資料3

顔を対象としたフェイクメディアの生成と検出

越前 功
国立情報学研究所 情報社会相関研究系 研究主幹・教授
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

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，SynMedia Centerでの取り組み）

Fake or Real?

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年)，GROVER(フェイクニュース 2019年)
  - 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)
  - 架空の人物になりすまして，株価操作を目論む(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 Deepfakeによるアダルトビデオ公開・逮捕(2020年)

THE WALL STREET JOURNAL.

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

ウオールストリートジャーナル 2019/8/30

英国エネルギー企業のCEOが，フェイク音声で親会社のCEOになりすました電話を受け，2600万円を送金してしまった.

https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

FAST COMPANY

04-30-19

How to spot the realistic fake people creeping into your timelines

FastCompany 2019/4/30

AIで生成されたプロファイル画像を用いて，Maisy Kinsley（Bloombergのジャーナリスト）という偽のTwitterアカウントを作り，Teslaの株主に接触して個人情報を取得した上で，Teslaの株価操作を目論んだ.

https://www.fastcompany.com/90332538/how-to-spot-the-creepy-fake-faces-who-may-be-lurking-in-your-timelines-deepfaces

アウトライン

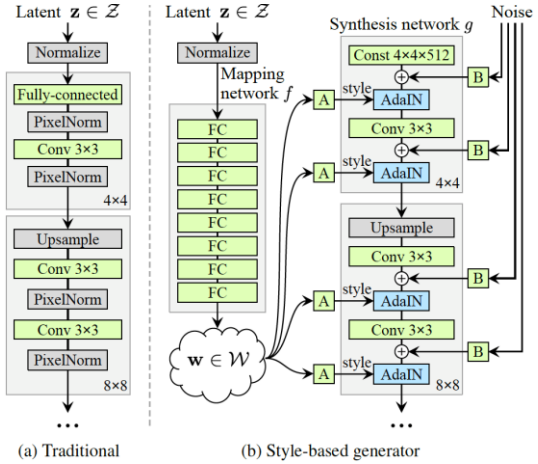
- イントロダクション，人間由来の情報を用いたフェイクメディアの生成
- 顔を対象としたフェイクメディアの生成手法
- 顔を対象としたフェイクメディアの検出手法
- インフォデミックの克服に向けて（CREST FakeMedia，SynMedia Centerでの取り組み）

顔を対象としたフェイクメディアの生成：5つのタイプ

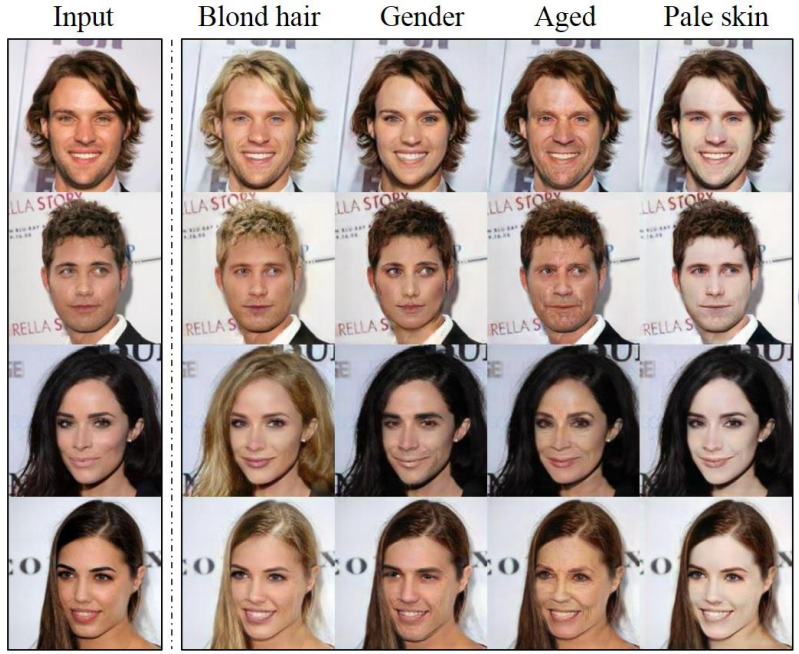
1. 顔全体の合成（Entire face synthesis）
   - ノイズ（潜在変数）から（実世界に存在しない）顔画像を生成する（StyleGAN，VQ-VAEなど）
2. 顔の属性操作（Attribute manipulation: hair, skin color, expression）
   - ターゲットの顔画像の髪の色，肌の色や表情などを変更した顔画像を生成する（StarGAN，ELEGANTなど）
3. 顔映像の表情操作（Facial reenactment）
   - 攻撃者の表情と，ターゲットの顔画像／映像を合成して，攻撃者の表情と同期したターゲットの顔映像を生成する（Face2Face，ICFaceなど）
4. 顔映像の話し方操作（Speaking manipulation）
   - 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）

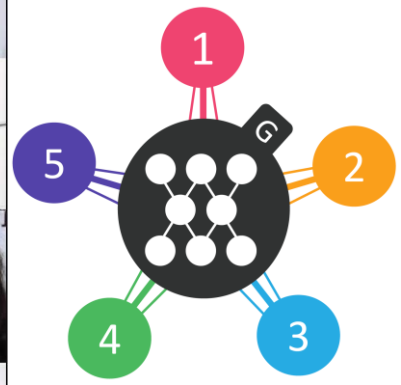
顔を対象としたフェイクデータの生成：5つのタイプ

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）
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.





StarGAN (Choi et al. 2018).
Image-to-image translation for multiple domains.

顔を対象としたフェイクデータの生成：5つのタイプ

3. 顔映像の表情操作（Facial reenactment）

- 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）

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

Face2Face (Thies et al. 2016).

Transferring facial movements of one person to the other one.

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

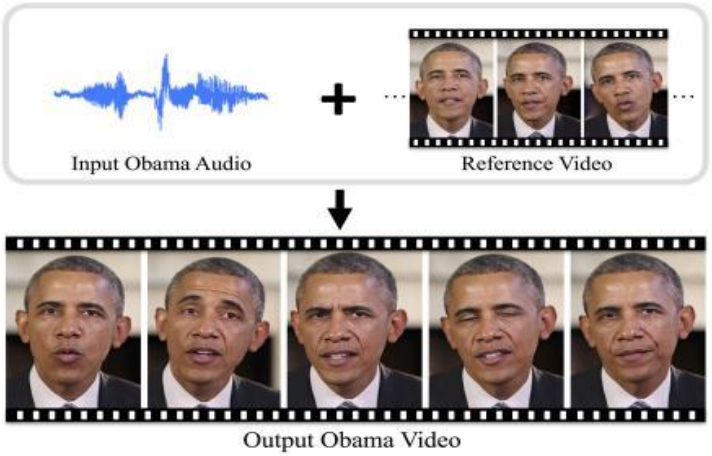
Neural Talking Head Models
(Zakharov et al. 2019)

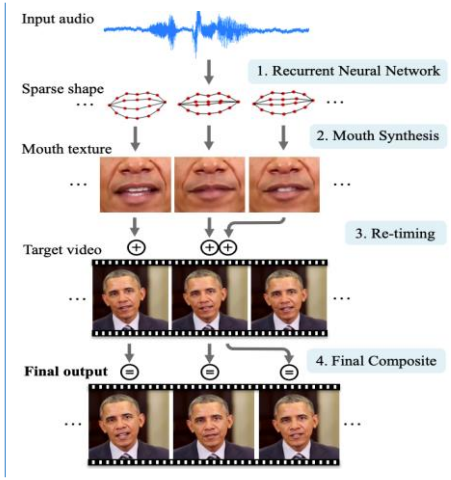
顔を対象としたフェイクデータの生成：5つのタイプ

4. 顔映像の話し方操作（Speaking manipulation）

- 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）

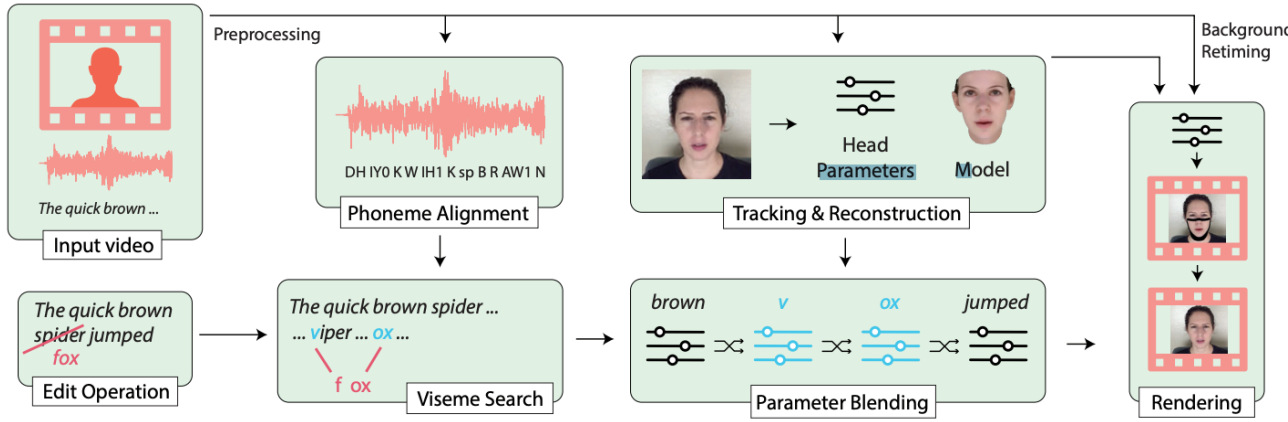
Synthesized speech (attacker) + image/video (victim)
→ forged video





Synthesizing Obama
(Suwajanakorn et al. 2017)

Modified text (attacker) + video (victim)
→ forged video



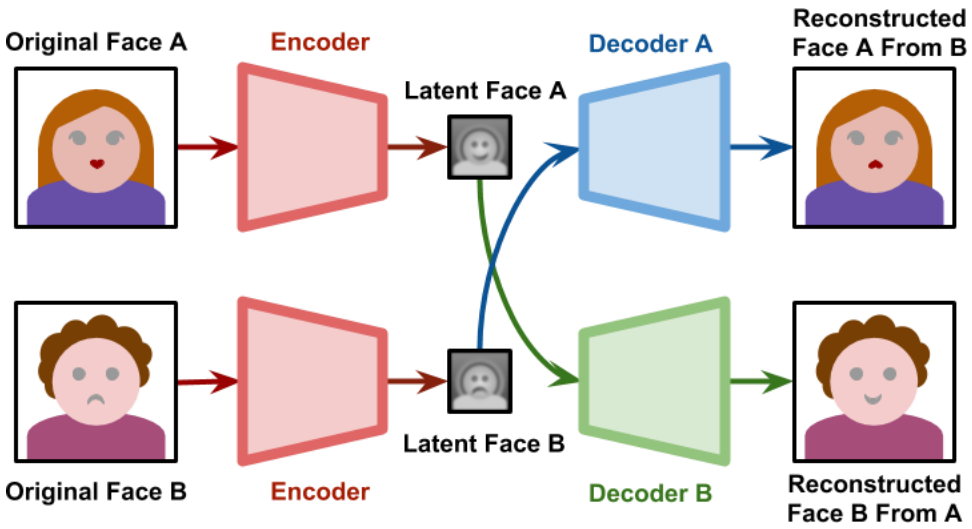
Text-based Editing of Talking-head Video
(Fried et al. 2019)

顔を対象としたフェイクデータの生成：5つのタイプ

5. 顔の入れ替え（Face swap）

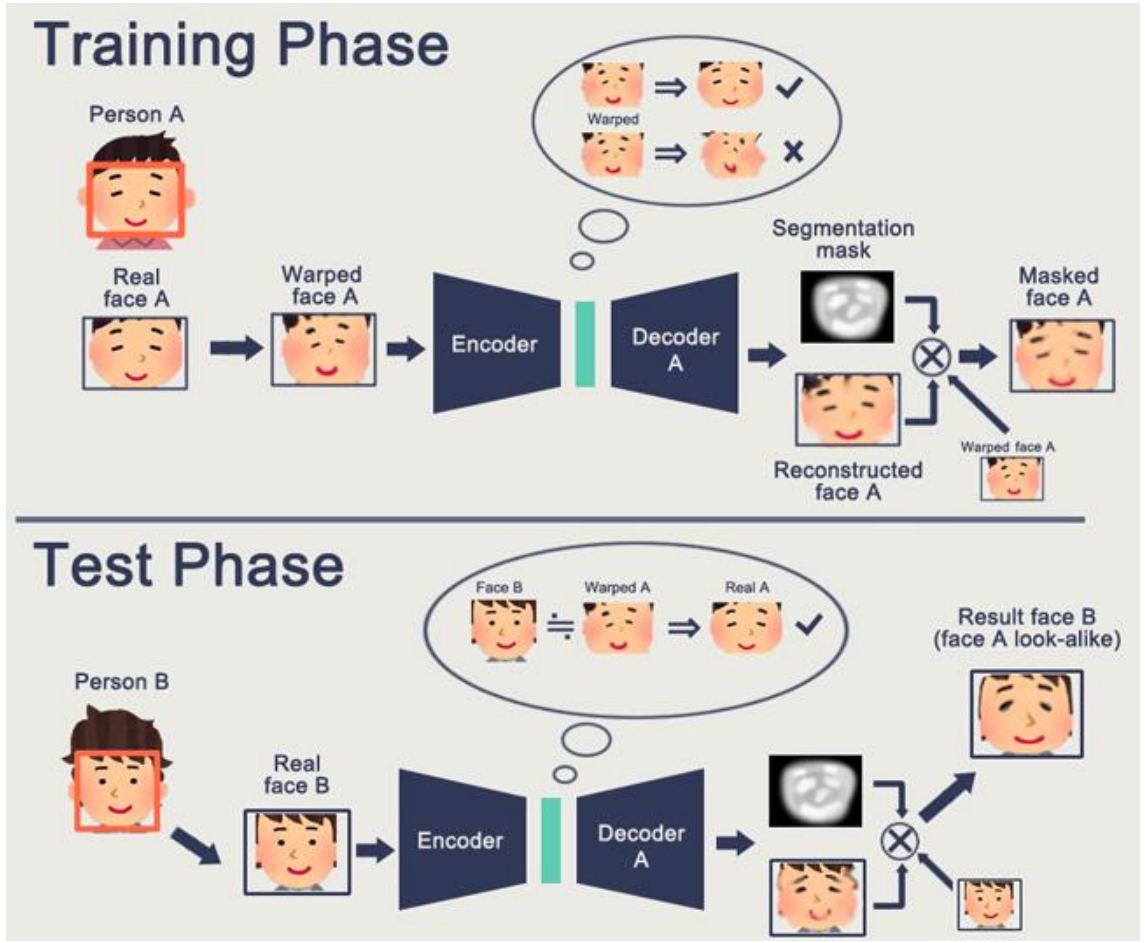
- ソースとなる映像の顔部分をターゲットの顔と入れ替える（Faceswapなど）

Deep learning based face swap



Original Deepfake (Faceswap)[1]

Image: Alan Zucconi



Faceswap – GAN[2]

Image: shaoanlu

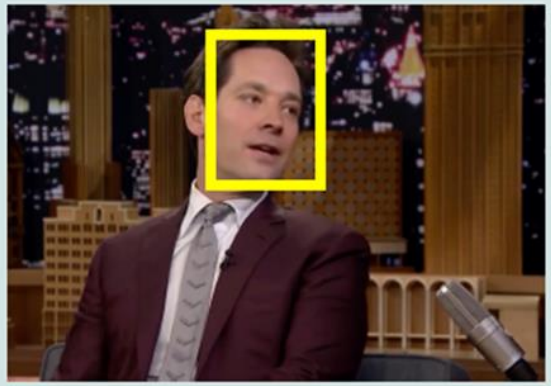
アウトライン

- 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，SynMedia Centerでの取り組み）

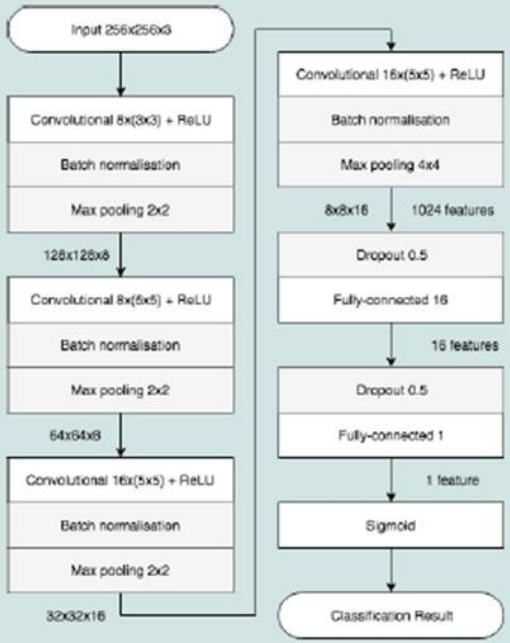
Mesonet: フェイク顔映像を検出するコンパクトな深層学習モデル

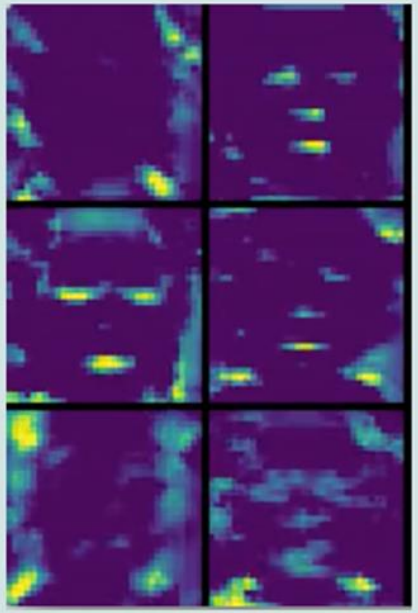
Input video

1 - Face detection, alignment and extraction



2 - Frame prediction using a deep learning network





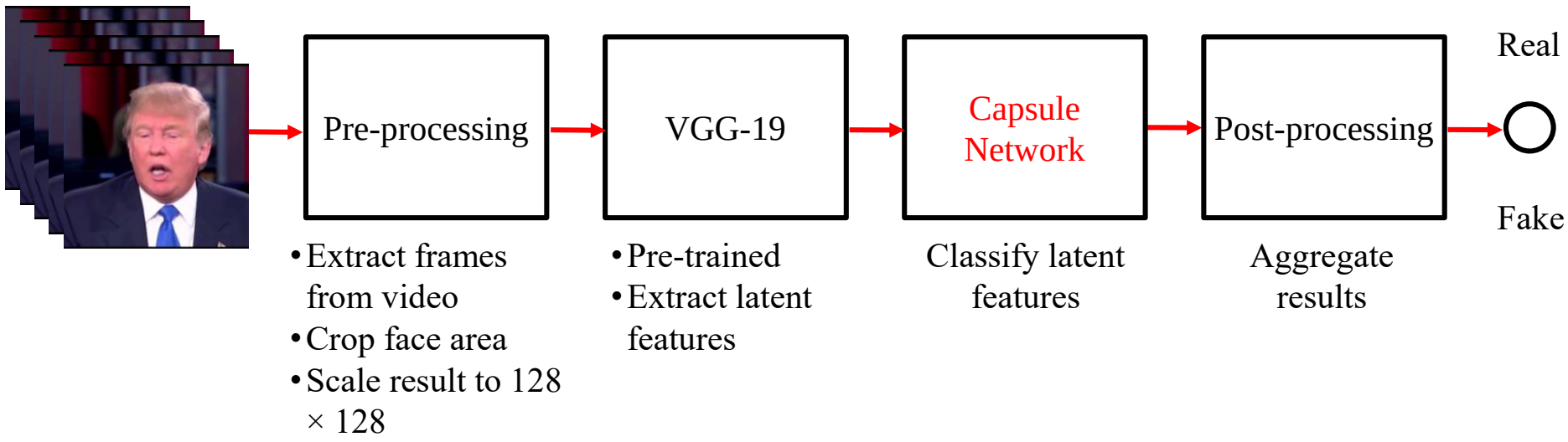
3 - Aggregation over time and decision

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: 456)

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).

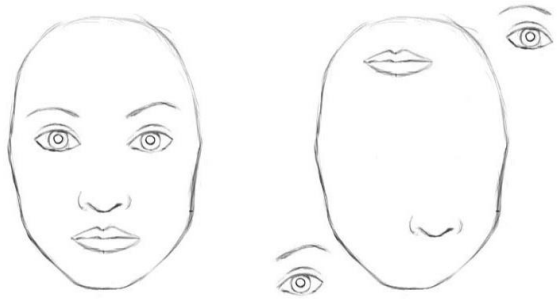
- Combine VGG19 with Capsule Network as a countermeasure



Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos” ICASSP 2019 (number of citations: 190)

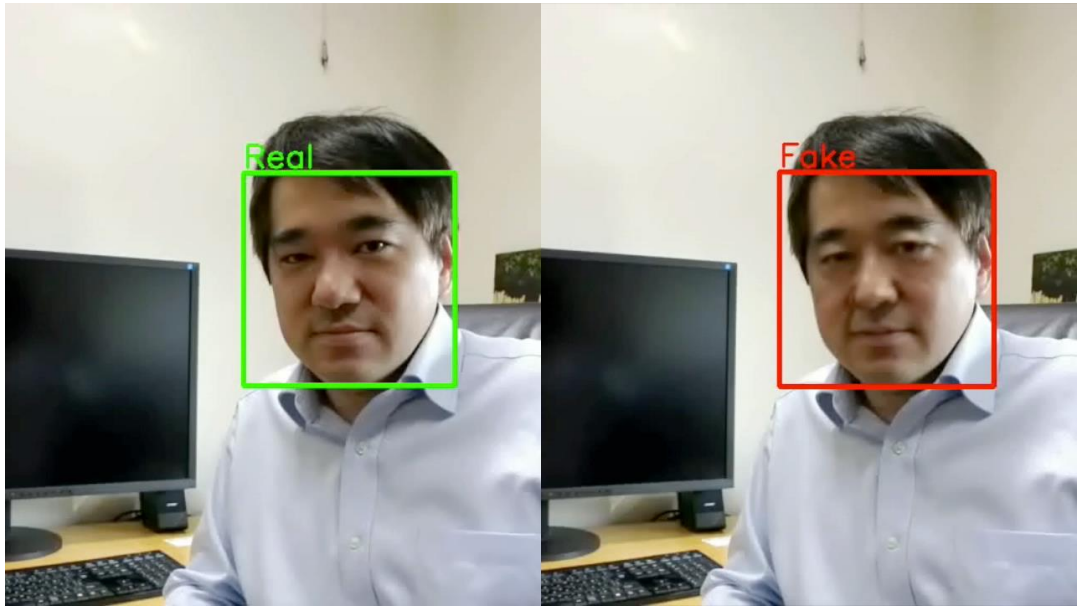
Why capsule networks?

- **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)

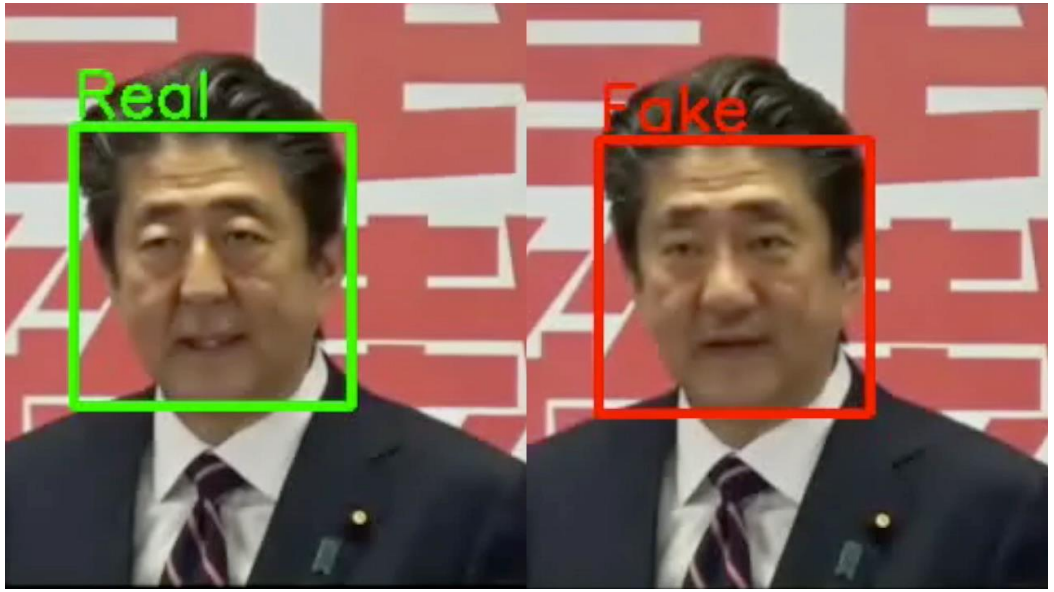


Swap faces using deepfake!

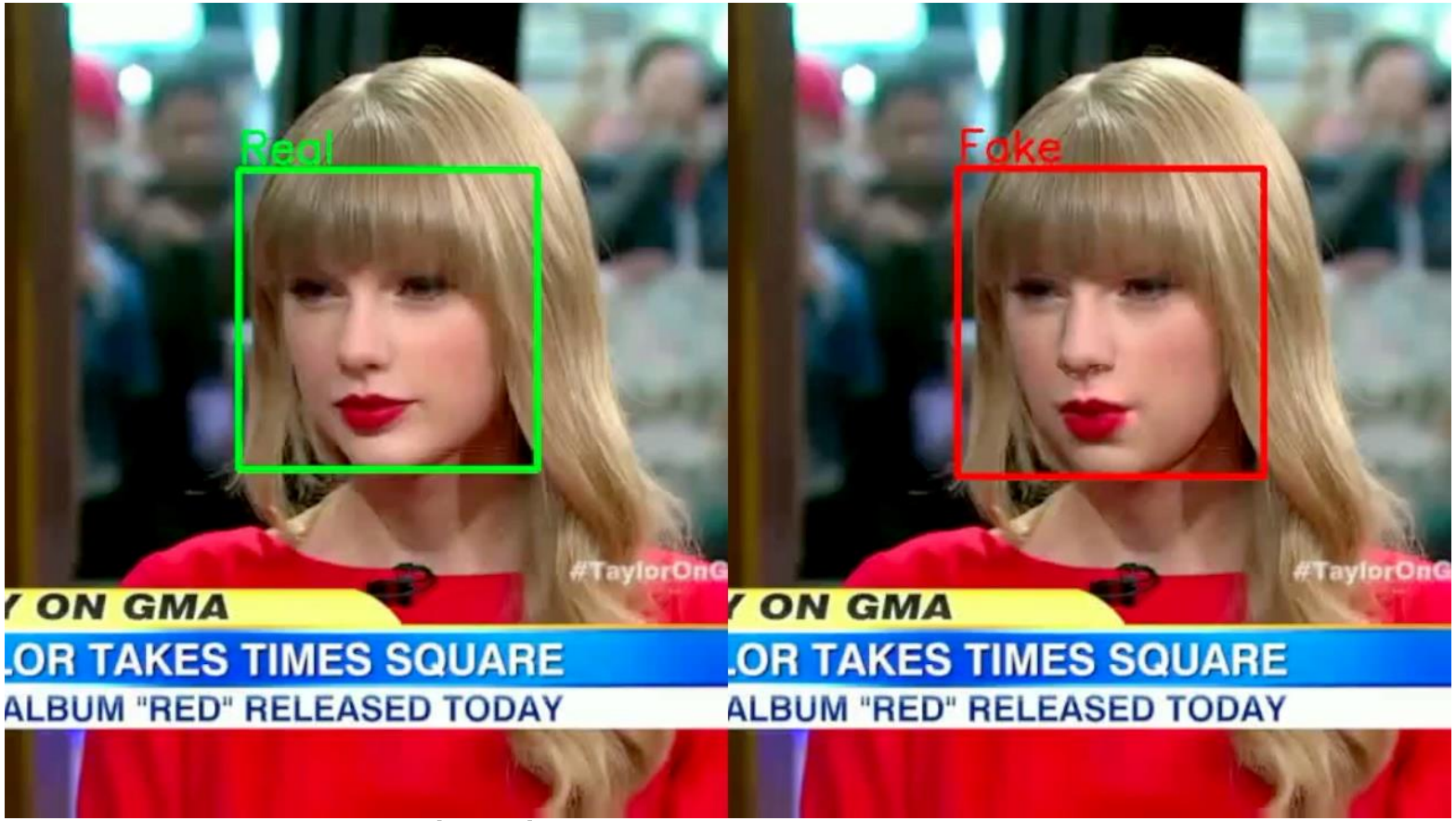
Our Deepfake dataset

	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

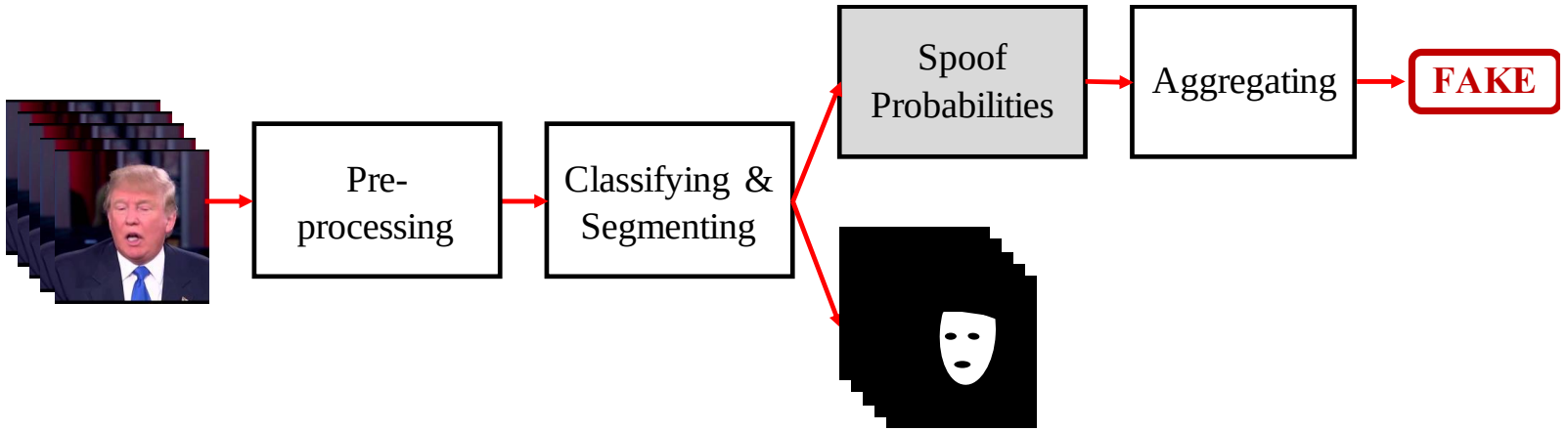
	Real (frame)	Forged (frames)
Train	7,040	7,040
Dev	1,500	1,500
Eval	1,500	1,500

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: 163)


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

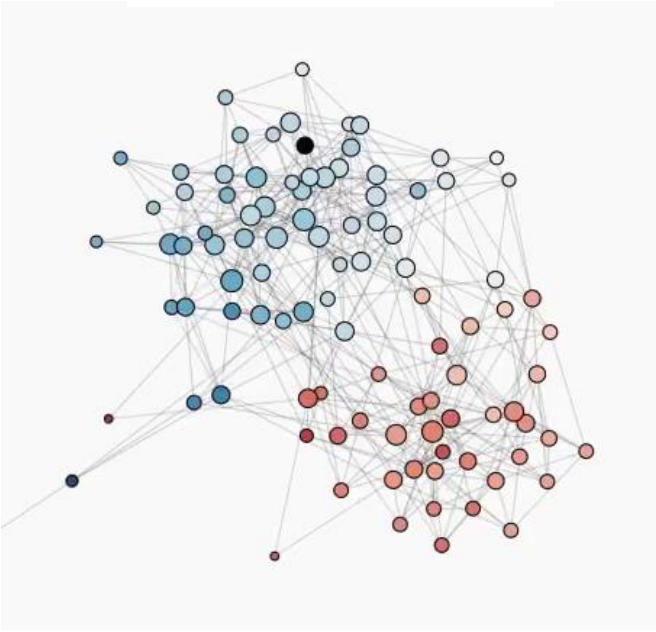
アウトライン

- 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，SynMedia Centerでの取り組み）

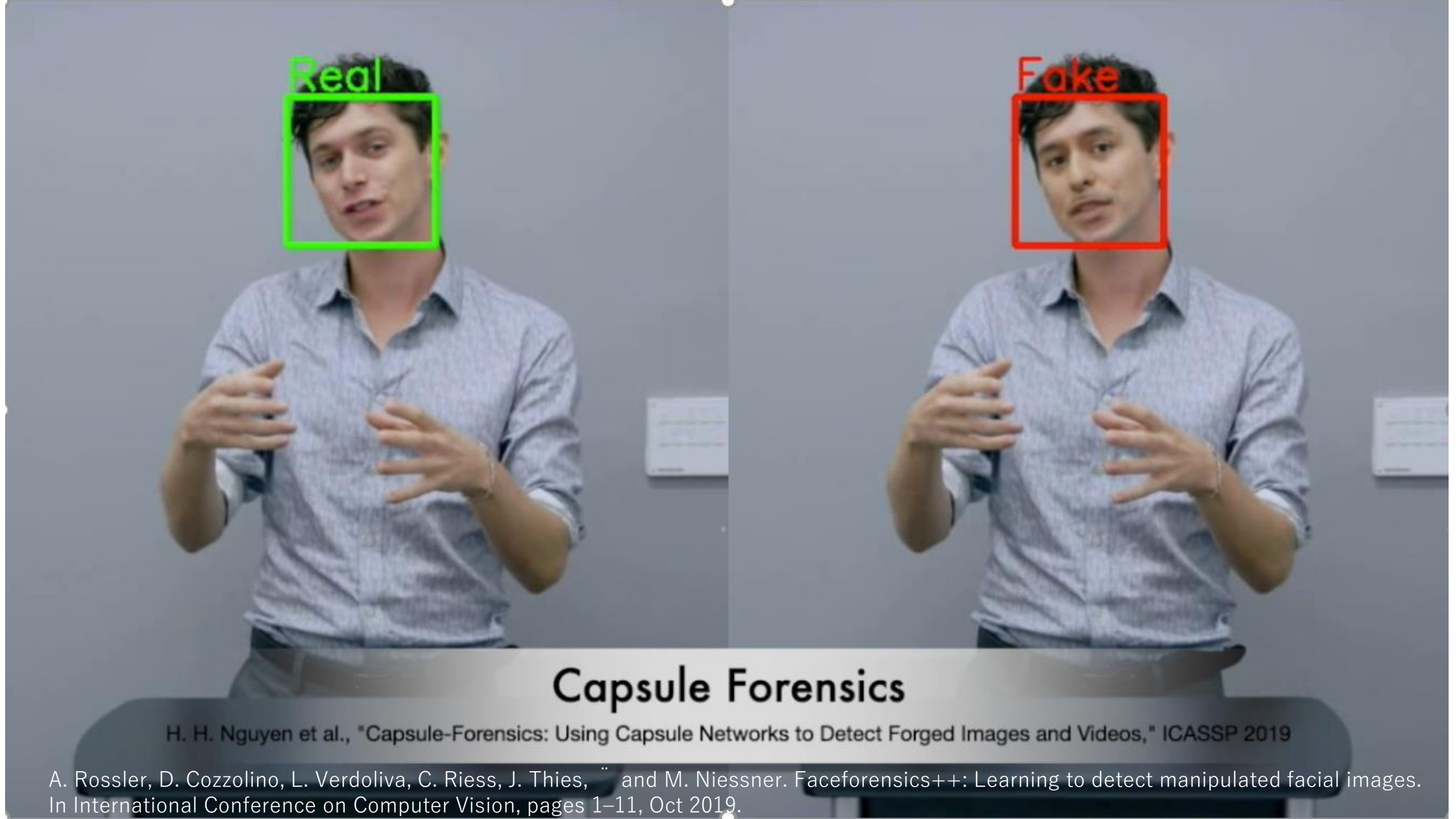
フェイクメディア(FM)とインフォデミック

CREST FakeMedia

- 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（FM）の生成が可能に(2018年～)
  - Deepfake(フェイク顔 2018年)，GROVER（フェイクニュース 2019年）
  - 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）
- 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
- 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



SNSを模した意見形成モデルによるエコーチェンバーの発生
（笹原和俊 東工大准教授）



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.

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

MC型FMの脅威が現実になってきた

- 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

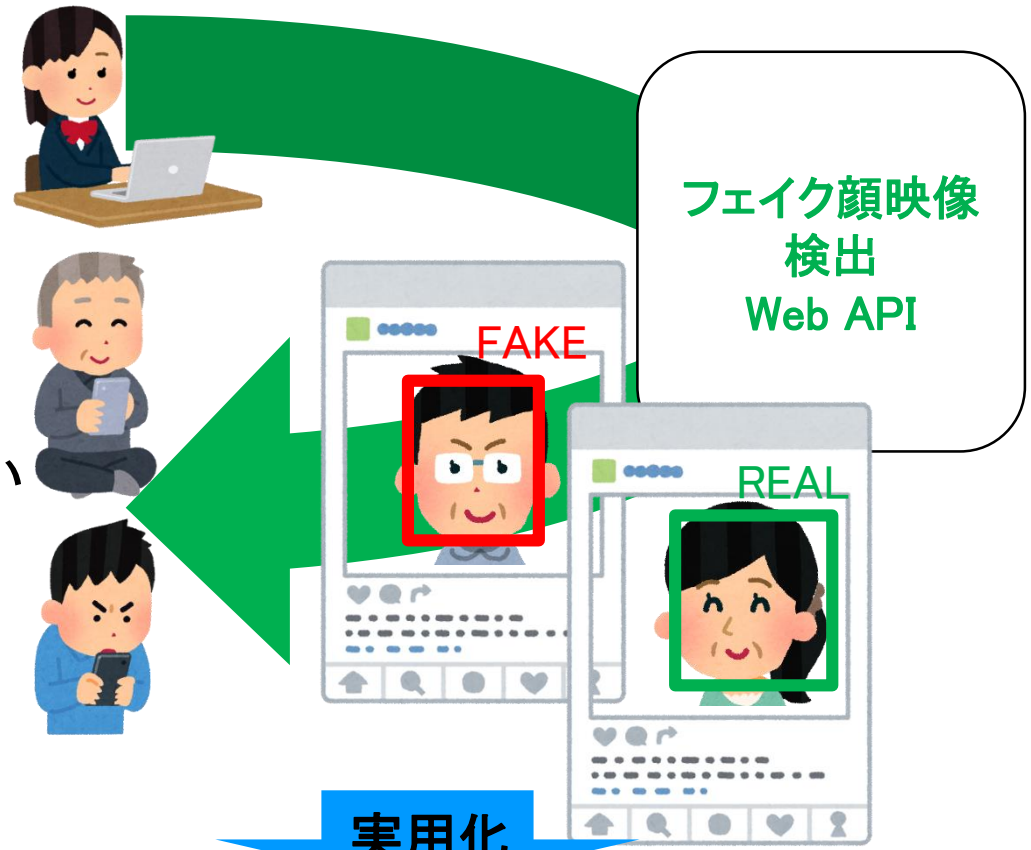
フェイク顔映像検出AIaaSの開発

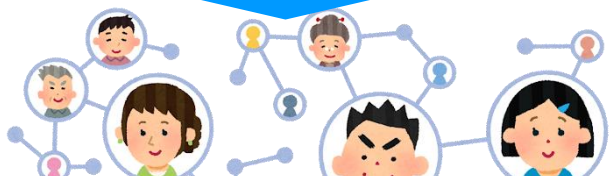
JST A-Stepトライアウト

Deepfake自動検出に興味を持っている企業は多い、しかし
- Deepfake検出には高度な深層学習技術が必要
- 深層学習モデルの学習にはGPUサーバの様なリソースへの投資も必要になる
- 機械学習の知識や設備のない会社にとっては、参入障壁が高い

そこで、学習済みの深層学習モデル、および、前処理・後処理も包含し、即利用可能な状態のサーバプログラムとAPIを作成
- サーバを利用→ 単独でAI as a serviceとして利用可
- APIを利用→既存サービスと組み合わせる事も可能

山岸CRESTと連携して，Deepfake自動検出モデルを利用できる”Deepfake detection AIaaS (AI as a service)”のAPIを開発
- プレスリリース・パートナー企業を募集中





誰でも簡便・迅速に
ファクトチェックが可能に！

1. JST A-Stepトライアウト（標準），FY2021（研究代表者 越前）「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」
2. 山岸CREST(日仏共同提案) VoicePersonae: 声のアイデンティティクローニングと保護

フェイク顔映像検出AIaaSの開発
–SYNTHETIQ: 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

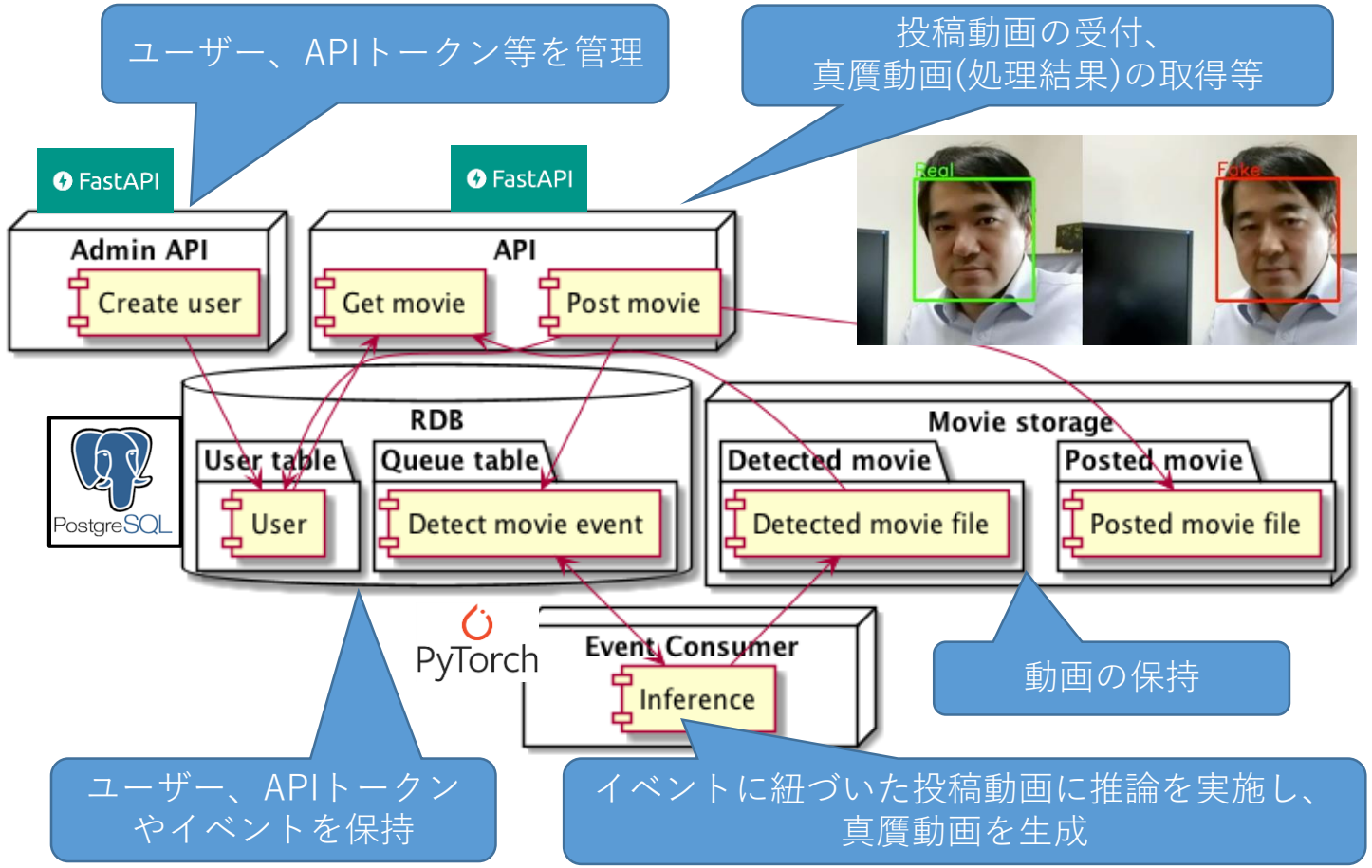
NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

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。



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

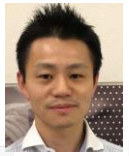
音声情報処理，画像・映像処理，自然言語処理，コンピュータビジョン

馬場口 阪大教授・工学研究科長

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

参考：
SynMedia Centerウェブサイト

SynMedia Center
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
Global Research Center for Synthetic Media,
National Institute of Informatics

ENGLISH　SITE MAP

ホーム　SynMedia Centerについて　メンバー　研究成果　インターン募集

シンセティックメディア国際研究センター（SynMedia Center）は，人間中心のAI社会を実現するために，顔，音声，身体，自然言語などの多様なモダリティを対象とした，シンセティックメディアの生成，不正な目的で生成されたシンセティックメディア（フェイクメディア）の検知，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ことをミッションとしています．

Topics　トピックス一覧へ

2021/07/29　[Recruit] インターン　募集

NII 国立情報学研究所
CREST FakeMedia
ELAB Content Security
国立情報学研究所 山岸研究室

ホーム > SynMedia Centerについて

SynMedia Centerについて

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，本物と見紛うシンセティックメディア（Synthetic media）の生成が可能になりつつあります．シンセティックメディアは，バーチャルアバターなどのコミュニケーション分野や，落語音声合成などのエンターテイメント分野を始めとした様々な分野で活用が期待されており，高品質なシンセティックメディア生成技術の確立が期待されています．一方で，シンセティックメディアの負の側面として，詐欺や思考誘導，世論操作を行う目的で，愉快犯や攻撃者が，フェイク映像，フェイク音声，フェイク文書といったフェイクメディアを生成，流通させる可能性があり，社会問題となっています．

シンセティックメディア国際研究センター（SynMedia Center）は，人間中心のAI社会を実現するために，顔，音声，身体，自然言語などの多様なモダリティを対象とした，シンセティックメディアの生成，不正な目的で生成されたシンセティックメディア（フェイクメディア）の検知，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ため，研究施設（センター）を設置し，研究展開を図ります．

フェイクメディア検出の課題

- SNS共有や再配信のタイミングでリサイズや再圧縮が生じる

  →補間や符号化によるノイズにより品質劣化し，真贋判定精度が低下

- 多種多様なフェイクメディア生成手法が出現

  →機械学習モデルは，未知の手法で生成されたフェイクメディアの検出は苦手

  →定期的な学習データのアップデート・モデル学習が必要

- 画像を対象とした自動ファクトチェックの必要性

  →クエリー画像から類似のオリジナル画像を検索し，オリジナル画像とクエリー画像の比較から，オリジナル画像の真贋判定，改ざん箇所を推定

  →フェイクメディアの生成手法は多岐に渡るため，将来的に機械学習による真贋判定とファクトチェックを相補的に活用することが重要