Personalized Facial Action Unit Detection Using Multi-task Network Cascades

Cheng-Hao Tu

Advisor: Jane Yung-jen Hsu, Ph.D.

2018.01.09

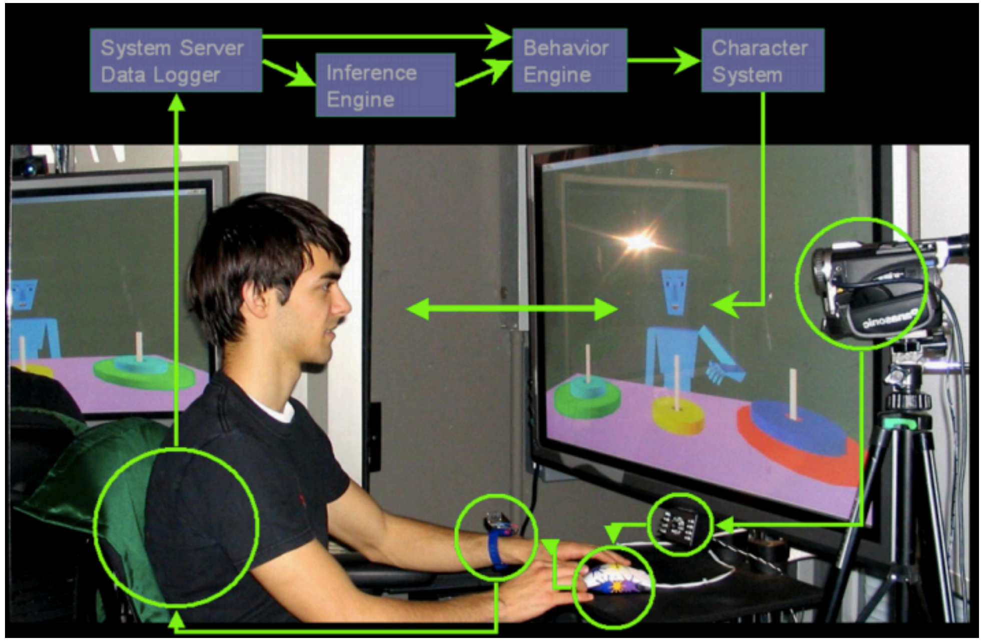
Outline

- Introduction and Motivation
- Related Work
- Problem Statement
- Methodology
- Experiments
- Conclusion

Introduction and Motivation

Motivation

- Facial Expression is a fast and natural non-verbal channel conveying our emotions and intentions.
- Machines can adjust the provided services according to users' current emotions.



A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," IJHCS, vol. 65, no. 8, pp. 724–736, 2007.

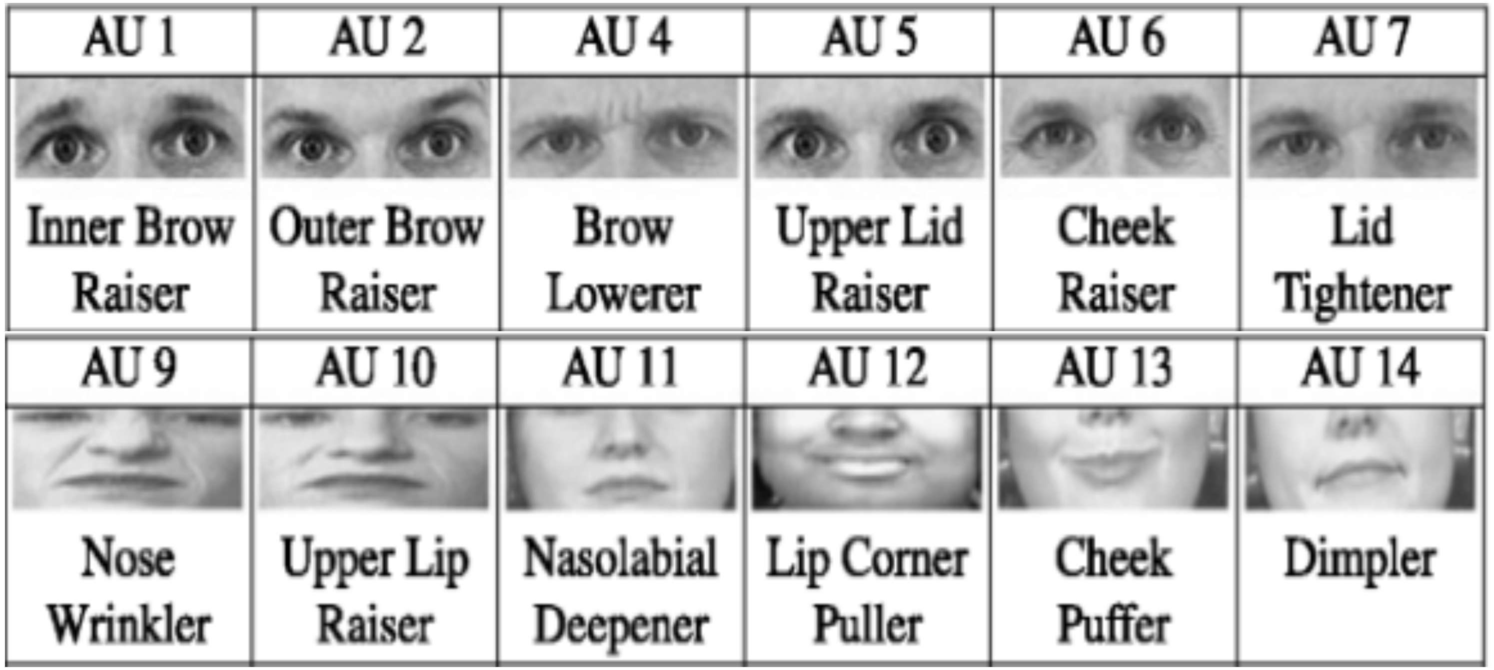
Emotion Recognition

Convolutional Neural Networks → Happy

- Require enormous independent annotated data for each emotion.

Facial Action Coding System

- Muscle contractions of facial parts are defined as Facial Action Units (AUs) that describe more than 7,000 observed facial expressions



Facial Action Coding System

- Muscle contractions of facial parts are defined as Facial Action Units (AUs) that describe more than 7,000 observed facial expressions

Emotion Recognition From AUs

AU Detection → AU6 AU12

1. Emotional FACS Rules
2. Machine Learning

→ Happy

- Combine with Human Knowledge
- Dimension Reduction

Emotion Recognition From AUs

AU Detection

AU6
AU12

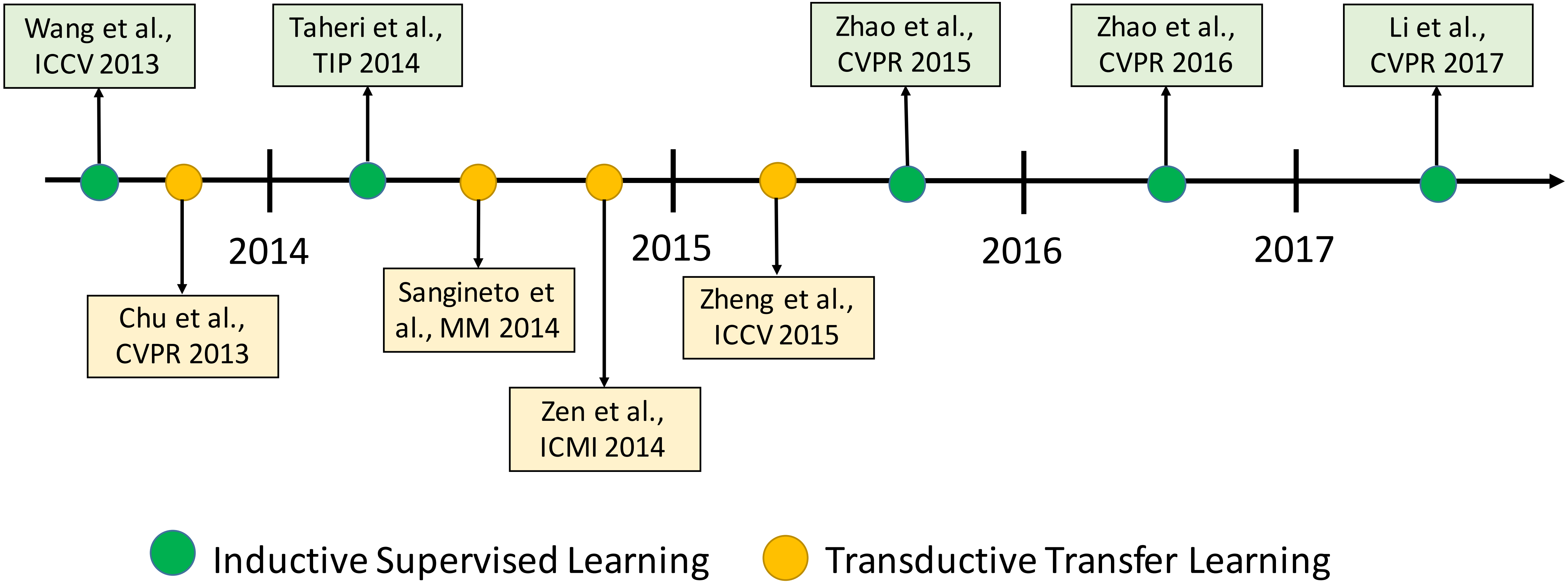
1. Emotional FACS Rules
2. Machine Learning

Happy

- 6 months of FACS training for a coder
- Coding 1 minutes of video takes over an hour

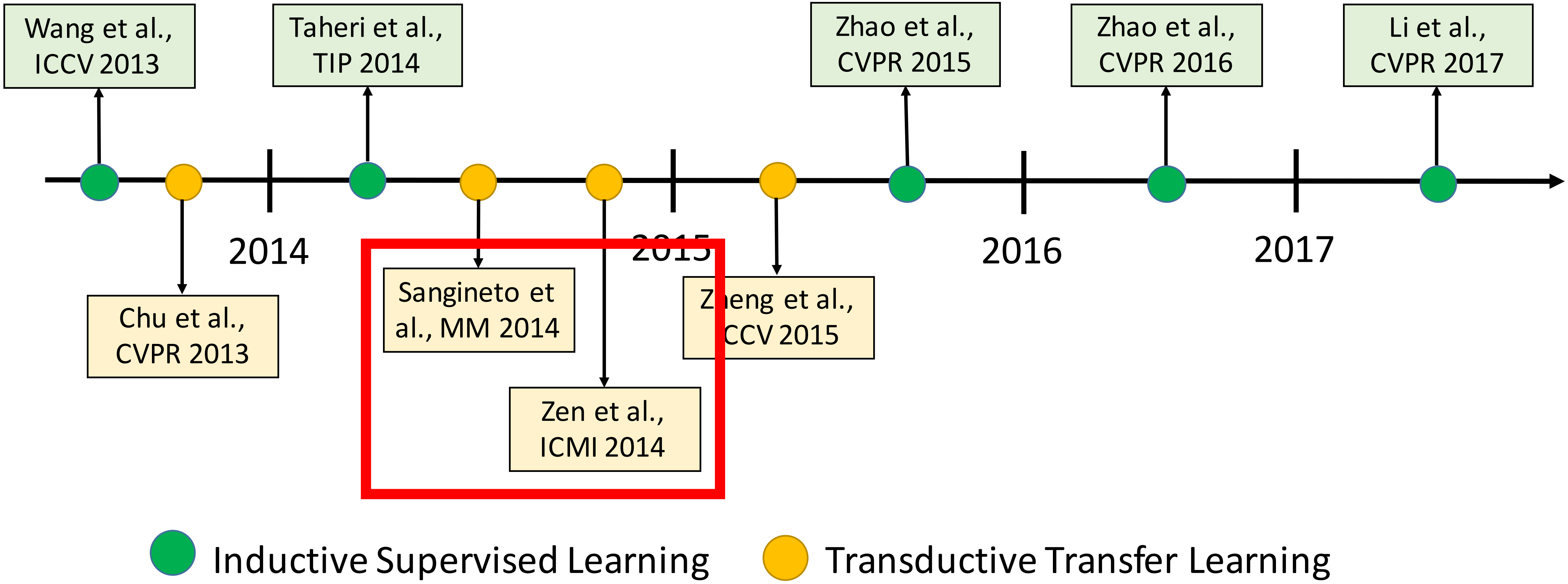
Related Work

Timeline



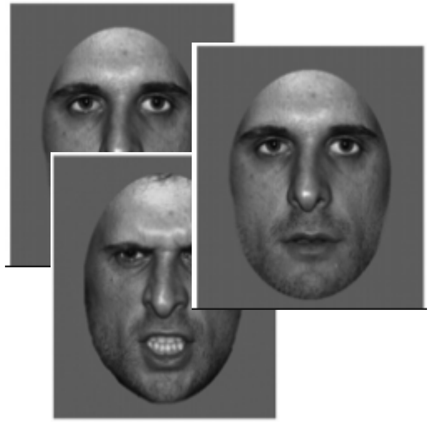
Wang et al., ICCV 2013
Taheri et al., TIP 2014
Zhao et al., CVPR 2015
Zhao et al., CVPR 2016
Li et al., CVPR 2017
2014
2015
2016
2017
Chu et al., CVPR 2013
Sangineto et al., MM 2014
Zen et al., ICMI 2014
Zheng et al., ICCV 2015
Inductive Supervised Learning
Transductive Transfer Learning

Timeline



Wang et al., ICCV 2013
Taheri et al., TIP 2014
Zhao et al., CVPR 2015
Zhao et al., CVPR 2016
Li et al., CVPR 2017
2014
2015
2016
2017
Chu et al., CVPR 2013
Sangineto et al., MM 2014
Zen et al., ICMI 2014
Z eng et al.,
CCV 2015
Inductive Supervised Learning
Transductive Transfer Learning

Support Vector Transductive Parameter Transfer (SVTPT)



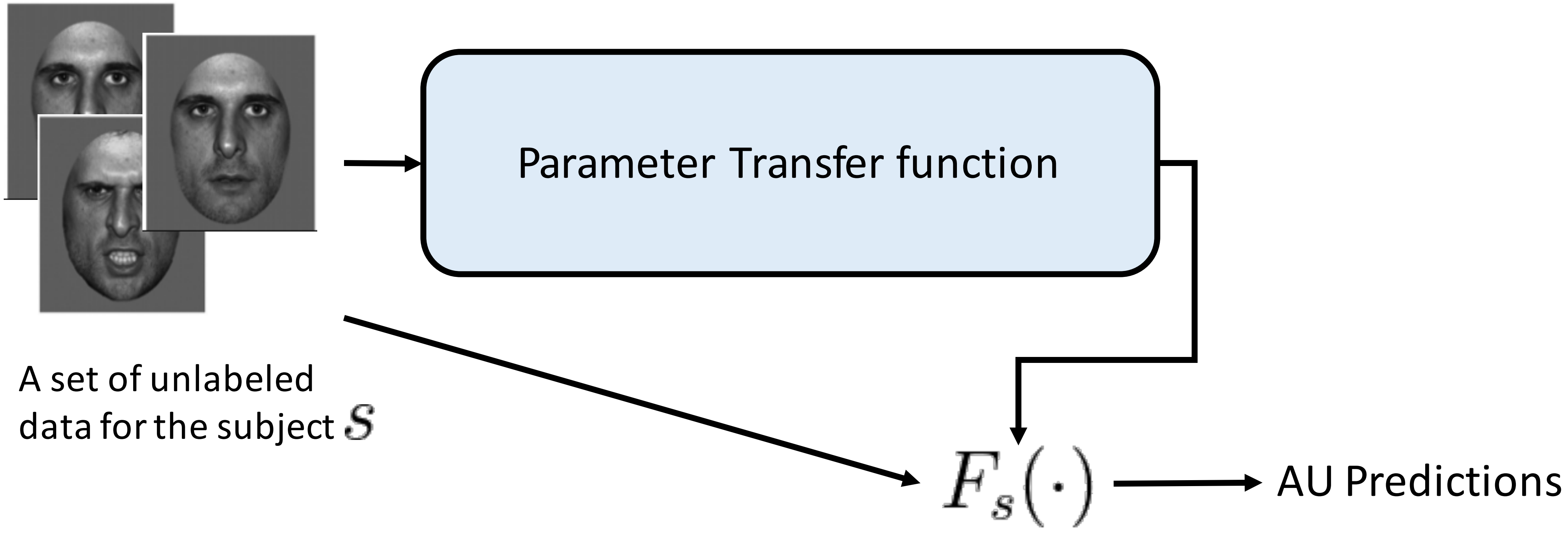
Parameter Transfer function

$\rightarrow F_s(\cdot)$

A set of unlabeled data for the subject S

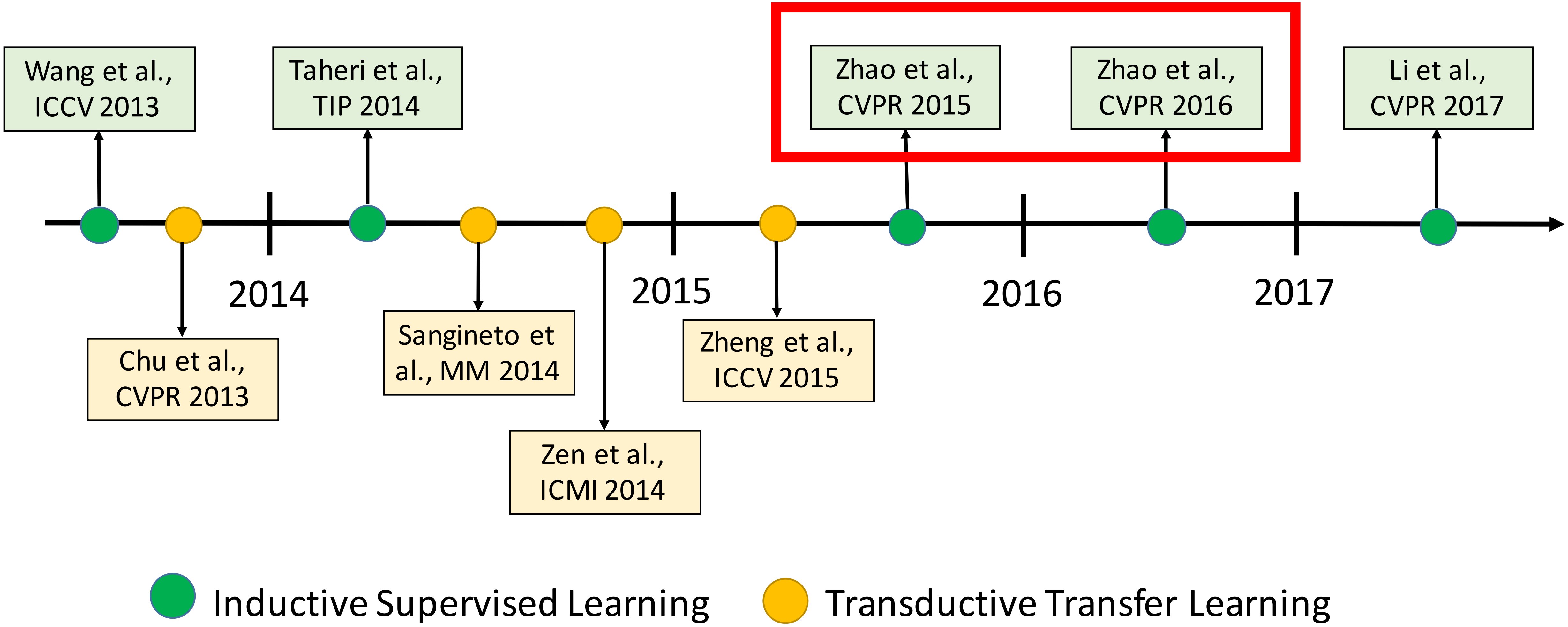
G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In Proc. ICMI, 2014.

Support Vector Transductive Parameter Transfer (SVTPT)



G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In Proc. ICMI, 2014.

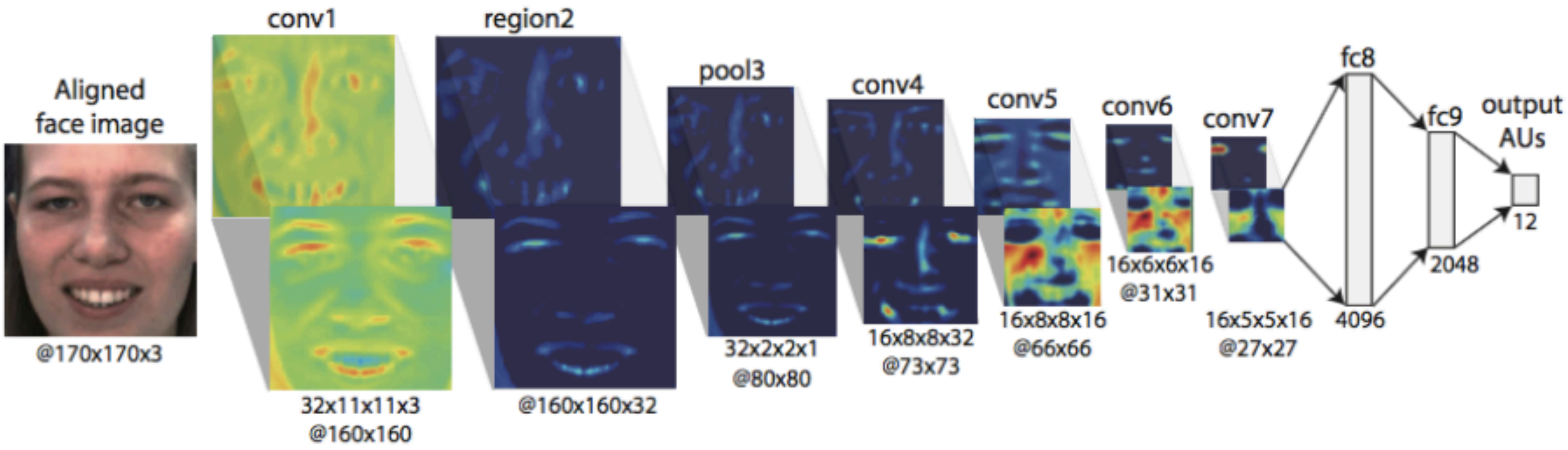
Timeline



Wang et al., ICCV 2013
Chu et al., CVPR 2013
2014
Taheri et al., TIP 2014
Sangineto et al., MM 2014
Zen et al., ICMI 2014
2015
Zheng et al., ICCV 2015
Zhao et al., CVPR 2015
2016
Zhao et al., CVPR 2016
2017
Li et al., CVPR 2017
Inductive Supervised Learning
Transductive Transfer Learning

Deep Region and Multi-label Learning (DRML)

- The 2016 state-of-the-art method that adopts deep neural networks for AU detection



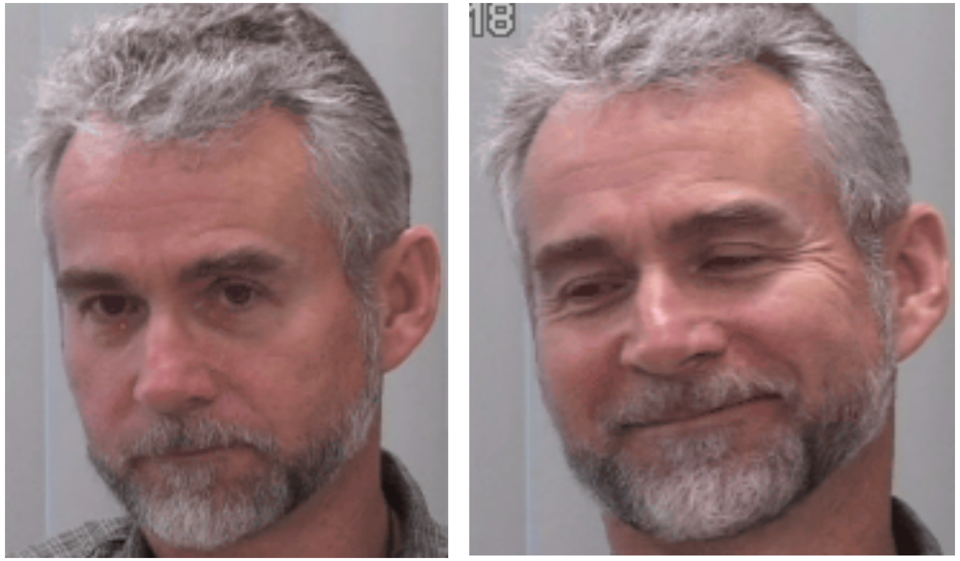
K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In Proc. CVPR, 2016.

Problem Statement

Individual Differences

- Appearance of AUs vary with facial shapes, ages and races, which makes AU detection challenging.

AU15: Lip Corner Depressor occurs?



AU 6 : Cheek Raiser occurs?
AU12: Lip Corner Puller occurs ?

Performance Drops

- Train DRML on the BP4D dataset (teenagers), and Test on the McMaster-UNBC dataset (adults and elders)

F1 Score	AU4	AU6	AU7	AU10	AU12
BP4D	0.416	0.766	0.719	0.807	0.823
McMaster-UNBC	0.046	0.246	0.134	0.027	0.310
Decreasing Rate	88.94%	68.02%	81.36%	78.00%	62.33%

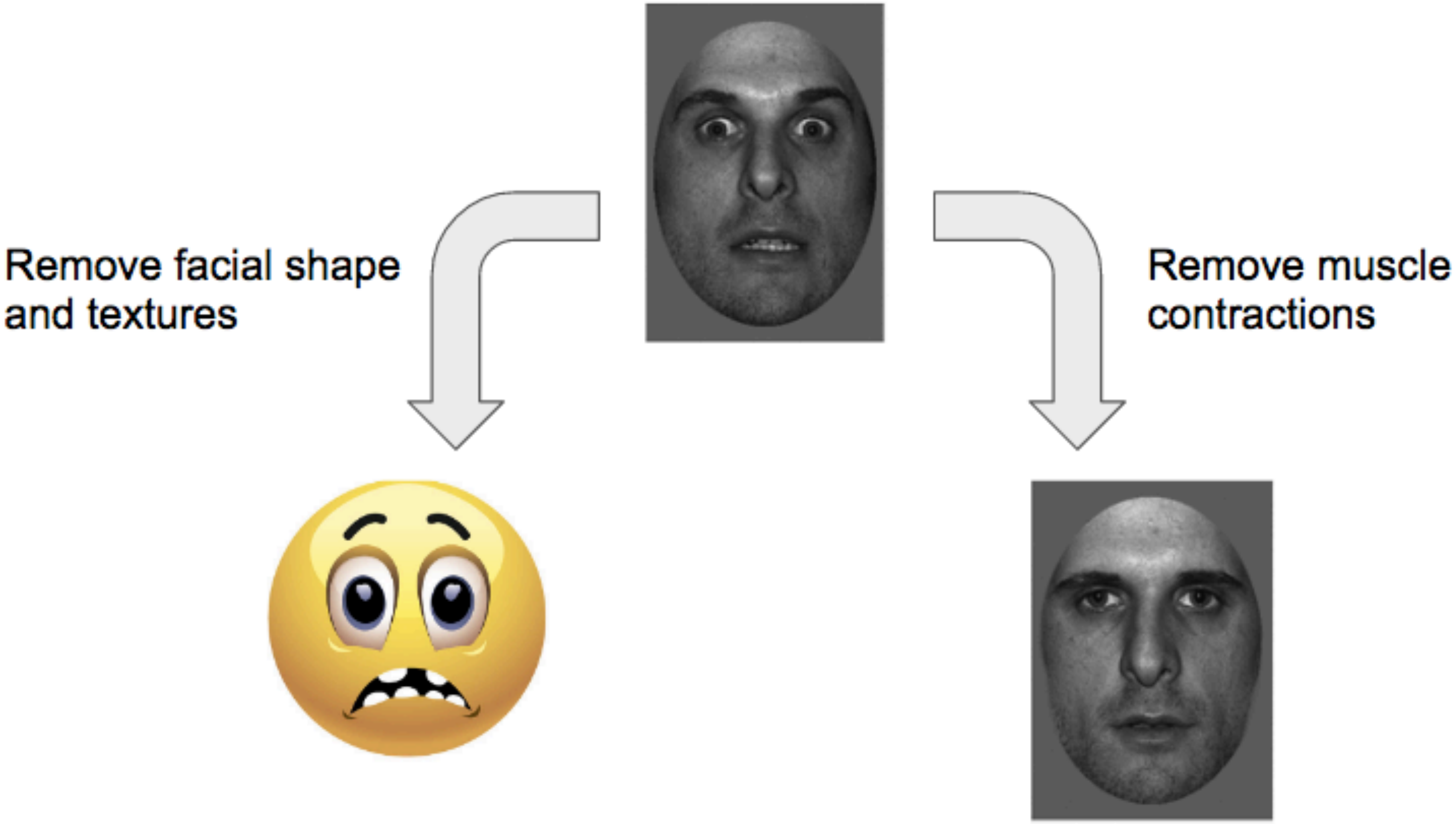
We retrain models using the codes provided by the author: https://github.com/zkl20061823/DRML

Lack of Subject Variations in AU Datasets

Dataset Name	Labels	Number of Subjects	Number of Samples
AMFED	AUs, Interest	<= 242	242 videos (1 mins)
DISFA	AUs	27	27 videos (4 mins)
BP4D	AUs	41	328 videos (with 148562 frames)
UNBC-McMaster	AUs, Pain	25	200 videos (with 48398 frames)

- Annotating AUs for huge amount of data takes a lot of time

Components in Expressional Faces

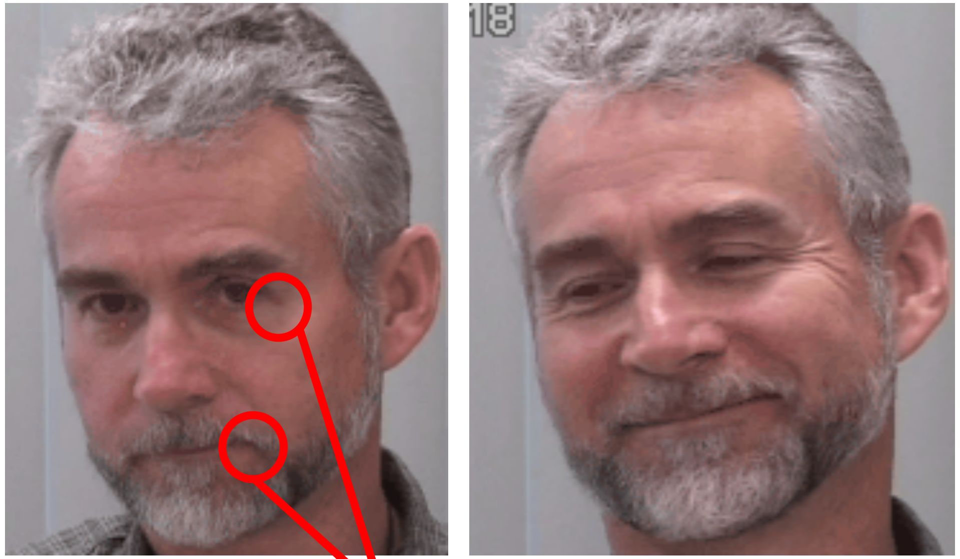


Utilizing Neutral Faces

- Neutral faces are suitable to describe ones' personal appearance features

AU15: Lip Corner Depressor occurs

Subject's lip drops naturally



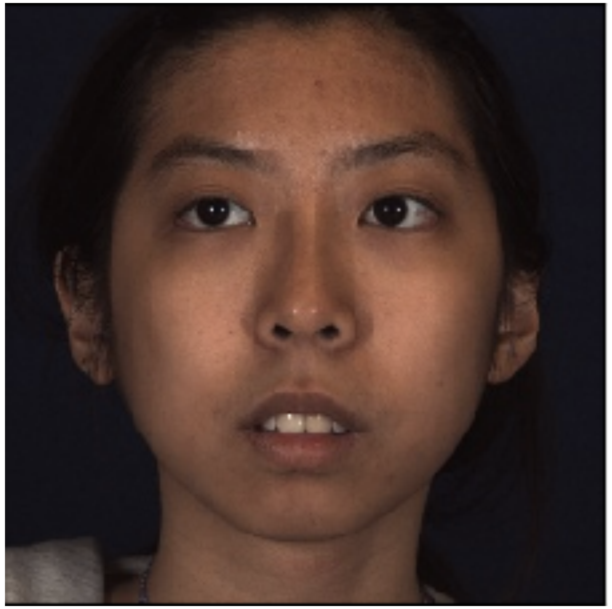
Mustache and Wrinkles

Problem Definition

- Given: A set $\mathcal{Y}$ of AUs that we would like to detect.
- Input: A face image $\mathbf{x}_{main}$ and a neutral face image $\mathbf{x}_{aux}$, and both face images have a common identity.
- Output: For each AUs in $\mathcal{Y}$, whether the AU occurs in the face image $\mathbf{x}_{main}$

Problem Definition (An Example)

- Given: A set $\mathcal{Y} = \{\mathrm{AU}_{12}, \mathrm{AU}_{17}\}$
- Input: $\mathbf{x}_{main}$ $\mathbf{x}_{aux}$

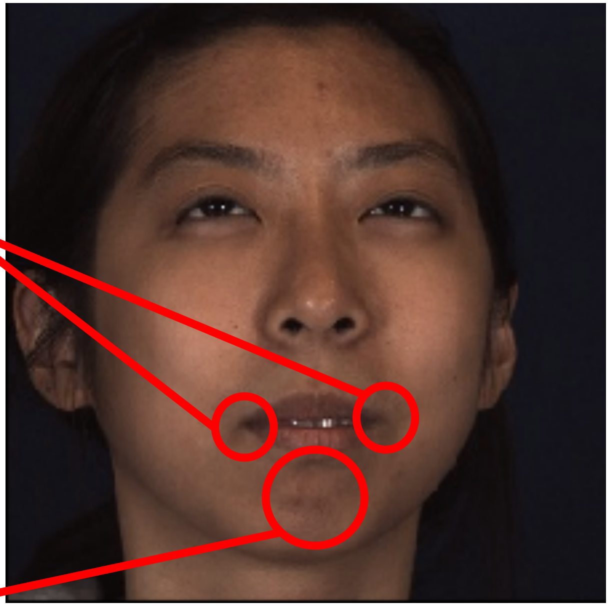


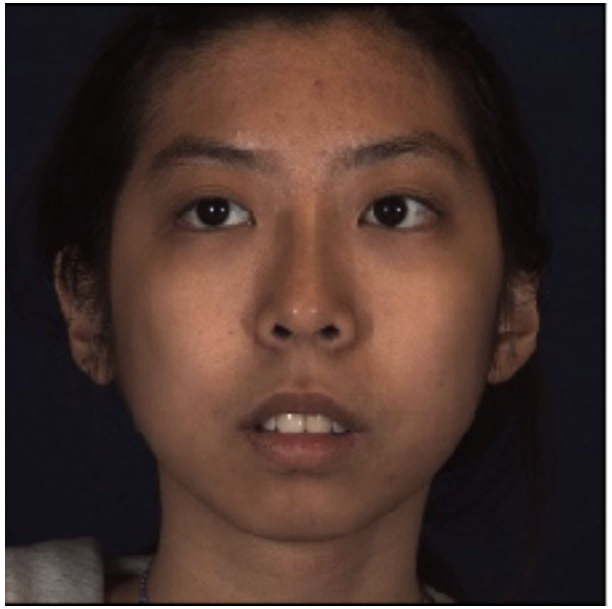
Problem Definition (An Example)

- Given: A set $\mathcal{Y} = \{\mathrm{AU}_{12}, \mathrm{AU}_{17}\}$
- Output: $\mathbf{x}_{main}$ $\mathbf{x}_{aux}$

AU12: Lip Corner Puller Occurs

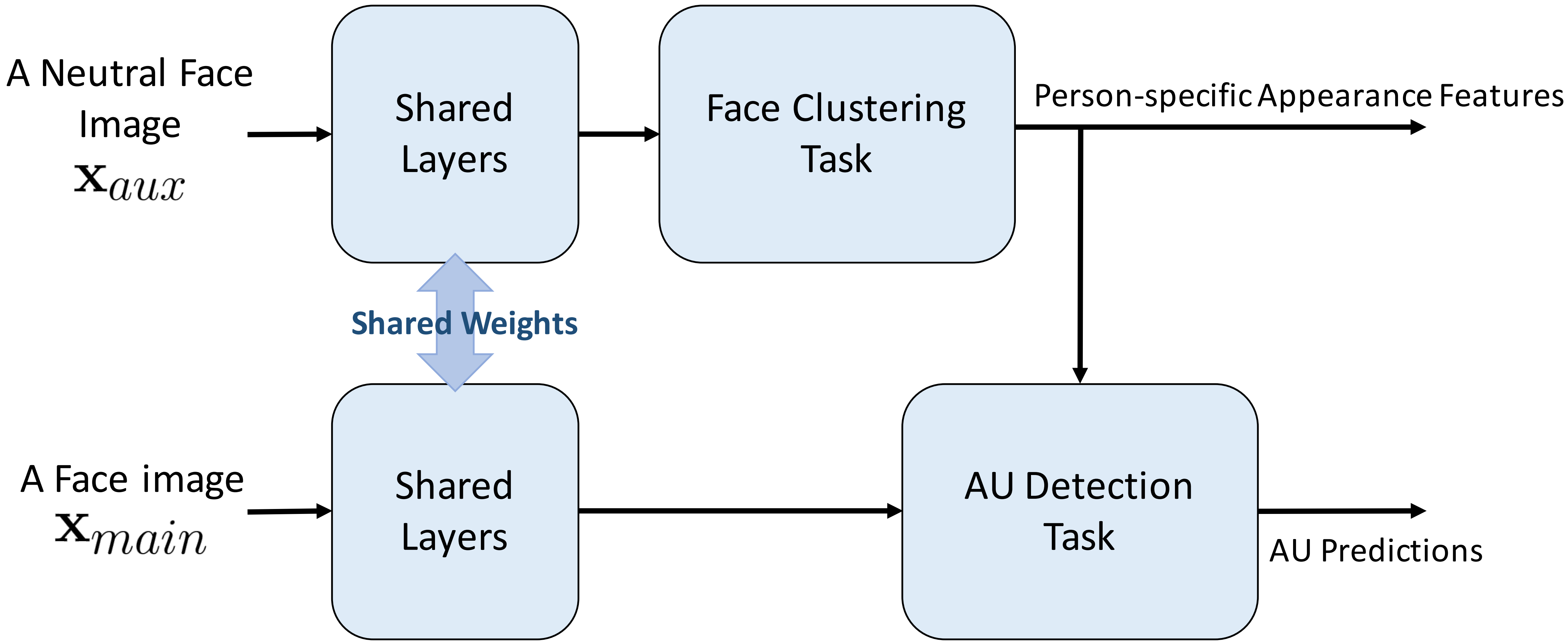
AU17: Chin Raiser Occurs

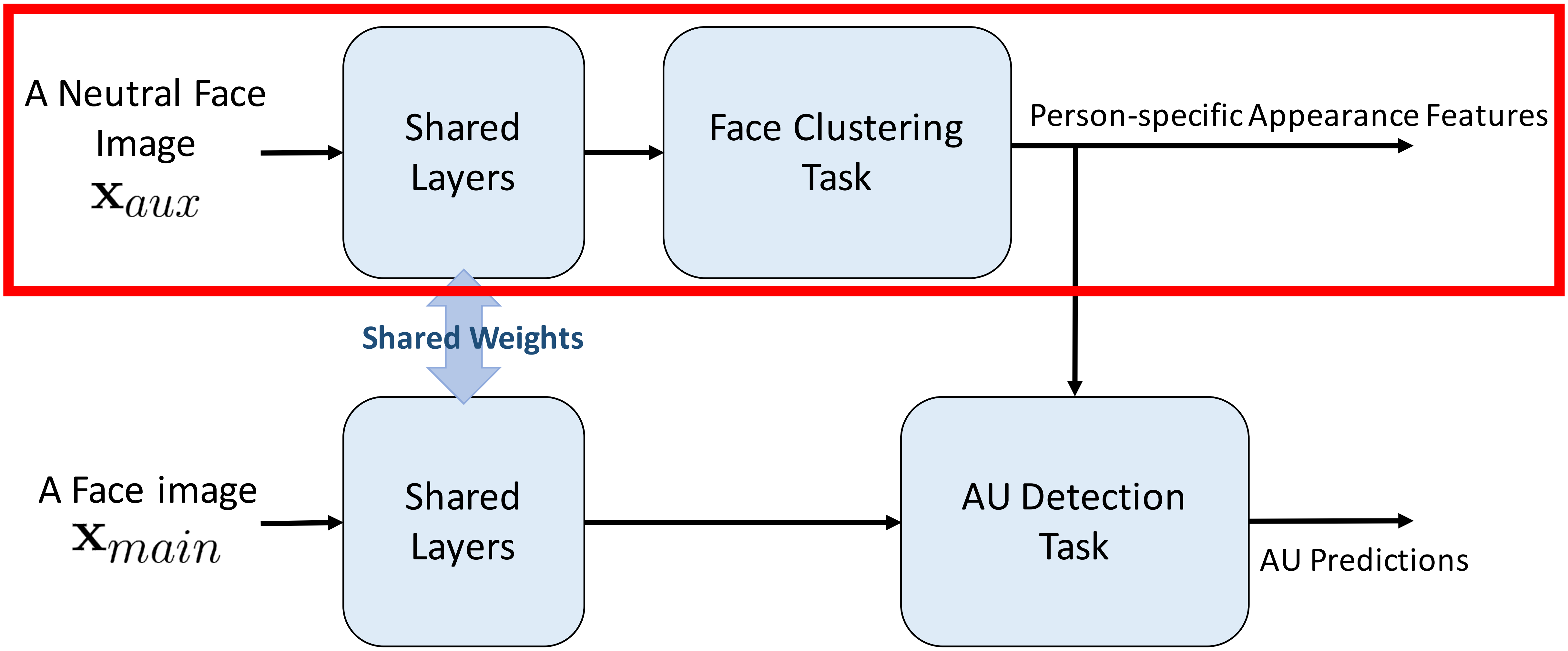




Methodology

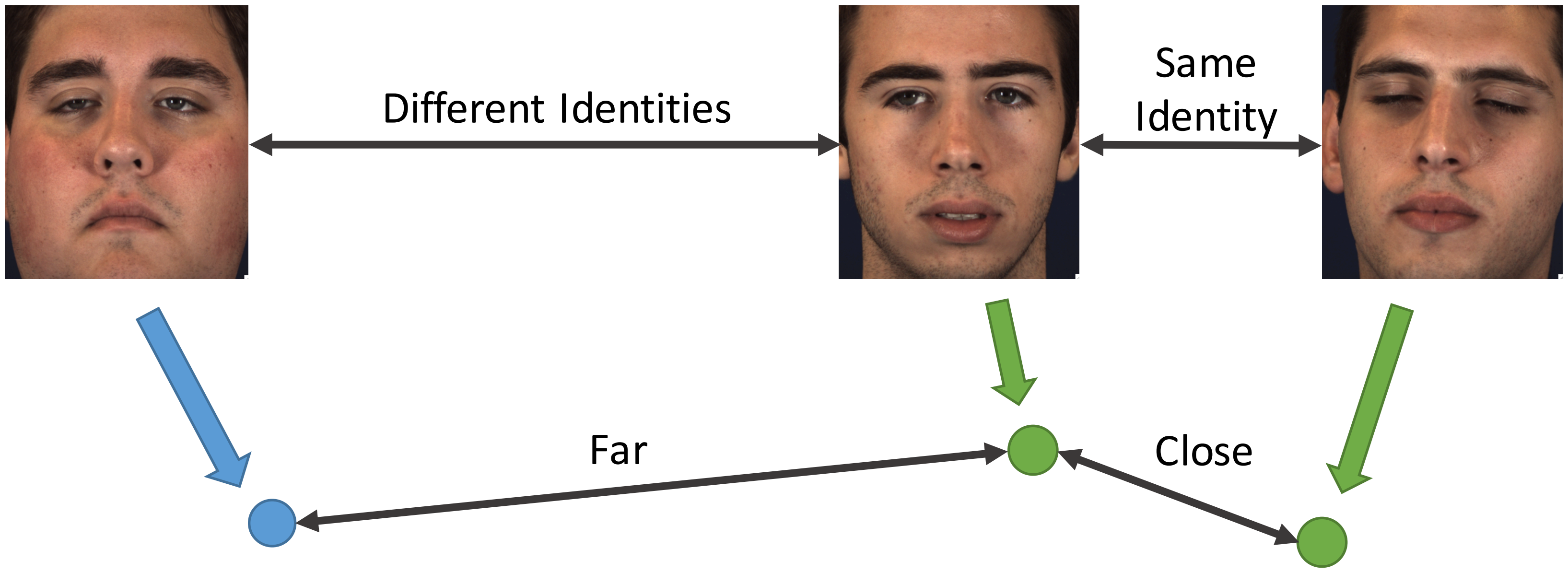
Structures of Network Cascades




A Neutral Face Image
$\mathbf{x}_{aux}$
Shared Layers
Face Clustering Task
Person-specific Appearance Features
Shared Weights
A Face image
$\mathbf{x}_{main}$
Shared Layers
AU Detection Task
AU Predictions

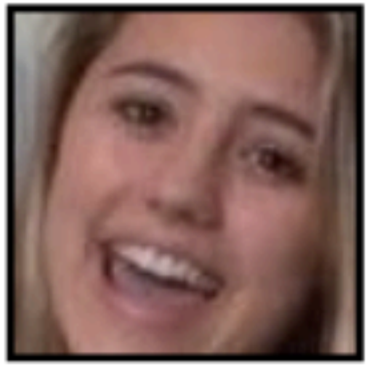
Face Clustering

- This stage aims to extract **person-specific expression-invariant appearance features** that can be used to distinguish faces from different people .



Identity-Annotated Datasets

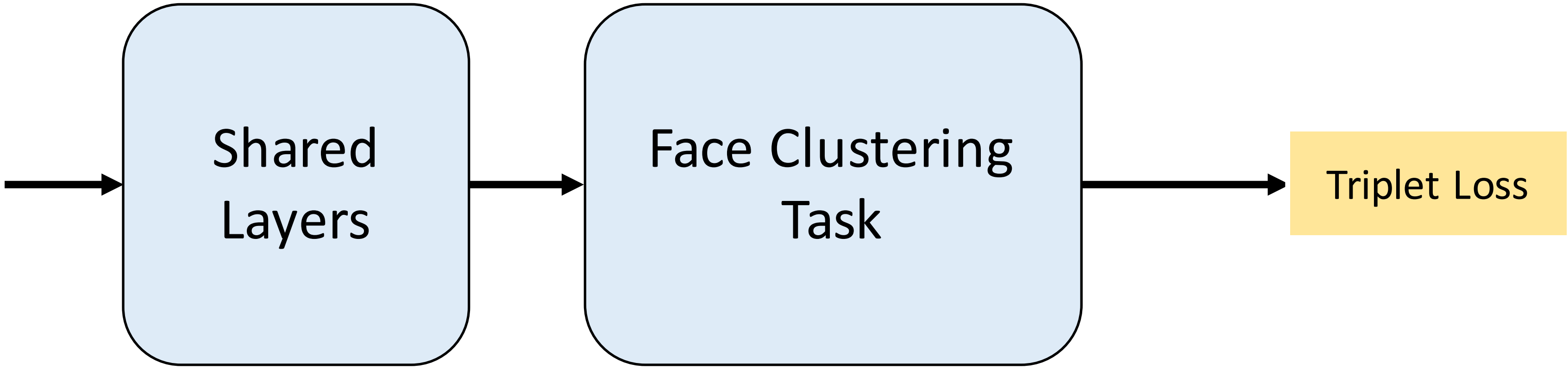
Dataset Name	Number of Subjects	Number of Samples
LFW	5,749	13,233
WDRef	2,995	99,773
CelebA	10,177	202,599
VGG FACE	2,622	2.6M







Face Clustering Network

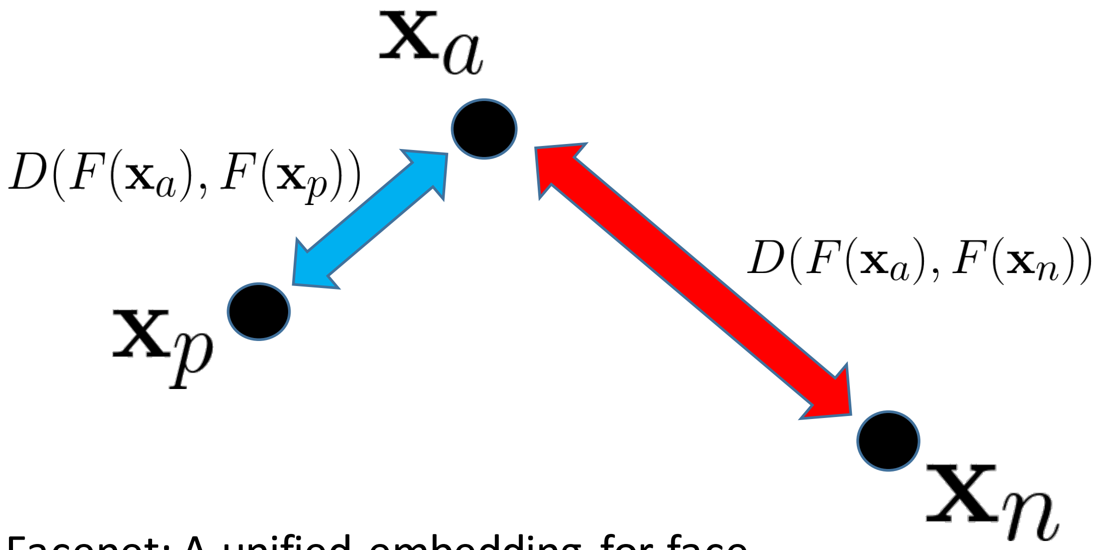


Triplet Loss (Schroff et al. 2015CVPR)

➢ We train this branch $F(\cdot)$ using the triplet loss defined as following

$$triplet\_loss(\mathcal{T}) = max(0, k + D(F(\mathbf{x}_a), F(\mathbf{x}_p)) - D(F(\mathbf{x}_a), F(\mathbf{x}_n)))$$

$$\mathcal{T} = (\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n)$$



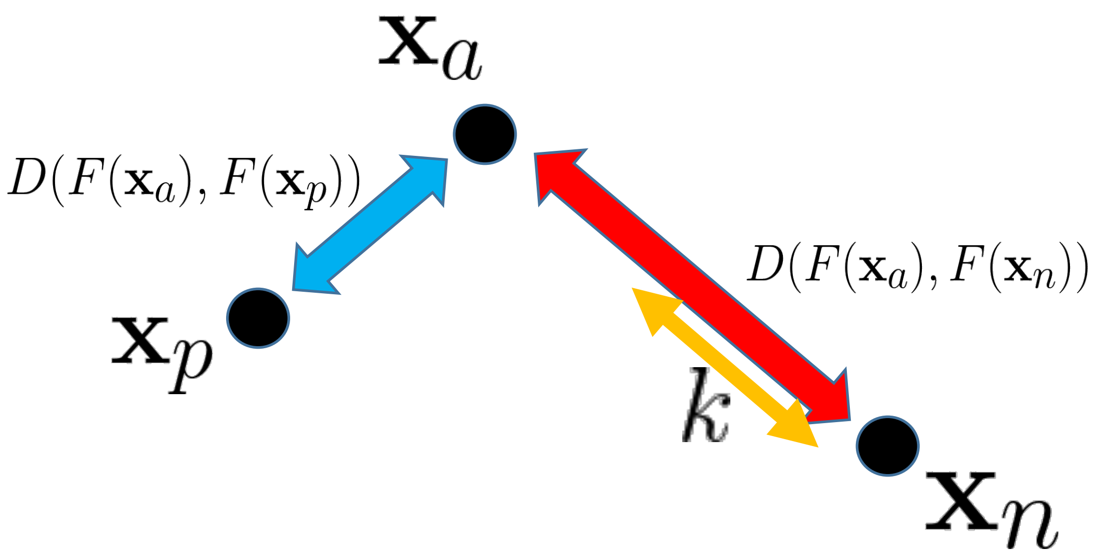
F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proc. CVPR, 2015.

Triplet Loss (Schroff et al. 2015CVPR)

- We train this branch $F(\cdot)$ using the triplet loss defined as following

$$triplet\_loss(\mathcal{T}) = max(0, k + D(F(\mathbf{x}_a), F(\mathbf{x}_p)) - D(F(\mathbf{x}_a), F(\mathbf{x}_n)))$$

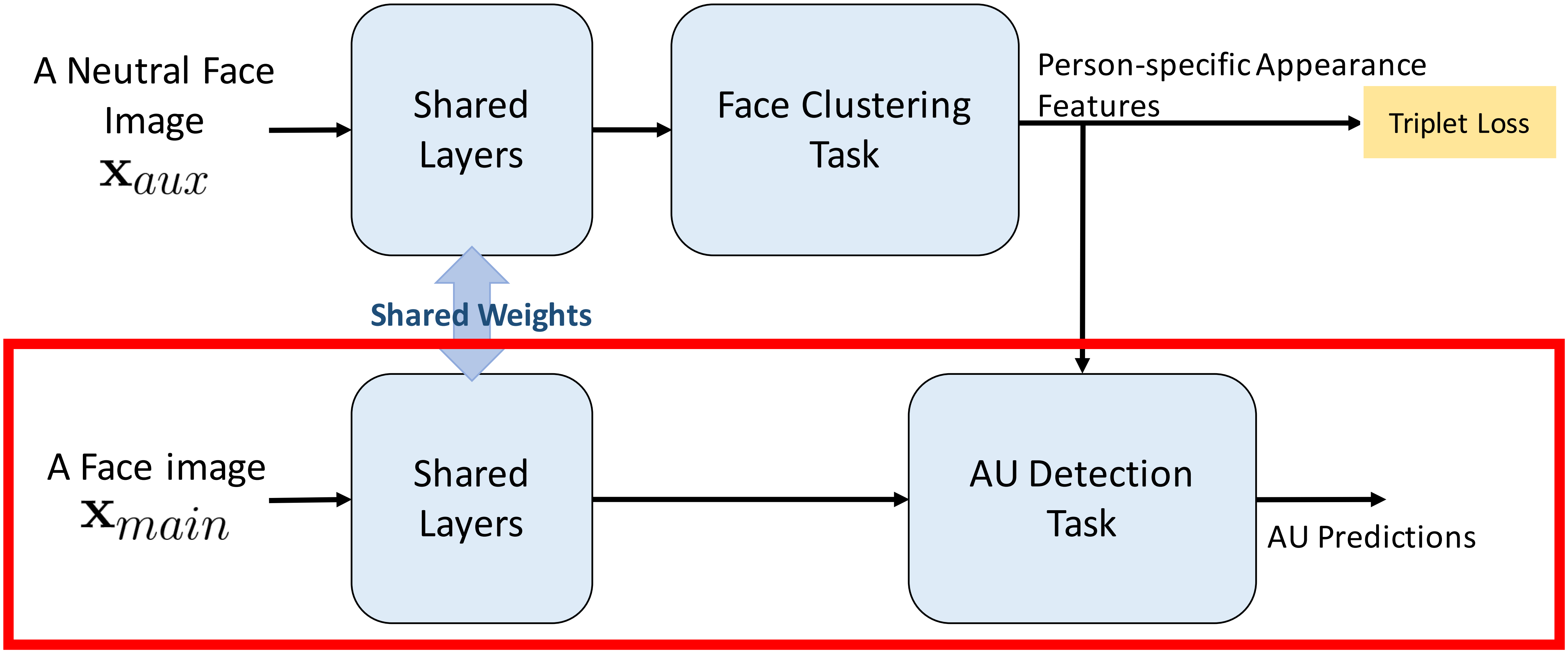
$$\mathcal{T} = (\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n)$$



$$k + D(F(\mathbf{x}_a), F(\mathbf{x}_p)) - D(F(\mathbf{x}_a), F(\mathbf{x}_n)) < 0$$

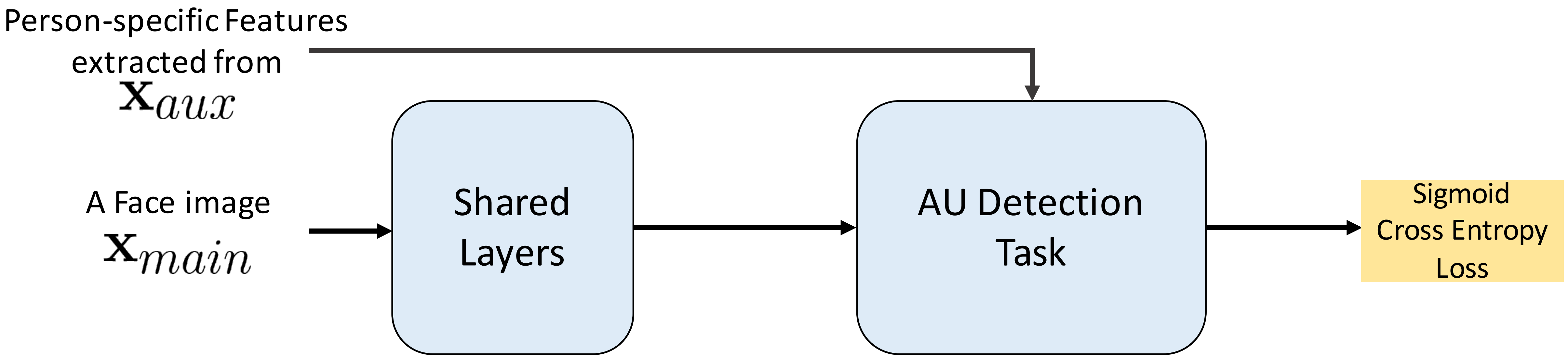
$$k + D(F(\mathbf{x}_a), F(\mathbf{x}_p)) < D(F(\mathbf{x}_a), F(\mathbf{x}_n))$$

F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proc. CVPR, 2015.


A Neutral Face Image
$\mathbf{x}_{aux}$
Shared Layers
Face Clustering Task
Person-specific Appearance Features
Triplet Loss
Shared Weights
A Face image
$\mathbf{x}_{main}$
Shared Layers
AU Detection Task
AU Predictions

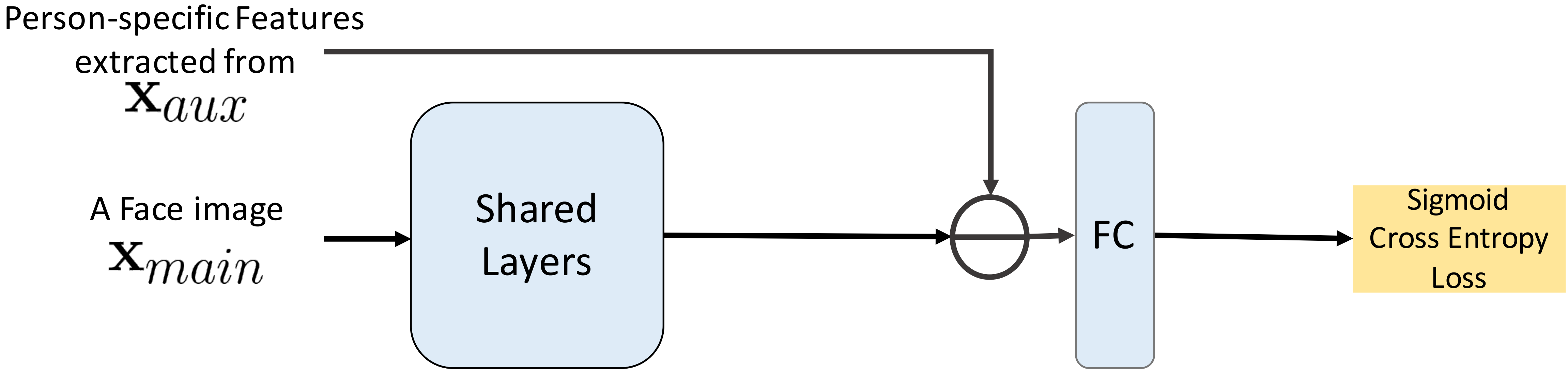
AU Detection Network

- We assume that people with similar appearance features will have similar appearance patterns of AUs
- We combine $\mathbf{x}_{main}$ and person-specific appearance features to predict AUs



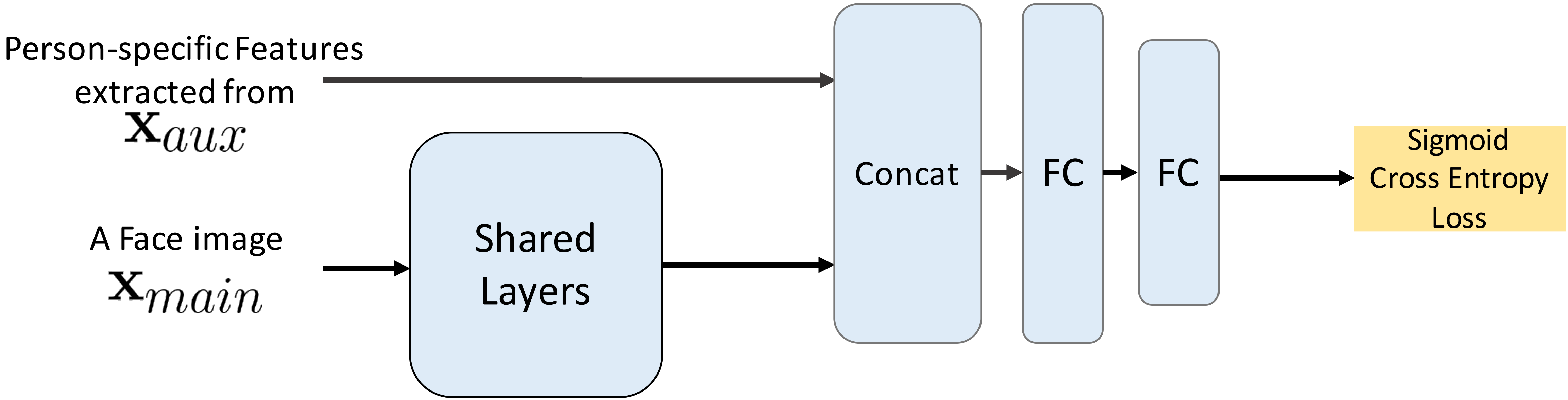
Combining with Person-specific Features

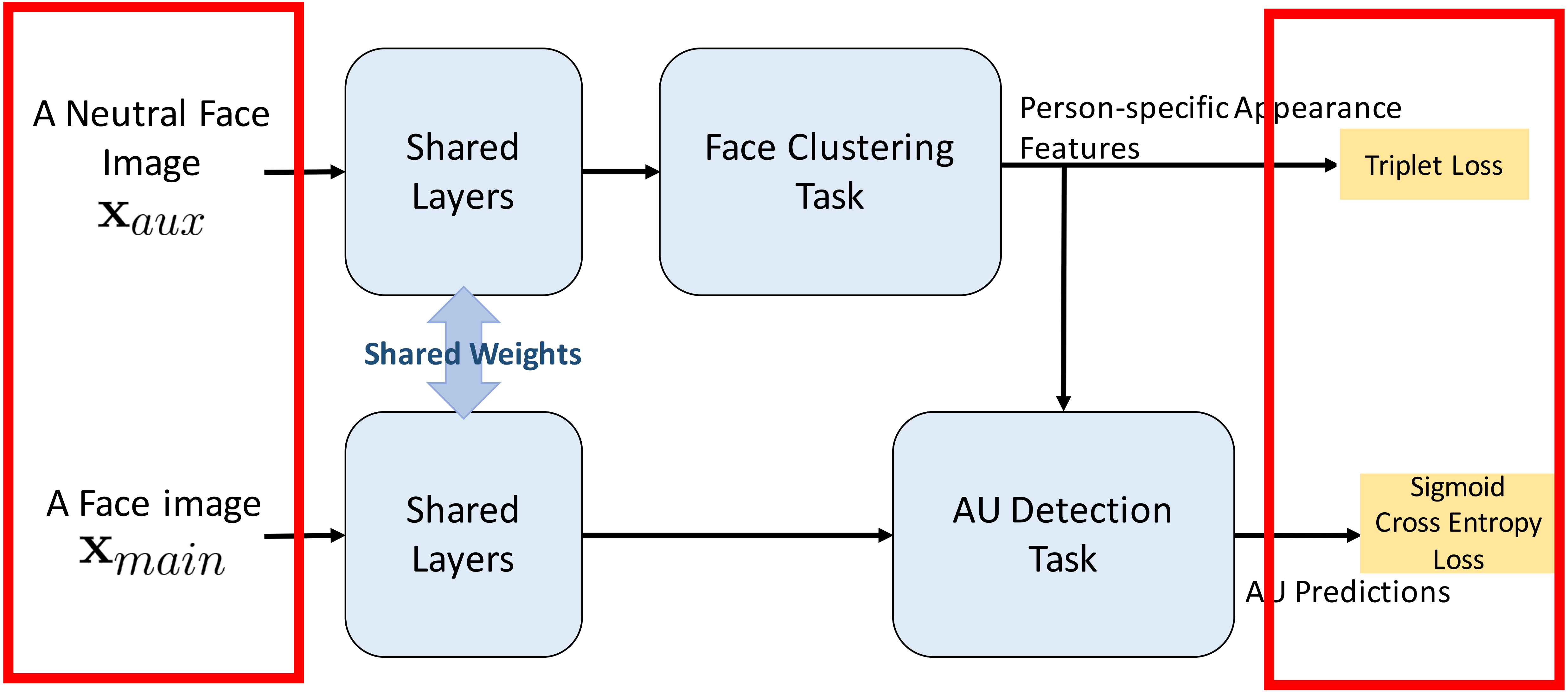
- We adopt two ways to combine with person-specific features
  - Identity Normalization that subtracts person-specific features from $\mathbf{x}_{main}$ features.



Combining with Person-specific Features

- We adopt two ways to combine with person-specific features
  - Identity Normalization that subtracts person-specific features from $\mathbf{x}_{main}$ features.
  - Concatenate the two features to learn their relations from data.




A Neutral Face Image
$\mathbf{x}_{aux}$
Shared Layers
Face Clustering Task
Person-specific Appearance Features
Triplet Loss
Shared Weights
A Face image
$\mathbf{x}_{main}$
Shared Layers
AU Detection Task
Sigmoid Cross Entropy Loss
AU Predictions

Optimizing the whole network cascades

- We combine two types of datasets
  - The Identity-annotated Dataset $\mathcal{D}_{id}$
  - The AU-annotated Dataset (with identity labels) $\mathcal{D}_{id+au}$

$\mathcal{D}_{id}$

Dataset Name	Number of Subjects	Number of Samples
LFW	5,749	13,233
WDRef	2,995	99,773
CelebA	10,177	202,599
VGG FACE	2,622	2.6M

$\mathcal{D}_{id+au}$

Dataset Name	Labels	Number of Subjects	Number of Samples
AMFED	AUs, Interest	<= 242	242 videos
DISFA	AUs	27	27 videos
BP4D	AUs	41	148,562 frames
McMaster-UNBC	AUs, Pain	25	48,398 frames

Optimizing the whole network cascades

- The first half of our training batch

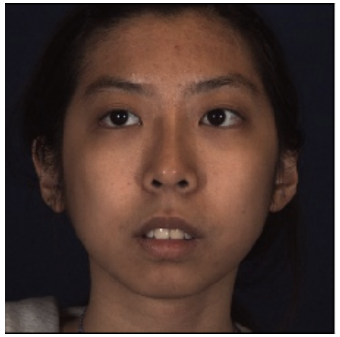
$\mathbf{x}_{main}$

Face images from $\mathcal{D}_{id+au}$

Sigmoid Cross Entropy Loss

$\mathbf{x}_{aux}$

The subject's neutral faces



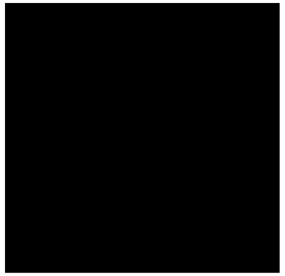
Triplet Loss

Optimizing the whole network cascades

- The second half of our training batch

$\mathbf{x}_{main}$

Zero images

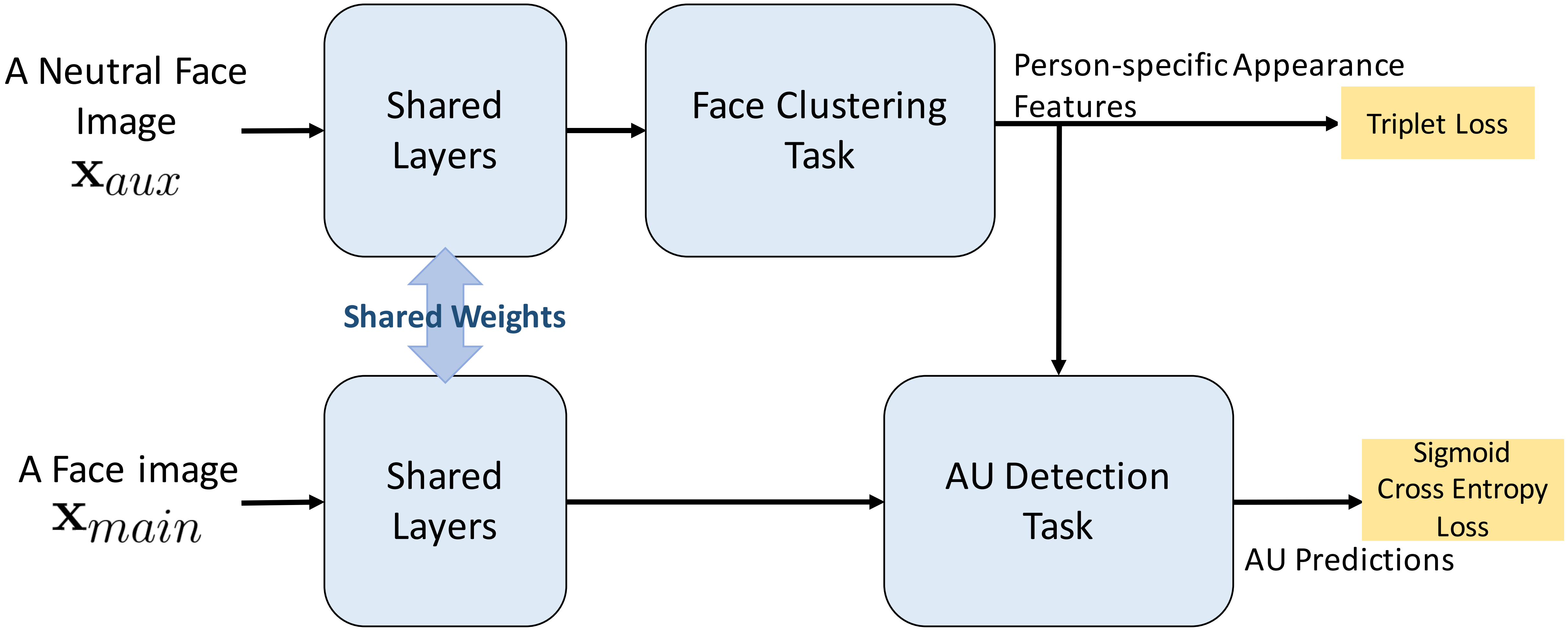


~~Sigmoid Cross Entropy Loss~~

$\mathbf{x}_{aux}$

Face images from $\mathcal{D}_{id}$

Triplet Loss


A Neutral Face Image
$\mathbf{x}_{aux}$
Shared Layers
Face Clustering Task
Person-specific Appearance Features
Triplet Loss
Shared Weights
A Face image
$\mathbf{x}_{main}$
Shared Layers
AU Detection Task
Sigmoid Cross Entropy Loss
AU Predictions

Experiments

Network Structure

- We adopt the 4 convolutional layers from **LightCNNA** architecture, which is designed for face clustering tasks, as our shared layers

- The network pre-trained on the CASIA-WebFace dataset is provided on the author's github repository

X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. arXiv preprint, 2015.

Datasets

- For the $\mathcal{D}_{id}$, we adopt the CelebA dataset that contains face images of celebrities collected from the Internet.

Dataset Name	Number of Subjects	Number of Samples
CelebA	10,177	202,599

- For the $\mathcal{D}_{id+au}$, we adopt the BP4D dataset that contains 328 videos from 41 subjects that are 18 to 29 years of ages.

Dataset Name	Number of Subjects	Number of Samples
BP4D	41	148,562

Action Unit Labels

AU Name	AU1	AU2	AU4	AU6	AU7
Description	Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Cheek Raiser	Lid Tightener

AU Name	AU10	AU12	AU14	AU15	AU17	AU23	AU24
Description	Upper Lip Raiser	Lip Corner Puller	Dimpler	Lip Corner Depressor	Chin Raiser	Lip Tightener	Lip Pressor

BP4D 3-fold Random Splits Results

	AlexNet	DRML	SVTPT	Ours+Sub	Ours+Concat	ROINet
au1	0.399	0.413	0.393	**[0.505]**	0.504	0.362
au2	0.269	0.347	0.349	0.359	**[0.385]**	0.316
au4	0.400	0.416	0.375	**[0.506]**	0.501	0.434
au6	0.694	0.766	0.647	**[0.772]**	0.764	0.771
au7	0.646	0.719	0.724	**[0.742]**	0.711	0.737
au10	0.781	0.807	0.750	0.829	0.827	**[0.850]**
au12	0.812	0.823	0.796	0.851	0.865	**[0.870]**
au14	0.529	0.607	0.482	**[0.630]**	0.557	0.626
au15	0.234	0.311	0.392	0.422	0.430	**[0.457]**
au17	0.510	0.568	0.577	0.608	**[0.623]**	0.580
au23	0.270	0.342	0.330	0.421	**[0.451]**	0.383
au24	0.302	0.352	0.404	0.465	**[0.486]**	0.374
avg.	0.487	0.539	0.518	**[0.593]**	0.592	0.564

BP4D to DISFA Scenarios

- The DISFA dataset contains 27 subjects that are 18 to 29 years of ages.
- We use the 3 models trained on the BP4D 3-fold random splits.

Dataset Name	Number of Subjects	Number of Samples
DISFA	27	130,814

BP4D to DISFA Results

	AlexNet	DRML	SVTPT	Ours+Sub	Ours+Concat
au1	0.127	0.112	0.124	0.201	**[0.246]**
au2	0.096	0.040	0.112	0.255	**[0.299]**
au4	0.270	0.329	0.131	0.373	**[0.393]**
au6	0.335	0.326	0.259	0.496	**[0.524]**
au12	0.461	0.488	0.443	0.661	**[0.666]**
avg.	0.258	0.259	0.214	0.397	**[0.426]**

au1	68.17%	72.88%	68.45%	60.20%	51.19%
au2	64.31%	88.47%	67.91%	28.97%	22.34%
au4	32.50%	20.91%	65.07%	26.28%	21.56%
au6	51.73%	57.44%	59.97%	35.75%	31.41%
au12	43.23%	40.70%	44.35%	22.33%	23.01%
avg.	51.99%	56.08%	61.15%	34.71%	29.90%

BP4D to UNBC-McMaster Scenarios

- The UNBC-McMaster dataset contains 200 videos from 25 subjects that are self-identified as having a problem with shoulder pain.
- We use the 3 models trained on the BP4D 3-fold random splits

Dataset Name	Number of Subjects	Number of Samples
UNBC-McMaster	25	48,398

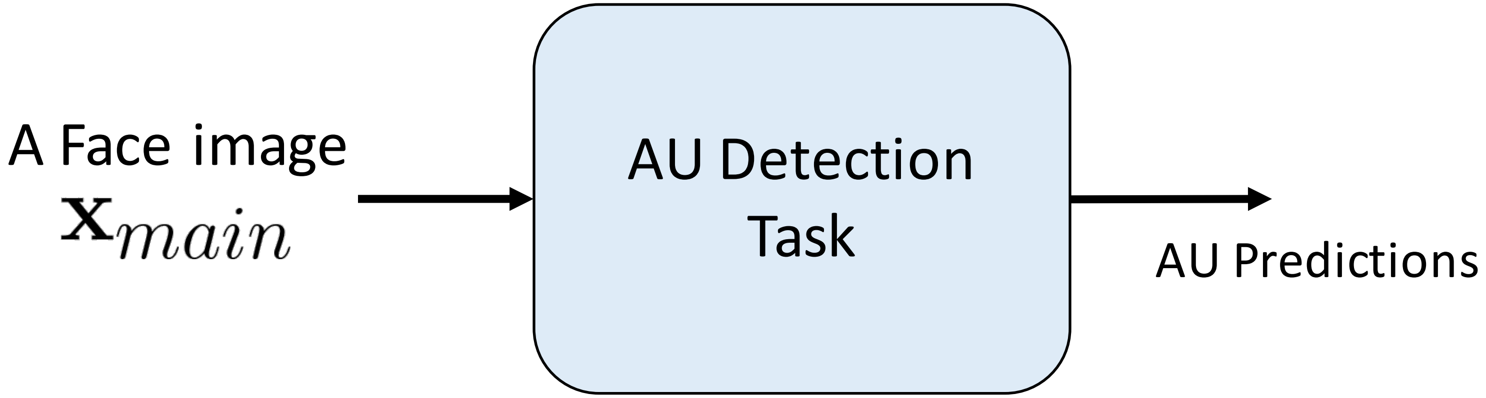
BP4D to UNBC-McMaster Results

	AlexNet	DRML	SVTPT	Ours+Sub	Ours+Concat
au4	0.037	0.046	0.061	**[0.097]**	0.084
au6	0.254	0.245	0.206	**[0.330]**	0.294
au7	**[0.148]**	0.134	0.116	0.128	0.132
au10	**[0.045]**	0.027	0.020	0.028	0.024
au12	0.279	0.310	0.254	**[0.421]**	0.394
avg.	0.153	0.153	0.131	**[0.201]**	0.186

au4	90.75%	88.94%	83.73%	80.83%	83.23%
au6	63.40%	68.02%	68.16%	57.25%	61.52%
au7	77.09%	81.36%	83.98%	82.75%	81.43%
au10	94.24%	96.65%	97.33%	96.62%	97.10%
au12	65.64%	62.33%	68.09%	50.53%	54.45%
avg.	78.22%	79.46%	80.26%	73.60%	75.55%

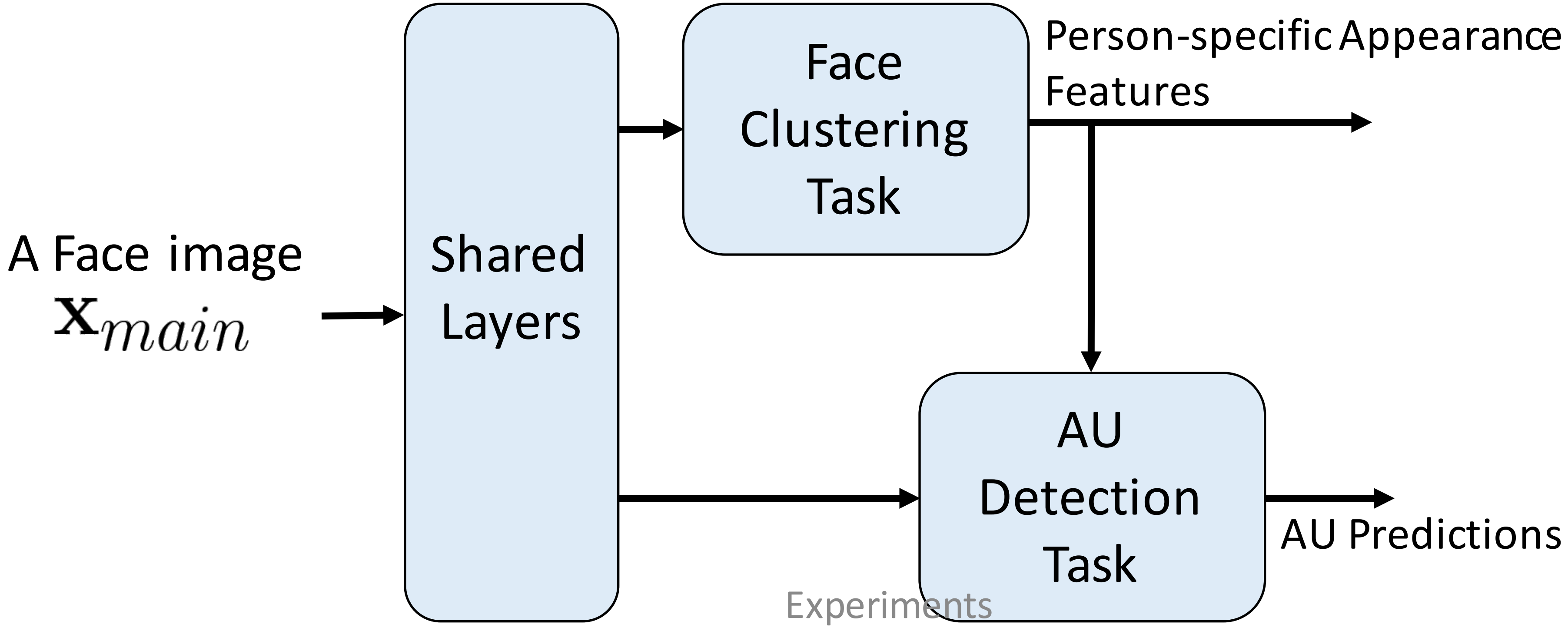
Ablation Study

- We consider 3 different stages of the proposed method.
  - Fine-tuned LightCNNA Network on AU detection (FLightCNNA)



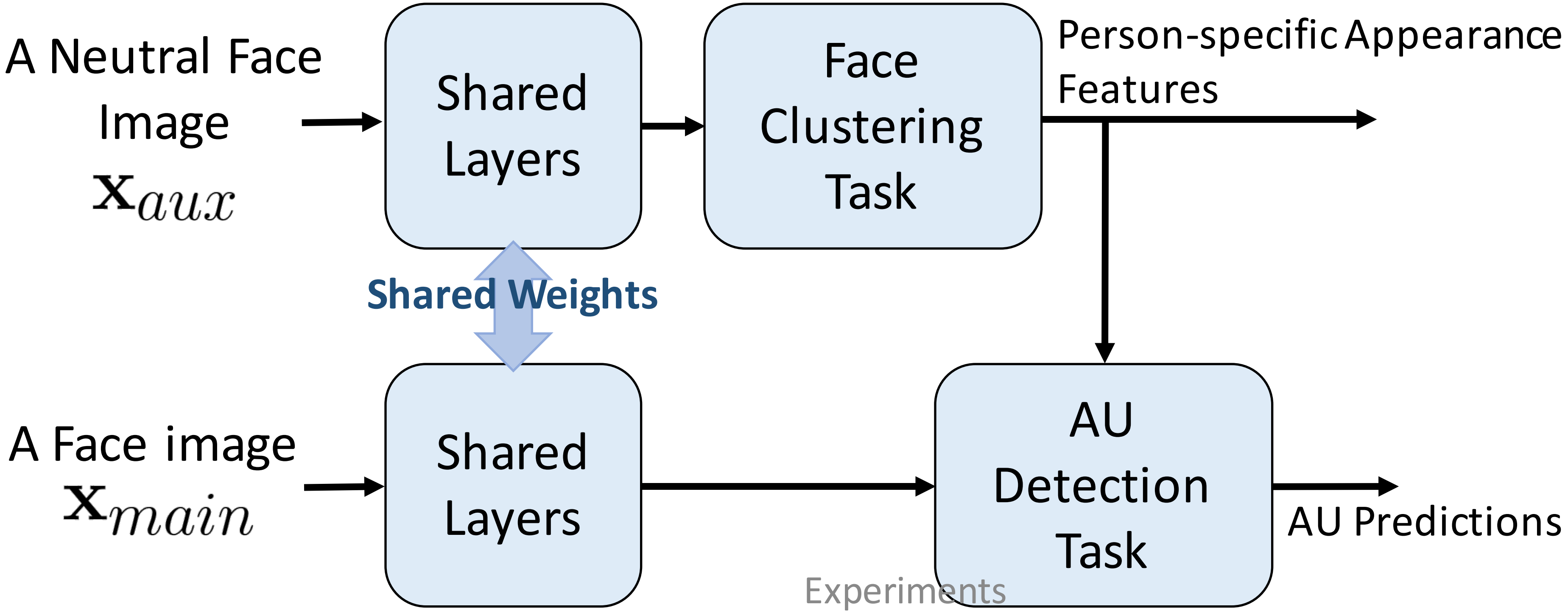
Ablation Study

- We consider 3 different stages of the proposed method.
  - Fine-tuned LightCNNA Network on AU detection (FLightCNNA)
  - Adding the face clustering branch (Ours(single))



Ablation Study

- We consider 3 different stages of the proposed method.
  - Fine-tuned LightCNNA Network on AU detection (FLightCNNA)
  - Adding the face clustering branch (Ours(single))
  - Combining with neutral faces (Ours)



Ablation Study Results on BP4D

	FLightCNNA	Ours(single) +Sub	Ours(single) +Concat	Ours+Sub	Ours+Concat
au1	0.436	**[0.533]**	0.521	0.505	0.504
au2	0.33	**[0.411]**	0.376	0.359	0.385
au4	0.500	**[0.518]**	0.476	0.506	0.501
au6	0.761	**[0.791]**	0.766	0.772	0.764
au7	0.726	0.729	0.717	**[0.742]**	0.711
au10	0.800	0.816	**[0.831]**	0.829	0.827
au12	0.833	0.848	0.861	0.851	0.865
au14	0.594	**[0.644]**	0.620	0.630	0.557
au15	0.307	**[0.452]**	0.428	0.422	0.430
au17	0.544	**[0.625]**	0.614	0.608	0.623
au23	0.339	0.440	0.449	0.421	**[0.451]**
au24	0.393	0.465	0.472	0.465	**[0.486]**
avg.	0.547	**[0.606]**	0.594	0.593	0.592

Ablation Study Results on DISFA

	FLightCNNA	Ours(single) +Sub	Ours(single) +Concat	Ours+Sub	Ours+Concat
au1	0.118	0.206	0.237	0.201	**[0.246]**
au2	0.079	**[0.359]**	0.278	0.255	0.299
au4	0.369	0.370	0.387	0.373	**[0.393]**
au6	0.413	**[0.548]**	0.534	0.496	0.524
au12	0.521	0.652	0.659	0.661	**[0.666]**
avg.	0.300	**[0.427]**	0.419	0.397	0.426

au1	72.94%	61.35%	54.51%	60.20%	51.19%
au2	76.06%	12.65%	26.06%	28.97%	22.34%
au4	26.20%	28.57%	18.70%	26.28%	21.56%
au6	45.73%	30.72%	30.29%	35.75%	31.41%
au12	37.45%	23.11%	23.46%	22.33%	23.01%
avg.	51.68%	31.28%	30.60%	34.71%	29.90%

Ablation Study Results on UNBC-McMaster

	FLightCNNA	Ours(single) +Sub	Ours(single) +Concat	Ours+Sub	Ours+Concat
au4	0.063	**[0.121]**	0.090	0.097	0.084
au6	0.281	**[0.334]**	0.301	0.330	0.294
au7	0.124	**[0.140]**	0.126	0.128	0.132
au10	**[0.028]**	0.025	0.025	**[0.028]**	0.024
au12	0.290	0.398	0.418	**[0.421]**	0.394
avg.	0.157	**[0.204]**	0.192	0.201	0.186

au4	87.40%	76.64%	81.09%	80.83%	83.23%
au6	63.07%	57.77%	60.70%	57.25%	61.52%
au7	82.92%	80.80%	82.43%	82.75%	81.43%
au10	96.50%	96.94%	96.99%	96.62%	97.10%
au12	65.19%	53.07%	51.45%	50.53%	54.45%
avg.	79.02%	73.04%	74.53%	73.60%	75.55%

Conclusion

Summary

- We propose to extract person-specific appearance features for AU detection using face clustering tasks
- Our experimental results show that our methods outperform state-of-the-art ones in terms of average performance

Future Work

- Our network cascades can be improved by combining with face landmark localization to investigate discriminative facial regions for AUs

- Primary Emotion Classification can be combined into our network cascades to predict emotions for different applications

Thank You!