

Face Presentation Attacks: Physical and Digital

Prof. Xiaoming Liu

Michigan State University



MICHIGAN STATE UNIVERSITY

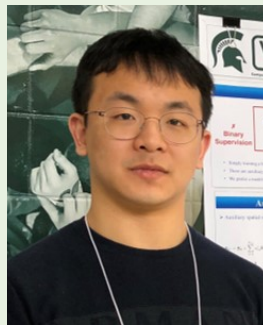


Computer Vision Lab

Acknowledgement



Dr. Anil Jain



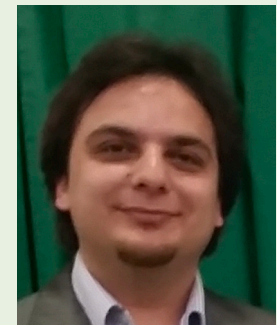
Yaojie Liu



Joel Stehouwer



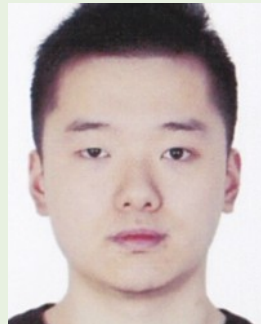
Dr. Amin Jourabloo



Dr. Yousef Atoum



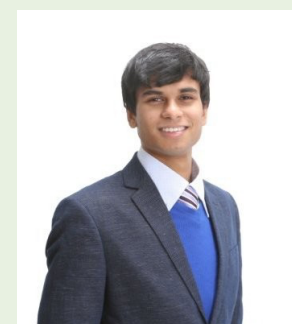
Dr. Feng Liu



Xiaohong Liu



Vishal Asnani



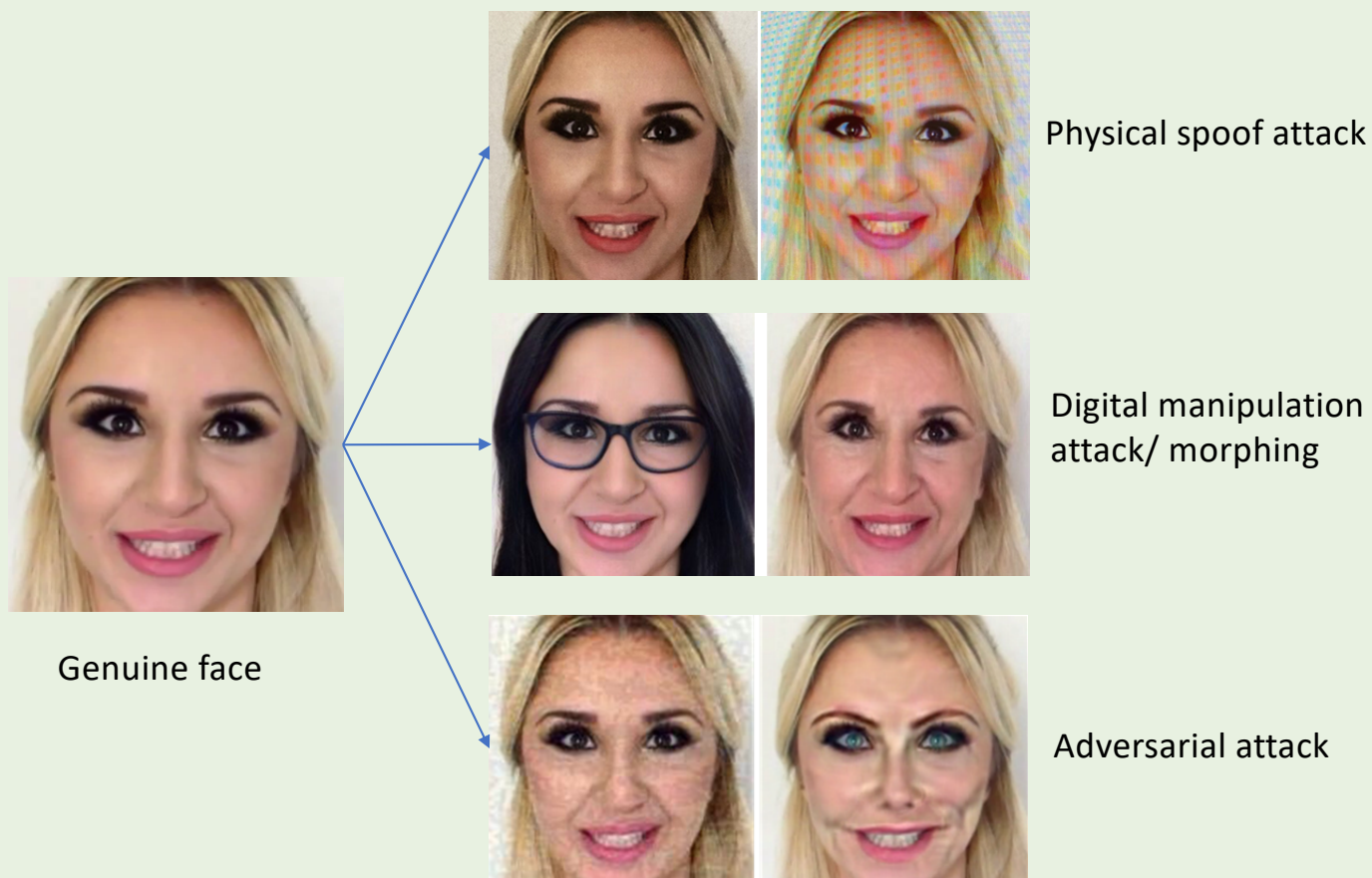
Debayan Deb

Acknowledgement

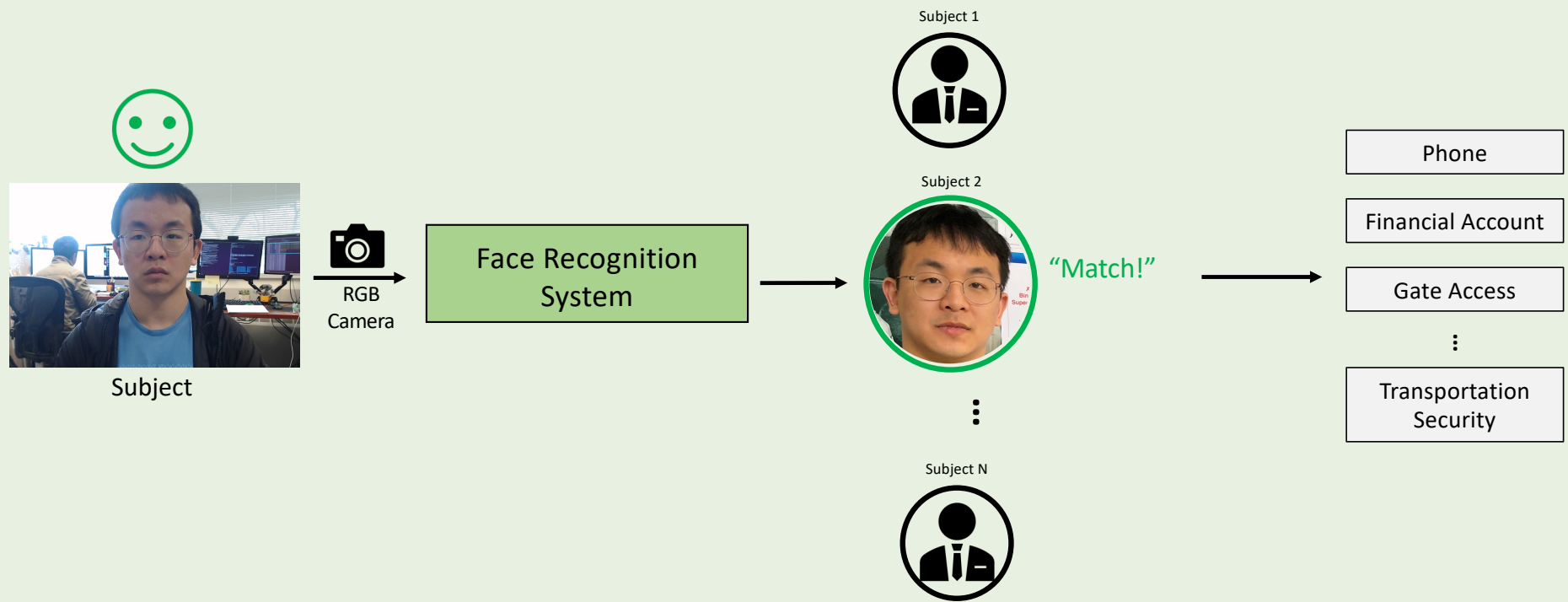
The research of face presentation attack detection is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.



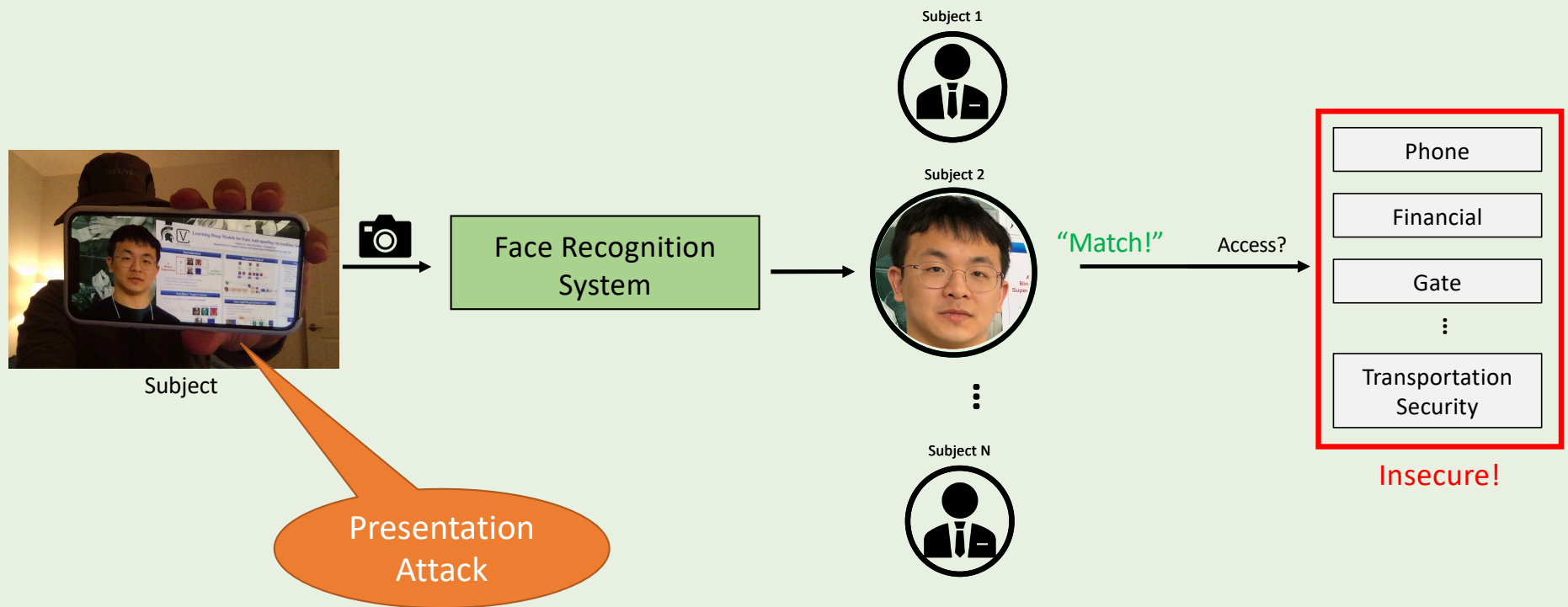
Face Attacks



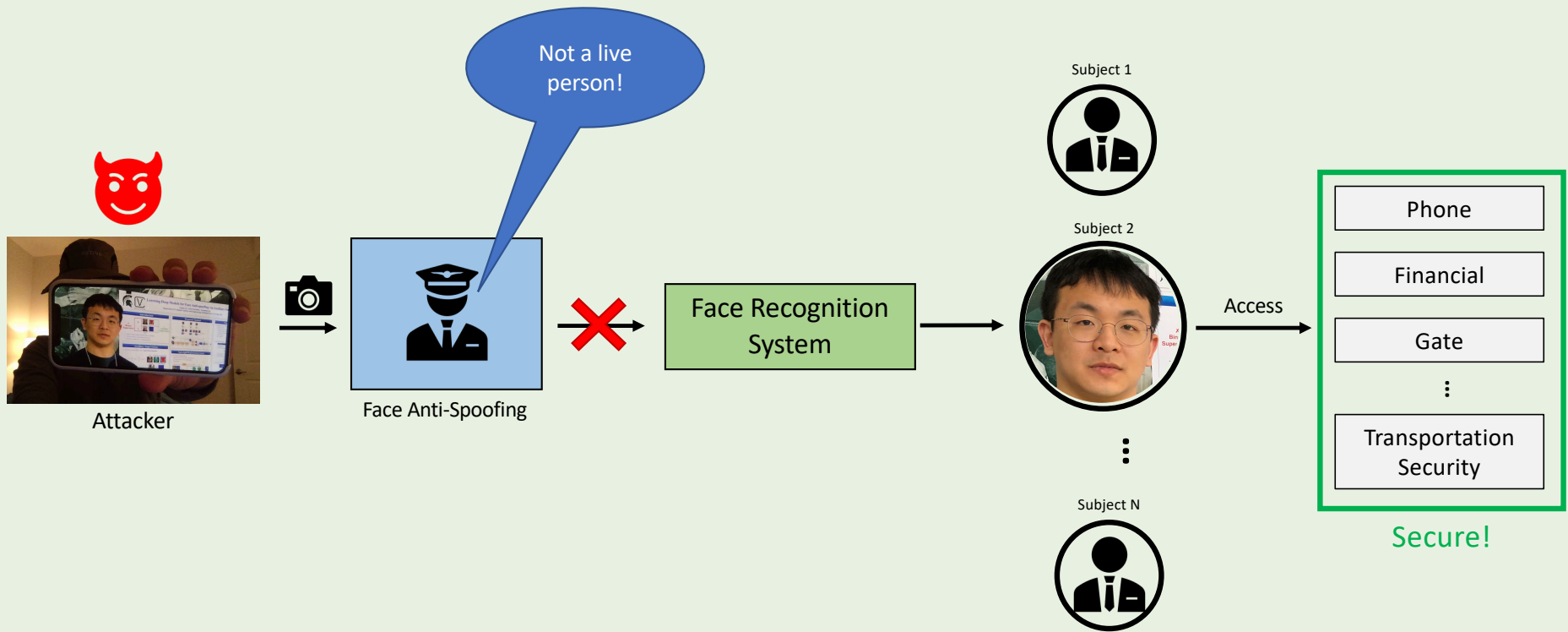
A General Face Recognition Flow



Is This Secure?

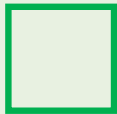


Face Anti-Spoofing

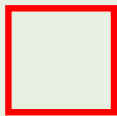


Face Anti-Spoofing

- Face anti-spoofing is the technique to distinguish the source of a face: whether it's coming from a **real human** or from a **spoofing material**

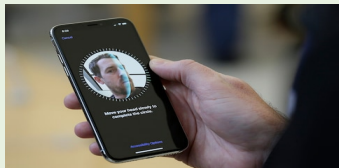
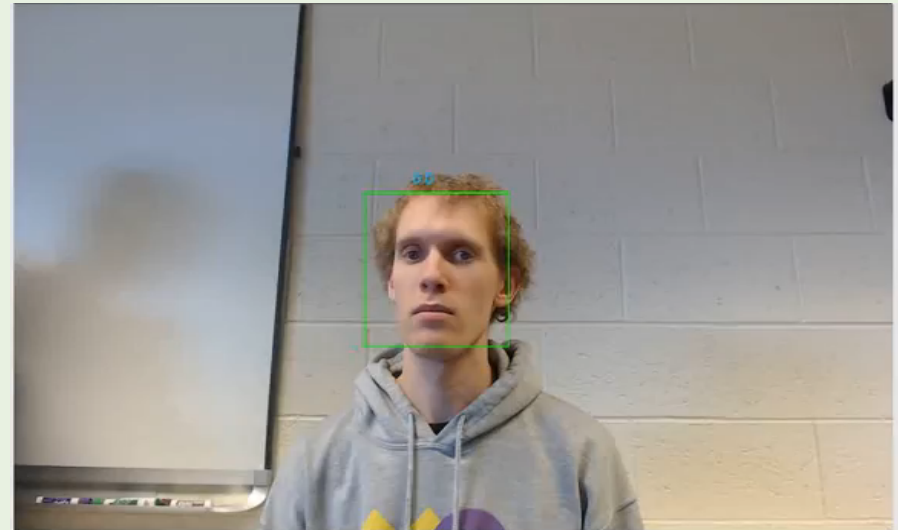


Detected face is live

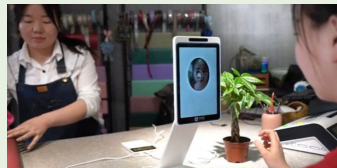


Detected face is spoof

- Finding a monocular RGB camera-based solution



Phone Unlock



Quick Purchase



Border Control



Building Access

- <https://www.seattletimes.com/business/technology/should-you-buy-an-iphone-x-only-if-you-like-to-hold-the-future-in-your-hand/>
- <http://xue.haishua.com/18454.html>
- <https://www.donga.com/news/inter/article/all/20181206/93178774/1>
- <https://www.jetwayipc.com/product/frg1-3288s/>



The Development

- Interaction-based methods (2006-2010)
- Texture-based methods (2010-2017)
- Deep-learning-based methods (2017-)

Texture-based Methods

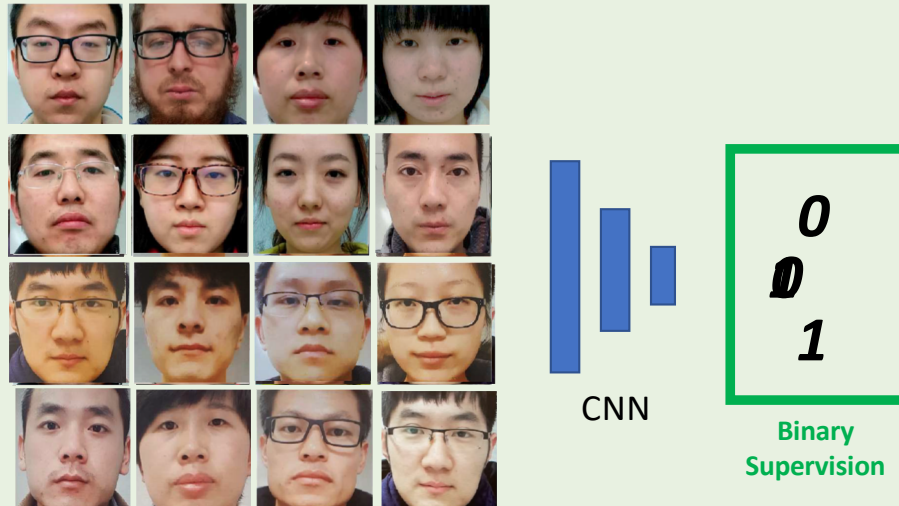
- J. Maatta, et. al., Face Spoofing Detection from Single Images using Micro-Texture Analysis, *IJCB*, 2011.
- J. Galbally, et. al., Face Anti-Spoofing Based on General Image Quality Assessment, *ICPR*, 2014.
- Z. Boulkenafet, et. al., Face Anti-Spoofing Based on Color Texture Analysis, *ICIP*, 2015
- S. Liu, et. al., 3D Mask Face Anti-Spoofing with Remote Photoplethysmography, *ECCV*, 2016.
- Z. Boulkenafet, et. al., Face Anti-Spoofing Using Speeded Up Robust Features and Fisher Vector Encoding, *IEEE Signal Processing Letters*, 2017.
- A. Agarwal, et. al., Face anti-spoofing using Haralick features, *BTAS*, 2016.
- K. Patel, et. al., Secure face unlock: Spoof detection on smartphones, *TIFS*, 2016.
- K. Patel, et. al., Live face video vs. spoof face video: Use of moire patterns to detect replay video attacks, *ICB*, 2015.

The Development

- Interaction-based methods (2006-2010)
- Texture-based methods (2010-2017)
- **Deep-learning-based methods (2017-2020)**

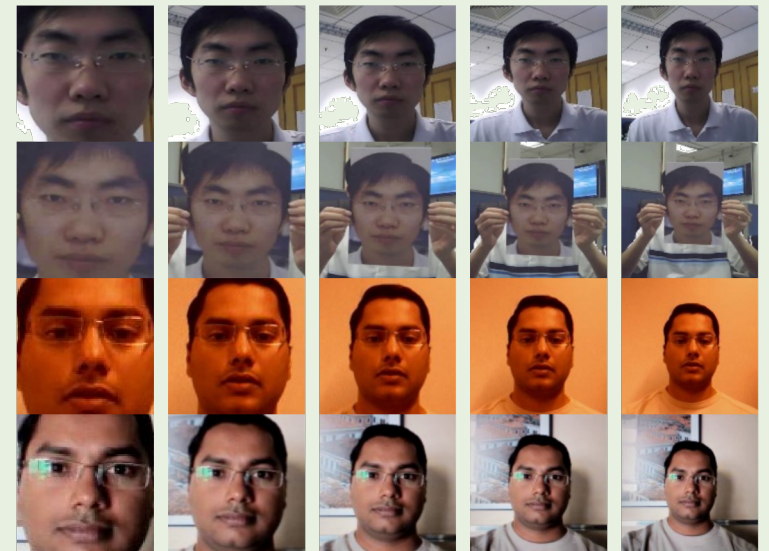
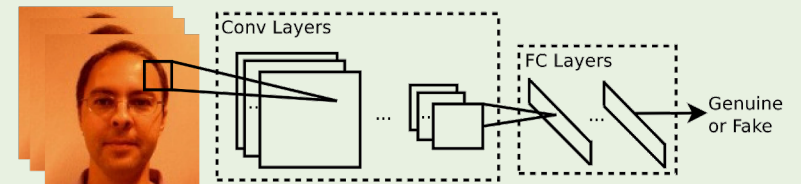
Direct FAS

- CNN is trained to do a binary classification: live vs spoof



Direct FAS

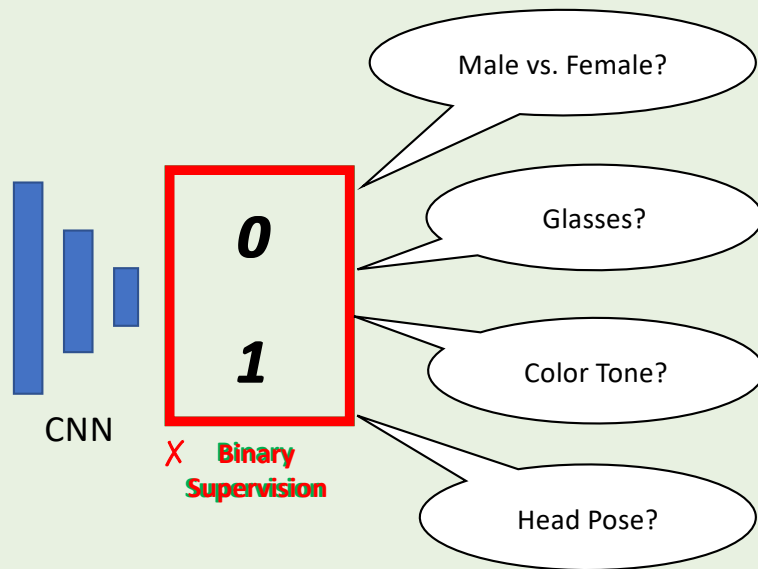
- MLP / CNN feature + SVM classifier
- Search different input to improve performance
 - Features (LBP, IQM)
 - Face scales
 - Color spaces (RGB, HSV, YCbCr)



1. Yang et. al., Learn Convolutional Neural Network for Face Anti-Spoofing. arXiv 2014.
2. Xu et. al., Learning temporal features using LSTM-CNN architecture for face anti-spoofing. ACPR 2015.

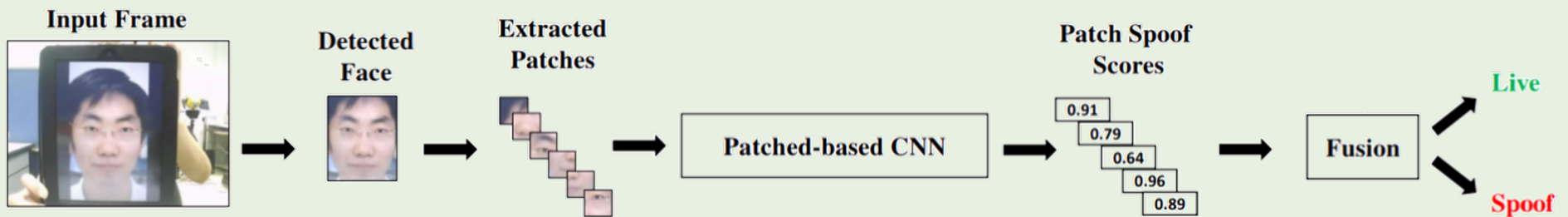


Drawbacks



Patch-based CNNs

- CNN is trained to do a binary classification for each face patch

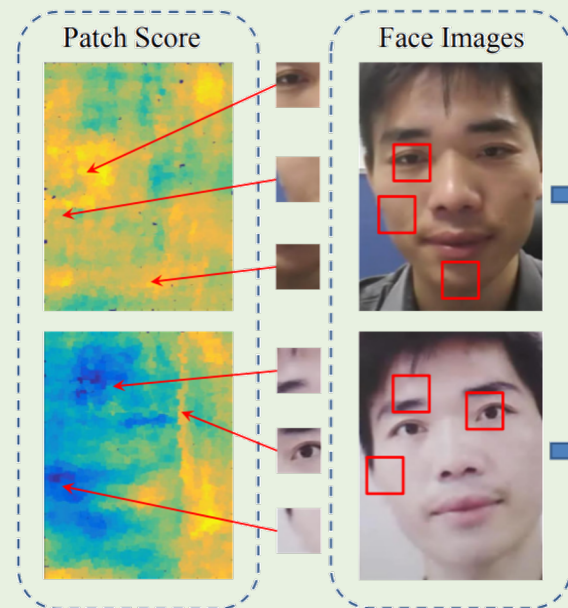


1. Yousef Atoum et. al., Face Anti-Spoofing Using Patch and Depth-Based CNNs, IJCB, 2017
2. Gustavo Botelho de Souza et. al., On the Learning of Deep Local Features for Robust Face Spoofing Detection, SIBGRAPI, 2018
3. Xiao Yang et. al., Face Anti-Spoofing: Model Matters, So Does Data, CVPR, 2019
4. DRL-FAS: A Novel Framework Based on Deep Reinforcement Learning for Face Anti-Spoofing, arXiv, 2020
5. Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach, arXiv, 2020



Patch-based CNNs

- Benefits
 - Mitigate overfitting (+ training samples)
 - Light-weight network
- Challenges
 - All patches matter?

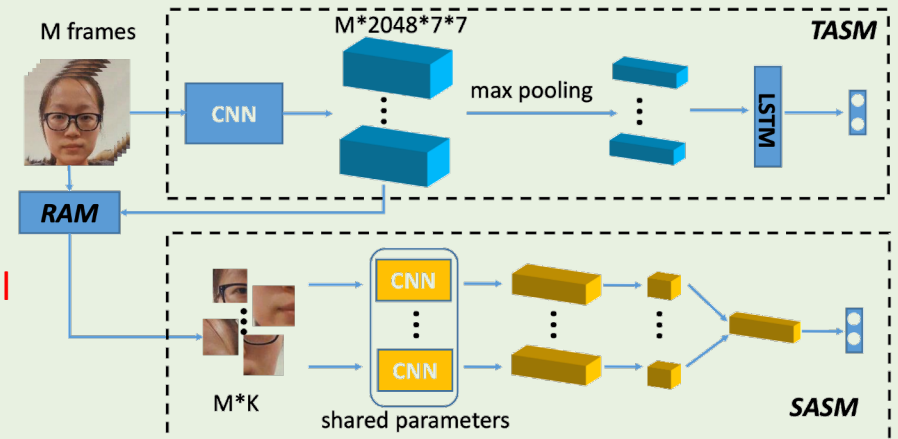


1. Yousef Atoum et. al., Face Anti-Spoofing Using Patch and Depth-Based CNNs, IJCB, 2017



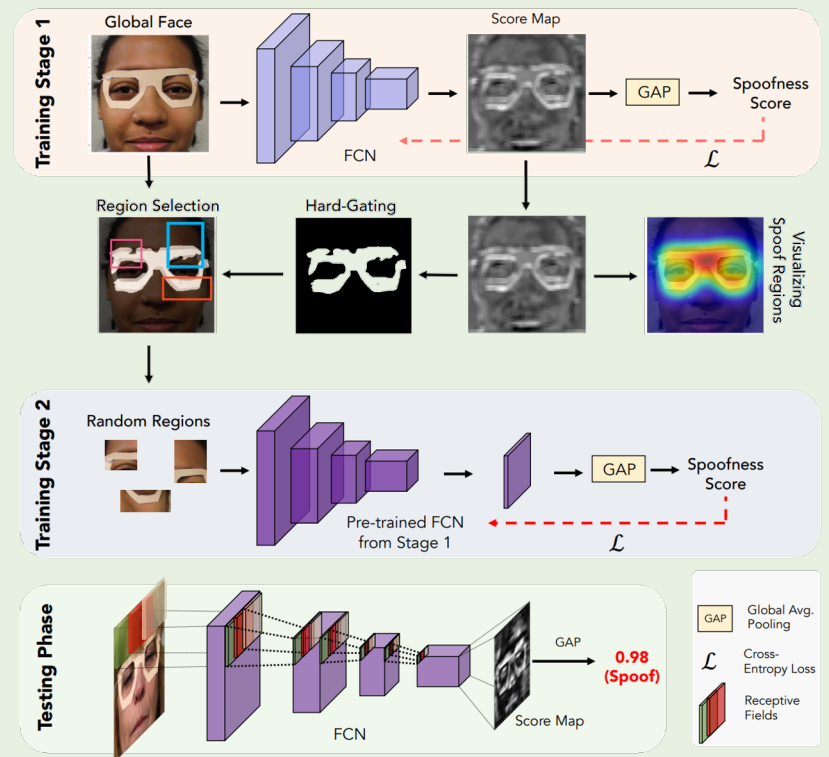
Patch-based CNNs

- Benefits
 - Mitigate overfitting (+ training samples)
 - Light-weight network
- Challenges
 - **Not all patches matter**
Select more important patches based on global information



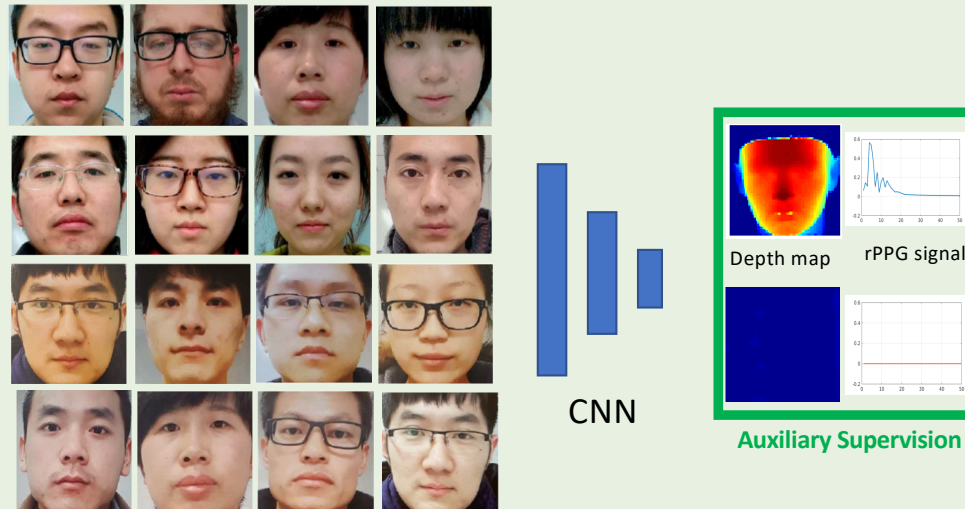
Patch-based CNNs

- Benefits
 - Mitigate overfitting (+ training samples)
 - Light-weight network
- Challenges
 - Not all patches matter
Select more important patches based on global information



Auxiliary FAS

- CNN is trained to do auxiliary tasks, which can help face anti-spoofing

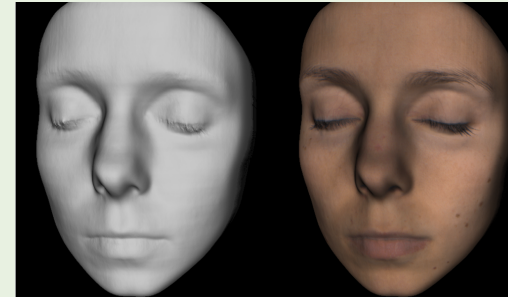


1. Face anti-spoofing using patch and depth-based CNNs. IJCB 2017.
2. Learning deep models for face anti-spoofing: binary or auxiliary supervision. CVPR 2018.
3. Face de-spoofing: anti-spoofing via noise modeling. ECCV 2018.
4. Exploiting temporal and depth information for multi-frame face anti-spoofing, arXiv 2019
5. Aurora guard: real-time face anti-spoofing via light reflection, arXiv 2019
6. Meta Anti-spoofing: Learning to Learn in Face Anti-spoofing, arXiv 2019
7. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. CVPR 2019
8. Deep tree learning for zero-shot face anti-spoofing. CVPR 2019



Facial Depth Estimation

- Faces have rich 3D information
- Spoof attacks are more flat
- Good depth estimation → Good anti-spoofing
- Still about learning texture features



Facial depth



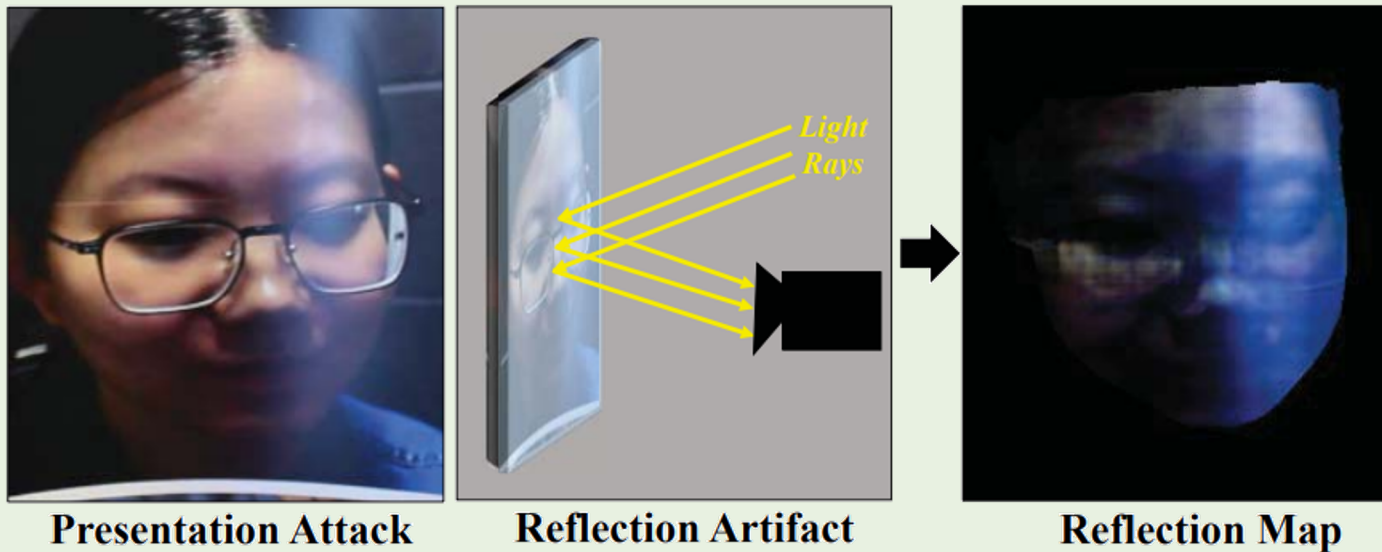
Flat Surface

1. Figure 1 source: <https://www.spiedigitallibrary.org/journals/journal-of-biomedical-optics/volume-24/issue-06/066002/Three-dimensional-maps-of-human-skin-properties-on-full-face/10.1117/1.JBO.24.6.066002.full?SSO=1>
2. Figure 2 source: <https://www.pinterest.com/pin/59250551334974173/>



Reflection Estimation

- Can CNN learn specific tasks that contain anti-spoofing information?

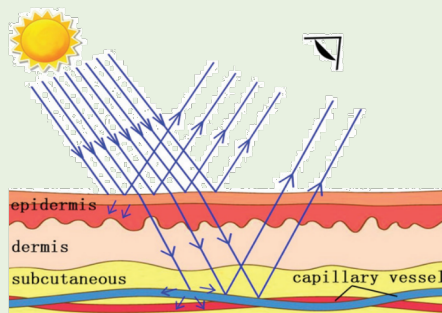


1. T. Kim. BASN: Enriching Feature Representation Using Bipartite Auxiliary Supervisions for Face Anti-Spoofing. *ICCVW 2019*
2. Z. Yu, et. al., Face Anti-Spoofing with Human Material Perception, *ECCV 2020*

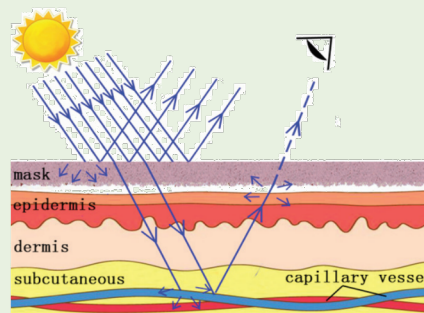


rPPG Estimation

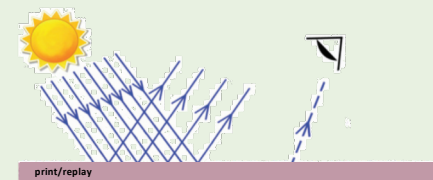
- Remote photoplethysmography: heartbeat measurement from human skin using a non-contact camera



Live Face



3D Mask Spoof Face



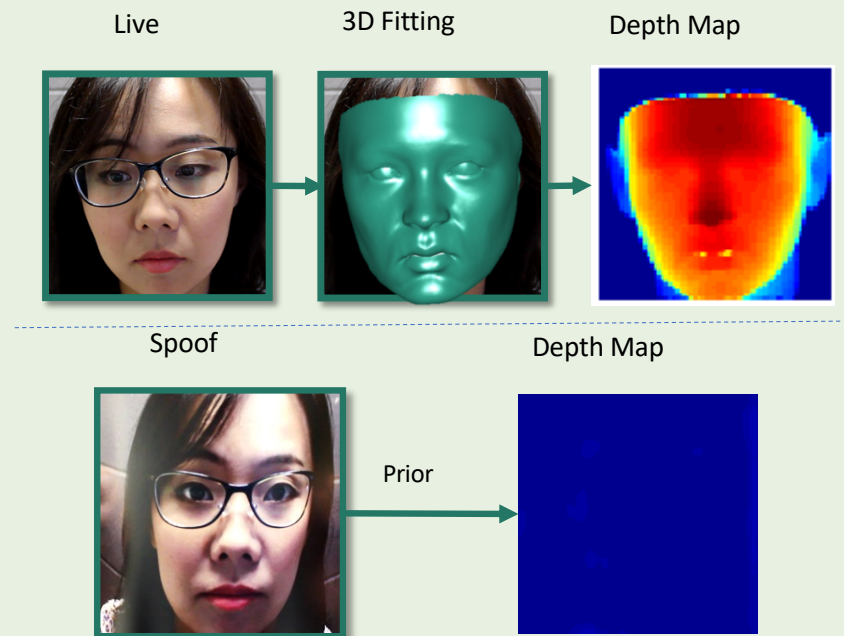
Print/Replay Spoof Face

1. Liu et. al., Learning deep models for face anti-spoofing: binary or auxiliary supervision, CVPR 2018.
2. Liu et. al., 3D Mask Face Anti-spoofing with Remote Photoplethysmography, ECCV 2016
3. Liu et. al., Remote Photoplethysmography Correspondence Feature for 3D Mask Face Presentation Attack Detection, ECCV 2018



How to Obtain Label?

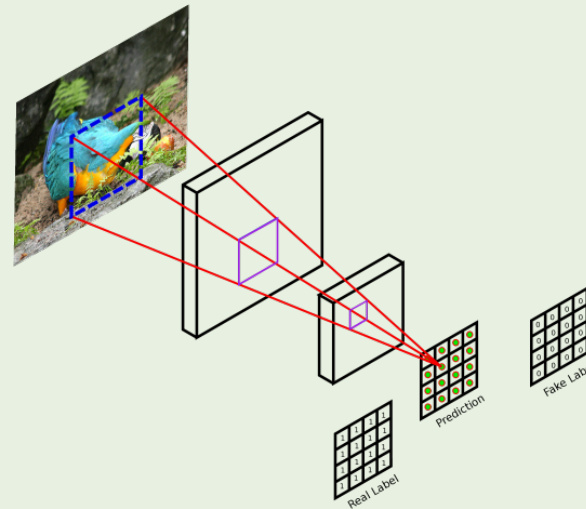
- Depth map for example
 - live faces: 3D face fitting* + z-buffering rendering
 - spoof faces: zero maps



1. Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *CVPR 2018*
2. Y. Liu, A. Jourabloo, W. Ren, and X. Liu. Dense Face Alignment. *ICCVW 2017*.

Why It Works?

- Local responses
- Multi-scale features
- Addition knowledge ($> 0/1$ map)

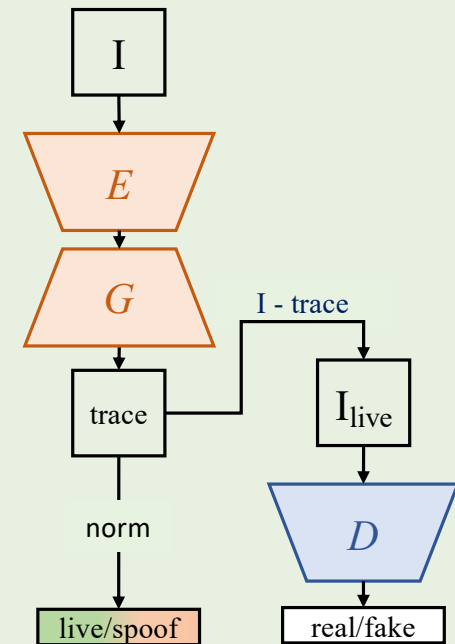


1. Revisiting Pixel-Wise Supervision for Face Anti-Spoofing, arXiv, 2020



Generative FAS

- CNN is trained to generate some type of image to extract FAS feature
- Generate:
 - Data augmentation
 - Spoof trace



1. Y. Liu, et. al. "On Disentangling Spoof Traces for Generic Face Anti-Spoofing", ECCV 2020
2. A. Jourabloo, et. al. "Face De-Spoofing: Anti-Spoofing via Noise Modeling", ECCV 2018
3. K. Zhang, et. al., "Face Anti-Spoofing via Disentangled Representation Learning", ECCV 2020
4. J. Stehouwer, et. al., "Noise Modeling, Synthesis and Classification for Generic Object Anti-Spoofing", CVPR 2020
5. H. Feng, et. al., "Learning Generalized Spoof Cues for Face Anti-spoofing", arXiv, 2020
6. Y. Liu, et. al. "Physics-Guided Spoof Trace Disentanglement for Generic Face Anti-Spoofing", under PAMI review



FAS Generalization

The testing scenarios are different with the training phase.

- Environment (Lighting, Indoor/outdoor, etc.)
- Camera/Image quality
- Subjects (Age, Race, etc.)
- Spoof types

Training-Testing Difference

The testing scenarios are different with the training phase.

- Environment (Lighting, Indoor/outdoor, etc.)
- Camera/Image quality
- Subjects (Age, Race, etc.)

- Spoof types

Cross-database Domain
Adaption

Training-Testing Difference

The testing scenarios are different with the training phase.

- Environment (Lighting, Indoor/outdoor, etc.)
- Camera/Image quality
- Subjects (Age, Race, etc.)

- Spoof types

Unknown Spoof
Detection

Cross-database Domain Adaption

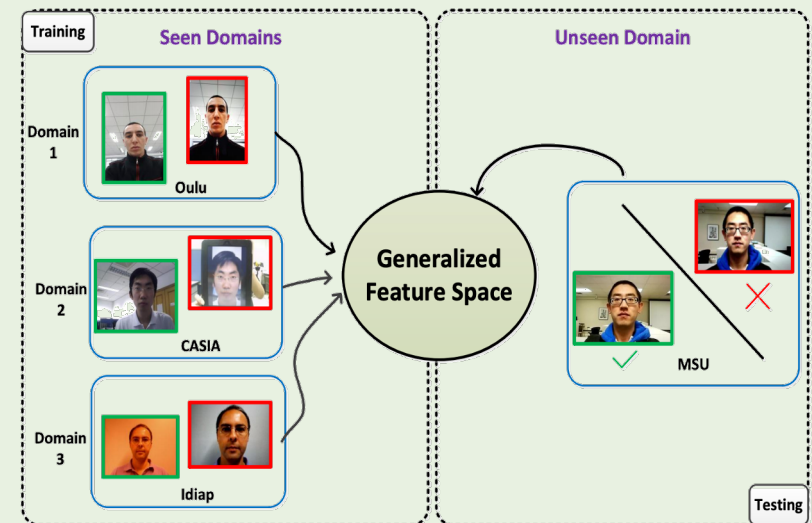
- Enforce features to be domain-invariant
 - Domain adaption [1,2]
 - Metric learning [3,5,6]
 - Meta learning [7,8]

1. Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing, TIFS, 2018
2. Unsupervised Domain Adaptation for Face Anti-Spoofing, TIFS 2018
3. Multi-adversarial Discriminative Deep Domain Generalization, CVPR, 2019
4. Domain Adaptation in Multi-Channel Autoencoder based Features for Robust Face Anti-Spoofing, ICB 2019
5. Improving Cross-database Face Presentation Attack Detection via Adversarial Domain Adaptation, ICB 2019
6. Single-Side Domain Generalization for Face Anti-Spoofing, CVPR 2020
7. Regularized Fine-grained Meta Face Anti-spoofing, AAAI 2020
8. Learning Meta Model for Zero- and Few-shot Face Anti-spoofing, AAAI 2020



Metric learning

- Adversarial learning
 - learn target features such that discriminator cannot correctly predict the domain
 - remove unrelated features
- Triplet loss
 - learn target features such that live samples from different domains are similar
 - find shared features



1. Multi-adversarial Discriminative Deep Domain Generalization, CVPR, 2019
2. Improving Cross-database Face Presentation Attack Detection via Adversarial Domain Adaptation, ICB 2019
3. Single-Side Domain Generalization for Face Anti-Spoofing, CVPR 2020



Unknown Attack Detection

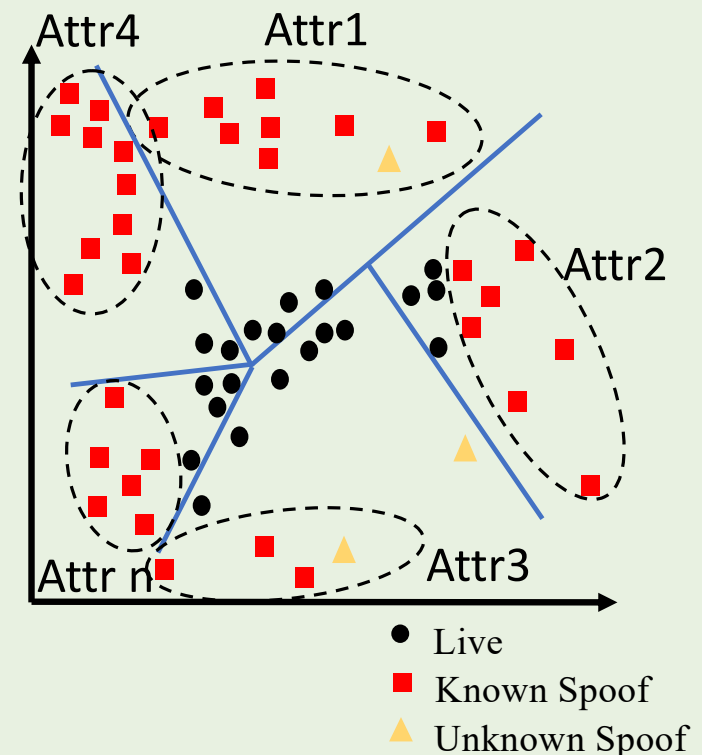
- One-class classifier
 - One-class SVM
 - Gaussian Mixture Model
 - AutoEncoder
- Zero-shot learning

1. An Anomaly Detection Approach to Face Spoofing Detection: A New Formulation and Evaluation Protocol, IEEE Access, 2017
2. Unknown Presentation Attack Detection with Face RGB Images, ICB, 2018
3. Deep Anomaly Detection for Generalized Face Anti-Spoofing, CVPRW, 2019
4. Deep Tree Learning for Zero-shot Face Anti-Spoofing, CVPR 2019



Deep Tree Learning for Zero-shot Face Anti-Spoofing

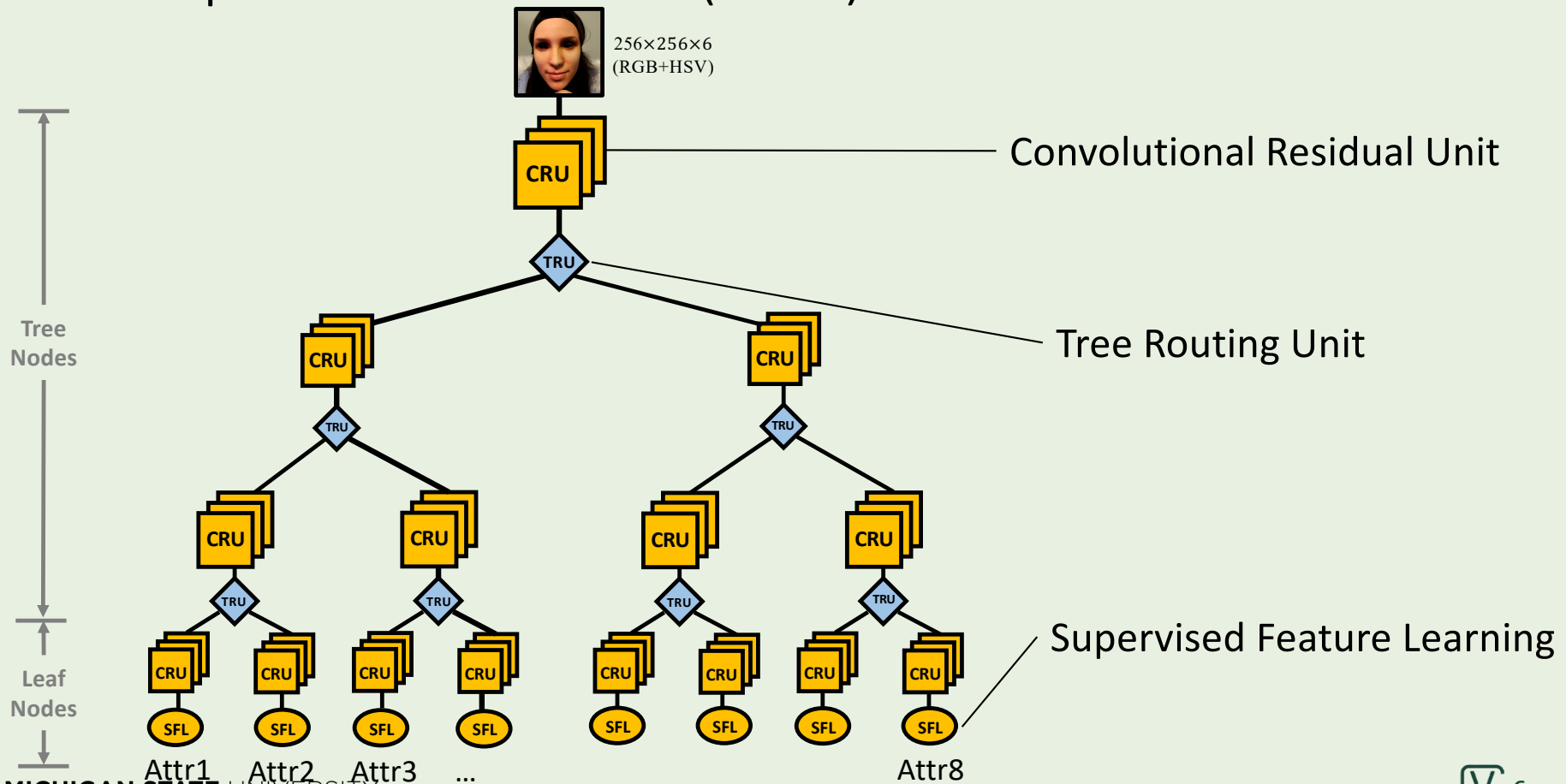
- Previous methods only model the live
- Learning semantic spoof attributes



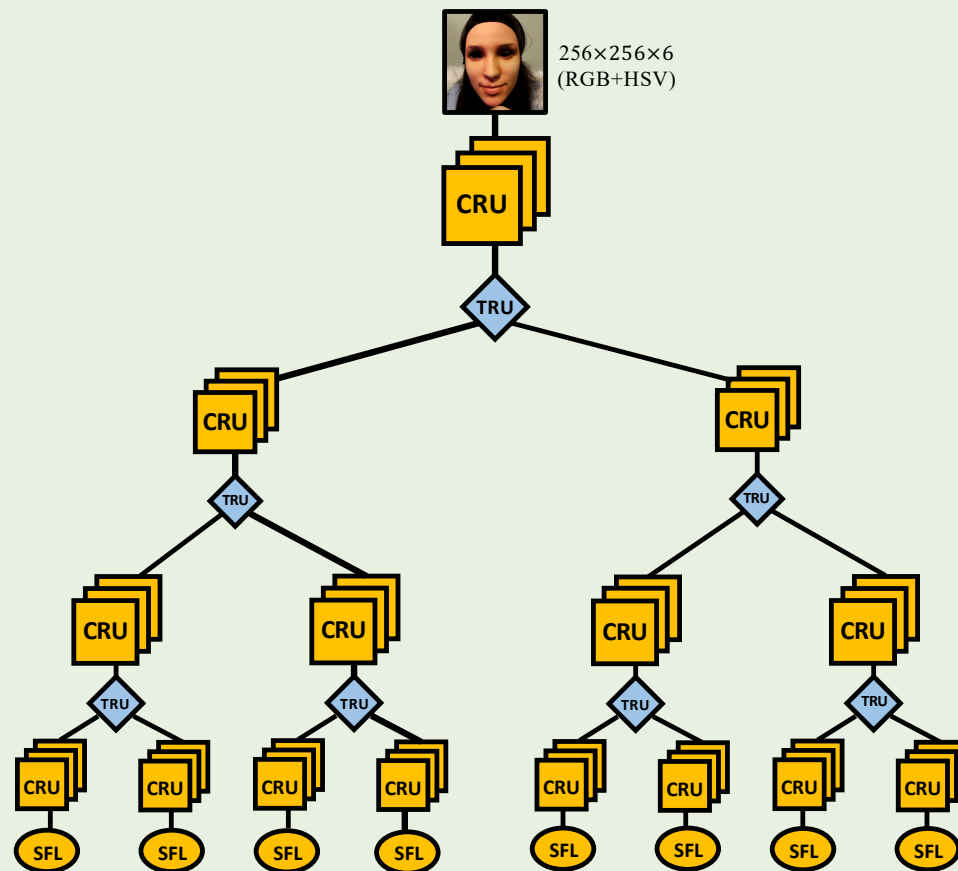
1. Liu et. al., Deep Tree Learning for Zero-shot Face Anti-Spoofing, CVPR 2019



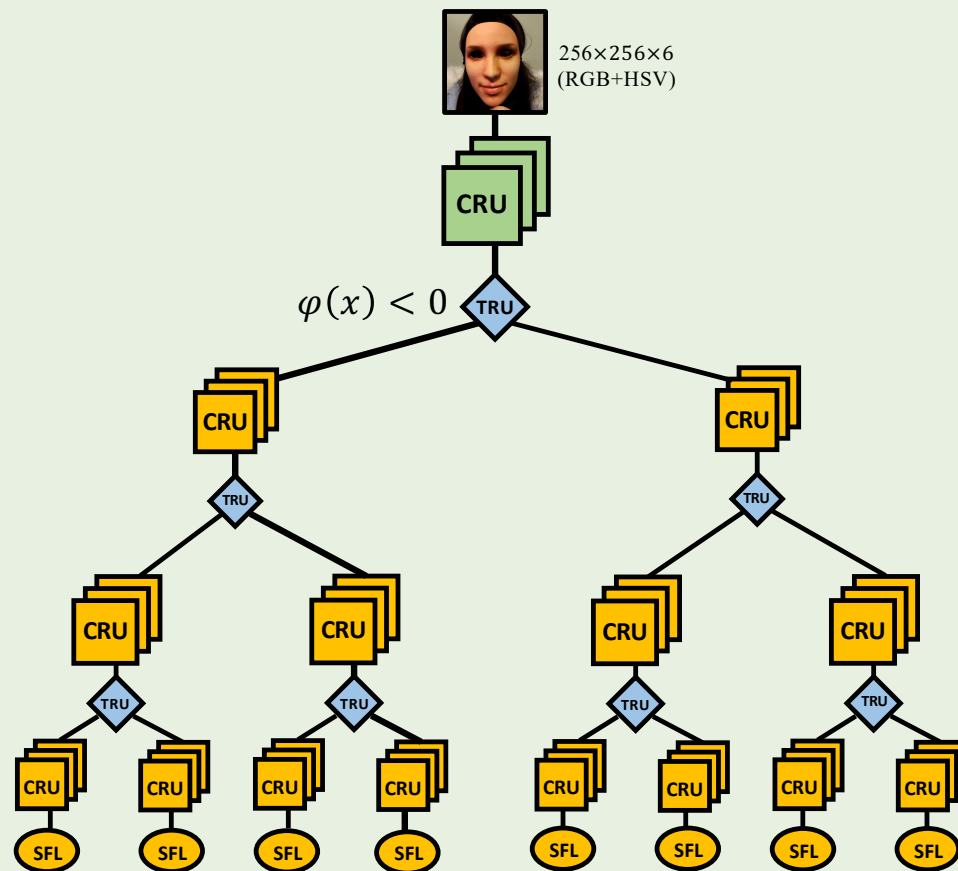
Deep Tree Networks (DTN)



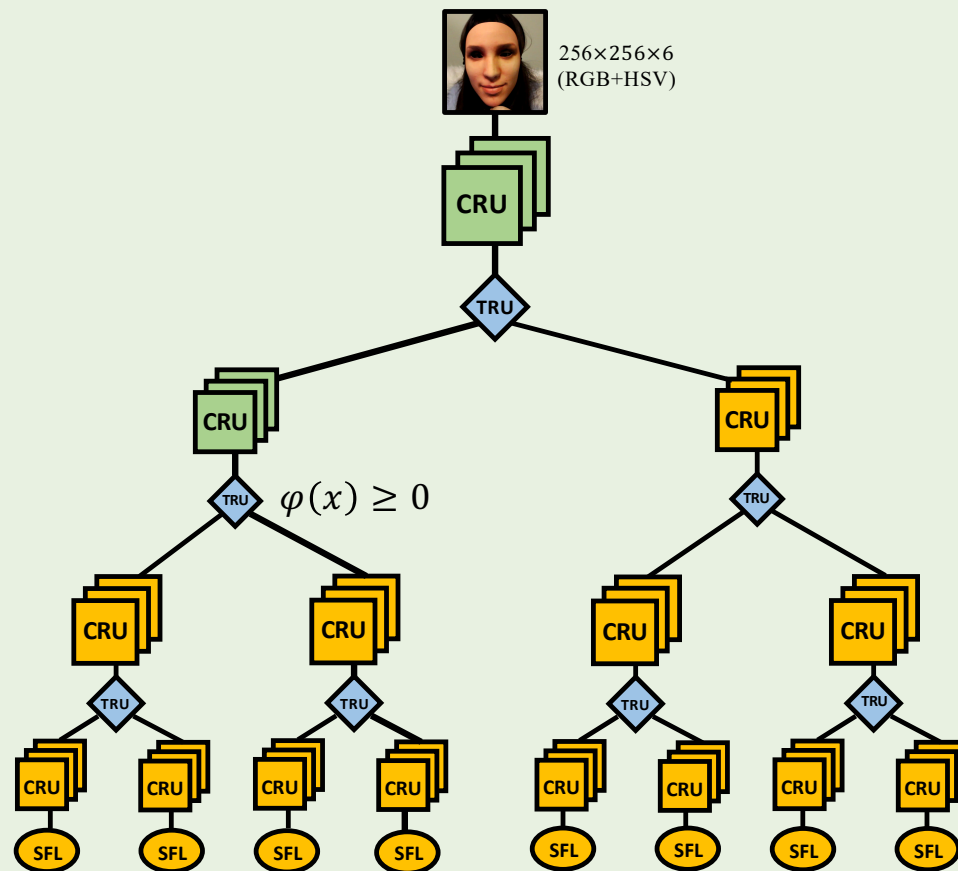
Deep Tree Networks (DTN)



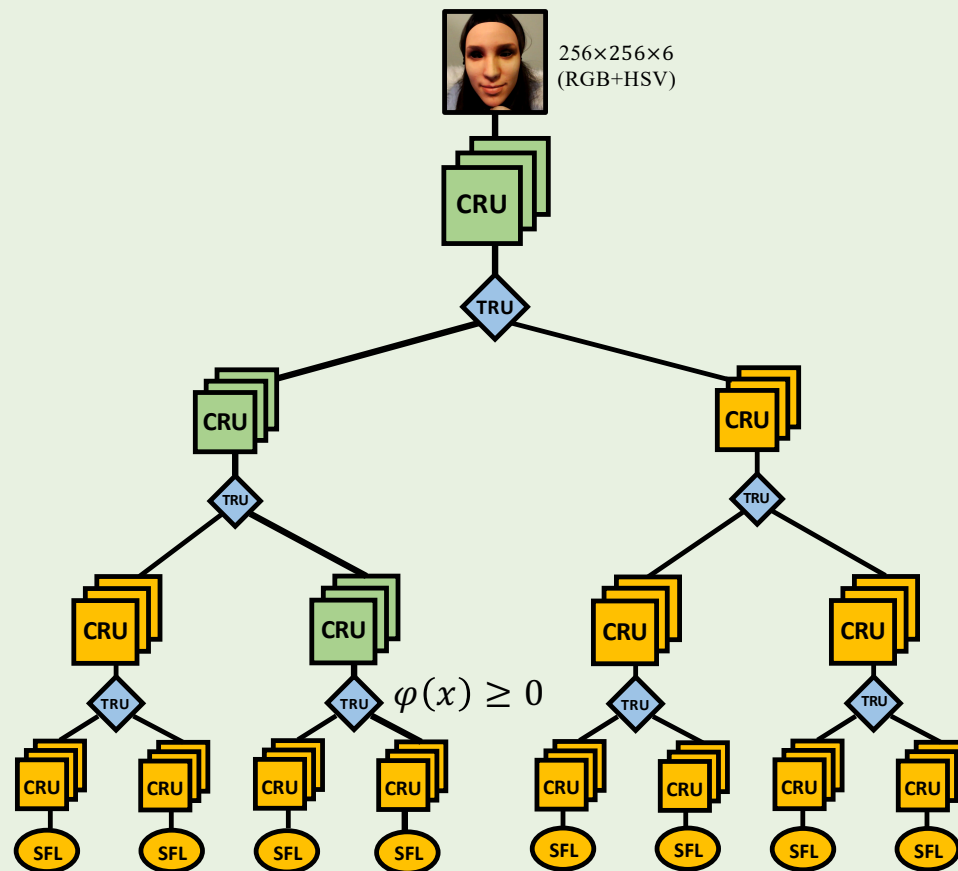
Deep Tree Networks (DTN)



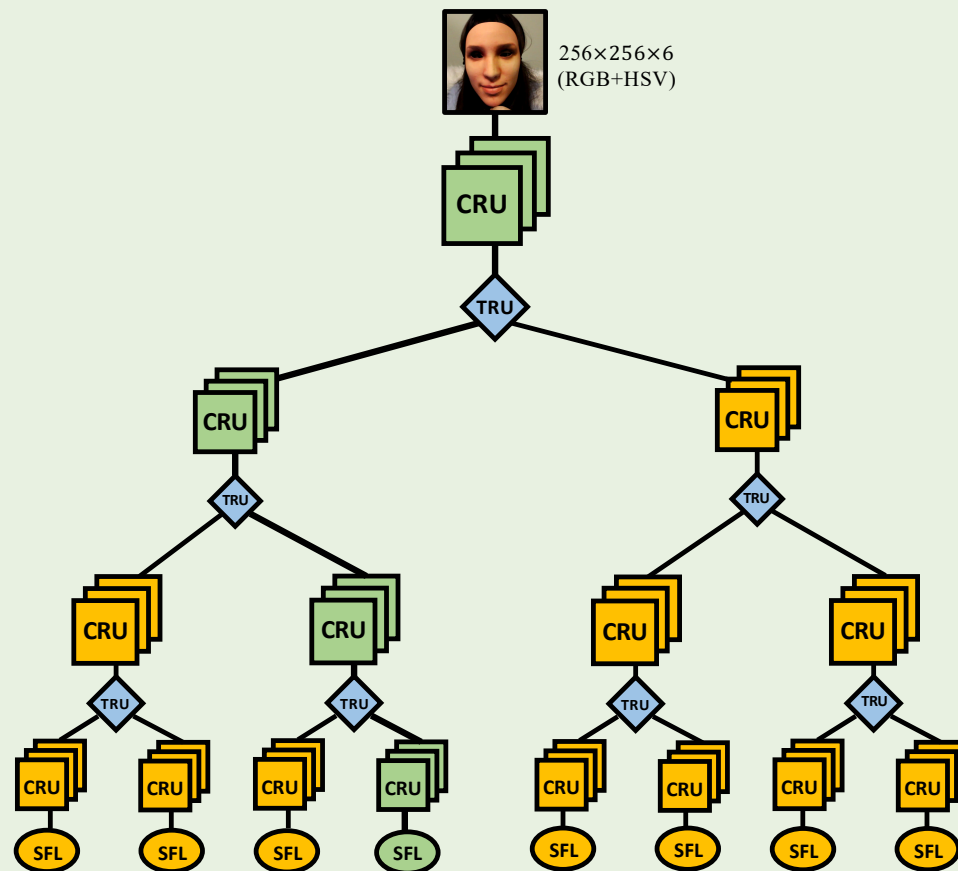
Deep Tree Networks (DTN)



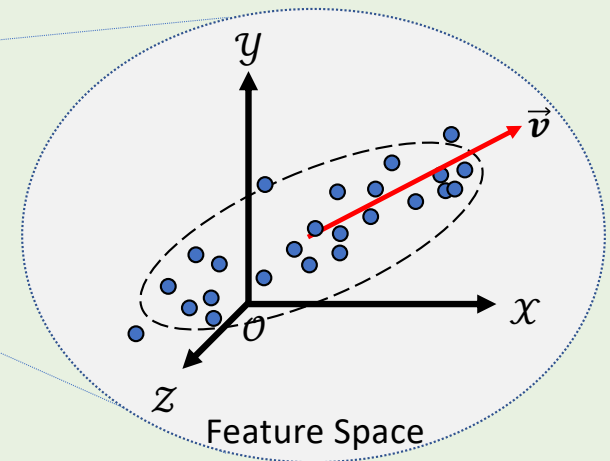
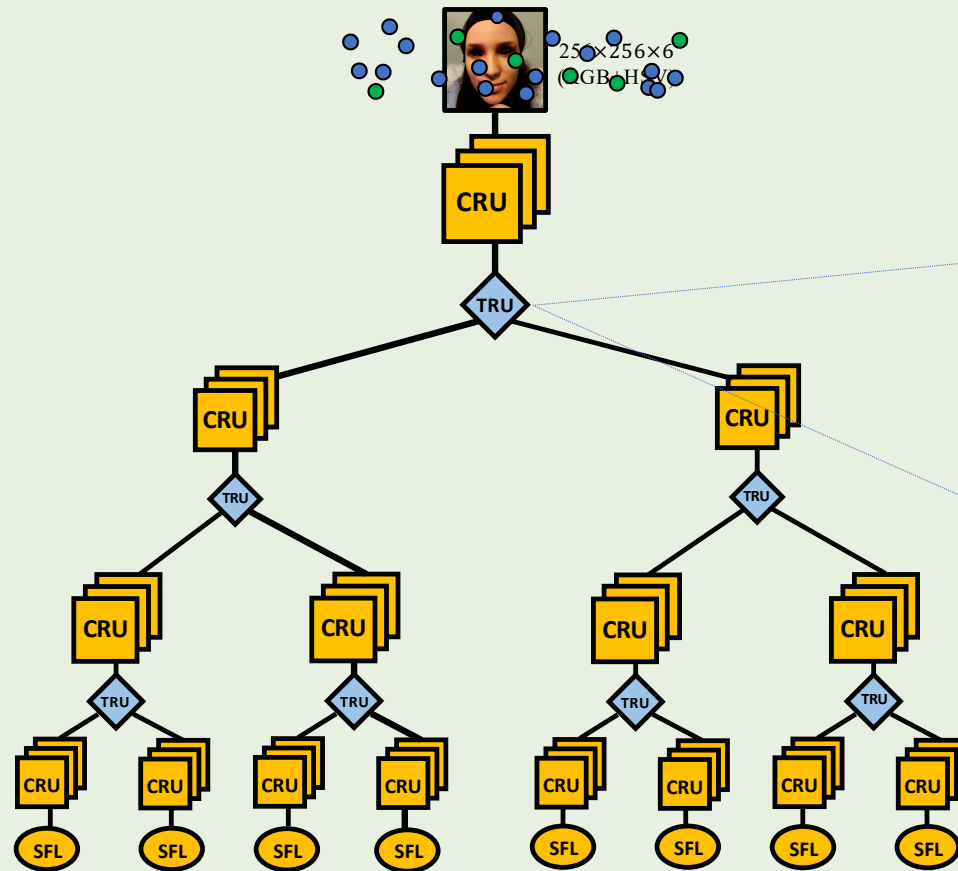
Deep Tree Networks (DTN)



Deep Tree Networks (DTN)



Training TRU

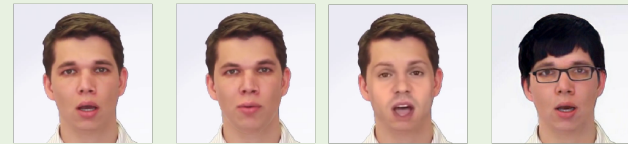


Future Works

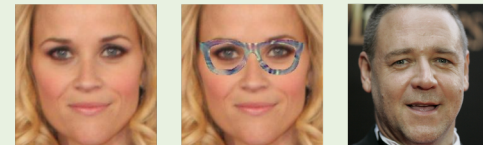
- FAS generalization (still poor performance)
- Extreme lighting conditions
- Unknown attacks
- A joint model to handle multiple attacks



Generic Anti-spoofing



Face manipulation attack



Adversarial attack

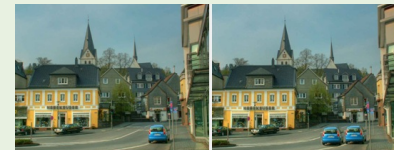


Image forgery detection

Digital Manipulation Detection

- Introduction of digital attacks
 - Problem, Facial manipulation types, Challenges
- Benchmark databases
- Face manipulation detection methods
 - Dynamic methods, Static methods
- Future Work

Outline

- **Introduction of digital attacks**
 - **Problem, Facial manipulation types, Challenges**
- Benchmark databases
- Face manipulation detection methods
 - Dynamic methods, Static methods
- Future Work

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)

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

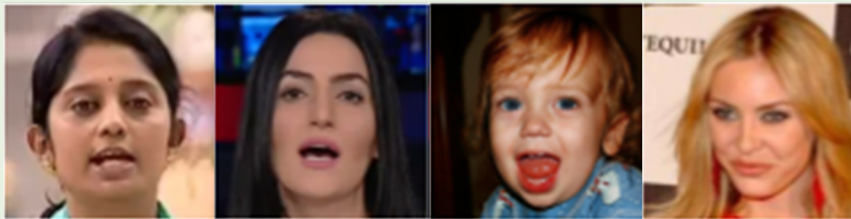


(b) Expression swap

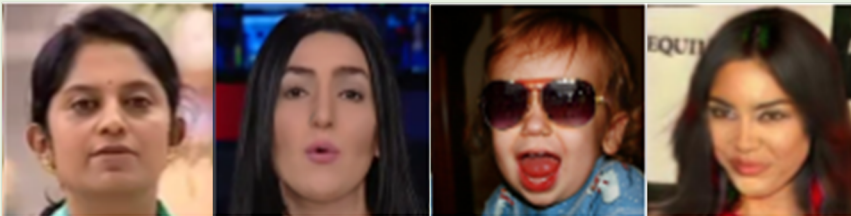
Dang et al. On the Detection of Digital Face Manipulation. In CVPR, 2020.

Facial Manipulation Types

Real



Fake



(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

Subject #1



Subject #2



(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.

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 Work

Databases Comparison

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

Outline

- Introduction of digital attacks
 - Problem, Facial manipulation types, Challenges
- Benchmark databases
- **Face manipulation detection methods**
 - **Dynamic methods**, Static methods
- Future Work

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.

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:

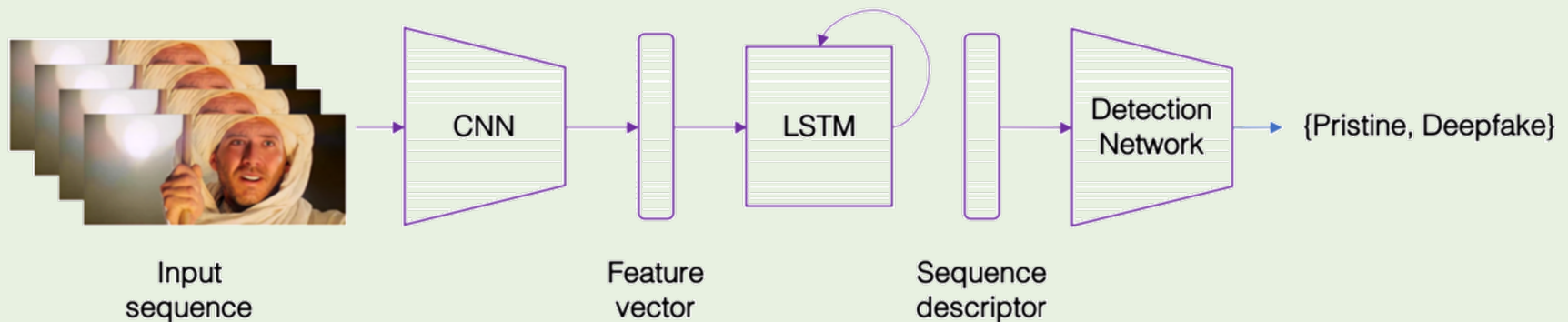
- Spatial-aware and temporal-aware

Con:

- Fake feature can be immersed during long aggregation

Deepfake Video Detection Using Recurrent Neural Networks

- 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.



Outline

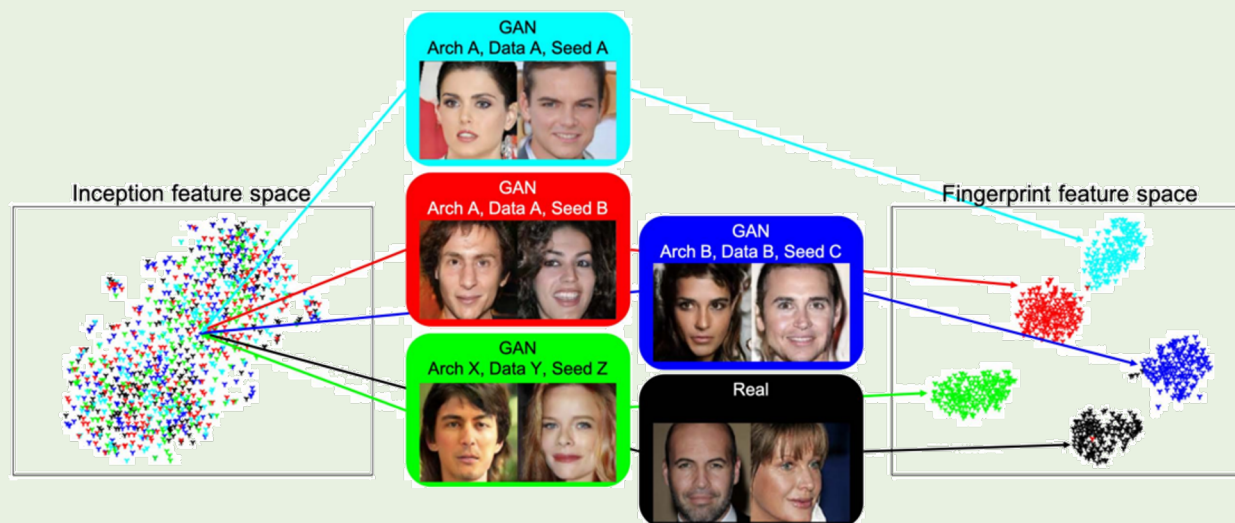
- Introduction of digital attacks
 - Problem, Facial manipulation types, Challenges
- Benchmark databases
- **Face manipulation detection methods**
 - Dynamic methods, **Static methods**
- Future Work

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)

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.



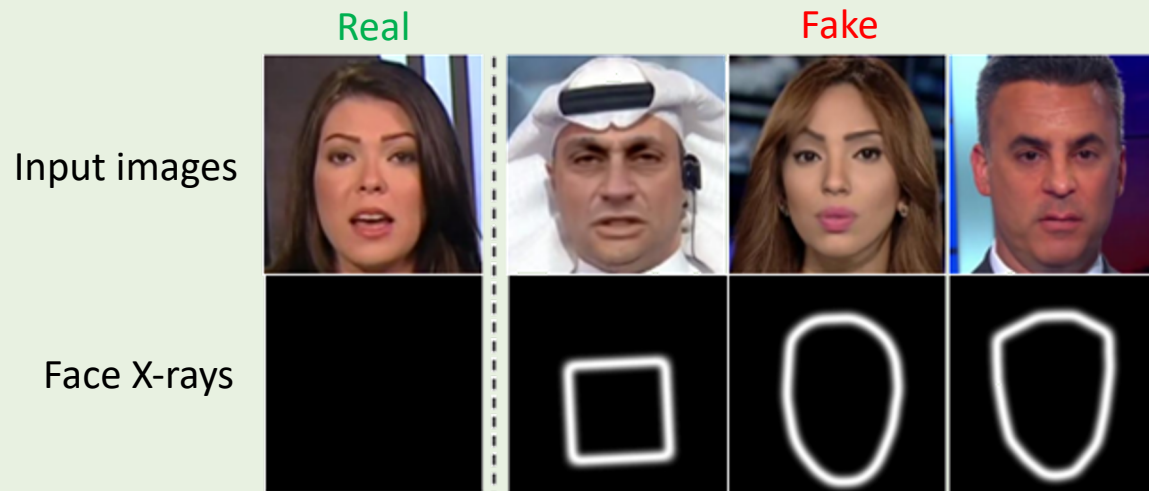
Yu et al. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In ICCV, 2019.

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)

Face X-ray for More General Face Forgery Detection



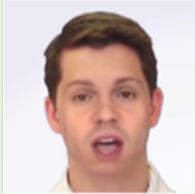


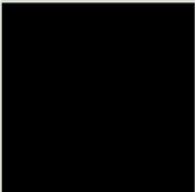




- Most existing manipulation methods share a common step: blending the altered face into a background image.
- Face X-ray reveals whether the input image can be decomposed into the blending of two images from different sources.



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



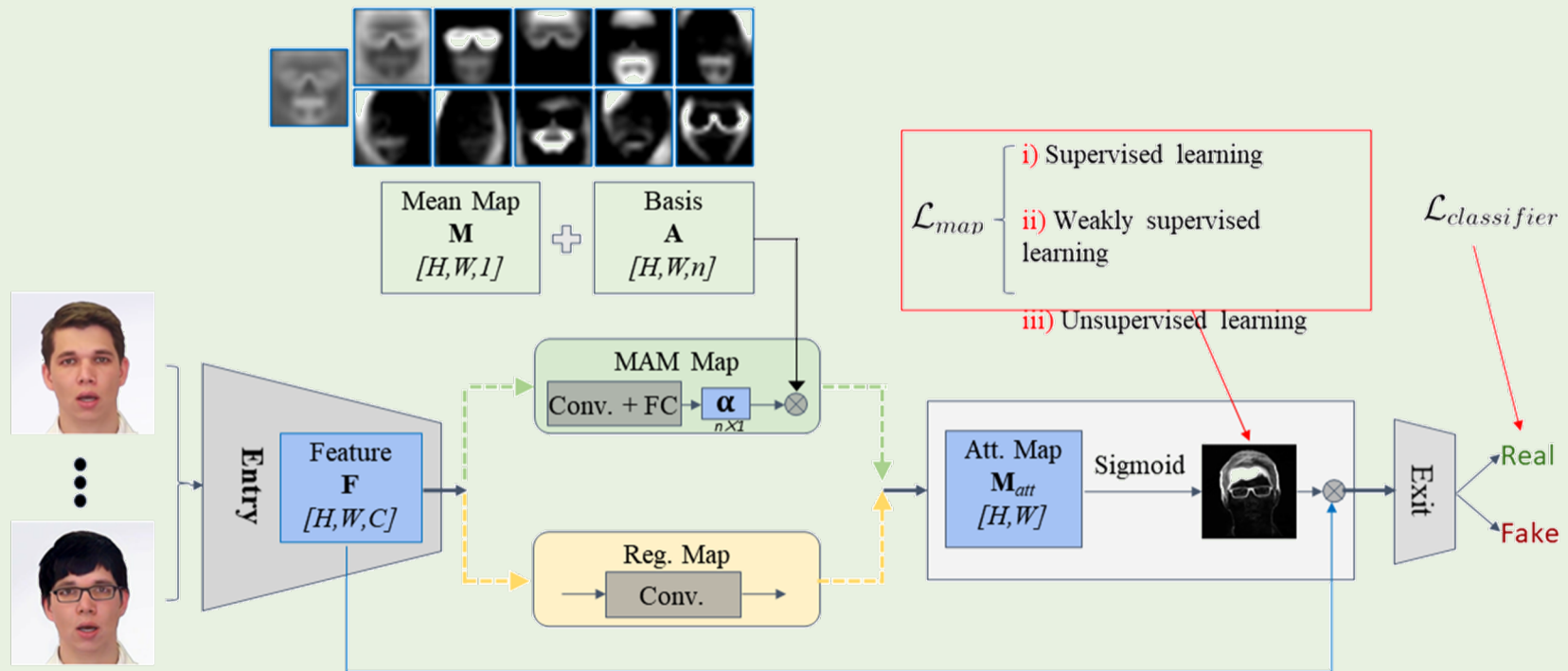
On the Detection of Digital Face Manipulation

Fake Type	Real	Expression swap	Identity swap	Attribute manipulation	Entire face synthesis
Input Sample					
Binary Prediction	Real	Fake	Fake	Fake	Fake
Attention Map					

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.

Experimental Results ---- manipulation localization

Source image															
Manipulated image															
Ground-truth manipulated mask															
Estimated attention map															
IINC score	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.34	0.25	0.36	0.61	0.40	0.44	0.37	0.40
PBCA score	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.66	0.73	0.86	0.66	0.47	0.84	0.68	0.22
	(a) Real			(b) Entire synthesis			(c) Attribute manipulation			(d) Expression swap			(e) Identity swap		

Outline

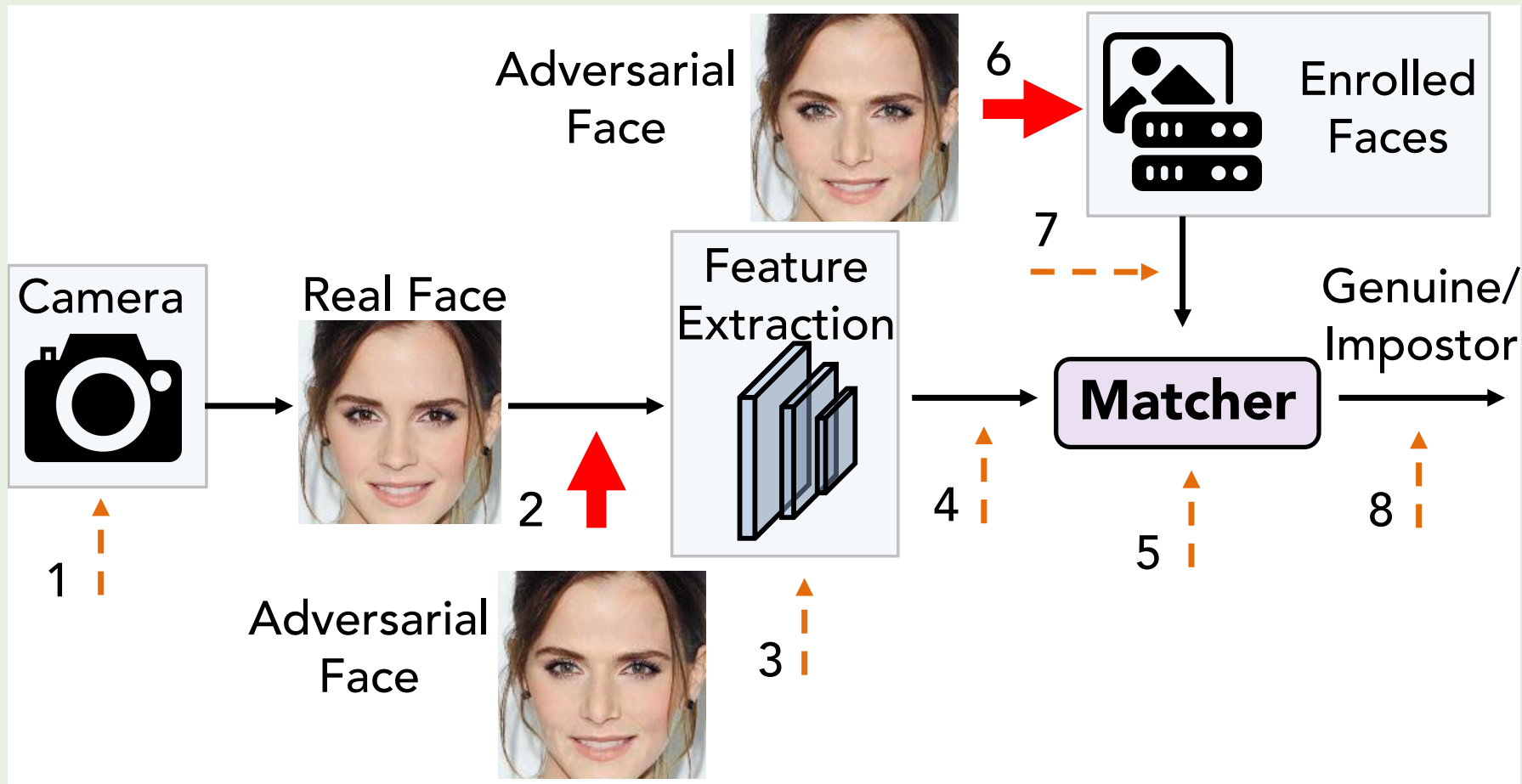
- Introduction of digital attacks
 - Problem, Facial manipulation types, Challenges
- Benchmark databases
- Face manipulation detection methods
 - Dynamic methods, Static methods
- **Future Work**

Future Directions

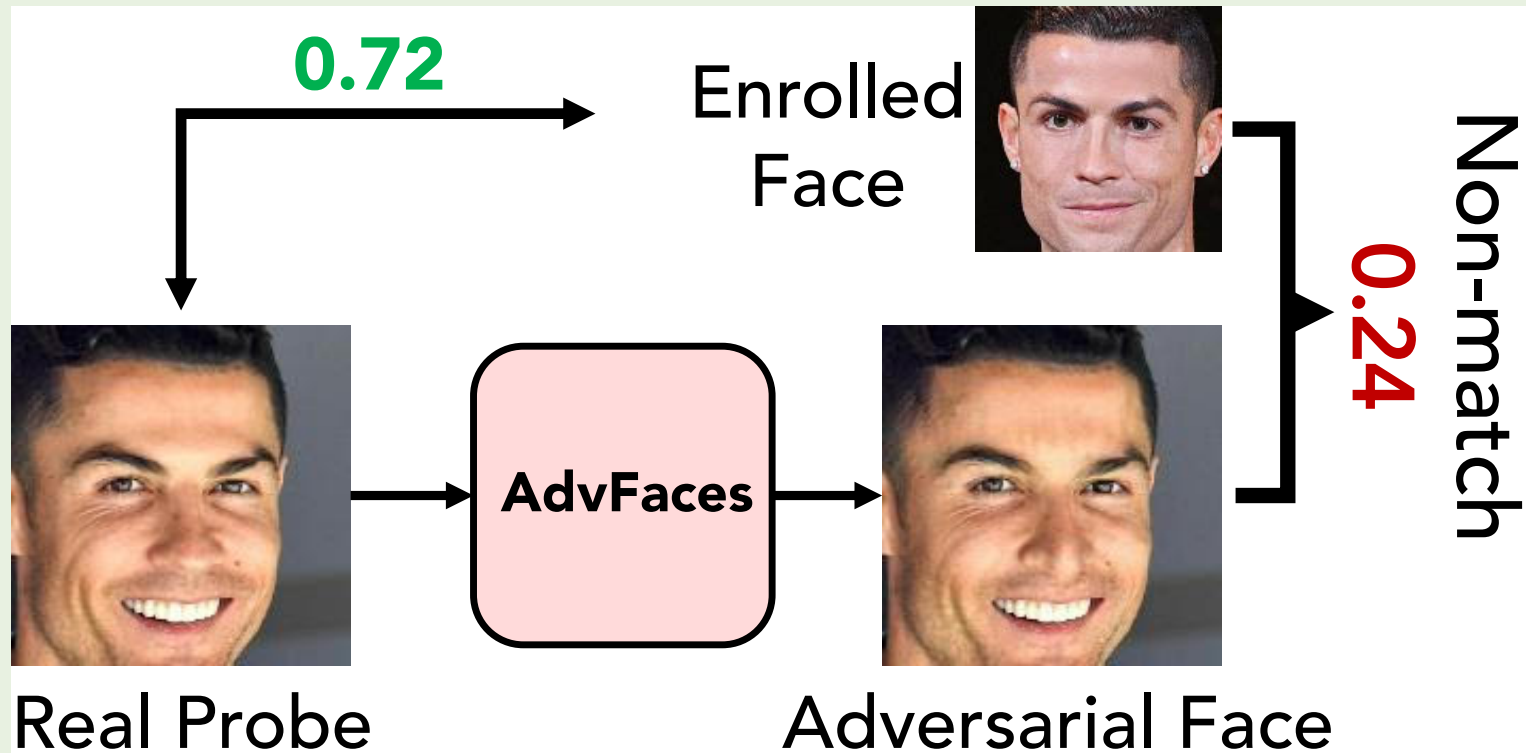
- Facial manipulation techniques are continuously improving. More research on generalization ability of forgery detection against **unseen** manipulation types.
- Challenging when performed in **uncontrolled** scenarios. Fake imagery on social network are usually suffering from large variations in compression, resizing, noise, etc.
- Fusion of other **modalities** such as text or audio can be valuable to improve the detectors.



Adversarial Attack on AFR Systems



Obfuscation Attack



Cosine similarity scores $\in [-1, 1]$ via ArcFace . **Score** > 0.28 is accepted as genuine (threshold @ 0.1% FAR)

Adversarial Face



Gallery

Face Matcher	Match Score	Decision
FaceNet	0.87	Match
SphereFace	0.92	Match
ArcFace	0.79	Match
COTS-A	0.93	Match
COTS-B	0.83	Match



Real Probe

Matching Threshold determined at 0.1% False Accept Rate

Adversarial Face



Gallery

Face Matcher	Match Score	Decision
FaceNet	0.03	Non-Match
SphereFace	0.34	Non-Match
ArcFace	0.27	Non-Match
COTS-A	0.12	Non-Match
COTS-B	0.32	Non-Match



Adversarial Probe
via AdvFaces*

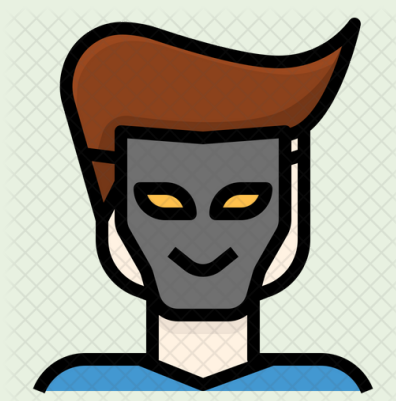
Matching Threshold determined at 0.1% False Accept Rate

* D. Deb, J. Zhang, and A. K. Jain. "AdvFaces: Adversarial Face Synthesis". In *IEEE IJCB*, 2020.

Requirements of an Adversarial Face Generator



01 Visually Realistic



02 Successful Attack

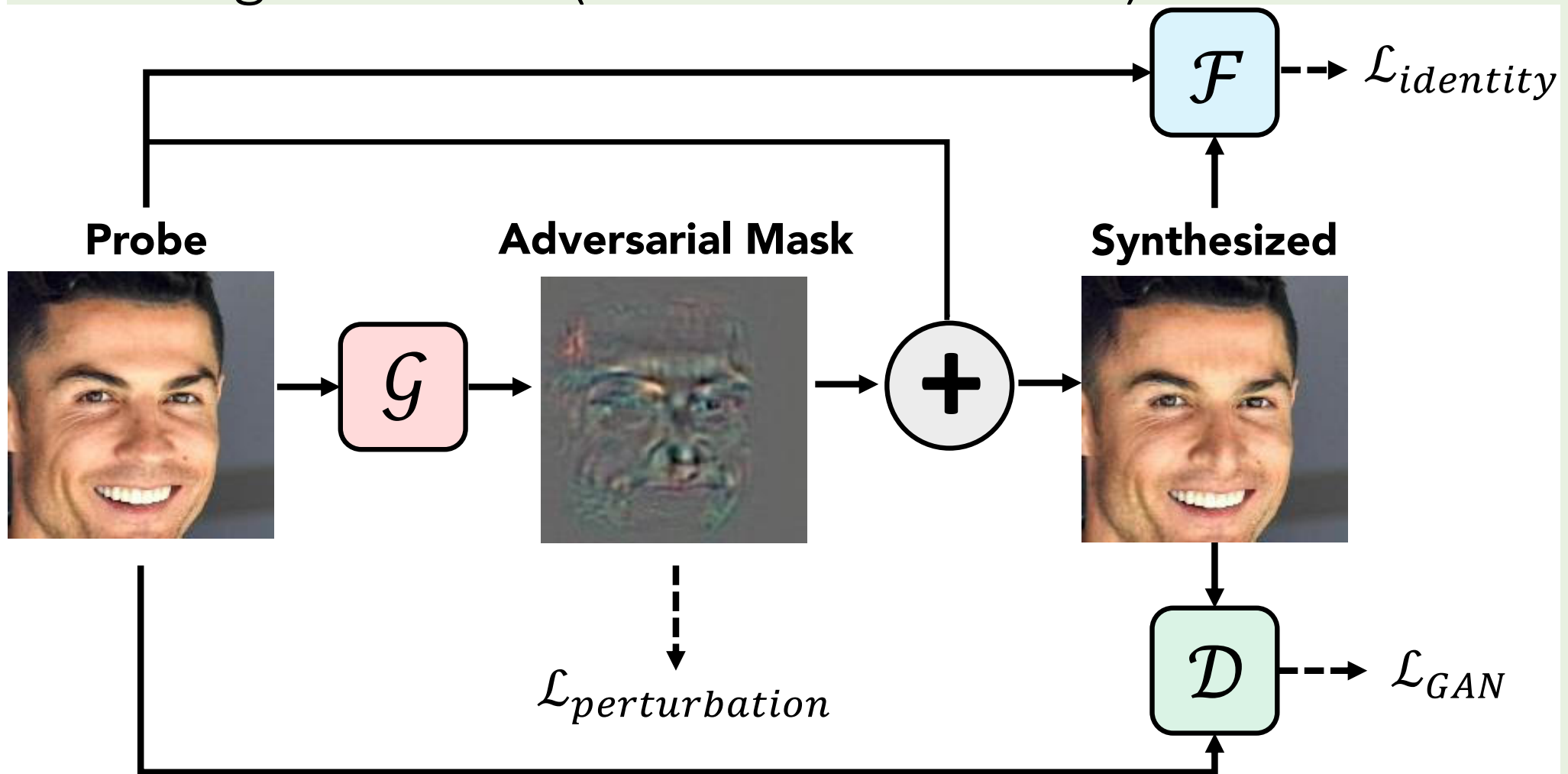


03 Controllable



04 Transferable

Training AdvFaces (Obfuscation Mode)



Examples of Obfuscation Attack

Gallery

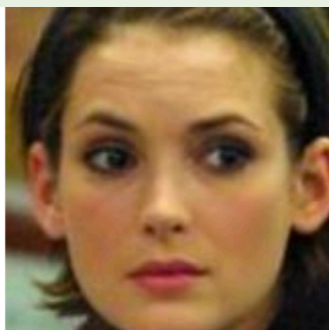
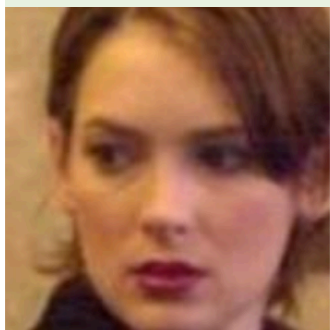
Probe

AdvFaces

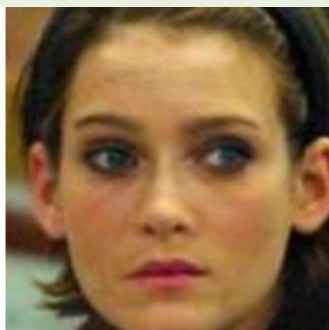
GFLM

PGD

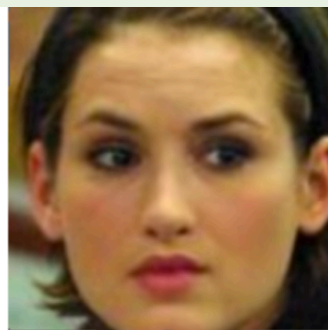
FGSM



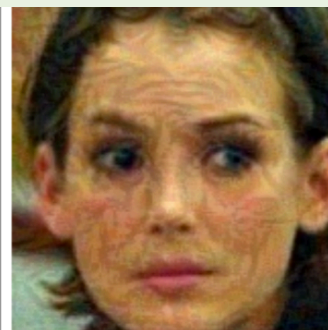
0.68



0.14



0.26



0.27



0.04



0.38



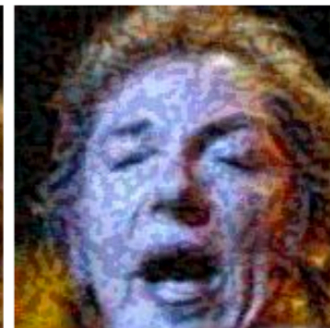
0.08



0.12



0.21



0.02

Cosine similarity scores $\in [-1, 1]$ via ArcFace . **Score > 0.28** is accepted as genuine (threshold @ 0.1% FAR)

Obfuscation Attack on 5 SOTA AFR Systems

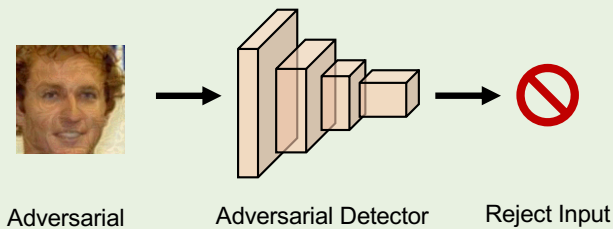
	AdvFaces	GFLM	PGD	FGSM
Attack Success Rate (%) @ 0.1% False Accept Rate				
<i>FaceNet*</i>	99.67	23.34	99.70	99.96
SphereFace	97.22	29.49	99.34	98.71
ArcFace	64.53	03.43	33.25	35.30
COTS-A	82.98	08.98	18.74	32.48
COTS-B	60.71	05.05	01.49	18.75
Structural Similarity				
	0.95 ± 0.01	0.82 ± 0.12	0.29 ± 0.06	0.25 ± 0.06
Computation Time (s)				
	0.01	3.22	11.74	0.03

* White-box matcher utilized for training AdvFaces and baselines

Defense Against Adversarial Faces

Detection

(Dedicated detector that distinguishes between real and adversarial faces)

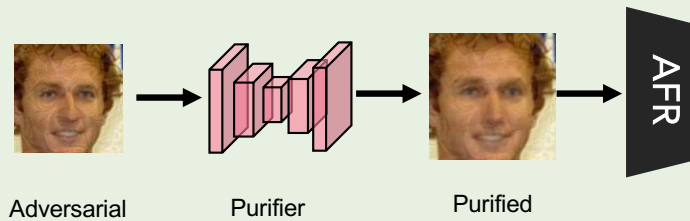


Independent of the deployed AFR system

Potentially poor generalization performance on unseen attack types

Purification

(Dedicated purifier that attempts to remove adversarial noise)

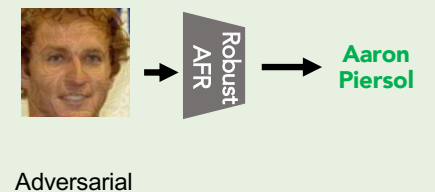


Interpretable: visual inspection of perturbed pixels

Challenging to model the manifold of real faces

Robustness

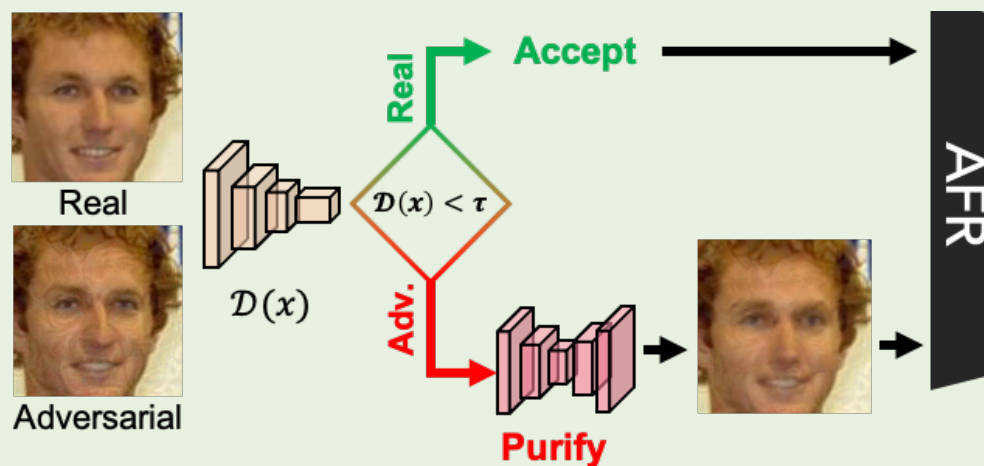
(Train AFR system robust to adversarial noise)



No additional model needed in AFR pipeline

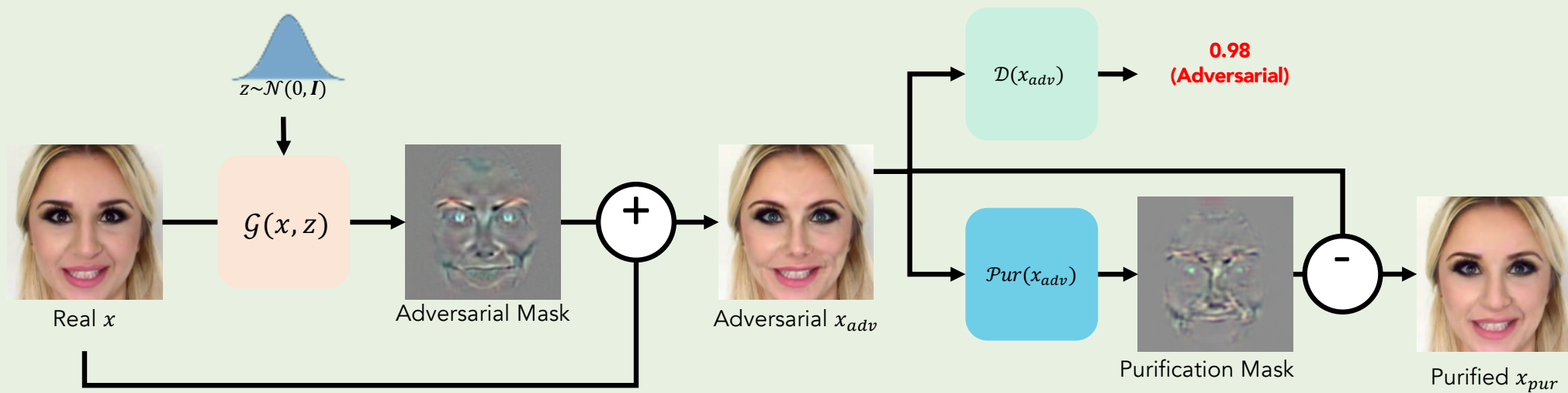
Robustness comes at the cost of general FR performance

FaceGuard: Adversarial Face Detection



D. Deb, X. Liu, and A. K. Jain. "FaceGuard: A Self-Supervised Defense Against Adversarial Face Images". In *arXiv preprint arXiv:2011.14218*, 2020.

Training FaceGuard



Detection Results

Real

FGSM

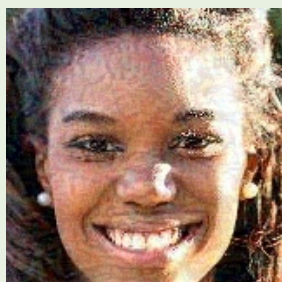
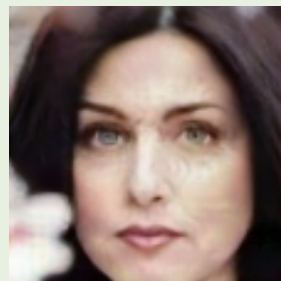
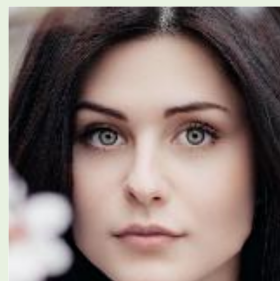
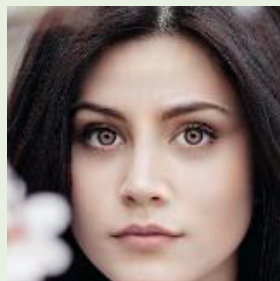
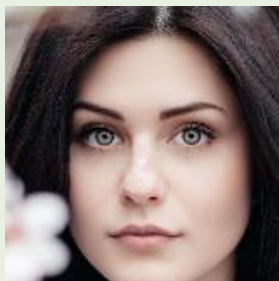
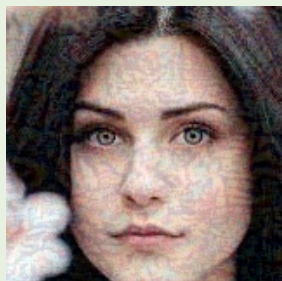
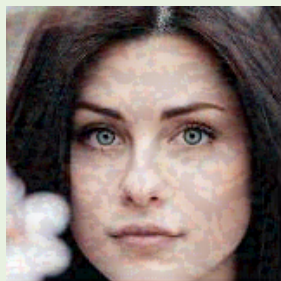
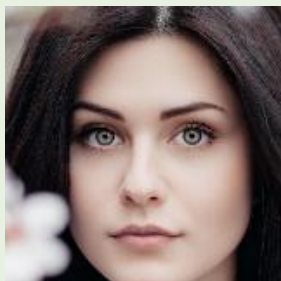
PGD

DeepFool

AdvFaces

GFLM

SemanticAdv



**Detection
Accuracy:**

99.85%

99.85%

99.85%

99.84%

99.61%

99.85%

Evaluated on 9,164 real and 9,164 adversarial face per attack type in LFW dataset.

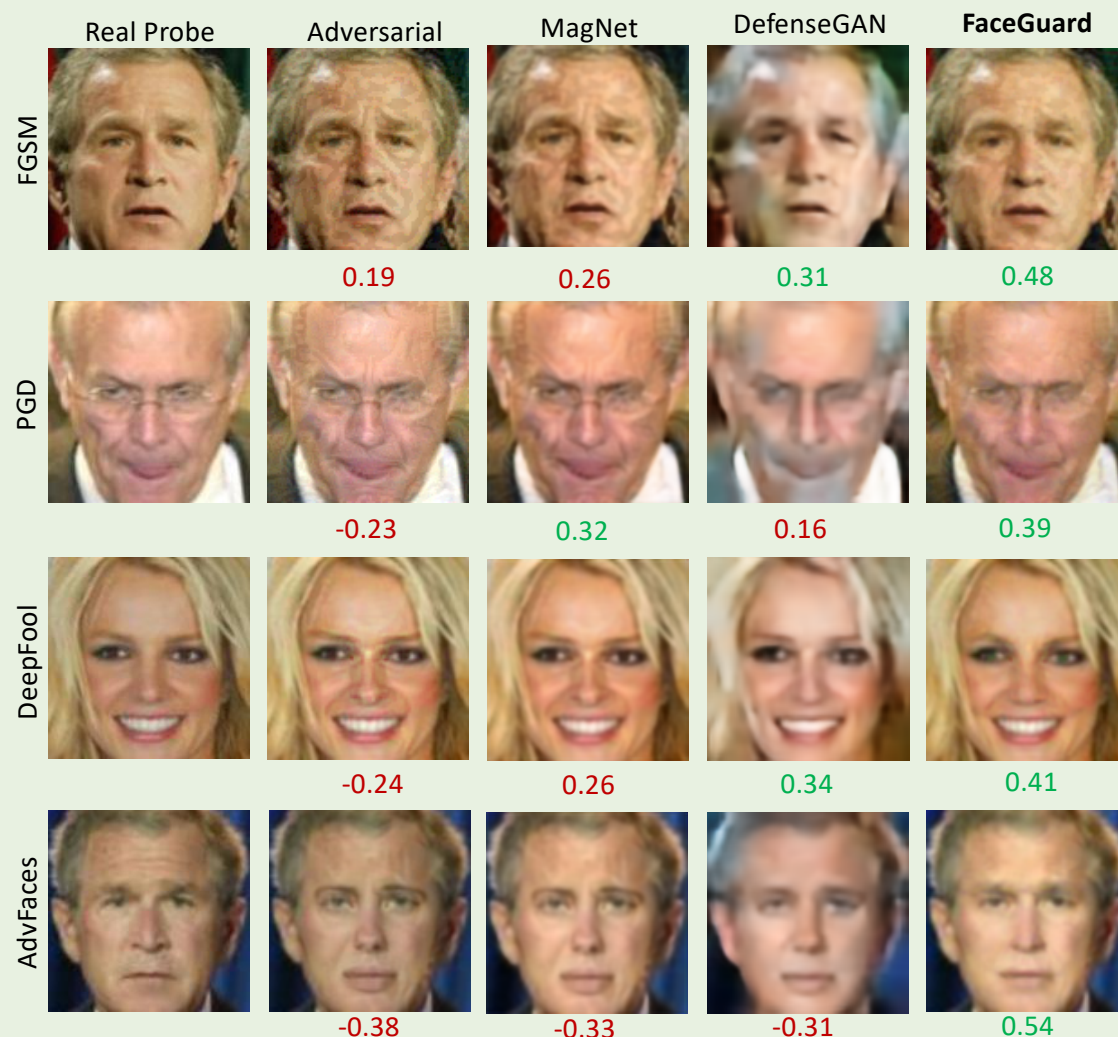
Purification Results

TAR (%) @ 0.1% FAR

Defenses	Approach	Adversarial Pairs 3M pairs
No-Defense	-	34.27
Adv. Training	Robustness	11.23
Rob-GAN	Robustness	13.89
L2L	Robustness	16.76
MagNet	Purification	38.32
DefenseGAN	Purification	39.21
NRP	Purification	61.44
<i>FaceGuard</i>	Purification	77.46

Cosine similarity scores $\in [-1, 1]$ via ArcFace.
ArcFace achieves TAR = 99.82% @ 0.1% FAR
under bona fide pairs in LFW.

Score > 0.28 is accepted as genuine.



Future Directions

1. Explainability meets Adversarial Learning
 - Can we explain face recognition systems by understanding how learning-based adversarial synthesizers craft attacks for any AFR system?
2. Secondary Adversarial Attacks on Defense Systems
 - How can we prevent further attacks against prevailing defense systems?
3. Designing a metric for evaluating “robustness against adversarial faces” in AFR systems
 - We can then evaluate different AFR systems based on **both** accuracy and robustness
 - Decide which AFR vendors to choose depending on operational requirement

Conclusions

- Emerging problems in biometrics and computer vision
- Increasing importance for practical applications
- Many common issues among different attack types
 - Robustness to PIE,
 - low-level image processing,
 - printing,
 - unseen types,
 - generalization,
 - explainability

Thanks
Questions?



MICHIGAN STATE UNIVERSITY



Computer Vision Lab