Biometric Recognition and Identification: A Useful Testbed for Artificial Intelligence

Rama Chellappa
Bloomberg Distinguished Professor
Departments of Electrical and Computer Engineering and Biomedical Engineering
Johns Hopkins University
Baltimore, Maryland, USA

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U. S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

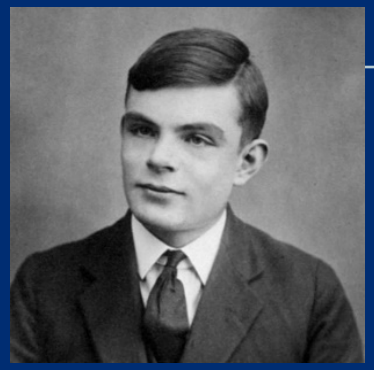
Outline

- Biometric Recognition and AI
- Bias detection and mitigation in face recognition
- Face detection and recognition at range
- Gait recognition
- Facial privacy protection

Thanks to

- Prithviraj Dhar, Ayush Gupta, Yuxiang Guo, Siyuan Huang, Zhaoyang Wang, Basudha Pal, Vishal Patel, Cheng Peng, Ram Prabhakar, and Maitreya Suin

The Life of AI



"Can machines think?"

50's 60's 70's 80's 90's 00's 10's 2024

Credit to Dr. Robert Hummel

Search

Theorem proving
Heuristic search
Planning
Chess, Go
Search engines

Problem Solving → **Game Strategies** → **Browsers**

Common-sense reasoning

Language

Knowledge representation
Machine translation

Natural Language Processing (NLP) → **Language Understanding**

Rule-based
Statistical
Speech to text
Text analytics
Voice assistants, LLMs

Vision

Computer Vision

Scene understanding
Robotics
Autonomous vehicles
Deep learning
Generative AI

Expertise

Expert Systems → **Recommenders**

Prospector
Medical diagnosis
Configuration design

AI tried everything else before finally settling on data

Good: What can AI do? A lot!

Optical character recognition
Fingerprint, face matching
Homeland security, Defense
Autonomous driving
Intelligent agents
Human computer interaction
Commerce
Education
Art, Music
Medicine
Material discovery
Deepfakes, spreading misinformation
…..

Biometric Recognition

Medical Imaging

Self-driving car

???

AI and Biometrics- Living Together for Five Decades

- 1970's
  - Planning for facial feature detection, face recognition using distances between fiducial points
  - Michael Kelly, Visual identification of people by computer, PhD thesis in Computer Science, July 1970, AIM-130, Stanford, Takeo Kanade, 1973
- 1980's
  - Neural networks for face detection, recognition
- 1990's
  - PCA, LDA, SexNet, dynamic link architecture, Bayesian methods, iris
- 2001-2009
  - Facial attributes, Sparse representations, remote face recognition, gait
- 2010-2019
  - Deep learning, deep fakes, presentation attacks
- 2020-
  - Remote face recognition, body, gait-based biometrics

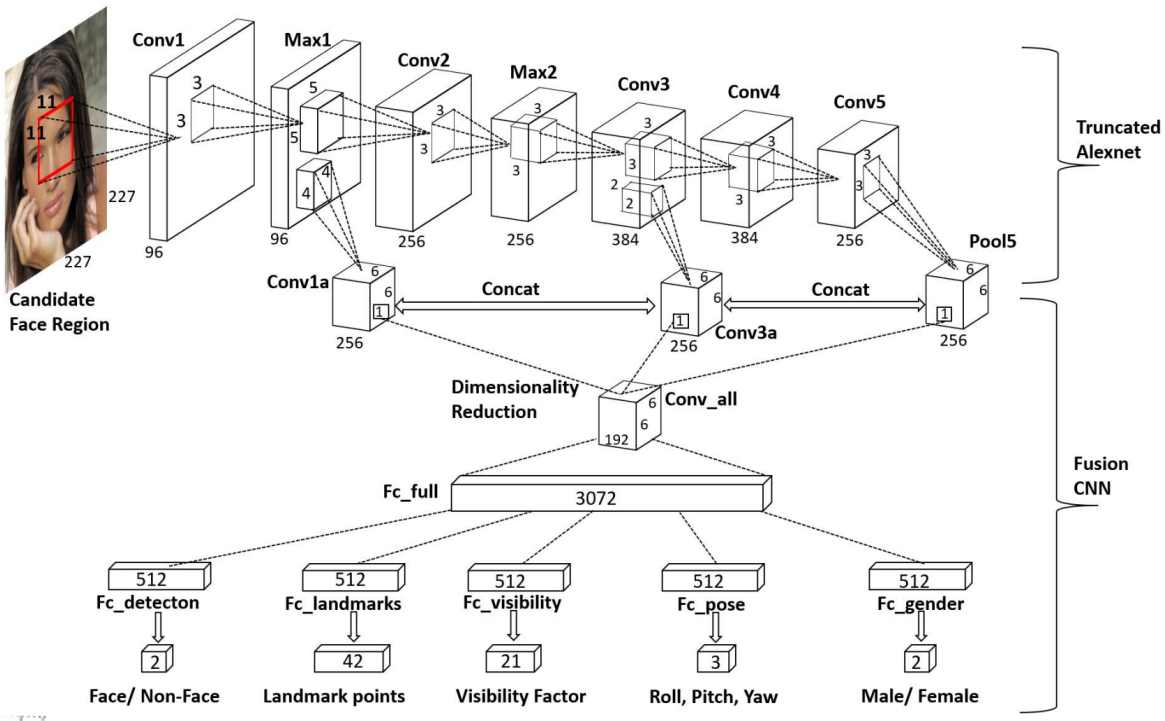
AI and Biometrics Share the Same Challenges/Concerns

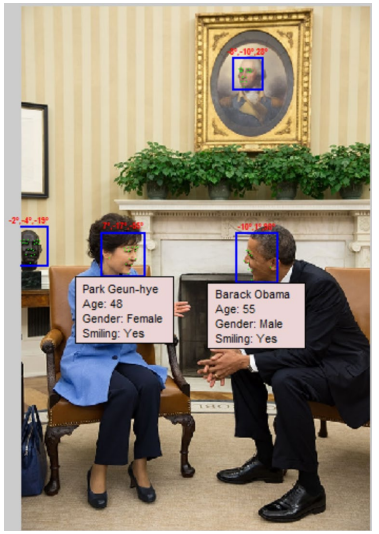
- Bias, Fairness (Biometrics2AI)
  - Face recognition
- Robustness to adversarial attacks (Both ways)
  - Adversarial attacks, presentation attacks, deepfakes
- Interpretability (AI2Biometrics)
  - Black box intelligence
- Domain adaptation/generalization (AI2Biometrics)
  - Visible to thermal, robustness to atmospheric turbulence, high pitch angles
- Generative AI
  - Markov random fields, GANs, Diffusion models, LLMs, …

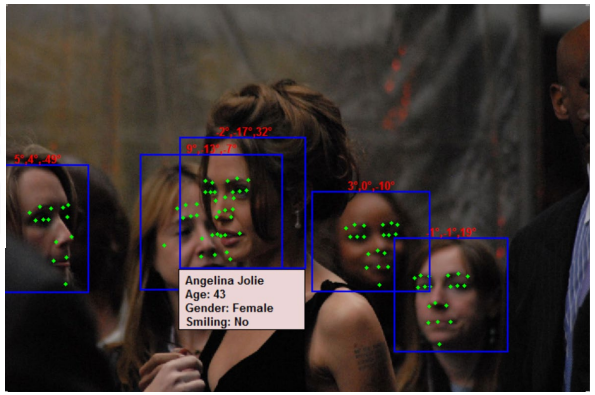
Deep Learning System for Unconstrained Face Recognition

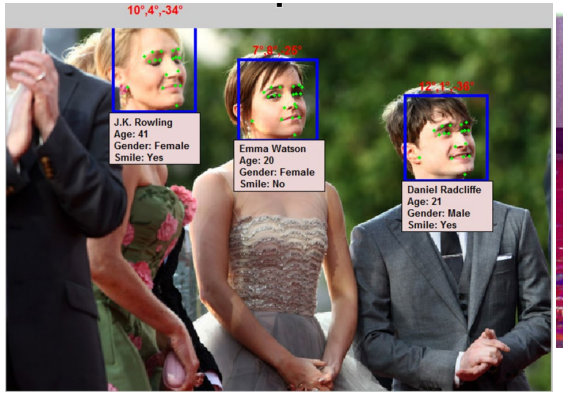
- 2014 – 2020, Supported by the IARPA JANUS program
- UMD (Lead) with CMU, Columbia, JHU, UB, UCCS, UTD.
- Multi-task learning in deep networks
  - Face and gender detection, pose and age estimation, fiducial extraction
- Network of networks
  - Fusion of short and tall networks
- Current template size is 384 floats (1536 bytes or 12288 bits)
  - Hashing reduces size to 3072 bits
- State-of-the art performance on face verification, search, clustering tasks using relatively small training data set.
- Implications to forensics (Collaborations with Jonathon Phillips, and Alice O'Toole) – Proc. National Academy of Sciences, May 28, 2018.
- Integrated into routine operations of USG agencies

Multi-task Learning for Face Recognition







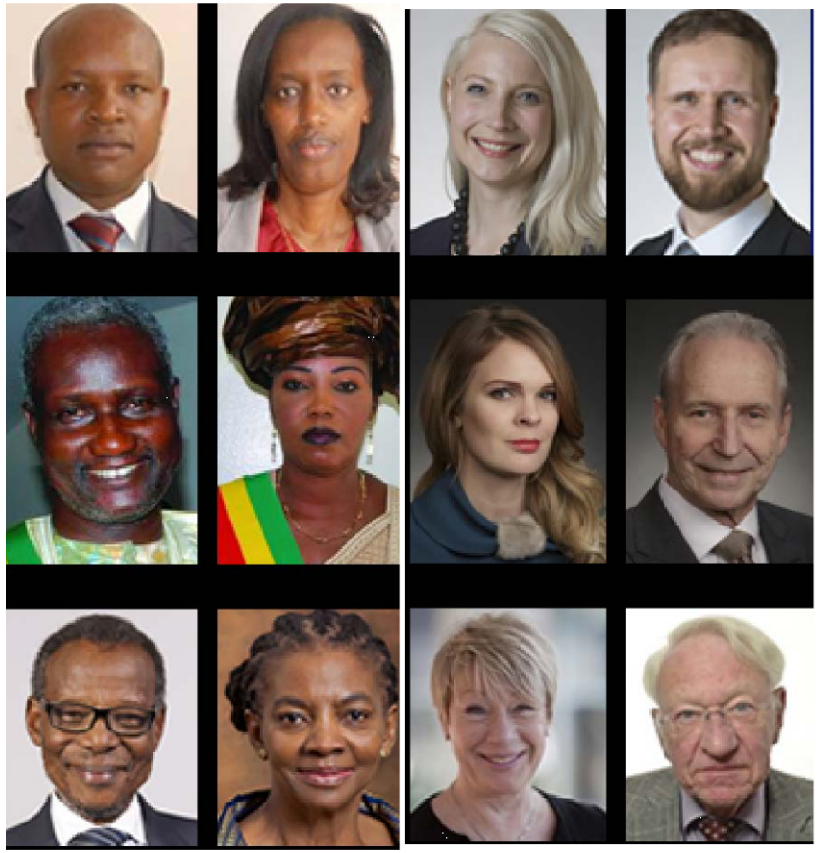


IJB-B (Multi-shot Videos)

CS6 (single-shot surveillance Videos)

The PPB dataset

AFRICAN SCANDINAVIAN



6.3%

20.8%

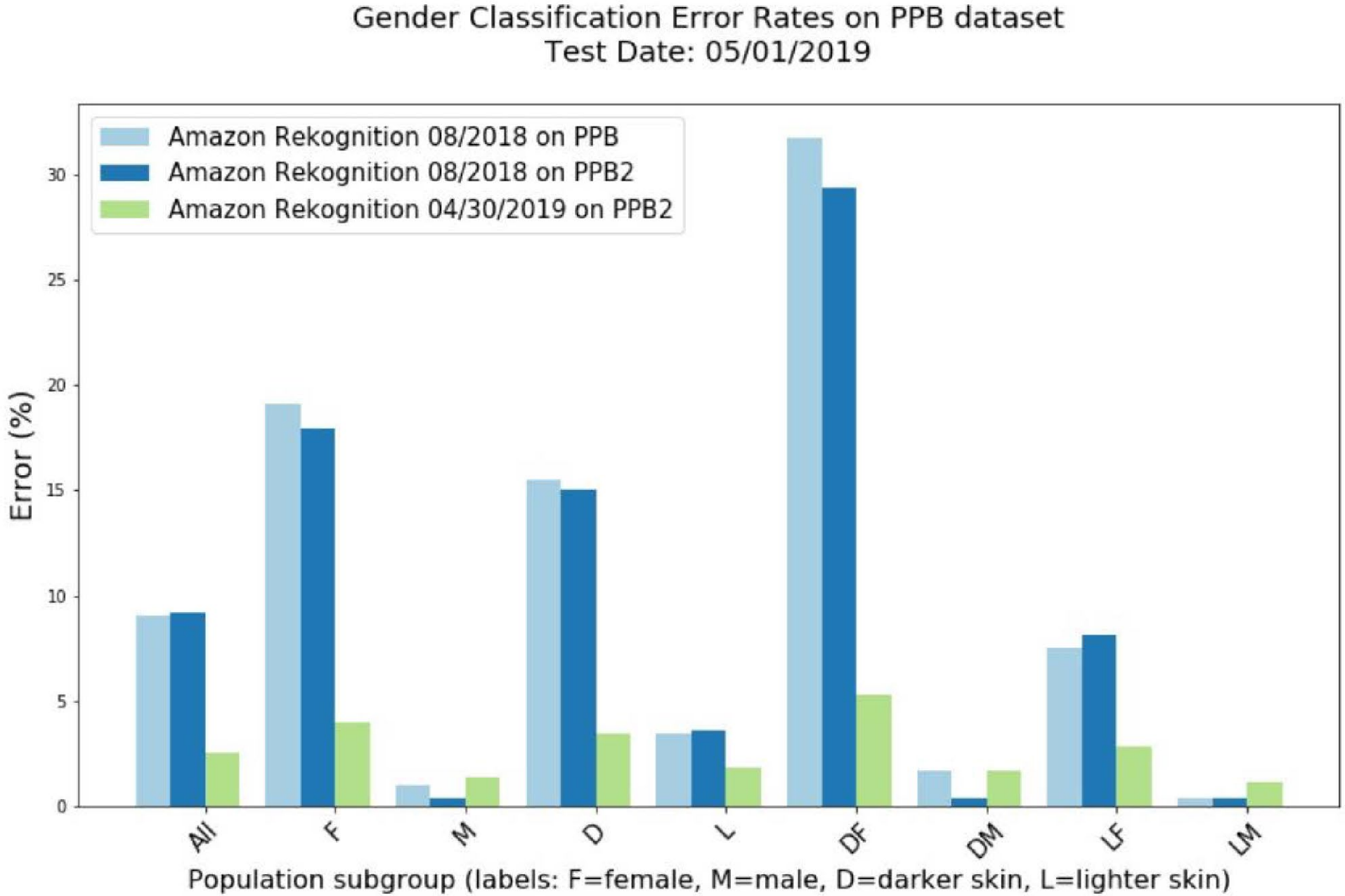
Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
A	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
B	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
C	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

PPB: Pilot Parliaments Benchmark

GenderShades.Org

[Buolamwini 2018]

[Buolamwini & Gebru 2018]


Gender Classification Error Rates on PPB dataset
Test Date: 05/01/2019
Amazon Rekognition 08/2018 on PPB
Amazon Rekognition 08/2018 on PPB2
Amazon Rekognition 04/30/2019 on PPB2
Error (%)
0
5
10
15
20
25
30
All
F
M
D
L
DF
DM
LF
LM
Population subgroup (labels: F=female, M=male, D=darker skin, L=lighter skin)

What is bias in computer vision?

- Better performance on a particular demographic
- “Correlation” of the output to certain sensitive attributes such as race, gender etc. in case of tasks like face recognition.
- Types of bias: Data bias, Algorithmic bias, Cognitive bias

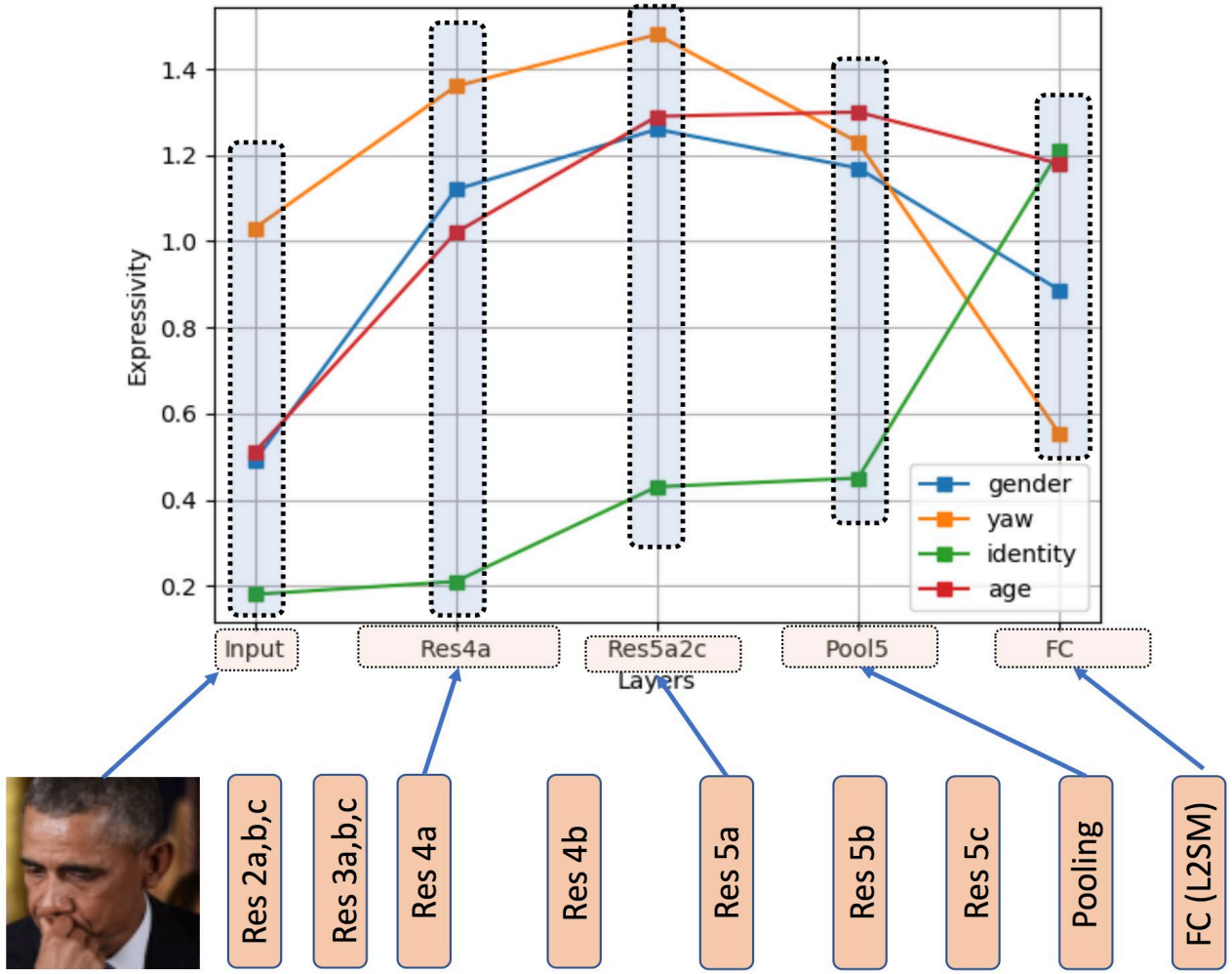
Expressivity of Facial Attributes

- Expressivity of an entity = the ease with which that entity can be predicted using a given set of features.

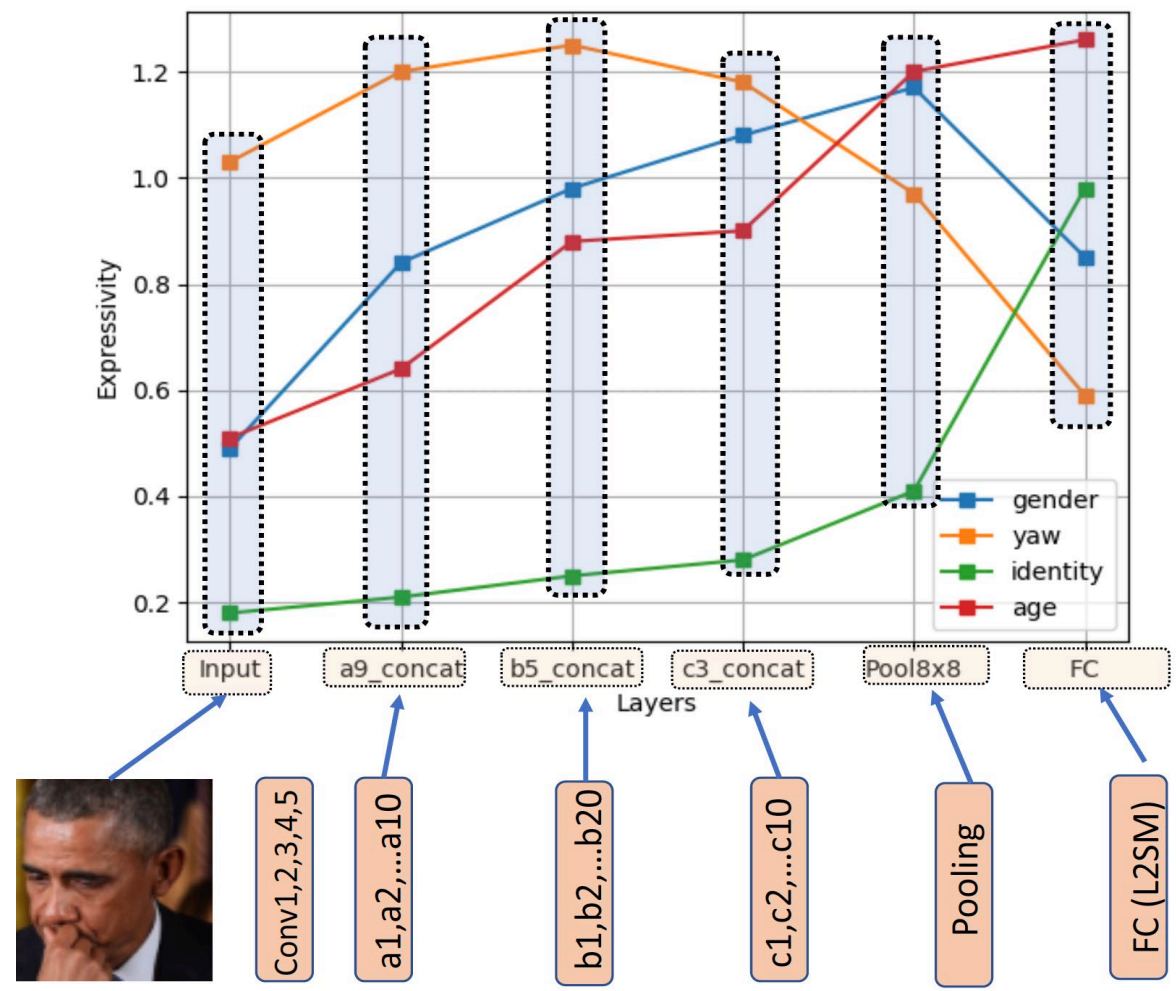
- We compute expressivity of facial attributes (yaw, age, gender, identity) in a given set of face descriptors

- To compute expressivity, we approximate the mutual information (MI) between features and attributes, by using an existing approach called Mutual Information Neural Estimation (MINE) [Belghazi et. al, ICML 2018].

Expressivity of Yaw, Gender and Age



Crystalface (Resnet-101)



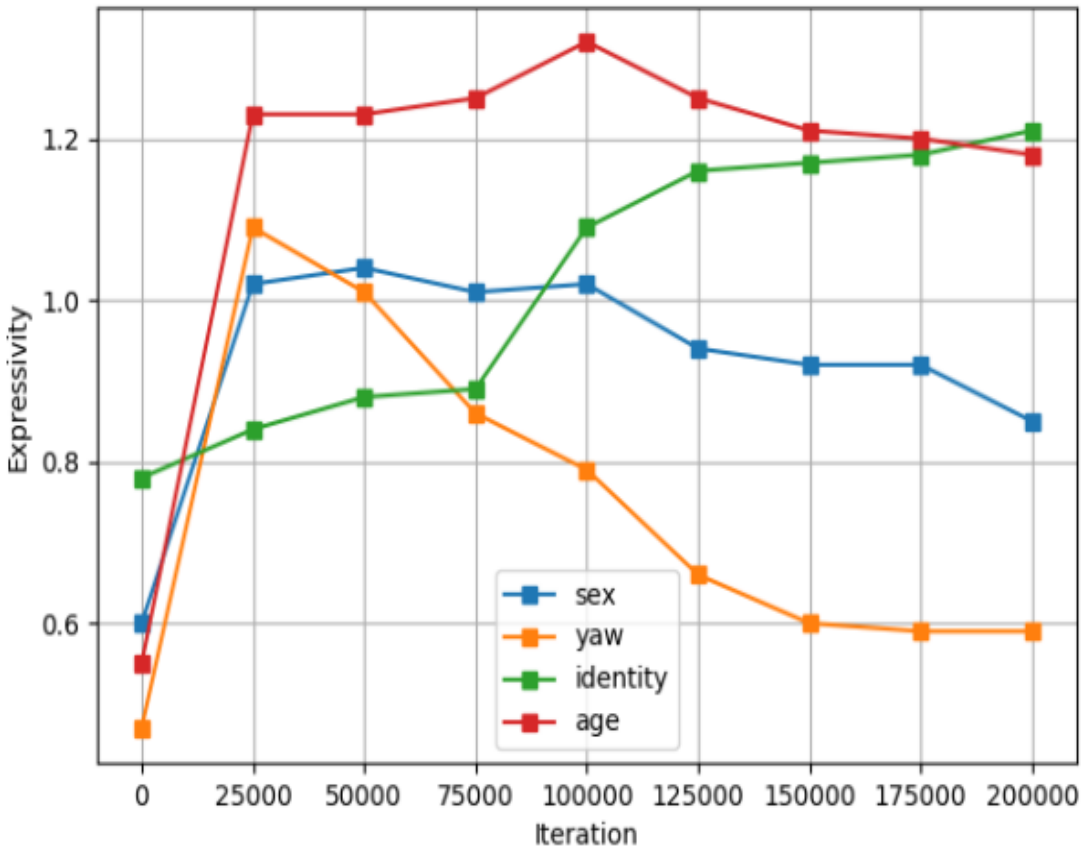
Inception Resnet-v2

Results: Hierarchical course of feedforward pass

- Expressivity of yaw, sex and age is high and that of identity is the lowest in the shallower layers. (*Yaw and sex are high level face features as compared to identity, which cannot be extracted using shallow layers*).
- In the final (fully connected) layer, expressivity of yaw and sex attain their lowest values, whereas identity and age have very high expressivity.
  (*Identity and age are more fine-grained features compared to other attributes*).
- Comparing the expressivity values of all attributes except identity in the final layer, we can infer that for identity recognition, yaw is the least important and age is the most important attribute.

Results

CrystalFace



Inception Resnet v2

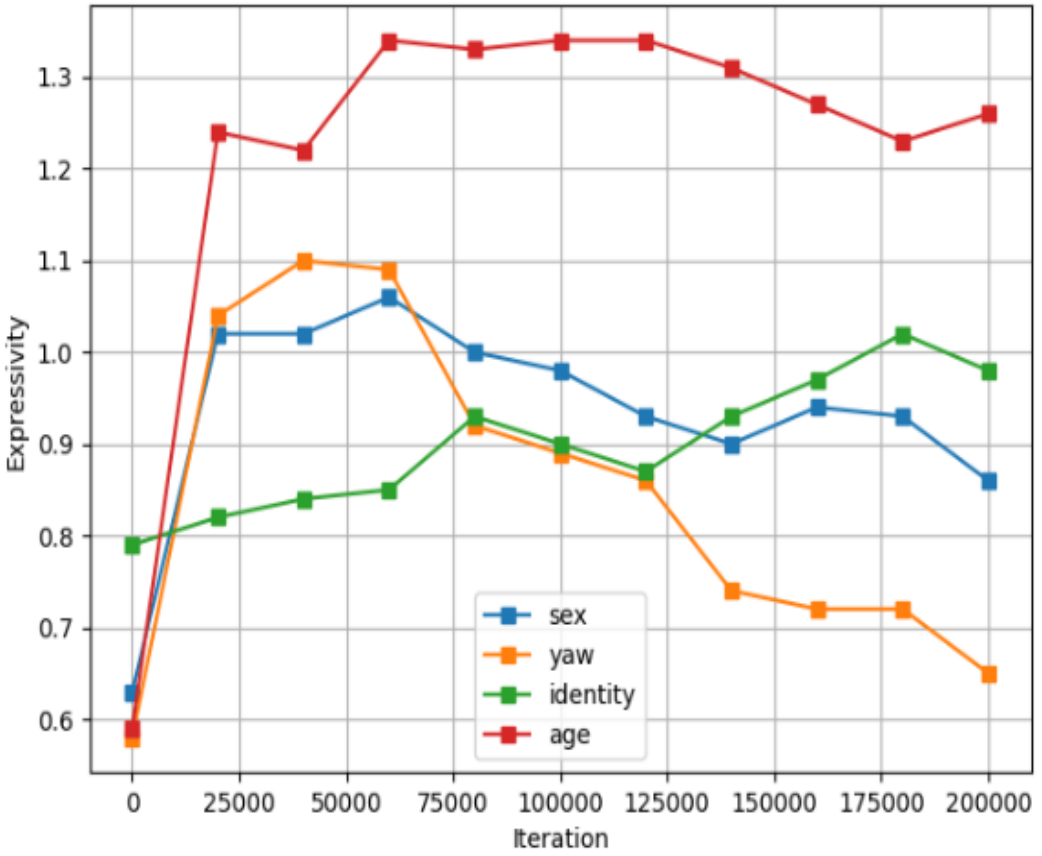


Fig. 2: Expressivity of identity, age, sex and yaw in final layer (FC-L2S) features of a.) Network A and b.) Network B. Decreasing expressivity of task irrelevant attributes (yaw, age, sex) is a part of training. Observed rate of decrease : Age < Sex < Yaw

Results: Temporal course of training

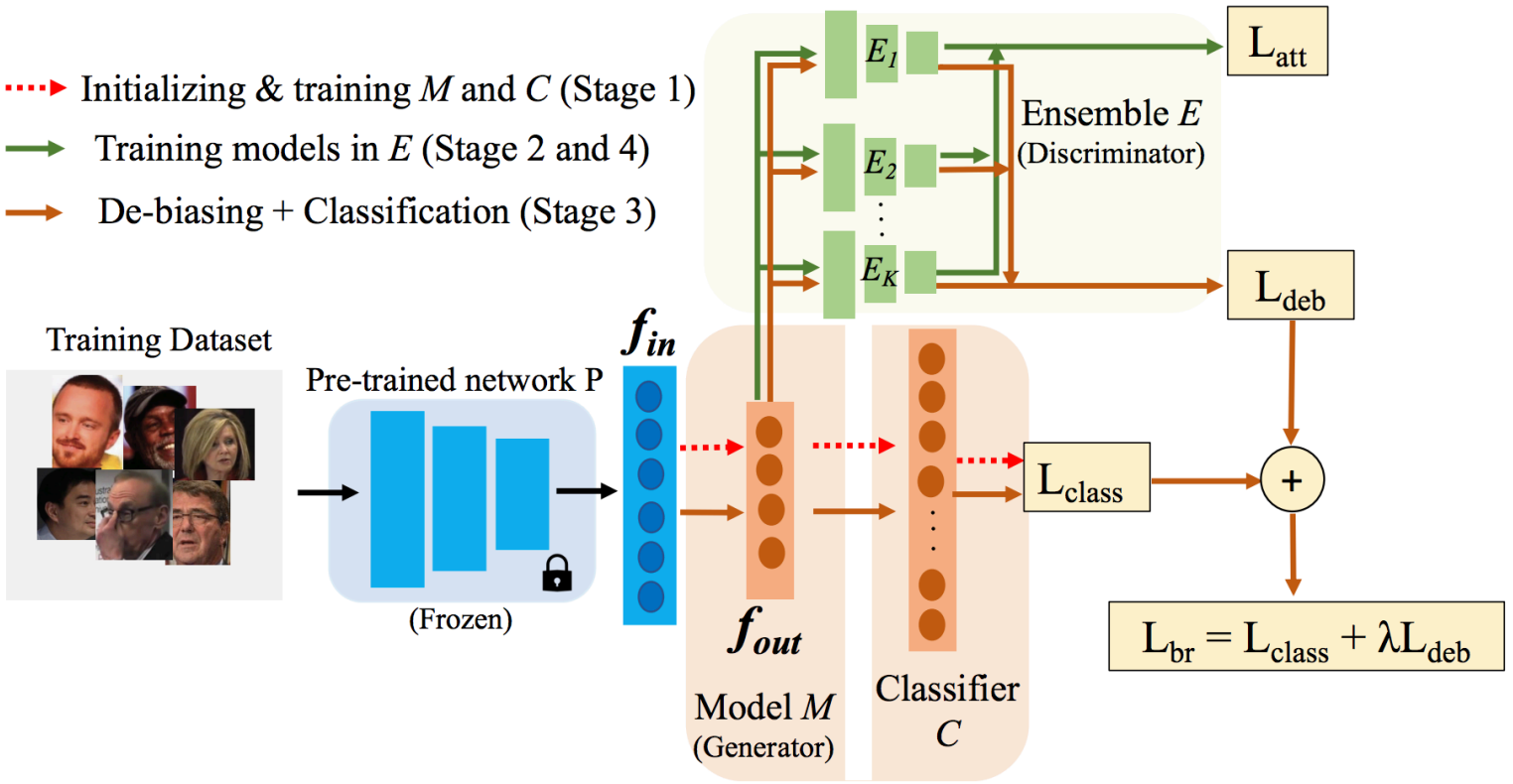
- We find that the expressivity of yaw, age and sex reach their peak values in the first 25000 iterations for CrystalFace and 40000 iterations for Inception Resnet v2, to learn the general concept of facial pose, age and sex.

- We find that the yaw expressivity decreases rapidly as the training proceeds, showing that making features almost agnostic to pose variance is an essential part of the training process.

- Expressivity of age and sex decreases slightly after their corresponding expressivity peaks are attained, during the course of training. However, compared to yaw, the rate of decay in the expressivity of age and sex is low. This shows that age and sex are more important for identity recognition than face yaw.

- Observing the expressivity values in the final iteration, we can infer that for identity recognition, the following is the order of relevance of attributes for which the network does not receive any supervision : Age > Sex > Yaw. The opposite of this order is observed in the rate by which the expressivity values of yaw, age and sex decreases, i.e. the rate of decrease is : Age < Sex < Yaw. This is true for both CrystalFace and Inception Resnet v2.

Protected Attributes Suppression System (PASS)

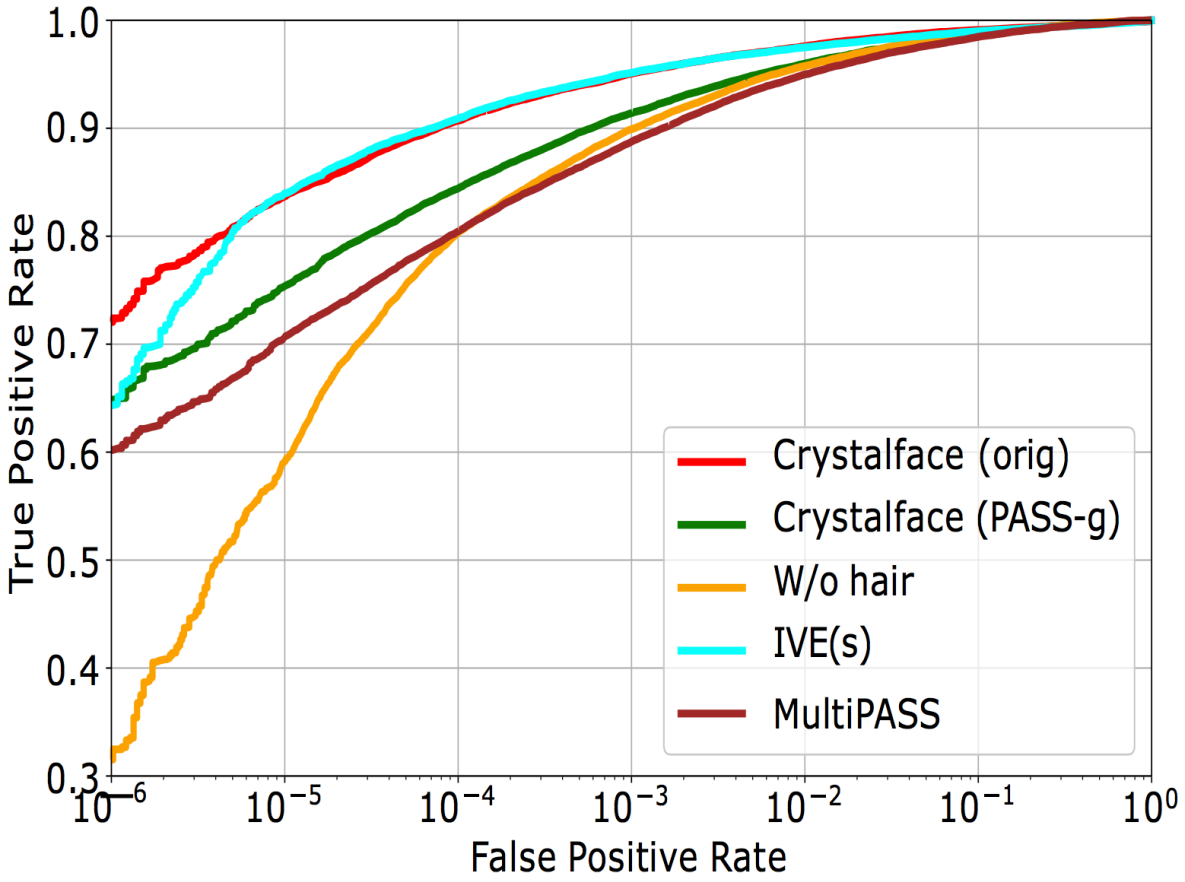


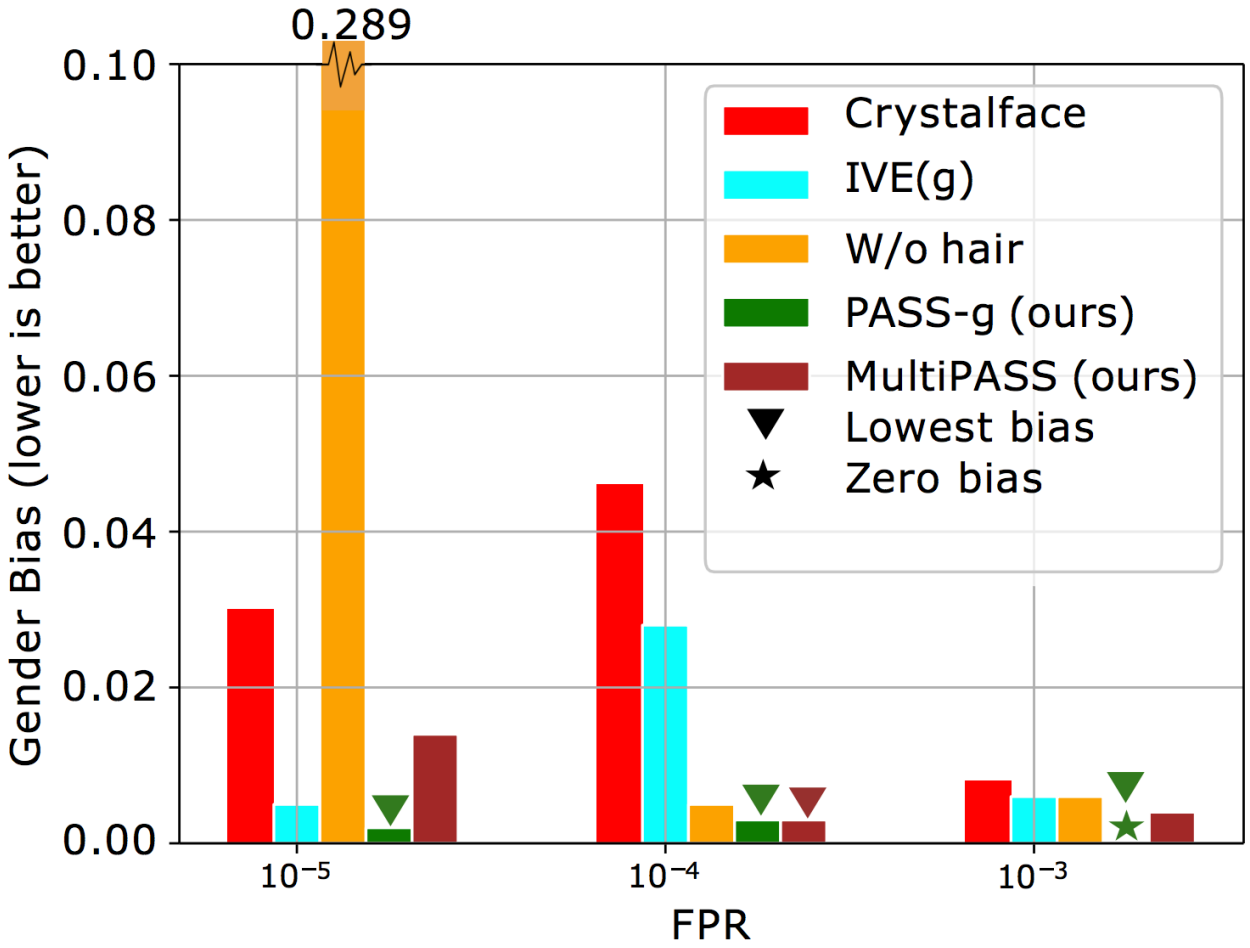
L_{class}: Classification loss for training **M** to classify identities

L_{att}: Classification loss for **E** (discriminator) to classify sensitive attribute (gender/race)

L_{deb}: Adversarial loss to discourage **M** from encoding gender/race information

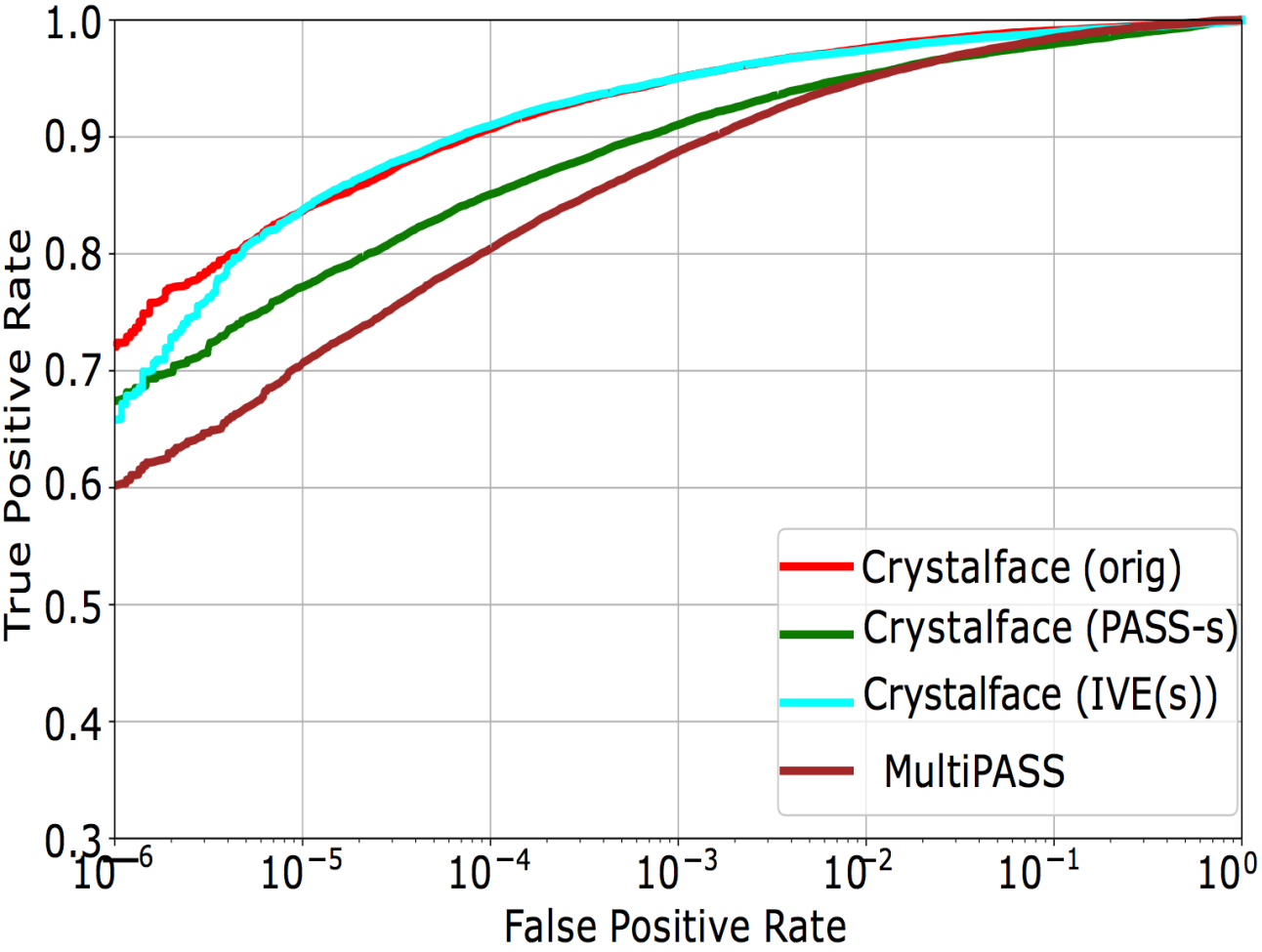
Results (Crystalface – Gender Bias)

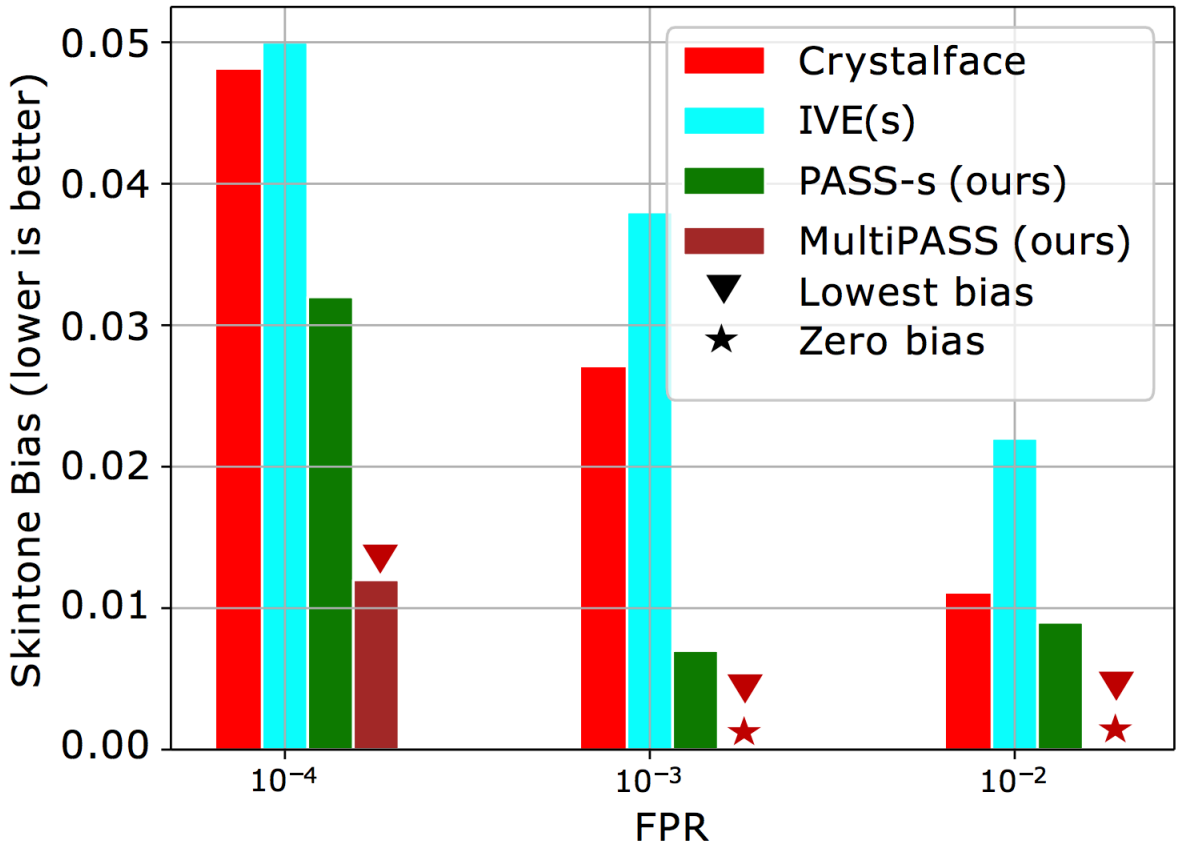




Crystalface: JANUS-UMD face matcher
IVE: Incremental variable elimination

Results (Crystalface – Skintone Bias)





Bias-performance Tradeoff

- Most adversarial de-biasing systems demonstrate a drop in face verification performance.
- An ideal face recognition system should demonstrate high bias reduction and low drop in performance.
- To measure this tradeoff between reduction in bias and drop in verification performance, we propose a new metric called Bias Performance Coefficient:

$$\mathrm{BPC}^{(F)} = \underbrace{\frac{\mathrm{Bias}^{(F)} - \mathrm{Bias}_{deb}^{(F)}}{\mathrm{Bias}^{(F)}}}_{\text{\% drop in bias}} - \underbrace{\frac{\mathrm{TPR}^{(F)} - \mathrm{TPR}_{deb}^{(F)}}{\mathrm{TPR}^{(F)}}}_{\text{\% drop in TPR}}$$

PASS/MultiPASS Systems Achieve High BPCs

Crystalface – Gender bias analysis

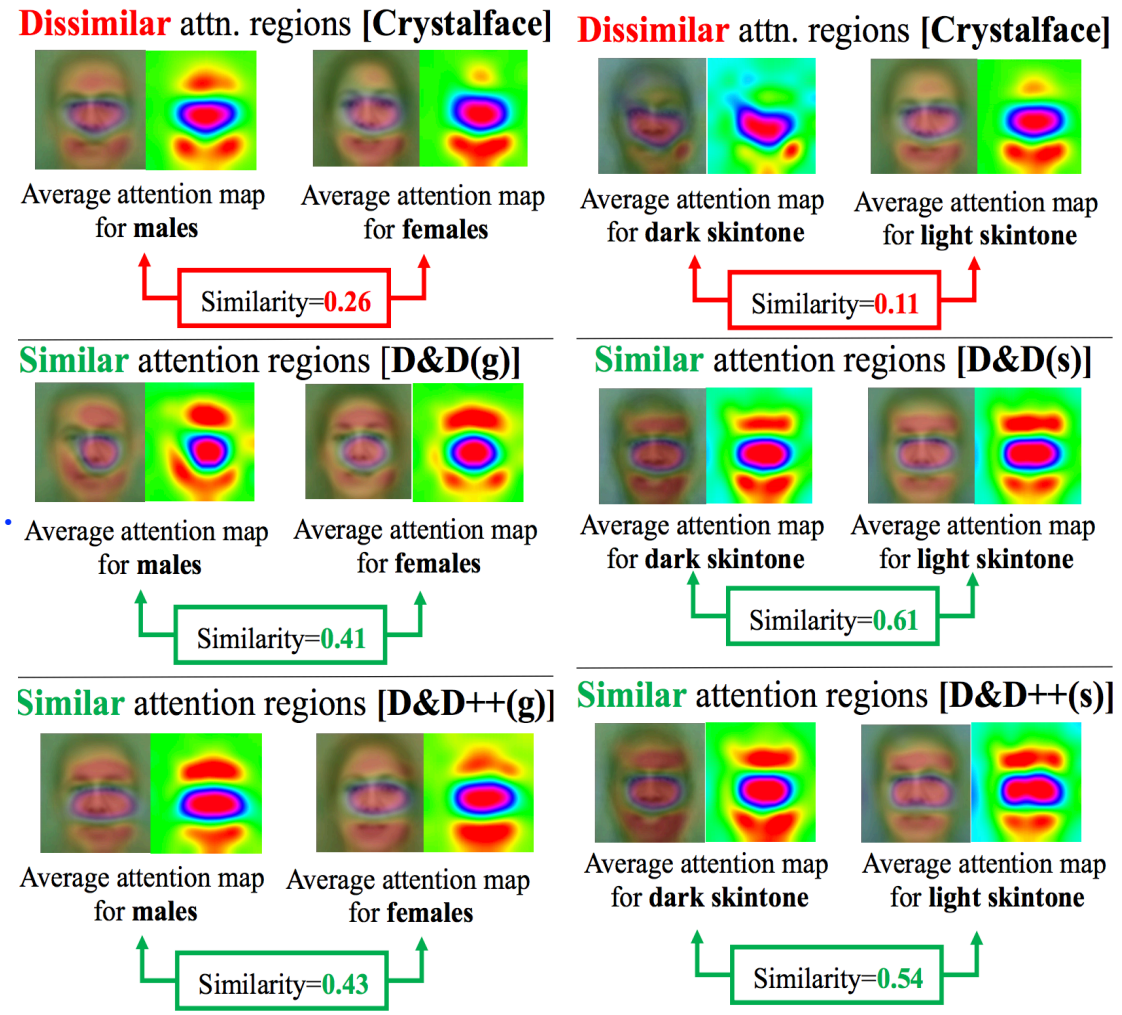
FPR		10^{-5}		10^{-4}		10^{-3}	
Network	Acc-g(↓)	TPR	BPC_g (↑)	TPR	BPC_g(↑)	TPR	BPC_g(↑)
Crystalface[36]	86.73	0.833	0.000	0.910	0.000	0.951	0.000
W/o hair[3]	86.04	0.589	-8.926	0.780	0.823	0.899	0.195
IVE(g)[42]	86.10	0.833	<u>0.833</u>	0.910	0.391	0.951	0.250
PASS-g	<u>80.54</u>	0.761	**0.847**	0.839	**0.857**	0.921	**0.968**
MultiPASS	**76.31**	0.708	0.383	0.809	<u>0.823</u>	0.881	<u>0.426</u>

Crystalface – Skin tone bias analysis

FPR		10^{-4}		10^{-3}		10^{-2}	
Network	Acc-st (↓)	TPR	BPC_{st}(↑)	TPR	BPC_{st}(↑)	TPR	BPC_{st}(↑)
Crystalface[36]	89.30	0.910	0.000	0.950	0.000	0.974	0.000
IVE(s)[42]	88.26	0.910	-0.041	0.950	-0.407	0.974	-1.000
PASS-s	<u>83.84</u>	0.844	<u>0.261</u>	0.914	<u>0.702</u>	0.919	<u>0.125</u>
MultiPASS	**79.44**	0.809	**0.639**	0.881	**0.927**	0.968	**0.994**

Equalizing the Activation Maps

D&D(g) and **D&D++(g)** generate more similar attention maps for male and female frontal faces.



D&D(s) and **D&D++(s)** generate more similar attention maps for dark and light frontal faces.

Increasing Data Diversity as a solution to bias

- Diffusion models are biased in generation. Some control mechanisms or even vanilla conditioning is better than just using unconditional diffusion models.
- Vanilla balancing of data is not always sufficient to reduce bias [1]
- Augmenting using generated images to diversify dataset and improve performance has been explored
- We conduct experiments to see the effects of bias when conditional diffusion models are used.
- Diffusion generated data effectively supplement the underrepresented dataset, serving as a regularization technique to enhance feature learning.

[1] Wang, T., Zhao, J., Yatskar, M., Chang, K.W. and Ordonez, V., 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5310-5319).

Bias Reduction Using Synthesis

- Augmenting original datasets using conditional diffusion model generated images helps reduce bias in downstream target classification tasks in face images.
- The increase in overall accuracy and the reduction in bias is more pronounced in smaller datasets than larger datasets.

Pal, B., Roy, A., Kathirvel, R.P., O'Toole, A.J. and Chellappa, R., 2024, September. DiversiNet: Mitigating Bias in Deep Classification Networks across Sensitive Attributes through Diffusion-Generated Data. In *2024 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-10). IEEE.

DiversiNet

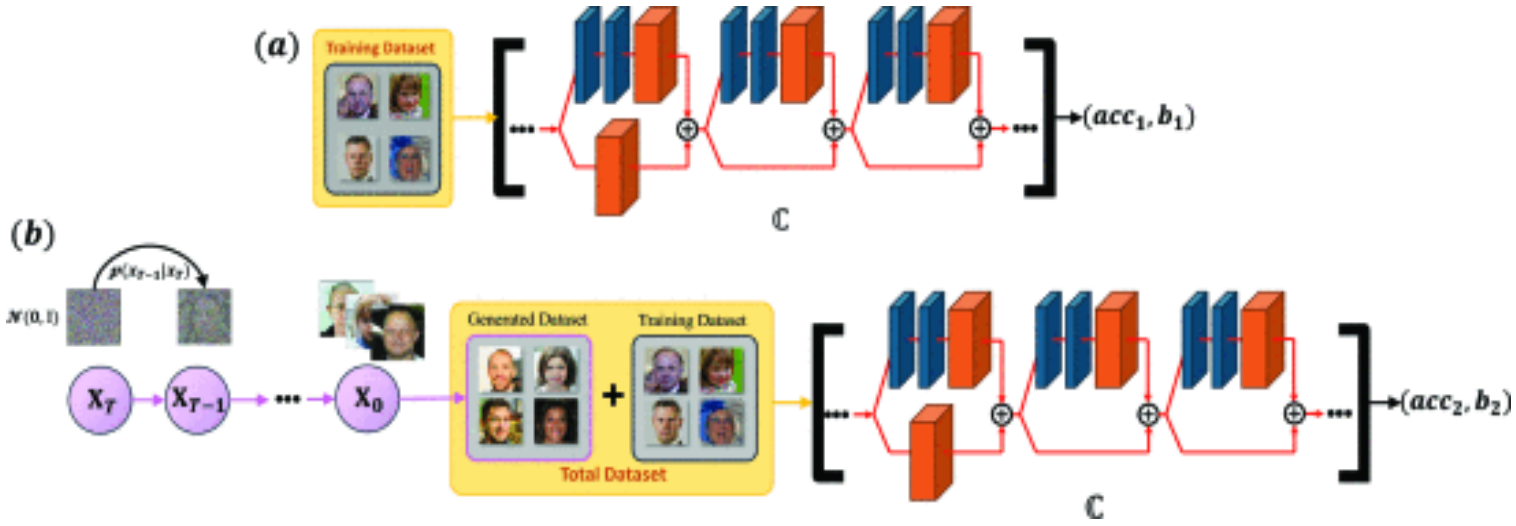


Fig 1: The overall block diagram for the proposed approach. In (a) the sensitive attribute classifier is trained only on original images from the dataset. In (b) the sensitive attribute classifier is trained on original data+diffusion-generated data. For a small dataset regime, we show that upon adding data generated by a conditional diffusion model, if we retrain the sensitive attribute classifier, we can obtain an increase in accuracy and a reduction in bias (**acc1**<**acc2** while **b1** > **b2**)

Implementation Details :

- We form smaller, balanced datasets from the original training sets given the perceived difficulty of the tasks. These reduced datasets aid in detecting more pronounced biases and accuracy changes resulting from augmentation effects.
- In single attribute scenarios, three class-conditioned diffusion models generate an equal number of images, incorporating class label information into the UNet architecture.
- For multi-attribute tasks, where we address bias within attributes or auxiliary attributes, the diffusion model is conditioned on two attributes simultaneously to produce an equal number of images per class.

Results

Dataset	Task	Bias (↓)			BPC [5](↑)			Overall Accuracy (↑)		
		Baseline-I	Baseline-II	**Ours**	Baseline-I	Baseline-II	**Ours**	Baseline-I	Baseline-II	**Ours**
FFHQ	Gender clf	0.0340	0.0437	**0.0204**	0	-0.2826	**0.4051**	93.92%	94.21%	**94.44%**
	Age clf	0.0700	0.0537	**0.0195**	0	0.2353	**0.7373**	89.88%	90.13%	**91.24%**
	Race clf	✗	✗	✗	✗	✗	✗	✗	✗	✗
FairFace	Gender clf	0.0652	0.0798	**0.0438**	0	-0.2187	**0.3560**	84.97%	85.40%	**88.08%**
	Age clf	0.2821	0.0243	**0.0096**	0	0.9433	**0.9849**	80.47%	83.12%	**82.28%**
	Race clf	0.1786	0.1844	**0.0559**	0	-0.0286	**0.6965**	76.56%	76.86%	**77.29%**

Tab 1: Comparison of gender, age, and racial bias analysis and overall accuracy of classifiers (clf denotes classification) trained on only small original dataset (Baseline I), trained on original data + cGAN generated data (Baseline-II) and those trained on original data + conditional diffusion-generated data. (Best in bold)

$$\mathrm{BPC}(A_t) = \frac{\mathrm{Bias}^{(A_p)} - \mathrm{Bias}_{deb}^{(A_p)}}{\mathrm{Bias}^{(A_p)}} - \frac{\mathrm{TPR}^{(A_p)} - \mathrm{TPR}_{deb}^{(A_p)}}{\mathrm{TPR}^{(A_p)}}$$

$$\mathrm{Bias}(A_t) = \frac{\sum_{i=1}^{A_{p,.}} \sum_{j=i+1}^{A_{p,r}} (\mathcal{T}_i - \mathcal{T}_j) \cdot w_i \cdot w_j}{\sum_{i=1}^{A_{p,.}} \sum_{j=i+1}^{A_{p,r}} w_i \cdot w_j}$$

Datasets	**Protected attributes** (A_p)	**Target classification task** (A_t)	**Bias (↓)**		**BPC (↑)**		**Overall accuracy (%) (↑)**	
			Baseline-I	DiversiNet (ours)	Baseline-I	DiversiNet (ours)	Baseline-I	DiversiNet (ours)
FFHQ	Age-Gender	Smile	0.0221	**0.0201**	0	**0.1163**	94.31	**95.18**
		Age	0.0700	**0.0058**	0	**0.9318**	89.88	**91.24**
		Gender	0.0117	**0.0029**	0	**0.7598**	94.06	**94.99**
		Glasses	0.0225	**0.0208**	0	**0.0892**	98.09	**98.33**
		Hair Color	0.1224	**0.1089**	0	**0.0901**	80.73	**81.40**
FF	Age-Race	Age	0.2821	**0.0580**	0	**0.8184**	80.73	**82.78**
		Race	0.1786	**0.1322**	0	**0.2581**	76.56	**77.32**
		Gender	0.0652	**0.0385**	0	**0.4335**	84.97	**87.09**
	Gender-Race	Age	0.2821	**0.0134**	0	**1.0545**	80.74	**89.99**
		Race	0.1786	**0.0482**	0	**0.7502**	76.56	**77.49**
		Gender	0.0652	**0.0476**	0	**0.2721**	84.97	**85.26**
	Age-Gender	Age	0.2821	**0.1342**	0	**0.5338**	80.74	**81.52**
		Race	0.1786	**0.0249**	0	**0.8629**	76.56	**76.74**
		Gender	0.0652	**0.0029**	0	**0.9712**	84.97	**86.33**

Tab 2: Analyzing bias and performance across multiple attributes shows simultaneous bias mitigation in all Ap. Quantitative evaluation of our method on FFHQ and FairFace datasets consistently demonstrates its effectiveness. In our assessment, ↑ indicates higher BPC and overall accuracy, signifying superior systems, while ↓ indicates preferable lower bias values. (Best in bold)

Some concerns to be addressed

- Diffusion models are themselves biased! [2] Conditioning reduces it which is why we saw good results in the previous case
- However, it can get really complicated to have multi-class conditioning and is also computationally not effective.
- We introduce a novel approach to address this issue by debiasing the attributes in the images generated by diffusion models. Our approach involves disentangling facial attributes by localizing the means within the latent space of the diffusion model using Gaussian mixture models (GMM).

[2] Perera, M.V. and Patel, V.M., 2023, September. Analyzing bias in diffusion-based face generation models. In 2023 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-10). IEEE.

GAMMA-FACE

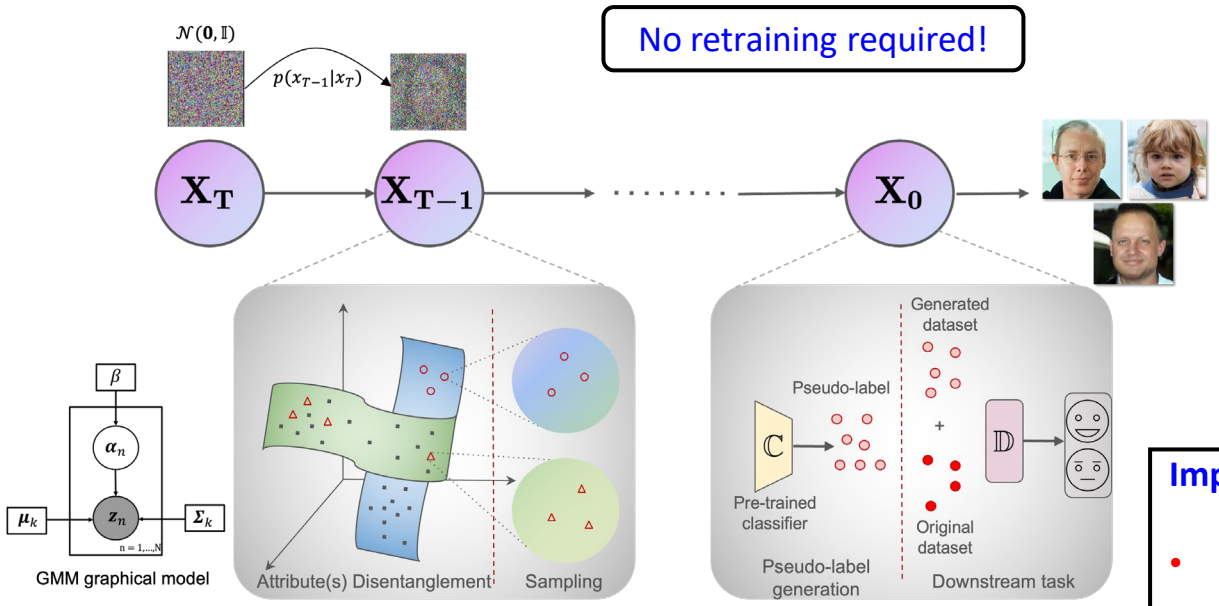
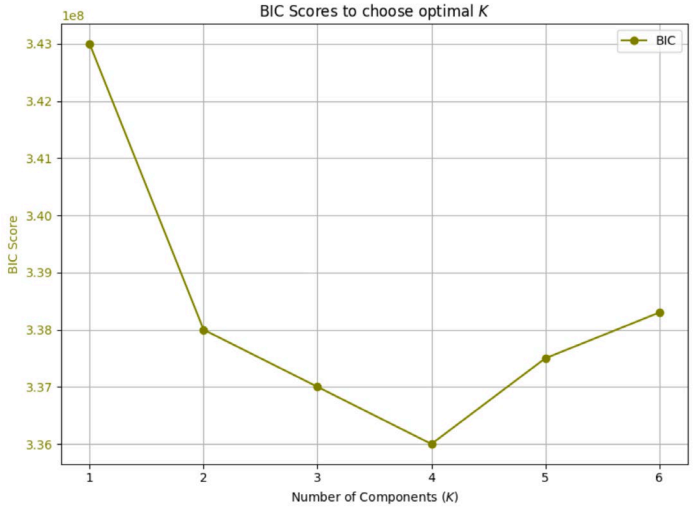


Fig 2: The reverse diffusion phase of an unconditional diffusion model where we use a Gaussian Mixture Model (GMM) with parameters μk, Σk, and π as represented in the graphical model to disentangle latent codes into K components. Images are sampled uniformly from each component, and pre-trained classifiers (C) provide pseudo-labels for attributes. To mitigate bias and improve performance, we retrain our downstream classifier (D) using original image-label pairs (dark red dots) along with generated images and pseudo-labels for the target attribute (lighter red dots).



Implementation Details:

- We use off-the-shelf ResNet-101 classifiers for binary and three-class downstream classification tasks.
- A GMM with K=4 components (selected via BIC) is fit to the latent code. The component means are used as noise means during the reverse diffusion process to generate images.
- We generate equal number of samples from each GMM component and retrain the A_t classifiers using original training images plus generated images with pseudo-labels. Evaluation is performed on the original test set.

Results

	FairFace								
	$A_t = g \mid A_p = a, r$			$A_t = r \mid A_p = a, g$			$A_t = a \mid A_p = r, g$		
Method	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)
[34]	0.187	1.36	81.67	0.237	**1.27**	78.10	0.112	1.502	78.62
[50]	0.142	1.38	84.14	0.163	1.393	79.3	**0.097**	**1.43**	80.13
[8]	0.169	1.53	82.28	0.218	1.781	75.5	0.130	1.62	76.51
Ours	**0.088**	**1.29**	**86.5**	**0.102**	1.36	**80.23**	0.128	1.510	**81.00**
	FFHQ								
	$A_t = s \mid A_p = a, g$			$A_t = h \mid A_p = a, g$			$A_t = gl \mid A_p = a, g$		
Method	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)
[34]	0.015	**1.48**	91.68	0.221	1.787	84.03	0.028	0.995	96.50
[50]	0.0064	1.61	93.16	0.153	1.798	**88.87**	0.031	1.008	97.29
[8]	0.019	1.77	91.58	0.192	1.84	82.11	0.040	1.156	96.10
Ours	**0.0056**	1.52	**94.84**	**0.146**	**1.756**	82.81	**0.0208**	**0.987**	**98.70**

Tab 3: Quantitative assessment of GAMMA-FACE and existing debiasing techniques on FairFace and FFHQ dataset using bias evaluation metrics: Bias (B), Bias Amplification (BA) and Overall accuracy (Acc.). The up-arrow (↑) indicates higher Acc. while the down-arrow (↓) indicates lower B and BA values are preferable.

	FairFace					
	$A_t = g \mid A_p = a, r$		$A_t = r \mid A_p = a, g$		$A_t = a \mid A_p = r, g$	
Method	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)
Baseline	0	0.886	0	0.798	0	**0.769**
Ours	**0.085**	**0.801**	**0.118**	**0.740**	**0.454**	0.783
	FFHQ					
	$A_t = s \mid A_p = a, g$		$A_t = h \mid A_p = a, g$		$A_t = gl \mid A_p = a, g$	
Method	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)
Baseline	0	0.782	0	0.95	0	1.814
Ours	**0.673**	**0.698**	**0.4244**	**0.912**	**0.128**	**0.918**

Tab 4: Quantitative assessment of GAMMA-FACE and existing debiasing techniques on FairFace and FFHQ dataset using bias evaluation metrics: Bias Performance Coefficient (BPC) and KL divergence (KL). The up-arrow (↑) indicates higher BPC while the down-arrow (↓) indicates lower KL values are preferable.

Images Generated by ‘Debiased’ Model

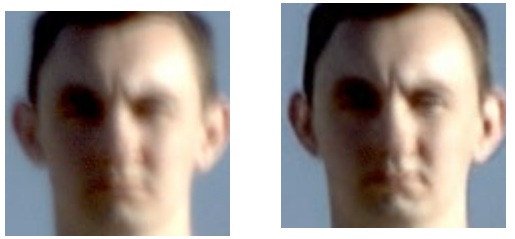
Fig. 3: An exemplar set of face images generated using debiased DDPM. **Top:** Generated from high quality FFHQ dataset. **Bottom:** Generated from FairFace dataset.

Domain Shift May Kill AI!

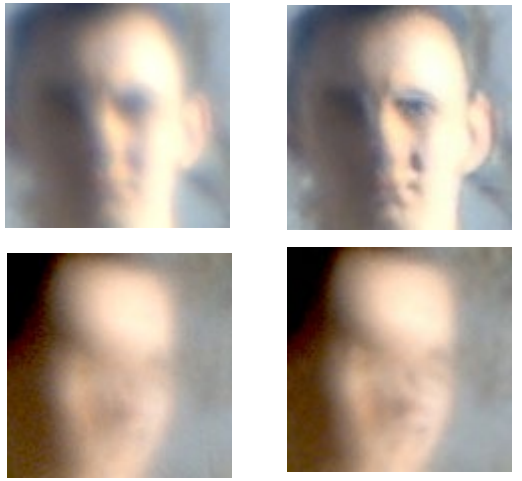
- Domain shift arises due to distribution differences between training and test data
- An old problem
- Domain adaptation assumes training and test data are concurrently available
- Domain generalization assumes the availability of just training data
- In biometric recognition, domain shift is caused by illumination, pose, expression, range variations
- Domain adaptation via synthesis – Physics-based models, GANs, diffusion models, …

Face Detection at Range

- Remote face data set
  - Dataset is divided into three ranges: 300m, 650m and 1000m
  - 2211 probes for each range
  - 100 subjects
  - Face detection algorithm is applied to detect faces
  - 1447, 188, 80 probes are detected for range 300m, 650m and 1000m respectively out of 2211
  - Restoration performs well in 300m while results in 650m and 1000m could be improved



- 1447 probes detected out of 2211
- 188 probes detected out of 2211
- 80 probes detected out of 2211



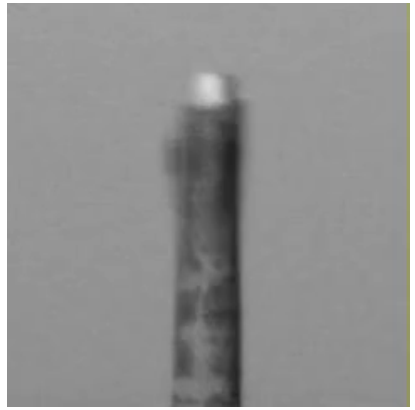
Approximated Turbulence Model

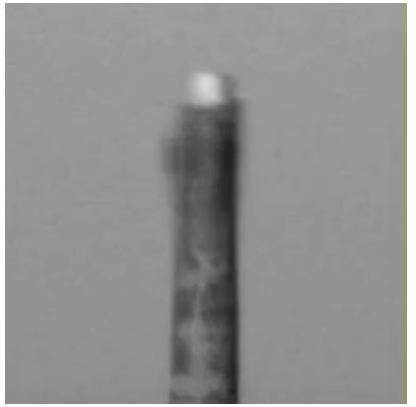
Two kinds of degradation in atmospheric turbulence: blur and deformation. Effects of the turbulent flow of air and changes in temperature, density of air particles, humidity and carbon dioxide level, the captured image is blurry and deformed due to variations in the refractive **index.**

- .
-

$$\tilde{I}_k = D_k(H_k(I)) + n_k$$

Observed frame — Deformation operator — Blurring operator — Ground truth — Noise





Real turbulence

Approximate turbulence

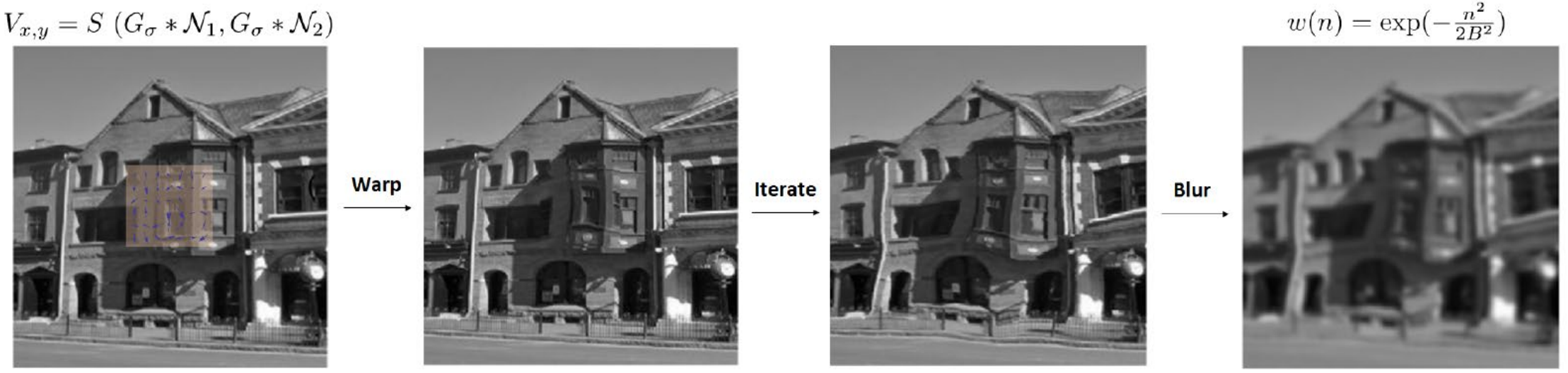
Synthetic Atmospheric Turbulence Images

- For each point (x, y), generate a random motion vector field in the patch centered at x

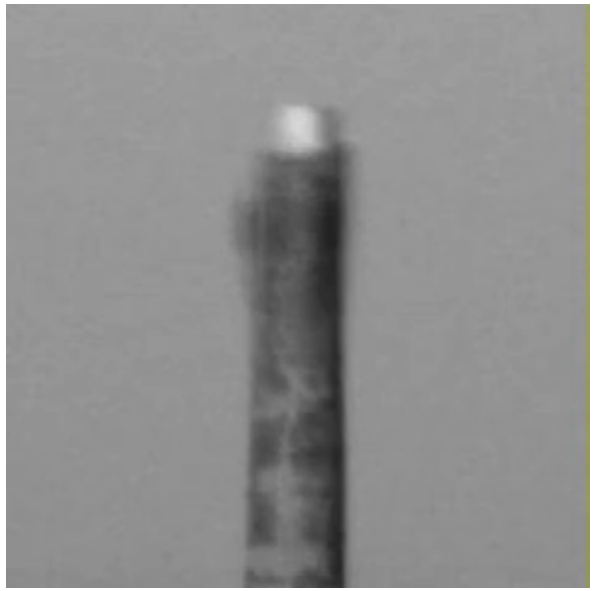
$$V_{x,y} = S(G_\sigma * \mathcal{N}_1, G_\sigma * \mathcal{N}_1)$$

Where G_σ is the Gaussian kernel with standard deviation σ, S is the strength value, $\mathcal{N}_1$ and $\mathcal{N}_2$ are randomly selected from a normal distribution.

- After warp the image with the motion vector field, apply Gaussian blur to the image



Conventional Pipeline



Original video

Subsampled video

Stabilized video

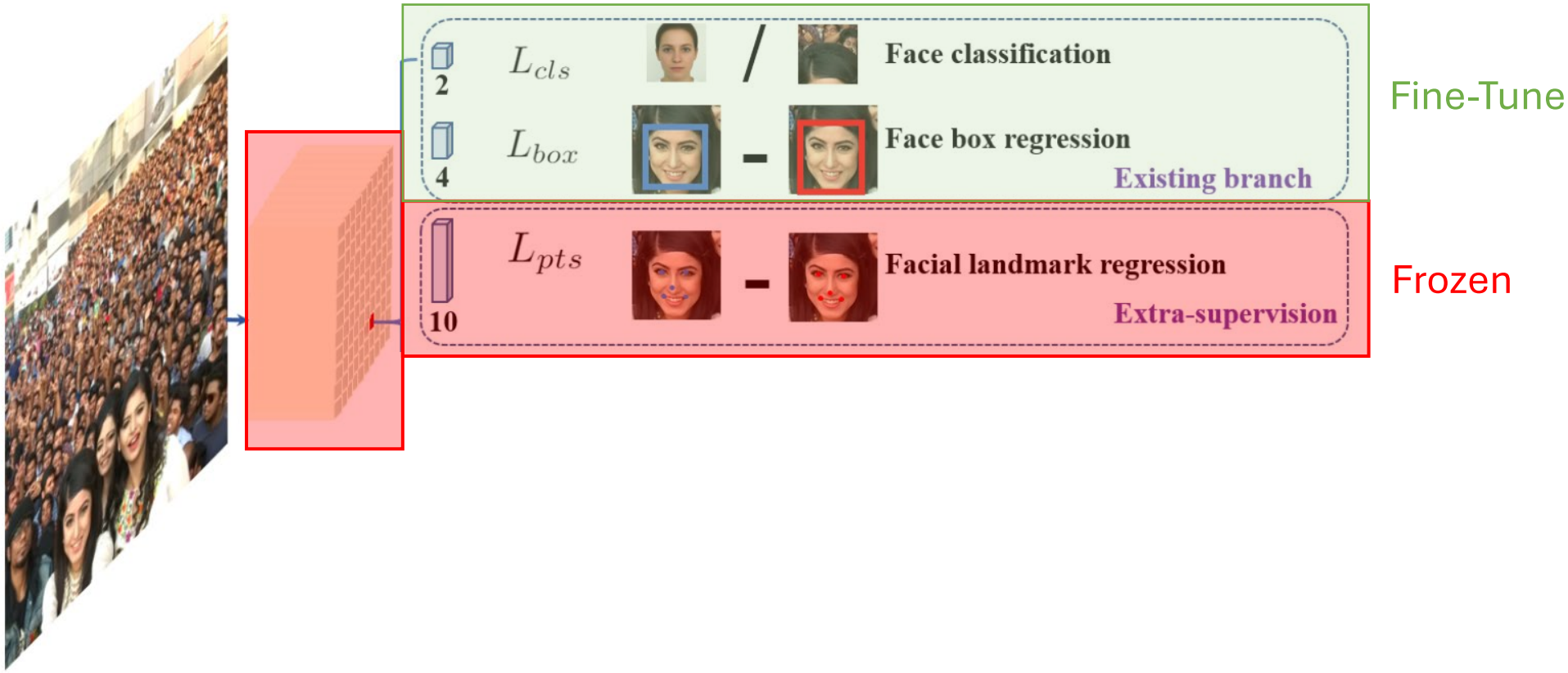
Registered video

Registration is computational costly!

Lau, C. P., Lai, Y. H., & Lui, L. M. (2019). Restoration of atmospheric turbulence-distorted images via RPCA and quasiconformal maps. Inverse Problems, 35(7), 074002.

RetinaFace_finetune

- We use a pretrained RetinaFace and finetune with unconstrained biometric data

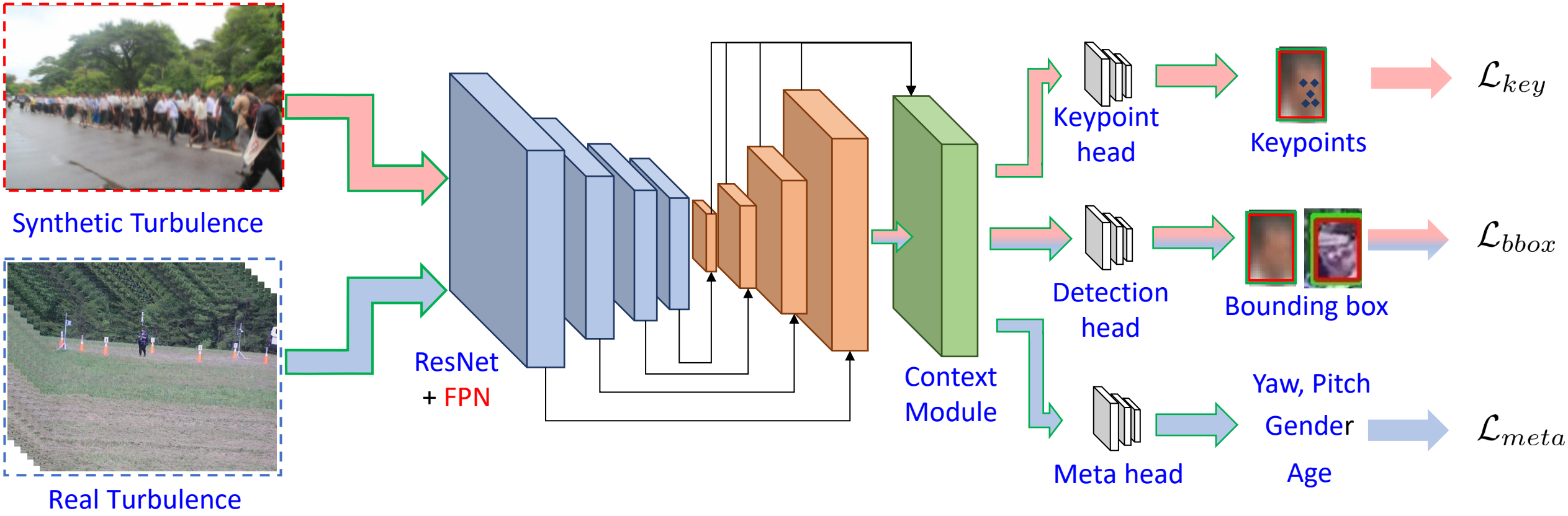


Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5203-5212).

ATDetect

- Observation: Among the three baseline models, from scratch is the best. However, we could not apply keypoints detection as there are no annotations.
- On the other hand, there are many other meta data from the video, such as headpose angles, gender and age information
- We use both the synthetic dataset WIDER-AT and the real unconstrained data to train the model
- An AT Face Detection dataset, WIDER-AT, is constructed from WIDER FACE. A total 16 combinations of blur and deformation are applied to the images
- This can improve the model generalizing to multiple objects
- Moreover, keypoint annotations exist in the WIDER-AT dataset
- To get a rich representation, additional heads for different downstream tasks, such as headpose estimation, gender classification and age prediction, are added in the model

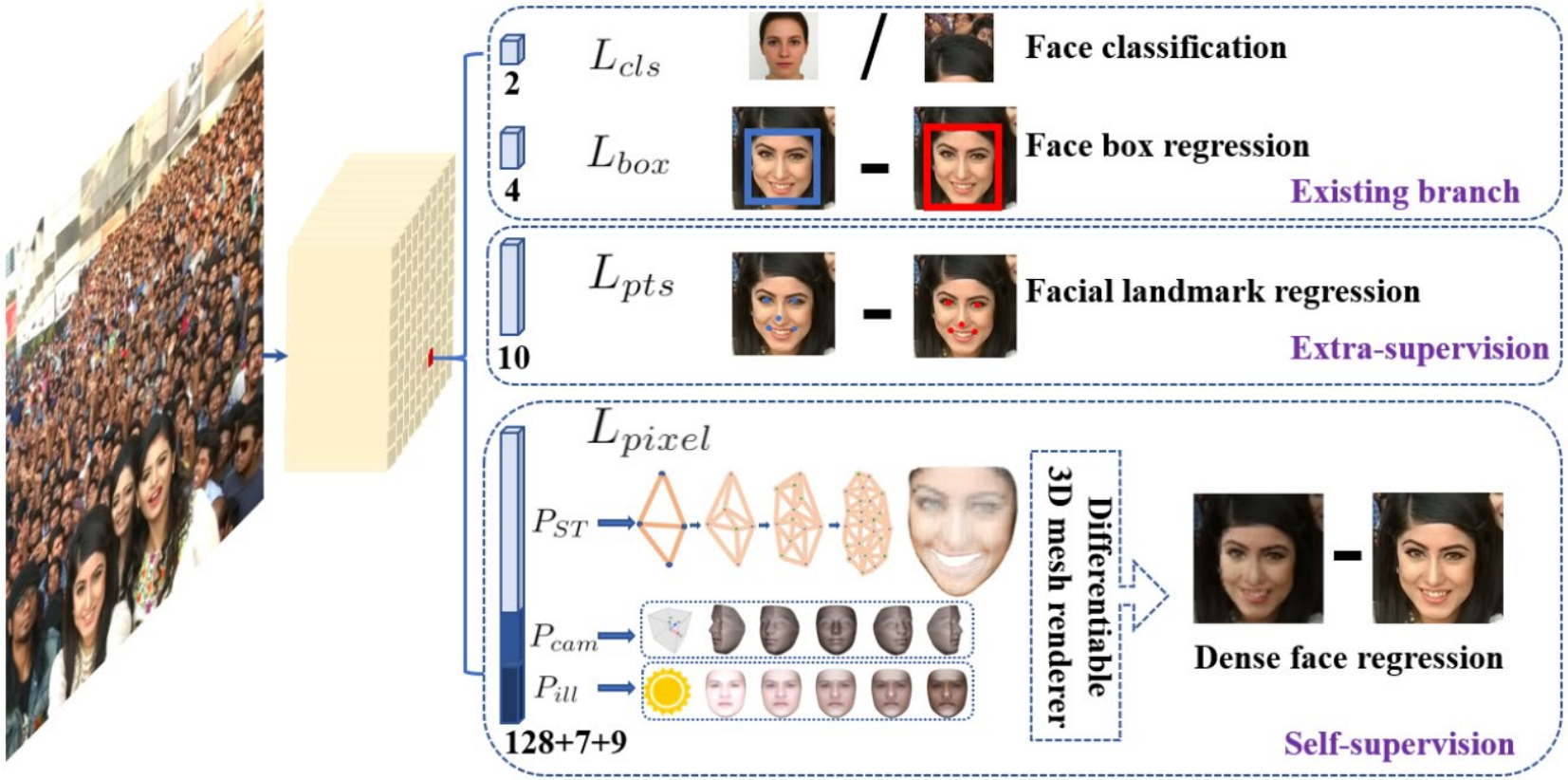
ATDetect Pipeline



Lau, Chun Pong, Maitreya Suin, and Rama Chellappa. "ATDetect: Face detection and keypoint extraction at range and altitude." *IJBC 2023*

ATDetect

- A RetinaFace with ResNet50 backbone is trained with the WIDER-AT dataset
- A baseline model is also trained with the original WIDER dataset



Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5203-5212).

Face Detection

100m

200m

Subjects consented
to publication

Face Detection

400m

500m

Subjects consented
to publication

Face Detection

Close Range

UAV

Subjects consented
to publication

Average Precision (AP)

- If the ratio of the intersection of a detected region with an annotated face region is greater than 0.5, a score of 1 is assigned to the detected region, and 0 otherwise.

	100m	270m	370m	490m
ATDetect	90.11	83.39	91.65	84.63

	600m	1000m	Close range (face)	UAV	Average
ATDetect	81.68	76.42	96.1	56.57	79.97

Average Precision (AP)

- If the ratio of the intersection of a detected region with an annotated face region is greater than 0.5, a score of 1 is assigned to the detected region, and 0 otherwise.
- ATDetect now operates directly on body chips. Significantly more efficient than operating on entire image.

	100m	200m (face)	300m	400m (face)	500m (face)	Close range (face)	Moog	Average
ATDetect	85.56	83.60	74.93	74.70	79.39	82.12	66.81	82.22

Face recognition and Identification at Altitude and Range

- A dual-path approach evaluated at 100m -1000m on a remote data set. Performed under IARPA JANUS transition effort 2019-2020.

- ATFaceGAN
  - C.P.Lau, C. D. Castillo, and R. Chellappa, “ATFaceGAN: Single Face Semantic Aware Image Restoration and Recognition from Atmospheric Turbulence”, IEEE Trans. on Biometrics, Behaviors and Identity Science, vol. 3, pp. 240-251, April 2021.

- More results from my group can be found in
  - M. Suin, N. G. Nair, C.P. Lau, V.M. Patel and R. Chellappa, "Diffuse and Restore: A Region-Adaptive Diffusion Model for Identity-Preserving Blind Face Restoration", Proc. Winter Conference on Applications on Computer Vision, Kona, Hawaii, Jan. 2024.
  - M. Suin and R. Chellappa, “CLR-Face: Conditional Latent Refinement for Blind Face Restoration Using Score-Based Diffusion Models”, Proc. Intl. Jt. Conf. on Artificial Intelligence, Jeju, Korea, August 2024.

- Prof. Vishal Patel’s group is doing amazing work on this problem for the BRIAR project.

Problem Formulation

$$\tilde{I} = D(H(I)) + n,$$

- A more challenging and practical setting: one frame is available to reconstruct the latent clean image
- Build a restoration function *G* to restore the distorted face image, i.e. $G(\tilde{I}) = I$
- The **Wasserstein GAN with gradient penalty** is employed.
- Denote blurry image and deformed image as I_b and I_d respectively. Build a deblur function G_d and a deformation correction G_b to remove undesired blur and deformation, i.e. $G_d(\tilde{I}) = I_d$, $G_b(\tilde{I}) = I_b$ and $G(\tilde{I}) = G_d(G_b(\tilde{I})) = I$.
- Then the **turbulence is decomposed into blur and deformation**.
- The "mixing" of deformation and blur in realistic turbulence face images is very fast and we could not be sure whether deformation precedes blur or blur precedes deformation.
- **Commutative constraint** is enforced, i.e. $D(H(I)) = H(D(I)) = \tilde{I}$.

Data Augmentation

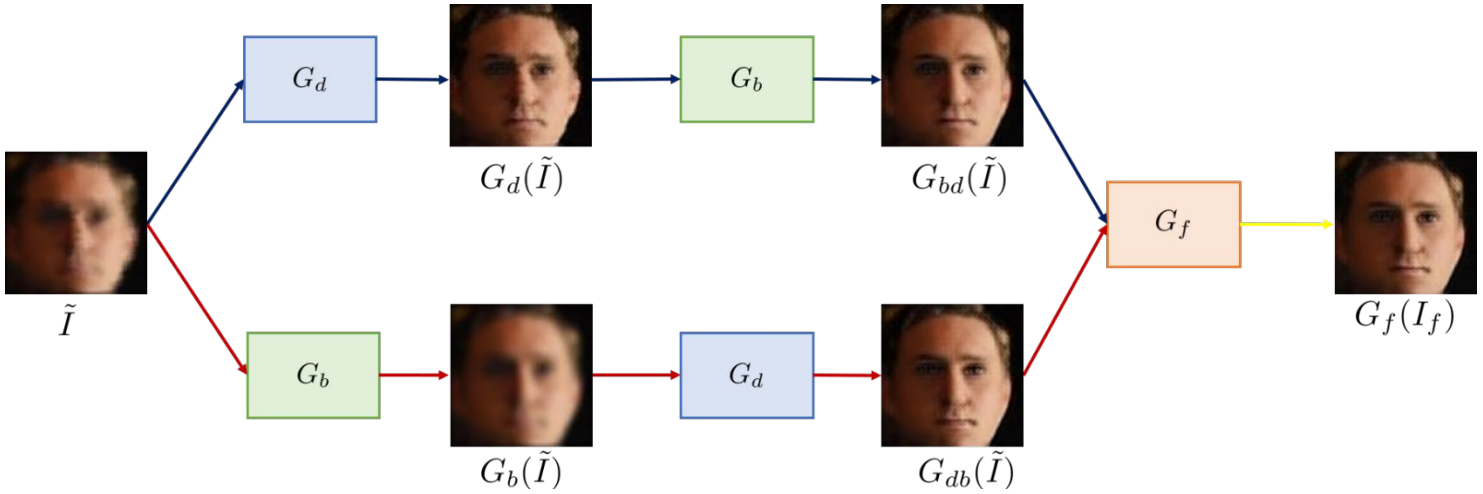
- 10000 aligned face images are picked from UMDFaces
- Use Gaussian blurring kernel with different variance as H
- Construct D as follows:

M points are selected in a face image I. For each point (x, y), a $N \times N$ patch $P^N_{x,y}$ centered at (x, y) is considered. A random motion vector field $V_{x,y}$ is obtained in $P^N_{x,y}$. Mathematically,

$$V_{x,y} = \eta(G_\sigma * \mathcal{N}_1, G_\sigma * \mathcal{N}_2),$$

where G_σ is the Gaussian kernel with standard deviation σ, η is the strength value, $\mathcal{N}_1$ and $\mathcal{N}_2$ are randomly selected from a Gaussian distribution. The overall motion vector field is generated after M iterations as follows, $V = \sum_{i=1}^{M} V_{(x,y)_i}$. Then this motion vector field would be our deformation operator D as $D(I) = I \boxplus V$, where $\boxplus$ is the warping operator.

Proposed Method



- Using Wasserstein GAN with gradient penalty
- Split the turbulence degradation due to blur and deformation in the training stage
  Introduce deblur function G_d and deformation correction function G_b.
- Enforce novel constraint: commutative constraint:
  Denote $G_{bd} = G_b \circ G_d$ and $G_{db} = G_d \circ G_b$. Mathematically, $G(\tilde{I}) = G_f(I_f)$, where G_f is a image fusion function and I_f is the pixelwise average of the restored image pair $(G_{bd}(\tilde{I}), G_{db}(\tilde{I}))$.

Loss Functions

- Content Loss: $\mathcal{L}_{con} = \|G_b(\tilde{I}) - I_b\|_2^2 + \|G_d(\tilde{I}) - I_d\|_2^2$
- Commutative Constraint: $\mathcal{L}_{cc} = \|G_{db}(\tilde{I}) - I)\|_2^2 + \|G_{bd}(\tilde{I}) - I)\|_2^2$
- Fusion Loss: $\mathcal{L}_f = \|G_f(I_f) - I\|_2^2$
- Adversarial Loss:

$$\mathcal{L}_{Dis}^{\mathcal{I}_i} = \mathbb{E}_{\tilde{I}\sim\tilde{\mathcal{I}}}[D_i(G_i(\tilde{I}))] - \mathbb{E}_{I_i\sim\mathcal{I}_i}[D_i(I_i))] + \lambda_{WGAN}\cdot\mathbb{E}_{\hat{I}_i\sim\widehat{\mathcal{I}}_i}[(\|\nabla_{\hat{I}_i} D_i(\hat{I})\|_2 - 1)^2],$$

$$\mathcal{L}_{Dis}^{\mathcal{I}_j} = \mathbb{E}_{I_j\sim\mathcal{I}_j}[D_f(G_j(I_j))] - \mathbb{E}_{I\sim\mathcal{I}}[D_f(I))] + \lambda_{WGAN}\cdot\mathbb{E}_{\hat{I}_j\sim\widehat{\mathcal{I}_j}}[(\|\nabla_{\hat{I}} D_f(\hat{I})\|_2 - 1)^2],$$

$\mathcal{L}_{Gen}^{\mathcal{I}_k} = -\mathbb{E}_{I_k\sim\mathcal{I}_k}[D_k(G_k(I_k))]$, where $\widehat{\mathcal{I}}_i$ is the distribution obtained by randomly interpolating between real images I_i and restored images $G_i(\tilde{I})$, $i \in \{b, d\}$, $j \in \{bd, db, f\}$ and $k \in \{b, d, bd, db, f\}$. For convenience of notation, $I_{bd} = I_{db} = \tilde{I}$, $\mathcal{I}_{bd} = \mathcal{I}_{db} = \tilde{\mathcal{I}}$ and $D_{bd} = D_{db} = D_f$.

- Perceptual Loss:
$\mathcal{L}_p^{\mathcal{I}_i} = \|\phi_l(G_i(\tilde{I})) - \phi_l(I_i)\|_2^2, \quad i \in \{b, d\}, \quad \mathcal{L}_p^{\mathcal{I}_j} = \|\phi_l(G_j(I_j)) - \phi_l(I)\|_2^2, \quad j \in \{bd, db, f\}$, where $\phi_l(\cdot)$ is the features of the l^{th} layer of a pretrained CNN.
- Full Loss Function: $\mathcal{L} = \mathcal{L}_{adv} + \lambda_{con}\mathcal{L}_{con} + \lambda_{cc}\mathcal{L}_{cc} + \lambda_f\mathcal{L}_f + \lambda_p\mathcal{L}_p$

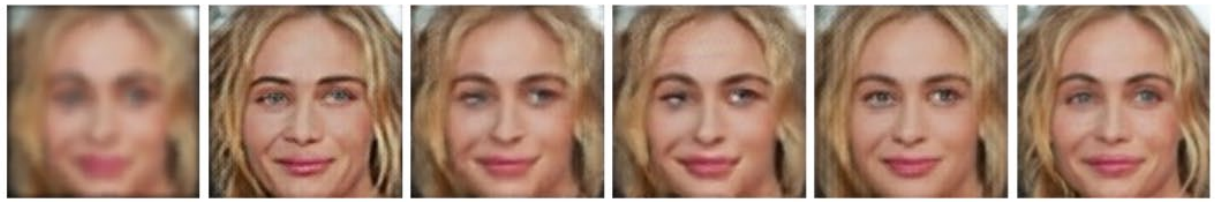
Results

- The first row is the synthetic atmospheric turbulence degraded images. The second row is the corresponding restored images. The third row is the latent groun dtruth images

Ablation Study

Table 1: Ablation study tested with LFW dataset

Method	One generator	Decompose into two generators	Add commutative constraint	Add Perceptual loss
PSNR	25.09	25.21	25.50	**26.53**
SSIM	0.882	0.878	0.886	**0.908**



(a) (b) (c) (d) (e) (f)

Figure 3: Ablation study. (a) is the distorted image and (f) is the sharp image. (b) only contains one generator. (c) is split into G_d and G_b. (d) adds the commutative constraints and (f) adds perceptual loss.

Qualitative and Quantitative Evaluation

(a) (b) (c) (d) (e)

Figure 4: Visual performance comparison with state-of-the-art methods. (a) is the distorted image. (b) Kupyn et al. [17]. (c) Shen et al. [32]. (d) Ours. (e) Groundtruth.

Table 2: Quantitative performance comparison with state-of-the-art methods on LFW dataset

Distorted		Kupyn et al. [17]		Shen et al. [32]		Ours	
PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
24.17	0.878	23.89	0.867	19.60	0.768	**26.53**	**0.908**

Table 3: Face verification results on the LFW dataset.

Method	Sharp	Distorted	Kupyn et al. [17]	Shen et al. [32]	Ours
Accuracy	0.998	0.726	0.783	0.647	**0.799**

Performance of the Disentangled Representation

(a) (b) (c) (d) (e)

Figure 5: Visual performance comparison of the deblur function G_d and deformation correction G_b with the LFW dataset. (a) Blurry image. (b) Restored image of (a) by G_d. (c) Deformed image. (d) Restored image of (c) by G_b. (e) Groundtruth.

Table 4: PSNR, SSIM and face verification results for LFW dataset with G_b and G_d.

	I_b	I_d	$G_d(I_b)$	$G_b(I_d)$
PSNR	25.33	29.78	**28.72**	**29.93**
SSIM	0.895	0.958	**0.931**	**0.961**
Accuracy	0.793	0.649	**0.817**	**0.809**

Gait-based Human Recognition

- Brief review of earlier work
  - DARPA HumanID (2000-2003)
  - DHS (2006-2008)
  - ARL CTA on Advanced Sensors (2001-2007)
- Drawbacks
  - Pre-deep learning
  - Lack of robustness to viewing angles, shoes, surface, clothing…
  - Orchestrated datasets (CASIA, USF, UMD)
- Proposed approaches
  - Based on deep learning
  - Fuses geometry, motion and data
  - Preliminary results on CASIA-B and BRIAR data
  - CASIA-B data uses contours and images (done before March 15th)

HumanID-Gait

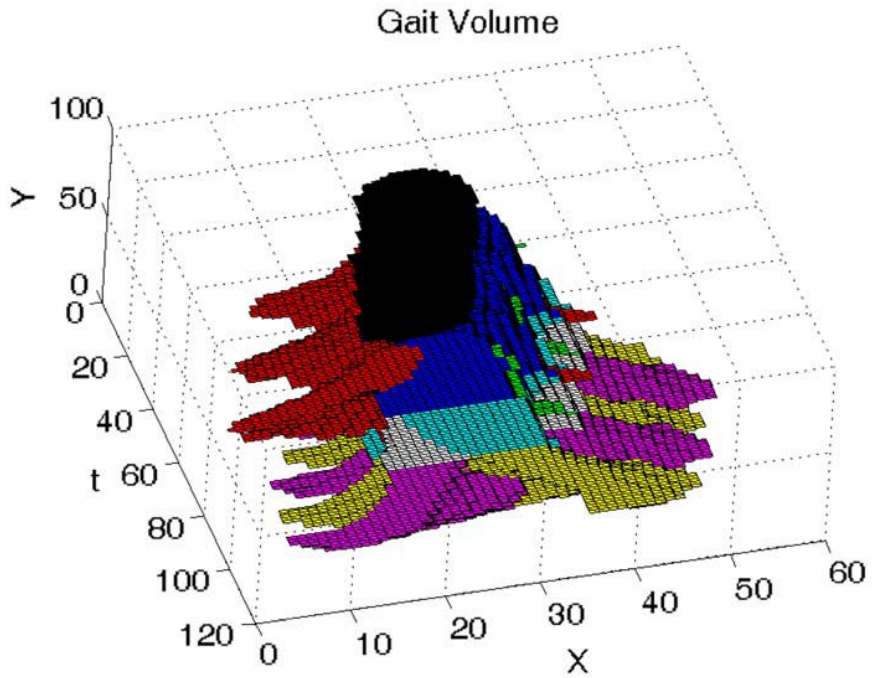
- Developed an HMM-based approach for gait-based human identification
- Best performance in the program, until improved by Sudeep Sarkar using a population HMM
- Tested on USF dataset (122 subjects), CMU MOBO dataset.
- View-robustness using a cardboard model for humans
- Several papers and a book (Pre-JANUS)
  - H. Moon, R. Chellappa, and A. Rosenfeld, "3D Object Tracking Using Shape-Encoded Particle Propagation", International Conference on Computer Vision, Vancouver, Canada, pp. II:307-314, 2001.
  - N. Cuntoor, A. Kale and R. Chellappa, "Combining Multiple Evidences for Gait Recognition," Proc. International Conf. on Acoustics, Speech and Signal Processing, Hong Kong, Vol. 3, pp. 33-36, April 2003.
  - T. Yamamoto and R. Chellappa, "Shape and Motion Driven Particle Filtering for Human Body Tracking", Proc. Intl. Conf. on Multimedia and Expo, Baltimore, MD Vol. 3, pp. 61-64, July 2003.
  - A. Sundaresan, A. Roy Chowdhury, and R. Chellappa, "A Hidden Markov Model Based Framework for Recognition of Humans from Gait Sequences," Proc. Intl. Conf. on Image Processing, Vol. 2, pp. 93-96, Barcelona, Spain, Sept. 2003.
  - A. Kale, A. Roy Chowdhury, and R. Chellappa, "Towards View Invariant Gait Recognition Algorithm", Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance, Miami, FL, pp. 143-150, July 2003.
  - A. Kale, A. Roy Chowdhury, and R. Chellappa, "Fusion of Gait and Face for Human Identification", Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Processing, Montreal, Canada, May 2004.
  - Amit Kale, et al, "Identification of Humans Using Gait", IEEE Trans. Image Processing, vol. 13, pp. 1163-1173, Sept. 2004.
  - *Human Identification Based on Gait*, Springer, Mark Nixon, Tieniu Tan and Rama Chellappa, 2005. ISBN: 978-0-387-29488-9.

HumanID-Face

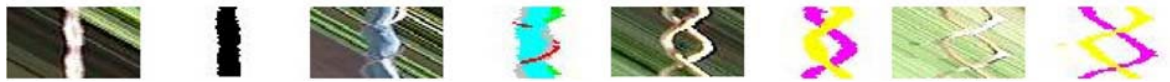
- Video-based face recognition
- NIST data collected by Patrick Grother (30 subjects walking towards a camera); CMU data of subjects walking on a treadmill
- Fusion of face and gait
- Several papers and a book (Pre-JANUS)
  - S. Zhou and R. Chellappa, "A Robust Algorithm for Probabilistic Human Recognition from Video", International Conference on Pattern Recognition, Quebec City, Canada, vol. I, pp. 226-229, 2002.
  - S. Zhou and R. Chellappa, "Probabilistic Human Recognition from Video", European Conf. on Computer Vision, Copenhagen, Denmark, pp. 681-697, May 2002.
  - S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic Recognition of Human Face from Video", Computer Vision and image Understanding, Special Issue on Face Recognition, Vol. 91, pp. 214-245, July 2003.
  - G. Aggarwal, A. Roy Chowdhury, and R. Chellappa, "A System Identification Approach for Video-Based Face Recognition", Proc. Intl. Conf. on Pattern Recognition, Cambridge, UK, August 2004.
  - S. Zhou, R. Chellappa and B. Moghaddam, "Visual Tracking and Recognition Using Appearance-Based Modeling in Particle Filters", IEEE Transactions on Image Processing, vol. 13, pp. 1491-1506, Nov. 2004.
  - *S. Zhou and R. Chellappa, Unconstrained Face Recognition,* Springer, Shaohua Zhou and Rama Chellappa, 2005. ISBN: 978-0-387-29486-5
  - *W. Zhao and R. Chellappa, Face Processing: Advanced Modeling and Methods, Elsevier, 2005, ISBN:* 9780080488844.
  - Y. C. Chen, V. M. Patel, S. Shekhar, R. Chellappa and P. J. Phillips, "Video-based Face Recognition via Joint Sparse Representation", Proc. IEEE Computer Society Conf. on Face and Gestures, Shanghai, China, April 2013.
  - W. Zou, P.C. Yuen, and R. Chellappa, "A Low-Resolution Face Tracker Robust to Illumination Variations", IEEE Trans. on Image Processing, vol. 22, pp. 1726-1739, May 2013
  - J.C. Chen, V.M. Patel, H.T. Ho and R. Chellappa, "Dictionary-based Video Face Recognition Using Dense Multi-Scale Facial Landmark Features", Proc. Intl. Conf. on Image Proc., Paris, France, Oct. 2014.
  - Y.C. Chen, V.M. Patel, P. J. Phillips and R. Chellappa, "Adaptive Representations for Video-based Face Recognition Across Pose", Proc. Winter Conference on Applications of Vision", CO, March 2014.
  - V. M. Patel, Y. C. Chen, R. Chellappa, and P. J. Phillips, "Dictionaries for Image and Video-based Face Recognition" Invited Paper , 30th Anniversary Issue, Jl. Opt. Society of America, vol. 31, pp. 1090-1103, May 2014

Gait-Pattern

(a)



(b)



(c)

Several papers

Y. Ran, Q. Zheng, R. Chellappa and T. Strat, “Applications of a Simple Characterization of Human Gait in Surveillance”, IEEE Trans. Syst., Man and Cybernetics, vol. 40, pp. 1009-1020, Nov. 2010.

H. Moon and R. Chellappa, "3D Shape-Encoded Particle Filter for Object Tracking and its Application to Human Body Tracking", EURASIP Jl. On Image and Video Processing, Jan. 2008.

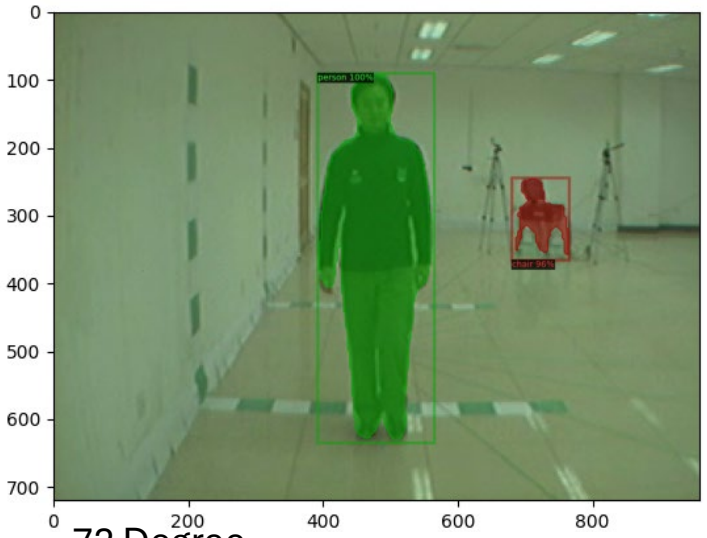
.

Proposed Approaches

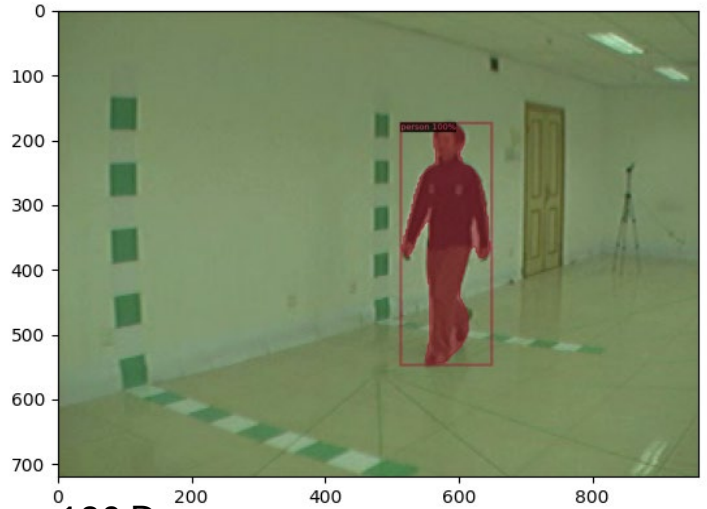
- Fine-tune pre-trained models for gait-recognition
- Based on global and component representations
- Gaitset, Gaitpart and GaitGL
- Incorporate turbulence models
- Incorporate representations from Gait-pattern
- Deeper models (DeepGait v2 , SwinGate)
- Fusion

Detectron2 Segmentation Masks

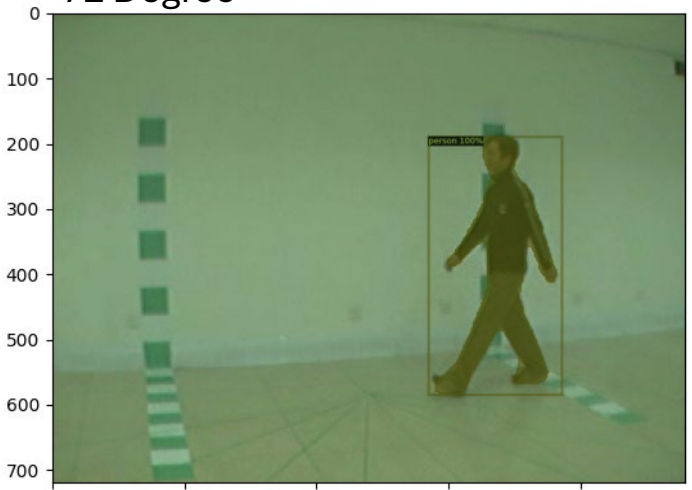
0 Degree



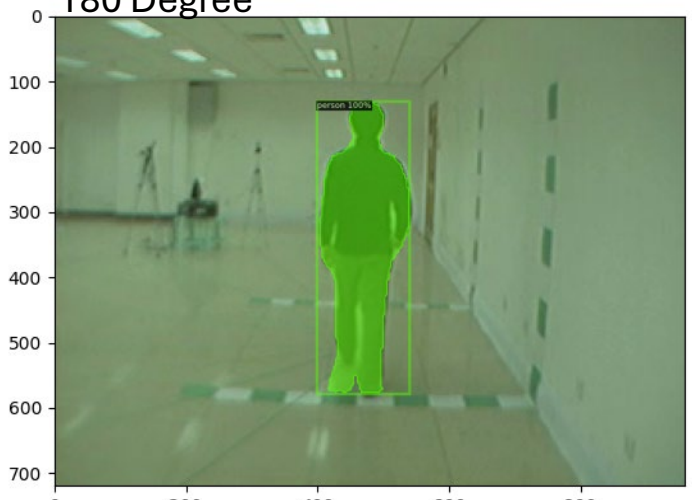
36 Degree



72 Degree



180 Degree



Ground Truth

0 Degree

36 Degree

72 Degree

180 Degree

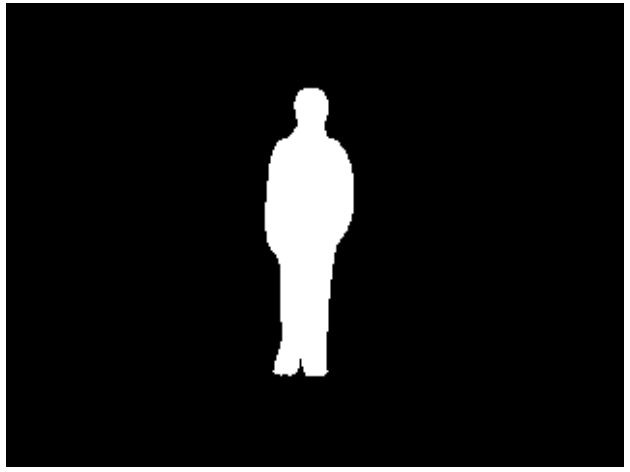
Segmentation Masks

0 Degree

36 Degree

72 Degree

180 Degree



SOTA Methods

Model	NM	BG	CL	Input Size
GaitSet(GroundTruth)	96.273(95.900)	90.667(89.825)	76.992(75.427)	64x44
GaitSet(Detectron2)	89.917(88.918)	78.633(76.824)	63.248(61.064)	64x44
GaitPart(GroundTruth)	96.479(96.127)	91.477(90.652)	80.231(78.691)	64x44
GaitPart(Detectron2)	90.124(89.136)	80.727(79.064)	67.025(61.064)	64x44
GaitGL(GroundTruth)	97.364(97.164)	94.687(94.256)	84.686(83.364)	64x44
GaitGL(Detectron2)	91.331(90.536)	86.008(84.746)	76.950(75.036)	64x44

GaitSet: Chao, Hanqing, et al. "Gaitset: Regarding gait as a set for cross-view gait recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
GaitPart: Fan, Chao, et al. "Gaitpart: Temporal part-based model for gait recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
GaitGL: Lin, Beibei, Shunli Zhang, and Xin Yu. "Gait recognition via effective global-local feature representation and local temporal aggregation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Atmospheric Turbulence Image Degradation

$$\tilde{I}_k = D_k(H_k(I)) + n_k$$

$\tilde{I}_k$ is the observed distorted images,

I is the latent clear image,

H_k is a space-invariant point spread function (PSF),

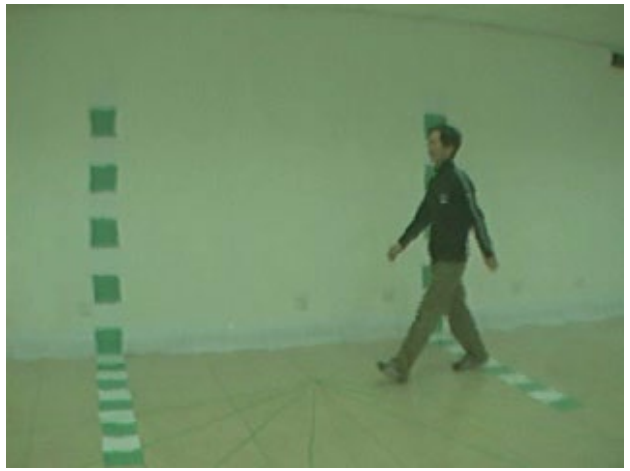
D_k is the deformation operator, which is assumed to deform randomly,

n_k is the sensor noise.

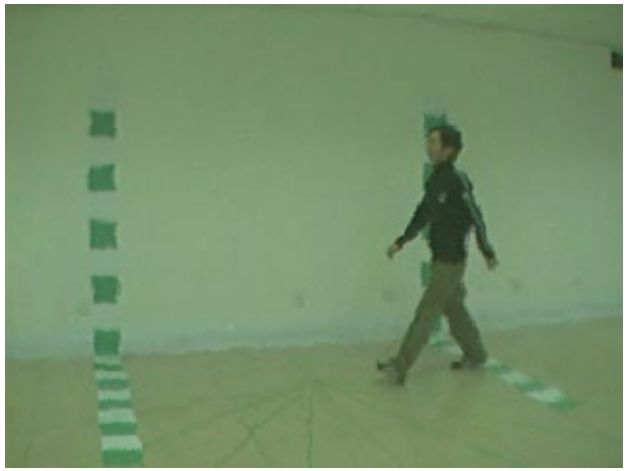
Lau, Chun Pong, Carlos D. Castillo, and Rama Chellappa. "ATFacegan: Single face semantic aware image restoration and recognition from atmospheric turbulence." *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.2 (2021): 240-251.

Image Distortion

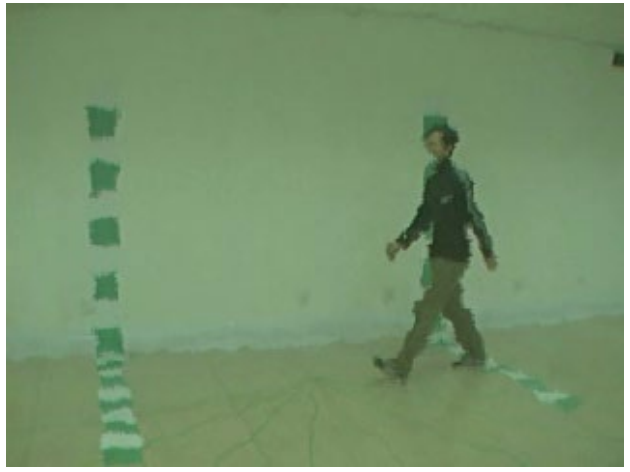
Iteration 10000 Stride 0.1



Iteration 5000 Stride 0.3



Iteration 10000 Stride 0.3



Iteration 10000 Stride 0.5

Image Distortion

Iteration 10000 Stride 0.1

Iteration 5000 Stride 0.3

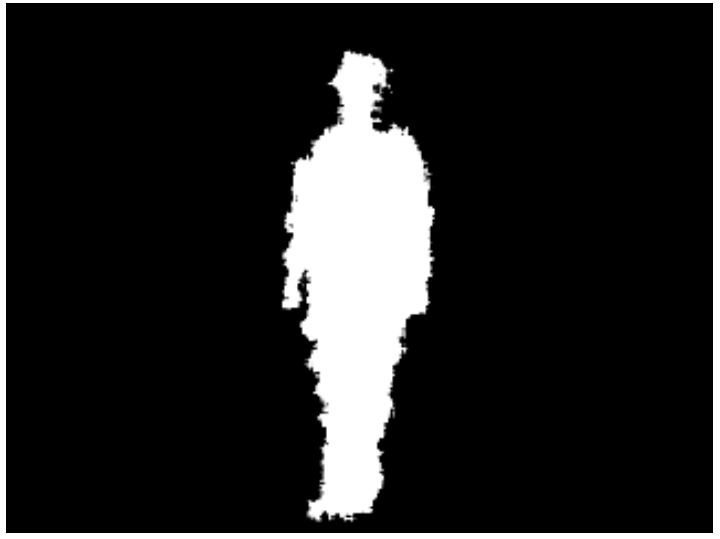
Iteration 10000 Stride 0.3

Iteration 10000 Stride 0.5

Detectron2 segmentation masks on degraded image

GroundTruth

Degraded Image

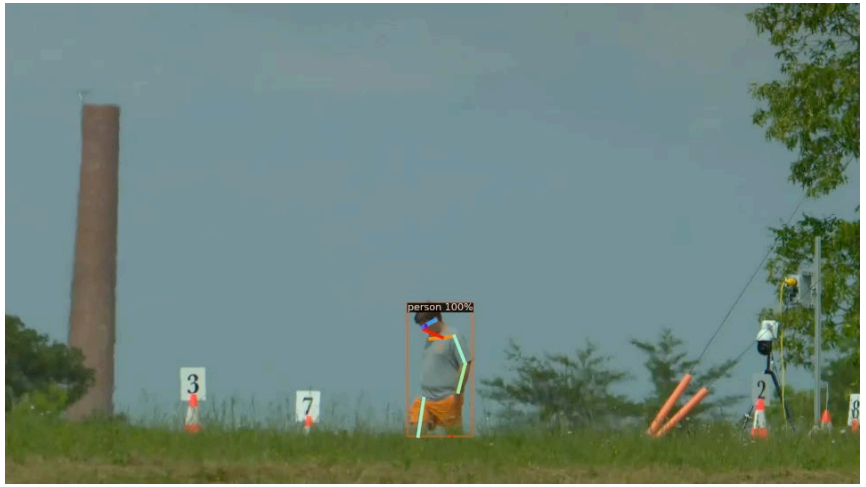


Detectron2 segmentation masks on degraded image

Gait Recognition on degraded masks

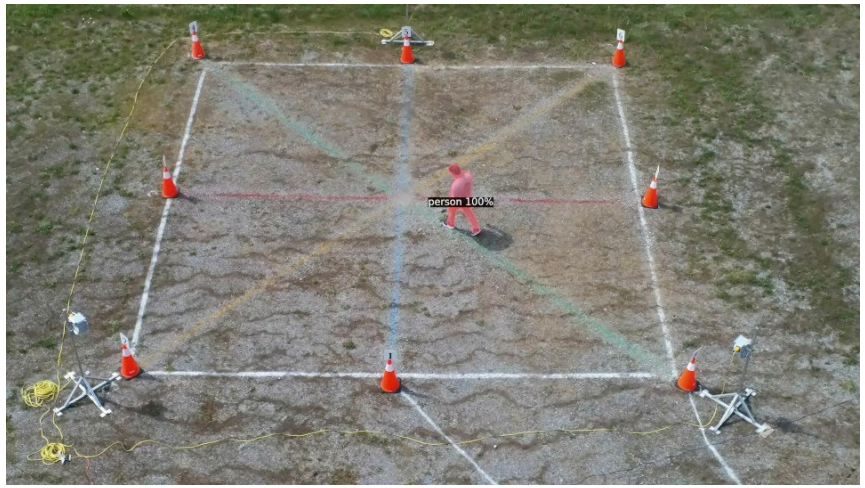
Model	NM	BG	CL	Input Size
GaitGL(GroundTruth)	97.364(97.164)	94.687(94.256)	84.686(83.364)	64x44
GaitGL(Detectron2)	91.331(90.536)	86.008(84.746)	76.950(75.036)	64x44
GaitGL(GroundTruth Deformation Stride 0.5 Iteration 10000)	85.025(83.609)	78.082(76.063)	58.752(56.064)	64x44
GaitGL(Detectron2 Deformation Stride 0.5 Iteration 10000)	84.347(82.900)	74.825(72.717)	60.810(58.209)	64x44

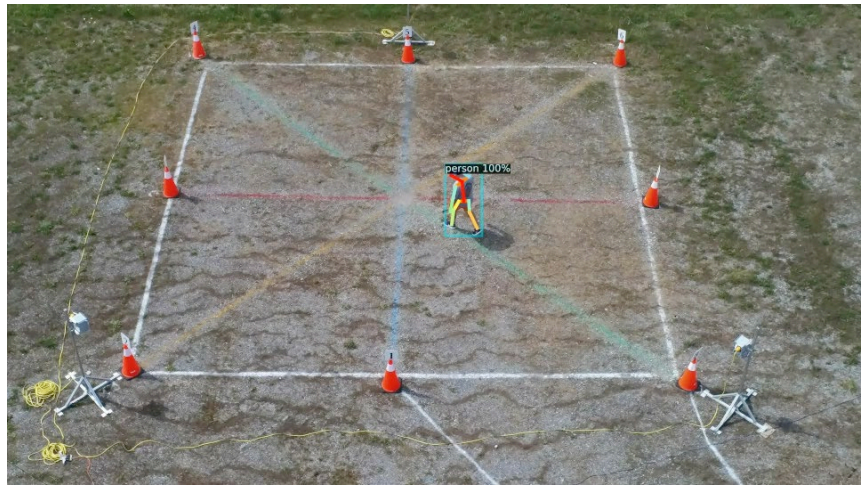
Detectron2 Instance Segmentation and Keypoint Detection at 100m and 200 m



Detectron2 Instance Segmentation and Keypoint Dtection at 400m and 500m

Detectron2 Instance Segmentation and Keypoint Detection on Close Range and UAV data





Gait recognition

- Objective
  - Biometric feature for identification of individuals based on walking patterns
- Advantages
  - Compared with other biometrics, gait can be captured from a distance
  - Robust to subject-related co-variates (clothing, disguise, etc.)
- Challenges
  - Long range
  - Mild to severe atmospheric turbulence
  - Variation in viewing angle and lighting
  - Occlusion

Gait Success Examples

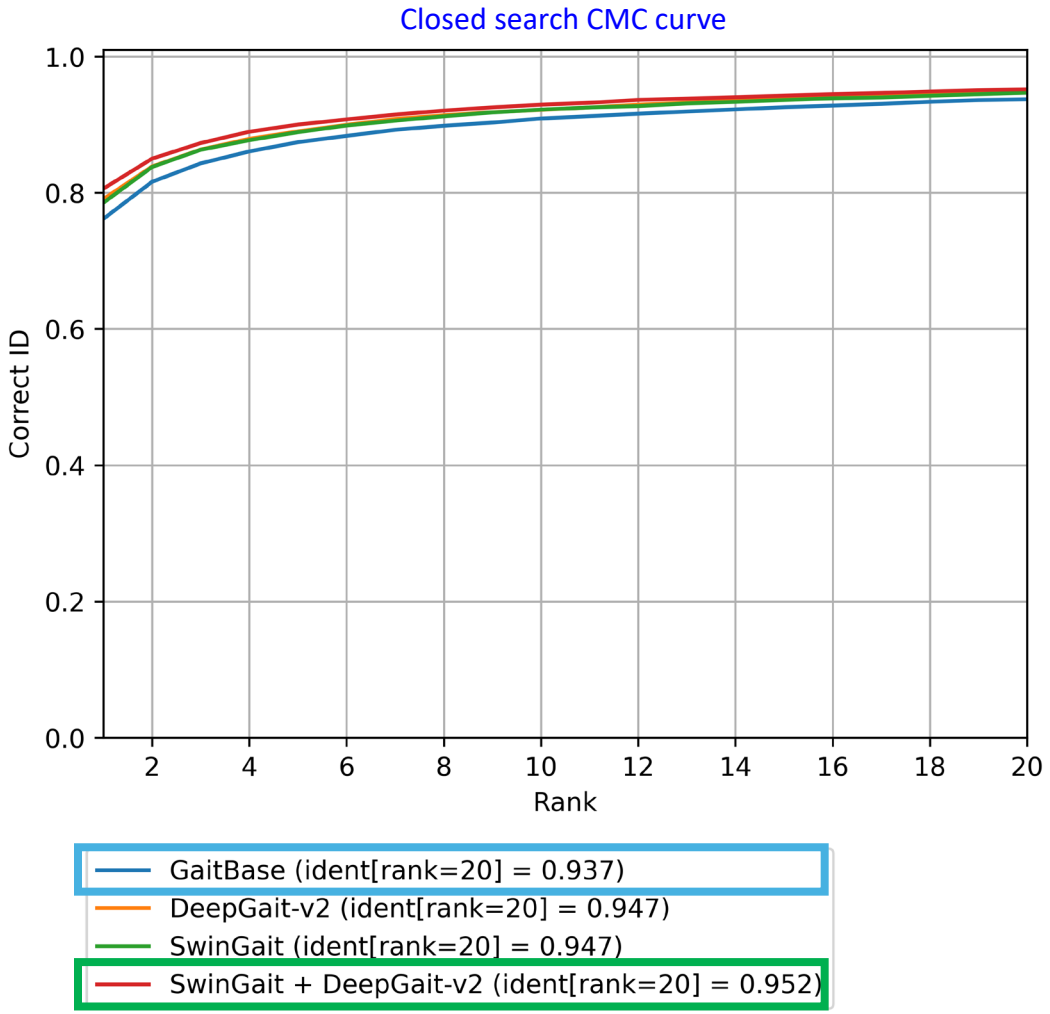
800m distance

Close range - 9m Elevation & 46 pitch angle

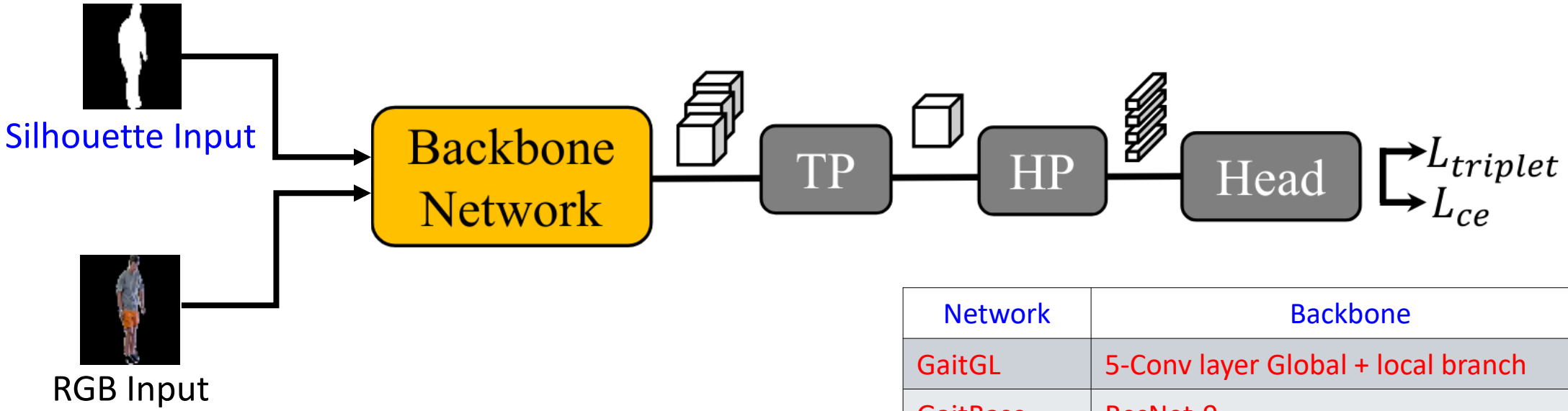
Subjects have consented for publication

Recent Progress

- Deeper models for gait
  - Early months – GaitGL
  - Nov. 2023 – GaitBase
  - May 2024 – DeepGaitv2 + SwinGait
- Varying sampling frequency
  - Accuracy vs Speed
- Transformer-based gait model
- Gait model ensemble



Gait Pipeline



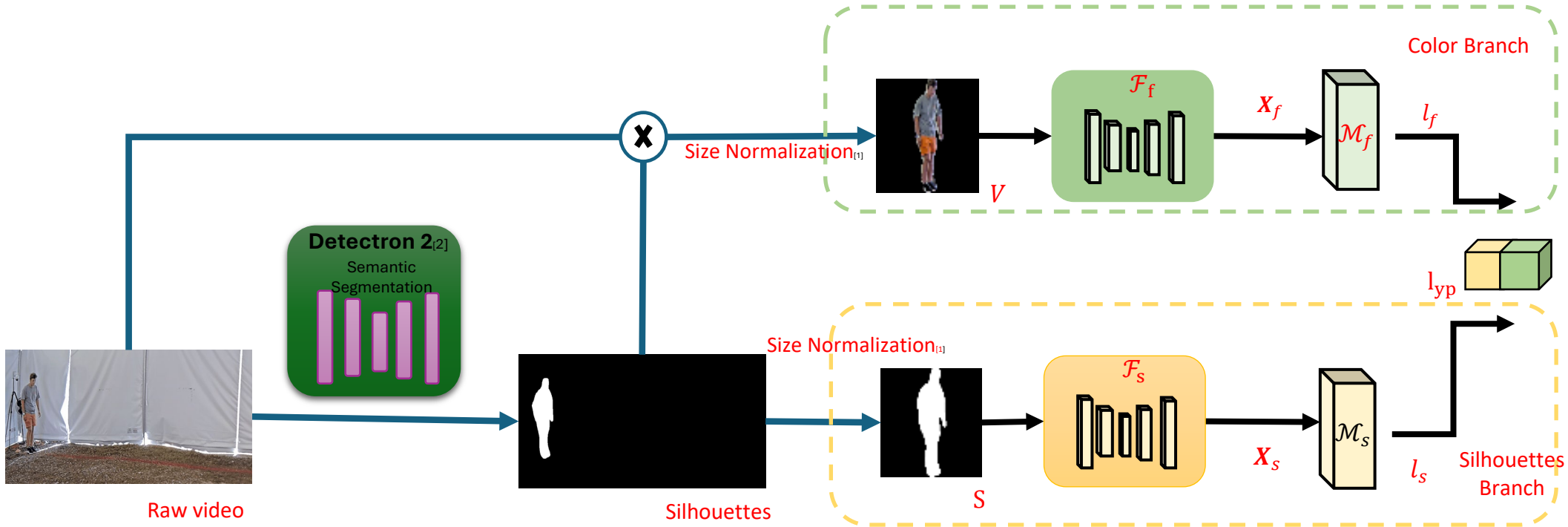
Network	Backbone
GaitGL	5-Conv layer Global + local branch
GaitBase	ResNet-9
DeepGaitv2	4-stage Residual Conv blocks
SwinGait	2-stage Residual Conv blocks + 2-stage transformer

- Temporal pooling (TP) and Horizontal pooling (HP)
- Losses: Cross Entropy + Triplet loss
- Triplet loss selection
  - Each iteration, eight people are selected. For each people, there are eight videos are collected to serve as positive and negative pairs
  - During training process, we select 100 continuous frame with random start point to train the model robustly.

https://arxiv.org/pdf/2303.03301

Early Months – GaitGL

- GaitGL: End-to-end Gait Recognition Using Silhouettes and RGB
- Backbone network: Conv layers and Global + Local branches



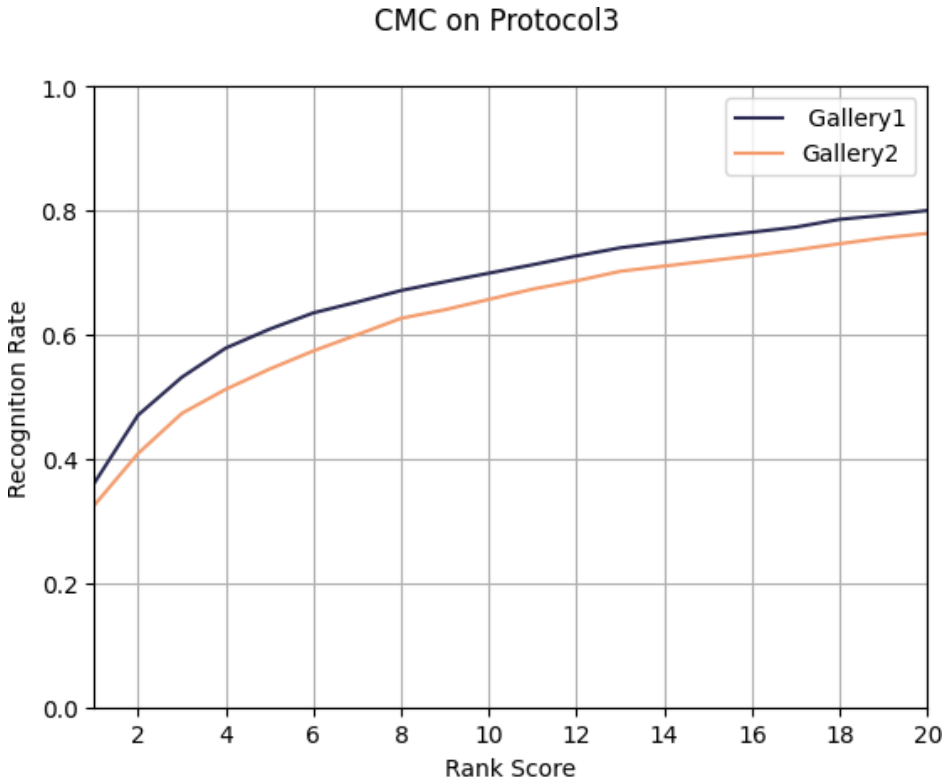
[3]

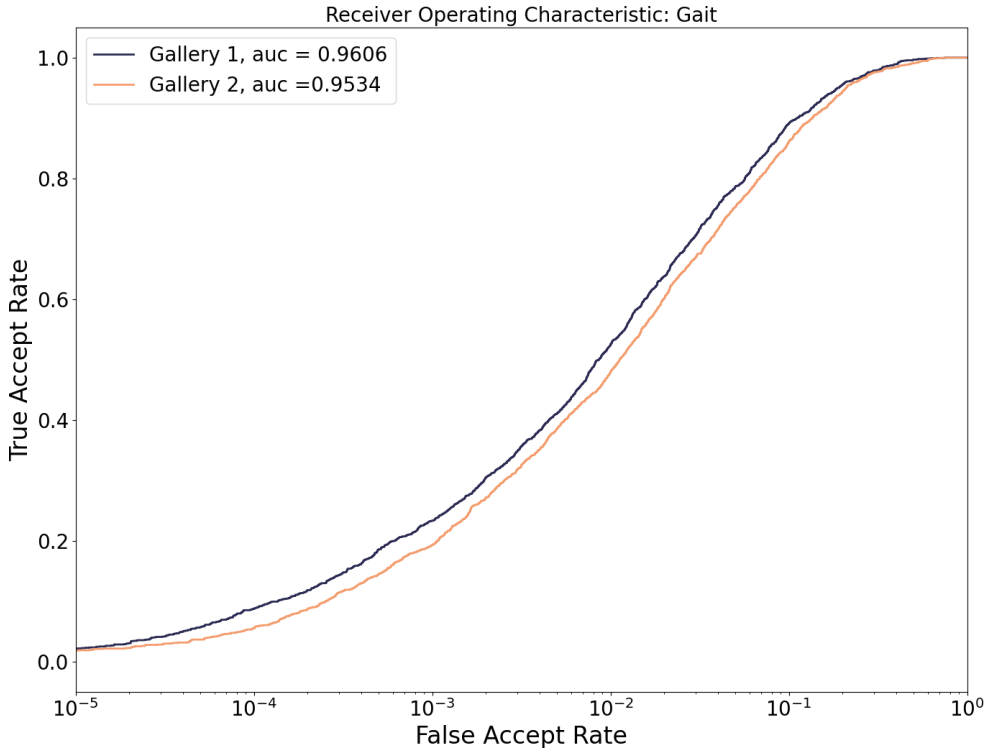
[1]Iwama, H., Okumura, M., Makihara, Y., & Yagi, Y. (2012). The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, *7*(5), 1511-1521.

[2]Detectron2： http://github.com/facebookresearch/detectron2

[3] Guo, Y., Peng, C., Lau, C. P., & Chellappa, R. (2023, January). Multi-modal human authentication using silhouettes, gait and rgb. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 1-7). IEEE.

GaitGL Evaluation





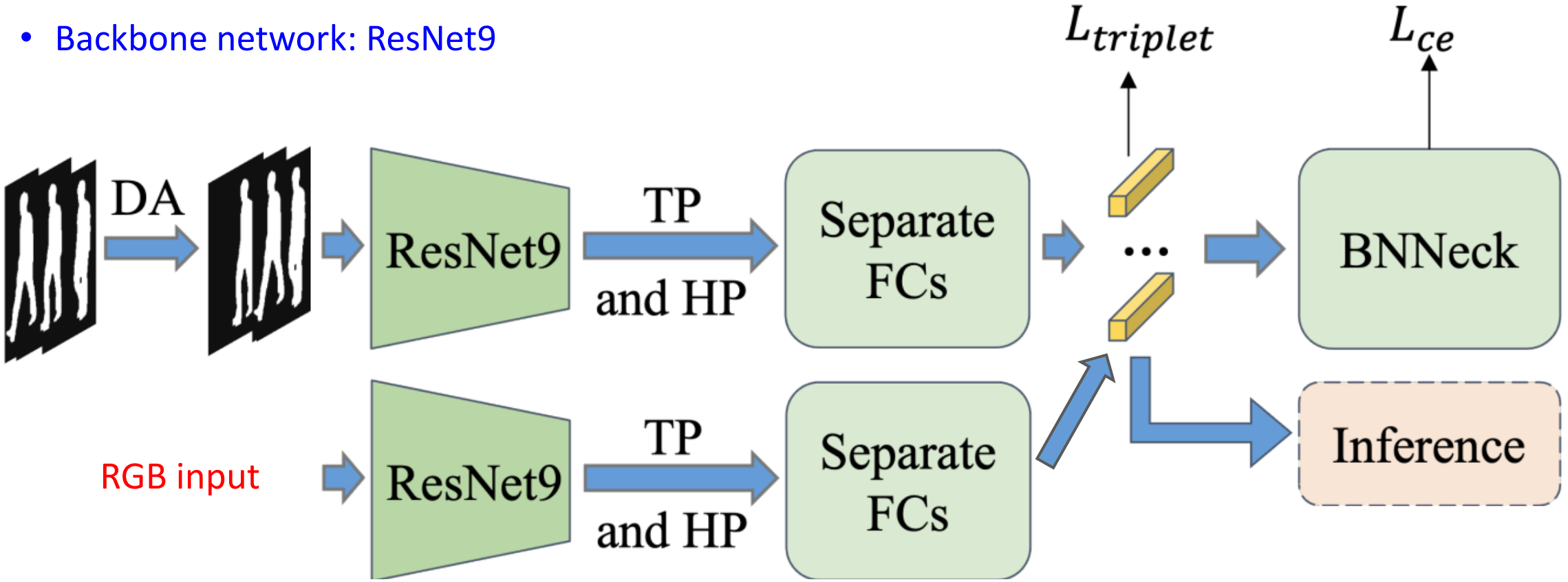
Gallery	**Rank 1 (%)**	Rank 20 (%)	TAR (%) @ FAR = 10^{-3}
1	36.06	79.98	23.35
2	32.50	76.29	19.29

Protocol 3.1 - Gait mission sigset evaluation

62% Rank-20 on full protocol 3.1 --> Needs improvement!

June 2023 – Nov. 2024 – GaitBase

- Backbone network: ResNet9



DA – Data augmentation

TP – Temporal pooling

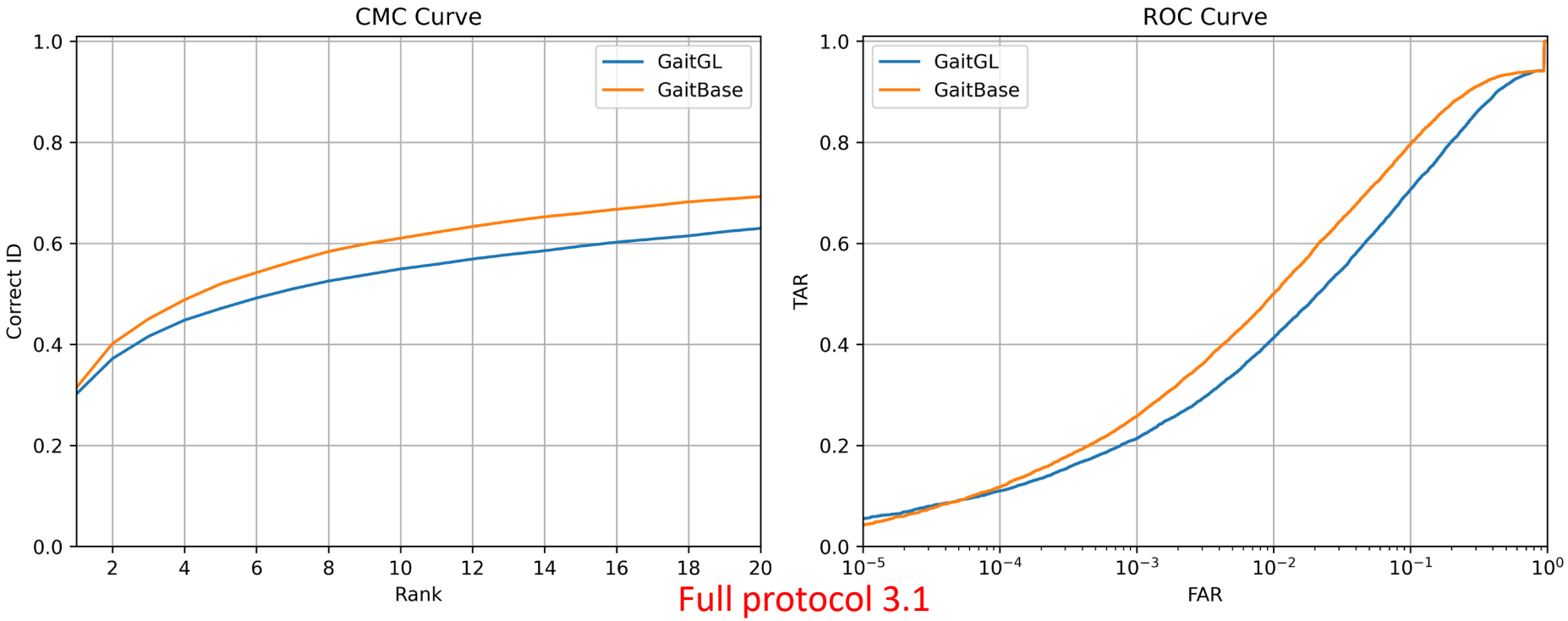
HP – Horizontal pooling

[1] Fan et al. “Opengait: Revisiting gait recognition towards better practicality”, CVPR 2023

Better Backbone

- Flexible architecture to various feature extraction backbones

GaitGL [1] → GaitBase [2]



Full protocol 3.1

Protocol 3.1– Gait Mission sigset – Gallery 1

Backbone	Rank 1 (%)	Rank 20 (%)	TAR (%) @ FAR = 10^{-3}
GaitBase	47.03	88.38	44.68
GaitGL	36.06	79.98	23.35

[1]Lin, B., Zhang, S., & Yu, X. (2021). Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 14648-14656).[2]Detectron2: http://github.com/facebookresearch/detectron2、

[2] Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., & Yu, S. (2023). OpenGait: Revisiting Gait Recognition Towards Better Practicality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9707-9716).

Smaller Template

- Feature size is relatively large comparing than other modalities.
  - Output size: 128*256 (RGB 64 + Silhouette 64)
  - Solutions: Reduce the embedding size:

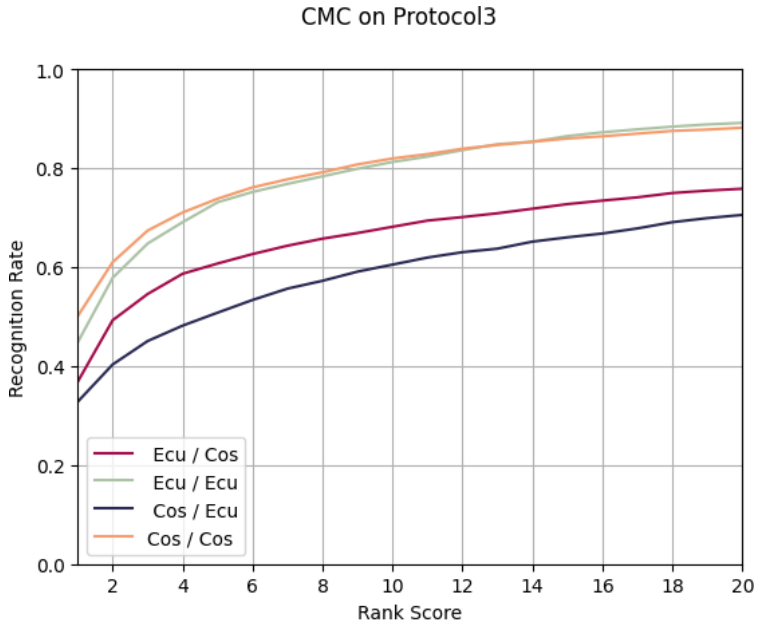
Protocol 3.1– Gait Mission (GaitBase) - Ablation study on template size

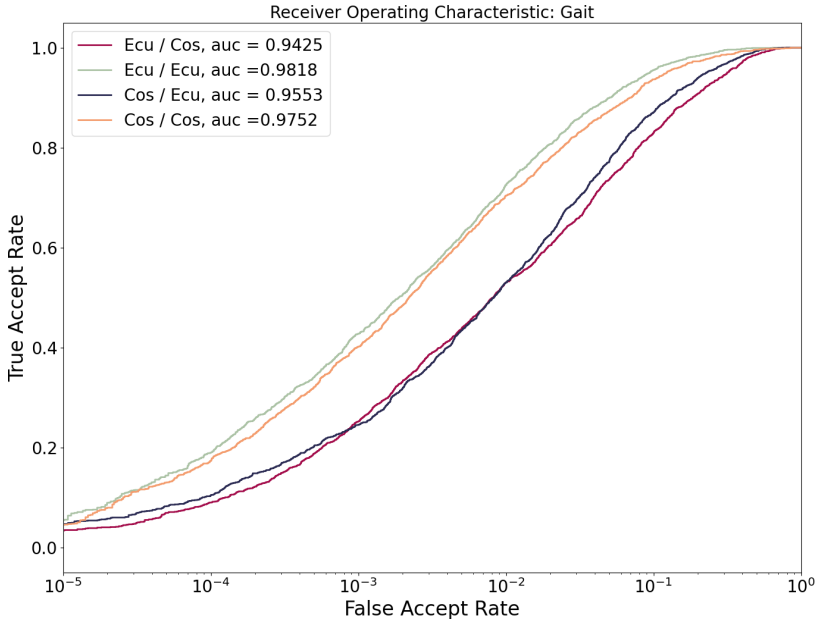
Template size	Modality	Rank 1 (%)	Rank 20 (%)	TAR (%) @ FAR = 10^{-3}
32 *256	RGB+Sil	**47.03**	88.38	**44.68**
16*256	RGB+Sil	44.79	**89.15**	42.77
	Sil	37.26	75.5	24.55
8*256	RGB+Sil	45.06	84.94	38.95
	Sil	40.37	76.98	28.37
4*256	RGB+Sil	38.13	81.51	31.70
	Sil	37.53	76.27	26.30

Euclidian to Similarity

Other modalities apply Cosine similarity rather than Euclidian distance to measure the features. To have similar distance range, we train a model using Cosine similarity.

Triplet: Euclidian → Cosine Classification: Cross Entropy → Arcface [1]





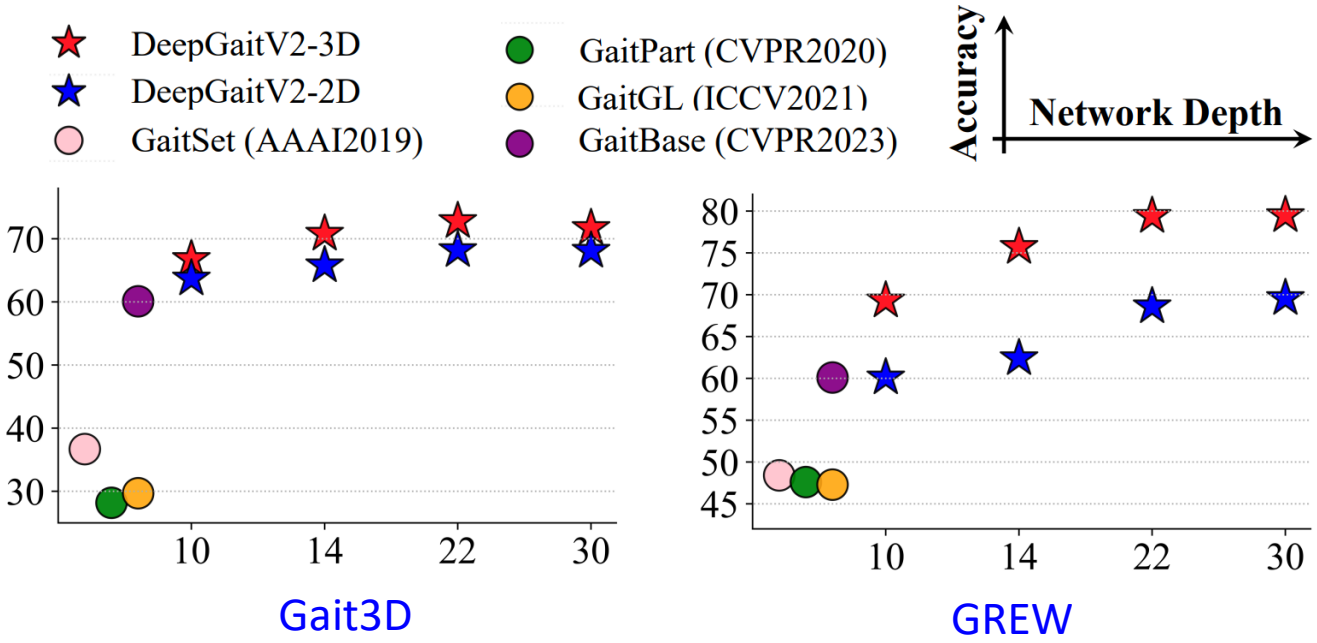
Protocol 3.1– Gait Mission (GaitBase with template 16)

Train / Test	Rank 1 (%)	Rank 20 (%)	TAR (%) @ FAR = 10^{-3}
Ecu / Ecu	44.79	**89.15**	**42.77**
Ecu / Cos	36.82	75.83	25.21
Cos / Ecu	32.79	70.54	15.82
Cos / Cos	**50.08**	88.16	40.21

[1] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).

DeepGait-v2

- Deeper models are better!
  - Shallow gait models are inefficient for diverse real-world gait scenarios
  - Challenging scenarios, such as long distance, turbulence, need deeper model to understand

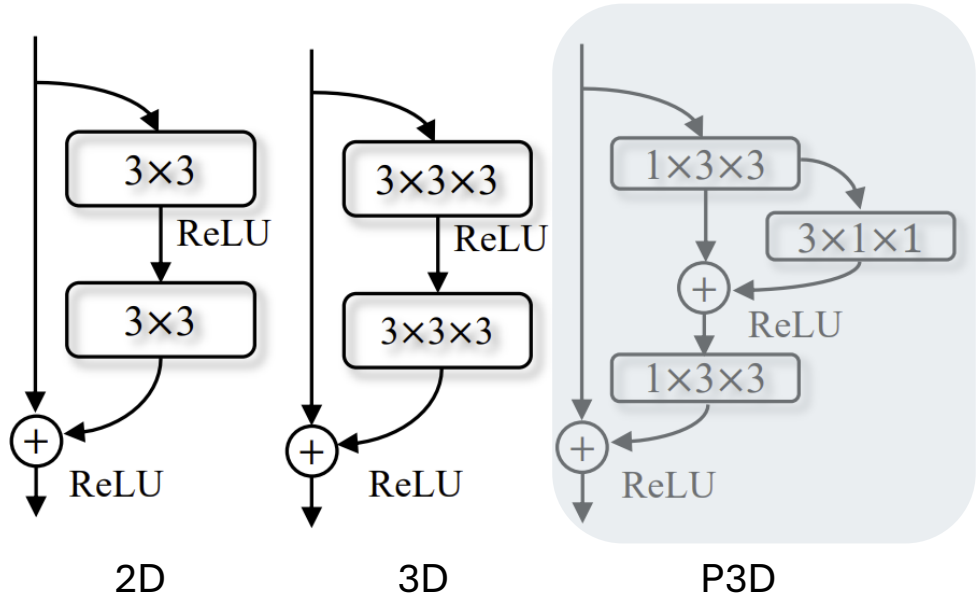


Gait3D GREW

[1] Fan et al. “Exploring Deep Models for Practical Gait Recognition”, arXiv 2024

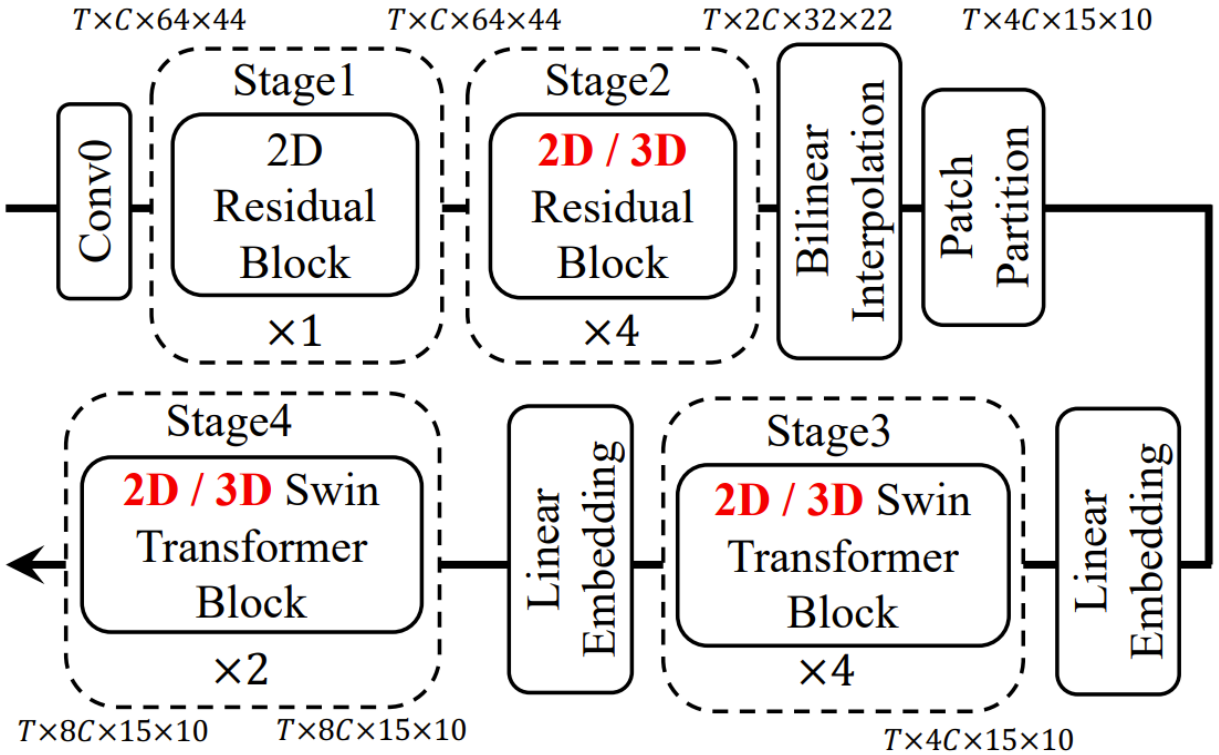
DeepGait-v2 Architecture

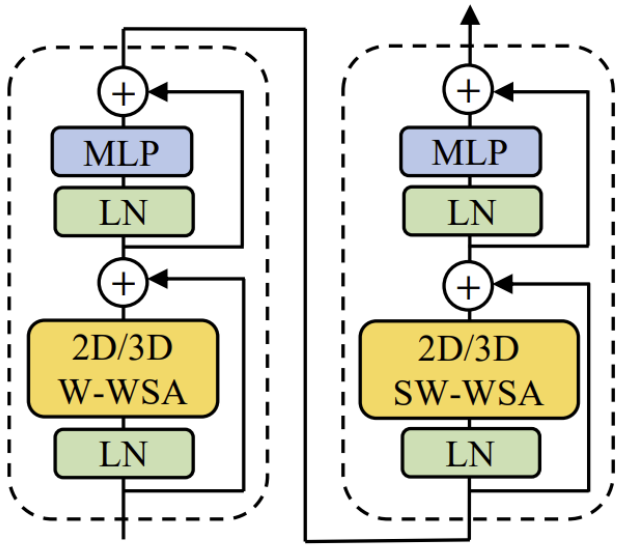
Layers	Output feature map size	DeepGaitV2-3D				
		Block structure	10-layer	14-layer	22-layer	30-layer
Conv0	$(T, C, 64, 44)$	3×3, stride 1				
Stage1	$(T, C, 64, 44)$	$\begin{bmatrix} 3 \times 3, C \\ 3 \times 3, C \end{bmatrix}$	$\times 1$	$\times 1$	$\times 1$	$\times 1$
Stage2	$(T, 2C, 32, 22)$	$\begin{bmatrix} 3 \times 3 \times 3, 2C \\ 3 \times 3 \times 3, 2C \end{bmatrix}$	$\times 1$	$\times 2$	$\times 4$	$\times 4$
Stage3	$(T, 4C, 16, 11)$	$\begin{bmatrix} 3 \times 3 \times 3, 4C \\ 3 \times 3 \times 3, 4C \end{bmatrix}$	$\times 1$	$\times 2$	$\times 4$	$\times 8$
Stage4	$(T, 8C, 16, 11)$	$\begin{bmatrix} 3 \times 3 \times 3, 8C \\ 3 \times 3 \times 3, 8C \end{bmatrix}$	$\times 1$	$\times 1$	$\times 1$	$\times 1$
TP	$(1, 8C, 16, 11)$	Temporal Pooling				
HP	$(1, 8C, 16, 1)$	Horizontal Pooling				
Head	$(1, 8C, 16, 1)$	Flatten, 16 separate fully-connected layers and BNNecks				



SwinGait Architecture

- Utilizing effectiveness of transformers to generate efficient gait descriptors
- Divide input into non-overlapping patches as *Tokens*
- Self-attention on linear embedding of tokens

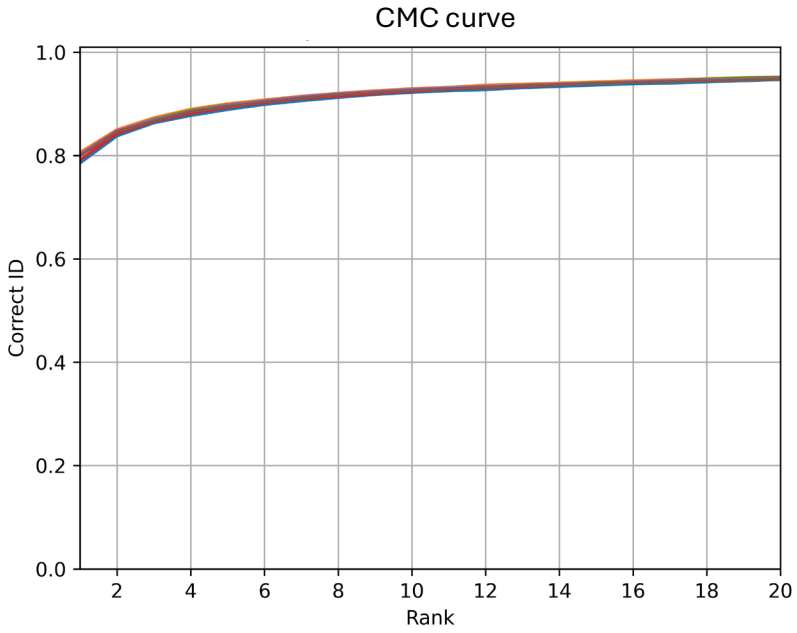


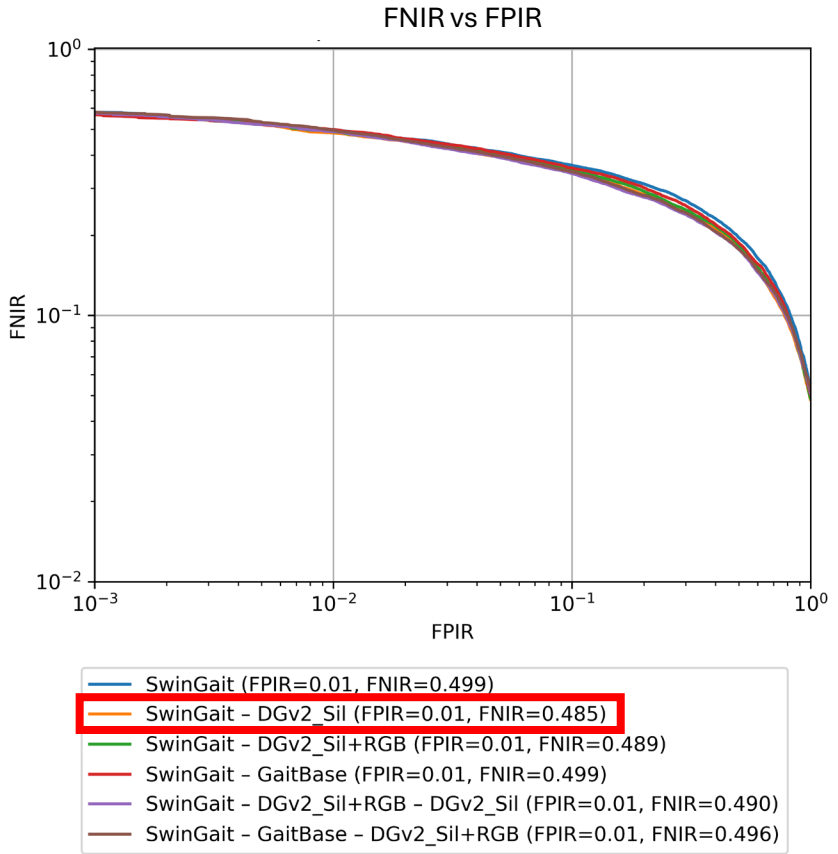


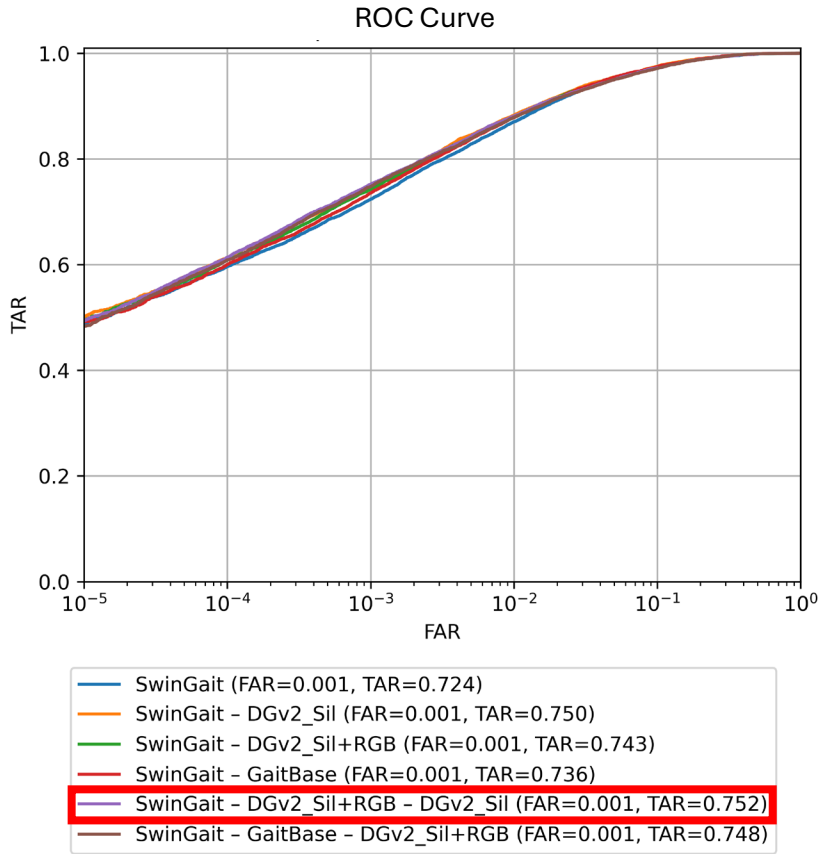
- T – number of frames
- C – number of channels

Gait Ensemble Analysis

- Gait models ensemble analysis on full protocol 4.2.1. **Plot shows final fused accuracy with face and body features.**

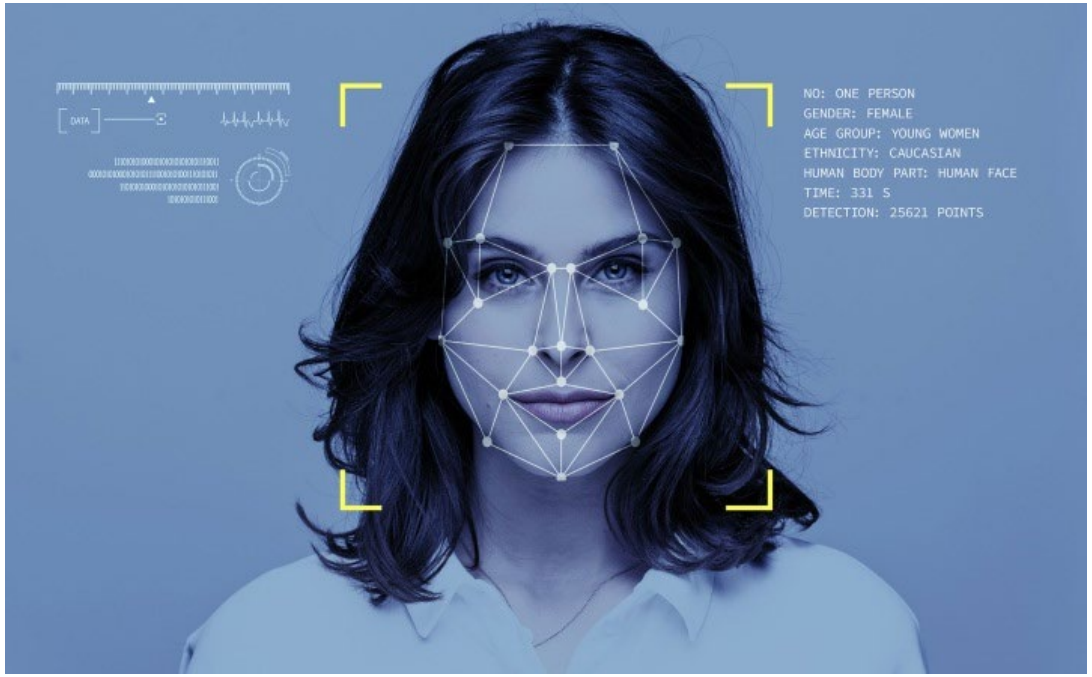






SwinGait + DeepGaitv2 > SwinGait or DeepGaitv2

How Safe is Your Facial Privacy?

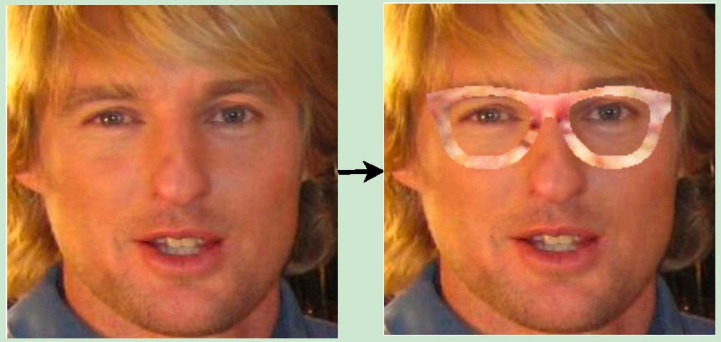


Visual Quality vs. Performance

(a) Noise-based: TIP-IM

Visual Quality ↓
Attack Performance ↑

(b) Patch-based: AdvGlasses

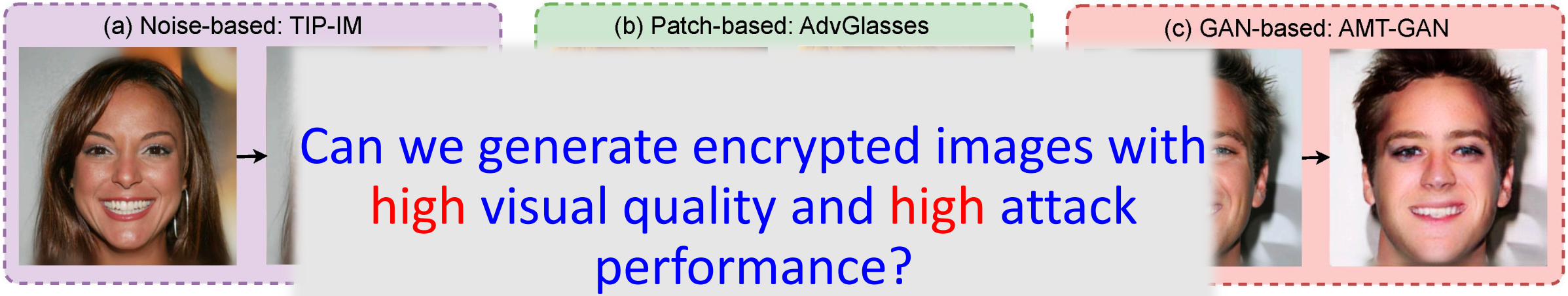


Visual Quality ↓
Attack Performance ↑

(c) GAN-based: AMT-GAN

Visual Quality ↑
Attack Performance ↓

Visual Quality vs. Performance



Can we generate encrypted images with high visual quality and high attack performance?

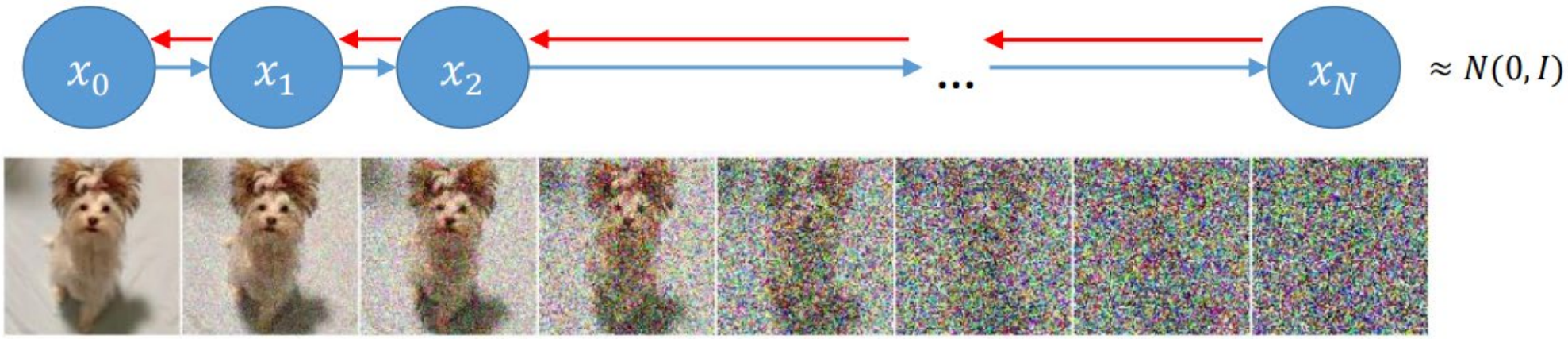
Visual Quality ↓
Attack Performance ↑

Visual Quality ↓
Attack Performance ↑

Visual Quality ↑
Attack Performance ↓

Denoising Diffusion Model (DDM)

- State-of-the-art image generation models

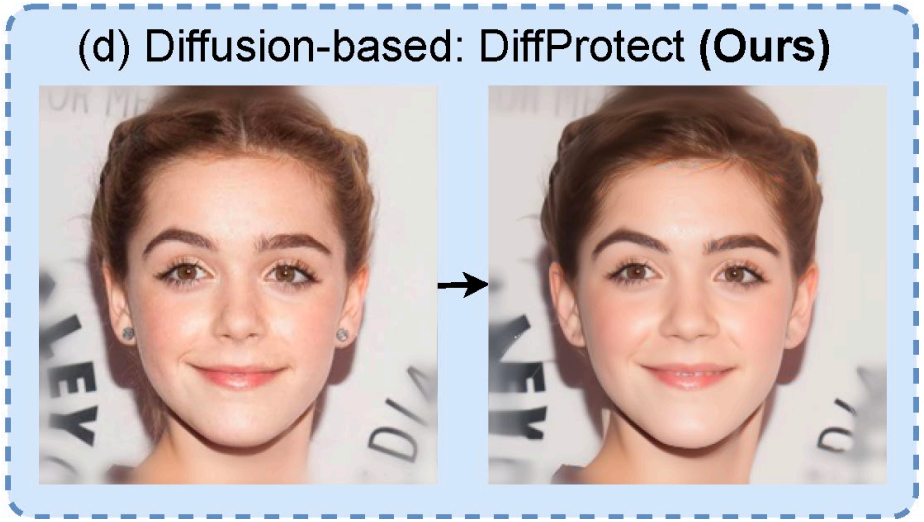


Diffusion process: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

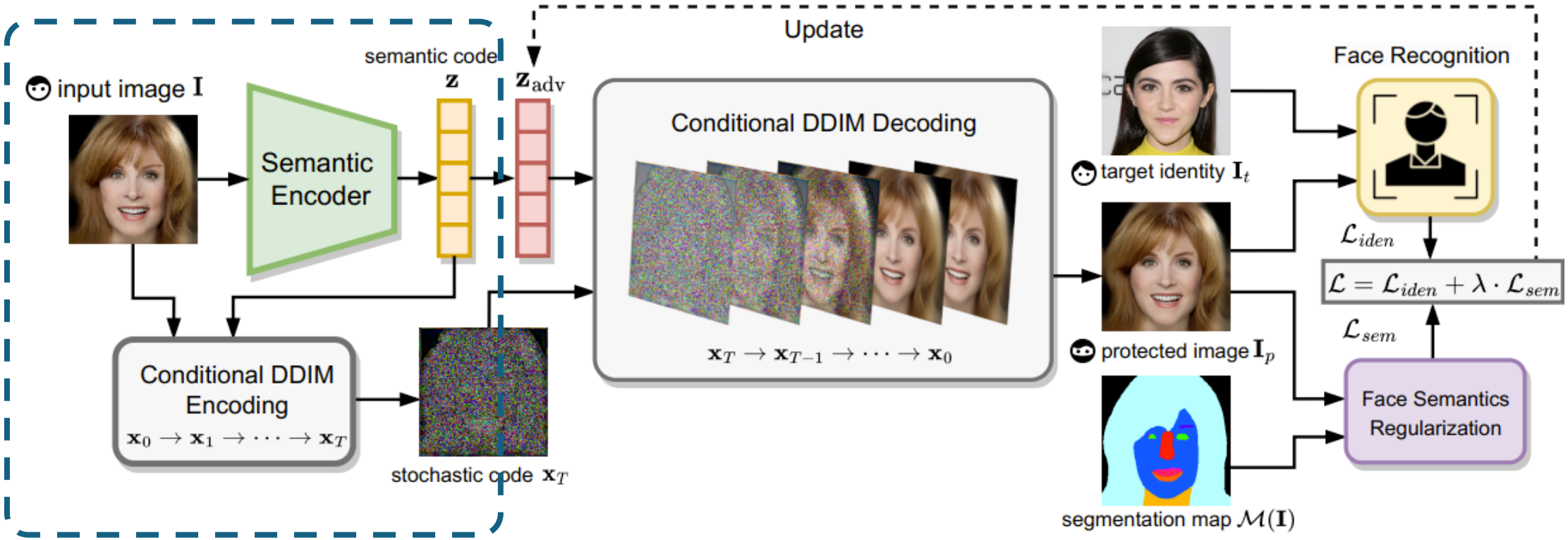
$= q(x_0|x_1) \dots q(x_{N-1}|x_N)q(x_N)$

DiffProtect

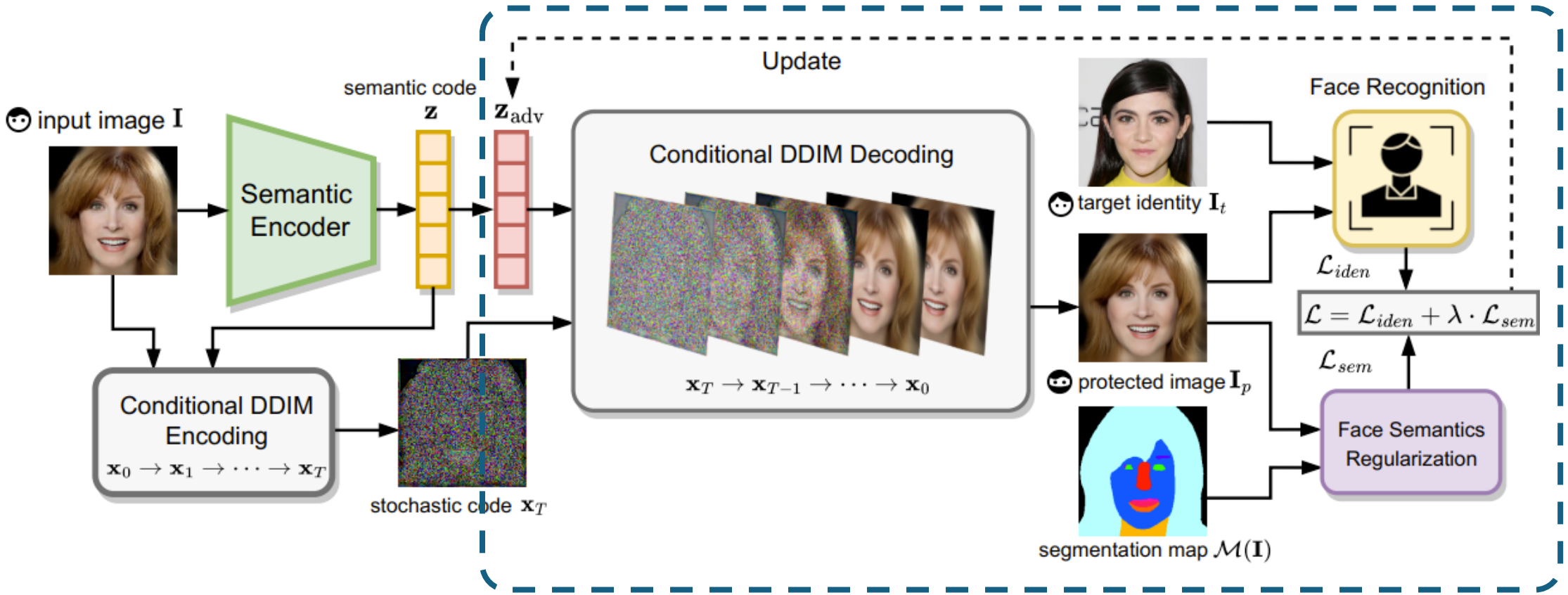
- Diffusion model for adversarial attacks
- Visual Quality ↑
- Attack Performance ↑



Overview


input image $\mathbf{I}$
Semantic Encoder
semantic code $\mathbf{z}$
$\mathbf{z}_{\text{adv}}$
Update
Conditional DDIM Decoding
$\mathbf{x}_T \rightarrow \mathbf{x}_{T-1} \rightarrow \cdots \rightarrow \mathbf{x}_0$
Conditional DDIM Encoding
$\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow \cdots \rightarrow \mathbf{x}_T$
stochastic code $\mathbf{x}_T$
target identity $\mathbf{I}_t$
protected image $\mathbf{I}_p$
segmentation map $\mathcal{M}(\mathbf{I})$
Face Recognition
$\mathcal{L}_{iden}$
$\mathcal{L} = \mathcal{L}_{iden} + \lambda \cdot \mathcal{L}_{sem}$
$\mathcal{L}_{sem}$
Face Semantics Regularization

Overview


input image $\mathbf{I}$
Semantic Encoder
semantic code $\mathbf{z}$
$\mathbf{z}_{\text{adv}}$
Update
Conditional DDIM Decoding
$\mathbf{x}_T \rightarrow \mathbf{x}_{T-1} \rightarrow \cdots \rightarrow \mathbf{x}_0$
Conditional DDIM Encoding
$\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow \cdots \rightarrow \mathbf{x}_T$
stochastic code $\mathbf{x}_T$
target identity $\mathbf{I}_t$
protected image $\mathbf{I}_p$
segmentation map $\mathcal{M}(\mathbf{I})$
Face Recognition
$\mathcal{L}_{iden}$
$\mathcal{L} = \mathcal{L}_{iden} + \lambda \cdot \mathcal{L}_{sem}$
$\mathcal{L}_{sem}$
Face Semantics Regularization

Main results

- DiffProtect outperforms state-of-the-art methods in terms of ASR and FID

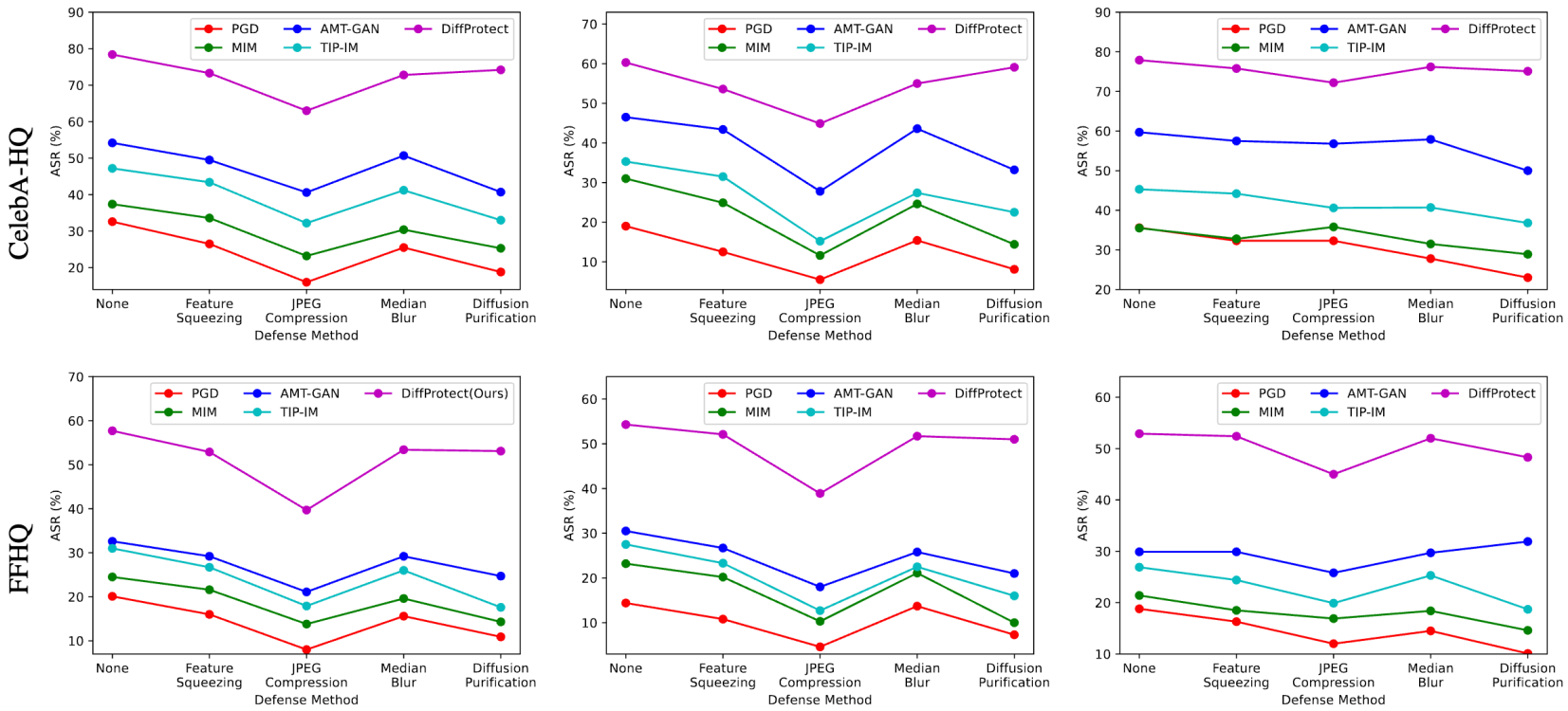
Methods	CelebA-HQ					FFHQ				
	ASR (%) ↑				FID ↓	ASR (%) ↑				FID ↓
	IRSE50	IR152	FaceNet	MobileFace		IRSE50	IR152	FaceNet	MobileFace	
No attack	7.3	3.8	1.1	12.7	0	4.4	2.5	1.7	5.2	0
PGD [35]	32.6	19.0	7.2	35.6	39.6	20.1	14.4	9.9	18.8	38.1
MIM [37]	37.4	31.0	11.3	35.5	69.9	24.5	23.2	13.9	21.4	62.1
TIP-IM [10]	47.2	35.3	15.3	45.3	71.6	31.0	27.5	12.5	26.9	62.8
AMT-GAN [11]	53.9	41.9	5.9	60.0	31.1	32.6	30.5	3.8	29.9	30.5
DiffProtect-fast ($\lambda = 0$)	68.4	49.8	19.5	72.1	26.7	50.8	47.6	18.4	47.0	26.4
DiffProtect ($\lambda = 0$)	**78.4**	**60.3**	**26.2**	**77.9**	27.6	**57.7**	**54.3**	**20.7**	**52.9**	26.0
DiffProtect ($\lambda = 0.2$)	67.7	48.7	19.2	69.3	**25.1**	46.2	45.4	16.5	44.3	**23.4**

Visualization

- DiffProtect produces good-looking protected images with natural and inconspicuous changes
- It works well across genders, ages, and races.

Attack Success Rate (ASR) under Defenses

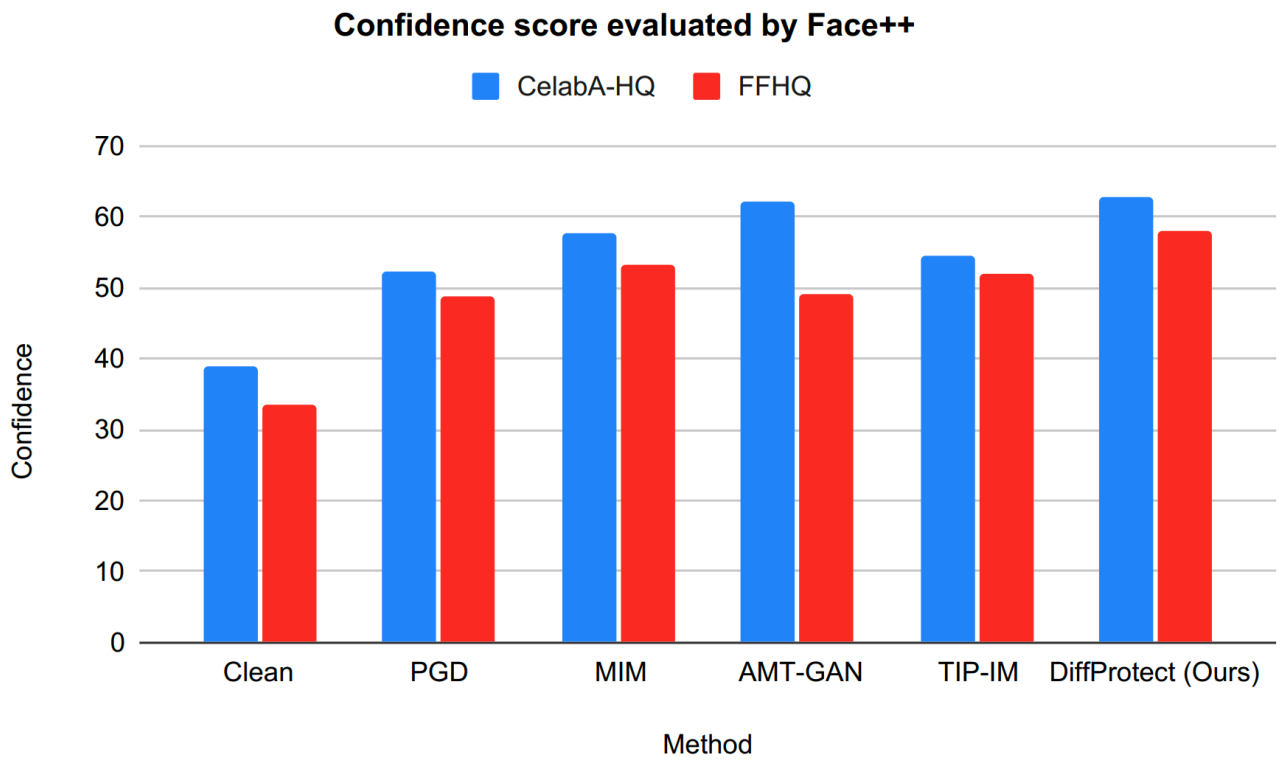
- Defense methods:
  - feature squeezing, JPEG compression, median blurring, and
- DiffProtect achieves the highest ASRs even under various defenses



(a) IRSE50 (b) IR152 (c) MobileFace

Evaluation on Commercial API

- DiffProtect achieves the highest confidence scores when evaluated by Face++.



Publications in 2023 - 2025

- Y. Guo, C. Peng, C. P. Lau and R. Chellappa, “Multi-Modal Human Authentication Using Silhouettes, Gait and RGB”, Proc. IEEE Conference on Face and Gestures, Kona, HI, Jan. 2023.
- A. Gupta and R. Chellappa, “You Can Run but not Hide: Improving Gait Recognition with Intrinsic Occlusion Type Awareness", Proc. Winter Conference on Applications on Computer Vision, Kona, Hawaii, Jan. 2024.
- Y. Guo, S. Huang, R. Prabhakar, C.L. Lau, R. Chellappa and C. Peng, “ Distillation-guided Representation Learning for Unconstrained Gait Recognition”, Proc. Intl. Jt. Conf. on Biometrics, Buffalo, NY, Sept. 2024.
- Z. Wang, J. Liu, R.P. Kathirvel, C.P. Lau and R. Chellappa, “HyperGait: A Video-based Multitask Network for Gait Recognition and Human Attribute Estimation at Range and Altitude”, Proc. Intl. Jt. Conf. on Biometrics, Buffalo, NY, Sept. 2024.
- Y. Guo, A. Shah, J. Liu, A. Gupta, R. Chellappa and C. Peng, “GaitContour: Efficient Gait Recognition based on a Contour-Pose Representation”, Proc. Winter Conference on Applications on Computer Vision, Tucson, AZ, Feb. 2025.
- S. Huang, R. Prabhakar Kathirvel, Y. Guo, C. P. Lau, and R. Chellappa, “Whole-body Detection, Identification and Recognition at Altitude and Range”, IEEE Transactions on Biometrics, Behavior and Identity Science, 2023
- A. Gupta and R. Chellappa, “MimicGait: A Model Agnostic approach for Occluded Gait Recognition using Correlational Knowledge Distillation”, Proc. Winter Conference on Applications on Computer Vision, Tucson, AZ, Feb. 2025.
- Z. Wang, J. Liu, J. Chen and R. Chellappa, “VM-Gait: Multi-Modal 3D Representation Based on Virtual Marker for Gait Recognition”, Proc. Winter Conference on Applications on Computer Vision, Tucson, AZ, Feb. 2025.
- S. Huang, R.P. Kathirvel, Y. Guo, R. Chellappa and C. Peng, “VILLS : Video-Image Learning to Learn Semantics for Person Re-Identification”, Proc. Winter Conference on Applications on Computer Vision, Tucson, AZ, Feb. 2025.

- AI and biometric recognition have similar challenges
  - Bias, vulnerabilities to adversarial attacks, privacy concerns, domain adaptation/generalization
  - Solutions to these concerns are generalizable to other AI applications (medicine, public health, security and defense)
- If AI Can Solve Biometrics, …
- What are AI easy and AI hard problems?