Microsoft

CVPR

微软亚洲研究院创研论坛

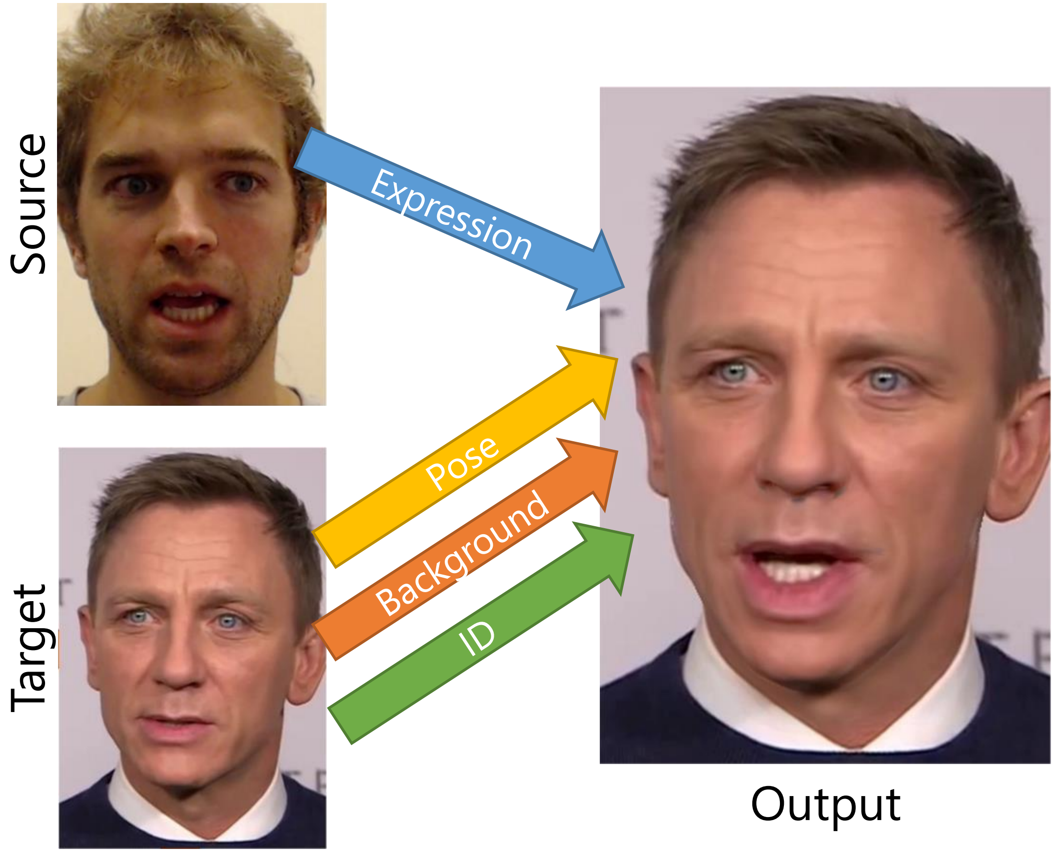
CVPR 2020 论文分享会

Face X-Ray for More General Face Forgery Detection

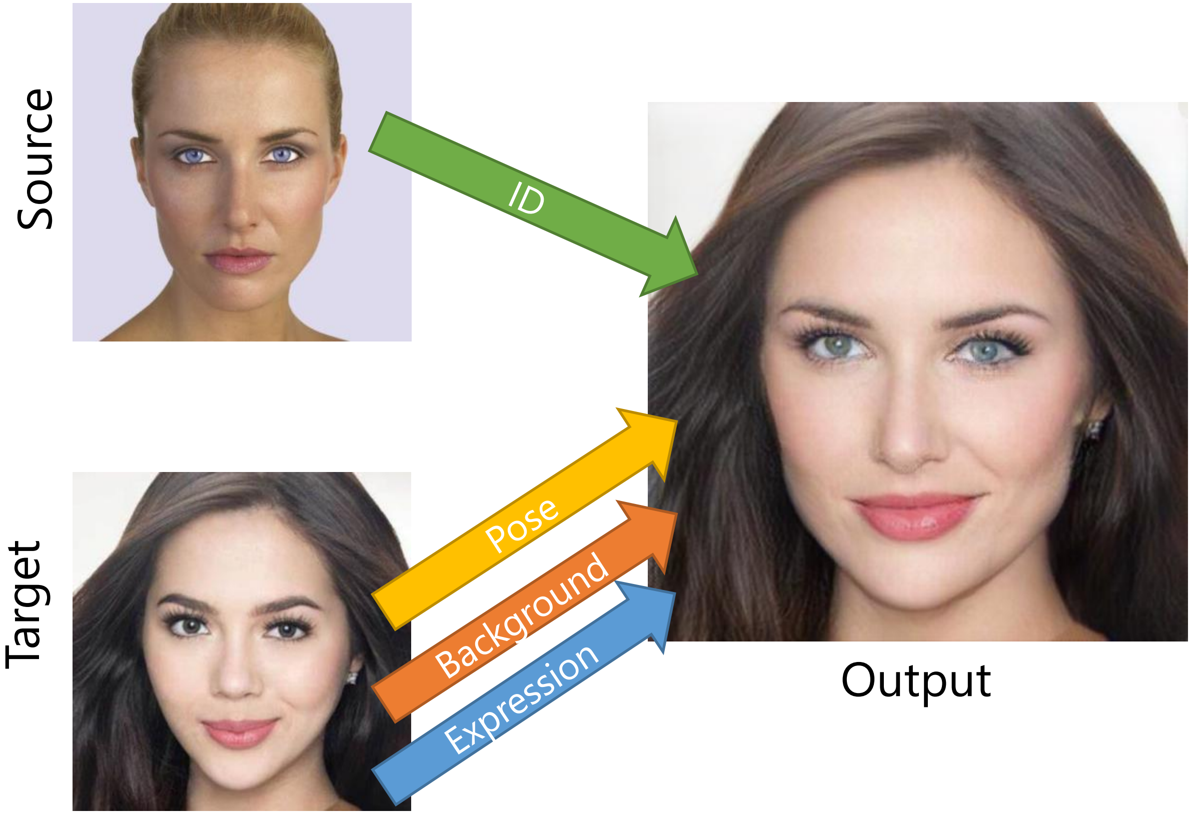
Jianmin Bao
Microsoft Research Asia
CVPR2020 Oral Presentation
Joint work with Lingzhi Li, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo

Face Forgery

Face Reenactment



Face Replacement



*Images of face reenactment are from paper "Face2Face: Real-time Face Capture and Reenactment of RGB Videos"

Deepfakes are eroding our trust

Source: Claire Wardle / New York Times

Fake videos appear real

Fake voices sound real

Thus, real media can't be trusted

Fake media production time	
10 years ago – by studios (CG + DSP)	Today – by anyone (AI/GANs + DSP)
Video: hundreds of hours Audio: dozens of hours	Video: hours Audio: minutes

Face manipulation detection dataset

- FaceForensics++ Datasets :
  - Real: 1000 videos downloaded from the Internet.
  - Fake: 1.5M images from 3000 videos containing 1000 identities.
  - Source: High quality videos, 70% training, 15% validation, 15% test.
  - Four Manipulation methods: DeepFake, FaceSwap, Face2Face, NeuralTexture.


Original
DeepFakes
Face2Face
FaceSwap

Manipulated Face Detection

- Human evaluation.
- Previous State-of-the-art: Binary classifiers for each kind of manipulation, backbone: XceptionNet.

AP	Deepfake	Face2Face	FaceSwap
Human*	77.22%	60.03%	77.25%
FaceForensics++*	98.76%	98.59%	98.53%

*The results are from paper: "Rössler, Andreas, et al. Faceforensics++: Learning to detect manipulated facial images, ICCV, 2019."

Has this problem been solved?

No!

Generalization Dilemma

- The naive baseline real/fake classifier failed to generalize to unknown face manipulation algorithms
- We present Face X-ray to tackle the generalization dilemma on unknown face manipulation algorithms

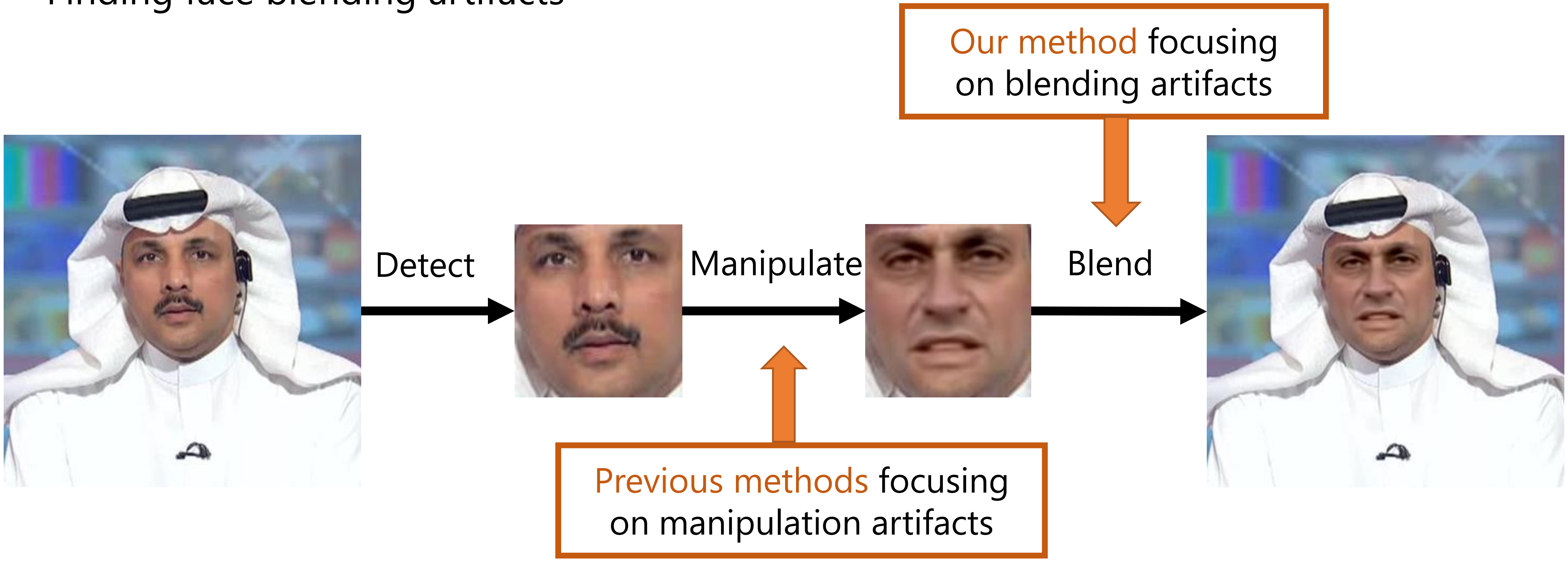
Training set			Binary classification (AP)		
Deepfake	Face2Face	FaceSwap	Deepfake	Face2Face	FaceSwap
√			99.13%	70.99%	51.12%
	√		82.25%	98.91%	63.82%
		√	65.40%	58.90%	99.20%

How we deal with the Generalization Dilemma?

Our idea: Looking for common defects

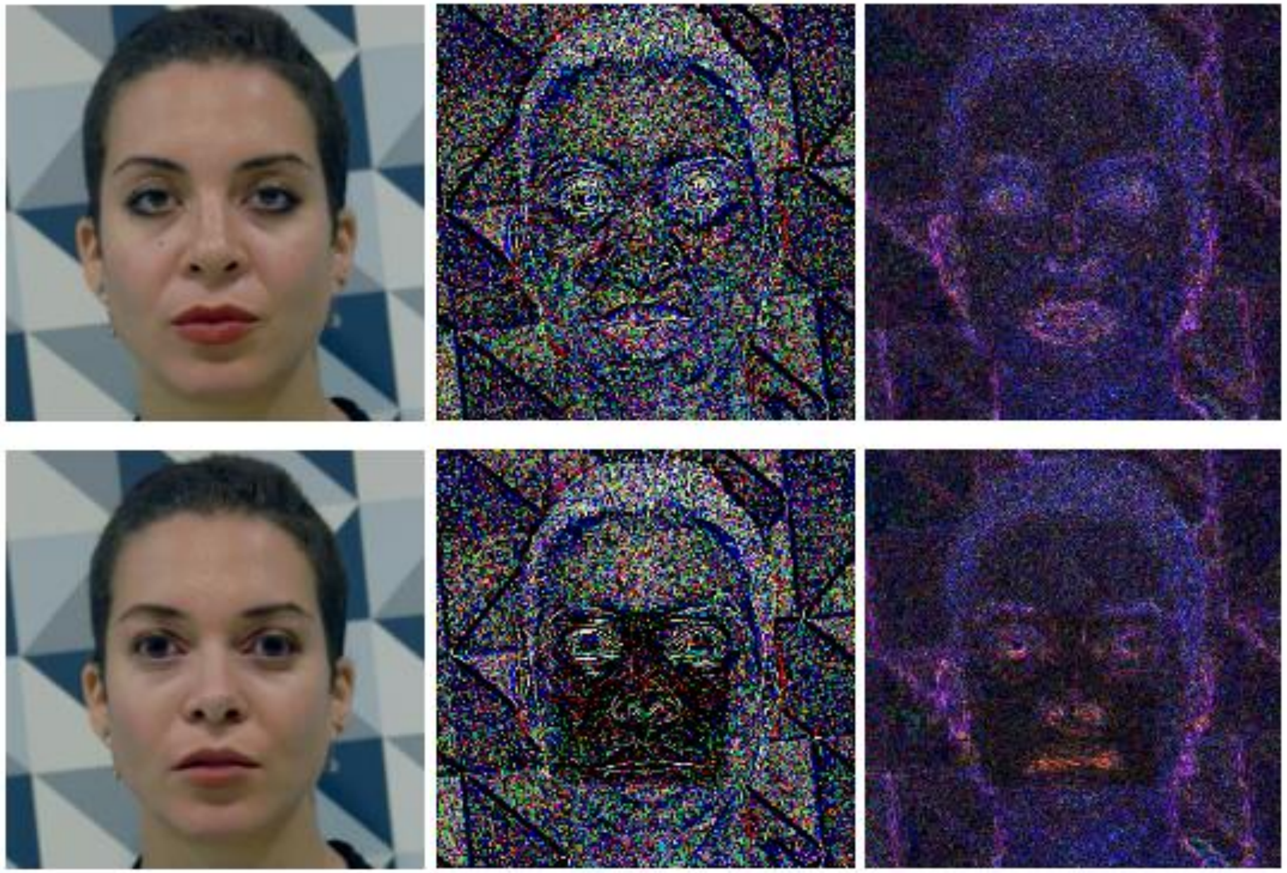
Our observation

- Face manipulation methods only generate inner face region
- Image blending exists in almost all face manipulation methods
- Finding face blending artifacts



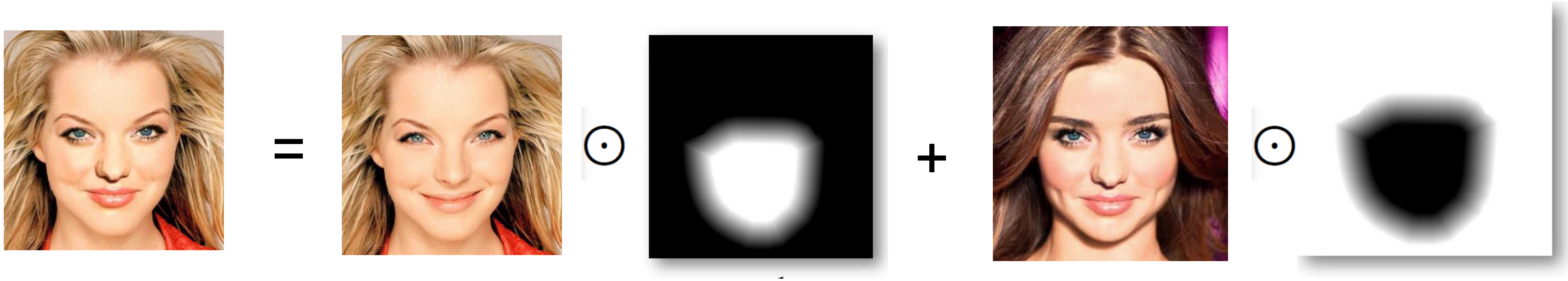
What is Blending Artifacts

- Each image has its own distinctive marks
- Introduced from hardware (e.g., sensor, lens) or software (e.g., compression, synthesis algorithm)
- There is the strong possibility that some marks change across the boundary.



Blending boundary definition

- A typical Blending Step



- We define blending boundary as $(1 - mask) \times mask \times 4$

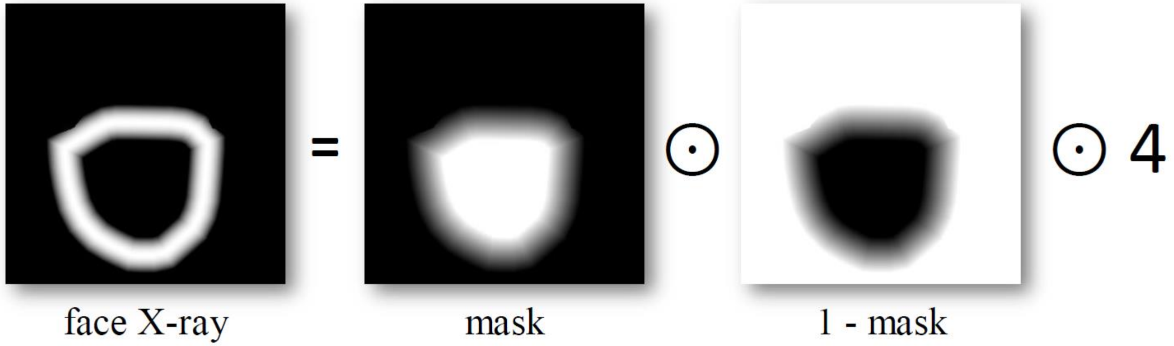
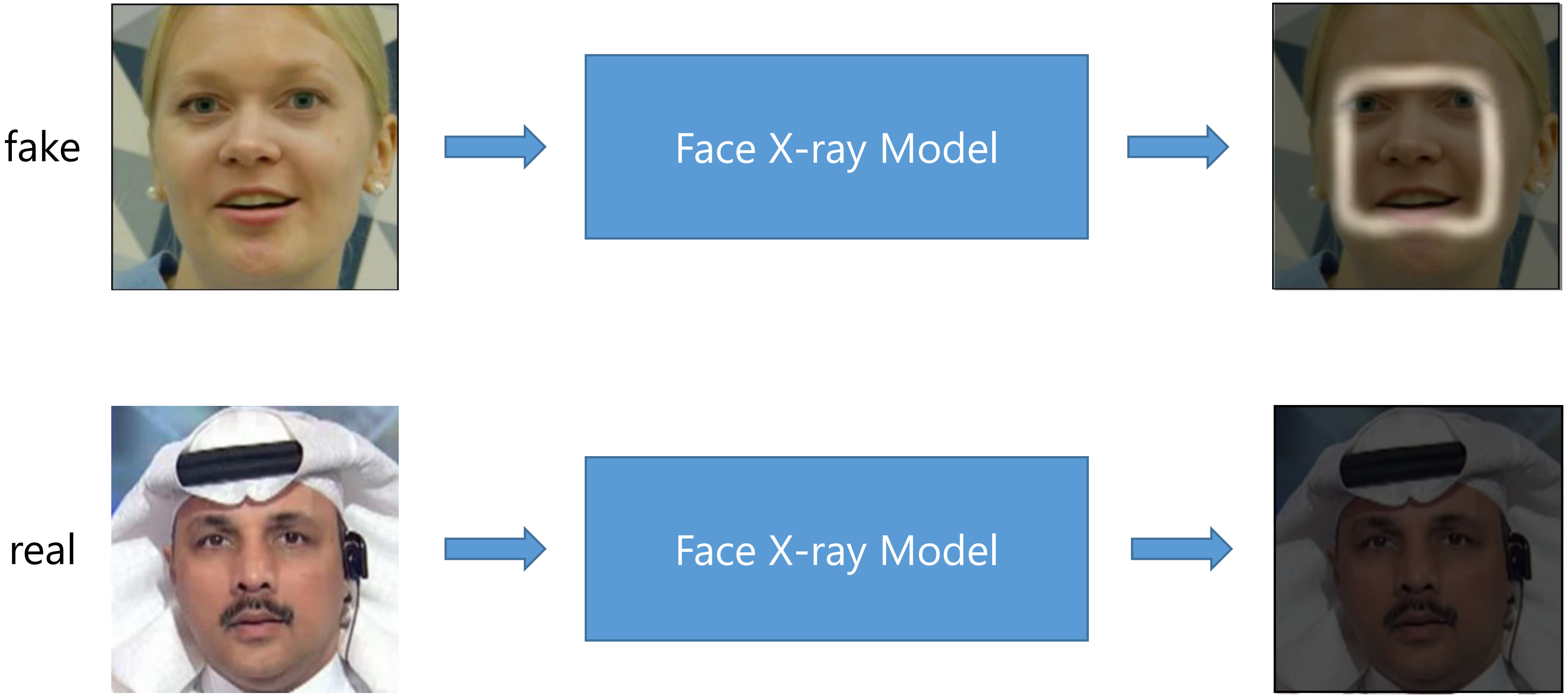


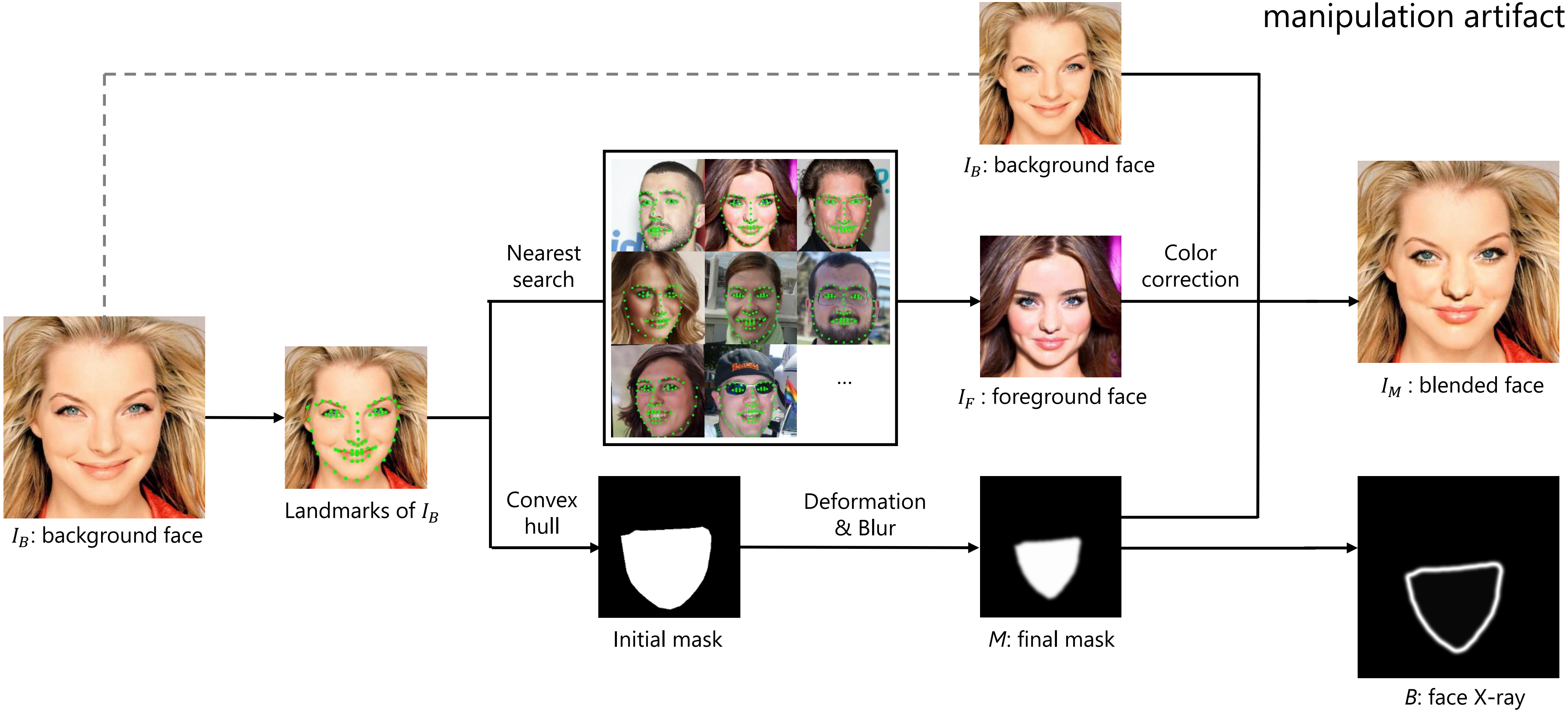
Illustration of our methods: Face X-ray

- Detecting blending boundary:

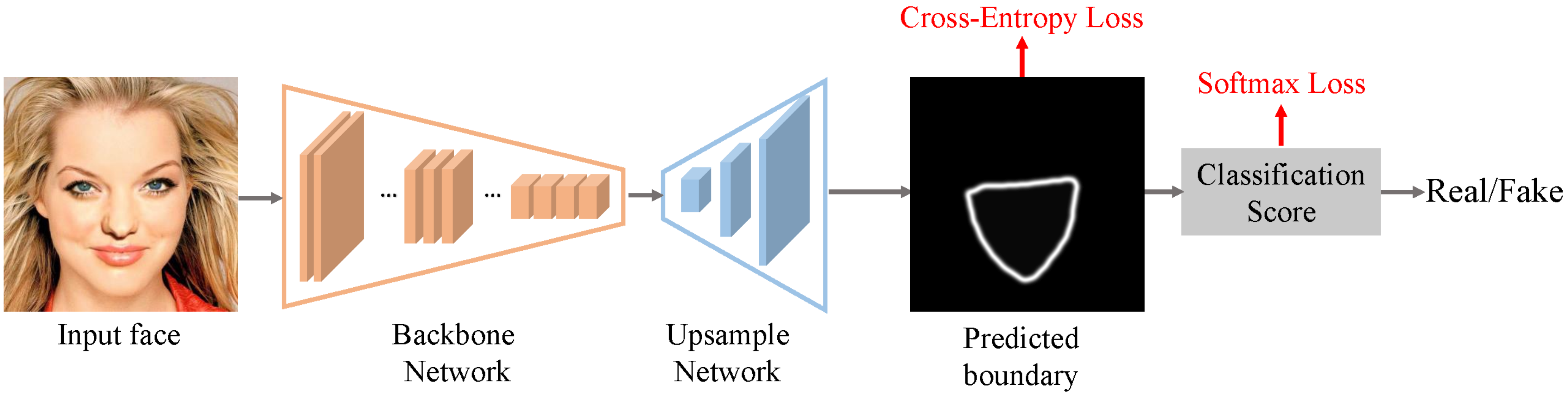


Training In A Self-Supervied Way

- Only utilize real images
- No other face manipulation data is required.
- Avoid falling into specific manipulation artifact



Our proposed framework



Experimental Results

Generalization Capability

- Training and testing with fake faces generated by different algorithms.

Model	Training set		Test set AUC				
	FS	BI	FS	DF	F2F	NT	ALL
Xception	√	-	99.36	87.56	61.70	68.71	74.91
HRNet	√	-	99.24	83.64	64.12	68.89	73.96
Face X-ray	√	-	99.20	98.52	92.29	86.63	93.13
	√	√	99.09	99.03	98.16	96.66	98.25
	-	√	99.21	99.17	**98.57**	**98.13**	**98.52**

Classification model has a terrible cross-algorithms generalization ability

Face X-ray can improve generalization performance significantly : △ **19.17**

Real-face based generated data further improve generalization performance : △ **5.12**

Comparable results with only the Blended image with real faces

Comparison with SOTA

- Our methods outperform SOTA methods by large margins.

	Training Set		**Detection accuracies**	
	F2F	FS	F2F	FS
LAE	√	-	90.93	63.15
FT-res	√	4 images	94.47	72.57
MTDS	√	-	92.77	54.07
Face X-ray	√	-	**97.73**	**85.69**

Ablation Study

- Training data(image and Face X-ray pair) generation with different settings.

	AUC	
	FF++	DFD
w/o mask deformation	93.92	85.89
w/o color correction	96.21	89.91
Face X-ray	**98.52**	**93.47**

Ablation Study

- Omethod remain effect to possion blending and a learning-based blending.

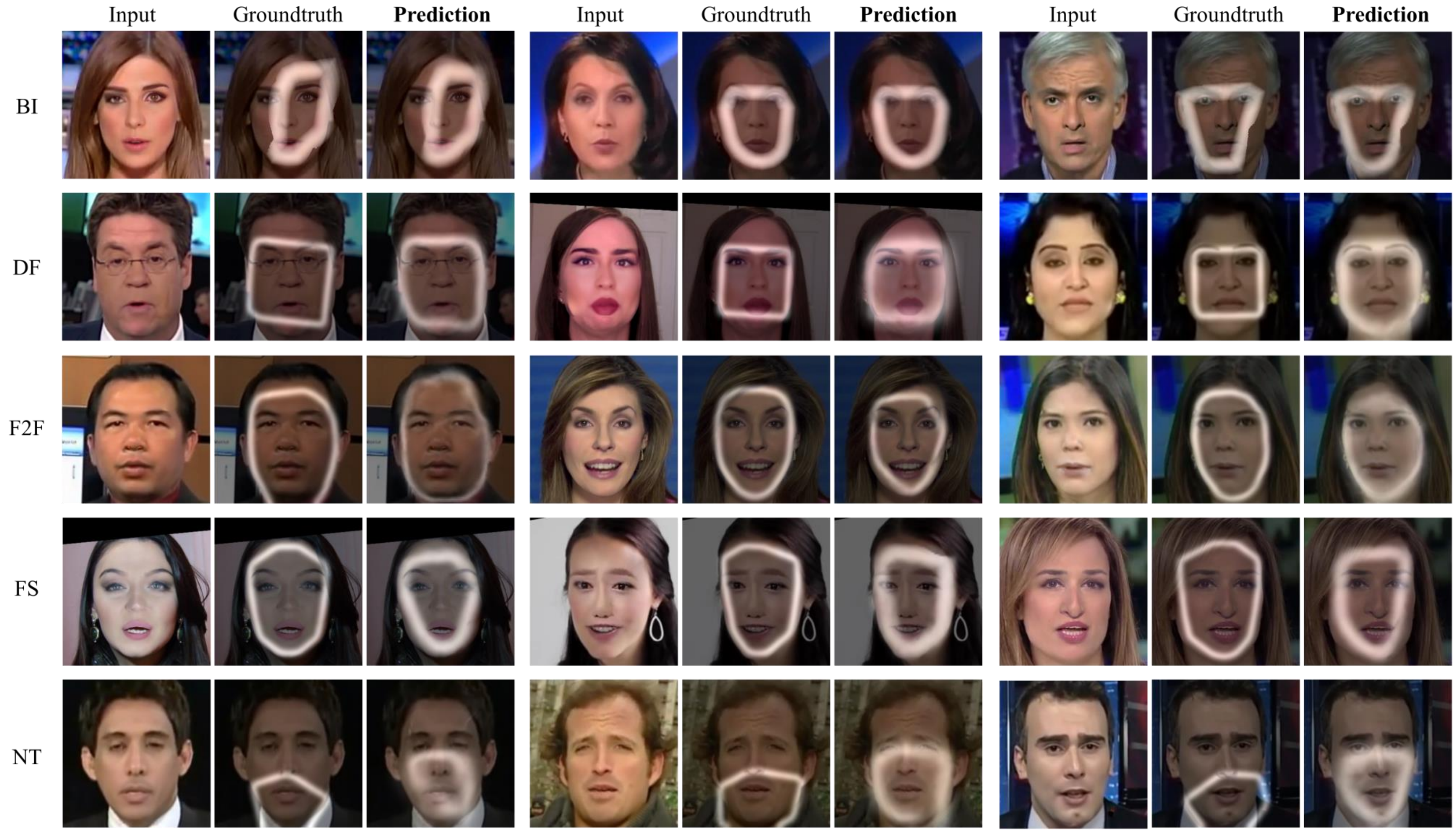
Blending type	AUC	AP	EER
Alpha blending	99.46	98.50	1.50
Possion blending	94.62	88.85	11.41
Deep Blending	99.90	98.77	1.36

Cross dataset results

- Training on FaceForensics++ dataset
- Testing on DFD, DFDC and Celeb-DF dataset

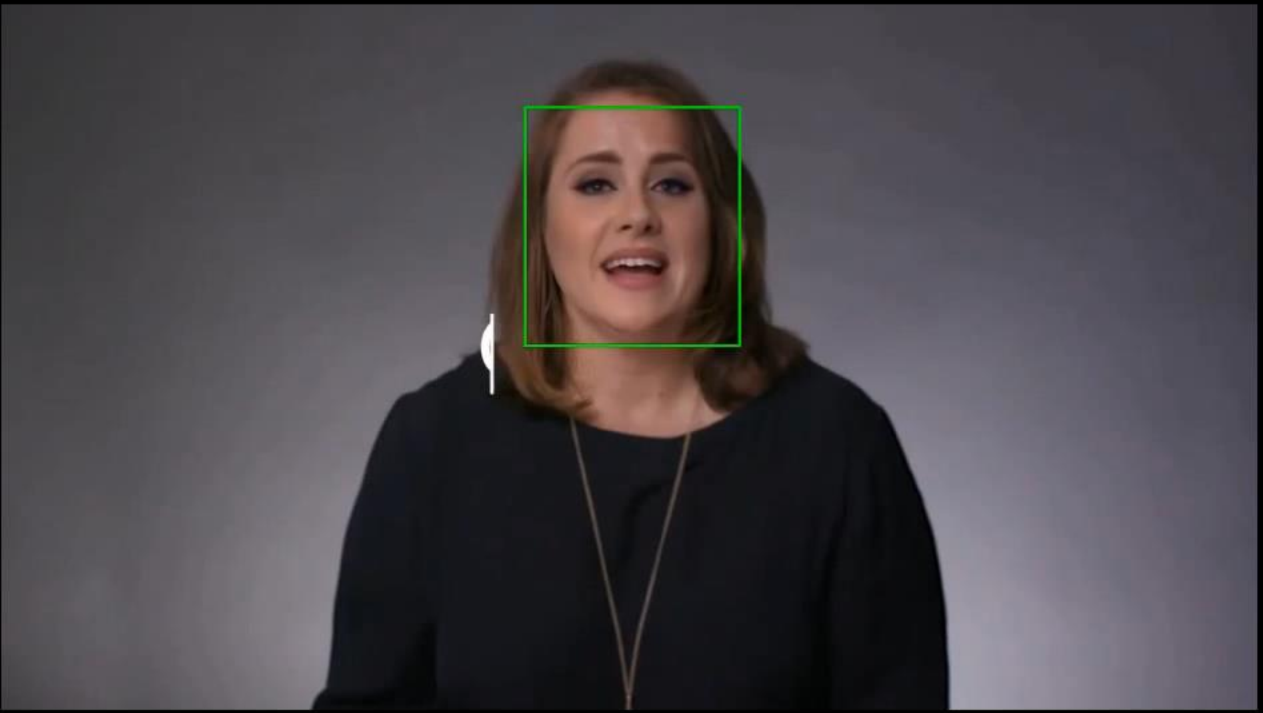
AP	DFD	DFDC	Celeb-DF
Binary classification	78.82%	50.83%	50.07%
Face X-ray	**93.34%**	**72.65%**	**73.33%**

Predicted blending boundary

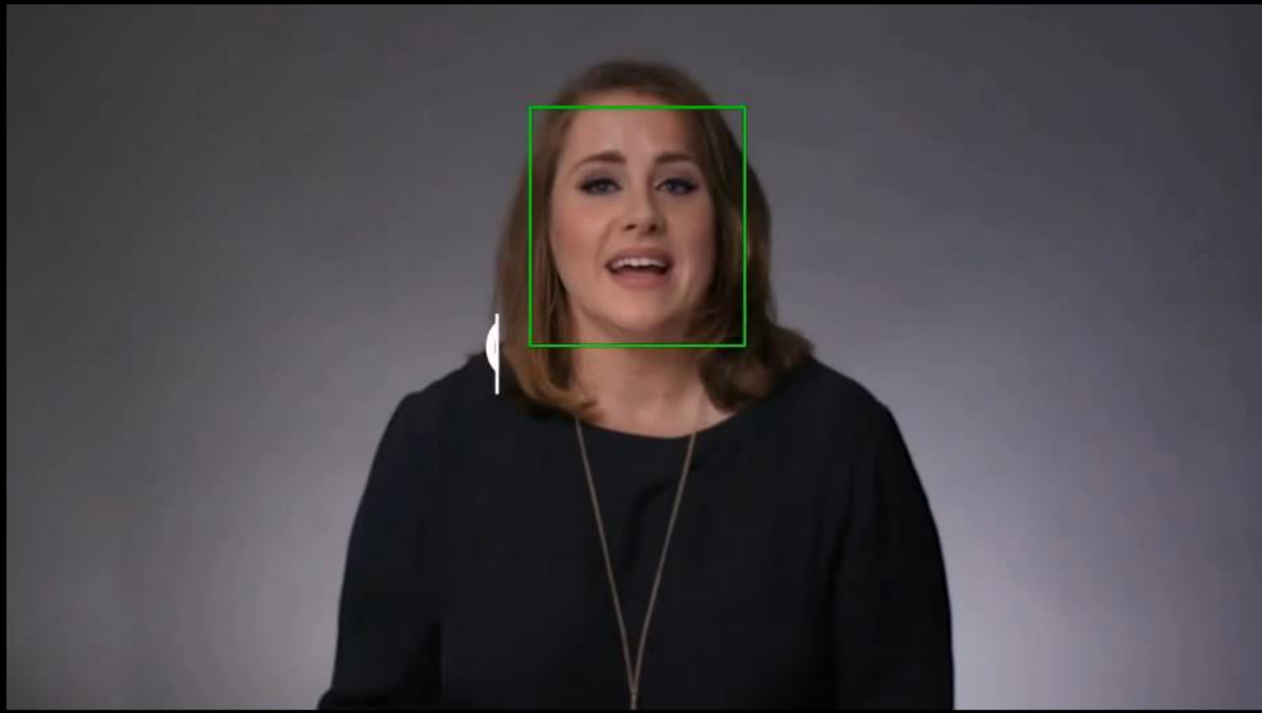


green box: real, red box: fake

FaceForensics++



Our predictions



Limitations

- Cat-mouse game
  - An image is entirely synthesis.
  - Adversarial samples to against our detector.
- Image/video compression
  - Suffer from performance drop when encounter low resolution Images.

AUC	No compression	Light compression	High compression
Face X-ray	98.52%	87.35%	61.6%

Conclusion

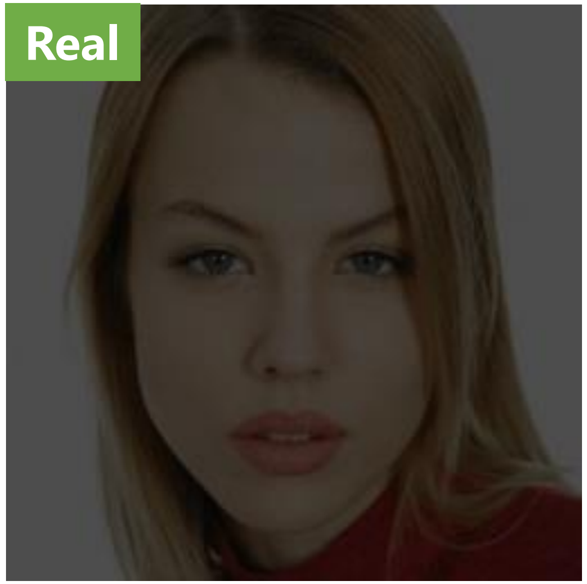
- Face X-ray: a novel framework for more general manipulated face detection.
- Our method not only distinguishes whether an image is forged but also identifies the location where two images are blending with each other.
- We train our framework in a self-supervised way that only utilize real images, making our model more robust and generalizable.

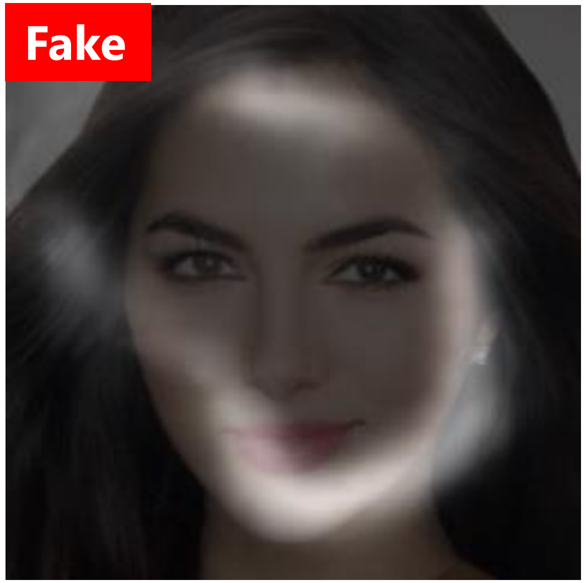
Real or Fake?




*Real images and source images are from https://www.bing.com. Fake Images are generated by our algorithm.

Real or Fake?





*Real images and source images are from https://www.bing.com. Fake Images are generated by our algorithm.

谢谢观看

THANK YOU