## Session 3:

# Digital Face Manipulation Detection

Yaojie Liu and Xiaoming Liu

MICHIGAN STATE UNIVERSITY

Computer Vision Lab

IJCB 2020

# Outline

- Introduction of digital attacks

    -- Problem, Facial manipulation types, Challenges

- Benchmark databases
- Face manipulation detection methods

    -- Dynamic methods, Static methods

- Future Direction

# Outline

- **Introduction of digital attacks**

  **-- Problem, Facial manipulation types, Challenges**

- Benchmark databases
- Face manipulation detection methods

  -- Dynamic methods, Static methods

- Future Direction

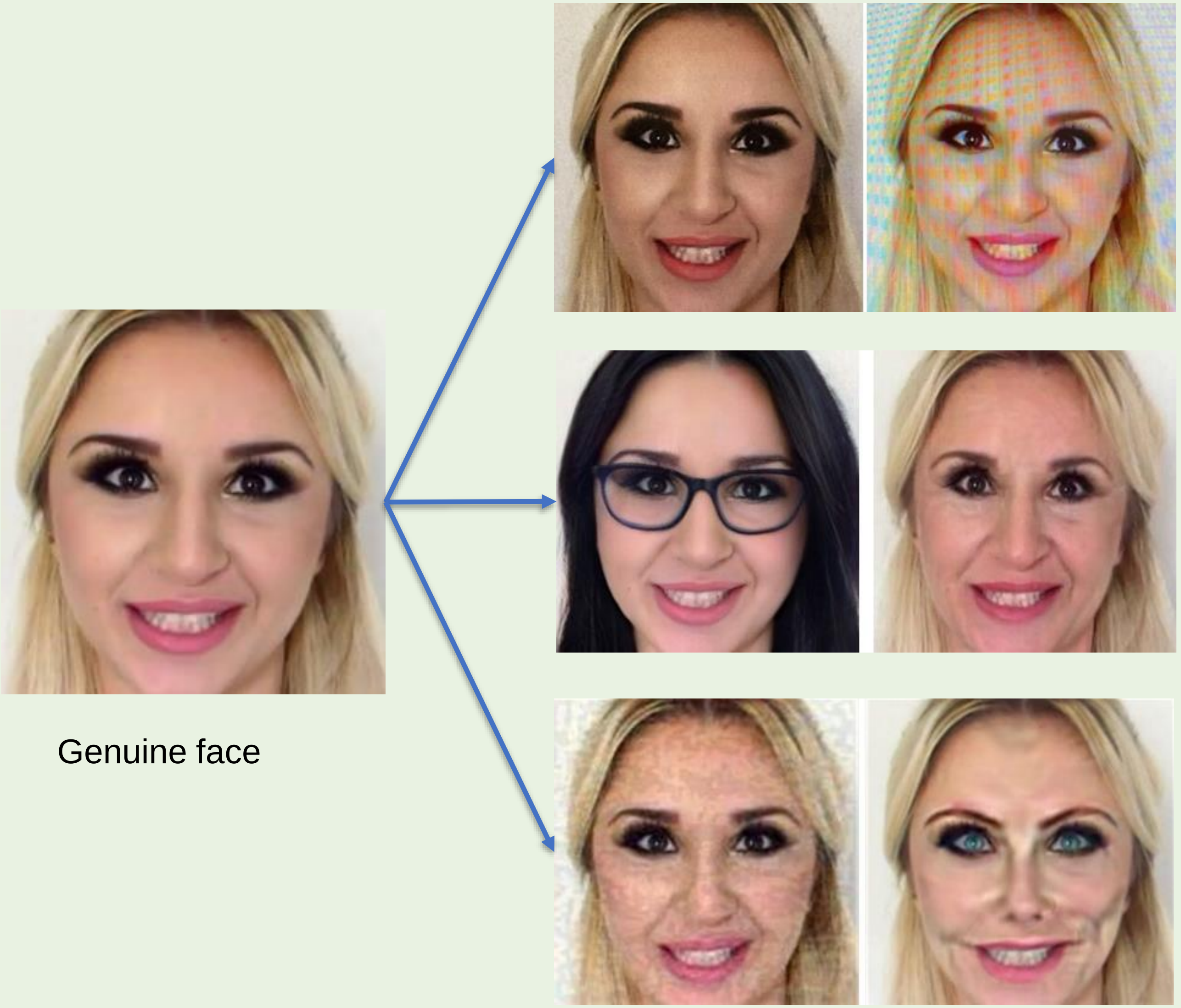
# Problem

- Manipulation of faces has become ubiquitous, and raise concerns especially in social media content.
  -- Advances in deep learning enable a rapid dissemination of “fake news”.


Deepfake (by Facebook)

Fake News (by The Telegraph)

# Face Attacks



Physical spoof attack

Digital manipulation attack

Adversarial attack

Genuine face

# New Software

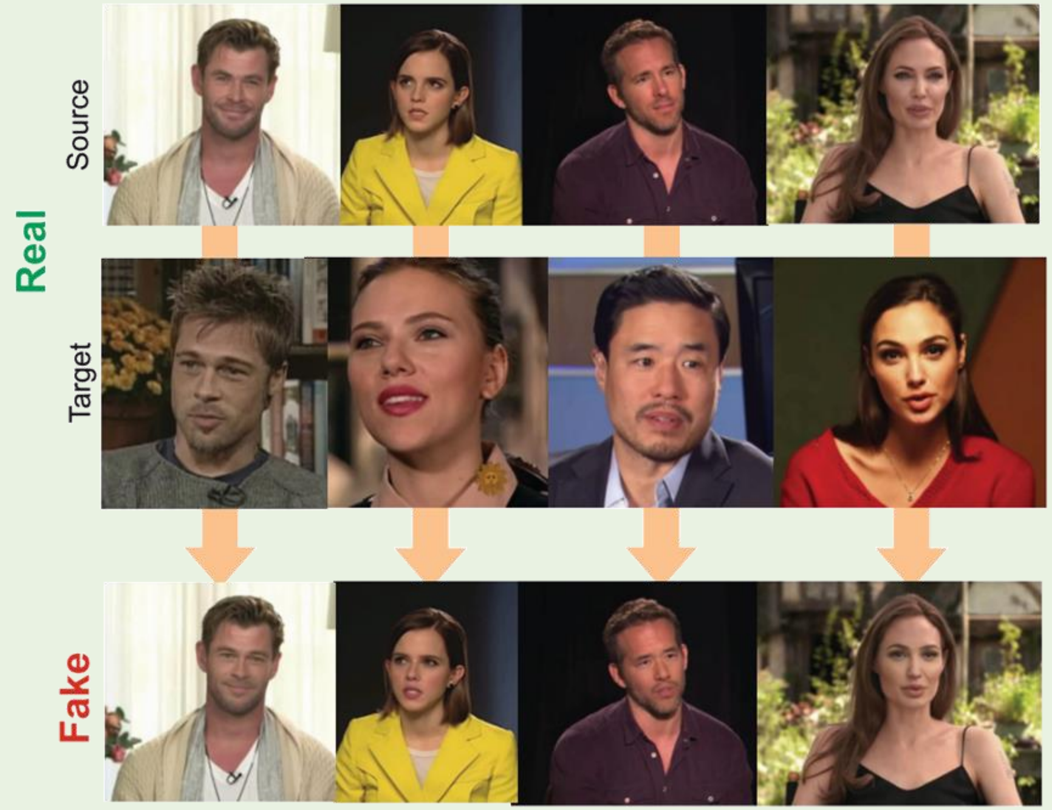
Apps are released to public to create their own fake images and videos, e.g., FaceApp and ZAO.

FaceAPP

ZAO

FaceAPP: https://faceapp.com/app ZAO: https://apps.apple.com/cn/app/zao/id1465199127

# Facial Manipulation Types



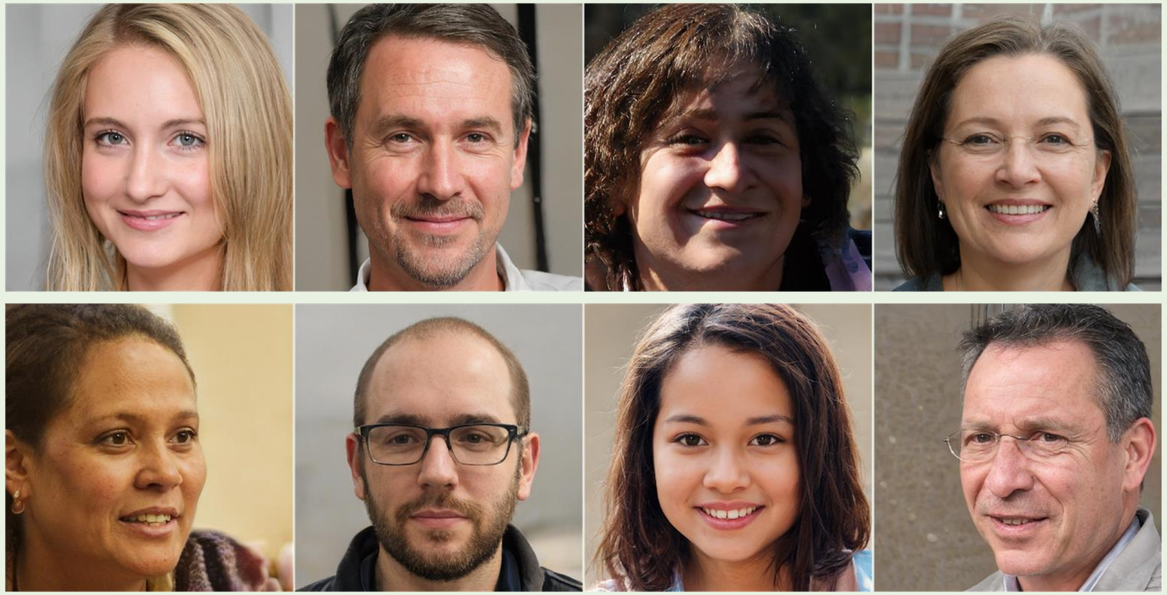

(a) Identity swap


(b) Expression swap

Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

# Facial Manipulation Types


(c) Attribute manipulation

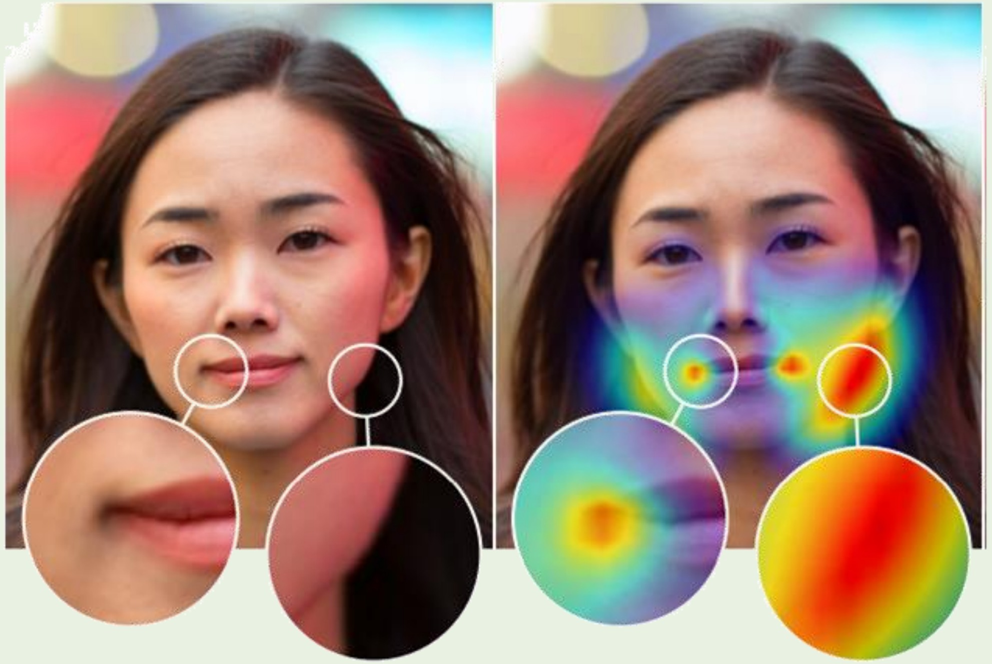


(d) Entire Face synthesis*

* https://www.thispersondoesnotexist.com/

Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

# Facial Manipulation Types



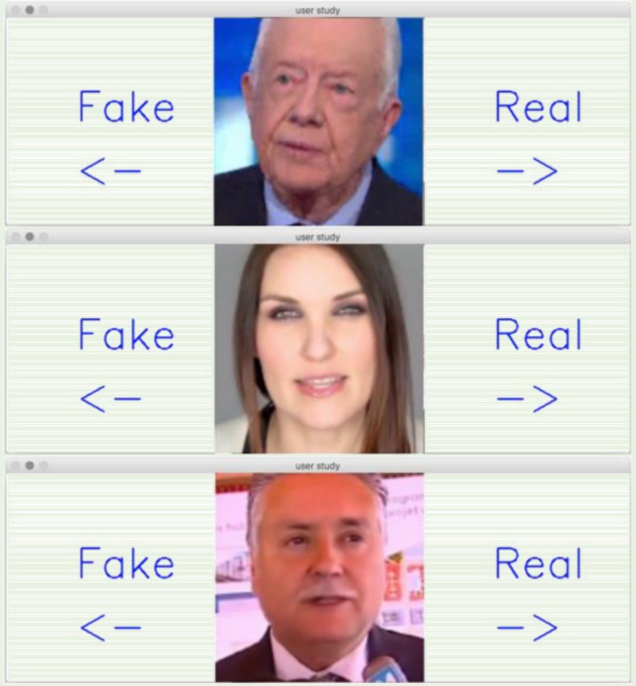
(e) Photoshopped faces

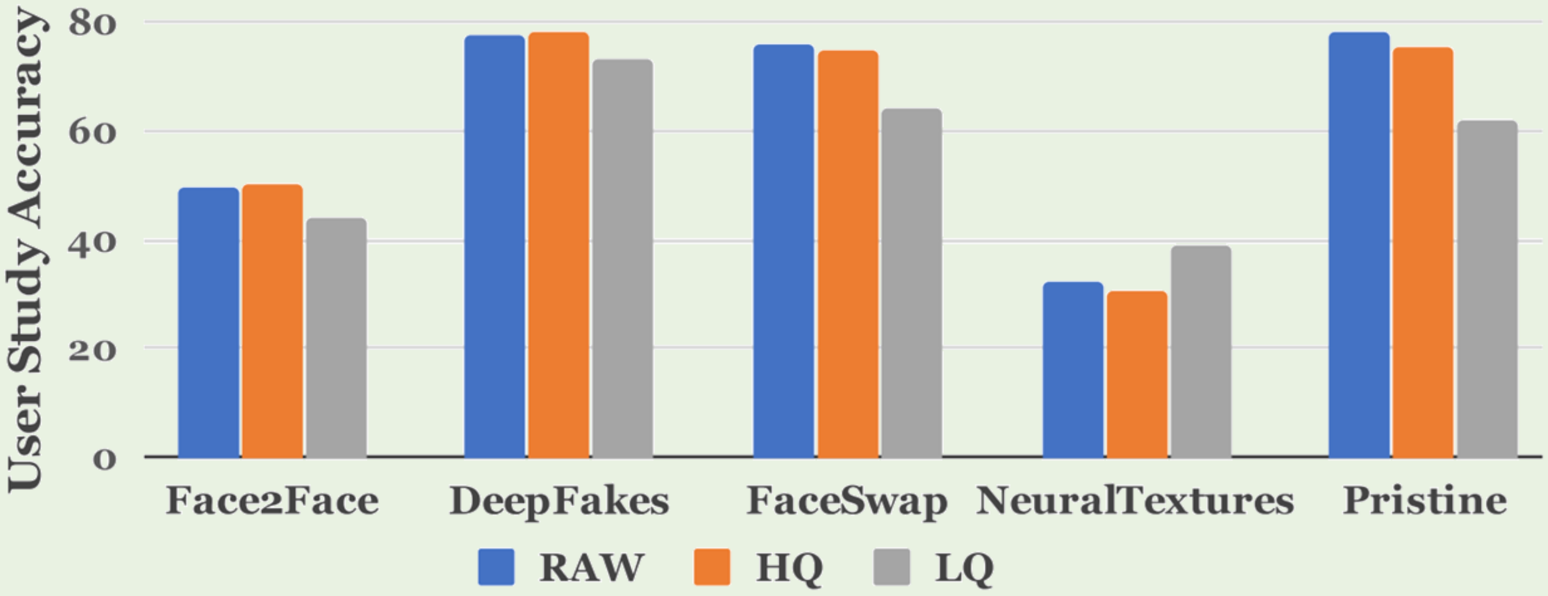

(f) Morphed faces

Wang et al. Detecting Photoshopped Faces by Scripting Photoshop. In ICCV, 2019.
Raja et al. Morphing Attack Detection - Database, Evaluation Platform and Benchmarking. Arxiv, 2020.

# Human Study on Face Forgery Detection

- 204 participants.
- On different video qualities.



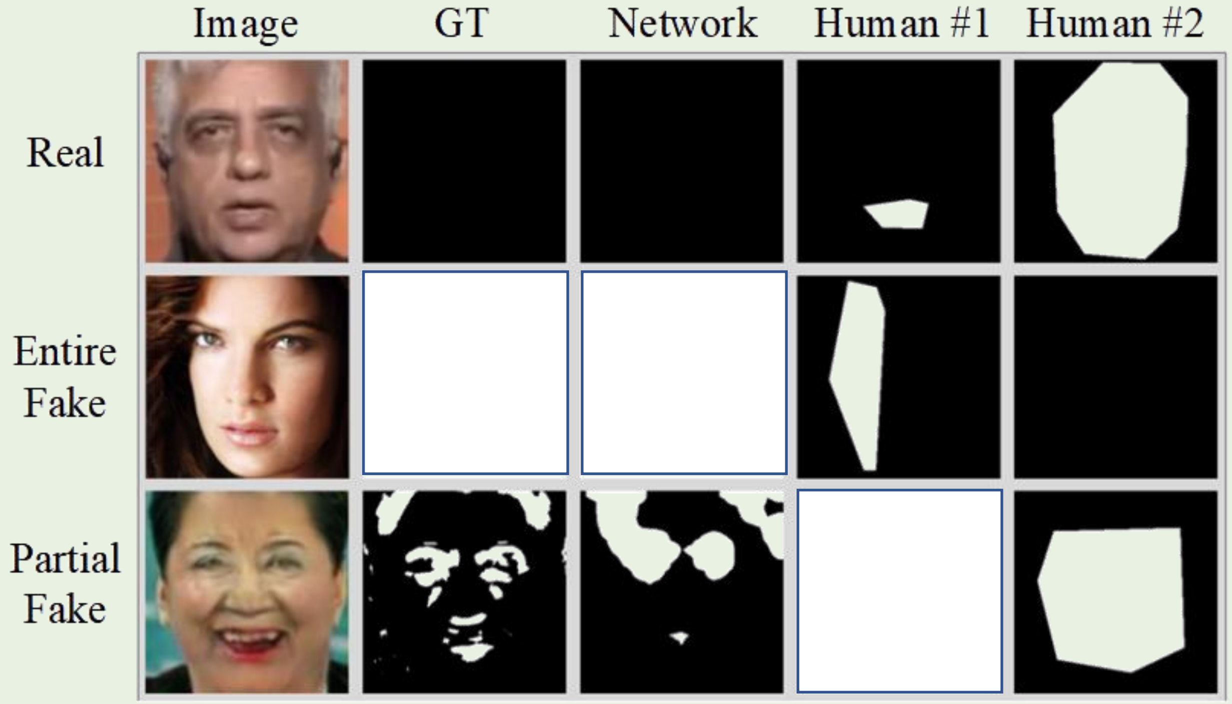




Andreas et al. Faceforensics++: Learning to detect manipulated facial images. In ICCV, 2019.

# Human Study on Face Forgery Detection

- 10 participants.
- Forgery detection and localization of the manipulated regions.

| | Human | Network |
|---|---|---|
| ACC | 68.18% | 97.27% |
| AUC | 81.71% | 99.29% |
| EER | 30.00% | 3.75% |
| TDR (0.01%) | 42.50% | 85.00% |
| Localization Accuracy | 58.20% | 90.93% |




Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

# Challenges

- The lack of **diverse** training data is a bottleneck for training deep networks for manipulation detection.
- Most works are trained for **known** face manipulation techniques. How to capture more intrinsic forgery evidence to improve the **generalizability?**
- Less attention has been paid to the identification of manipulated faces in video by taking advantage of the **temporal** information.
- Besides manipulation detection, there are few methods focusing on **localizing** the manipulated region.

# Outline

- Introduction of digital attacks

  -- Problem, Facial manipulation types, Challenges

- **Benchmark databases**
- Face manipulation detection methods

  -- Dynamic methods, Static methods

- Future Direction

# Databases

| Database | Number of real samples (videos) | Number of fake samples (videos) | Fake generation method | Source of real data | Year |
|---|---|---|---|---|---|
| UADFV | 49 | 49 | FaceSwap | Youtube | 2019 |
| FaceForensics++ | 1,000 | 6,000 | FaceSwap, Face2Face, Neural textures, Deepfakes | Youtube, actors | 2019 |
| Deepfake Detection Challenge (DFDC) | 19,154 | 100,000 | FaceSwap, autoencoder, GAN, Neural talking heads | Actors | 2019 |
| Deepfake TIMIT | 430 | 640 | FaceSwap GAN | VidTIMIT | 2019 |
| Diverse Fake Face Dataset (DFFD) | 58,703 images | 240,336 images | FaceSwap, Deepfake, GANs | FFHQ, CelebA | 2019 |
| Celeb-DF | 890 | 5,639 | Deepfake | Youtube | 2020 |
| Deepforensics 1.0 | 50,000 | 10,000 | FaceSwap | Actors | 2020 |
| Deep Fakes Dataset (http://cs.binghamton.edu/~ncilsal2/DeepFakesDataset/) | 142 | | Deepfake | various sources | 2020 |

# Outline

- Introduction of digital attacks

  -- Problem, Facial manipulation types, Challenges

- Benchmark databases

- **Face manipulation detection methods**

  -- **Dynamic methods**, Static methods

- Future Direction

# Dynamic Methods

- Inconsistent motion (head or lip movement detection, optical flow)
  - Exposing deep fakes using inconsistent head poses
  - Speaker inconsistency detection in tampered video
  - Deepfake video detection through optical flow-based CNN

- Feature aggregation
  - Deepfake video detection using recurrent neural networks
  - Recurrent strategies for face manipulation detection in videos
  - Deepfake detection with automatic face weighting

# Dynamic Methods ---- Inconsistent Motion

Pro:

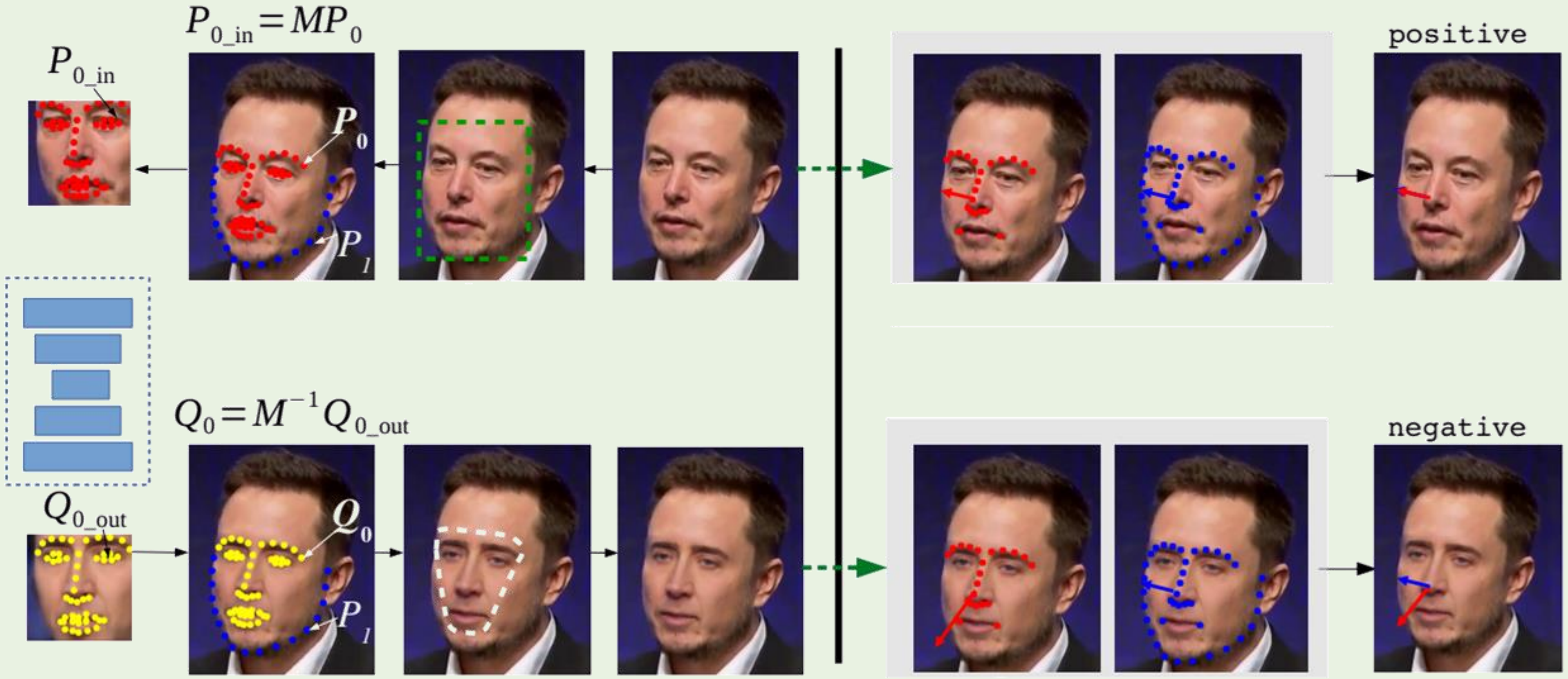
- Tracing the inconsistent motion (e.g., eye, lips and head) makes the detection explainable.

Con:

- May fail when dealing with extremely realistic synthetic images and videos.

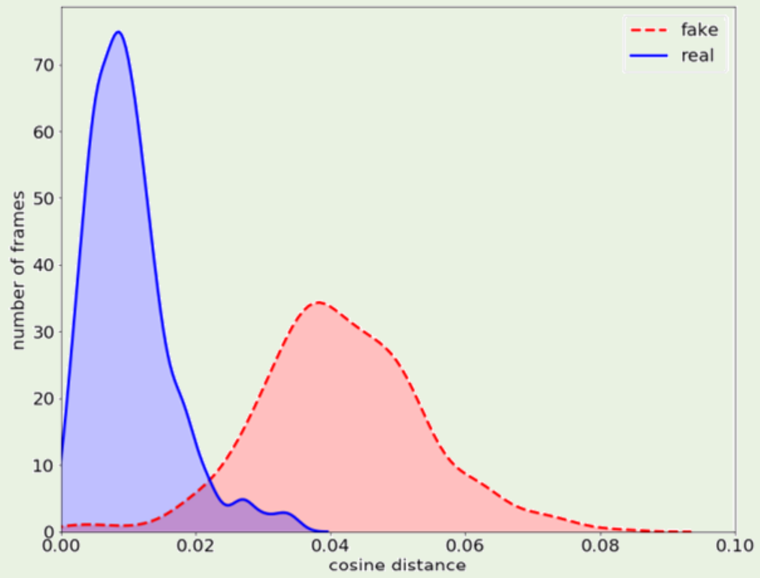
# Exposing Deep Fakes Using Inconsistent Head Poses

- Splicing synthetic face regions in Deepfake introduce errors, which can be revealed when 3D head poses are estimated.
- One SVM classifier is developed based on this inconsistent cue.

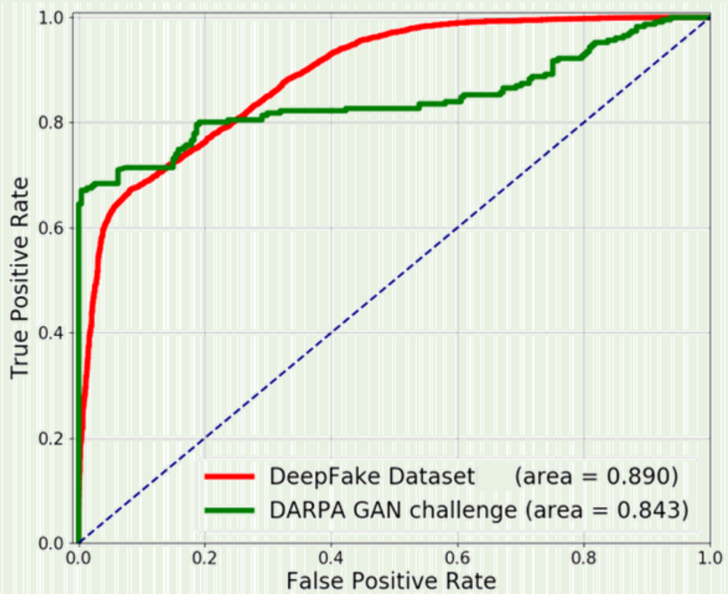



Xin et al. Exposing deep fakes using inconsistent head poses. In ICASSP, 2019.

# Exposing Deep Fakes Using Inconsistent Head Poses




Distribution of the cosine distance between head orientation vectors for fake and real face images.

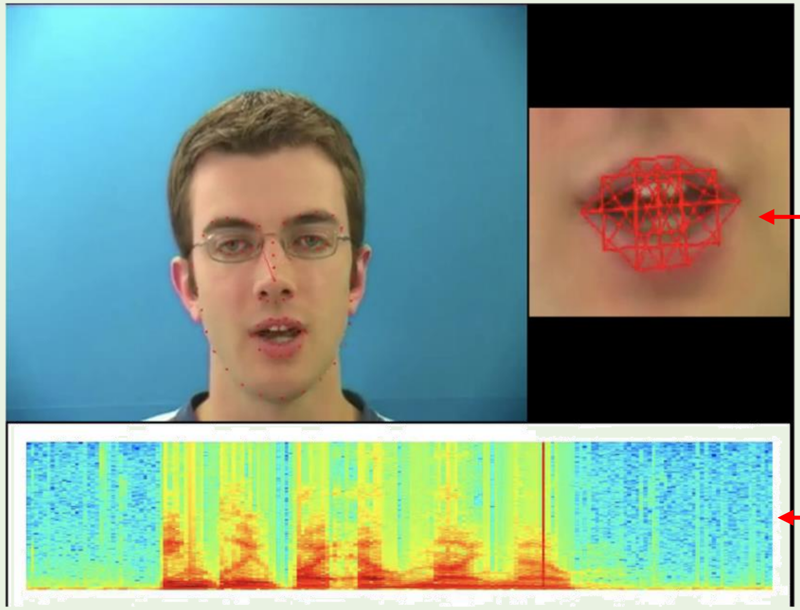



ROC of the SVM classification on DeepFake and DARPA datasets.

Xin et al. Exposing deep fakes using inconsistent head poses. In ICASSP, 2019.

# Speaker Inconsistency Detection in Tampered Video

- Audio-visual tampering in a video of talking person.
- Combining mel-frequency cepstral coefficients (audio features) and distances between mouth landmarks (visual features) for detecting the inconsistencies between video and audio tracks.



detected visual features

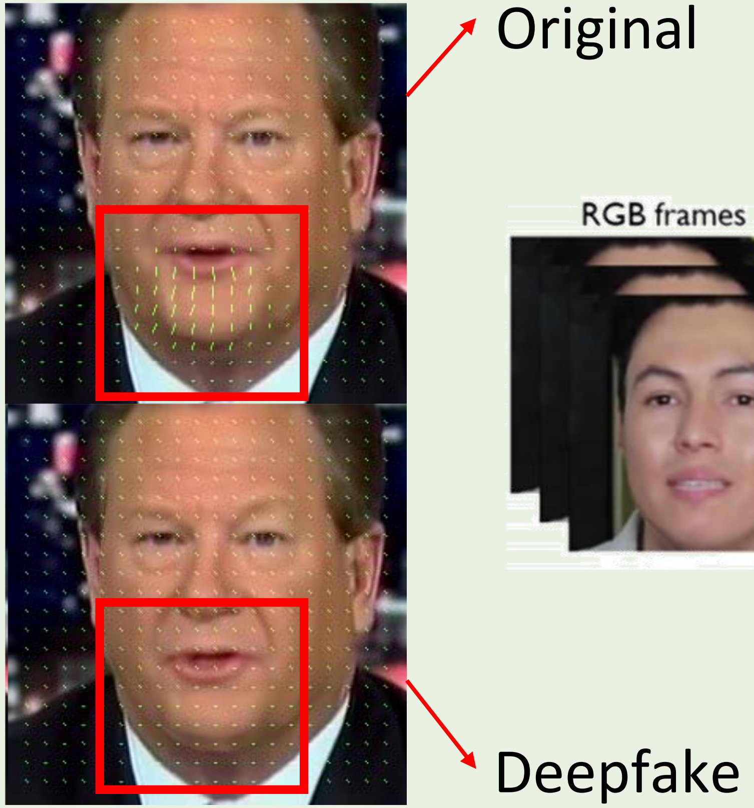
detected MFCC

| Database | Type of data | Train | Test | Total |
|---|---|---|---|---|
| VidTIMIT | subjects | 22 | 21 | 43 |
| | time (hours) | 0.25 | 0.26 | 0.51 |
| | genuine | 220 | 210 | 430 |
| | tampered | 2 | 995 | 2,033 |
| AMI | subjects | 42 | 36 | 54 |
| | time (hours) | 3.82 | 2.28 | 6.1 |
| | genuine | 613 | 364 | 977 |
| | tampered | 2,732 | 1,934 | 4,666 |
| GRID | subjects | 17 | 16 | 33 |
| | time (hours) | 14.01 | 13.19 | 27.2 |
| | genuine | 17,000 | 15,890 | 32,891 |
| | tampered | 79,479 | 75,646 | 155,125 |

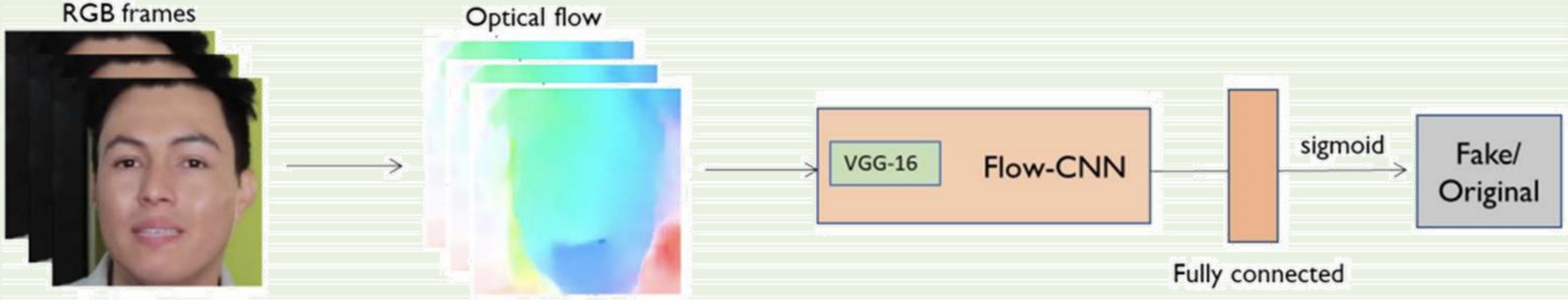
Korshunov et al. Speaker inconsistency detection in tampered video. In EUSIPCO, 2018.

# Deepfake Video Detection via Optical Flow based CNN

- Using inter-frame dissimilarities as clue for forgery detection.




Architecture




Irene et al. Deepfake video detection through optical flow based CNN. In ICCVW, 2019.

# Dynamic Methods

- Inconsistent motion (head or lip movement detection, optical flow)
  - Exposing deep fakes using inconsistent head poses
  - Speaker inconsistency detection in tampered video
  - Deepfake video detection through optical flow-based CNN

- Feature aggregation
  - Deepfake video detection using recurrent neural networks
  - Recurrent strategies for face manipulation detection in videos
  - Deepfake detection with automatic face weighting

# Dynamic Methods ---- Feature Aggregation

- Use CNN to extract frame-level features
- Use RNN to check the consistency among all frame-level features

Pro:

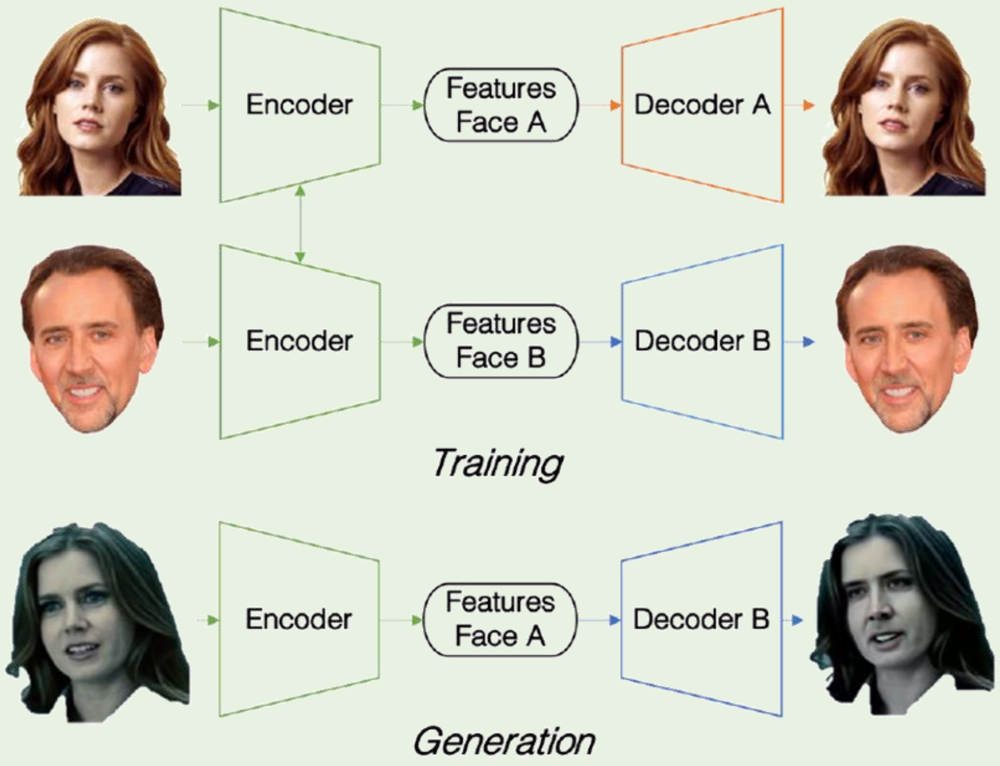
- Spatial-aware and temporal-aware

Con:

- Fake feature can be immersed during long aggregation

# Deepfake Video Detection Using Recurrent Neural Networks

- What makes deepfakes possible is finding a way to force both latent faces to be encoded in the same space.

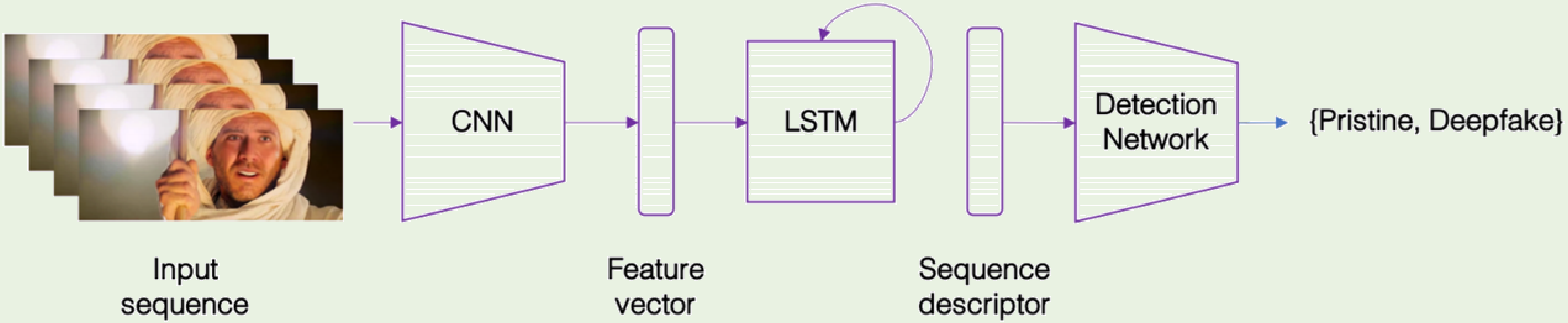



- When we want to do a new faceswapp, we encode the input face and decode it using the target face decoder.

David et al. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

# Deepfake Video Detection Using Recurrent Neural Networks

- CNN obtains a set of features for each frame.
- Concatenate the features of consecutive frames and pass them to LSTM for analysis.




David et al. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

# Deepfake Video Detection Using Recurrent Neural Networks

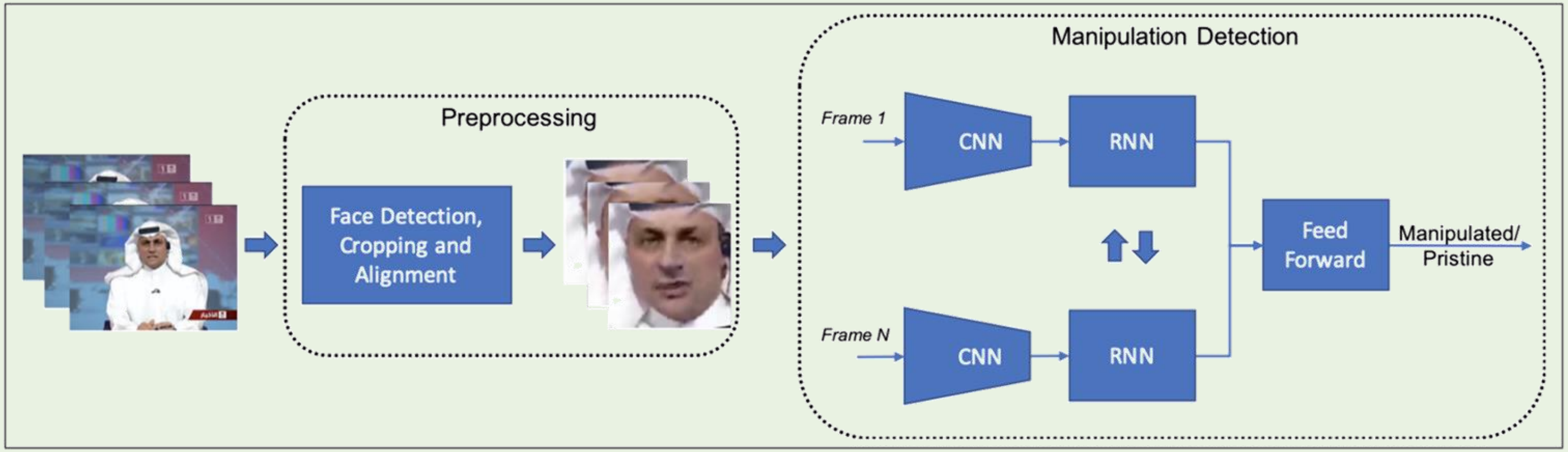
- Deepfake manipulation detection results

| Model | Training acc. (%) | Validation acc. (%) | Test acc. (%) |
|---|---|---|---|
| Conv-LSTM, 20 frames | 99.5 | 96.9 | 96.7 |
| Conv-LSTM, 40 frames | 99.3 | 97.1 | 97.1 |
| Conv-LSTM, 80 frames | 99.7 | 97.2 | 97.1 |

David et al. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

# Recurrent Strategies for Face Manipulation Detection in Videos

- **Temporal discrepancies** are expected to occur in images, since manipulations are performed on a frame-by-frame basis.
- Low-level artifacts caused by manipulations on faces are expected to further manifest themselves as temporal artifacts with inconsistent features across frames.




Sabir et al. Recurrent convolutional strategies for face manipulation detection in videos. In CVPRW, 2019.

# Recurrent Strategies for Face Manipulation Detection in Videos

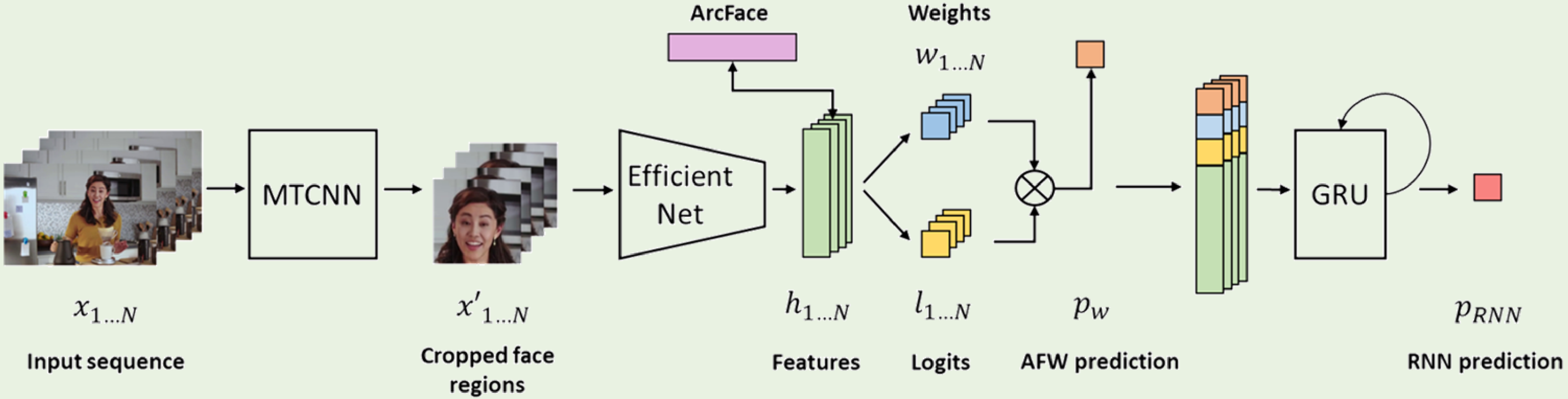
Accuracy for manipulation detection on all manipulation types. The FF++ is the baseline and DenseNet with alignment and bidirectional recurrent network performs the best.

| Manipulation | Frames | FF++ | ResNet50 | DenseNet | ResNet50 + Alignment | DenseNet + Alignment | ResNet50 + Alignment + BiDir | DenseNet + Alignment + BiDir |
|---|---|---|---|---|---|---|---|---|
| Deepfake | 1 | 93.46 | 94.8 | 94.5 | 96.1 | 96.4 | - | - |
| | 5 | - | 94.6 | 94.7 | 96.0 | 96.7 | 94.9 | **96.9** |
| Face2Face | 1 | 89.8 | 90.25 | 90.65 | 89.31 | 87.18 | - | - |
| | 5 | - | 90.25 | 89.8 | 92.4 | 93.21 | 93.05 | **94.35** |
| FaceSwap | 1 | 92.72 | 91.34 | 91.04 | 93.85 | 96.1 | - | - |
| | 5 | - | 90.95 | 93.11 | 95.07 | 95.8 | 95.4 | **96.3** |

Sabir et al. Recurrent convolutional strategies for face manipulation detection in videos. In CVPRW, 2019.

# Deepfakes Detection with Automatic Face Weighting

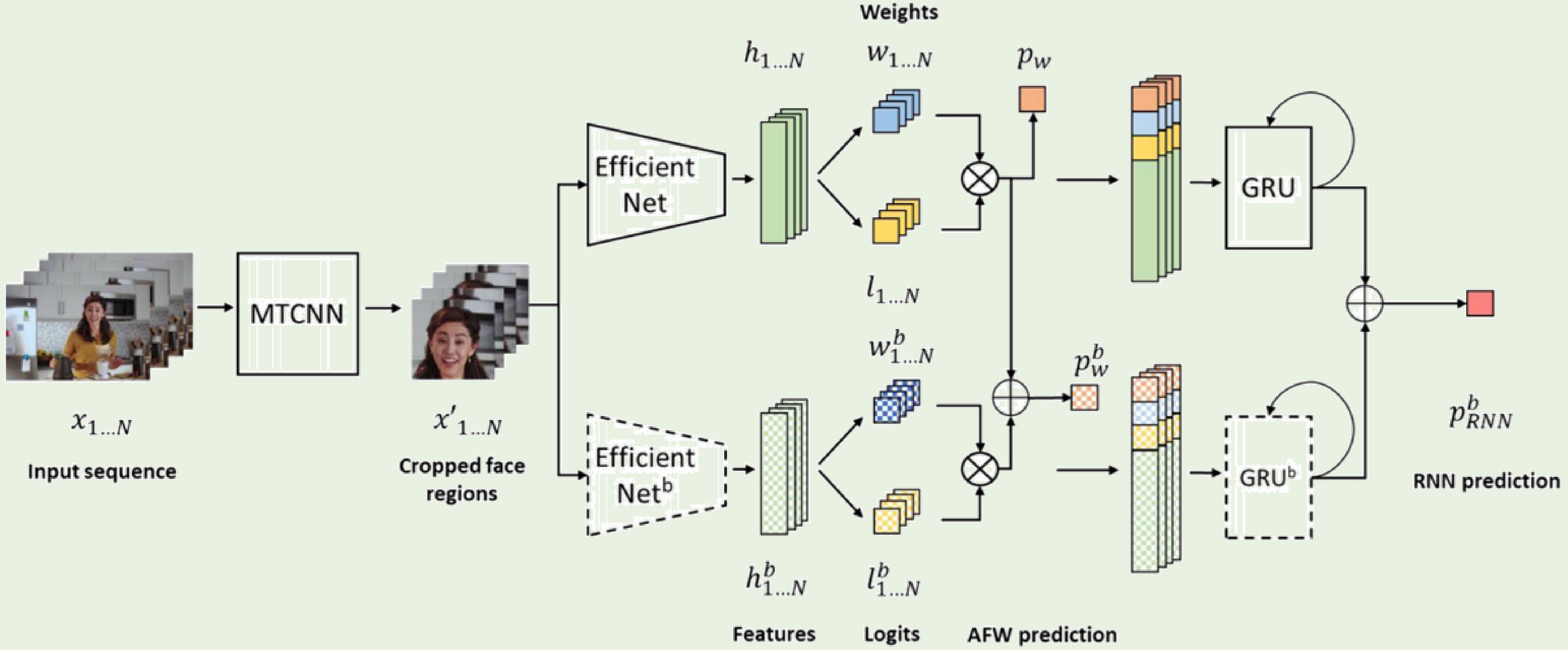
The network automatically selects the most reliable frames to detect these manipulations with a weighting mechanism combined with a Gated Recurrent Unit that provides a probability of a video being real or fake.




Montserrat et al. Deepfakes detection with automatic face weighting. In CVPRW, 2020.

# Deepfakes Detection with Automatic Face Weighting

Add a boosting network for more robust predictions.




Montserrat et al. Deepfakes detection with automatic face weighting. In CVPRW, 2020.

# Deepfakes Detection with Automatic Face Weighting

- Accuracy of the presented method and previous works.

| Method | Validation | Test |
|---|---|---|
| **Conv-LSTM** | 66.05% | 70.78% |
| **EfficientNet-b5** | 79.25% | 80.62% |
| **Xception** | 78.42% | 80.14% |
| **Ours** | 92.61% | 91.88% |

- The log-likelihood error of our method with and without boosting network and test augmentation.

| Method | Log-likelihood |
|---|---|
| **Baseline** | 0.364 |
| **+ Boosting Network** | 0.341 |
| **+ Test Augmentation** | 0.321 |

Montserrat et al. Deepfakes detection with automatic face weighting. In CVPRW, 2020.

# Outline

- Introduction of digital attacks

  -- Problem, Facial manipulation types, Challenges
- Benchmark databases
- **Face manipulation detection methods**

  -- Dynamic methods, **Static methods**
- Future Direction

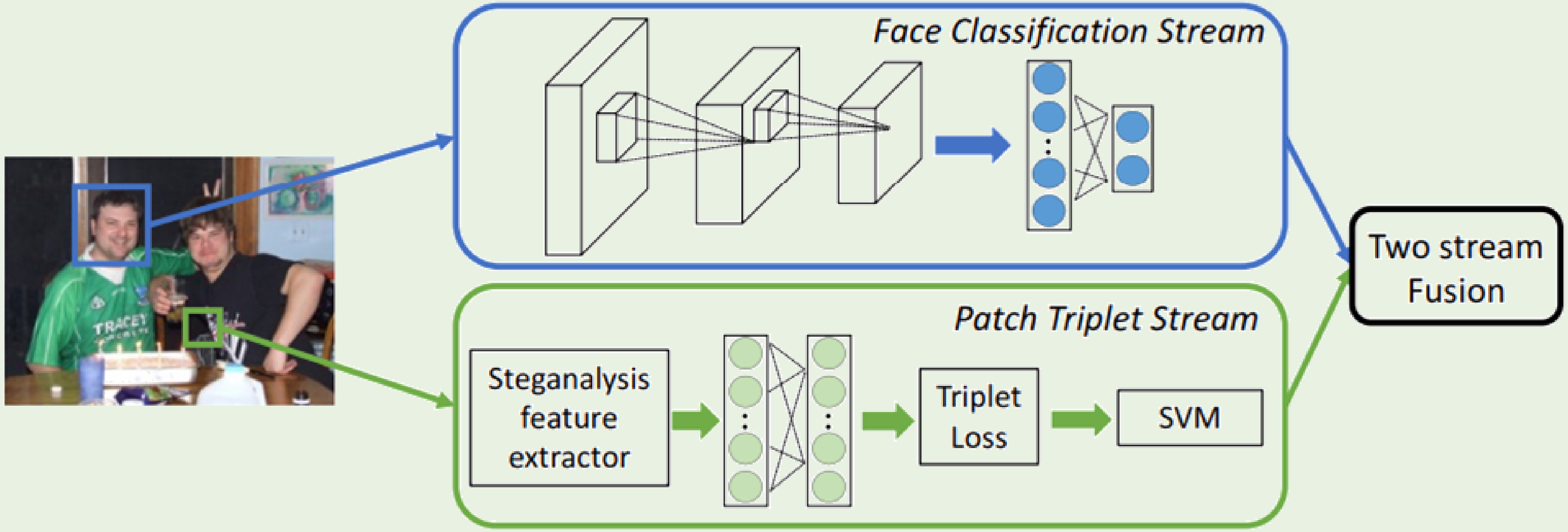
# Static Methods

- CNN binary classification only
  - Two-stream neural networks for tampered face detection
  - Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

- Joint binary classification and manipulated region localization
  - Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos (segmentation)
  - Face X-ray for more general face forgery detection (face X-ray)
  - On the Detection of Digital Face Manipulation (attention)

# Two-Stream Neural Networks for Tampered Face Detection

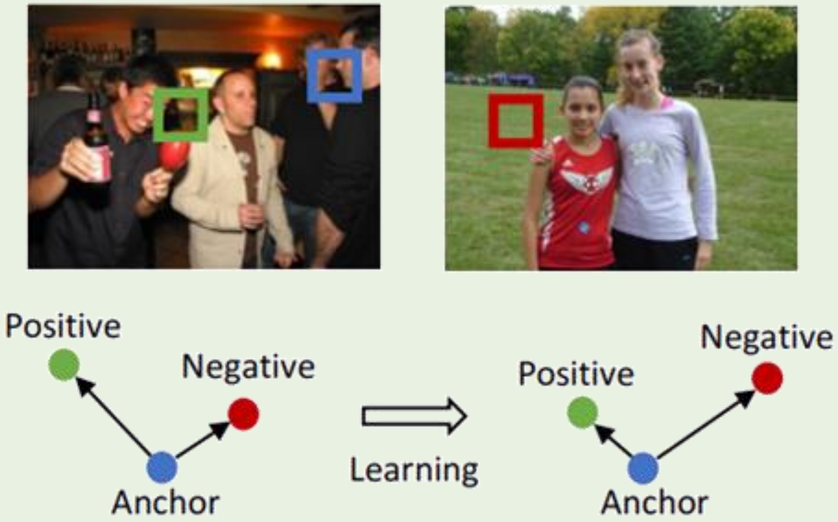
- The classification stream is trained on tampered and authentic images and serves as a tampered face classifier.
- The patch triplet stream captures low-level camera characteristics and local noise residuals.

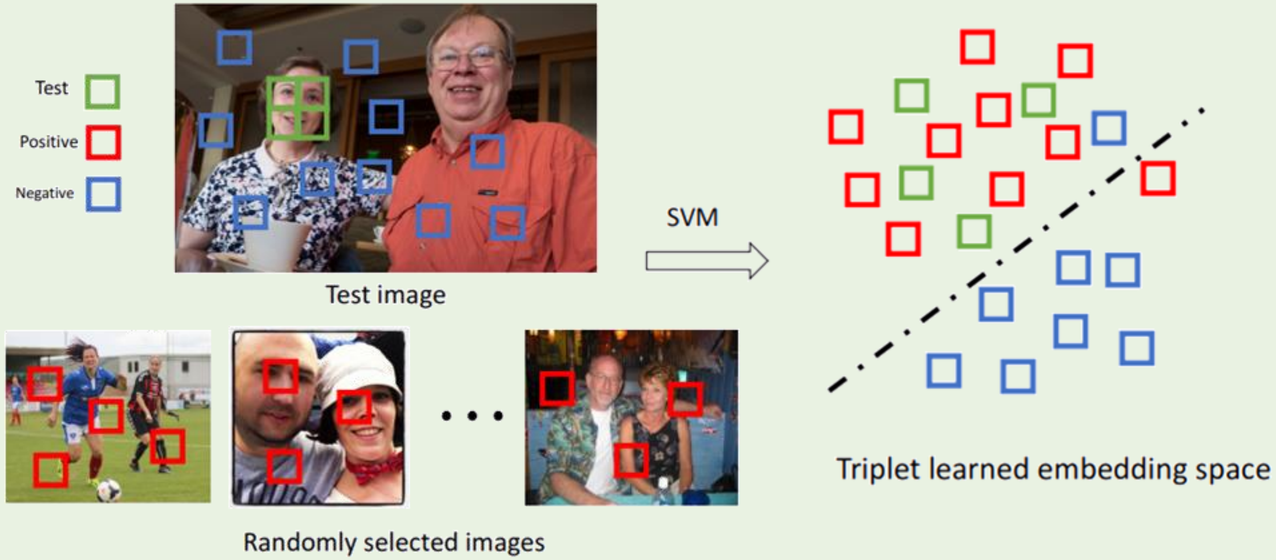



Zhou et al. Two-stream neural networks for tampered face detection. In CVPRW, 2017.

# Two-Stream Neural Networks for Tampered Face Detection

- The triplet network is designed to determine whether two patches come from the same image.
- Leveraging clues hidden in the in-camera processing for tampered face detection.

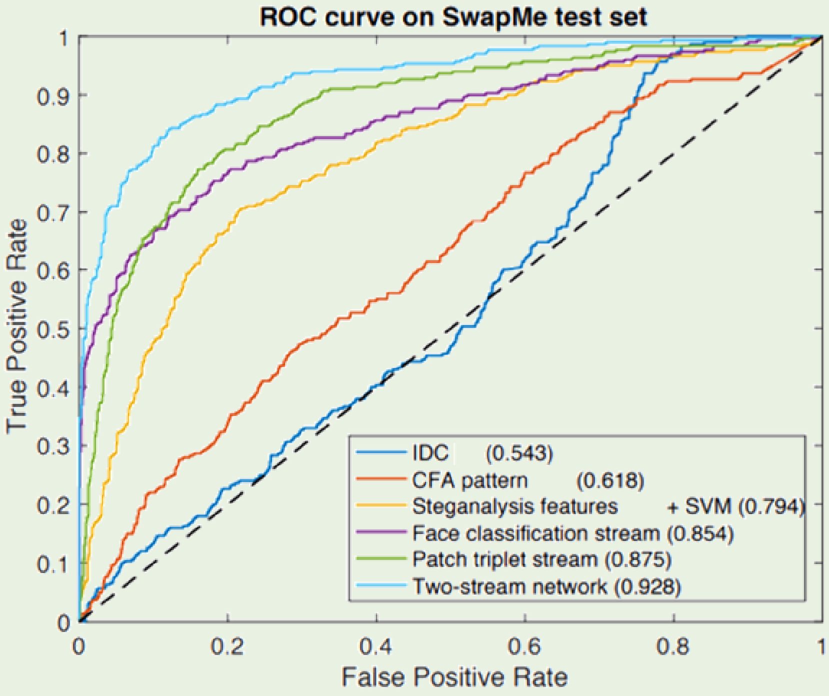



Zhou et al. Two-stream neural networks for tampered face detection. In CVPRW, 2017.

# Two-Stream Neural Networks for Tampered Face Detection

- ROC comparison between two-stream network and baselines.
- AUC for different methods.

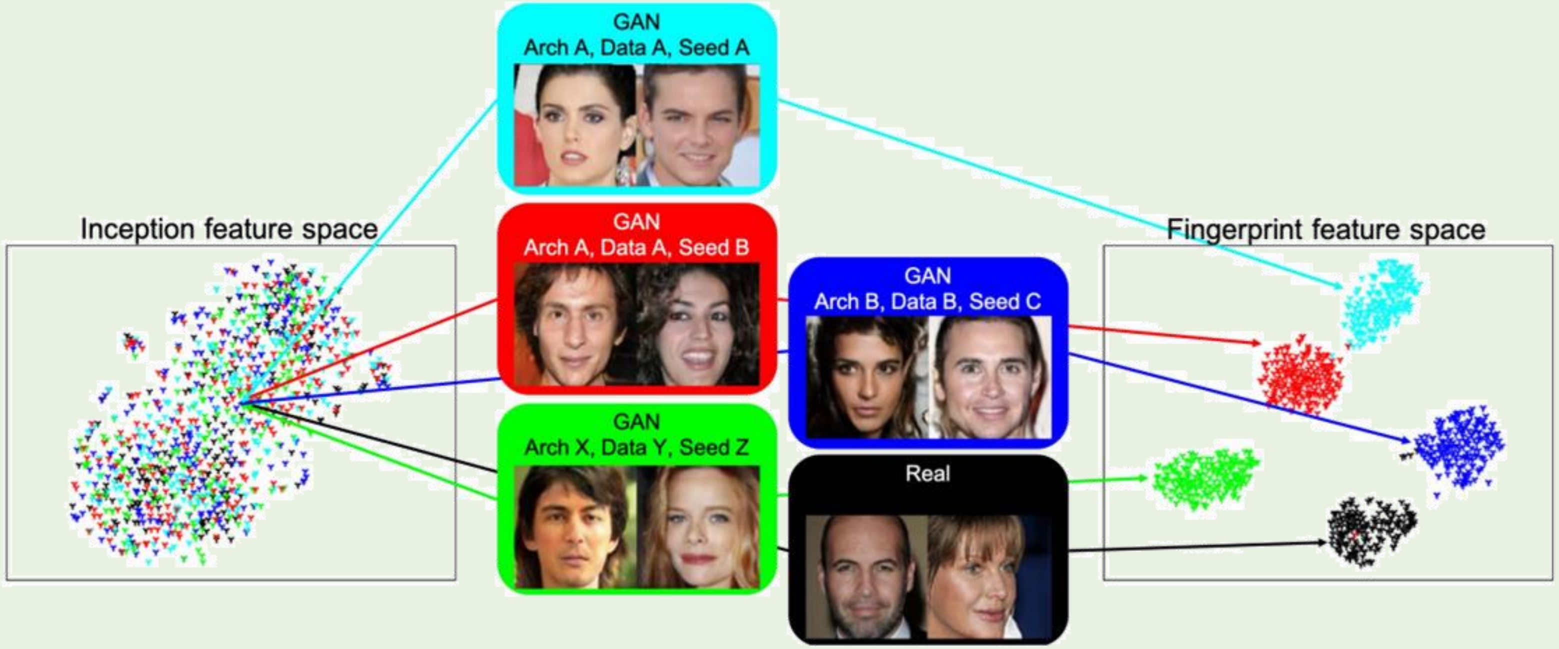



| Methods | AUC |
|---|---|
| IDC | 0.543 |
| CFA Pattern | 0.618 |
| Steganalysis features+SVM | 0.794 |
| Face classification stream | 0.854 |
| Patch triplet stream | 0.875 |
| Two-stream network | **0.927** |

Zhou et al. Two-stream neural networks for tampered face detection. In CVPRW, 2017.

# Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

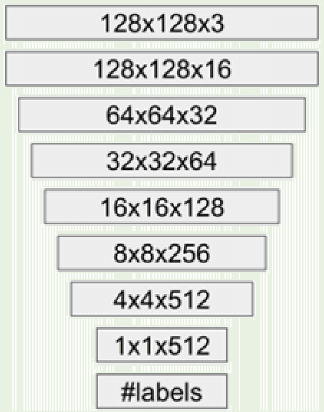
- Learning GAN fingerprints towards image attribution and using them to classify an image as real or GAN-generated.
- For GAN-generated images, we further identify their sources.

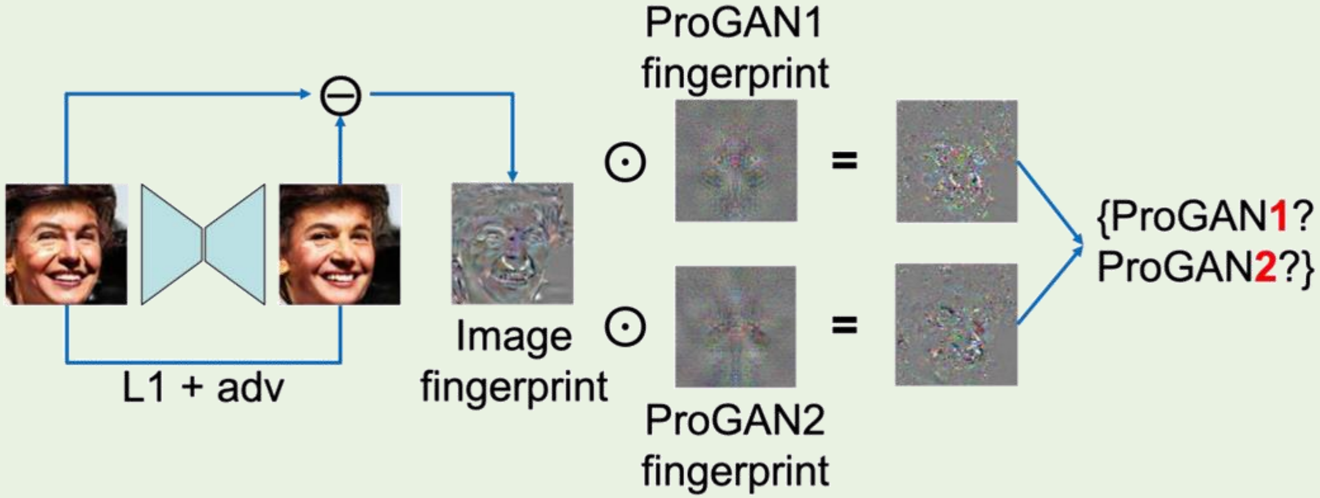



Yu et al. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In ICCV, 2019.

# Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

- We train an attribution classifier that can predict the source of an image: real or from a GAN model.
- We implicitly represent image fingerprints as the final classifier features and represent GAN model fingerprints as the corresponding classifier parameters.
- Fingerprint visualization module by an AutoEncoder reconstruction network.



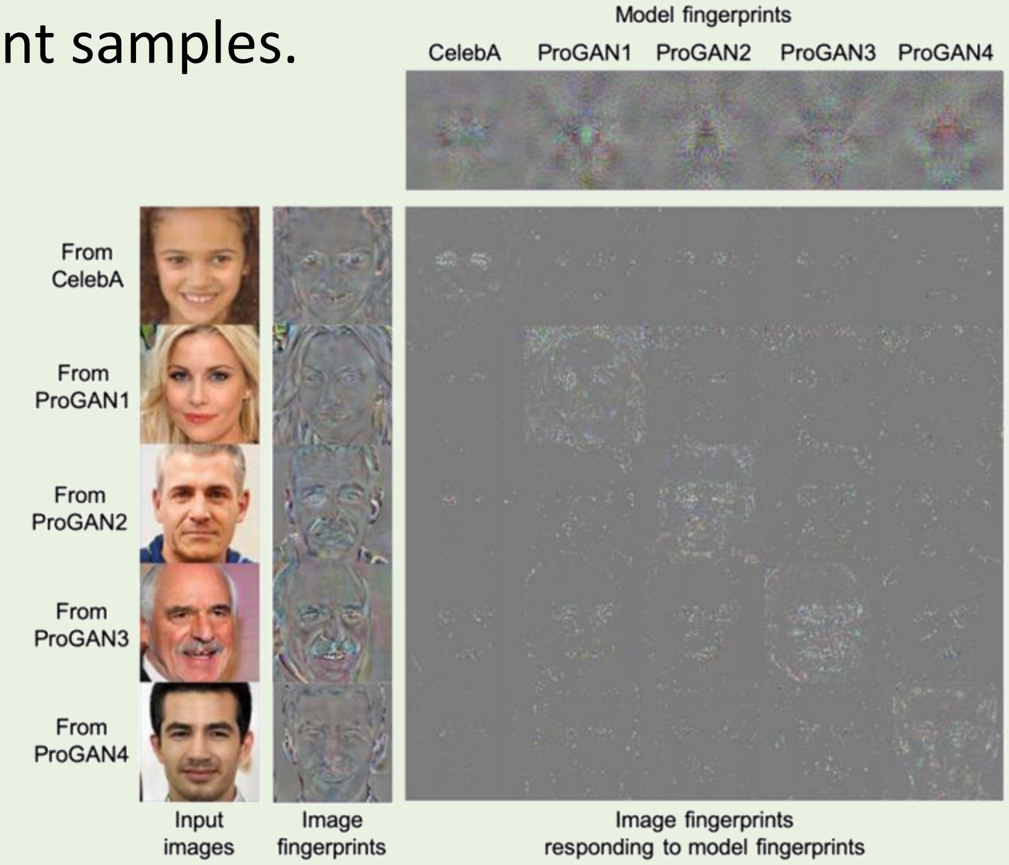




Yu et al. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In ICCV, 2019.

# Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

- Classification results on real, ProGAN, SNGAN, GramerGAN, and MMDGAN data.
- Visualization of model and image fingerprint samples.

| Metric | Method | CalebA | LSUN |
|---|---|---|---|
| Accuracy | kNN | 28.00 | 36.30 |
| | Eigenface | 53.28 | - |
| | PRNU | 86.61 | 67.84 |
| | Fingerprint | **99.43** | **98.58** |
| FD ratio | Inception | 2.36 | 5.27 |
| | Fingerprint | **454.76** | **226.59** |




Yu et al. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In ICCV, 2019.

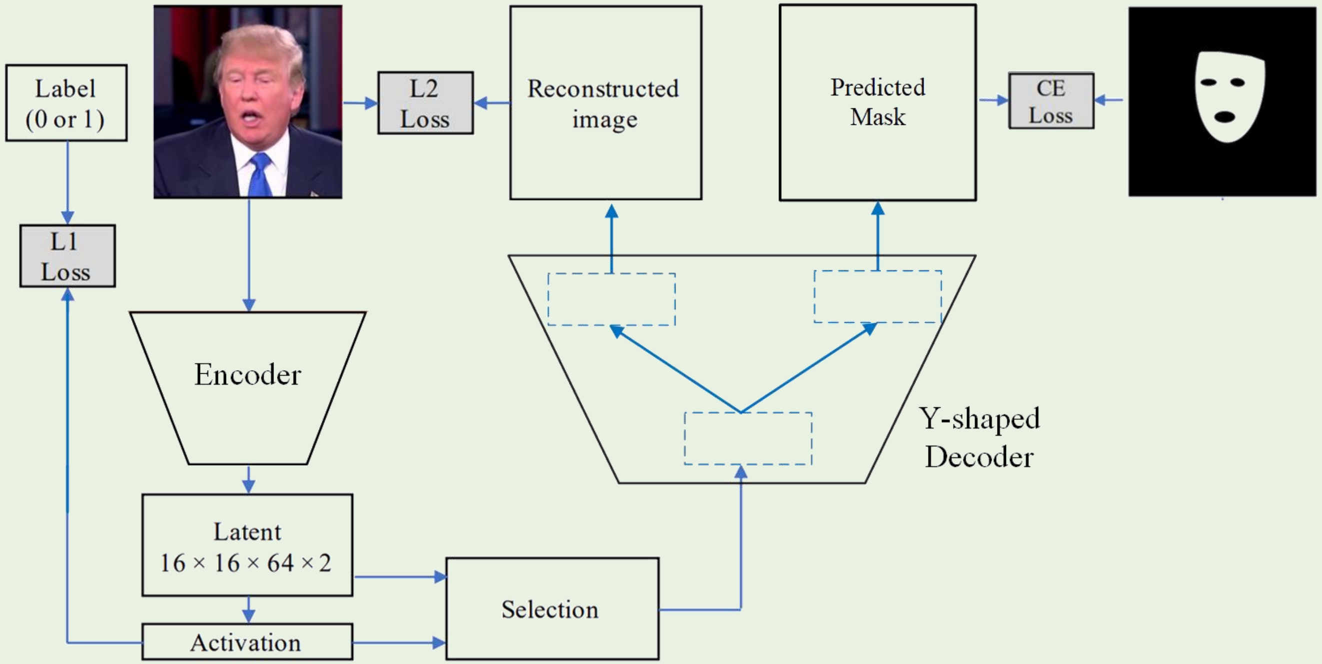
# Static Methods

- CNN binary classification only
  - Two-stream neural networks for tampered face detection
  - Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

- Joint binary classification and manipulated region localization
  - Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos (segmentation)
  - Face X-ray for more general face forgery detection (face X-ray)
  - On the Detection of Digital Face Manipulation (attention)

# Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos

- A multi-task learning approach for simultaneously performing classification and segmentation of manipulated facial images.
- The information gained from classification, segmentation and reconstruction is shared among them, thereby improving the overall performance.




Neguyen et al. Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. In BTAS, 2019.

# Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos

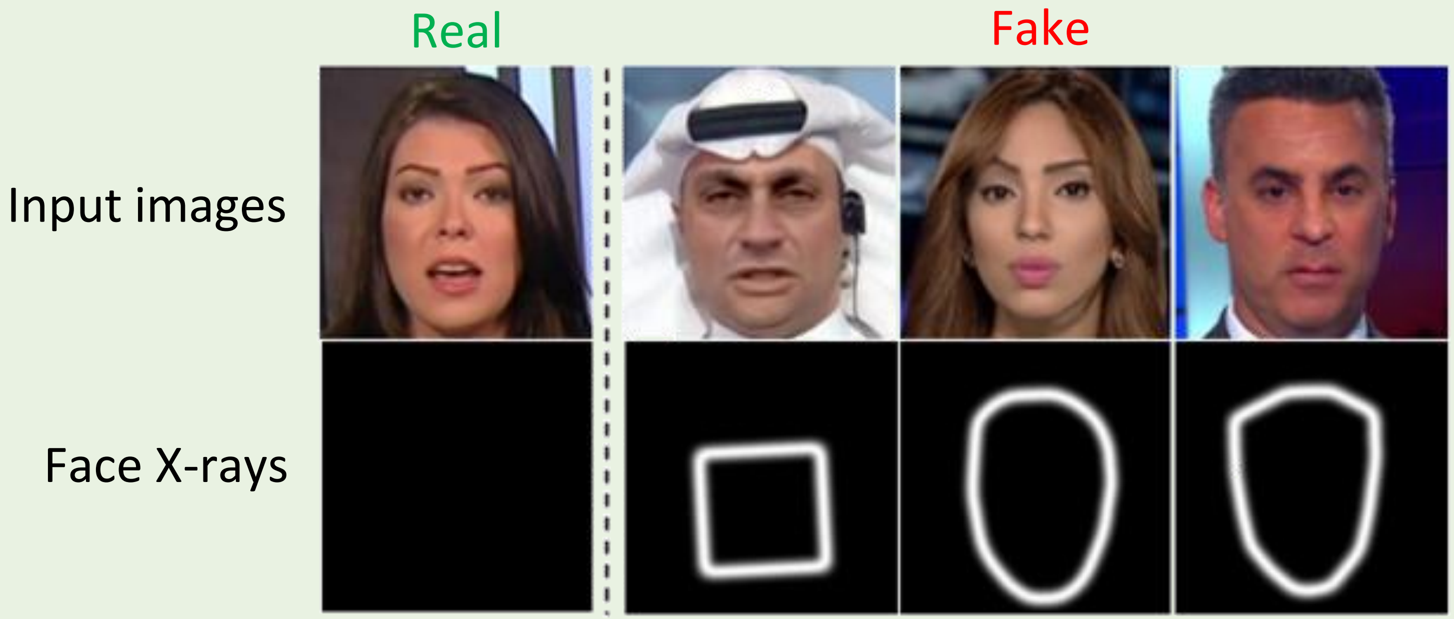
- Classification and segmentation on FaceForensics++ datasets.
- Proposed method without segmentation branch (No_seg), without reconstruction branch (No_recon), complete proposed method (Proposed_all).

| Method | Classification | | Segmentation |
|---|---|---|---|
| | Accuracy (%) | EER (%) | Accuracy (%) |
| FT_Res | 82.30 | 14.53 | - |
| FT | 88.43 | 11.60 | - |
| No_seg | **93.63** | 7.20 | - |
| No_recon | 93.40 | **7.07** | 89.21 |
| Proposed_all | 92.77 | 8.18 | **90.27** |

Neguyen et al. Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. In BTAS, 2019.

# Face X-ray for More General Face Forgery Detection

- Most existing manipulation methods share a common step: blending the altered face into a background image.
- Face X-ray reveals whether the input image can be decomposed into the blending of two images from different sources.



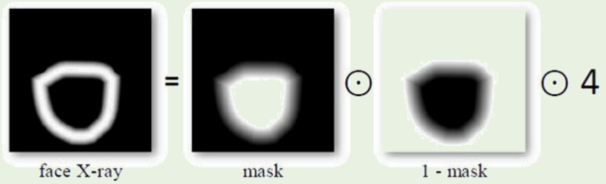

Li et al. Face X-ray for more general face forgery detection. In CVPR, 2020.

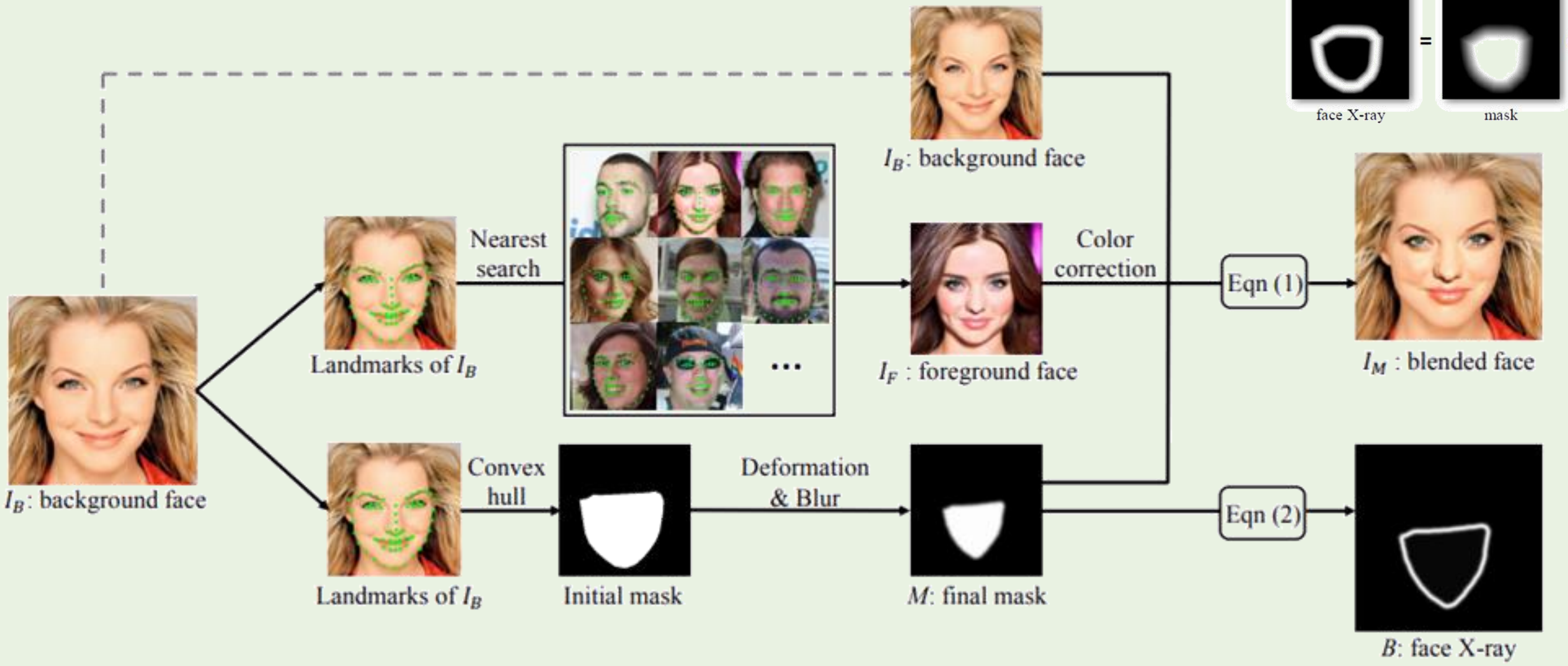
# Face X-ray for More General Face Forgery Detection

- Training data generation from real images.

$$I_M = M \odot I_F + (1 - M) \odot I_B \qquad (1)$$

$$B_{i,j} = 4 \cdot M_{i,j} \cdot (1 - M_{i,j}) \qquad (2)$$




Li et al. Face X-ray for more general face forgery detection. In CVPR, 2020.

# Face X-ray for More General Face Forgery Detection

- HRNet predicts Face X-rays which is then used to classify the image as fake or real.
- Loss functions:

$$L = \lambda L_b + L_c$$

- Cross-entropy loss measures the accuracy of the predicted X-rays.

$$L_b = -\sum_{\{I,B\}\in D}(B_{i,j}\log(\hat{B}_{i,j} + (1 - B_{i,j})\log(1 - \hat{B}_{i,j}))$$

- For classification, the loss is

$$L_c = -\sum_{\{I,c\}\in D}(c\log(\hat{c}) + (1 - c)\log(1 - \hat{c})$$

Li et al. Face X-ray for more general face forgery detection. In CVPR, 2020.

# Face X-ray for More General Face Forgery Detection

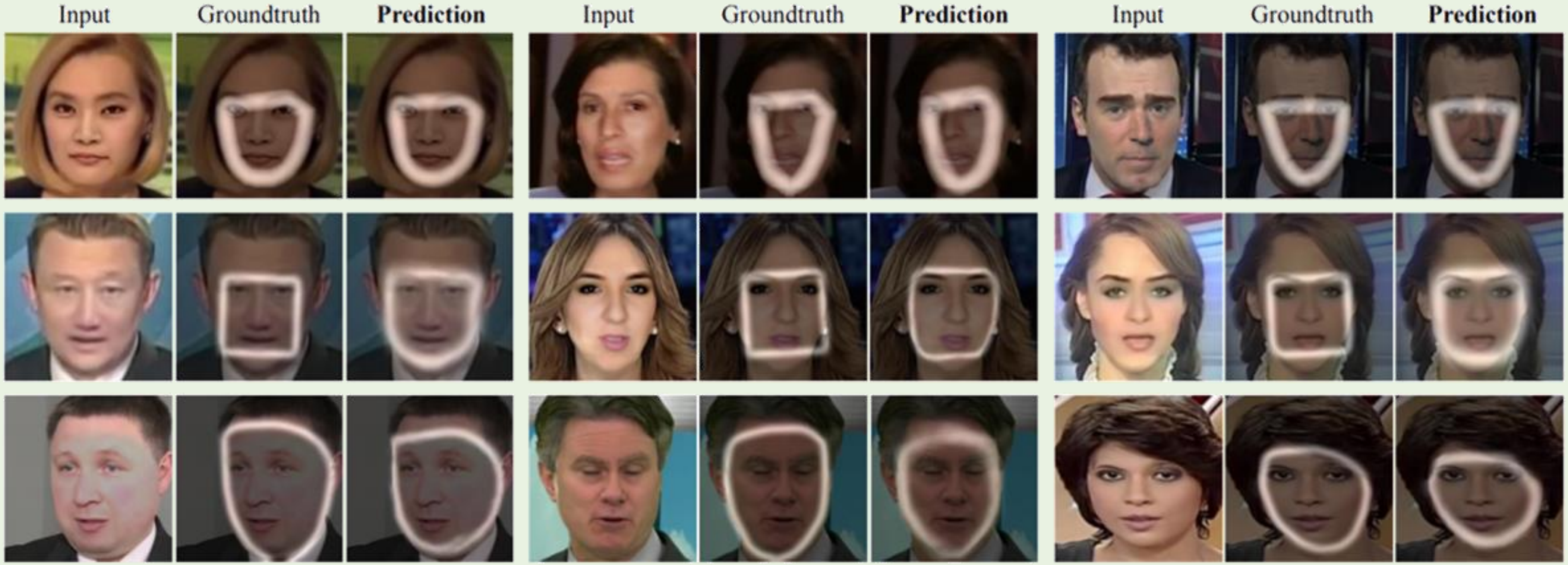
- Results on unseen datasets.
- Without using images from facial manipulation methods, already outperforms the baseline (Xception)

| Model | Training dataset | Test dataset | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | DFD | | | DFDC | | | Celeb-DF | | |
| | | AUC | AP | EER | AUC | AP | EER | AUC | AP | EER |
| Xception | FF++ | 87.86 | 78.82 | 21.49 | 48.98 | 50.83 | 50.45 | 36.19 | 50.07 | 59.64 |
| Face X-ray | BI | 93.47 | 87.89 | 12.72 | 71.15 | **73.52** | 32.62 | 74.76 | 68.99 | 31.16 |
| Face X-ray | FF++ and BI | **95.40** | **93.34** | **8.37** | **80.92** | 72.65 | **27.54** | **80.58** | **73.33** | **26.70** |

Li et al. Face X-ray for more general face forgery detection. In CVPR, 2020.

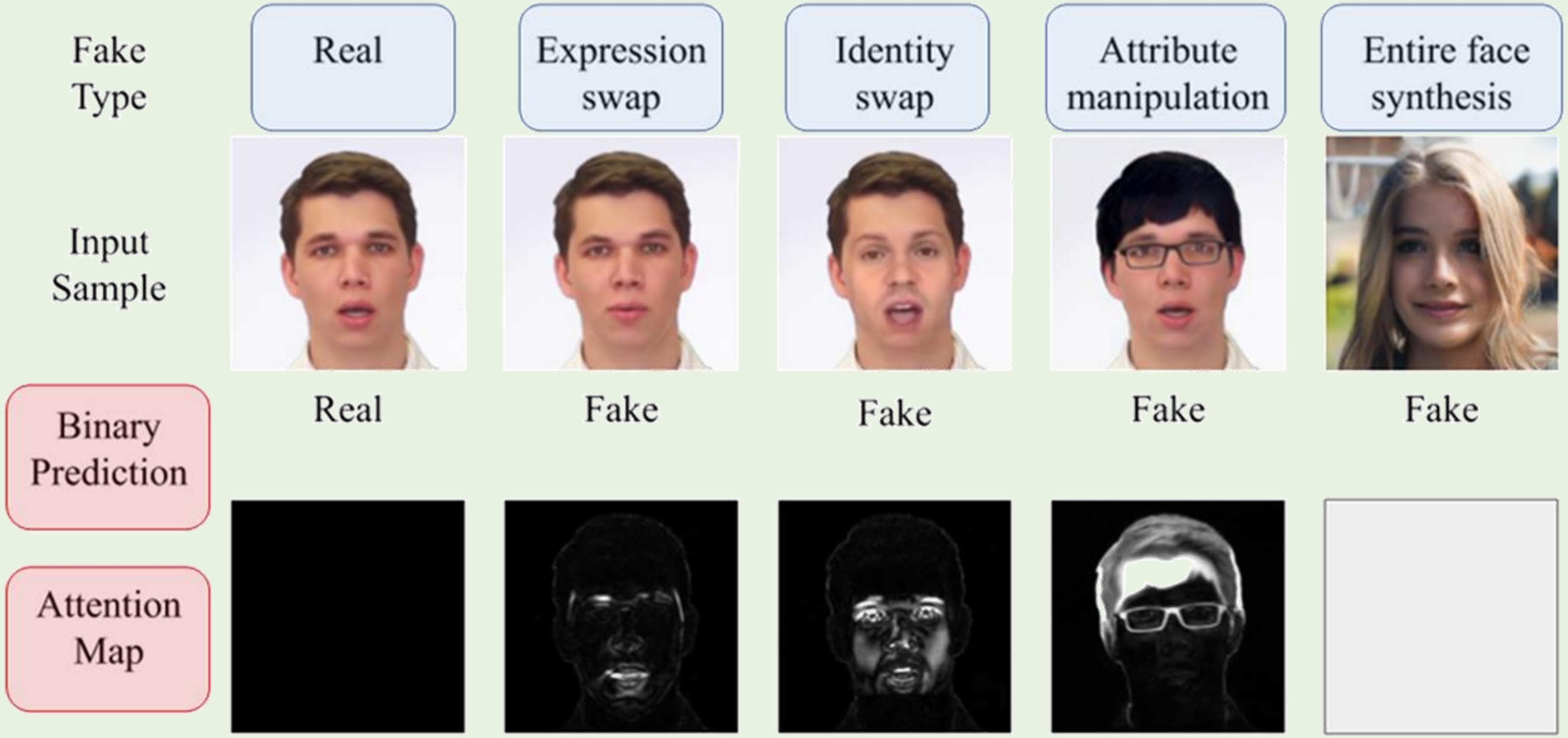
# Face X-ray for More General Face Forgery Detection

- Visual results on various facial manipulation samples.



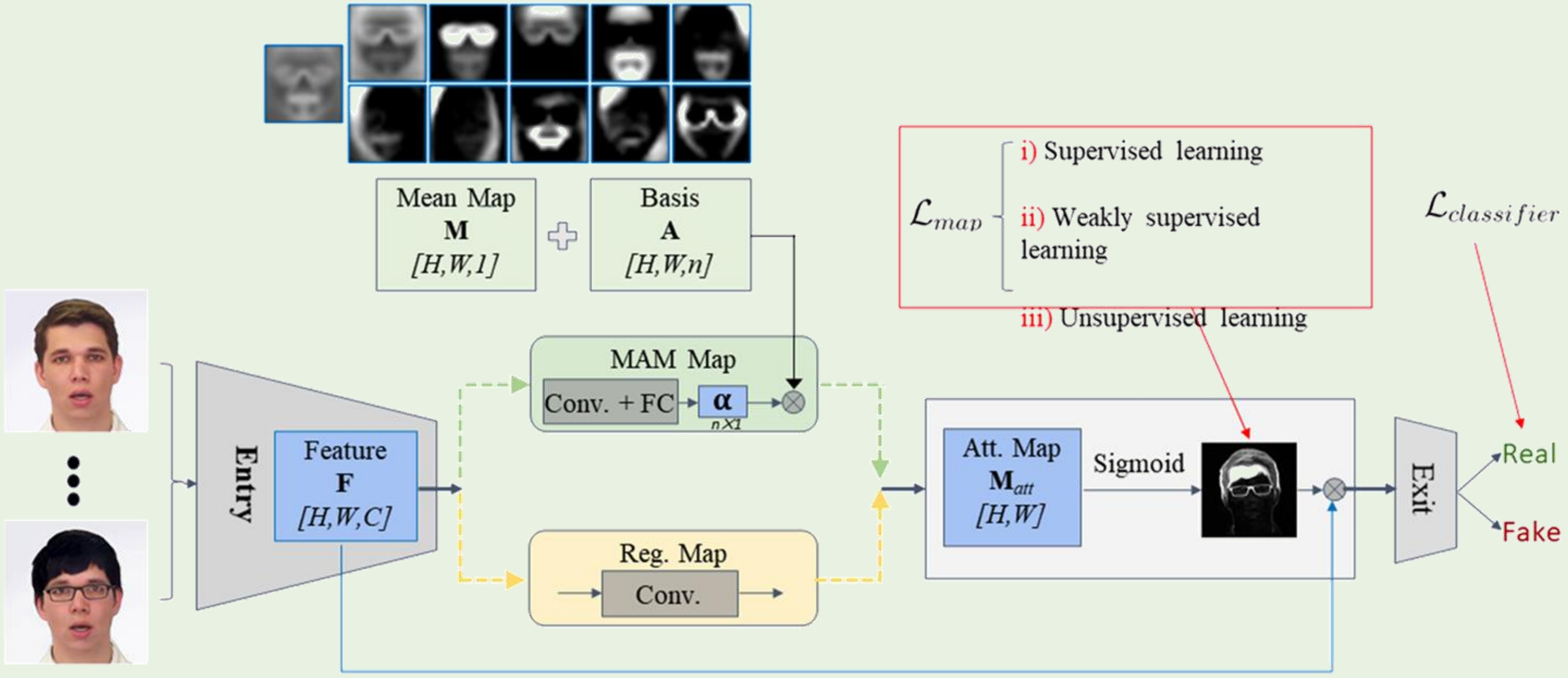

Li et al. Face X-ray for more general face forgery detection. In CVPR, 2020.

# On the Detection of Digital Face Manipulation




Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

# Proposed Method




Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

# Loss Functions

$$\mathcal{L} = \mathcal{L}_{classifier} + \lambda \mathcal{L}_{map}$$

Binary classification loss of Softmax

Attention map loss

- Supervised learning

$$\mathcal{L}_{\text{map}} = ||\mathbf{M}_{att} - \mathbf{M}_{gt}||_1$$

- Weakly supervised learning

$$\mathcal{L}_{\text{map}} = \begin{cases} |\text{Sigmoid}(\mathbf{M}_{att}) - 0|, & \text{if real} \\ |\max(\text{Sigmoid}(\mathbf{M}_{att})) - 0.75|. & \text{if fake} \end{cases}$$

- Unsupervised learning

Without any map supervision when $\lambda$ set to 0.

Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

# DFFD ---- comparison

Diverse Fake Face Dataset (DFFD)

| Dataset | Year | # Still images | | # Video clips | | # Fake types | | | | Pose variation |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Real | Fake | Real | Fake | Id. swap | Exp. swap | Attr. mani. | Entire syn. | |
| Zhou *et al.* [1] | 2018 | 2,010 | 2,010 | - | - | 2 | - | - | - | Unknown |
| Yang *et al.* [2] | 2018 | 241 | 252 | 49 | 49 | 1 | - | - | - | Unknown |
| Deepfake [3] | 2018 | - | - | - | 620 | 1 | - | - | - | Unknown |
| FaceForensics++ [4] | 2019 | - | - | 1,000 | 3,000 | 2 | 1 | - | - | $[-30°, 30°]$ |
| FakeSpotter [5] | 2019 | 6,000 | 5,000 | - | - | - | - | - | 2 | Unknown |
| **DFFD** (our) | 2019 | 58,703 | 240,336 | 1,000 | 3,000 | 2 | 1 | 28 + 40 | 2 | $[-90°, 90°]$ |


[1] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis.Two-stream neural networks for tampered face detection. CVPRW 2017

[2] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakesusing inconsistent head poses. ICASSP 2019

[3] Pavel Korshunov and S ́ebastien Marcel. DeepFakes: a newthreat to face recognition? assessment and detection. arXiv:1812.08685

[4] Andreas R ̈ossler, Davide Cozzolino, Luisa Verdoliva, Chris-tian Riess, Justus Thies, and Matthias Nießner. FaceForen-sics++: Learning to detect manipulated facial images. ICCV 2019

[5] Run Wang, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, Jian Wang,and Yang Liu. FakeSpotter: A simple baseline for spottingAI-synthesized fake faces. arXiv:1909.06122

# Experimental Results ---- ablation study

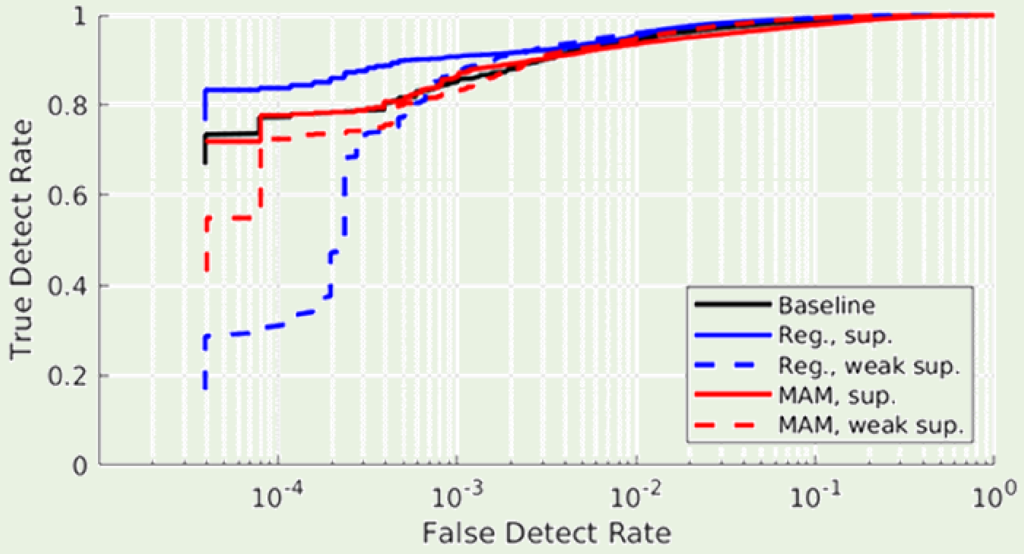
| Map Supervision | AUC | EER | $TDR_{0.01\%}$ | $TDR_{0.1\%}$ | PBCA |
|---|---|---|---|---|---|
| Xception | 99.61 | 2.88 | 77.42 | 85.26 | – |
| + Reg., *unsup.* | 99.76 | 2.16 | 77.07 | 89.70 | 12.89 |
| + Reg., *weak sup.* | 99.66 | 2.57 | 46.57 | 75.20 | 30.99 |
| + Reg., *sup.* | 99.64 | **2.23** | **83.83** | **90.78** | **88.44** |
| + Reg., *sup.* - map | **99.69** | 2.73 | 48.54 | 72.94 | **88.44** |
| + MAM, *unsup.* | 99.55 | 3.01 | 58.55 | 77.95 | 36.66 |
| + MAM, *weak sup.* | 99.68 | 2.64 | 72.47 | 82.74 | 69.49 |
| + MAM, *sup.* | 99.26 | 3.80 | 77.72 | 86.43 | 85.93 |
| + MAM, *sup.* - map | 98.75 | 6.24 | 58.25 | 70.34 | 85.93 |

**Ablation for benefit of the attention map, with various combinations of map generation methods and supervisions.**

| Network | AUC | EER | $TDR_{0.01\%}$ | $TDR_{0.1\%}$ | PBCA |
|---|---|---|---|---|---|
| Xception | 99.61 | 2.88 | 77.42 | 85.26 | - |
| Xception + Reg. | **99.64** | **2.23** | **83.83** | **90.78** | **88.44** |
| Xception + MAM | 99.26 | 3.80 | 77.72 | 86.43 | 85.93 |
| VGG16 | 96.95 | 8.43 | 0.00 | 51.14 | - |
| VGG16 + Reg. | 99.46 | 3.40 | 44.16 | 61.97 | **91.29** |
| VGG16 + MAM | **99.67** | **2.66** | **75.89** | **87.25** | 86.74 |

**Our attention layer in two backbone networks**

# Experimental Results ---- forgery detection




Forgery detection ROCs of the XceptionNet


Binary classification accuracy for different fake types

# Experimental Results ---- forgery detection

| Methods | Training data | UADFV [3] | Celeb-DF [8] |
|---|---|---|---|
| Two-stream [1] | Private data | 85.1 | 55.7 |
| Meso4 [2] | Private data | 84.3 | 53.6 |
| MesoInception4 [2] | | 82.1 | 49.6 |
| HeadPose [3] | UADFV | 89.0 | 54.8 |
| FWA [4] | UADFV | 97.4 | 53.8 |
| VA-MLP [5] | Private data | 70.2 | 48.8 |
| VA-LogReg [5] | | 54.0 | 46.9 |
| Multi-task [6] | FF | 65.8 | 36.5 |
| Xception-FF++ [7] | FF++ | 80.4 | 38.7 |
| Xception | DFFD | 75.6 | 63.9 |
| Xception | UADFV | 96.8 | 52.2 |
| Xception | UADFV, DFFD | 97.5 | 67.6 |
| Xception+Reg. | DFFD | 84.2 | 64.4 |
| Xception+Reg. | UADFV | **98.4** | 57.1 |
| Xception+Reg. | UADFV, DFFD | **98.4** | **71.2** |

**AUC (%) on UADFV and Celeb-DF**


[1] Zhou et al.Two-stream neural networks for tampered face detection. CVPRW 2017

[2] Afchar et al. MesoNet: a compact facial video forgery detection network. WIFS 2018

[3] Yang et al. Exposing deep fakes using inconsistent head poses. ICASSP 2019

[4] Li et al. Exposing deepfake videos by detecting face warping artifacts. CVPRW 2019

[5] Matern et al. Exploiting visual artifacts to expose deepfakes and face manipulations. WACVW 2019

[6] Nguyen et al. Multi-task learning for detecting and segmenting manipulated facial images and videos. BTAS 2019

[7] R ̈ossler et al. FaceForen-sics++: Learning to detect manipulated facial images. ICCV 2019

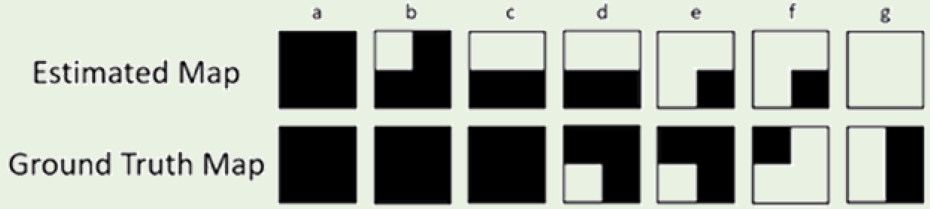
[8] Li et al. Celeb-DF: A new dataset for deepfake forensics. CVPR 2020

# Experimental Results ---- manipulation localization

Metrics for evaluating the attention maps: IoU, Cosine Similarity, Pixel-wise Binary Classification Accuracy (PBCA) and proposed Intersection Non-Containment (IINC).

$$\text{IINC} = \frac{1}{3 - |\mathbf{U}|} * \begin{cases} 0 & \text{if } \overline{\mathbf{M}_{\text{gt}}} = 0 \text{ and } \overline{\mathbf{M}_{\text{att}}} = 0 \\ 1 & \text{if } \overline{\mathbf{M}_{\text{gt}}} = 0 \text{ xor } \overline{\mathbf{M}_{\text{att}}} = 0 \\ (2 - \frac{|\mathbf{I}|}{|\mathbf{M}_{\text{att}}|} - \frac{|\mathbf{I}|}{|\mathbf{M}_{\text{gt}}|}) & \text{otherwise,} \end{cases}$$

I and U are the intersection and union between $\mathbf{M}_{gt}$ and $\mathbf{M}_{att}$ .

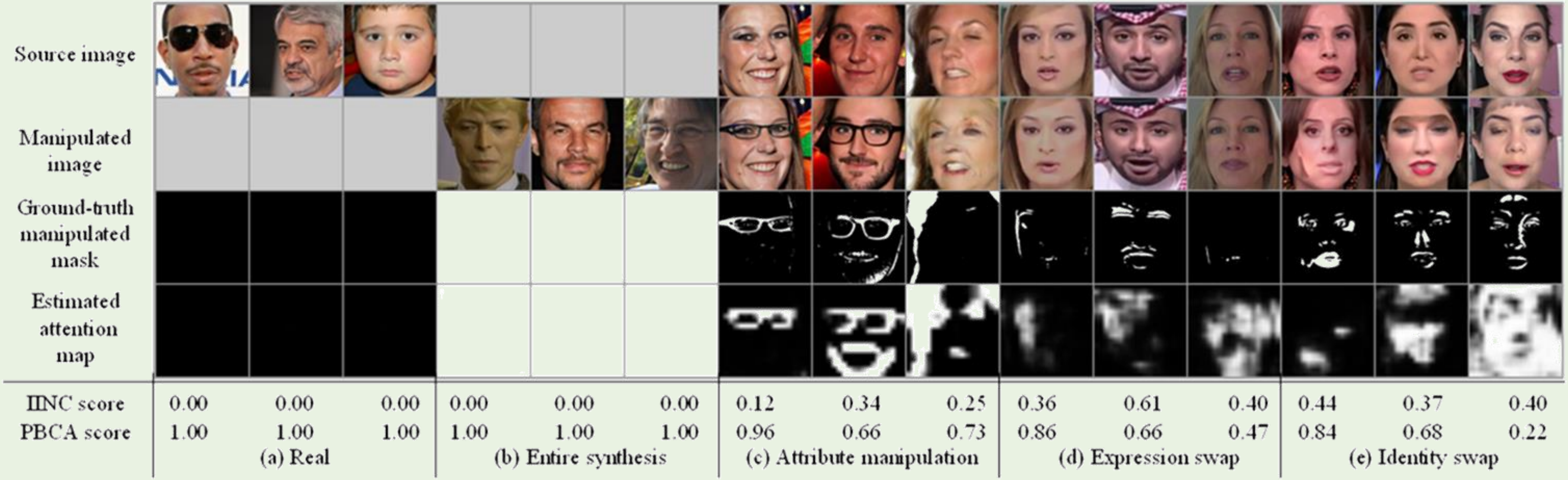



| | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| Cosine | - | - | - | 1 | 0.43 | 0.33 | 0.29 |
| IOU | - | 0.00 | 0.00 | 0.00 | 0.33 | 0.25 | 0.50 |
| PBCA | 1.00 | 0.75 | 0.50 | 0.25 | 0.50 | 0.50 | 0.50 |
| IINC | 0.00 | 0.36 | 0.40 | 0.89 | 0.30 | 0.33 | 0.25 |

| Data | IINC ↓ | IoU ↑ | Cosine Similarity ↓ | PBCA ↑ |
|---|---|---|---|---|
| All Real | 0.015 | – | – | 0.998 |
| All Fake | 0.147 | 0.715 | 0.192 | 0.828 |
| Partial | 0.311 | 0.401 | 0.429 | 0.786 |
| Complete | 0.077 | 0.847 | 0.095 | 0.847 |
| All | 0.126 | – | – | 0.855 |

**Evaluating manipulation localization with 4 metrics**

# Experimental Results ---- manipulation localization



# Outline

- Introduction of digital attacks

    -- Problem, Facial manipulation types, Challenges

- Benchmark databases
- Face manipulation detection methods

    -- Dynamic methods, Static methods

- **Future Directions**

# Future Directions

- Facial manipulation techniques are continuously improving. More research on generalization ability of forgery detection against **unseen** manipulation types.

- Challenging when performed in **uncontrolled** scenarios. Fake imagery on social network are usually suffering from large variations in compression, resizing, noise, etc.

- Fusion of other **modalities** such as text or audio can be valuable to improve the detectors.