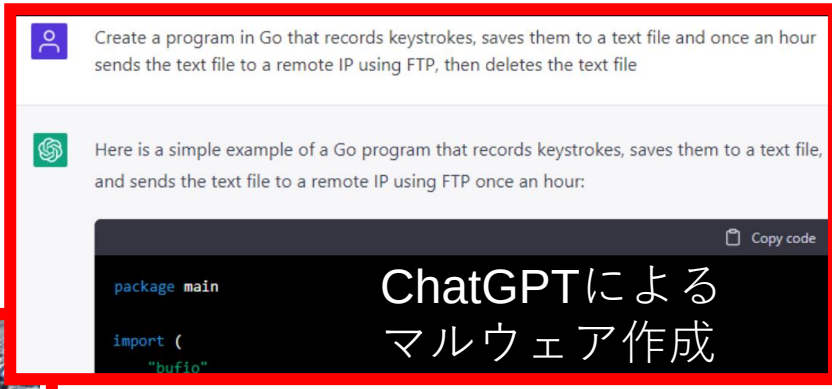


A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, ¨ and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In International Conference on Computer Vision, pages 1–11, Oct 2019.

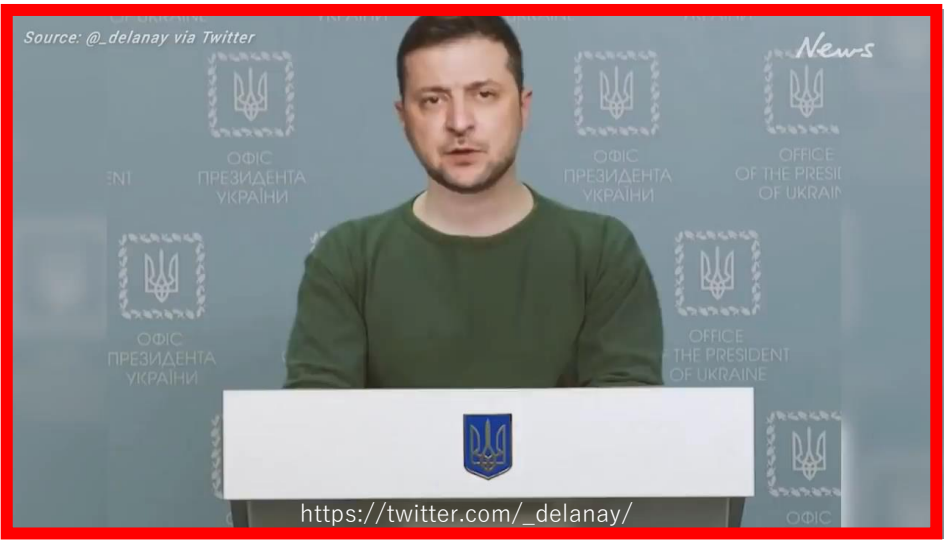
人間由来の情報を用いたフェイクメディアの生成

- 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

- イントロダクション，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）

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

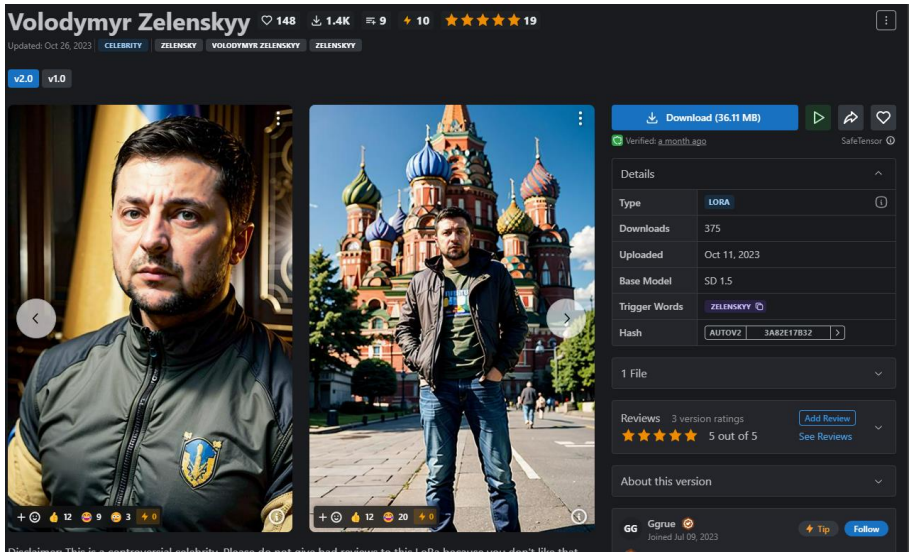
1. 顔全体の合成（Entire face synthesis）

- 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）

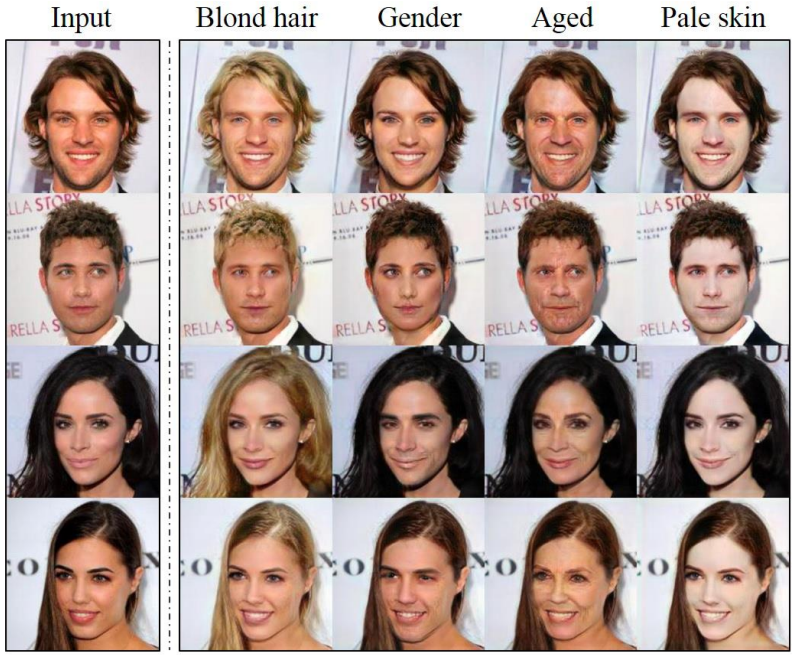
2. 顔の属性操作（Attribute manipulation: hair, skin color, expression）

- 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 2[1] (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.

Sora（Open AI）

暖かく光るネオンとアニメーションの街の看板で埋め尽くされた東京の通りを歩くスタイリッシュな女性。黒いレザージャケットに赤いロングドレス、黒いブーツを履き、黒い財布を持っている。サングラスに赤い口紅。彼女は自信に満ち、さりげなく歩いている。通りは湿っていて反射し、色とりどりのライトの鏡のような効果を生み出している。多くの歩行者が歩いている。

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

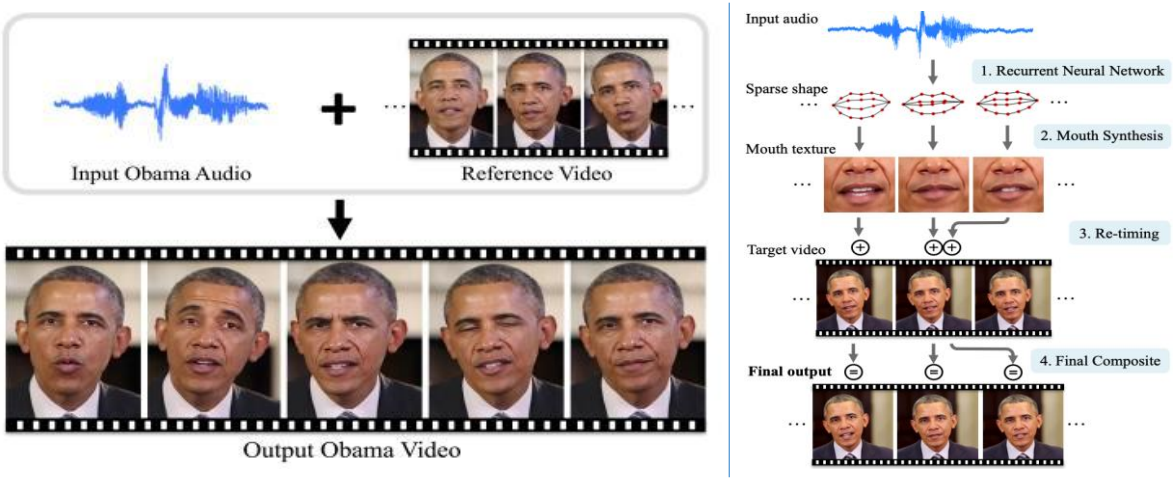
Neural Talking Head Models
(Zakharov et al. 2019)

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



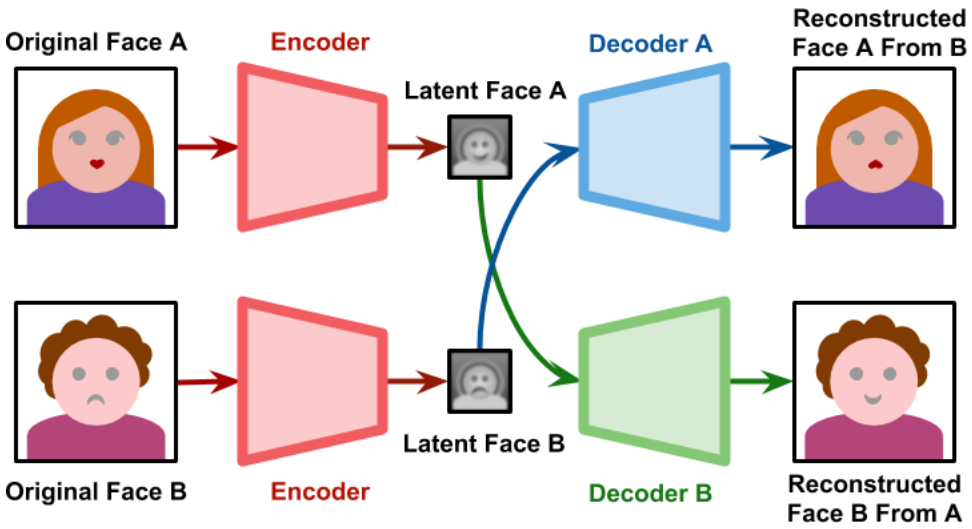
Synthesizing Obama
(Suwajanakorn et al. 2017)

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）

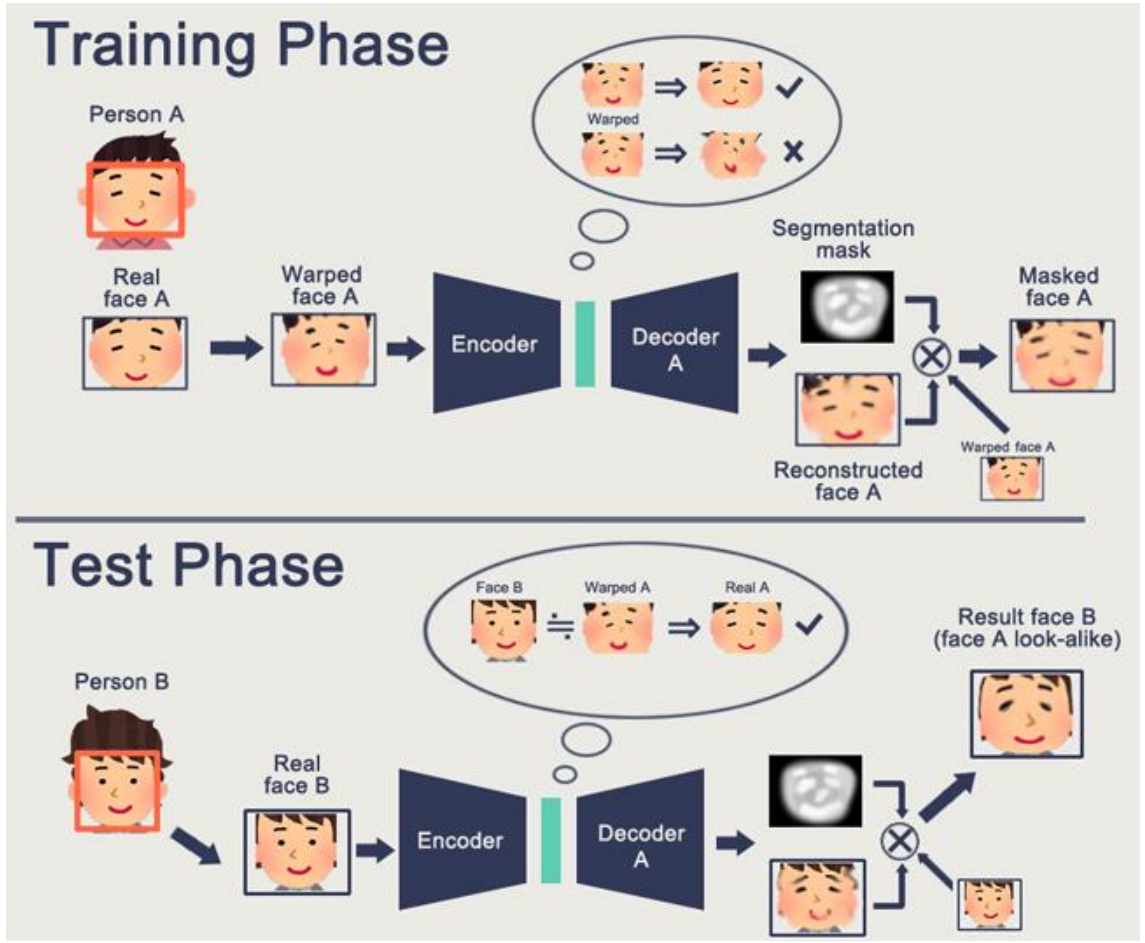
- ソースとなる映像の顔部分をターゲットの顔と入れ替える（Faceswapなど）

Deep learning based face swap



Original Deepfake (Faceswap)[1]

Image: Alan Zucconi



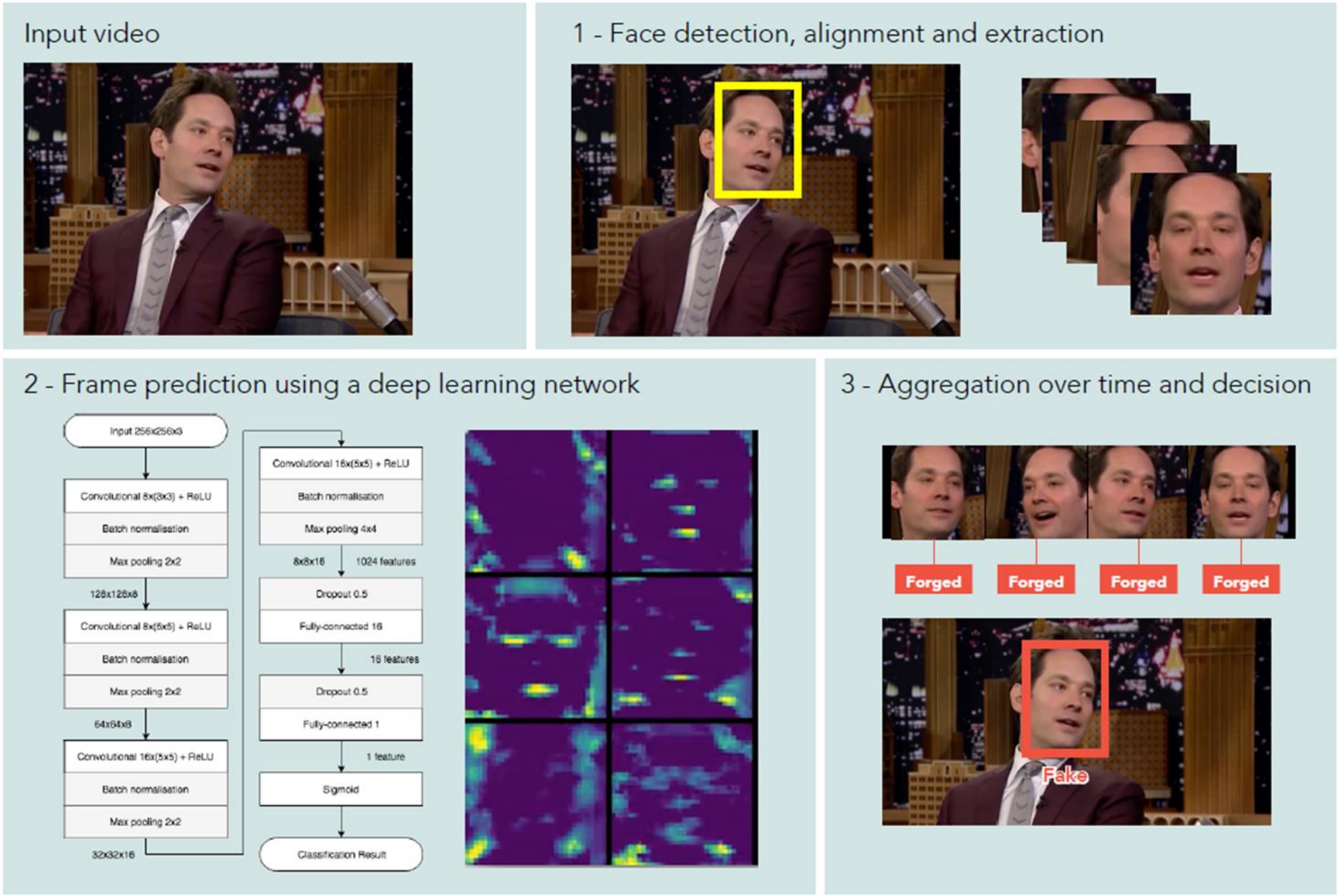
Faceswap – GAN[2]

Image: shaoanlu

アウトライン

- イントロダクション，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）

Mesonet: フェイク顔映像を検出するコンパクトな深層学習モデル

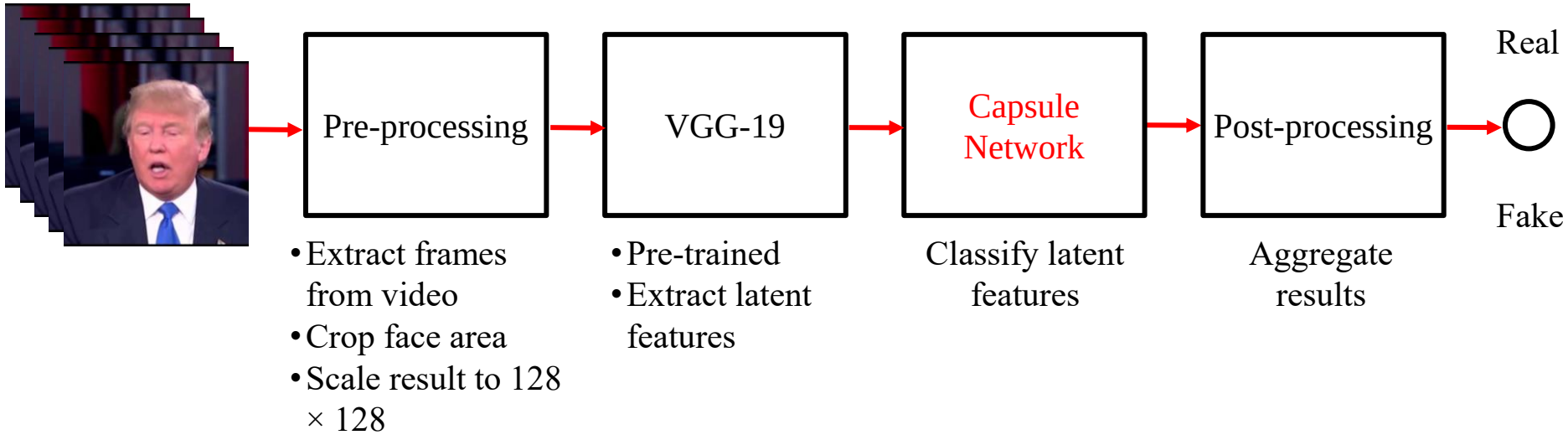


D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, " MesoNet: a Compact Facial Video Forgery Detection Network, " Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS 2018), pp.1-7, December 2018 (number of citations: 1,250)

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).

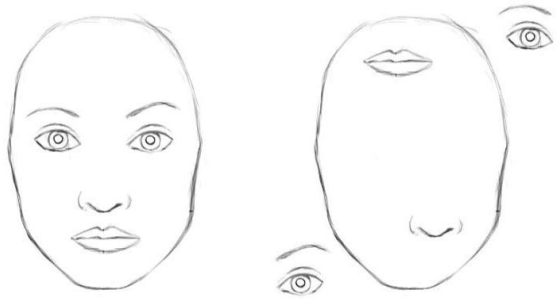
- Combine VGG19 with Capsule Network as a countermeasure



Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos” ICASSP 2019 (number of citations: 561)

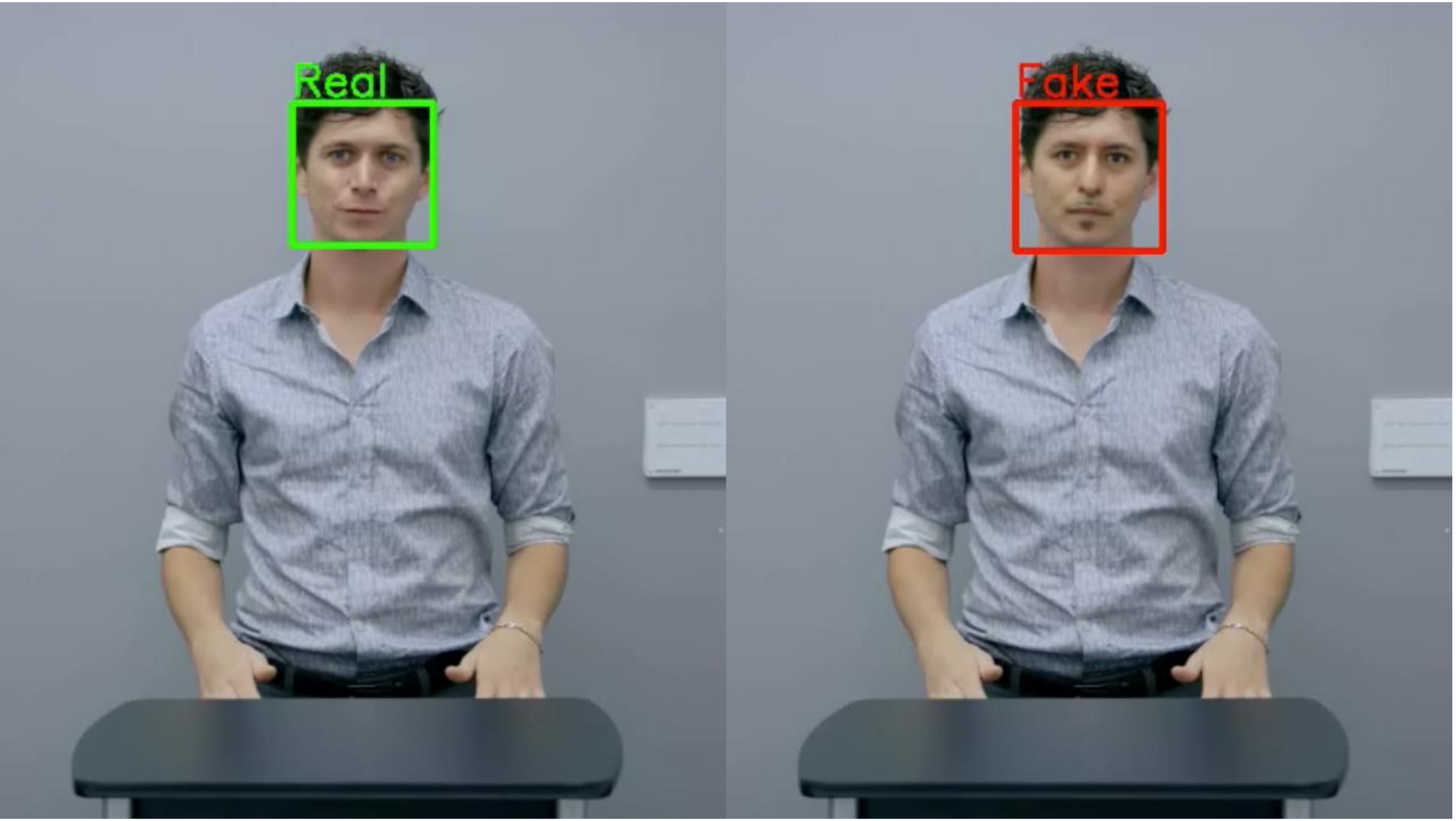
Why capsule networks?

- **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.

Detection Results (Faceswap)

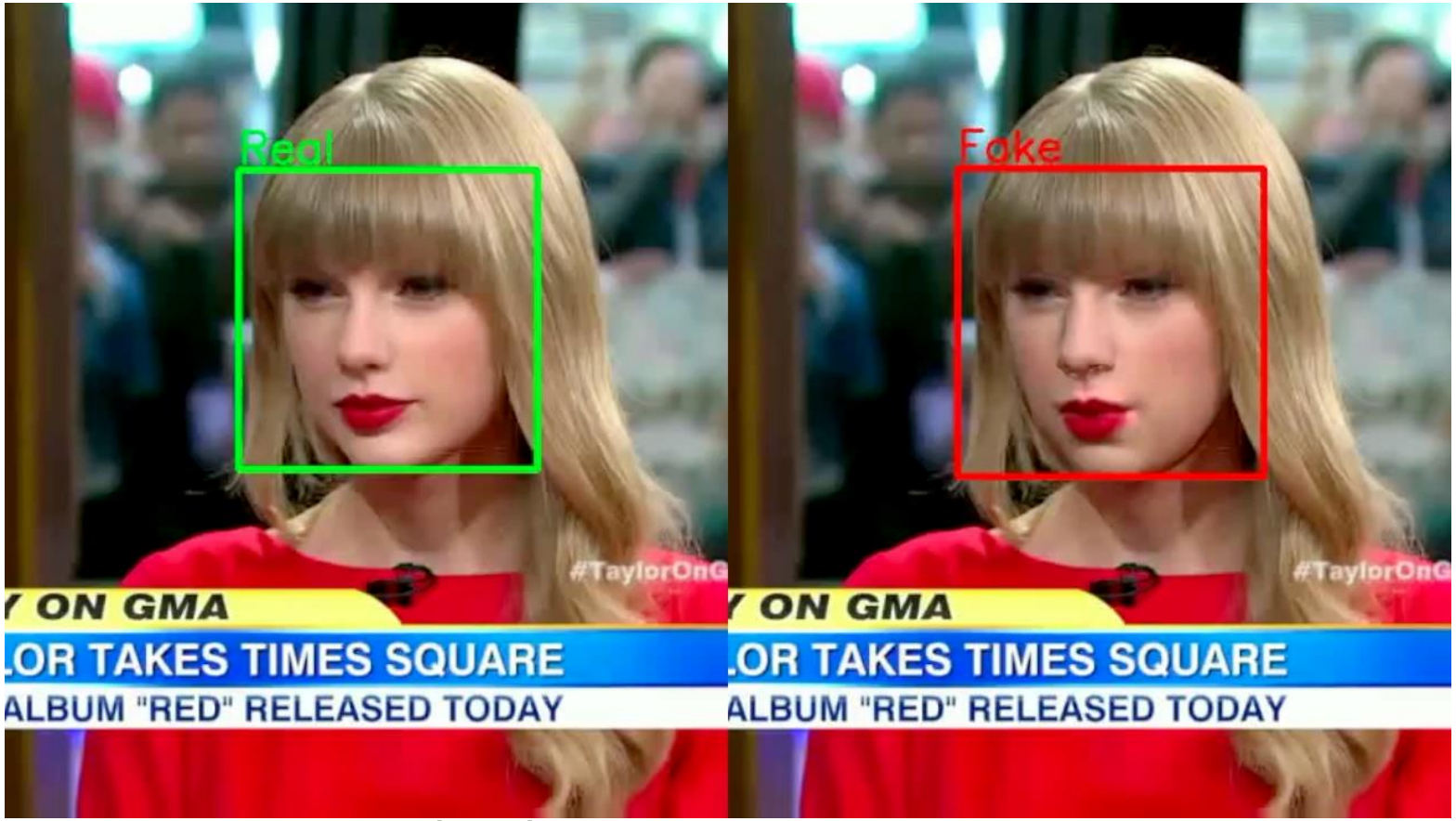


Our Deepfake dataset

	Real (frame)	Forged (frames)
Train	4,600	6,525
Dev	511	725
Eval	2,889	4,259

EER: 1.42%

Detection Results (Face2Face)



FaceForensics dataset

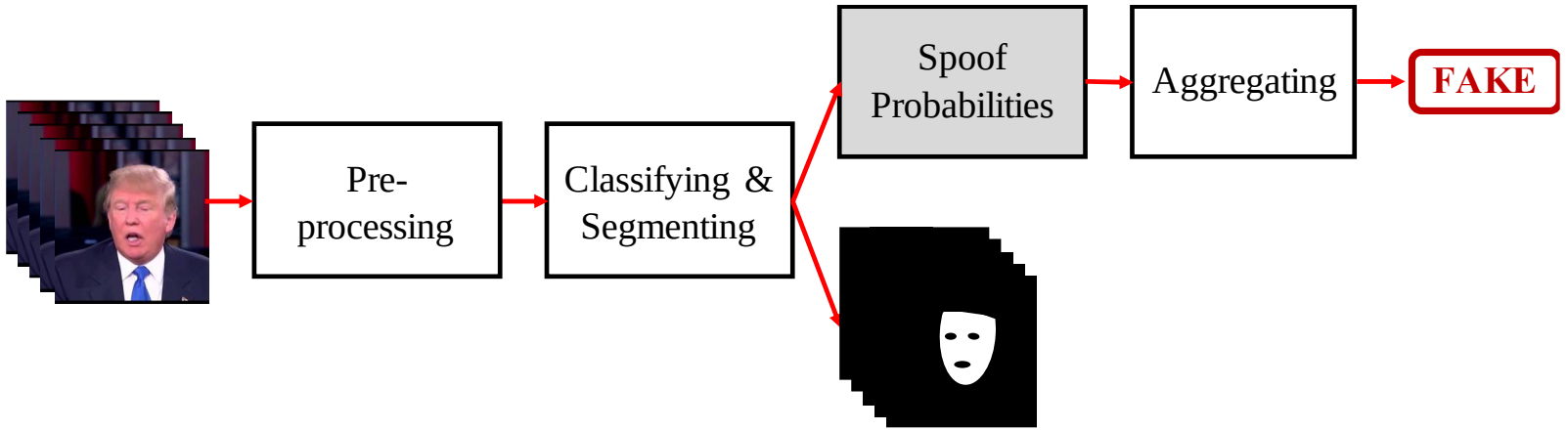
	Real (frame)	Forged (frames)
Train	7,040	7,040
Dev	1,500	1,500
Eval	1,500	1,500

EER

No compression: 0.67%
Light compression: 2.67%
Strong compression: 17.0%

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.**

Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, Isao Echizen, “Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos"Proc. of the BTAS 2019,8 pages, September 2019 (number of citations: 437)


Fake
COUNTDOWN TO NOVEMBER 4
#5 OBAMA UNPLUGGED
MSNBC
COUNTDOWN with KEITH OLBERMANN

アウトライン

- イントロダクション，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）

CREST FakeMedia

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 准教授）

研究背景：フェイクメディア(FM)とインフォデミック

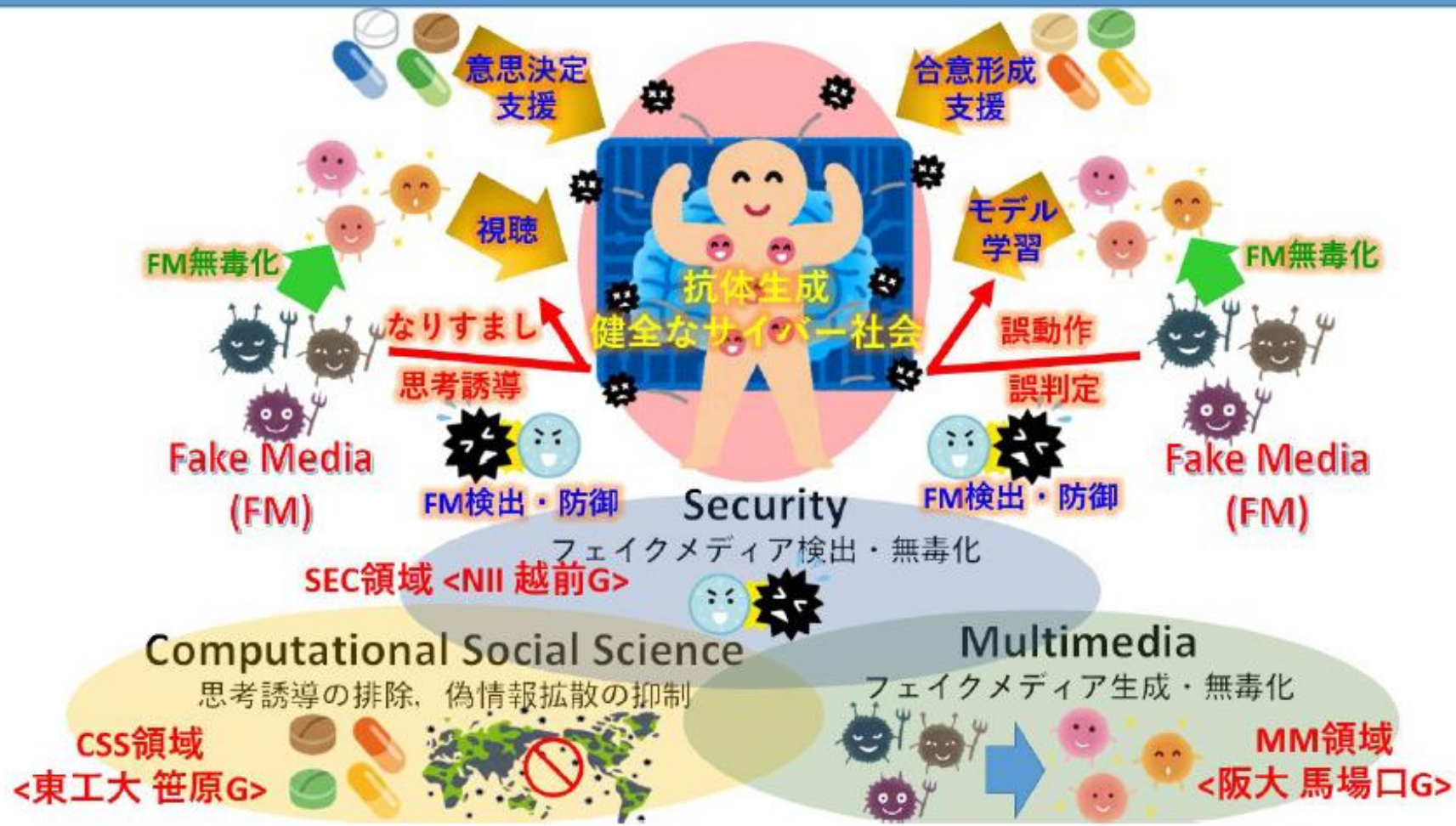
- COVID-19とインフォデミック
  - 社会に恐怖や混乱を引き起こす不確かな情報の氾濫
    - 科学的根拠のない予防法や治療法に関わるフェイクニュース
    - 望遠カメラ撮影により意図的に密集状態を演出
  - 愉快犯や攻撃者：多様なFMを駆使して，インフォデミックを意図的に発生させる可能性
    - メディアクローン（MC）型FM：本物に限りなく近いが本物ではない
    - プロパガンダ（PG）型FM：世論操作のためにメディアを意図的に加工
    - 敵対的サンプル（AE）型FM：AIを誤動作・誤判定させる

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だけではなく，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
    通常の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！**



研究実施項目

- 3種類のFMを想定して，4つの研究実施項目に取り組む
  - 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FM（例：DeepfakeやBERTにより生成されたメディア）
  - プロパガンダ（PG）型FM
    世論操作などを目的として素材となるメディアを意図的に編集して生成したFM
  - 敵対的サンプル（AE）型FM
    人間には識別困難だが，AIを誤動作・誤判定させることを目的に生成したFM
- 研究実施項目
  1. 多様なモダリティによる高度なFM生成技術（主担当：MM領域 馬場口）
     映像（顔，身体など），音声，文書などの多様なモダリティに対するFM生成技術の確立
  2. FM検出・防御技術（主担当：SEC領域 越前）
     1で生成した多様なFMを対象とした高度な検出・防御技術の確立
  3. FM無毒化技術（主担当：MM領域，SEC領域）
     思考誘導，誤動作・誤判定が生じないようにFMを無毒化し，通常のメディアとして活用する技術の確立
  4. インフォデミックを緩和し多様な意思決定を支援する情報技術（主担当：CSS領域 笹原）
     1～3の要素技術を最大限に活かし，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する

SEC：Security
MM：MultiMedia
CSS：Computational Social Science

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

42編

うちQ1ジャーナル 21編

Top 10% 論文 18編（Scopus）

Top 3% 論文 10編（Scopus）

査読付き国際会議論文

100編

うちCORE A*/A論文 29編

（CORE A*: 15, CORE A: 14）

ニュースリリース

4件

NII SynMedia Center設置

SYNTHETIQ VISION開発・実用化等

メディア掲載

238件

招待講演・依頼講演

62件

うち国際講演 26件

書籍・著書・解説記事

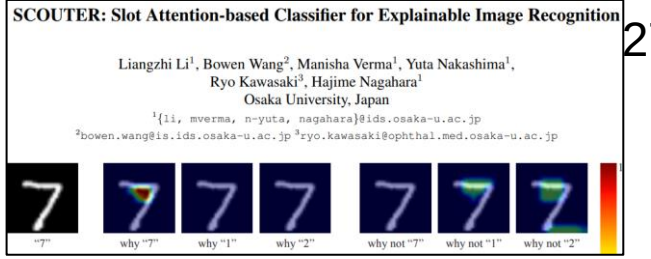
28編

口頭発表・ポスター・デモ

62件

受賞

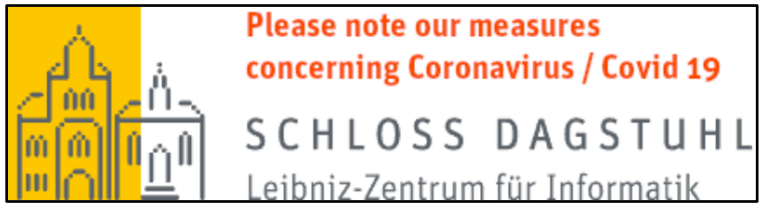
11件



CVPR, ICCV, CHI (CORE A*)に採択

COMMUNICATIONS OF THE ACM

CACM Regional Special Issueに掲載 (June 2023)

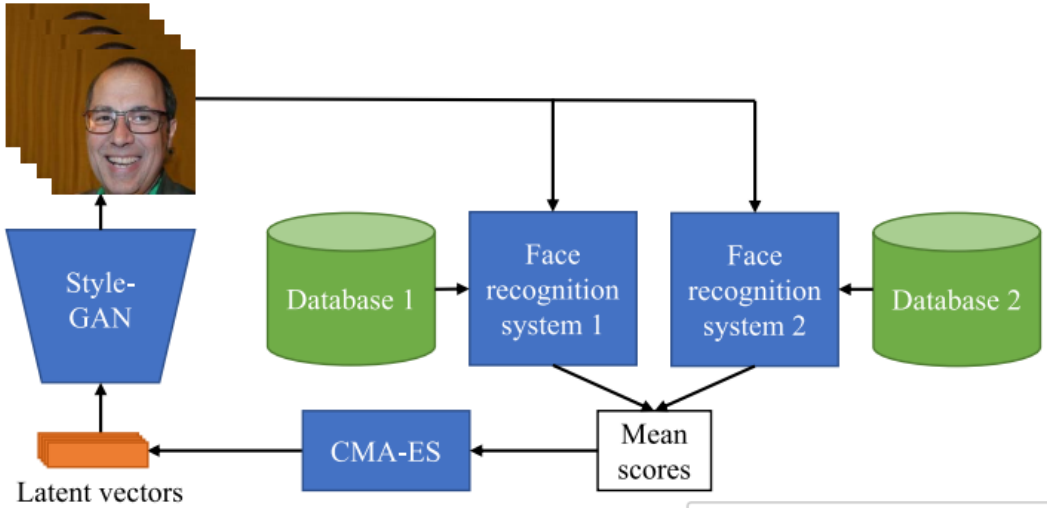


Media Forensics and the Challenge of Big Dataに参加 (Jan. 2023)

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

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と一致度が高いと判定された登録顔

Master faceの例

逐次更新により登録顔が密集しているエリアに収束する

H. H. Nguyen, S. Marcel, J. Yamagishi and I. Echizen, "**Master Face Attacks on Face Recognition Systems**," in **IEEE Transactions on Biometrics, Behavior, and Identity Science**, vol. 4, no. 3, pp. 398-411, July 2022

マスター静脈(Master Vein)による生体認証システムへのウルフ攻撃

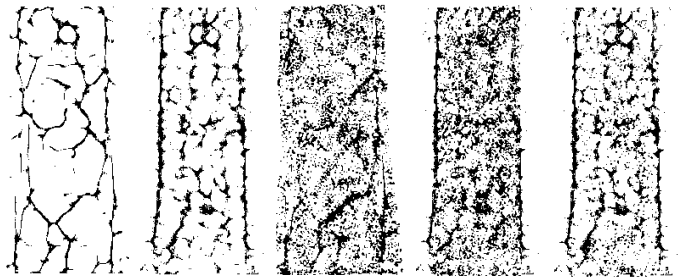
- 静脈識別システムに登録された複数の静脈特徴と類似するマスター静脈（Master Vein）を機械学習モデルで生成
- Handcrafted featureベースの指静脈識別（Miura's system）においてMaster Vein Attackが有効であることを示した
- WACV 2023 (CORE A)に採択

Background

- Finger vein recognition systems have been deployed in ATMs.
- Some systems still use traditional hand-crafted features and do not have proper countermeasure methods deployed.

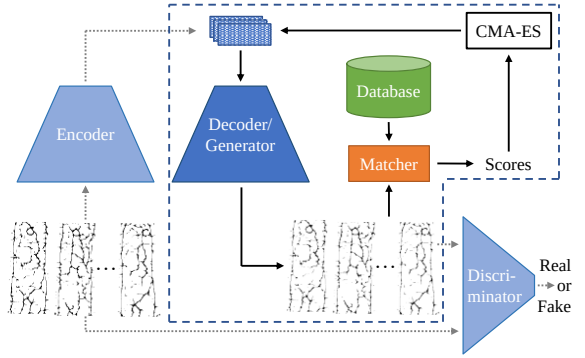
→ They may be vulnerable to master vein attacks.

Master Vein Examples

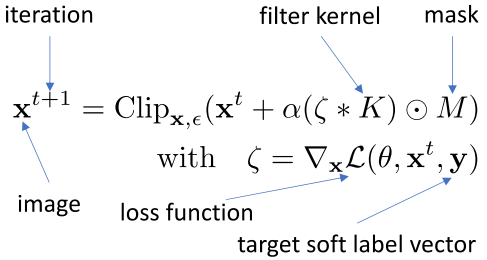


a. Original image　b. LVE[3]　c. AdvML　d. Combination　e. Combination with top labels

Proposed Methods



Latent variable evolution (LVE)-based attack



Adversarial machine learning (AdvML)-based attack

Results (FARs) on **cross-database** and **cross-system** attacks

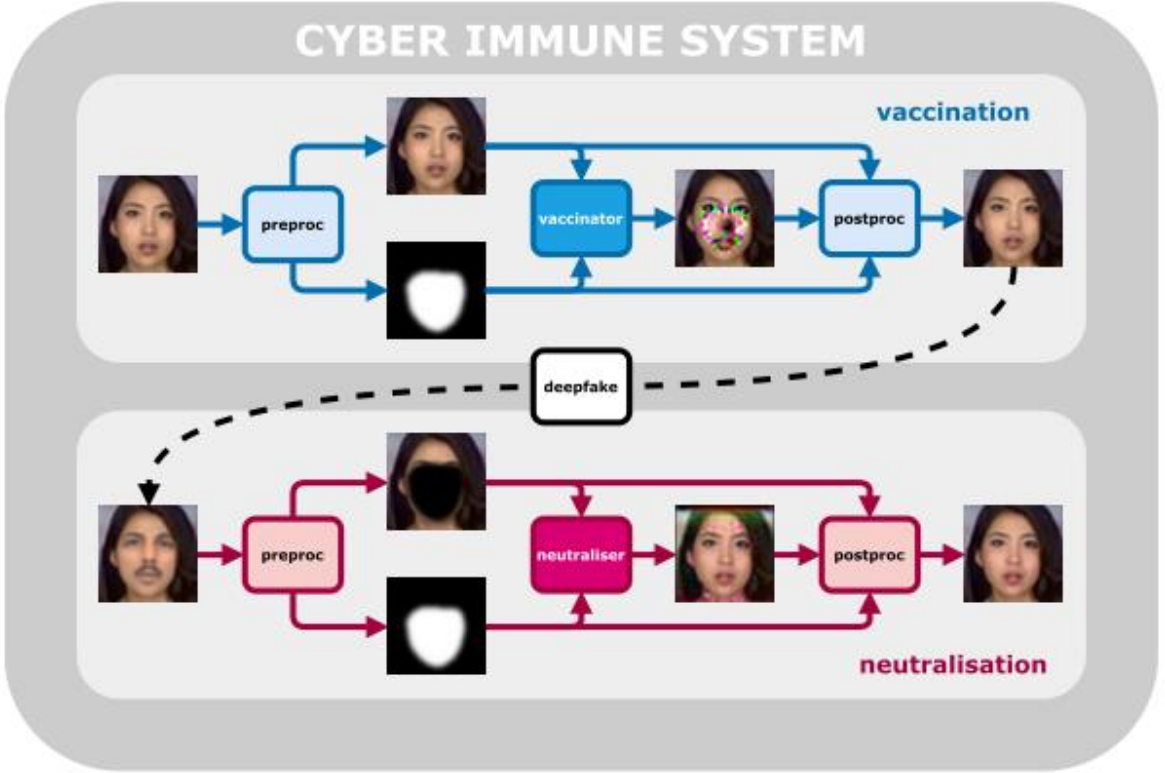
Attack \ Matcher	Miura's system (Partial matching)	Miura's system (Full matching)	ResNeXt 50	ResNet 18	Mobile NetV3-L
Bona fide	04.07	03.13	8.22	7.28	8.10
LVE[1] (WGAN)	**38.84**	**43.86**	0.18	0.10	0.18
LVE[2] (β-VAE)	**15.08**	02.92	0.00	0.00	0.00
LVE[3] (Comb.)	**20.84**	**19.54**	0.54	0.00	0.01
AdvML (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

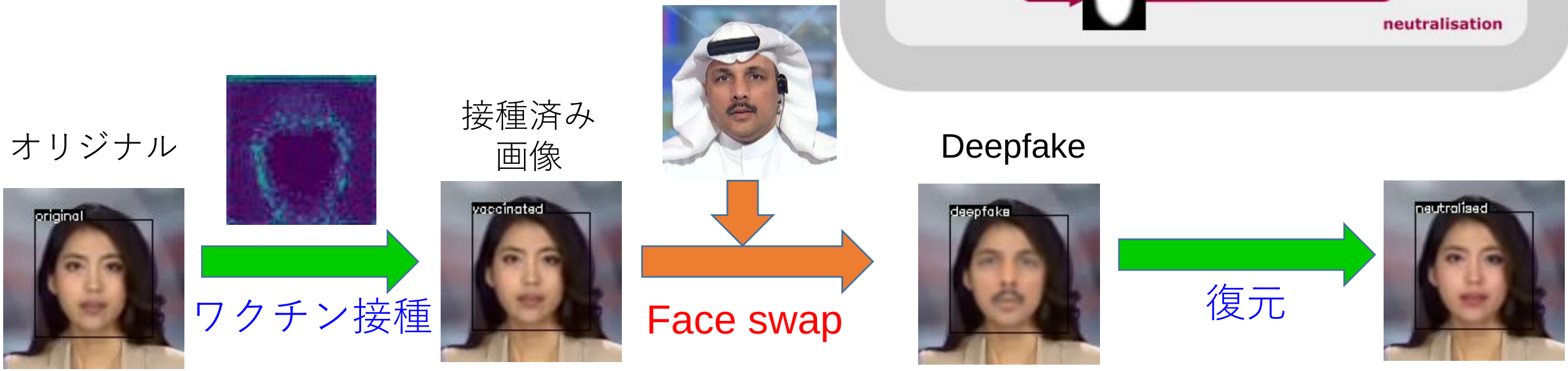
- **Miura's system** is **vulnerable** to master vein attacks.
- The combination of the LVE-based and AdvML-based methods achieved the best results.
- CNN-based FVRSs are more robust against master vein attacks.

H. H. Nguyen, T. N. Le, J. Yamagishi, and I. Echizen, "**Analysis of Master Vein Attacks on Finger Vein Recognition Systems**," **Winter Conference on Applications of Computer Vision (WACV) 2023**.

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





C.-C. Chang, H. H. Nguyen, J. Yamagishi, I. Echizen, “Cyber Vaccine for Deepfake Immunity” IEEE Access, September 2023

ワクチン接種と非接種における復元画像の比較

オリジナル
顔画像

unvaccinated samples

ワクチン接種
→Faceswap
→復元

neutralised samples with vaccination

ワクチン非接種
→Faceswap
→Inpainting-
based method

neutralised samples without vaccination

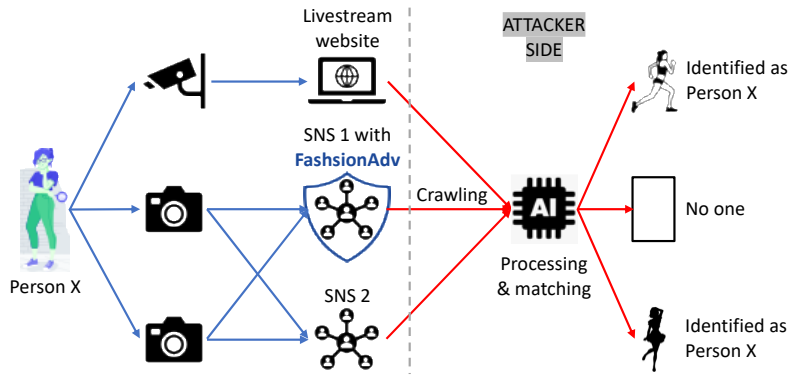
オリジナル顔との
類似度
(FaceNet)

200枚の顔画像（FaceForensics++）における類似度の平均値：0.99（ワクチン接種），0.57（ワクチン非接種）

Cyber Vaccine: 敵対的ノイズの付加によるプライバシー保護

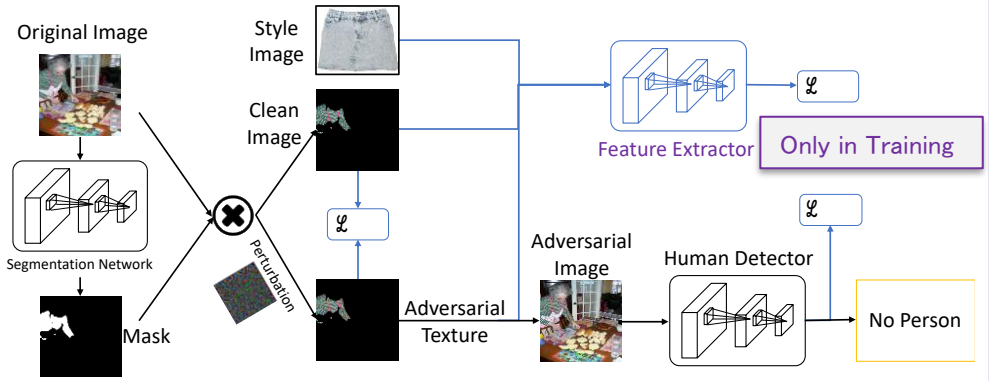
- 自身が共有したコンテンツを不用意に第三者に解析されないようにメディア処理にロバストな敵対的サンプルを重畳する
- Person segmentationを解析例として，衣服領域にノイズ重畳することで解析不能にする手法を提案
- CVPR Workshop on Media Forensics 2021に採択
- 本研究を応用して，写真内のランドマークの解析から写真の位置情報を推測する分析を不能にする手法を提案：WIFS2022に採択

Background



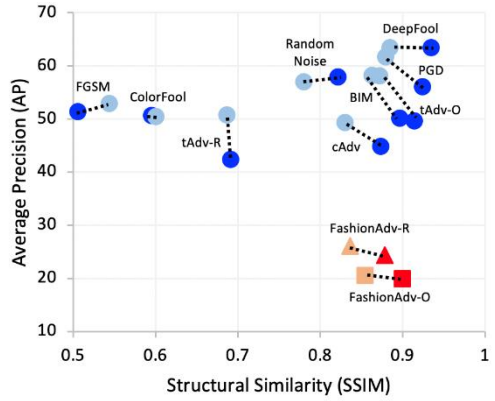
_ Shared images can be accessed and crawled illegally to train recognition systems
_ FashionAdv prevents it by making people disappear from detectors

Proposed Approach

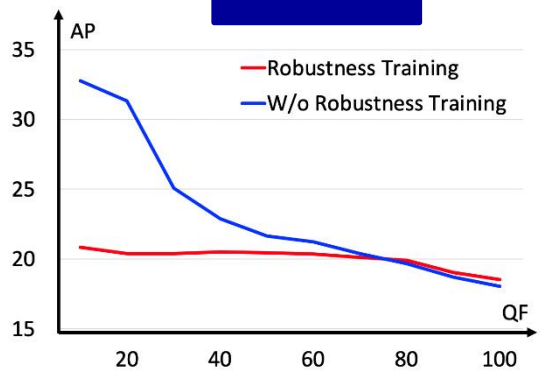


_ FashionAdv synthesizes adversarial textures from fashion images and blend them into original cloth-regions
_ FashionAdv aims to balance robustness and naturalness

Results



FashionAdv achieves competitive naturalness and significantly outperforms SOTA on robustness on JPEG compression (QF 80 & 40).



Robustness on JPEG compression at different Quality Factors (QF)

Marc T., Le T.-N., Nguyen, H. H., Yamagishi, J., & Echizen, I., “Fashion-Guided Adversarial Attack on Person-Instance Segmentation”, CVPR Workshop 2021

フェイクメディア（FM）検出の性能向上を目指した高品質な大規模データセットの構築

- 従来のデータセットは実世界のFMから乖離（被写体1名，室内で撮影）
- 画像内の複数objectを同時検出するモデルが近年提案されており，FM検出も同時検出に対応したモデルが提案される可能性
- 複数の被写体による自然な画像を対象としたFMデータセットを提案
- FM検出に加えて，改ざん領域を推定するsegmentationタスクにも対応
- データセット公開，7000回以上のダウンロード
- ICCV2021(CORE A*)に採択

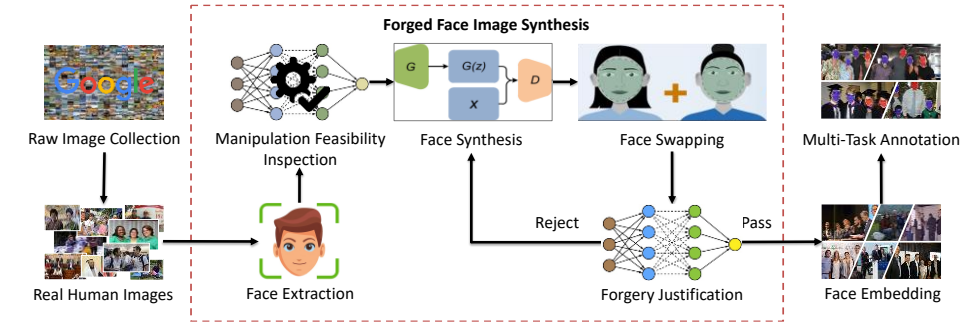
Background

- It is extremely difficult to point out forged faces among many faces in natural scenes.

Contributions

- Address new tasks of multi-face forgery detection and segmentation in-the-wild
- Present new dataset: 115k images with 334k faces
- Provide benchmark suite

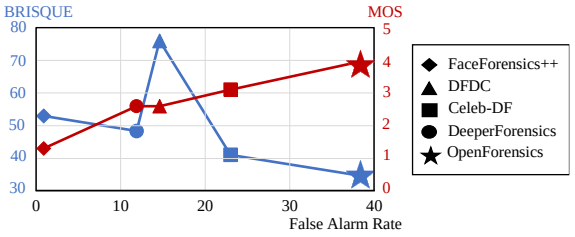
Dataset Generation

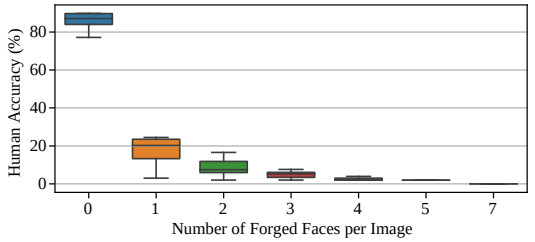


❖ **Test-Challenge set with data augmentation:**

User Study

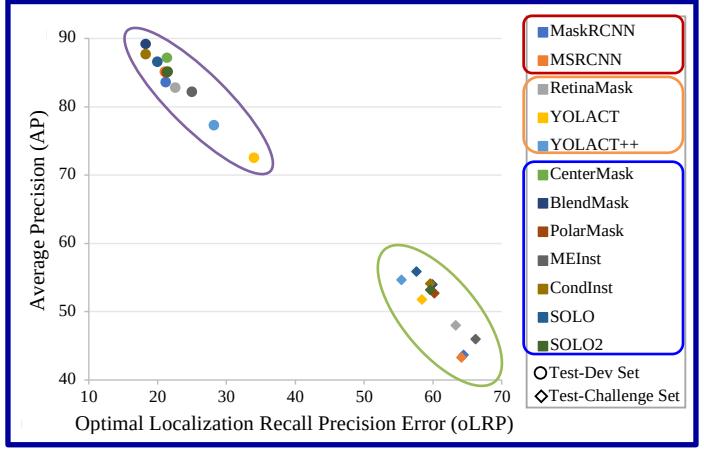
- 3,000 images (5 datasets) was used in experiments
- 200 participants (80 experts and 120 non-experts)





- OpenForensics can trick human (highest justification error) with highest realism
- More fake faces cause more missed detection

Benchmark



Le T.-N., Nguyen, H. H., Yamagishi, J., & Echizen, I., “OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild”, International Conference on Computer Vision (ICCV), 2021 **(CORE A*, # of citations: 46)**

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

国内でも生成AIの脅威が深刻化（2021年～）

- 詐欺，詐称
  - フェイク音声で企業の幹部になりすまし，現金を搾取（2019年）
  - フェイク顔でイーロン・マスクになりすまし，Zoom参加（2020年）
  - **国内 機械学習モデルで生成・配布したサンプル顔画像を，利用企業が自社の宣伝に不正利用（2021年）**
- 思考誘導，世論操作
  - 架空の人物になりすまして，株価操作を目論む（2019年）
  - **国内 加藤官房長官の地震直後の記者会見の表情を改ざん（2021年）**
- 特定個人に対する名誉毀損，いじめ
  - 国内 Deepfakeによるアダルトビデオ公開・逮捕（2020年）
  - **娘のライバルを蹴落とすため，母親がライバルのDeepfake生成（2021年）**

読売オンライン，2021/6/13, 架空の顔で「お客様の声」「大満足」ＡＩで生成、９０サイトで宣伝に悪用
https://www.yomiuri.co.jp/national/20210613-OYT1T50073/

読売オンライン，2021/4/28,加藤官房長官がフェイクの笑み、ＡＩで悪意の改変
https://www.yomiuri.co.jp/national/20210411-OYT1T50038/

BBC NEWS | JAPAN

娘のライバル蹴落とすため……「ディープフェイク」でわいせつ動画作成の母親逮捕

https://www.bbc.com/japanese/56411511

2021年3月16日

BBC 2021年3月
娘のライバルを蹴落とすため，母親がライバルのフェイク映像を作成し，コーチに送信・逮捕
誰でもFMの生成が可能になりつつある

NIIシンセティックメディア国際研究センター

SynMedia Center

人間中心のAI社会を実現するために，多様なメディアの生成，メディアの信頼性確保，意思決定支援のための研究開発を，実世界の課題を取り上げながら，国際的な拠点として推進する

山岸CREST(日仏共同提案)

JST-ANR VoicePersonae Project

VoicePersonae: 声のアイデンティティクローニングと保護

音声合成，声質変換，音声強調の統合による話者アイデンティのモデル化，利活用と保護

越前CREST

CREST FakeMedia

インフォデミックを克服するソーシャル情報基盤技術

フェイクメディアの脅威に対する適切な対処，多様なコミュニケーションを支援

Global Research Center for Synthetic Media

シンセティックメディア生成

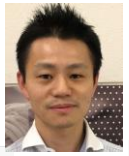
音声情報処理，画像・映像処理，自然言語処理，コンピュータビジョン

馬場口 阪大特任教授

フェイクメディア検知

デジタルフォレンジクス，情報セキュリティ，プライバシー

貴家 都立大教授



笹原 東工大准教授

メディアの信頼性確保，意思決定支援

計算社会科学，社会心理学，ELSI

水野 NII准教授

EURECOM

AVIGNON UNIVERSITÉ

idiap RESEARCH INSTITUTE

universität wien

IIT JAMMU

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP. HỒ CHÍ MINH

INDIANA UNIVERSITY

多様なメディアの利活用と信頼性確保を学際的・国際的体制で追究

新たな科学技術分野と研究潮流の創生，国内外の学術機関との連携，産学官連携を通じた実社会適用を推進

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

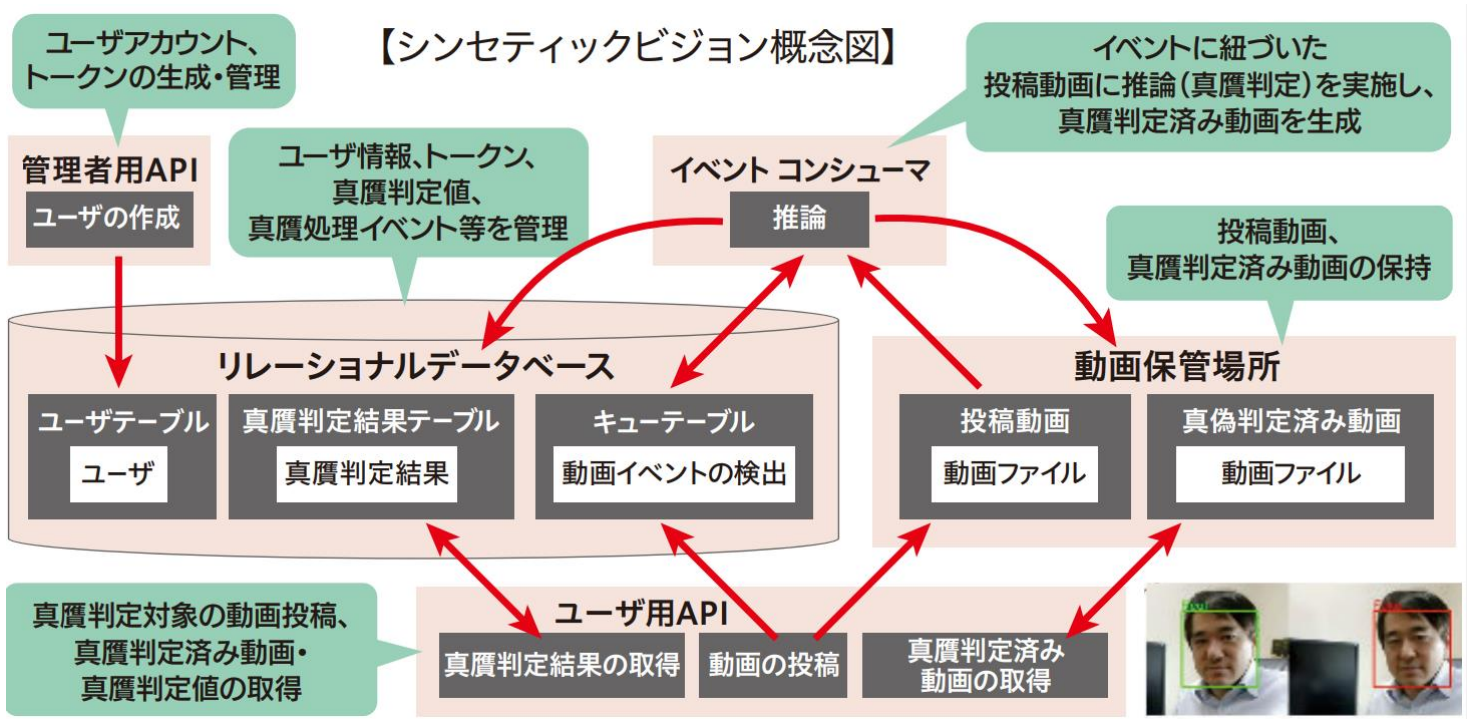
NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

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：濵口 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。



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

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（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「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（ＪＳＴ、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 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 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）

NEWS RELEASE

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

2023 年（令和 5 年）5 月 24 日

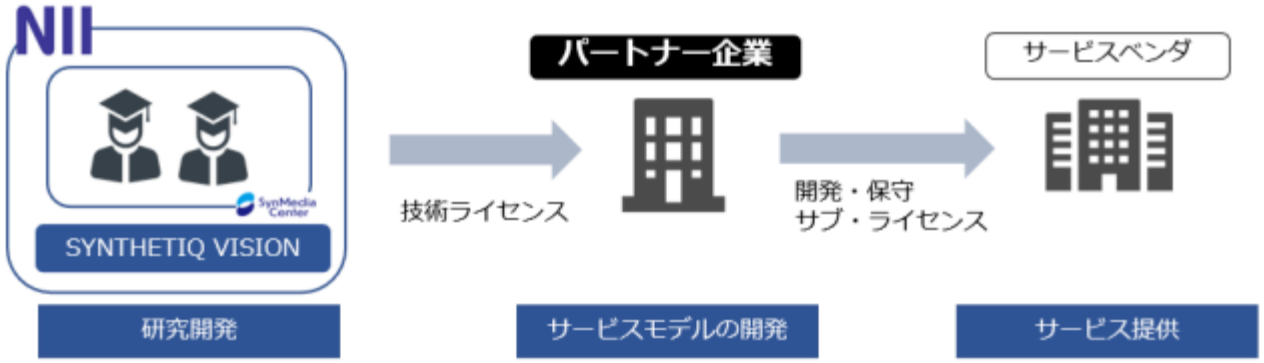
国立情報学研究所によるフェイク顔映像の真贋自動判定プログラム「SYNTHETIQ VISION」のライセンス事業者を募集

～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（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「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 で作り出された映像）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
グラム「SYNTHETIQ VISION」の技術移転を円滑に進めることを目的として
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

偽・誤情報の拡散に対する技術的対策について

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
- ✓透明性やアカウンタ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

多種多様な偽誤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・誤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

偽・誤情報の拡散に対する技術的対策について

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，メインテナンスしていくのか

Q & A