Facial Expression Recognition and Editing with Limited Data

Hui Ding
Department of Electrical and Computer Engineering
University of Maryland, College Park

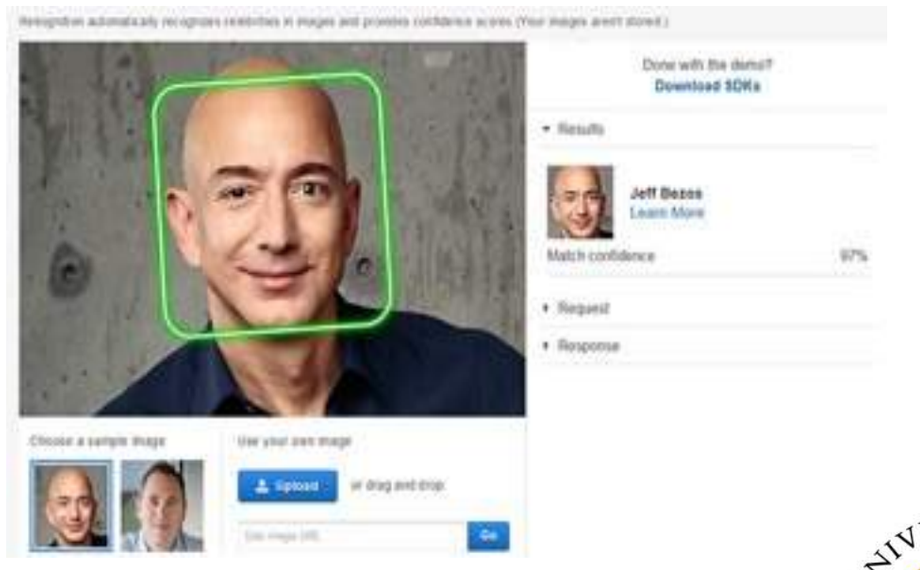
Advisory Committee:
Professor Rama Chellappa
Professor Gang Qu
Professor Min Wu
Professor Behtash Babadi
Professor Ramani Duraiswami

Deep Learning is Changing our Lives

Self-Driving Car

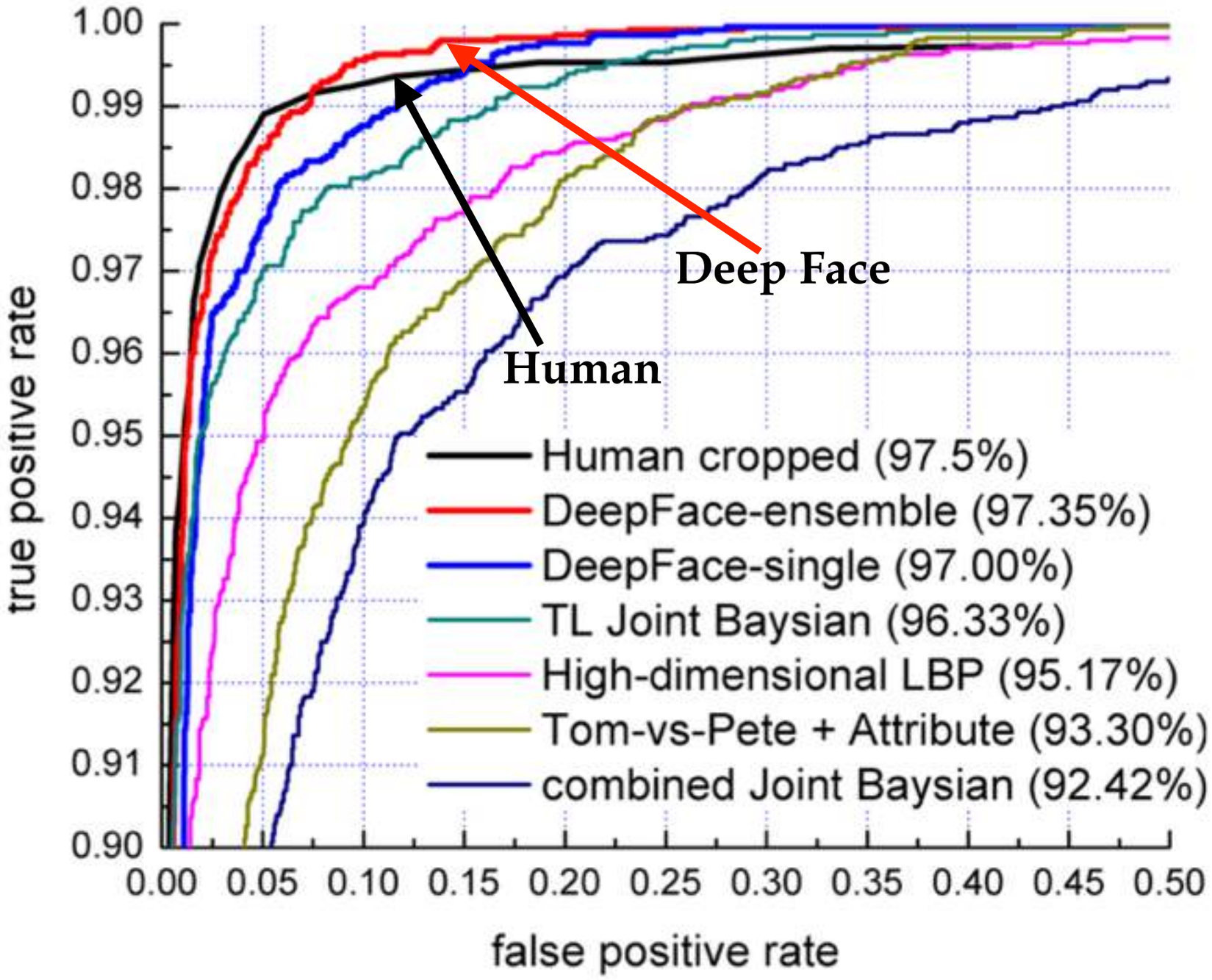
AlphaGo

Machine Translation



Face Recognition

Deep Face Recognition is Successful



Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." CVPR, 2014.

Deep Facial Expression Recognition is Relatively Unexplored

The First Challenge: Small Datasets

Hard to train an accurate expression classifier

Face Datasets	# Images
CASIA-WebFace	494,414
VGG Face	2,600,000
Facebook	**4,400,000**
MS-Celeb-1M	10,000,000

Expr. Datasets	# Images
CK+	1,308
OULU-CASIA	1,440
TFD	4,178
SFEW	1,322

The Second Challenge: In-the-wild Conditions

Occlusion and pose decrease the model performance greatly

The Third Challenge: No Fine-Grained Dataset

Hard to collect expression datasets with fine-grained labels

Agenda

- ✦ Transfer Learning (Small Datasets)
  - FaceNet2ExpNet
- ✦ Robust Model Design (Occlusion, Pose)
  - Occlusion Robust Deep Network
  - Unaligned Attribute Classifier
- ✦ Generative Model (Fine-Grained)
  - ExprGAN

Agenda

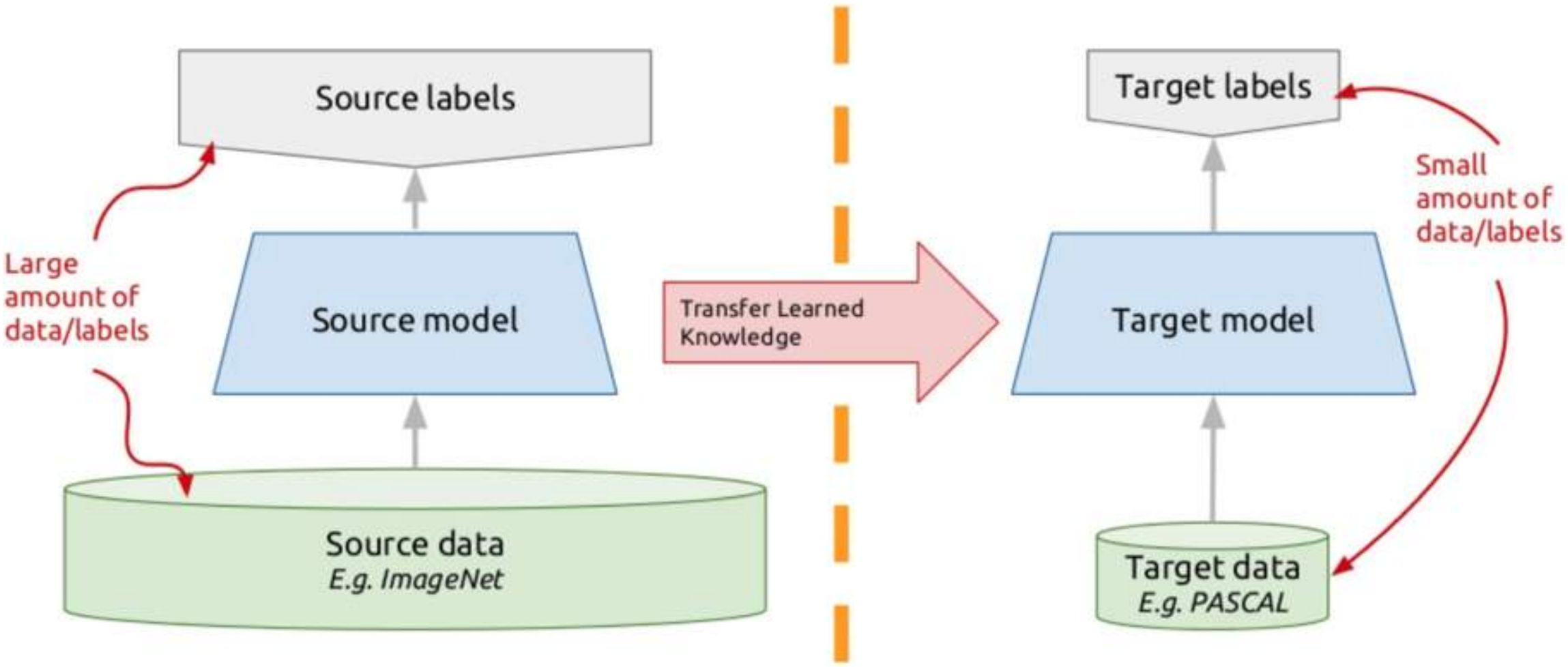
- ✦ Transfer Learning (Small Datasets)
  - FaceNet2ExpNet
- ✦ Robust Model Design (Occlusion, Pose)
  - Occlusion Robust Deep Network
  - Unaligned Attribute Classifier
- ✦ Generative Model (Fine-Grained)
  - ExprGAN

How to train an accurate expression classifier for small datasets?

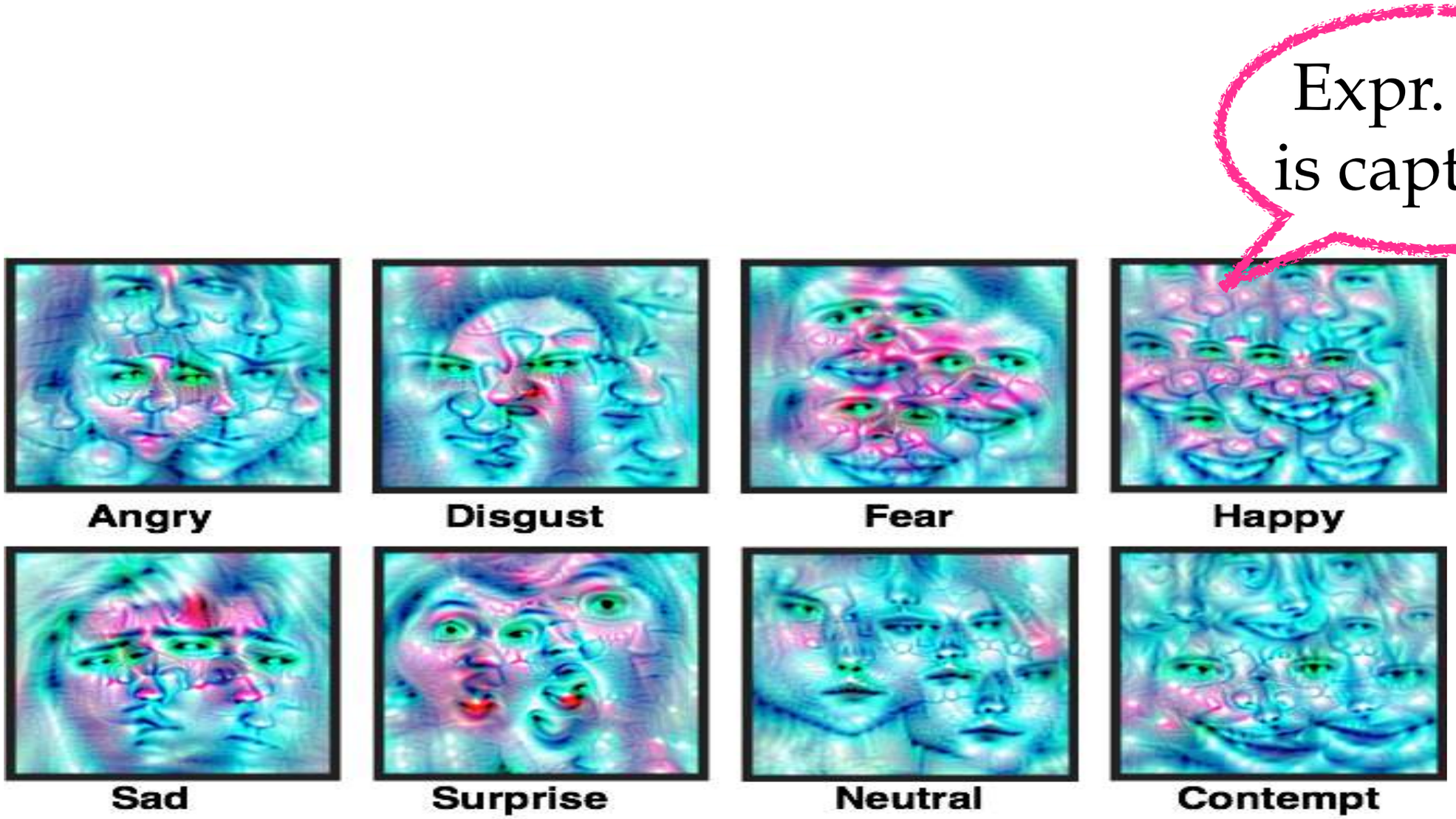
FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition

Hui Ding, Shaohua Kevin Zhou and Rama Chellappa, IEEE International Conference on Automatic Face Gesture Recognition (FG), 2017.

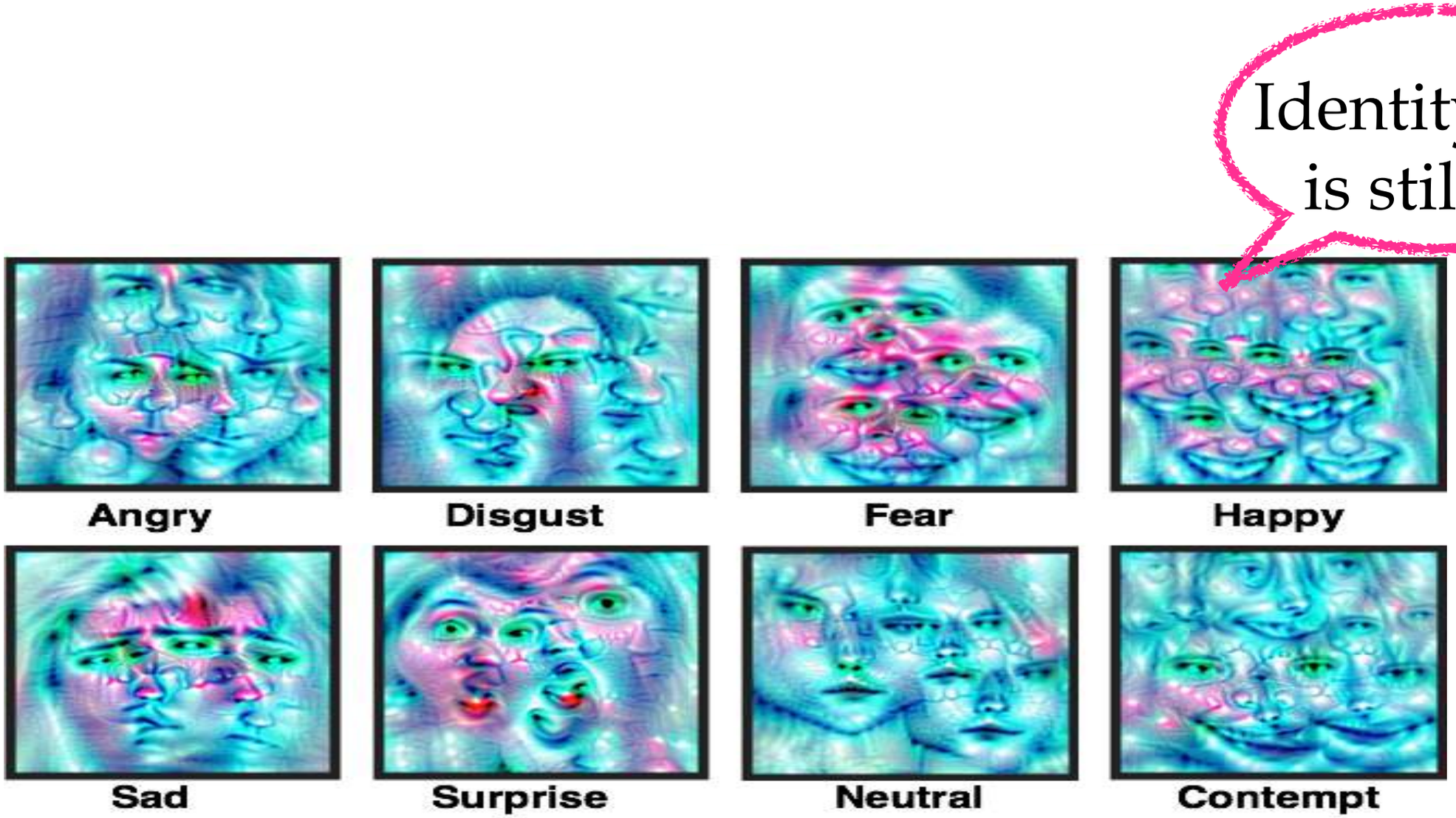
Conventional Transfer Learning



Feature Visualization



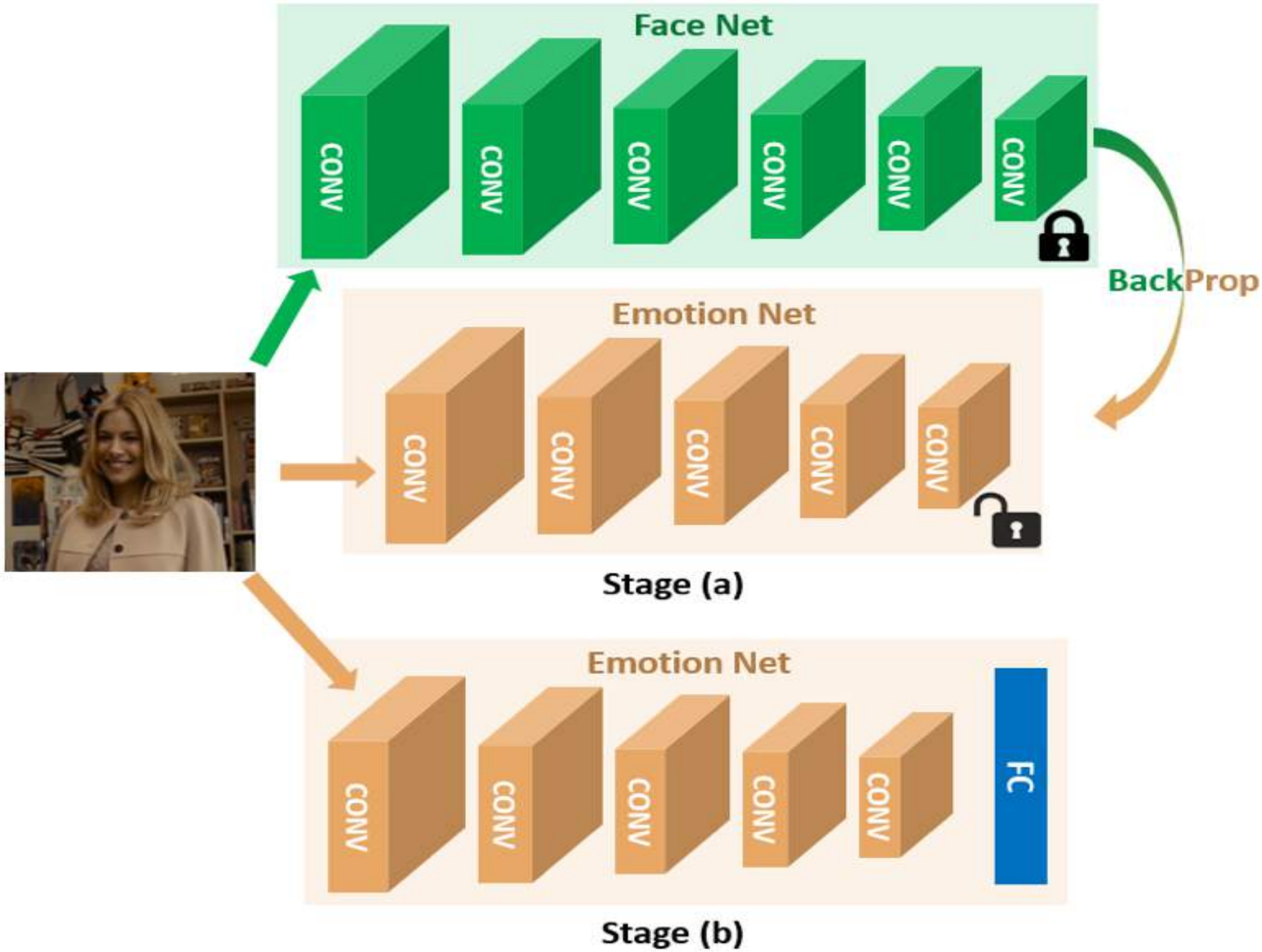
Feature Visualization



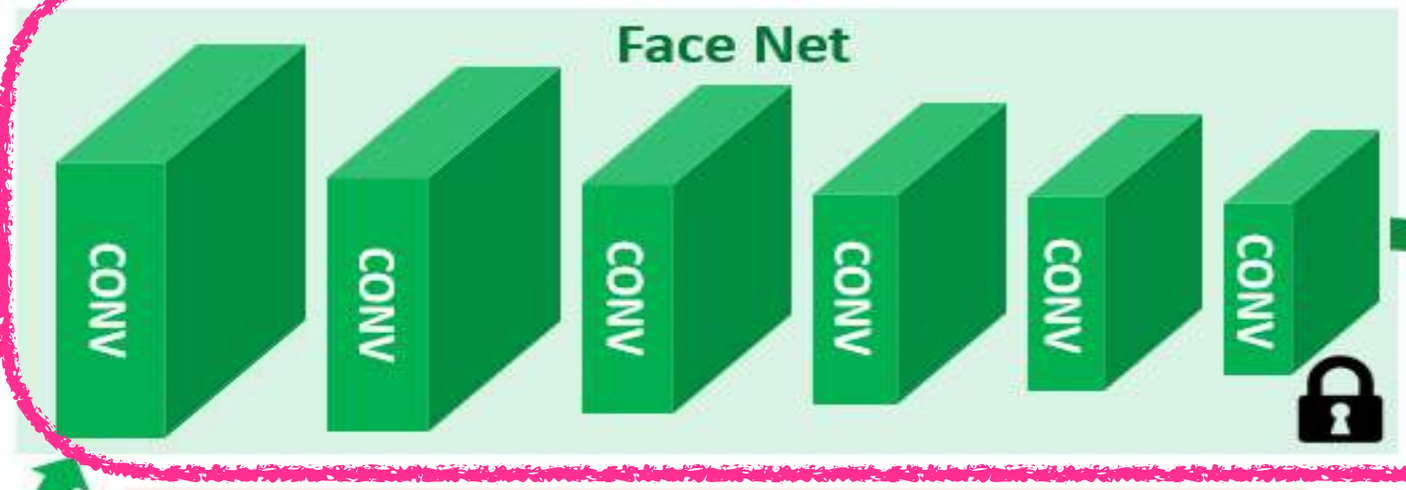
Motivation

Can we utilize the face recognition network to help the training of the expression recognition network without the redundant identity information?

FaceNet2ExpNet



1.Fine Tune Face Net



BackProp

Stage (a)

Stage (b)

Face Net

CONV CONV CONV CONV CONV CONV

BackProp

Emotion Net

CONV CONV CONV CONV CONV

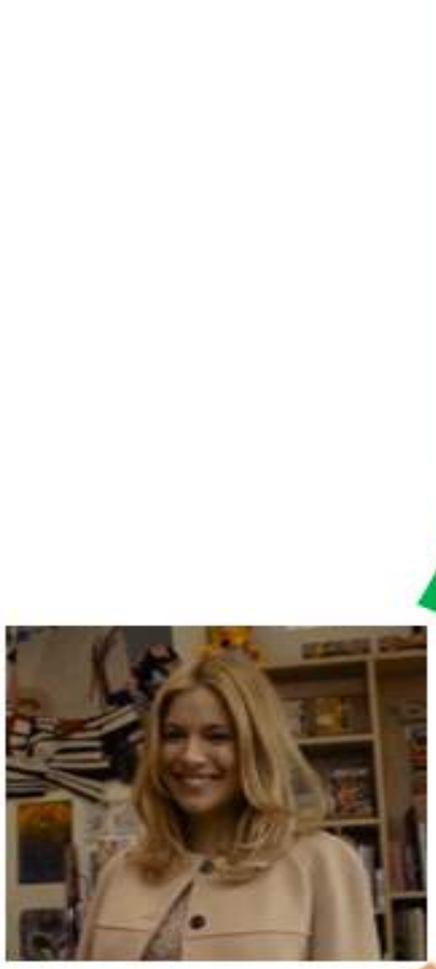
Stage (a)

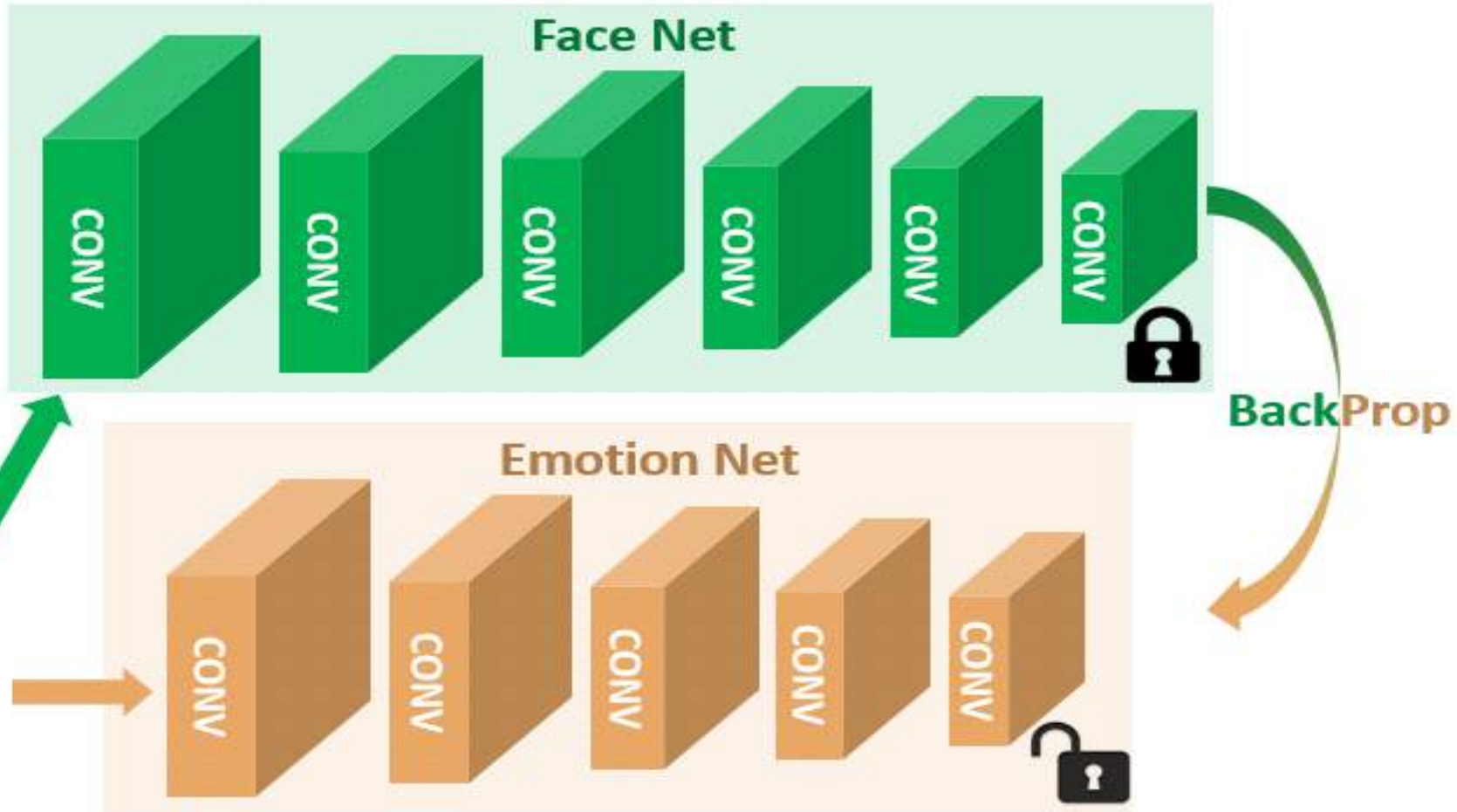
2.Train Conv. Layers by Matching Feature Maps

Emotion Net

CONV CONV CONV CONV CONV FC

Stage (b)



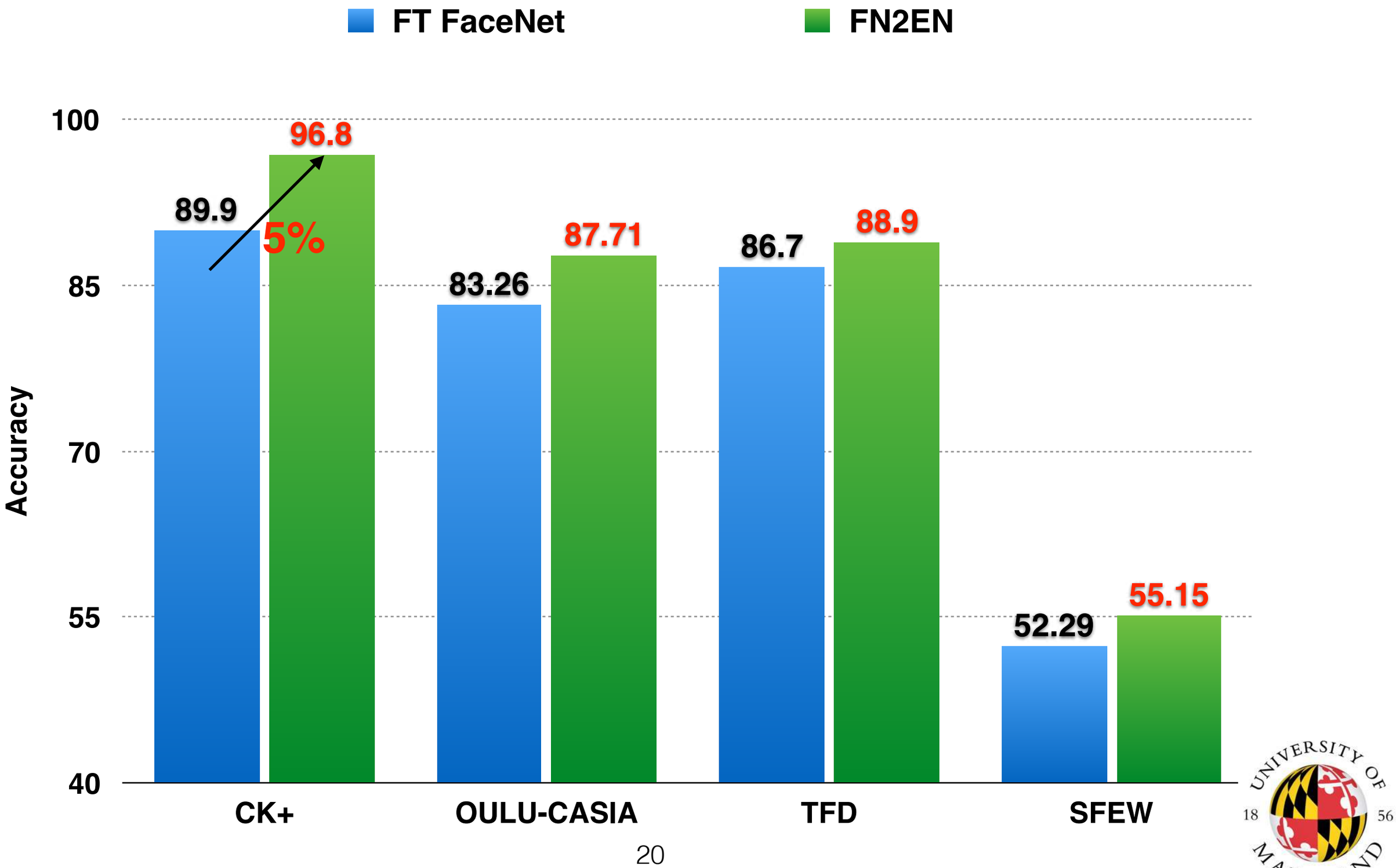


Stage (a)

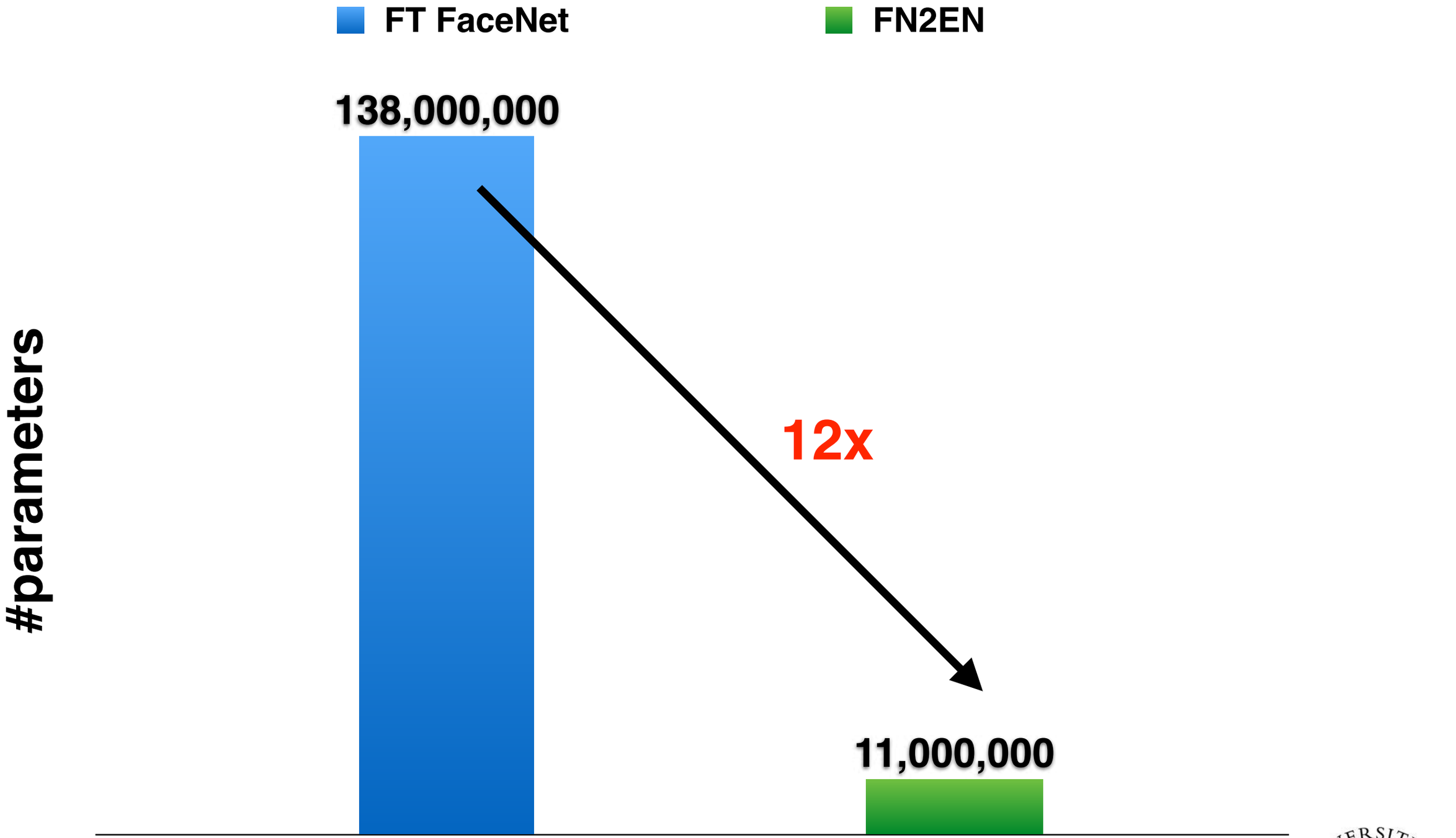
Stage (b)

3. Jointly Train Conv. and FC Layers

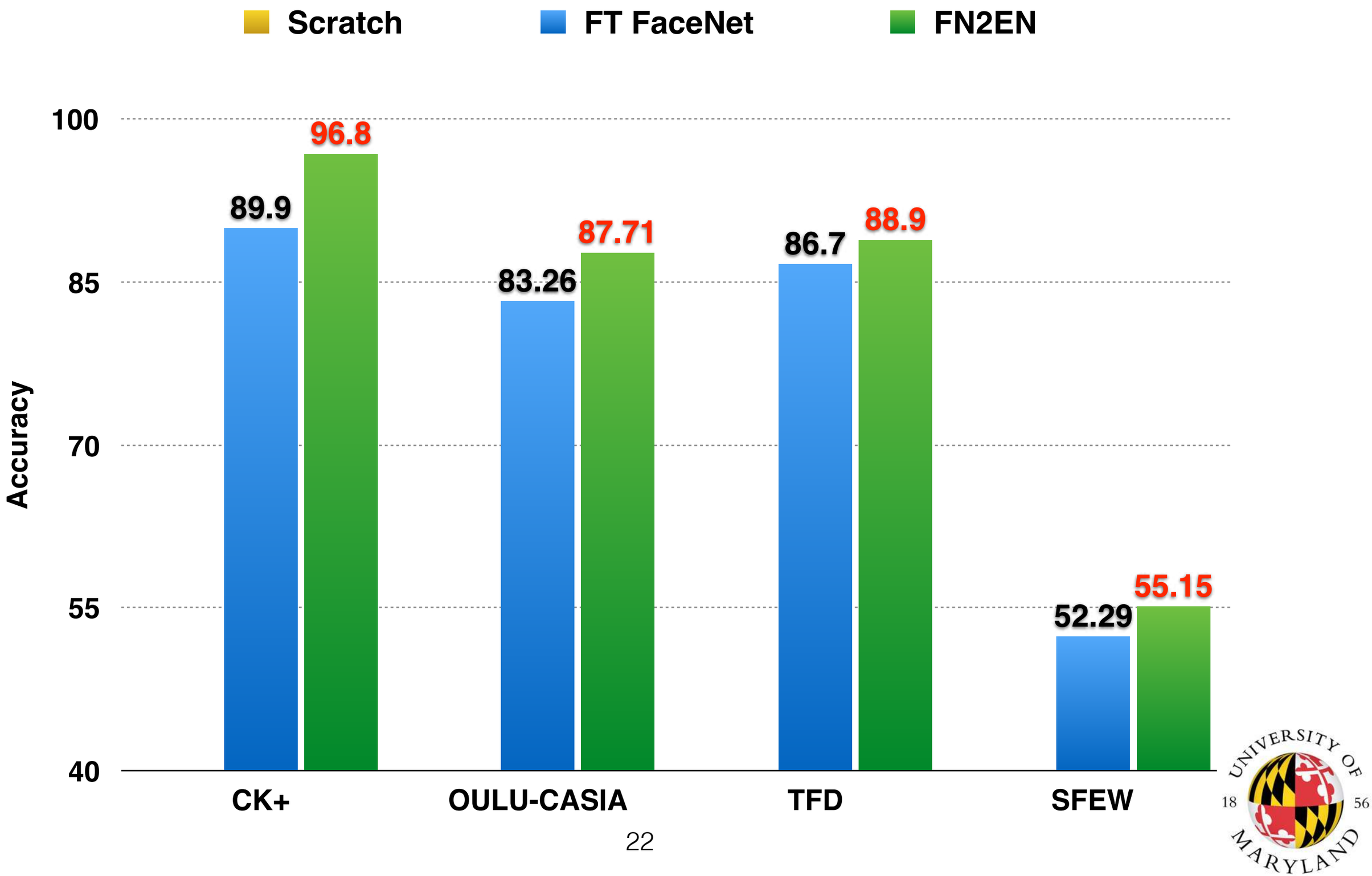
Recognition Accuracy Comparison



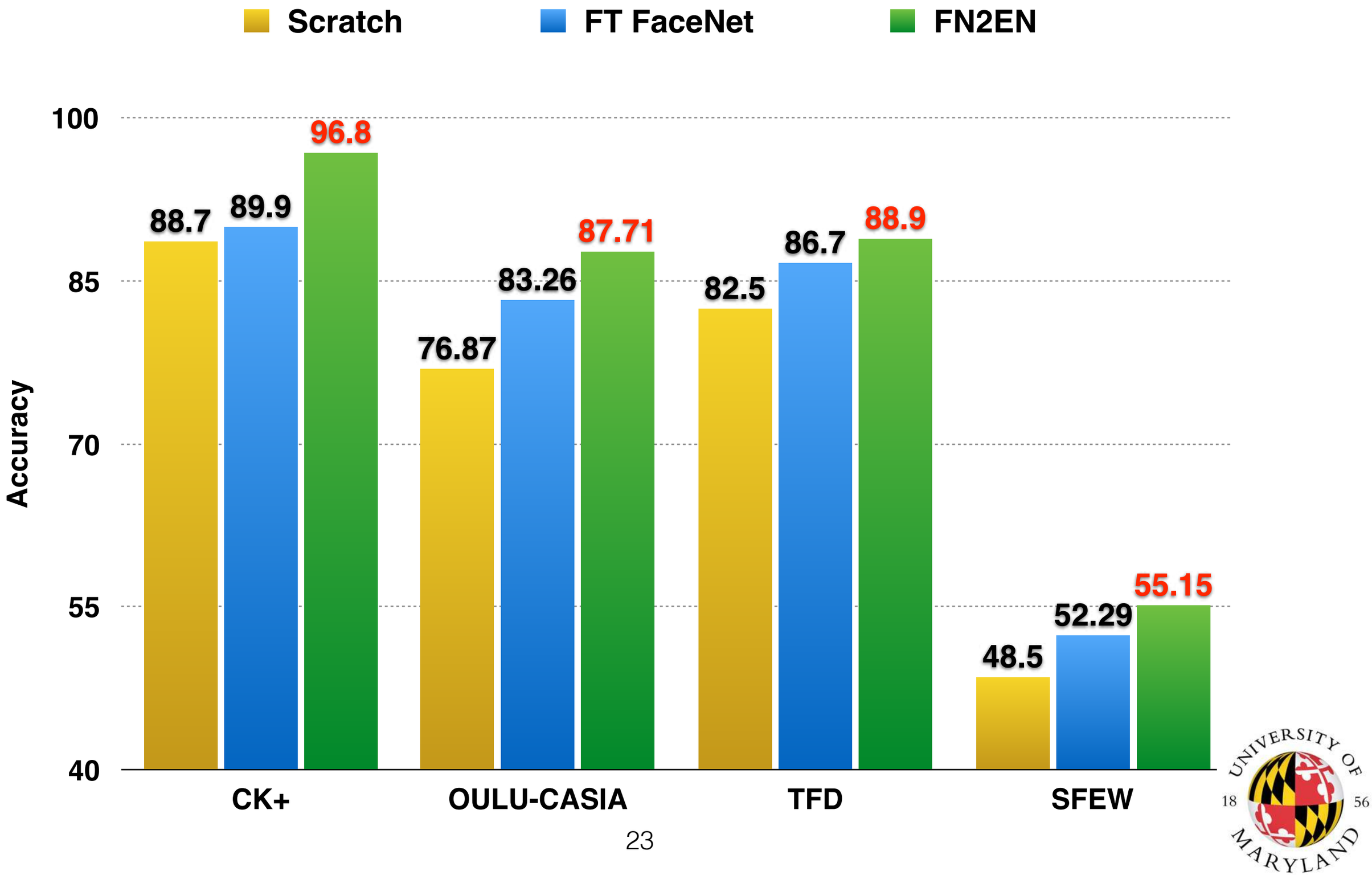
Model Size Comparison



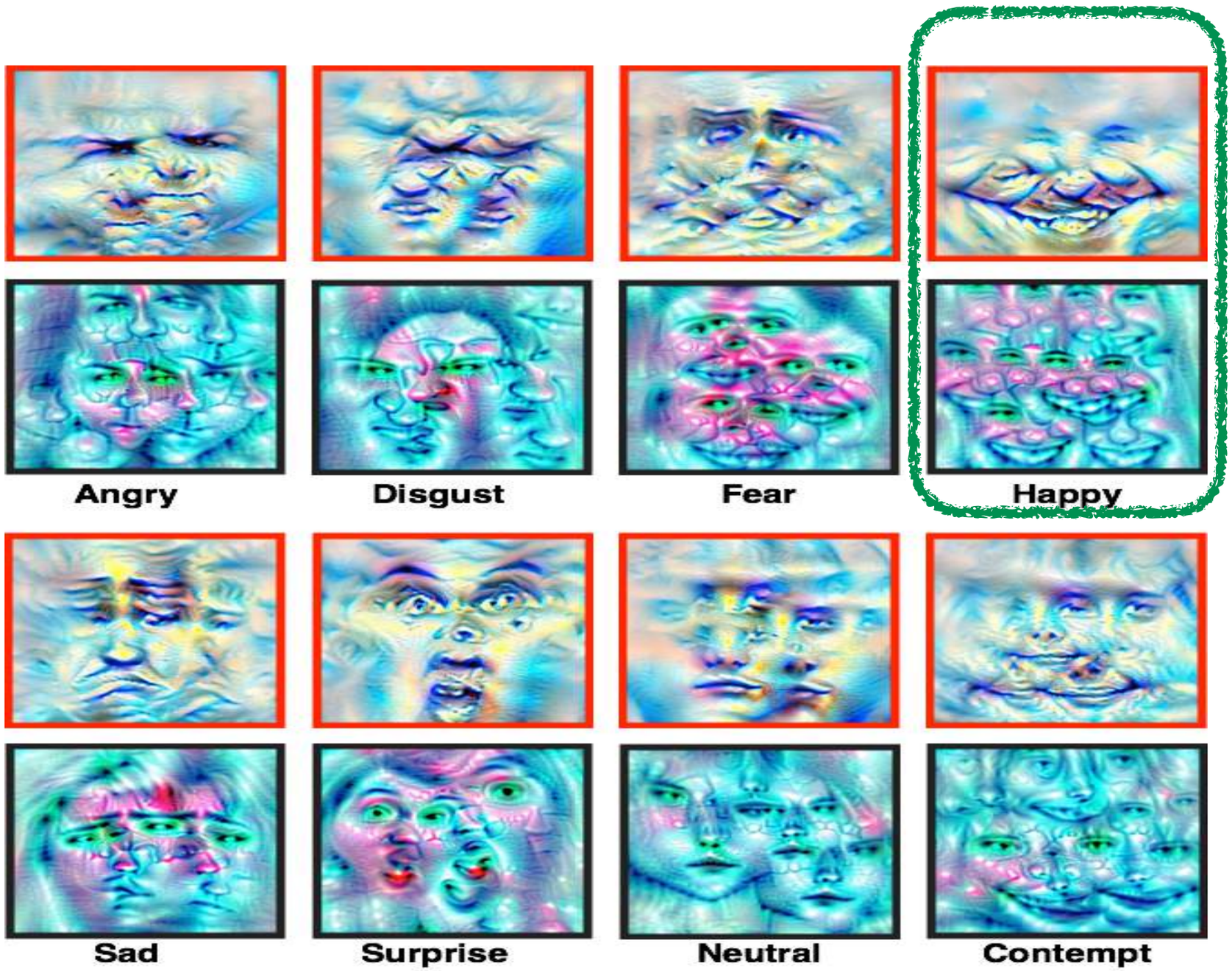
Recognition Accuracy Comparison



Recognition Accuracy Comparison



Feature Visualization



FaceNet2ExpNet

Fine-tune FaceNet

FaceNet2ExpNet

Fine-tune FaceNet

Classification Accuracy
CK+

Method	Average Accuracy	#Exp. Classes
CSPL [40]	89.9%	Six Classes
AdaGabor [53]	93.3%	
LBPSVM [54]	95.1%	
3DCNN-DAP [43]	92.4%	
BDBN [1]	96.7%	
STM-ExpLet [2]	94.2%	
DTAGN [3]	97.3%	
Inception [4]	93.2%	
LOMo [55]	95.1%	
PPDN [7]	97.3%	
FN2EN	**98.6%**	
AUDN [42]	92.1%	Eight Classes
Train From Scratch (BN)	88.7%	
VGG Fine-Tune (baseline)	89.9%	
FN2EN	**96.8%**	

Classification Accuracy OULU-CASIA

Method	Average Accuracy
HOG 3D [56]	70.63%
AdaLBP [46]	73.54%
Atlases [57]	75.52%
STM-ExpLet [2]	74.59%
DTAGN [3]	81.46%
LOMo [55]	82.10%
PPDN [7]	84.59%
Train From Scratch (BN)	76.87%
VGG Fine-Tune (baseline)	83.26%
FN2EN	**87.71%**

Classification Accuracy
TFD

Method	Average Accuracy
Gabor + PCA [58]	80.2%
Deep mPoT [59]	82.4%
CDA+CCA [60]	85.0%
disRBM [18]	85.4%
bootstrap-recon [61]	86.8%
Train From Scratch (BN)	82.5%
VGG Fine-Tune (baseline)	86.7%
FN2EN	**88.9%**

Classification Accuracy
SFEW

Method	Average Accuracy	Extra Train Data
AUDN [42]	26.14%	None
STM-ExpLet [2]	31.73%	
Inception [4]	47.70%	
Mapped LBP [13]	41.92%	
Train From Scratch (BN)	39.55%	
VGG Fine-Tune (baseline)	41.23%	
FN2EN	**48.19%**	
Transfer Learning [6]	48.50%	FER2013
Multiple Deep Network [5]	52.29%	
FN2EN	**55.15%**	

Expression Recognition for Frontal Faces

CK+

FaceNet2ExpNet

OULU CASIA

Expression Recognition for In-the-wild Faces


RAF


AffectNet

Li et al. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." CVPR. 2017.

Mollahosseini et al. "Affectnet: A database for facial expression, valence, and arousal computing in the wild." IEEE Transactions on Affective Computing. 2017.

Agenda

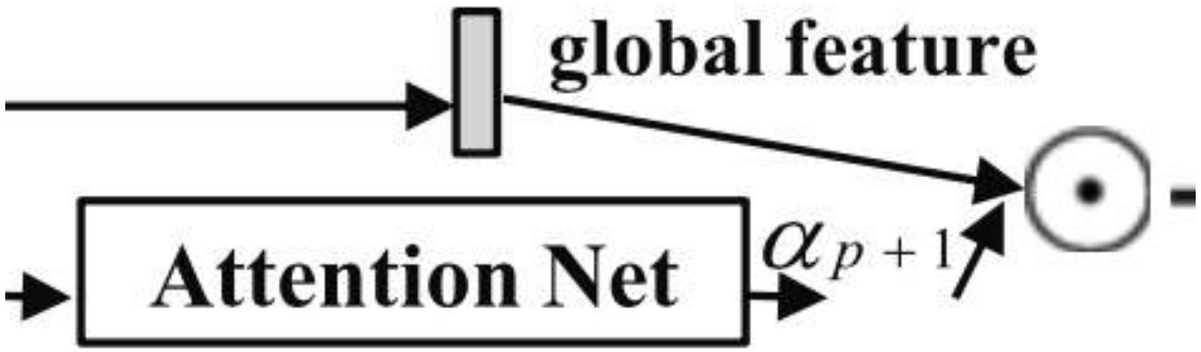
- ✦ Transfer Learning (Small Datasets)
  - FaceNet2ExpNet
- ✦ Robust Model Design (Occlusion, Pose)
  - Occlusion Robust Deep Network
  - Unaligned Attribute Classifier
- ✦ Generative Model (Fine-Grained)
  - ExprGAN

Occlusion Adaptive Deep Network for Robust Facial Expression Recognition

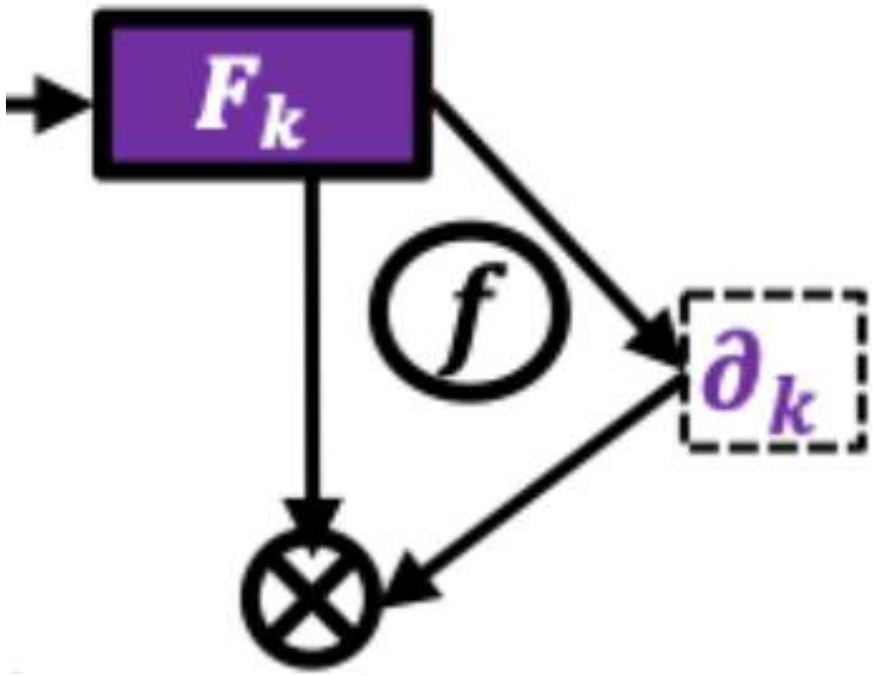
Hui Ding, Peng Zhou, and Rama Chellappa, Submitted to IJCB 2020

Related Works

Common goal: learn features from non-occluded regions

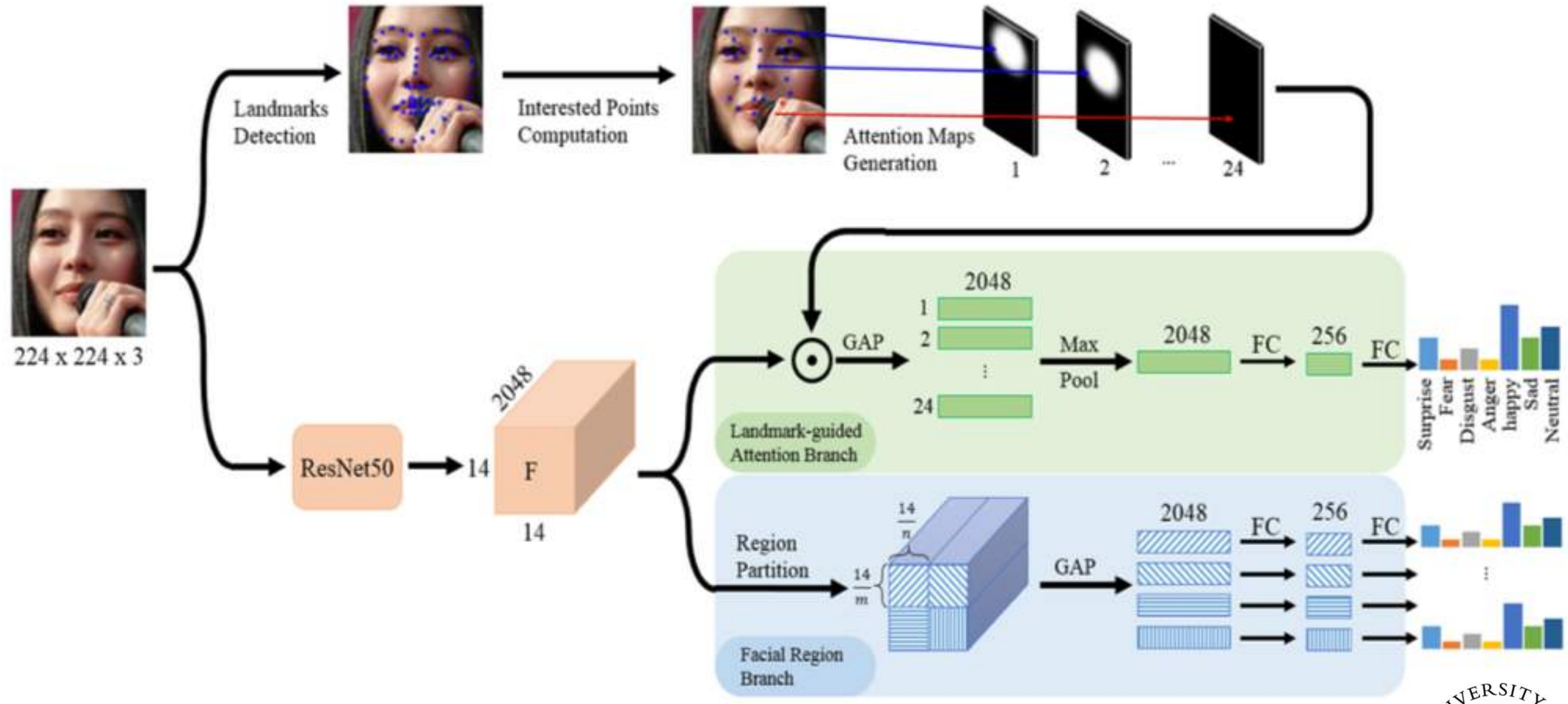


gACNN (2019)

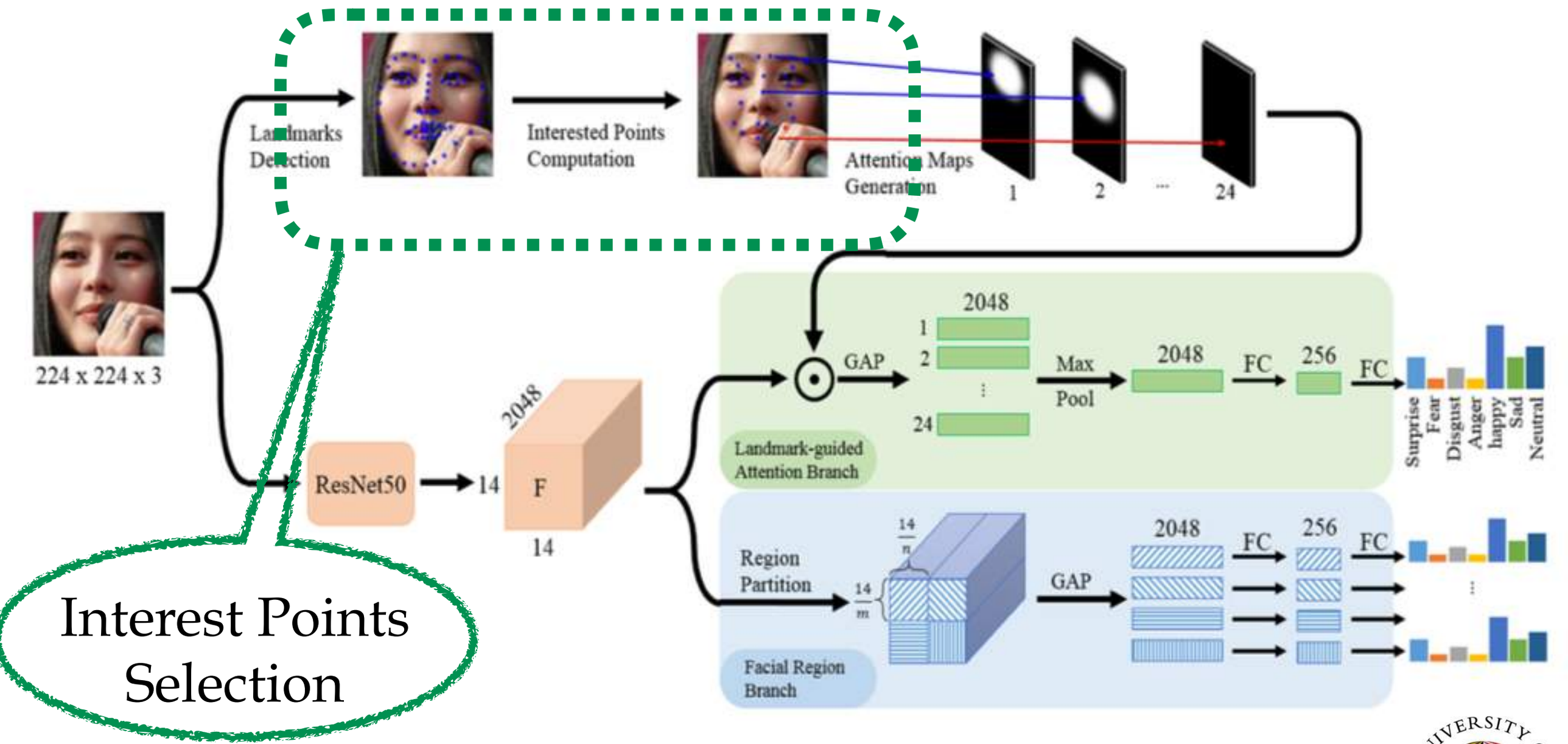


RAN (2020)

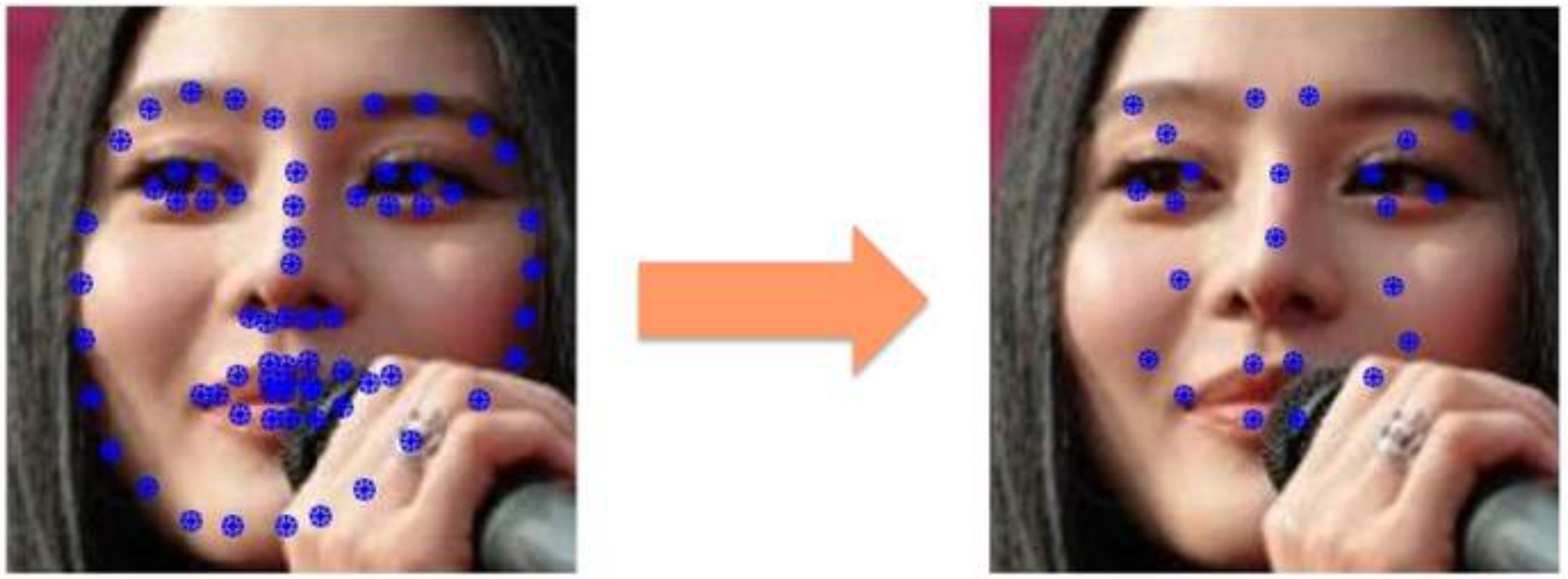
Occlusion Adaptive Deep Network (OADN)



Landmark-guided Attention Branch (LAB)



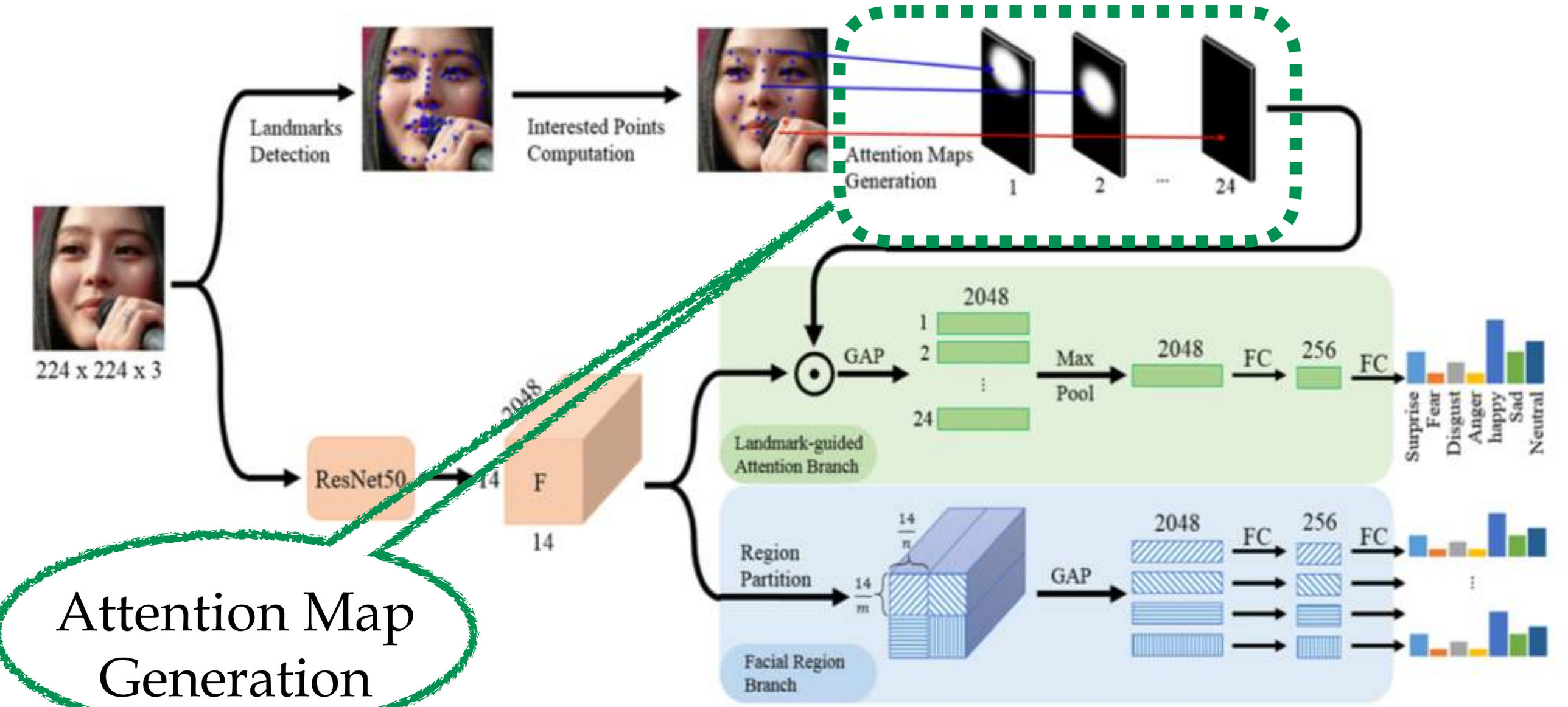
Interest Points Selection



(a) Original 68 detected landmarks

(b) Recomputed 24 points

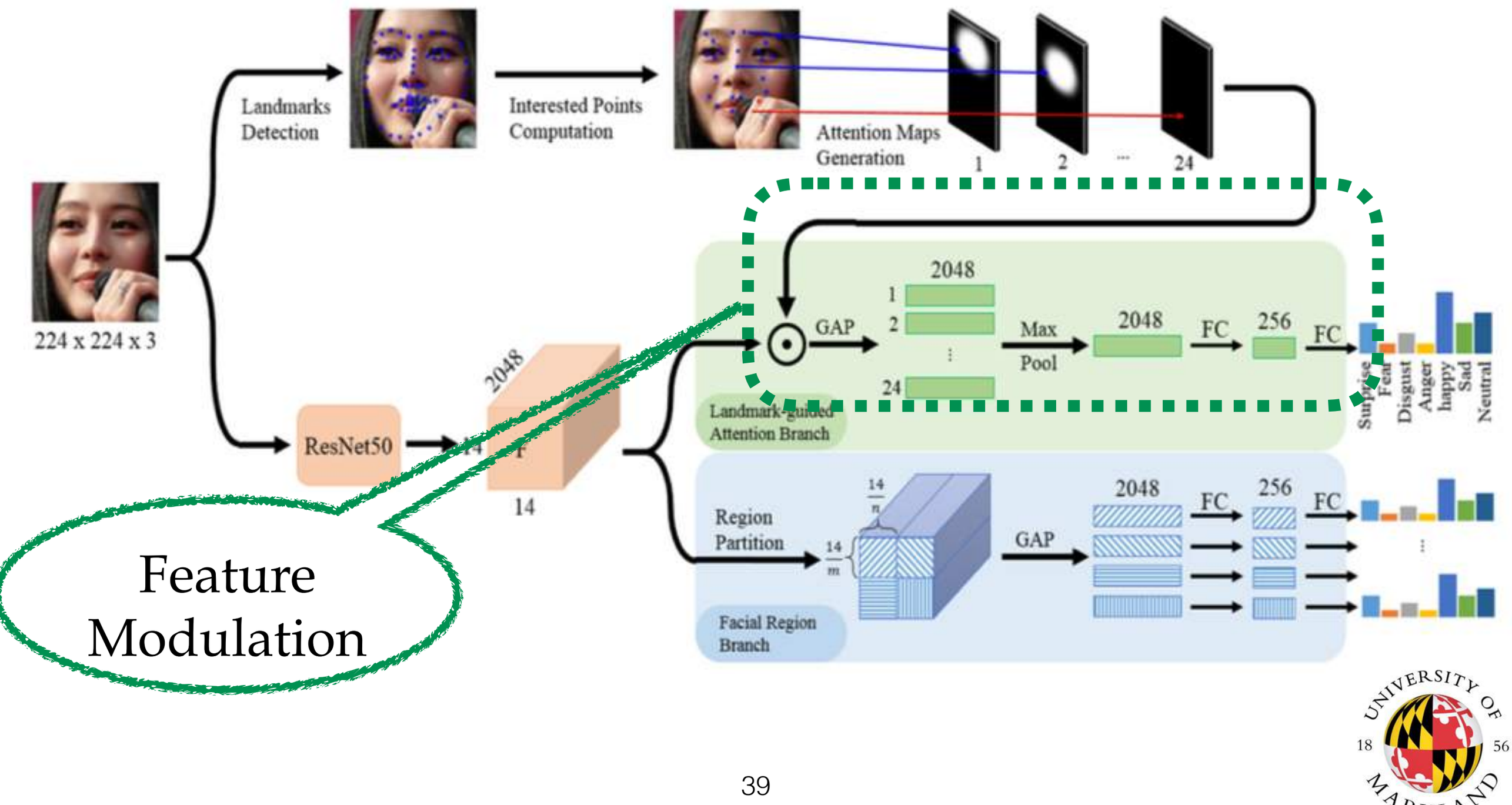
Landmark-guided Attention Branch (LAB)



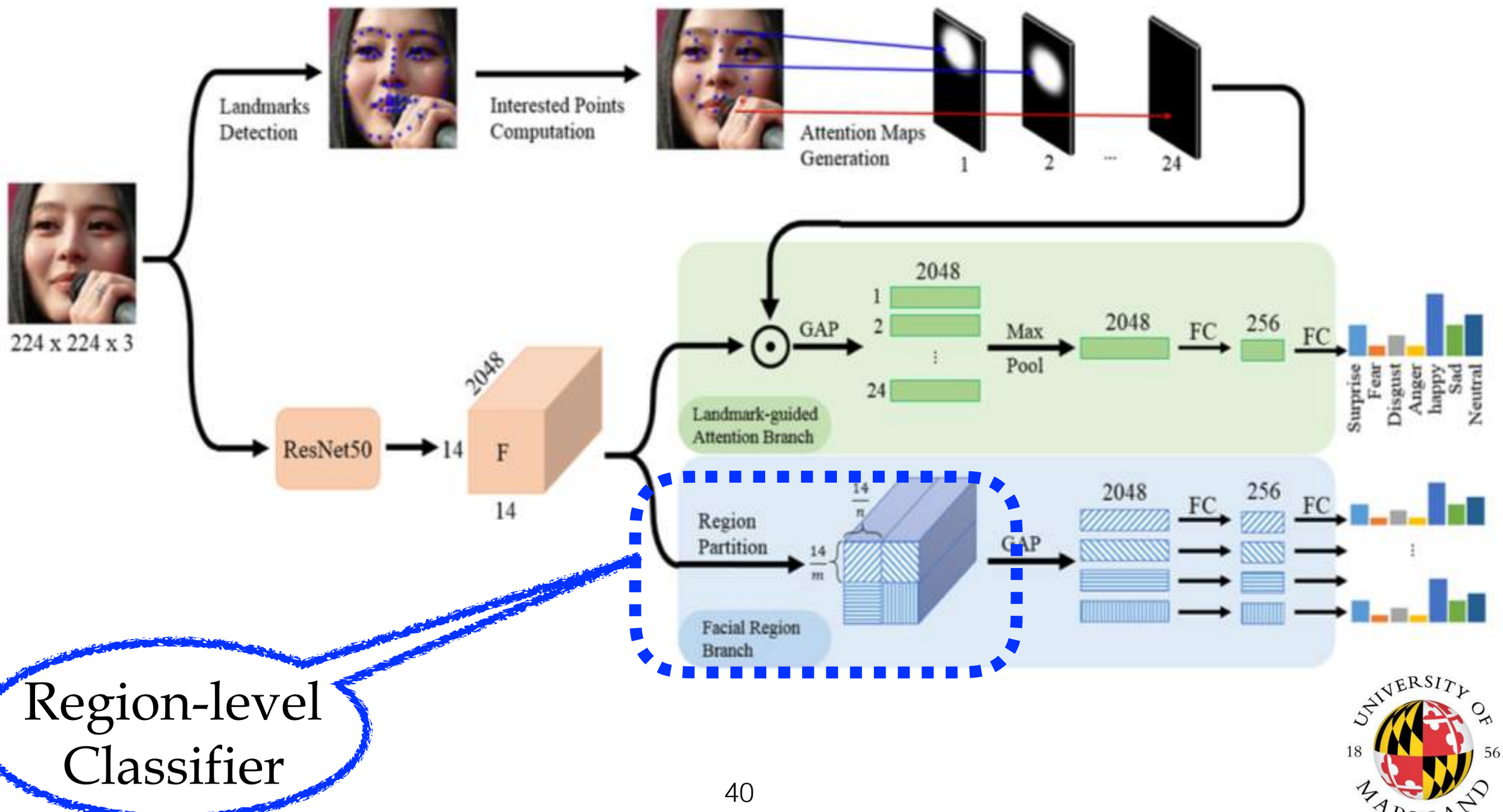
Attention Map Generation

$$p_i = \begin{cases} (x_i, y_i) & if \;\; s_i^{conf} \geq T \\ 0 & else \end{cases}$$

Landmark-guided Attention Branch (LAB)



Facial Region Branch (FRB)



Training Loss

$$L = \lambda L_{LAB} + (1 - \lambda) L_{FRB}$$

$$L_{LAB} = -\sum_{i=1}^{C} y_i \log \hat{y}_i$$

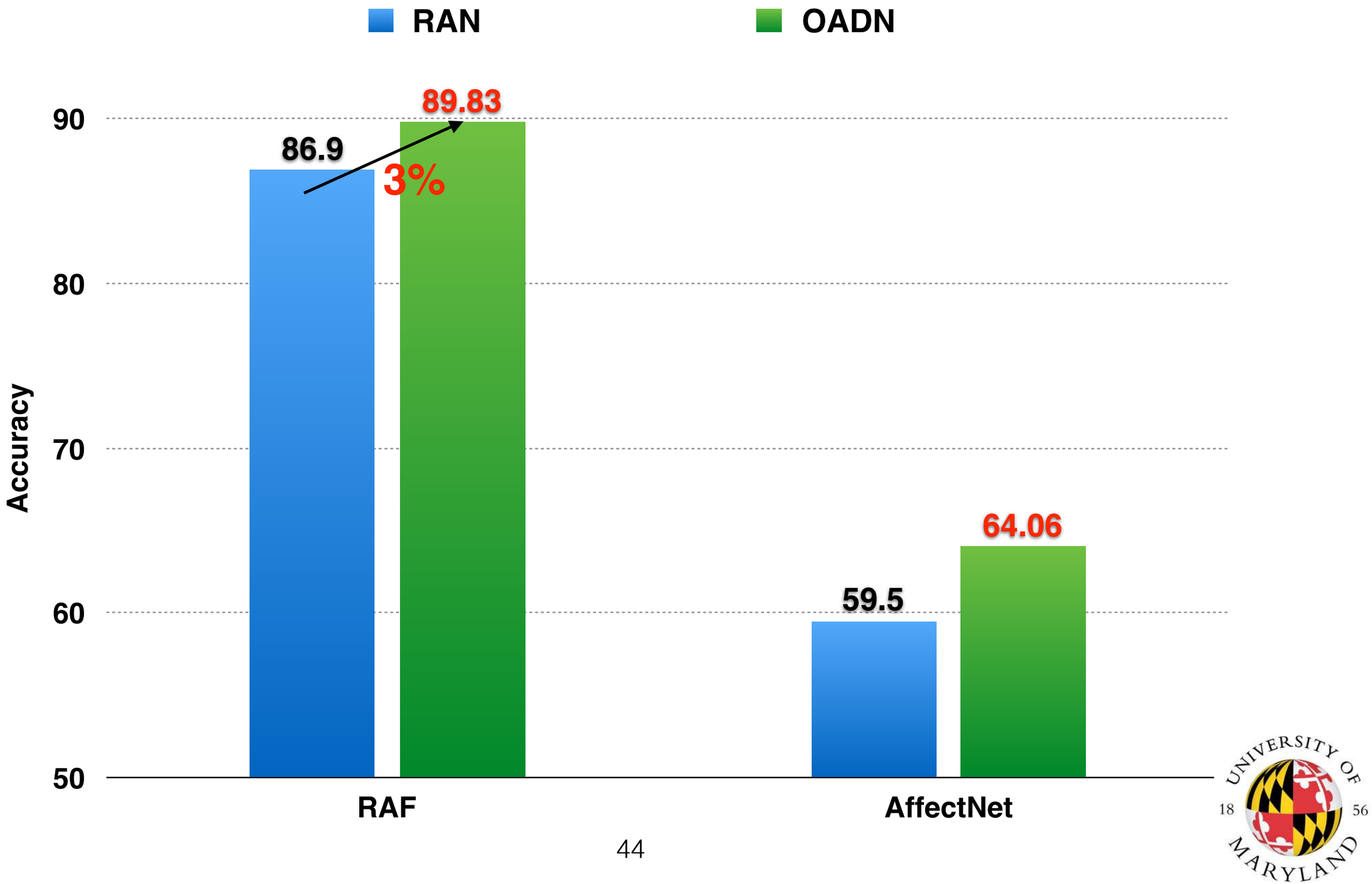
$$L_{FRB} = -\sum_{i=1}^{C} \sum_{j=1}^{K} y_i \log \hat{y}_{i,j}^{R}$$

Interest Points Selection Results

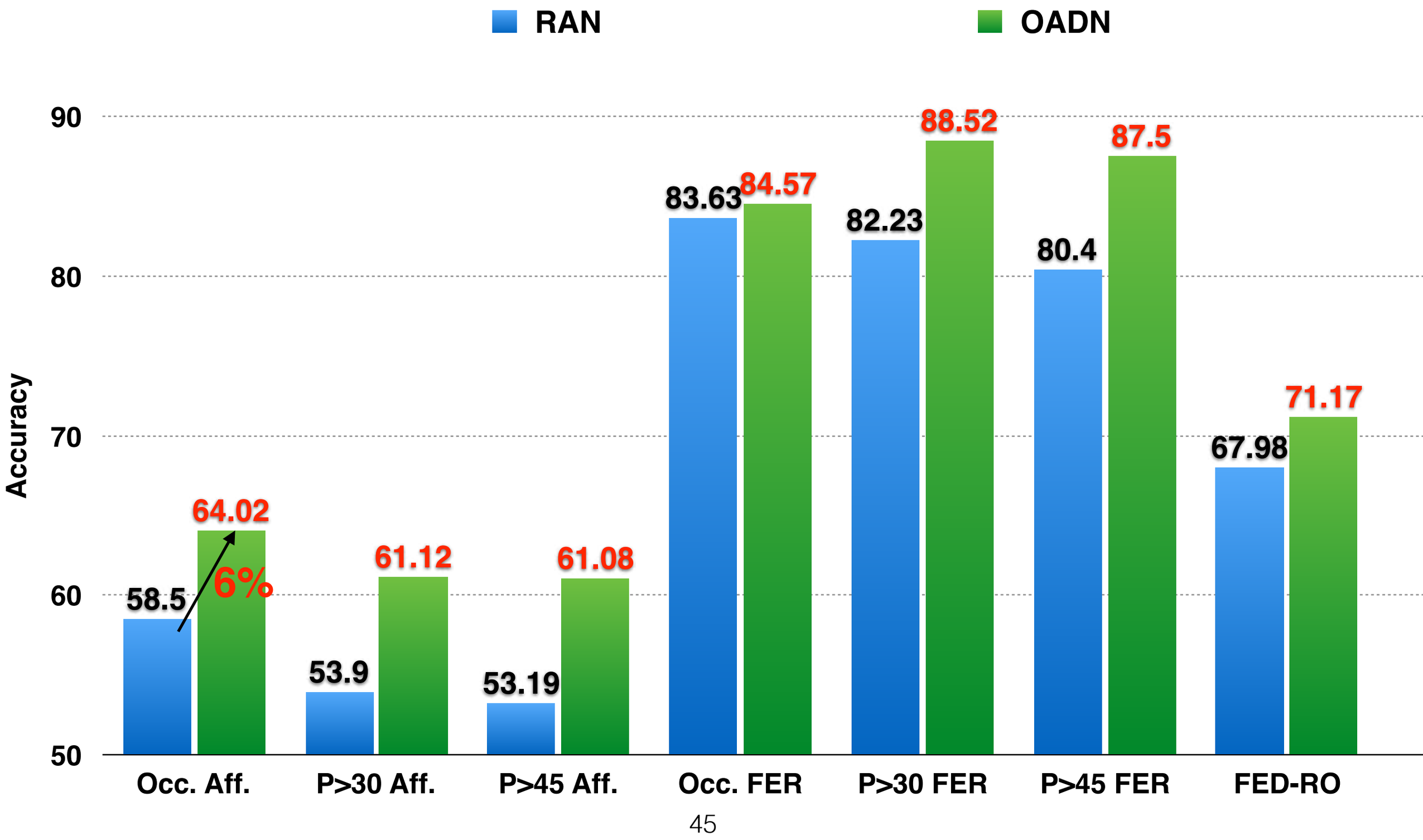
Experiments

Datasets	Train	Test	In-the-wild	Occlusion Specific	Pose Specific
RAF	12,271	3,068	Yes		
AffectNet	280,000	3,500	Yes		
Occlusion-AffectNet	N/A	682		Yes	
Pose>30 AffectNet	N/A	1,949			Yes
Pose>45 AffectNet	N/A	985			Yes
Occlusion-FER	N/A	605		Yes	
Pose>30 FER	N/A	1,171			Yes
Pose>45 FER	N/A	634			Yes
FED-RO	N/A	400		Yes	

Recognition Accuracy Comparison on Occlusion and Pose Datasets

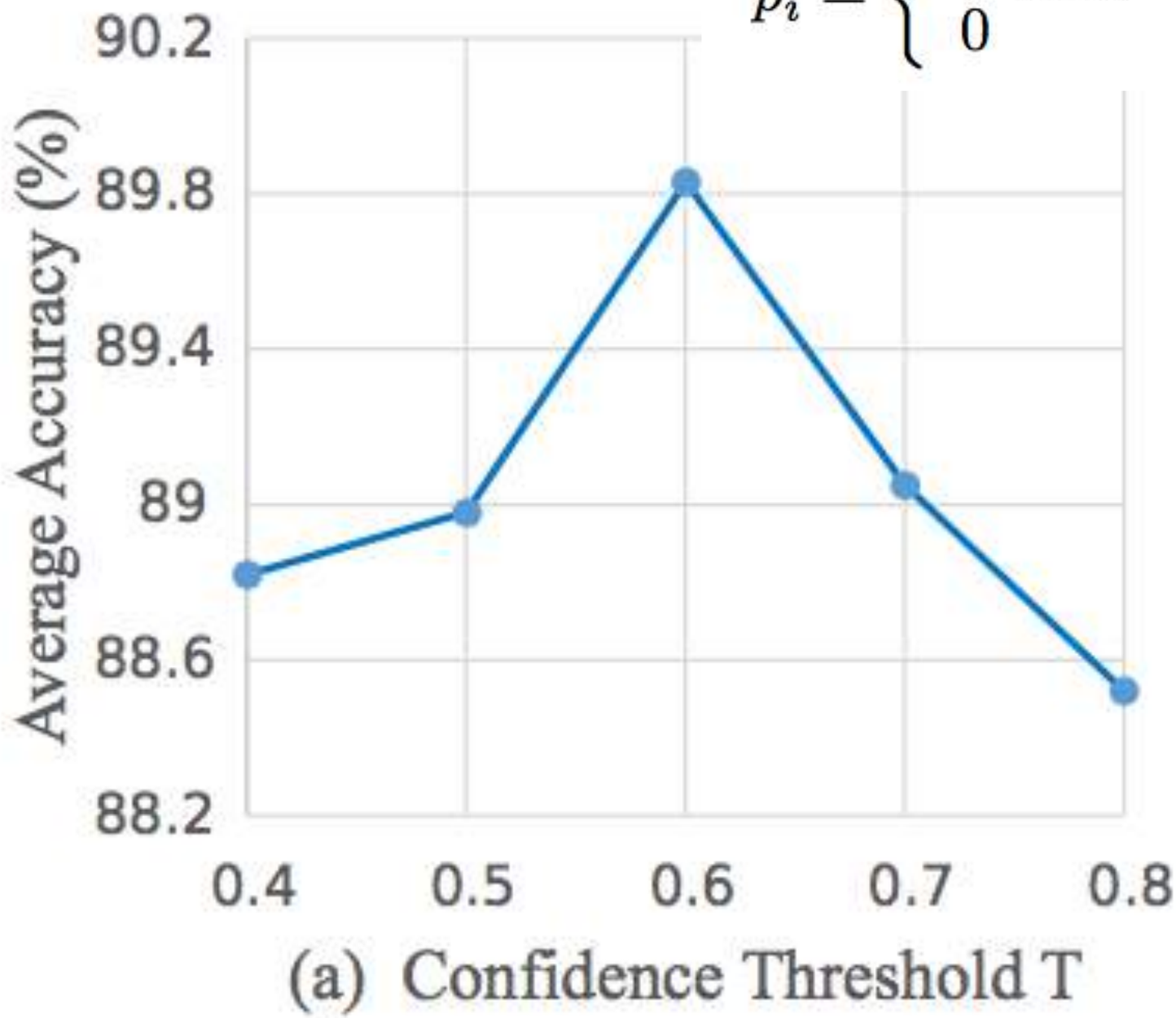


Recognition Accuracy Comparison on Occlusion and Pose Datasets

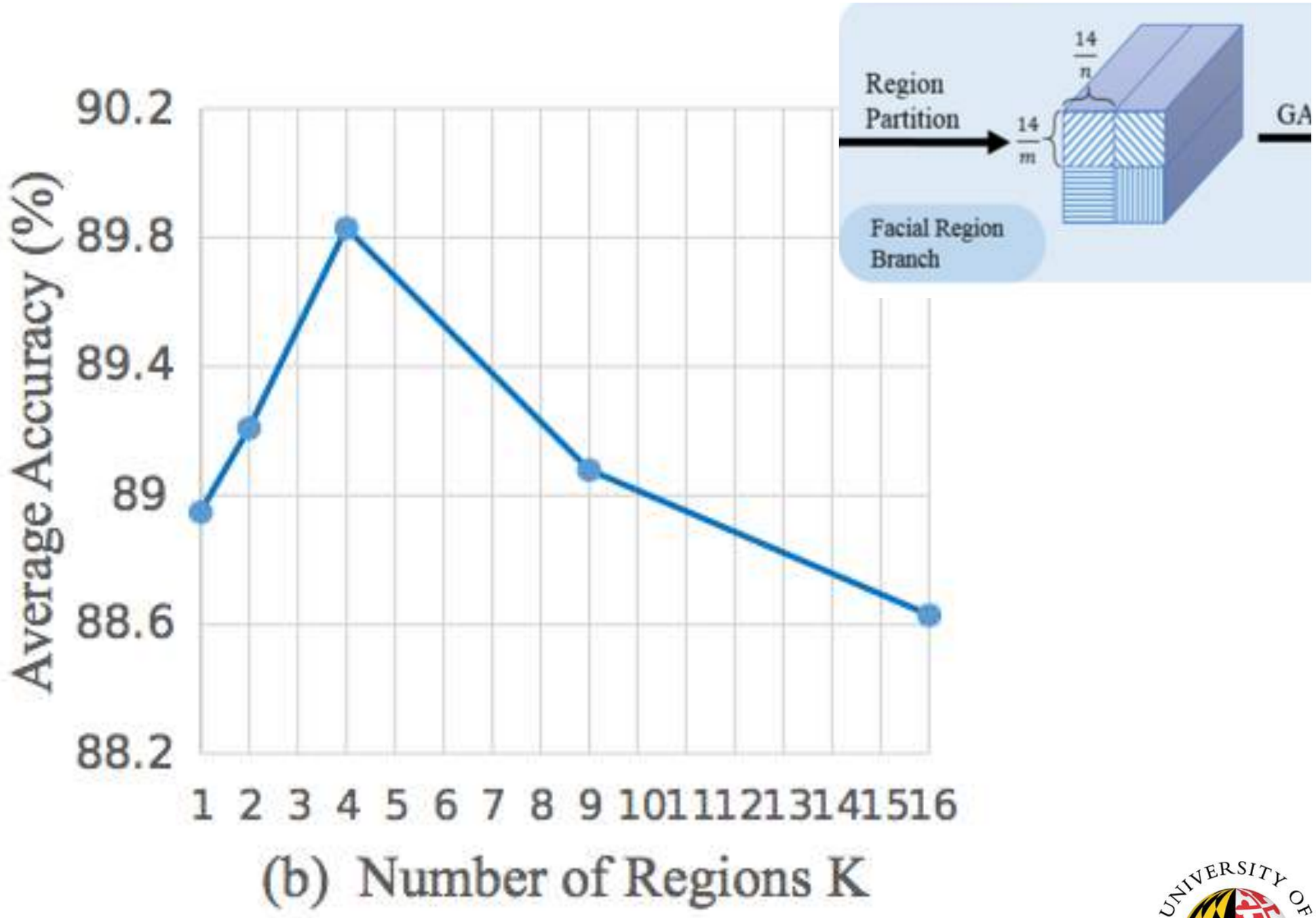


Ablation Study

$$p_i = \begin{cases} (x_i, y_i) & if \ \ s_i^{conf} \geq T \\ 0 & else \end{cases}$$

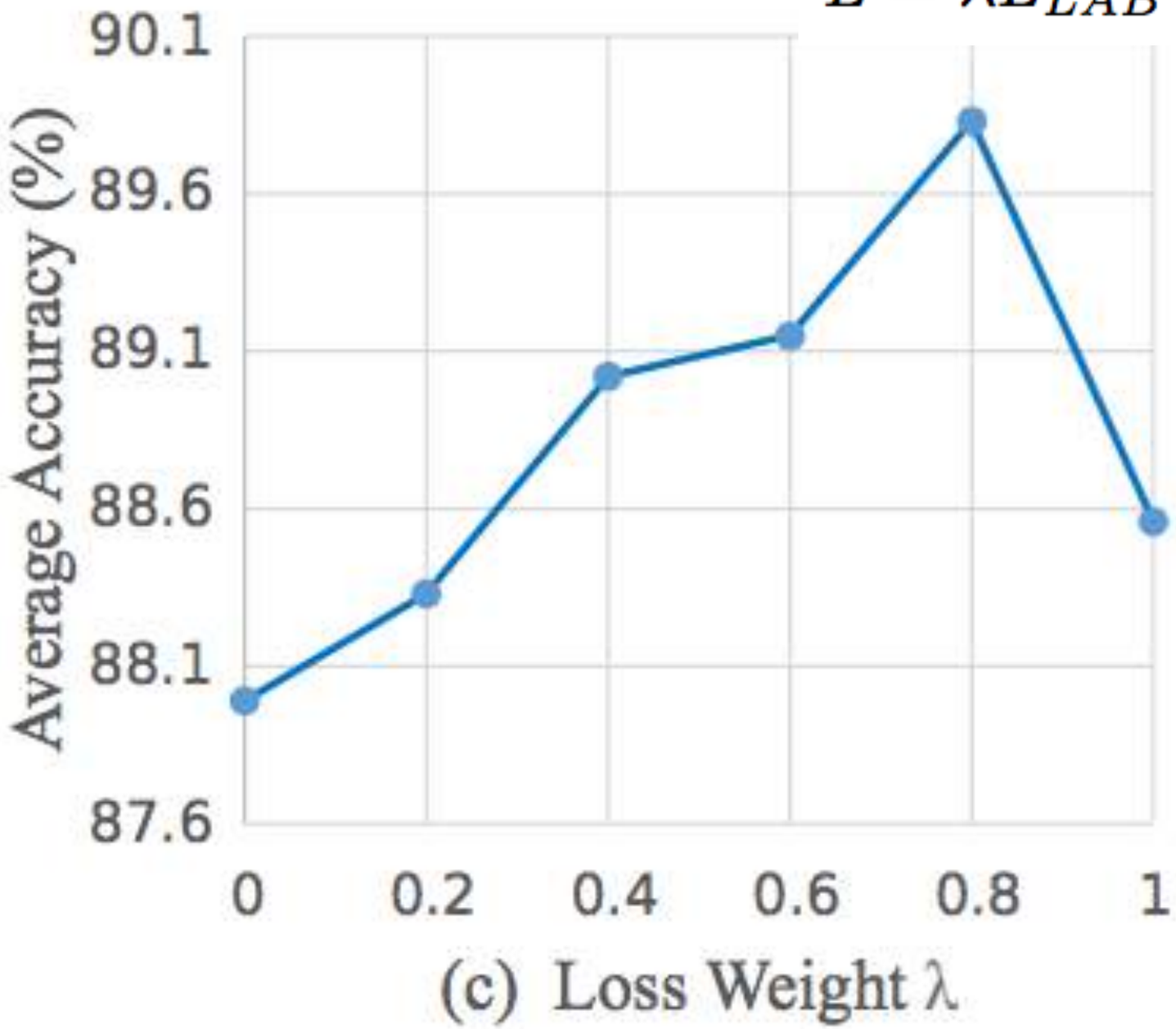


Ablation Study



Ablation Study

$$L = \lambda L_{LAB} + (1 - \lambda) L_{FRB}$$



Expression Recognition Results

gACNN	Happy	Happy	Neutral	Sad	Fear
OADN	Sad	Sad	Sad	Fear	Surprise

From Expression Recognition to Attributes Classification

Agenda

- ✦ Transfer Learning (Small Datasets)
  - FaceNet2ExpNet
- ✦ Robust Model Design (Occlusion, Pose)
  - Occlusion Robust Deep Network
  - Unaligned Attribute Classifier
- ✦ Generative Model (Fine-Grained)
  - ExprGAN

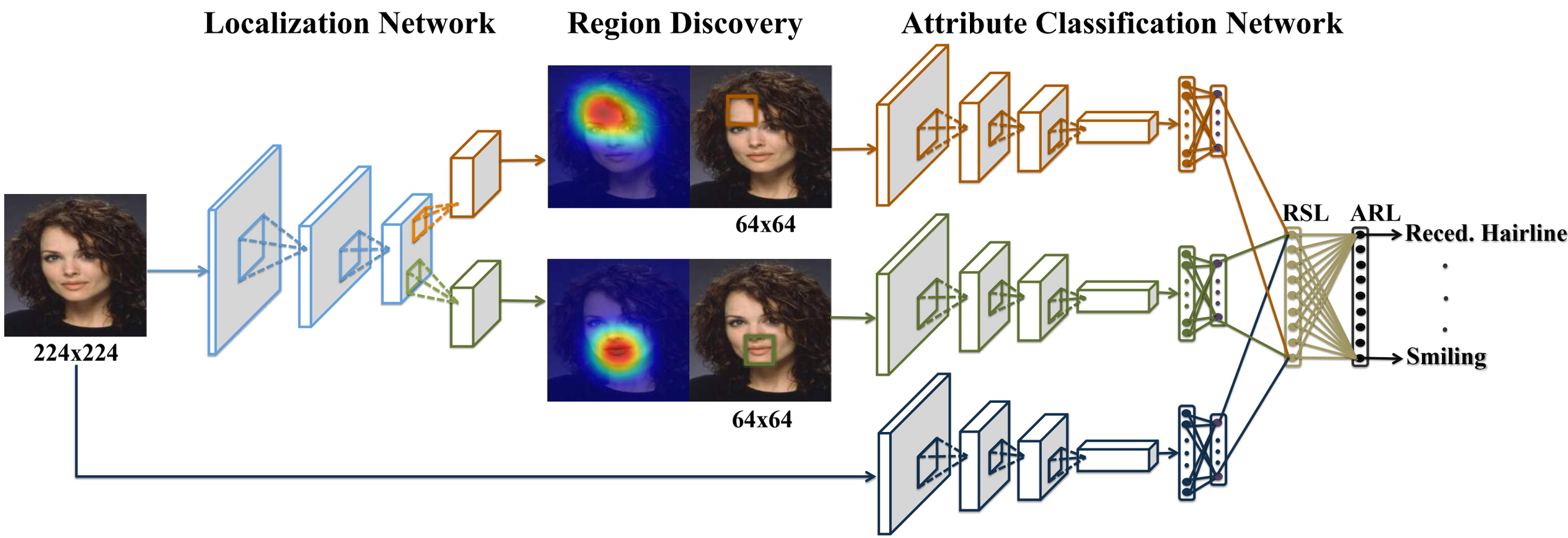
A Deep Cascade Network for Unaligned Face Attribute Classification

Hui Ding, Hao Zhou, Shaohua Kevin Zhou and Rama Chellappa, AAAI, 2018.

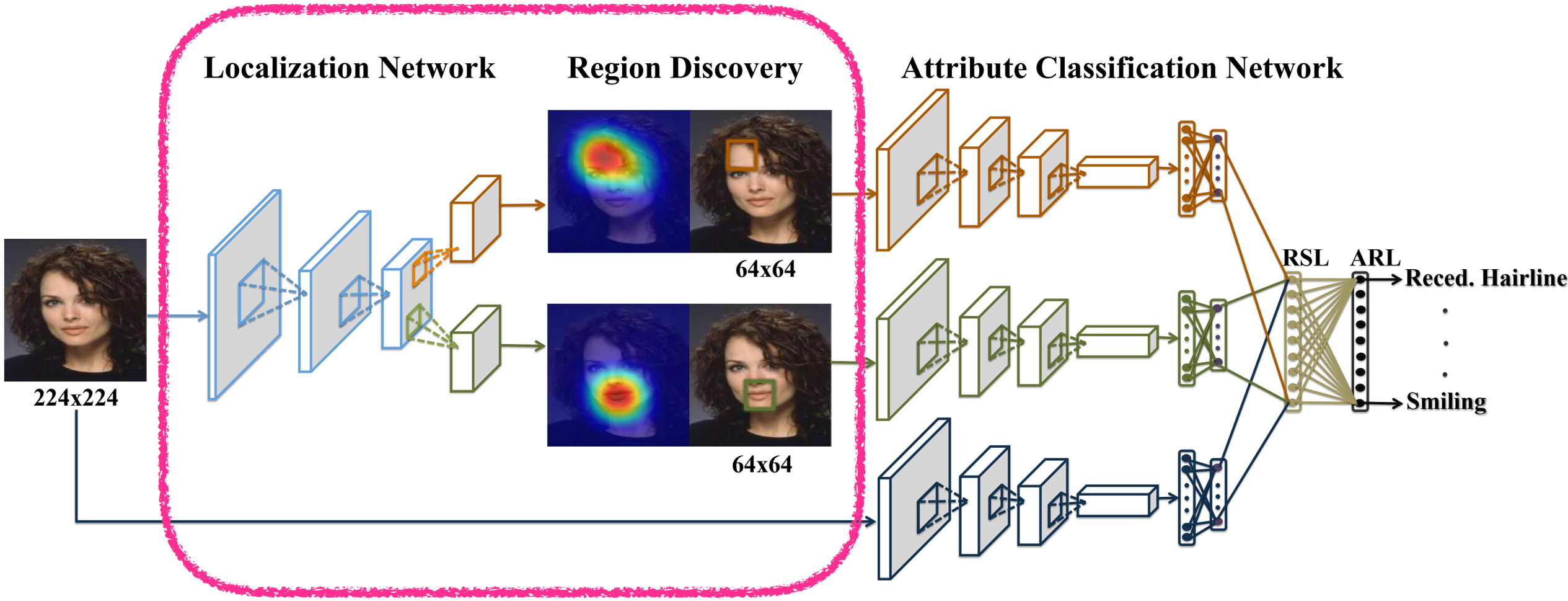
Motivation

Attend to the most related regions for attributes recognition

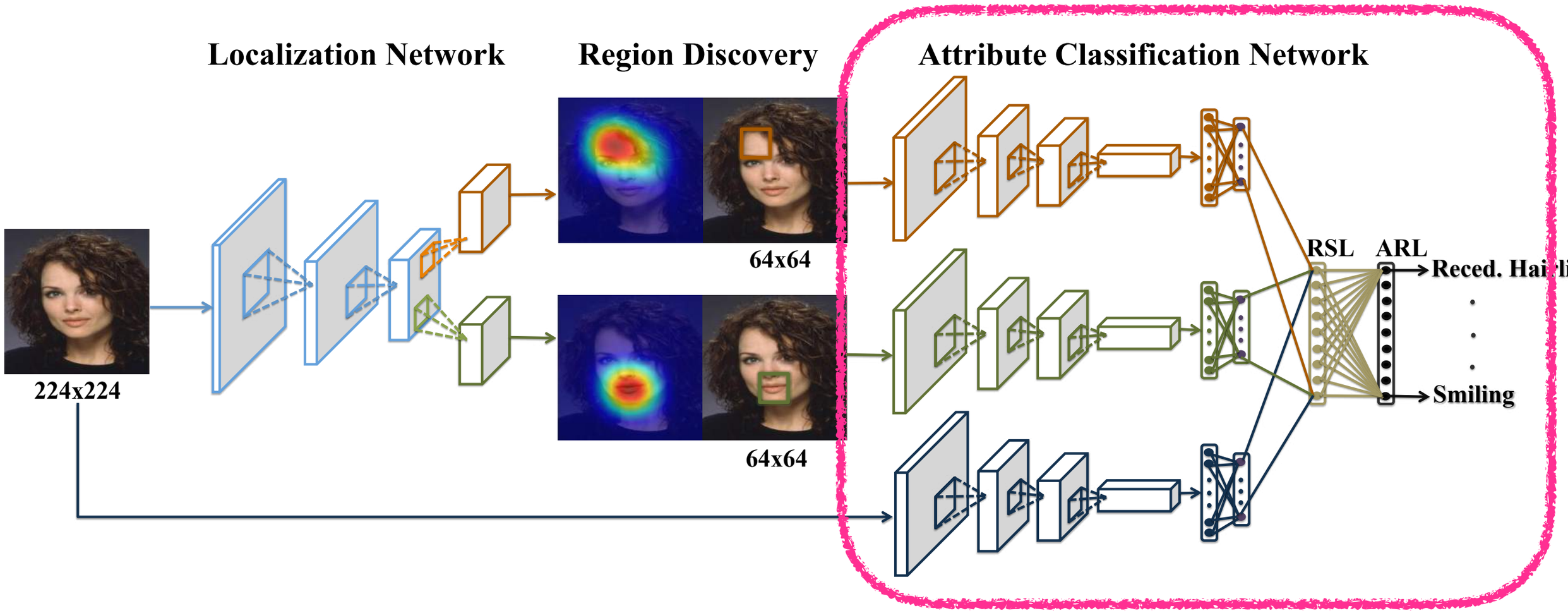
Unaligned Attribute Classifier (UAC)



Region Localization Network



Attribute Classification Network



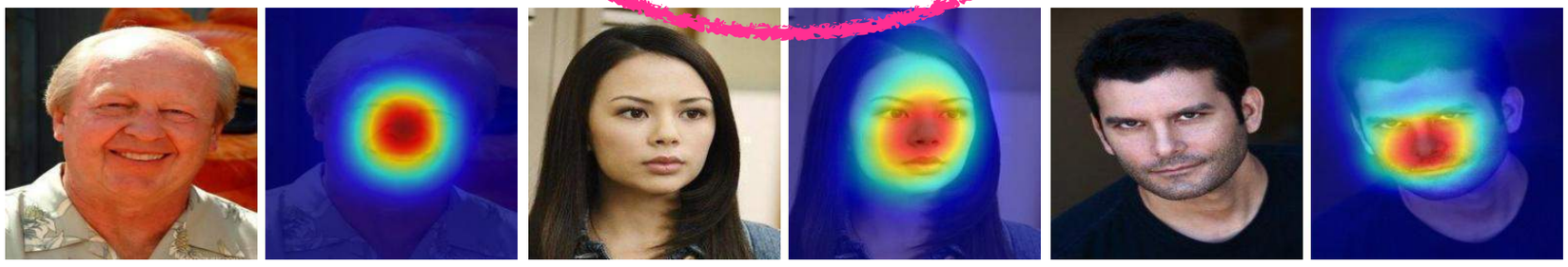
Face Region Localization Results

Bald

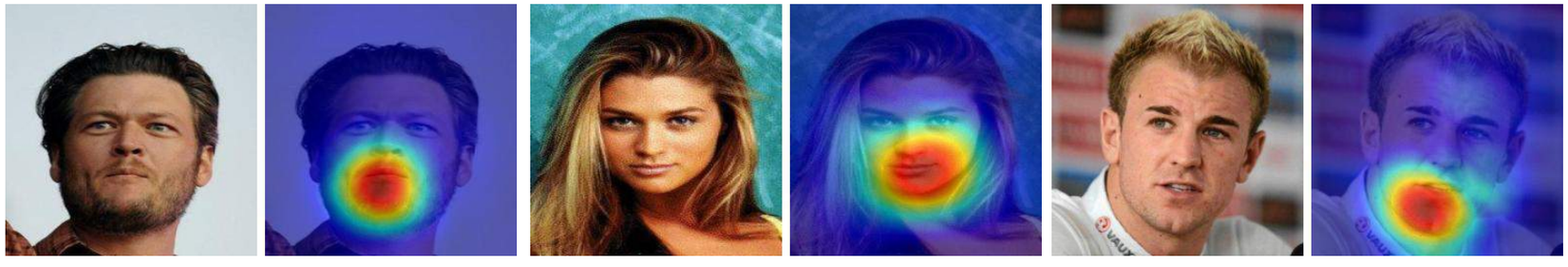
Wavy Hair

Arched Eyebrow

Big Nose

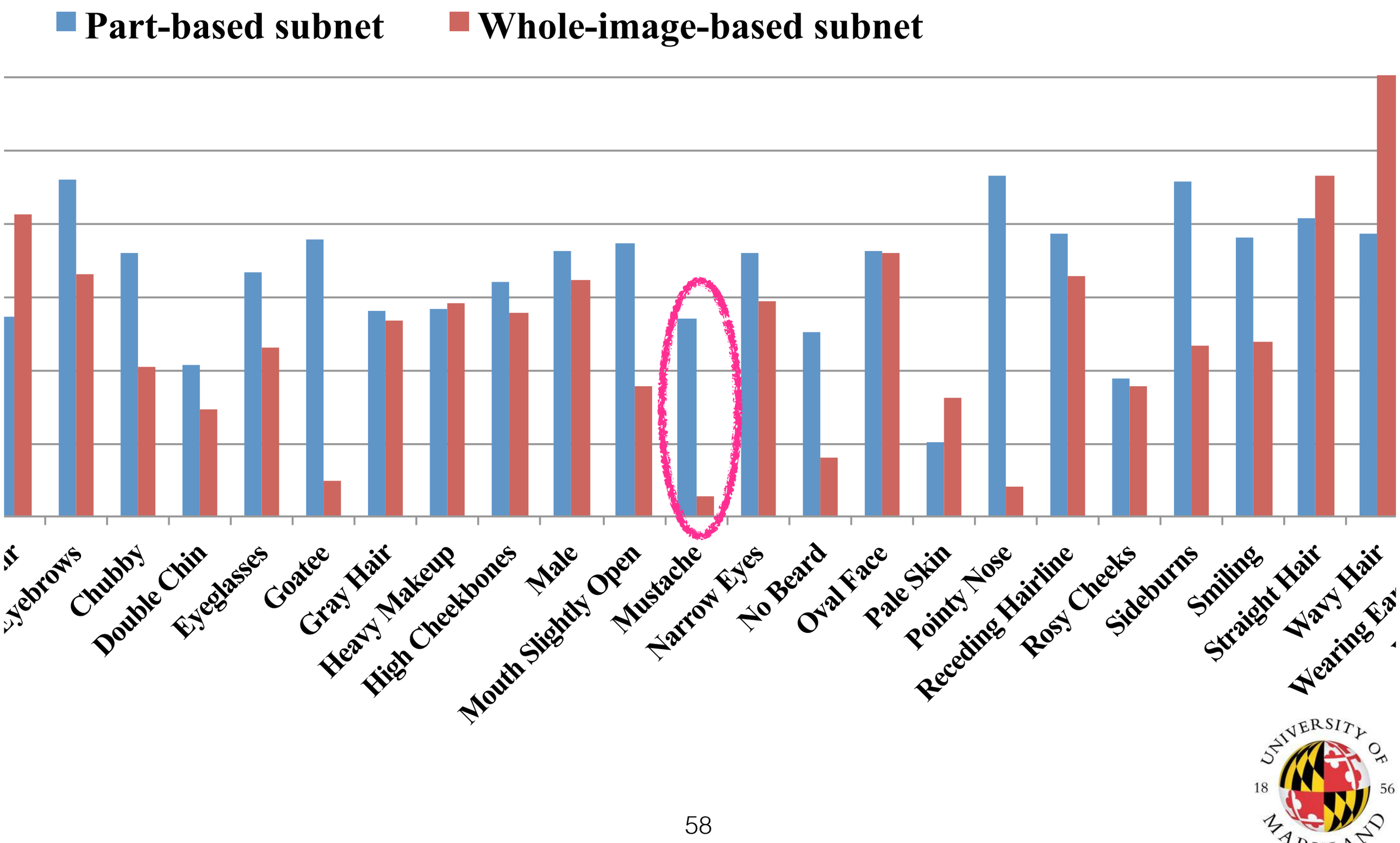


5 O Clock Shadow

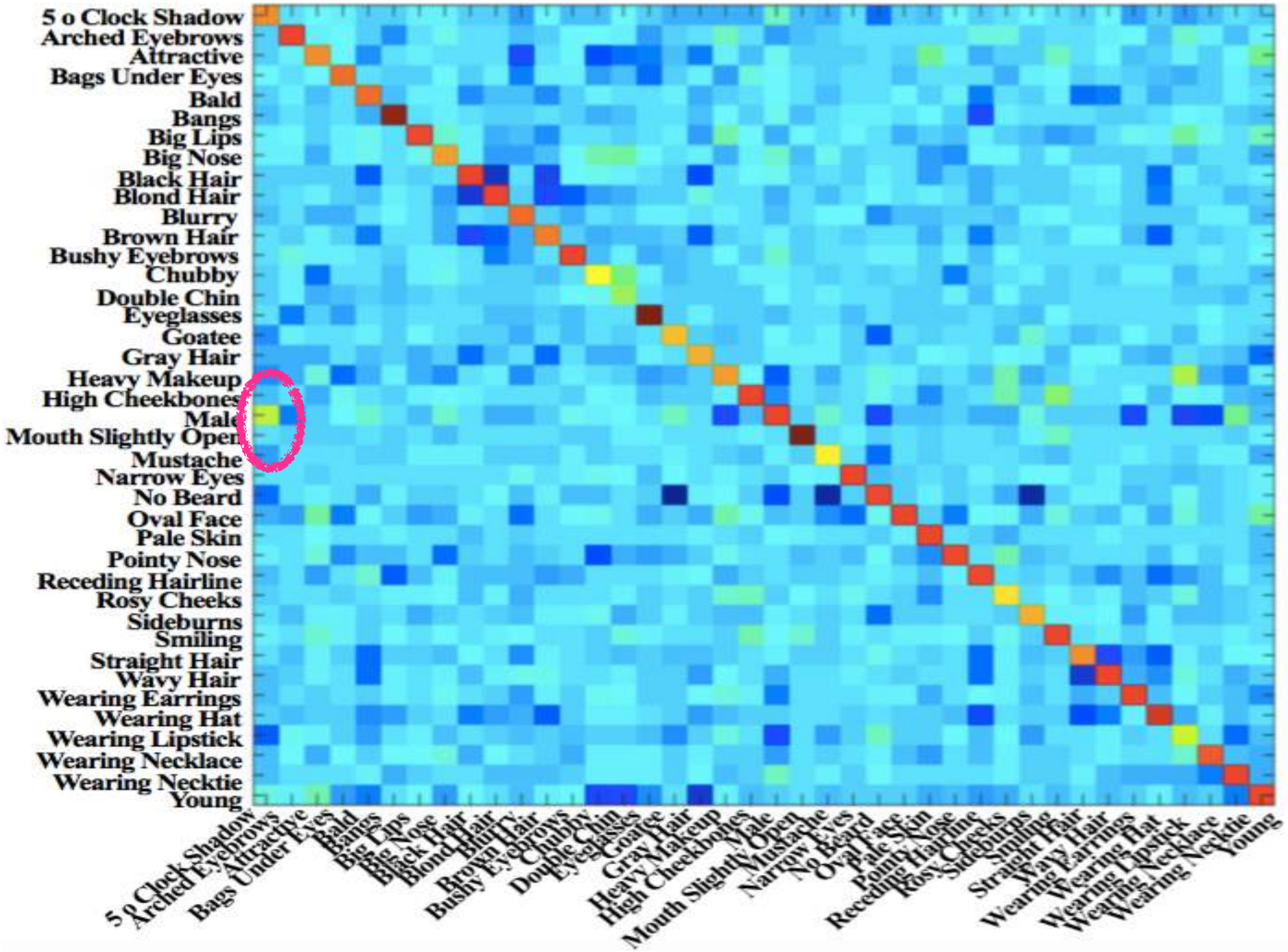


Heavy Makeup

Region Switch Layer



Attribute Relation Layer



Attributes Classification Accuracy

		5 o Clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
uCelebA	LNets+ANet [25]	91.00	79.00	81.00	79.00	98.00	95.00	68.00	78.00	88.00	95.00	84.00	80.00	90.00	91.00	92.00	99.00	95.00	97.00	90.00	87.00	98.00
	Part-only	93.90	81.86	81.88	84.07	98.72	95.71	70.63	83.48	87.97	95.16	95.83	87.53	91.73	95.05	95.92	99.46	97.19	97.93	90.26	86.20	96.65
	Whole-only	93.95	81.43	82.06	84.11	98.57	95.45	70.66	82.91	89.08	95.52	96.01	88.63	92.32	95.12	95.98	99.40	96.90	98.07	90.67	86.57	97.10
	PaW	**94.64**	**83.01**	**82.86**	**84.58**	**98.93**	**95.93**	**71.46**	**83.63**	**89.84**	**95.85**	**96.11**	**88.50**	**92.62**	**95.46**	**96.26**	**99.59**	**97.38**	**98.21**	**91.53**	**87.44**	**98.39**

		Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		Average
uCelebA	LNets+ANet [25]	92.00	95.00	81.00	95.00	66.00	91.00	72.00	89.00	90.00	96.00	92.00	73.00	80.00	82.00	99.00	93.00	71.00	93.00	87.00		87.30
	Part-only	93.55	96.63	86.96	95.71	73.03	96.86	76.40	92.87	94.77	97.63	91.98	82.53	81.29	89.07	98.75	92.96	87.13	96.69	86.51		90.46
	Whole-only	93.24	96.59	87.19	95.40	74.48	96.85	76.06	92.95	94.83	97.50	91.61	82.18	82.63	89.13	98.50	93.58	87.14	96.77	87.14		90.60
	PaW	**94.05**	**96.90**	**87.56**	**96.22**	**75.03**	**97.08**	**77.35**	**93.44**	**95.07**	**97.64**	**92.73**	**83.52**	**84.07**	**89.93**	**99.02**	**94.24**	**87.70**	**96.85**	**88.59**		**91.23**

From Expression Recognition to Attributes Classification

- Challenge 1:
  training datasets are small

From Expression Recognition to Attributes Classification

- ✦ Challenge 1: training datasets are small
- ✓ FaceNet2ExpNet: 12x smaller, high accuracy

From Expression Recognition to Attributes Classification

- ✦ Challenge 1:
  training datasets are small
- ✦ Challenge 2:
  occlusion and pose

- ✓ FaceNet2ExpNet:
  12x smaller, high accuracy

From Expression Recognition to Attributes Classification

- ✦ Challenge 1: training datasets are small
  - ✓ FaceNet2ExpNet: 12x smaller, high accuracy
- ✦ Challenge 2: occlusion and pose
  - ✓ OADN/UAC: occlusion robust no need of face alignment

Agenda

- ✦ Transfer Learning (Small Datasets)
  - FaceNet2ExpNet
- ✦ Robust Model Design (Occlusion, Pose)
  - Occlusion Robust Deep Network
  - Unaligned Attribute Classifier
- ✦ Generative Model (Fine-Grained)
  - ExprGAN

ExprGAN: Facial Expression Editing with Controllable Expression Intensity

Hui Ding, Kumar Sricharan and Rama Chellappa, AAAI 2018, Oral.

Face Generation

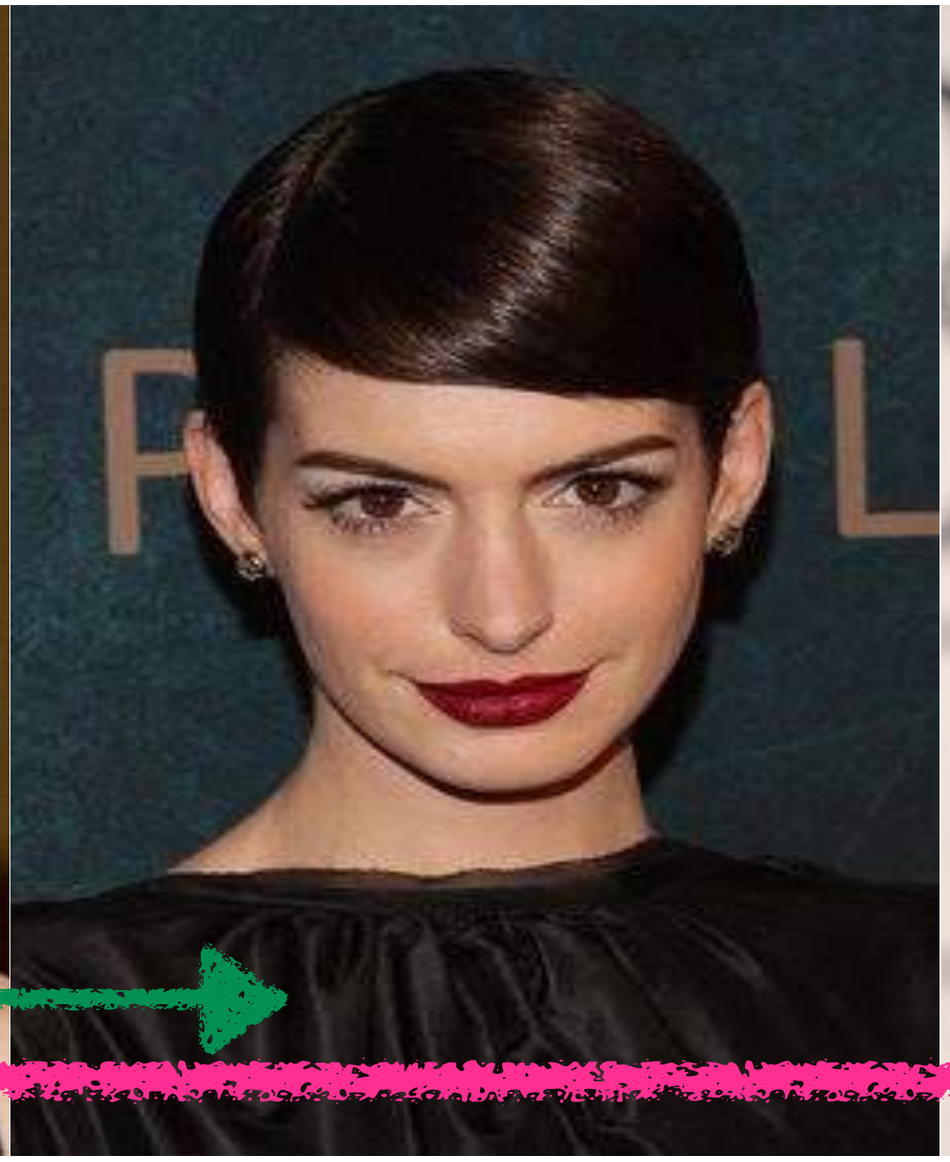
Ian Goodfellow's twitter

Expression Generation

Deep Belief Network	Gated Boltzmann Machine	Variational Auto-encoder
		
2008	2014	2016

Expression Editing is Multi-modal

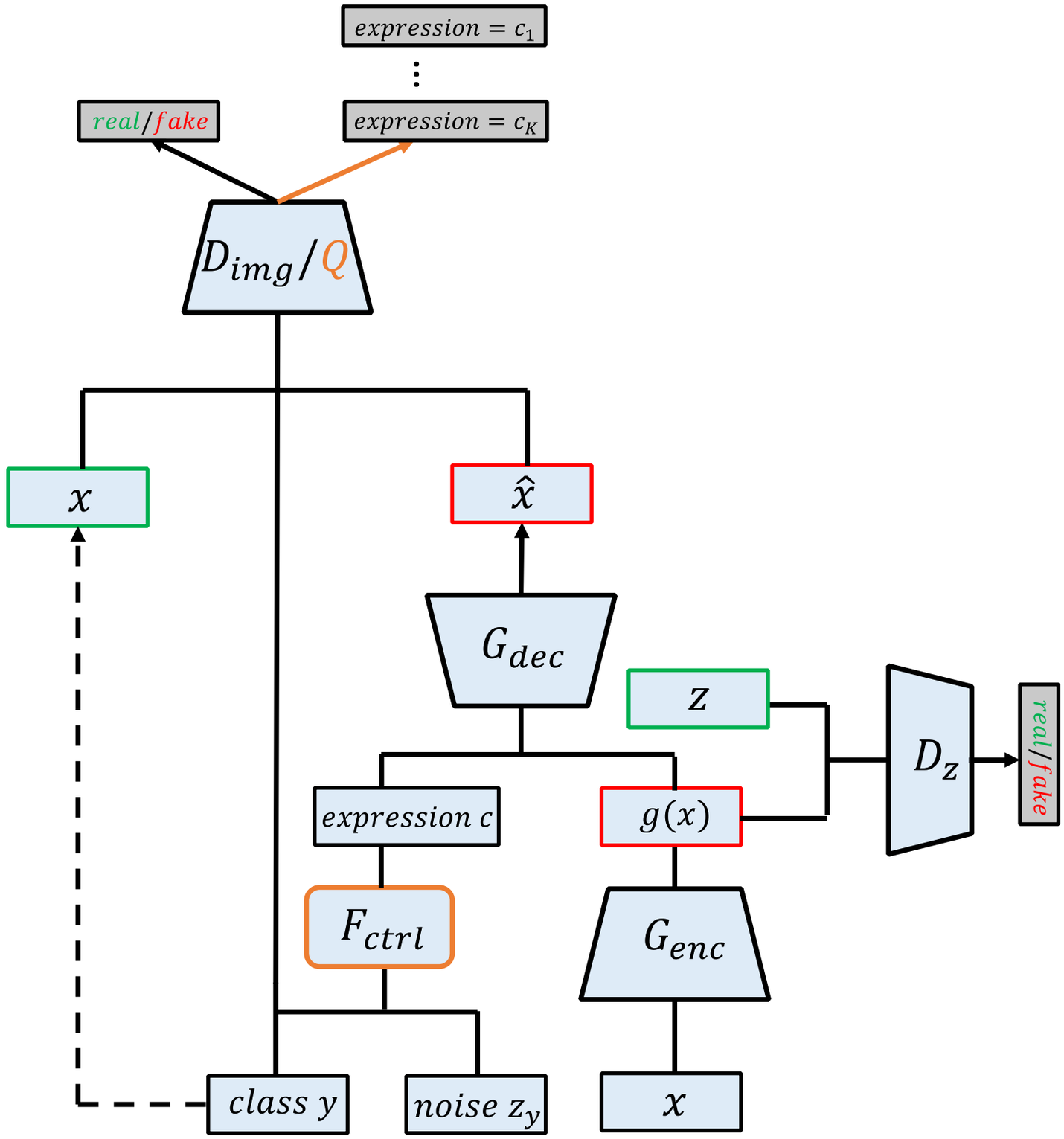
Expression Editing is Multi-modal



What is ExprGAN

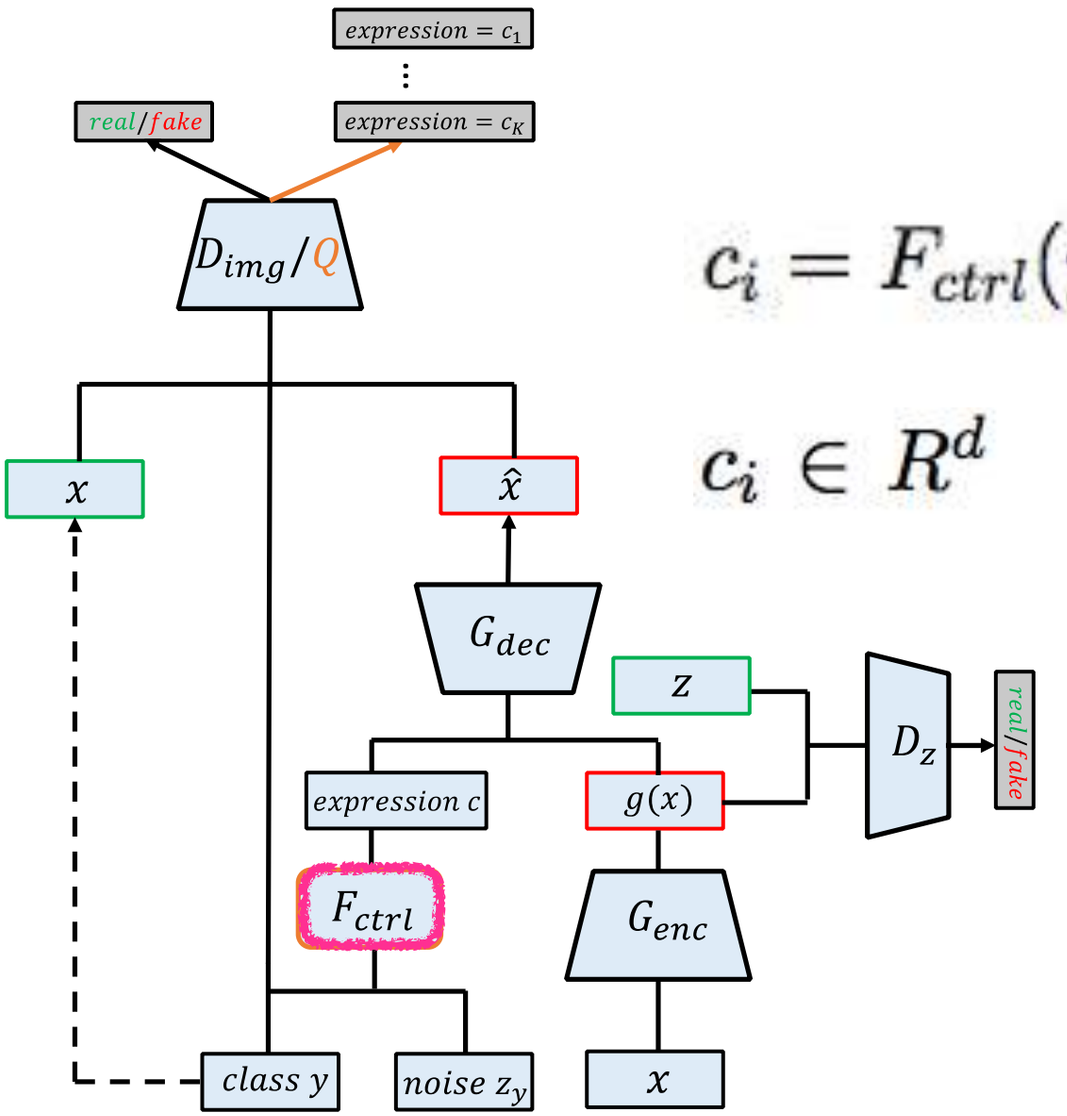
— First GAN-based model for facial expression editing

What is ExprGAN



How to generate images displaying different expression intensities when we only have training images labeled with categories?

Expression Controller Module



$$c_i = F_{ctrl}(y_i, z_y) = |z_y| \cdot (2y_i - 1) \quad i = 1, 2, \ldots, K$$

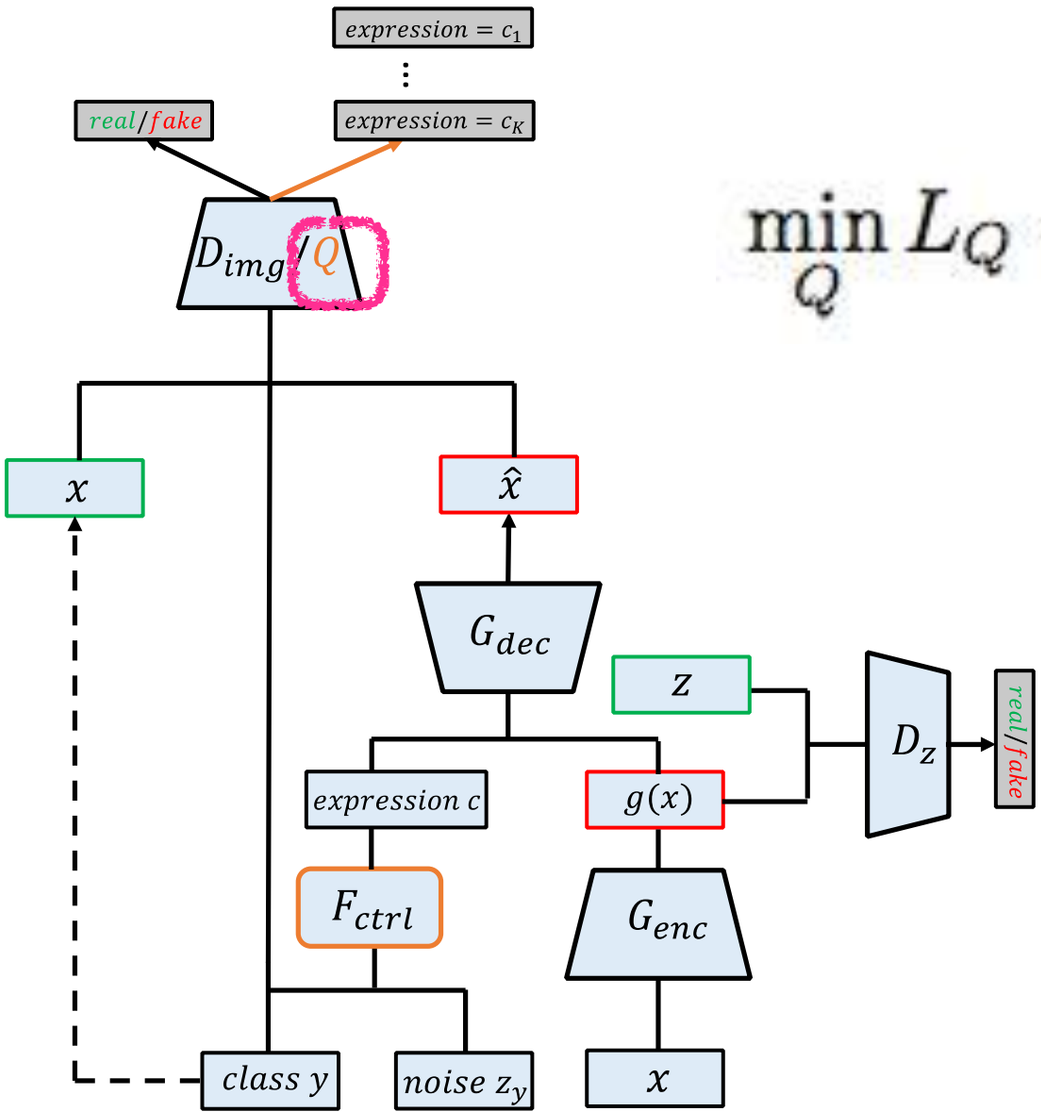
$$c_i \in R^d$$

y_i	c_i
1	0.5 0.2
0	-0.1 -0.4

One-hot Label

Expr. Code

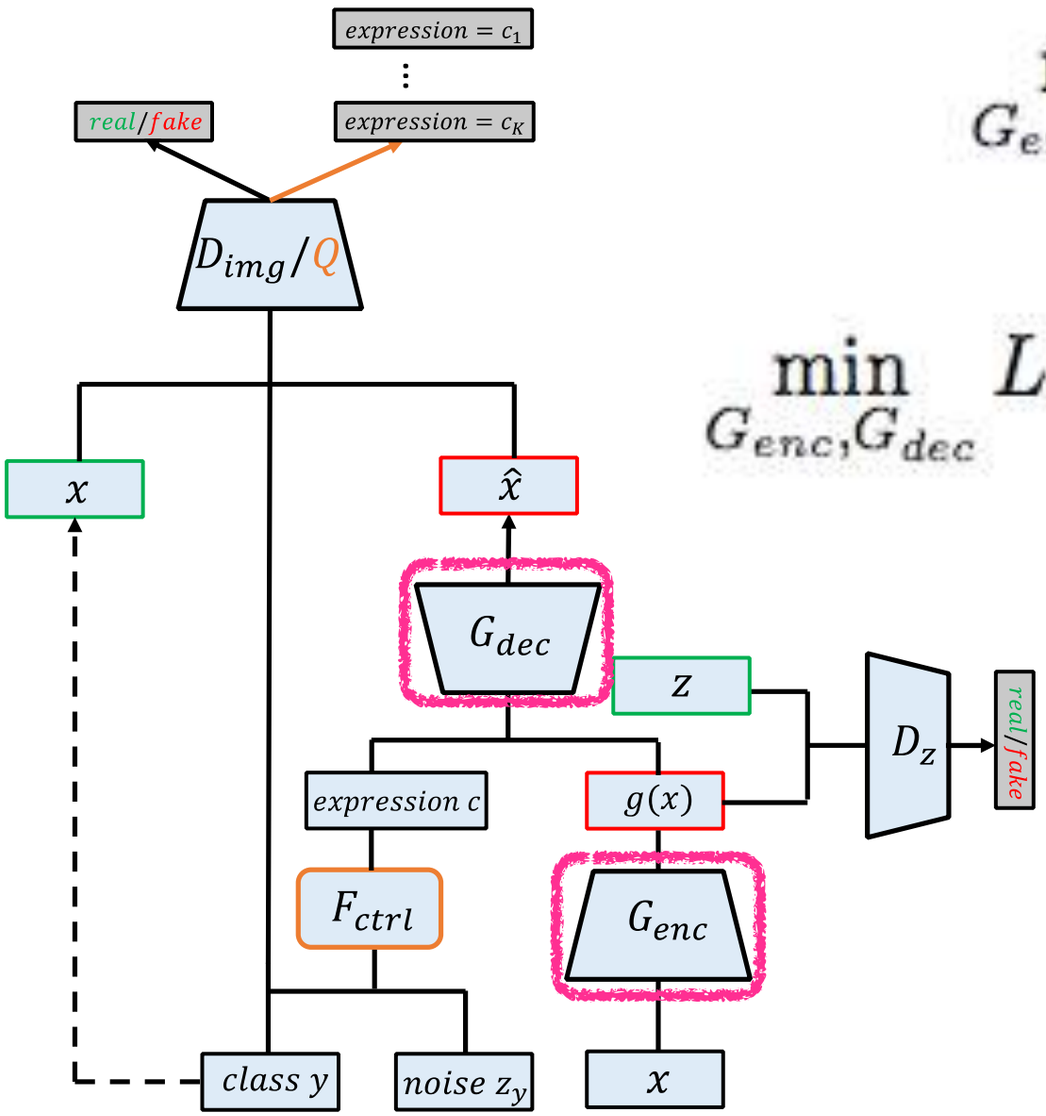
Expression Regularization Network



$$\min_Q L_Q = -\mathbb{E}_{c\sim P(c|y),\hat{x}\sim G_{dec}(g(x),c)}[\log Q(c|\hat{x},y)]$$

$$\begin{aligned}
&I(c;\hat{x}|y)\\
&= H(c|y) - H(c|\hat{x},y)\\
&= \mathbb{E}_{\hat{x}\sim G_{dec}(g(x),c)}[\mathbb{E}_{c'\sim P(c'|\hat{x},y)}[\log P(c'|\hat{x},y)]] + H(c|y)\\
&= \mathbb{E}_{\hat{x}\sim G_{dec}(g(x),c)}[D_{KL}(P(\cdot|\hat{x},y)||Q(\cdot|\hat{x},y))+\\
&\quad \mathbb{E}_{c'\sim P(c'|\hat{x},y)}[\log Q(c'|\hat{x},y)]] + H(c|y)\\
&\geq \mathbb{E}_{\hat{x}\sim G_{dec}(g(x),c)}[\mathbb{E}_{c'\sim P(c'|\hat{x},y)}[\log Q(c'|\hat{x},y)]] + H(c|y)\\
&= \mathbb{E}_{c\sim P(c|y),\hat{x}\sim G_{dec}(g(x),c)}[\log Q(c|\hat{x},y)] + H(c|y)
\end{aligned}$$

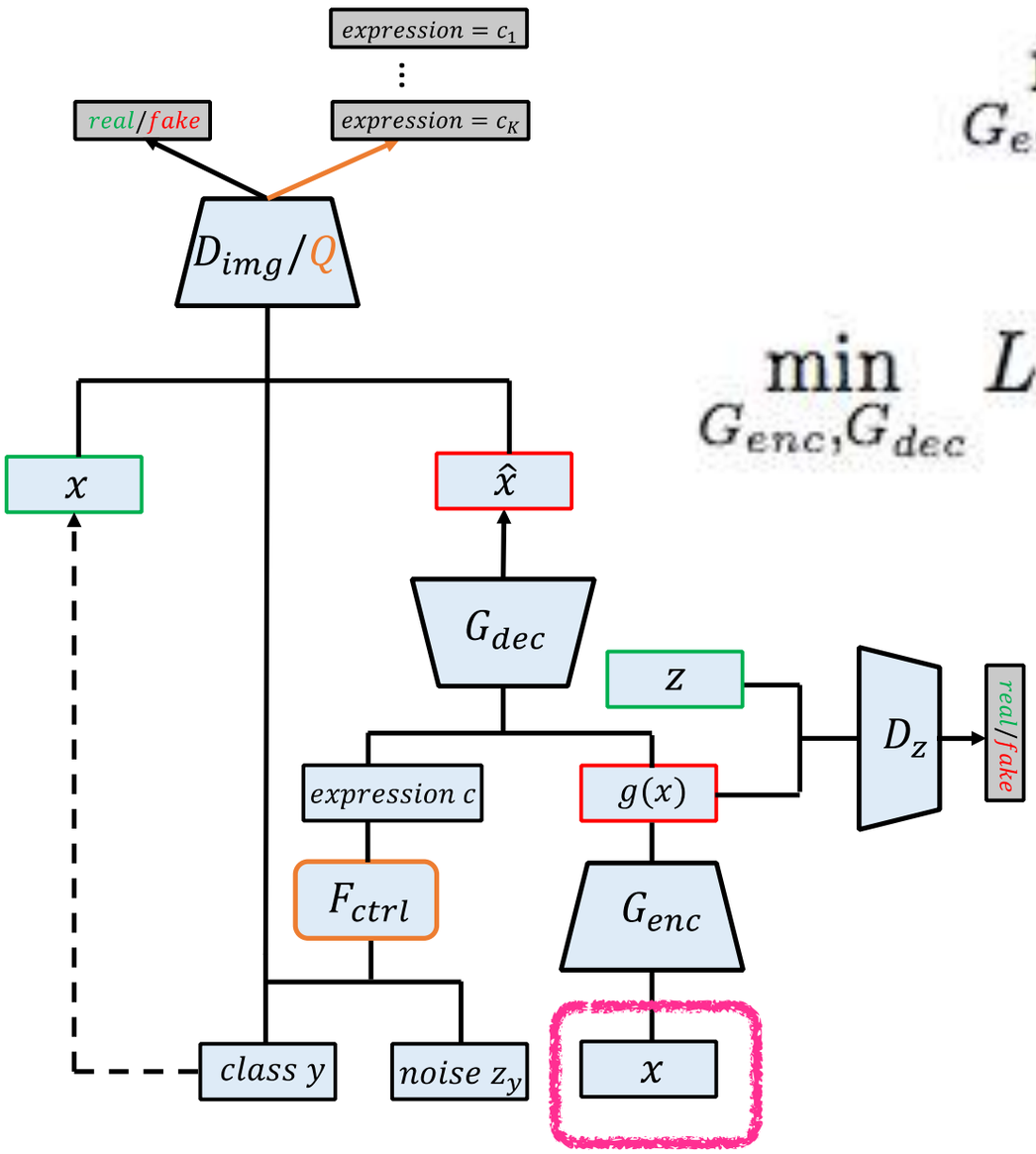
Generator Network: Encoder and Decoder



$$\min_{G_{enc},G_{dec}} L_{pixel} = L_1(G_{dec}(G_{enc}(x), c), x)$$

$$\min_{G_{enc},G_{dec}} L_{id} = \sum_l \beta_l L_1(\phi_l(G_{dec}(G_{enc}(x), c)), \phi_l(x))$$

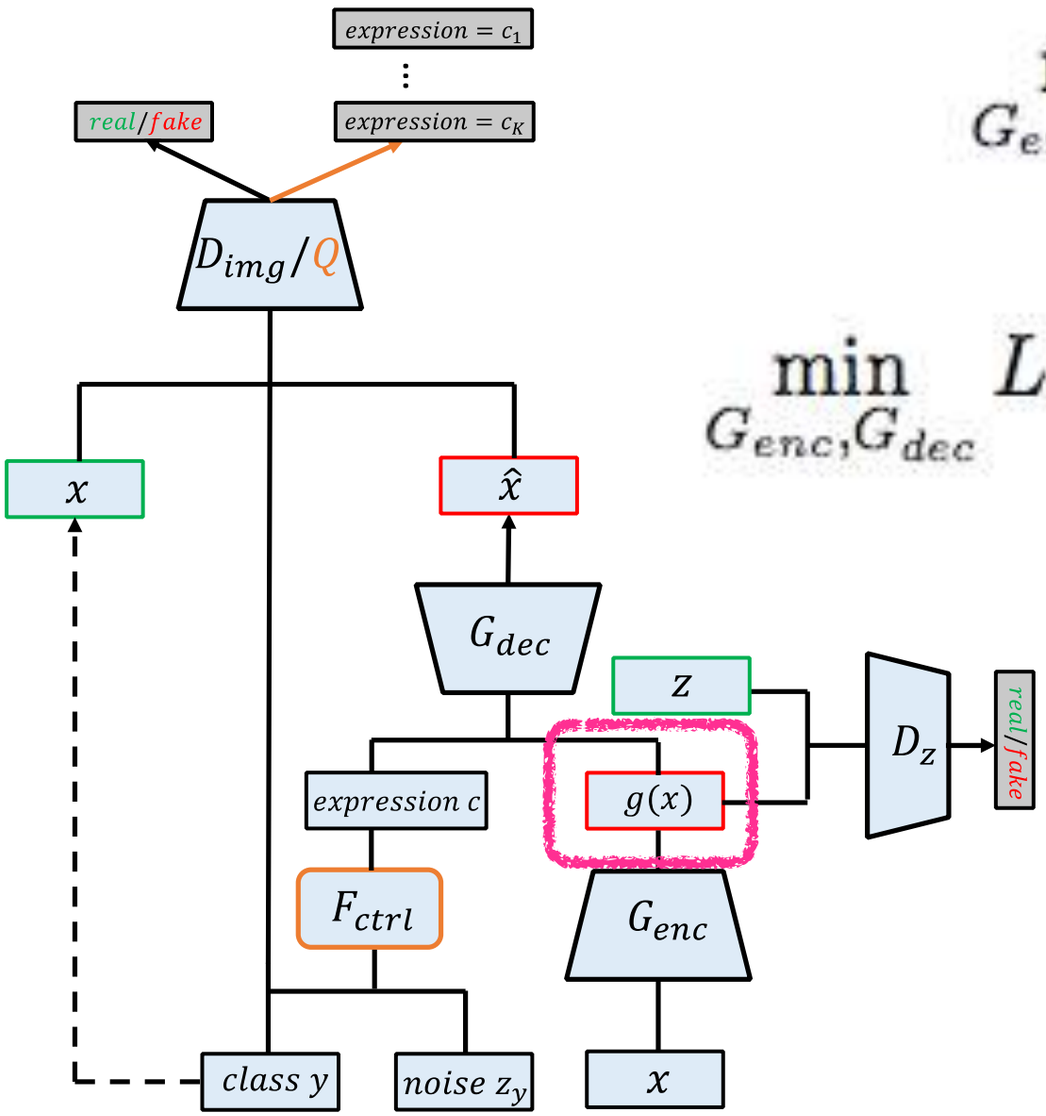
Generator Network: Encoder and Decoder



$$\min_{G_{enc},G_{dec}} L_{pixel} = L_1(G_{dec}(G_{enc}(x), c), x)$$

$$\min_{G_{enc},G_{dec}} L_{id} = \sum_l \beta_l L_1(\phi_l(G_{dec}(G_{enc}(x), c)), \phi_l(x))$$

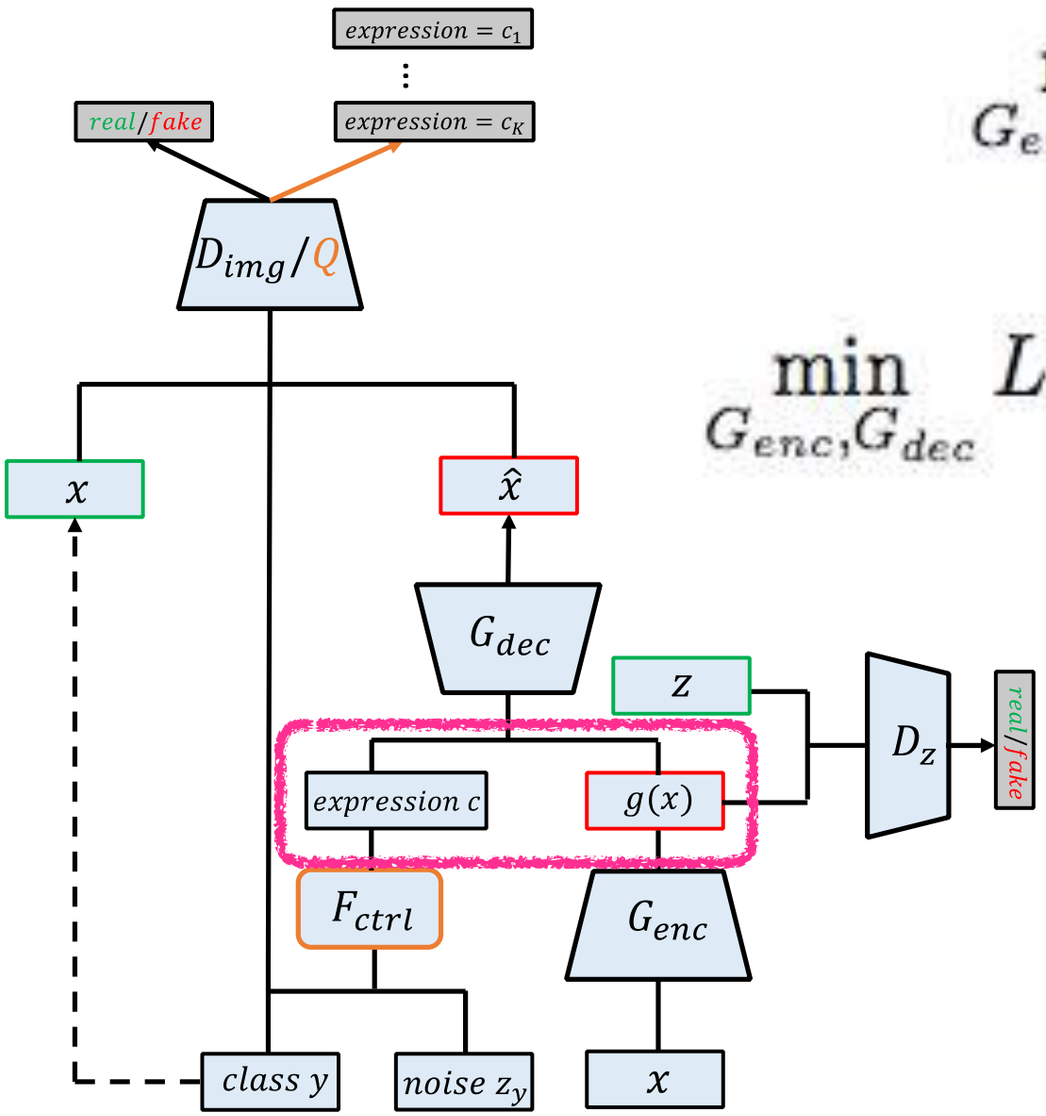
Generator Network: Encoder and Decoder



$$\min_{G_{enc},G_{dec}} L_{pixel} = L_1(G_{dec}(G_{enc}(x), c), x)$$

$$\min_{G_{enc},G_{dec}} L_{id} = \sum_l \beta_l L_1(\phi_l(G_{dec}(G_{enc}(x), c)), \phi_l(x))$$

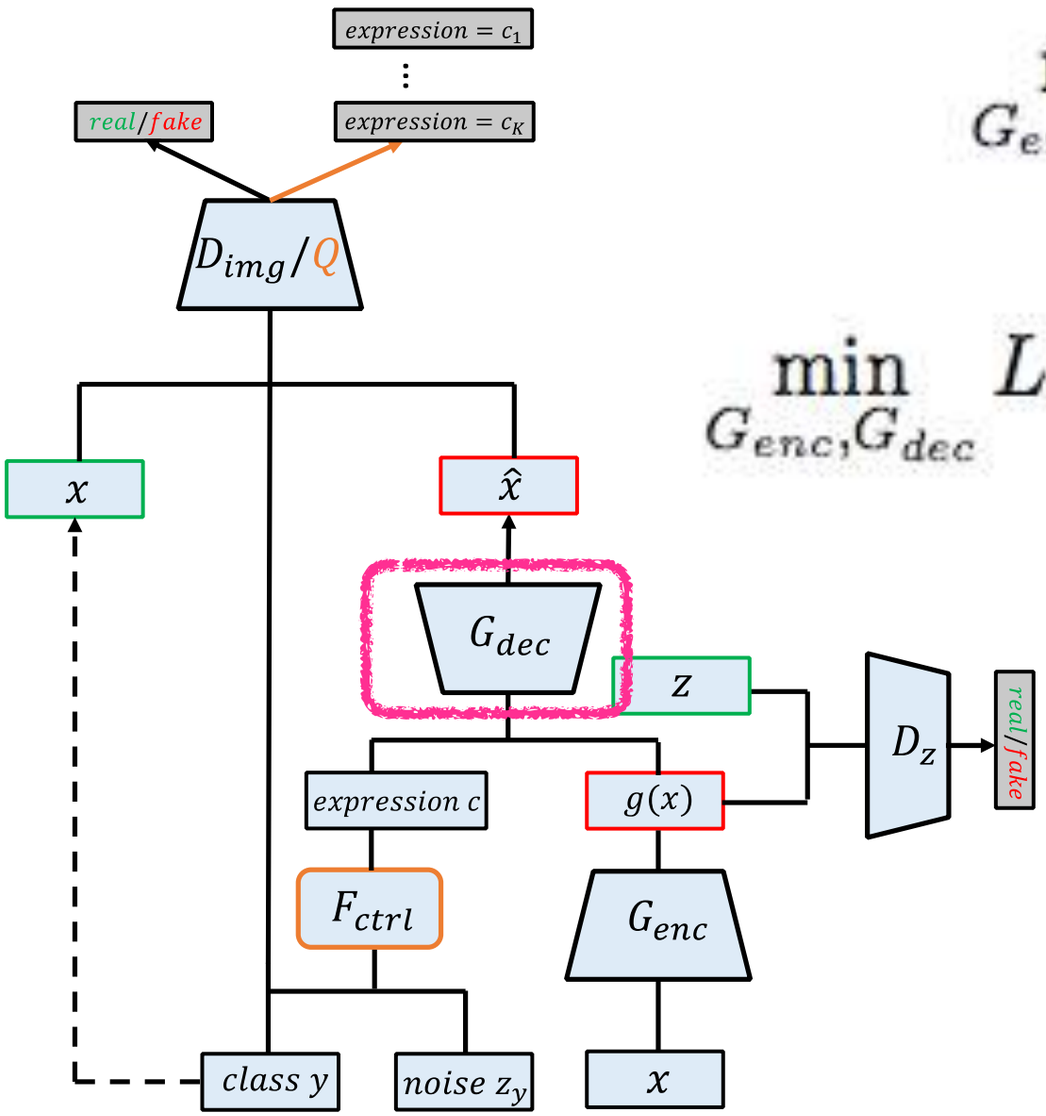
Generator Network: Encoder and Decoder



$$\min_{G_{enc},G_{dec}} L_{pixel} = L_1(G_{dec}(G_{enc}(x), c), x)$$

$$\min_{G_{enc},G_{dec}} L_{id} = \sum_l \beta_l L_1(\phi_l(G_{dec}(G_{enc}(x), c)), \phi_l(x))$$

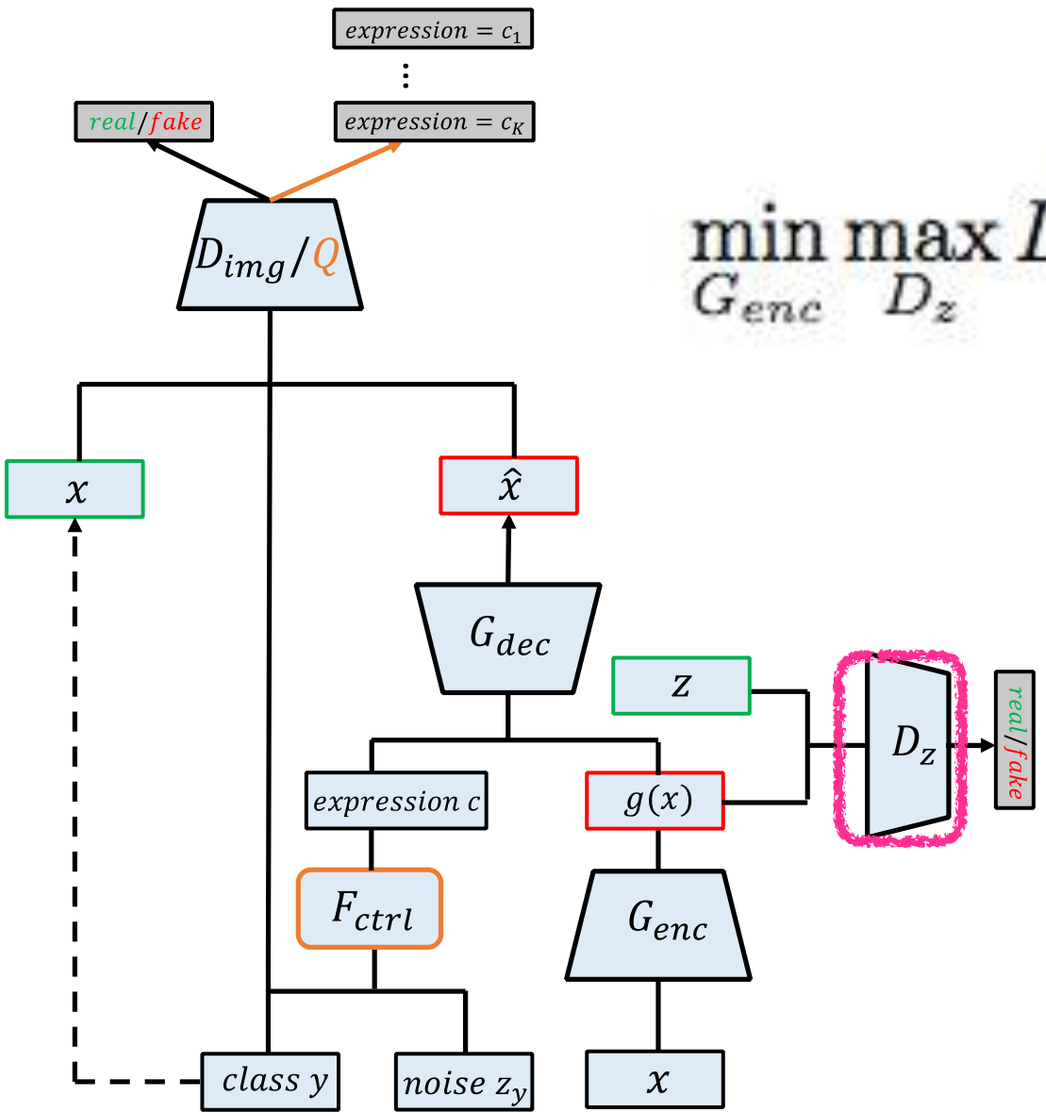
Generator Network: Encoder and Decoder



$$\min_{G_{enc},G_{dec}} L_{pixel} = L_1(G_{dec}(G_{enc}(x), c), x)$$

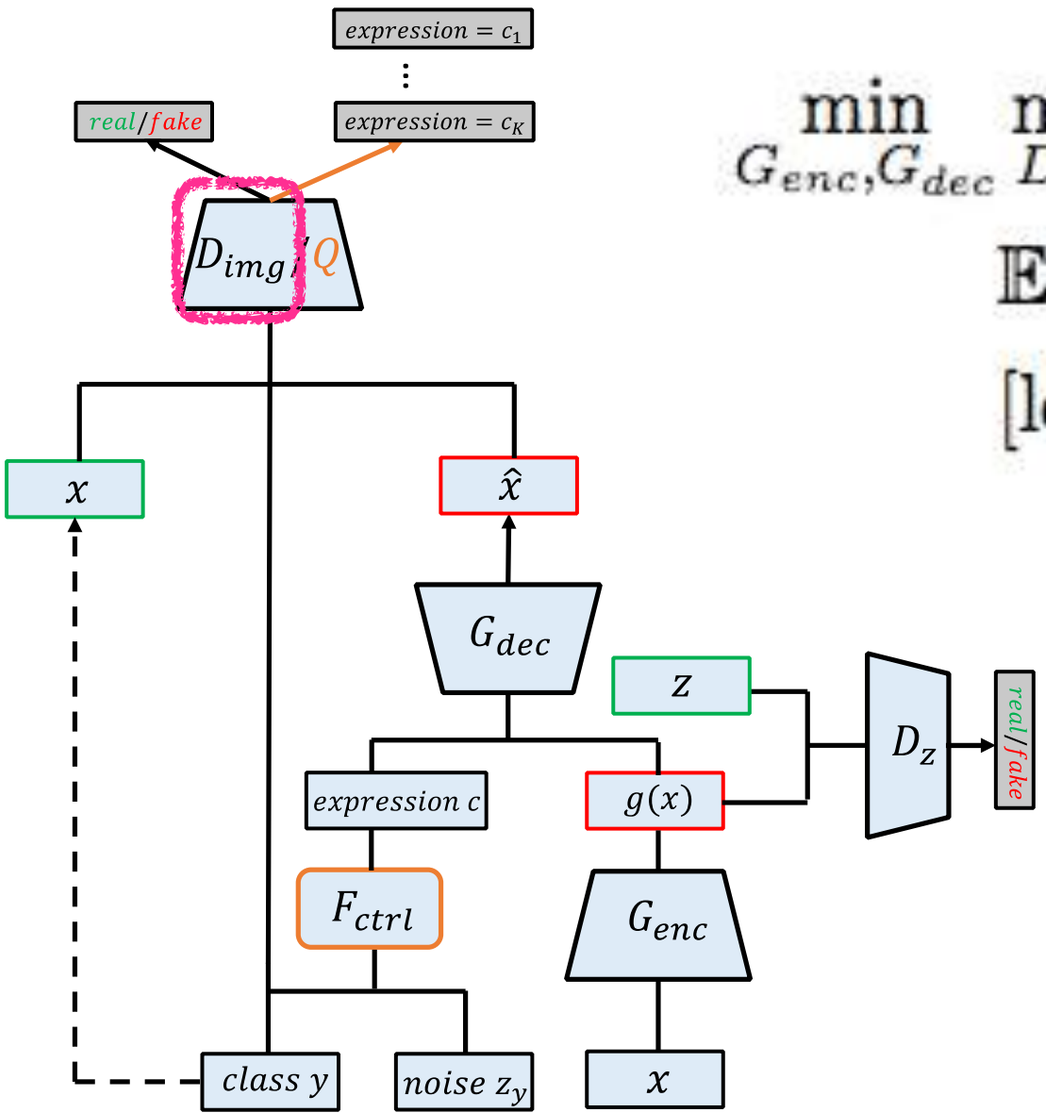
$$\min_{G_{enc},G_{dec}} L_{id} = \sum_l \beta_l L_1(\phi_l(G_{dec}(G_{enc}(x), c)), \phi_l(x))$$

Discriminator on Identity Representation



$$\min_{G_{enc}} \max_{D_z} L^z_{adv} = \mathbb{E}_{z \sim P_z(z)}[\log D_z(z)] + \mathbb{E}_{x \sim P_{data}(x)}[\log(1 - D_z(G_{enc}(x))]$$

Discriminator on Image



$$\min_{G_{enc},G_{dec}} \max_{D_{img}} L_{adv}^{img} = \mathbb{E}_{x,y\sim P_{data}(x,y)}[\log D_{img}(x,y)] +$$

$$\mathbb{E}_{x,y\sim P_{data}(x,y),z_y\sim P_{z_y}(z_y)}$$

$$[\log(1 - D_{img}(G_{dec}(G_{enc}(x), F_{ctrl}(z_y, y)), y))]$$

Overall Objective Function

$$\min_{G_{enc},G_{dec},Q} \max_{D_{img},D_z} L_{ExprGAN} = L_{pixel} + \lambda_1 L_{id} + \lambda_2 L_Q + \lambda_3 L_{adv}^{img} + \lambda_4 L_{adv}^{z} + \lambda_5 L_{tv}$$

Difficult to Train the Model with Limited Data

Mode Collapse

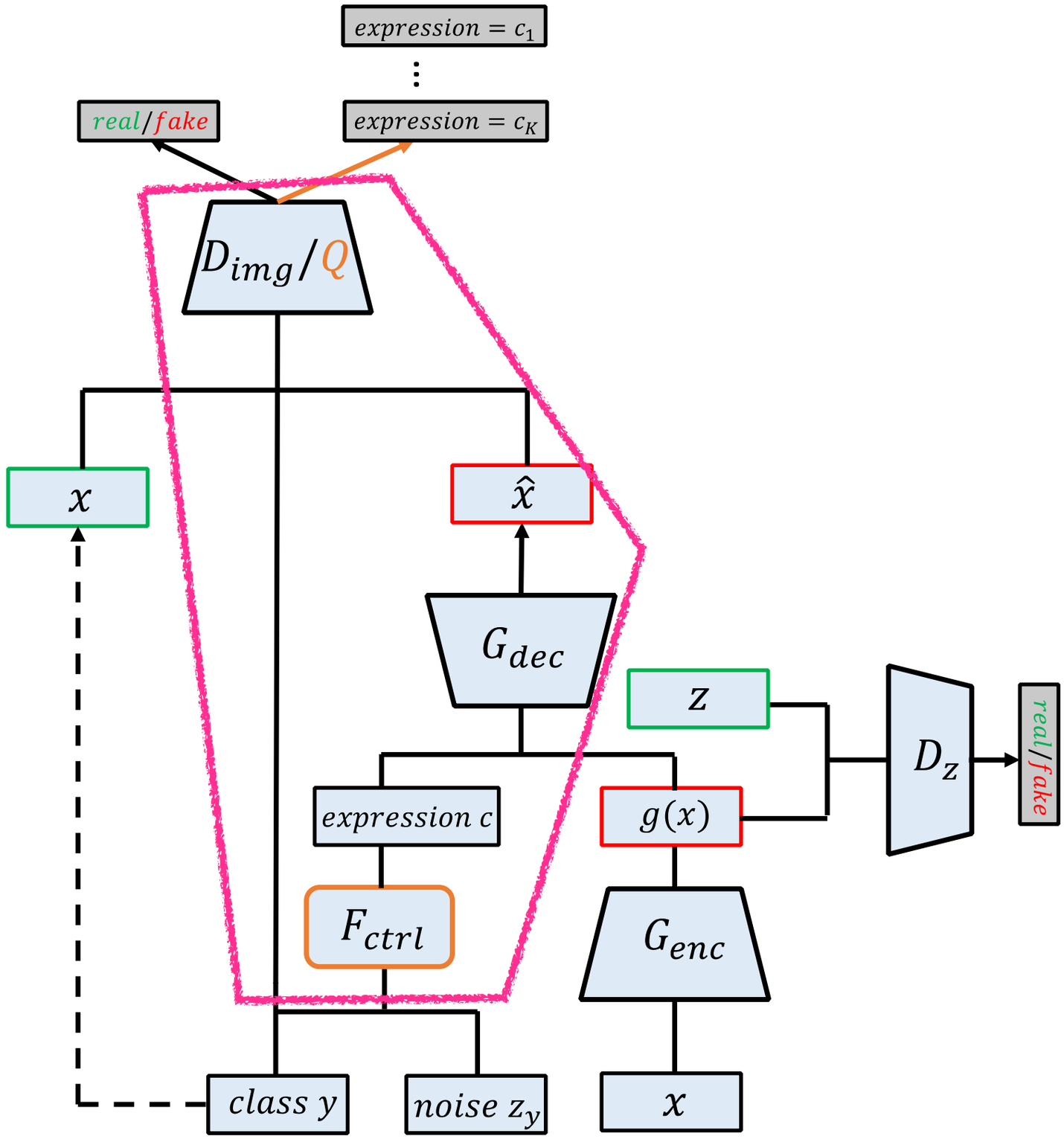
Limited Training Data?

Curriculum Training

Curriculum Training

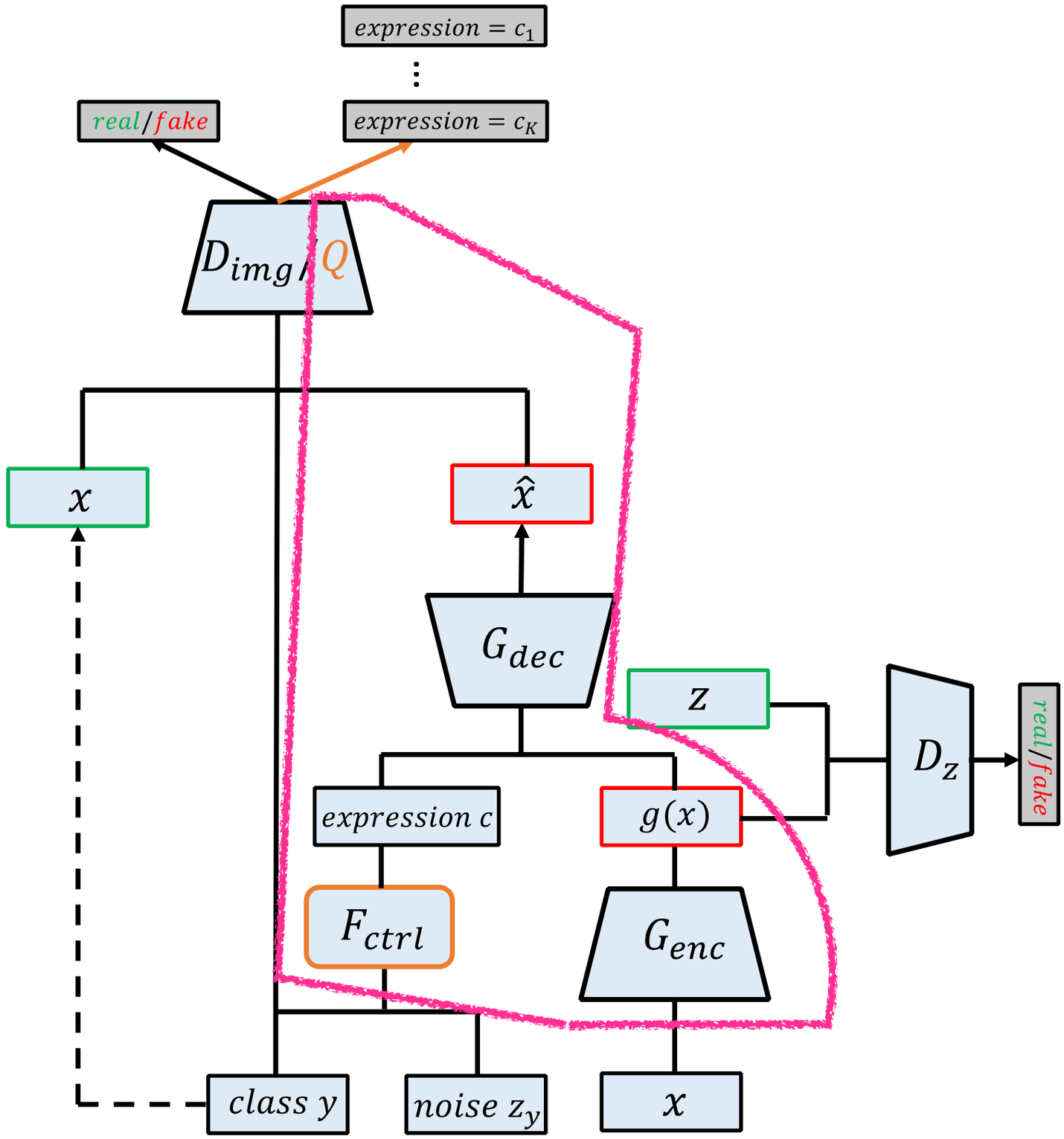
- Controller Learning
- Image Reconstruction
- Image Refinement

Controller Learning Stage



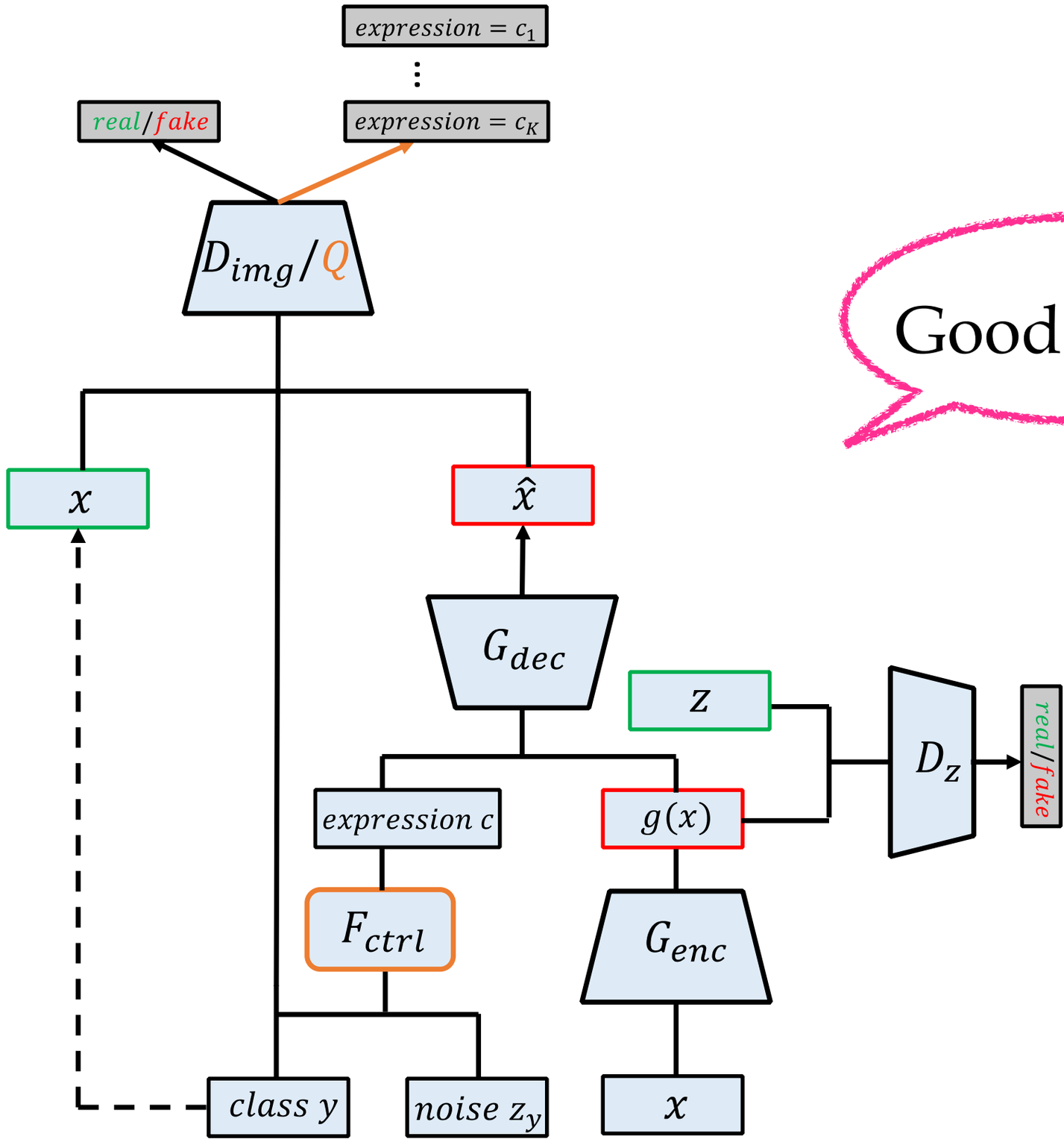
Images from Controller Learning Stage

Image Reconstruction Stage



Images from Reconstruction Stage

Image Refining Stage



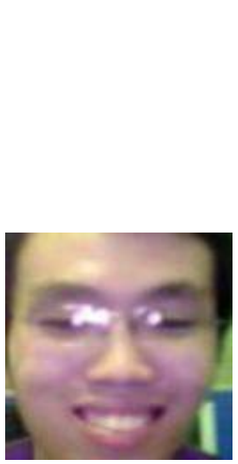
Good to Go!

Images from Refining Stage

Dataset

Dataset	Angry	Disgust	Fear	Happy	Sad	Surprise	Total
Oulu-CASIA	**240**	**240**	**240**	**240**	**240**	**240**	**1,440**

Expression Editing

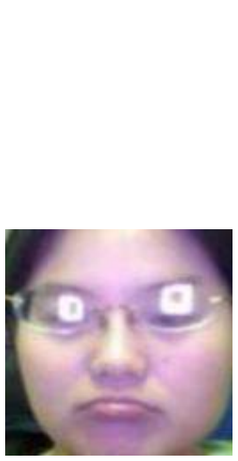


Happy

GT

ExprGAN

CAAE



Sad

GT

ExprGAN

CAAE

Zhang, Zhifei, Yang Song, and Hairong Qi. "Age Progression/Regression by Conditional Adversarial Autoencoder." CVPR (2017)

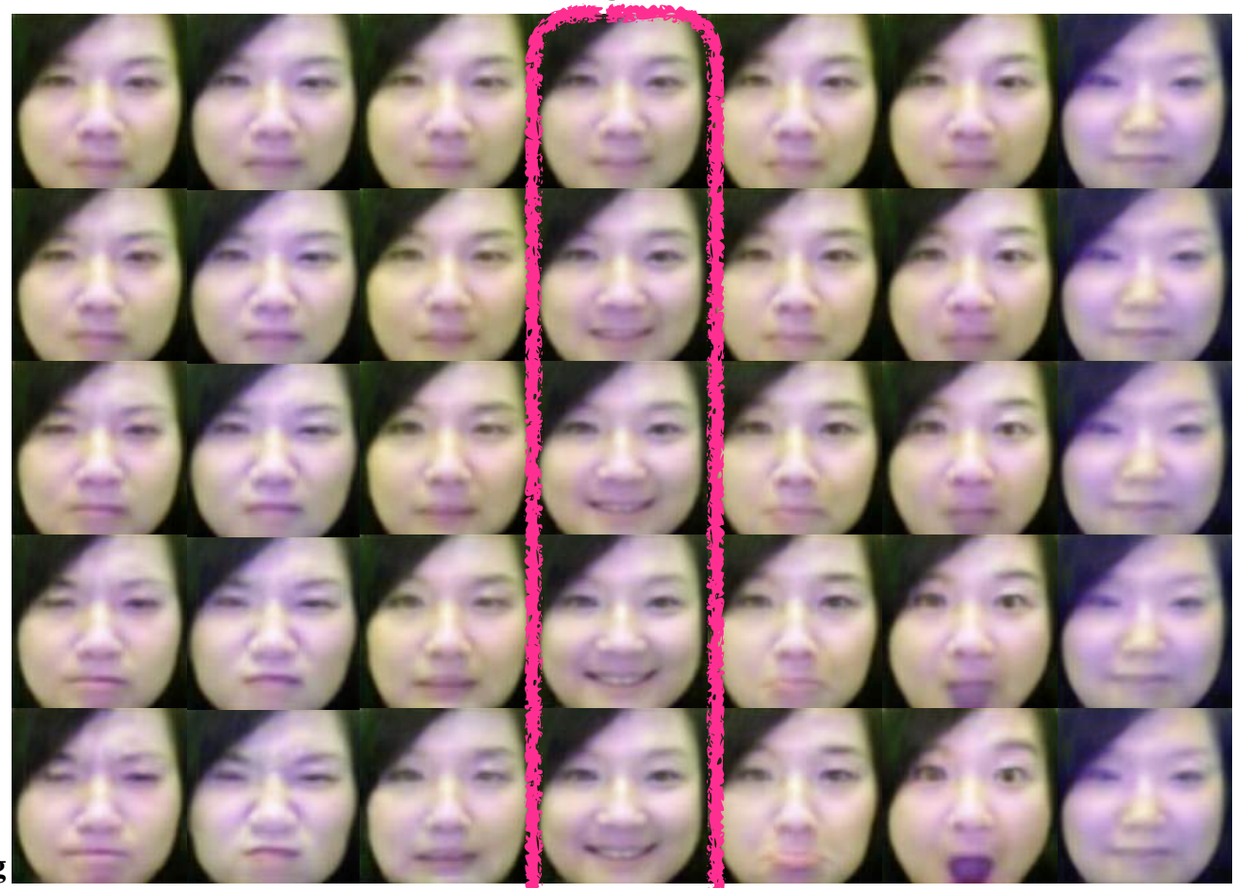
Expression Editing with Controllable Intensity

Disgust

Sad

Weak

Strong



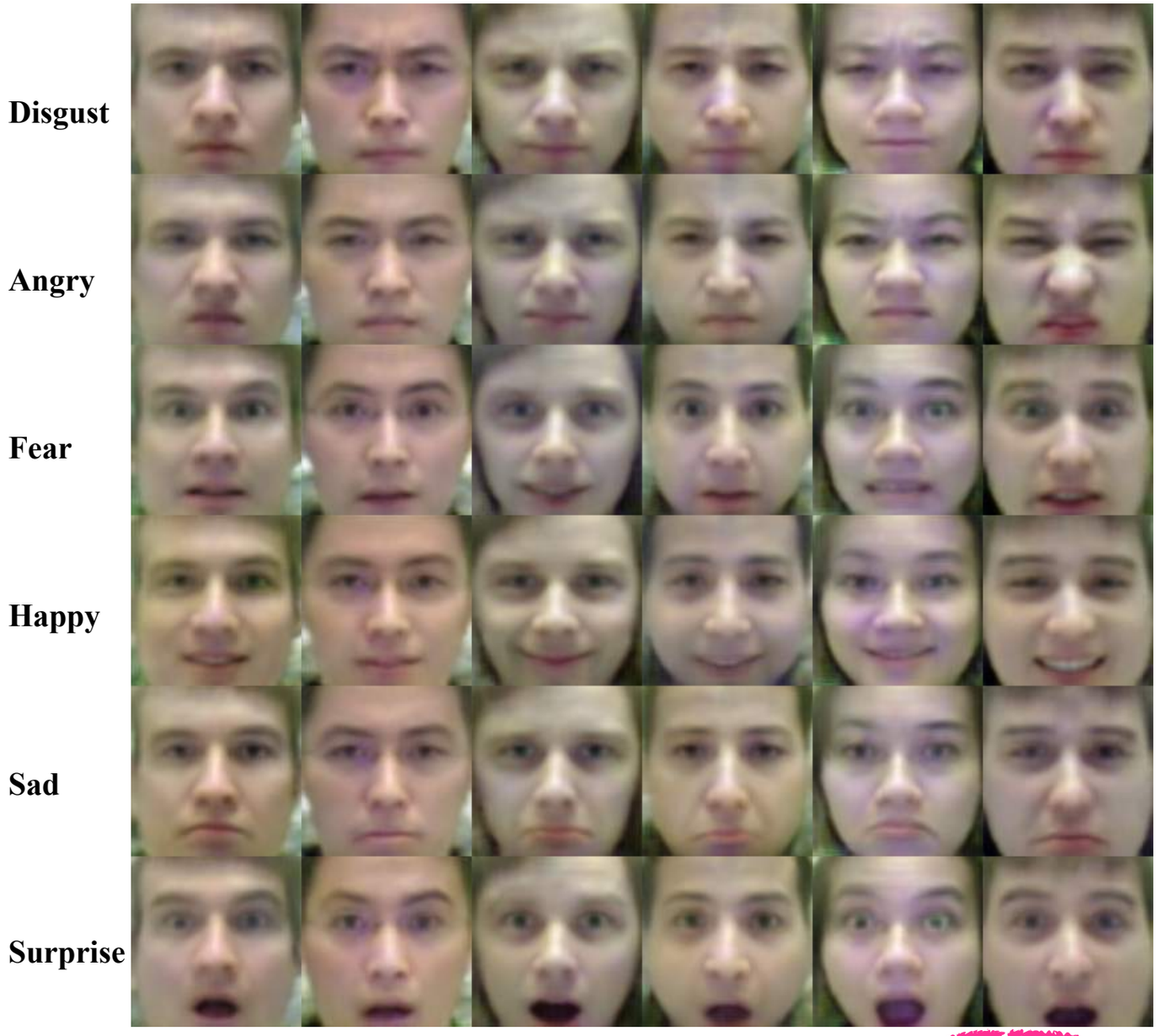
Angry Disgust Fear Happy Sad Surprise Neutral

Angry Disgust Fear Happy Sad Surprise Neutral

Expression Transfer

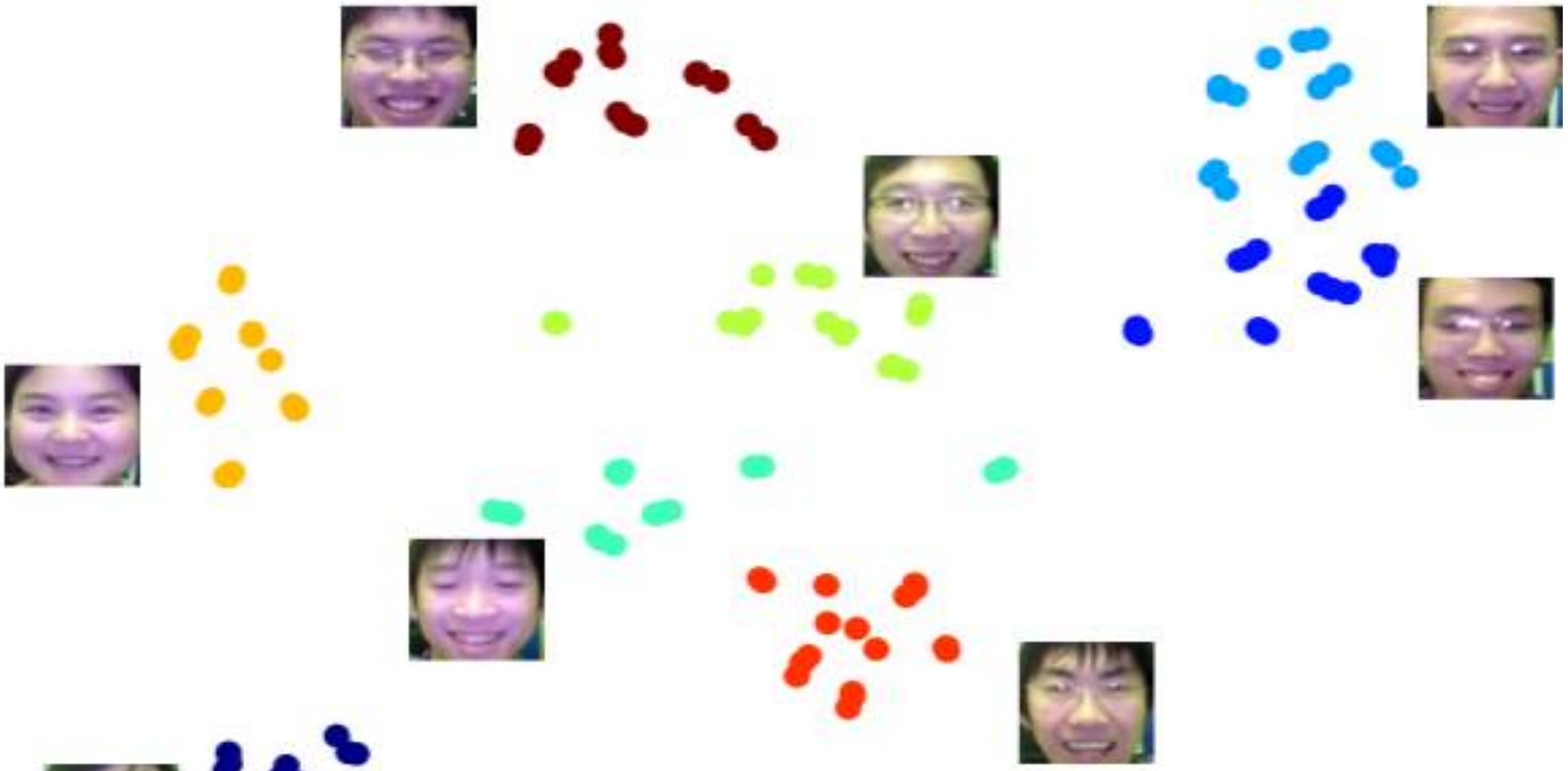
IdA ExprB IdA+ExprB IdA ExprB IdA+ExprB

Synthetic Images for Data Augmentation

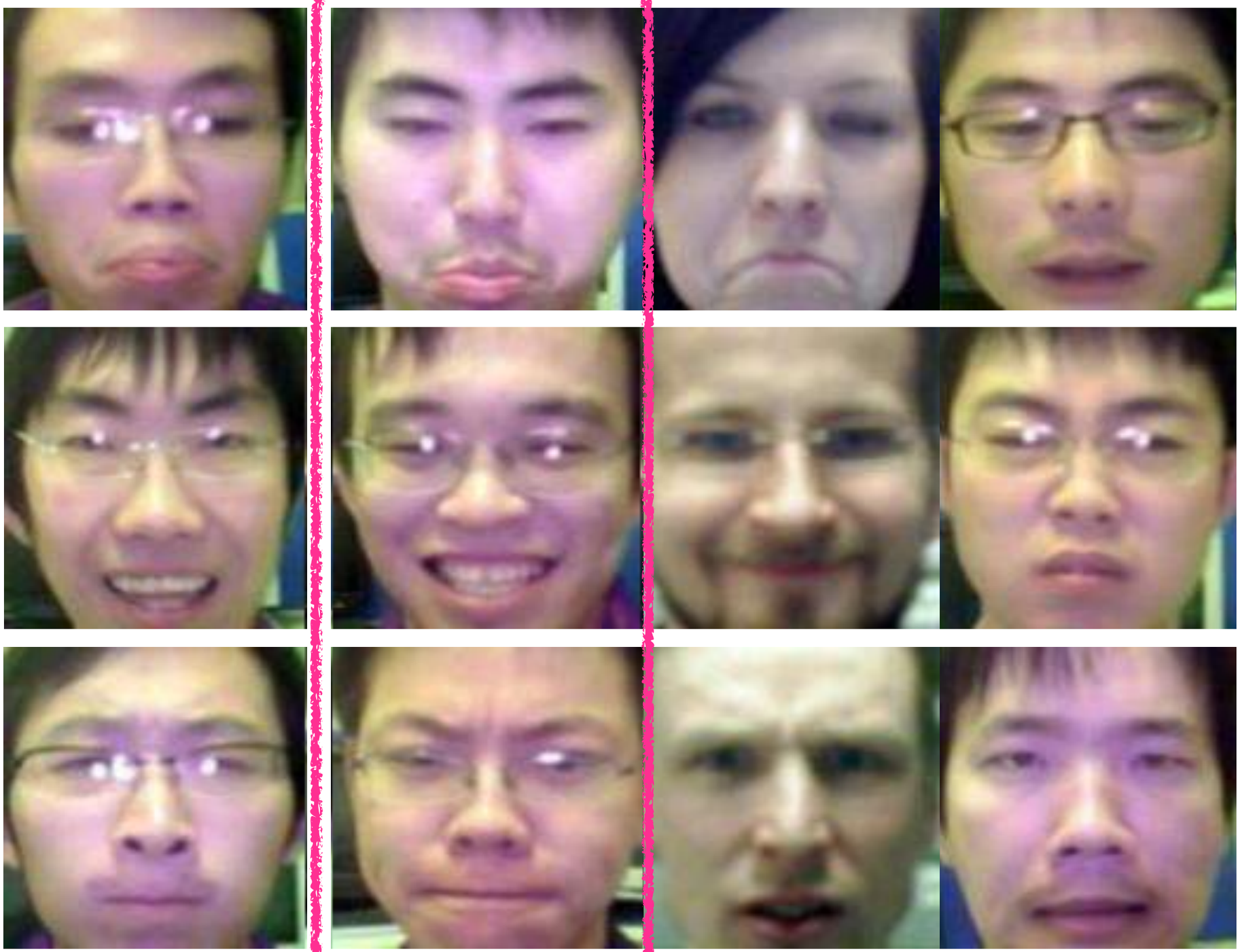


# Syn. Images	0	$3K$	$6K$	$30K$	$60K$
Accuracy (%)	77.78	78.47	81.94	**84.72**	84.72

Identity Feature Visualization



Expression Feature Visualization



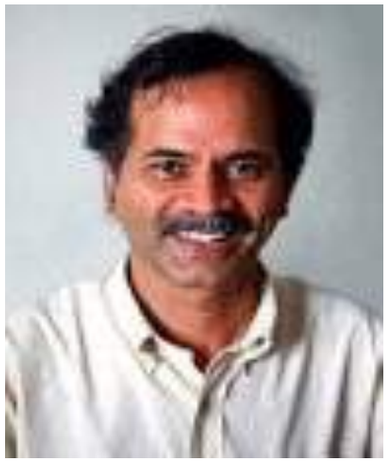
Summary

- ✦ Transfer Learning (Small Datasets)
  - FaceNet2ExpNet
- ✦ Robust Model Design (Occlusion, Pose)
  - Occlusion Robust Deep Network
  - Unaligned Attribute Classifier
- ✦ Generative Model (Fine-Grained)
  - ExprGAN

Future

Self-supervised learning will play an important role

Many Thanks to my Advisors, Collaborators and Friends



Professor **Rama Chellappa**
University of Maryland
College Park

Professor S. Kevin Zhou
Chinese Academy of Sciences
previously Siemens Healthineers

Dr. Kumar Sricharan
Intuit
previously Palo Alto
Research Center

Dr. Haoxiang Li
Wormpex AI Research
previously Adobe Research

Dr. Qian Yu
qcraft.ai
previously Waymo

Thanks my family

Publications

- **Hui Ding**, Peng Zhou and Rama Chellappa, "Occlusion Adaptive Deep Network for Robust Facial Expression Recognition", submitted to IJCB 2020

- **Hui Ding**, Jingjing Zheng and Rama Chellappa, "Facial Region-based Attention Network for Unaligned Expression Recognition", submitted to IJCB 2020

- **Hui Ding**, Hao Zhou, Shaohua Kevin Zhou and Rama Chellappa, "A Deep Cascade Network for Unaligned Face Attribute Classification", Association for the Advancement of Artificial Intelligence (AAAI), 2018.

- **Hui Ding**, Kumar Sricharan and Rama Chellappa, "ExprGAN: Facial Expression Editing with Controllable Expression Intensity", Association for the Advancement of Artificial Intelligence (AAAI), 2018

- **Hui Ding**, Shaohua Kevin Zhou and Rama Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition", IEEE International Conference on Automatic Face Gesture Recognition (FG), 2017

Thank you!

Codes & Models: www.huiding.org