Design of Unbiased, Adaptive and Robust AI Systems

November 22 , 2021

Rama Chellappa
Artificial Intelligence for Engineering and Medicine (AIEM),IAA, CIS, CLSP, Malone Center, and MINDS
Departments of Electrical and Computer Engineering and Biomedical Engineering (School of Medicine)
Johns Hopkins University
College Park Professor, University of Maryland

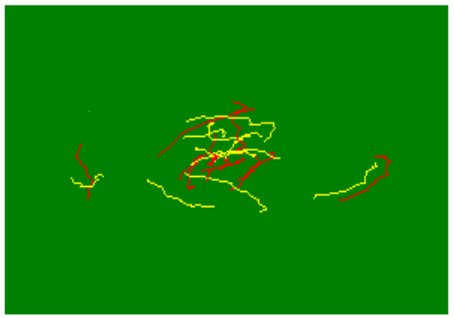
Outline

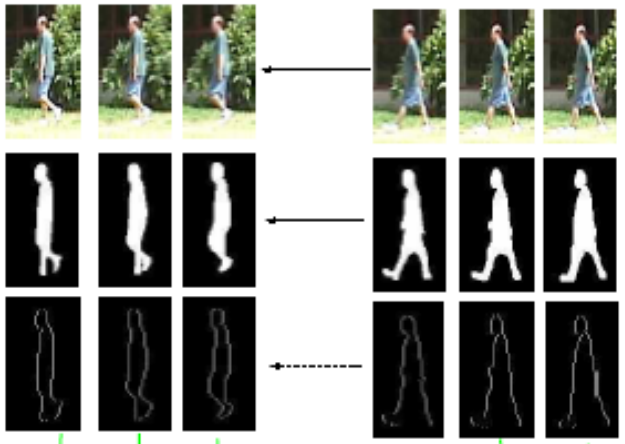
- Impact of mathematics/statistics on computer vision, machine learning and AI
- A brief history of AI
- Three at scale case studies
  - Unconstrained face verification and recognition
  - Vehicle detection and re-identification
  - Action detection in untrimmed videos
- Challenges in AI and partial solutions
  - Bias
  - Domain adaptation/generalization
  - Vulnerability to attacks
- Looking ahead
  - AI's impact on medicine and healthcare
  - Some open problems for math/stat folks

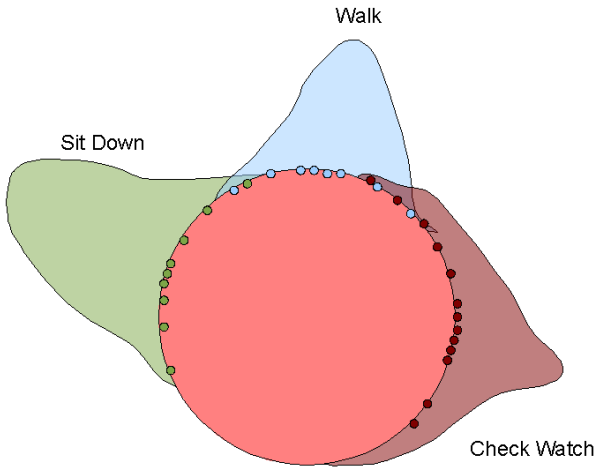
1970's - 2010s –The golden decades for math/statistics, computer vision and AI

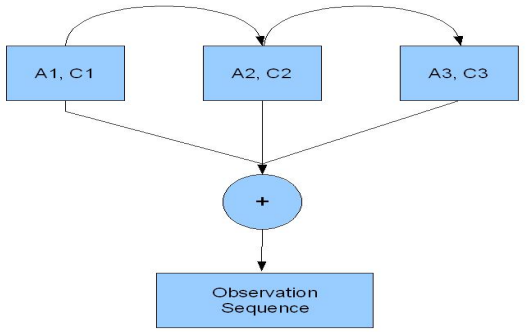
- MRF representations, estimation methods, neighborhood selection rules, texture synthesis, classification, image restoration
- Performance bounds, robust statistics for computer vision and machine learning
- Invariants
- Mumford-Shah segmentation algorithm
- Calculus of variations
- Lie group

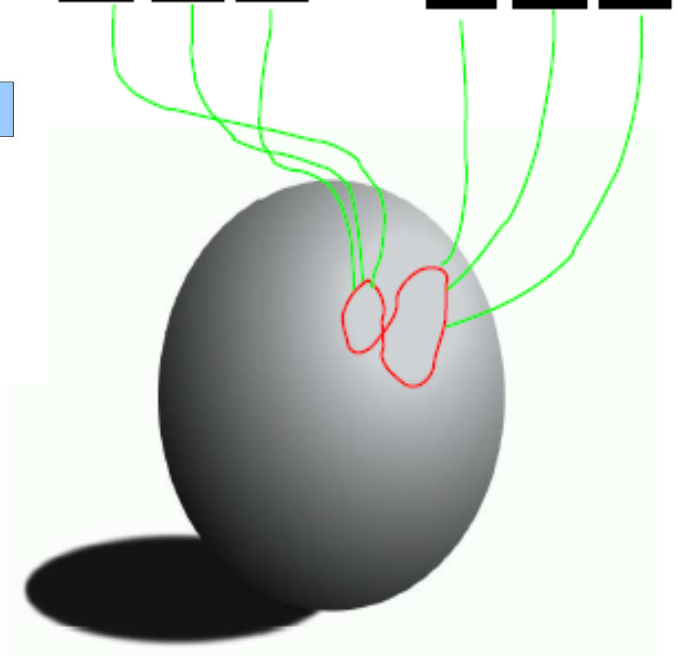
Statistics on manifolds











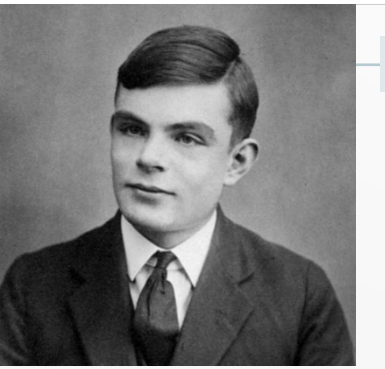
More on math/statistics

- Fisher-Rao metric
- Shape statistics (Mardia, Srivastava)
- Dictionary learning on statistical manifolds
- Model order selection
  - Bayes information criterion
- Object recognition
- Bayesian graphical models – shallow hierarchy – uncertainty in AI
- Simulated annealing, particle filter, MCMC
- Statistics on special manifolds

Statistics is struggling

- Statistical methods were mostly absent when compressive sensing and sparse representations were popular (2005-2012)
  - Statistics likes *l2* more than *l1* and *l0*!
- When hierarchical models are considered
  - Multi-resolution time series models, MRFs have challenging inference problems
- Statistical methods for hierarchical and non-linear models (Deep learning) are even more challenging!

Directions in AI



"Can machines think?"

50's 60's 70's 80's 90's 00's 10's 2020

Theorem proving

Heuristic search

Planning

Chess, Go

Search Engines

Search — **Problem Solving** → **Game Strategies** → **Browsers**

Common-sense reasoning

Knowledge representation

Machine translation

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

Rule-based
Eliza: Text response

SHRDLU: Blocks world

Statistical

Speech to Text

Text analytics

Voice assistants

Vision — **Computer Vision**

Optical character recognition

Scene Understanding

Robotics

Object Recognition

Deep Learning

Synthesis: GANs

Expertise — **Expert Systems** → **Recommenders**

Prospector

Medial diagnosis

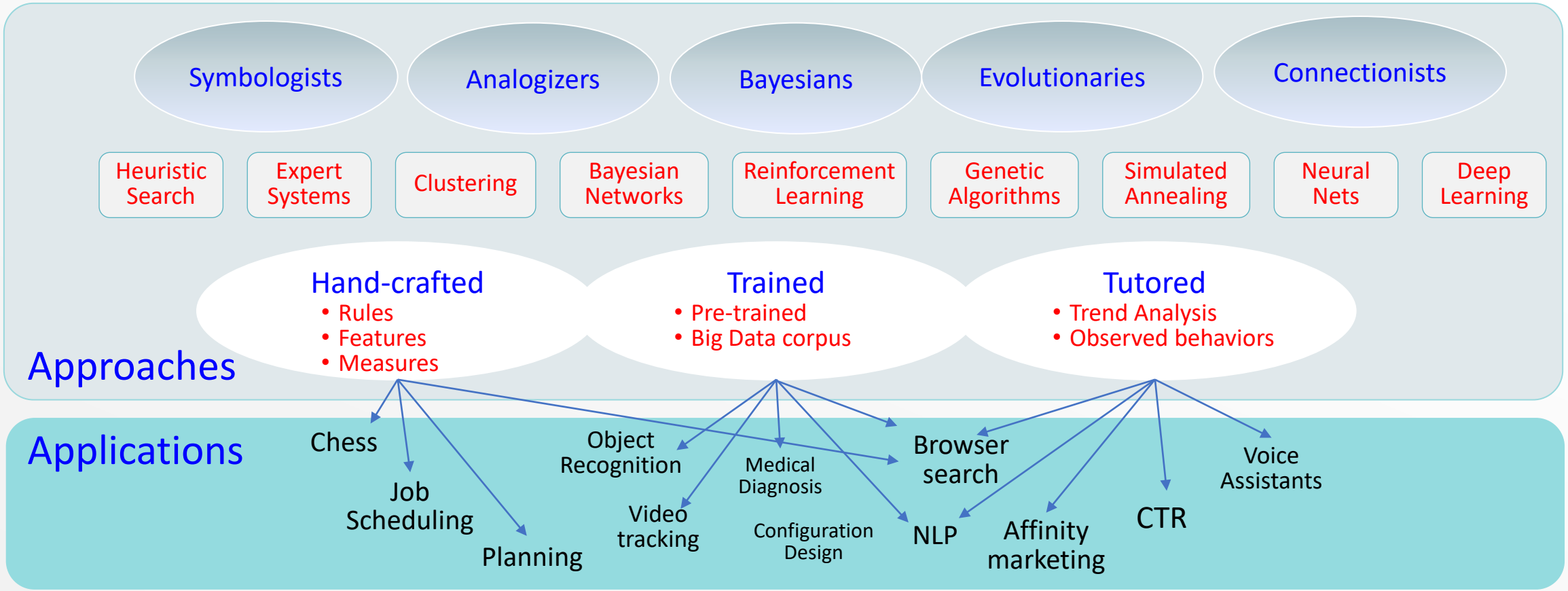
Configuration Design

Affinity Marketing

CTR

Credit to Robert Hummel

Many AI techniques have led to many applications



Credit to Robert Hummel

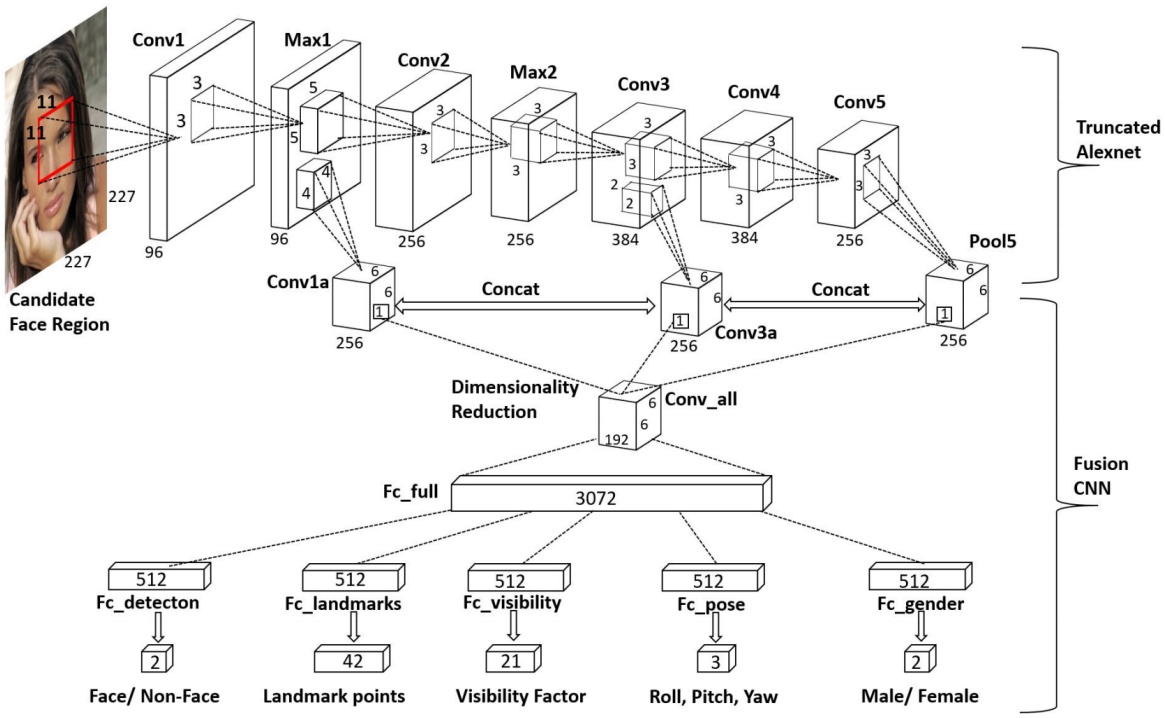
Deep learning - miracle or mirage?

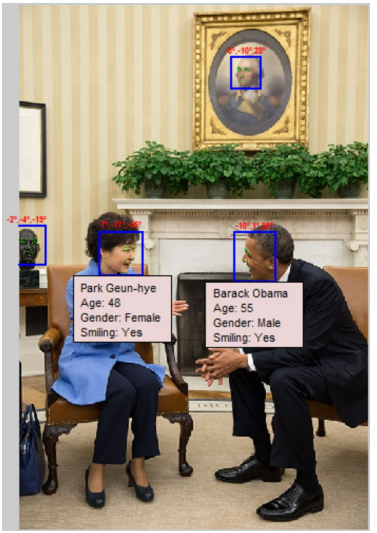
- Since 2012, computer vision has become a one-trick pony
- Even more troublesome – AI and deep learning are used synonomously
- Impressive performance on many tasks
  - Object/ face detection, classification, verification
  - For face verification at 10-7 false acceptance rate, > 90% true acceptance rate on faces in the wild (IARPA JANUS program)
- Not there yet
  - For action detection, probability of miss for the best systems are 0.34 and 0.53 for known and unknown facilities and 37 actions. 0.6 for detecting surprise activities

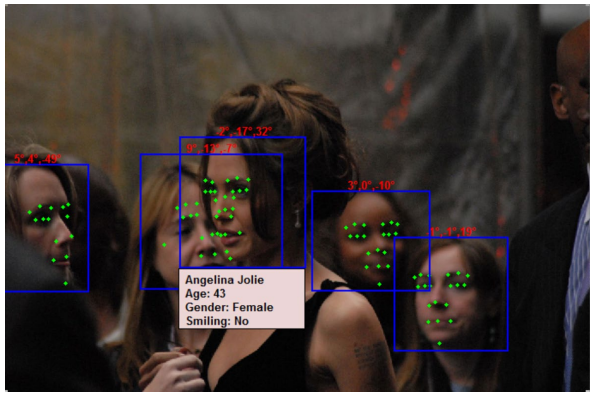
Unconstrained face verification

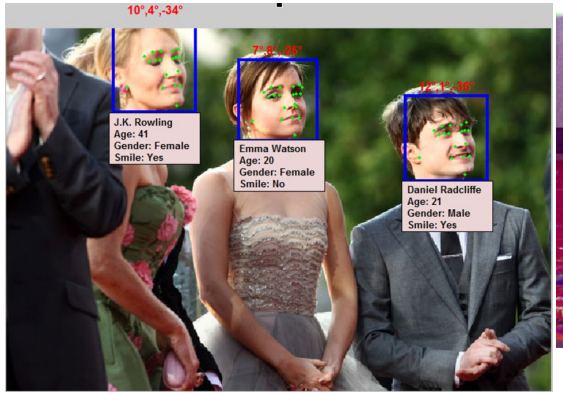
- 2014 – 2020, Supported by the IARPA JANUS program
- UMD (Lead) with CMU, Columbia, JHU, UB, UCCS, UTD.
- Multi-task learning in deep networks
  - Face and gender detection, pose and age estimation, fiducial extraction
- Network of networks
  - Fusion of short and tall networks
- Current template size 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.
- Drs. Rajeev Ranjan, Ankan Bansal, Hui Ding contributed to this effort.

Hyperface architecture (PAMI 2017)







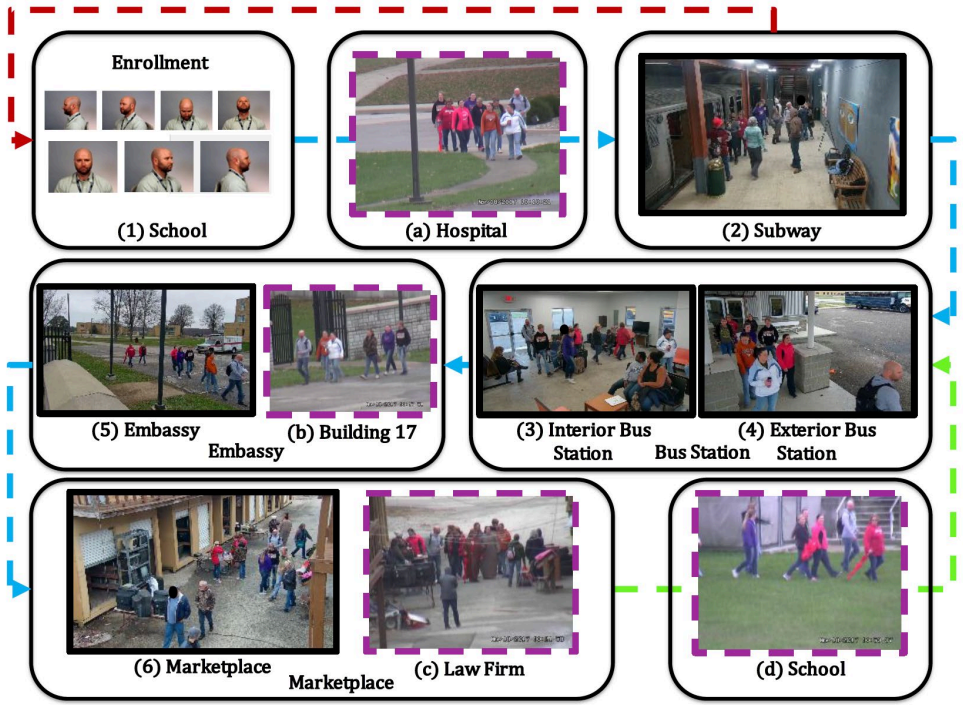


Ranjan, et al., T-PAMI, 2018

Unconstrained video-based face identification

- Recognize the identity of the target face in a video

IJB-B (Multi-shot Videos)

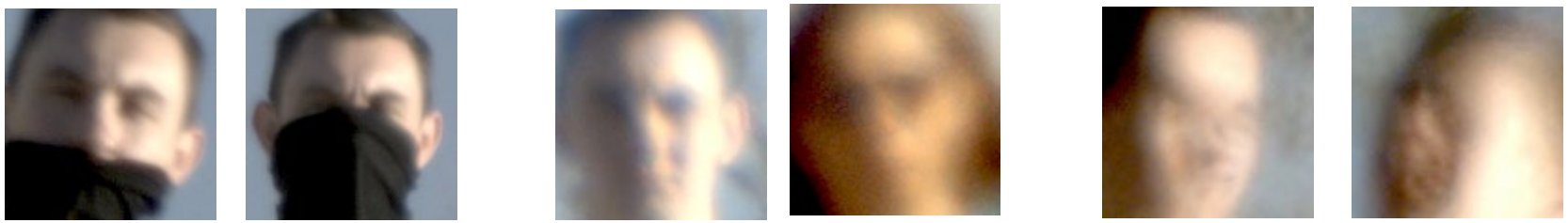


CS6 (single-shot surveillance Videos)

CS6 video dataset
An unconstrained video-based face recognition dataset.
Galleries: high-resolution still images. Probes: low quality, remotely captured surveillance videos.
202 subjects from 1421 images and 398 single-shot surveillance videos.
We focus on surveillance-to-single , surveillance-to-booking and surveillance-to-surveillance identification protocols.
Zheng, et al, T-BIOM 2019, Ranjan, et al., T-BIOM 2019

Other problems we have Worked on

- Large gallery face dataset
  - The large-scale dataset used in this work consists of 14,728,804 images from 4,233,581 distinct individuals of non-U.S. individuals. On average there are 3.47 images per individual - the number of images for each individual varies from a minimum of 1 to a maximum of 139, with most of the individuals having only 1 image.
- Faces at 300m 650m 1000m



- IARPA Janus Benchmark Multi-Domain Face (IJB-MDF) dataset.
  - Source domain is in the visible domain and the target domains are from four SWIR cameras.

Lau, et al, T-BIOM, 2021, Lau, et al., JST Special Issue on Deep learning, 2020

AICity challenges (2019-2021)

- Algorithms developed for multiple tasks (vehicle re-id, vehicle counting and anomaly detection)
- Datasets for challenge are carefully curated, with annotations
  - High Resolution: Most images are 1920x1080
  - High Frame Rate: Usually 10 frames per second

Clean curated data

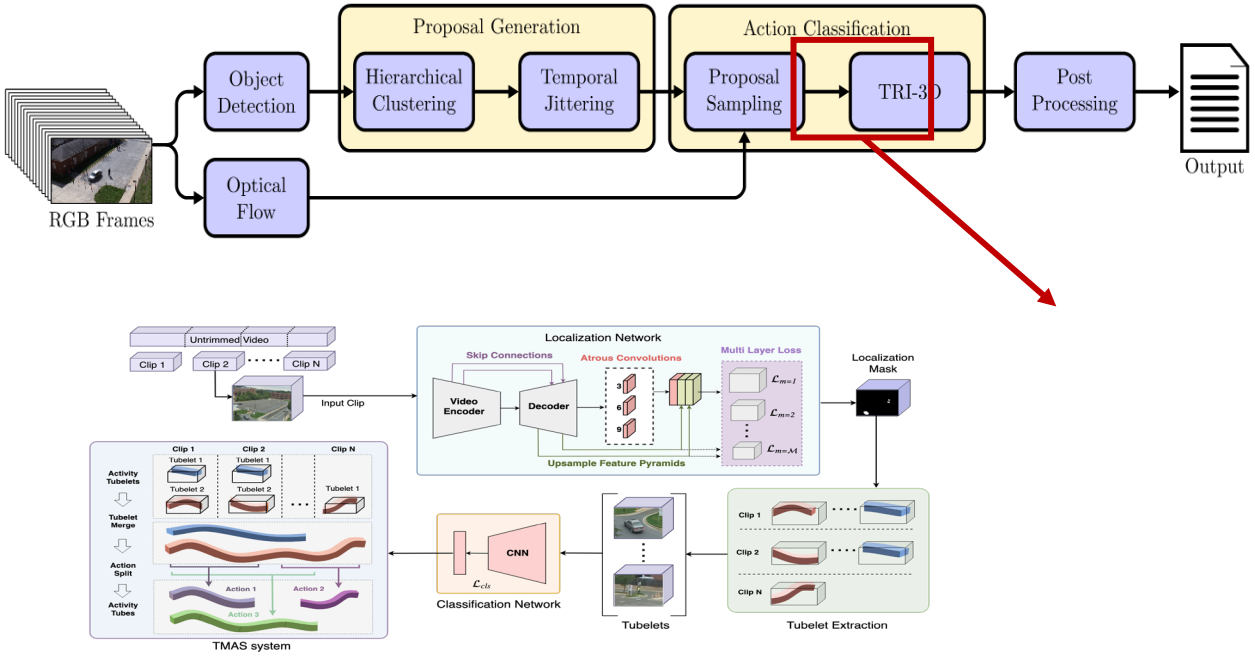
Challenges on trafficView/CATT data

- Operational cameras tend to be lower quality/less maintained by local governments than curated challenge datasets

    - Lower Resolution: Most images are 320x240
    - Lower Frame Rate
    - Artifacts such as compression artifacts, video tearing, etc
    - Various Scales
    - (large vehicles + closer to camera vs small vehicles + further from camera)

Activity detection in nntrimmed videos-DIVA (JHU, UMD, UCF, CMU, Columbia)

UMD-JHU system: Modular, generalizable, proposal-based system for action detection. Uses objects and motion for activity detection.
UCF system: Takes untrimmed RGB videos as input and provides spatio-temporal localization of activities present in videos
Columbia System: Takes the UMD/JHU proposals as input and replace the TRI-3D classifier with a 2-stage classifier to detect long-tail events (12 few-shot classes). For surprise activity detection, takes the UMD/JHU proposals as input, and combines a few-shot visual-based module and a text-based cross-modal module.
CMU Surprise activity retrieval system: Uses UMD/JHU's pipeline. Takes both visual and text surprise queries as input.
Retrieves activity cuboids for each surprise query



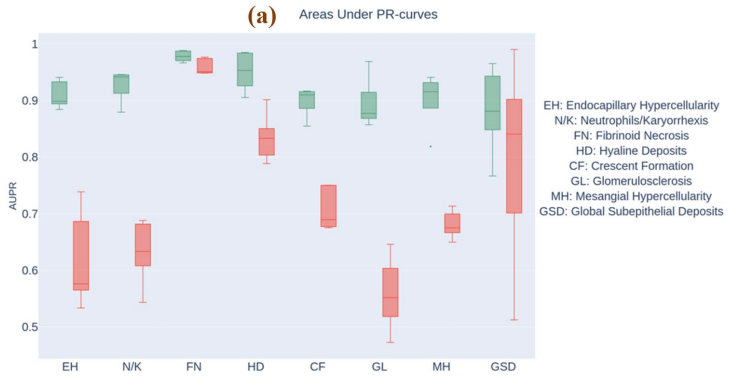
Best results: https://acev.nist.gov/sdl All p_miss numbers are at TFA=0.02
KF (EO): p_miss: 0.34, relative time: 0.41
KF (IR): p_miss: 0.68, relative time: 0. 38.
UF (Known activities): p_miss: 0.54, relative time: 0.68
UF (Surprise activities) p_miss: 0.78, relative time: 0.61.

Gleason, et al., WACV 2019, WACV2020

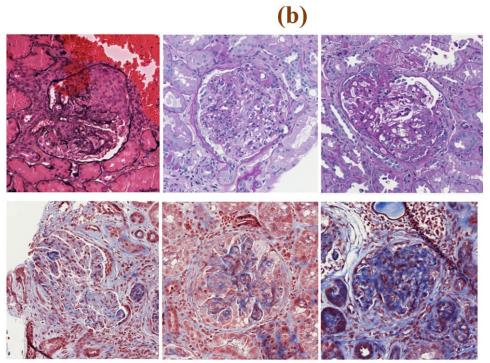
p_miss is the portion of activities where the system did not detect the activity for at least 1 second. TFA is the portion of time that the system detected an activity when in fact there was none

Challenges of using deep learning for pathology

- High cross-institutional data variation in staining protocols, patient population, etc.
- Models trained by one lab may not work well for other labs



(a) DenseNet-121 achieves good performances on data from the same lab (green color) but performs poorly when applied to data from another lab (red color).



(b) Glomerular images from one lab (top) and another lab (bottom) have different colors, brightness, and texture due to variations in slide thickness, reagent, etc.

Slide from Prof. Hien Nguyen, Univ. of Houston

Despite being successful, deep learning-based methods have issues

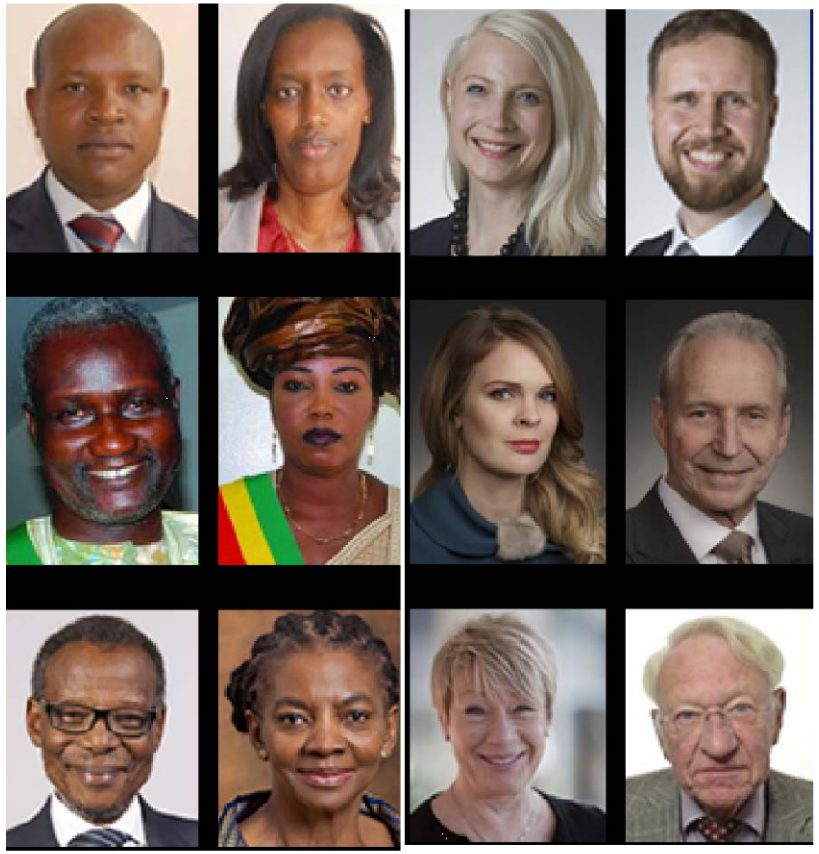
- While seen as a non-linear mapping between data and labels, lack of analytical results is worrisome.
- Learning millions of parameters from relatively small data is a statistical blasphemy!
- Tightly clings to training data and does not generalize well
- No performance measure to say why and when it works
- We can pile on…

Open problems and partial solutions

- Bias
  - Face recognition, vehicle re-identification
- Domain adaptation
  - Vehicle re-identification, activity detection
- Adversarial attacks and robust defenses
  - Patch attacks on object detection

The PPB dataset

AFRICAN SCANDINAVIAN



PPB: Pilot Parliaments Benchmark

6.3%

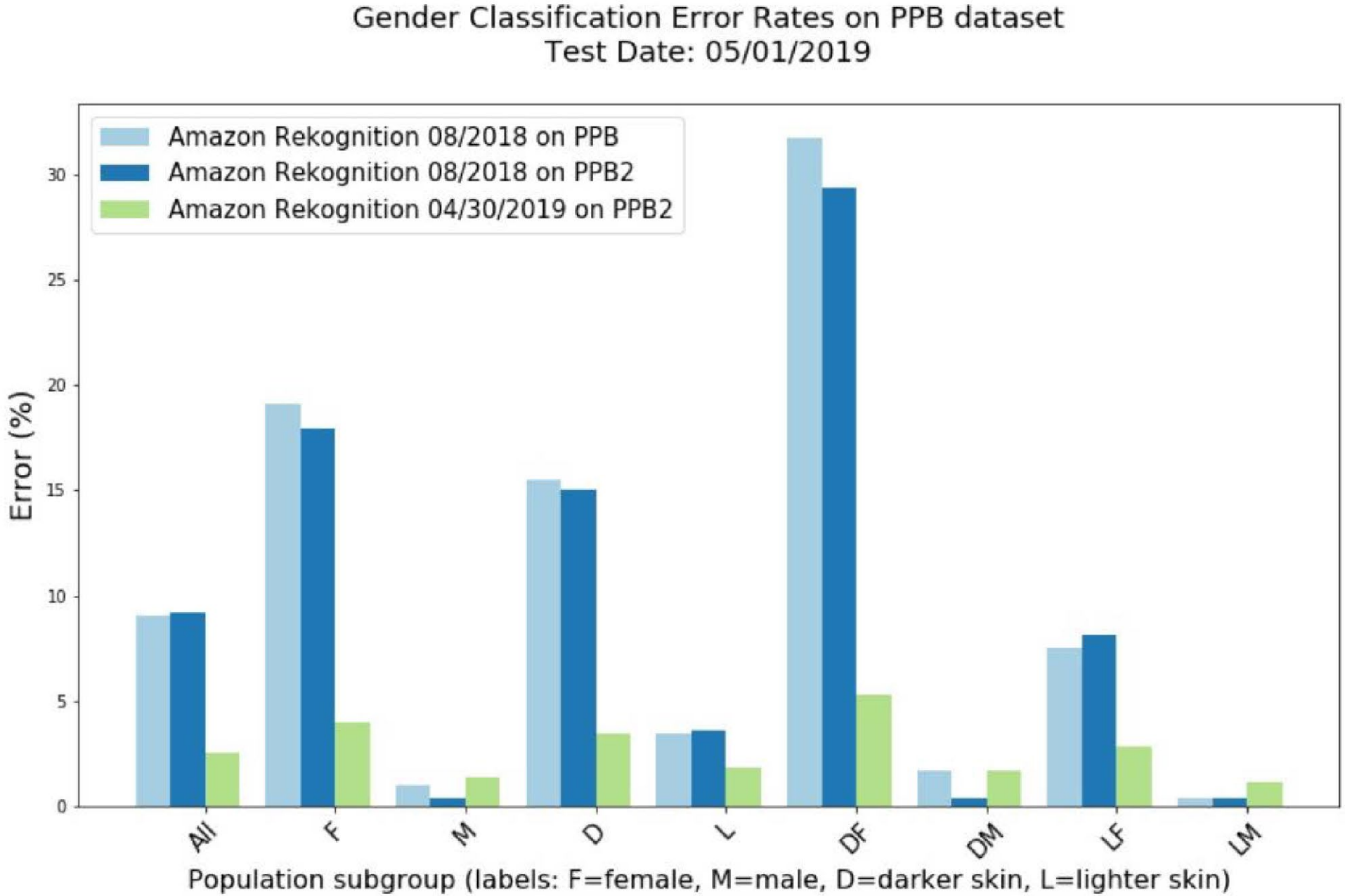
20.8%

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

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)

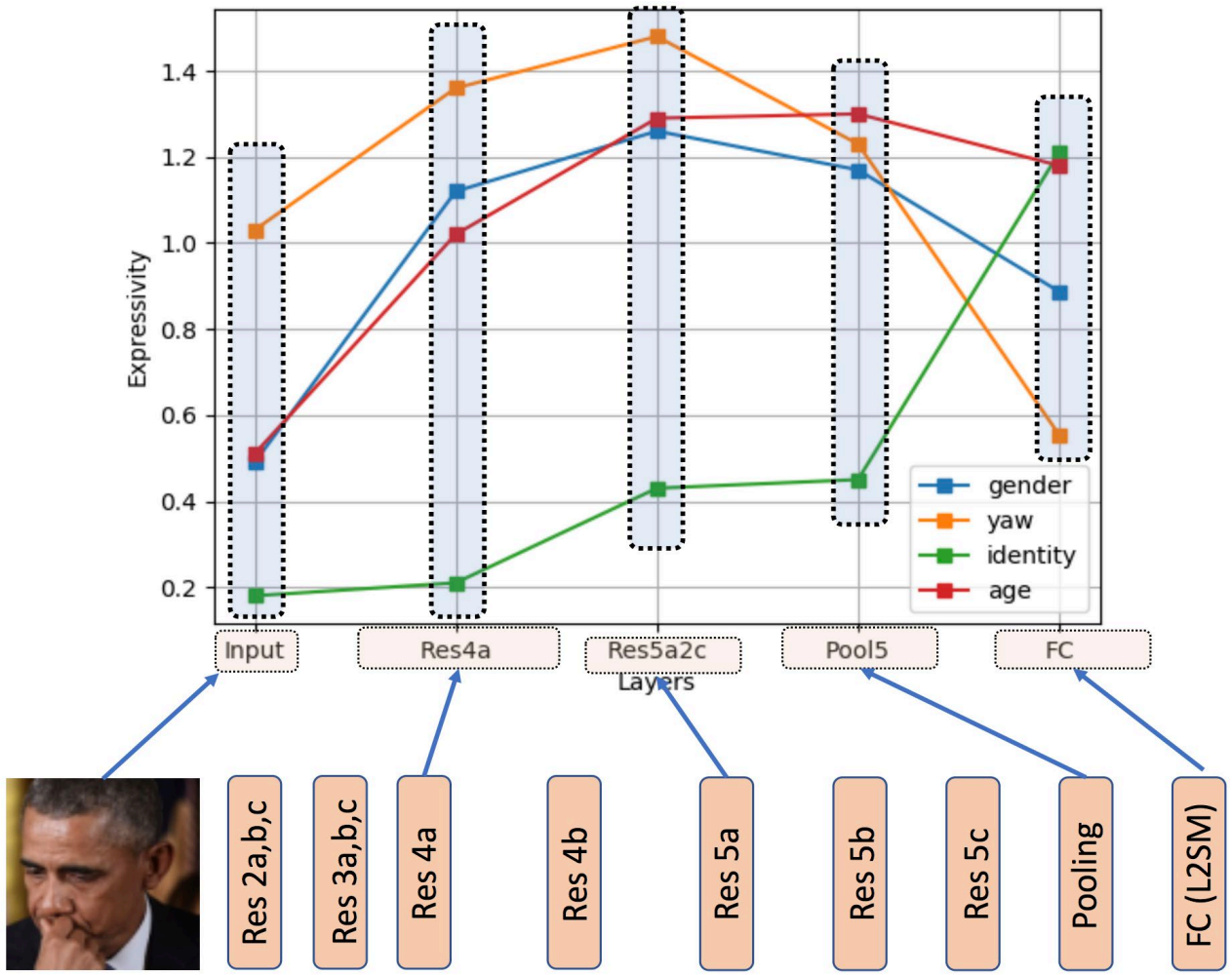
Expressivity of facial attributes

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

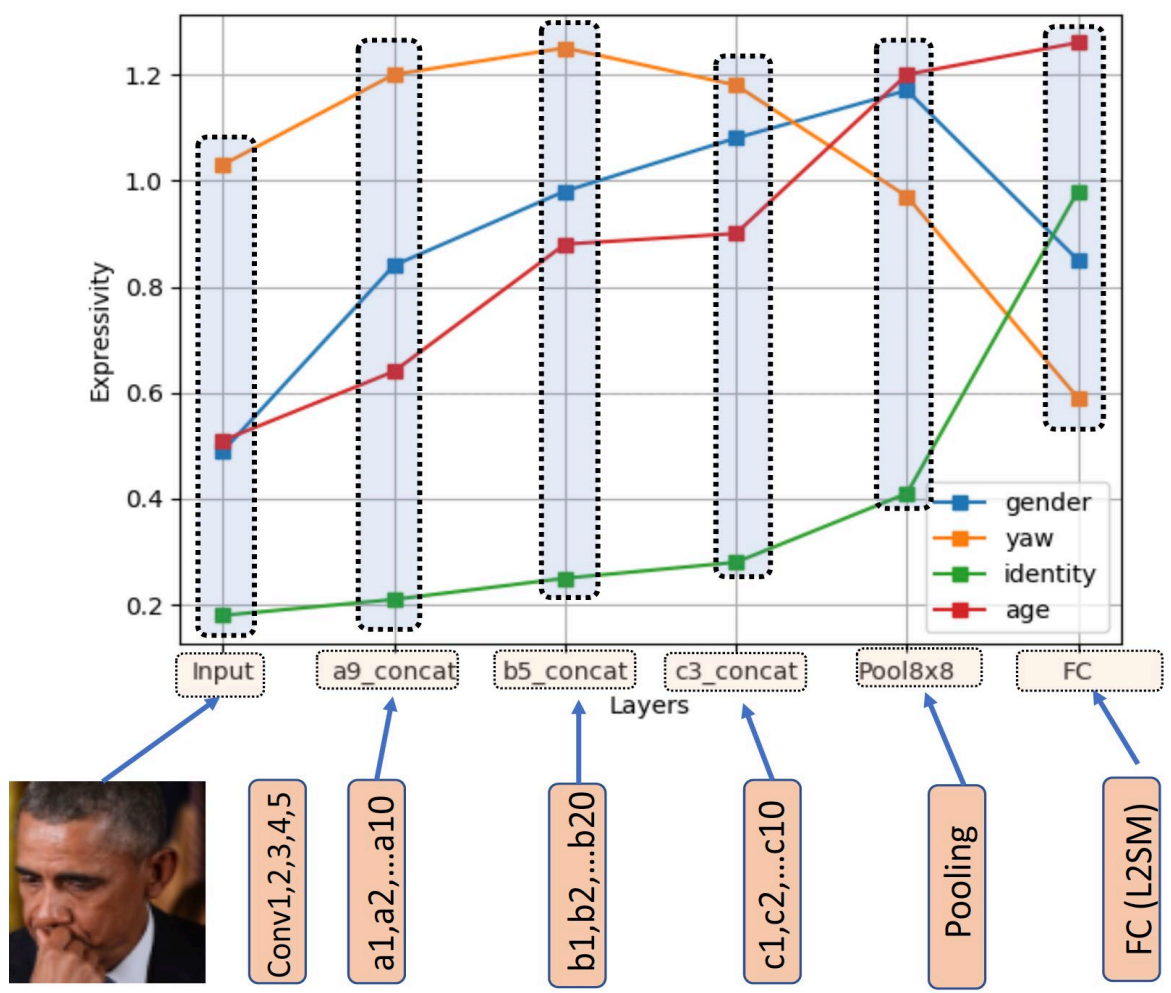
- We compute expressivity of facial attributes (yaw, age, gender, identity) in a given set of face descriptors

- 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 (Dhar, et al., FG2020



Crystalface (Resnet-101)

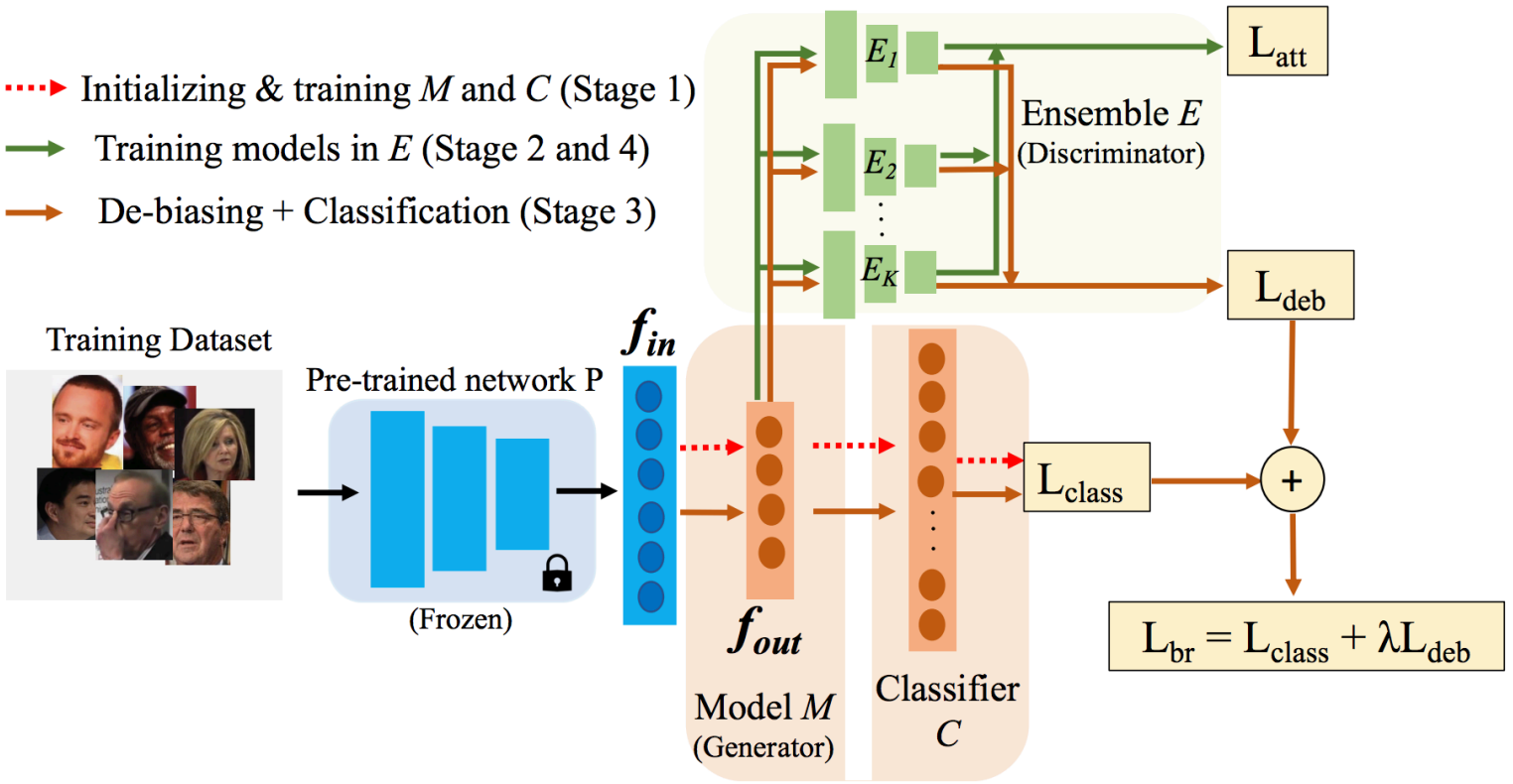


Inception Resnet-v2

Key takeaways

- Face recognition features implicitly encode attributes like yaw, gender and age.
- During the training process, the expressivity of identity increases while that of yaw, gender and age decreases, thus showing that *un-learning is a part of learning*. Expressivity of yaw, especially, decreases very rapidly.
- Rate of un-learning: **Age < Gender < Yaw** (opposite to the order of attribute-wise relevance)

Protected Attributes Suppression System (PASS)



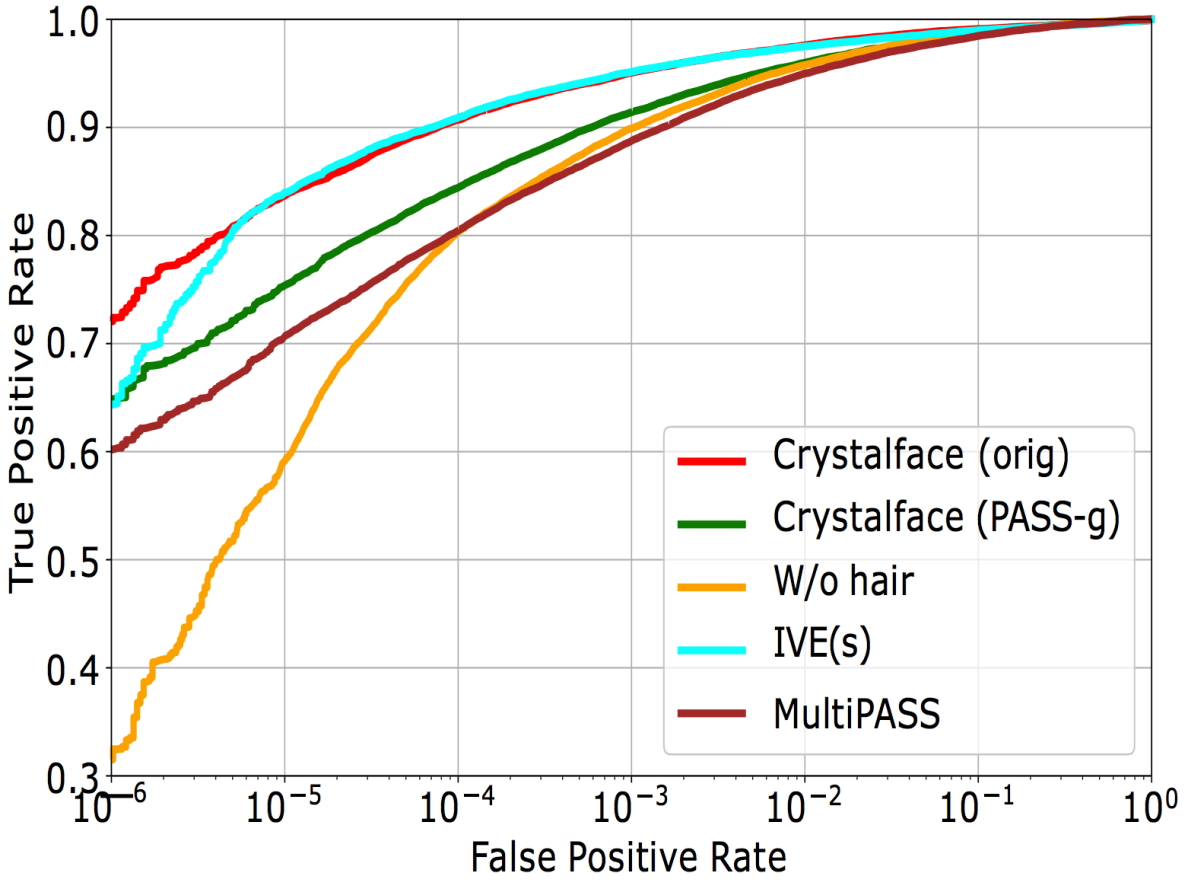
Dhar, et al, ICCV 2021

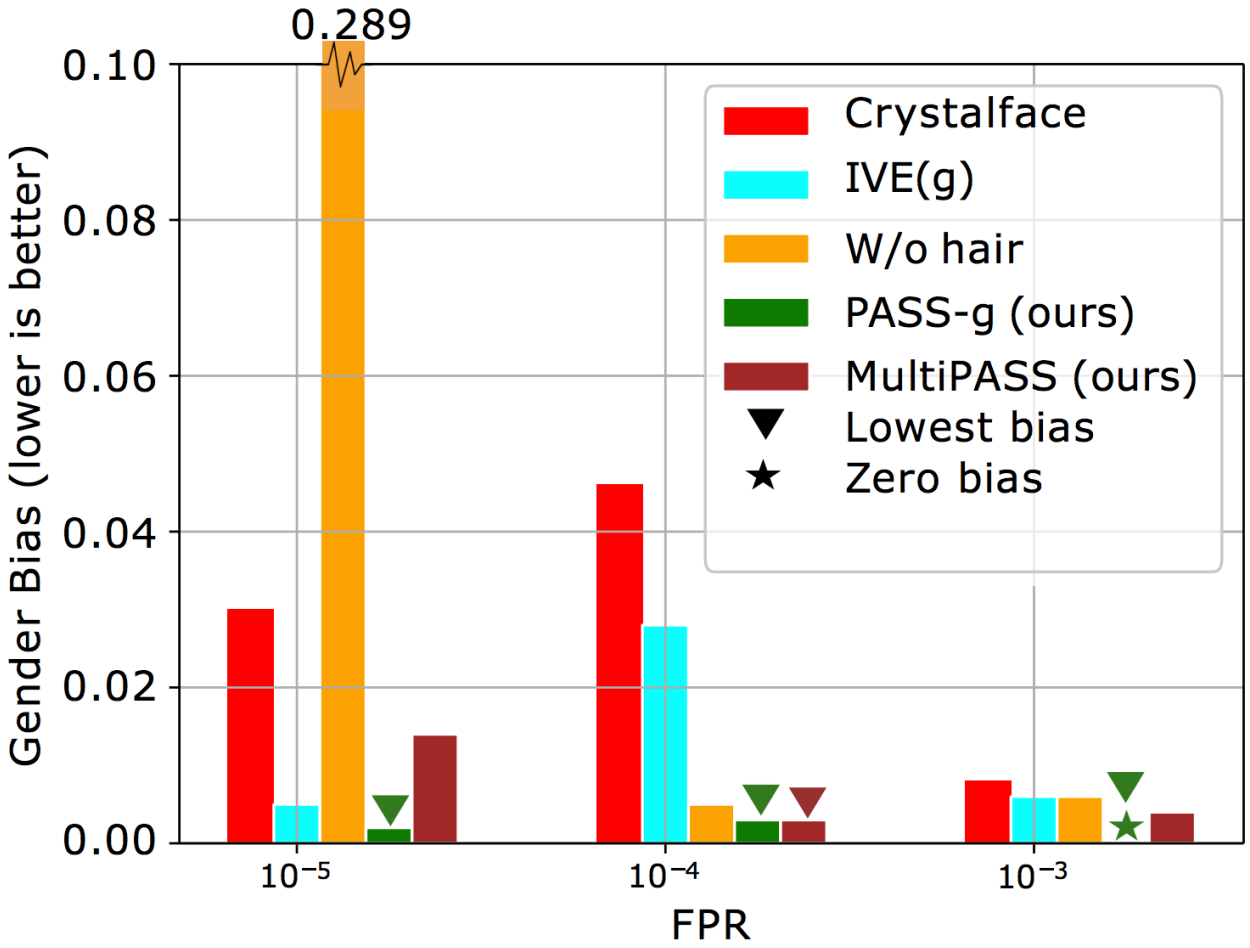
L_{class}: Classification loss for training **M** to classify identities

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

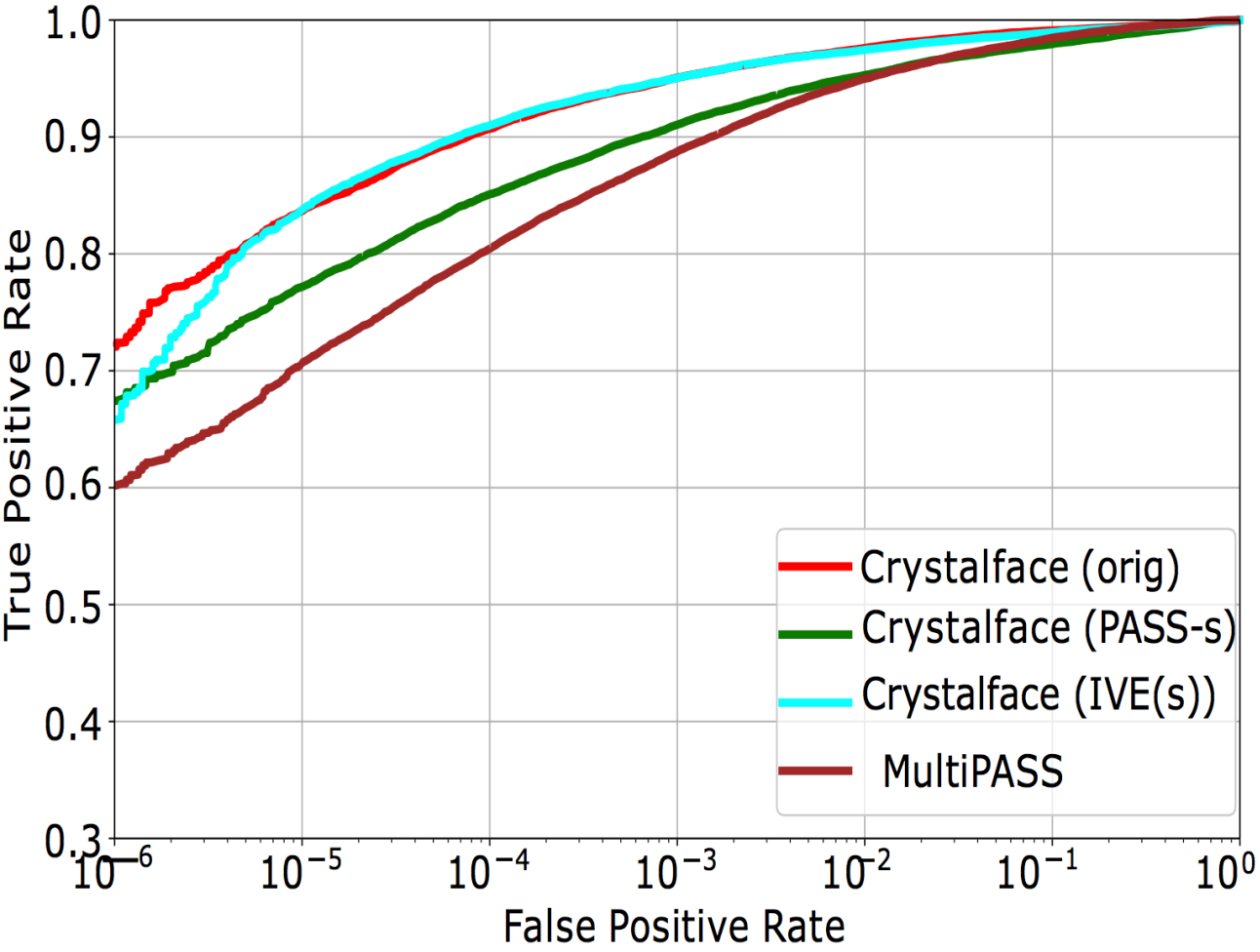
Results (Crystalface – Gender bias)

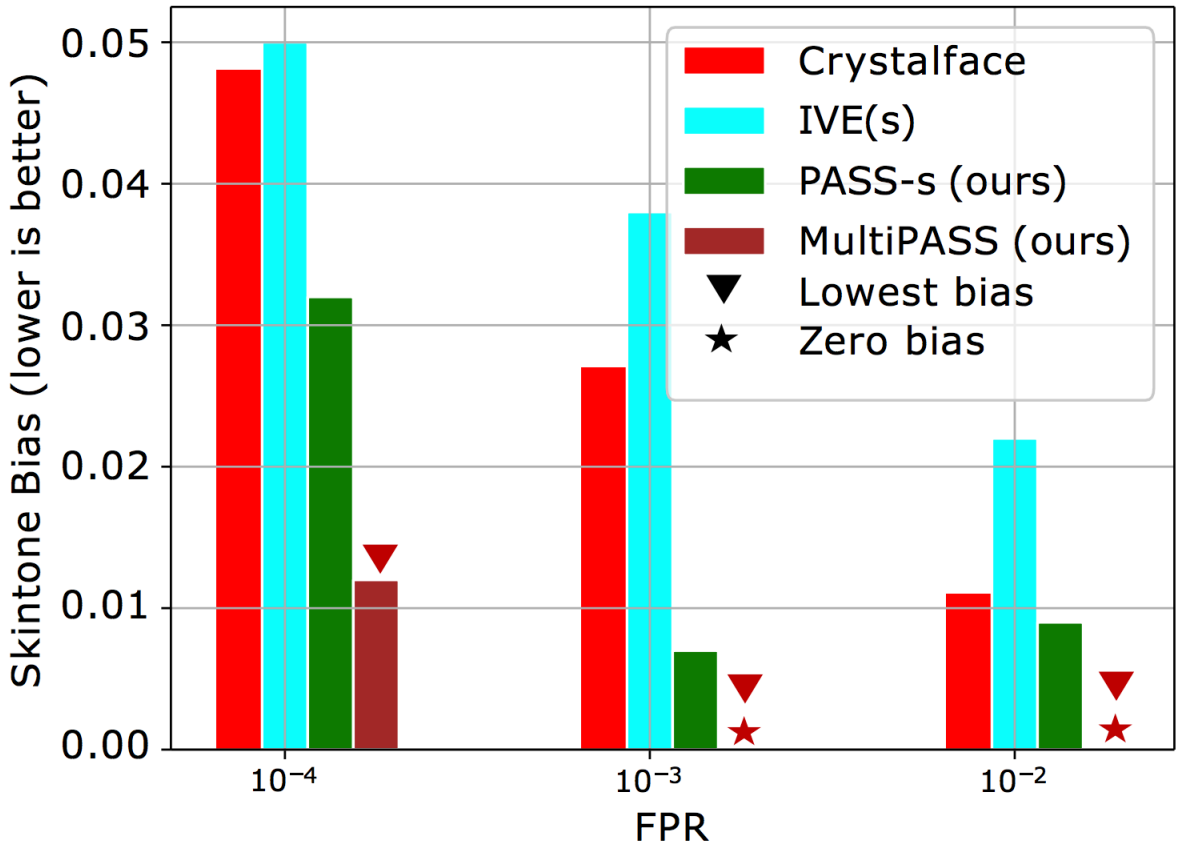




Crystalface: JANUS-UMD face matcher
IVE: Incremental variable elimination

Results (Crystalface – Skin tone 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

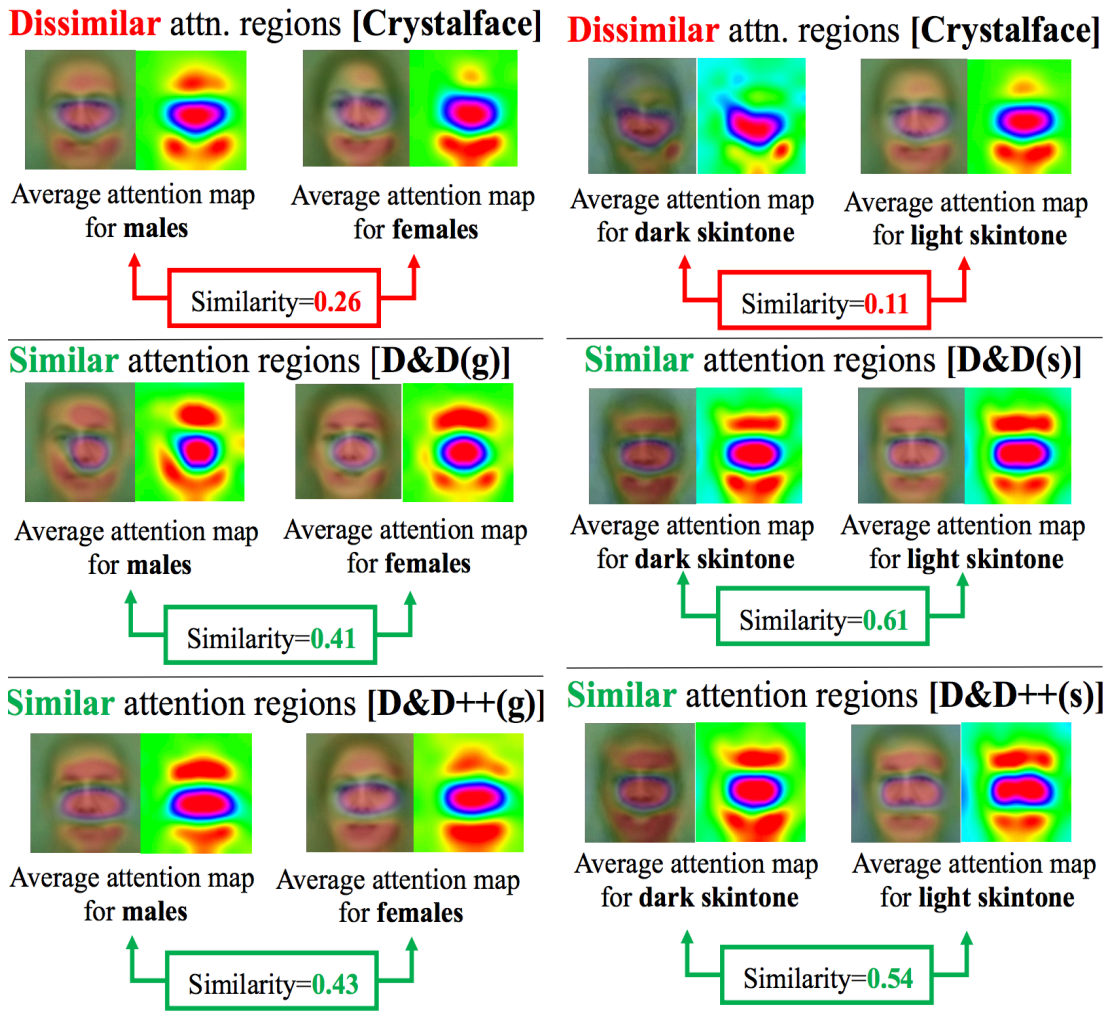
FPR		10^{-5}		10^{-4}		10^{-3}	
Network	Acc-g(↓)	TPR	BPC_g (↑)	TPR	BPC_g(↑)	TPR	BPC_g(↑)
Crystalface[36]	86.73	0.833	0.000	0.910	0.000	0.951	0.000
W/o hair[3]	86.04	0.589	-8.926	0.780	0.823	0.899	0.195
IVE(g)[42]	86.10	0.833	<u>0.833</u>	0.910	0.391	0.951	0.250
PASS-g	<u>80.54</u>	0.761	**0.847**	0.839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

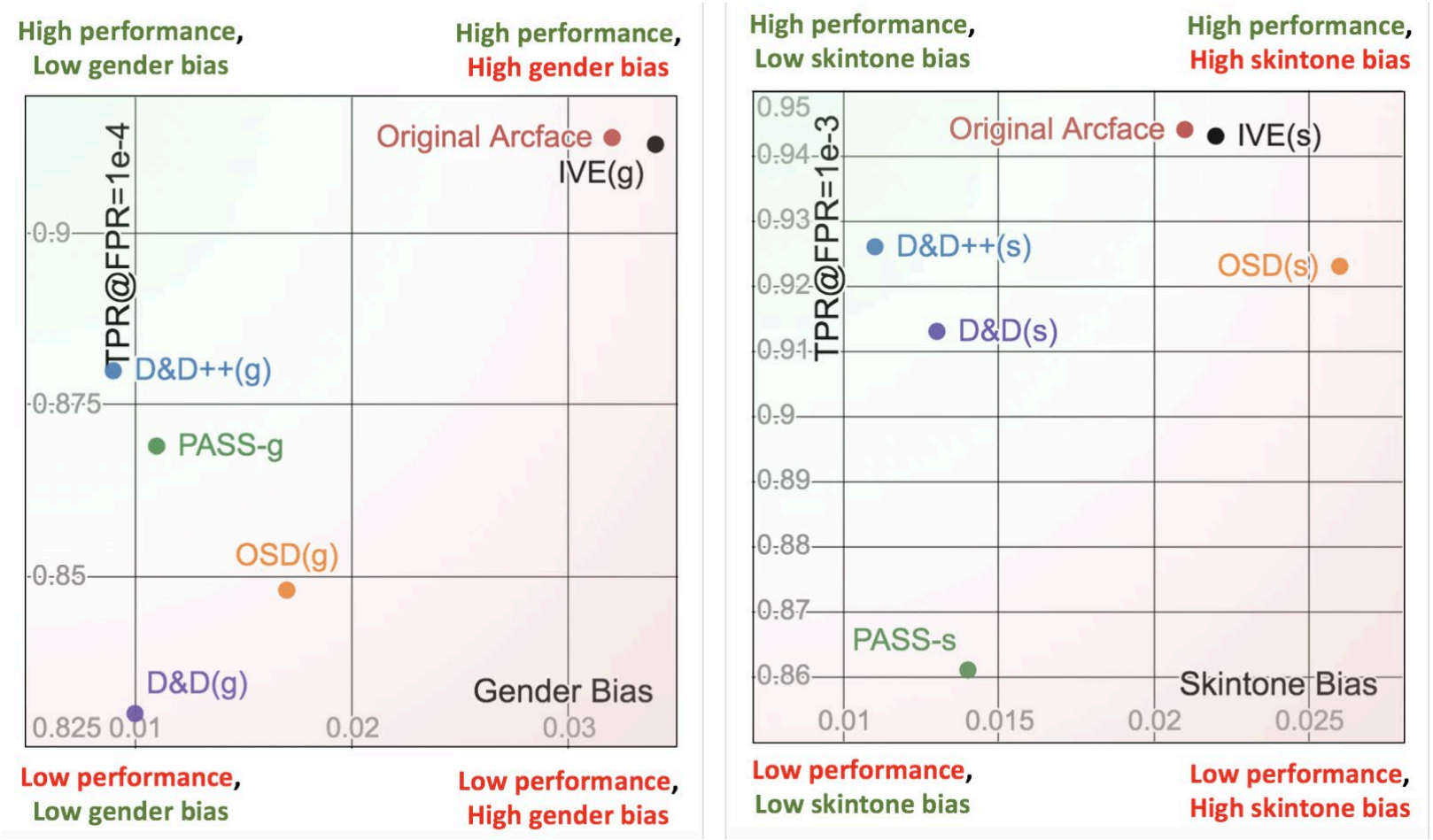
Qualitative results

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.

Bias performance tradeoff (TPR v/s bias)



Adversarial de-basing methods like PASS reduce bias but also reduce verification performance, but D&D++ reduces bias and minimizes the drop in verification performance

Domain adaptation: Motivation (Saenko et al., ECCV '10)

Source domain
Data: X, Labels: Y

Target domain
Data: X', Labels: Y'

Transfer Learning[1]
- $P(Y|X) \neq P(Y'|X'), P(X) \approx P(X')$

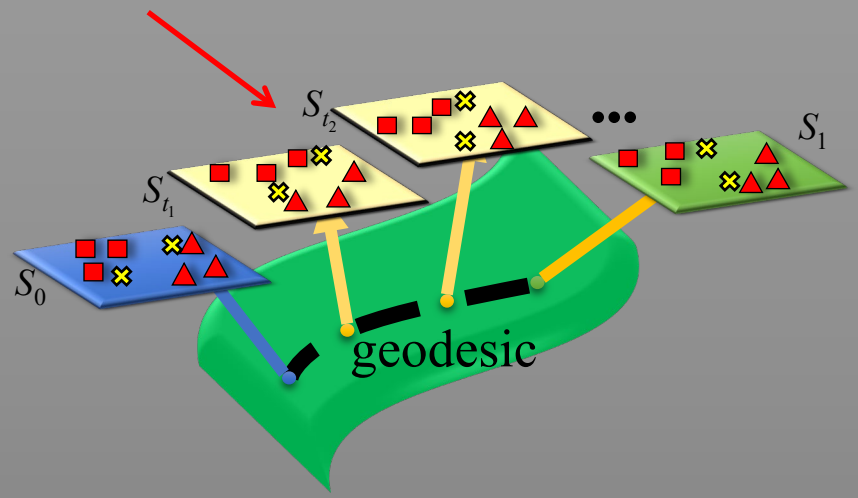
Domain adaptation
- $P(X) \neq P(X'), P(Y|X) \approx P(Y'|X')$

Image credit: Saenko et al., ECCV 2010, Bergamo et al., NIPS 2010

[1] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowledge and Data Engineering, 22:1345 –1359, October 2010.*

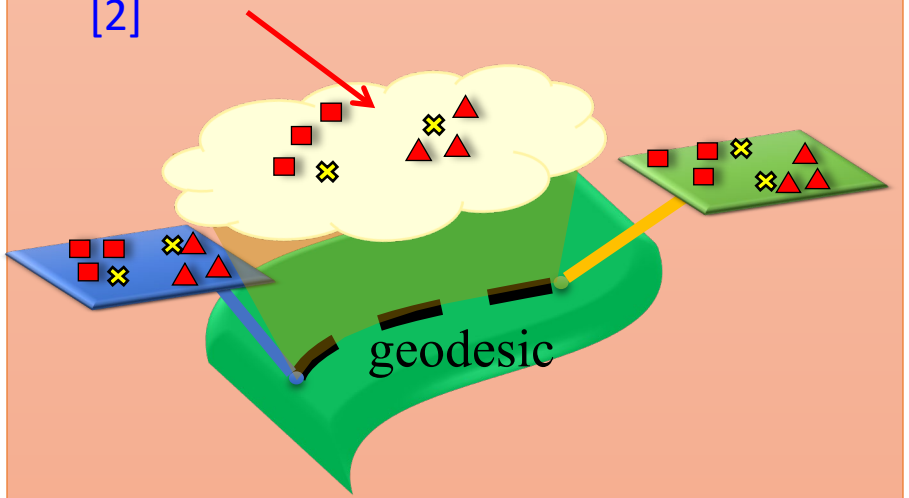
Finite vs infinite intermediate subspaces for domain adaptation

Finite intermediate subspaces [1]



- samples a limited number of intermediate subspaces
- concatenates the subspace projection as the final features for learning.
- train a discriminative learner on the projected source data

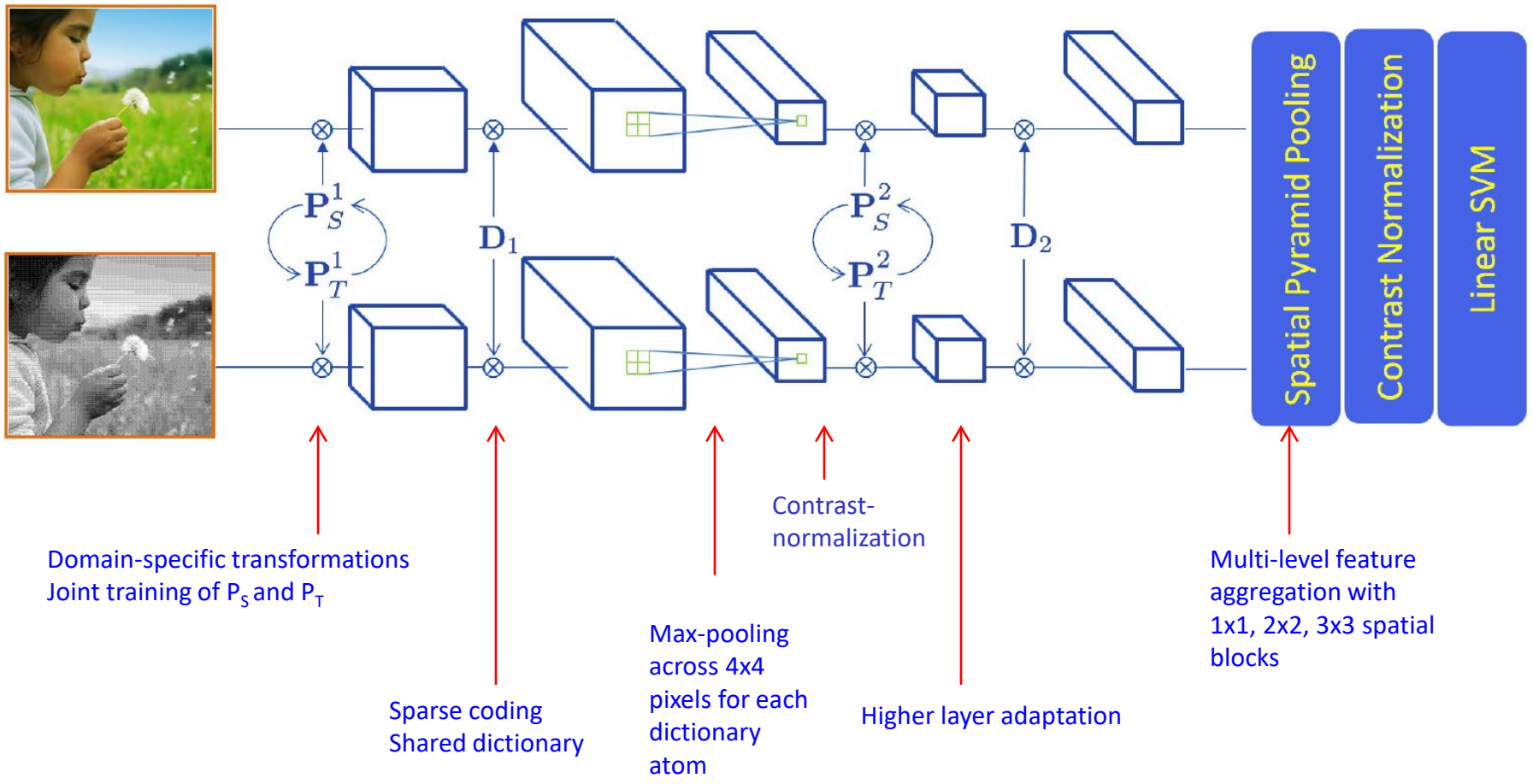
Infinite intermediate subspaces [2]



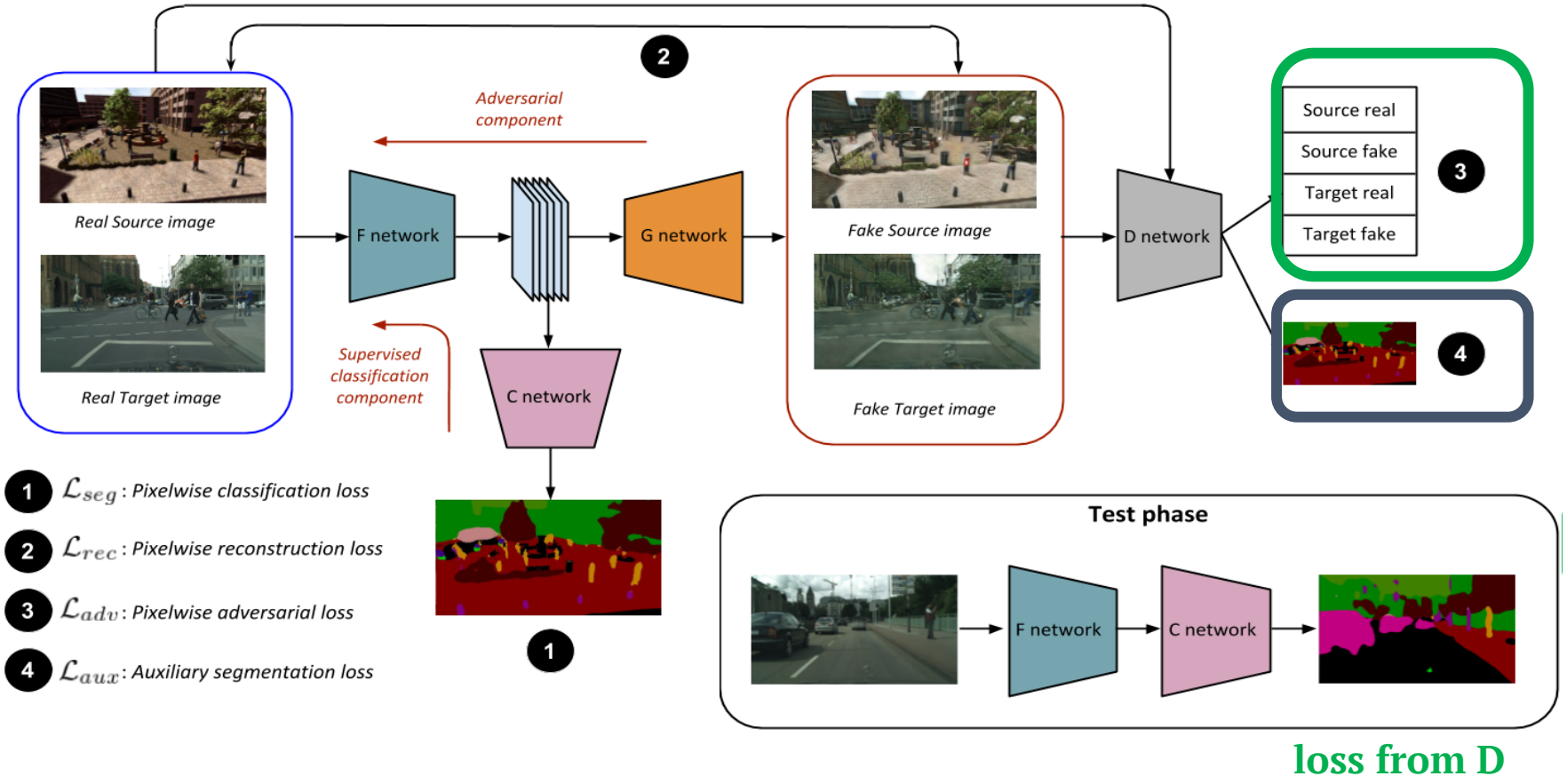
- Samples infinite intermediate subspaces
- Integrates the distance of sample projections along the geodesic

[1] R. Gopalan, R. Li, and R. Chellappa, "Domain Adaptation for Object Recognition: An Unsupervised Approach", ICCV, 2011, PAMI 2014
[2] Gong et al., Generalized Kernel flow, CVPR 2012

Unsupervised domain adaptation using hierarchical non-linear dictionaries



Synthetic to real domain adaptation for semantic segmentation

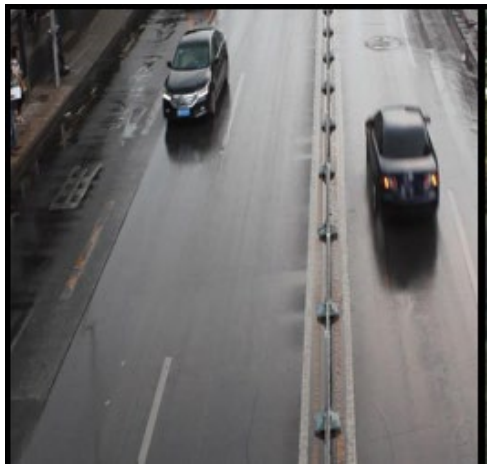


Domain adaptation on traffic camera data

- Motivation
  - Absence of annotations
  - Significantly lower resolution cameras
  - Suffering from artifacts such as video tear and compression artifacts
  - Taking advantage of annotated dataset
    - Adapt annotated datasets (source) to CATT data (target) using cycle consistency GANs
    - Train object detectors on the adapted source data
    - Deploy on the actual target data

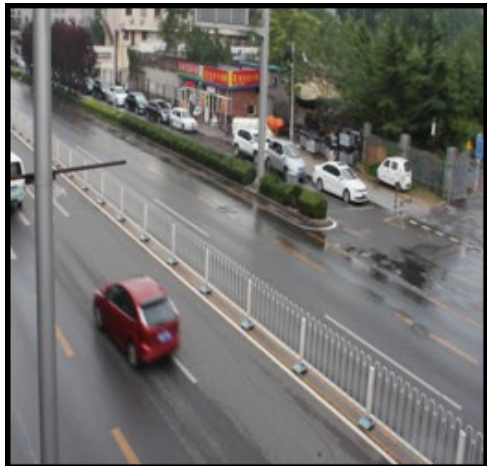
Qualitative results (UA-DETRAC to CATT v1)

Original Transferred





Original Transferred

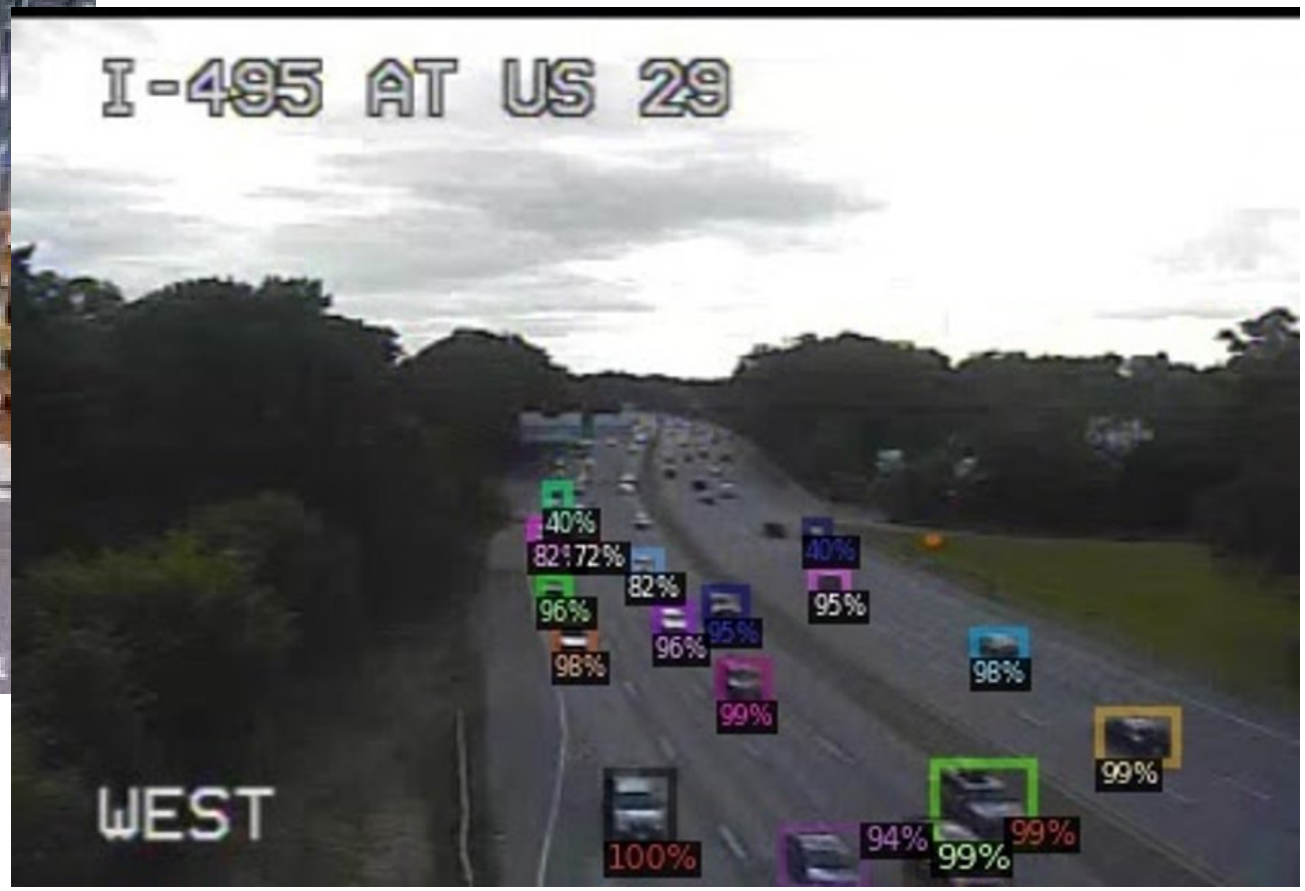




Quantitative results

	Model	mAP
Base Models	FasterRcnnR101-base	64.8
	RetinaNetR101-base	62.24
CityCam models	CityCam-FasterRcnnR101	73.92
	CityCam-RetinaNetR101	69.95
UADETRAC models	UA-DETRAC-FasterRcnnR101	37.86
	UA-DETRAC-RetinaNetR101	36.21
Domain Adapted UADETRAC models	DomainAdpt-FasterRcnnR101	61.49
	DomainAdpt-RetinaNetR101FPN	60.62
Domain Adapted + CityCam UADETRAC models	**DomainAdpt+CityCam-FasterRcnnR101**	**77.83**
	DomainAdpt+CityCam-RetinaNetR101	**75.84**

Detection results



City-scale multi-camera vehicle re-identification (Khorramshai, et al, ICCV 2019, ECCV 2020)

- Objective
  - Retrieve all images of a particular vehicle identity in a large gallery set, composed of vehicle images captured by a network of traffic cameras in different locations, time, weather condition and varying orientation.
- Challenges
  - Vehicles with different identities can be of same make, model, year and color.

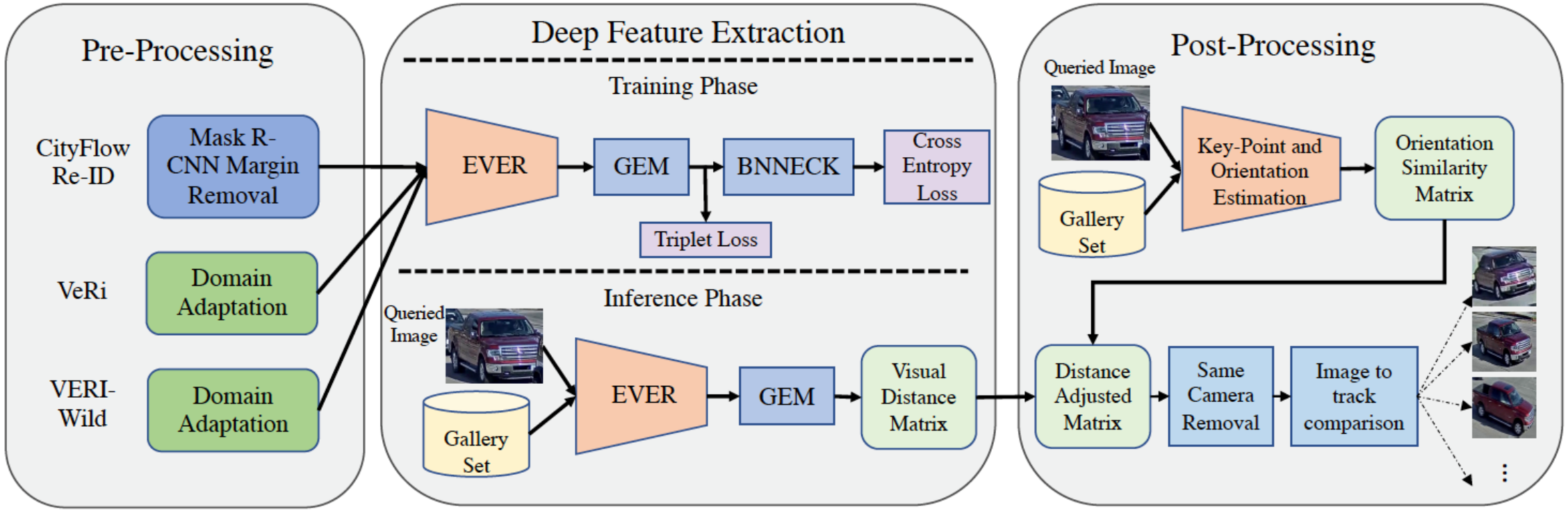



  - Vehicle's appearance and orientation can extremely vary from camera to camera



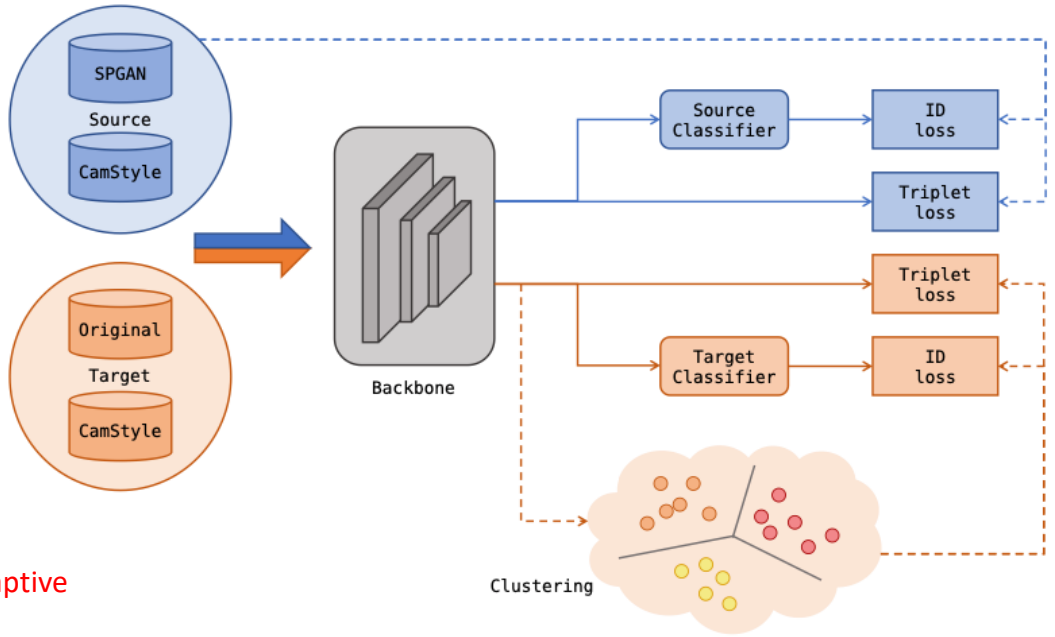

City scale multi-camera vehicle re-identification

Vehicle Re-Identification is the task of locating all instances of a particular vehicle identity in a gallery set consisting of a large volume of vehicle images which have been captured under diverse conditions using a network of traffic cameras.



Unsupervised domain adaptation training for Re-id

- Transfer knowledge from source domain to target domain
  - Train a Re-ID model on the source domain
  - Mine pseudo-labels from the target domain
    - Extract features from the target domain samples using the model trained on source domain
    - Use a clustering method to group extract features from target domain
    - Fine-tune the Re-ID model on the mined pseudo-labels
      - For a limited number of epochs
      - With a small learning rate



Gu, Jianyang, et al. "1st place solution to visda-2020: Bias elimination for domain adaptive pedestrian re-identification." *arXiv preprint arXiv:2012.13498* (2020).

City-scale multi-camera vehicle tracking and multi-camera vehicle re-identification

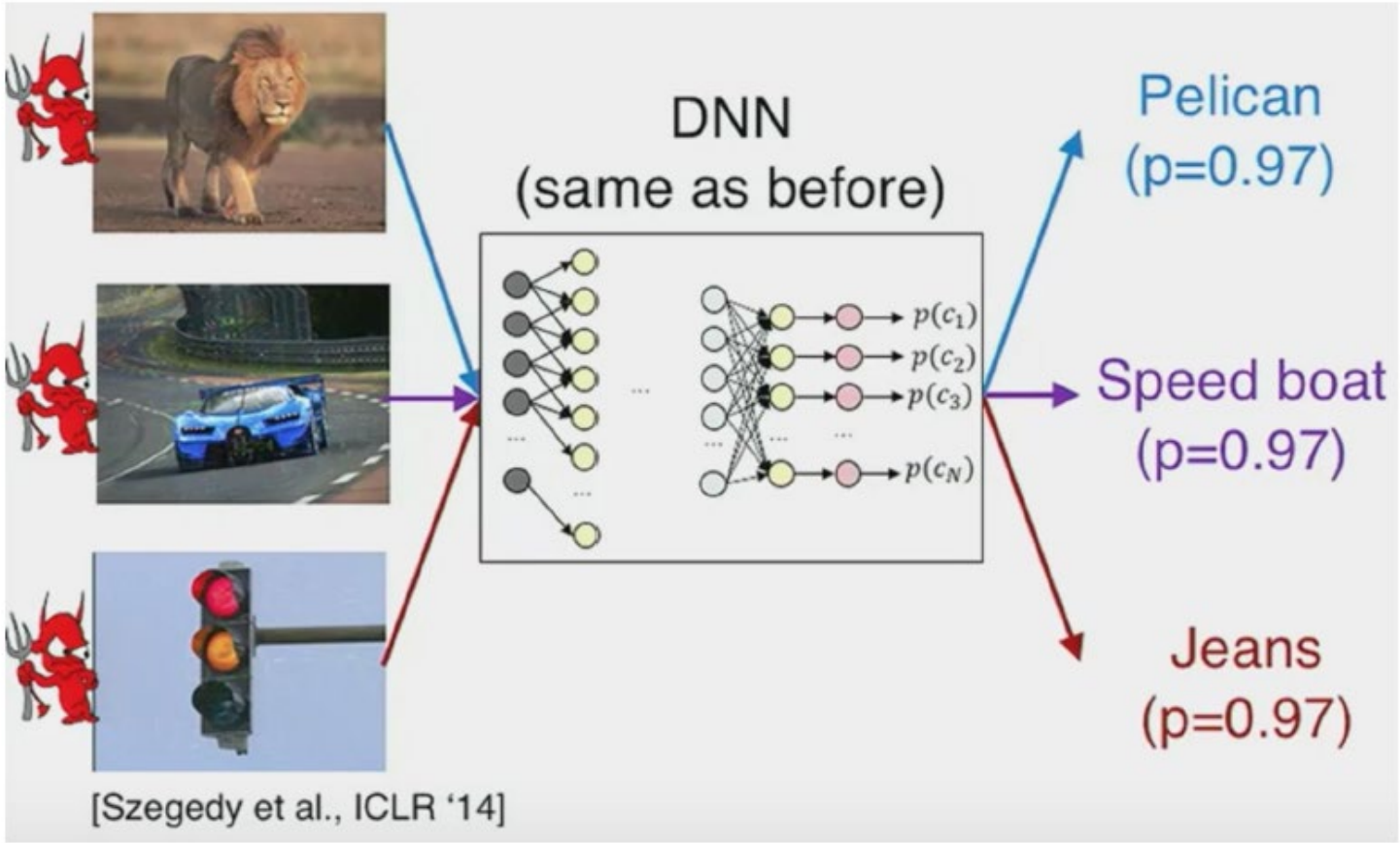
Rank	Team Name	Score (IDF1)
1	Cyber Tracking	0.8188
2	Mcmt	0.8095
3	Fivefive	0.7787
4	CyberHu	0.7651
5	**aiem2021 (Ours)**	**0.7189**
6	Track_mtmc	0.7061
7	FraunhoferIOSB	0.6910
8	Starwars	0.6575
9	Aiforward	0.5812
10	Janus Wars	0.5763

Rank	Team Name	Score (mAP)
1	DMT	0.7445
2	NewGeneration	0.7151
3	CyberHu	0.6650
4	For Azeroth	0.6555
5	IDo	0.6373
6	KeepMoving	0.6364
7	MegVideo	0.6252
8	aiem2021 (Ours)	0.6216
9	CyberCoreAI	0.6134
10	Janus Wars	0.6083

(Evaluation Metric: mean Average Precision of top 100 results)

Adversarial examples

- The classifier misclassifies adversarially manipulated images



Samagouei, et al, ICLR 2018, Lin, et al, NeurIPS, 2020

Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Attack types

- **White-box:** the adversary knows all the parameters of the model that is trained with loss $J(\mathbf{x}, y)$.
  - FGSM: Given an image $\mathbf{x}$ and its true class y, FGSM finds a perturbation $\boldsymbol{\delta}$ by moving the input pixels in the direction of the sign of the gradient components, by a fixed amount ϵ.

$$\boldsymbol{\delta} = \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}, y))$$

  - Rand+FGSM: First adds a random noise to the input image to make the attack stronger against defense methods that assume an FGSM attack:

$$\mathbf{x}' = \mathbf{x} + \alpha \cdot \text{sign}(\mathcal{N}(\mathbf{0}^n, \mathbf{I}^n)) \qquad \tilde{\mathbf{x}} = \mathbf{x}' + (\epsilon - \alpha) \cdot \text{sign}(\nabla_{\mathbf{x}'} J(\mathbf{x}', y))$$

  - CW: is an optimization-based technique to find the optimal perturbation.

$$\begin{aligned} \min_{\boldsymbol{\delta} \in \mathbb{R}^n} \quad & ||\boldsymbol{\delta}||_p + c \cdot f(\mathbf{x} + \boldsymbol{\delta}) \\ \text{s.t.} \quad & \mathbf{x} + \boldsymbol{\delta} \in [0,1]^n, \end{aligned}$$

    - f is an objective function that drives the example **x** to be misclassified

- **Black-box:** the adversary does not know the parameters of the model.
  - Train a substitute network that mimics the behavior of the target model, perform a white-box attack on the substitute model and transfer it to the target model.

Other attacks: Patch attacks on object detectors

Projected Gradient descent (PGD attack), the multi-step variant of FGSM attack

DARPA GARD program

Defending Object Detectors Against Patch Attacks

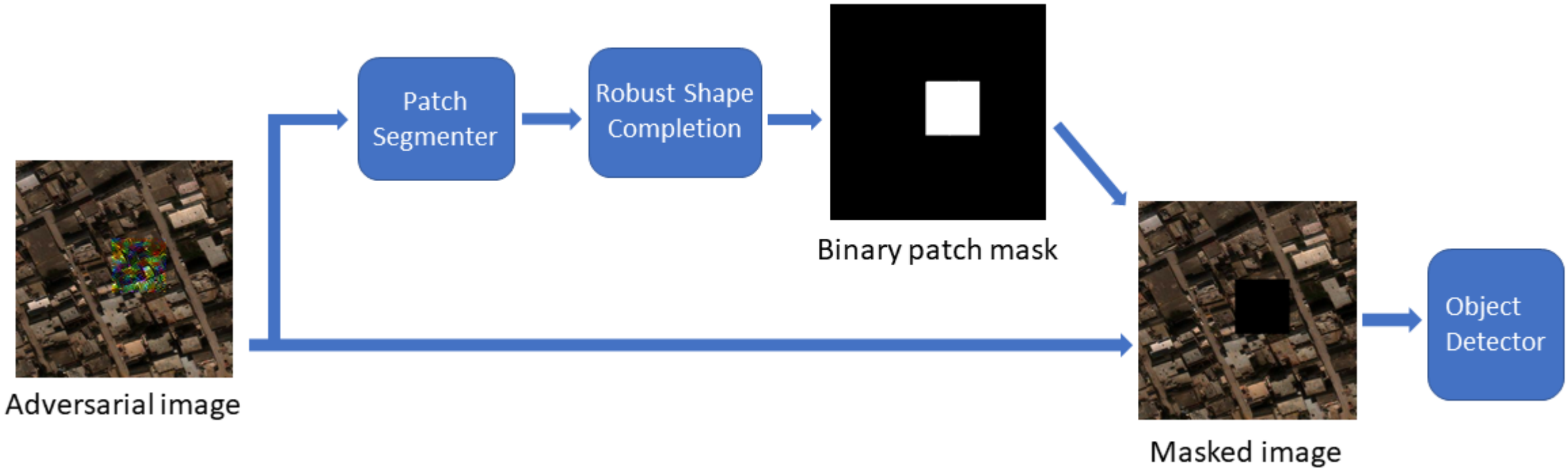
- Object detection plays a key role in many security critical systems
  - E.g., autonomous driving, security surveillance, identity verification, and robot manufacturing
- Adversarial patch attacks pose serious threats to real-world object detection systems
  - Attacker can arbitrarily distort pixels within a region of bounded size.
  - Easy to implement in the physical world.

(a) Predictions on clean image.

(b) Predictions on adversarial image.

Segment and complete (SAC) defense

- Adversarial patches are localized but affect predictions on whole images because object detectors utilize spatial context for reasoning.
- Our strategy: “detect and remove”
- SAC: a general, efficient, and effective defense for object detectors against patch attacks



Experiments and results

- Dataset: MS-COCO and xView
- Evaluation metrics: mean Average Precision (mAP) at Intersection over Union (IoU) 0.5
- SAC achieves superior robustness under attacks of various patch sizes without decreasing clean performance

Table 1: mAP (%) under non-adaptive attacks with different patch sizes.

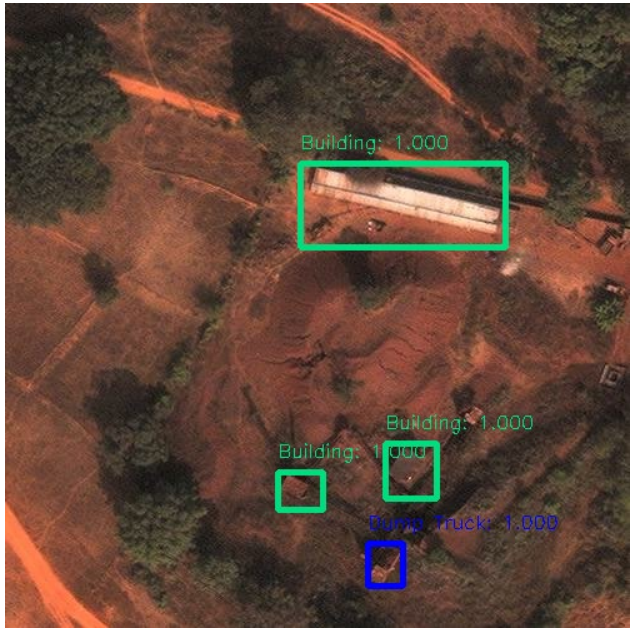
Dataset	Method	Clean	75×75	100×100	125×125
MS-COCO	Undefended	49.0	19.8±0.8	14.4±0.5	9.9±0.4
	LGS [29]	42.7	36.8±0.1	35.2±0.4	32.8±0.7
	SAC (**Ours**)	49.0	47.2±0.3	45.9±0.4	44.1±0.3
Dataset	Method	Clean	50×50	75×75	100×100
xView	Undefended	27.18	8.40±1.32	7.06±0.34	5.28± 0.86
	LGS [29]	19.08	11.86±0.42	10.86±0.26	9.77±0.37
	SAC (**Ours**)	27.18	25.80±0.34	23.75±1.24	23.21±0.04

(a) Ground-truth labels on clean image.

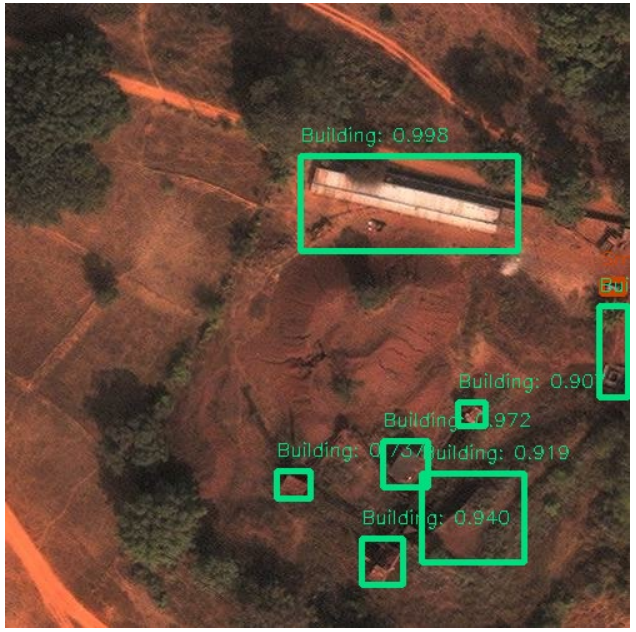
(b) Predictions on clean image.

(c) Predictions on adversarial image.

(d) Predictions on SAC masked image.

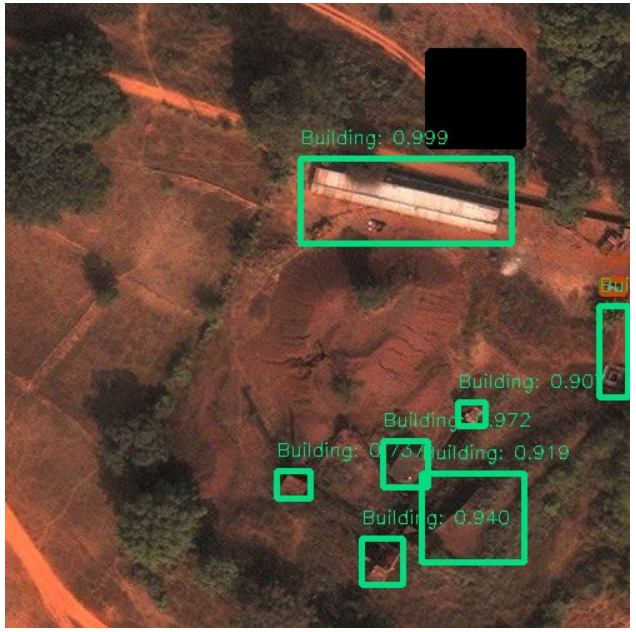


(a) Ground-truth labels on clean image.



(b) Predictions on clean image.

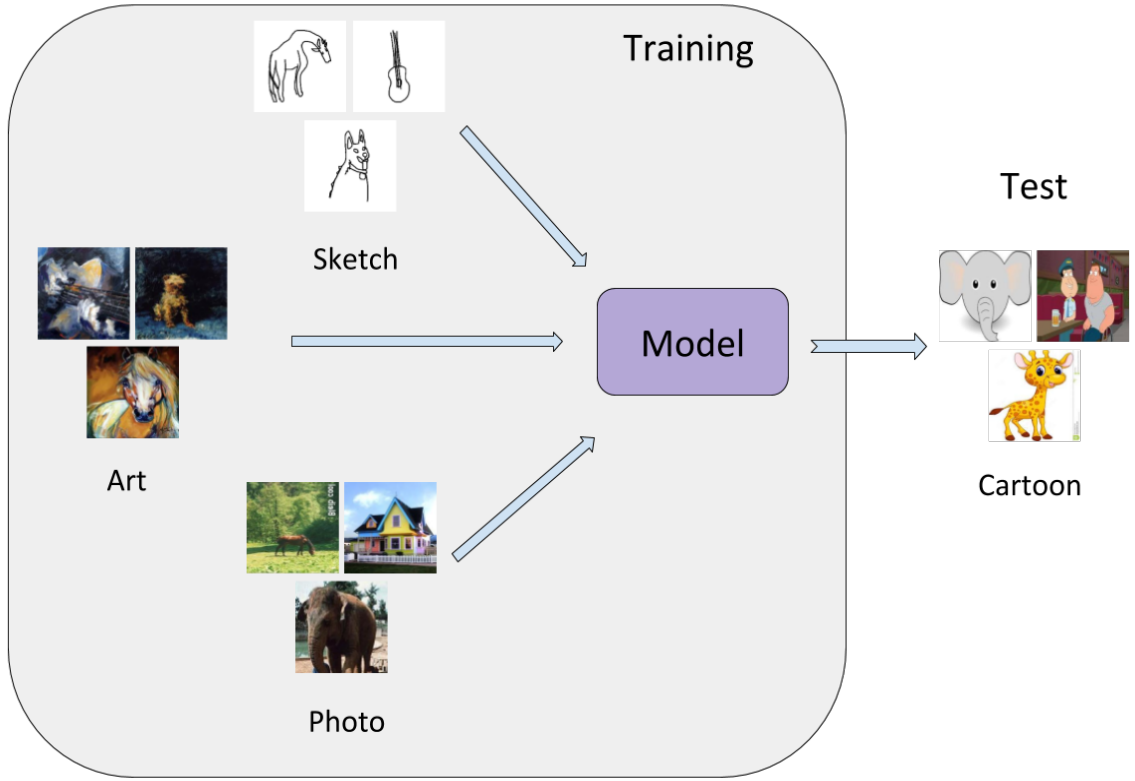
(c) Predictions on adversarial image.



(d) Predictions on SAC masked image.

Problem 1: Domain generalization

Domain generalization involves generalizing to novel test domains using variations in multiple source domains



Y. Balaji, S. Sankaranarayanan and R. Chellappa, "MetaReg: Towards Domain Generalization Using meta-regularization", Proc. Neural and Information Processing Systems, Montreal, Dec. 2018.

Problem 2: Model pruning/optimization

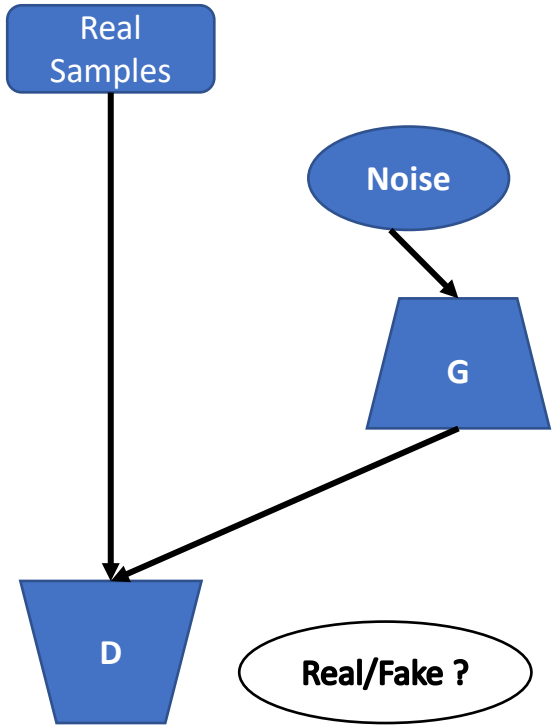
- Existing methods for model pruning/optimization are heuristic
- Is there a BIC for deep networks?
- Which parameters are statistically insignificant and what harm they would cause if removed?
- Learning from noisy data (errors-in-variable formulation)
- We need rigorous hypothesis testing procedures.
- Neural architecture search is an active area of research, but mostly adhoc.

Problem 3: Mathematical models for deep learning

- Five-year MURI supported by ONR
- Rich Baraniuk (PI), Moshe Vardi, Stan Osher, Ron DeVore, Ryan Tibshirani, Rob Nowak, Tom Goldstein and Rama Chellappa
- Research problems
  - Mathematical analysis approach to explain DN representations and interpret DN functionality
  - Function classes that are optimally approximated by DNs, and quantifying the rate of approximation error decay in terms of DN architecture
  - Mathematically characterize the class of functions learned via overparameterized and data-interpolating DNs, as well as the limitations of such models
  - Model the interplay between the choice of DN architecture and training algorithms
  - Mathematical interpretation of the implicit regularization properties that underlie popular training techniques like early stopping, gradient sampling, and preconditioning
  - Training DNs with fewer, but selectively chosen, training examples
  - statistical methods to rigorously quantify the uncertainty in network outputs, to identify the input features and network structures that resulted in a network output and estimate prediction intervals to test the significance and confidence of a DN's decisions
  - Understanding the behaviors of deep networks using scalable and parallelizable formal methods

Problem 4: Analysis of generative adversarial networks

1. Traditional GAN architecture involves a Generator (G) and Discriminator (D) pair.
2. Both are modeled as deep networks (since DCGAN)
3. Optimize D to identify the generator's fakes compared to real samples
4. Optimize G to fool the discriminator in thinking that G produced a real sample
5. Min-max adversarial game between G and D



$$\min_G \max_D \mathbf{E}_{x \sim p_{data}}(\log(D(x)) + \mathbf{E}_{z \sim p_{noise}} \log(1 - D(G(z)))$$

Goodfellow 2014

Problem 5: Choosing the best subsets for training from a much larger pool of training data

- Setup
  - Given a **fixed** classifier architecture
  - A set of **labeled** training data points from ***L*** different classes
- Objective
  - Iterative algorithm
  - At teach time instance *t*, select a subset of the training data to resume training on
- Selection criteria
  - The samples in the selected batch must be such that the classifier is *uncertain* about classifying them (or *certain* but *wrong* in its classification)
  - The batch must have a *balanced* selection from all classes
  - The batch should be sufficiently *diverse*.
  - The batch should be *representative* of the training samples.

Problem 6: Prediction of critical events from heterogeneous data

- Inputs: a) Clinical, claims and specialized JHM research data on the particular patient; b) Similar data from prior system (consortium) experience projected from PMAP onto a clinical cohort database (registry); c) Outputs of video and speech processing algorithms; and d) Expert knowledge about the etiology of the health or disease condition.
- Outputs: (1) the prediction, prevention, monitoring, and intervention of frailty and dementia, (2) the definition, measurement, and promotion of physical, physiological, and psychological well-being, and (3) the identification of robust signals, biomarkers, and processes of frailty and dementia.
- Bayesian hierarchical models (Zeger, Nishumura)
- More needs to be done!

Schematic of the information flow within the JHU Precision Medicine Analytics Platform (PMAP). The AI suite will analyze the integrated data and return the results to the clinician and patient to improve their interactive and collaborative shared decision making.

Looking ahead - 1

- AI is here to stay
  - The definition of AI is broad and all encompassing.
- At scale problems will reveal the warts in AI!
- If AI does not adapt, it is not learning.
- Designers should consider
  - Bias or perceived lack of fairness
    - More than 20 ways bias can be introduced; there are more than 10 metrics for fairness.
    - Task dependent.
    - Bias vs performance tradeoff
  - Domain adaptation/generalization
  - Robustness to adversarial attacks
  - Move from black-box decision making (interpretability)
- Human-centric AI
- Synthetic AI –AI via imagination
  - “What is now proved was once only imagined”. William Blake.
- Rigorous math will be a good medicine for the alchemy of deep learning.

Looking ahead -2

- AI's impact on medicine and healthcare
- Johns Hopkins Artificial Intelligence and Technology Collaboratory for Aging Research – a five-year, $20 million effort
  - Peter Abadir, Rama Chellappa, Greg Hager and Jeremy Walston
- Almost weekly conversations with pathologists, endocrinologists, eye doctors, …
- AI's role in prediction, prevention and diagnosis of deceases will fundamentally change how medicine will be practiced and care will be delivered.