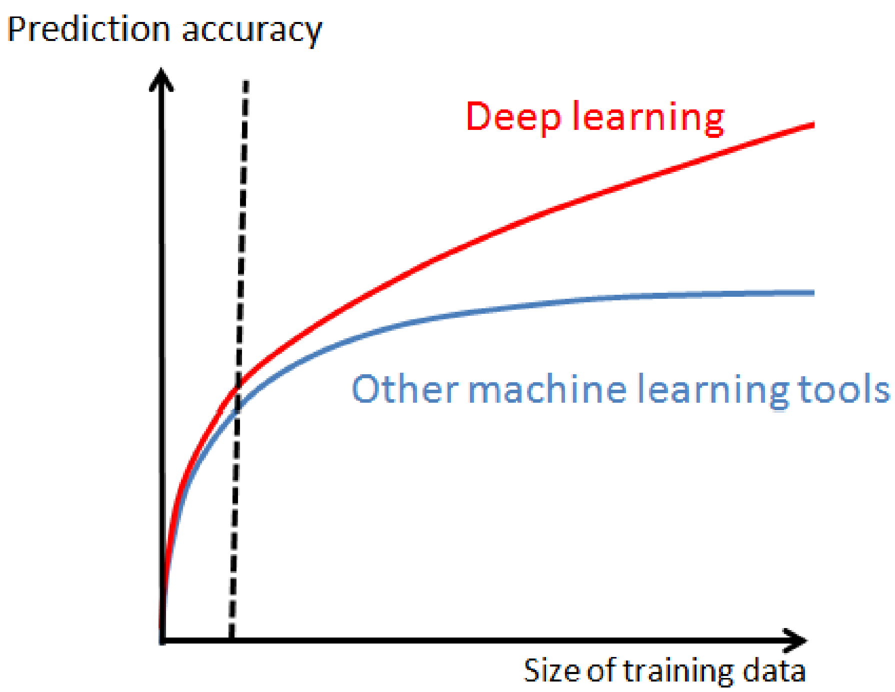
香港中文大學
The Chinese University of Hong Kong

DeepID: Deep Learning for Face Recognition

Xiaogang Wang
Department of Electronic Engineering,
The Chinese University of Hong Kong

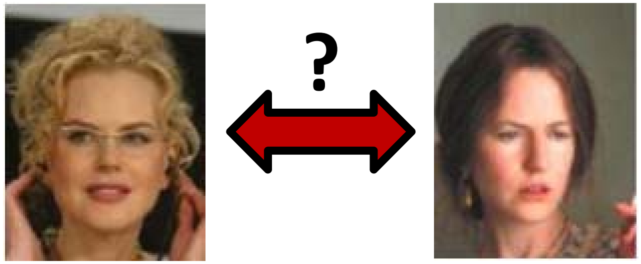
Machine Learning with Big Data

- Machine learning with small data: overfitting, reducing model complexity (capacity), adding regularization
- Machine learning with big data: underfitting, increasing model complexity, optimization, computation resource



Face Recognition

- Face verification: binary classification
  - Verify two images belonging to the same person or not



- Face identification: multi-class classification
  - classify an image into one of N identity classes

Labeled Faces in the Wild (2007)

Random guess (50%)

Best results
without deep learning

MSRA TL Joint Bayesian (96.33%)

Human funneled (99.20%)

CUHK deep learning result (99.53%)

Google deep learning result (99.6%)

Learn face representations from

face verification, identification, multi-view reconstruction

Properties of face representations

sparseness, selectiveness, robustness

Sparsify the network

sparseness, selectiveness

Applications of face representations

face localization, attribute recognition

Learn face representations from

face verification, identification, multi-view reconstruction

Properties of face representations

sparseness, selectiveness, robustness

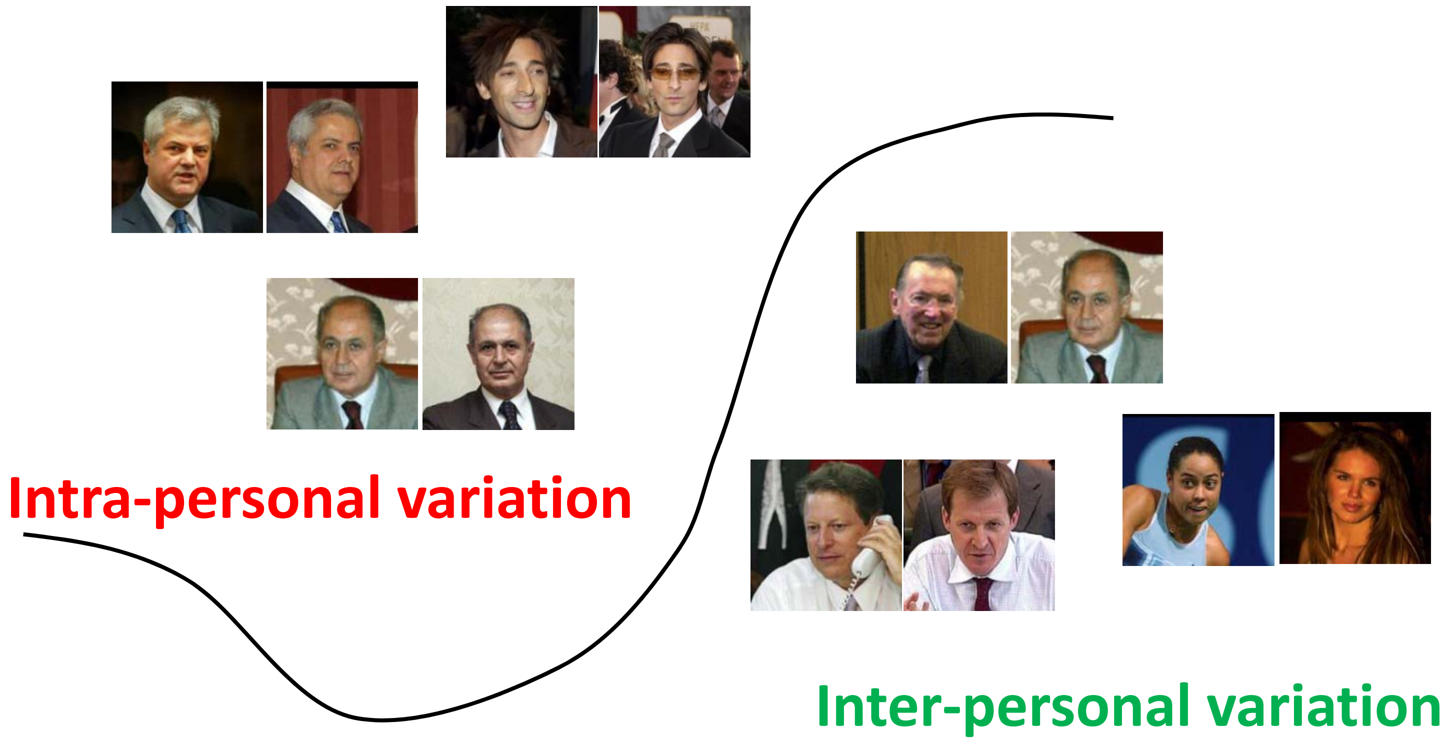
Sparsify the network

sparseness, selectiveness

Applications of face representations

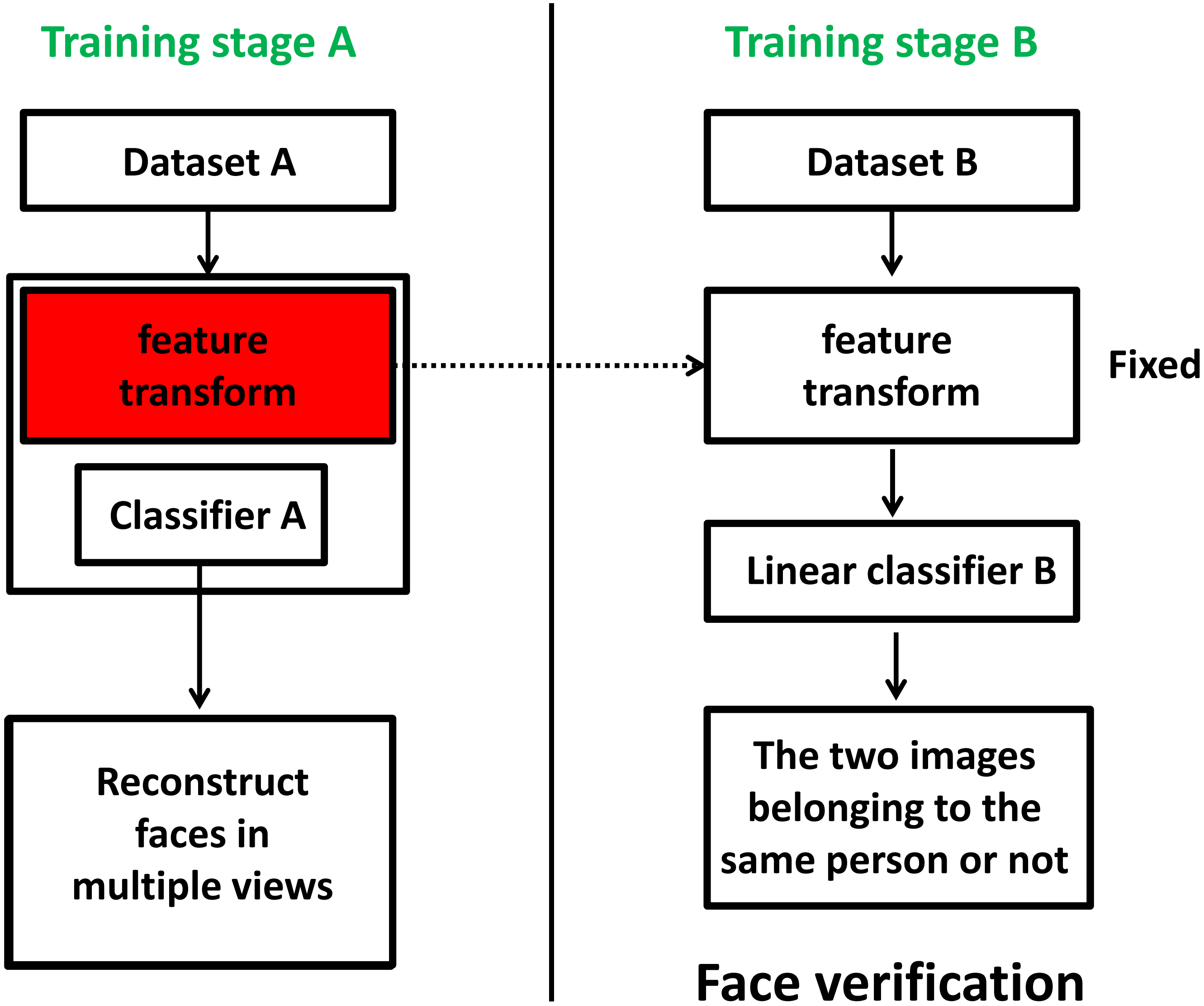
face localization, attribute recognition

Key challenge on face recognition



How to separate the two types of variations?

Learning feature representations



Learn face representations from

Prediction becomes richer

Prediction becomes more challenging

Supervision becomes stronger

Feature learning becomes more effective

Predicting binary labels (verification)

Predicting multi-class labels (identification)

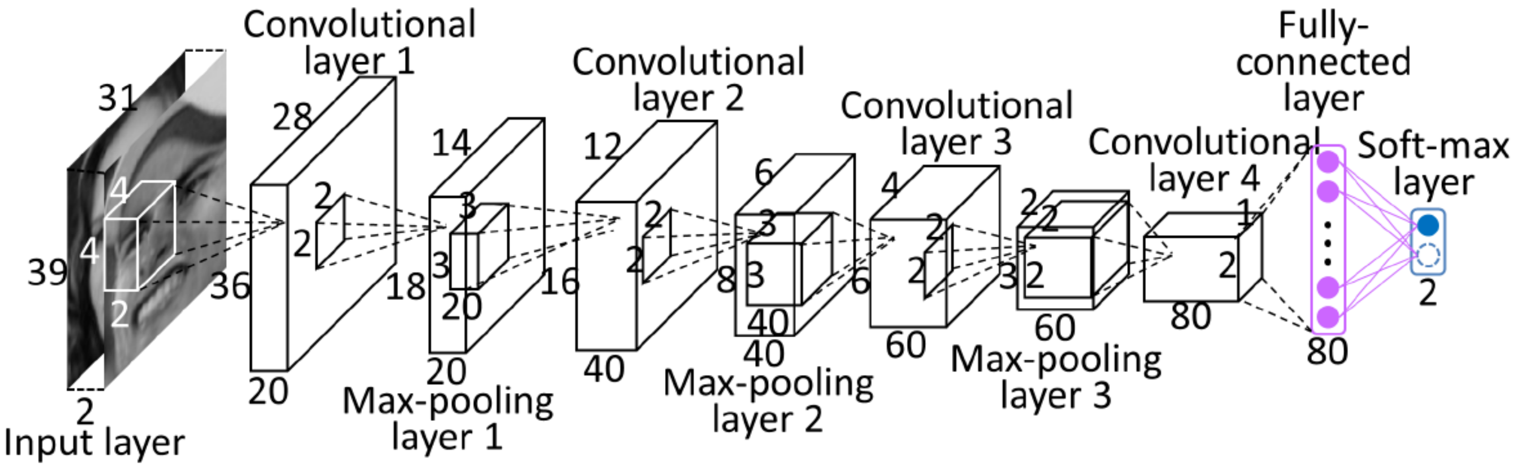
Predicting thousands of real-valued pixels (multi-view) reconstruction

Learn face representations with verification signal

- Extract relational features with learned filter pairs

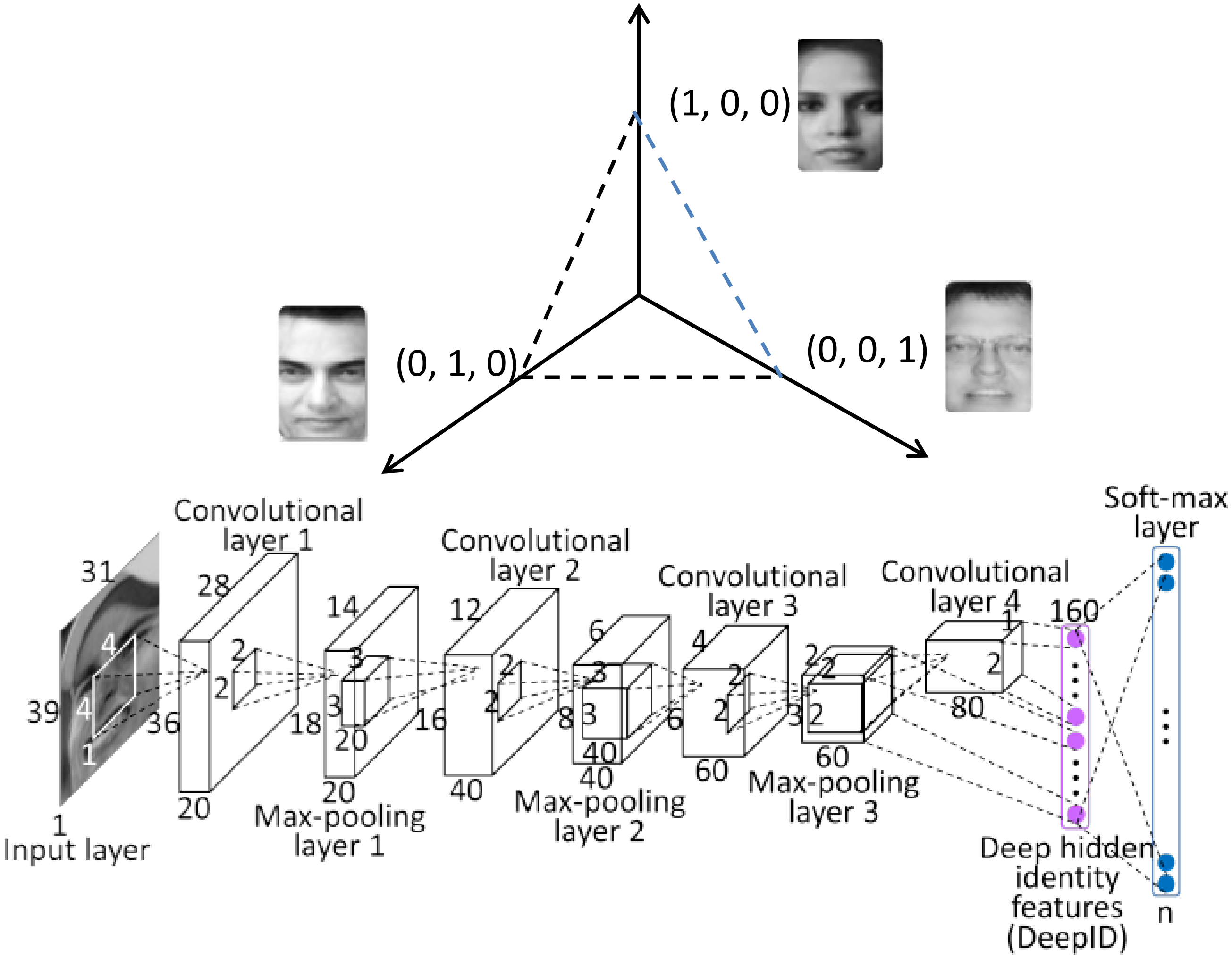
$$y^j = f\left(b^j + k^{1j} * x^1 + k^{2j} * x^2\right)$$

- These relational features are further processed through multiple layers to extract global features
- The fully connected layer is the feature representation



Y. Sun, X. Wang, and X. Tang, "Hybrid Deep Learning for Computing Face Similarities," Proc. ICCV, 2013.

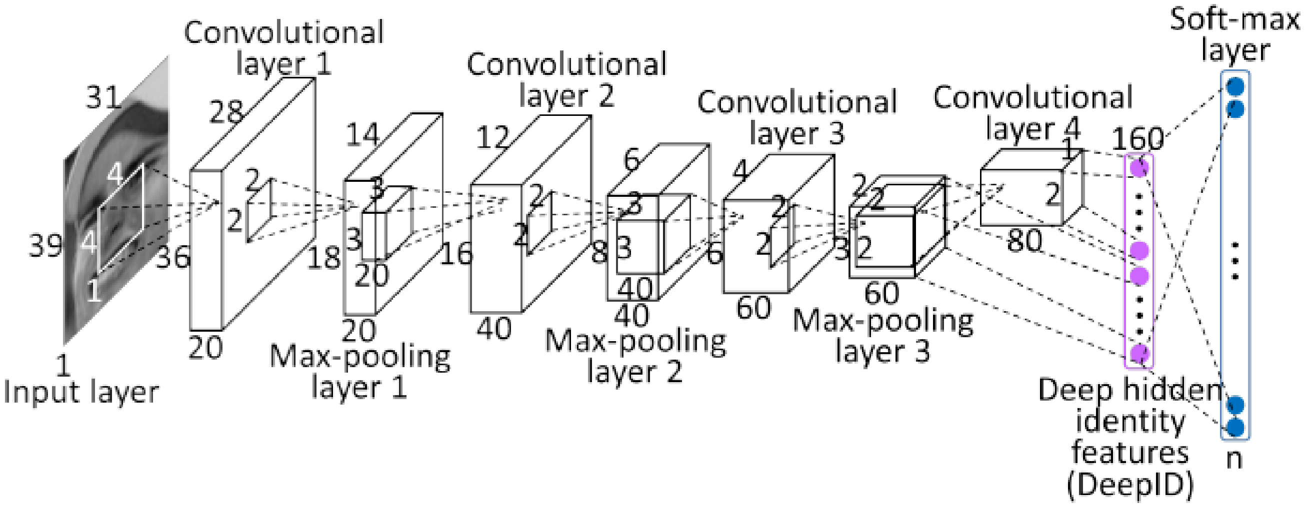
DeepID: Learn face representations with identification signal



Y. Sun, X. Wang, and X. Tang, “Deep Learning Face Representation from Predicting 10,000 classes,” Proc. CVPR, 2014.

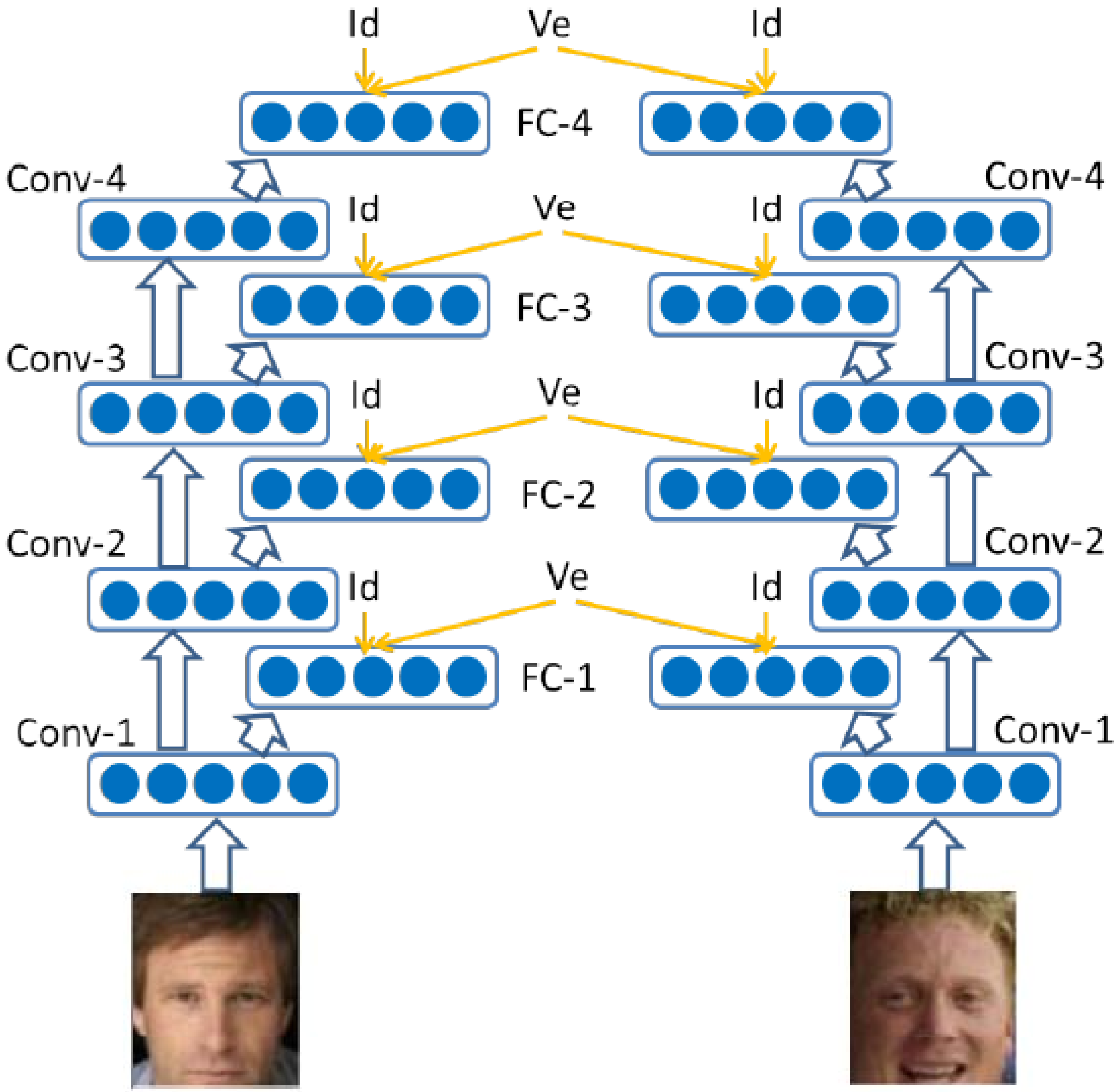
DeepID2: Joint Identification (Id)-Verification (Ve) Signals

$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

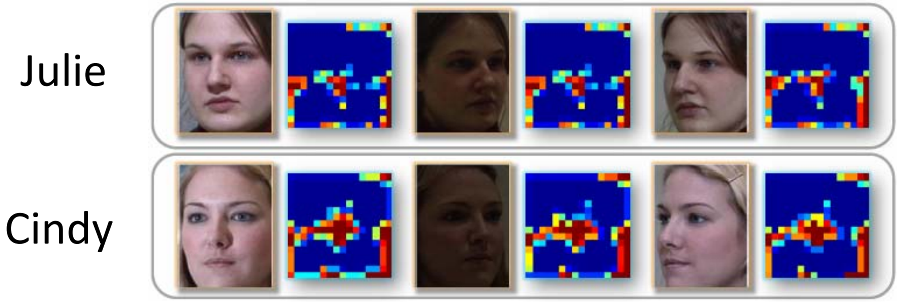


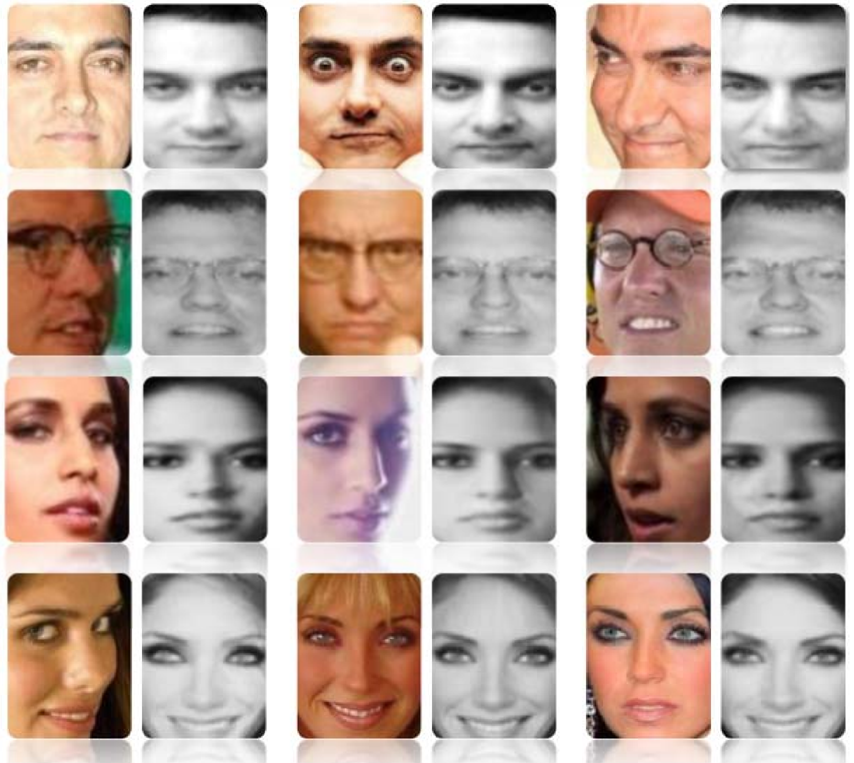
(Id)

Y. Sun, X. Wang, and X. Tang. NIPS, 2014.


Id
Ve
Id
FC-4
Conv-4
Id
Ve
Id
Conv-4
FC-3
Conv-3
Id
Ve
Id
Conv-3
FC-2
Conv-2
Id
Ve
Id
Conv-2
FC-1
Conv-1
Conv-1

Learning face representation from recovering canonical-view face images

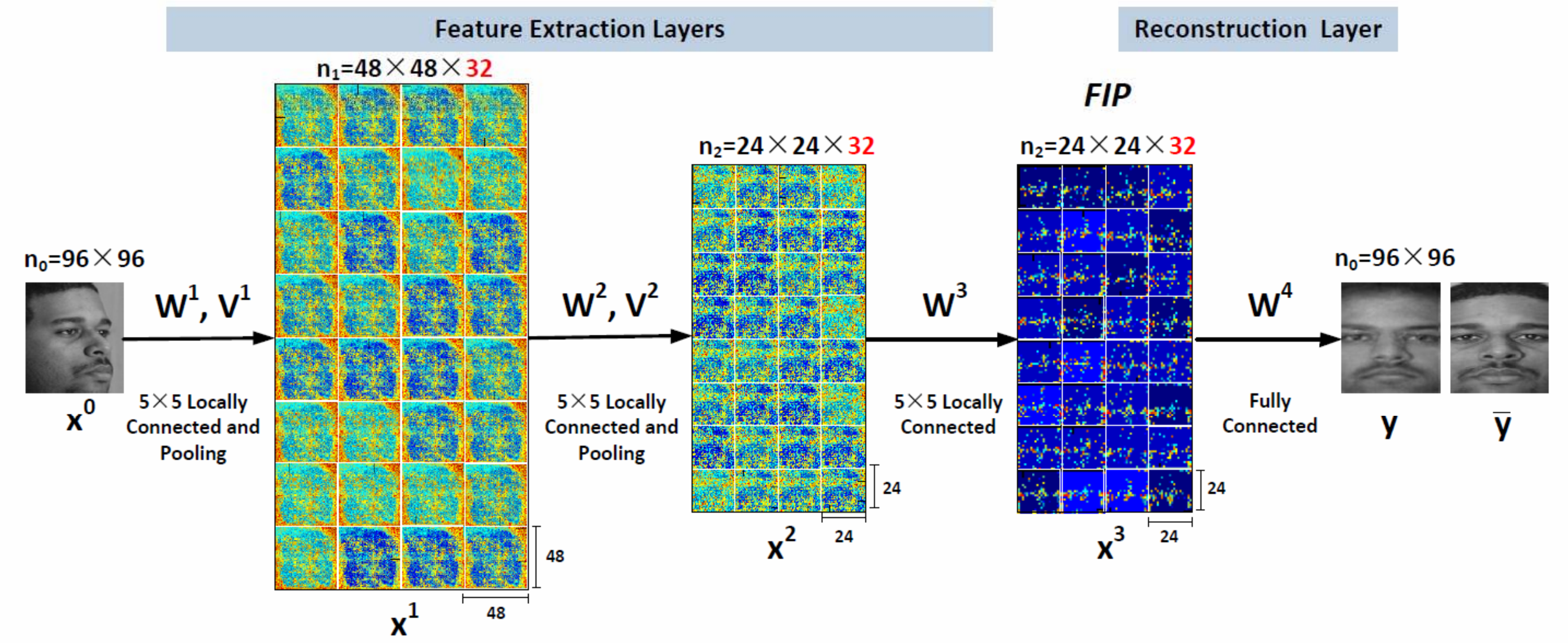




Reconstruction examples from LFW

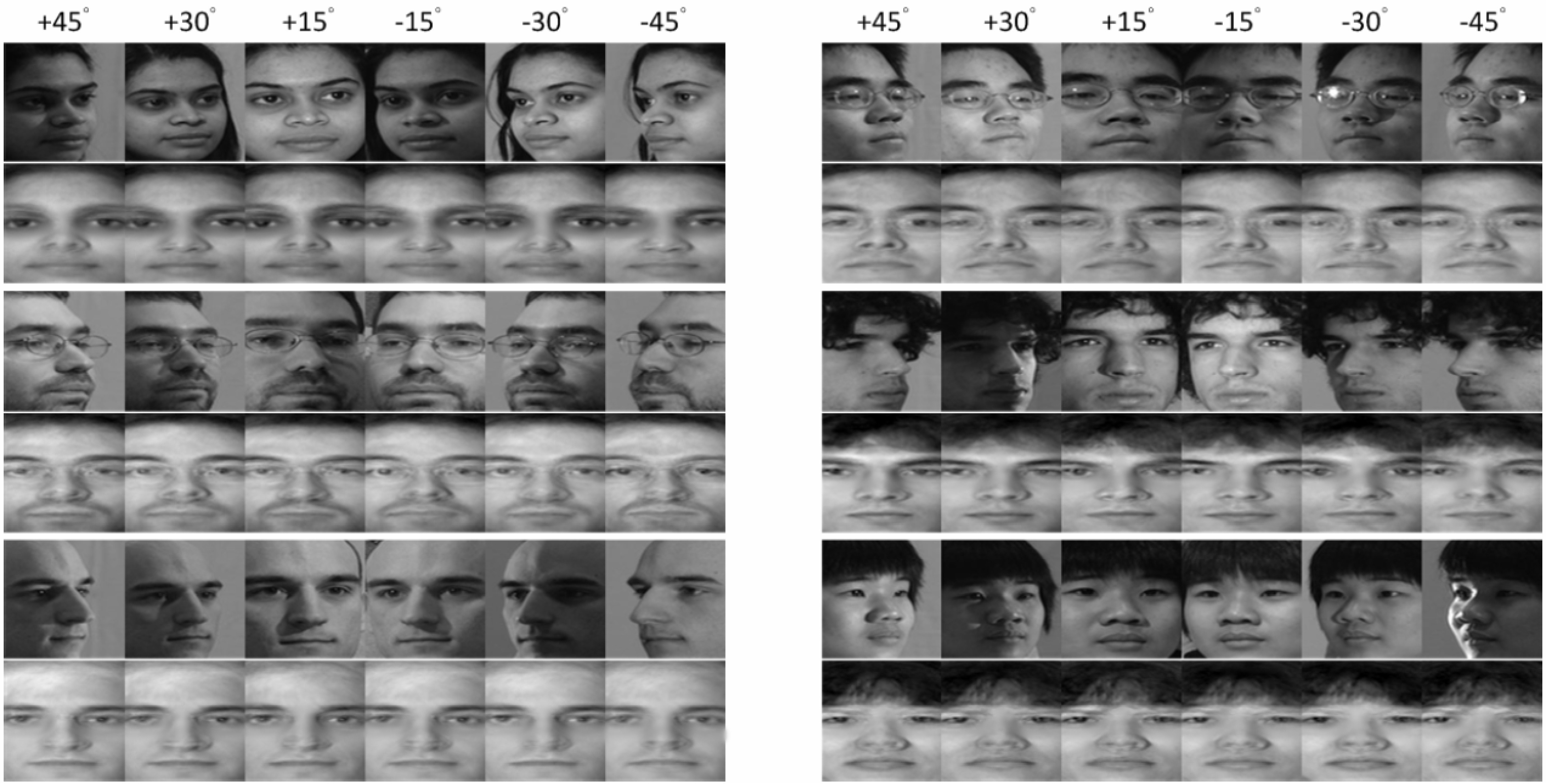
Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep Learning Identity Preserving Face Space," ICCV 2013.

- Disentangle factors through feature extraction over multiple layers
- No 3D model; no prior information on pose and lighting condition
- Model multiple complex transforms
- Reconstructing the whole face is a much strong supervision than predicting 0/1 class label



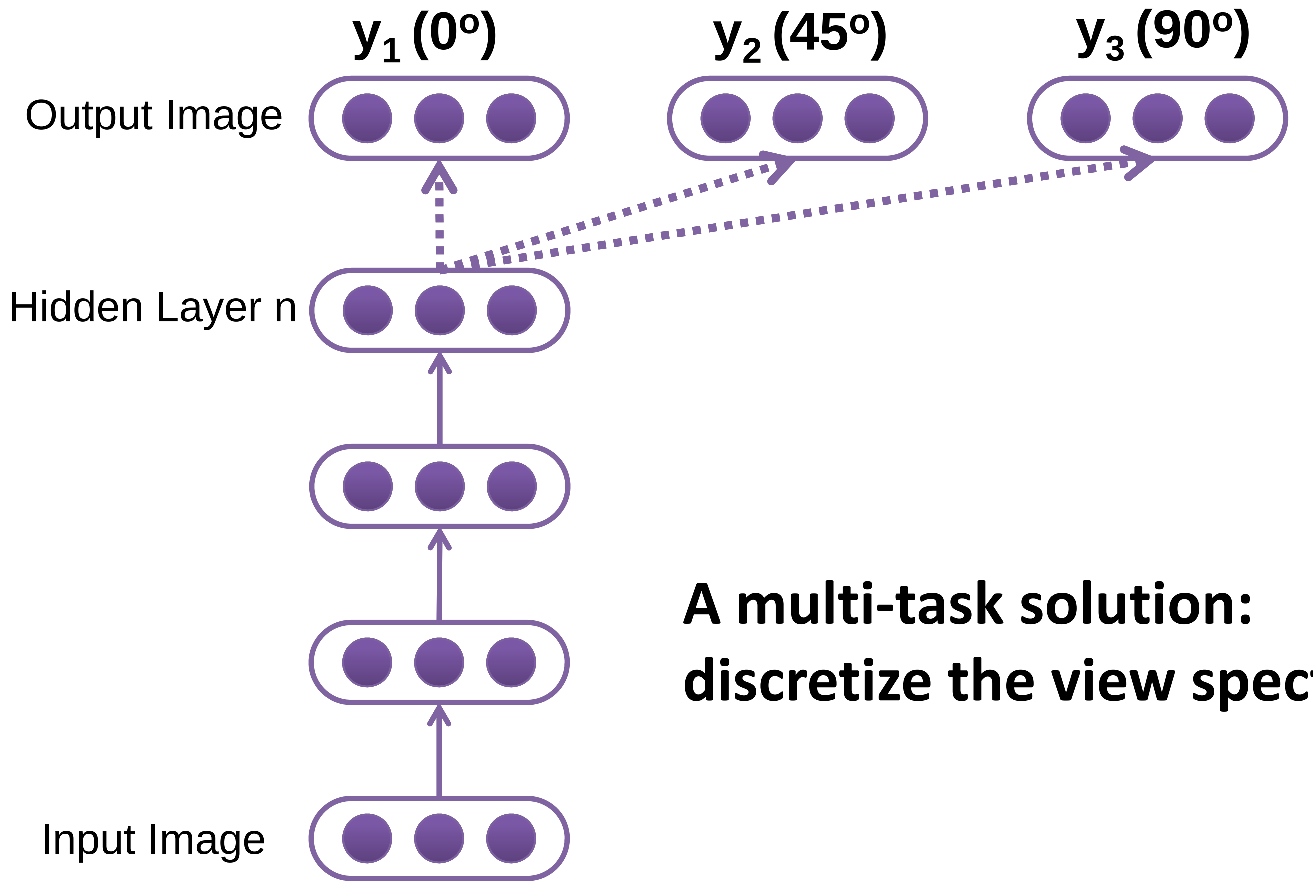
Arbitrary view

Canonical view


+45°
+30°
+15°
-15°
-30°
-45°
+45°
+30°
+15°
-15°
-30°
-45°

It is still not a 3D representation yet

Can we reconstruct all the views?

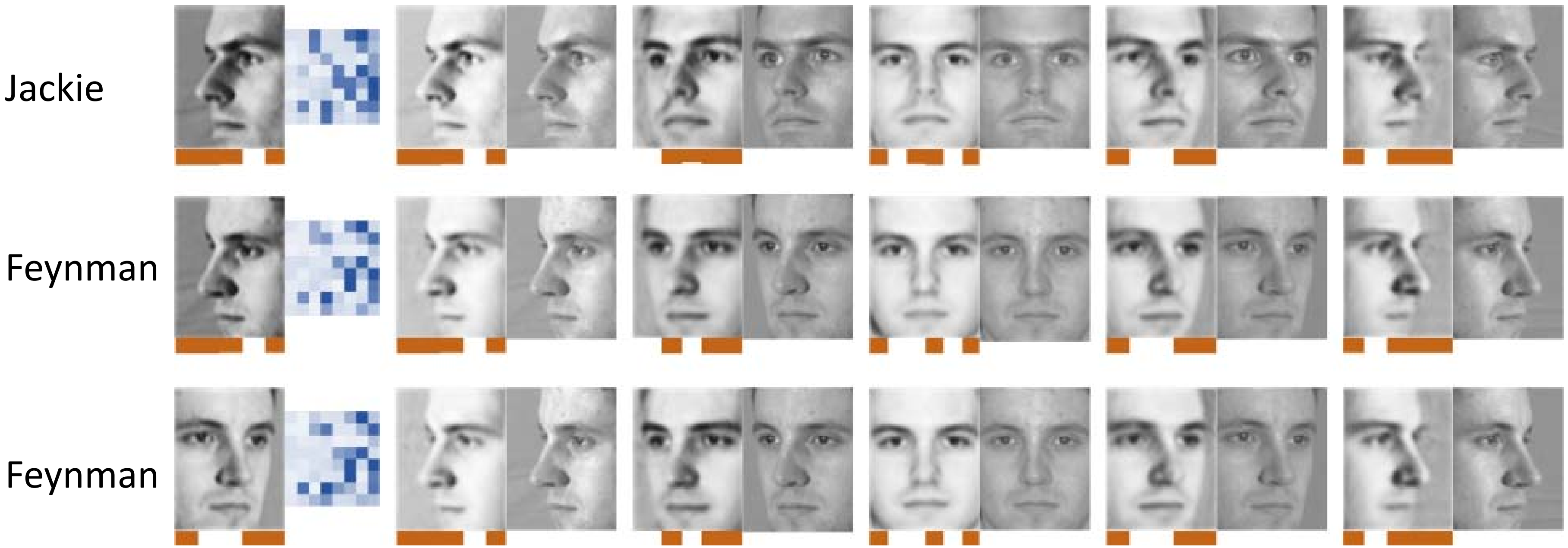


A multi-task solution: discretize the view spectrum

1. The number of views to be reconstructed is predefined, equivalent to the number of tasks
2. Cannot reconstruct views not presented in the training set
3. Encounters problems when the training data of different views are unbalanced
4. Model complexity increases as the number of views

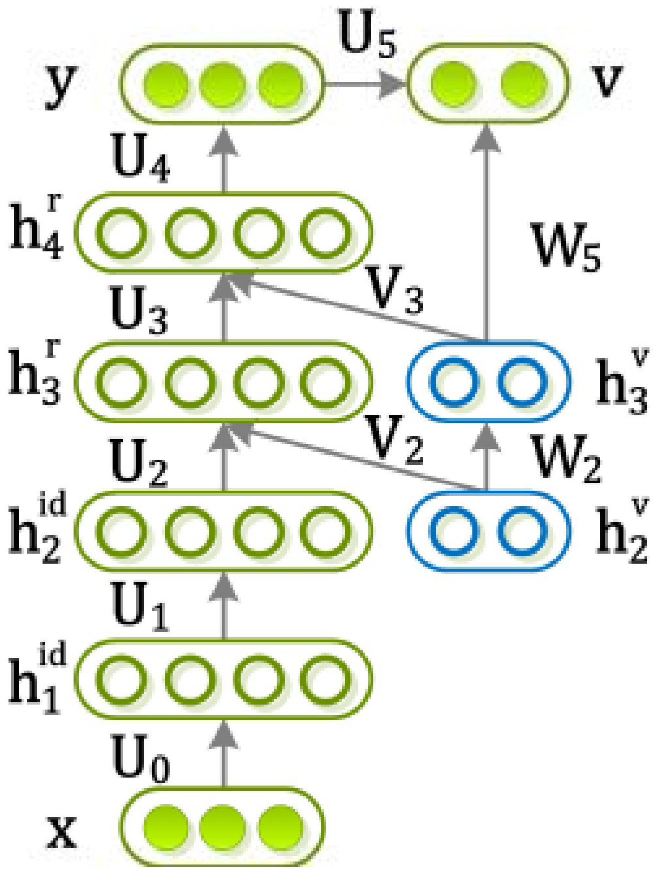
Deep learning multi-view representation from 2D images

- Given an image under arbitrary view, its viewpoint can be estimated and its full spectrum of views can be reconstructed
- Continuous view representation
- Identity and view represented by different sets of neurons



Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning and Disentangling Face Representation by Multi-View Perception,” NIPS 2014.

Network is composed of deterministic neurons and random neurons



x and y are input and output images of the same identity but in different views;

v is the view label of the output image;

h^{id} are neurons encoding identity features

h^v are neurons encoding view features

h^r are neurons encoding features to reconstruct the output images

Deep Learning by EM

- EM updates on the probabilistic model are converted to forward and backward propagation

$$\mathcal{L}(\Theta, \Theta^{old}) = \sum_{\mathbf{h}^v} p(\mathbf{h}^v | \mathbf{y}, \mathbf{v}; \Theta^{old}) \log p(\mathbf{y}, \mathbf{v}, \mathbf{h}^v | \mathbf{h}^{id}; \Theta)$$

- E-step: proposes s samples of $\mathbf{h}$

$$\mathbf{h}_s^v \sim \mathcal{U}(0,1)$$

$$w_s = p(\mathbf{y}, \mathbf{v} \,|\mathbf{h}^v; \Theta^{old})$$

- M-step: compute gradient refer to $\mathbf{h}$ with largest w_s

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \Theta} \simeq \frac{\partial}{\partial \Theta} \left\{ w_s \left(\log p(\mathbf{v} \,|\mathbf{y}, \mathbf{h}_s^v) + \log p(\mathbf{y} | \mathbf{h}^{id}, \mathbf{h}_s^v) \right) \right\}$$

	Avg.	0°	-15°	$+15^\circ$	-30°	$+30^\circ$	-45°	$+45^\circ$	-60°	$+60^\circ$
Raw Pixels+LDA	36.7	81.3	59.2	58.3	35.5	37.3	21.0	19.7	12.8	7.63
LBP [1]+LDA	50.2	89.1	77.4	79.1	56.8	55.9	35.2	29.7	16.2	14.6
Landmark LBP [6]+LDA	63.2	94.9	83.9	82.9	71.4	68.2	52.8	48.3	35.5	32.1
CNN+LDA	58.1	64.6	66.2	62.8	60.7	63.6	56.4	57.9	46.4	44.2
FIP [28]+LDA	72.9	94.3	91.4	90.0	78.9	82.5	66.1	62.0	49.3	42.5
RL [28]+LDA	70.8	94.3	90.5	89.8	77.5	80.0	63.6	59.5	44.6	38.9
MTL+RL+LDA	**74.8**	**93.8**	**91.7**	**89.6**	**80.1**	**83.3**	**70.4**	**63.8**	51.5	50.2
$\text{MVP}_{\mathbf{h}_1^{id}}$+LDA	61.5	92.5	85.4	84.9	64.3	67.0	51.6	45.4	35.1	28.3
$\text{MVP}_{\mathbf{h}_2^{id}}$+LDA	**79.3**	**95.7**	**93.3**	**92.2**	**83.4**	**83.9**	**75.2**	**70.6**	**60.2**	**60.0**
$\text{MVP}_{\mathbf{h}_3^{r}}$+LDA	72.6	91.0	86.7	84.1	74.6	74.2	68.5	**63.8**	**55.7**	**56.0**
$\text{MVP}_{\mathbf{h}_4^{r}}$+LDA	62.3	83.4	77.3	73.1	62.0	63.9	57.3	53.2	44.4	46.9

Face recognition accuracies across views and illuminations on the Multi-PIE dataset. The first and the second best performances are in bold.

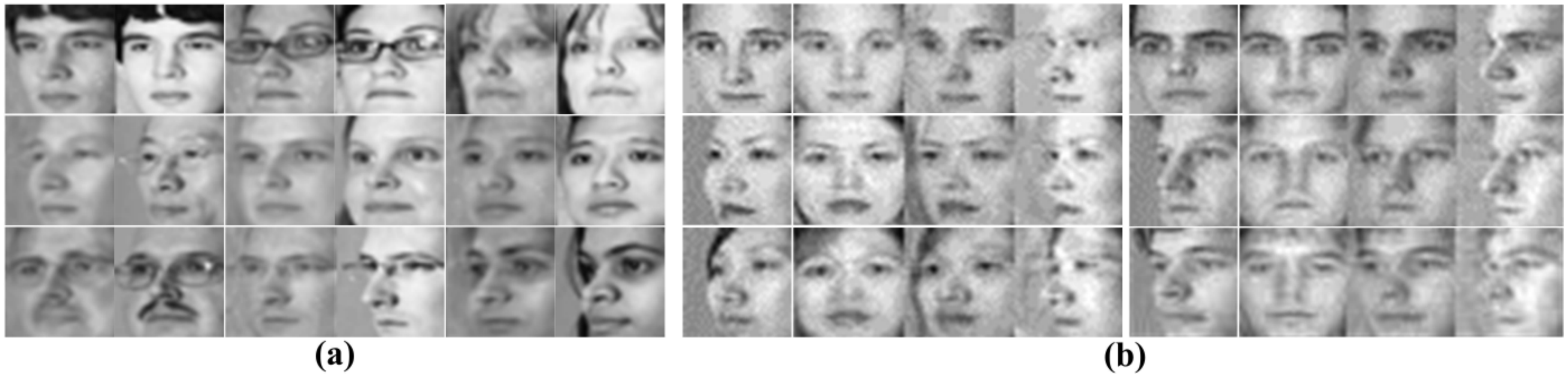
[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28:2037–2041, 2006.

[6] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, 2013.

[28] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity preserving face space. In *ICCV*, 2013.

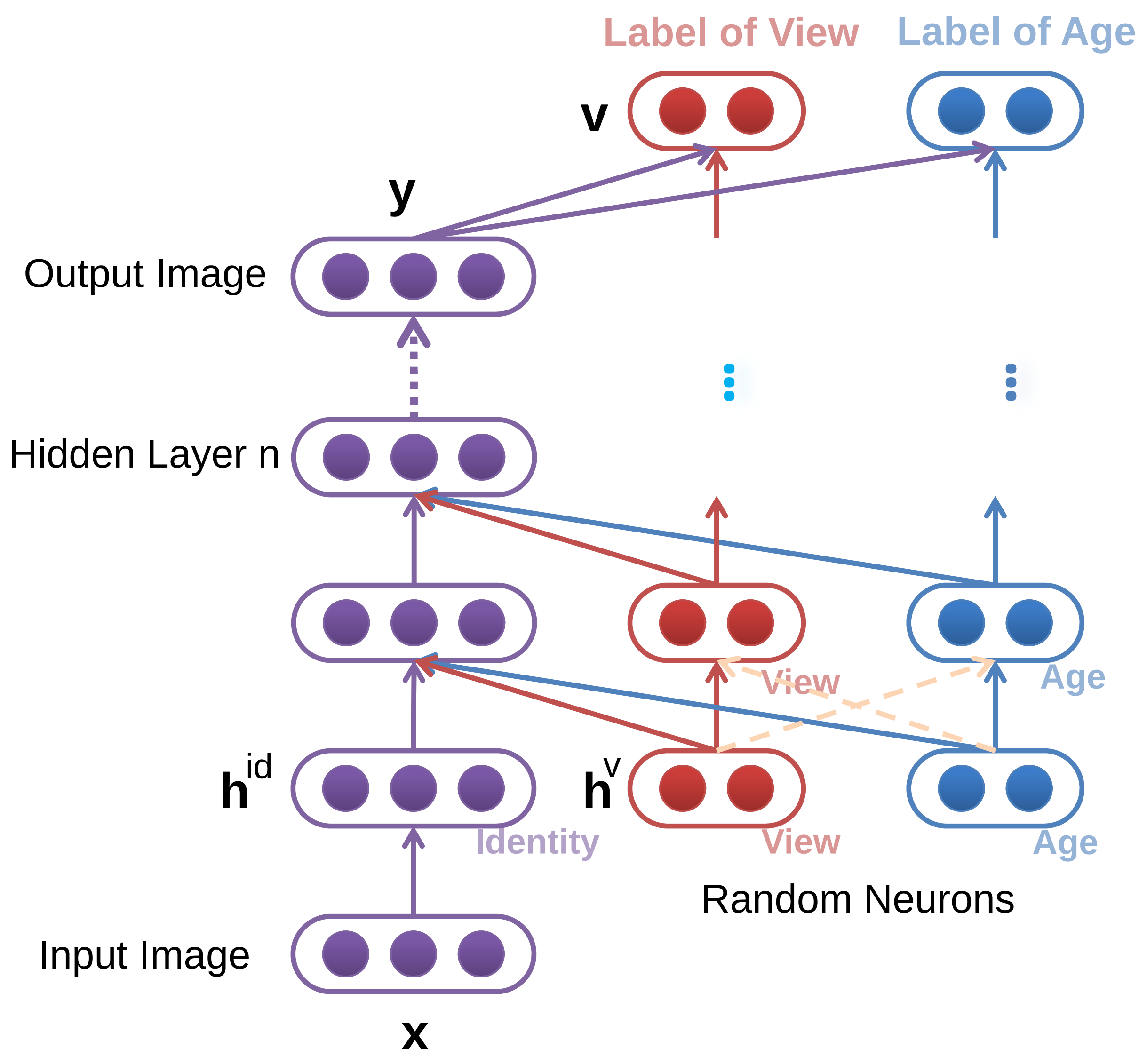
Deep Learning Multi-view Representation from 2D Images

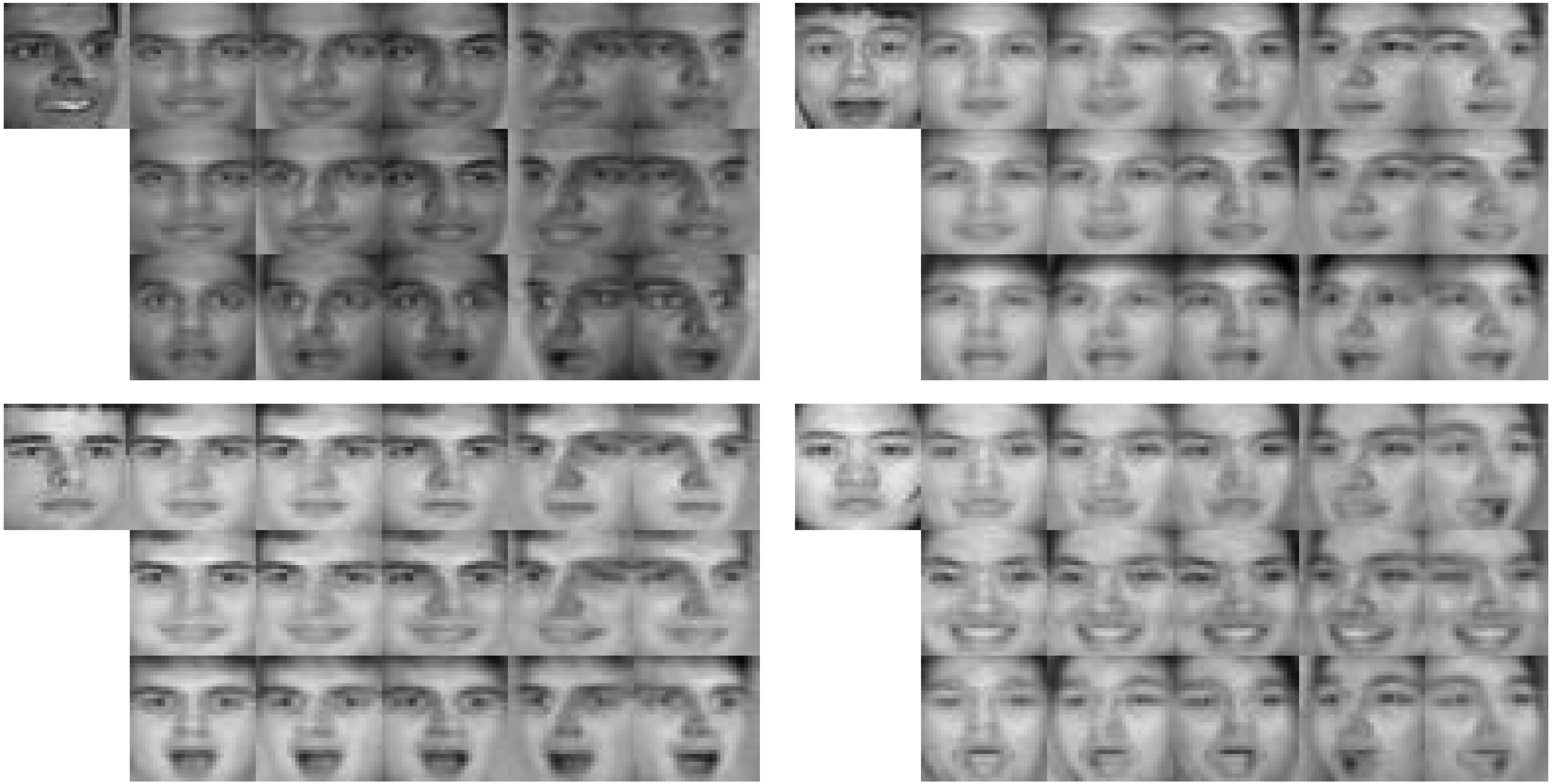
- Interpolate and predict images under viewpoints unobserved in the training set



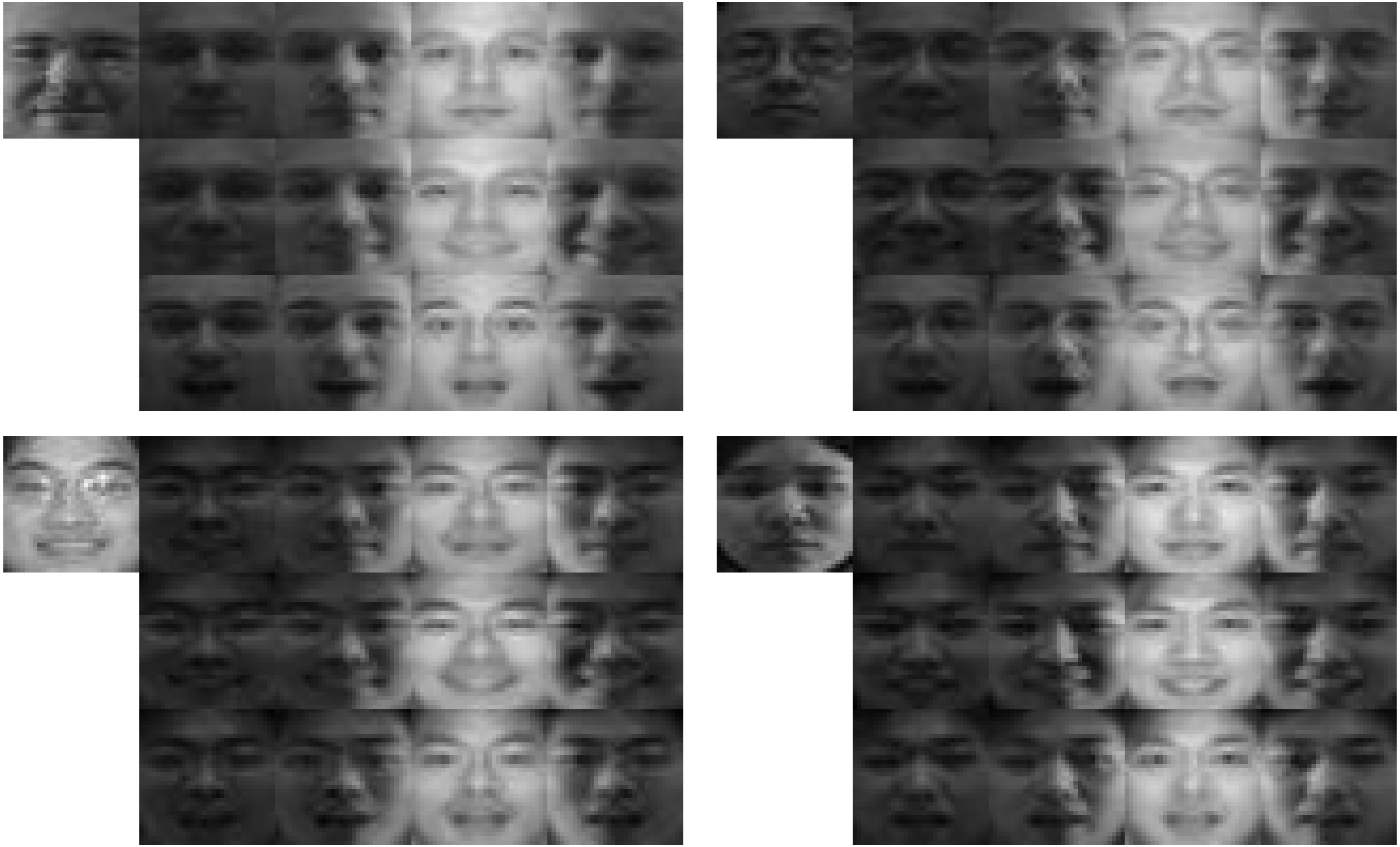
The training set only has viewpoints of 0^o, 30^o, and 60^o. (a): the reconstructed images under 15^o and 45^o when the input is taken under 0^o. (b) The input images are under 15^o and 45^o.

Generalize to other facial factors





Face reconstruction across poses and expressions



Face reconstruction across lightings and expressions

Learn face representations from

face verification, identification, multi-view reconstruction

Properties of face representations

sparseness, selectiveness, robustness

Sparsify the network

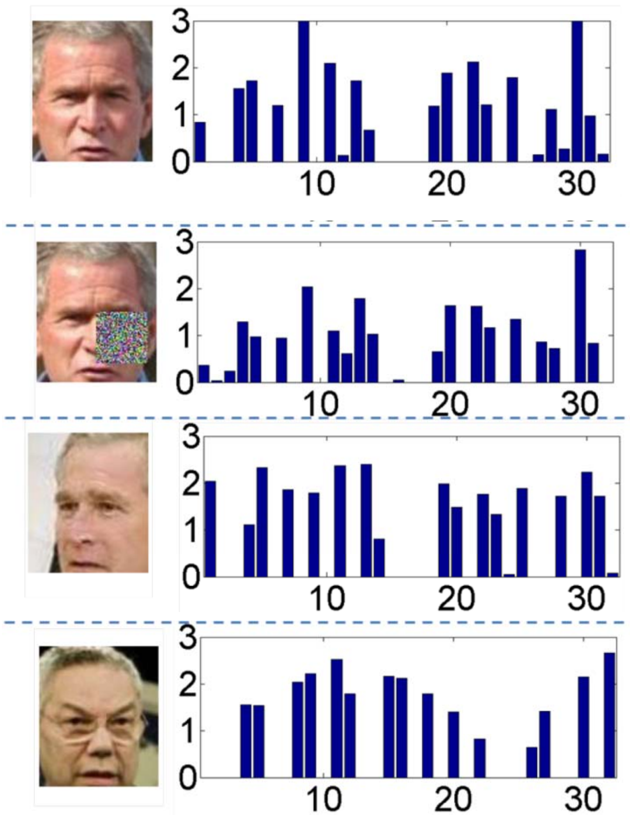
sparseness, selectiveness

Applications of face representations

face attribute recognition, face localization

Y. Sun, X. Wang, and X. Tang, CVPR 2015

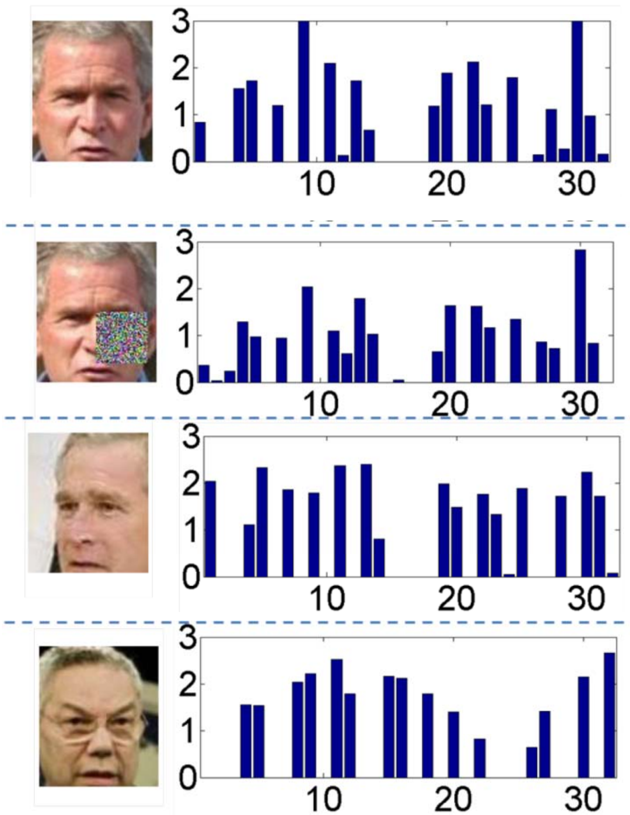
Deeply learned features are moderately sparse

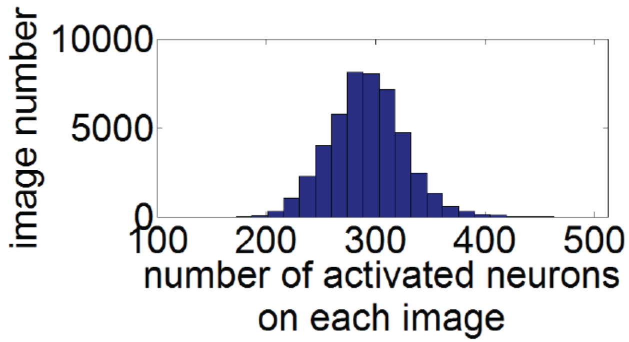


- The **binary codes** on activation patterns are very effective on face recognition
- Save storage and speedup face search dramatically
- Activation patterns are more important than activation magnitudes in face recognition

	Joint Bayesian (%)	Hamming distance (%)
Combined model (real values)	99.47	n/a
Combined model (binary code)	99.12	97.47

Deeply learned features are moderately sparse





1	0	1	1	0	0	6
0	1	0	0	1	1	

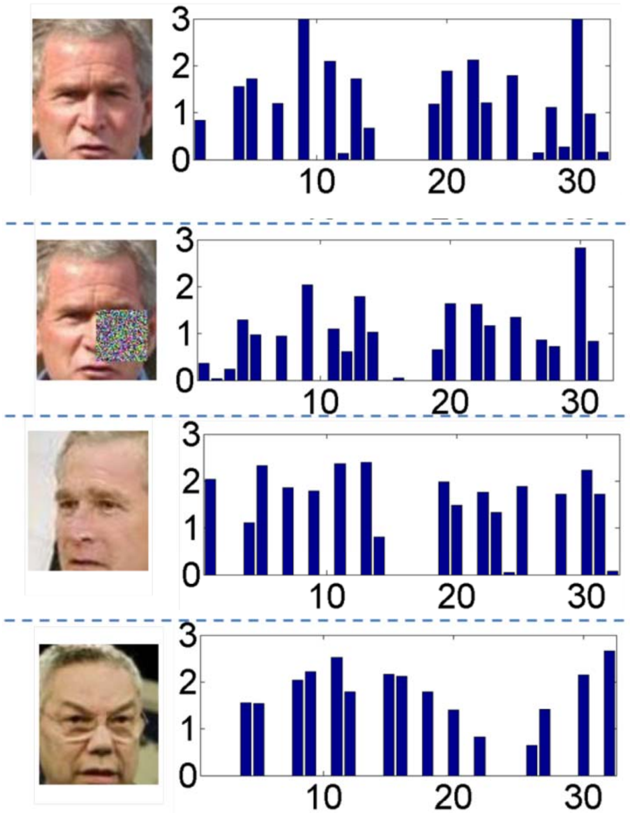
Moderately sparse

1	0	0	0	0	0	2
0	1	0	0	0	0	

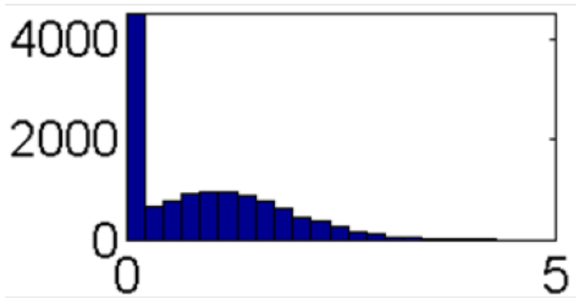
Highly sparse

- For an input image, about half of the neurons are activated
  - ✓ Maximize the Hamming distance between images

Deeply learned features are moderately sparse



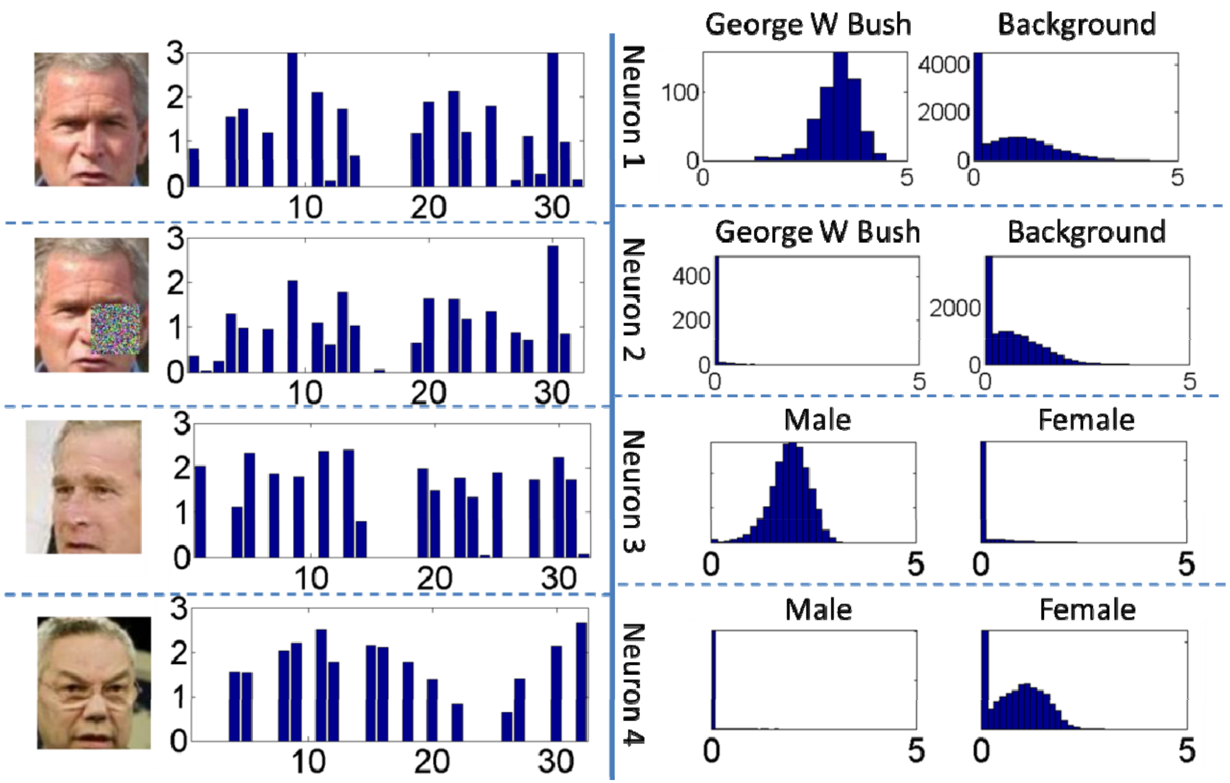
Responses of a particular neuron on all the images



- An neuron has response on about half of the images
  - ✓ Maximize the discriminative power (entropy) of a neuron on describing the image set

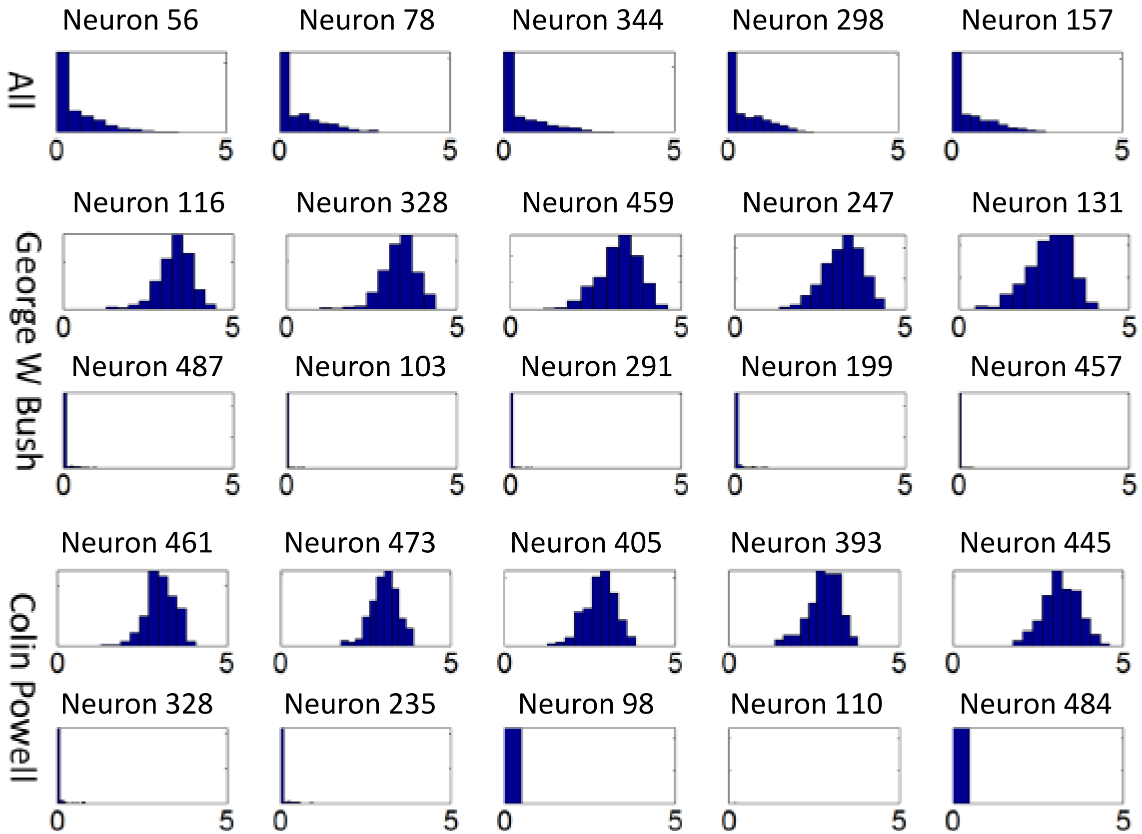
Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute

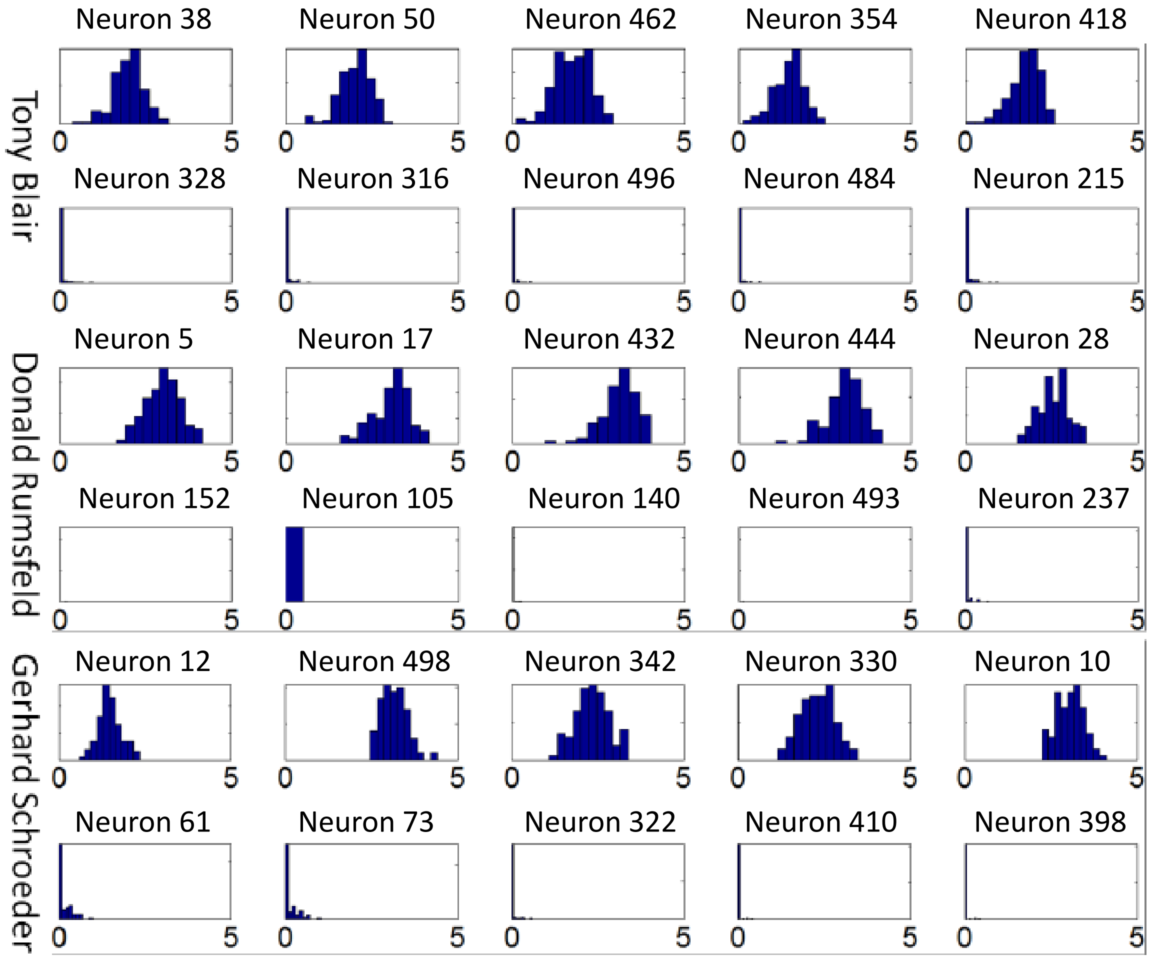


Deeply learned features are selective to identities and attributes

- Excitatory and inhibitory neurons (on identities)



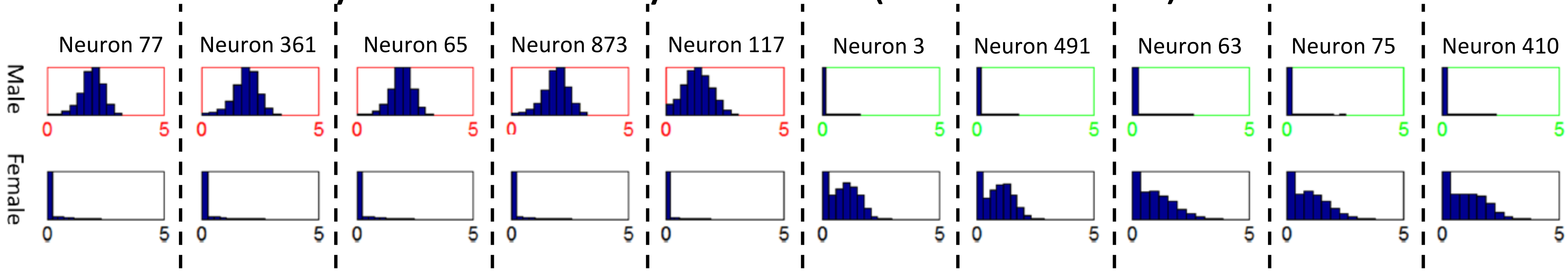
Histograms of neural activations over identities with the most images in LFW



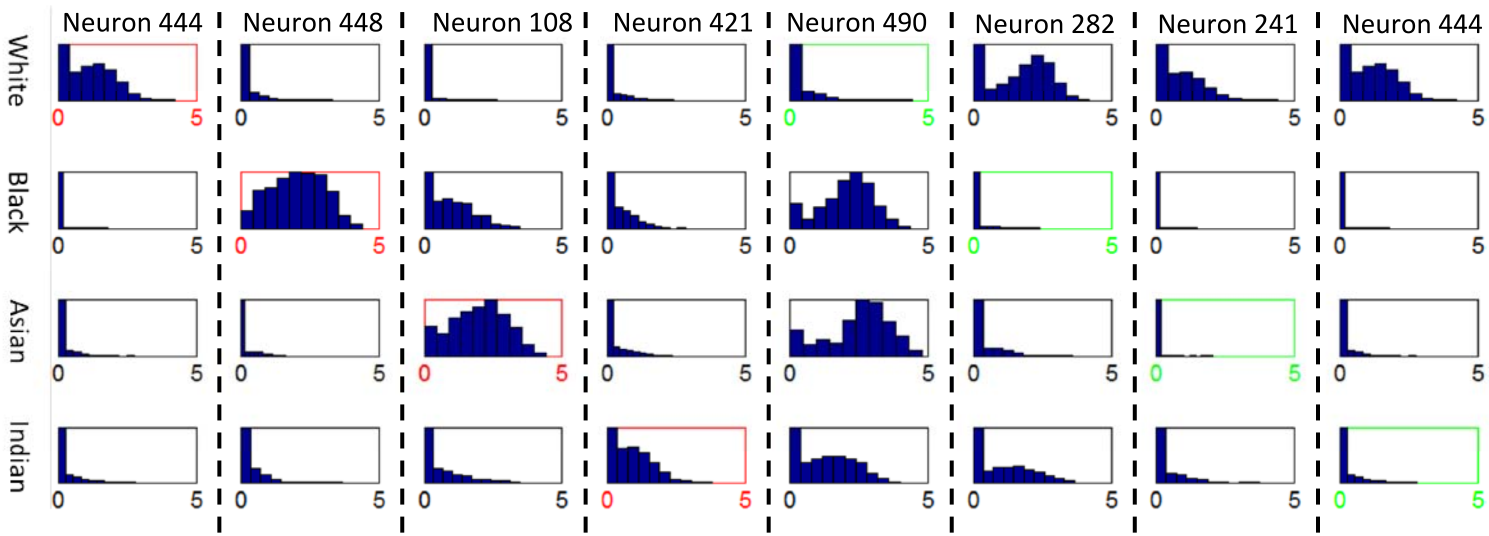
Tony Blair
Donald Rumsfeld
Gerhard Schroeder
Neuron 38
Neuron 50
Neuron 462
Neuron 354
Neuron 418
Neuron 328
Neuron 316
Neuron 496
Neuron 484
Neuron 215
Neuron 5
Neuron 17
Neuron 432
Neuron 444
Neuron 28
Neuron 152
Neuron 105
Neuron 140
Neuron 493
Neuron 237
Neuron 12
Neuron 498
Neuron 342
Neuron 330
Neuron 10
Neuron 61
Neuron 73
Neuron 322
Neuron 410
Neuron 398

Deeply learned features are selective to identities and attributes

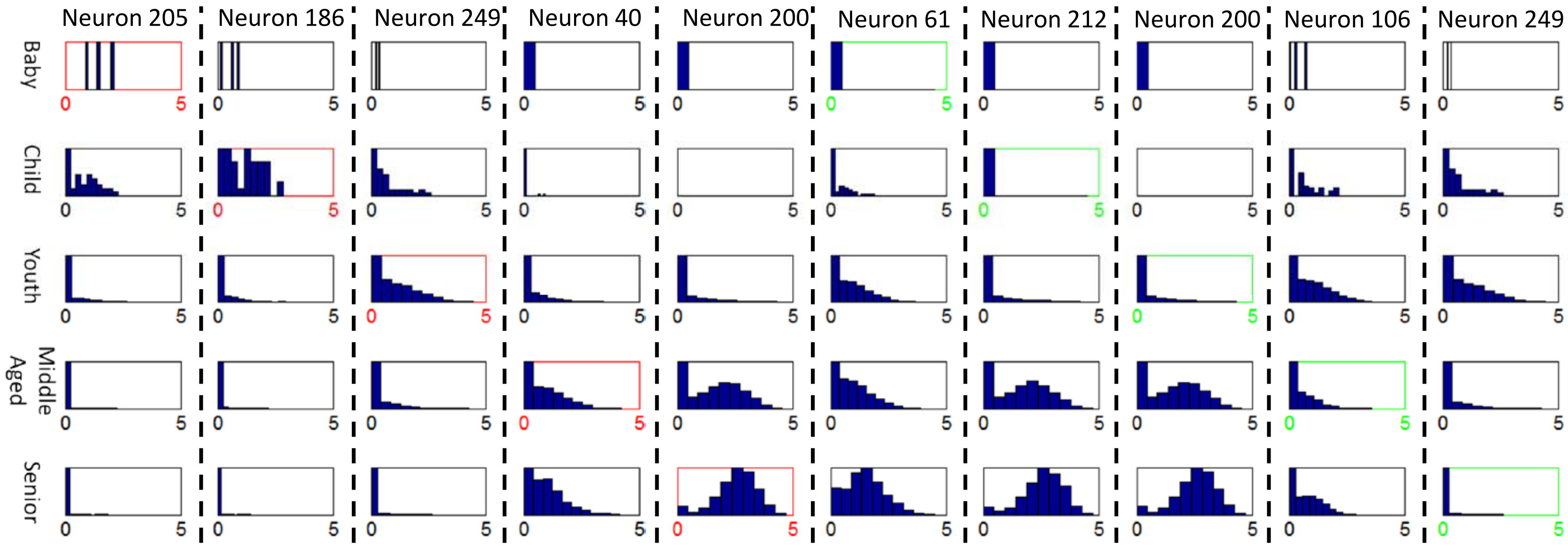
- Excitatory and inhibitory neurons (on attributes)



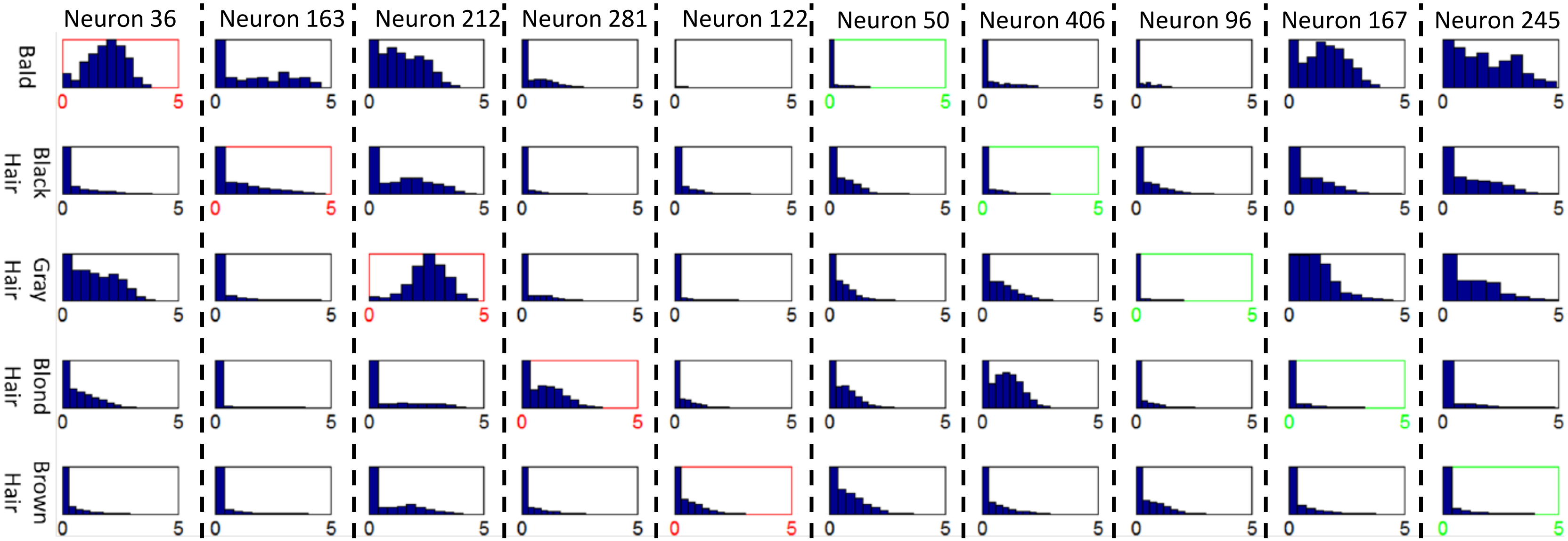
Histograms of neural activations over gender-related attributes (Male and Female)



Histograms of neural activations over race-related attributes (White, Black, Asian and India)



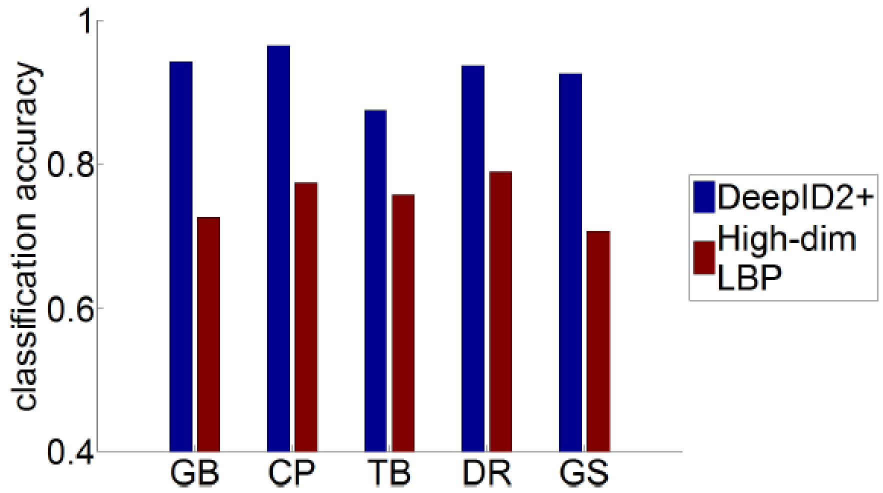
Histogram of neural activations over age-related attributes (Baby, Child, Youth, Middle Aged, and Senior)



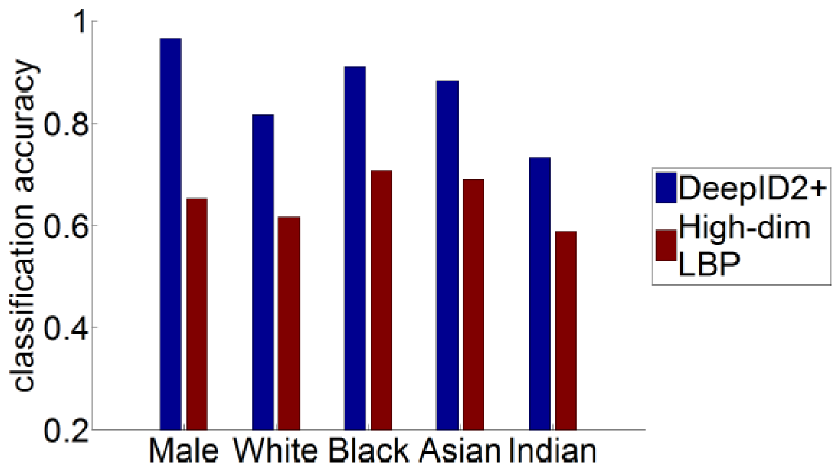
Histogram of neural activations over hair-related attributes (Bald, Black Hair, Gray Hair, Blond Hair, and Brown Hair.

Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute



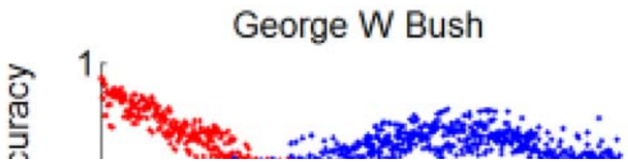
Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. GB, CP, TB, DR, and GS are five celebrities with the most images in LFW.



Attribute classification accuracy on LFW with one single DeepID2+ or LBP feature.

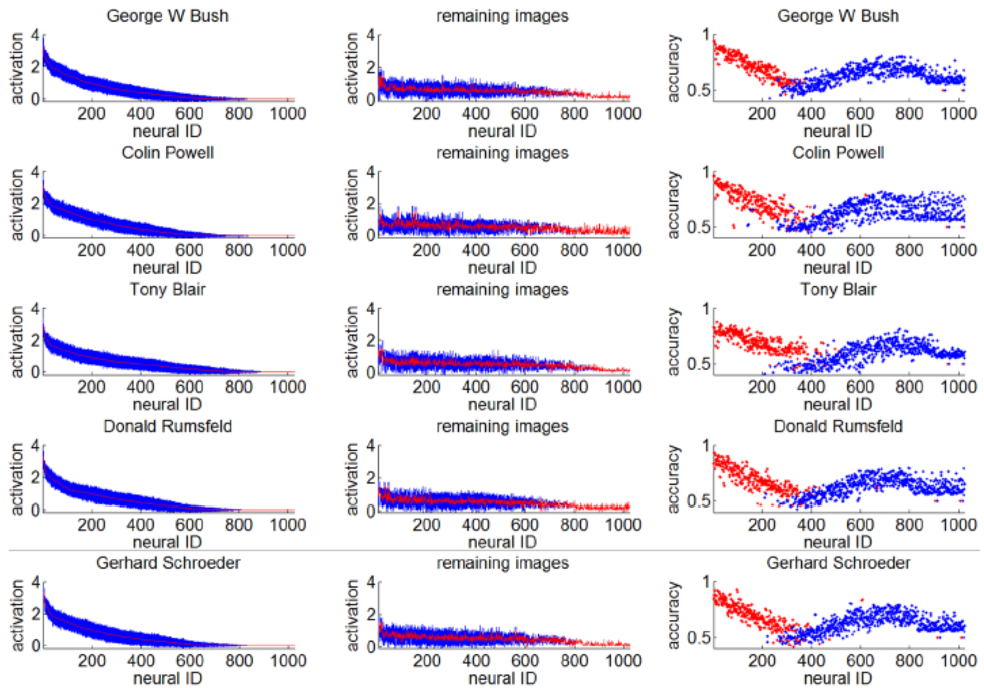
Excitatory and Inhibitory neurons

DeepID2+

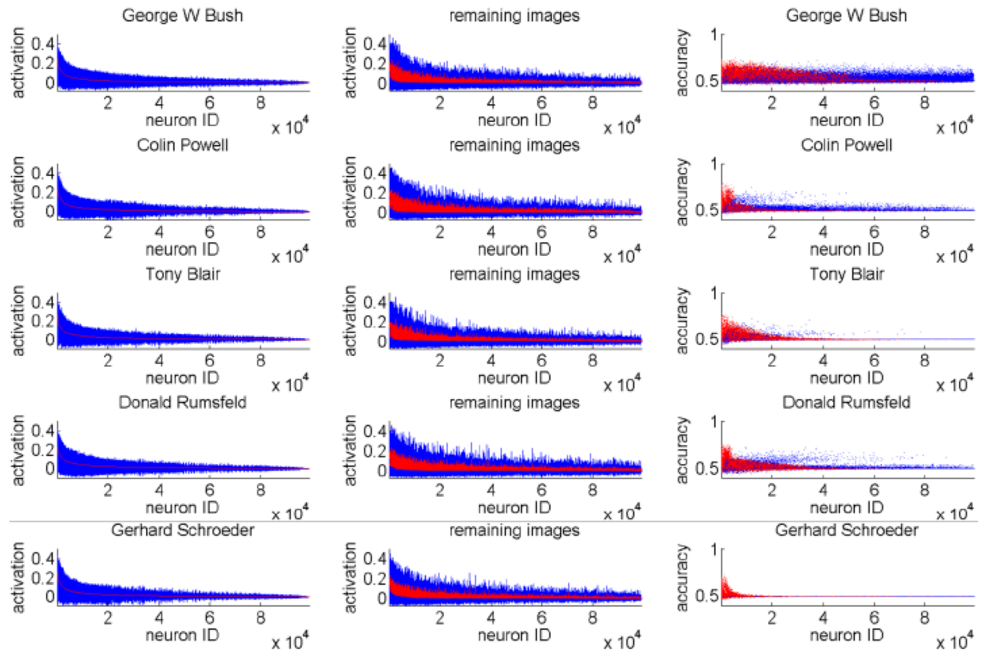


High-dim LBP

Excitatory and Inhibitory neurons

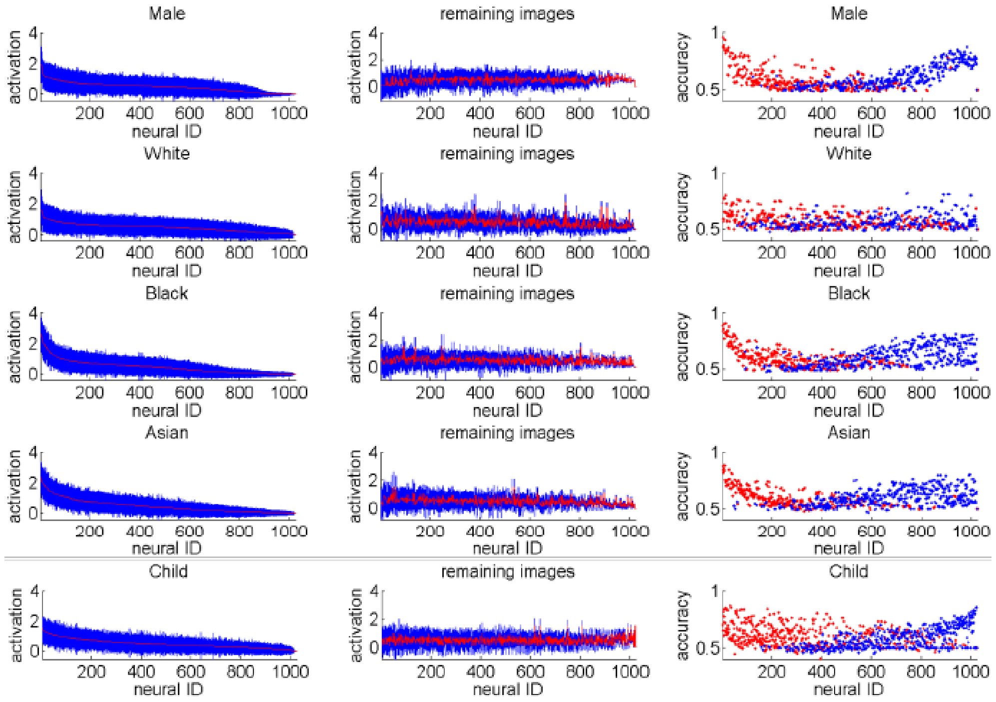


DeepID2+

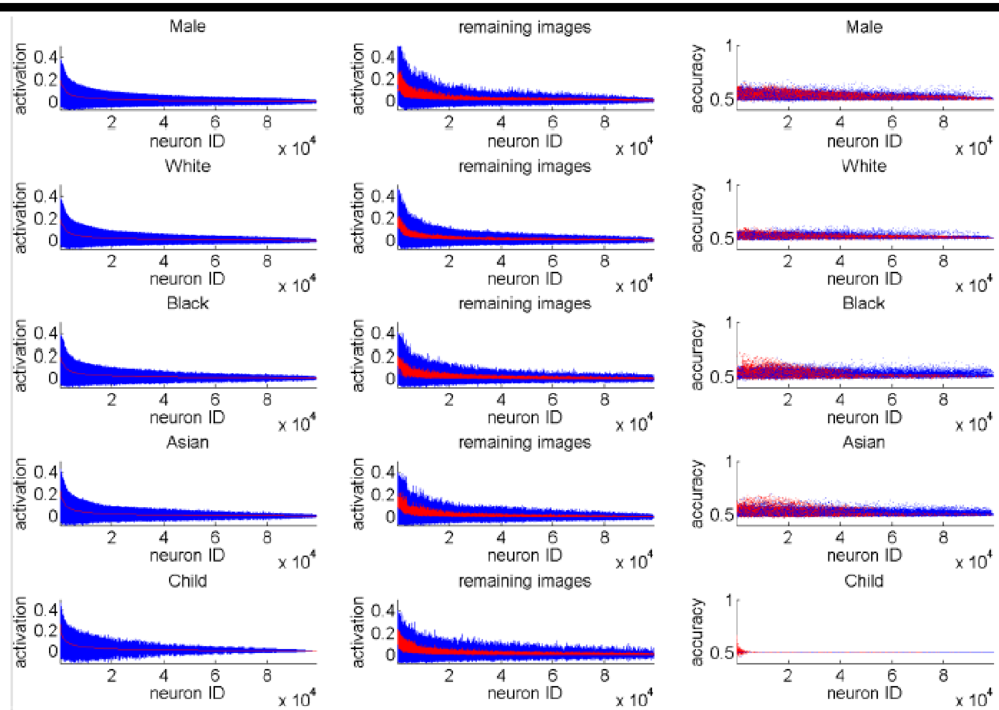


High-dim LBP

Excitatory and Inhibitory neurons



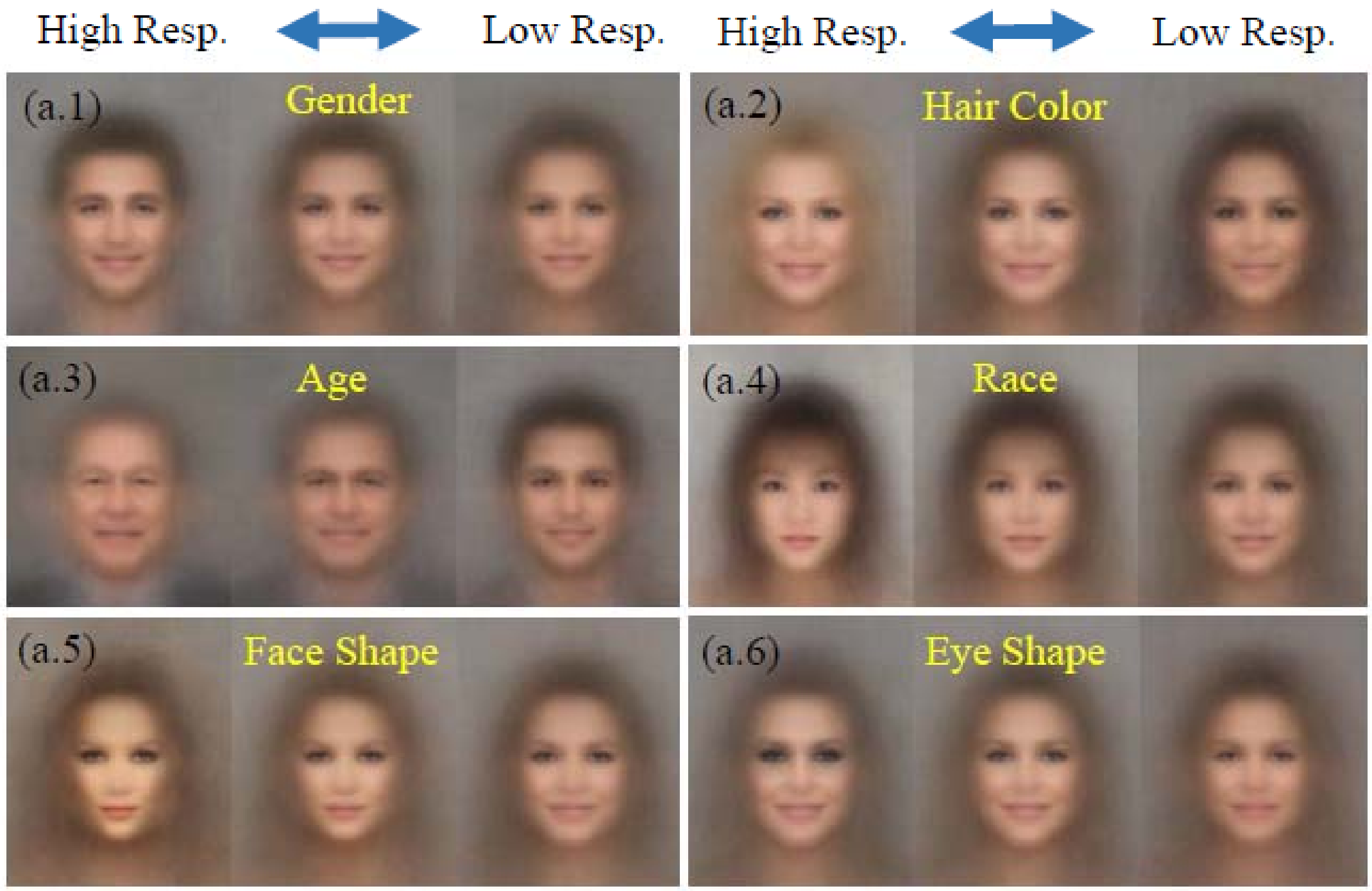
DeepID2+



High-dim LBP

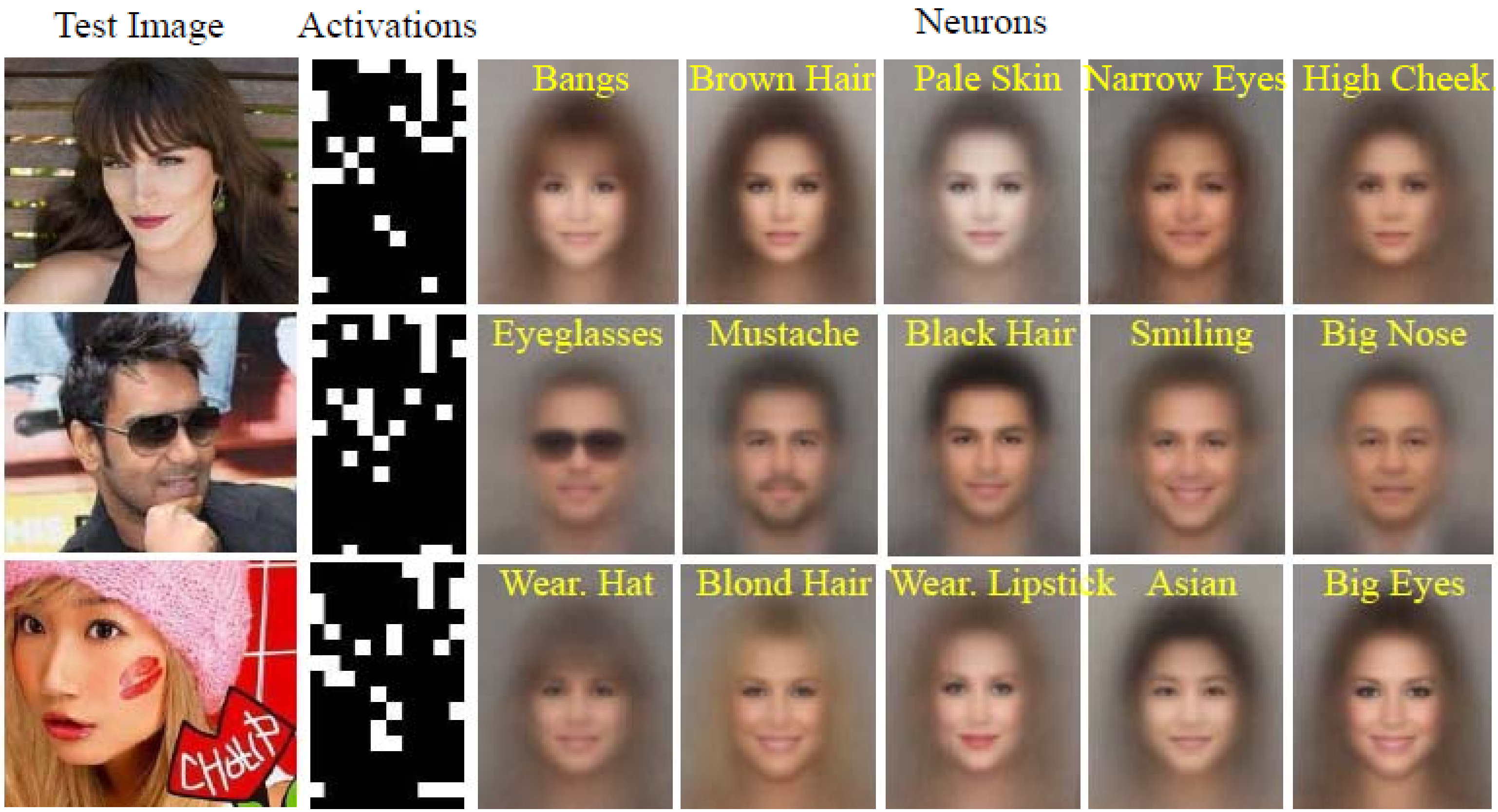
Deeply learned features are selective to identities and attributes

- Visualize the semantic meaning of each neuron



Deeply learned features are selective to identities and attributes

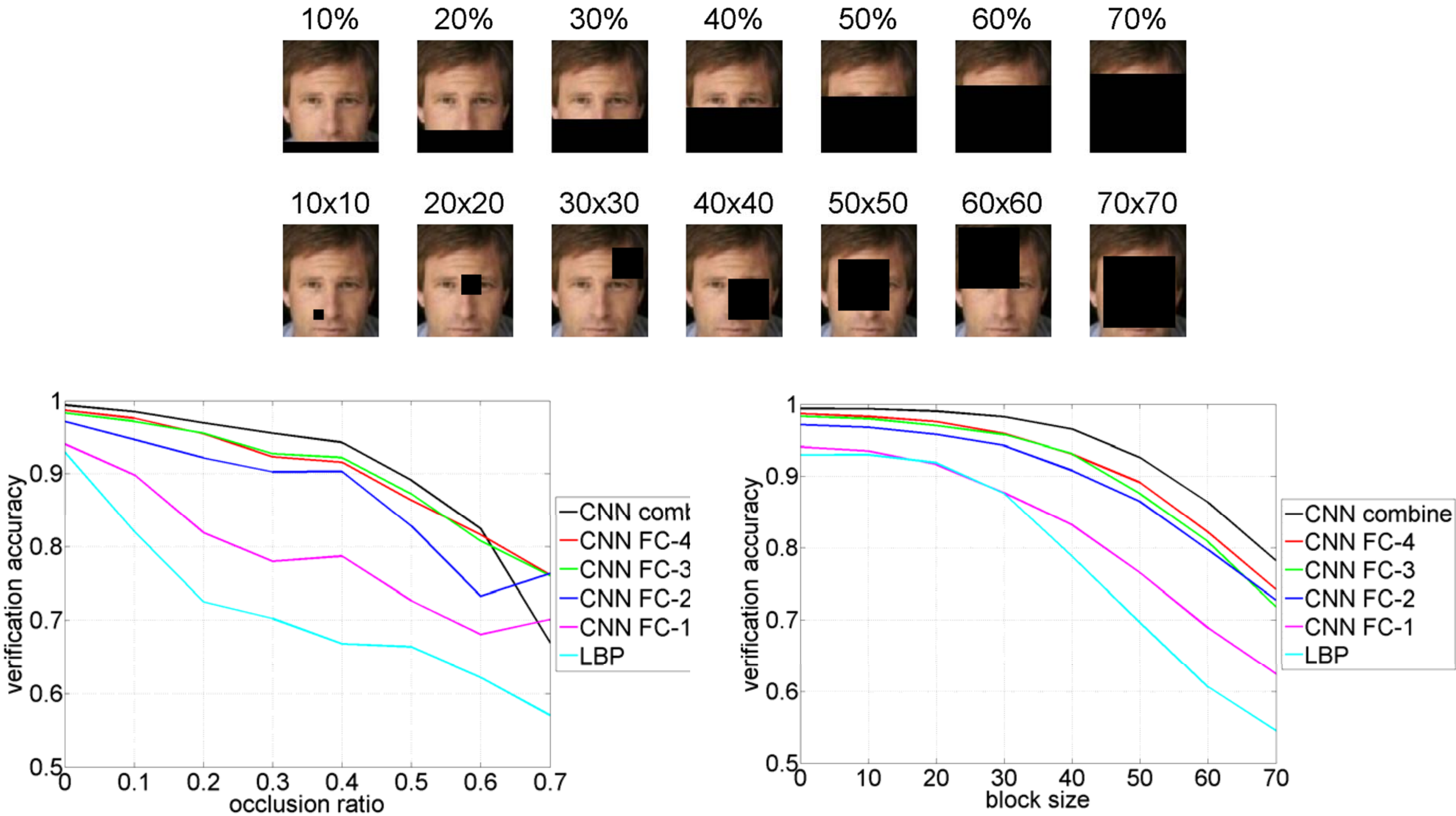
- Visualize the semantic meaning of each neuron



Neurons are ranked by their responses in descending order with respect to test images

Deeply learned features are robust to occlusions

- Global features are more robust to occlusions



Learn face representations from

face verification, identification, multi-view reconstruction

Properties of face representations

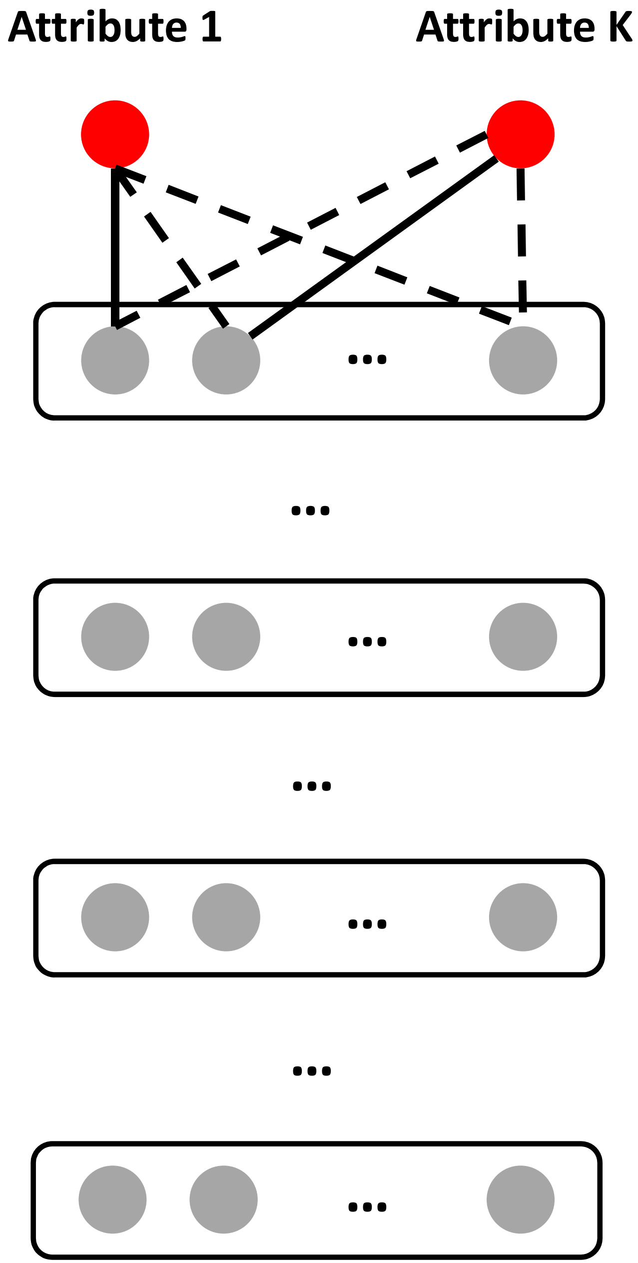
sparseness, selectiveness, robustness

Sparsify the network according to neural selectiveness

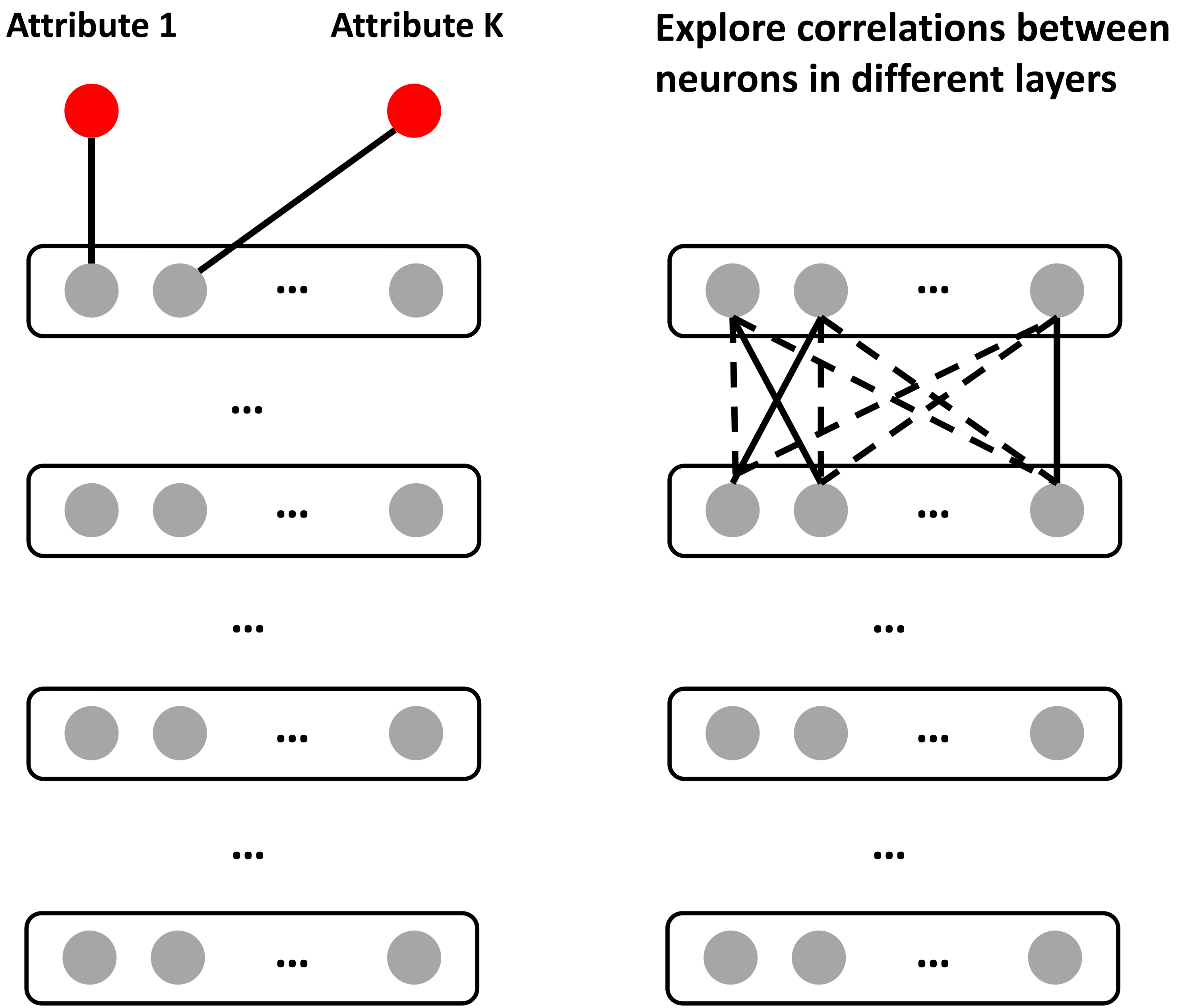
sparseness, selectiveness

Applications of face representations

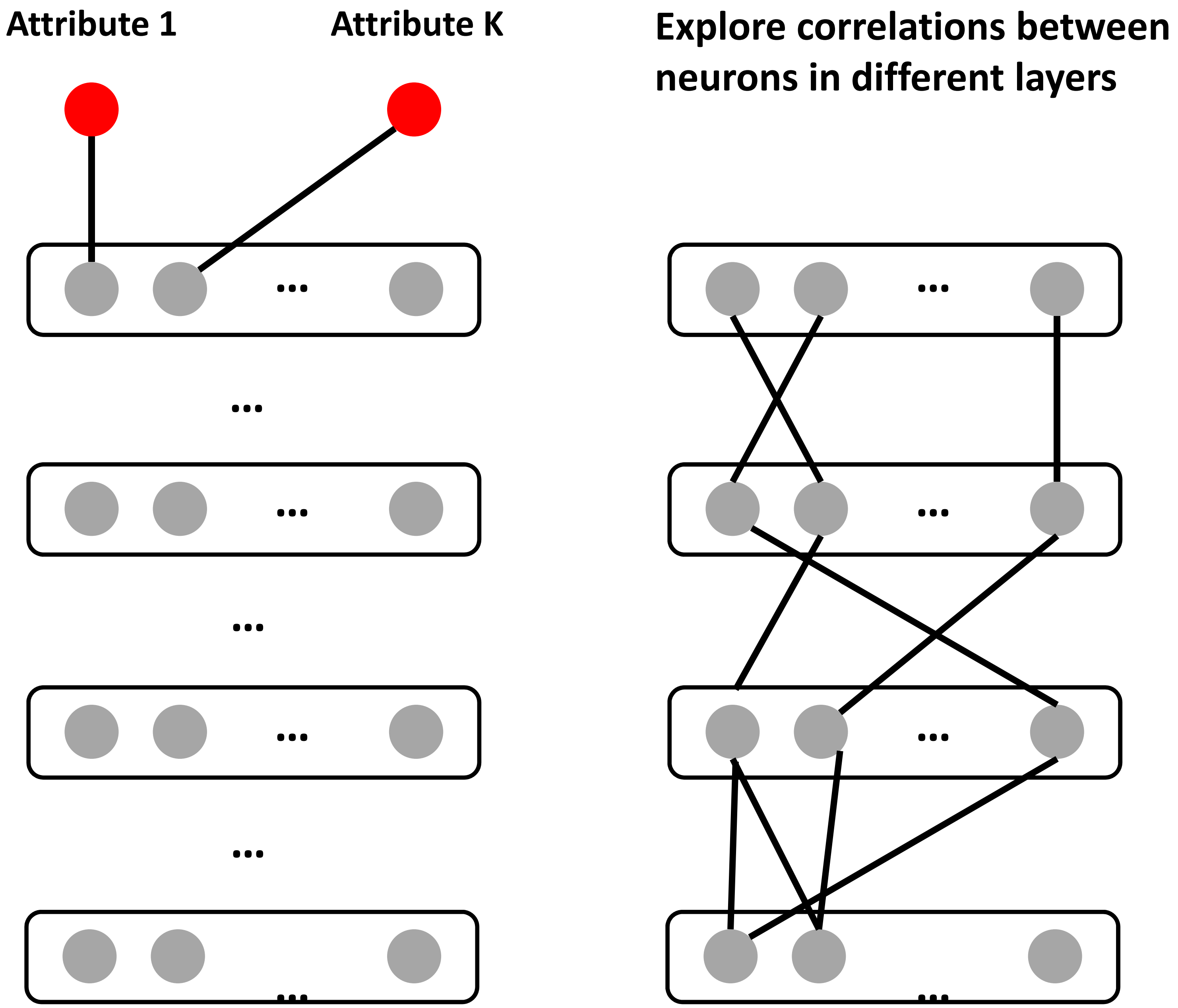
face localization, attribute recognition



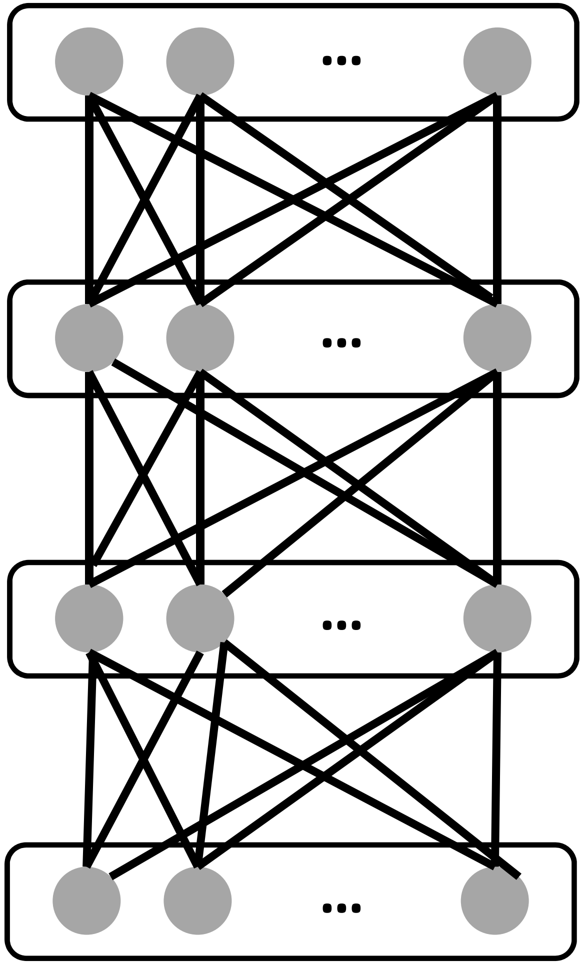
Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Sparsifying Neural Network Connections for Face Recognition,” arXiv:1512.01891, 2015



Attribute 1
Attribute K
Explore correlations between neurons in different layers


Attribute 1
Attribute K
Explore correlations between neurons in different layers
...

Alternatively learning weights and net structures



1. Train a dense network from scratch
2. Sparsify the top layer, and **re-train** the net
3. Sparsify the second top layer, and **re-train** the net

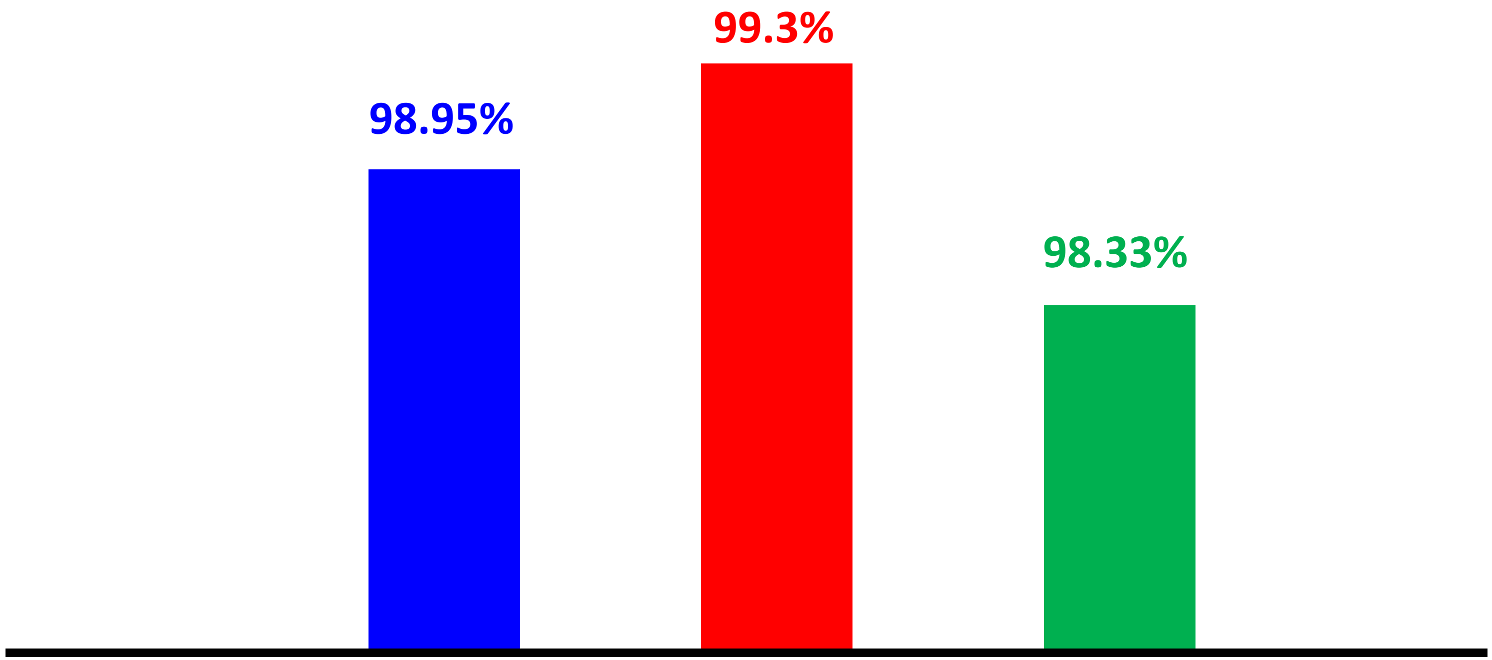
...

Conel, JL. The postnatal development of the human cerebral cortex. Cambridge, Mass: Harvard University Press, 1959.

Original deep neural network

Sparsified deep neural network and only keep 1/8 amount of parameters after joint optimization of weights and structures

Train the sparsified network from scratch



The sparsified network has enough learning capacity, but the original denser network helps it reach a better intialization

Learn face representations from

face verification, identification, multi-view reconstruction

Properties of face representations

sparseness, selectiveness, robustness

Sparsify the network according to neural selectiveness

sparseness, selectiveness

Applications of face representations

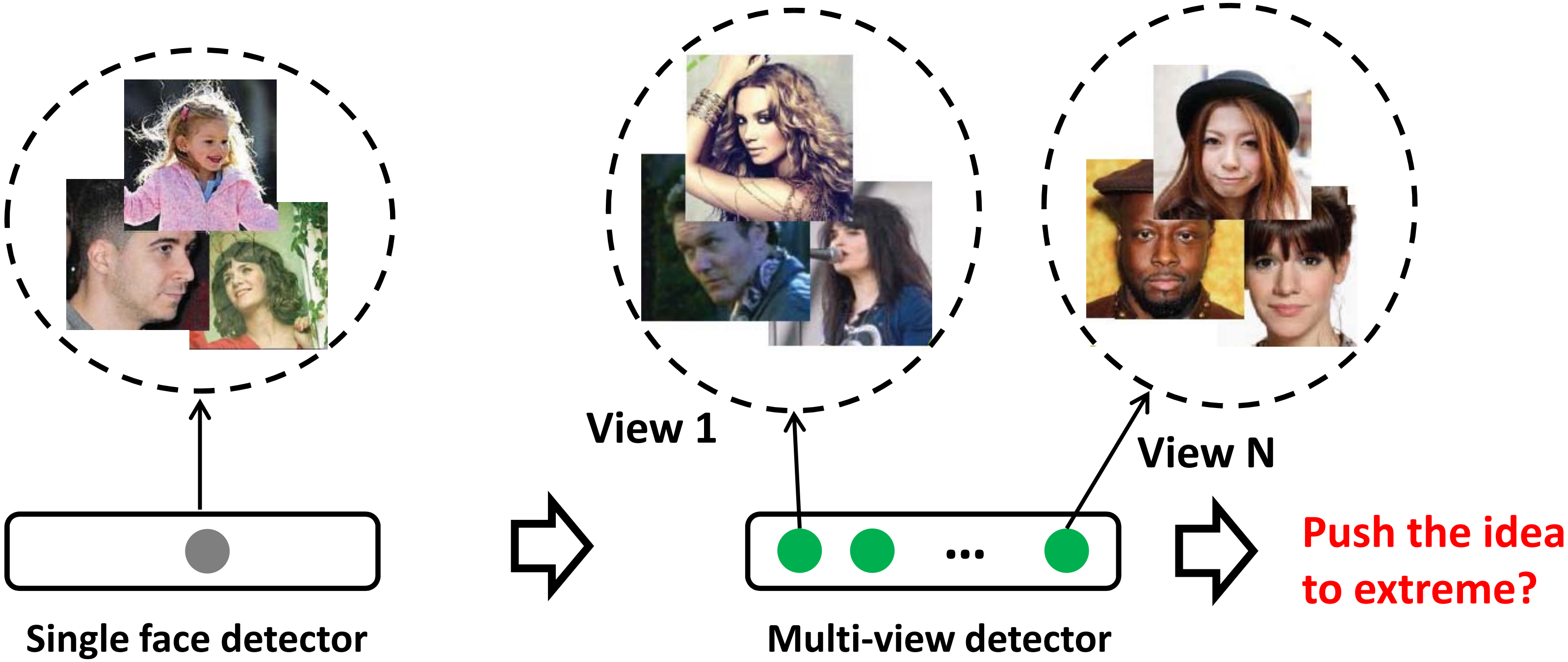
face localization, attribute recognition

DeepID2 features for attribute recognition

- DeepID2 features can be directly used for attribute recognition
- Use DeeID2 features as initialization (pre-trained result), and then fine tune on attribute recognition
- Multi-task learning face recognition and attribute prediction does not improve performance, because face recognition is a much stronger supervision than attribute prediction
- Average accuracy on 40 attributes on CelebA and LFWA datasets

	CelebA	LFWA
FaceTracer [1] (HOG+SVM)	81	74
Training CNN from scratch with attributes	83	79
Directly use DeepID2 features	**84**	**82**
DeepID2 + fine-tuning	**87**	**84**

Features learned from face recognition can improve face localization?



Hard to handle large variety especially on views

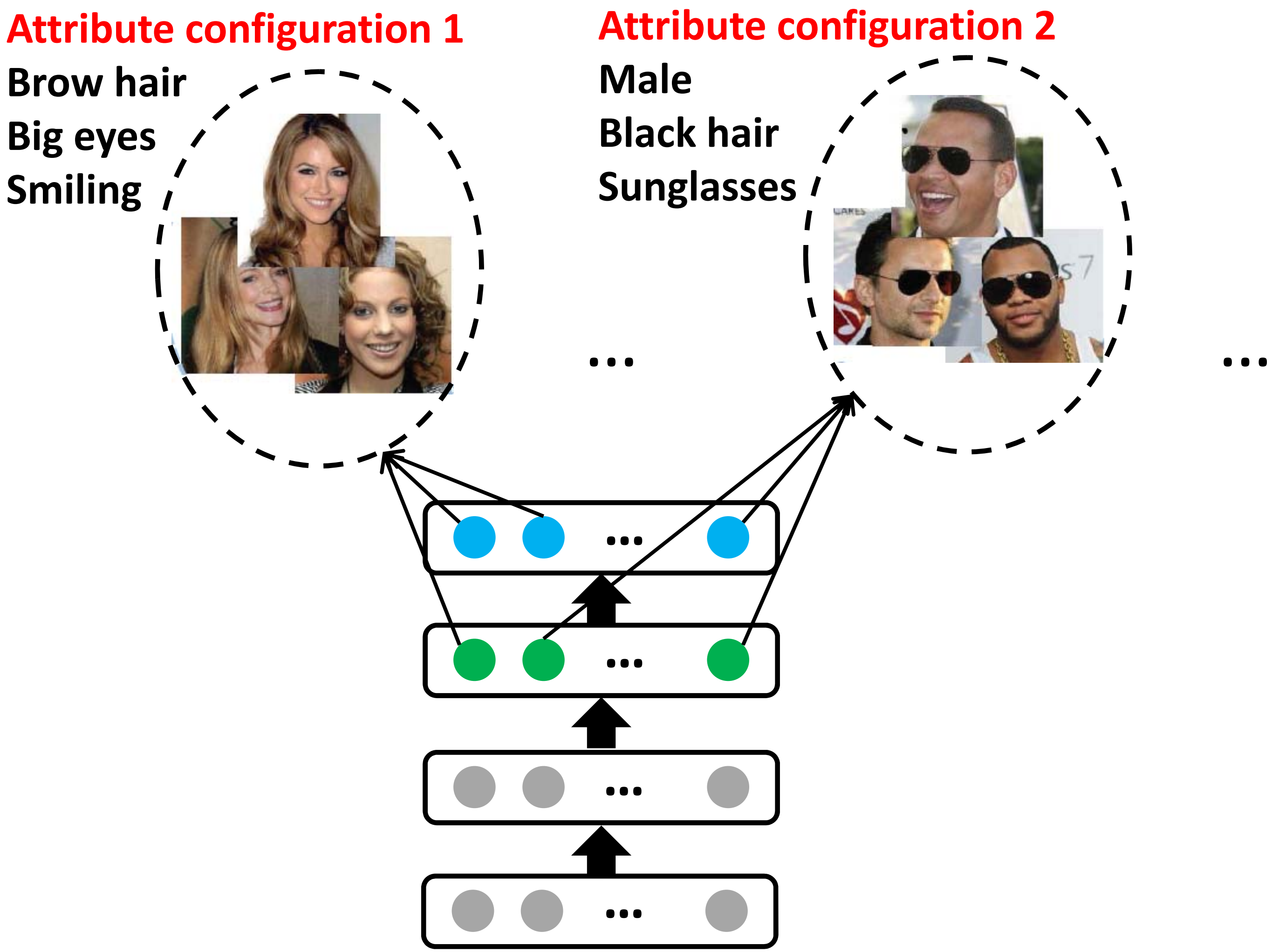
View labels are given in training;
Each detector handles a view

Viewpoints → Gender, expression, race, hair style → Attributes

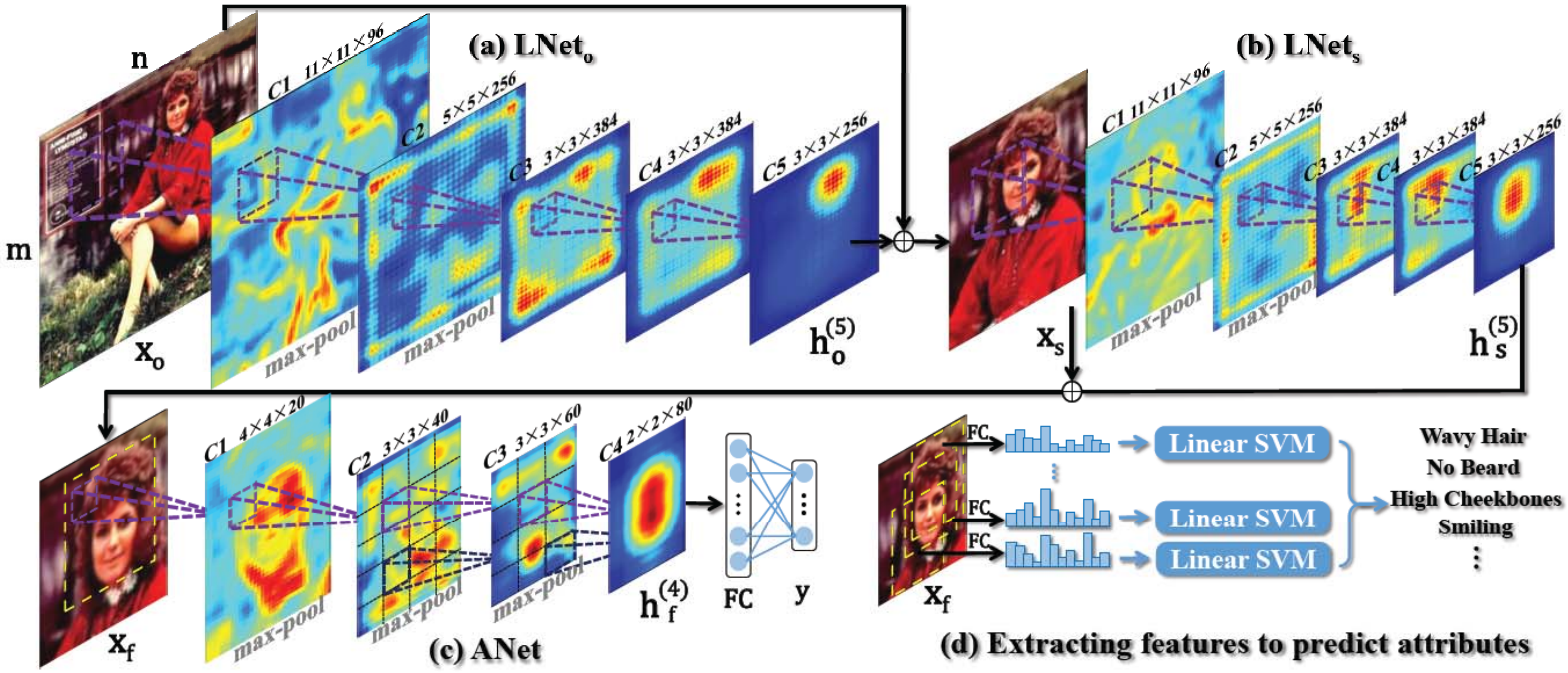
Neurons have selectiveness on attributes

A filter (or a group of filters) functions as a detector of a face attribute

When a subset of neurons are activated, they indicate existence of faces with an attribute configuration



The neurons at different layers can form many activation patterns, implying that the whole set of face images can be divided into many subsets based on attribute configurations

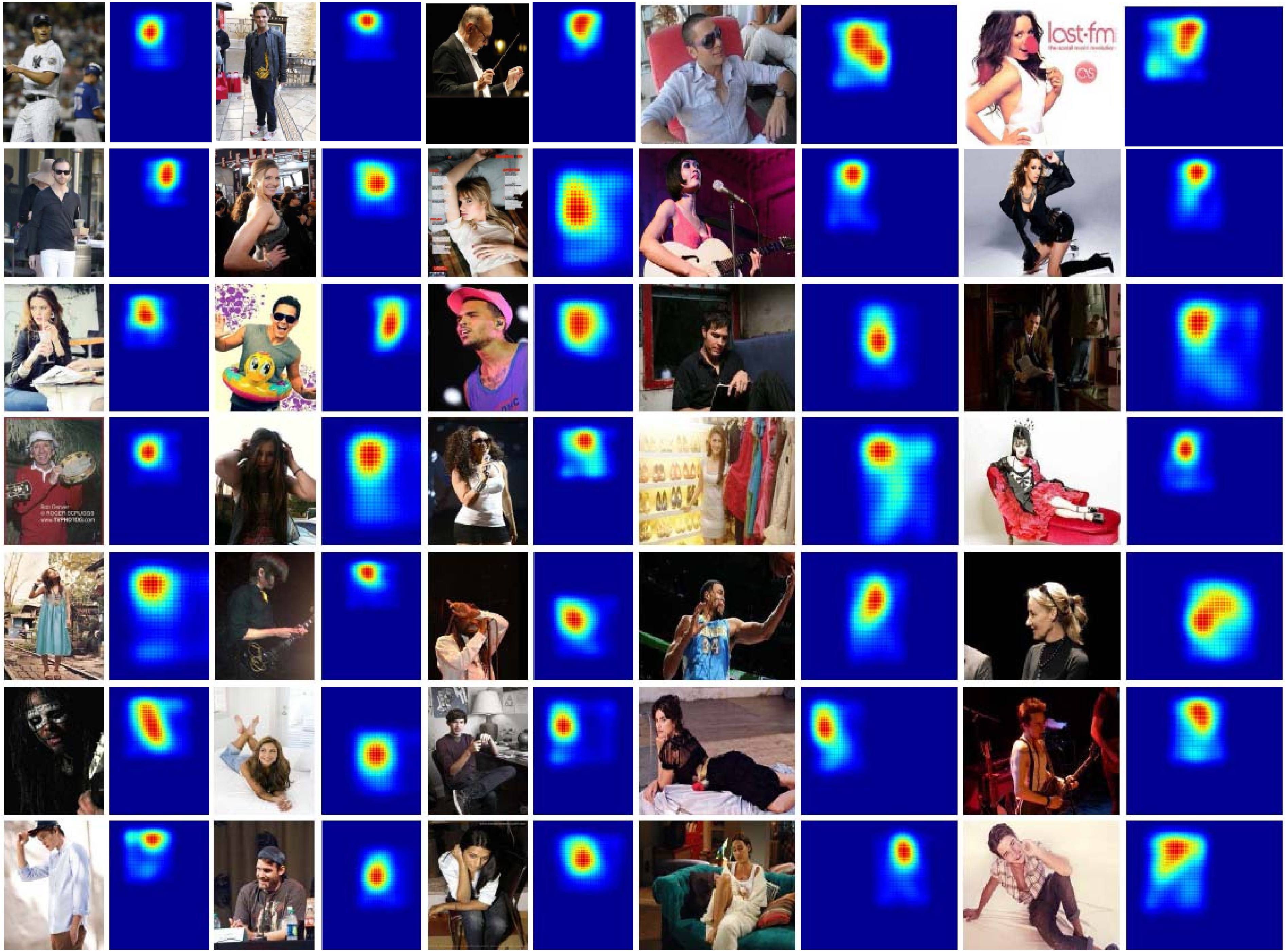


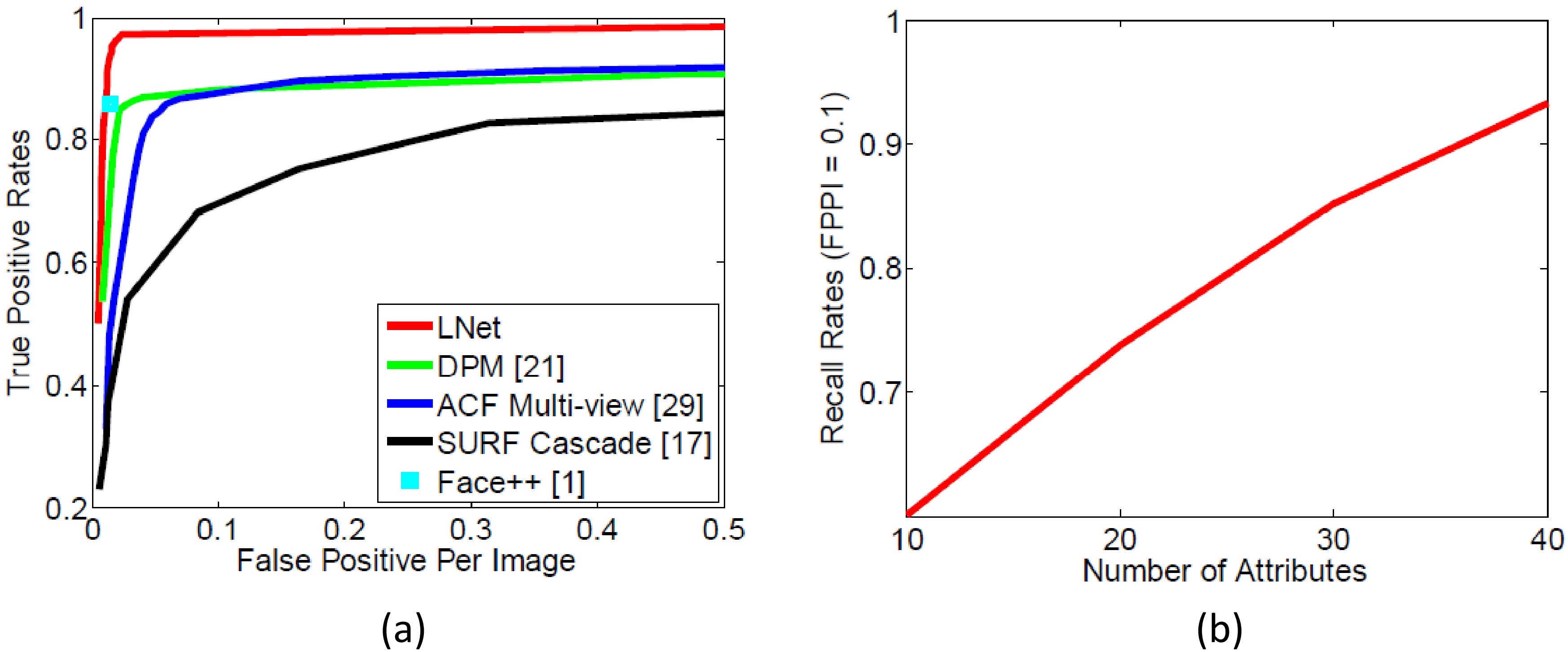
LNet localizes faces

LNet is pre-trained with face recognition and fine-tuned with attribute prediction

By simply averaging response maps and good face localization is achieved

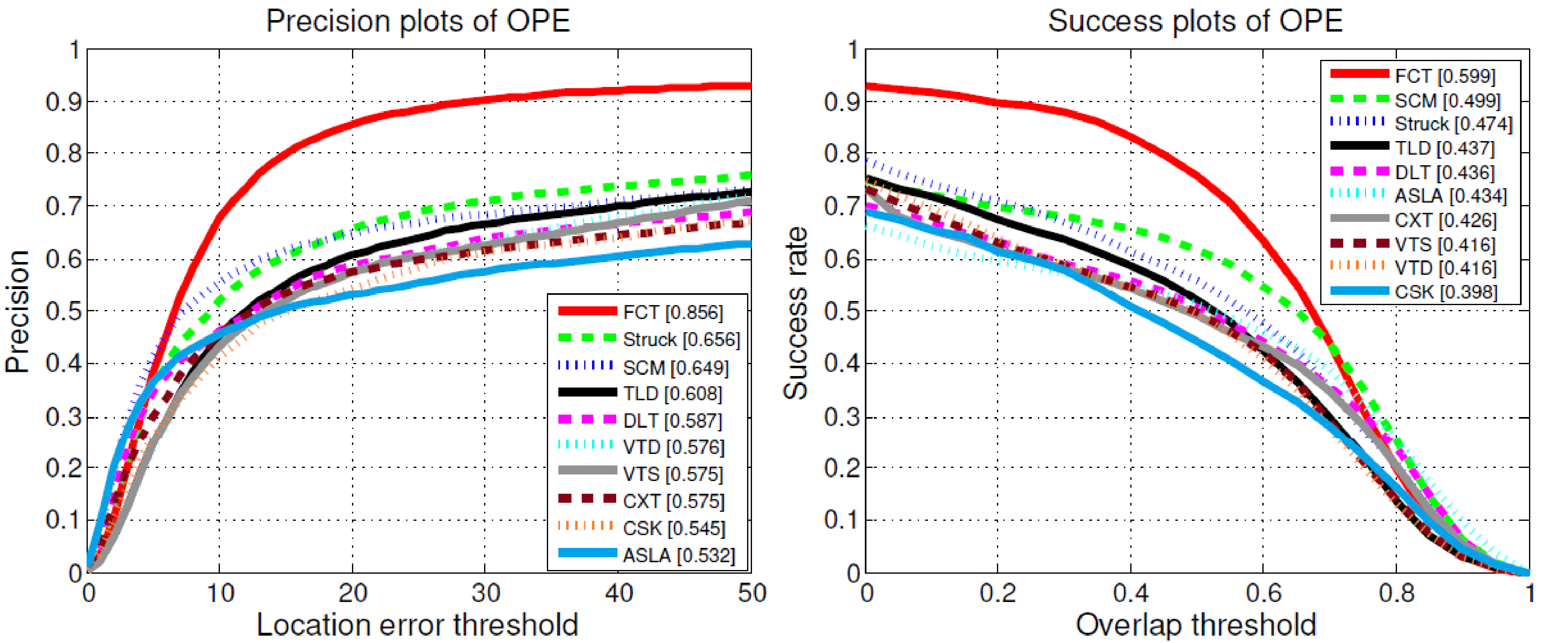
Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” ICCV 2015





(a) ROC curves of LNet and state-of-the-art face detectors
(b) Recall rates w.r.t. number of attributes (FPPI = 0.1)

Attribute selectiveness: **neurons serve as detectors**
Identity selectiveness: **neurons serve as trackers**



L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual Tracking with Fully Convolutional Networks," ICCV 2015.

Conclusions

- Face representation can be learned from the tasks of verification, identification, and multi-view reconstruction
- Deeply learned features are moderately sparse, identity and attribute selective, and robust to data corruption
- The net can be sparsified substantially by alternatively optimizing the weights and structures
- Because of these properties, the learned face representation are effective for applications beyond face recognition, such as face localization and attribute prediction

Collaborators

Yi Sun

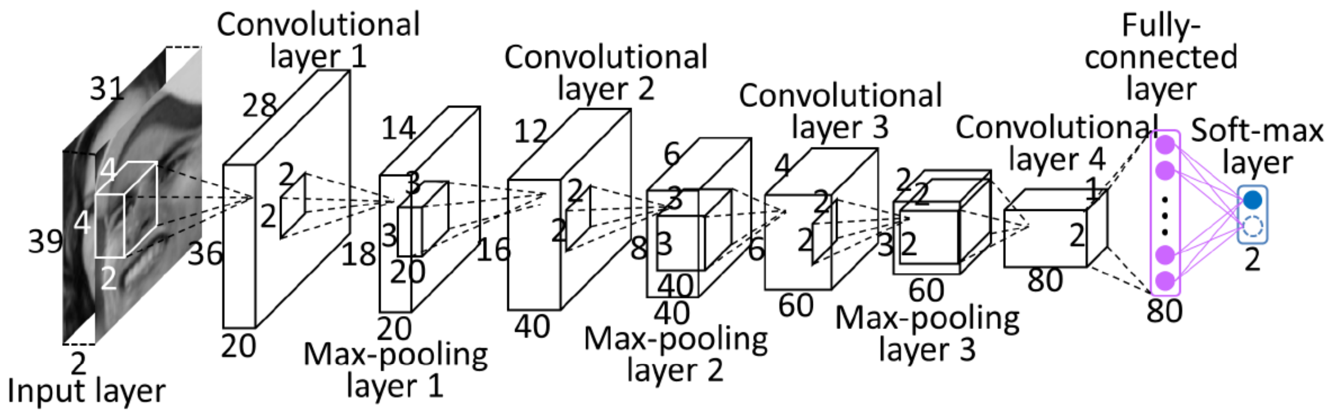
Ziwei Liu

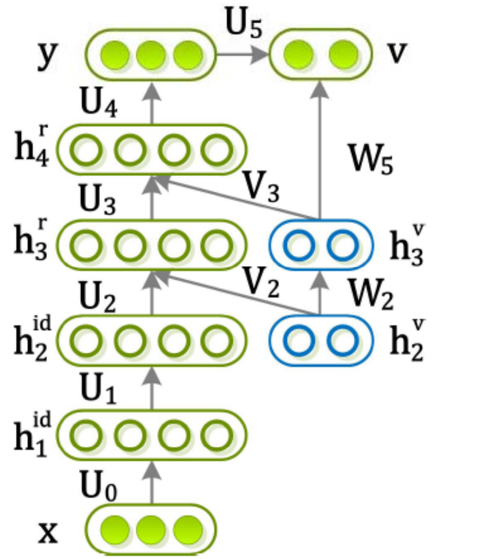
Zhenyao Zhu

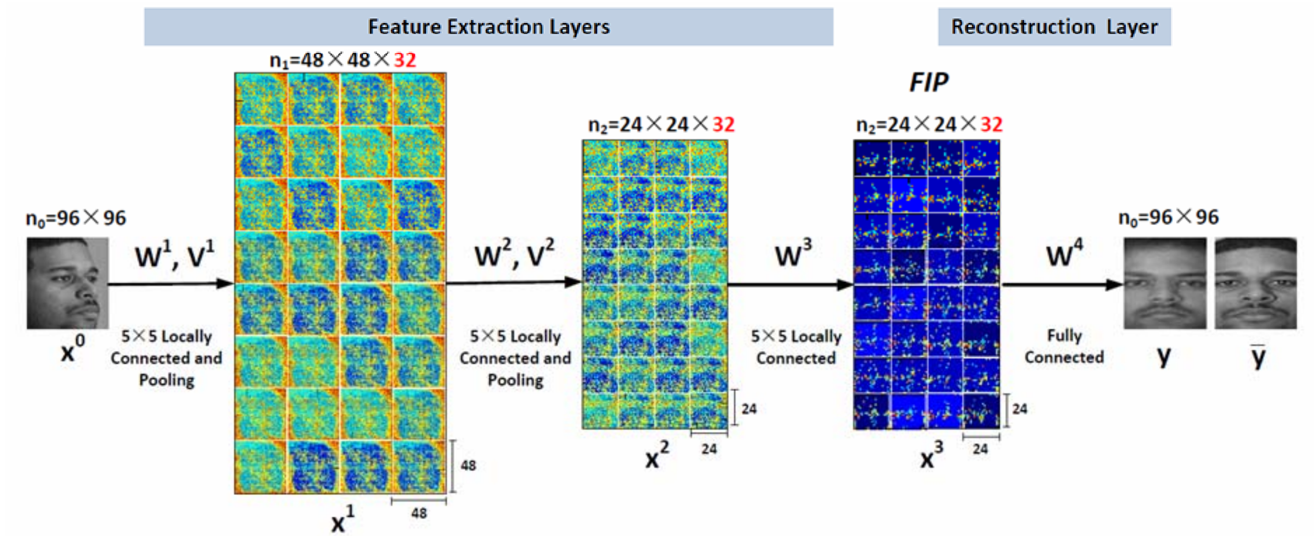
Ping Luo

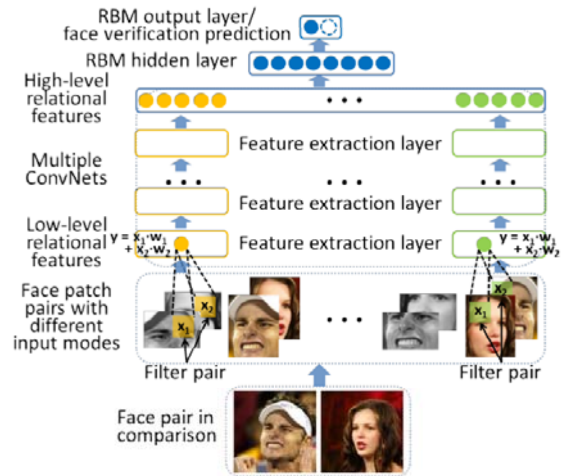
Xiaoou Tang

Thank you!









http://mmlab.ie.cuhk.edu.hk/

http://www.ee.cuhk.edu.hk/~xgwang/