

Towards Robust and Secure Face Recognition: Defense Against Physical and Digital Attacks

Dissertation Defense

Presenter: Debayan Deb

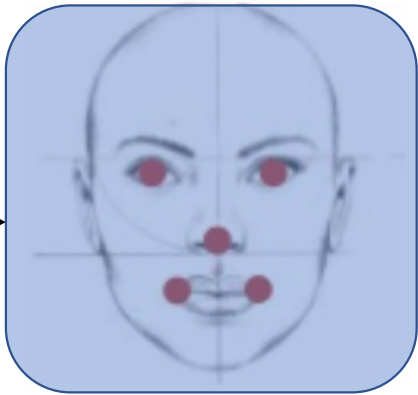
Advisor: Dr. Anil K. Jain

Date: April 22, 2021

Automated Face Recognition (AFR) Pipeline



Input Face



Face Alignment



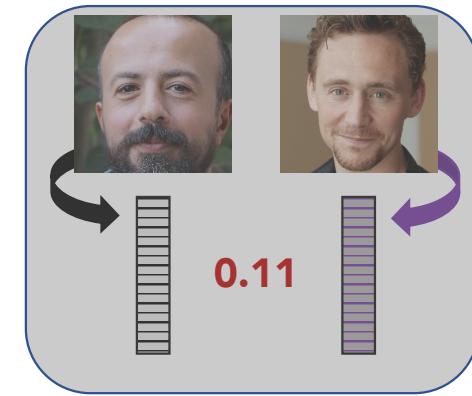
Cropped and Aligned Face



Feature
Extraction



Face Feature
Vector



Similarity
Measurement

AFR Performance over Time

Methods	Year	Face Representation	LFW
Eigenfaces	1991	Holistic	60.02
Fisherfaces	2013	Holistic	87.47
HighDim-LBP	2013	Local	95.17
Joint Bayesian	2012	Local	96.33
DeepFace	2014	Learned	97.35
DeepID	2014	Learned	97.45
VGGFace	2015	Learned	98.95
FaceNet	2015	Learned	99.63
SphereFace	2017	Learned	99.42
CosFace	2018	Learned	99.73
ArcFace	2019	Learned	99.83

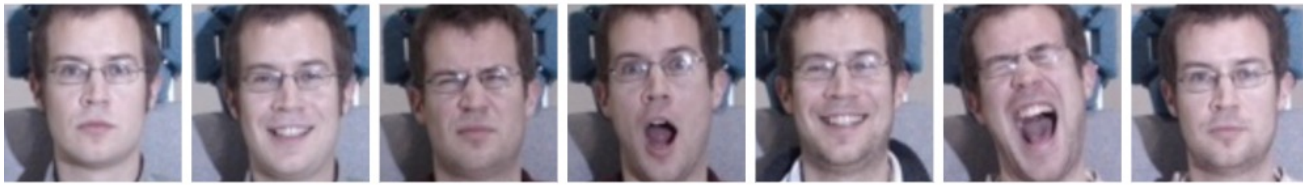
Sources of Error



(a) Pose



(b) Illumination



(c) Expression

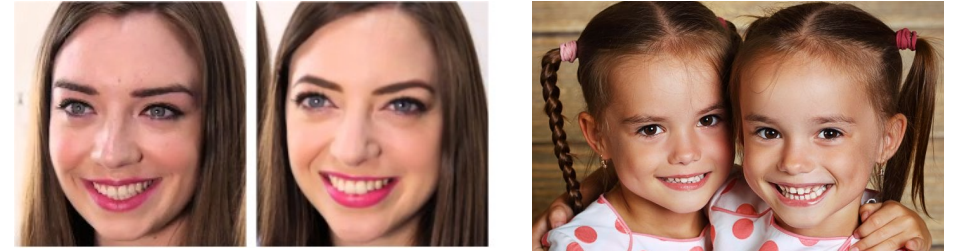


(d) Occlusion



(e) Resolution

Twins



Döpplegangers



Aging

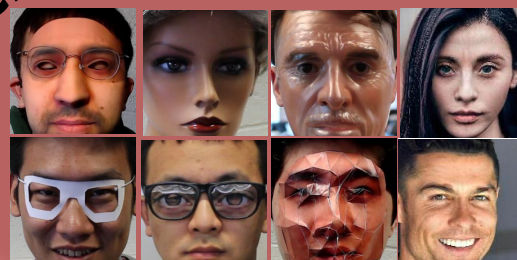


Genuine/Real/Bona Fides

Human
Faces



Face Attacks

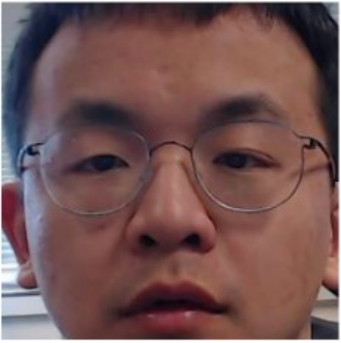


Detecting Presentation Attacks



Deb and Jain: "Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach." *IEEE TIFS*, 2020

Face Spoofs



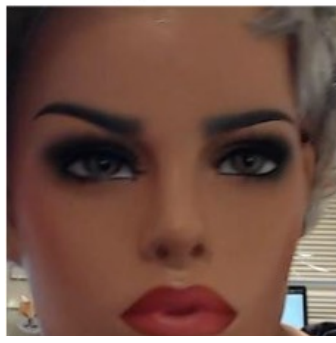
Live



Paper Mask



Half Mask



Mannequin



Paper Cut



Impersonation



Replay



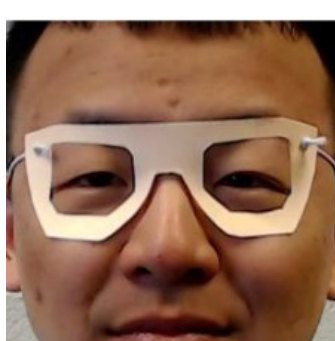
Silicone Mask



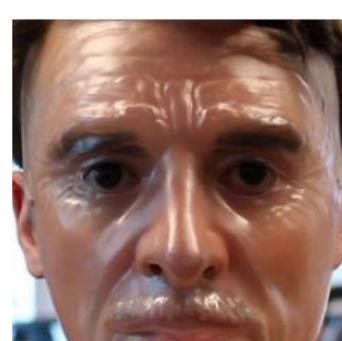
Print



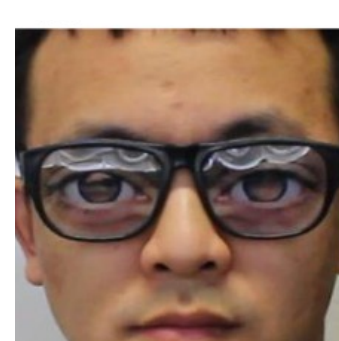
Cosmetic



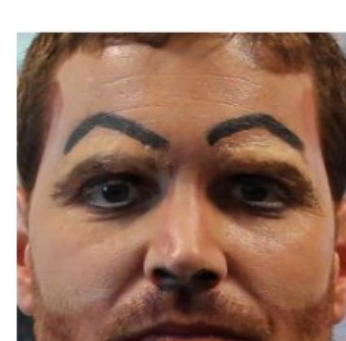
Paper Glasses



Transparent



FunnyEye



Obfuscation

All spoofs except Mannequin, Impersonation, Transparent, and Obfuscation belong to the same person in Live.

Limitations of Prevailing Face Anti-Spoofing Approaches

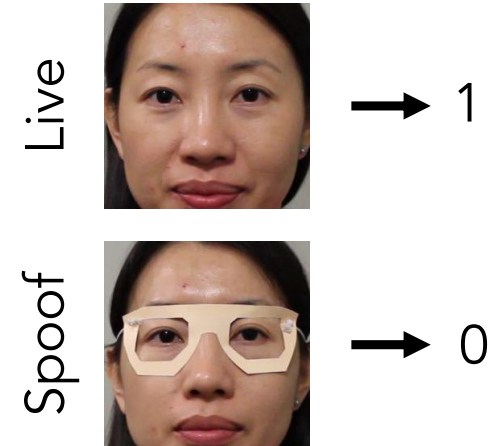


1. Generalizability to Unseen Spoof Types

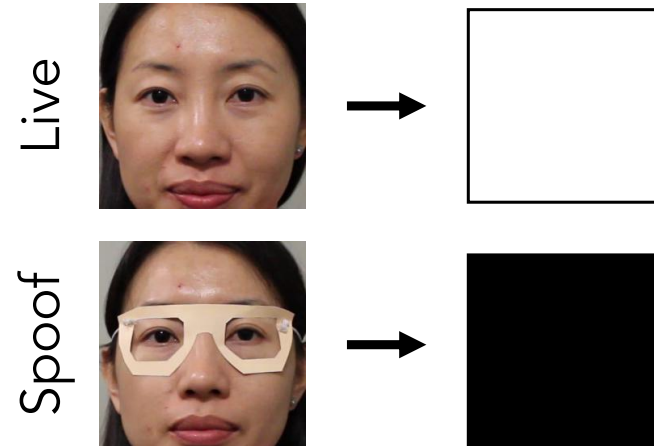
2. Interpretability

Training Supervision

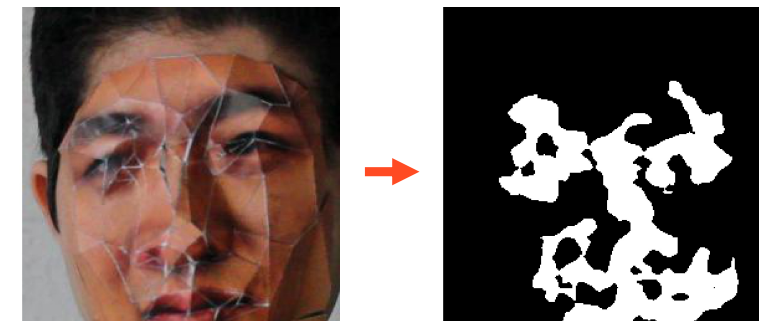
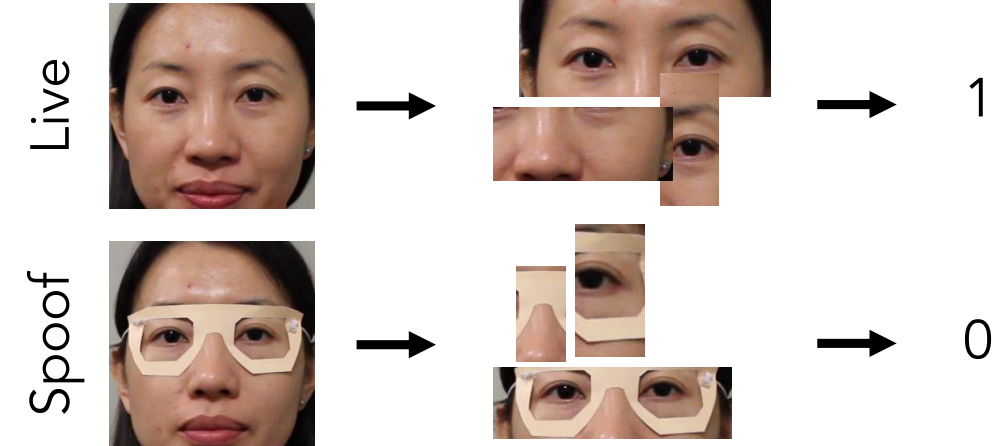
Global Supervision



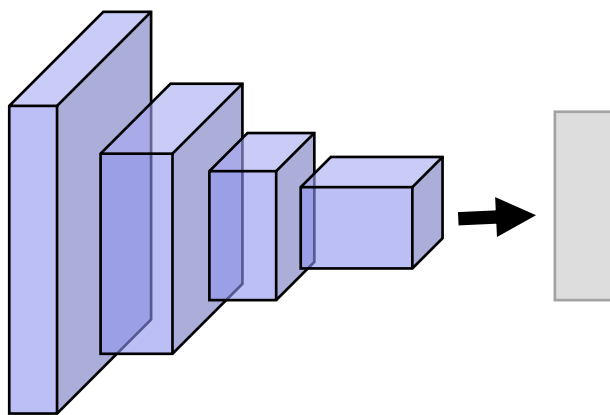
Pixel-level Supervision



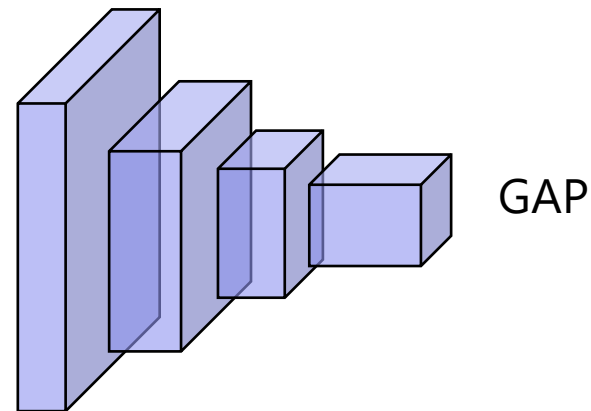
Regional Supervision



CNN vs. FCN

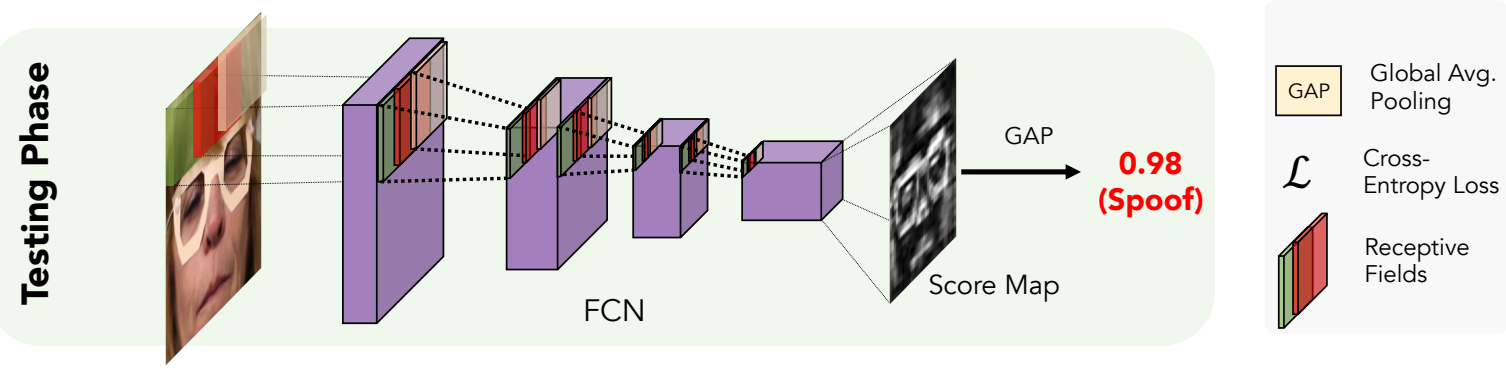
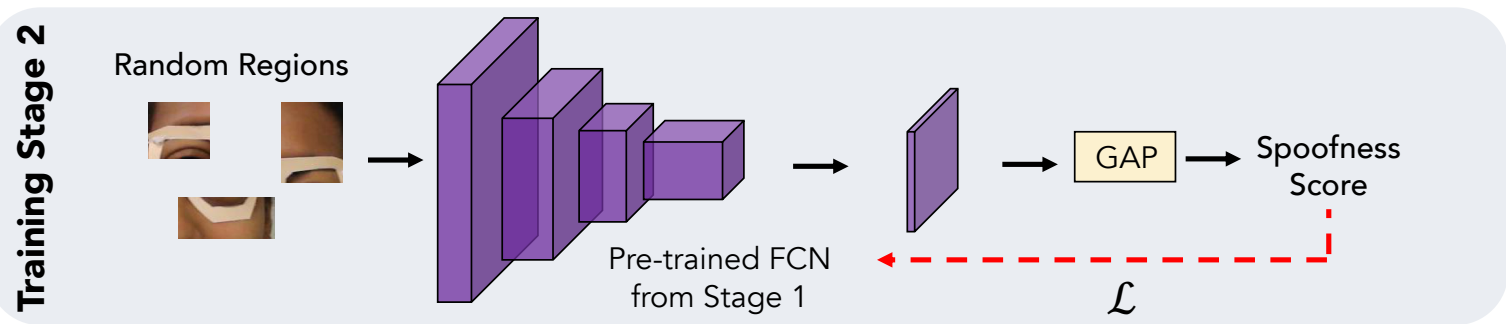
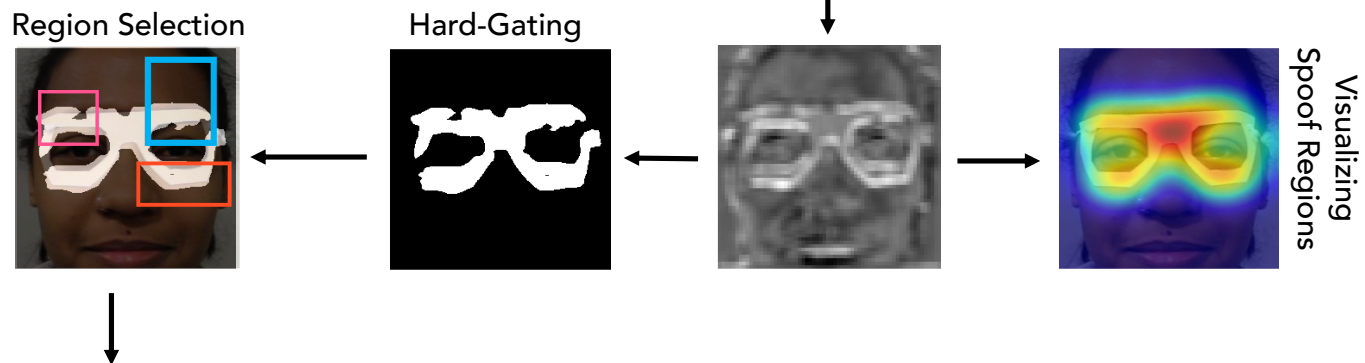
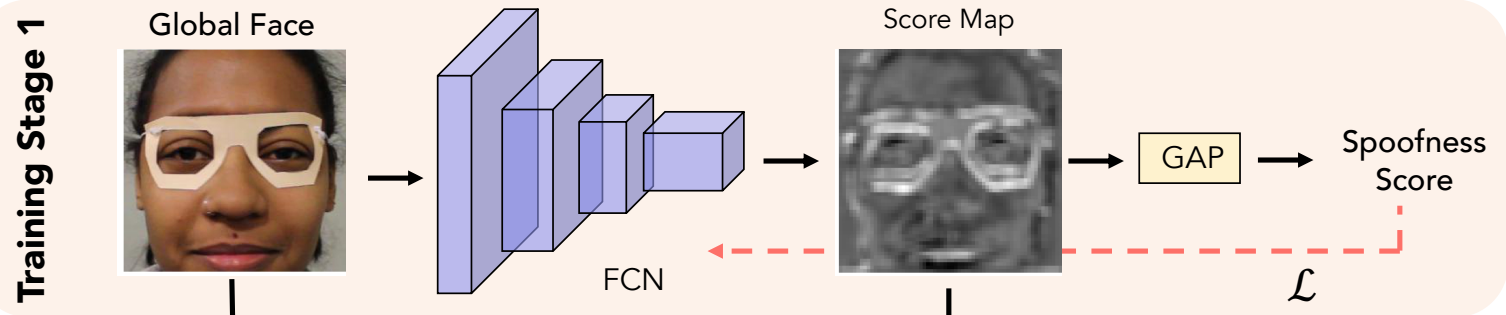


CNN



FCN

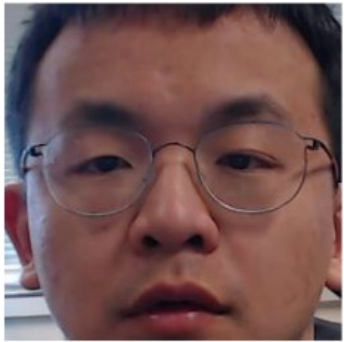
Method	Metric (%)	Replay	Obfuscation	Paper Glasses	Overall
CNN	ACER	13.3	47.1	32.2	30.8 $\pm$ 17.0
	EER	12.8	44.6	23.6	27.0 $\pm$ 13.2
FCN (Stage I)	ACER	11.2	52.2	12.1	25.1 $\pm$ 23.4
	EER	11.2	37.6	12.4	20.4 $\pm$ 12.1



Datasets

Dataset	Year	Statistics		PIE Variations			# Spoof Types					Total
		# Subj.	# Vids.	Pose	Expression Change	Illumination Change	Replay	Print	3D Mask	Makeup	Partial	
Replay-Attack	2012	50	1,200	Frontal	No	Yes	2	1	0	0	0	3
CASIA-FASD	2012	50	600	Frontal	No	No	1	1	0	0	0	2
3DMAD	2013	17	255	Frontal	No	No	0	0	1	0	0	1
MSU-MFSD	2015	35	440	Frontal	No	No	2	1	0	0	0	3
Replay-Mobile	2016	40	1,030	Frontal	No	Yes	1	1	0	0	0	2
HKBU-MARs	2016	35	1,009	Frontal	No	Yes	0	0	2	0	0	2
Oulu-NPU	2017	55	4,950	Frontal	No	Yes	2	2	0	0	0	4
SiW	2018	165	4,620	$[-90^\circ, 90^\circ]$	Yes	Yes	4	2	0	0	0	6
SiW-M	2019	493	1,630	$[-90^\circ, 90^\circ]$	Yes	Yes	1	1	5	3	3	13

Accurate Face Anti-Spoofing on Unknown Spoof Types



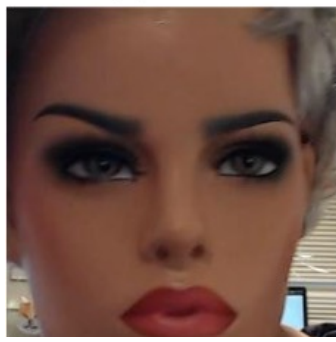
Live



Paper Mask



Half Mask



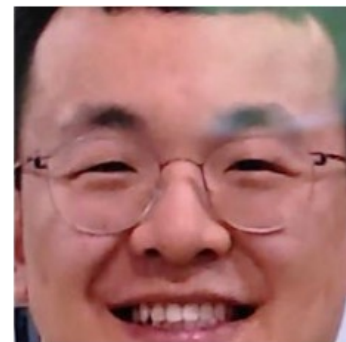
Mannequin



Paper Cut



Impersonation



Replay

TAR @ 2.0% FAR:

100%

96%

95%

95%

90%

72%



Silicone Mask

56%



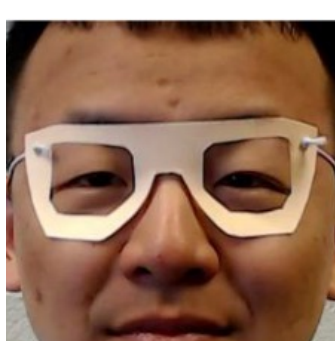
Print

51%



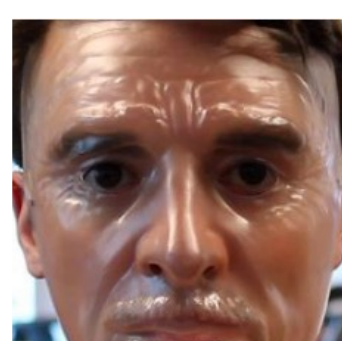
Cosmetic

44%



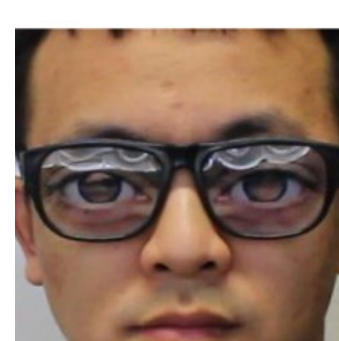
Paper Glasses

43%



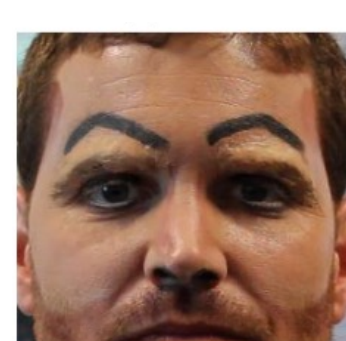
Transparent

39%



FunnyEye

33%

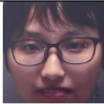
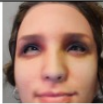
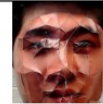
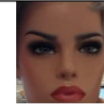
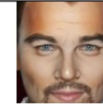


Obfuscation

31%

All spoofs except Mannequin, Impersonation, Transparent, and Obfuscation belong to the same person in Live

Results on SiW-M: Unknown Attack Protocol

		Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			
Method	Metric														Mean $\pm$ Std.
		Replay 99 vids.	Print 118 vids.	Half 72 vids.	Silicone 27 vids.	Trans. 88 vids.	Paper 17 vids.	Mann. 40 vids.	Obf. 23 vids.	Imp. 61 vids.	Cosm. 50 vids.	FunnyEye 160 vids.	Glasses 127 vids.	Paper Cut 86 vids.	
SVM+LBP	ACER	20.6	18.4	31.3	21.4	45.5	11.6	13.8	59.3	23.9	16.7	35.9	39.2	11.7	26.9 $\pm$ 14.5
	EER	20.8	18.6	36.3	21.4	37.2	7.5	14.1	51.2	19.8	16.1	34.4	33.0	7.9	24.5 $\pm$ 12.9
Auxiliary	ACER	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	23.6 $\pm$ 18.5
	EER	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	9.4	21.4	18.6	4.0	17.0 $\pm$ 17.7
DTN	ACER	9.8	6.0	15.0	18.7	36.0	4.5	7.7	48.1	11.4	14.2	19.3	19.8	8.5	16.8 $\pm$ 11.1
	EER	10.0	2.1	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	16.1 $\pm$ 12.2
CDC	ACER	10.8	7.3	9.1	10.3	18.8	3.5	5.6	42.1	0.8	14.0	24.0	17.6	1.9	12.7 $\pm$ 11.2
	EER	9.2	5.6	4.2	11.1	19.3	5.9	5.0	43.5	0.0	14.0	23.3	14.3	0.0	11.9 $\pm$ 11.8
Proposed	ACER	7.4	19.5	3.2	7.7	33.3	5.2	3.3	22.5	5.9	11.7	21.7	14.1	6.4	12.4 $\pm$ 9.2
	EER	6.8	11.2	2.8	6.3	28.5	0.4	3.3	17.8	3.9	11.7	21.6	13.5	3.6	10.1 $\pm$ 8.4
	TDR*	72.0	51.0	96.0	55.9	39.0	100.0	95.0	31.0	90.0	44.0	33.0	42.9	94.7	65.0 $\pm$ 25.9

*TDR evaluated at 2.0% FDR

Cross-Dataset Generalization

Half Total Error Rate (%)

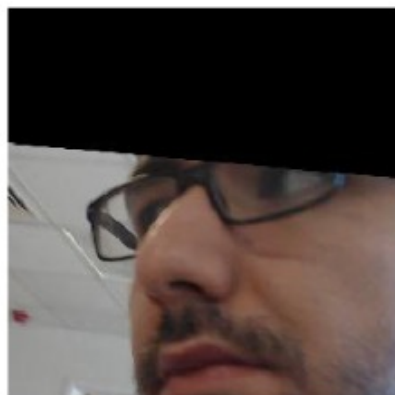
Method	CASIA → Replay	Replay → CASIA
CNN	48.5	45.5
Color Texture	47.0	49.6
FaceSpoofBuster	43.3	53.0
Auxiliary	27.6	28.4
De-Noising	28.5	41.1
Damer & Dimitrov	28.4	38.1
STASN	31.5	30.9
SAPLC	27.3	37.5
<i>SSR-FCN</i> (Proposed)	19.9	41.9

“CASIA→Replay” denotes training on CASIA and testing on Replay-Attack

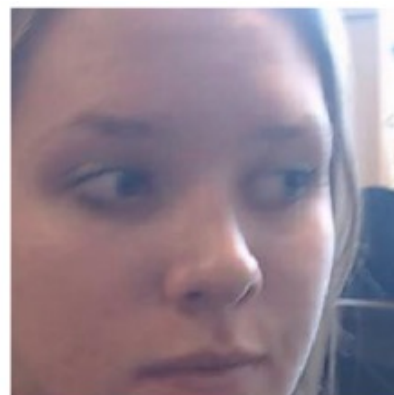
Failure Cases



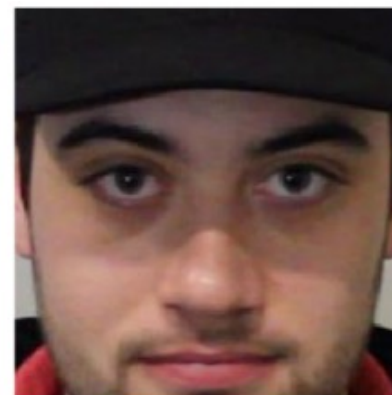
0.91



0.99

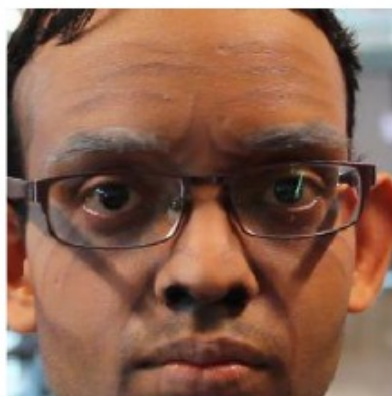


0.58

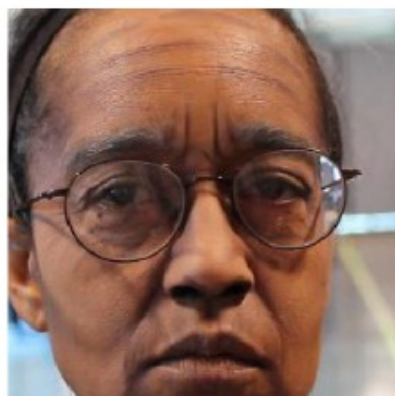


0.72

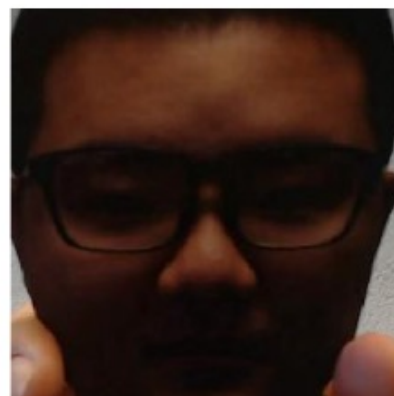
(a) Misclassified Lives as Spoofs



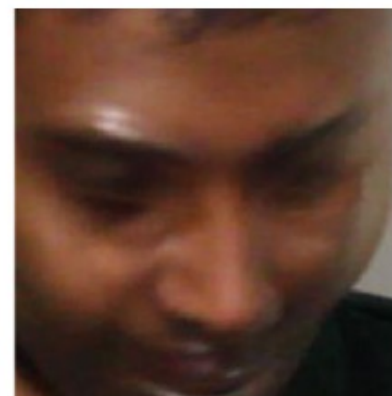
0.14



0.18



0.43

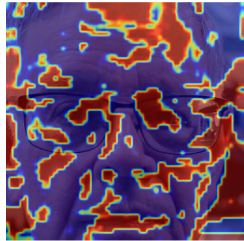


0.36

(b) Misclassified Spoofs as Lives

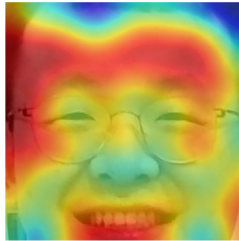
Visualization

Live



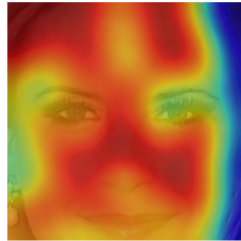
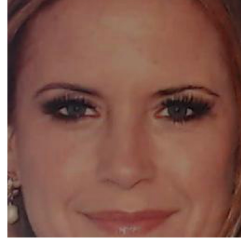
0.00

Replay



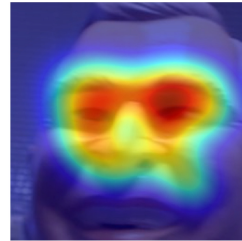
0.99

Print



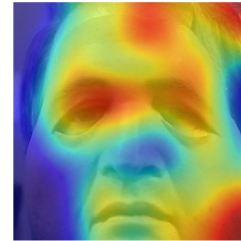
0.99

Half-Mask



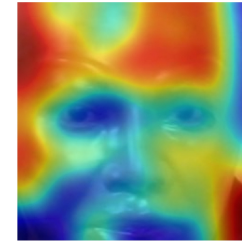
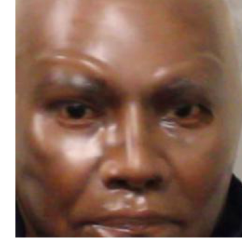
0.99

Silicone Mask



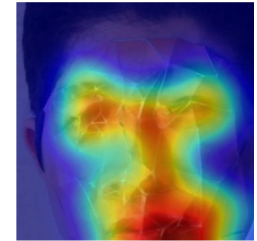
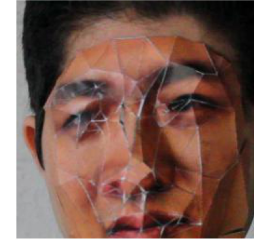
0.73

Trans. Mask



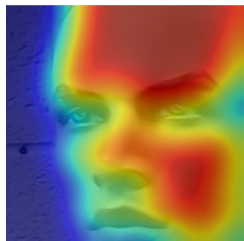
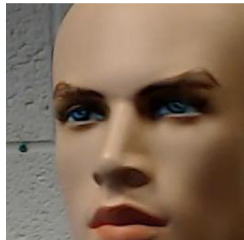
0.79

Paper Mask



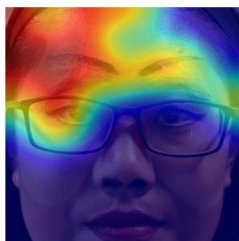
0.99

Mannequin



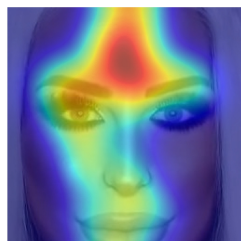
0.99

Obfuscation



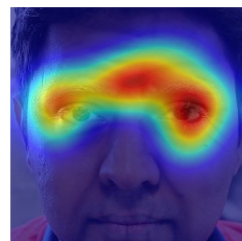
0.62

Impersonation



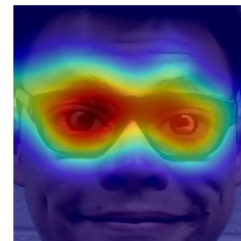
0.99

Cosmetic



0.91

FunnyEye



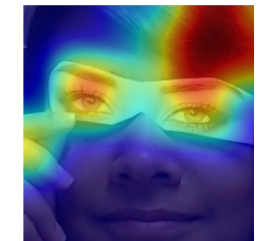
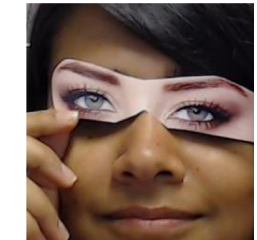
0.98

Paper Glasses



0.99

Paper Cut

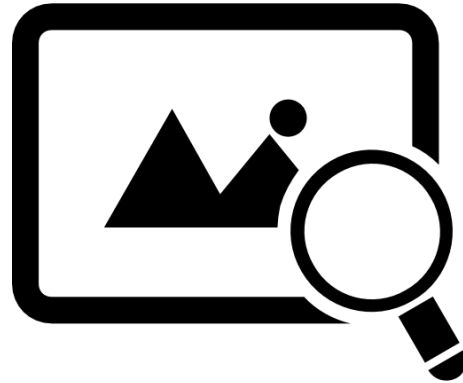


0.99

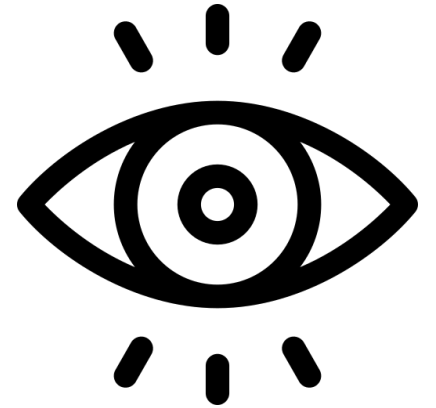
Summary



Global Guidance



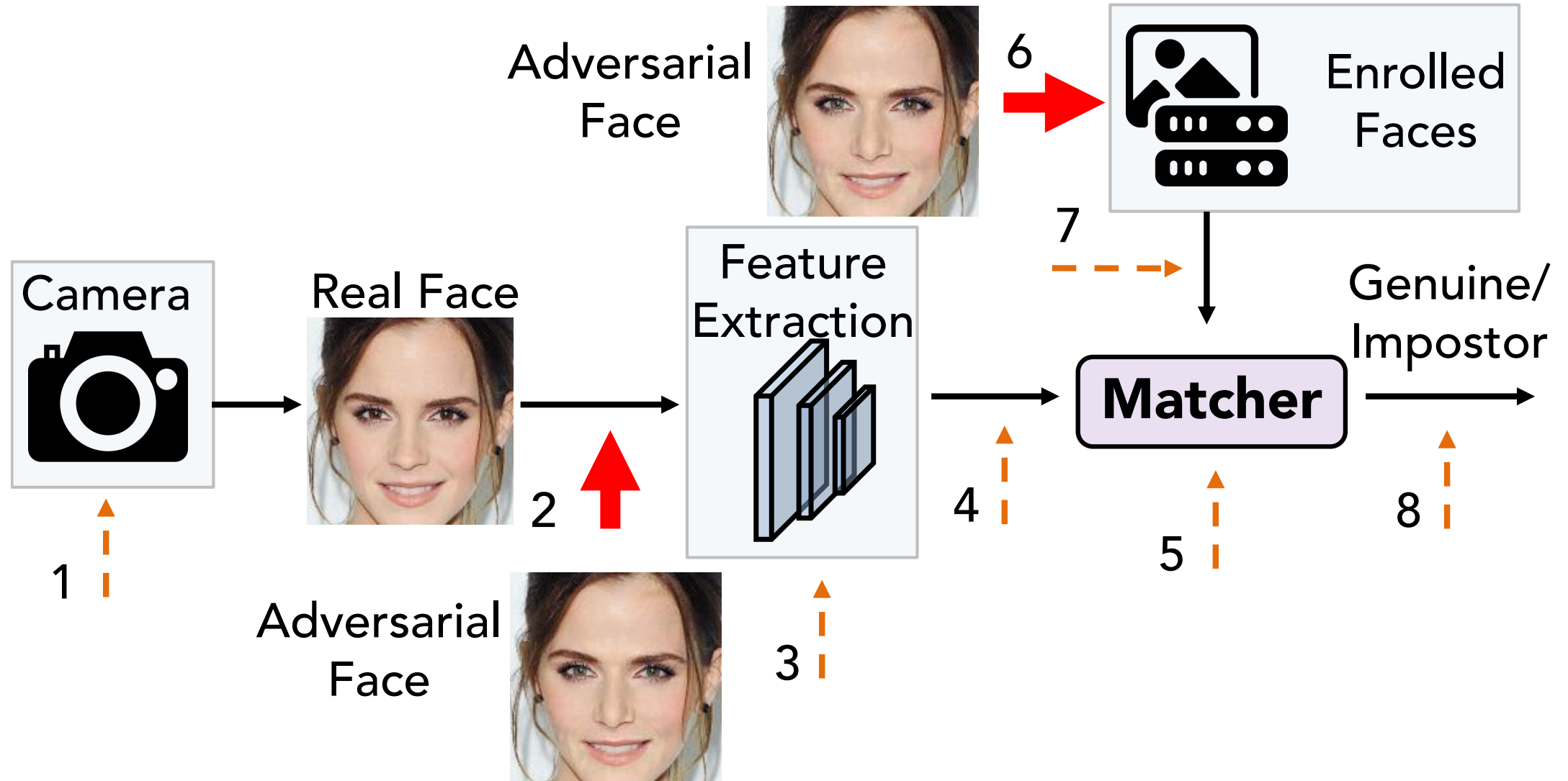
Local Features



Interpretable

Adversarial Attacks

Adversarial Attack on AFR Systems



Adversarial Faces

Enrolled Face



Input Probe



0.72

PGD



0.26

FGSM



0.14



0.78



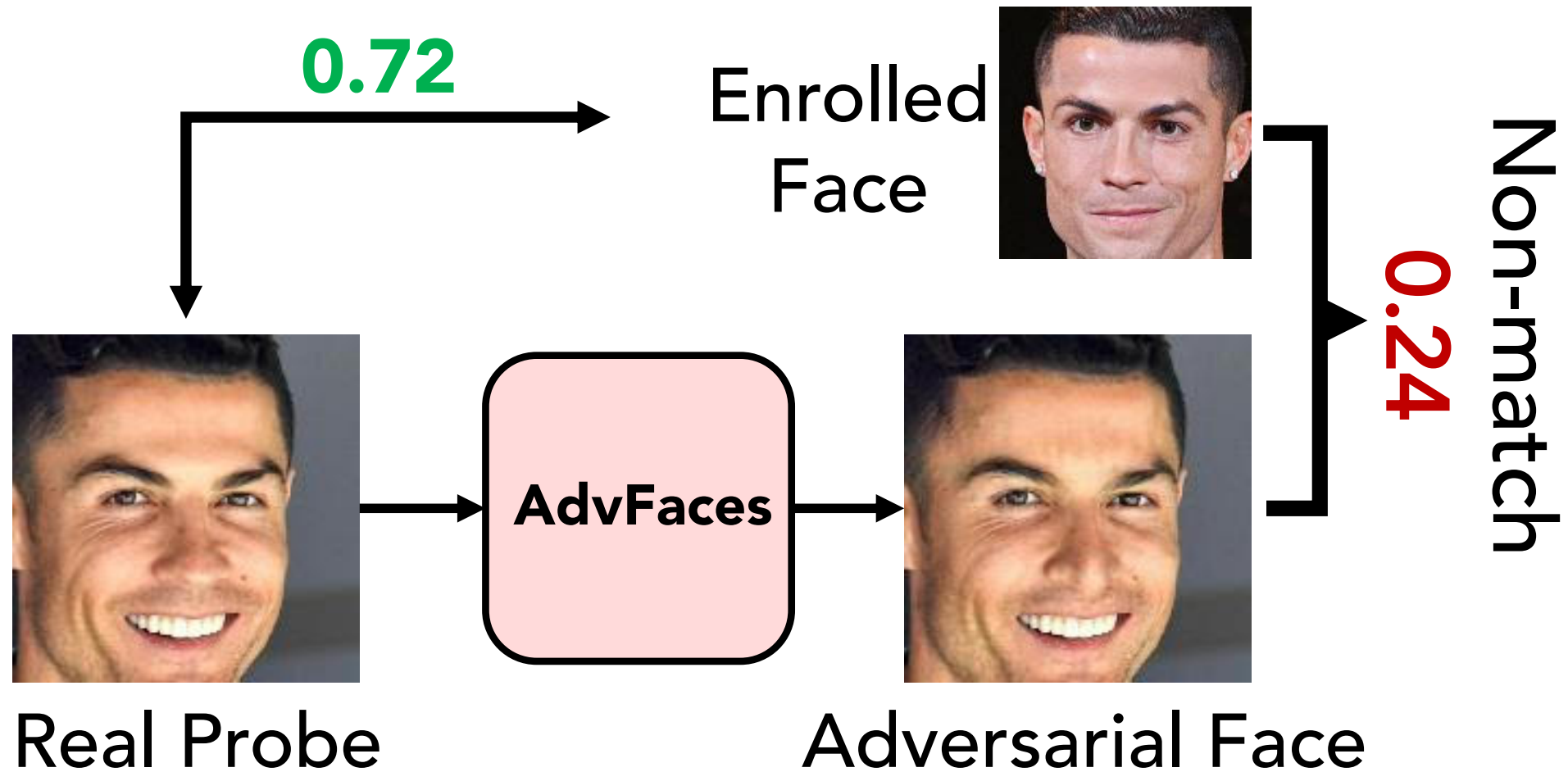
0.25



0.25

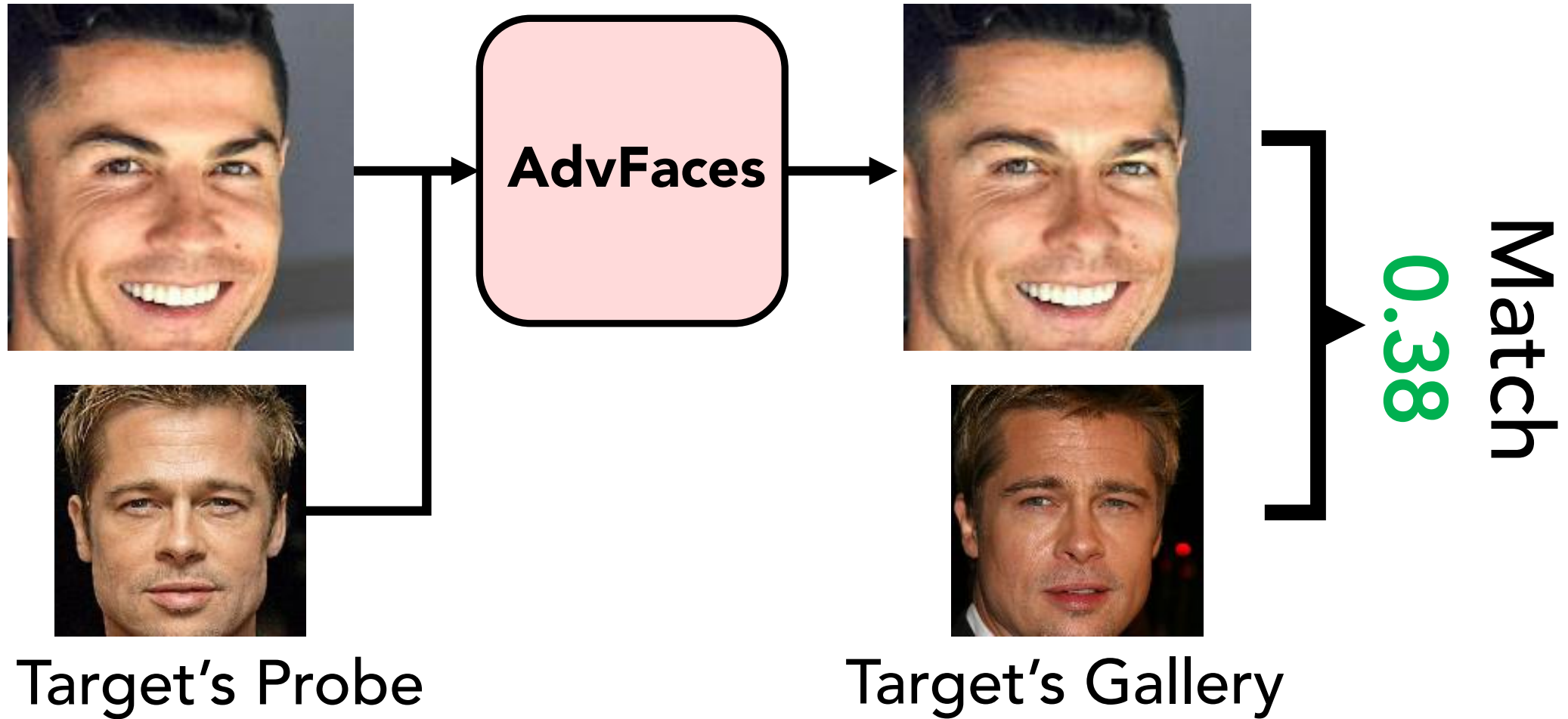
Cosine similarity scores $\in [-1, 1]$ via ArcFace . **Score > 0.28** is accepted as genuine (threshold @ 0.1% FAR)

Obfuscation Attack



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

Impersonation Attack

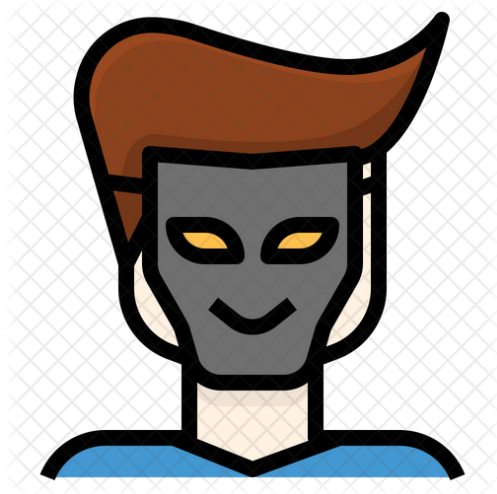


Cosine similarity scores $\in [-1, 1]$ via ArcFace . **Score** > 0.28 is accepted as genuine (threshold @ 0.1% FAR)

Requirements of an Adversarial Face Generator



01 Visually Realistic



02 Successful Attack

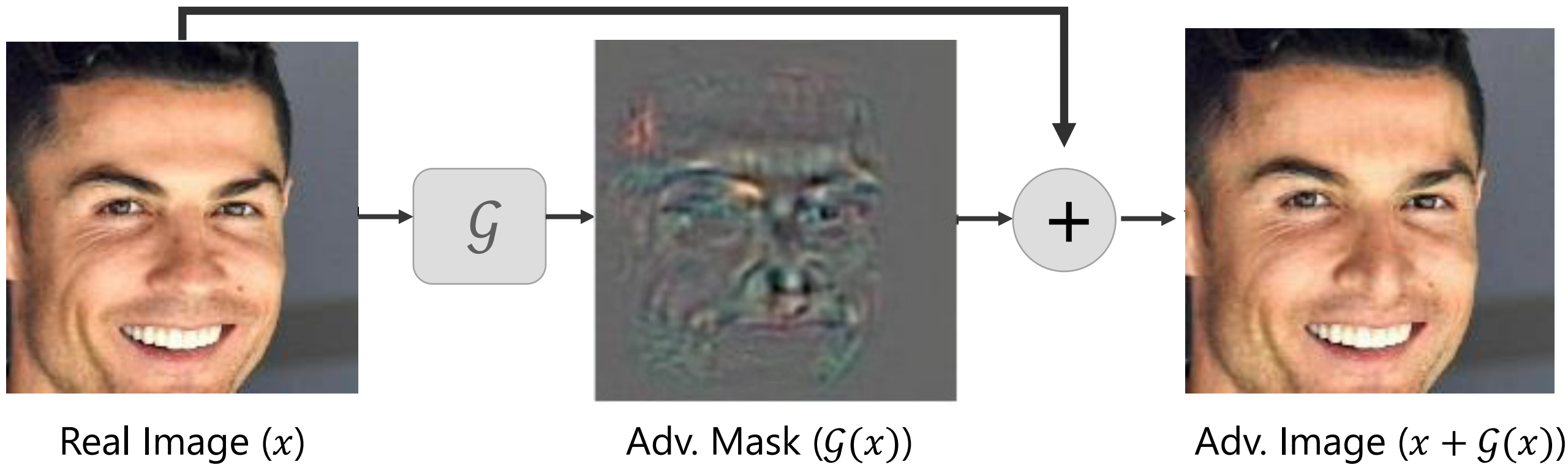


03 Controllable



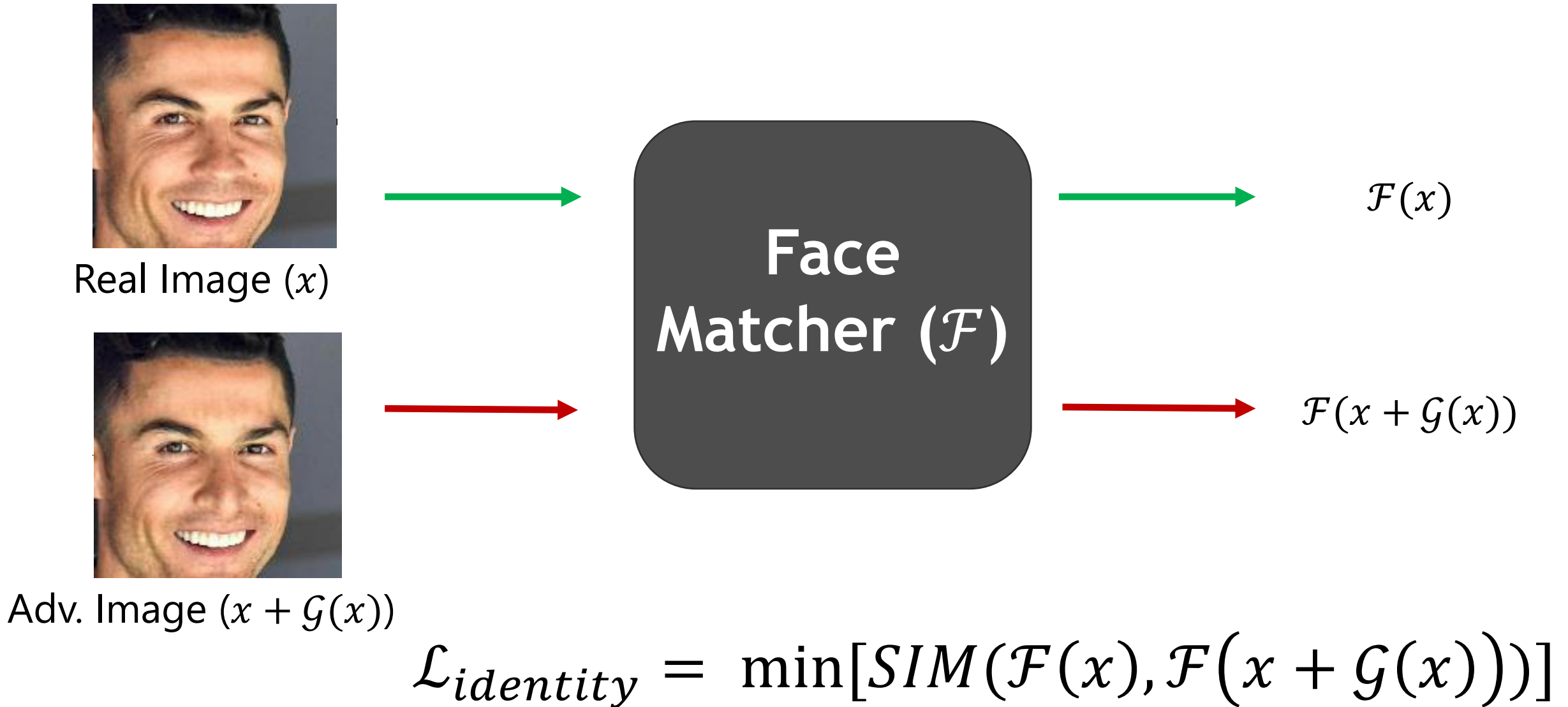
04 Transferable

1. Visual Realism

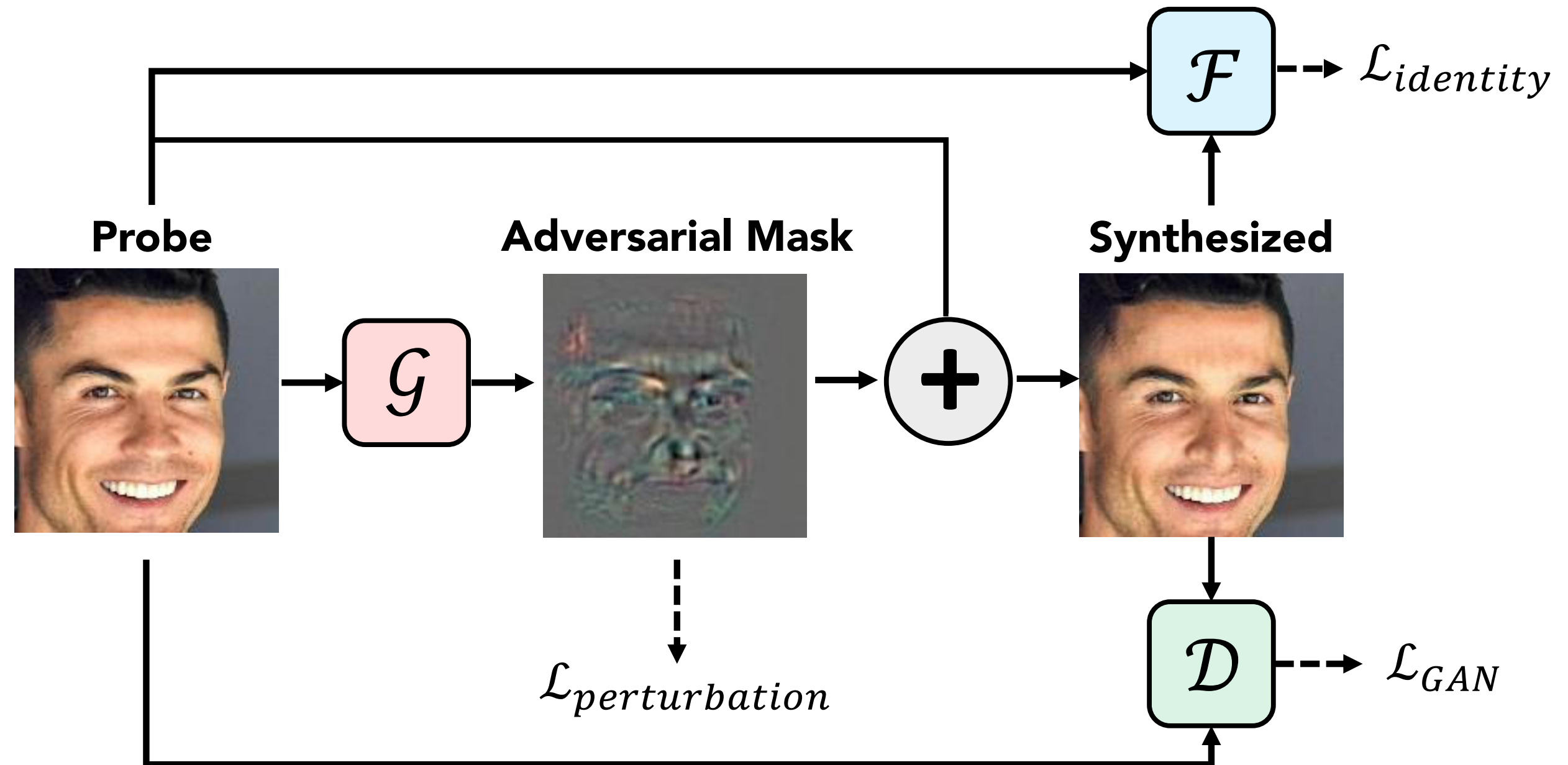


$$\mathcal{L}_{perturbation} = \min ||\max(\varepsilon, G(x))||_2 \quad , \varepsilon \in (0, \infty)$$

2. Attack Success (Obfuscation)



Training AdvFaces (Obfuscation)



Examples of Obfuscation Attack

Gallery

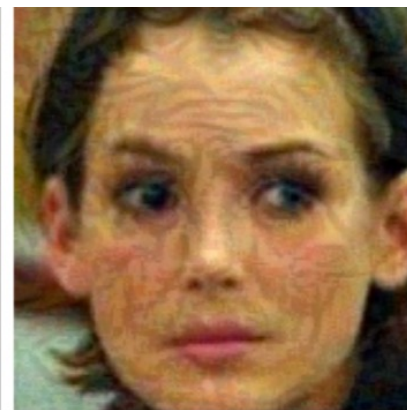
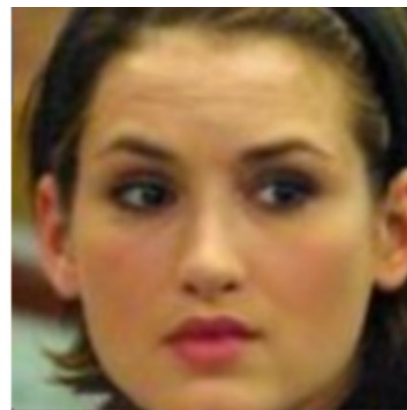
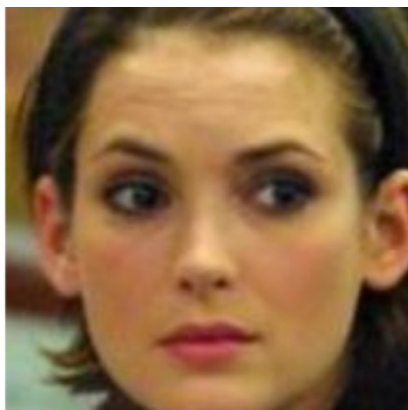
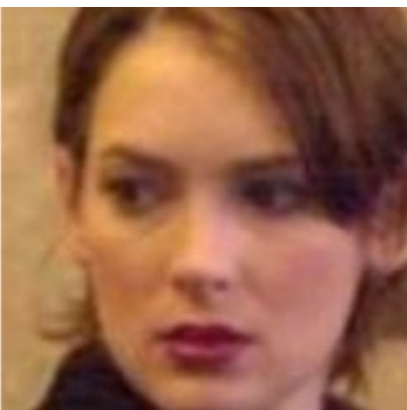
Probe

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)

What is AdvFaces Learning?

Probe



Adv. Mask



Visualization



Adv. Image



Evaluation Protocol



Enrolled



Adv. Probe

Obfuscation

$$\text{Attack Success Rate} = \frac{\# \text{ of Comparisons} < \tau}{\text{Total \# of Comparisons}}$$



Target
Enrolled



Adv. Probe

Impersonation

$$\text{Attack Success Rate} = \frac{\# \text{ of Comparisons} \geq \tau}{\text{Total \# of Comparisons}}$$

Datasets

Training: CASIA-WebFace



494,414 images of 10,574 subjects

Testing: LFW



13,233 images of 5,749 subjects
Filter -> 9,164 images of 1,680
subjects

Obfuscation Attack: 484,514 comparisons

Quantitative Results: Obfuscation

	AdvFaces	GFLM	PGD	FGSM
Attack Success Rate (%) @ 0.1% False Accept Rate				
FaceNet	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

Pose Variations

0.019



-0.172

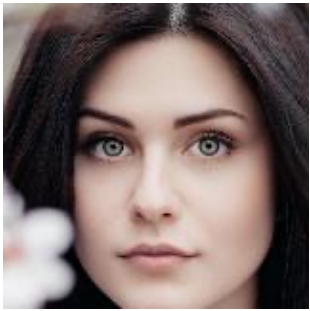


-0.327



FaceGuard: A Generalizable Defense Against Adversarial Faces

Real



AdvFaces



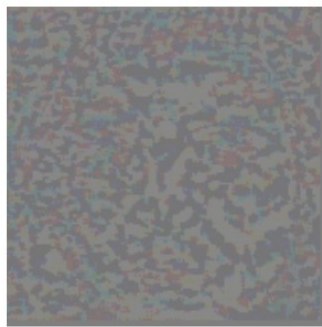
0.21



PGD



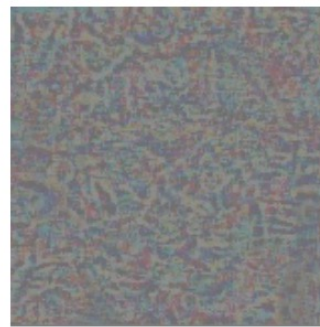
0.27



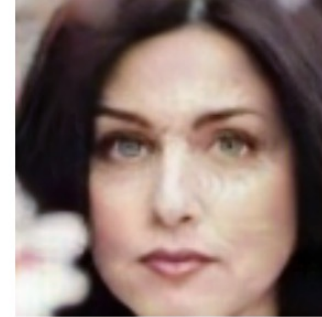
FGSM



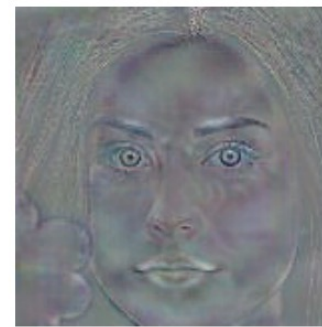
0.28



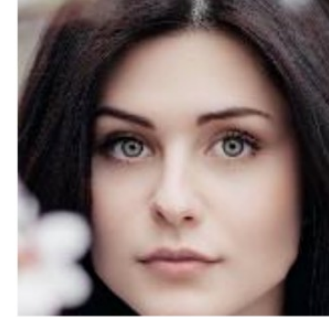
SemanticAdv



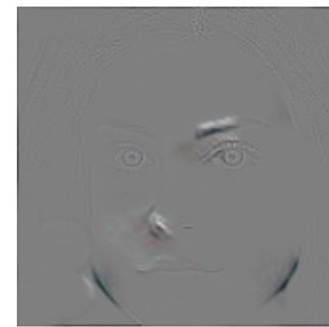
0.32



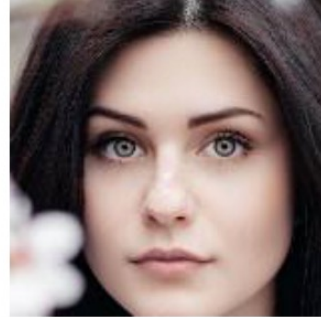
GFLM



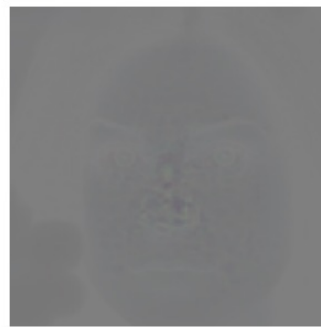
0.34



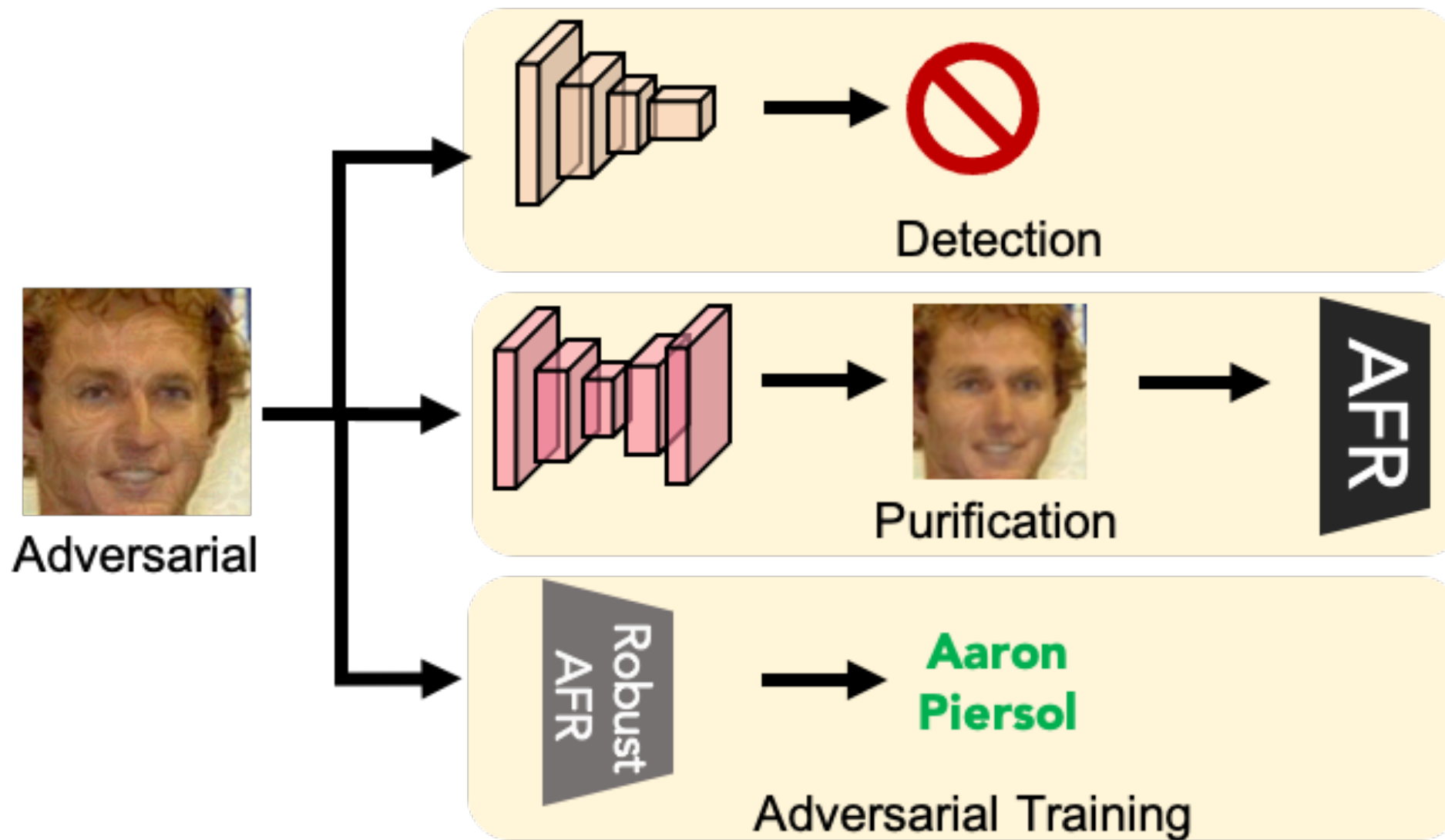
DeepFool



0.35



Prevailing Defenses

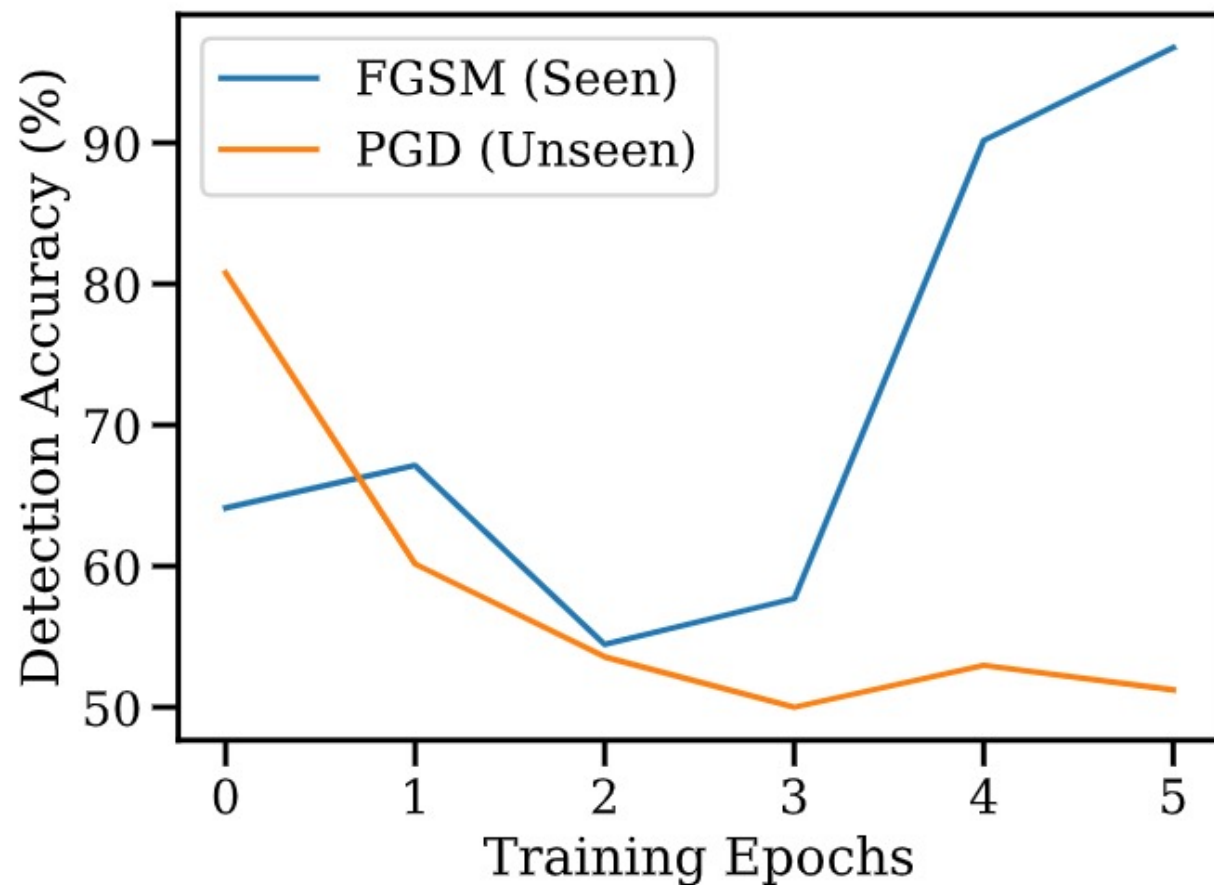


Prevailing Approaches Overfit

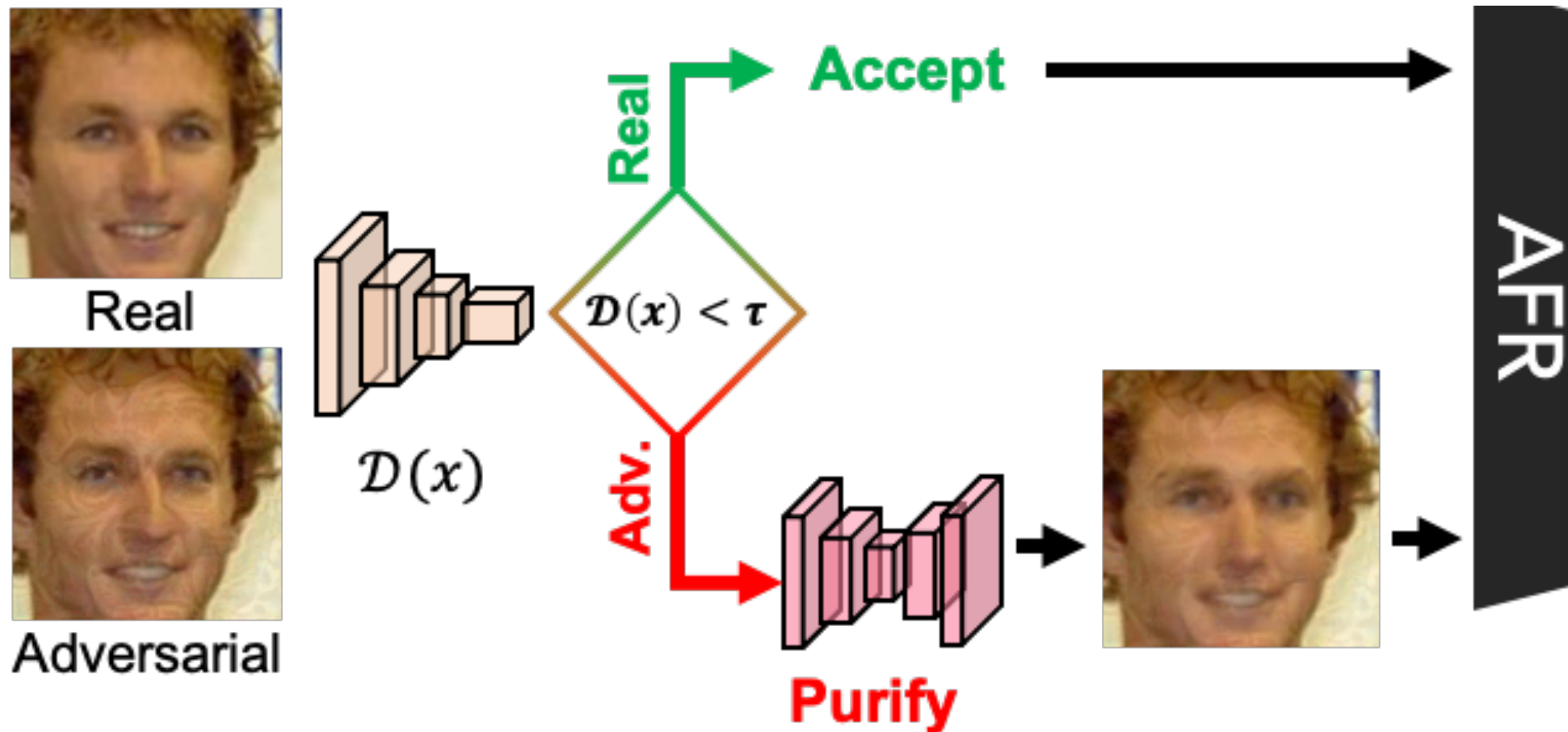
Robustness (Adversarial Training)



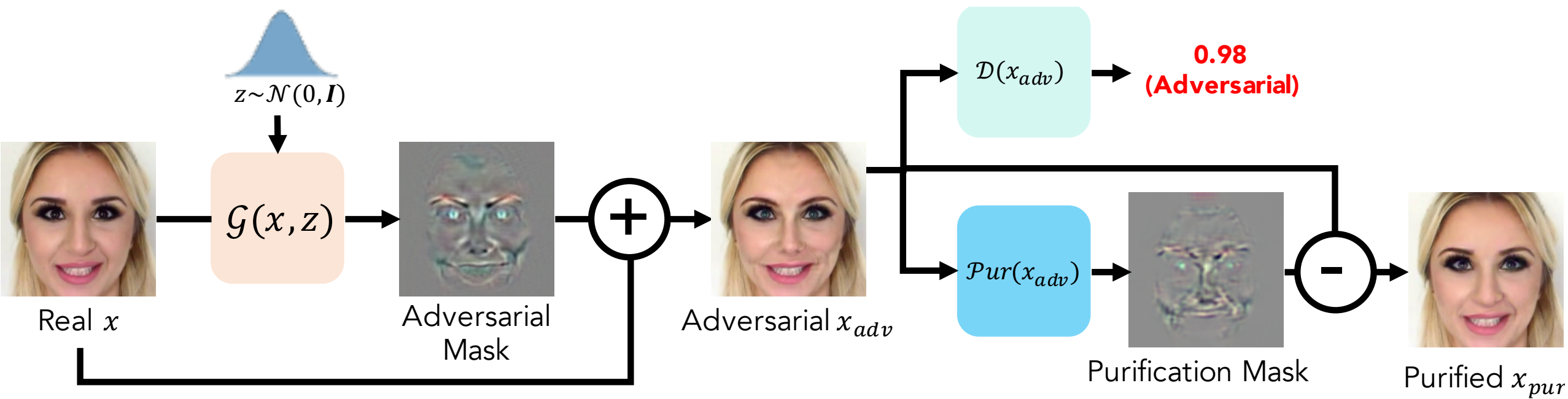
Binary CNN - Detection



Proposed FaceGuard



Proposed FaceGuard



Detector Results

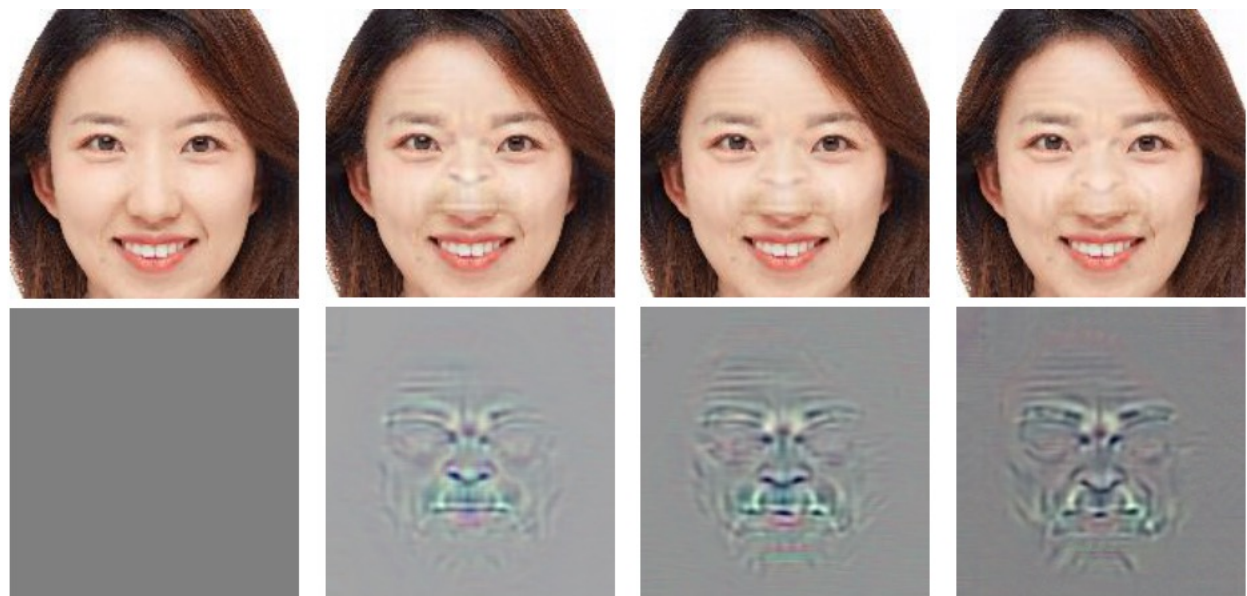
Detection Accuracy (%)		Year	FGSM	PGD	DpFl.	AdvF.	GFLM	Smnt.	Mean $\pm$ Std.
General	Gong <i>et al.</i>	2017	98.94	97.91	95.87	92.69	99.92	99.92	97.54 $\pm$ 02.82
	ODIN	2018	83.12	84.39	71.74	50.01	87.25	85.68	77.03 $\pm$ 14.34
	Steganalysis	2019	88.76	89.34	75.97	54.30	58.99	78.62	74.33 $\pm$ 14.77
Face	UAP-D	2018	61.32	74.33	56.78	51.11	65.33	76.78	64.28 $\pm$ 09.97
	SmartBox	2018	58.79	62.53	51.32	54.87	50.97	62.14	56.77 $\pm$ 05.16
	Goswami <i>et al.</i>	2019	84.56	91.32	89.75	76.51	52.97	81.12	79.37 $\pm$ 14.04
	Massoli <i>et al.</i> (MLP)	2020	63.58	76.28	81.78	88.38	51.97	52.98	69.16 $\pm$ 15.29
	Massoli <i>et al.</i> (LSTM)	2020	71.53	76.43	88.32	75.43	53.76	55.22	70.11 $\pm$ 13.35
	Agarwal <i>et al.</i>	2020	94.44	95.38	91.19	74.32	51.68	87.03	87.03 $\pm$ 16.86
Proposed FaceGuard		2021	99.85	99.85	99.85	99.84	99.61	99.85	99.81 $\pm$ 00.10

Purifier Results

Defenses	Year	Strategy 485K pairs	Real 3M pairs	Attacks
No-Defense	—	-	99.82	34.27
Adv. Training	2017	Robustness	96.42	11.23
Rob-GAN	2019	Robustness	91.35	13.89
Feat. Denoising	2019	Robustness	87.61	17.97
L2L	2019	Robustness	96.89	16.76
MagNet	2017	Purification	94.47	38.32
DefenseGAN	2018	Purification	96.78	39.21
Feat. Distillation	2019	Purification	94.64	41.77
NRP	2020	Purification	97.54	61.44
A-VAE	2020	Purification	93.71	51.99
<i>Proposed FaceGuard</i>	2021	Purification	99.81	77.46

Generator Quality

Within Iteration



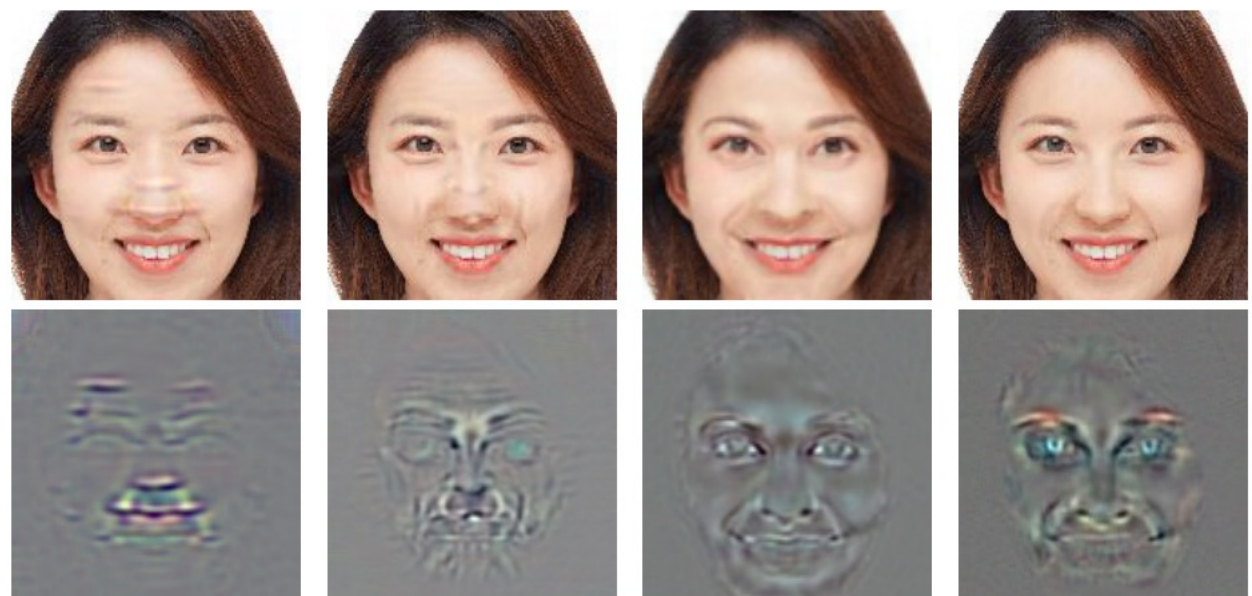
Input Probe ($\mathbf{x}$)

$\mathcal{G}(\mathbf{x}, \mathbf{z}_1)$

$\mathcal{G}(\mathbf{x}, \mathbf{z}_2)$

$\mathcal{G}(\mathbf{x}, \mathbf{z}_3)$

Across Iteration



Iteration: 5K

Iteration: 20K

Iteration: 60K

Iteration: 100K

Visualizing Perturbed Regions



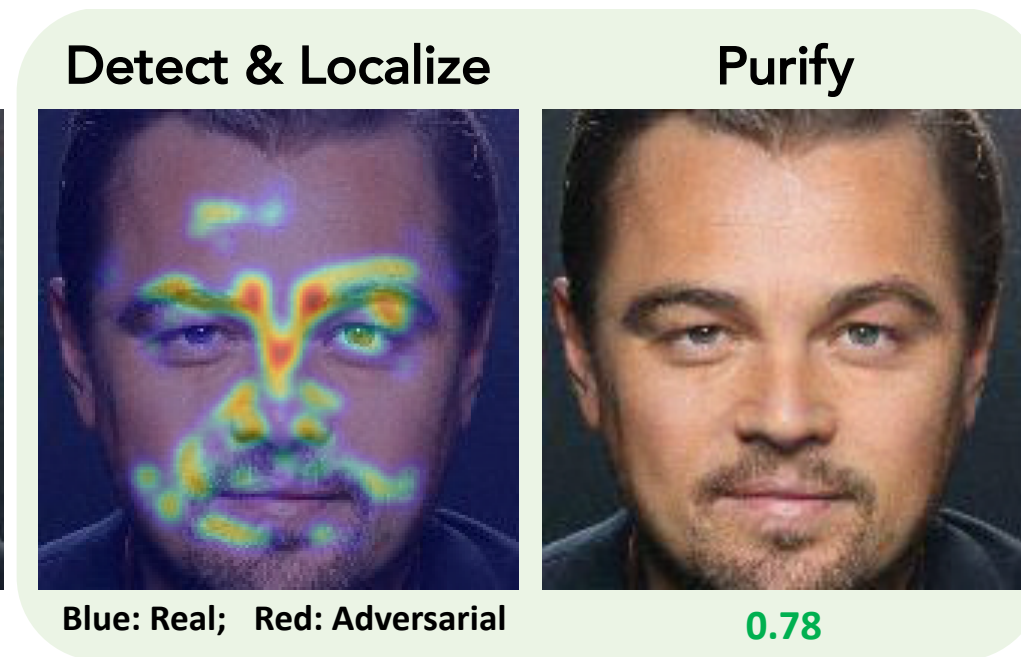
(a) Enrolled



0.85
(b) Probe

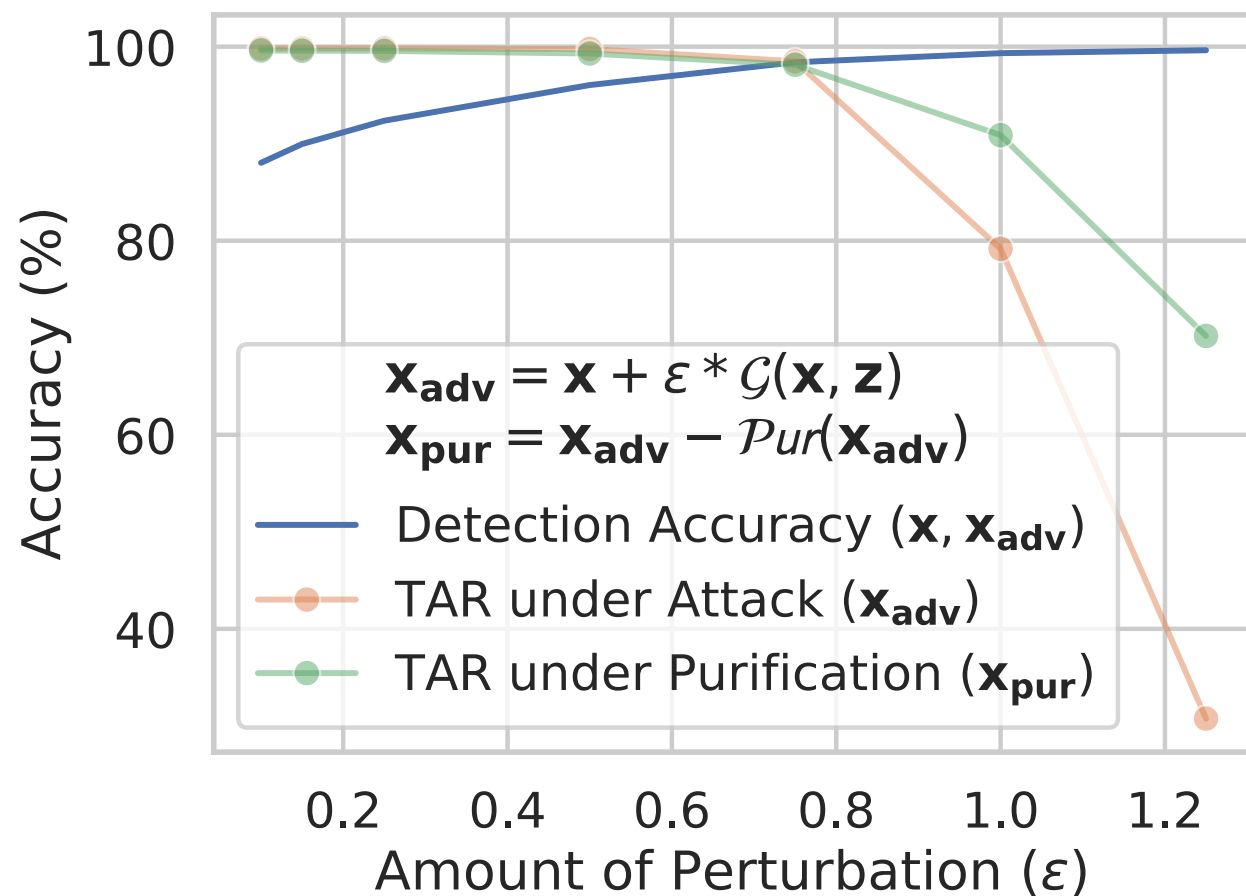
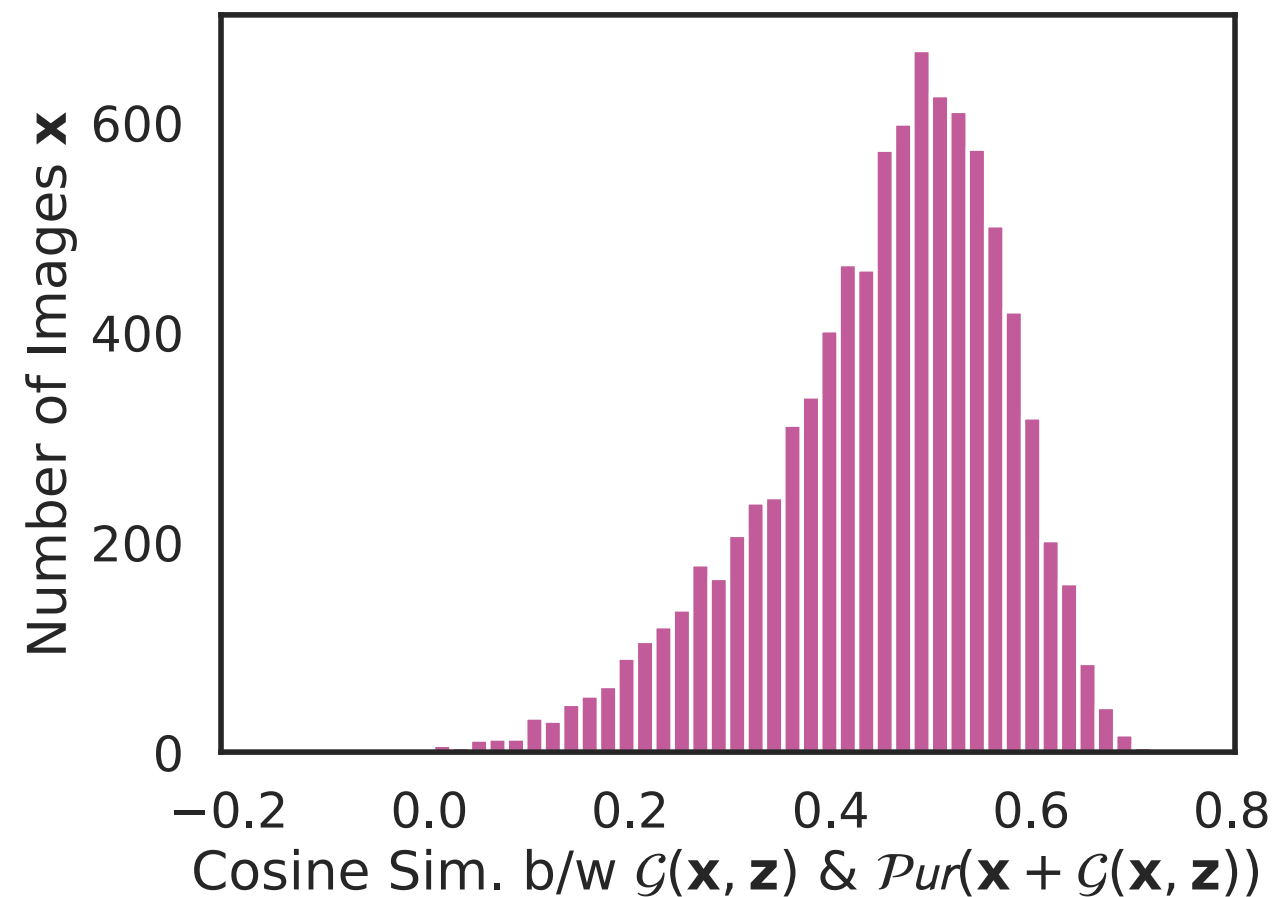


0.29
(c) AdvFaces

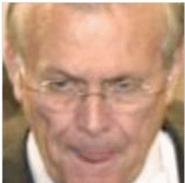
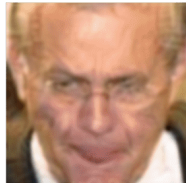
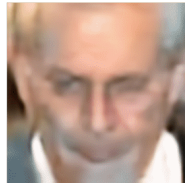
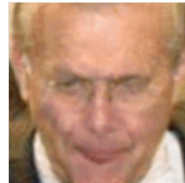
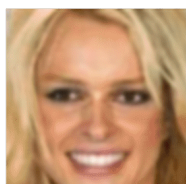
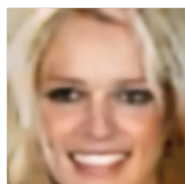
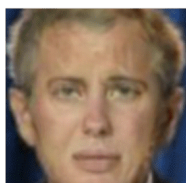
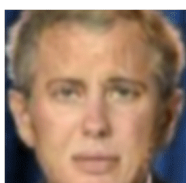
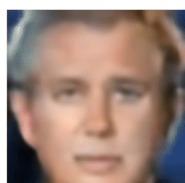
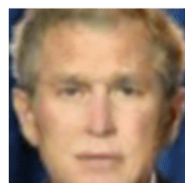
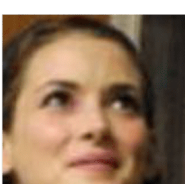
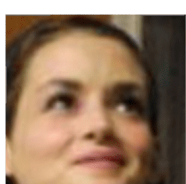
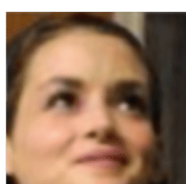
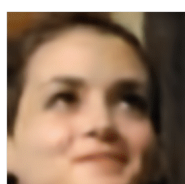
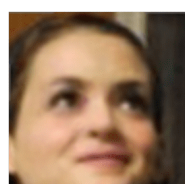
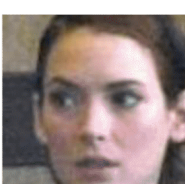
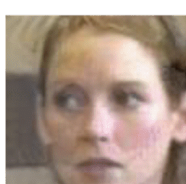
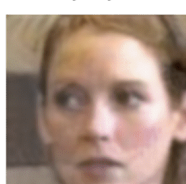
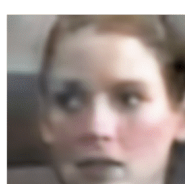
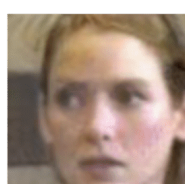


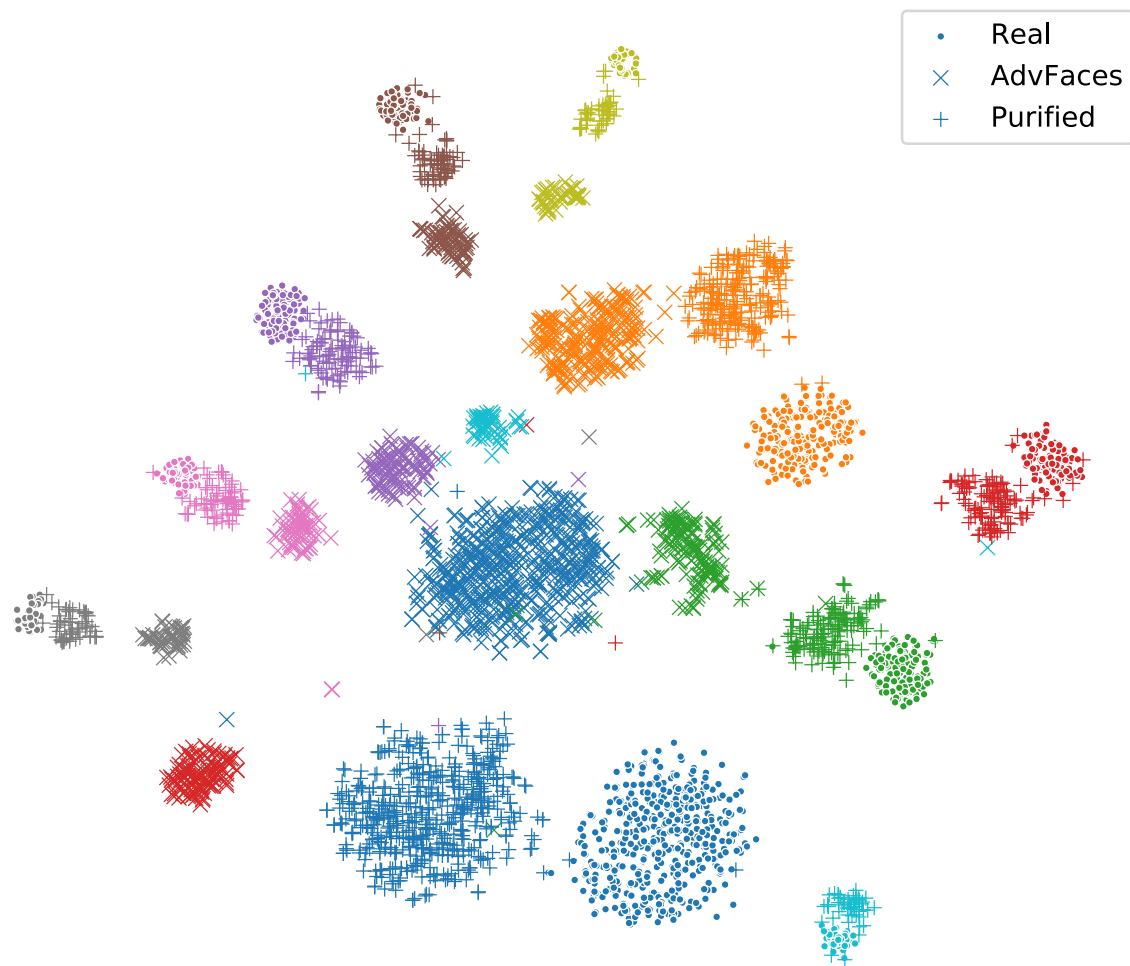
(d) Proposed FaceGuard

Purifier Quality

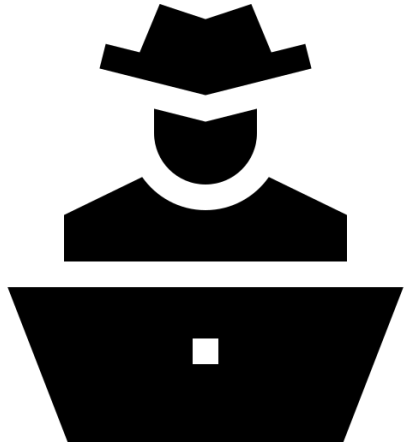


Purifier: Qualitative Results

	Real Probe	Adversarial	MagNet	DefenseGAN	FaceGuard
FGSM					
		0.19	0.26	0.31	0.48
PGD					
		-0.23	0.32	0.16	0.39
DeepFool					
		-0.24	0.26	0.34	0.41
AdvFaces					
		-0.38	-0.33	-0.31	0.54
GFLM					
		0.30	0.29	0.32	0.46
SemanticAdv					
		0.11	0.28	0.18	0.52



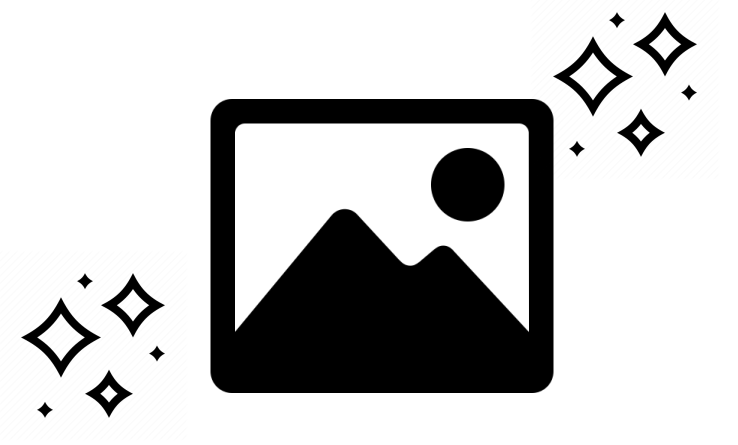
Summary



SOTA Adversarial Synthesizer



SOTA Adversarial Detector

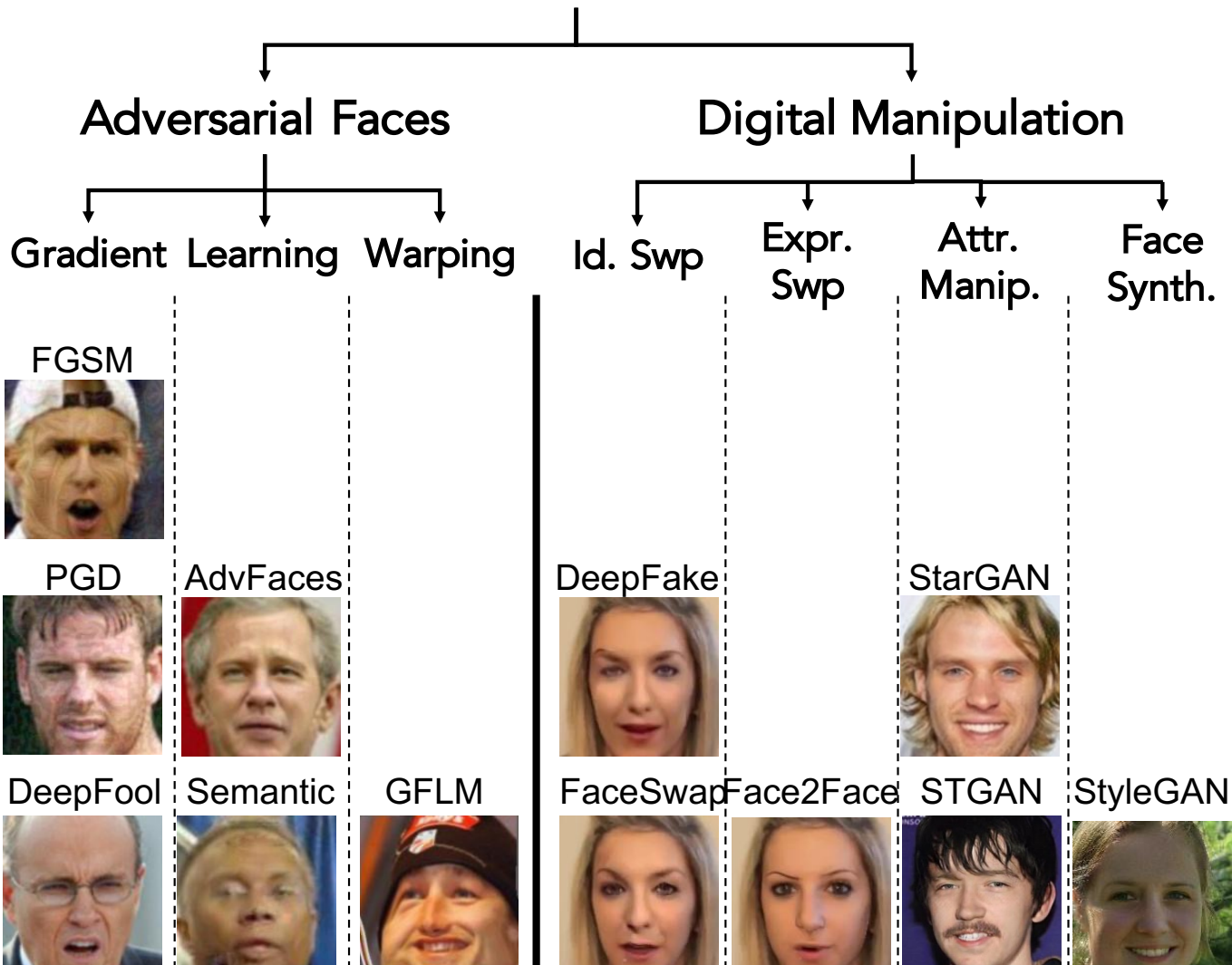


SOTA Adversarial Purifier

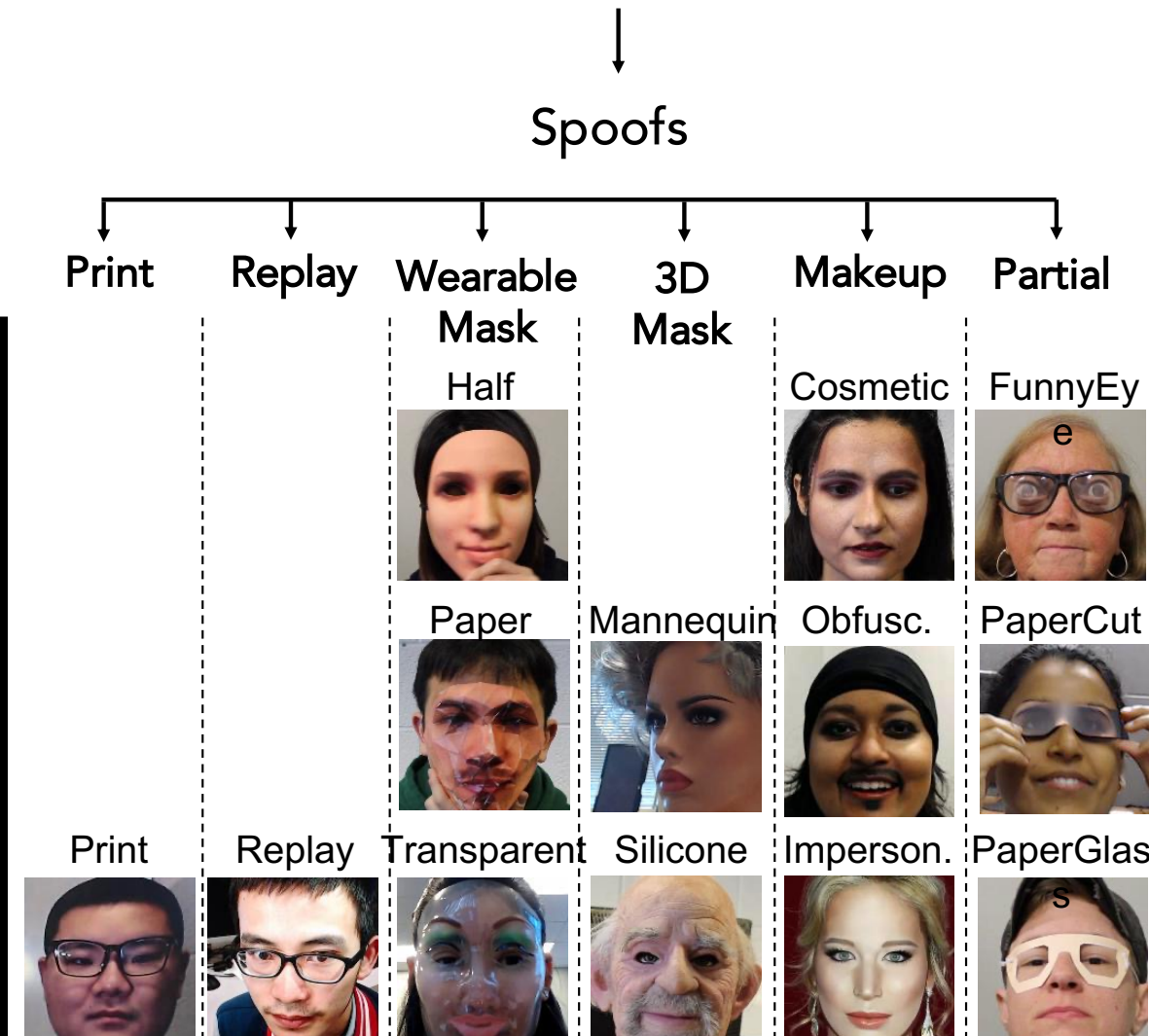
Unified Detector

Unified Detection of Digital and Physical Face Attacks

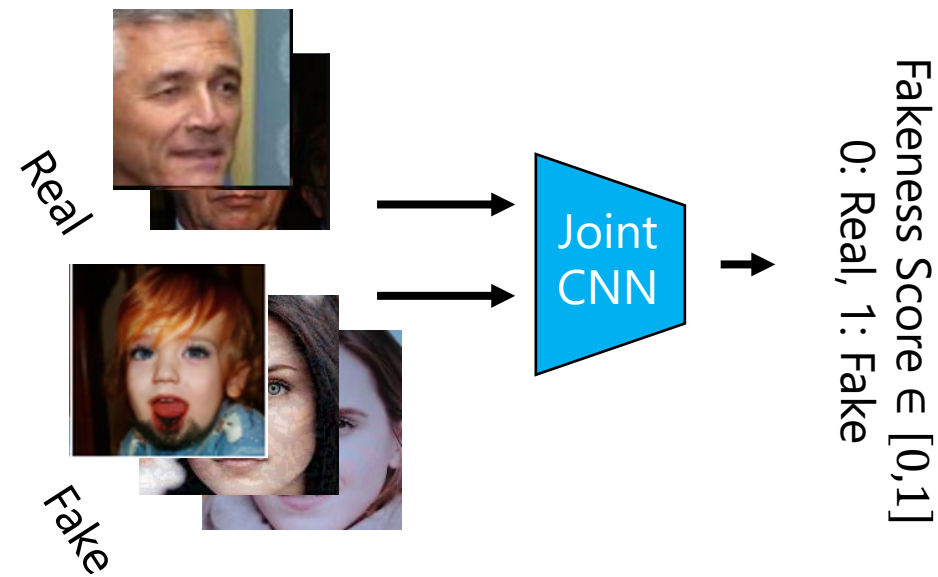
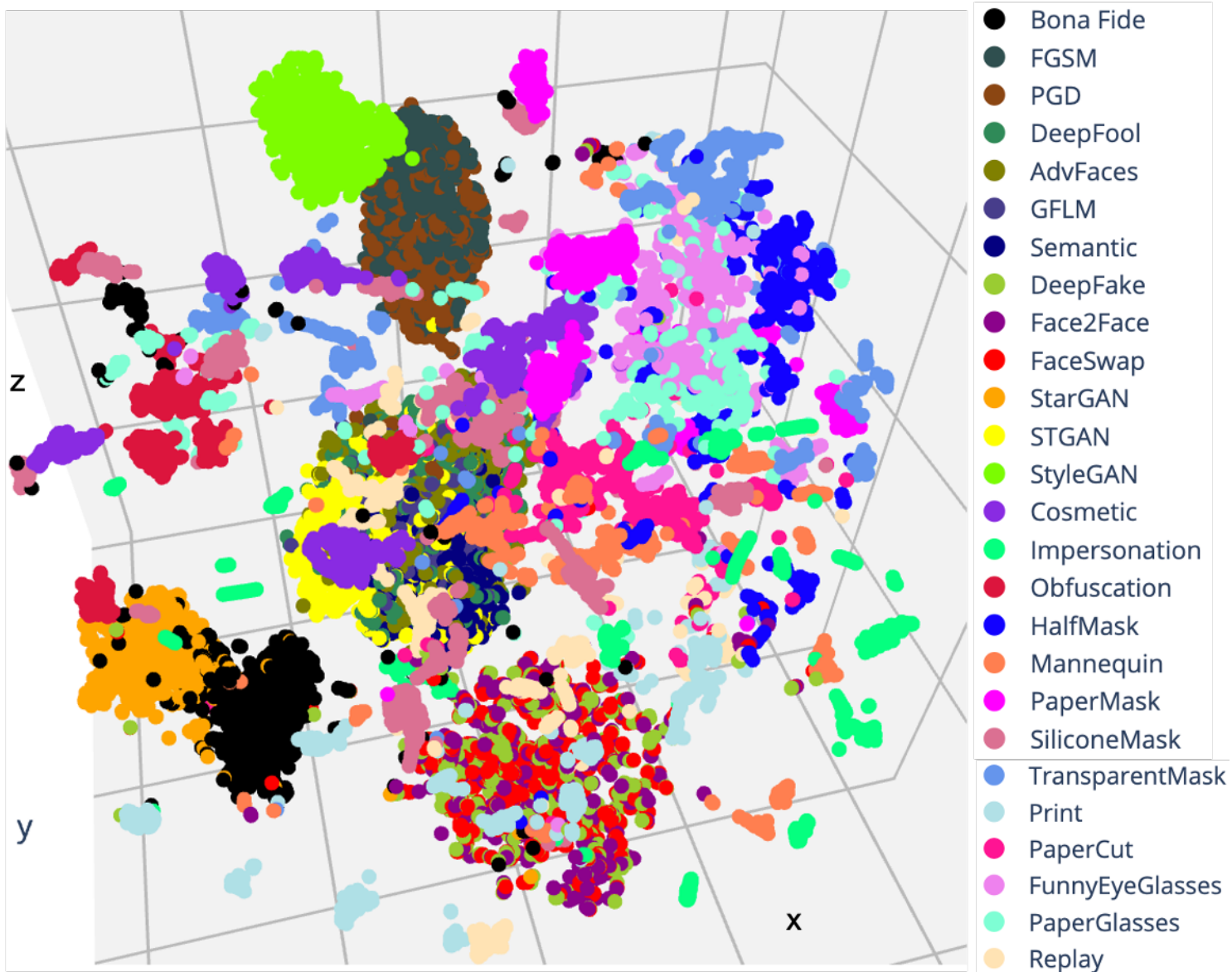
Digital Attacks



Physical Attacks

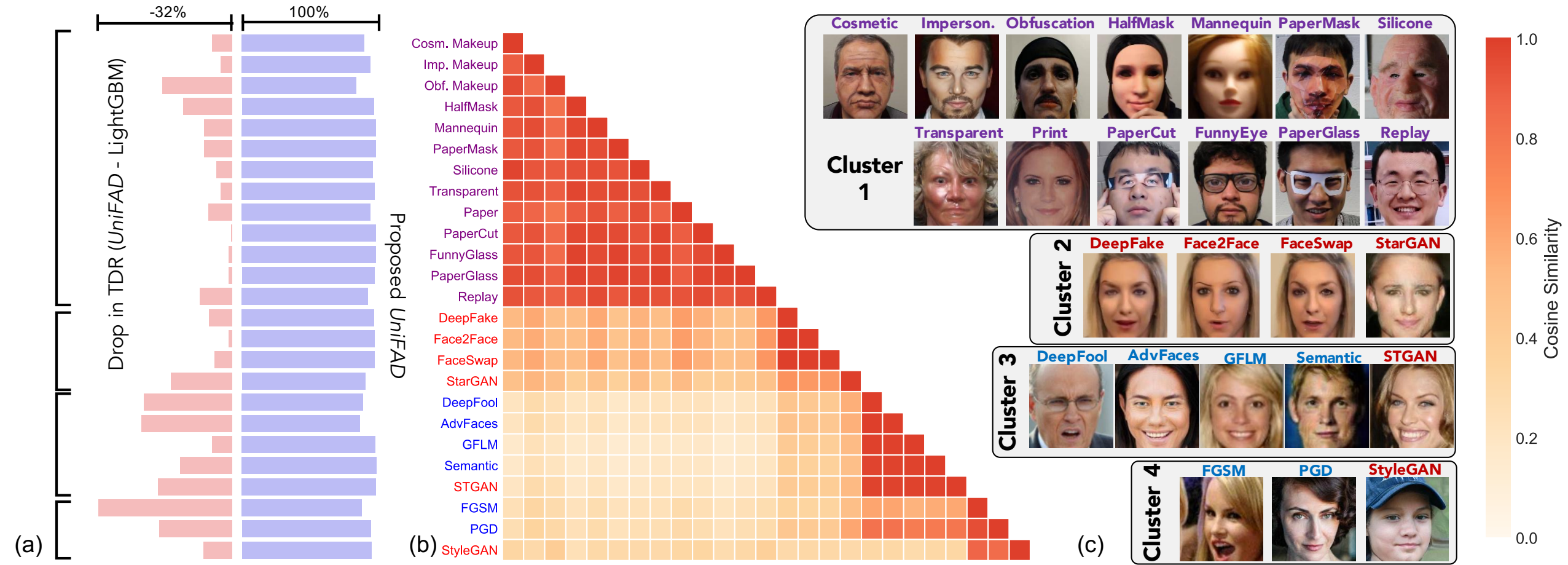


JointCNN

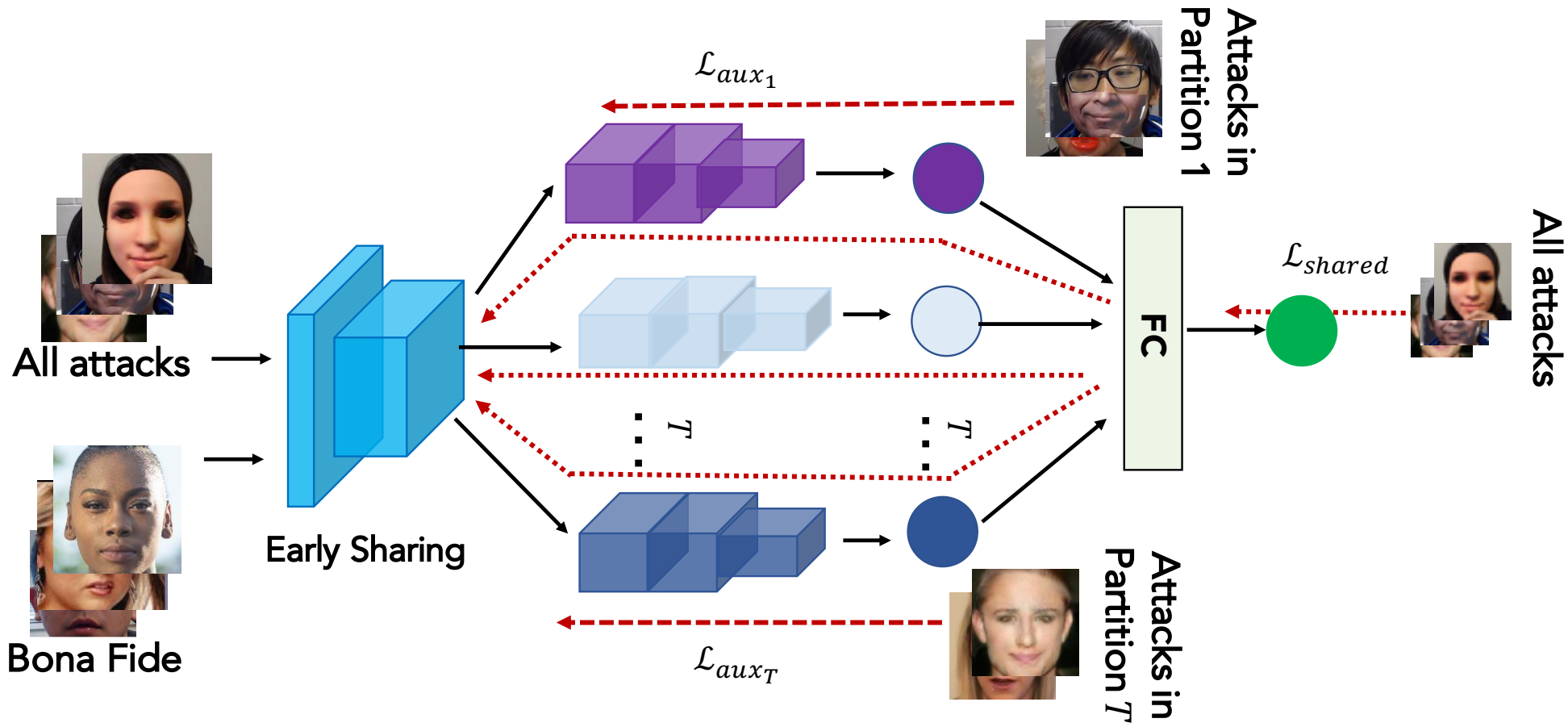
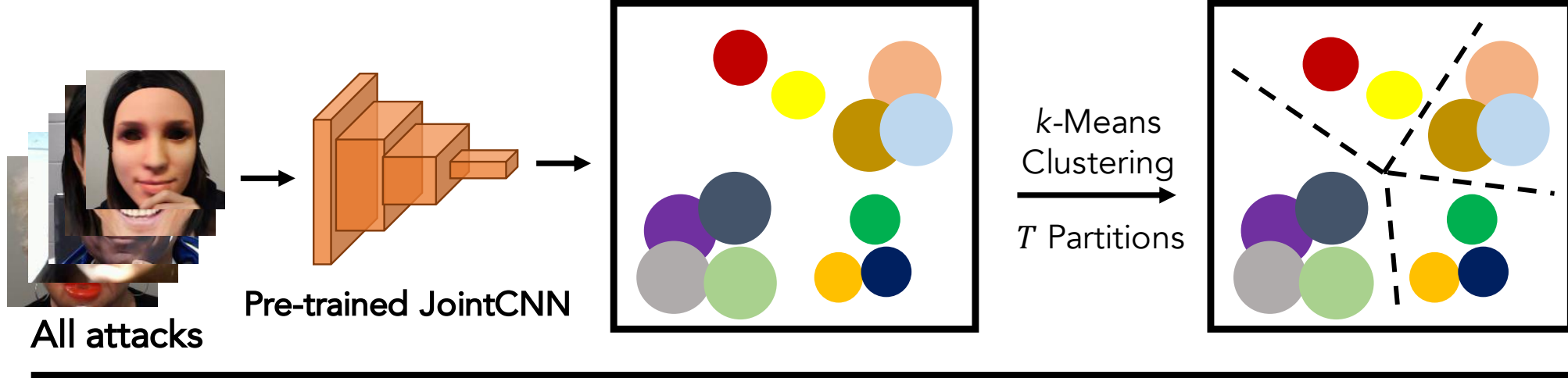


Overall TDR @ 0.2% FDR =
53.19%

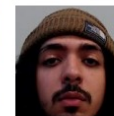
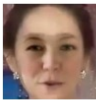
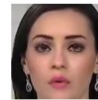
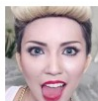
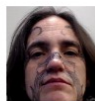
Unified Detection of Digital and Physical Face Attacks



Proposed UniFAD



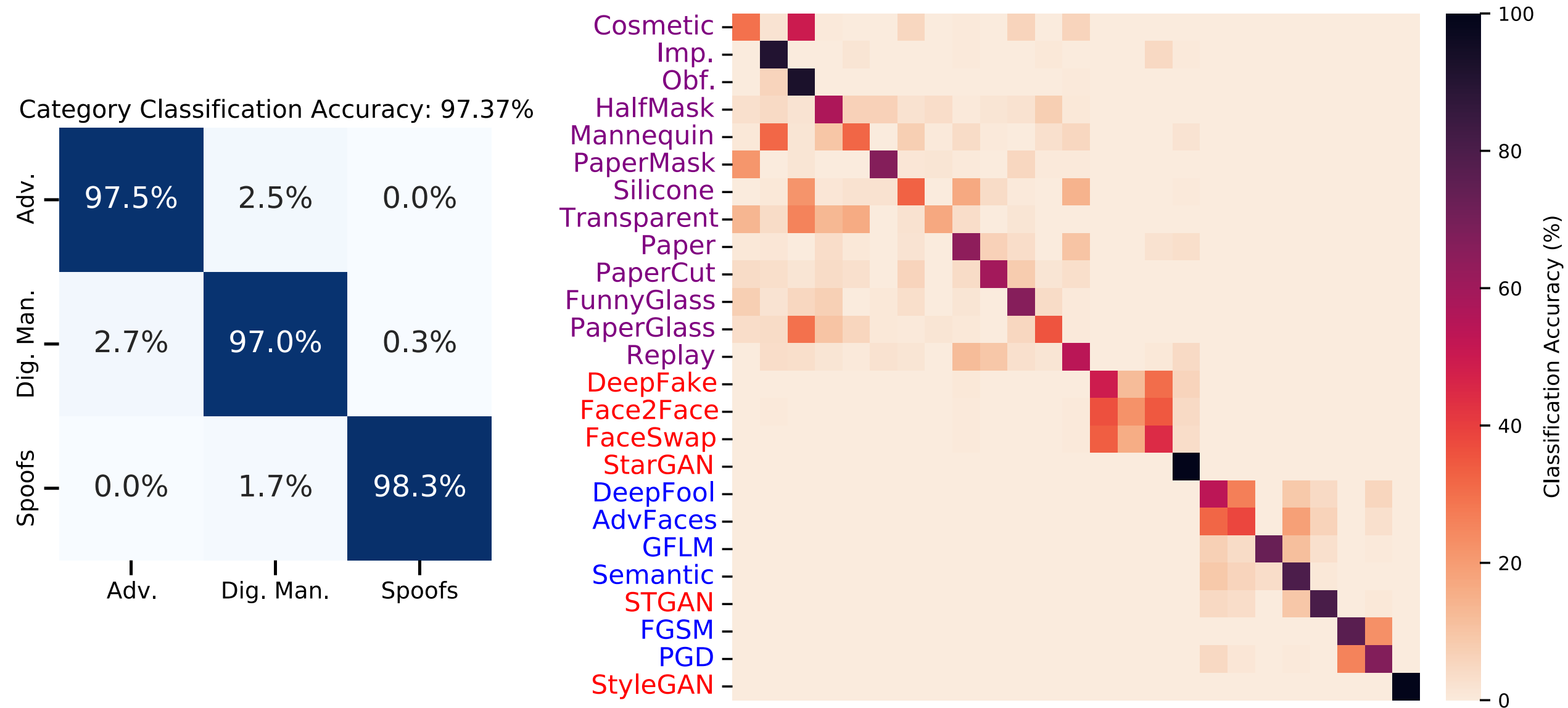
GrandFake Dataset

		# Total Samples	# Training	# Validation	# Testing	Examples								
Datasets	Bona Fides	CASIA-WebFace	10, 575	10, 075	500	0								
		CelebA	193, 506	134, 954	500	58, 052								
		LFW	9, 164	0	0	9, 164								
		FFHQ	20, 999	14, 200	500	6, 299								
		SiW-M	107, 494	74, 746	500	32, 248								
		Total	341, 738	233,975	2,000	105,763						CA-SIA	CelebA	LFW
Adversarial	FGSM	19, 739	10, 495	80	9, 164									
	PGD	19, 739	10, 495	80	9, 164									
	DeepFool	19, 739	10, 495	80	9, 164									
	AdvFaces	19, 739	10, 495	80	9, 164									
	GFLM	17, 946	8, 702	80	9, 164									
	Semantic	19, 739	10, 495	80	9, 164									
Attacks	Digital Manipulation	DeepFake	18, 165	10, 393	80	7, 692								
		Face2Face	18, 204	10, 385	80	7, 739								
		FaceSwap	14, 492	8, 299	80	6, 113								
		STGAN	29, 983	9, 903	80	20, 000								
		StarGAN	45, 473	10, 406	80	34, 987								
		StyleGAN	76, 604	10, 919	80	65, 605								
Spoofs	Cosmetic	2, 638	1, 766	80	792									
	Impersonation	9, 184	6, 348	80	2, 756									
	Obfuscation	3, 611	2, 447	80	1, 084									
	HalfMask	10, 486	7, 260	80	3, 146									
	Mannequin	5, 287	3, 620	80	1, 587									
	PaperMask	2, 550	1, 705	80	765									
	Silicone	5, 038	3, 446	80	1, 512									
	Transparent	11, 451	7, 935	80	3, 436									
	Print	10, 530	7, 290	80	3, 160									
	PaperCut	13, 178	9, 144	80	3, 954									
	FunnyEye	23, 470	16, 349	80	7, 041									
	PaperGlass	18, 563	12, 914	80	5, 569									
	Replay	12, 126	8, 408	80	3, 638									
	Total		447, 674	210,114	2,000						235,560			

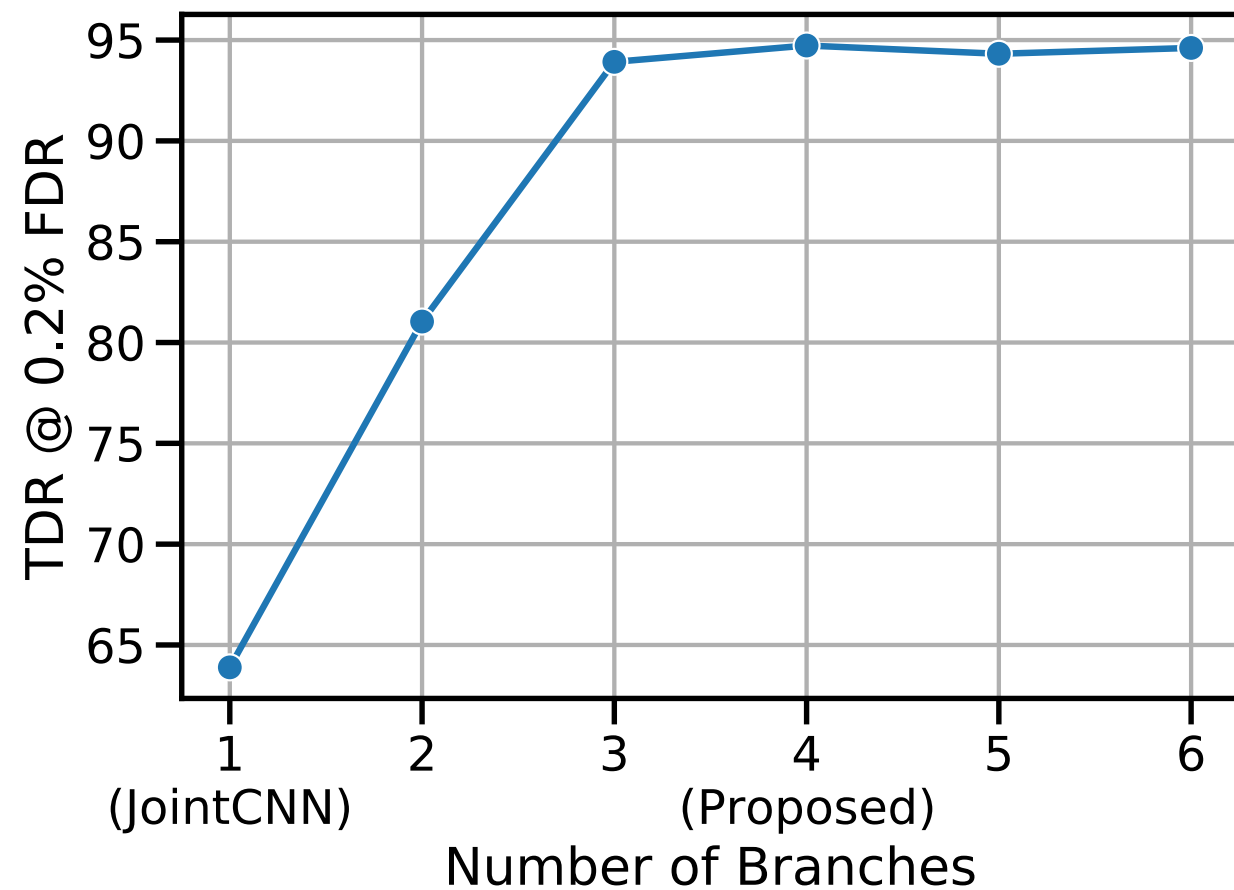
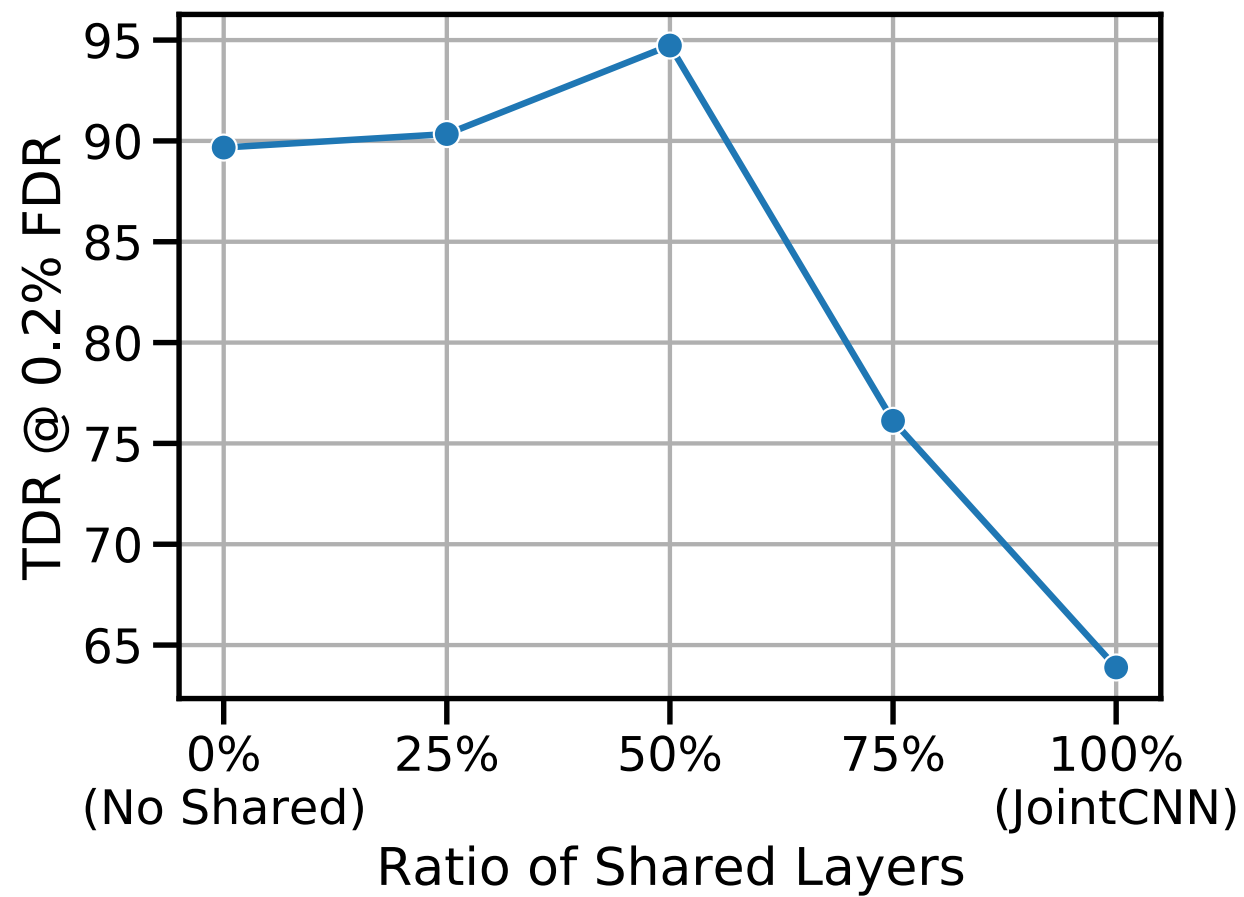
Quantitative Results

TDR (%) @ 0.2% FDR		Year	Proposed For	Adv.	Dig. Man.	Phys.	Overall	Time (ms)
w/o Re-train	FaceGuard	2020	Adversarial	99.91	22.28	00.58	29.64	01.41
	FFD	2020	Digital Manipulation	09.49	94.57	01.25	34.55	11.57
	SSRFCN	2020	Spoofs	00.25	00.76	93.19	22.71	02.22
	MixNet	2020	Spoofs	00.36	09.83	78.21	21.12	12.47
Baselines	FaceGuard	2020	Adversarial	99.86	41.56	04.35	56.69	01.41
	FFD	2020	Digital Manipulation	76.06	91.32	87.43	68.25	11.57
	SSRFCN	2020	Spoofs	08.23	27.67	89.19	43.26	02.22
	One-class	2020	Spoofs	04.81	45.96	79.32	39.40	07.92
	MixNet- <i>UniFAD</i>	2021	All	82.33	91.59	94.60	90.07	12.47
Fusion Schemes	Cascade	—	—	88.39	81.98	69.19	77.46	05.16
	Min-score	—	—	03.65	11.08	00.43	07.22	16.14
	Median-score	—	—	10.87	42.33	47.19	39.48	16.12
	Mean-score	—	—	14.53	47.18	61.32	38.23	16.12
	Max-score	—	—	85.32	61.93	56.87	73.89	16.13
	Sum-score	—	—	74.93	58.01	50.34	69.21	16.11
	LightGBM	—	—	76.25	81.28	88.52	85.97	17.92
<i>Proposed UniFAD</i>		2021	All	92.56	97.21	98.76	94.73	02.59

Attack Category & Type Classification



Architectural Analysis



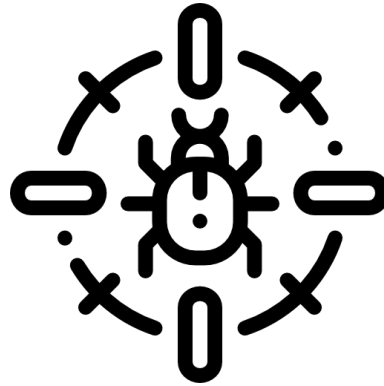
Failure Cases

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Transparent	Paper	Face2Face	StarGAN	AdvFaces	DeepFool	FGSM	StyleGAN
							
Final: 0.16	Final: 0.41	Final: 0.36	Final: 0.47	Final: 0.19	Final: 0.27	Final: 0.39	Final: 0.41
Brn1: 0.09	Brn1: 0.39	Brn1: 0.02	Brn1: 0.01	Brn1: 0.00	Brn1: 0.00	Brn1: 0.04	Brn1: 0.01
Brn2: 0.02	Brn2: 0.10	Brn2: 0.29	Brn2: 0.43	Brn2: 0.03	Brn2: 0.02	Brn2: 0.01	Brn2: 0.00
Brn3: 0.10	Brn3: 0.04	Brn3: 0.09	Brn3: 0.04	Brn3: 0.12	Brn3: 0.22	Brn3: 0.12	Brn3: 0.08
Brn4: 0.04	Brn4: 0.13	Brn4: 0.07	Brn4: 0.10	Brn4: 0.04	Brn4: 0.14	Brn4: 0.32	Brn4: 0.36

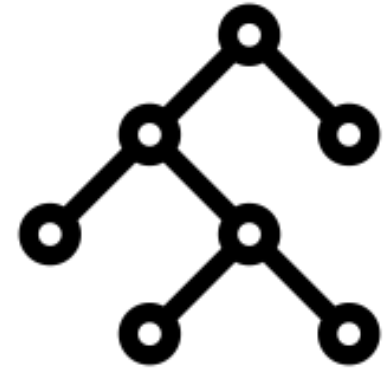
Summary



Unified Detection



SOTA Detection Performance



Free Attack Classifier

Publications

1. **D. Deb, X. Liu and A. K. Jain, "Unified Detection of Digital and Physical Face Attacks", arXiv:2104.02156, 2021.**
2. **D. Deb, X. Liu and A. K. Jain, "FaceGuard: A Self-Supervised Defense Against Adversarial Face Images", arXiv:2011.14218, 2021.**
3. J. J. Engelsma, D. Deb, K. Cao, A. Bhatnagar, P. S. Sudhish and A. K. Jain, "Infant-ID: Fingerprints for Global Good", to appear in the *IEEE PAMI*, 2021.
4. Deb, D. Aggarwal and A. K. Jain. "Child Face Age-Progression via Deep Feature Aging", in *IEEE ICPR*, 2021
5. **D. Deb and A. K. Jain. "Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach", in *IEEE TIFS*, 2020**
6. **D. Deb, J. Zhang and A. K. Jain, "AdvFaces: Adversarial Face Synthesis", in *IEEE IJCB*, 2020**
7. H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang and A. K. Jain, "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review", in *IJAC*, 2020
8. D. Deb, A. Ross, A. K. Jain, K. Prakah-Asante and K. Venkatesh Prasad, "Actions Speak Louder Than(Pass)words: Passive Authentication of Smartphone Users via Deep Temporal Features", in *IEEE ICB*, 2019.
9. J. J. Engelsma, D. Deb, A. K. Jain, P. S. Sudhish and A. Bhatnagar, "Infant-Prints: Fingerprints for Reducing Infant Mortality", in *IEEE CVPRW-CV4GC*, 2019
10. Y. Shi*, D. Deb* and A. K. Jain, "WarpGAN: Automatic Caricature Generation", in *IEEE CVPR*, 2019.
11. D. Deb, N. Nain and A. K. Jain, "Longitudinal Study of Child Face Recognition", in *IEEE ICB*, 2018
12. D. Deb, S. Wiper, S. Gong, Y. Shi, C. Tymoszek, A. Fletcher and A. K. Jain, "Face Recognition: Primates in the Wild", in *IEEE BTAS*, 2018.
13. D. Deb, E. Tabassi, T. Chugh and A. K. Jain, "Altered Fingerprints: Detection and Localization", in *IEEE BTAS*, 2018
14. D. Deb, T. Chugh, J. Engelsma, K. Cao, N. Nain, J. Kendall and A. K. Jain, "Matching Fingerphotos to Slap Fingerprint Images", arXiv:1804.08122, 2018.
15. D. Deb, J. Best-Powden and A. K. Jain. "Face Recognition Performance Under Aging" in *IEEE CVPRW*, 2017

Questions?



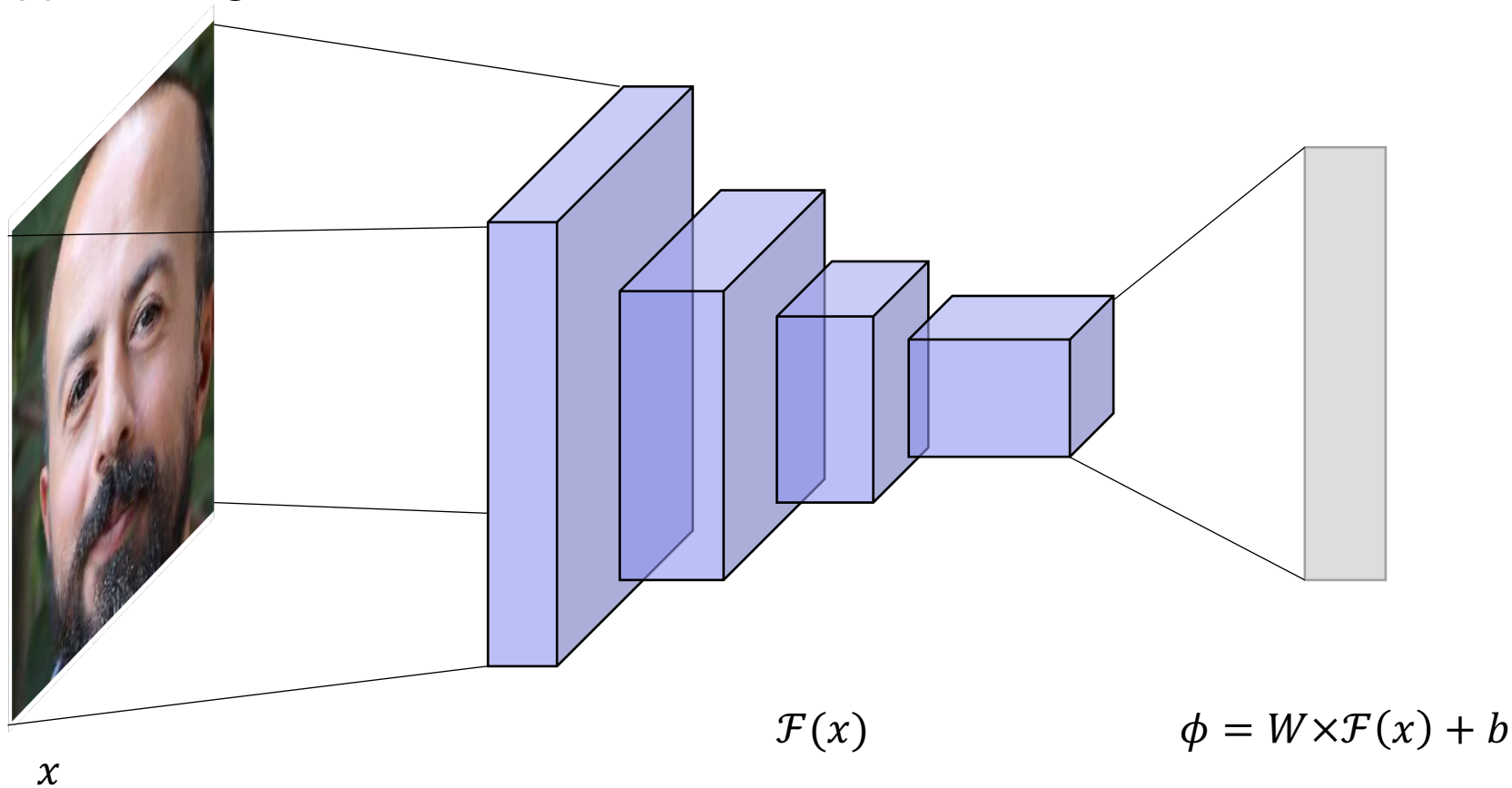
Extras

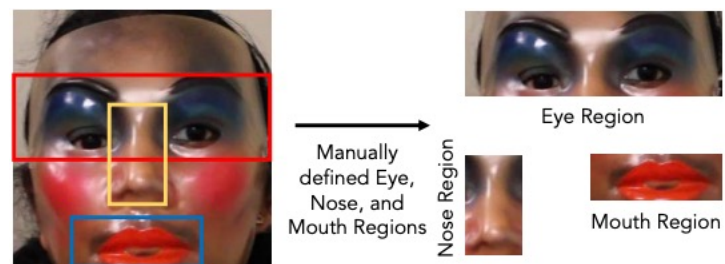
Face Feature Extraction

Cropped & Aligned Face

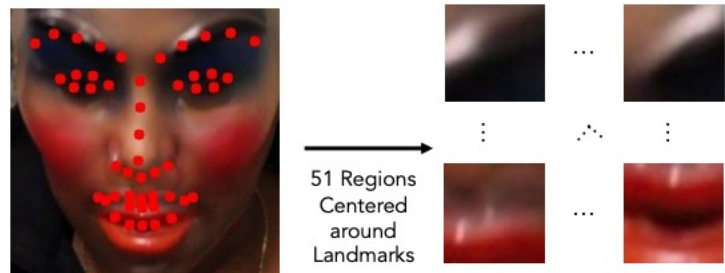
Convolutional Neural Net

Face Feature

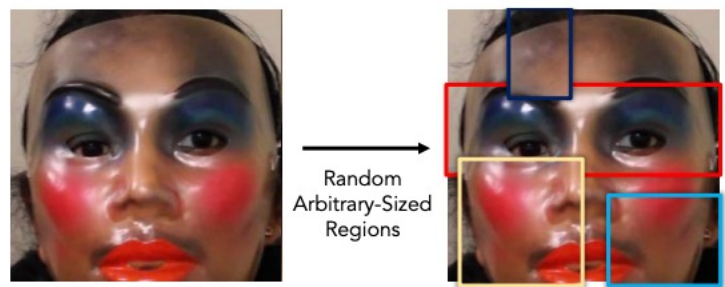




(a) Manually-defined Regions

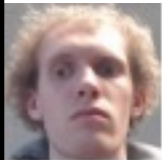
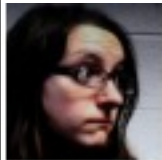
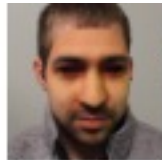
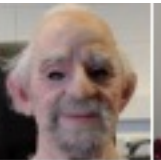
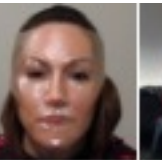
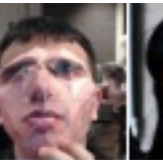
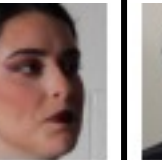
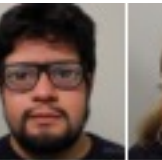


(b) Landmark Based



(c) Self-Supervised (Proposed)

Spoof-in-the-Wild-Multiple-Attacks (SiW-M) Dataset

		Globa		3D Masks					Makeup			Partial		
														
	Live	Replay	Print	Half Mask	Silicone	Transp arent	Paper Mask	Mann e-quin	Obfus cation	Imper sonati on	Cosm etic	Funny eye	Paper Glass	Paper Cut
# Subj	493	21	60	12	12	88	6	12	23	61	37	160	122	86
# Vids	660	99	118	72	27	88	17	40	23	61	50	160	127	86
# Frames	341,738	12,126	10,531	10,486	5,038	11,451	2,550	5,287	3,611	9,184	2,638	23,470	18,563	13,178

Largest Face Anti-Spoofing Dataset in terms of # Subj, # Videos, and # Spoof Types

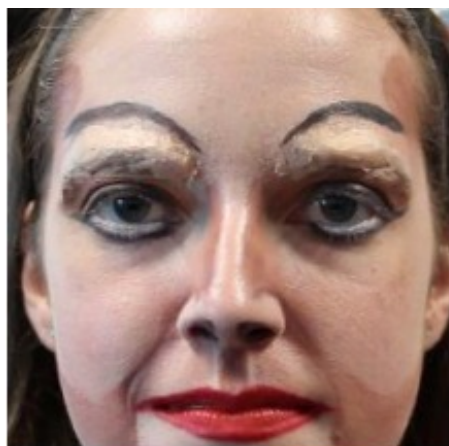
Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," In *CVPR*, 2019.

Results on SiW-M: Known Attack Protocol

Method	Metric (%)			Mask Attacks					Makeup Attacks			Partial Attacks			Mean $\pm$ Std.
		Replay	Print	Half	Silicone	Trans.	Paper	Mann.	Obf.	Imp.	Cosm.	Funny Eye	Glasses	Paper Cut	
Auxiliary	ACER	5.1	5.0	5.0	10.2	5.0	9.8	6.3	19.6	5.0	26.5	5.5	5.2	5.0	8.7 $\pm$ 6.8
	EER	4.7	0.0	1.6	10.5	4.6	10.0	6.4	12.7	0.0	19.6	7.2	7.5	0.0	6.5 $\pm$ 5.8
Proposed	ACER	3.5	3.1	1.9	5.7	2.1	1.9	4.2	7.2	2.5	22.5	1.9	2.2	1.9	4.7 $\pm$ 5.6
	EER	3.5	3.1	0.1	9.9	1.4	0.0	4.3	6.4	2.0	15.4	0.5	1.6	1.7	3.9 $\pm$ 4.4
	TDR*	55.5	92.3	69.5	100.0	90.4	100.0	85.1	92.5	78.7	99.1	95.6	95.7	76.0	87.0 $\pm$ 13.0

*TDR evaluated at 2.0% FDR

Subtlety of Obfuscation Makeups



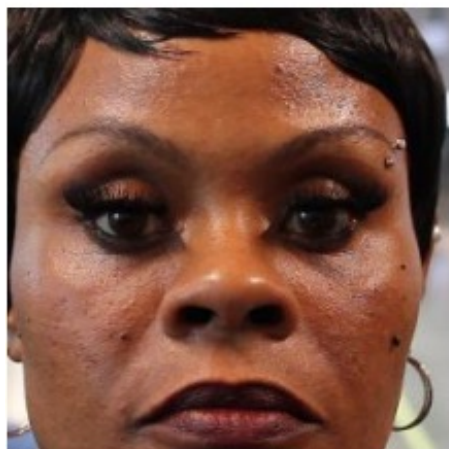
(a) Obfuscation



(b) Spoofness Score: 1.0



(c) Spoofness Score: 0.7



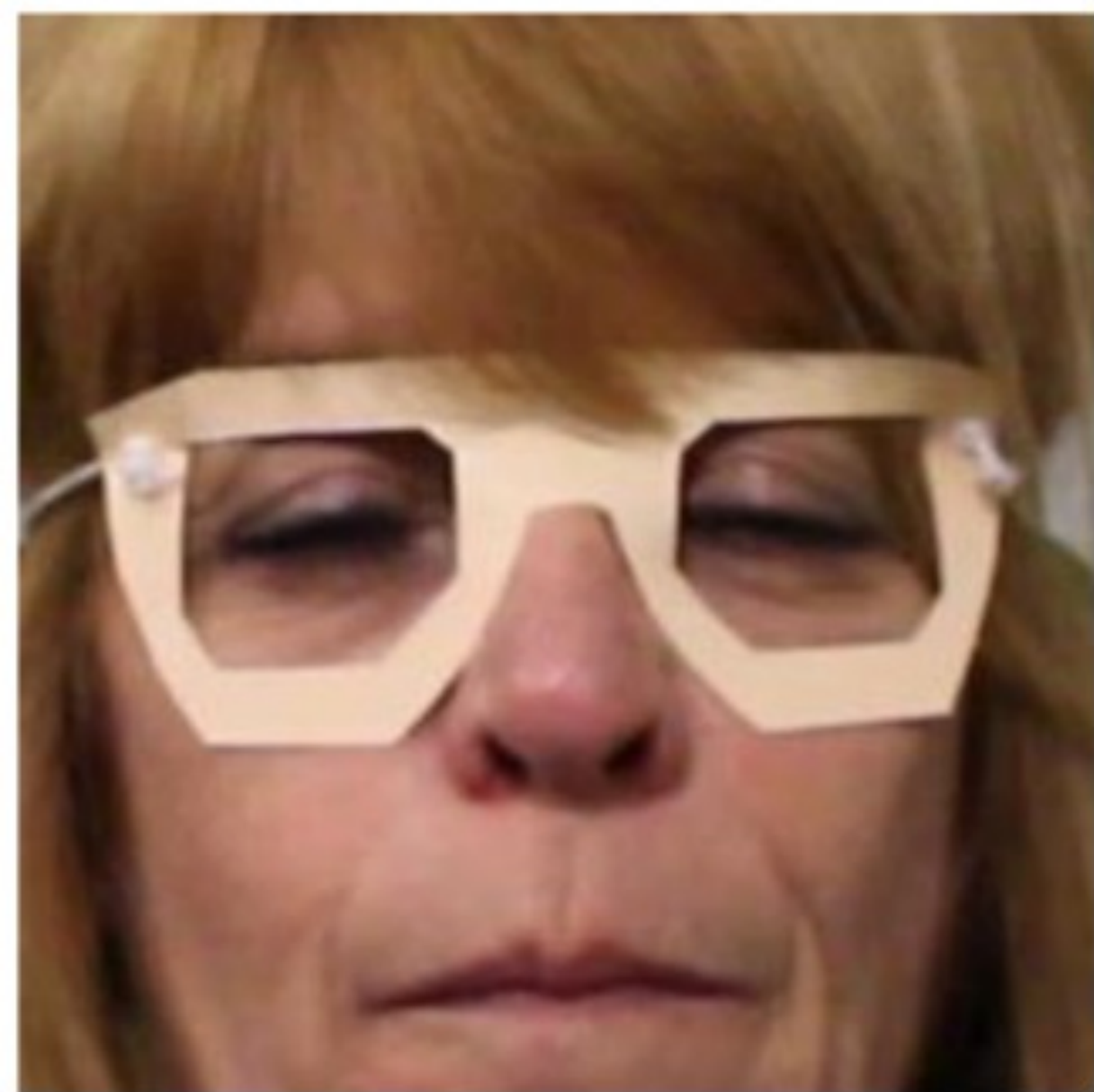
(d) Obfuscation



(e) Spoofness Score: 0.0



(f) Spoofness Score: 0.0

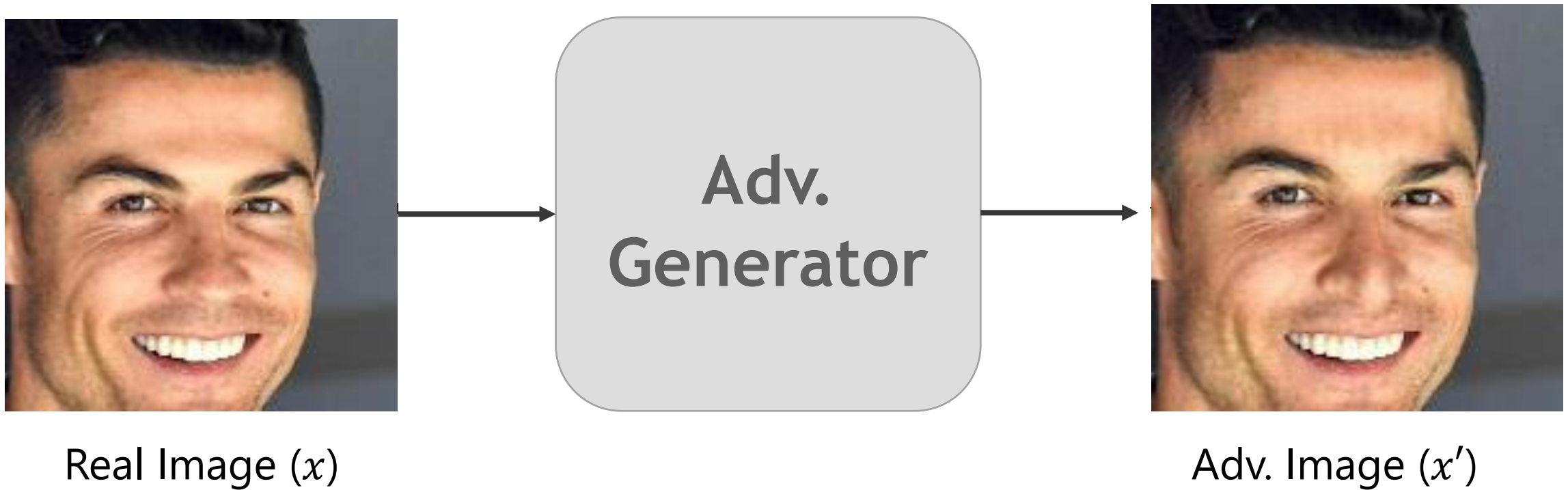


(a) Paper Eyeglasses



(b) Average Eyeglasses Score Map

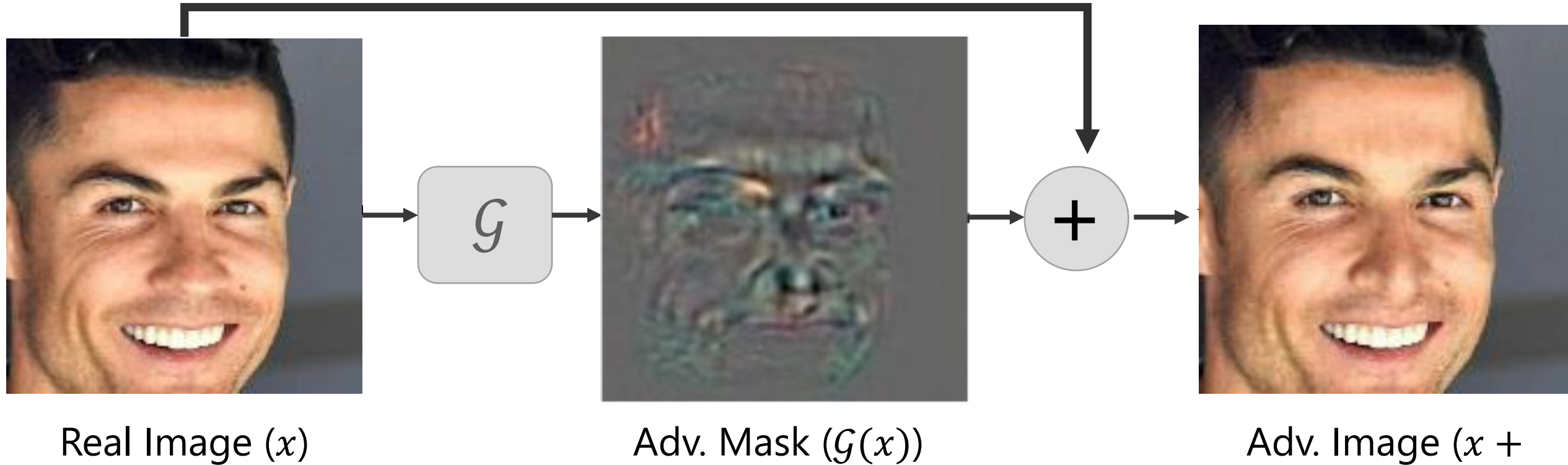
1. Visual Realism



$$\mathcal{L}_{perturbation} = \min ||x - x'||_2$$

$$\mathcal{L}_{perturbation} = \min ||x - (x + \Delta)||_2$$

1. Visual Realism



$$\mathcal{L}_{\text{perturbation}} = \min ||x - (x + \mathcal{G}(x))||_2$$

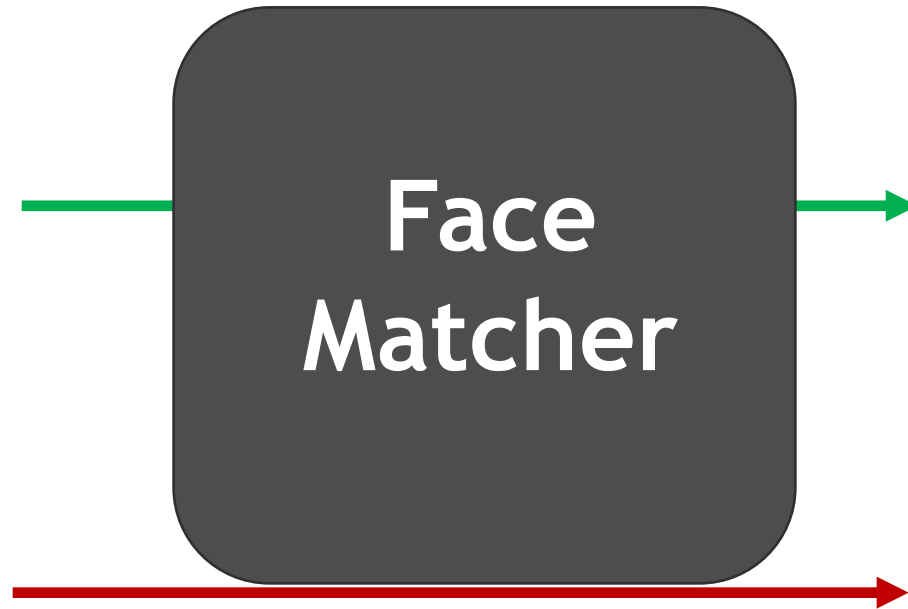
$$\mathcal{L}_{\text{perturbation}} = \min ||\max(\epsilon, \mathcal{G}(x))||_2, \epsilon \in (0, \infty)$$

2. Attack Success (Obfuscation)

Real Image



0.78 (Match)



0.12 (Non-Match)

Gallery



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

Ablation Study

Probe



w/o
Discriminator



w/o $\mathcal{L}_{identity}$



w/o
 $\mathcal{L}_{perturbation}$



with all



1.0

0.12



1.0

0.17

Quantitative Results: Impersonation

Impersonation Attack: 9,164 comparisons / fold

	AdvFaces	A³GN	PGD	FGSM
Attack Success Rate (%) @ 0.1% False Accept Rate				
<i>FaceNet</i>	<i>20.85 ± 0.40</i>	<i>05.99 ± 0.19</i>	76.79 ± 0.26	<i>13.04 ± 0.12</i>
SphereFace	20.19 ± 0.27	07.94 ± 0.19	09.03 ± 0.39	02.34 ± 0.03
ArcFace	24.30 ± 0.44	17.14 ± 0.29	19.50 ± 1.95	08.34 ± 0.21
COTS-A	20.75 ± 0.35	15.01 ± 0.30	01.76 ± 0.10	01.40 ± 0.08
COTS-B	19.85 ± 0.28	10.23 ± 0.50	12.49 ± 0.24	04.67 ± 0.16
Structural Similarity				
	0.92 ± 0.02	0.69 ± 0.04	0.77 ± 0.04	0.48 ± 0.75
Computation Time (s)				
	0.01	0.04	11.74	0.03

Examples of Impersonation Attack

Gallery



Target's
Probe



0.78

Probe



0.10

AdvFaces



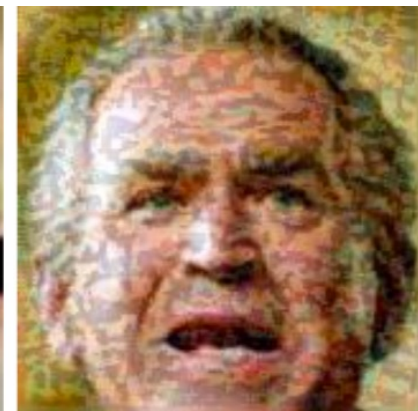
0.30

A³GN



0.29

FGSM



0.36



0.80



0.15



0.34



0.33



0.42

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

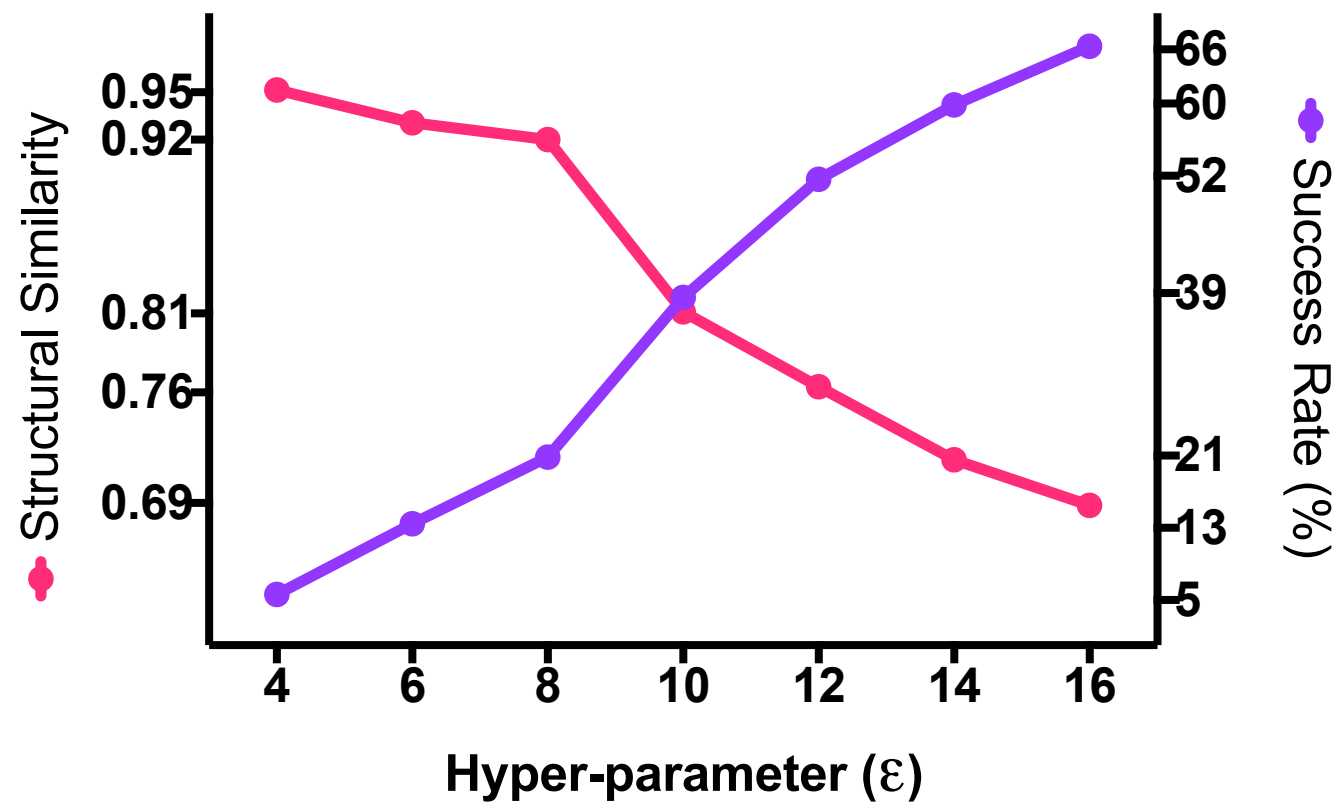
Controllability

$\epsilon = 4.0$

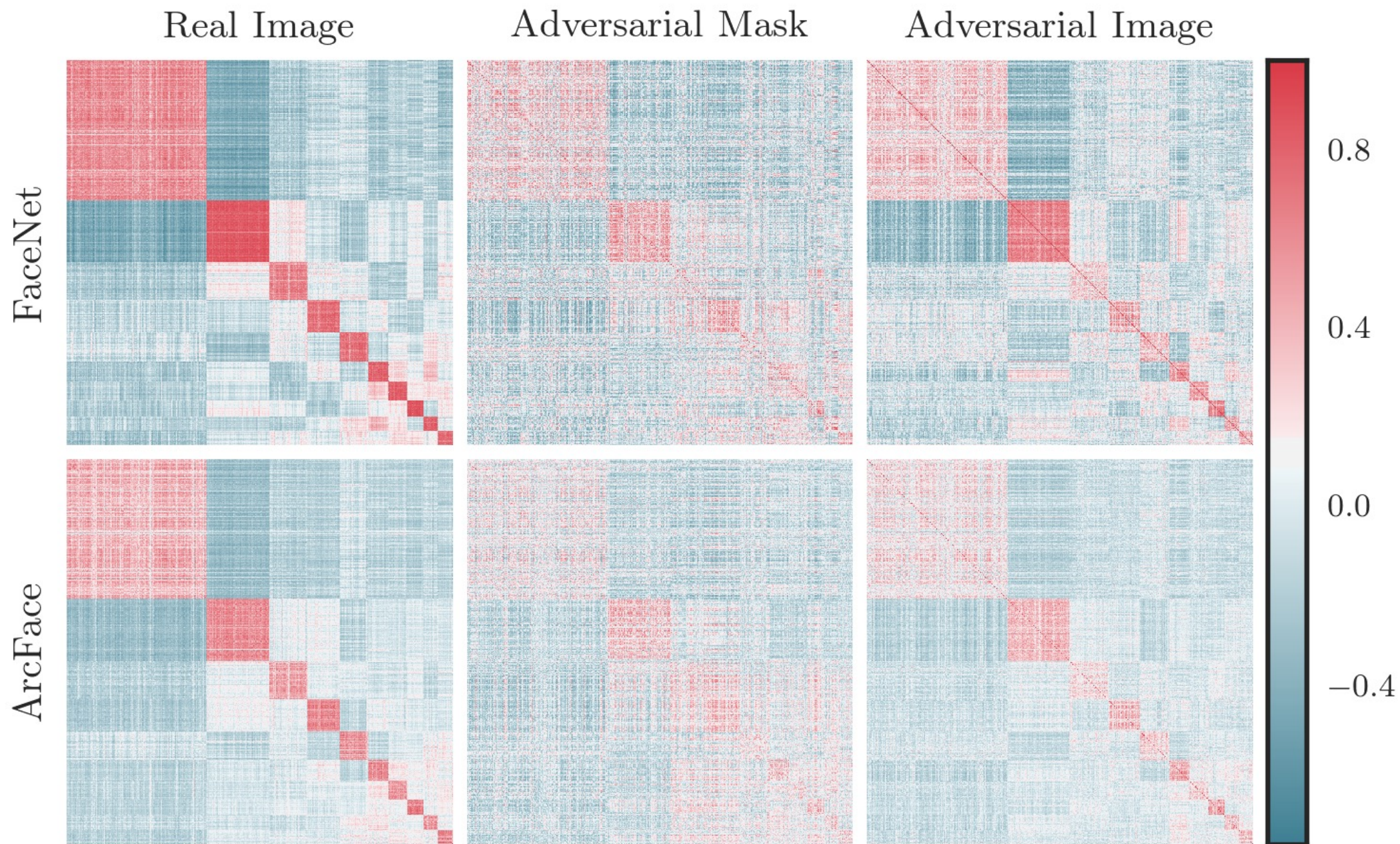
$\epsilon = 8.0$

$\epsilon = 10.0$

$\epsilon = 16.0$

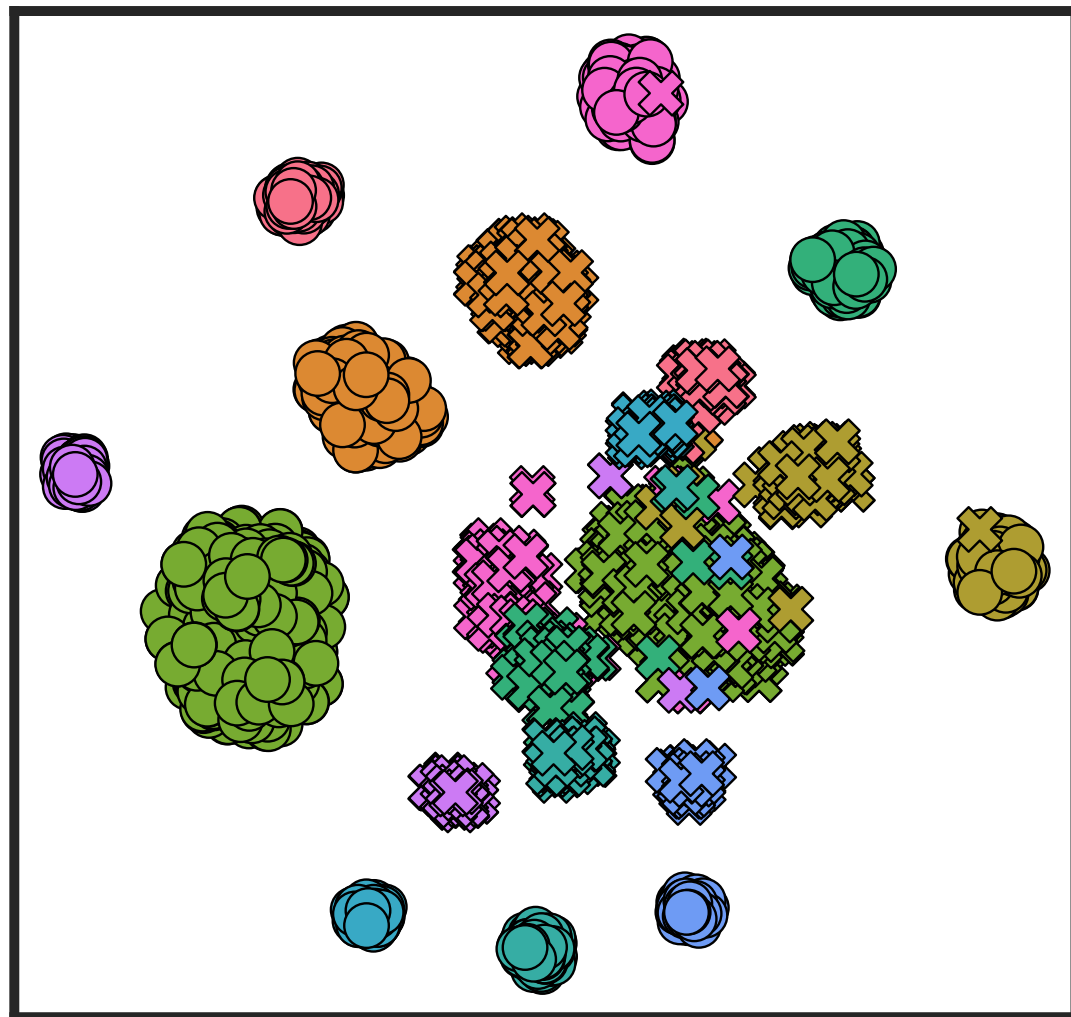


Transferability

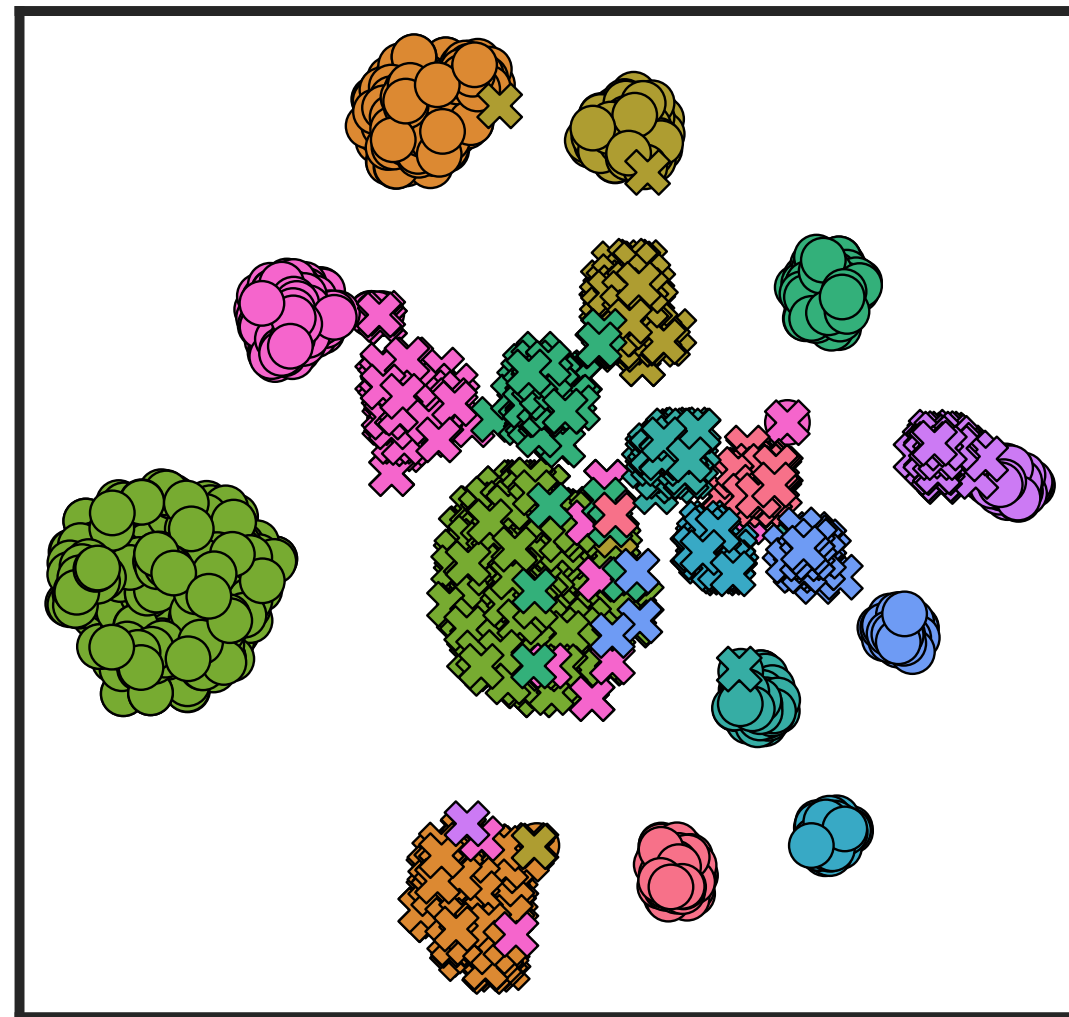


Transferability

FaceNet



ArcFace



Real Image



Adversarial Image (Obfuscation)

Additional Adversarial Detection Results

	Detection Accuracy (%)	Year	LFW	CelebA	FFHQ
General	Gong <i>et al.</i>	2017	97.54 $\pm$ 02.82	94.38 $\pm$ 04.48	96.89 $\pm$ 02.07
	ODIN	2018	77.03 $\pm$ 14.34	68.95 $\pm$ 19.64	74.63 $\pm$ 08.16
	Steganalysis	2019	74.33 $\pm$ 14.77	72.53 $\pm$ 11.30	71.09 $\pm$ 09.86
Face	UAP-D	2018	64.28 $\pm$ 09.97	63.19 $\pm$ 16.49	68.65 $\pm$ 08.73
	SmartBox	2018	56.77 $\pm$ 05.16	54.85 $\pm$ 09.33	57.19 $\pm$ 09.55
	Goswami <i>et al.</i>	2019	79.37 $\pm$ 14.04	74.70 $\pm$ 13.88	80.03 $\pm$ 09.24
	Massoli <i>et al.</i> (MLP)	2020	69.16 $\pm$ 15.29	61.78 $\pm$ 11.34	66.26 $\pm$ 10.06
	Massoli <i>et al.</i> (LSTM)	2020	70.11 $\pm$ 13.35	63.67 $\pm$ 16.21	69.58 $\pm$ 07.91
	Agarwal <i>et al.</i>	2020	87.03 $\pm$ 16.86	85.81 $\pm$ 15.64	86.70 $\pm$ 11.04
<i>Proposed FaceGuard</i>		2021	99.81 $\pm$ 00.10	98.73 $\pm$ 00.92	99.35 $\pm$ 00.09

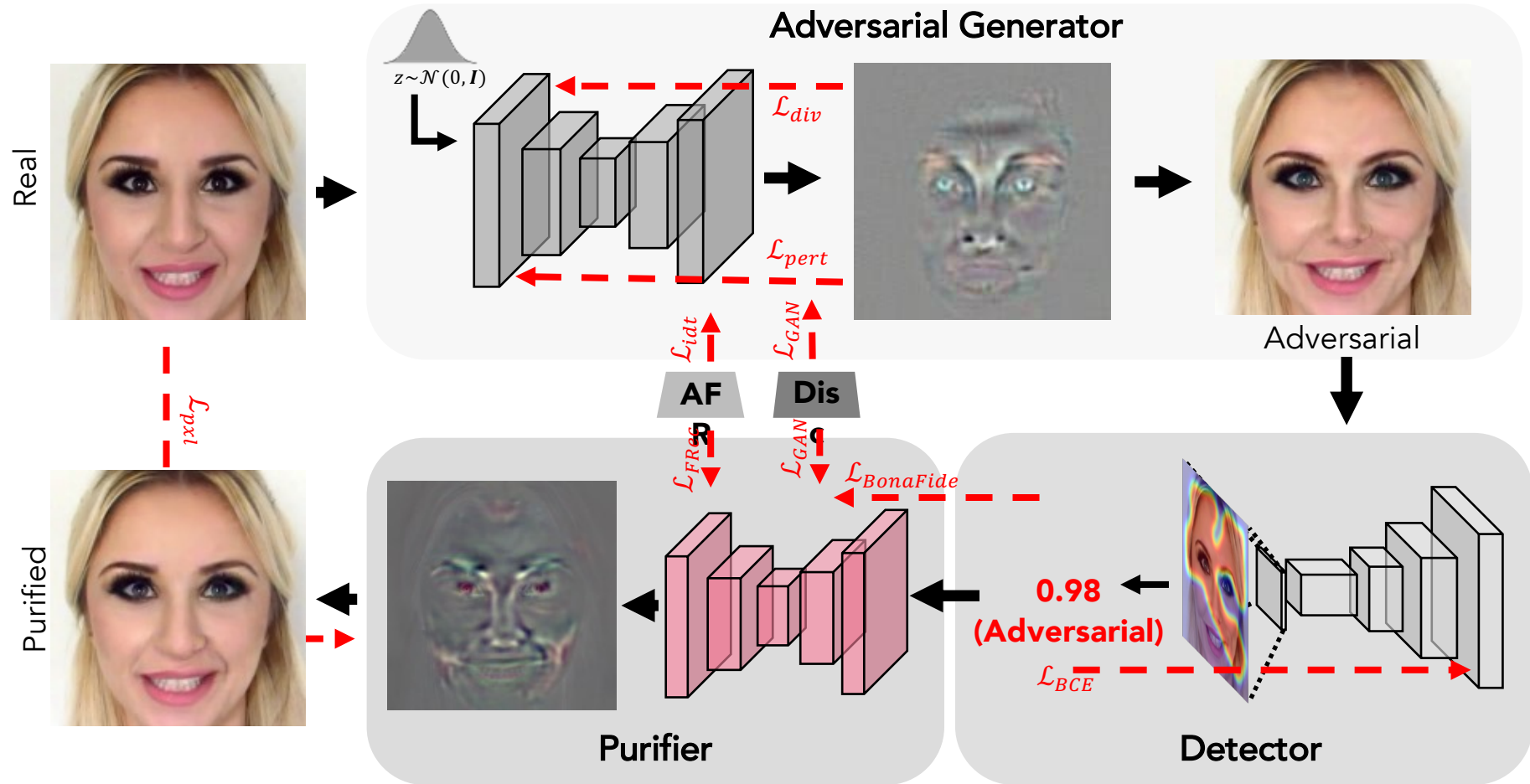
	Known			Unseen		
	FGSM	PGD	DeepFool	AdvFaces	GFLM	SemanticAdv
Gong <i>et al.</i>	94.51	92.21	94.12	68.63	50.00	50.21
UAP-D	63.65	69.33	56.38	60.81	50.12	50.28
SmartBox	58.79	62.53	51.32	54.87	50.97	62.14
Massoli <i>et al.</i> (MLP)	78.35	82.52	91.21	55.57	50.00	50.00
Massoli <i>et al.</i> (LSTM)	74.61	86.43	94.73	62.43	50.00	50.00

(a)

	Known			Unseen		
	AdvFaces	GFLM	SemanticAdv	FGSM	PGD	DeepFool
Gong <i>et al.</i>	81.39	96.72	98.97	84.46	57.00	72.32
UAP-D	68.78	54.31	77.46	51.64	50.32	52.01
SmartBox	54.87	50.97	62.14	58.79	62.53	51.32
Massoli <i>et al.</i> (MLP)	77.64	86.54	94.78	55.20	51.32	52.90
Massoli <i>et al.</i> (LSTM)	81.42	92.62	96.76	52.74	65.43	54.84

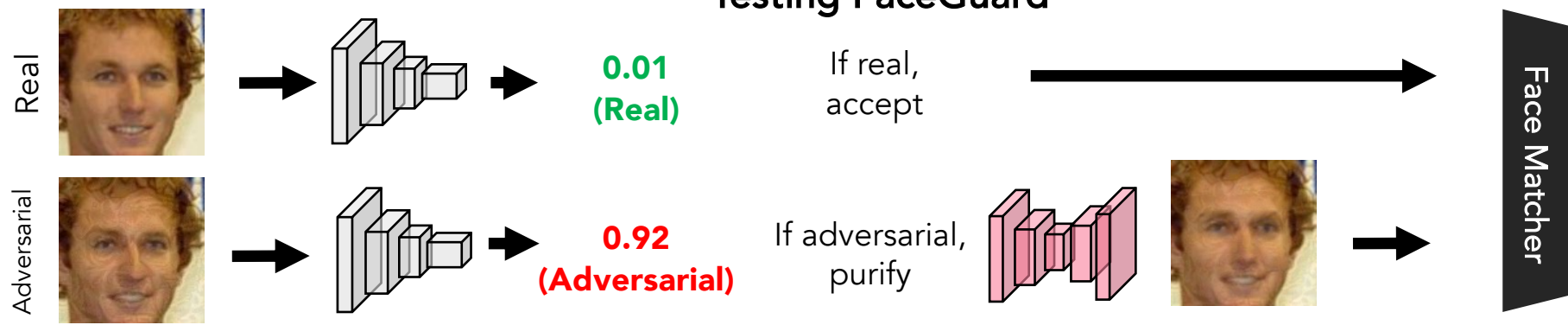
(b)

	Known					
	FGSM	PGD	DeepFool	AdvFaces	GFLM	SemanticAdv
Gong <i>et al.</i>	98.94	97.91	95.87	92.69	99.92	99.92
UAP-D	61.32	74.33	56.78	51.11	65.33	76.78
SmartBox	58.79	62.53	51.32	54.87	50.97	62.14
Massoli <i>et al.</i> (MLP)	63.58	76.28	81.78	88.38	51.97	52.98
Massoli <i>et al.</i> (LSTM)	71.53	76.43	88.32	75.43	53.76	55.22
	Unseen					
<i>Proposed FaceGuard</i>	99.85	99.85	99.85	99.84	99.61	99.85

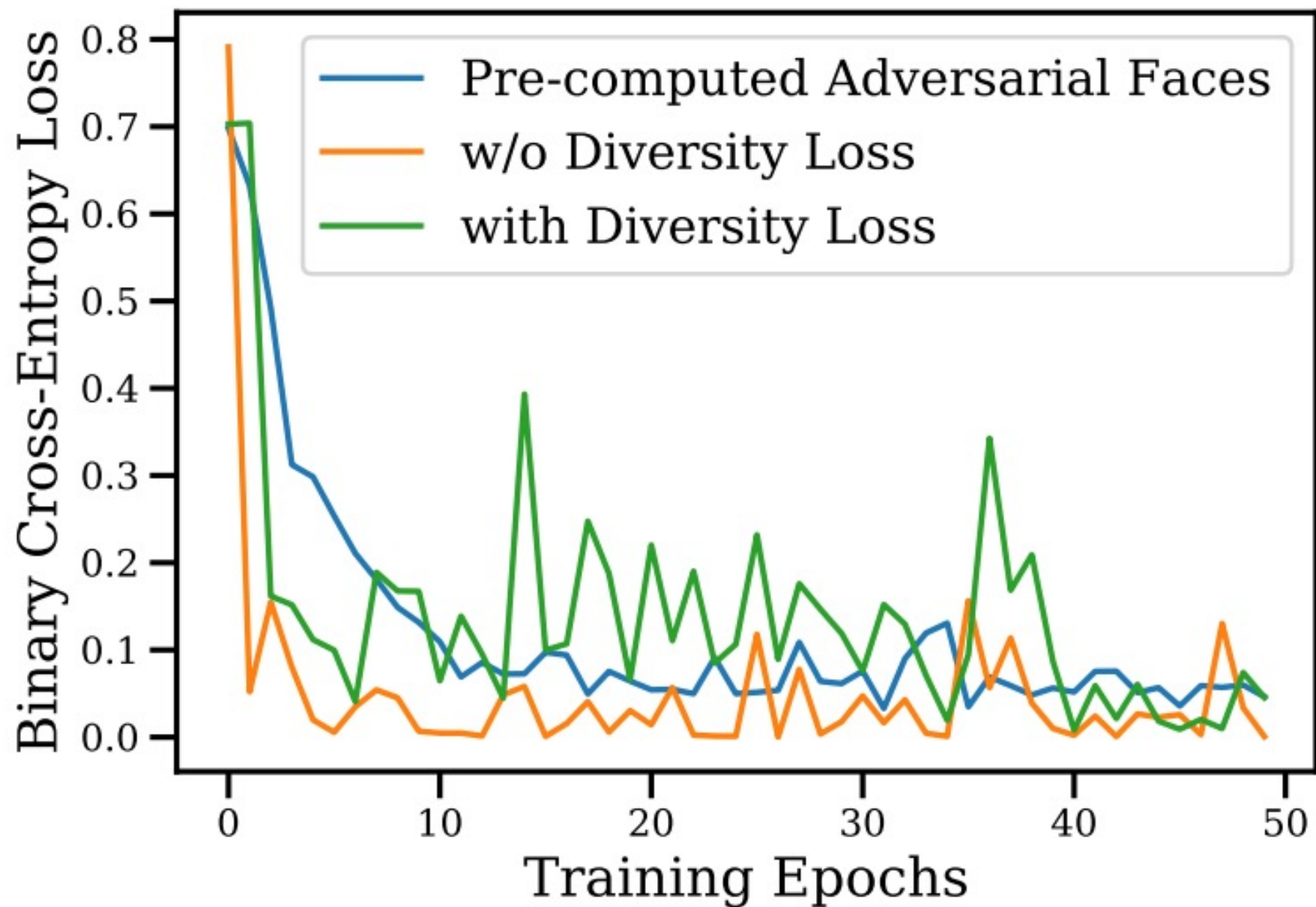


Training FaceGuard

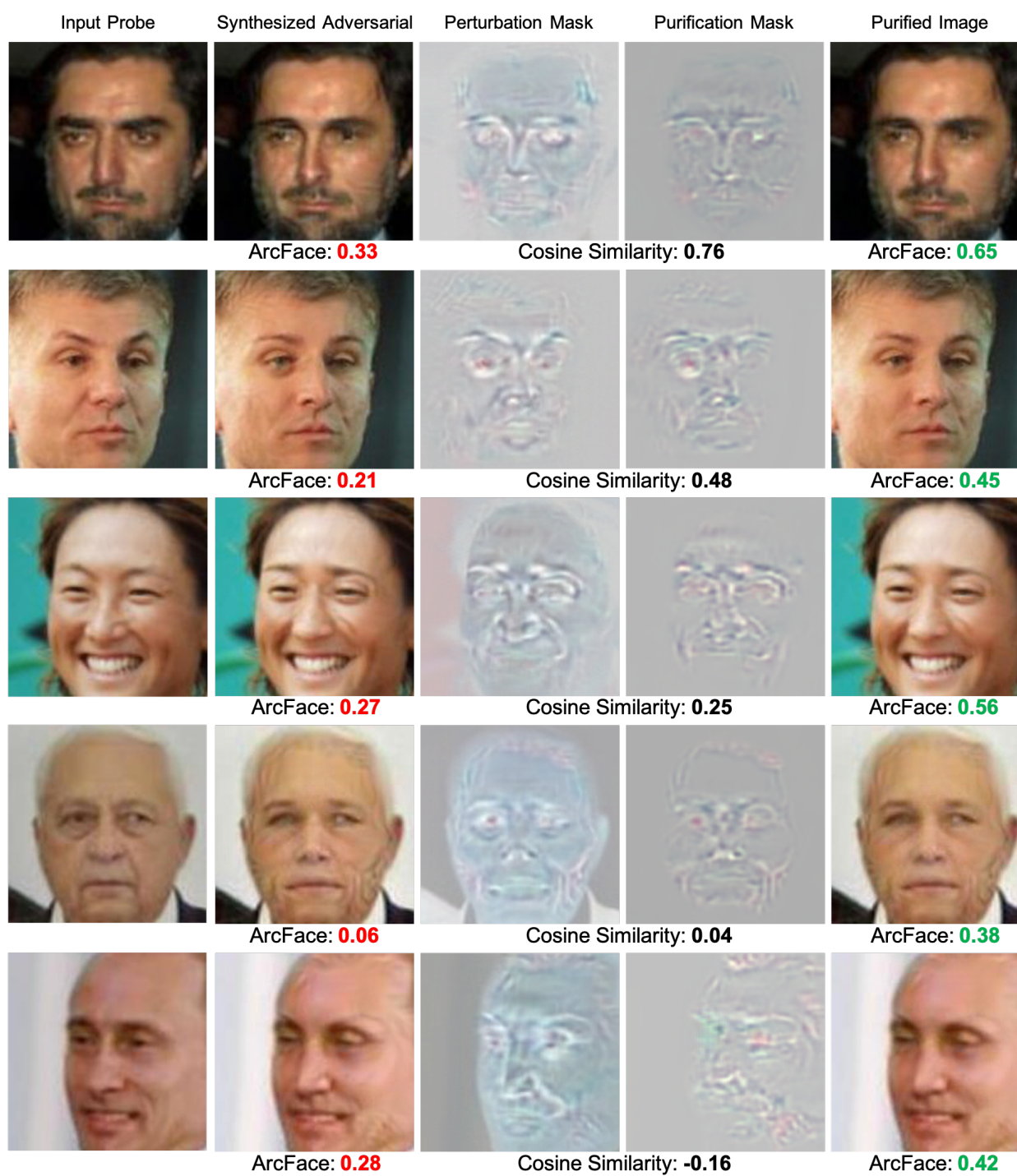
Testing FaceGuard



FaceGuard Loss Curves

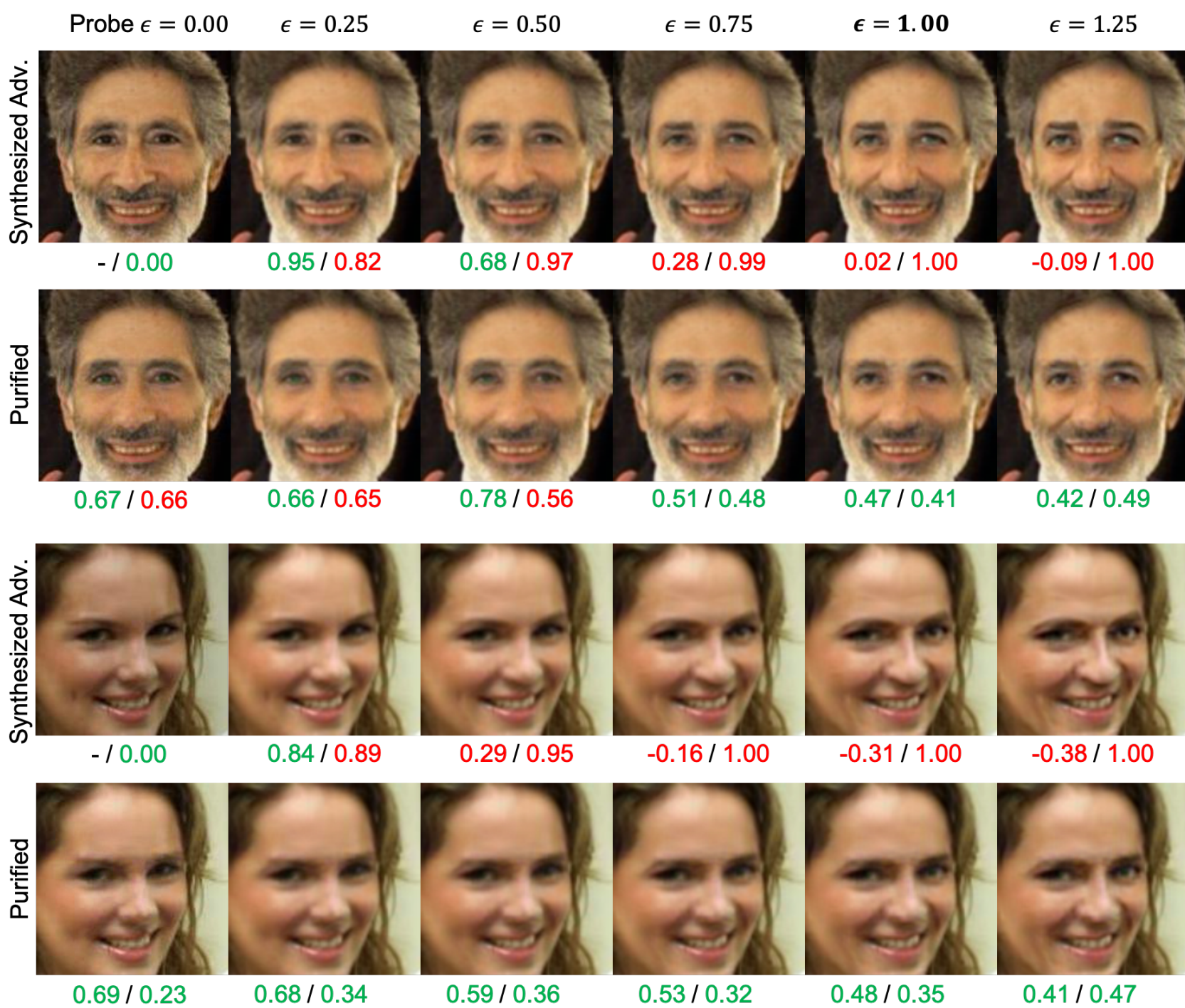


Generator vs. Purifier Masks

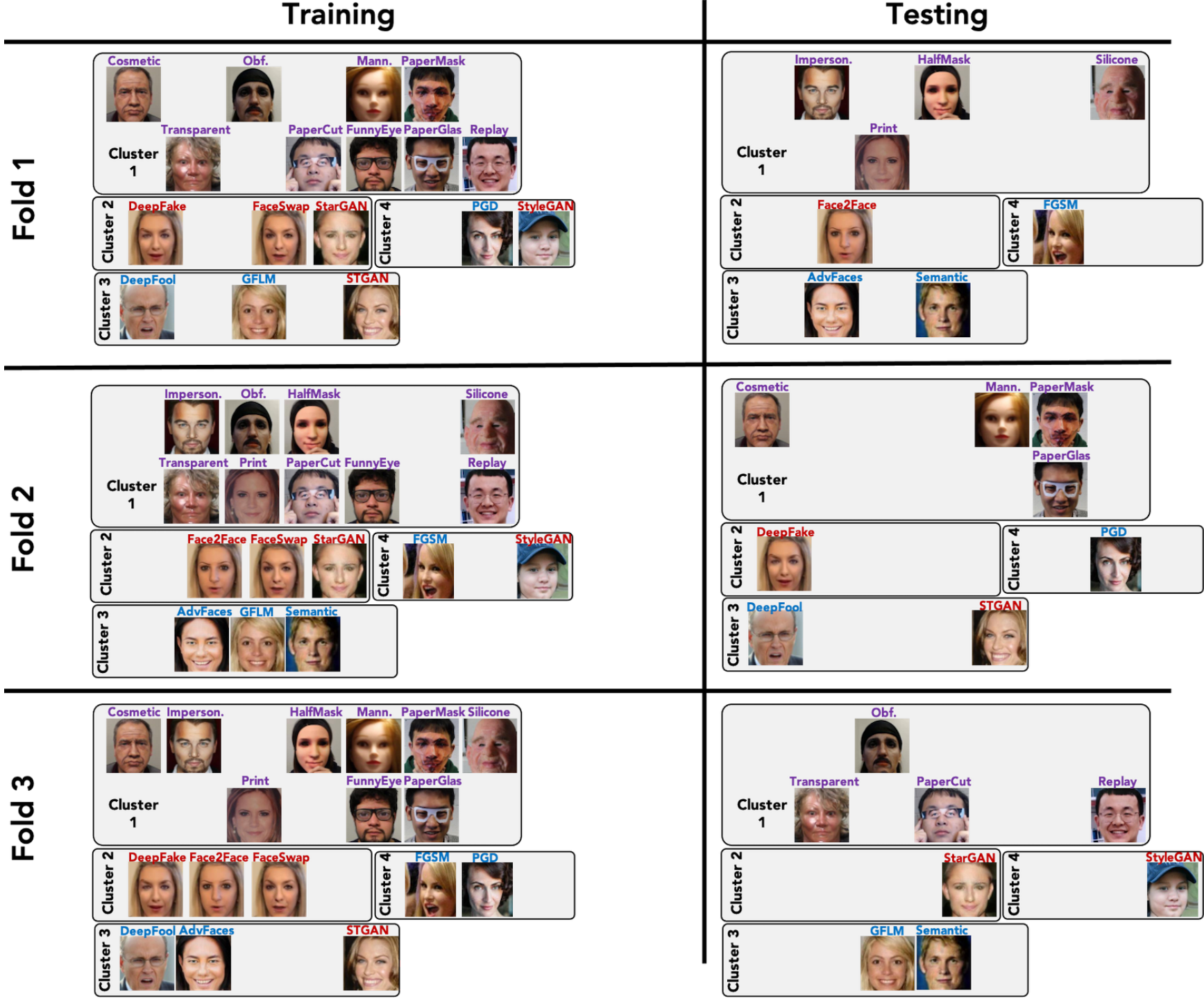


FaceGuard: Ablation Study

	Model	AdvFaces	Mean $\pm$ Std.
Gen. $\mathcal{G}$	Without $\mathcal{G}$	91.72	97.12 $\pm$ 04.54
	Without $\mathcal{L}_{div}$	95.42	98.23 $\pm$ 01.33
	With $\mathcal{G}$ and $\mathcal{L}_{div}$	99.84	99.81 $\pm$ 00.10
Det. $\mathcal{D}$	$\mathcal{D}$ as Discriminator	50.00	75.25 $\pm$ 21.19
	$\mathcal{D}$ via Pre-Computed $\mathcal{G}$	52.01	69.37 $\pm$ 19.91
	$\mathcal{D}$ as Online Detector	99.84	99.81 $\pm$ 00.10



Generalizability Study



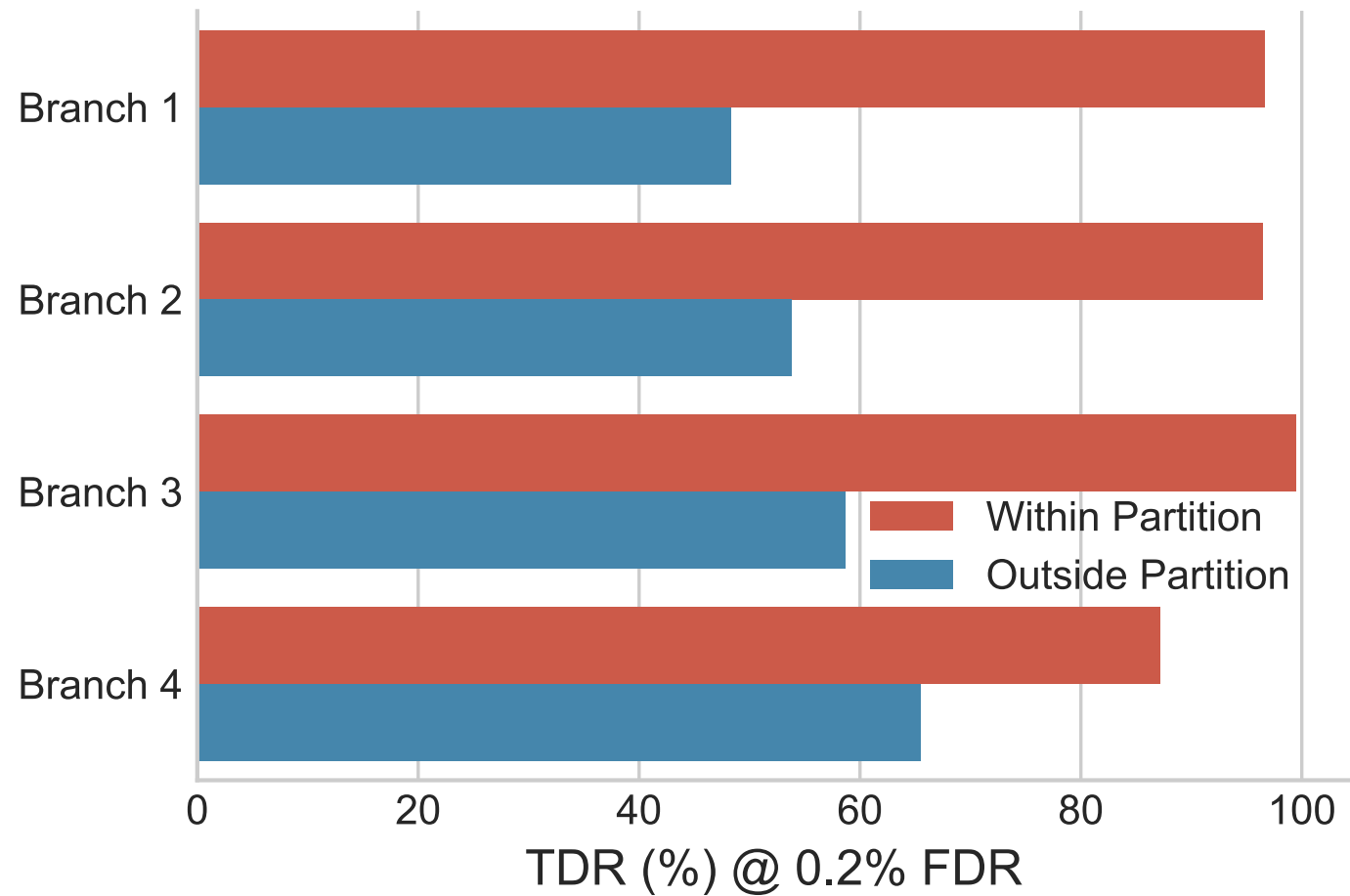
UniFAD: Generalizability Study

Method	Metric (%)	Fold 1	Fold 2	Fold 3	Mean $\pm$ Std.
FaceGuard [15]	TDR	41.38	54.19	36.82	44.13 $\pm$ 9.01
	Acc.	58.42	64.19	55.74	59.45 $\pm$ 4.32
FFD [12]	TDR	53.19	62.45	52.94	56.20 $\pm$ 5.42
	Acc.	66.15	69.33	67.86	67.78 $\pm$ 1.59
SSRFCN [14]	TDR	49.10	64.92	61.18	58.84 $\pm$ 8.26
	Acc.	60.07	72.77	69.83	66.57 $\pm$ 6.64
MixNet- <i>UniFAD</i>	TDR	67.19	73.18	72.74	71.04 $\pm$ 3.33
	Acc.	75.64	79.40	78.73	77.93 $\pm$ 2.00
LightGBM	TDR	51.65	65.73	67.91	61.76 $\pm$ 8.83
	Acc.	69.34	73.66	75.80	72.93 $\pm$ 3.29
Proposed <i>UniFAD</i>	TDR	76.18	83.19	82.67	80.68 $\pm$ 3.91
	Acc.	85.35	89.62	85.88	86.95 $\pm$ 2.32

UniFAD: Ablation Study

Model	Modules			Overall
	Shared Layers	Branching	k Means	TDR (%) @ 0.2% FDR
JointCNN	✓			63.89
$\mathcal{B}_{Semantic}$		✓		86.17
$\mathcal{B}_{Random}$		✓		53.95 $\pm$ 08.02
$\mathcal{B}_{kMeans}$		✓	✓	89.67
SharedSemantic	✓	✓		92.44
<i>Proposed</i>	✓	✓	✓	94.73

Inside vs. Outside Partition



Cascade

