Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues

Yuyang Qian[*,†1,2], Guojun Yin[†,‡1], Lu Sheng[‡3], Zixuan Chen[*1,4], and Jing Shao[1]

[1] SenseTime Research,
[2] University of Electronic Science and Technology of China,
[3] College of Software, Beihang University,
[4] Northwestern Polytechnical University

ECCV 2020

Background

- Target: Distinguish the fake human face (failed by naked eyes)

Face2Face*,2016

DeepFake*,2018

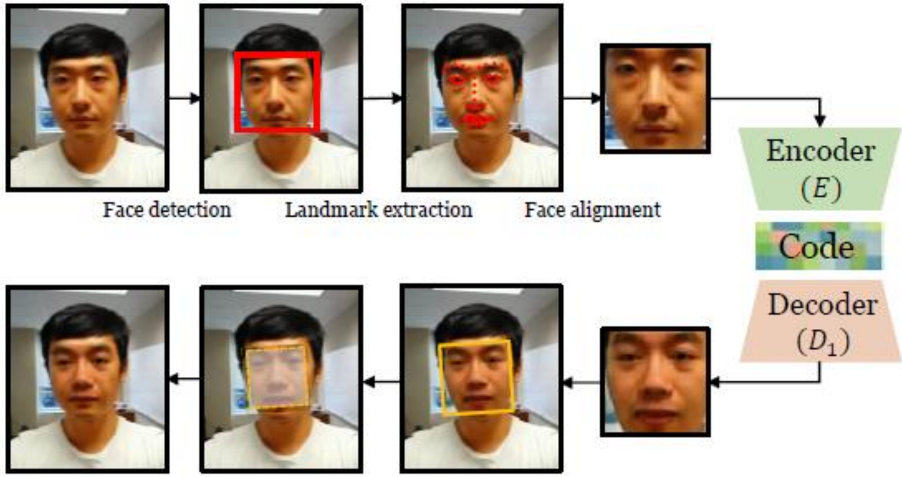
NuralTexture*,2019

FaceShifter,2020

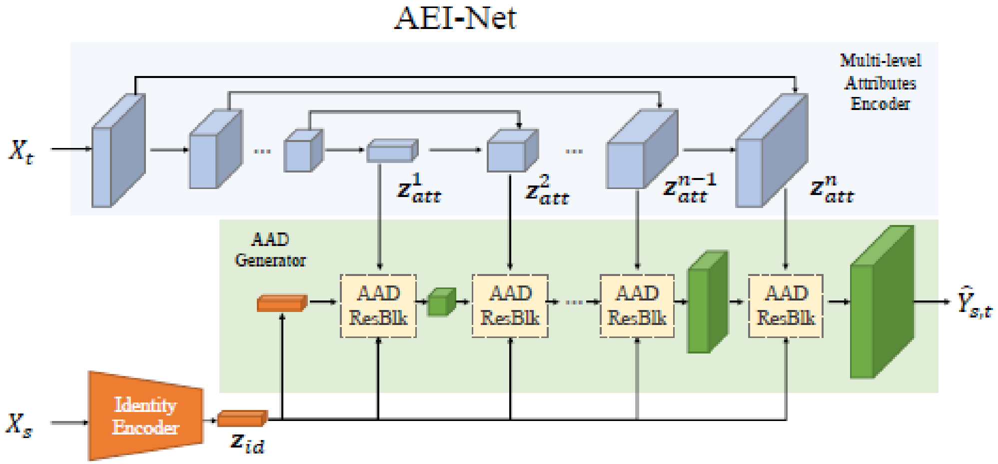
Celeb-DF,2020

FaceSwap*,
DFD,
DFDC,
DeeperForensics
(CVPR 2020),
......

- Classify: Manipulation or Reenactment / Blending or Synthesis



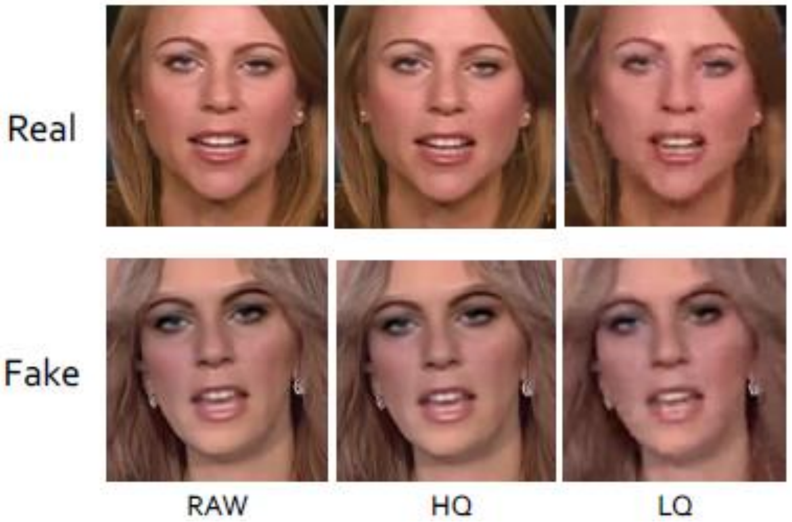
[Li et al., Celeb-DF: A Large-scale Challenging ···, 2020]



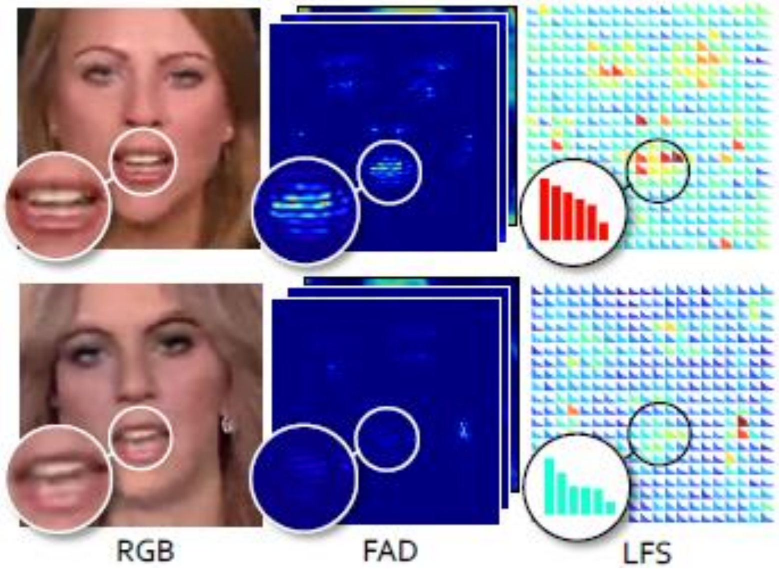
[Li et al., FaceShifter: Towards High Fidelity ···, 2020]

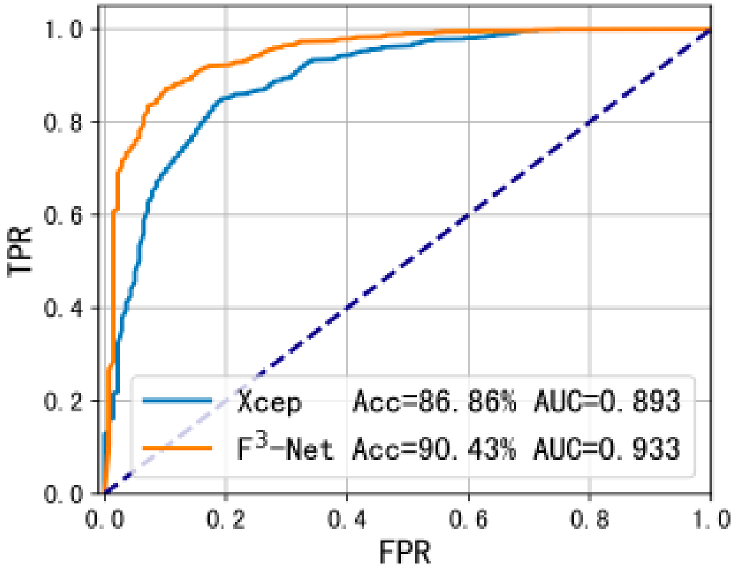
Motivation

- Related methods: Detect forgery artifacts
- Challenge: JEPG compression
- Insights: Mine forgery artifacts with the awareness of frequency
- CNN-compatible : frequency feature with shift-invariance and local consistency



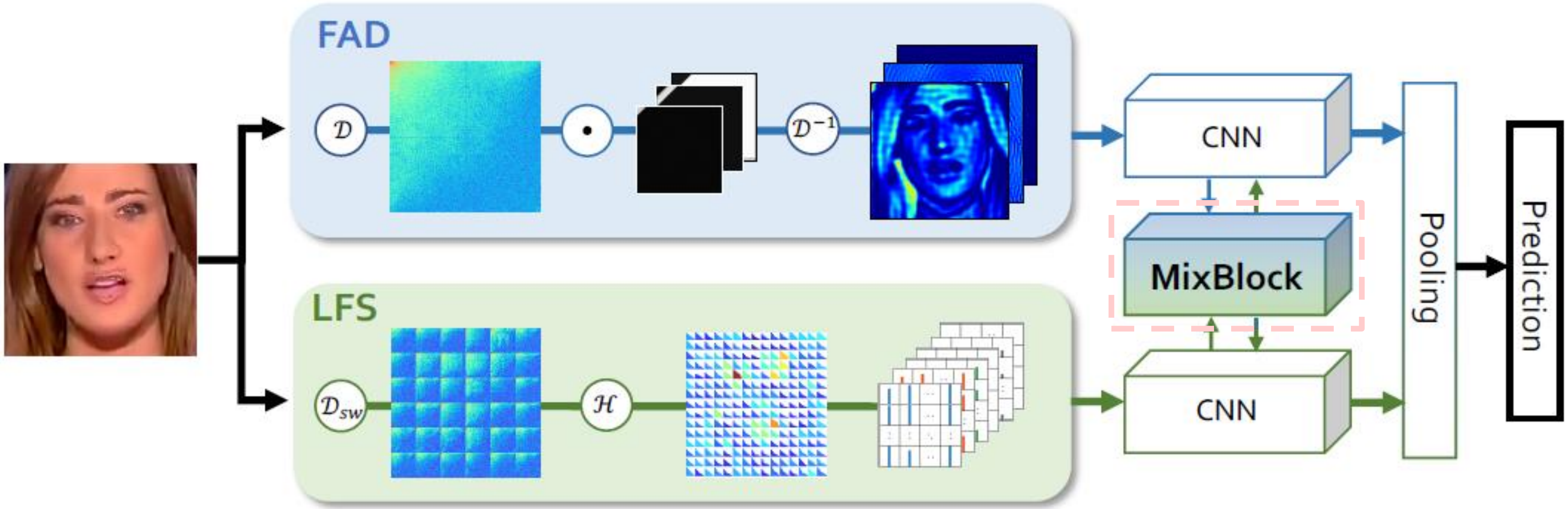
- FAD: Frequency-aware Decomposition
- LFS: Local Frequency Statistics





Overview of F3-Net

Learn subtle manipulation patterns through frequency-aware image decomposition



Extract local frequency statistics

Collaborative feature interaction

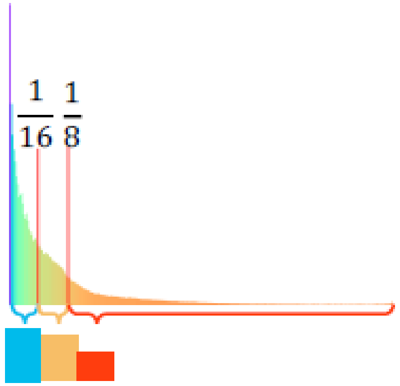
FAD: Frequency-Aware Decomposition

- Motive: Partition image in frequency and represent adaptively
- Combined filters: $\mathbf{f}_{base}^{i} + \sigma(\mathbf{f}_{w}^{i})$

Base filter

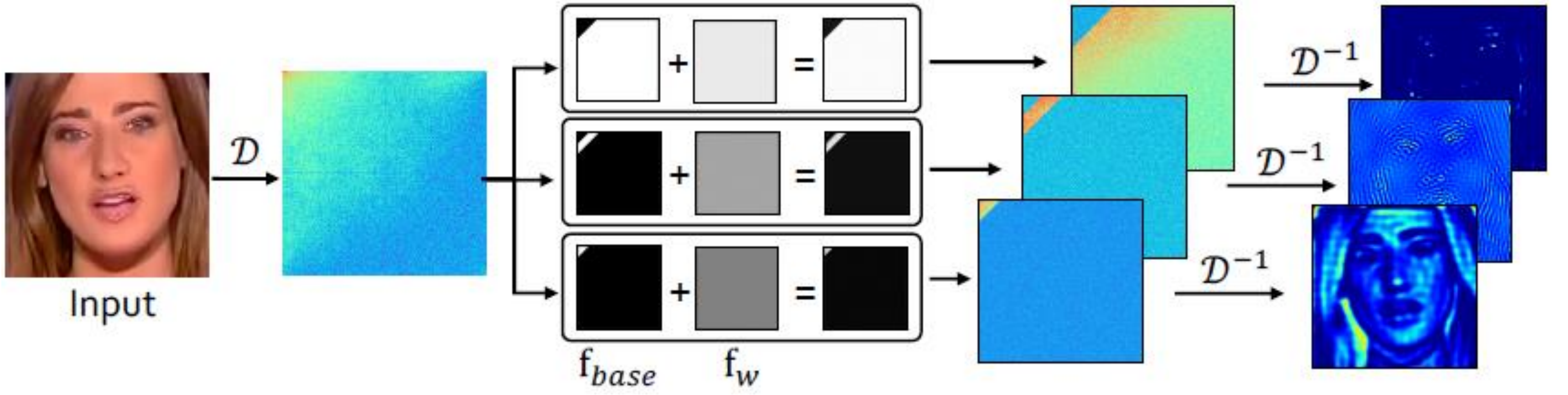
Learnable filter

Normalization in (-1, 1): $\sigma(x) = \frac{1-\exp(-x)}{1+\exp(-x)}$



- Formulation:

$$\mathbf{y}_i = \mathcal{D}^{-1}\{\mathcal{D}(\mathbf{x}) \odot [\mathbf{f}_{base}^{i} + \sigma(\mathbf{f}_{w}^{i})]\}, \quad i = \{1, \ldots, N\}$$



High band:

$[\frac{15}{16}, 1]$

Middle band:

$[\frac{7}{8}, \frac{15}{16}]$

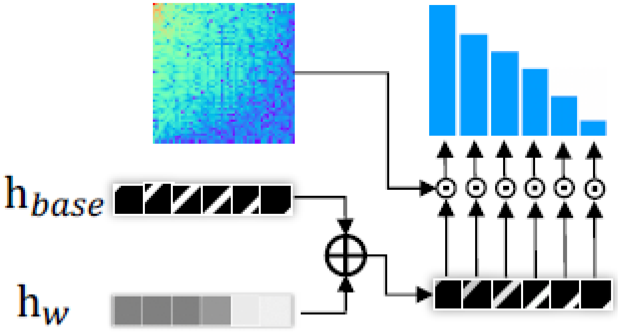
Low band:

$[0, \frac{7}{8}]$

LFS: Local Frequency Statistics

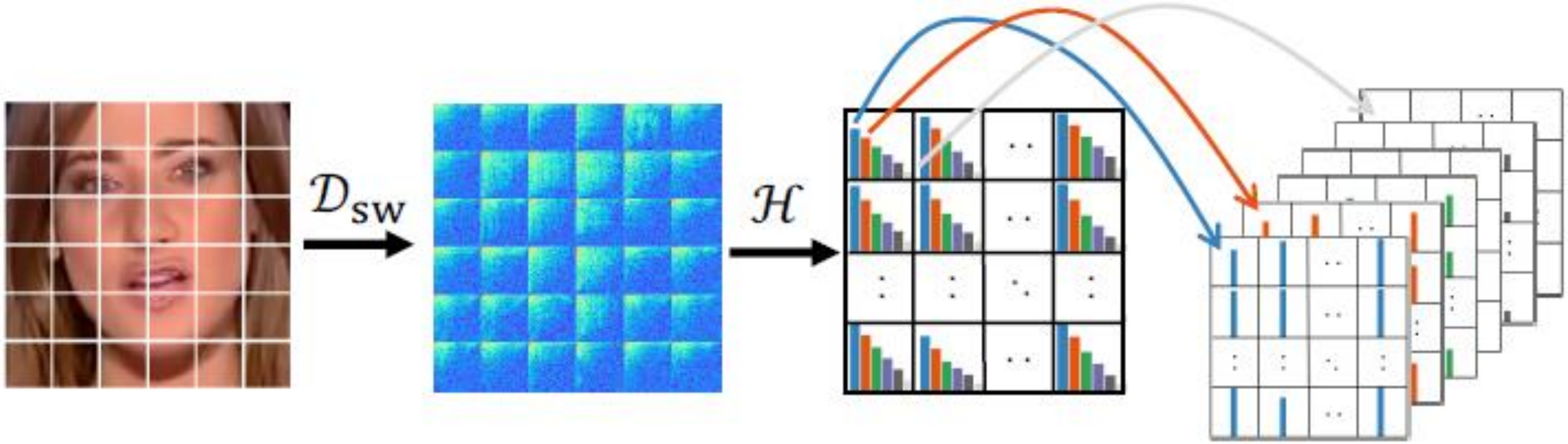
- Motive : Extract local frequency statistics to describe discrepancy
- Sliding Window DCT: Re-assemble back to a spatial map
- Local statistics in each frequency band:

$$\mathbf{q}_i = \log_{10} \|\mathcal{D}(\mathbf{p}) \odot [\mathbf{h}^i_{base} + \sigma(\mathbf{h}^i_w)]\|_1, \quad i = \{1, \ldots, M\}$$



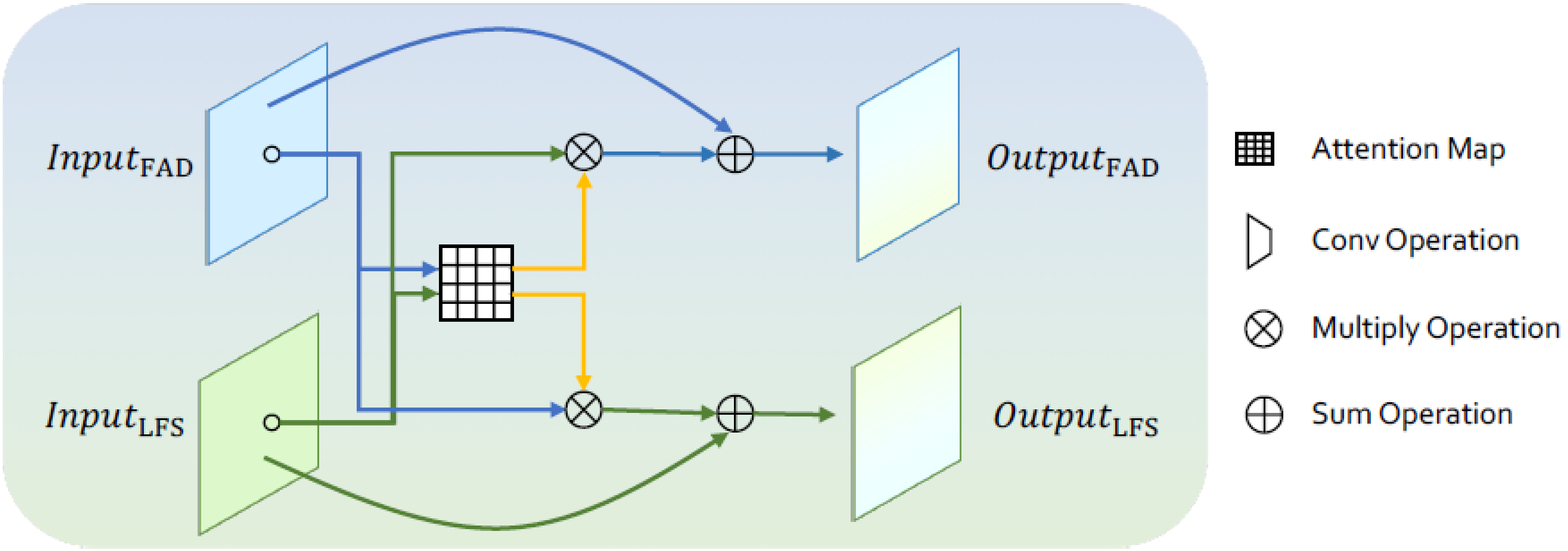
- Implementation details: *N = 6, Stride = 2*

  Input: 299 × 299 × 3 Output: 149 × 149 × 6



MixBlock: Two-stream Collaborative Learning

- Motive: Fuse two types of complementary clues FAD and LFS
- Cross-attention module: Augment the attentive features from one stream to another



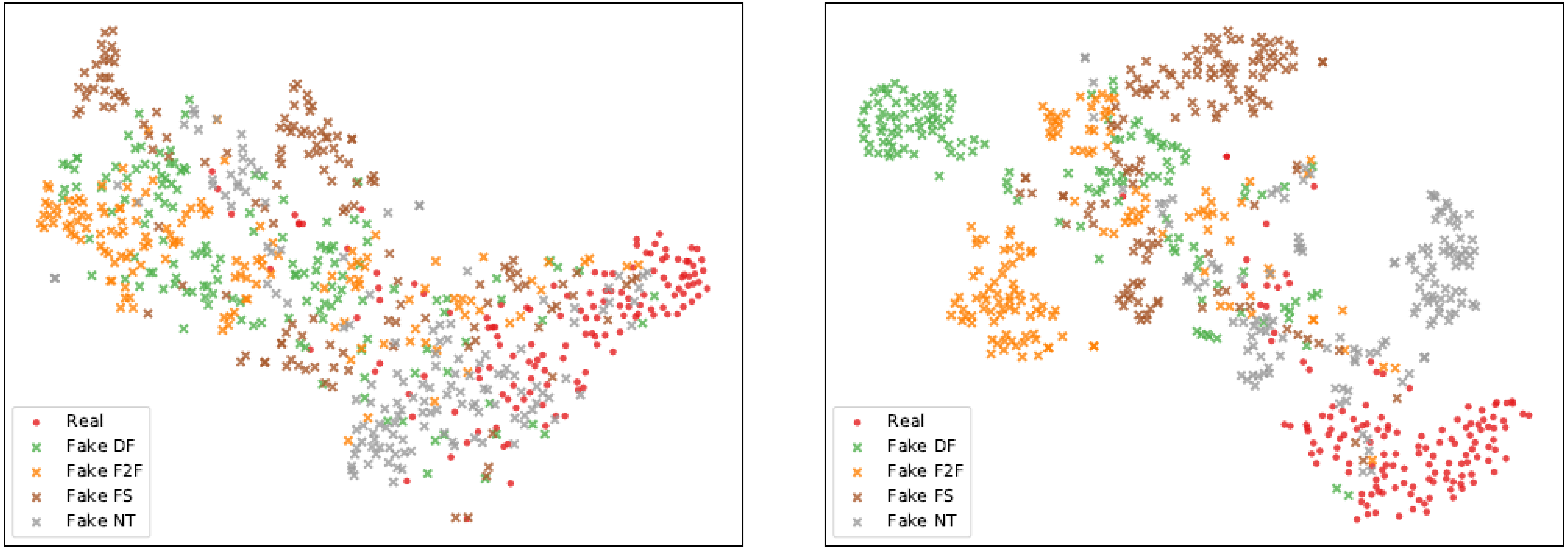
Results

- Quantitative results on FF++ dataset with all quality settings

Methods	Acc (LQ)	AUC (LQ)	Acc (HQ)	AUC (HQ)	Acc (RAW)	AUC (RAW)
Steg.Features [24]	55.98%	-	70.97%	-	97.63%	-
LD-CNN [14]	58.69%	-	78.45%	-	98.57%	-
Constrained Conv [6]	66.84%	-	82.97%	-	98.74%	-
CustomPooling CNN [49]	61.18%	-	79.08%	-	97.03%	-
MesoNet [3]	70.47%	-	83.10%	-	95.23%	-
Face X-ray [40]	-	0.616	-	0.874	-	-
Xception [12]	86.86%	0.893	95.73%	0.963	99.26%	0.992
Xception-ELA [27]	79.63%	0.829	93.86%	0.948	98.57%	0.984
Xception-PAFilters [10]	87.16%	0.902	-	-	-	-
F^3-Net (Xception)	**90.43%**	**0.933**	**97.52%**	**0.981**	**99.95%**	**0.998**
Optical Flow [5]	81.60%	-	-	-	-	-
Slowfast [20]	90.53%	0.936	97.09%	0.982	99.53%	0.994
F^3-Net(Slowfast)	**93.02%**	**0.958**	**98.95%**	**0.993**	**99.99%**	**0.999**

Embedding Visualization

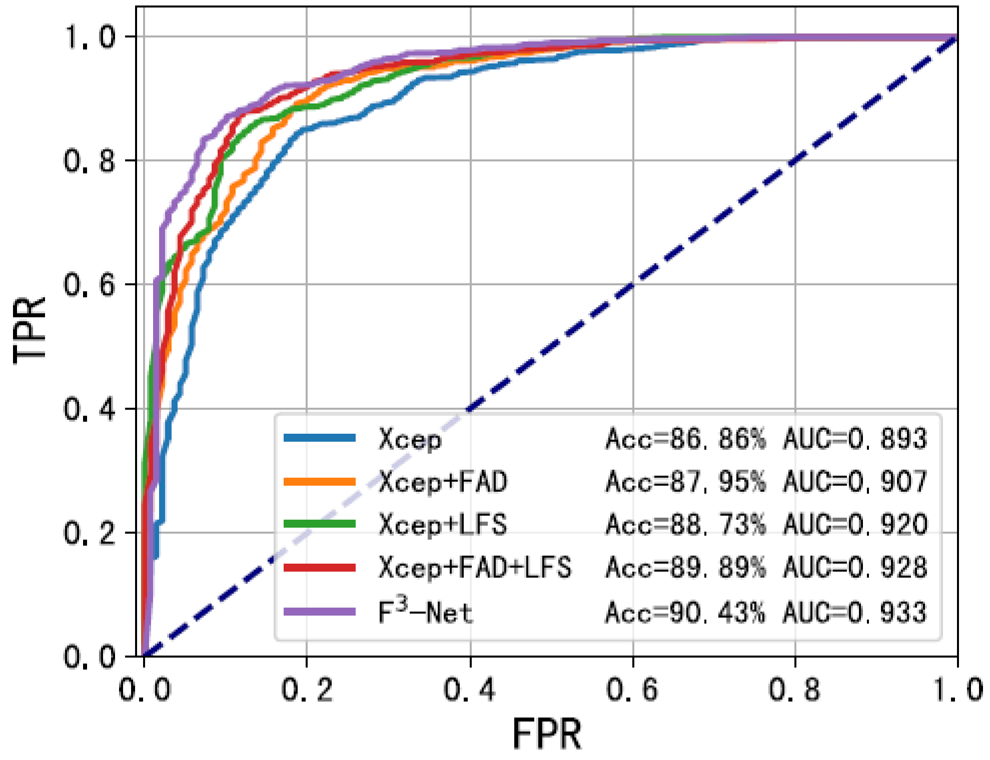
- The t-SNE embedding visualization of the baseline and F³-Net Baseline (Xception) in the left; F³-Net in the right



Ablation Study

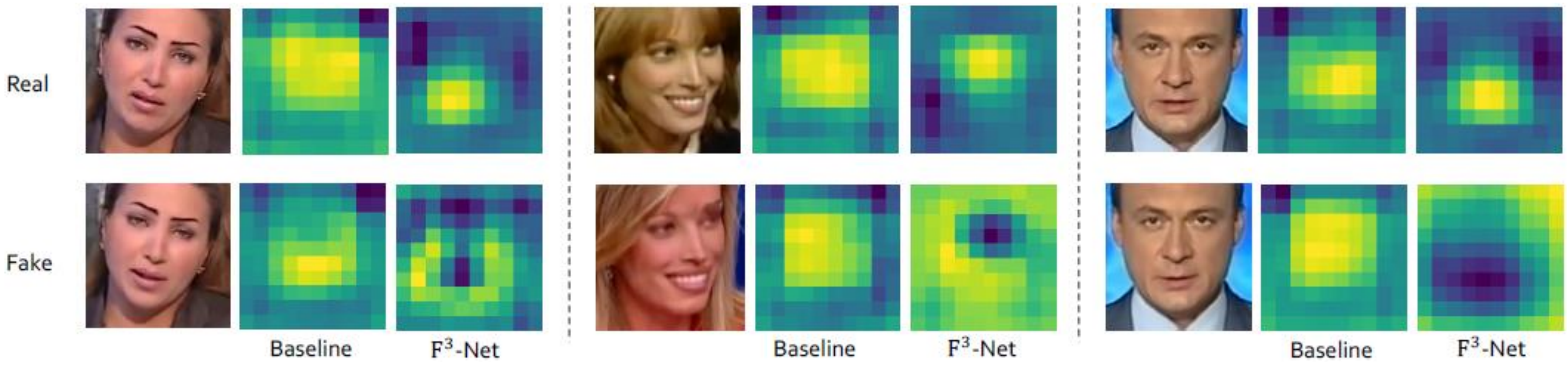
- Quantitatively evaluate F^3-Net and four variants:
  1) the baseline (Xception), 2) F^3-Net w/o LFS and MixBlock,
  3) F^3-Net w/o FAD and MixBlock, 4) F^3-Net w/o MixBlock

ID	FAD	LFS	MixBlock	Acc	AUC
1	-	-	-	86.86%	0.893
2	√	-	-	87.95%	0.907
3	-	√	-	88.73%	0.920
4	√	√	-	89.89%	0.928
5	√	√	√	**90.43%**	**0.933**



Visualization

- The visualization of feature map extracted by baseline and F^3-Net



Thank you!