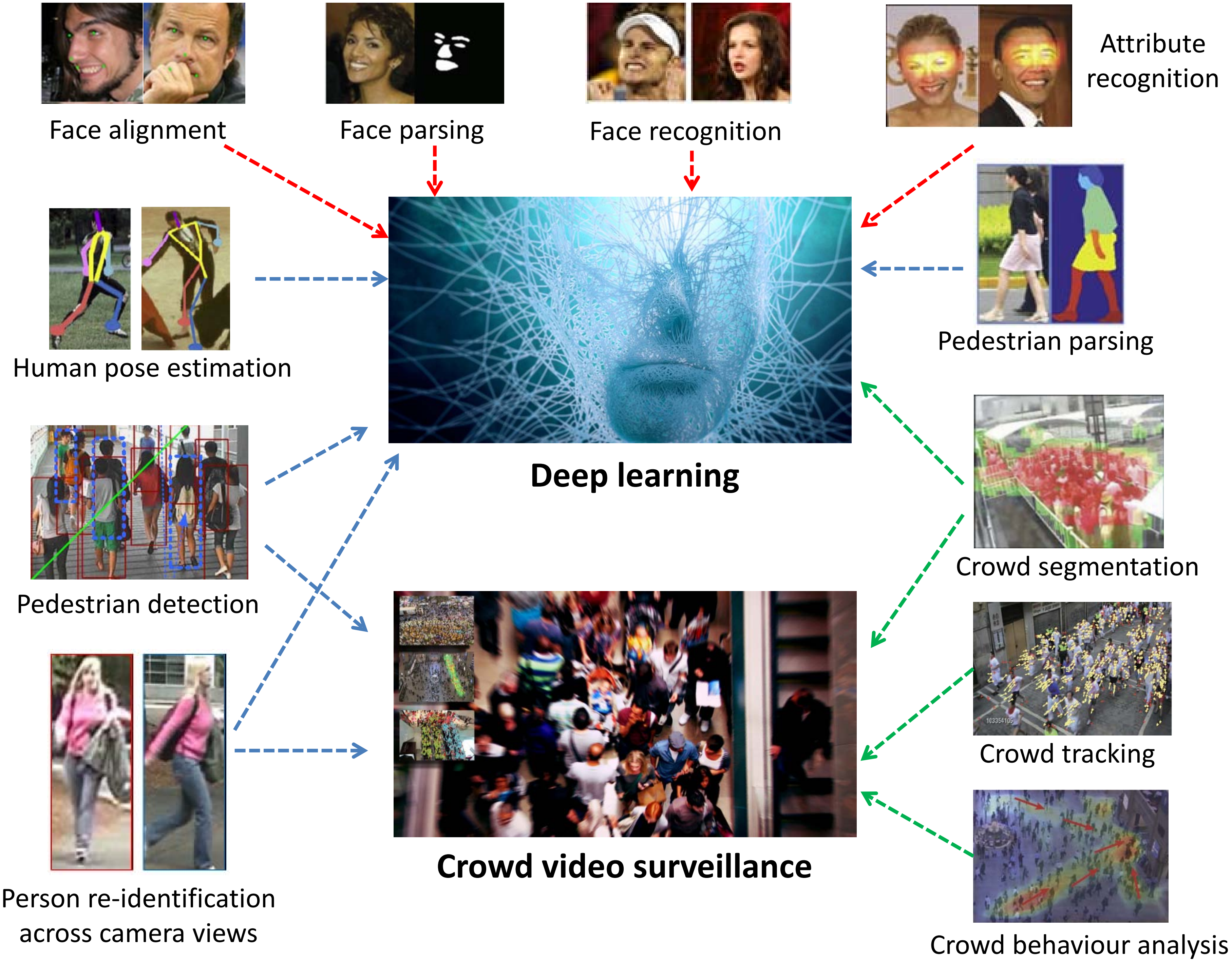
Deep Learning in Object Detection, Segmentation, and Recognition

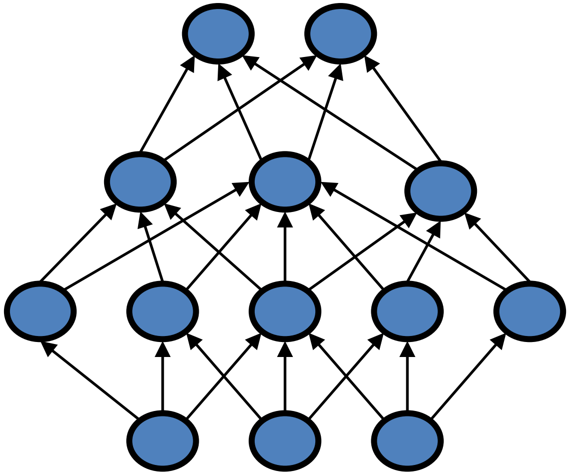
Xiaogang Wang

Department of Electronic Engineering,
The Chinese University of Hong Kong


Face alignment
Face parsing
Face recognition
Attribute recognition
Human pose estimation
Pedestrian parsing
Deep learning
Pedestrian detection
Crowd segmentation
Crowd tracking
Person re-identification across camera views
Crowd video surveillance
Crowd behaviour analysis

Neural network
Back propagation

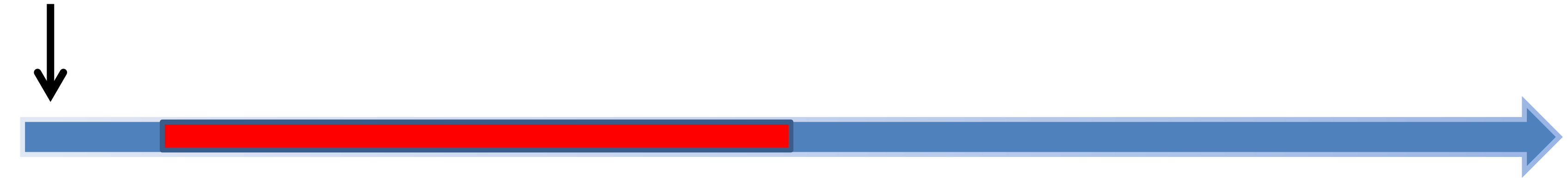
1986



- Solve general learning problems
- Tied with biological system

But it is given up…

- Hard to train
- Insufficient computational resources
- Small training sets
- Does not work well



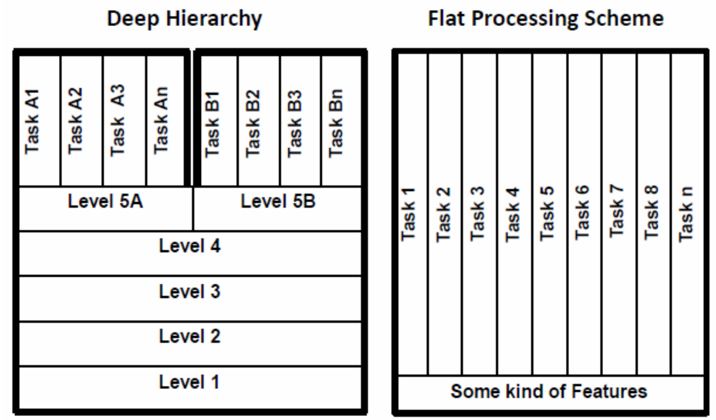
Neural network
Back propagation

1986

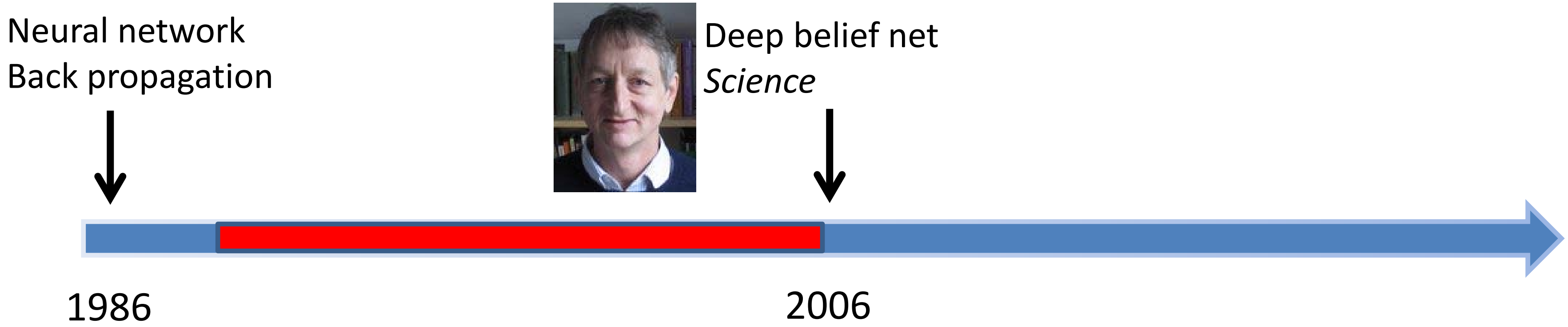
2006

- SVM
- Boosting
- Decision tree
- KNN
- …

- Loose tie with biological systems
- Flat structures
- Specific methods for specific tasks
  - Hand crafted features (GMM-HMM, SIFT, LBP, HOG)



Kruger et al. TPAMI’13

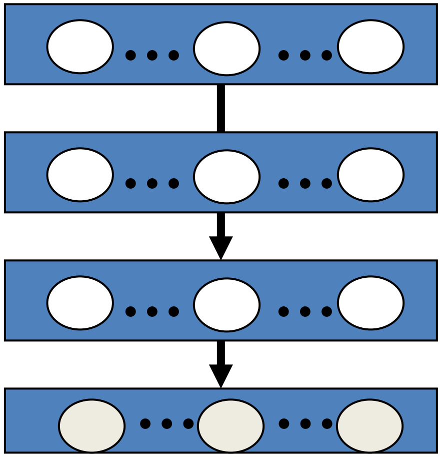


Neural network
Back propagation

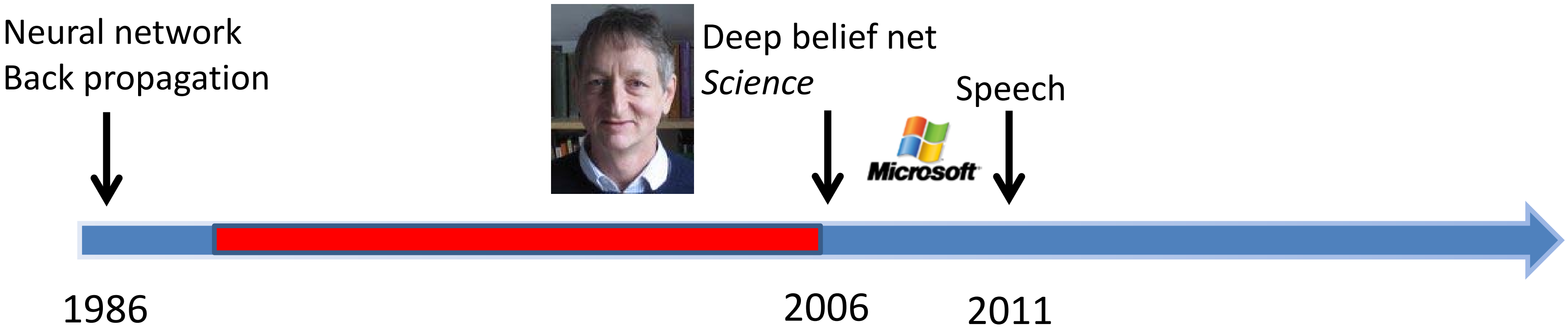
Deep belief net
Science

1986

2006



- Unsupervised & Layer-wised pre-training
- Better designs for modeling and training (normalization, nonlinearity, dropout)
- Feature learning
- New development of computer architectures
  - GPU
  - Multi-core computer systems
- Large scale databases



Neural network
Back propagation

Deep belief net
Science

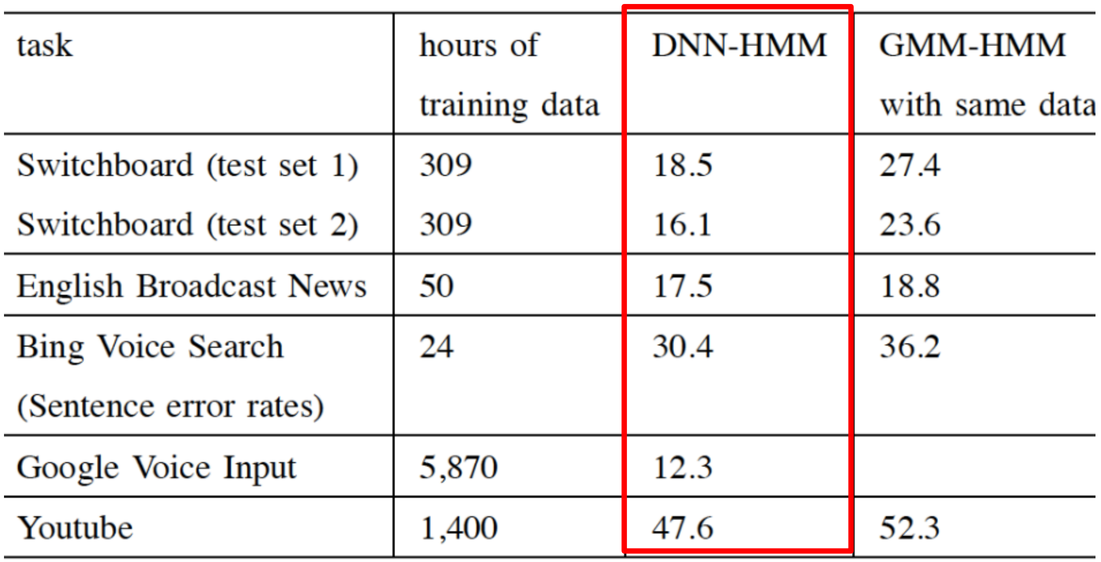
Speech

1986

2006

2011

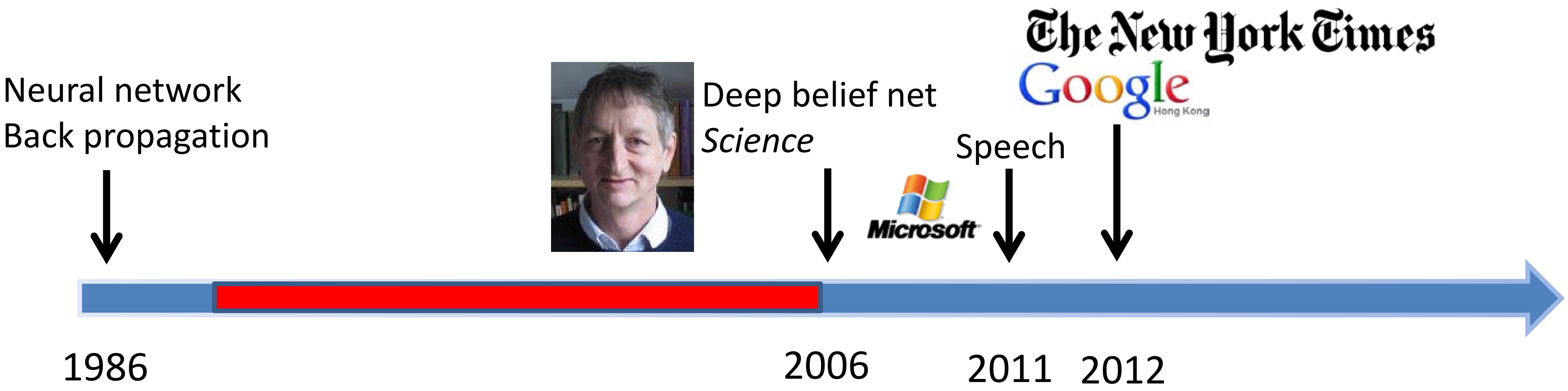
deep learning results



task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

Deep Networks Advance State of Art in Speech

Deep Learning leads to breakthrough in speech recognition at MSR.



Neural network
Back propagation

Deep belief net
Science

Speech

1986

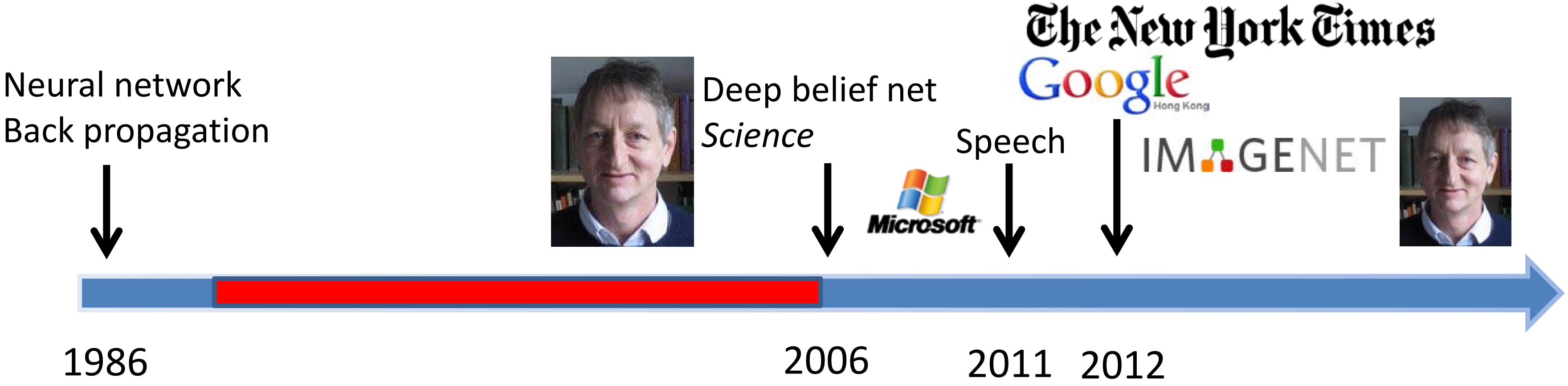
2006

2011

2012

Trained on:
10,000,000 YouTube videos
1 frame from each (200x200)

How Many Computers to Identify a **Cat**? **16000 CPU cores**



Neural network
Back propagation

Deep belief net
Science

Speech

1986

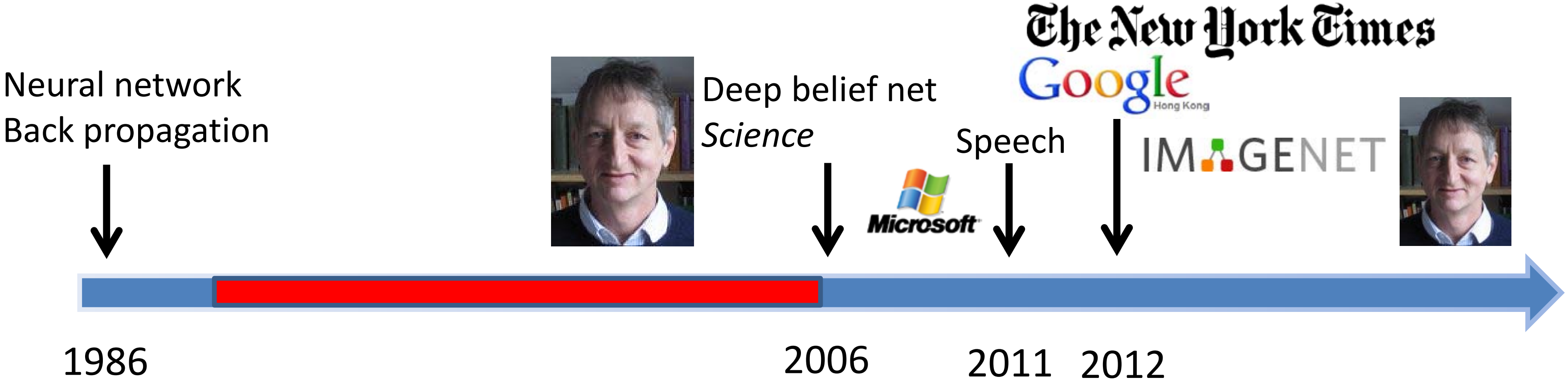
2006

2011

2012

Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

Object recognition over 1,000,000 images and 1,000 categories
(2 GPU)



Neural network
Back propagation

1986

Deep belief net
Science

2006

Speech

2011

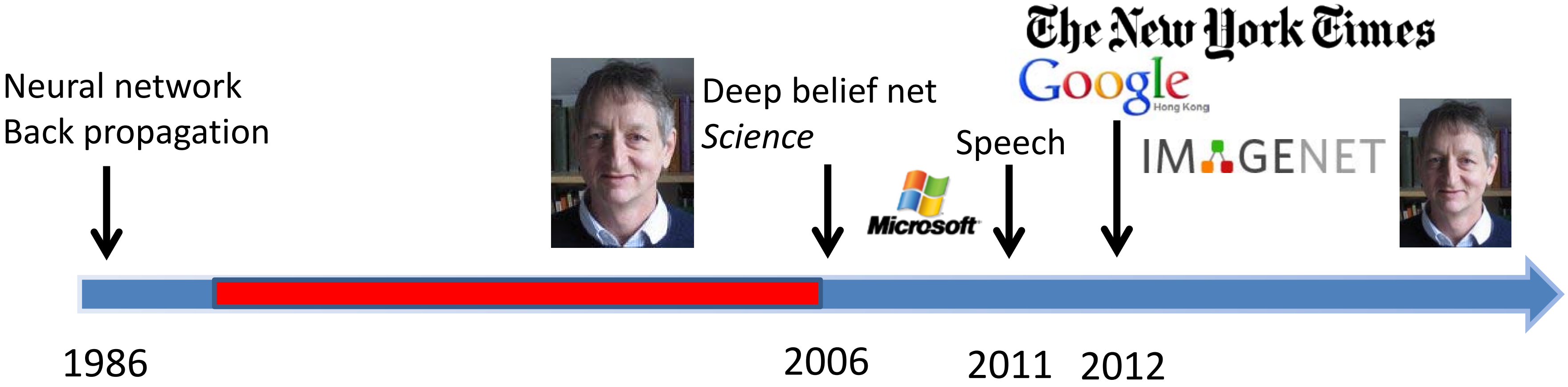
2012

- ImageNet 2013

Rank	Name	Error rate	Description
1	NYU	0.11197	Deep learning
2	NUS	0.12535	Deep learning
3	Oxford	0.13555	Deep learning

MSRA, IBM, Adobe, NEC, Clarifai, Berkley, U. Tokyo, UCLA, UIUC, Toronto ….

Top 20 groups all used deep learning



Neural network
Back propagation

Deep belief net
Science

Speech

1986

2006

2011 2012

- Google and Baidu announced their deep learning based visual search engines (2013)
  - Google
  - Baidu

Works Done by Us

Detection

- Pedestrian detection
- Facial keypoint detection

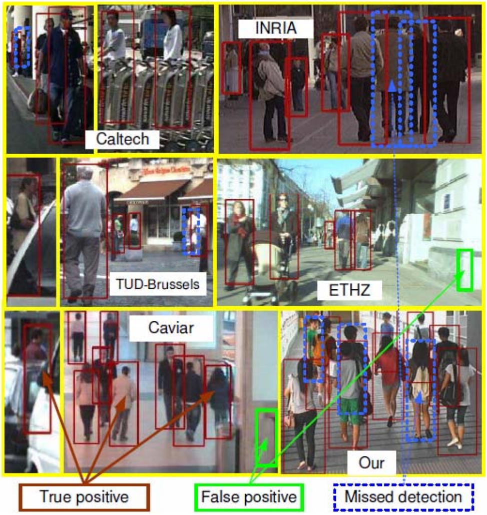
Segmentation

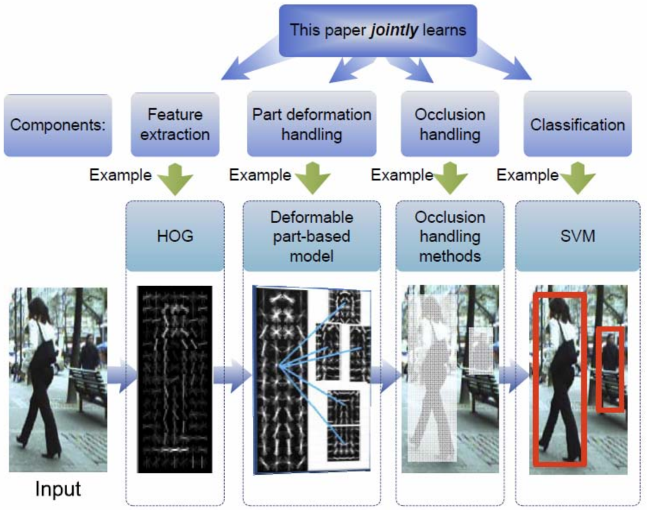
- Face parsing
- Pedestrian parsing

Recognition

- Face verification
- Face attribute recognition

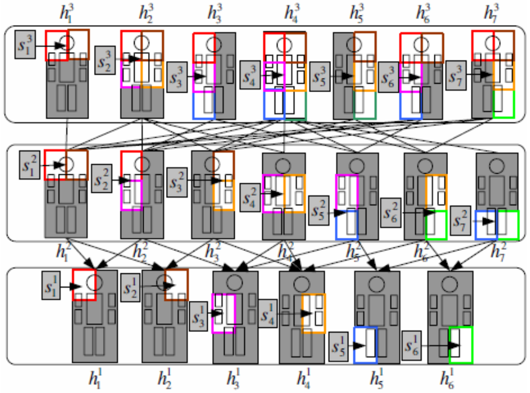
Pedestrian Detection



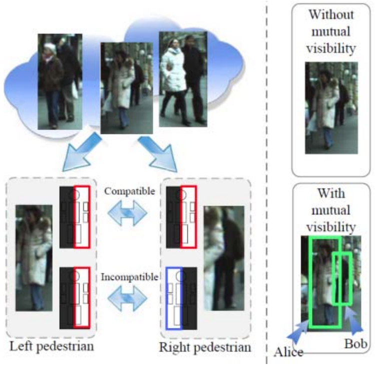

This paper *jointly* learns
Components:
Feature extraction
Part deformation handling
Occlusion handling
Classification
Example
Example
Example
Example
HOG
Deformable part-based model
Occlusion handling methods
SVM
Input

ICCV’13

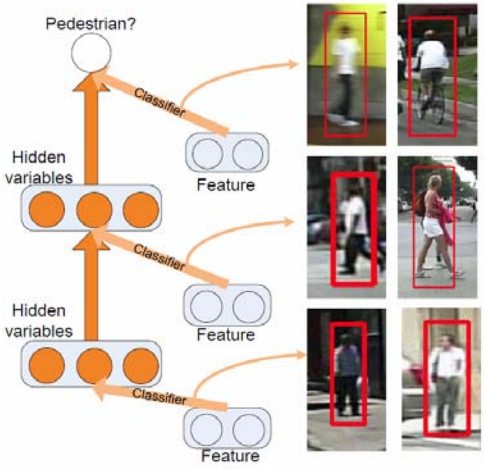
Improve state-of-the-art average miss detection rate on the largest Caltech dataset from 63% to 39%



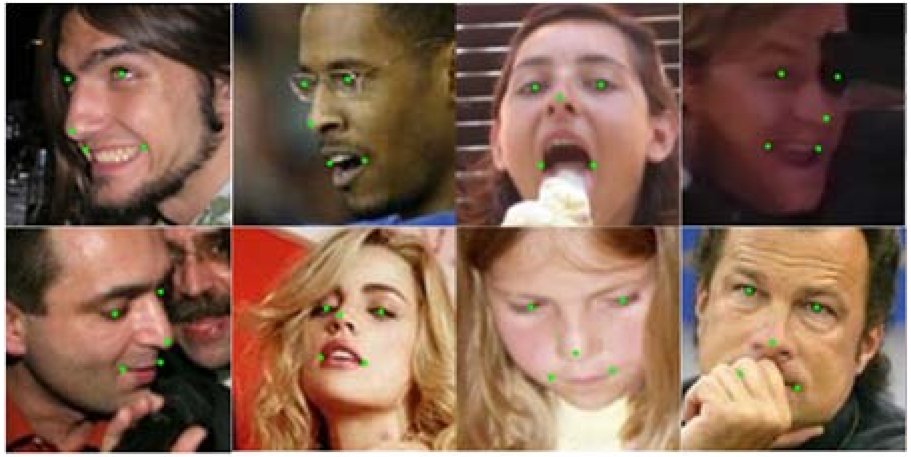
CVPR’12



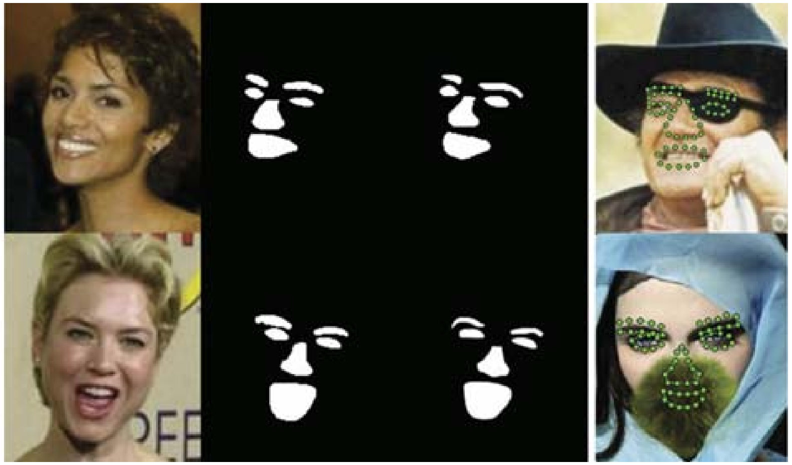
CVPR’13



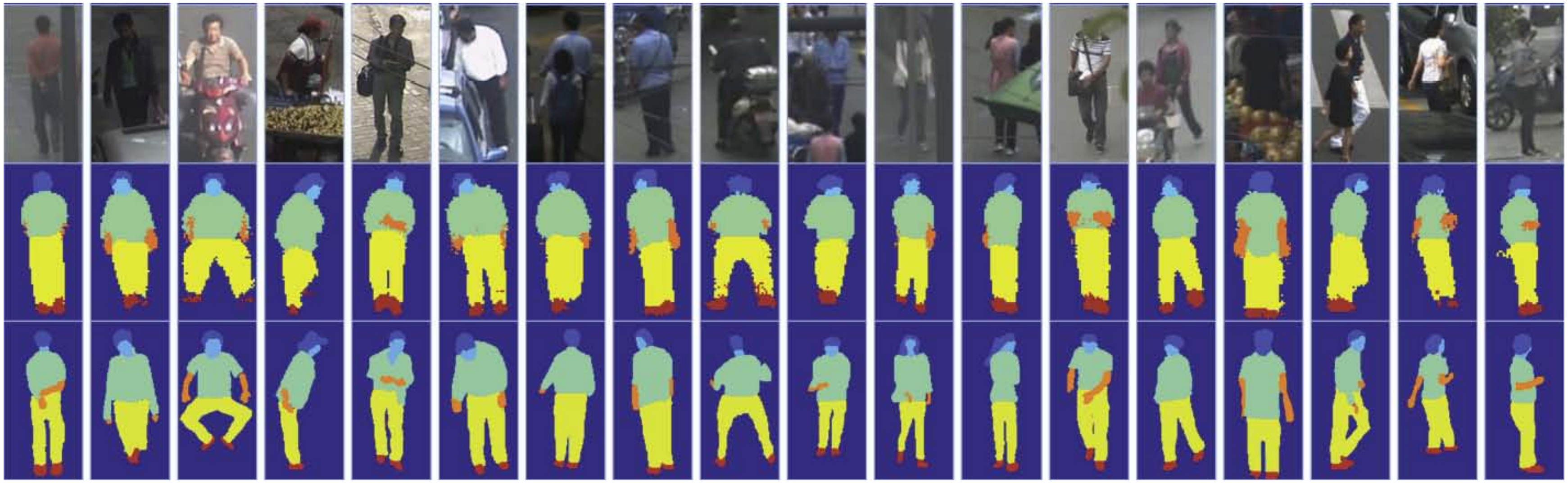
ICCV’13



Facial keypoint detection, CVPR'13
(2% average error on LFPW)



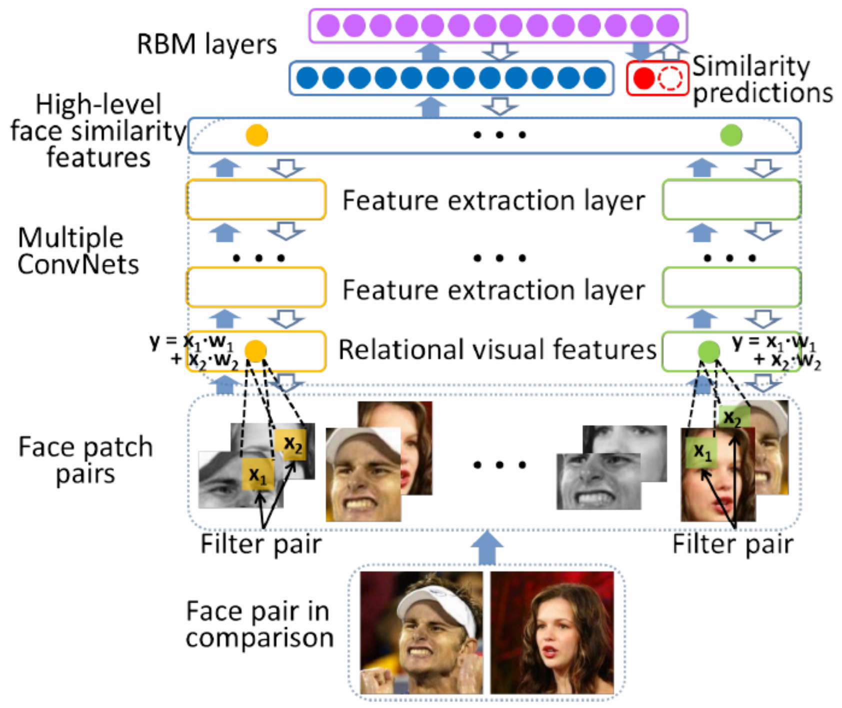
Face parsing, CVPR'12



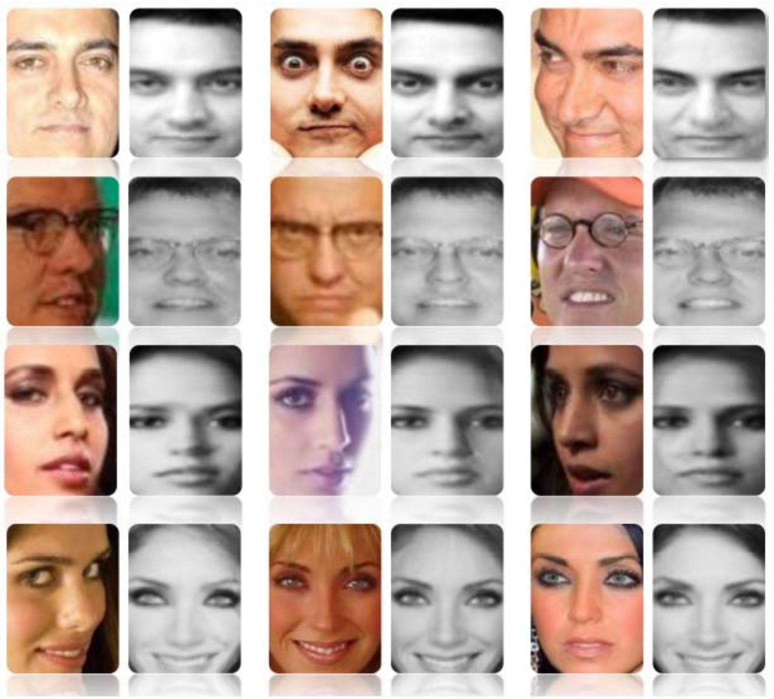
Pedestrian parsing, CVPR'12

Face Recognition and Face Attribute Recognition

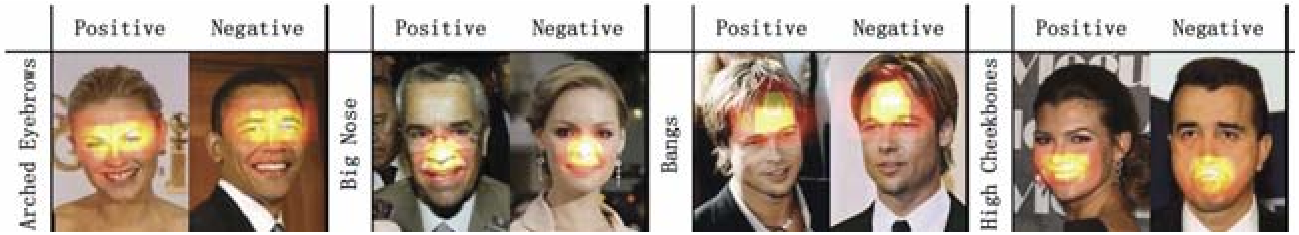
(LFW: 96.45%)



Face verification, ICCV’13



Recovering Canonical-View Face Images, ICCV’13



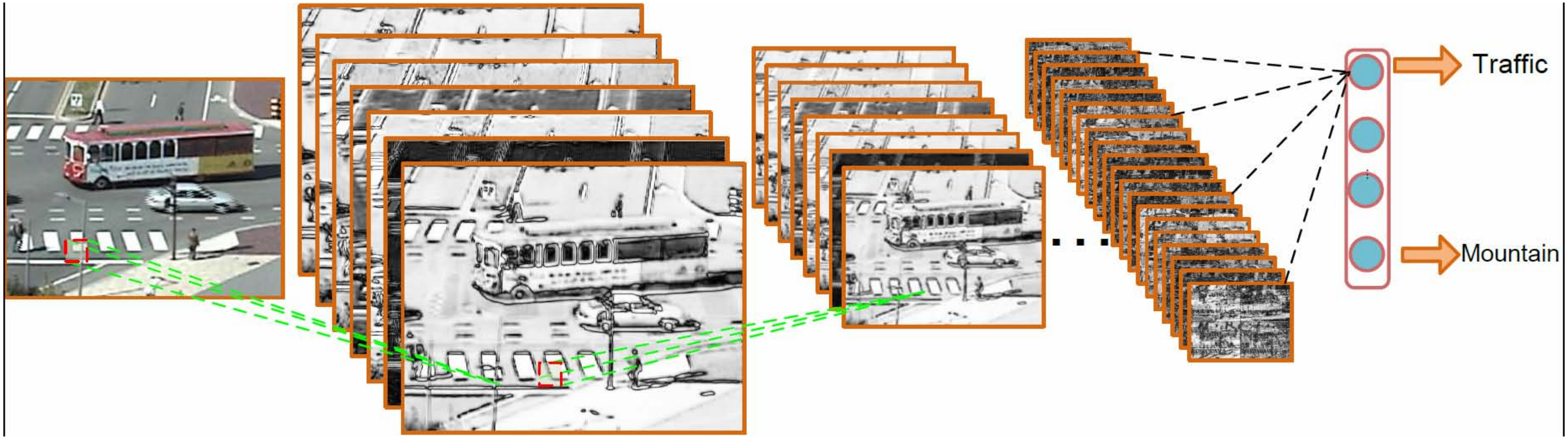
Face attribute recognition, ICCV’13

Introduction on Classical Deep Models

- Convolutional Neural Networks (CNN)
- Deep Belief Net (DBN)
- Auto-encoder

Classical Deep Models

- Convolutional Neural Networks (CNN)
  - LeCun'95



Convolution Pooling

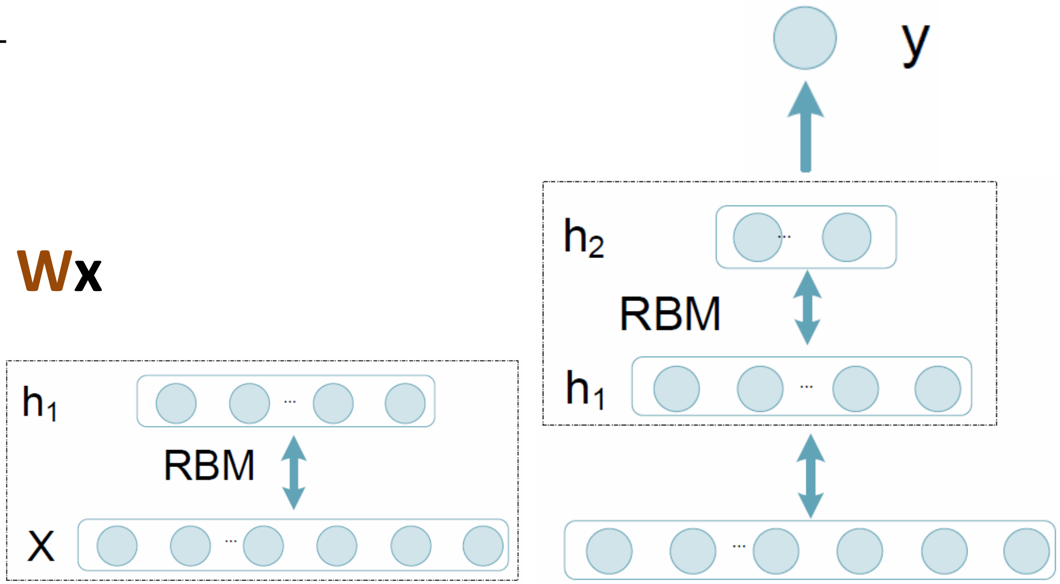
Classical Deep Models

- Deep belief net
  - Hinton'06

$$P(\mathbf{x},\mathbf{h}_1,\mathbf{h}_2) = p(\mathbf{x}|\mathbf{h}_1)\ p(\mathbf{h}_1,\mathbf{h}_2)$$

$$P(\mathbf{x},\mathbf{h}_1) = \frac{e^{-E(\mathbf{x},\mathbf{h}_1)}}{\sum_{\mathbf{x},\mathbf{h}_1} e^{-E(\mathbf{x},\mathbf{h}_1)}}$$

$$E(\mathbf{x},\mathbf{h}_1)=\mathbf{b}'\,\mathbf{x}+\mathbf{c}'\,\mathbf{h}_1+\mathbf{h}_1'\,\mathbf{W}\mathbf{x}$$



Classical Deep Models

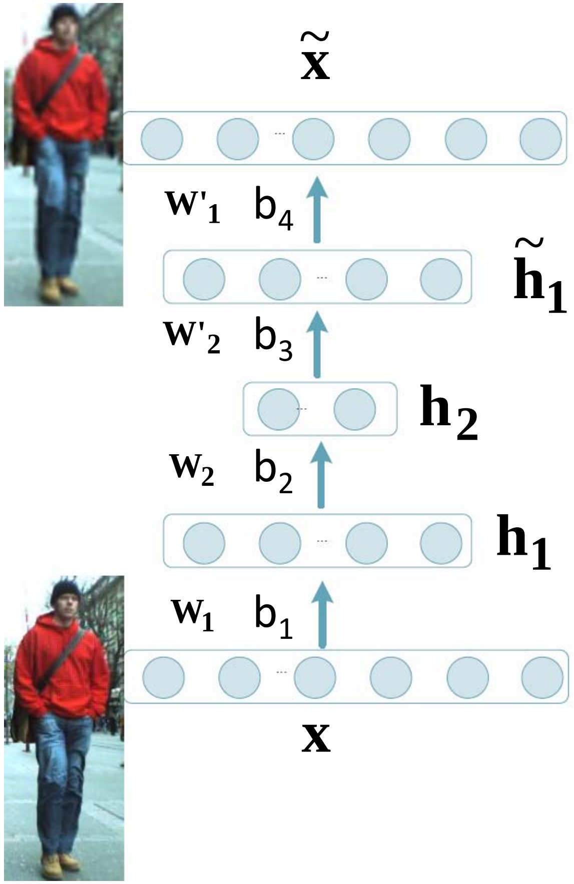
- Auto-encoder
  - Hinton'06

Encoding: $\mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{x}+b_1)$

$\mathbf{h}_2 = \sigma(\mathbf{W}_2\mathbf{h}_1+b_2)$

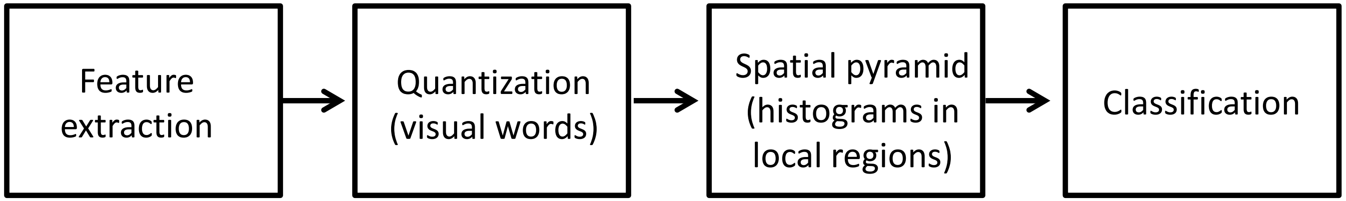
Decoding: $\tilde{\mathbf{h}}_1 = \sigma(\mathbf{W}'_2\mathbf{h}_2+b_3)$

$\tilde{\mathbf{x}} = \sigma(\mathbf{W}'_1\mathbf{h}_1+b_4)$



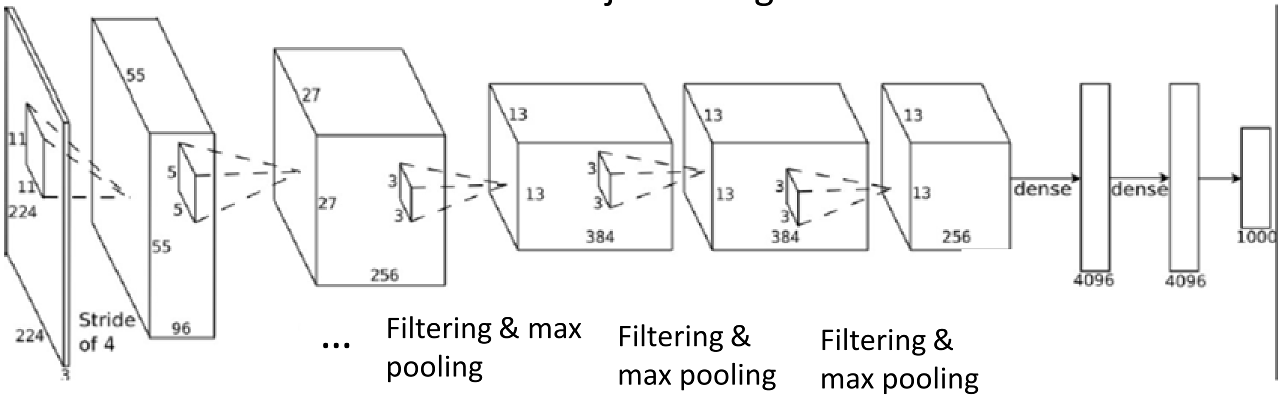
Opinion I

- How to formulate a vision problem with deep learning?
  - Make use of experience and insights obtained in CV research
  - Sequential design/learning vs **joint learning**
  - Effectively train a deep model (layerwise pre-training + fine tuning)



Feature extraction → Quantization (visual words) → Spatial pyramid (histograms in local regions) → Classification

Conventional object recognition scheme



… Filtering & max pooling

Filtering & max pooling

Filtering & max pooling

Feature extraction ↔ filtering

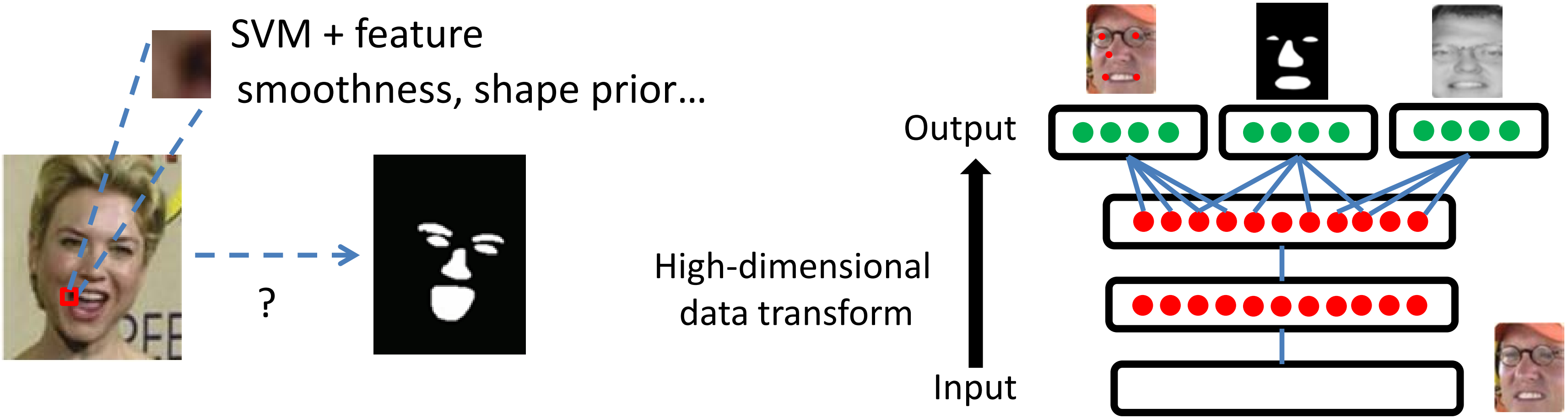
Quantization ↔ filtering

Spatial pyramid ↔ multi-level pooling

Krizhevsky NIPS'12

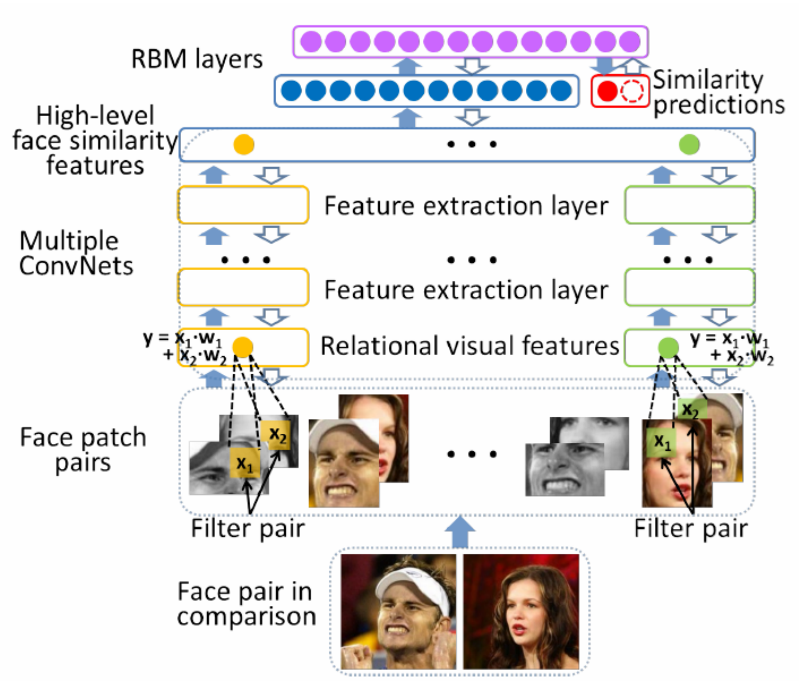
Opinion II

- How to make use of the large learning capacity of deep models?
  - **High dimensional data transform**
  - Hierarchical nonlinear representations



Opinion III

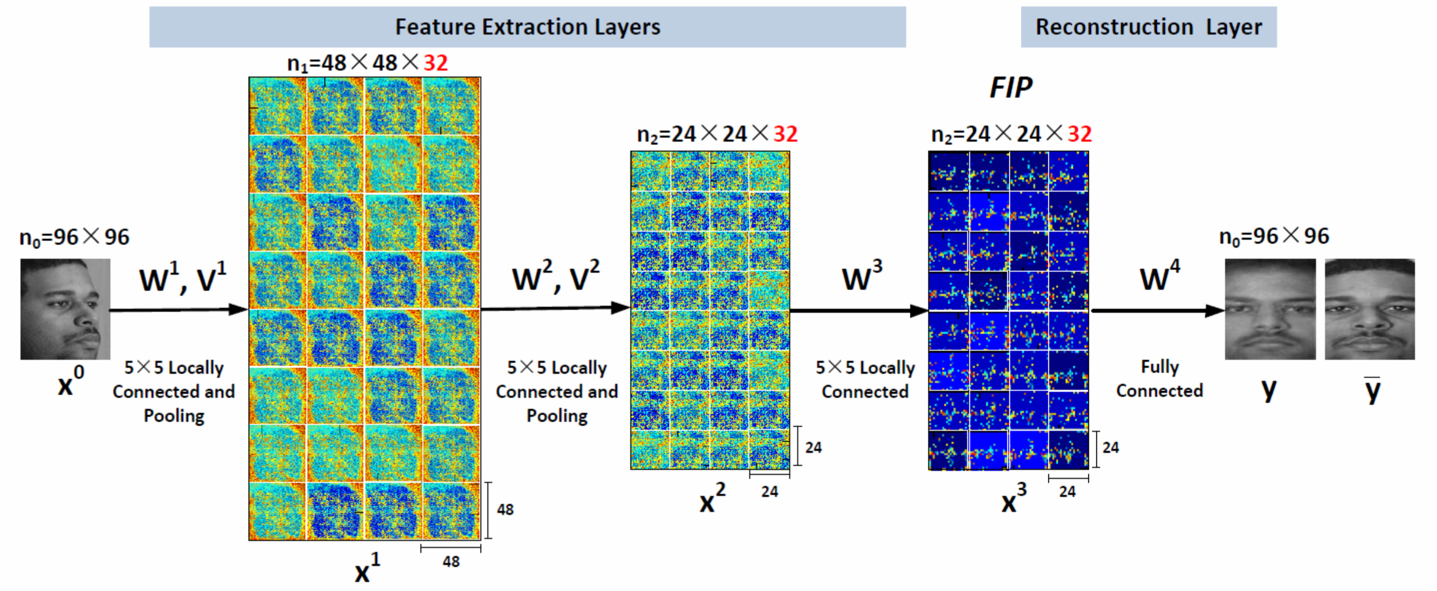
- Deep learning likes challenging tasks (for better generalization)
  - Make input data more challenging (augmenting data by translating, rotating, and scaling)
  - Make training process more challenging (dropout: randomly setting some responses to zero; dropconnect: randomly setting some weights to zero)
  - **Make prediction more challenging**



Y. Sun, X. Wang, and X. Tang, “Hybrid Deep Learning for Computing Face Similarities,” ICCV’13

Learning feature through face verification (predicting 0/1 label): 92.57% on LFW with 480 CNNs

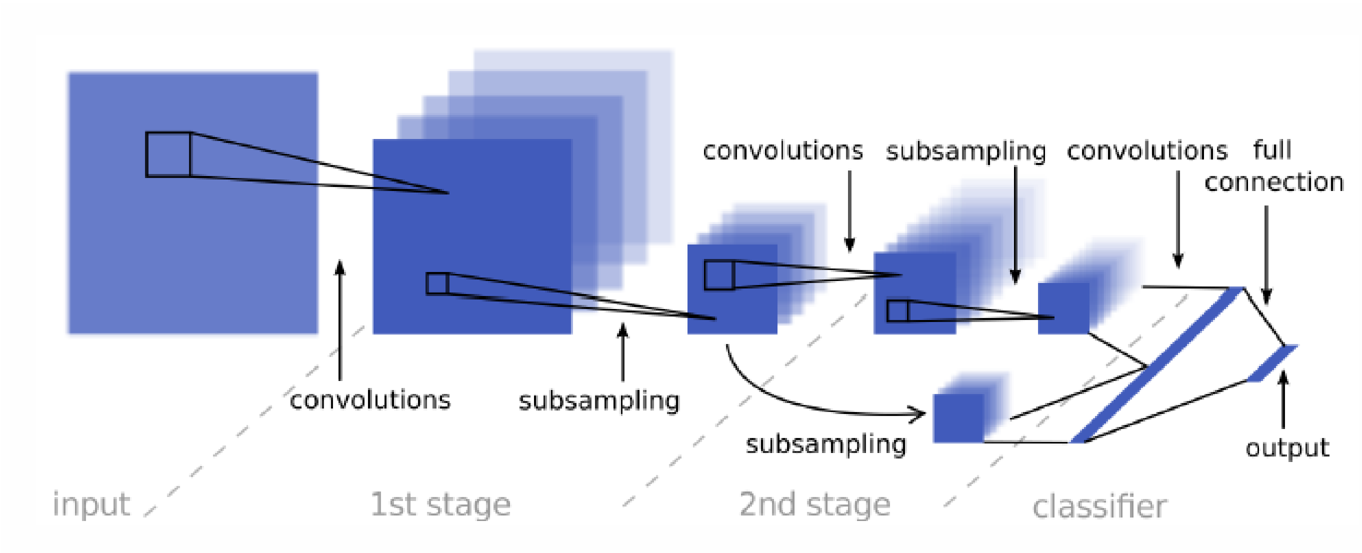
Learning feature through face reconstruction (predicting 9216 pixels): 96.45% on LFW with 4 CNNs



Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Indentify-Preserving Face Space,” ICCV 2013.

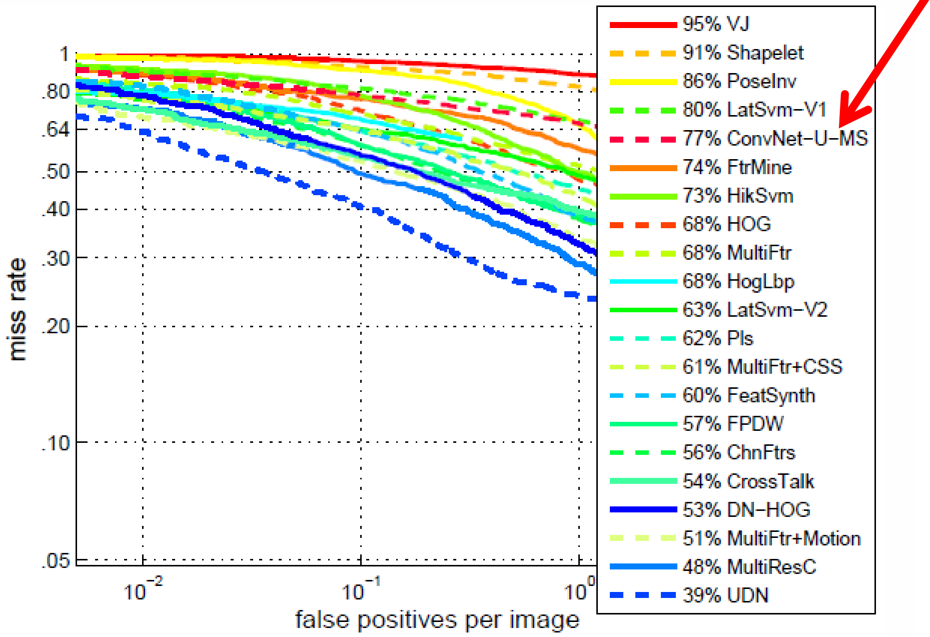
Joint Deep Learning

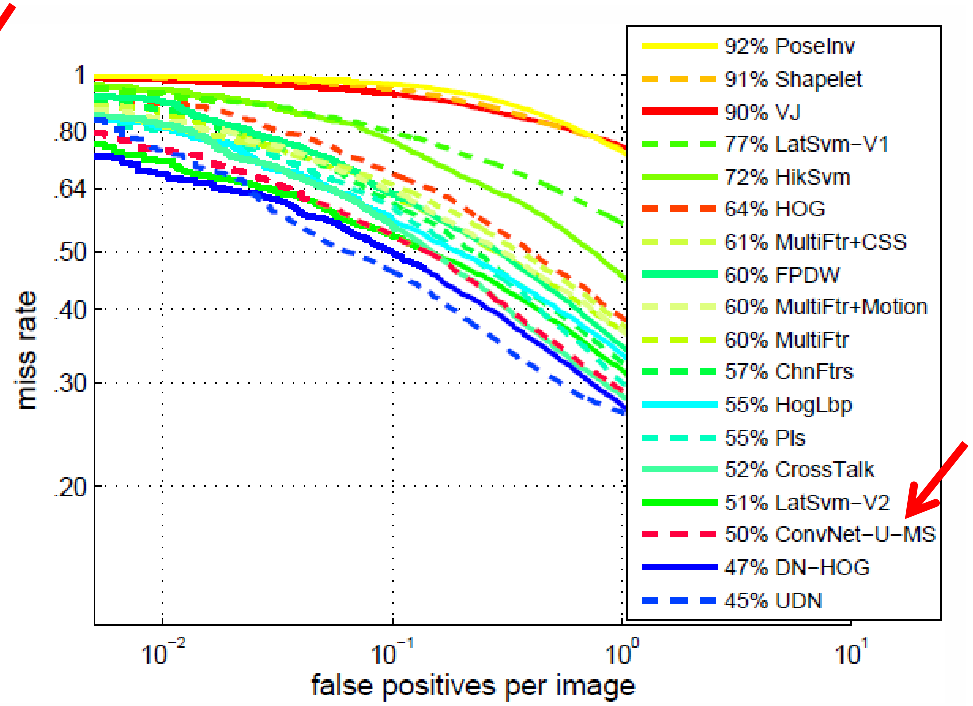
What if we treat an existing deep model as a black box in pedestrian detection?



ConvNet−U−MS

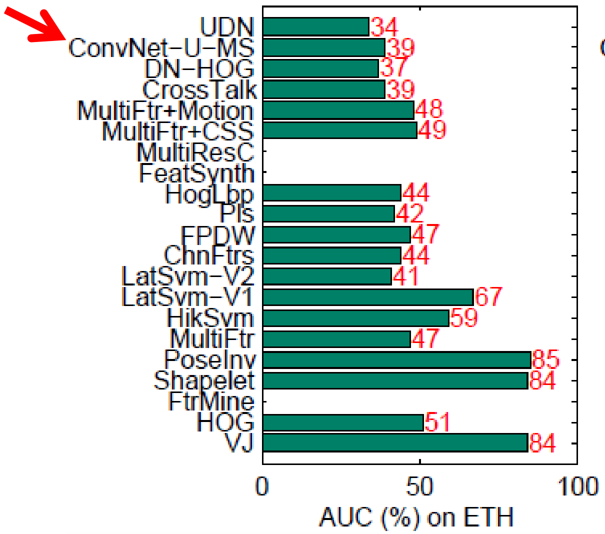
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” CVPR 2013.

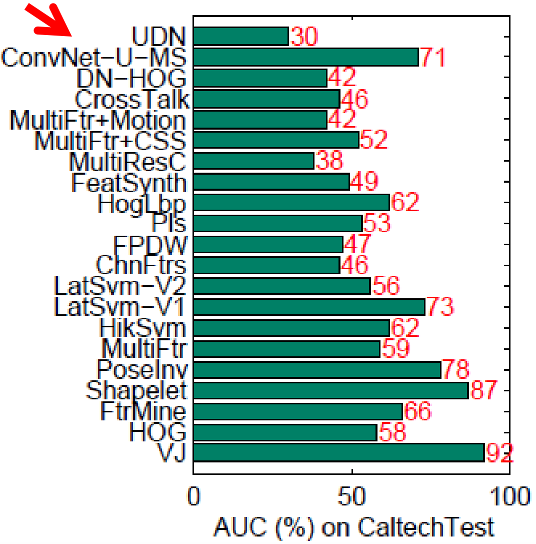


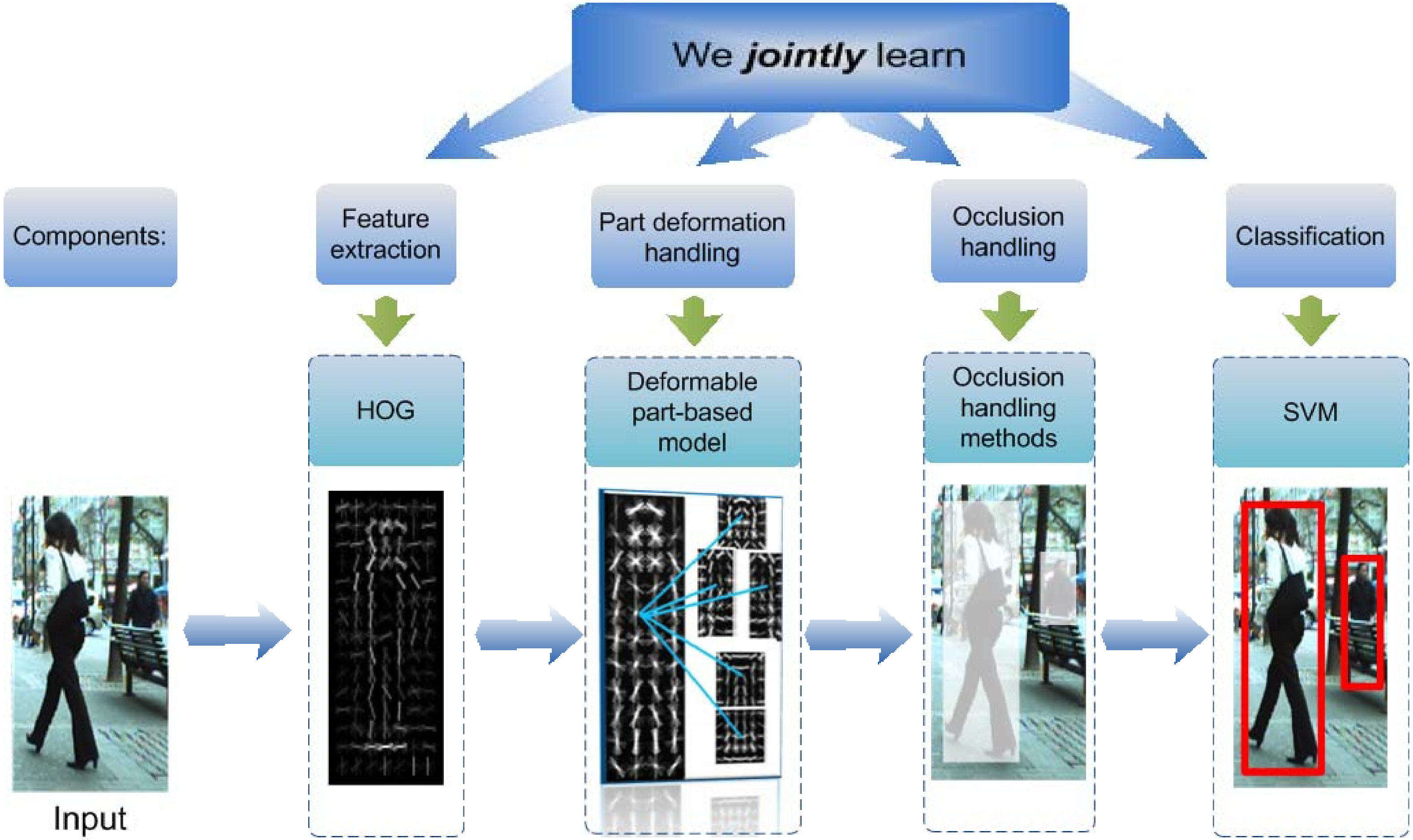


Results on Caltech Test

Results on ETHZ





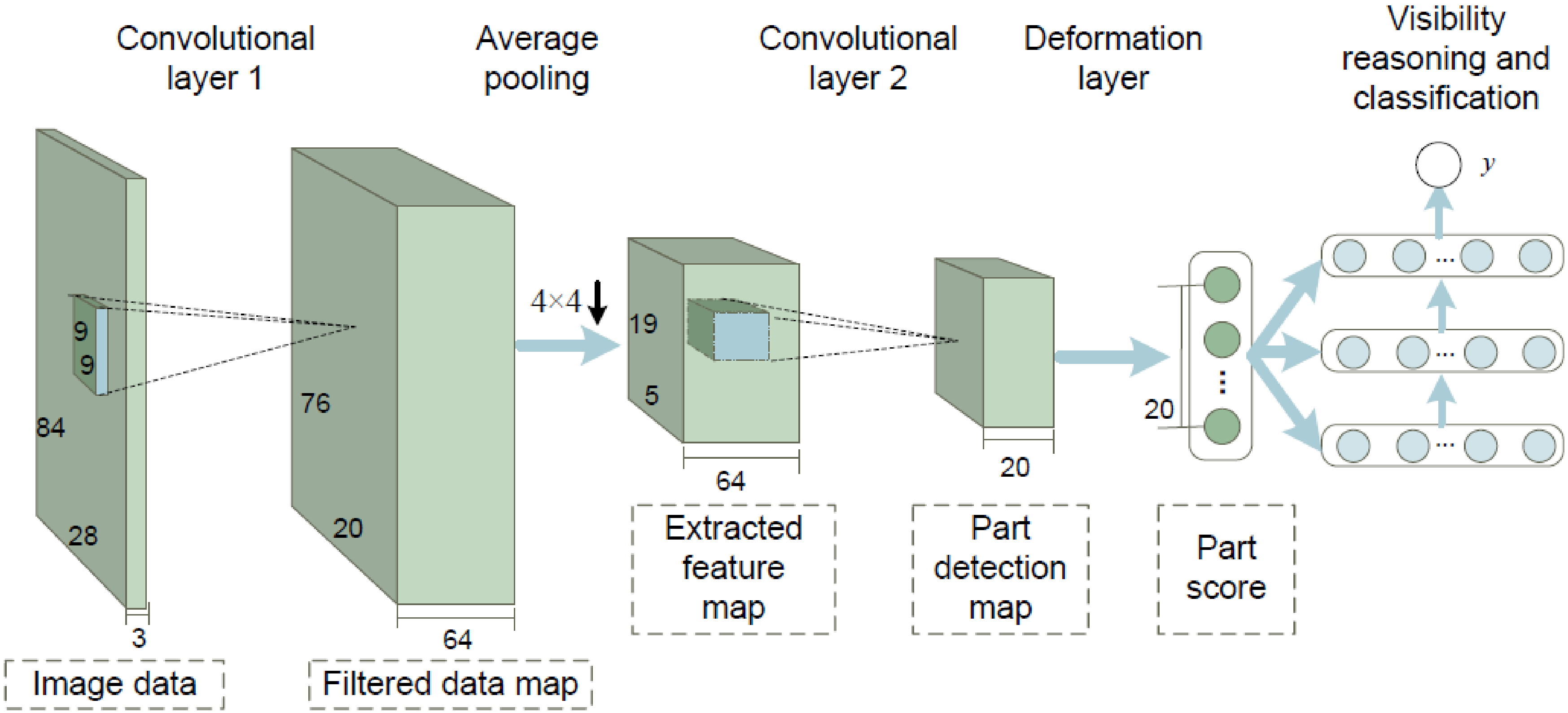


- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)

- P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)

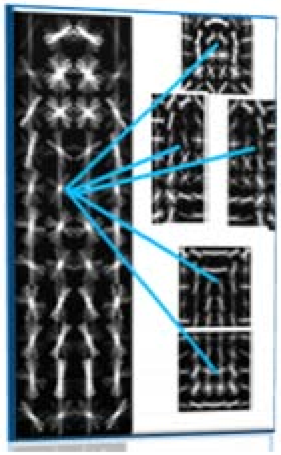
- W. Ouyang and X. Wang. A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling. CVPR, 2012.

Our Joint Deep Learning Model

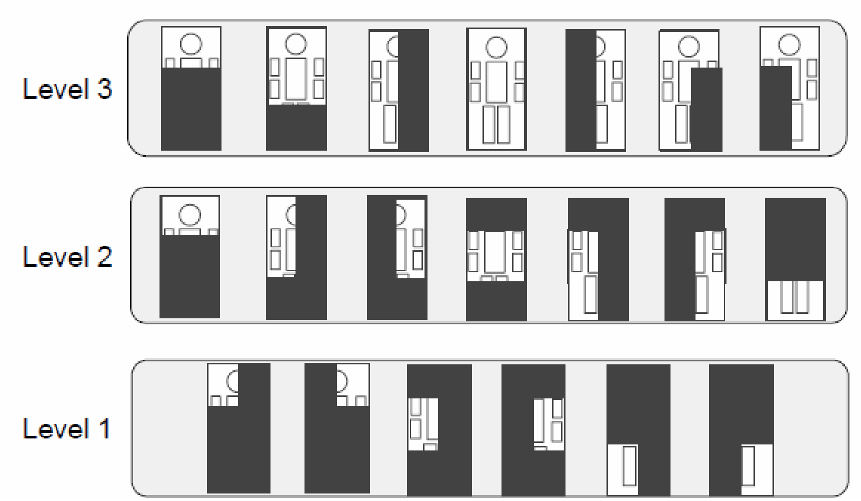


Modeling Part Detectors

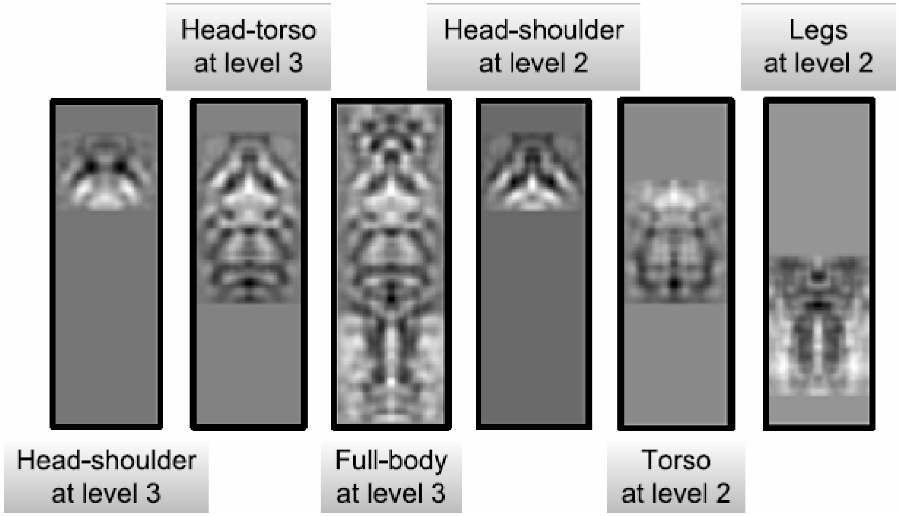
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

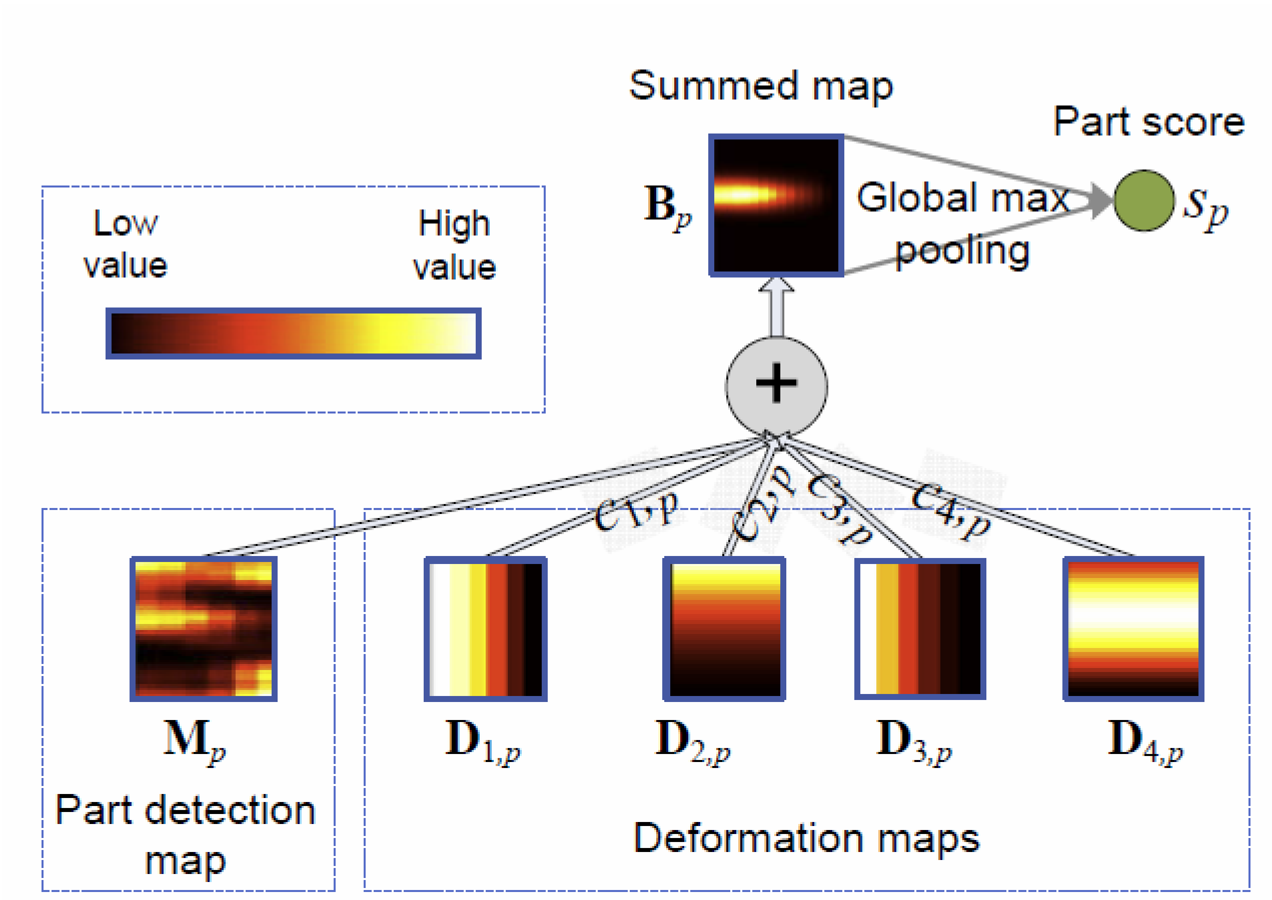


Part models

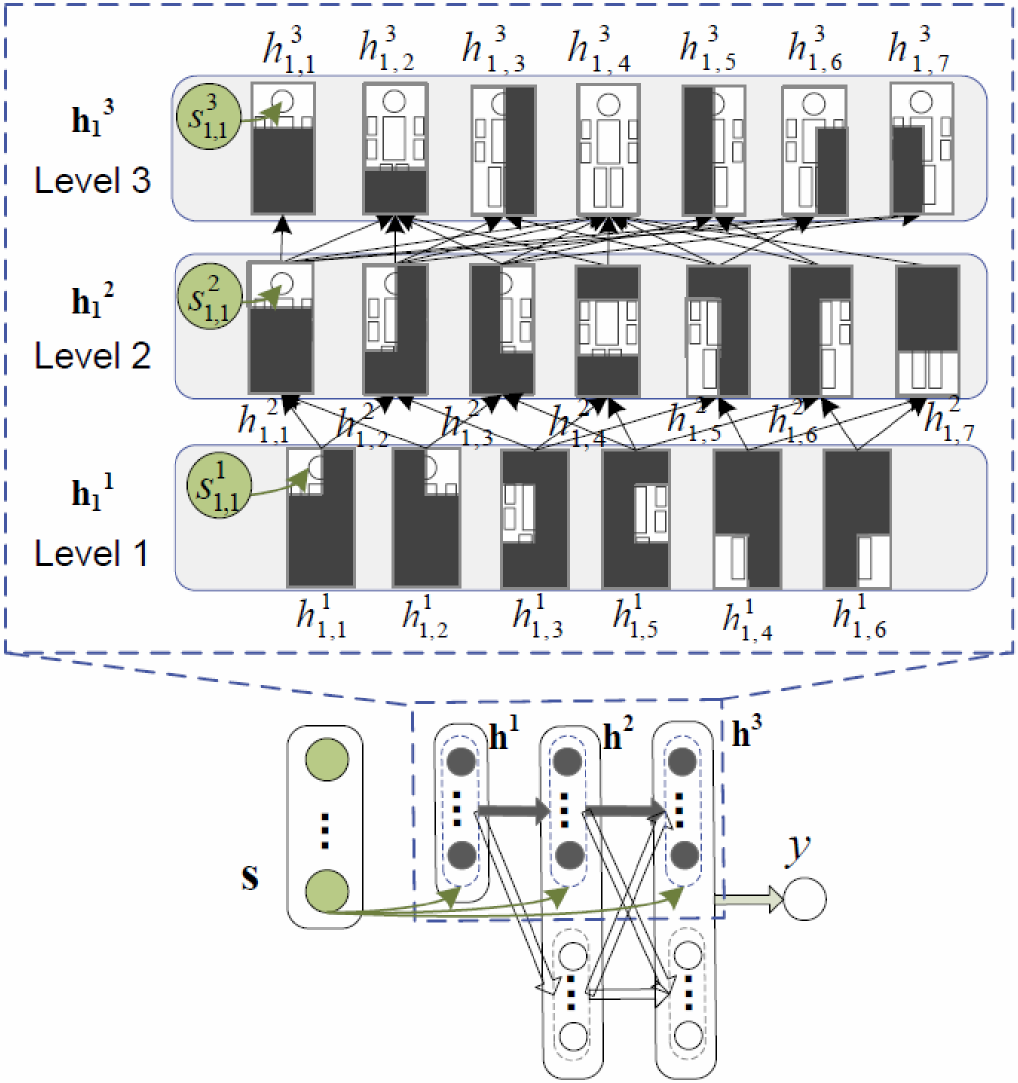


Learned filtered at the second convolutional layer

Deformation Layer

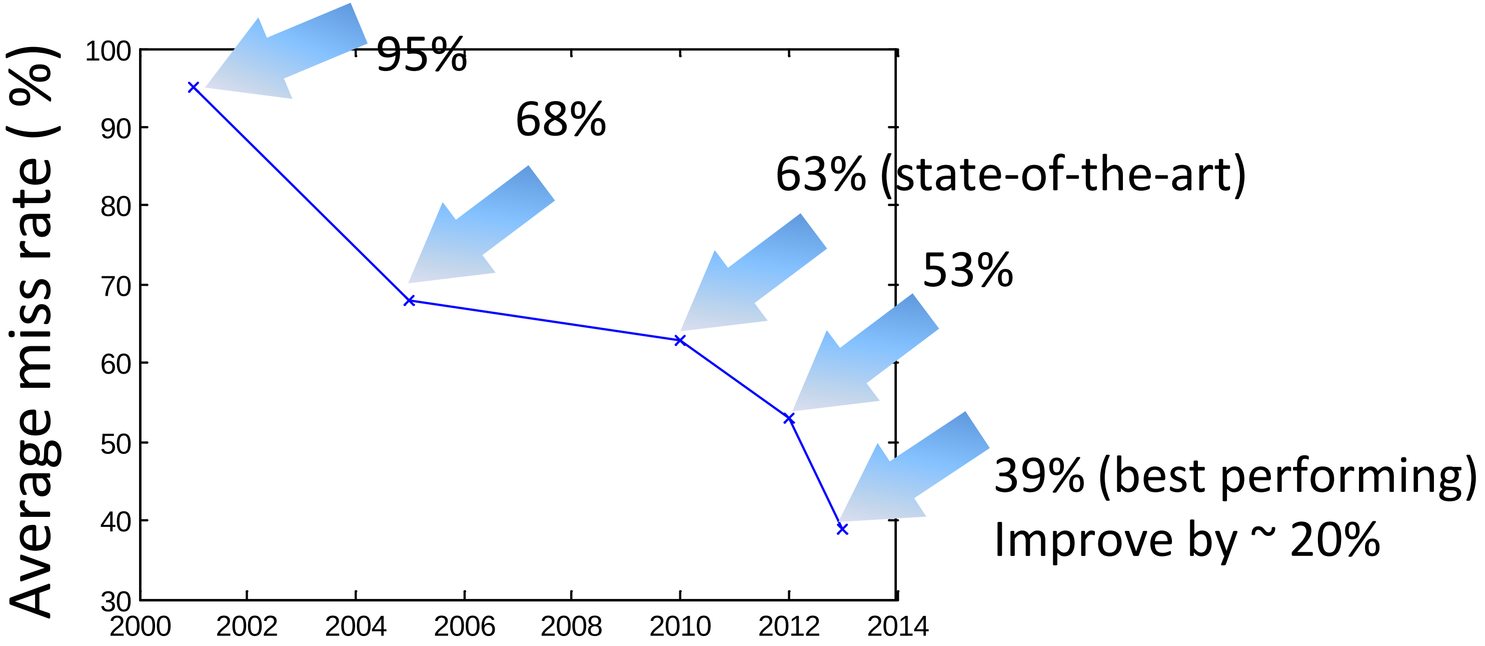


Visibility Reasoning with Deep Belief Net



Experimental Results

- Caltech – Test dataset (largest, most widely used)



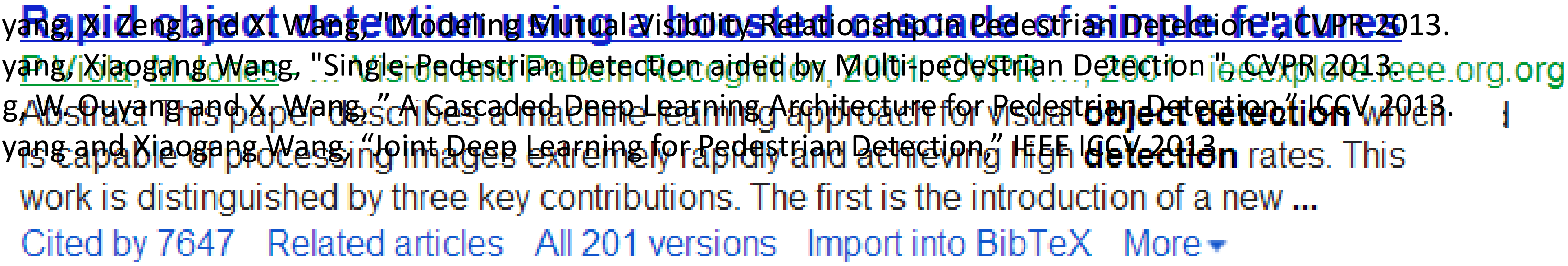
W. Ouyang and X. Wang, "A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling,“ CVPR 2012.

W. Ouyang, X. Zeng and X. Wang, "Modeling Mutual Visibility Relationship in Pedestrian Detection ", CVPR 2013.

W. Ouyang, Xiaogang Wang, "Single-Pedestrian Detection aided by Multi-pedestrian Detection ", CVPR 2013.

X. Zeng, W. Ouyang and X. Wang, ” A Cascaded Deep Learning Architecture for Pedestrian Detection,” ICCV 2013.

W. Ouyang and Xiaogang Wang, “Joint Deep Learning for Pedestrian Detection,” IEEE ICCV 2013.

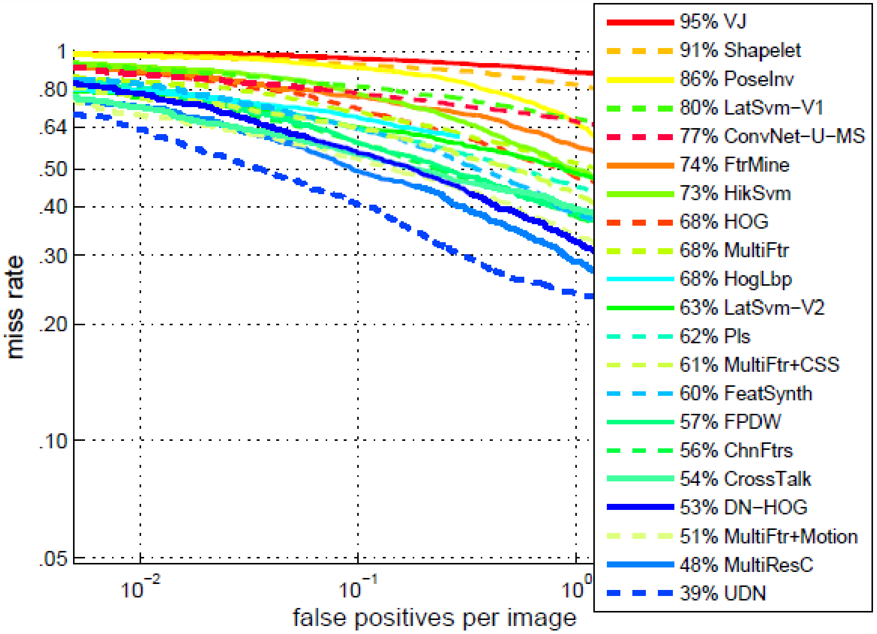


Rapid object detection using a boosted cascade of simple features

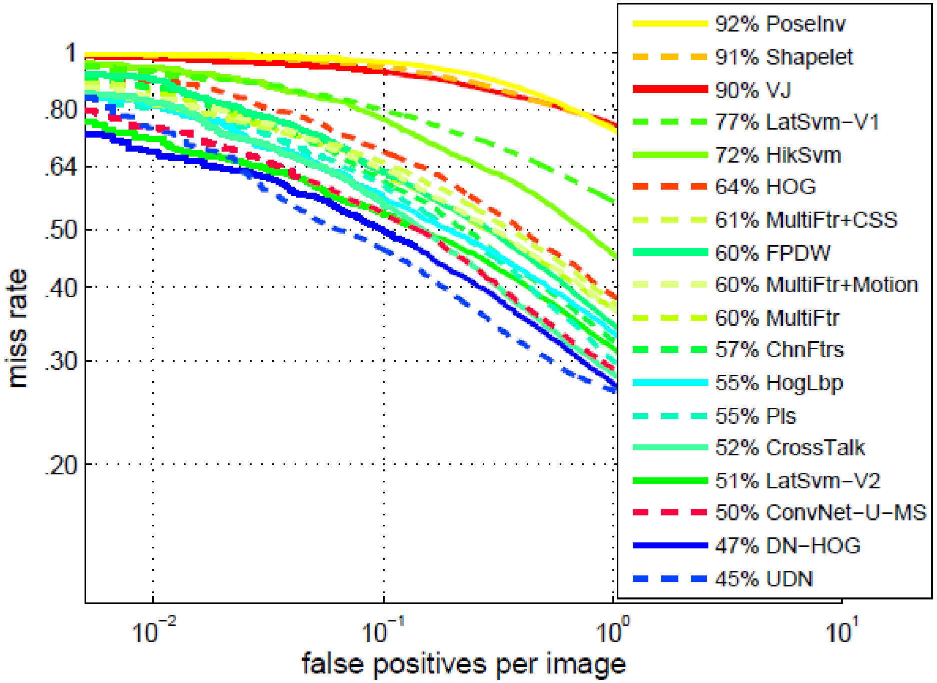
P Viola, M Jones - ... Vision and Pattern Recognition, 2001. CVPR ..., 2001 - ieeexplore.ieee.org

Abstract This paper describes a machine learning approach for visual **object detection** which is capable of processing images extremely rapidly and achieving high **detection** rates. This work is distinguished by three key contributions. The first is the introduction of a new ...

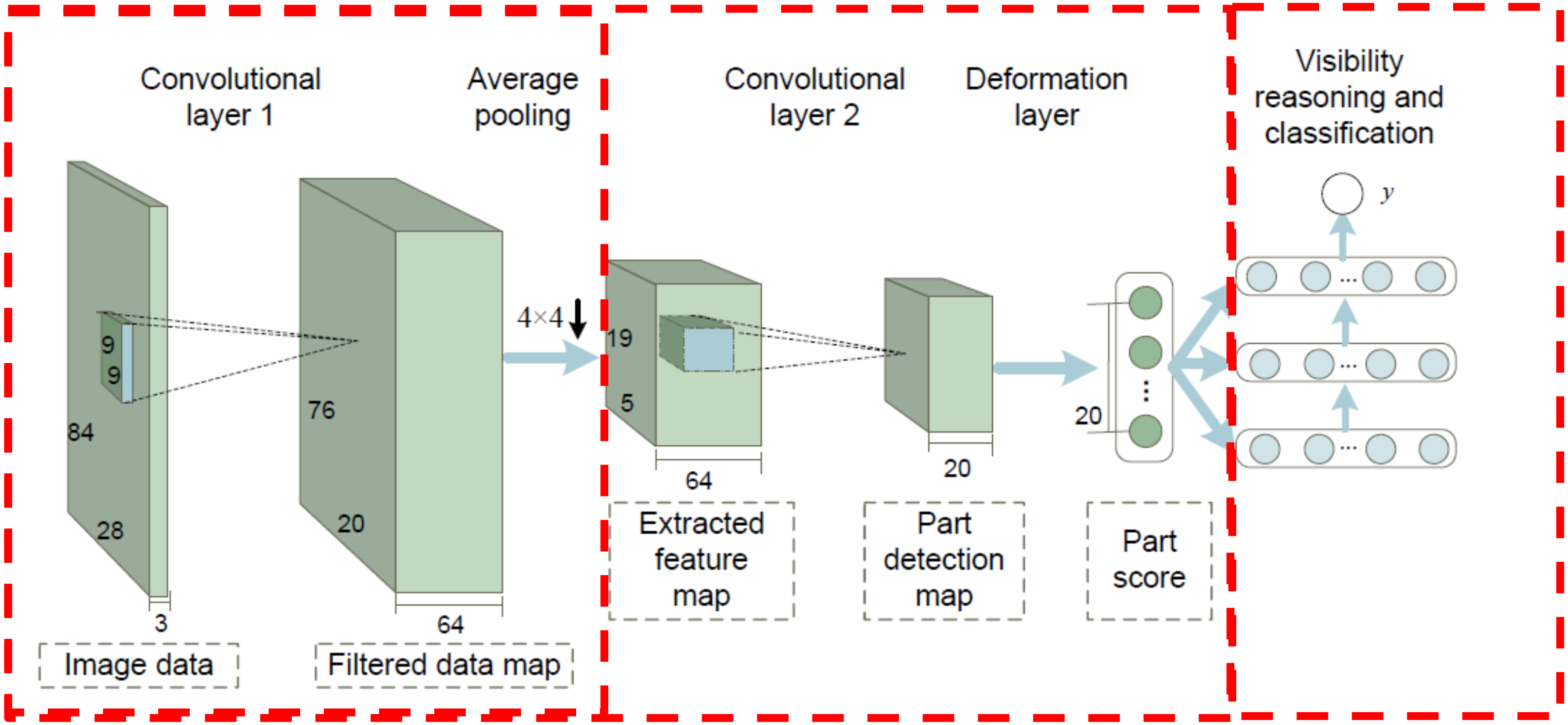
Cited by 7647 Related articles All 201 versions Import into BibTeX More

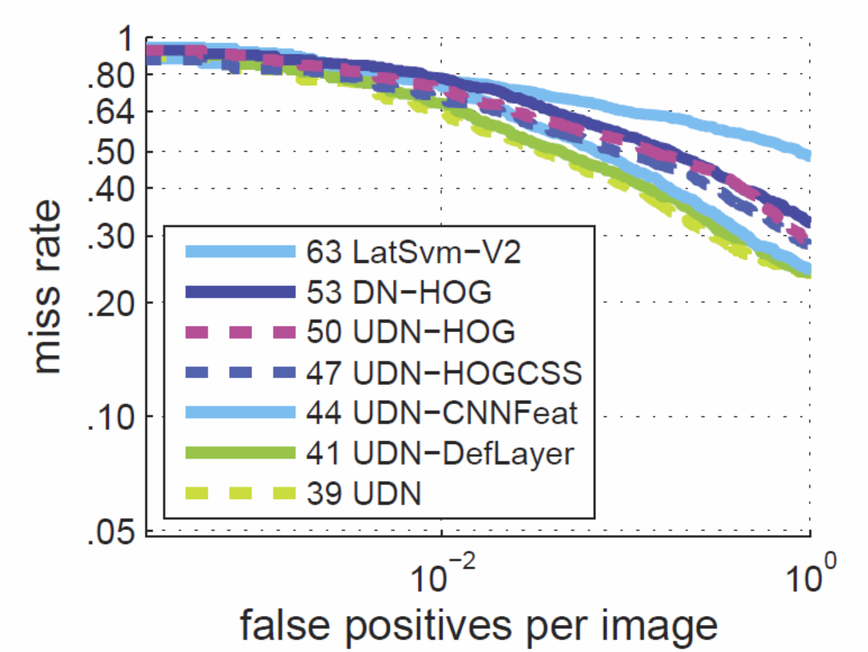


Results on Caltech Test



Results on ETHZ



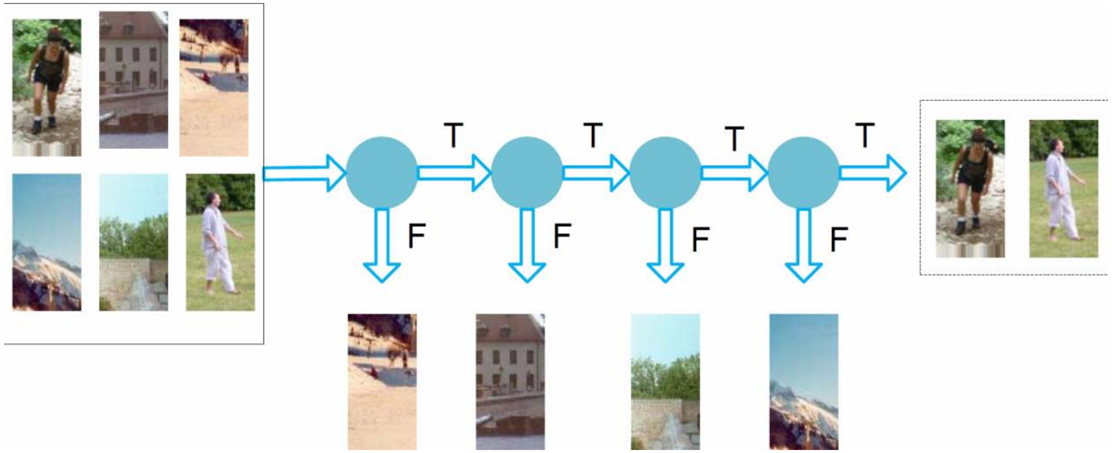


DN-HOG
UDN-HOG
UDN-HOGCSS
UDN-CNNFeat
UDN-DefLayer

Multi-Stage Contextual Deep Learning

Motivated by Cascaded Classifiers and Contextual Boost

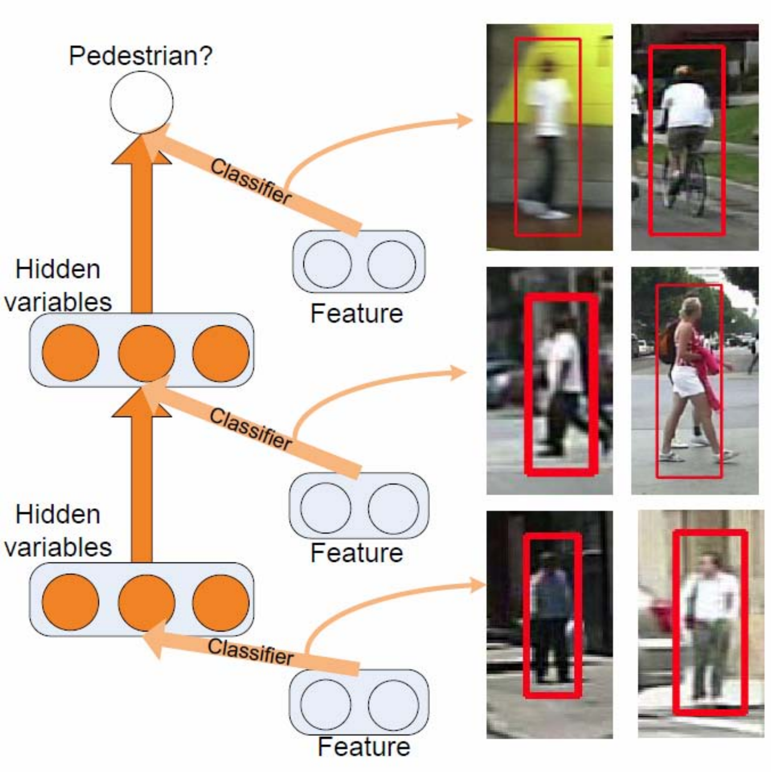
- The classifier of each stage deals with a specific set of samples
- The score map output by one classifier can serve as contextual information for the next classifier

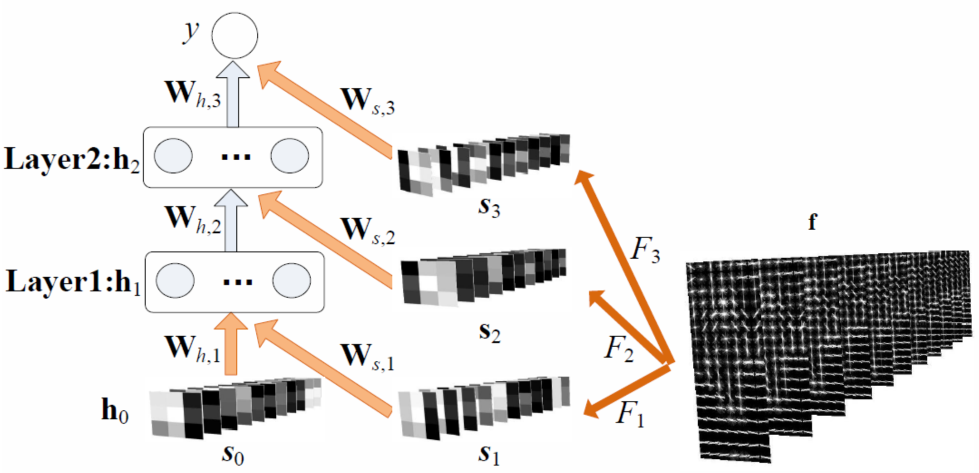


❖ Only pass one detection score to the next stage
❖ Classifiers are trained sequentially

Conventional cascaded classifiers for detection

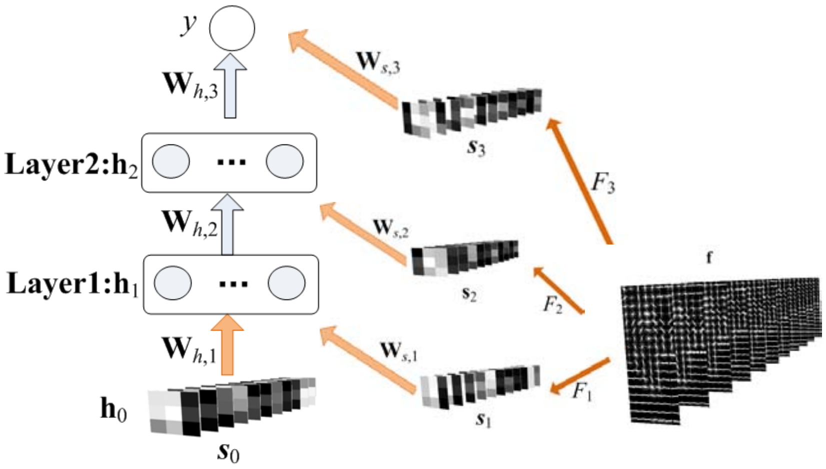
- Our deep model keeps the score map output by the current classifier and it serves as contextual information to support the decision at the next stage
- Cascaded classifiers are jointly optimized instead of being trained sequentially
- To avoid overfitting, a stage-wise pre-training scheme is proposed to regularize optimization
- Simulate the cascaded classifiers by mining hard samples to train the network stage-by-stage





Training Strategies

- Unsupervised pre-train $\mathbf{W}_{h,i+1}$ layer-by-layer, setting $\mathbf{W}_{s,i+1} = 0$, $\mathbf{F}_{i+1} = 0$
- Fine-tune all the $\mathbf{W}_{h,i+1}$ with supervised BP
- Train $\mathbf{F}_{i+1}$ and $\mathbf{W}_{s,i+1}$ with BP stage-by-stage
- A correctly classified sampled at the previous stage does not influence the update of parameters
- Stage-by-stage training can be considered as adding regularization constraints to parameters, i.e. some parameters are constrained to be zeros in the early training stages



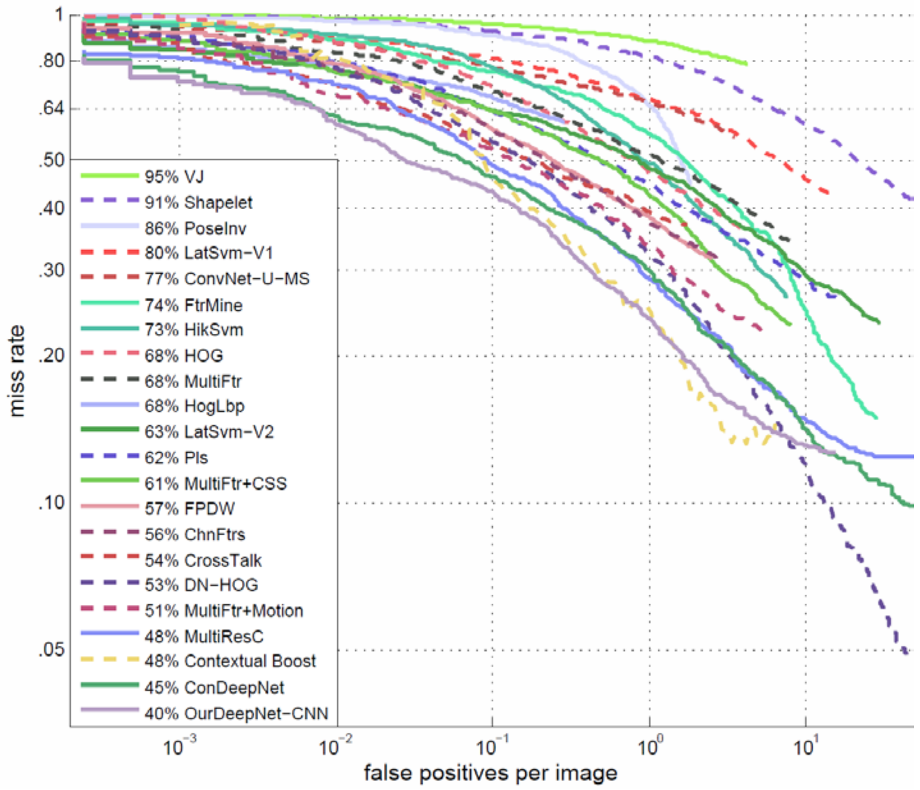
Log error function:

$$E = -l \log y - (1 - l) \log (1 - y)$$

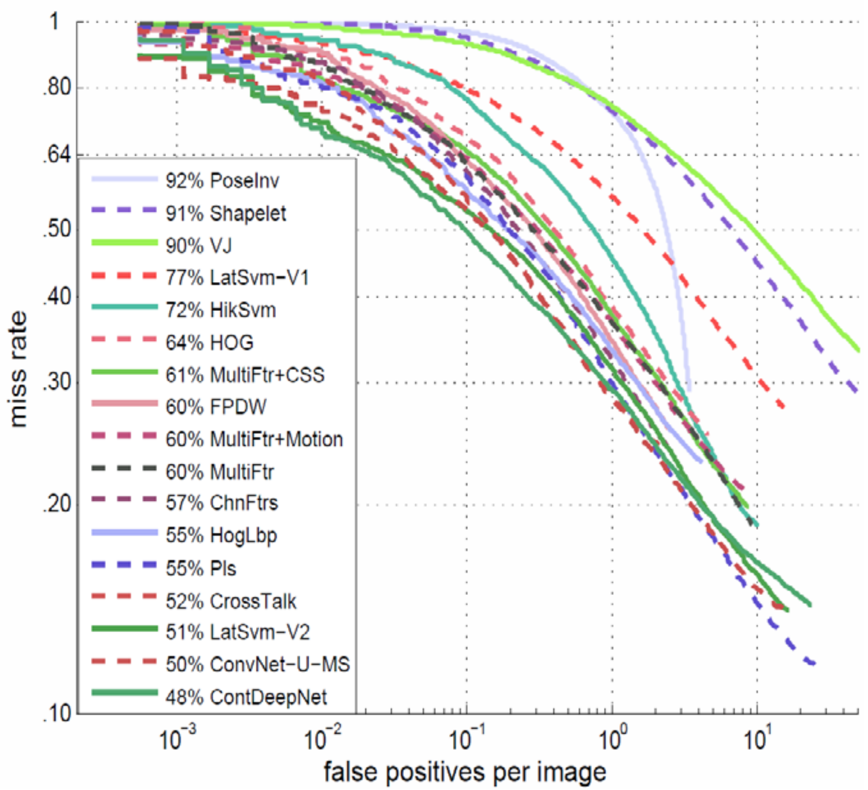
Gradients for updating parameters:

$$d\theta_{i,j} = -\frac{\partial E}{\partial \theta_{i,j}} = -\frac{\partial E}{\partial y}\frac{\partial y}{\partial \theta_{i,j}} = -(y - l)\frac{\partial y}{\partial \theta_{i,j}}$$

Experimental Results

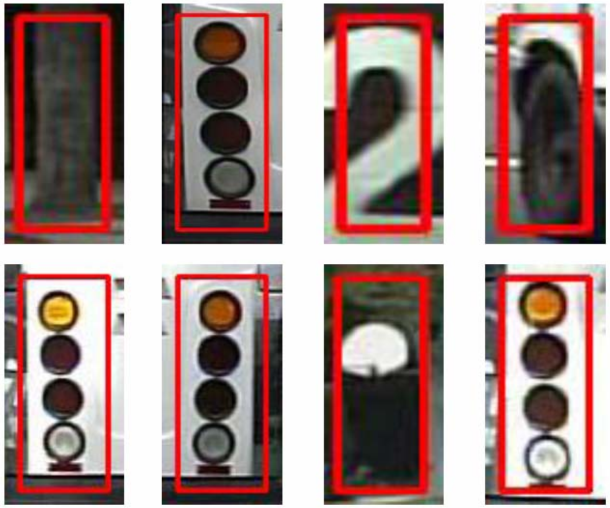


Caltech

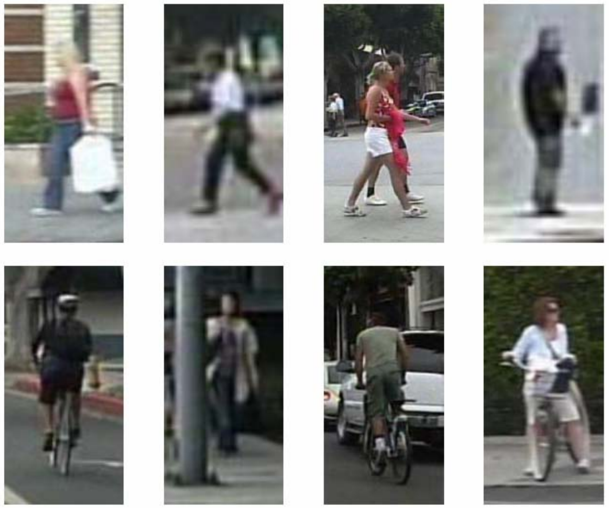


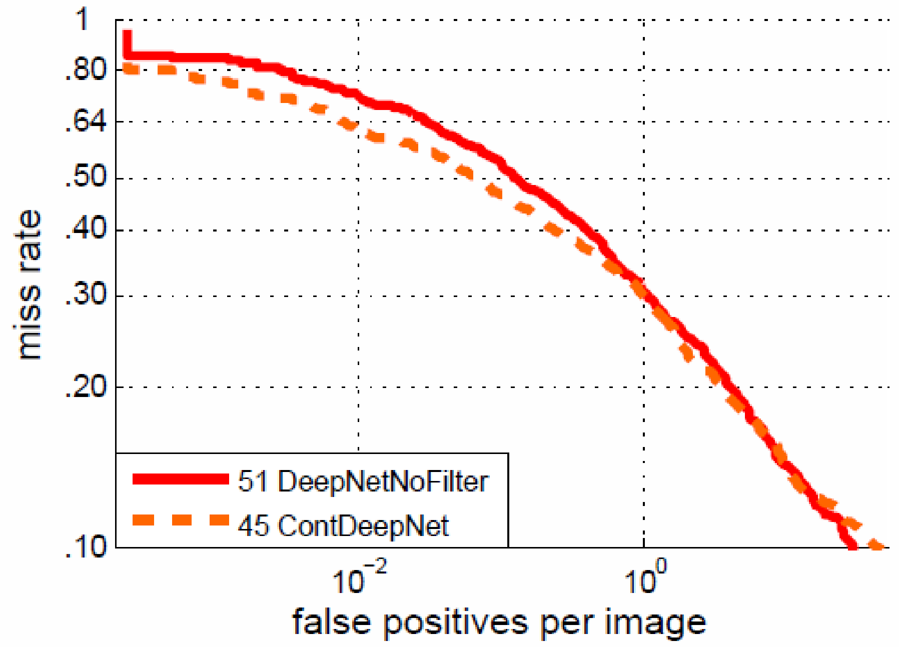
ETHZ

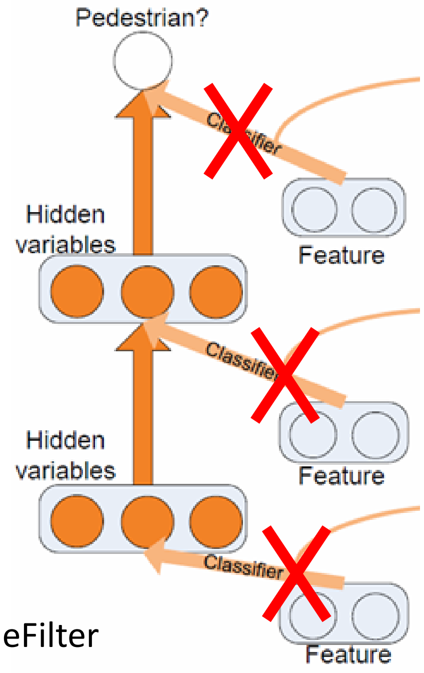
False positives of Net-NoneFilters



False negatives of Net-NoneFilters

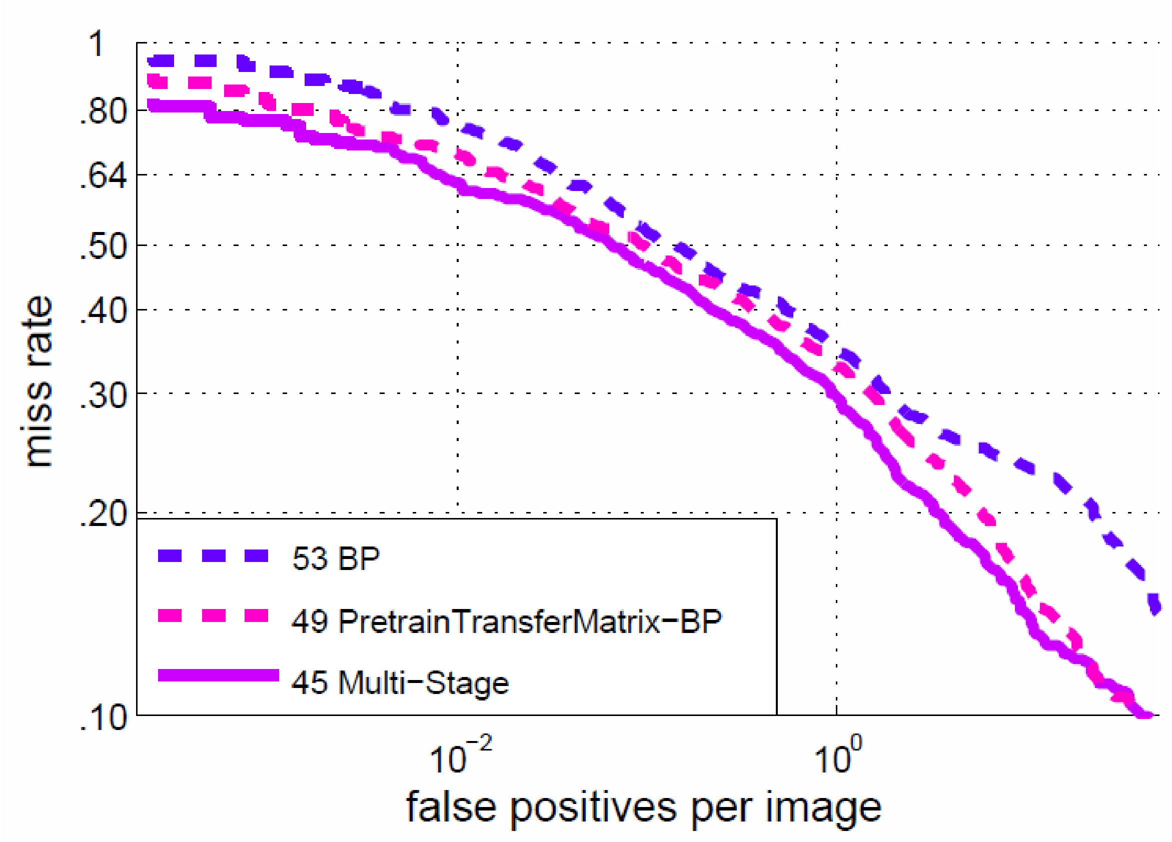






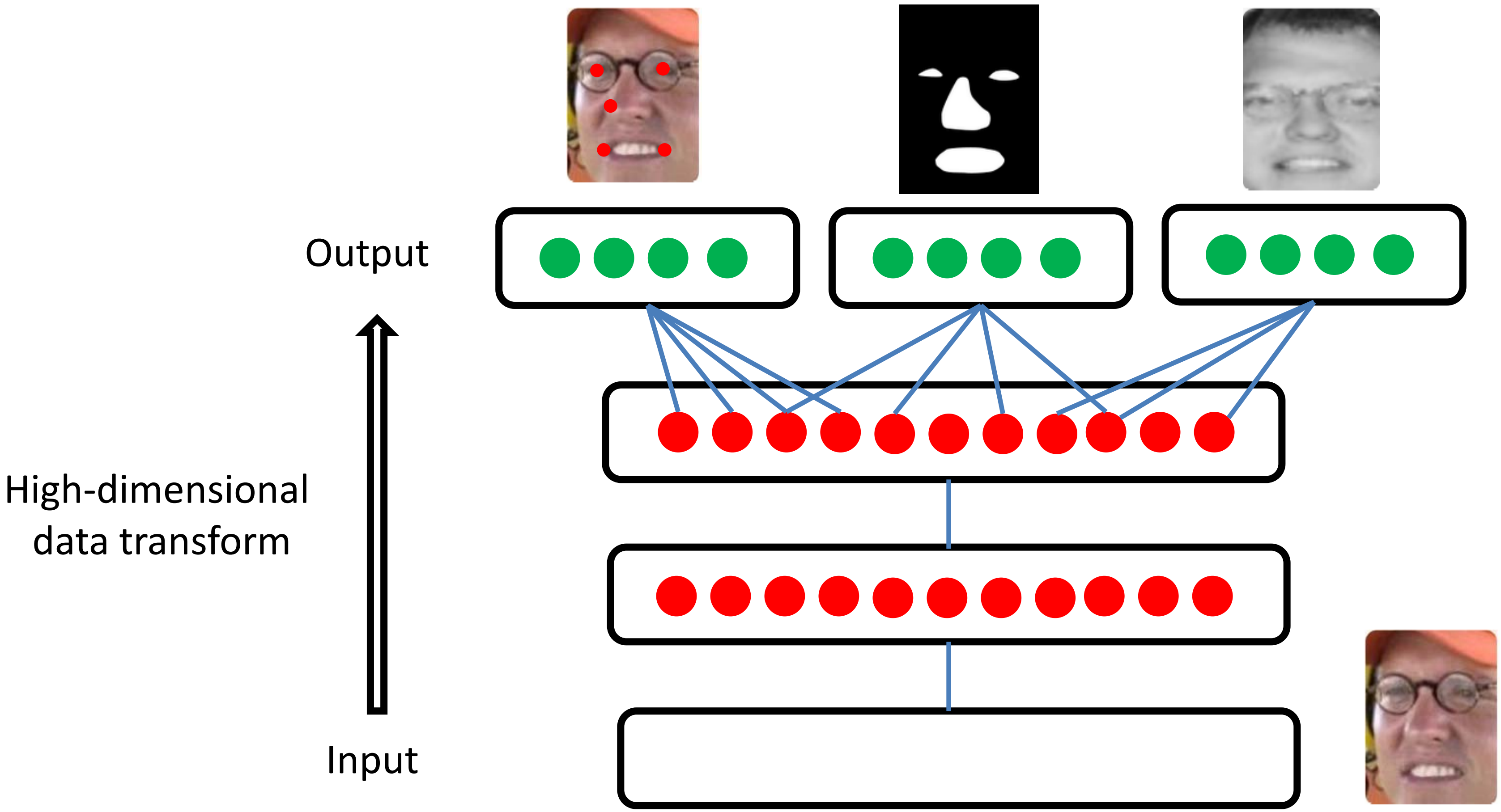
DeepNetNoneFilter

Comparison of Different Training Strategies



Network-BP: use back propagation to update all the parameters without pre-training
PretrainTransferMatrix-BP: the transfer matrices are unsupervised pertrained, and then all the parameters are fine-tuned
Multi-stage: our multi-stage training strategy

High-Dimensional Data Transforms



Output

High-dimensional data transform

Input

Facial keypoint detection: face image -> facial keypoint

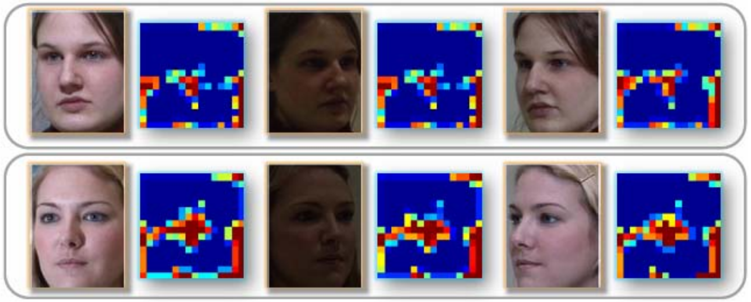
Face transform: face image in a arbitrary view -> face image in a canonical view

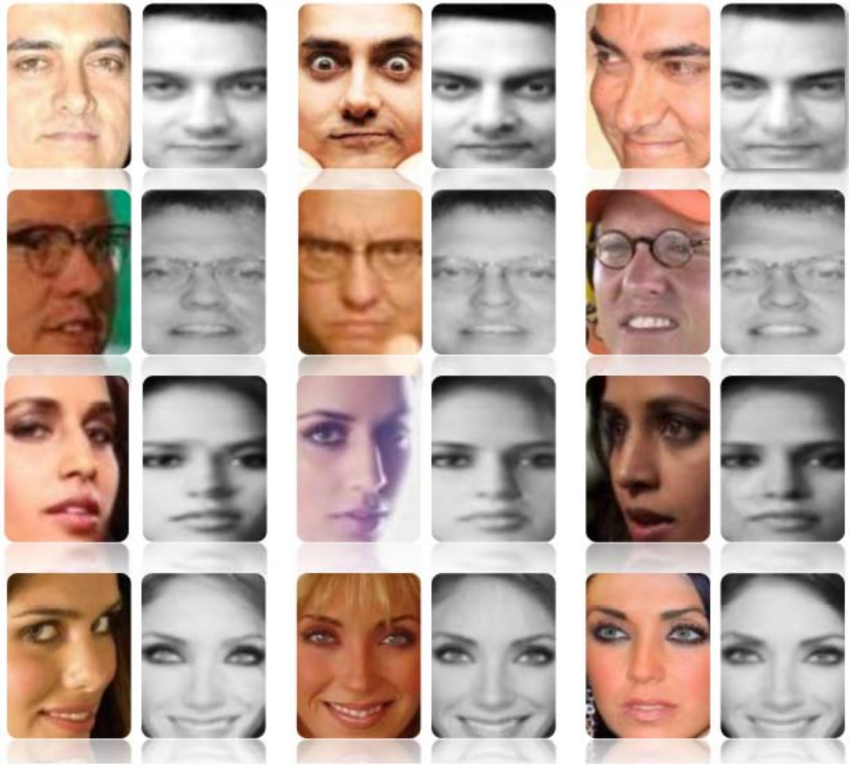
Face parsing: face image -> segmentation maps

Pedestrian parsing : pedestiran image -> segmentation maps

Recovering Canonical-View Face Images

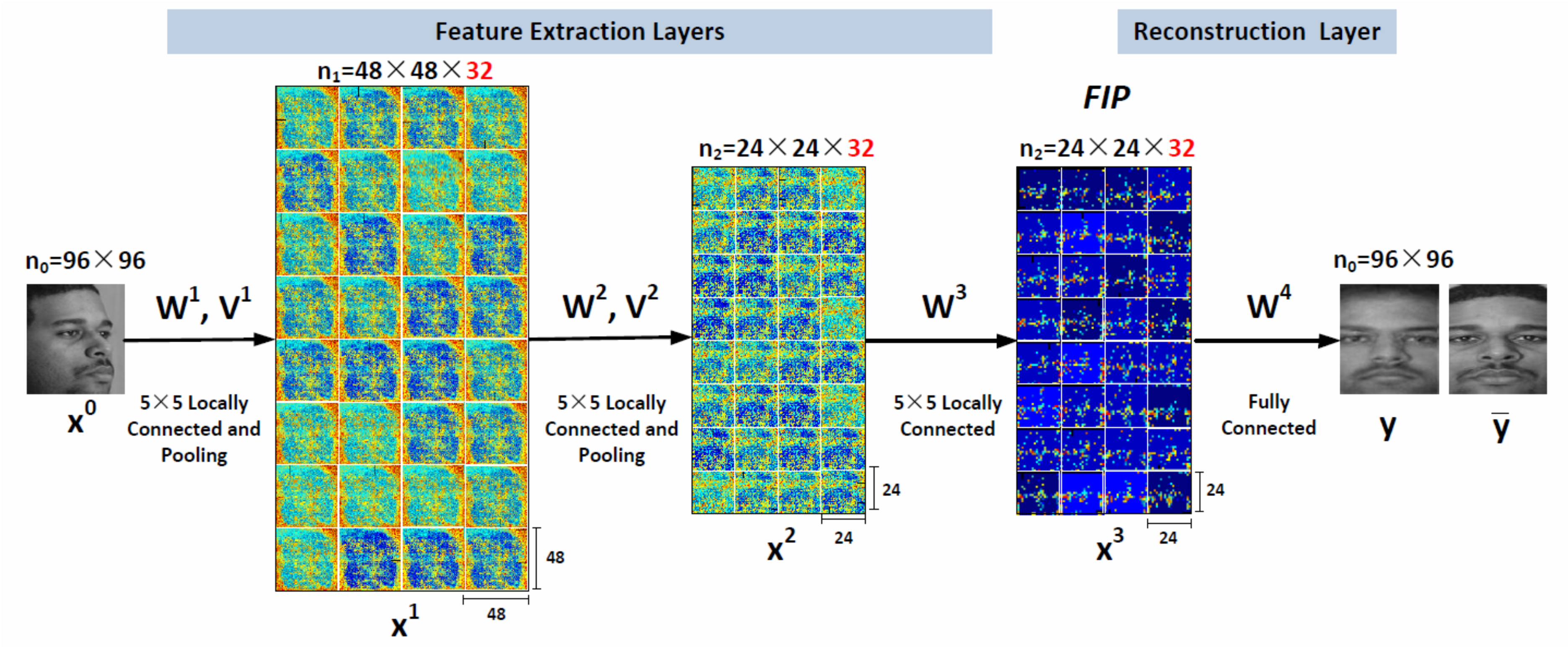
- Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Indentity-Preserving Face Space,” ICCV 2013.





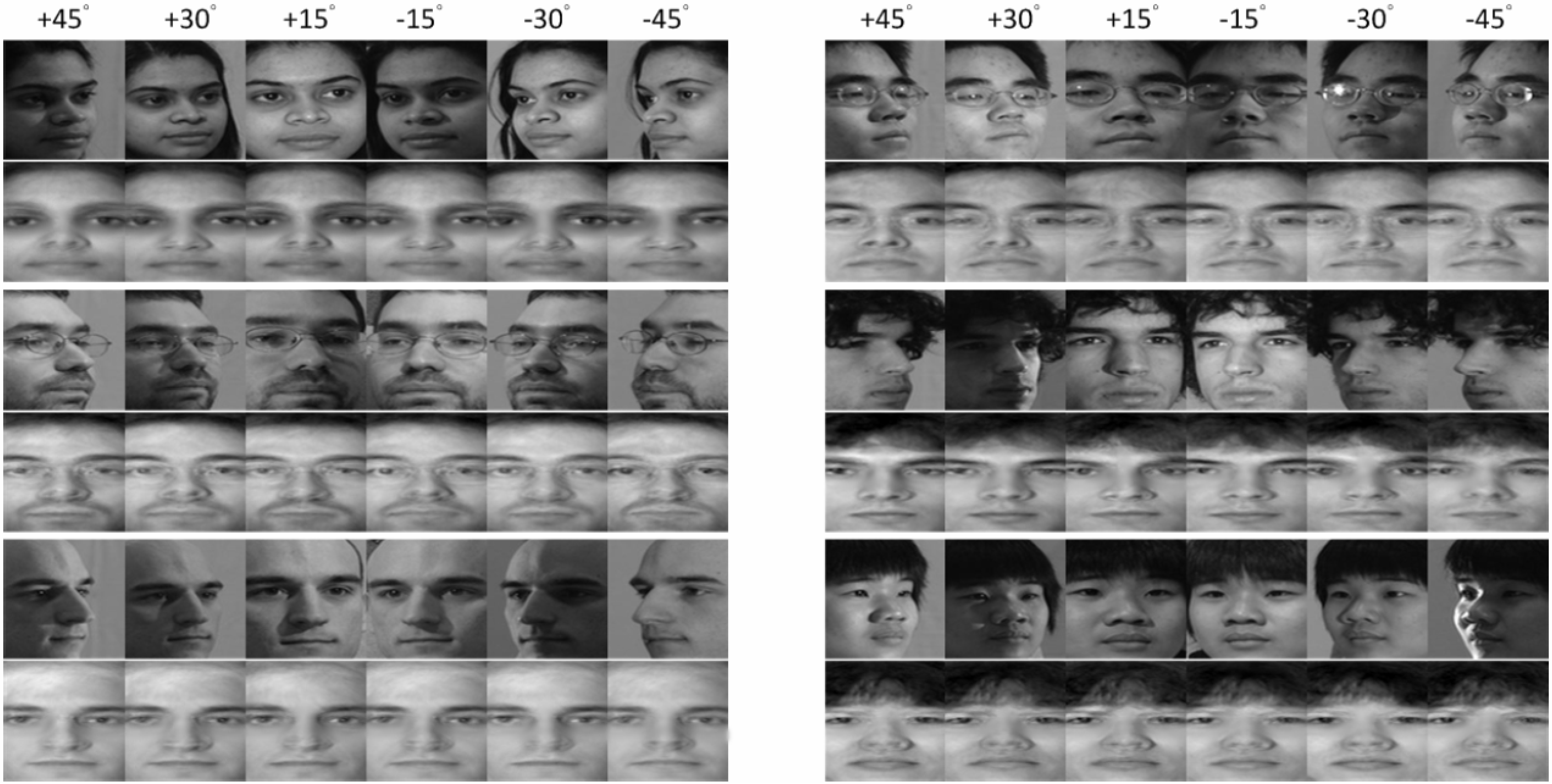
Reconstruction examples from LFW

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



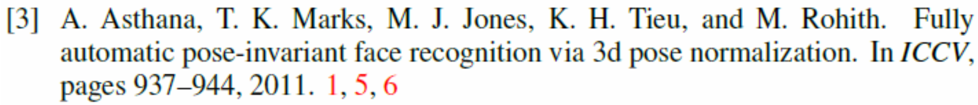
Arbitrary view

Canonical view

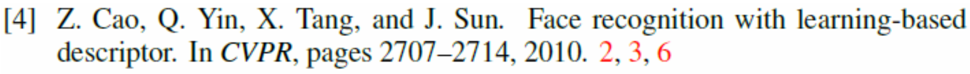

+45°
+30°
+15°
-15°
-30°
-45°
+45°
+30°
+15°
-15°
-30°
-45°

Comparison on Multi-PIE

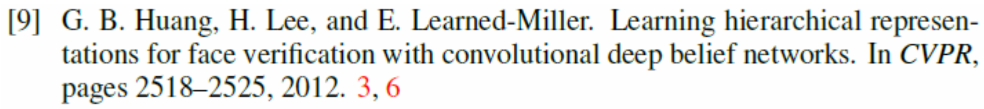
	-45°	-30°	-15°	+15°	+30°	+45°	Avg	Pose
LGBP [26]	37.7	62.5	77	83	59.2	36.1	59.3	√
VAAM [17]	74.1	91	95.7	95.7	89.5	74.8	86.9	√
FA-EGFC[3]	84.7	95	99.3	99	92.9	85.2	92.7	x
SA-EGFC[3]	93	**98.7**	99.7	**99.7**	**98.3**	93.6	97.2	√
LE[4] + LDA	86.9	95.5	99.9	**99.7**	95.5	81.8	93.2	x
CRBM[9] + LDA	80.3	90.5	94.9	96.4	88.3	89.8	87.6	x
Ours	**95.6**	**98.5**	**100.0**	**99.3**	**98.5**	**97.8**	**98.3**	x



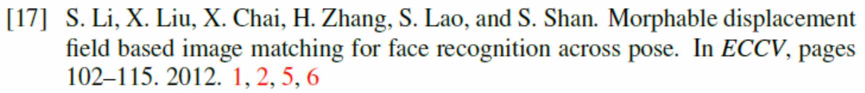
[3] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *ICCV*, pages 937–944, 2011. 1, 5, 6



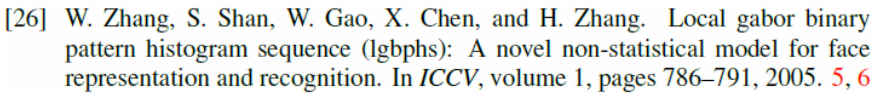
[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714, 2010. 2, 3, 6



[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012. 3, 6



[17] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *ECCV*, pages 102–115. 2012. 1, 2, 5, 6



[26] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *ICCV*, volume 1, pages 786–791, 2005. 5, 6

Comparison on LFW (without outside training data)

Methods	Accuracy (%)
PLDA (Li, TPAMI’12)	90.07
Joint Bayesian (Chen, ECCV’12, 5-point align)	90.9
Fisher Vector Faces (Barkan, ICCV’13)	93.30
High-dim LBP (Chen, CVPR’13, 27-point align)	93.18
Ours (5-point align)	**94.38**

Comparison on LFW (with outside training data)

Methods	Accuracy (%)
Associate-Predict (Yin CVPR’12)	90.57
Joint Bayesian (Chen, ECCV’12, 5-point align)	92.4
Tom-vs-Peter (Berg, BMVC’12, 90-point align)	93.30
High-dim LBP (Chen, CVPR’13, 27-point align)	95.17
Transfer learning joint Bayesian (Cao, ICCV’13, 27-point align)	96.33
Ours (5-point align)	**96.45**

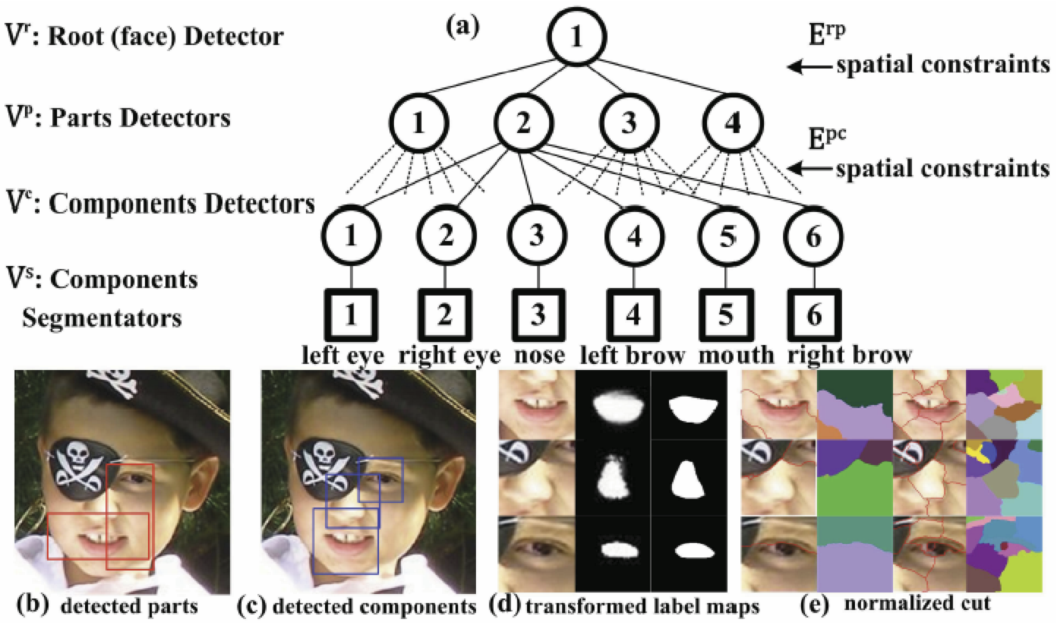
Face Parsing

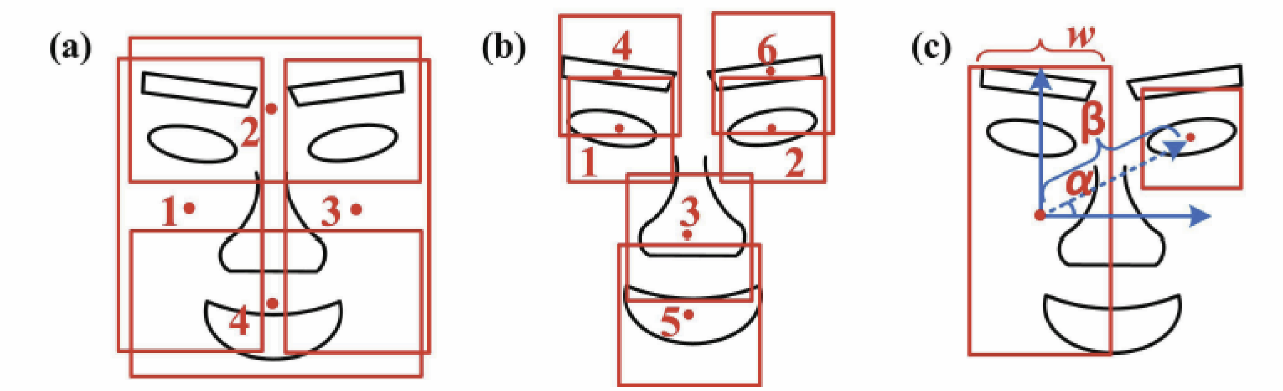
- P. Luo, X. Wang and X. Tang, “Hierarchical Face Parsing via Deep Learning,” CVPR 2012

Motivations

- Recast face segmentation as a cross-modality data transformation problem
- Cross modality autoencoder
- Data of two different modalities share the same representations in the deep model
- Deep models can be used to learn shape priors for segmentation

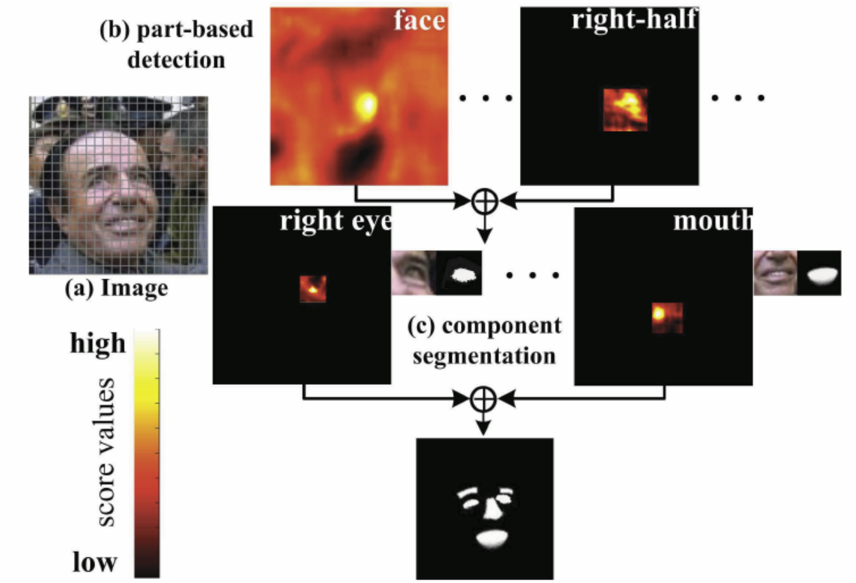
Hierarchical Representation of Face Parsing



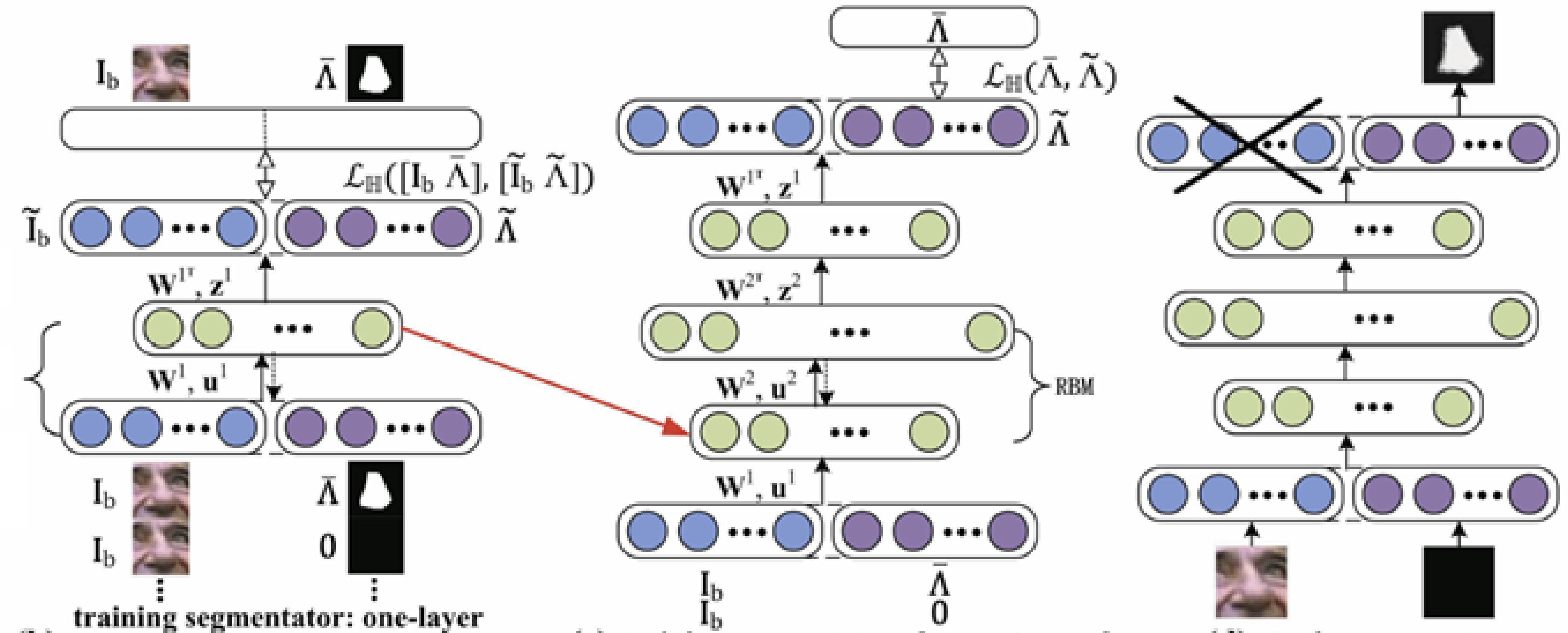


Joint Bayesian Formulation

- Detectors are trained with deep belief net (DBN) and segmentators are trained with deep autoencoder. Both have are generative models.
- Joint Bayesian framework for face detection, part detection, component detection, and component segmentation



Training Segmentators



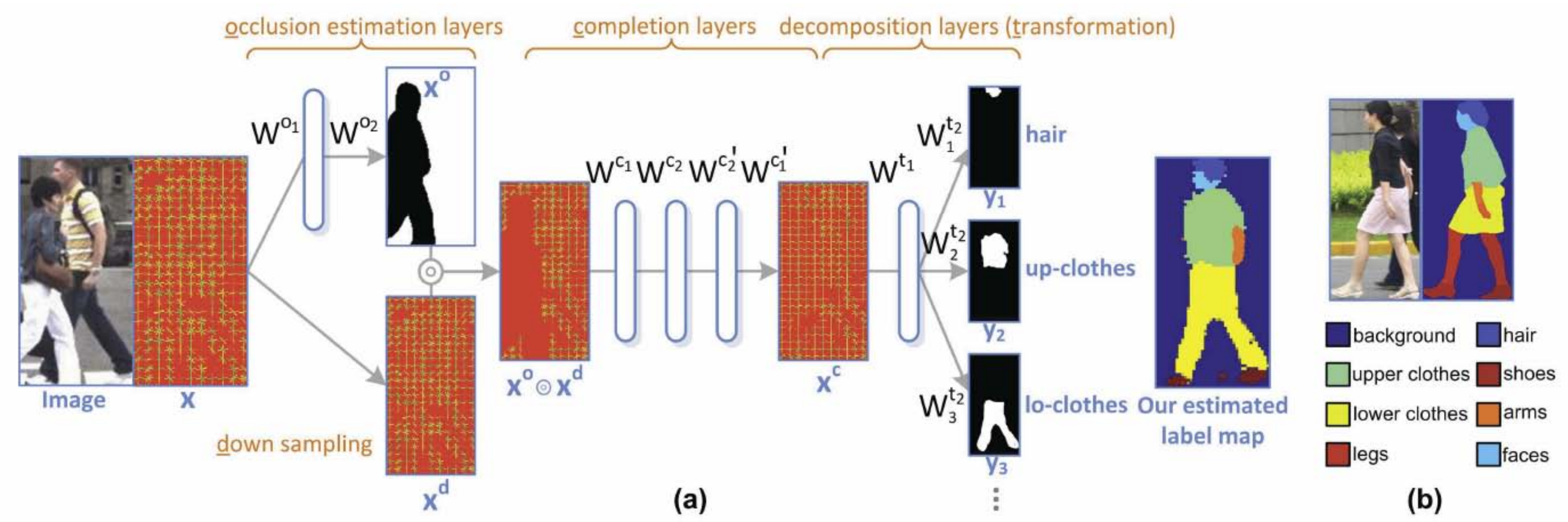
(b) training segmentator: one-layer denoising autoencoder

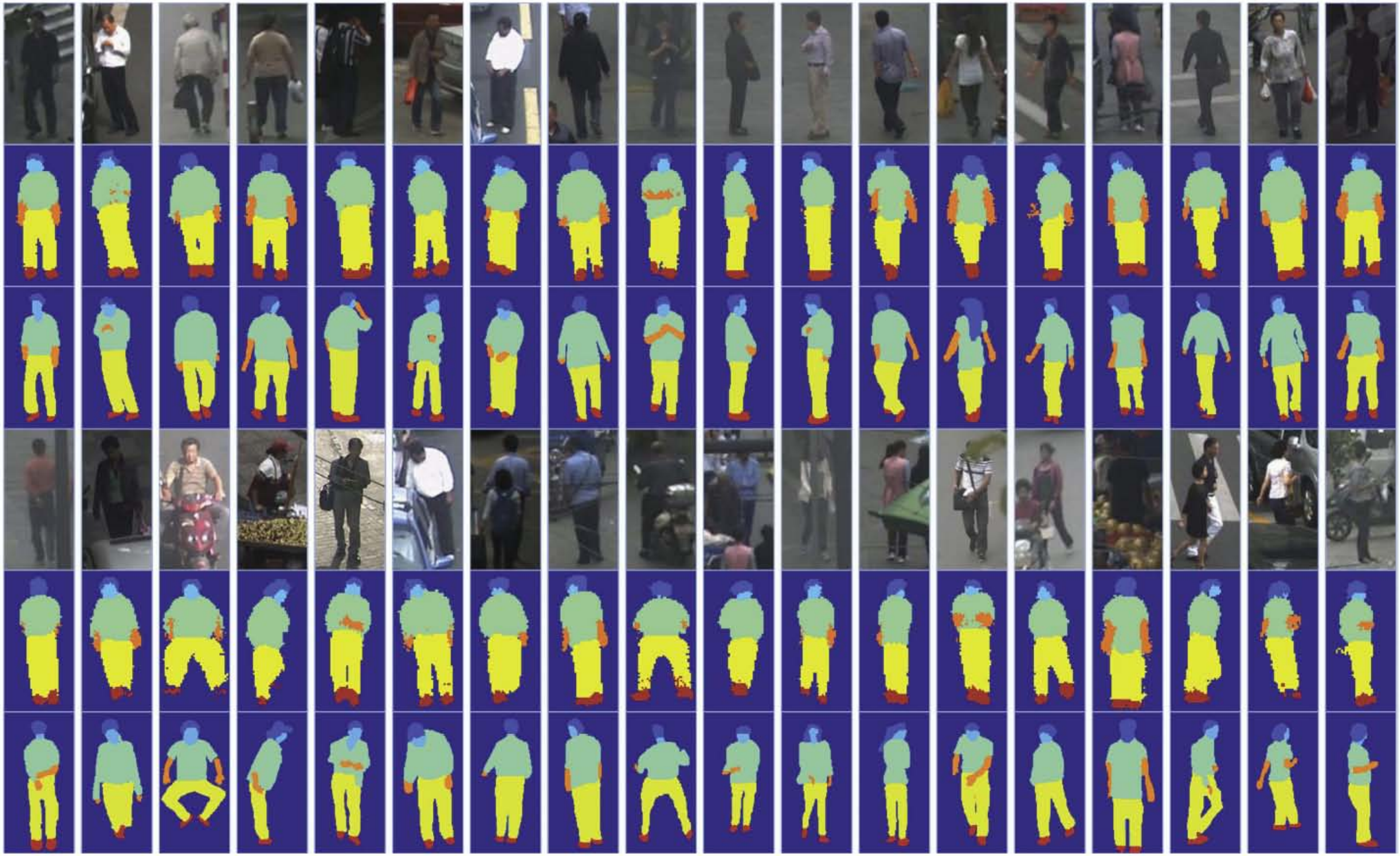
(c) training segmentator: deep autoencoder

(d) testing segmentator

Human Parsing

- P. Luo, X. Wang, and X. Tang, “Pedestrian Parsing via Deep Decompositional Network,” ICCV 2013

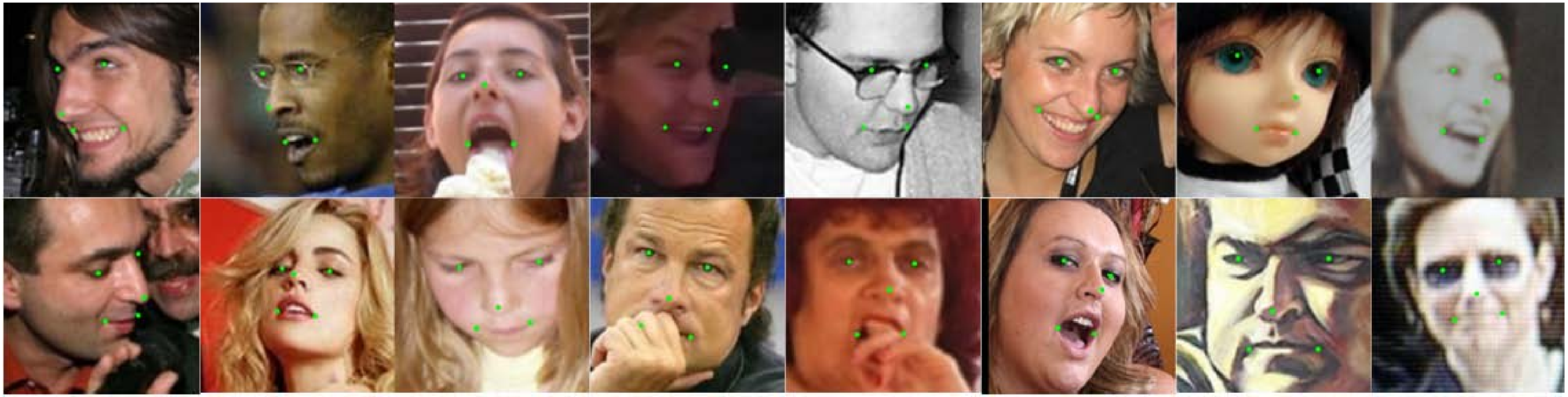


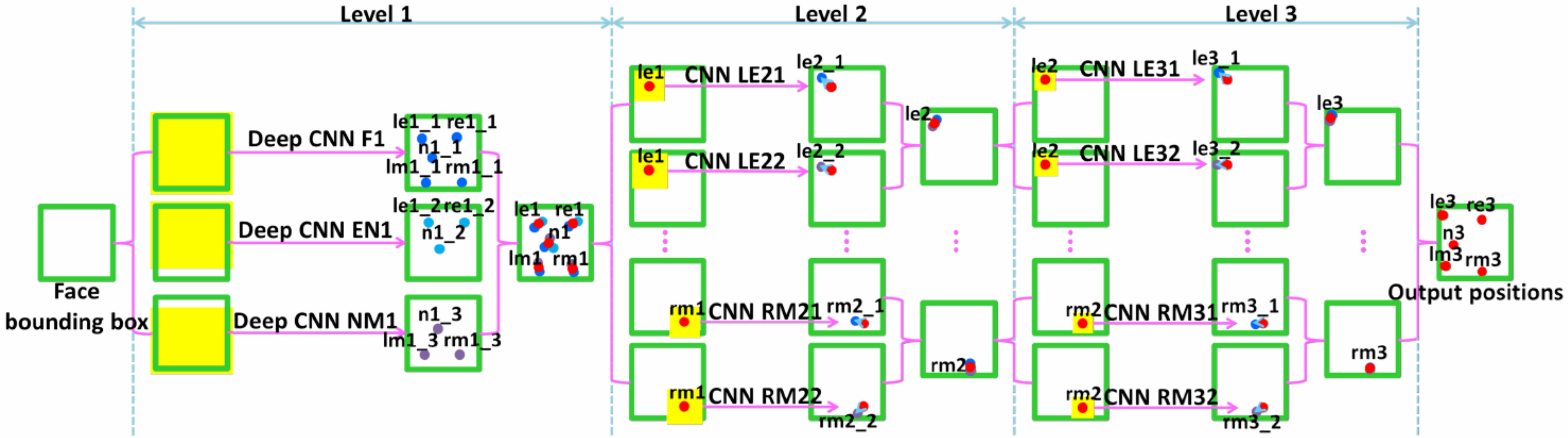


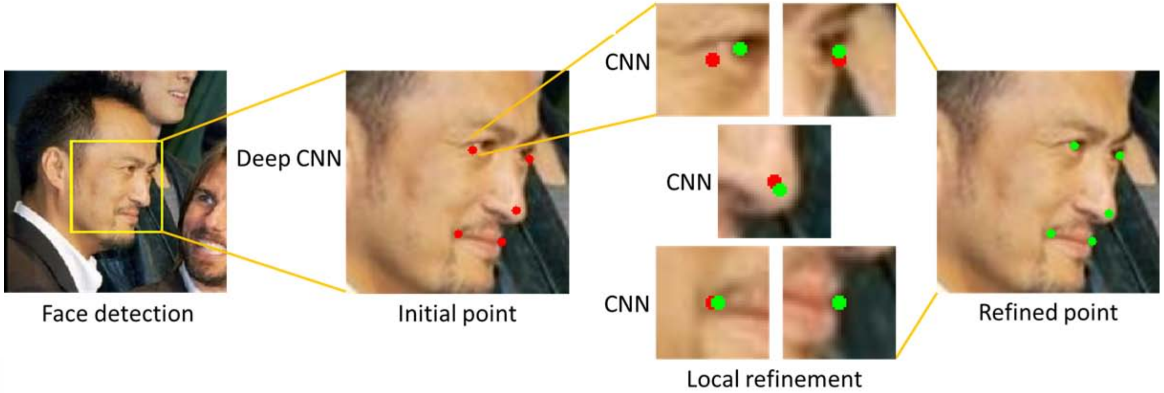
Second row: our result
Third row: ground truth

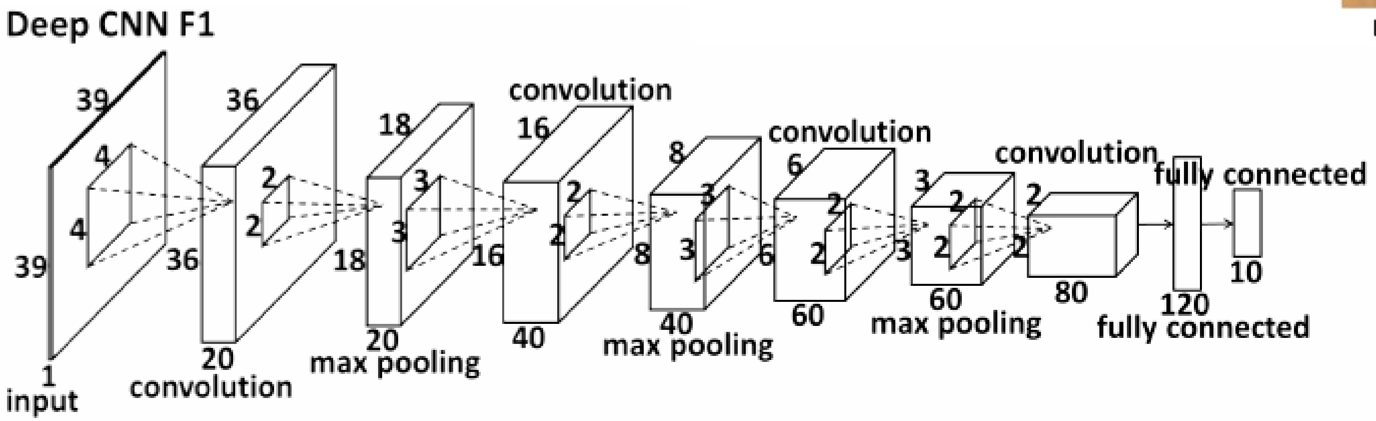
Facial Keypoint Detection

- Y. Sun, X. Wang and X. Tang, “Deep Convolutional Network Cascade for Facial Point Detection,” CVPR 2013




Level 1
Level 2
Level 3
Face
bounding box
Deep CNN F1
Deep CNN EN1
Deep CNN NM1
le1_1 re1_1
n1_1
lm1_1 rm1_1
le1_2 re1_2
n1_2
n1_3
lm1_3 rm1_3
le1 re1
n1
lm1 rm1
le1
CNN LE21
le1
CNN LE22
rm1
CNN RM21
rm1
CNN RM22
le2_1
le2_2
rm2_1
rm2_2
le2
rm2
le2
CNN LE31
le2
CNN LE32
rm2
CNN RM31
rm2
CNN RM32
le3_1
le3_2
rm3_1
rm3_2
le3
rm3
le3 re3
n3
lm3 rm3
Output positions


Deep CNN
CNN
CNN
CNN
Face detection
Initial point
Local refinement
Refined point

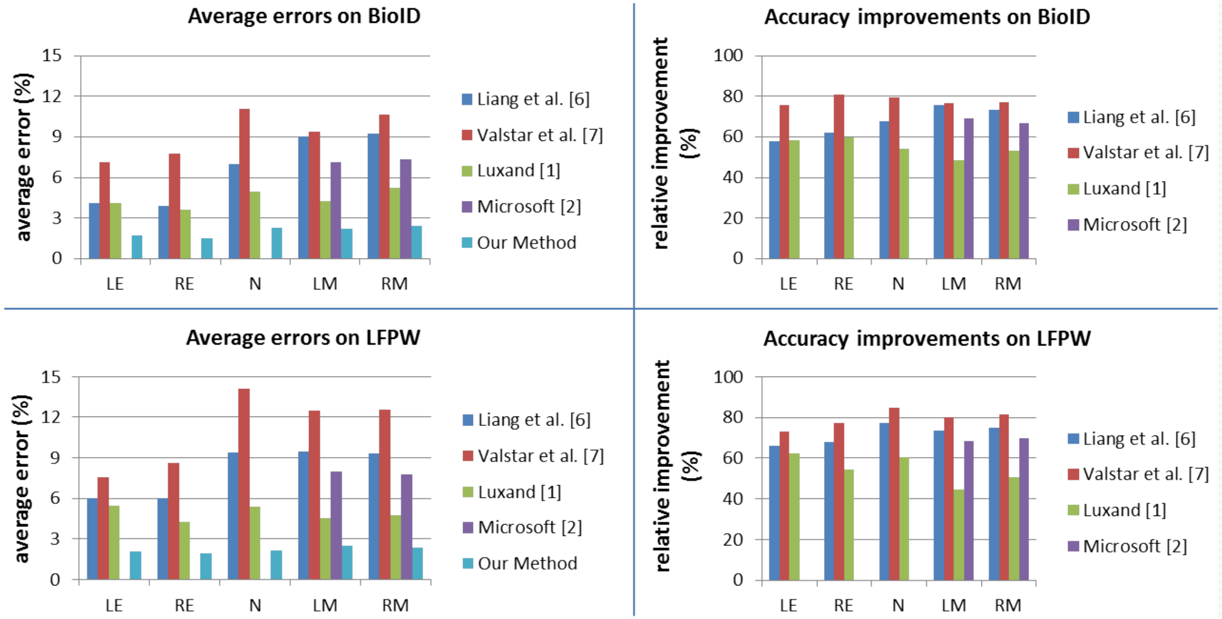

Deep CNN F1
39
39
4
4
1
input
36
36
2
2
20
convolution
18
18
3
3
20
max pooling
convolution
16
16
2
2
40
8
8
3
3
40
max pooling
convolution
6
6
2
2
60
3
3
2
2
60
max pooling
convolution
2
2
80
fully connected
120
fully connected
10

Benefits of Using Deep Model

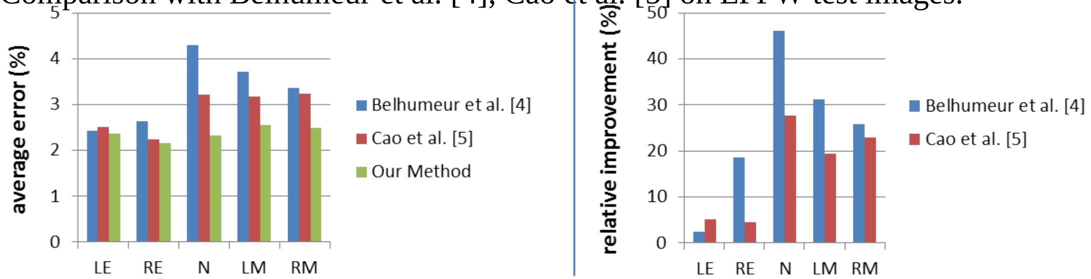
- Take the full face as input to make full use of texture context information over the entire face to locate each keypoint
- The first network of tacking the whole face as input needs **deep** structures to extract **high-level** features
- Since the networks are trained to predict all the keypoints simultaneously, the geometric constraints among keypoints are implicitly encoded

Comparison with Liang et al. [6], Valstar et al. [7], Luxand Face SDK [1] and Microsoft Research Face SDK [2] on BioID and LFPW.

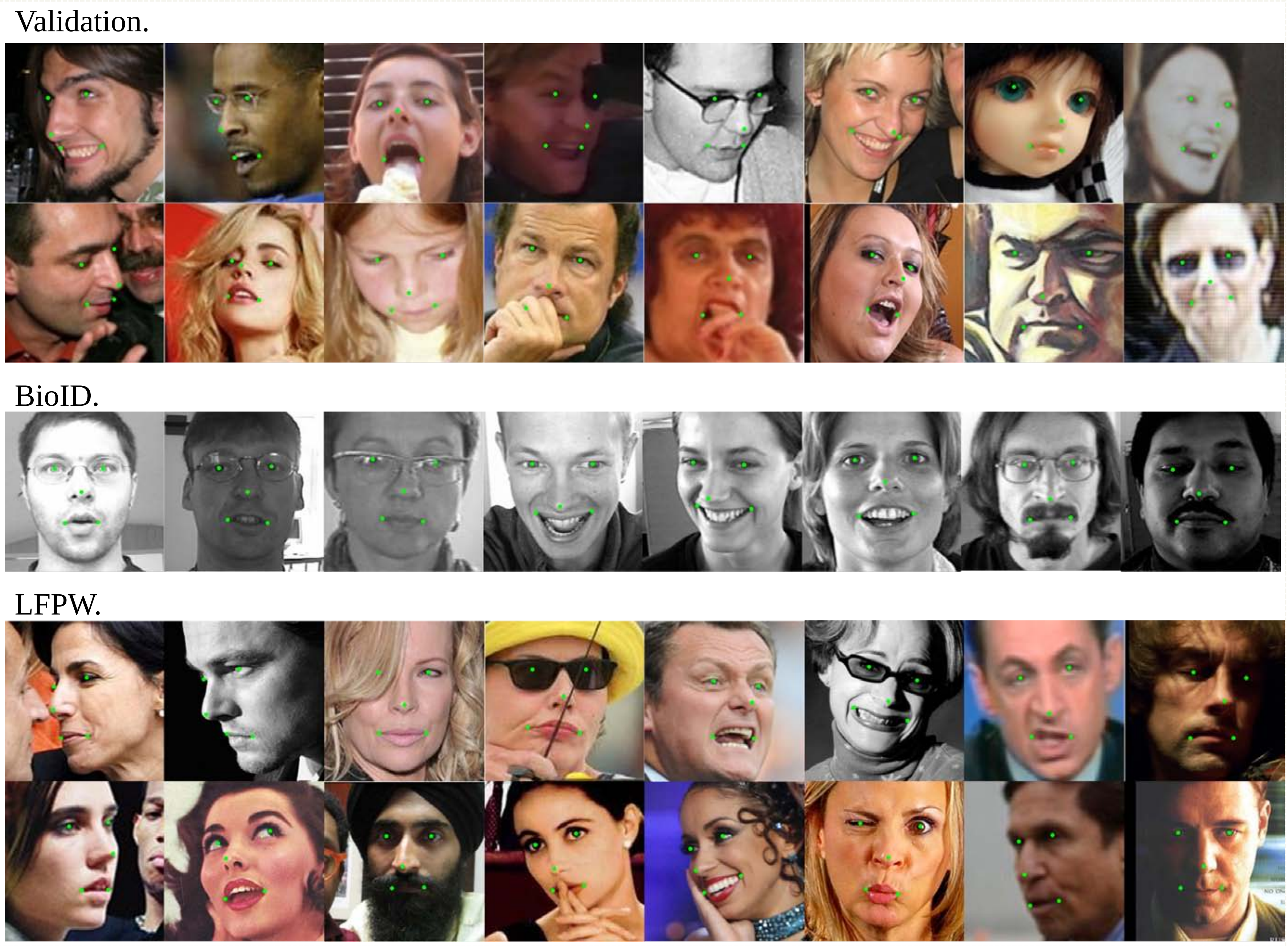
$$\text{Relative improvement} = \frac{\text{reduced average error}}{\text{average error of the method in comparison}}.$$



Comparison with Belhumeur et al. [4], Cao et al. [5] on LFPW test images.



1. http://www.luxand.com/facesdk/
2. http://research.microsoft.com/en-us/projects/facesdk/.
3. O. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In Proc. AVBPA, 2001.
4. P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In Proc. CVPR, 2011.
5. X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In Proc. CVPR, 2012.
6. L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In Proc. ECCV, 2008.
7. M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In Proc. CVPR, 2010.



Validation.

BioID.

LFPW.

Conclusions

- Deep learning can jointly optimize key components in vision systems
- Prior knowledge from vision research is valuable for developing deep models and training strategies
- Deep learning can solve some vision challenges as problems of high-dimensional data transform
- Challenging prediction tasks can make better use the large learning capacity and avoid overfitting

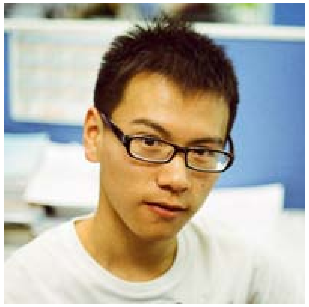
People working on deep learning in our group

Wanli Ouyang

Ping Luo

Yi Sun

Xingyu Zeng



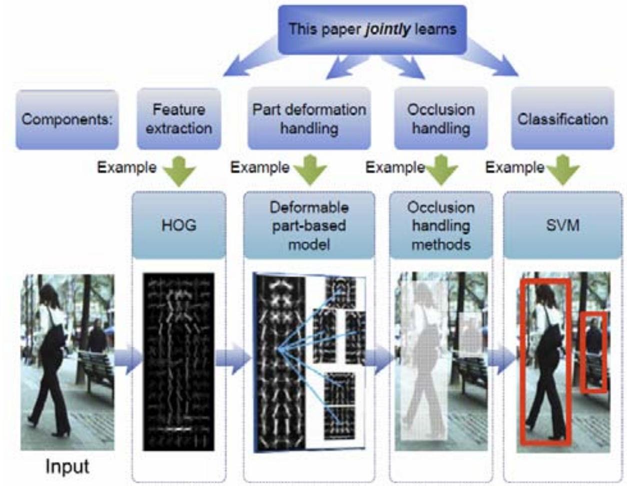
Zhenyao Zhu

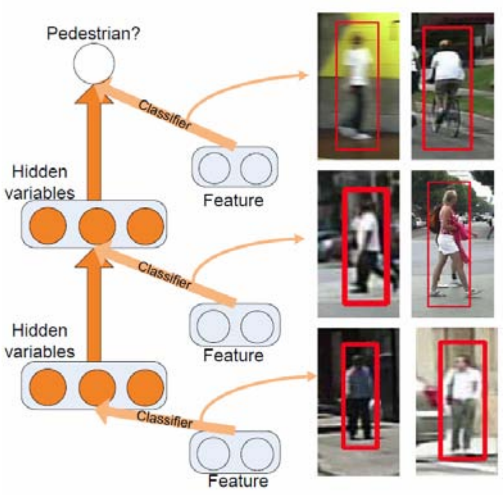
Acknowledgement

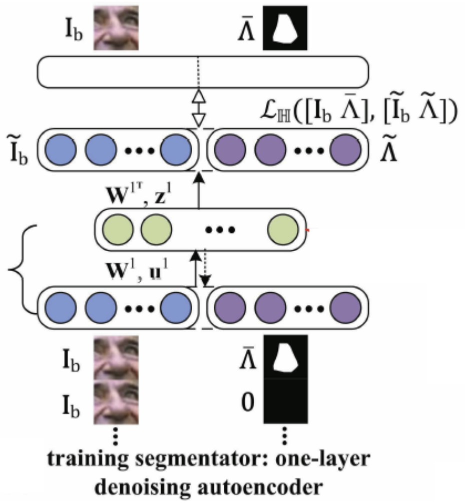
Hong Kong Research Grants Council

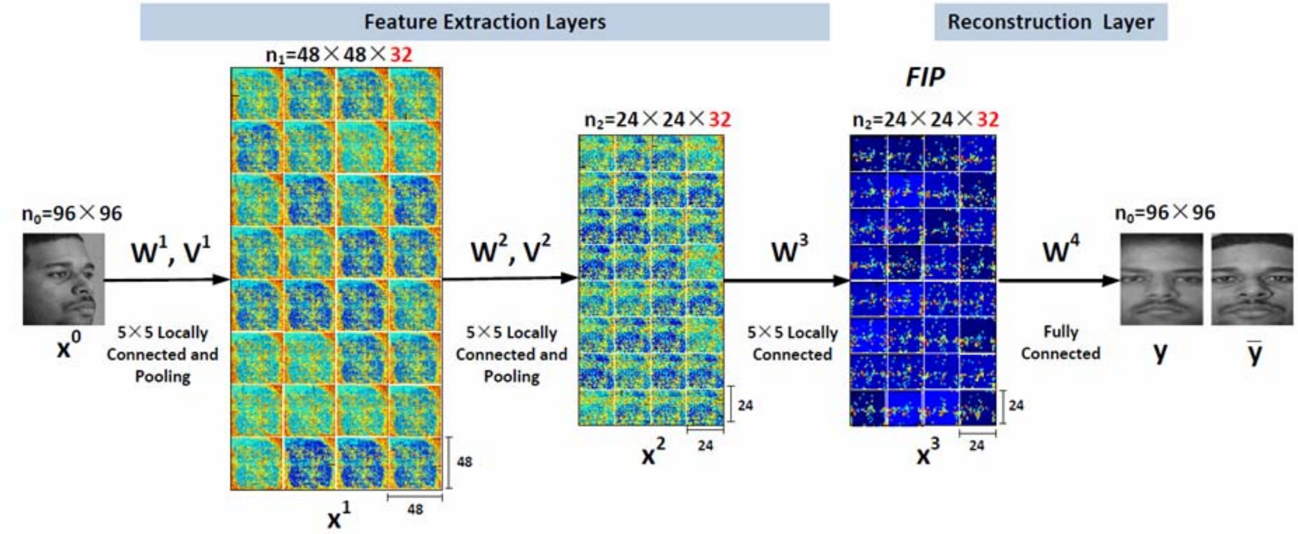
中国自然科学基金

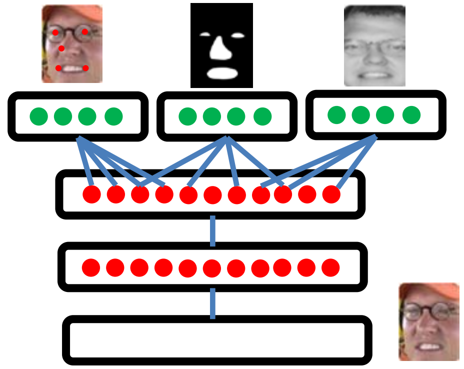
Thank you!


This paper jointly learns
Components:
Feature extraction
Part deformation handling
Occlusion handling
Classification
Example
HOG
Deformable part-based model
Occlusion handling methods
SVM
Input




I_b
$\bar{\Lambda}$
$\mathcal{L}_{\mathbb{H}}([I_b\ \bar{\Lambda}], [\tilde{I}_b\ \tilde{\Lambda}])$
$\tilde{I}_b$
$\tilde{\Lambda}$
W^{1T}, z^1
W^1, u^1
I_b
$\bar{\Lambda}$
I_b
0
training segmentator: one-layer denoising autoencoder





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

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