生成AIの著作権侵害等のリスクとその低減技術動向 シンポジウム Feb_26, 2024

生成AIに関するリスク低減技術について

早稲田大学　理工学術院　先進理工学部
応用物理学科　　**教授　森島繁生**

Deepfake

Deepfakeとは

深層学習を用いたメディア合成技術

特徴

①本物と見間違えるほどの写実性

②専門的知識が無くても作成可能

映像制作や広告作成の効率化が可能

- 俳優の顔や年齢の操作・故人の復活
- 様々な広告のバリエーションに対応

Synthesizing Obama：Learning Lip Sync from Audio, SIGGRAPH 2017.

You Won't Believe What Obama Says In This Video! ☺
url: https://www.youtube.com/watch?v=cQ54GDm1eL0
（注）この動画は注意喚起のためのもの

任意の動画コンテンツに対し顔交換を可能に

Simswap: An Efficient Framework For High Fidelity Face Swapping
(Chen, et al., *ACM MM*, 2020)

元動画

交換したい顔

生成結果

フレームごとに処理することで**動画にも拡張可能**

Deepfakeを活用したコンテンツ制作

インターネット広告配信の効率化

クライアント（年齢、性別、収入、趣味、購買
履歴などの属性）に合わせた広告自動配信

基本的な思想

①俳優・商品については忠実デジタル再現（モデルベース）

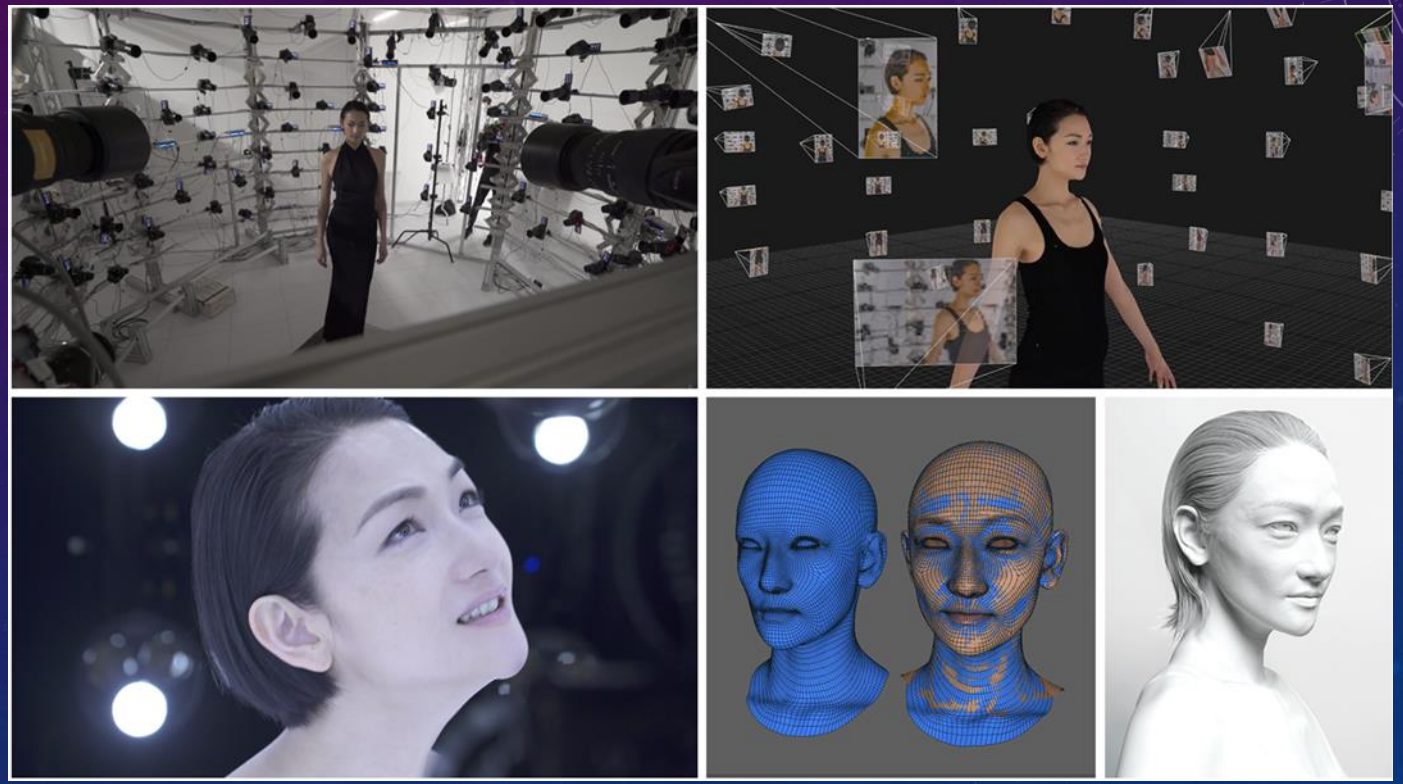
②それ以外の部分に関しては生成AI（プロンプト等）

広告作成のコストが削減

インタラクティブな広告配信が可能に

デジタルツインレーベル

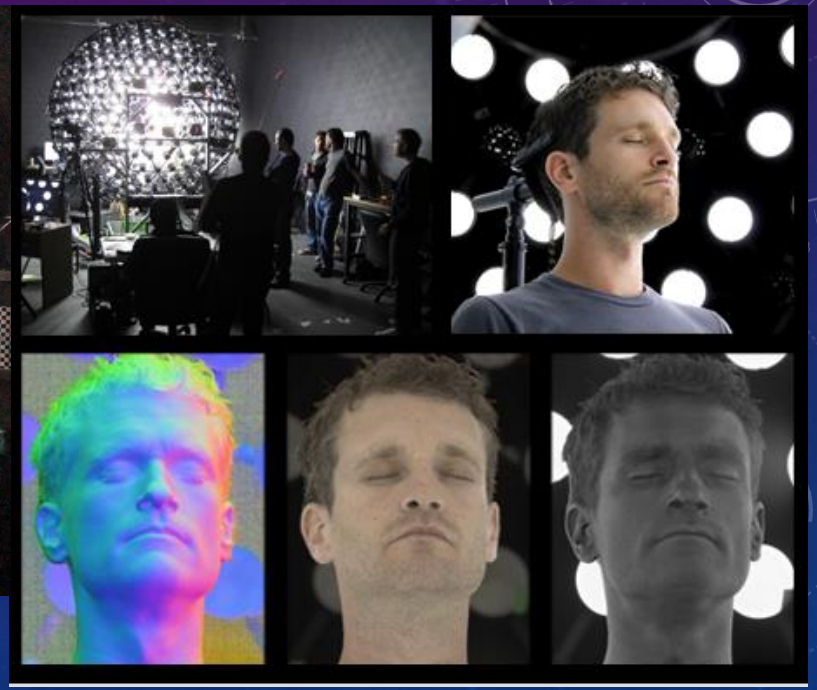
（広告業界でのDigital Human応用） By CA



Digital Human on the cutting edge（故人の蘇り）

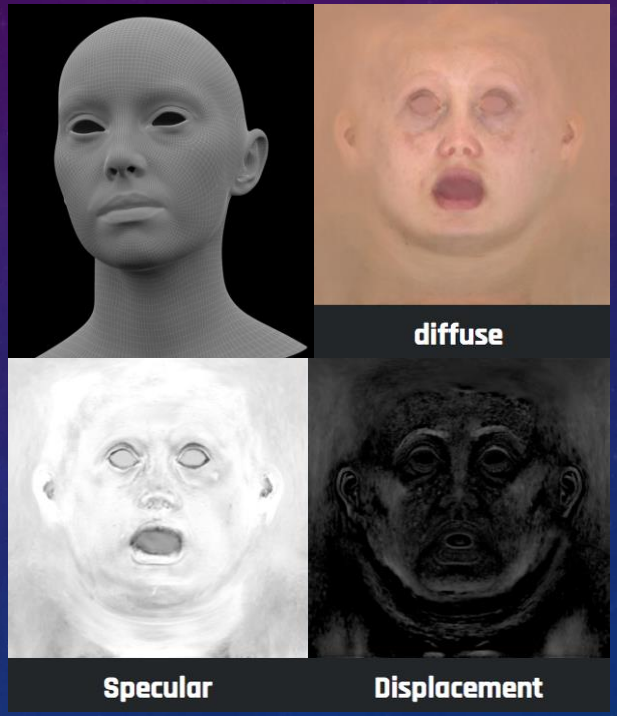
Weta Digital – ”Furious 7 VFX | Breakdown - Brian O'Conner | Weta Digital” (https://www.youtube.com/watch?v=ye7arp5IrAg)

How to Make a Digital Human:
Light Stage Capture System by USC ICT

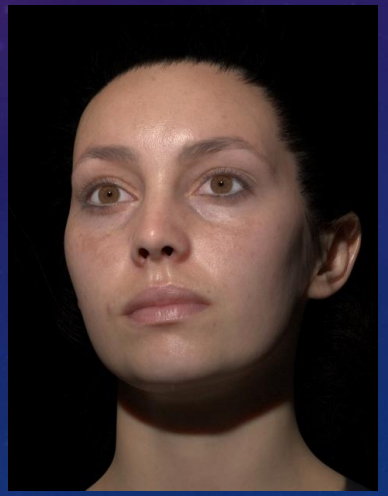


Light Stage導入実績：凸版、東映、CyberAgent(light cage)

Captured Images



3D Geometry and Reflectances



Rendering

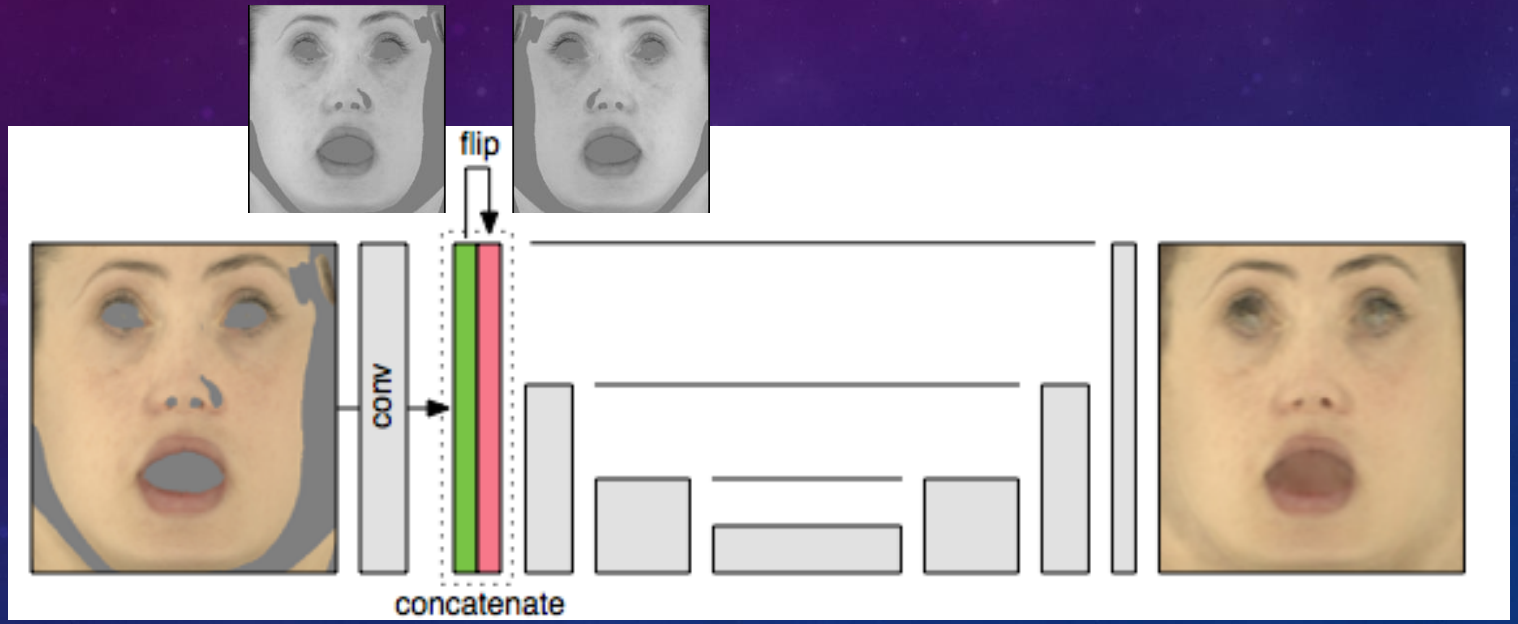
OUR GOAL: Easy and High-quality Digital Human

Selfie

Digital Human

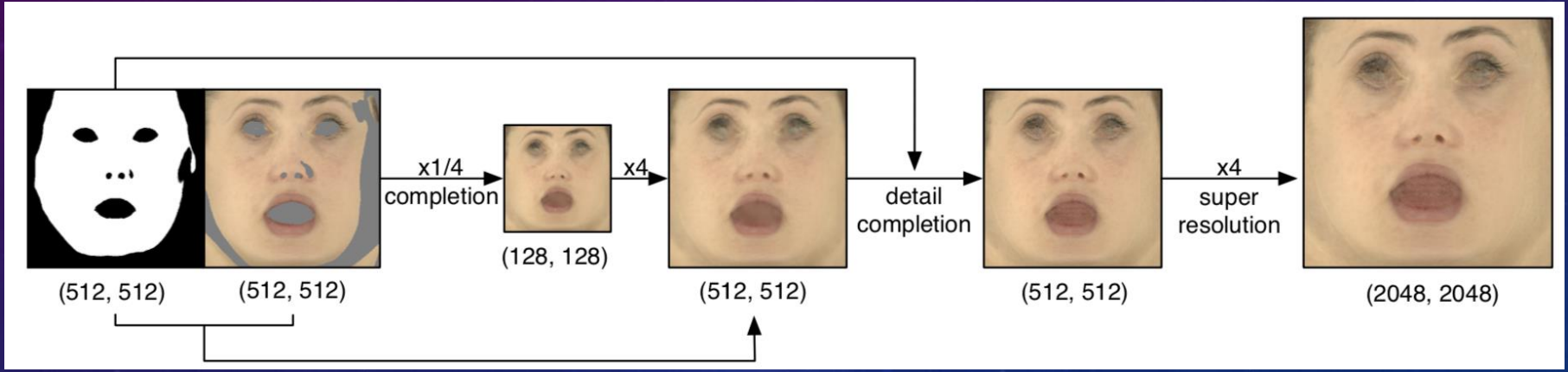
High-Fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image
Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Shigeo Morishima, Hao Li
ACM Transactions on Graphics (Proc. SIGGRAPH 2018), Vol. 37, Article 162

TEXTURE COMPLETION



U-net with latent flip

TEXTURE COMPLETION



input image

output textured 3d face
(illumination 1)

output textured 3d face
(illumination 2)

input image

output textured 3D face
(illumination 1)

output textured 3D face
(illumination 2)

input image

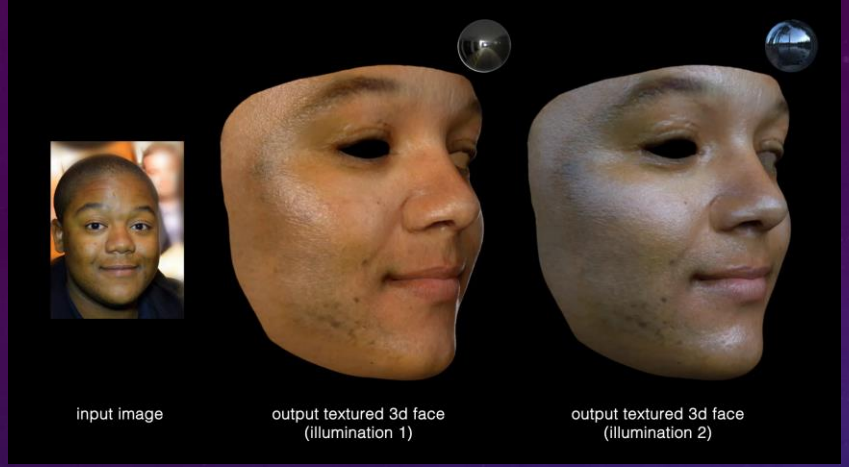
output textured 3d face
(illumination 2)

zoom-in

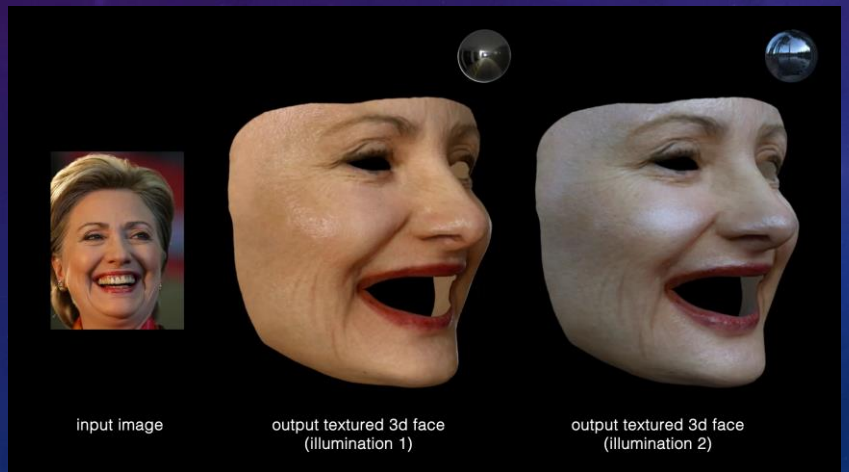
input image

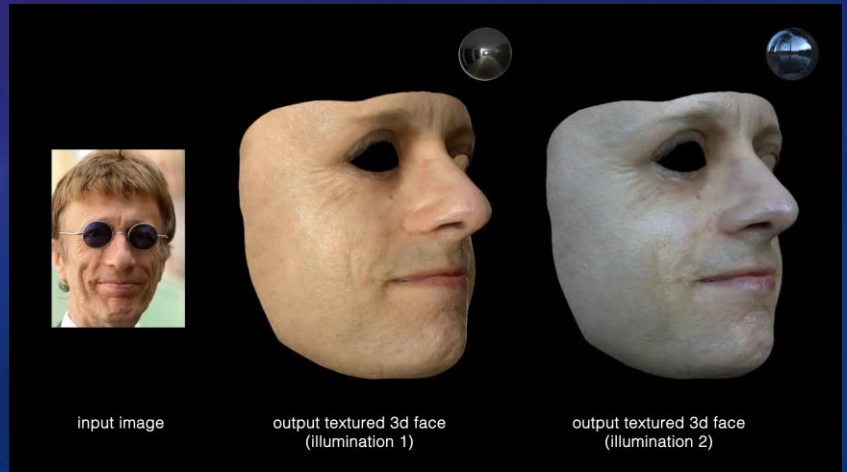
output textured 3d face
(illumination 2)

zoom-in


input image
output textured 3d face
(illumination 1)
output textured 3d face
(illumination 2)

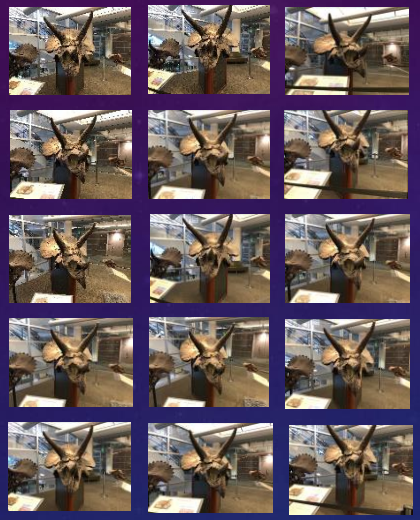

input image
output textured 3d face
(illumination 1)
output textured 3d face
(illumination 2)


input image
output textured 3d face
(illumination 1)
output textured 3d face
(illumination 2)

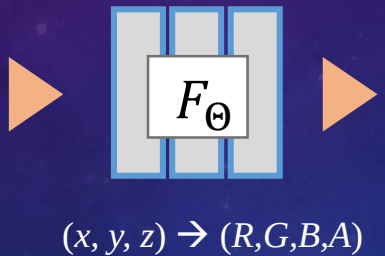

input image
output textured 3d face
(illumination 1)
output textured 3d face
(illumination 2)

空間を表現する技術：Neural Fields

入力画像群を元にニューラルネットワークを用いて3D表現を獲得する手法



入力画像群



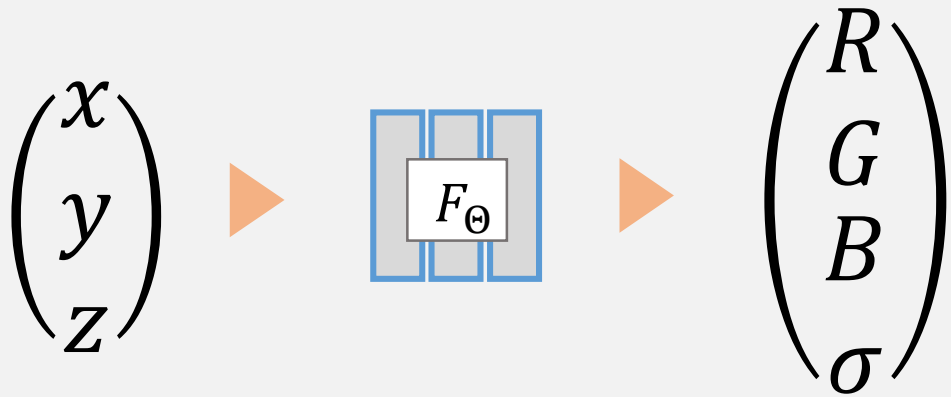
ネットワークの学習

任意視点生成

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis [ECCV 2020 Best Paper]

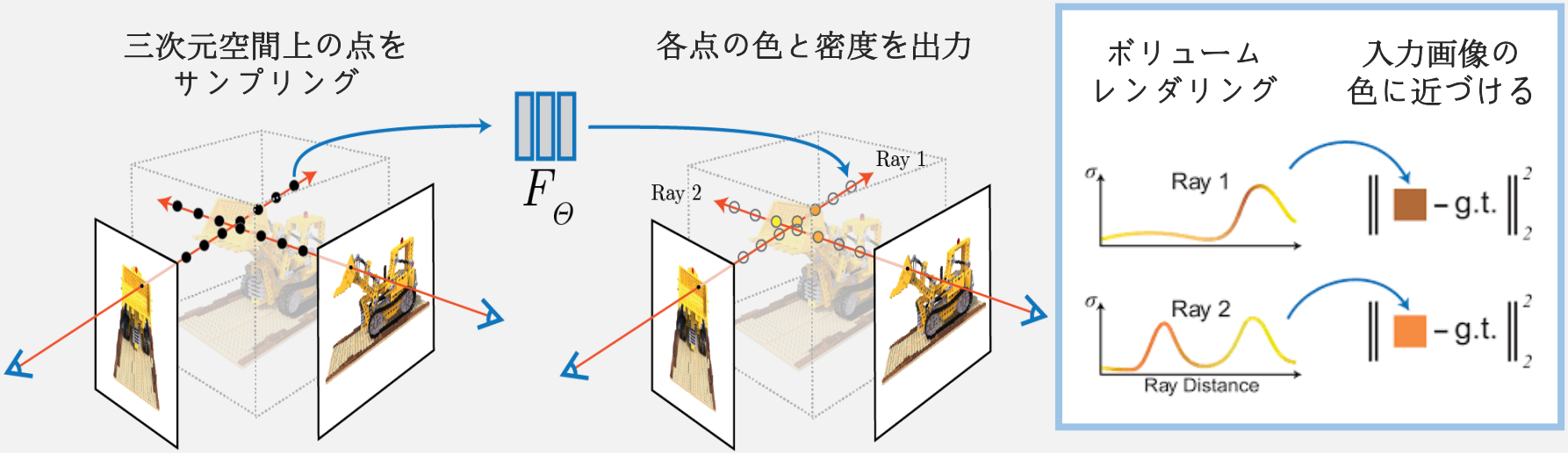
NeRFのネットワークの入出力(陰関数表現)

XYZ座標の値を色(RGB)と密度(σ)に変換するネットワーク



NeRFのレンダリング

XYZ座標の値を色(RGB)と密度(σ)に変換しボリュームレンダリング

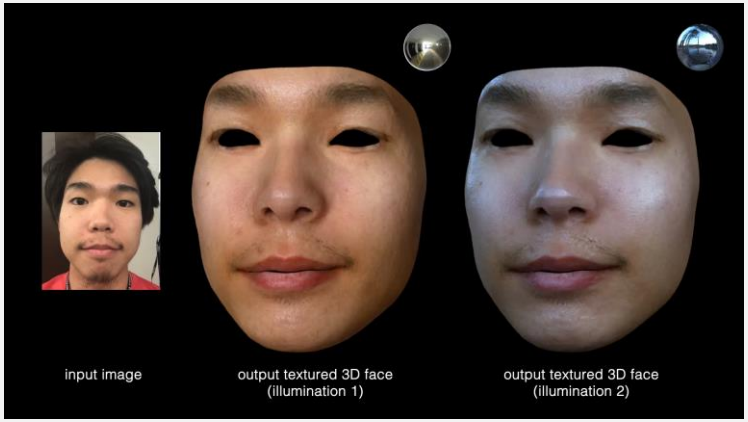


視線方向依存の効果も獲得可能

- 入力には，XYZ座標だけでなく視線方向も使用可能
- 鏡面反射のような視線方向に依存する効果もレンダリング手法の設計をすることなく学習可能

Neural Fieldsのいいところ

メッシュ表現



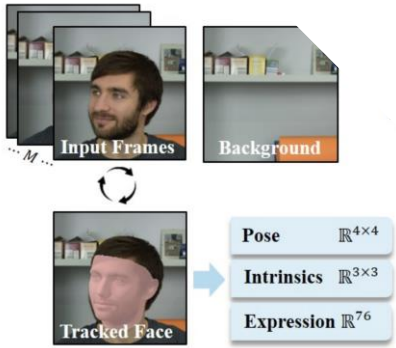
- 難しいレンダリングの設計が必要
  - 微細な反射，表面下散乱などなど
- パーツごとにモデル化が必要
  - 眼球・髪など

Neural Fields

- “リアルな”表現をすることが簡単
  - 入力した写真と同等のリアルさ
- 映ったものすべてを獲得可能

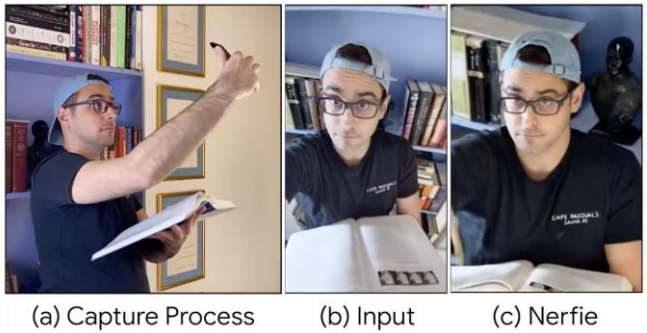
動画ベースの頭部モデル作成手法(撮影方法別)

単眼固定
(頭部の姿勢を利用)



NerFACE,
NeRFBlendShapeなど

単眼自由



Nerfies, HyperNeRFなど

多眼固定

NeRSembleなど

動画ベースの頭部モデル作成手法(撮影方法別)

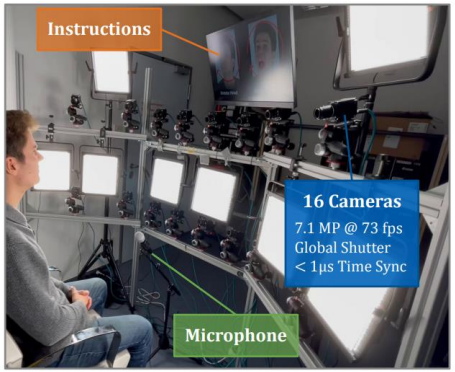
単眼固定
(頭部の姿勢を利用)

首から上のみ

単眼自由

欠損が発生しやすい

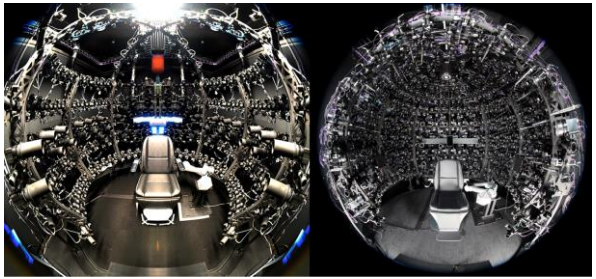
多眼固定



コストが高い

(補足)多眼固定

Multiface (Meta)

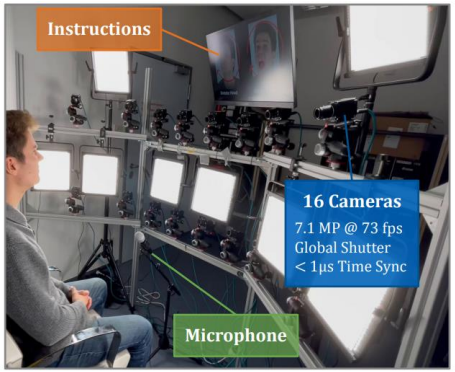


LightStage系
動画撮影可

DyNeRF

GoPro 18台
(信号で同期)

NeRSemble

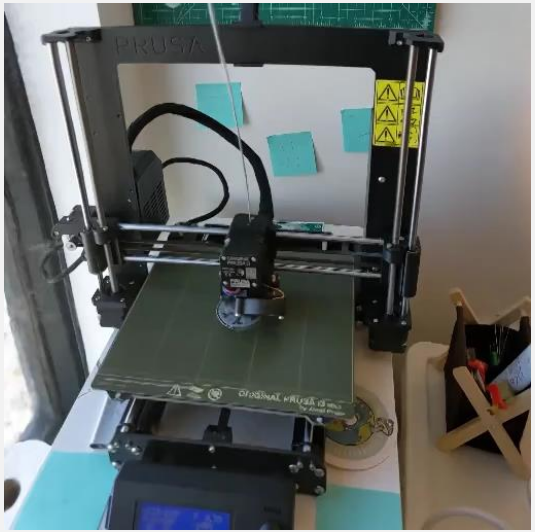


Global Shutterカメラ 16台
(信号で同期)

Neural Fieldsの主な課題とそれにトライした手法

動く物体の表現

- 通常のNeRFは静止した物体のみ
- HyperNeRFは変形のネットワークを使用し動きも学習可能



HyperNeRF: A Higher-Dimensional Representationfor Topologically Varying Neural Radiance Fields [ACM Trans. Graph. 2021]

Neural Fieldsの主な課題とそれにトライした手法

撮影条件の異なる画像での学習

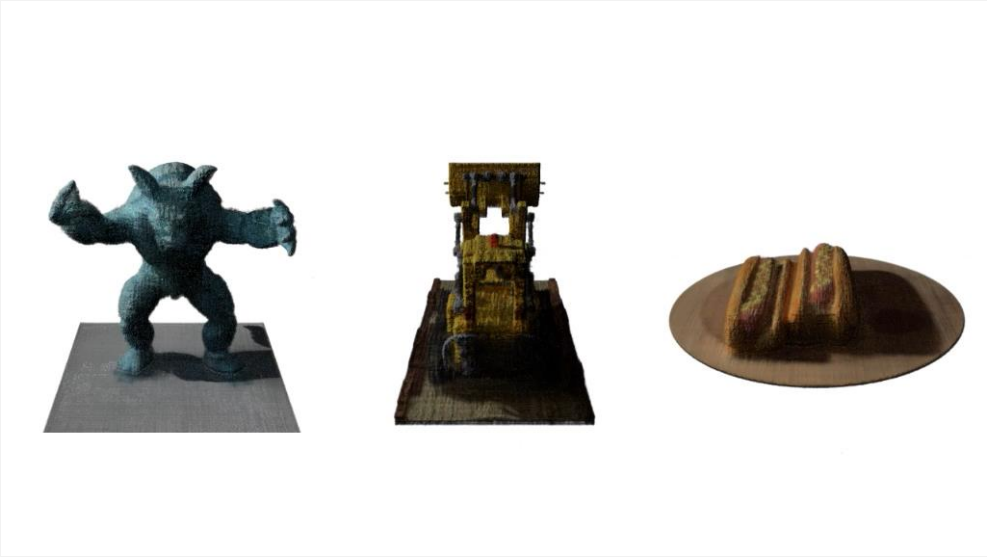
- 通常のNeRFはひとつのシーンしかNeural Field化できない
- NeRF-in-the-wildは観光客の撮影した写真群から物体の復元+様々な撮影環境を再現可能

NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections [CVPR 2021]

Neural Fieldsの主な課題とそれにトライした手法

再照明

- 通常のNeRFは撮影時の照明の再現しかできない
- NeRVは成分の分離と再照明が可能

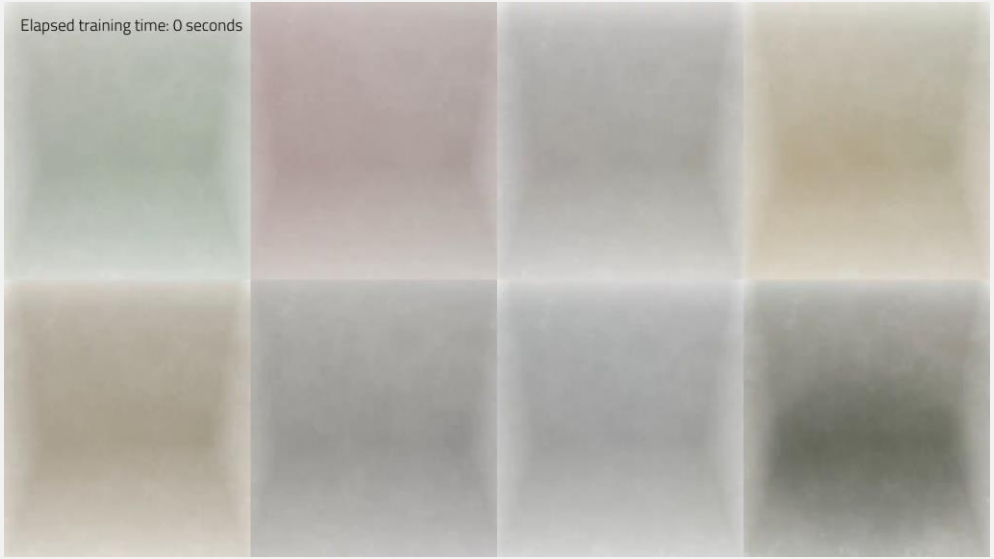


NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis [CVPR 2021]

Neural Fieldsの主な課題とそれにトライした手法

学習とレンダリングの重さ

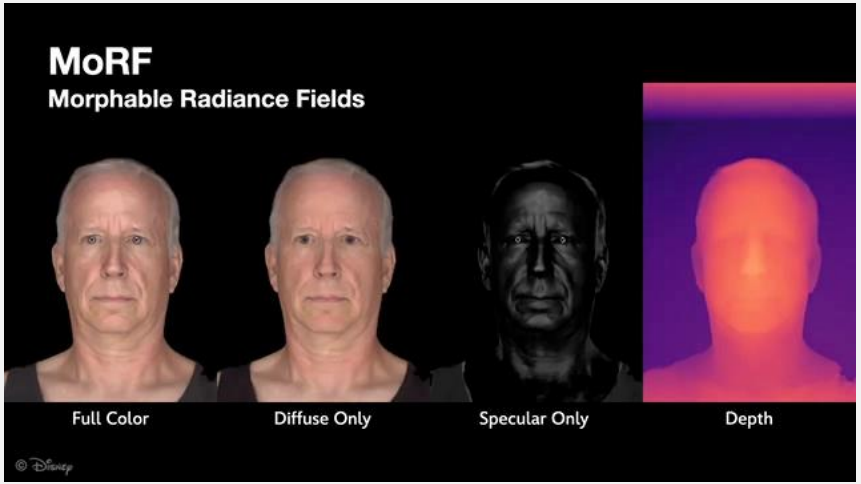
- 通常のNeRFは学習に数時間，１枚のレンダリングに数分
- Instant NGPは数秒で学習，リアルタイムでレンダリング



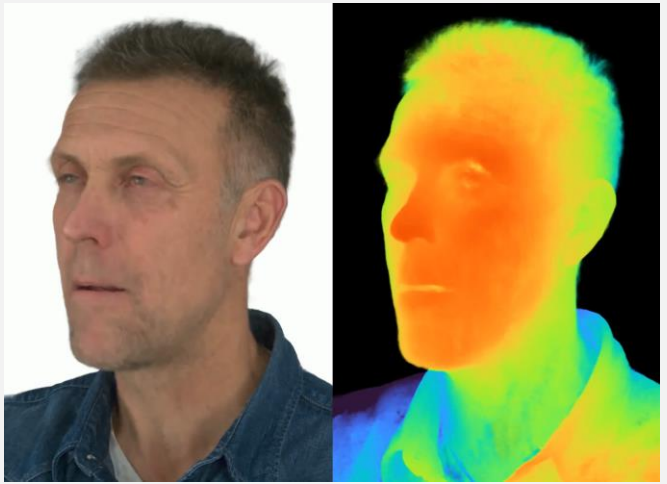
Instant Neural Graphics Primitives with a Multiresolution Hash Encoding [SIGGRAPH 2022]

NeRFの顔への応用

複数の顔のNeRFモデルを
モーフィング可能



顔の表情変化を忠実に表現可能



MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling [SIGGRAPH 2022]
NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads [SIGGRAPH 2023]

PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization


Shunsuke Saito[1,2*]


Zeng Huang[1,2*]


Ryota Natsume[3*]

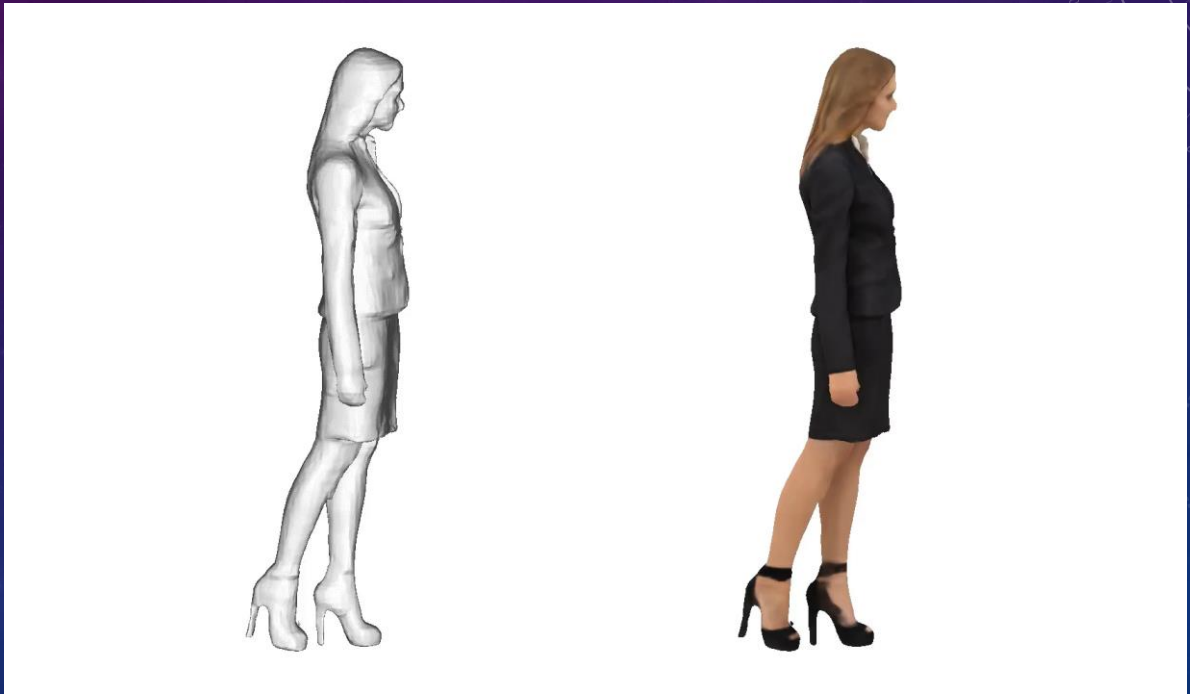

Shigeo Morishima[3]


Angjoo Kanazawa[4]


Hao Li[1,2,5]

2019年11月にICCV 2019で発表、4年3か月で引用数1138件。
現在、Neural Field分野の先駆的な研究として位置づけられる

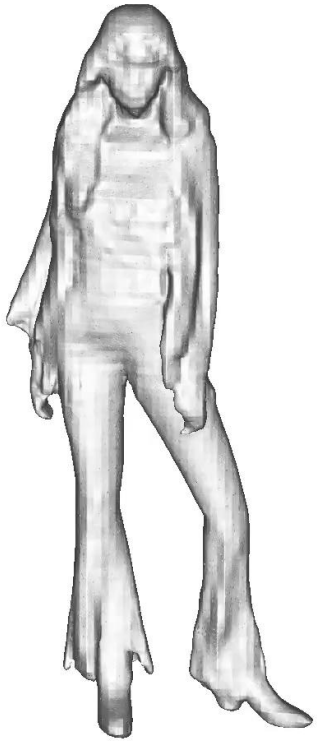
カメラアングル、照明条件、ポーズ、衣服の色・形状のバリエーション、衣服以外のアクセサリー、髪型のバリエーションに対応



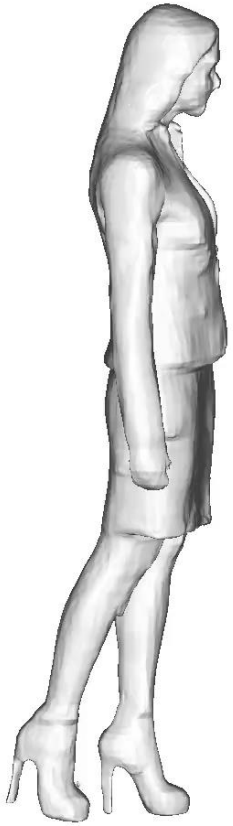
3D geometry estimation

Texture estimation

様々な服の形状に対応



複数枚画像の入力にも対応



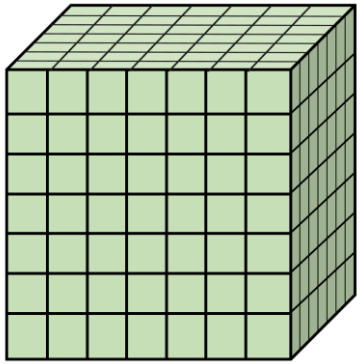
Key Idea: 詳細な情報を含む高解像度な三次元復元

Memory Efficiency and Spatial Alignment

Key Idea: 詳細な情報を含む高解像度な三次元復元

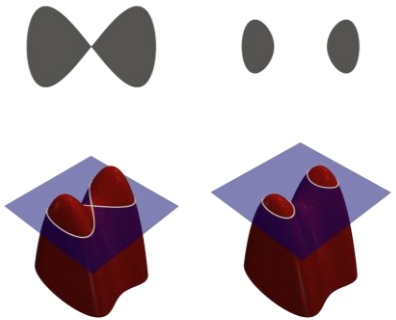
Memory Efficiency and Spatial Alignment

✕ 限定された解像度



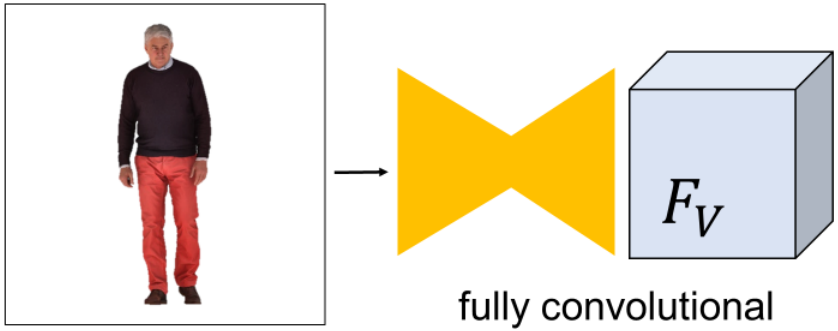
Voxel (Explicit)

✓ 連続値による表現



Implicit Surface

Pixel-Aligned Implicit Function (PIFu)



input

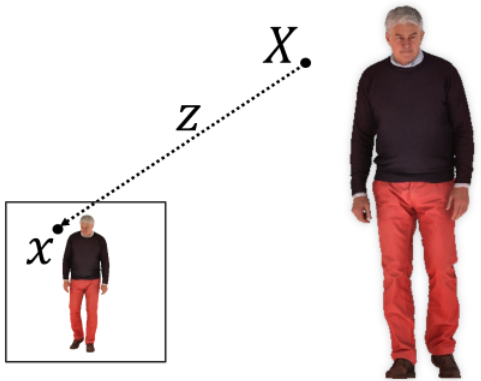
fully convolutional
image encoder

PIFu ($\forall x, \forall z$)

$$f(\text{image feature}, z) = F_V(x)$$

inside/outside

- ✓ メモリ効率が高い
- ✓ 連続値での扱い
- ✓ 局所的な詳細情報の保持
- ✓ 複雑な形状も復元可能



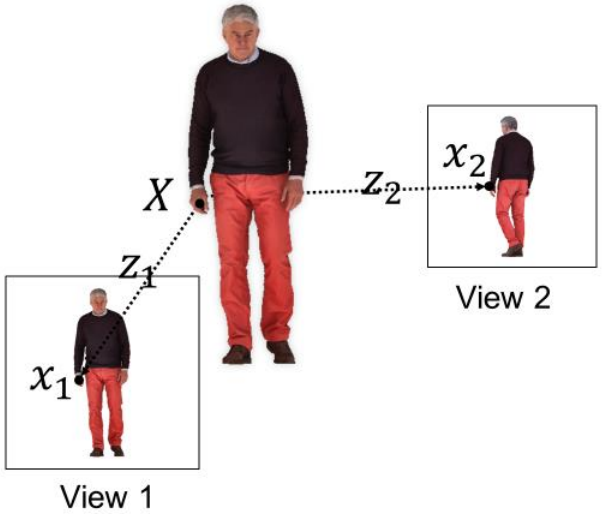
PIFu as a General Framework

Texture Inpainting

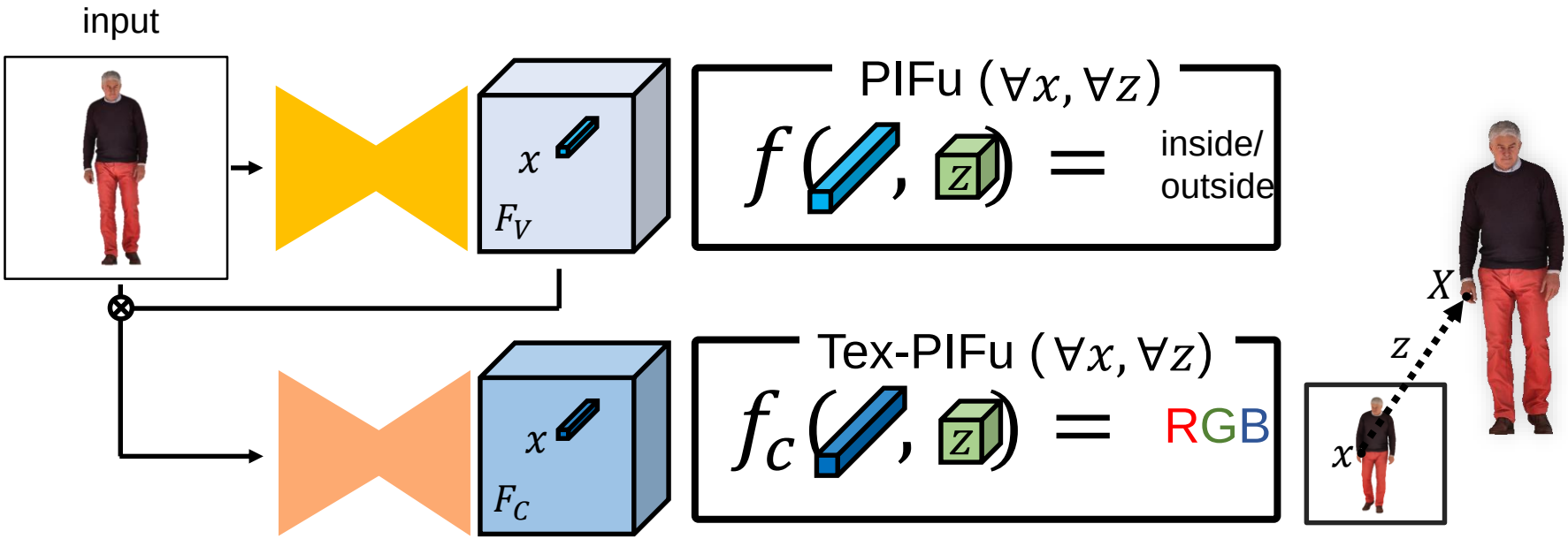
Tex-PIFu $(\forall x, \forall z)$

$f_c($  $) =$ RGB

Multi-view Incorporation



Training Pipeline

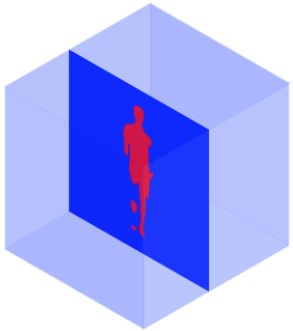


Test Time: Human Digitization Pipeline

n-view inputs
$(n \geq 1)$

Test Time: Human Digitization Pipeline

PIFu



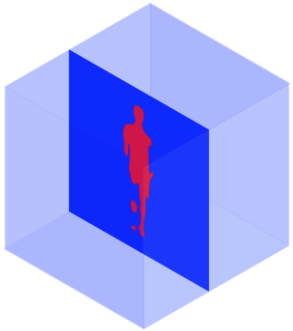
3D occupancy field

n-view inputs
$(n \geq 1)$

Test Time: Human Digitization Pipeline

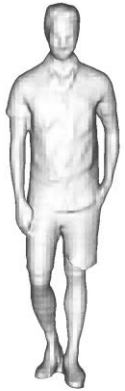
n-view inputs
$(n \geq 1)$

PIFu



3D occupancy field

Marching Cube

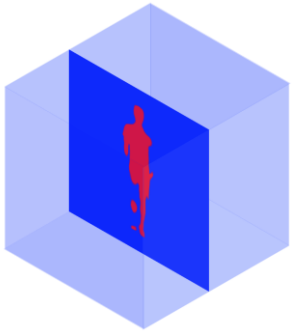


reconstructed geometry

Test Time: Human Digitization Pipeline

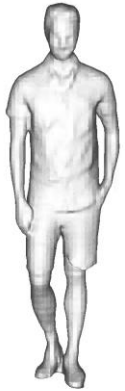
n-view inputs
$(n \geq 1)$

PIFu



3D occupancy field

Marching Cube



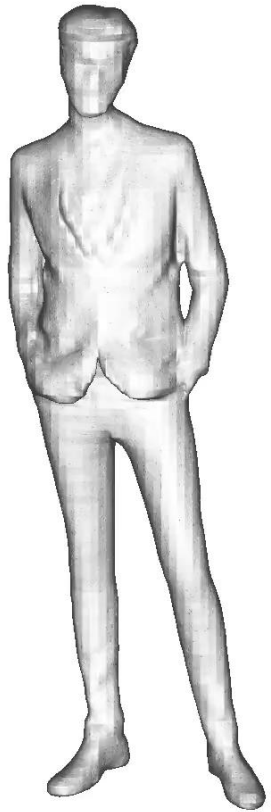
reconstructed geometry

Tex-PIFu

textured reconstruction

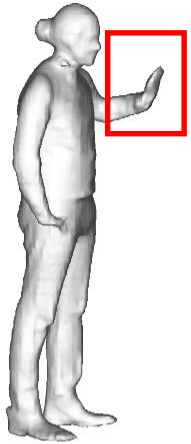
Results on Internet Photos

DeepFashion Dataset [Liu et al. 2016]



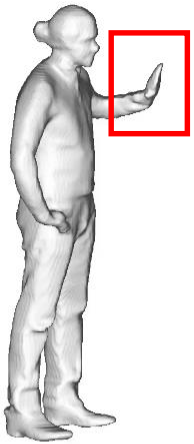
Multi-view Results

#views: 1

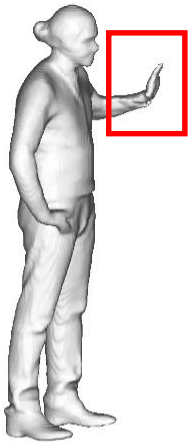


#views: 3

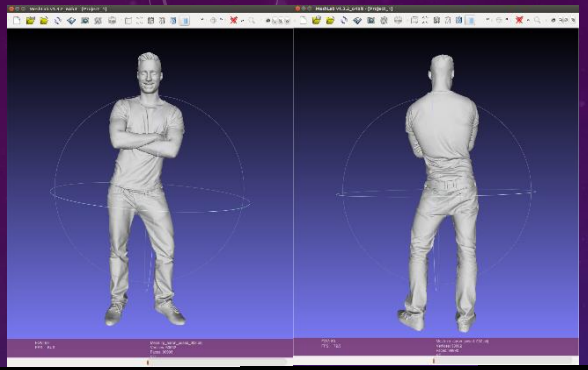
#views: 6

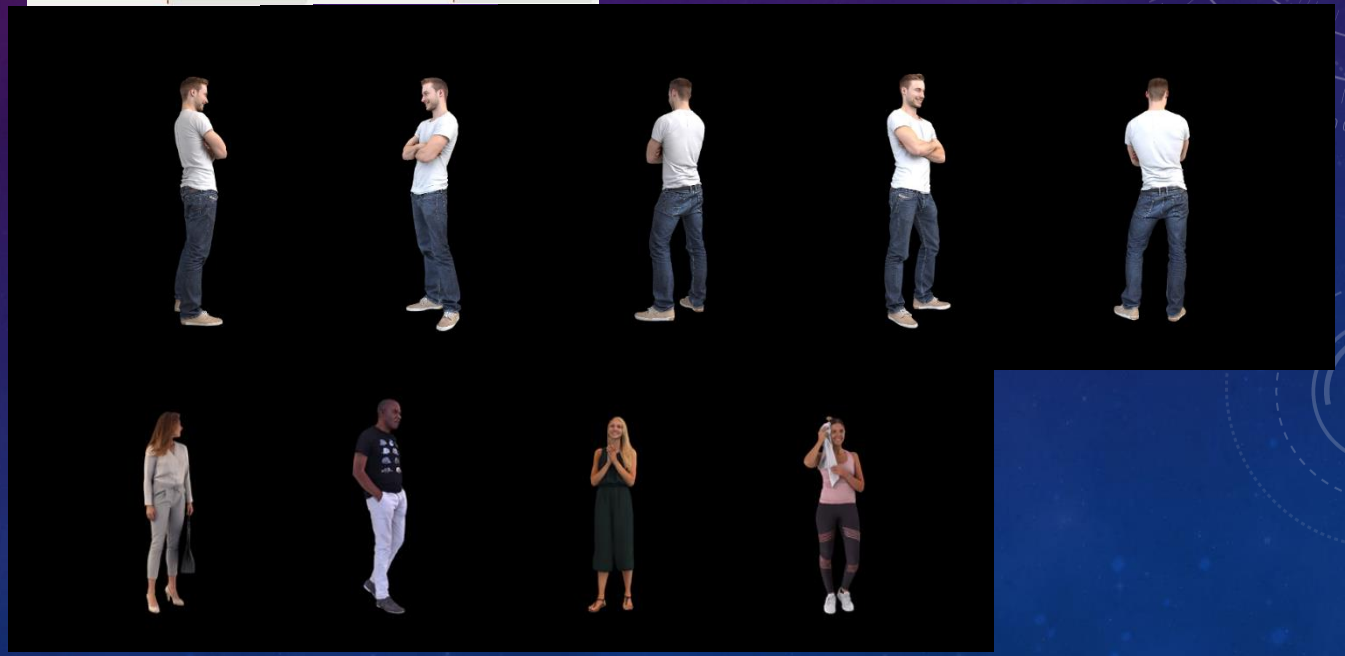


#views: 9



DATASET





PIFu-HD at CVPR 2020

[国際会議 CVPR 2019, ICCV 2019]

創作支援：Dancing Portrait

□ フォトリアルなキャラクタのダンス動画生成

・ 音楽に同期したフォトリアルなキャラクタの
ダンスアニメーション動画の自動生成

着衣全身写真1枚から高精細全身モデルを生成してダンスが可能に

着衣全身写真1枚

研究成果

自動生成

高精細全身モデル

ボーン自動挿入

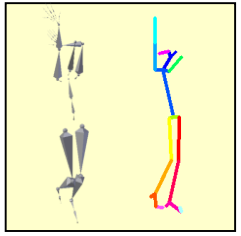
全身モデルによる
ダンスアニメーション

入力

ダンス動画

研究成果

自動生成



ダンスモーション

出力

ダンスキャラクタ自動生成フロー

1 枚の着衣全身画像を入力すると、自動的に音楽にシンクロしたダンス動画が出力される。なお②PIFu に基づいて生成されたキャラクタは、Unity 上でリアルタイム CG 合成される。

①背景削除

removebg

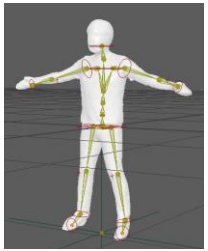
②三次元復元
Textrue 推定
（ PIFu）

研究成果

③テクスチャ抽出

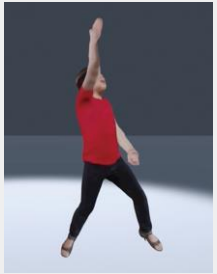
blender

④リギング



⑤リアルタイム
アニメーション
レンダリング

unity



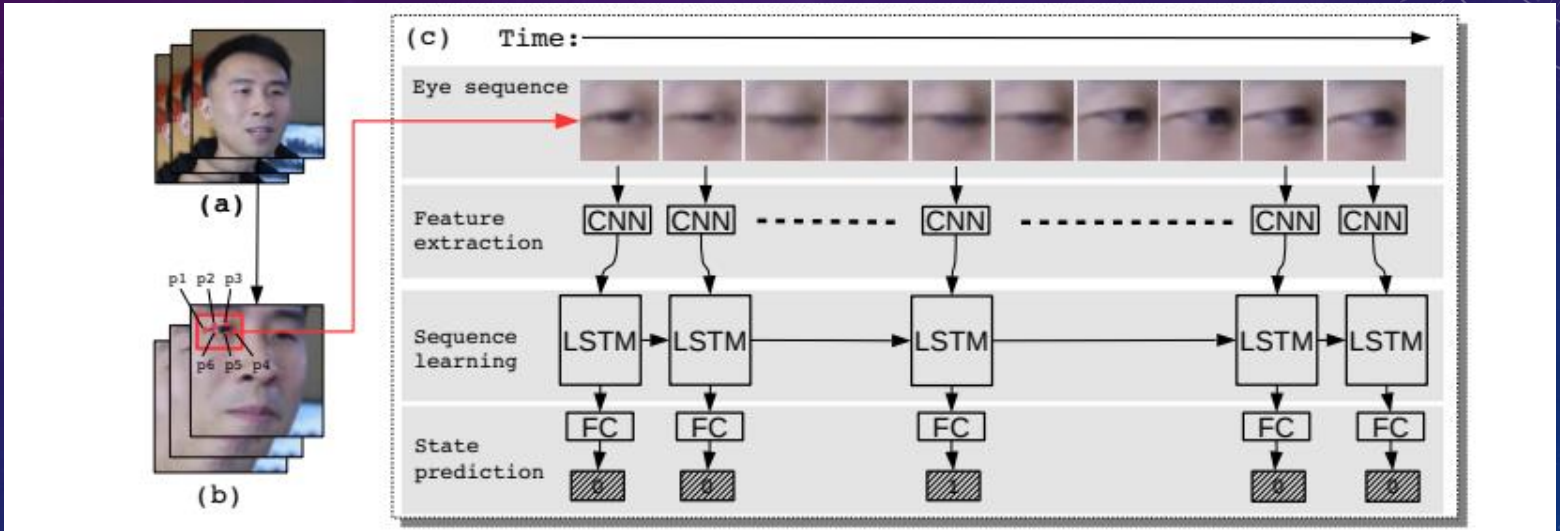
PIFu++&Face統合モデルによる歌唱合成

リスク低減技術（いたちごっこ）の現状

Deepfake検出（１）

不自然なアーティファクト検出

In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking
(Li et al., *IEEE WIFS*, 2018)



フレームごとの処理から生じる**まばたきの不自然さ**を検知

Deepfake検出（２）

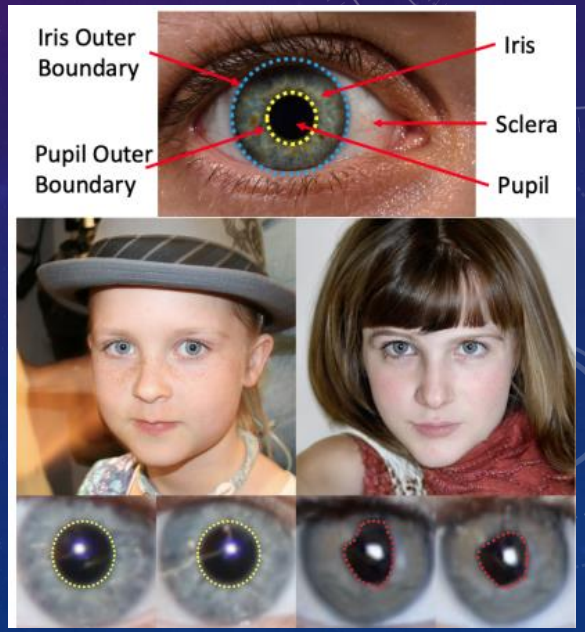
不自然なアーティファクト検出

Eyes Tell All: Irregular Pupil Shapes Reveal GAN-generated Faces
(Guo et al., ArXiv, 2021)

Deepfake生成によく用いる敵対的ネットワーク(GAN)は
実画像を学習することで**それらしい見た目の画像を生成**可能

しかし、**局所的な部分は正確に再現されていない**ことが多い

特に**目の虹彩の形が正確に再現されていない**ため
Deepfake検出のためのアーティファクトとして検知可能



Deepfake検出（３）

不自然なアーティファクト検出

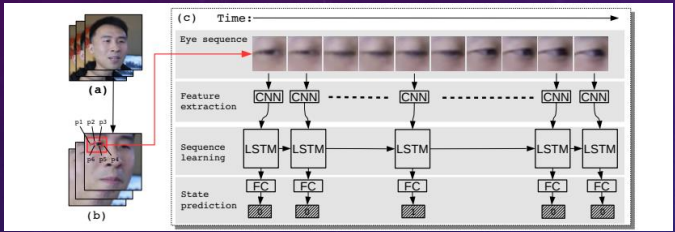
Face X-Ray for More General Face Forgery Detection (Li et al., *CVPR*, 2020)

顔交換によって生じるブレンド境界を検出

Deepfake検出の課題

不自然なアーティファクト検出

欠点：目に見える不自然さはすぐに対応され、より自然なDeepfakeが開発される



(例)

- 瞬きの不自然さを検出⇒瞬きをより自然にしたDeepfake
- 目の虹彩の形の不自然さを検出⇒目の虹彩の形を考慮したDeepfake
- ブレンド境界の検出⇒ブレンド境界が生じないようなDeepfake

すぐには対応できないような検出手法が望ましい

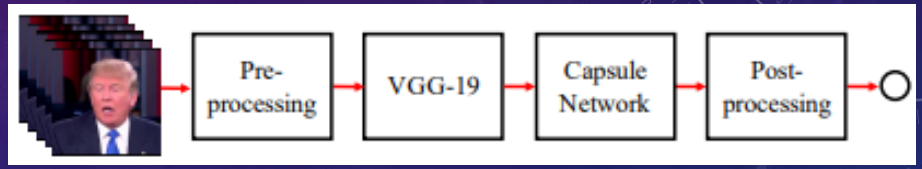
特徴量レベルでのDeepfake検出

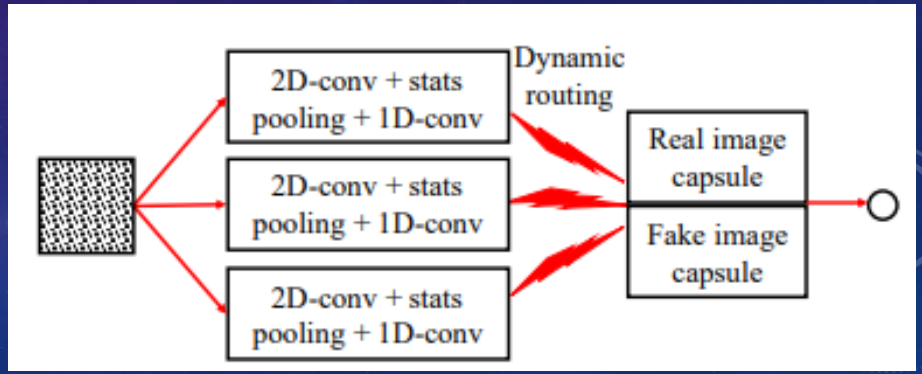
特徴量検出

Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos
(Nguyen et al., *ICASSP*, 2019)

深層学習を用いて、データセットから
本物と合成画像を識別する
有用な特徴量を学習

データにアクセスできれば、
新しいDeepfake手法にも対応可能





Deepfake検出回避

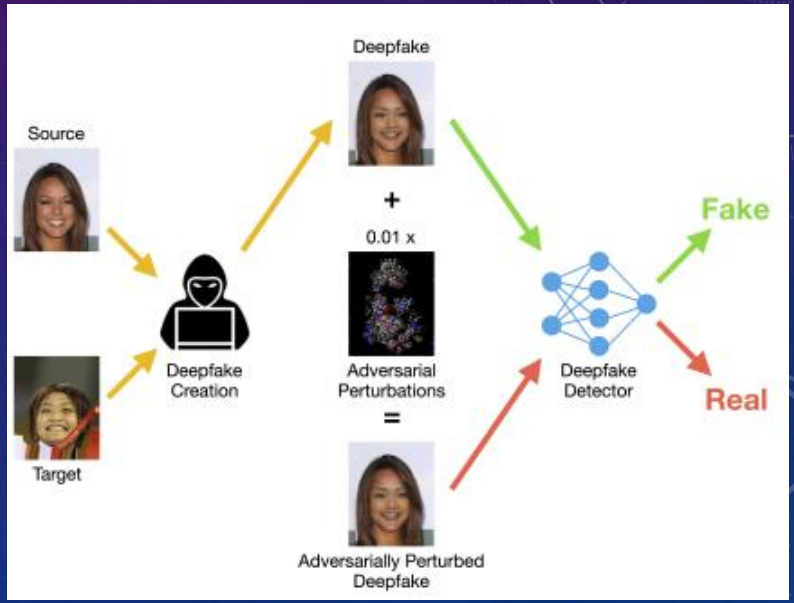
検出回避

Adversarial Perturbations Fool Deepfake Detectors (Gandhi et al., *IJCNN*, 2020)

検出器を誤動作させ間違った判別を
させるような**敵対的サンプル**を生成

Deepfake画像を実画像だと誤判定させるように
敵対的サンプルを最適化

Deepfake検出器に対するリスクを提唱



敵対的サンプル

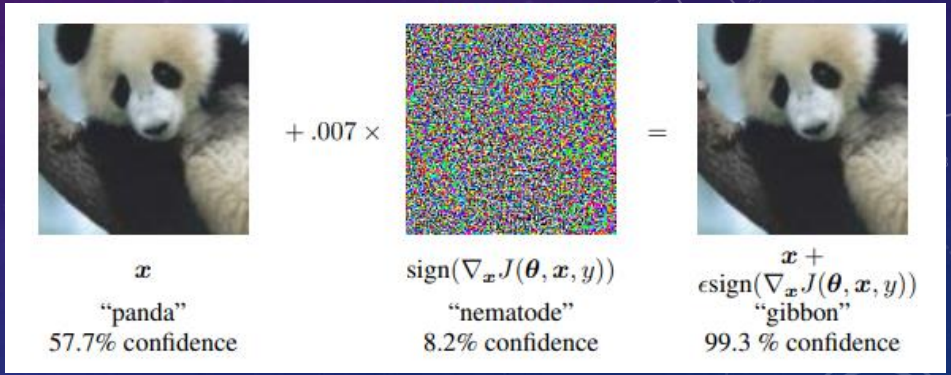
検出回避

敵対的サンプルとは？

人間の目では見えないような
微小なノイズを画像に加える

見た目は変わらないが認識モデルの
誤動作を引き起こす

(右図の例)
本来は“**パンダ**”と認識されるべき画像が、
ノイズを加えることで“**テナガザル**”と誤認識



Explaining and Harnessing Adversarial Examples
(Goodfellow et al., arXiv, 2014)

敵対的サンプルへの対策

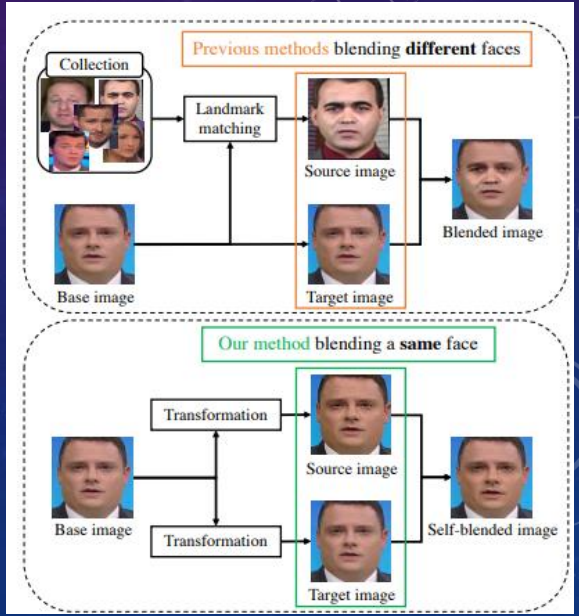
高品質なデータ生成

Detecting Deepfakes With Self-Blended Images (Shiohara et al., *CVPR*, 2022)

検出性能を上げるための**高品質な学習データ**を生成

異なる人物間のDeepfake画像だと**低品質なアーティファクト**が
含まれてしまい、検出性能を低下を引き起こす

同じ人物間のDeepfakeを生成することで、
検出に有用な高品質なアーティファクトのみを
学習させることが可能



Deepfake検出＋分析＋復元

偽造分析

判別だけでなくどのように偽造されたかを分析

(例)

- 偽造領域特定
- 元画像復元

偽造分析

Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos (Nguyen et al., *BTAS*, 2019)

偽造領域特定

どの部分が操作されたかを
Deepfake画像から推定

どのようなDeepfake手法が使用されたかの特定に役立つ

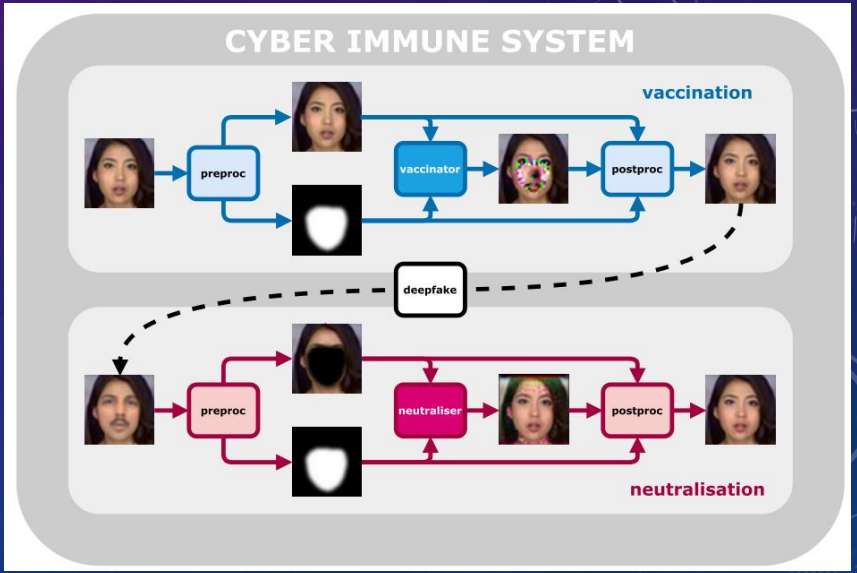
サイバーワクチン2023

偽造分析

Cyber Vaccine for Deepfake Immunity (Chang et al., *IEEE ACCESS*, 2023)

元画像復元

あらかじめ**元画像の情報を埋め込む**
ことで、Deepfake画像から**元画像を復元**



Deepfake生成防止（破壊）

敵対的サンプルを用いた生成防止

Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems (Ruiz et al., *CVPRW*, 2020)

敵対的サンプルを用いてDeepfake生成器を誤動作させる

Deepfake出力が破壊されるように敵対的摂動を最適化

Deepfake生成防止（破壊）

敵対的サンプルを用いた生成防止

TAFIM: Targeted Adversarial Attacks against Face Image Manipulations
（Aneja et al., *ECCV*, 2022）

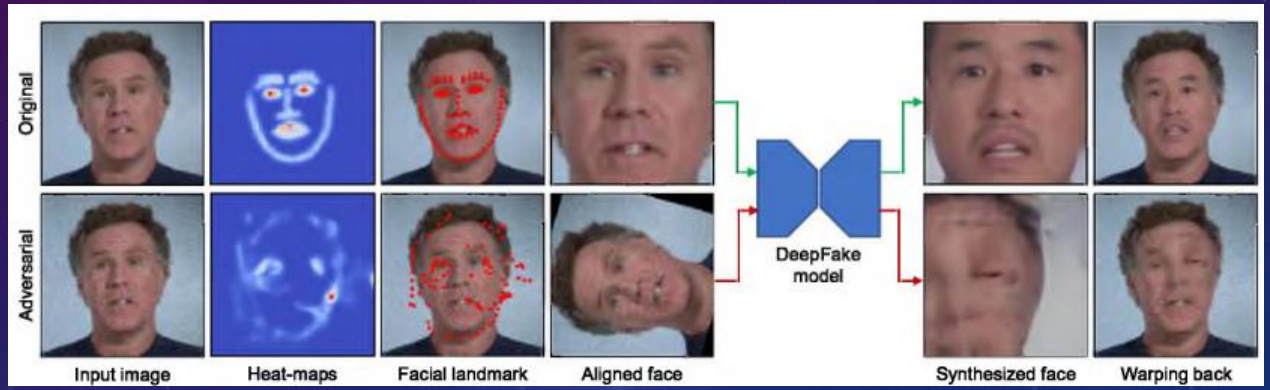
敵対的摂動を出力するモデルを学習

使用されている偽造アルゴリズム（３種）を判別可能
JPEG圧縮をシミュレートすることで、圧縮に対するノイズの頑健性を獲得

Deepfake生成防止

敵対的サンプルを用いた生成防止

Landmark Breaker: Obstructing DeepFake By Disturbing Landmark Extraction
(Sun et al., *WIFS*, 2020)



そもそもDeepfakeするためには**画像内の顔を検出する前処理の必要**がある

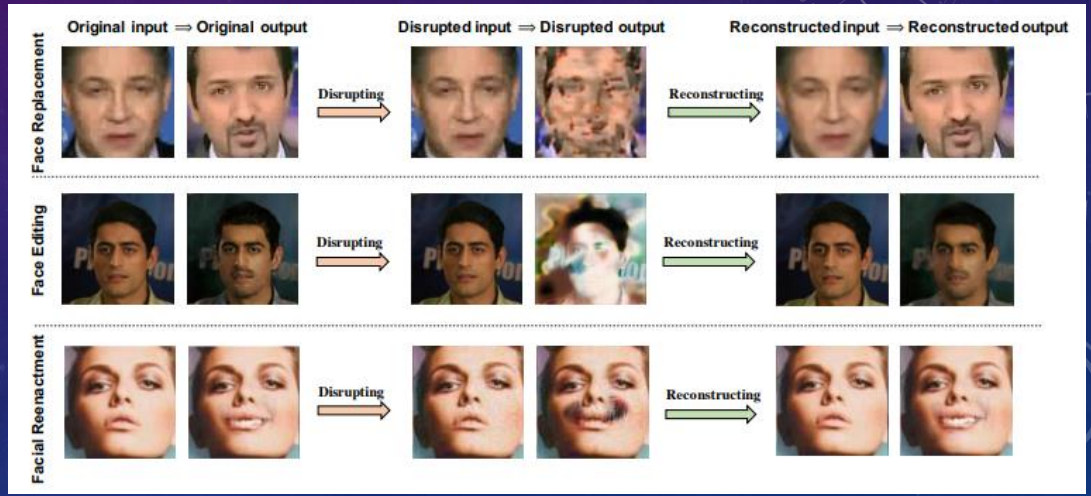
顔検出モデルを誤動作させる敵対的サンプルを生成することでDeepfake生成を防止

敵対的サンプルの無効化

MagDR: Mask-Guided Detection and Reconstruction for Defending Deepfakes
(Chen et al., *CVPR*, 2021)

敵対的サンプルを正常に戻すことで
Deepfake生成防止を無効化

再構成モデルを用いて
敵対的サンプルのノイズを除去



いたちごっこは、しばし留まることを知らず