Face Forgery Detection by 3D Decomposition

Xiangyu Zhu[1,2], Hao Wang[1], Hongyan Fei[3], Zhen Lei[1,2], Stan Z. Li[4]

[1]CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences,
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences,
[3]School of Automation and Electrical Engineering, University of Science and Technology Beijing,
[4]School of Engineering, Westlake University

Face Forgery

Deepfakes

Target

Face2Face

Pose
Background
Expression

Pose
Background
ID

Replacement
(Identity manipulation)

Original

Reenactment
(Expression manipulation)

FaceSwap

ID

Source

Expression

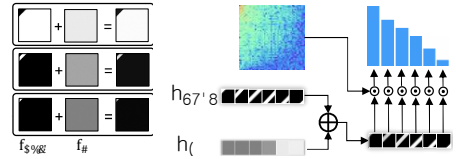
NeuralTextures

*Videos from *Faceforensics++*

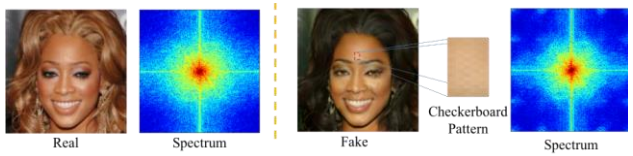
Goal: explore a novel decomposition

Hand-crafted **frequency** filters
(Stuchi *et al*. 2017)

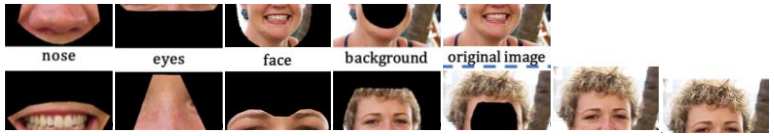
Learned **frequency** filters
(Qian *et al*. 2020)



Unique replications of **spectra**
(Zhang *et al*. 2019)

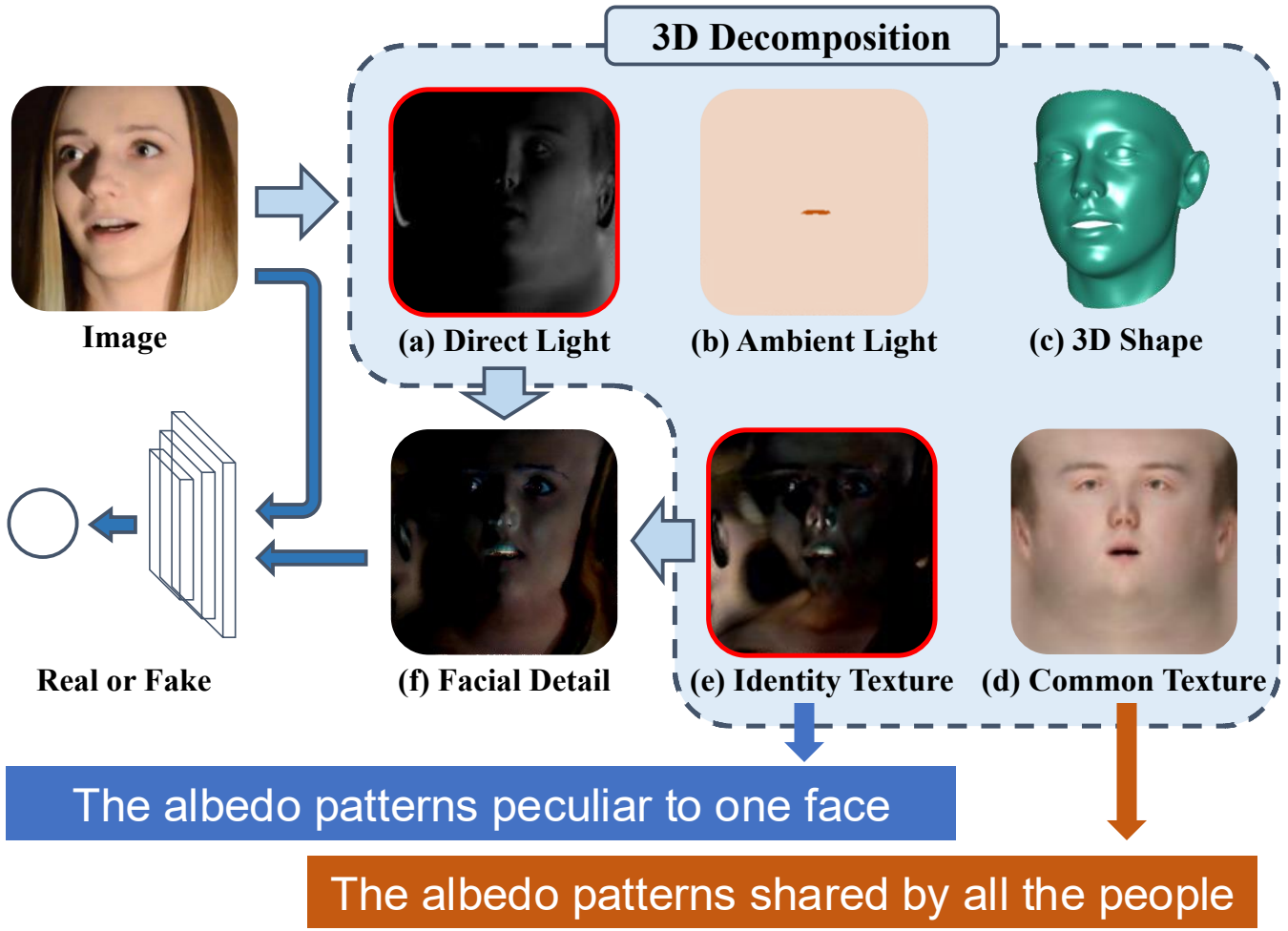


Spatial- and **frequency**-domain features
(Chen *et al*. 2020)

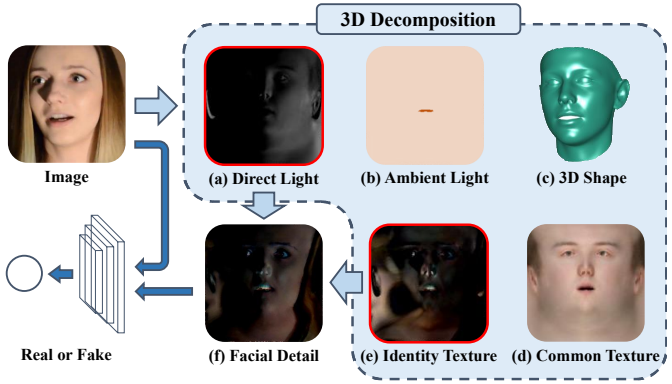


- **Difficult to decide which range of signals contains artifacts**
  - Different capture devices, environments, and compression algorithms
  - Large frequency distribution bias across datasets
- **Suffering from the generalization problem**

Idea: 3D Decomposition



3D Decomposition Implementation



- In computer graphics, a **face image** is generated by:

$$\mathbf{I}_{syn} = \mathit{Z\text{-}Buffer}(\mathbf{S}, \mathbf{C})$$

- Under the *Lambertian assumption*, the RGB of i[th] vertex is:

$$\mathbf{C}_i = \mathbf{Amb} * \mathbf{T}_i + \langle \mathbf{n}_i, \mathbf{l} \rangle \cdot \mathbf{Dir} * \mathbf{T}_i$$

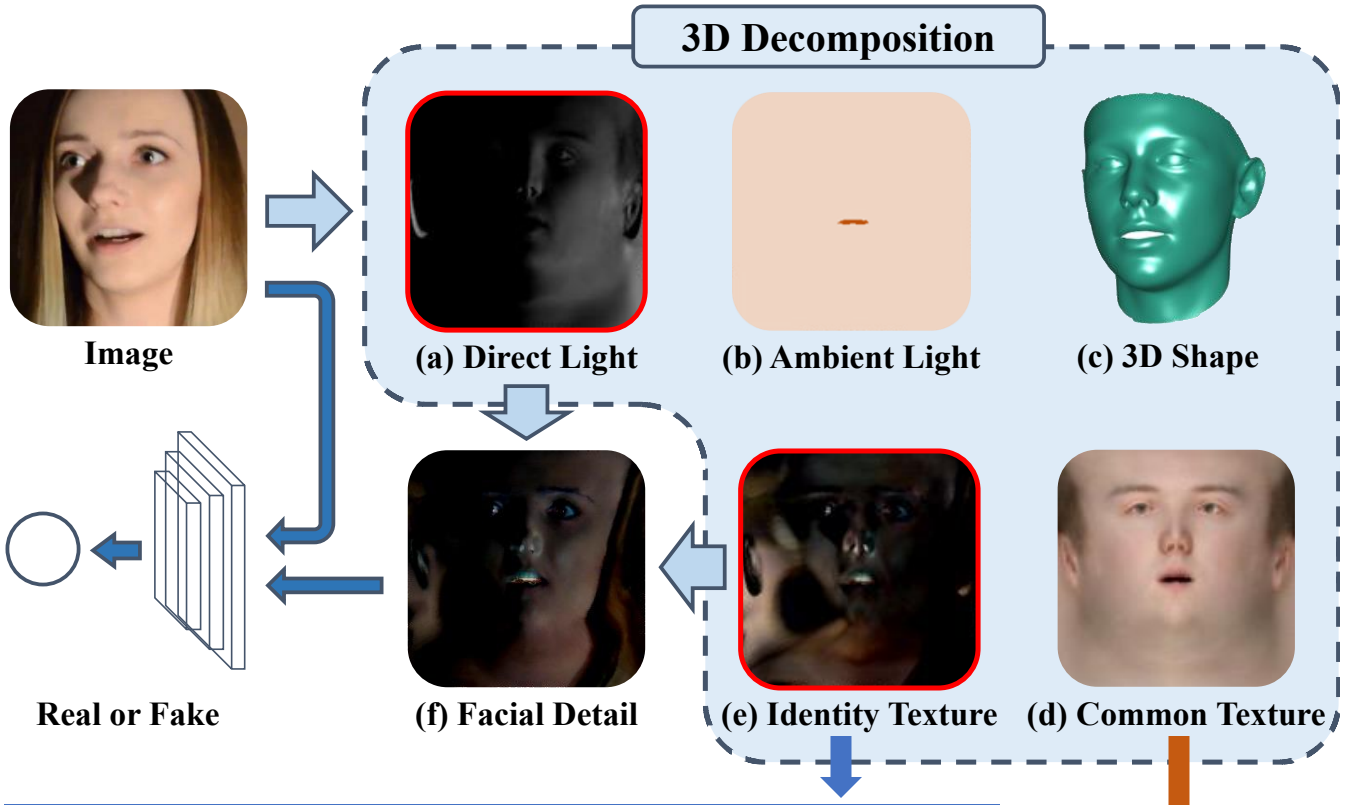
- Modeling the common texture by the **BFM PCA texture** model:

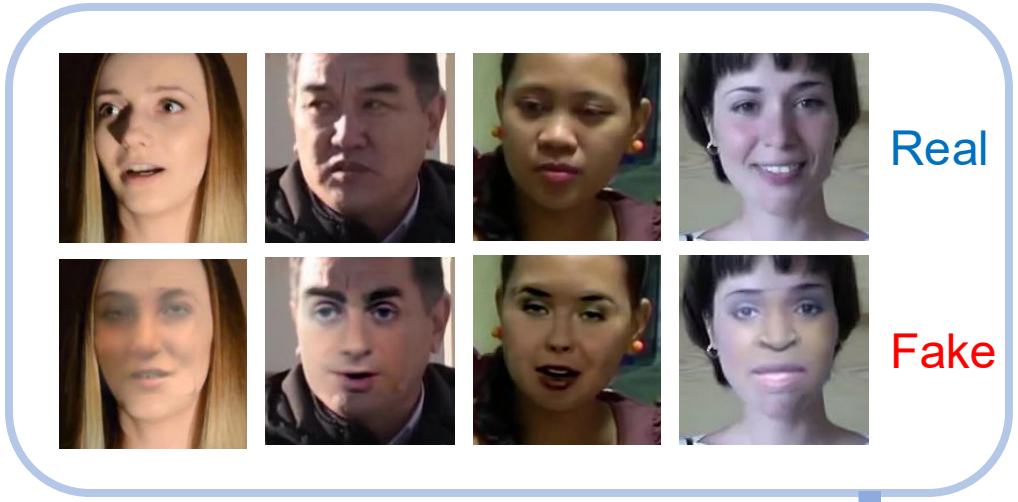
$$\mathbf{T} = \overline{\mathbf{T}} + \mathbf{B}\boldsymbol{\beta} + \mathbf{T}_{id}$$

- Optimizing the following loss to achieve the 3D face decomposition:

$$\arg \min_{\mathbf{S},\mathbf{Amb},\mathbf{Dir},\boldsymbol{\beta},\mathbf{T}_{id}} \|\mathbf{I} - \mathbf{I}_{syn}(\mathbf{S}, \mathbf{Amb}, \mathbf{Dir}, \boldsymbol{\beta}, \mathbf{T}_{id})\|$$

Assumption





- **Direct light**: decisive forgery clue with the observation on large artifacts under intense direct light
- **Identity texture**: hard to be simulated due to the rich variations across faces
- **3D shape**: its normalizing makes CNN concentrate on specific face regions and simplifies the detection
- **Ambient light** and **common texture**: easy to be faked and have little forgery clues

Facial Detail Generation

- Fitting texture and illumination by **PCA Texture** model and **Spherical Harmonic Reflectance**

$$\mathbf{I}(\mathbf{S}) = (\mathbf{H}\boldsymbol{\gamma}).*(\overline{\mathbf{T}} + \mathbf{B}\boldsymbol{\beta})$$

- **Facial detail**: Subtracting face image by the common texture under ambient light (in uv space)

$$\mathbf{FD} = UV(\mathbf{I} - (\mathbf{h}_1\gamma_1).*(\overline{\mathbf{T}} + \mathbf{B}\boldsymbol{\beta}), \mathbf{S})$$

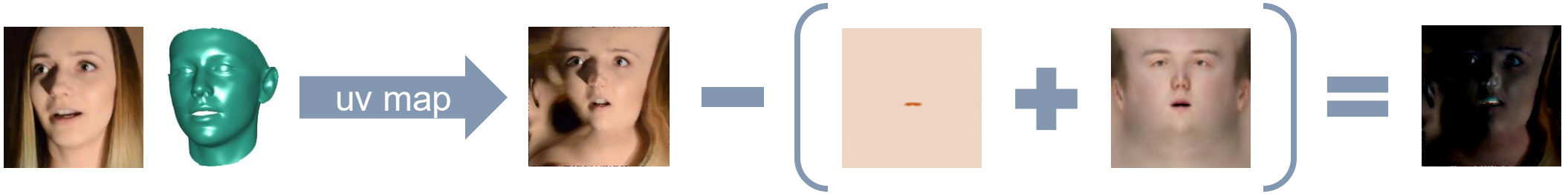
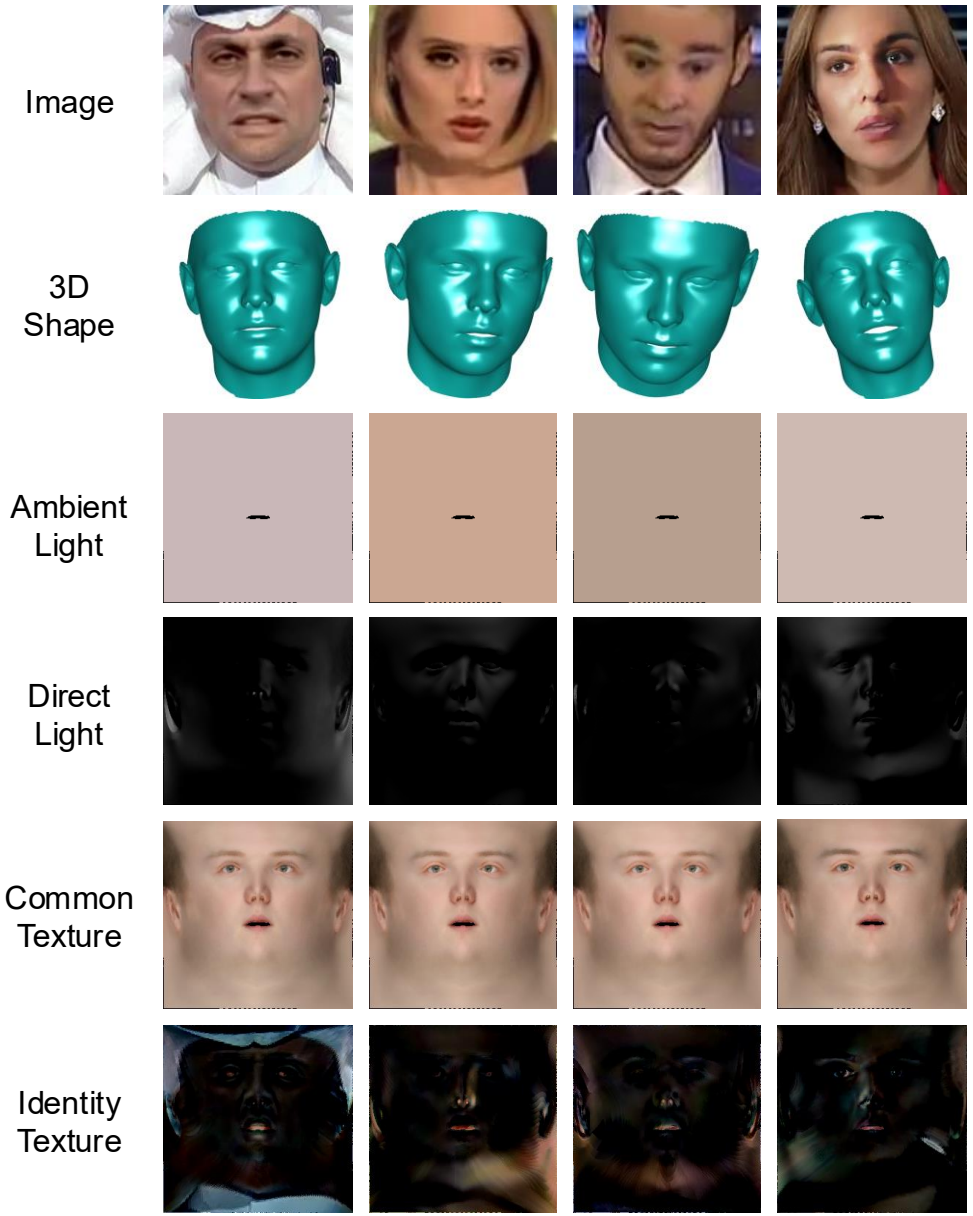
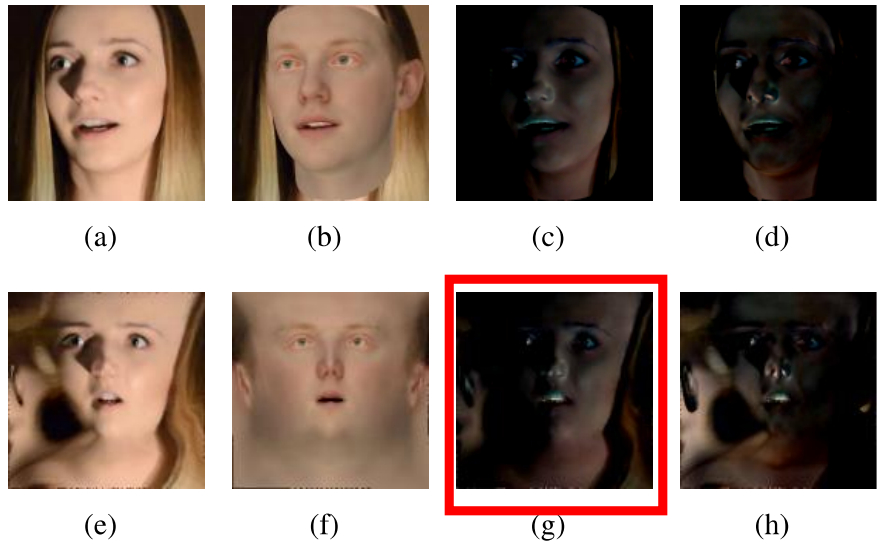


Image | 3D Shape | Image (uv space) | Ambient Light | Common Texture | Facial Detail

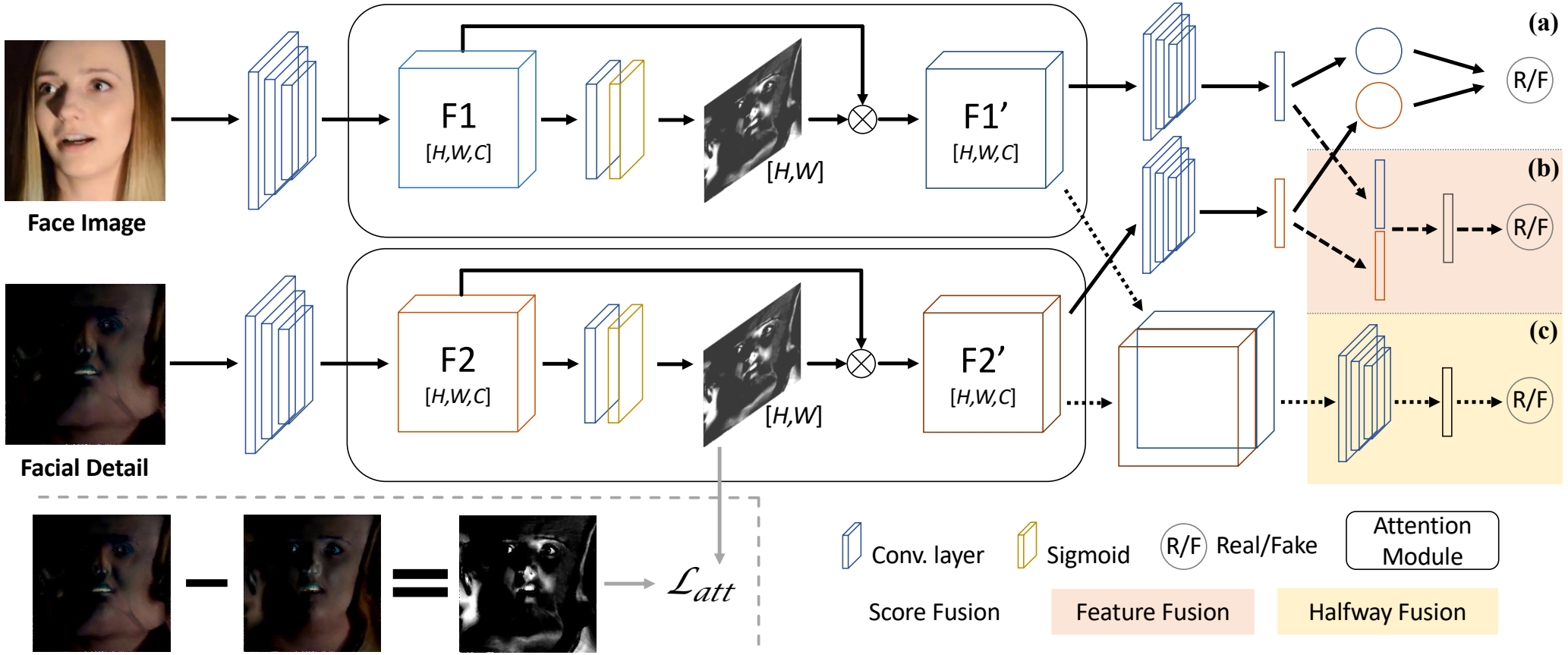
Component Selection





input	3D	amb	dir	ctex	itex	AUC
In-a	✓	✓	✓	✓	✓	99.13
In-b	✓	✓		✓		50.00
In-c	✓		✓		✓	99.29
In-d	✓				✓	99.14
In-e		✓	✓	✓	✓	98.93
In-f		✓		✓		50.00
In-g			✓		✓	**99.56**
In-h					✓	99.27

Forgery Detection with Facial Detail Net



- **Two-stream structure**: regarding **face image** and **facial detail** as two modalities
- **Half-way fusion**: concatenating the **intermediate** 2D feature maps of each stream
- **Detail-guided Attention**: supervised by the **facial detail difference**

Analysis of the Two-stream Network

Structure	FFpp			DFD			DFDC		
	AP	AUC	EER	AP	AUC	EER	AP	AUC	EER
Img	99.44	99.31	5.39	88.07	65.57	38.38	85.60	62.17	39.99
Detail	99.40	99.12	5.51	87.24	64.29	40.87	85.02	61.80	40.37
Img + Detail (HF)	**99.42**	**98.73**	**5.63**	**89.61**	**78.65**	**26.03**	**87.31**	**66.09**	**35.46**

The halfway fusion makes the fused features correspond to similar receptive fields

Further promotes the results on DFD and DFDC by the local fusion manner

Evaluation

- Cross-data

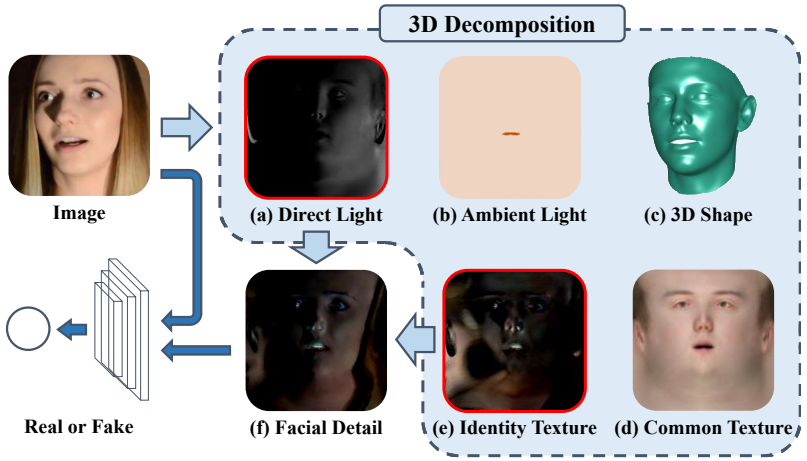
Model	Training dataset	DFD			DFDC		
		AP	AUC	EER	AP	AUC	EER
Xception [50]	FFpp	88.07	65.57	38.38	85.60	62.17	39.99
EfficientNetB4 Ensemble [9]	FFpp	89.35	72.82	34.86	85.71	63.03	38.86
FD^{2}Net	FFpp	**89.84**	**79.08**	**25.18**	**87.93**	**67.70**	**34.91**

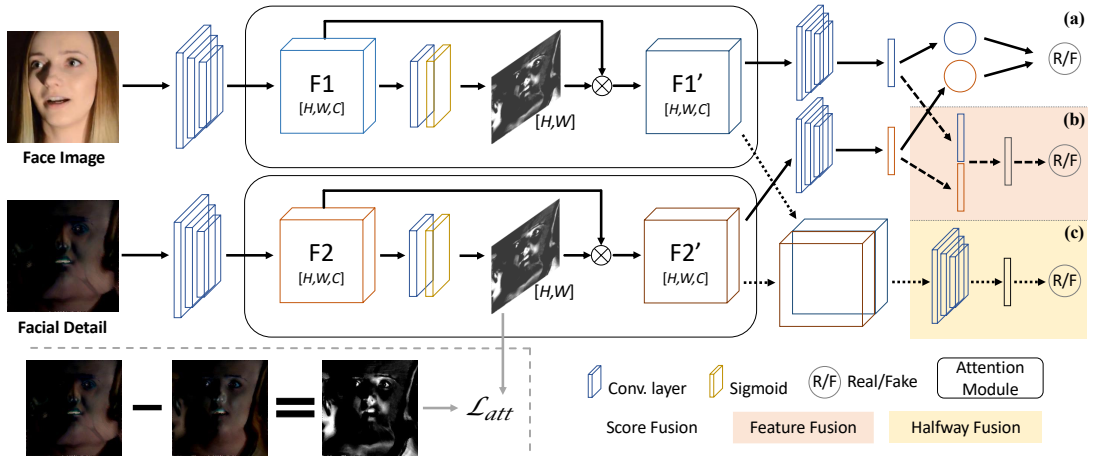
- Different Manipulation Methods

Model	Training data	Acc	
		F2F	FS
MesoInception4 [1]	F2F	84.56	56.71
VA-LogReg [39]		83.62	59.45
LAE [19]		90.34	62.51
Multi-task [40]		91.27	55.04
Face X-ray [36]		97.73	85.69
Xception + HP Filter		97.98	57.46
FD^{2}Net		**98.22**	**86.54**

Effectively extract more discriminative and general features, even from a different distribution of the training dataset

Conclusion





- A novel face forgery detection method by the **3D decomposition** of the face image
  - 3D shape, common texture, identity texture, ambient light, and direct
  - Critical forgery clues in **direct light** and **identity texture**
  - Propose the **facial detail** to highlight the subtle forgery patterns
- **FD^{2}Net**, a two-stream network with facial-detail-supervised attention module
- Improved **effectiveness** and **robustness** of the proposed FD^{2}Net
- A **novel direction** to explore the forgery clues by analyzing the physical generation of an image

Thank you!