香港中文大學
The Chinese University of Hong Kong

Examples of Deep Learning Applications

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

Our first deep learning project
– January 2011

Wanli Ouyang

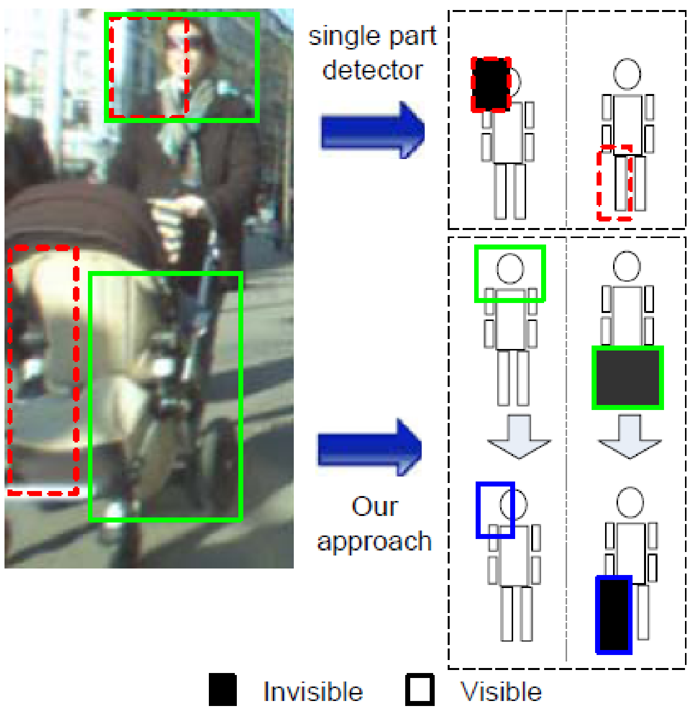
We wish to work on pedestrian detection

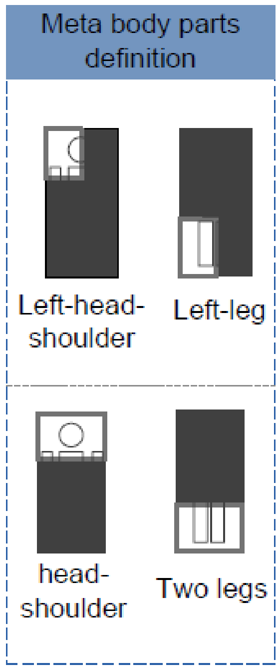
Where to start?

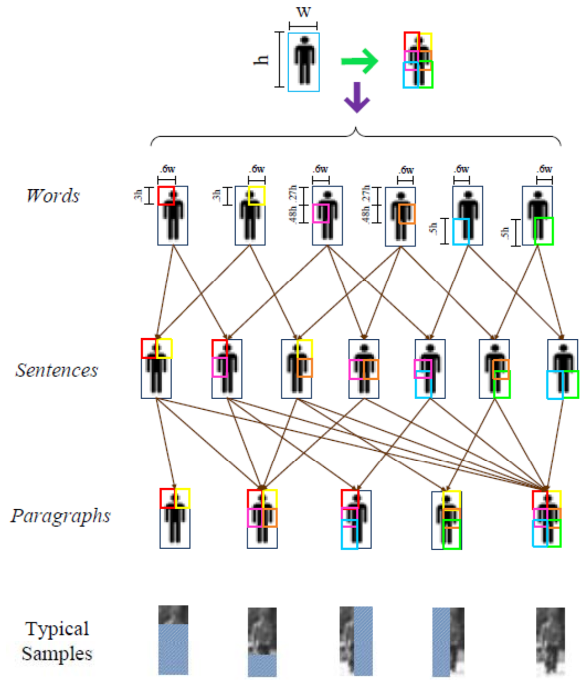
Our understanding of deep learning
- **DBN**
- **Unsupervised learning**
- **Model complex nonlinear relationship of variables**

Pedestrian detection

G. Duan, H. Ai, and S. Lao, “A structural filter approach to human detection,” in ECCV, 2010.





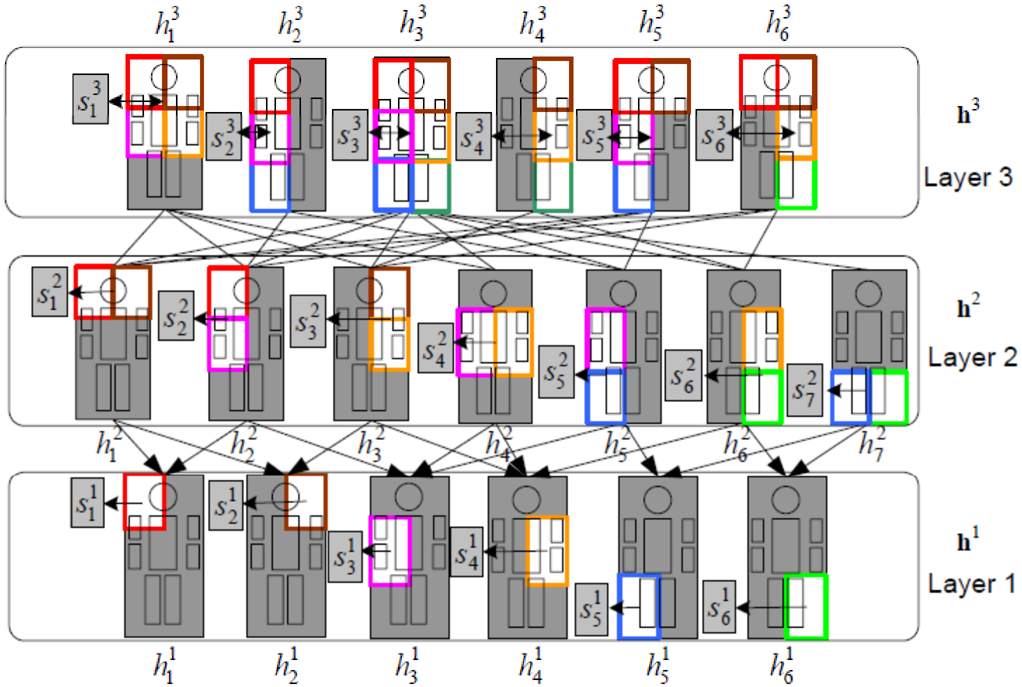


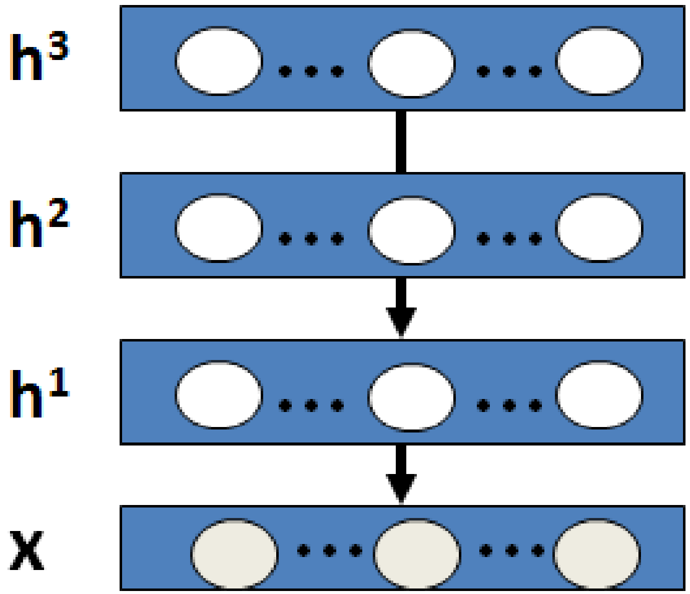
Use manually defined rules to describe the relationship between the visibility of a part and its overlapping larger parts and smaller parts, e.g. if the head or the torso was invisible, its larger part of upper-body should also be invisible.

Deep learning? **Deep belief net**

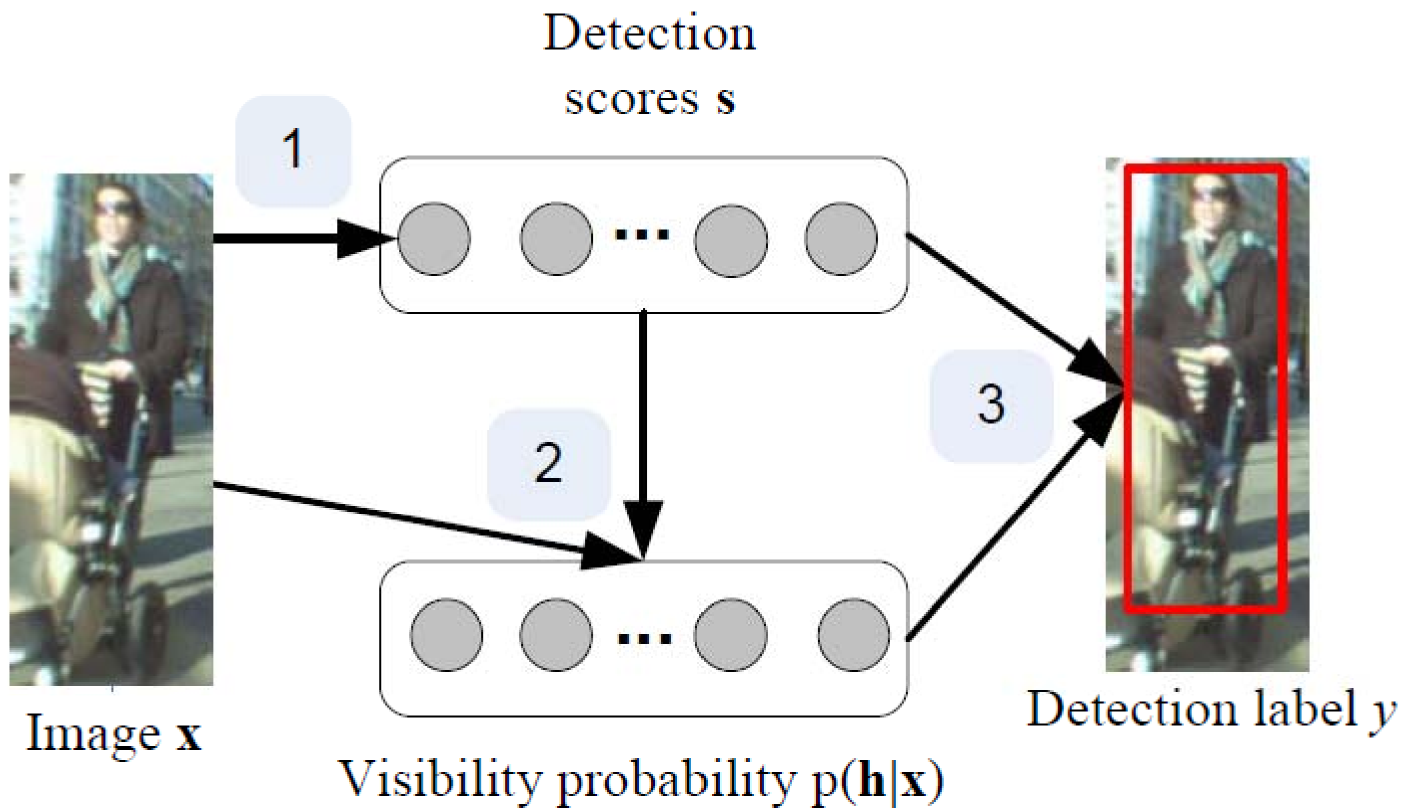
$$p(y|\mathbf{x}) = \sum_{\mathbf{h}} p(y, \mathbf{h}|\mathbf{x}) = \sum_{\mathbf{h}} p(y|\mathbf{h}, \mathbf{x})p(\mathbf{h}|\mathbf{x})$$

- **The hidden units in BDN have no physical meaning**
- **DBN is fully connected**





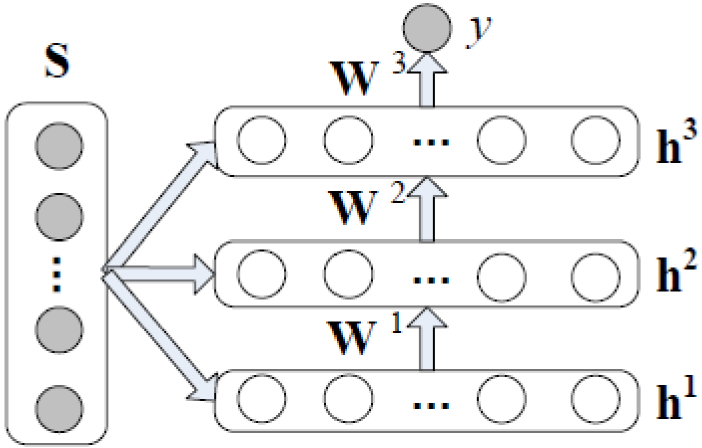
W. Ouyang and X. Wang, "A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling," CVPR 2012

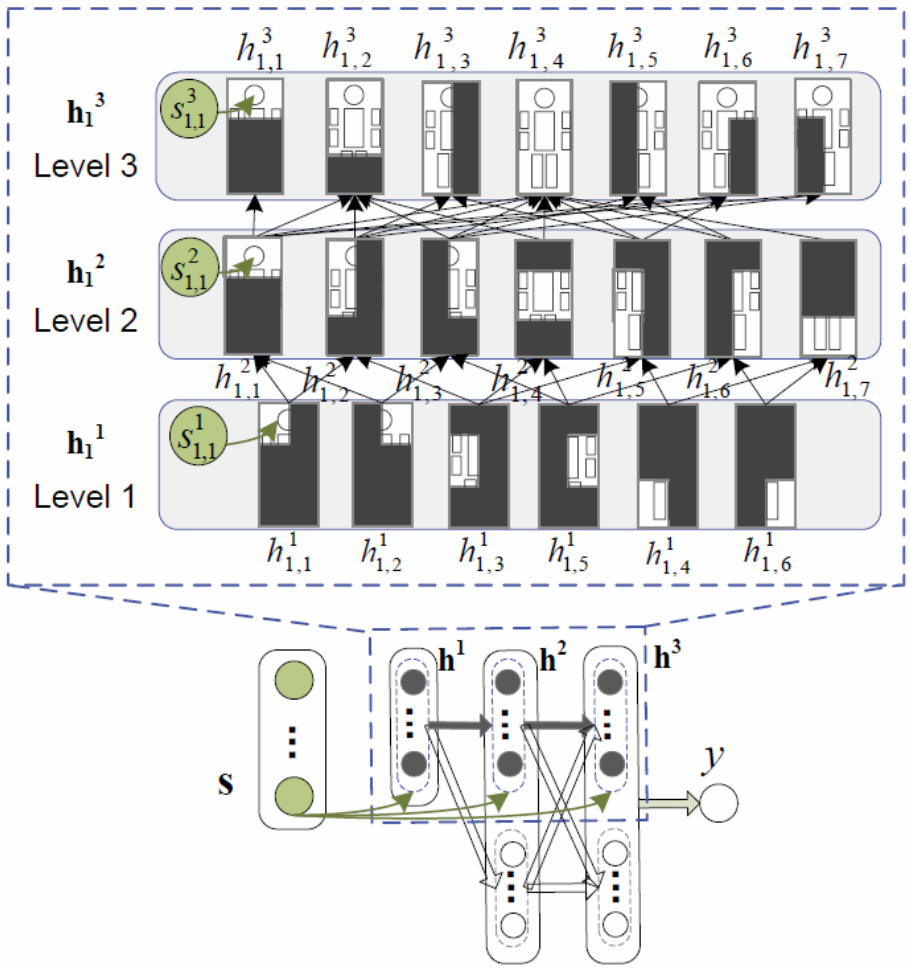


1. Part detection, 2. Visibility estimation,
3. Detection score integration

1. obtain the detection scores $\mathbf{s}$ by part detectors;
2. use $\mathbf{s}$ and $\mathbf{x}$ to estimate visibility probability $p(\mathbf{h}|\mathbf{x})$;
3. combine the detection scores $\mathbf{s}$ with the visibility probability $p(\mathbf{h}|\mathbf{x})$ to estimate the probability of an input window being pedestrian, c.f. (2) and (3).

- **Each hidden unit is associated with a part detection score and it indicates the visibility of a part**
- **DBN is designed considering the structure of human body**

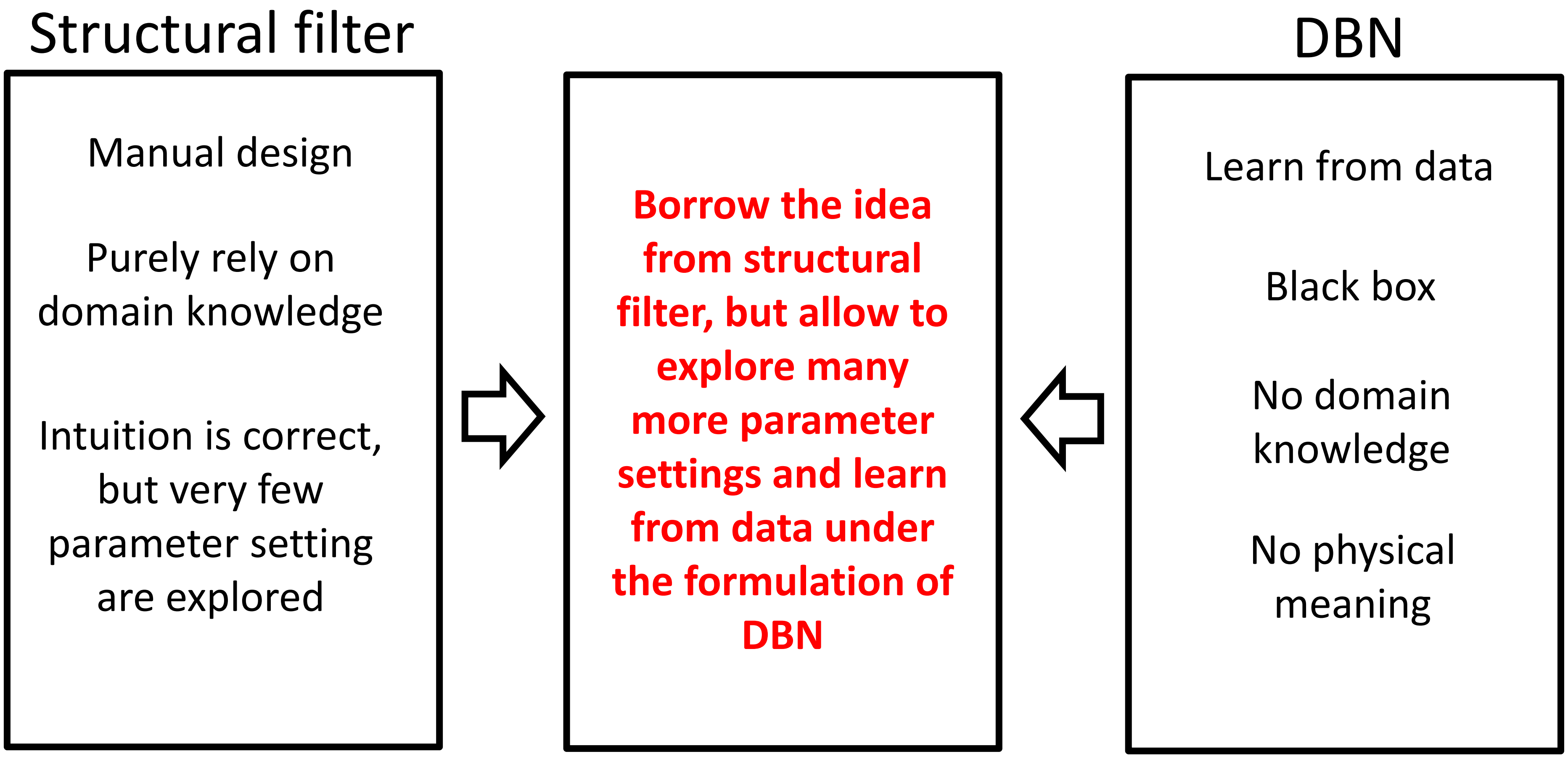




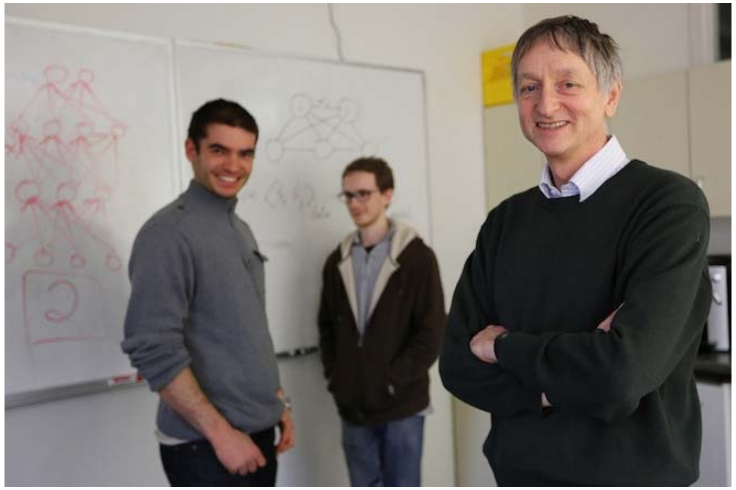
$$\mathbf{W}^l = \mathbf{W}^{l,0} + \tilde{\mathbf{W}}^l \circ \tilde{\mathbf{S}}^l$$

$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}^l_{*,j} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score


Structural filter
Manual design
Purely rely on domain knowledge
Intuition is correct, but very few parameter setting are explored
Borrow the idea from structural filter, but allow to explore many more parameter settings and learn from data under the formulation of DBN
DBN
Learn from data
Black box
No domain knowledge
No physical meaning

Deep Learning Won ImageNet Image Classification Challenge 2012

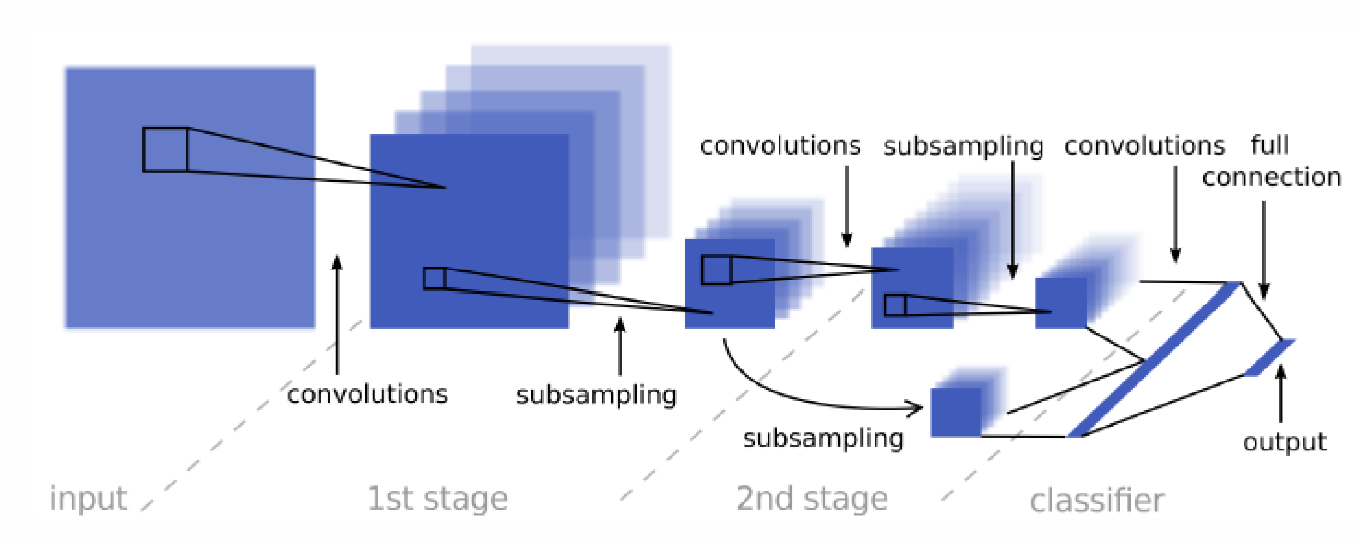


Our understanding of deep learning

- **Large scale supervised learning with CNN**
- **The key of deep learning is to learn feature representation**

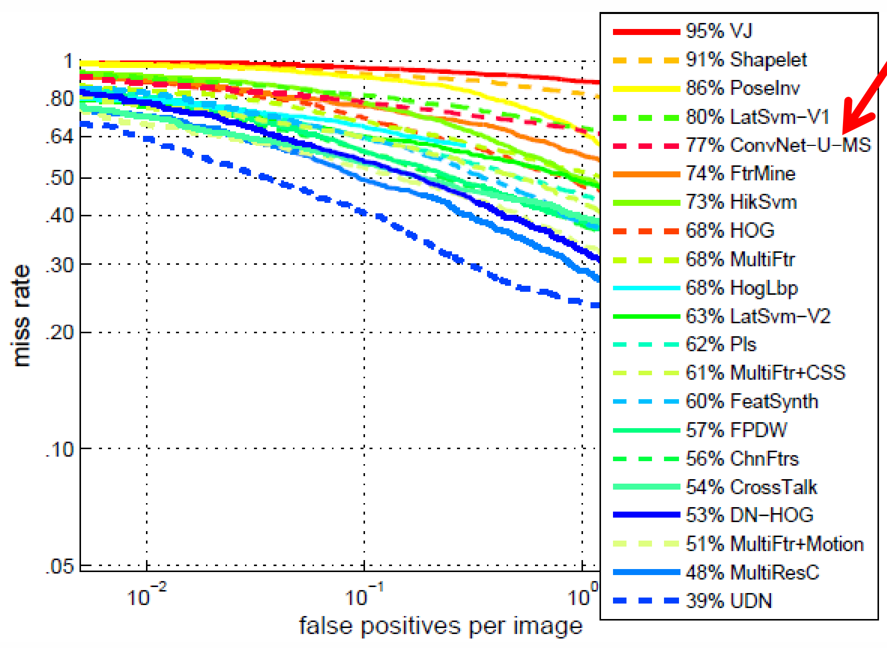
How to learn features in pedestrian detection?

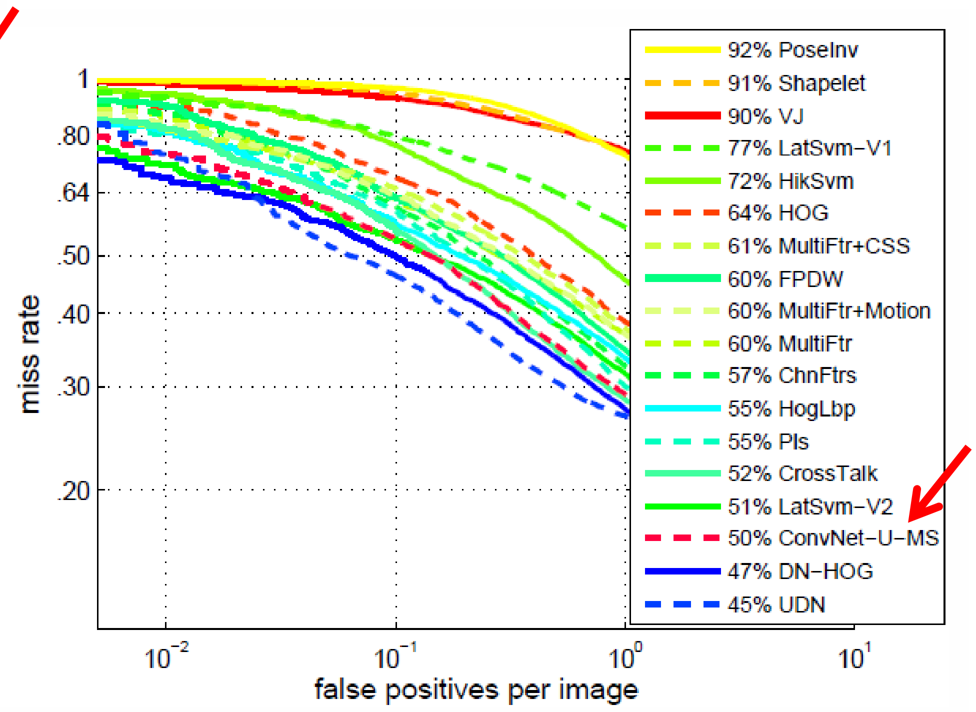
It may not be a good idea to treat deep learning as a black box



ConvNet−U−MS

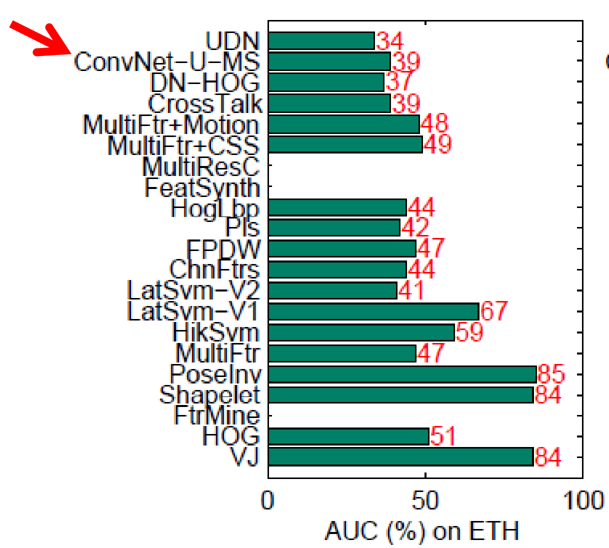
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, "Pedestrian Detection with Unsupervised Multi-Stage Feature Learning," CVPR 2013.

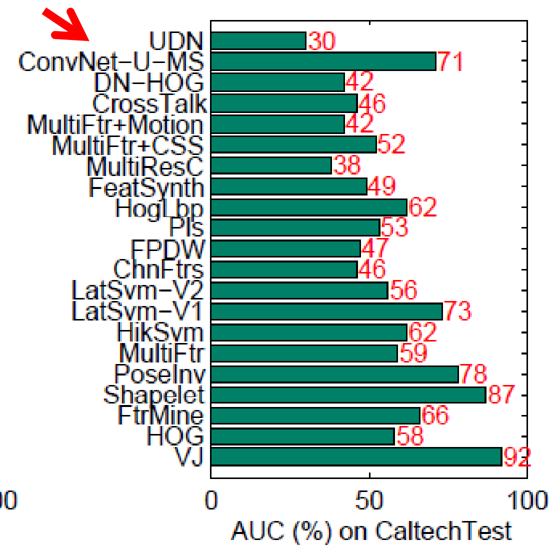




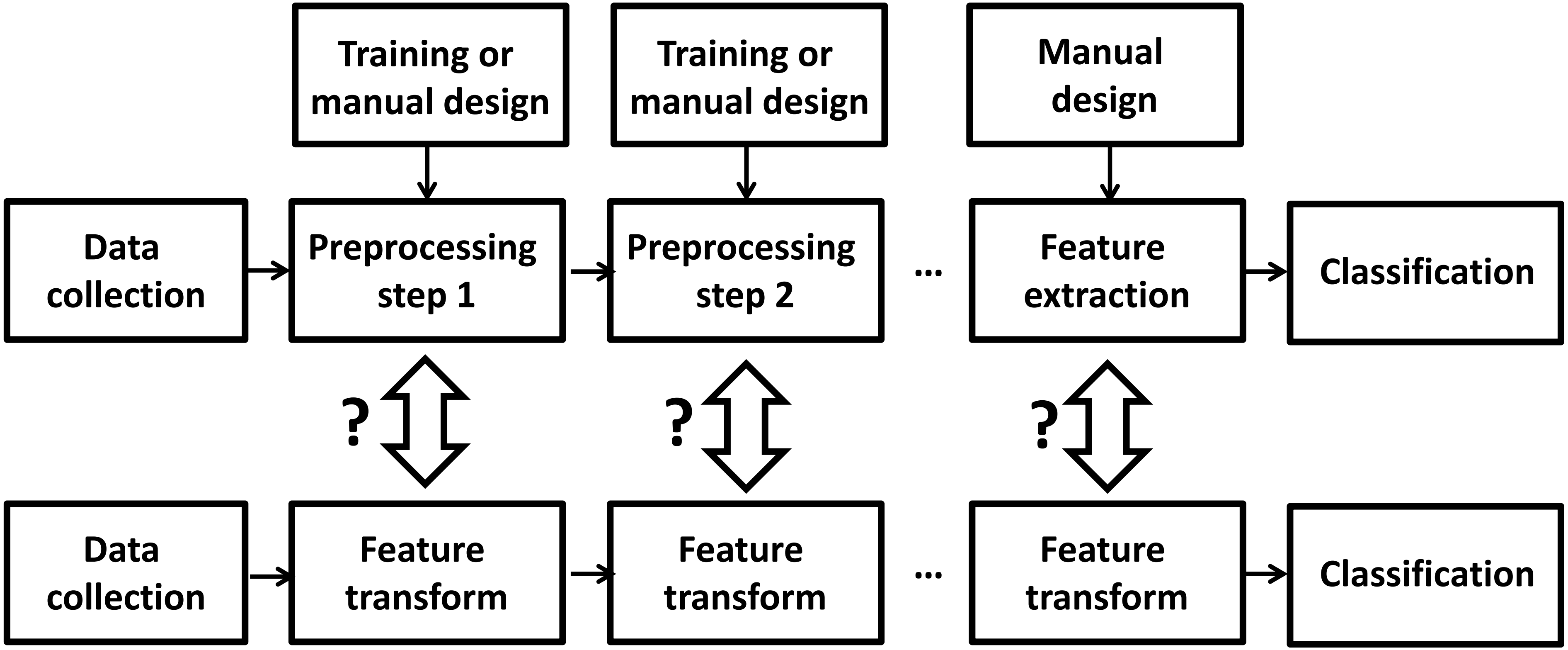
Results on Caltech Test

Results on ETHZ





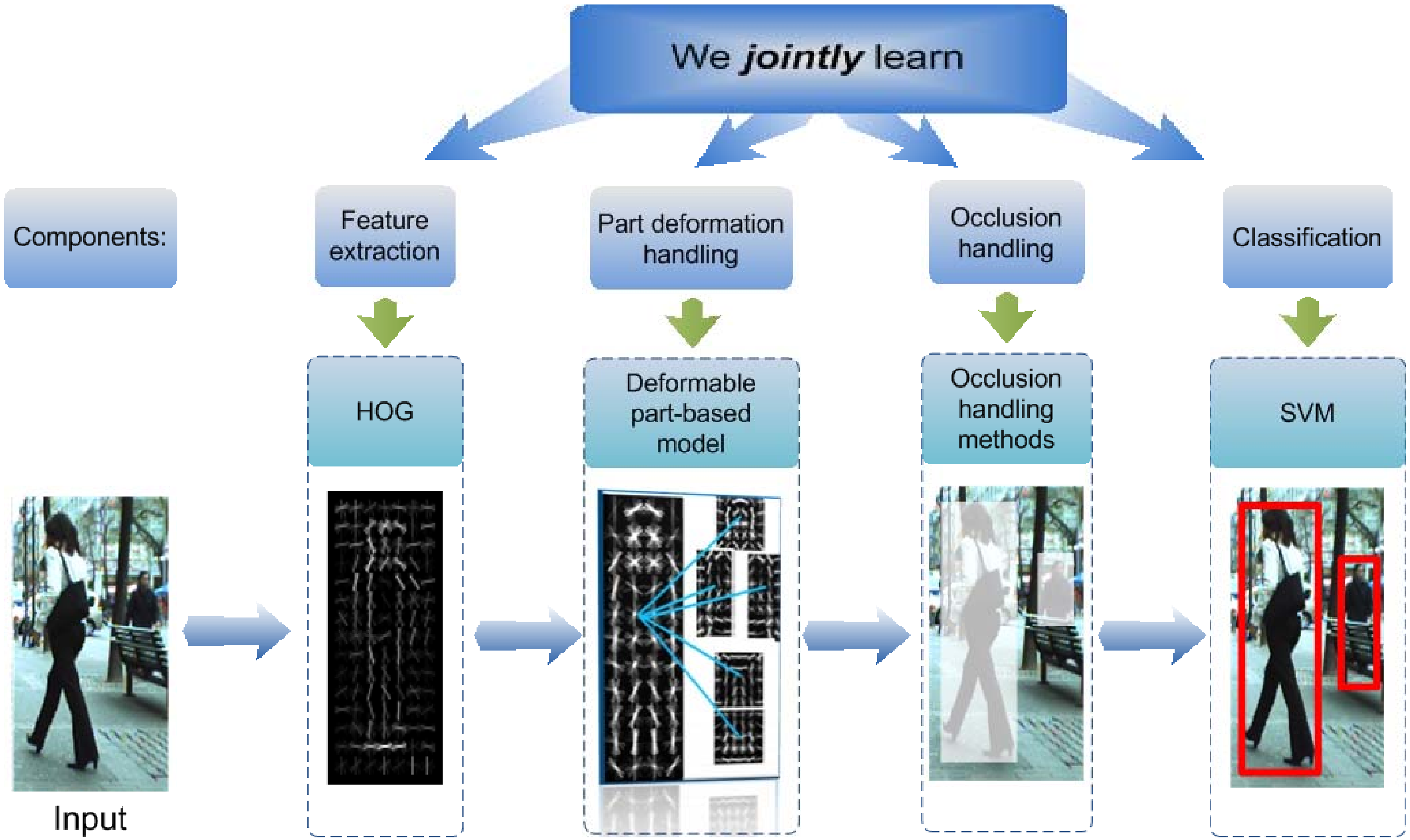
Bridge the connection between deep learning and conventional systems



End-to-end learning

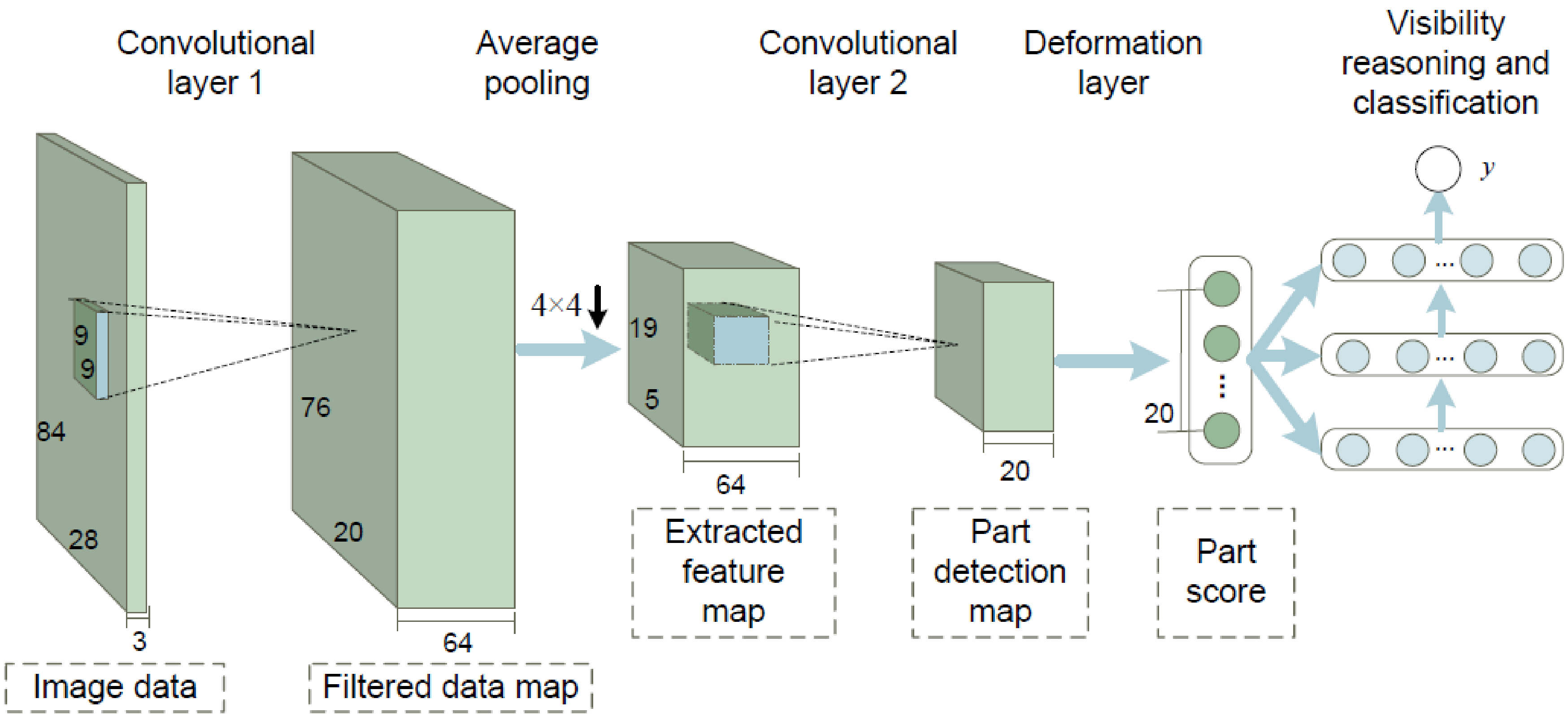
Deep learning is a framework/language but not a black-box model

Its power comes from joint optimization and increasing the capacity of the learner



- 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.

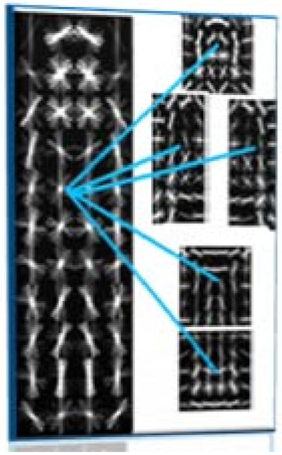
Our Joint Deep Learning Model



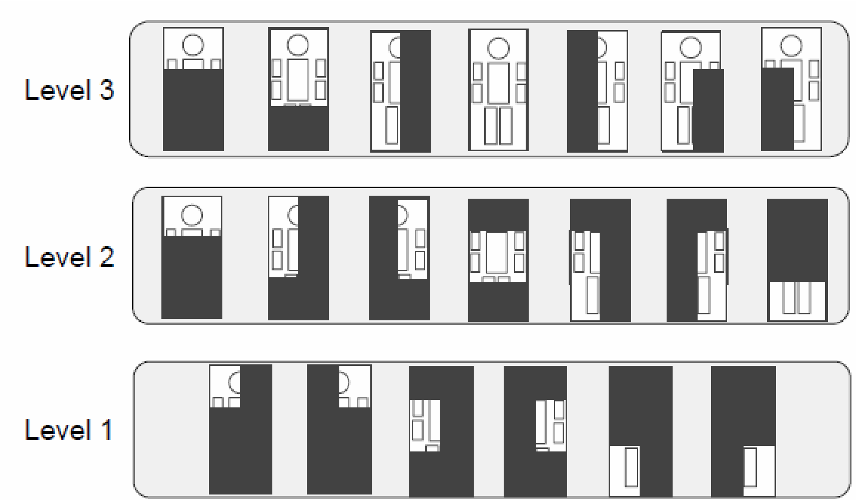
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling Part Detectors

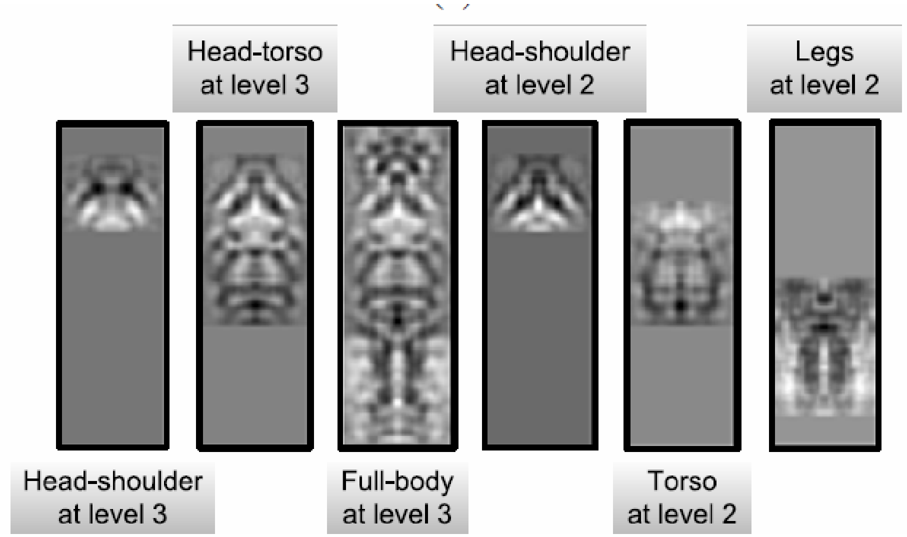
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

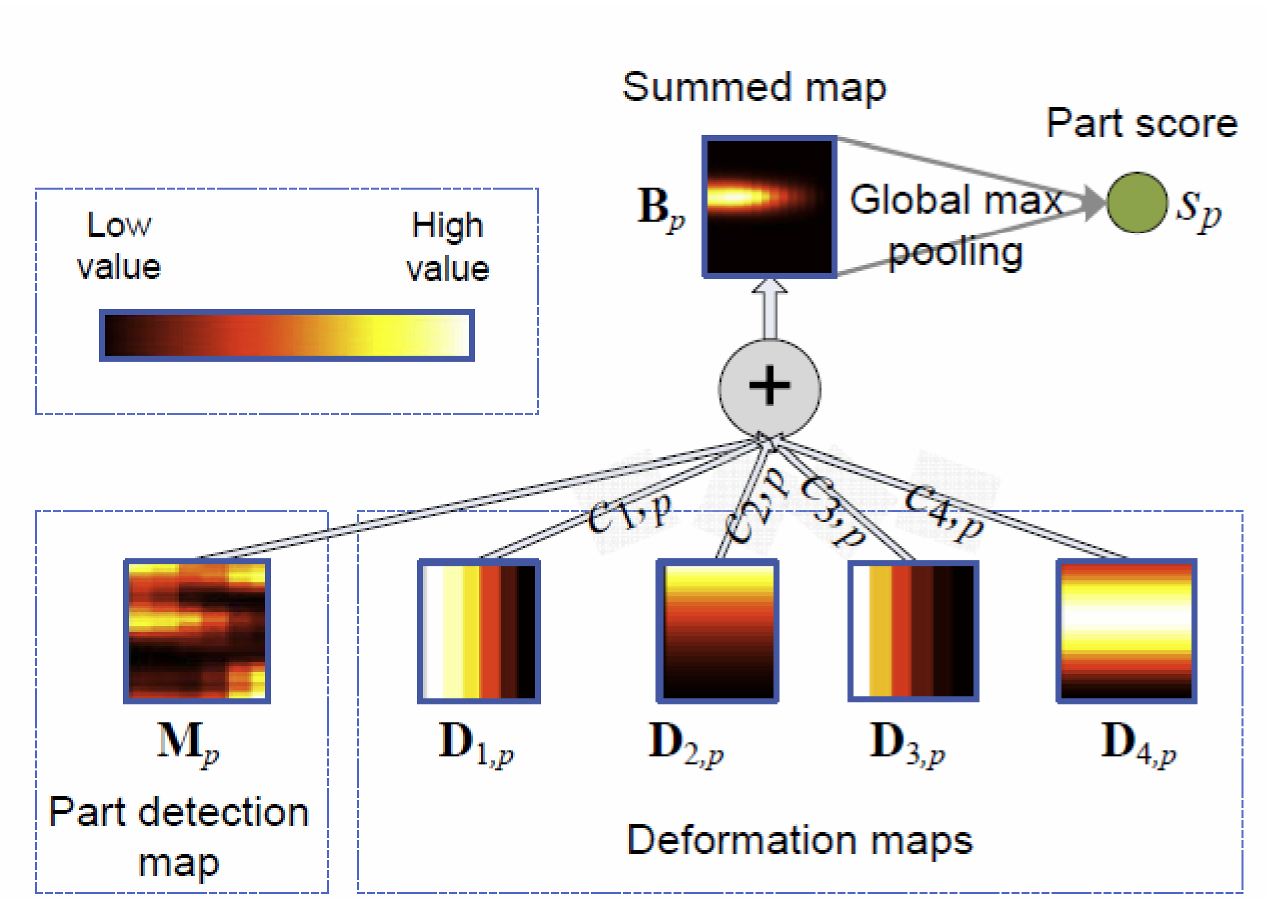


Part models

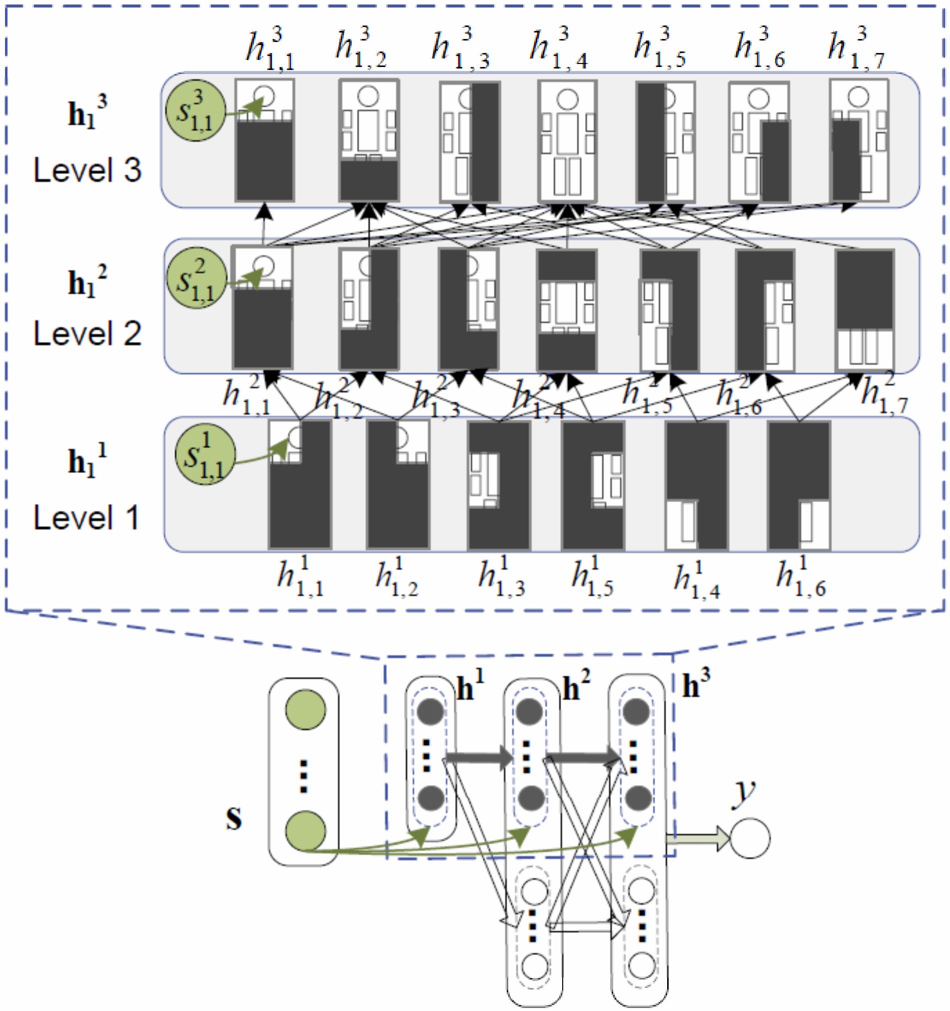


Learned filtered at the second convolutional layer

Deformation Layer



Visibility Reasoning with Deep Belief Net

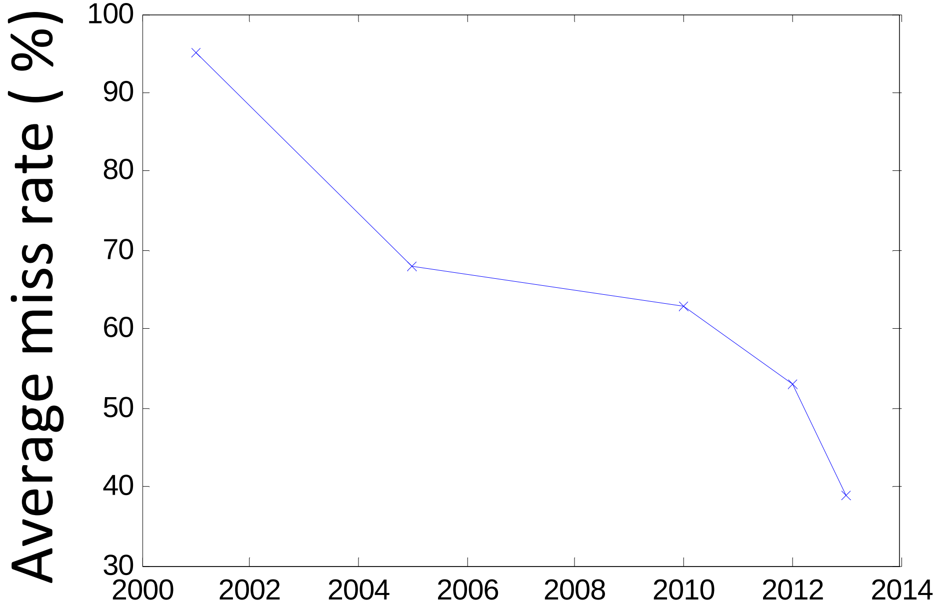


$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

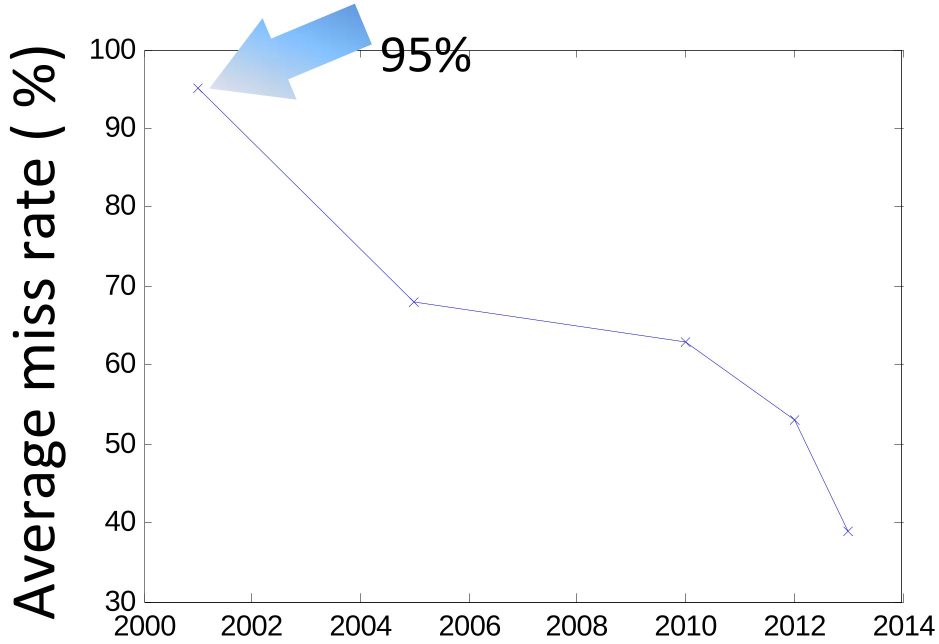
Experimental Results

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



Experimental Results

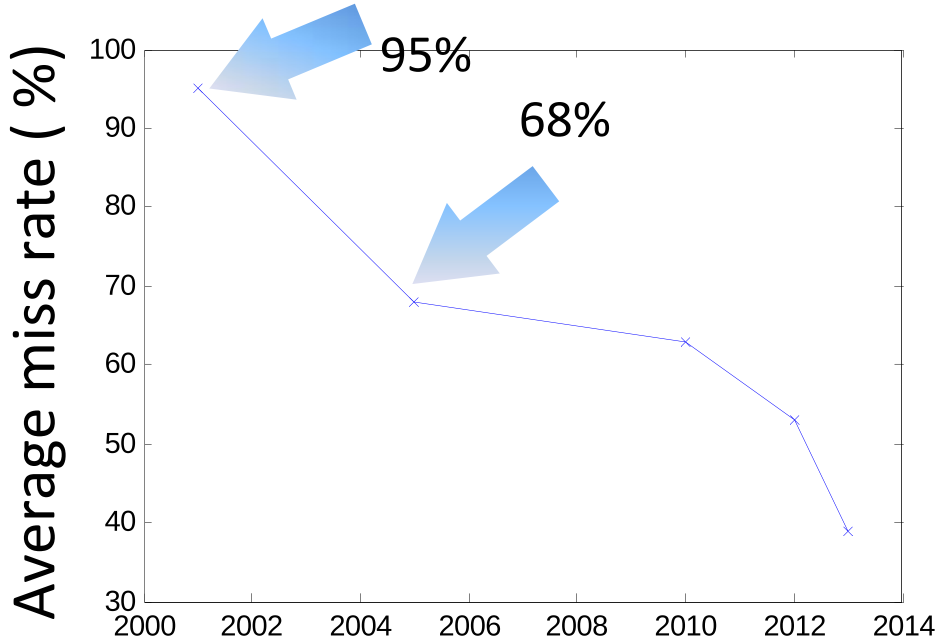
- Caltech – Test dataset (largest, most widely used)



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

Experimental Results

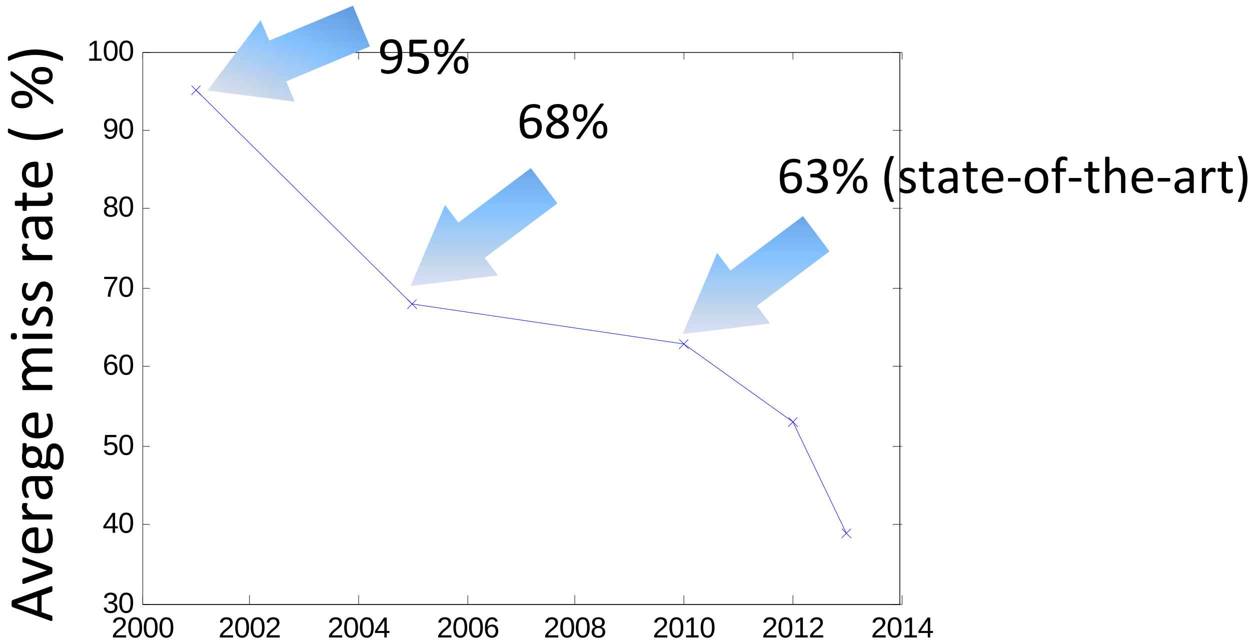
- Caltech – Test dataset (largest, most widely used)



Histograms of oriented **gradients** for **human detection**
N Dalal, B Triggs - … and Pattern Recognition, 2005. CVPR 2005 …, 2005 - ieeexplore.ieee.org
… We study the issue of feature sets for **human detection**, showing that lo- cally normalized **Histogram** of Oriented **Gradient** (HOG) de- scriptors provide excellent performance relative to other ex- isting feature sets including wavelets [17,22]. **…**
Cited by 5438 Related articles All 106 versions Import into BibTeX More

Experimental Results

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



Object detection with discriminatively trained part-based models

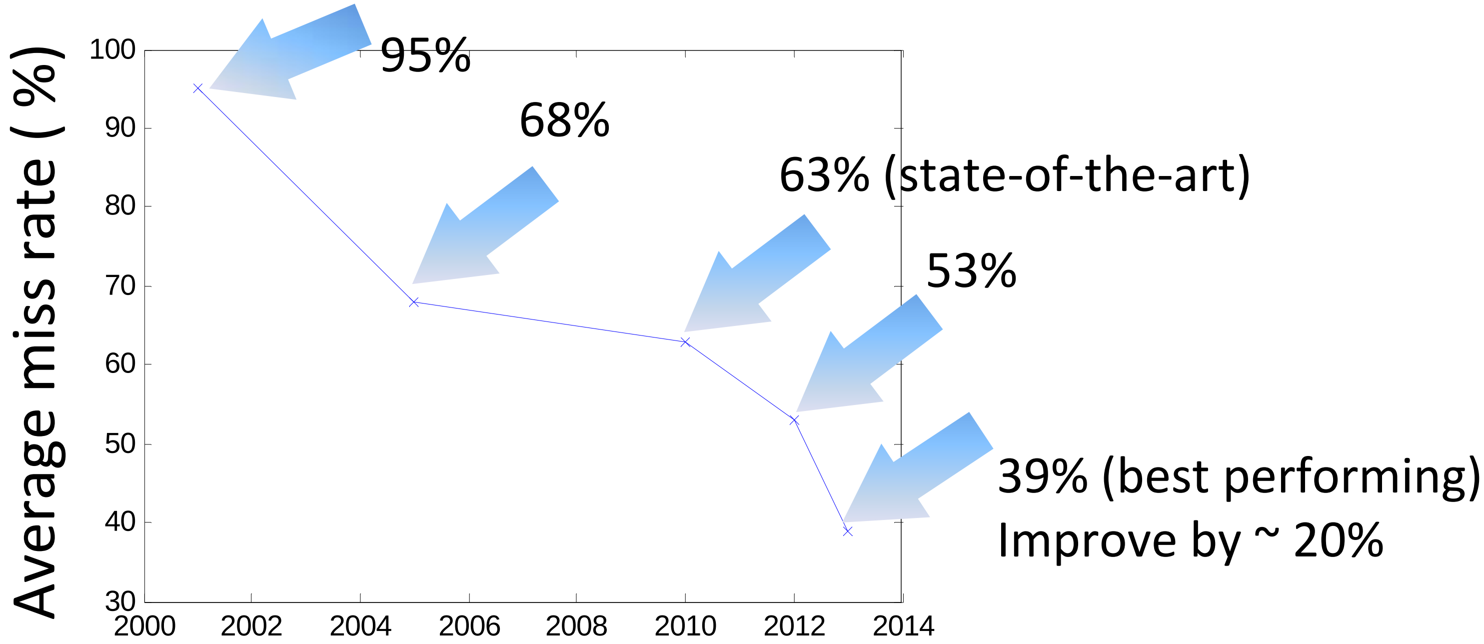
PF Felzenszwalb, RB Girshick… - Pattern Analysis and …, 2010 - ieeexplore.ieee.org

Abstract We describe an **object detection** system **based** on mixtures of multiscale deformable **part models**. Our system is able to represent highly variable **object** classes and achieves state-of-the-art results in the PASCAL **object detection** challenges. While **...**

Cited by 964 Related articles All 43 versions Import into BibTeX More

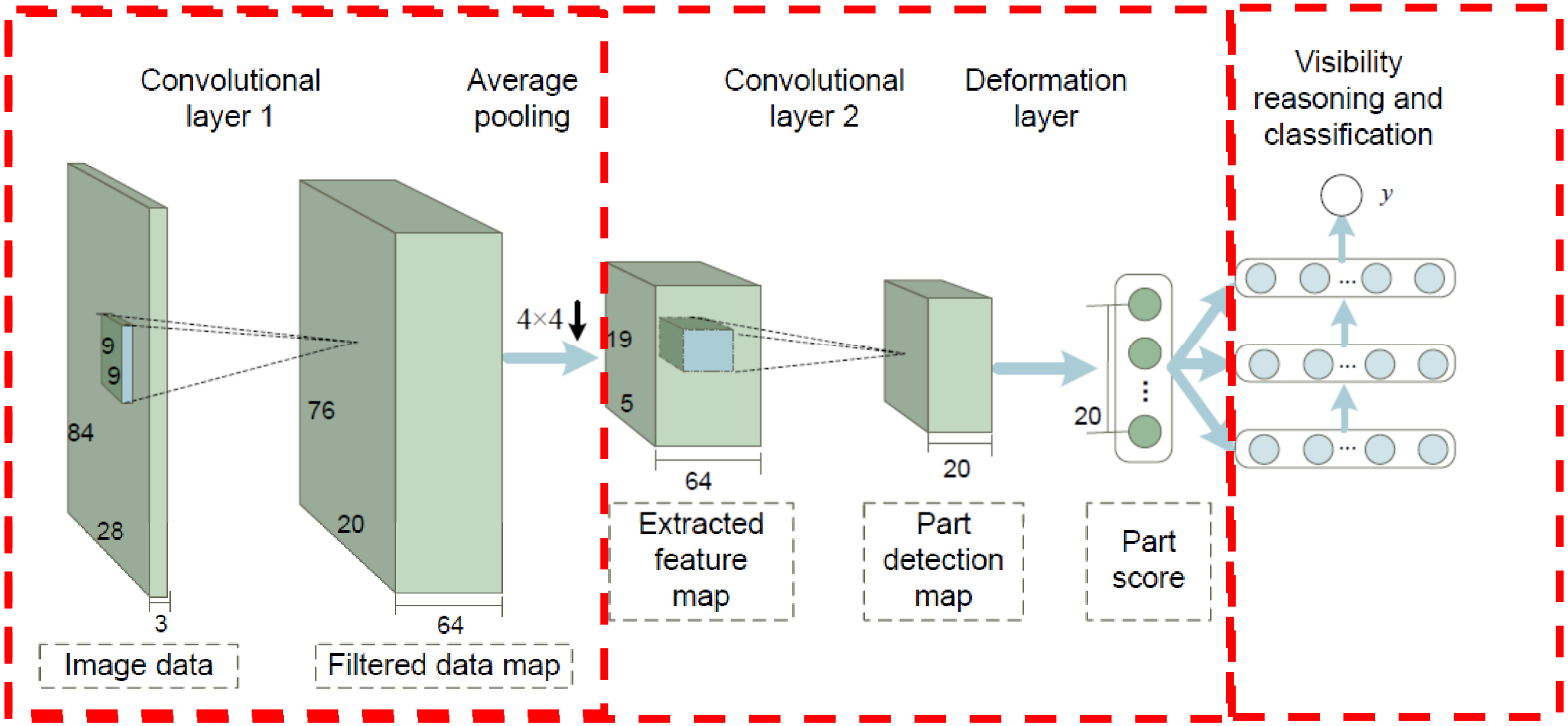
Experimental Results

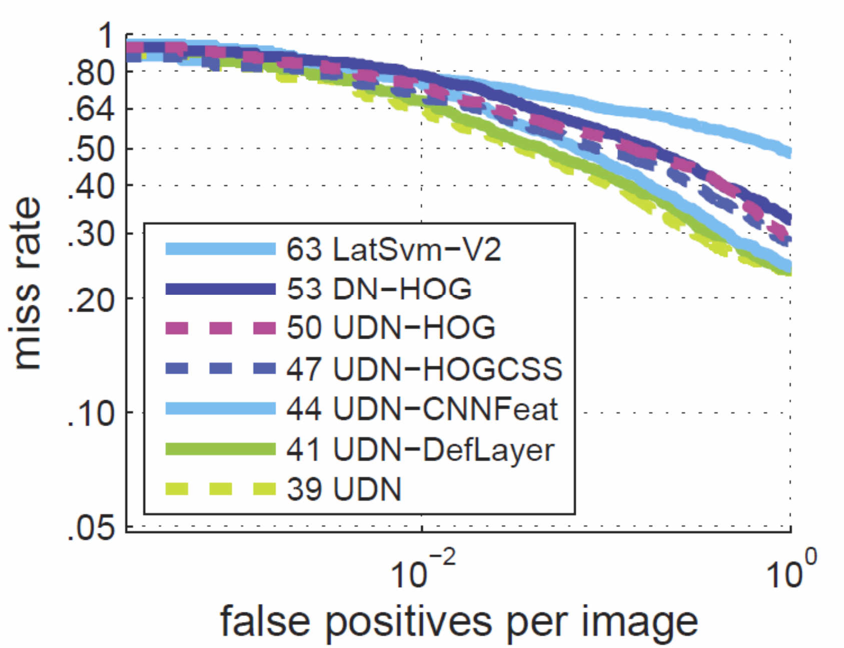
- Caltech – Test dataset (largest, most widely used)



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.





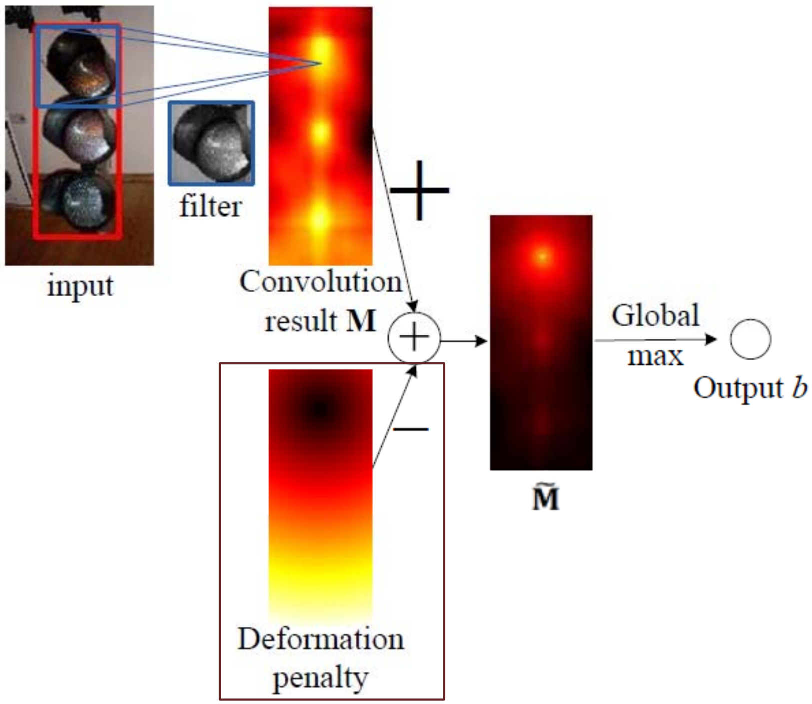
DN-HOG
UDN-HOG
UDN-HOGCSS
UDN-CNNFeat
UDN-DefLayer

Can this idea be generalized to general object detection in ImageNet?

Deformation of parts is widely observed in general objects


Deformation Layer [b]

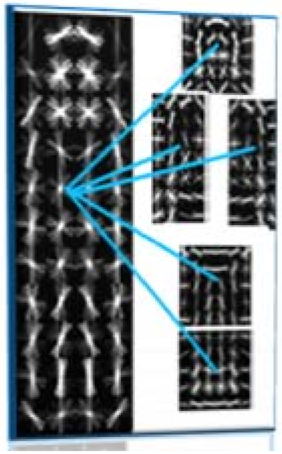
$$\mathbf{B}_p = \mathbf{M}_p + \boxed{\sum_{n=1}^{N} c_{n,p}\mathbf{D}_{n,p}} \qquad s_p = \max_{(x,y)} b_p^{(x,y)}$$



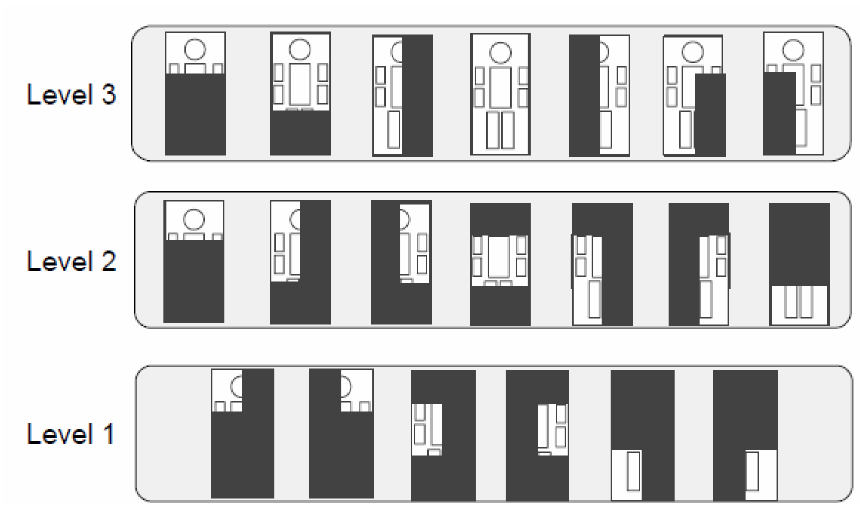
[b] Wanli Ouyang, Xiaogang Wang, "Joint Deep Learning for Pedestrian Detection ", ICCV 2013.

Modeling Part Detectors

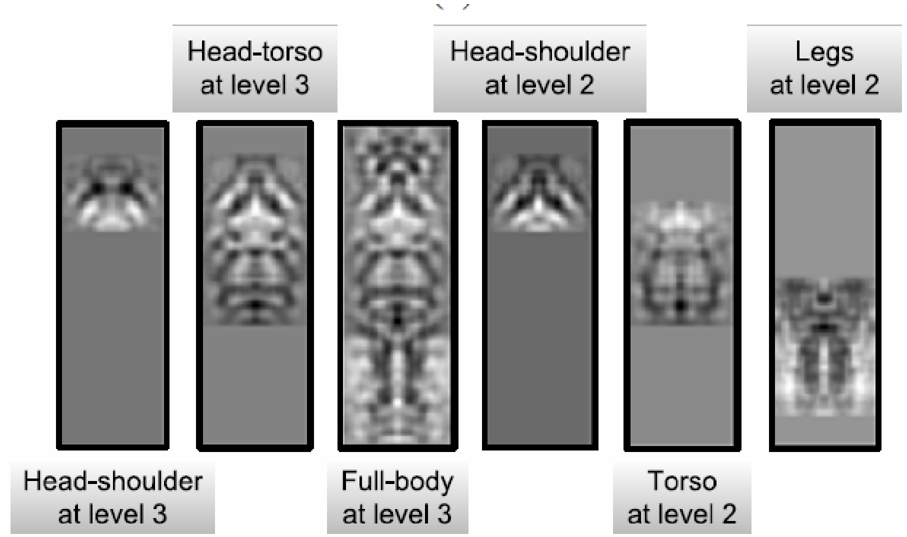
- Different parts have different sizes
- Design the filters with variable sizes



Part models learned from HOG



Part models



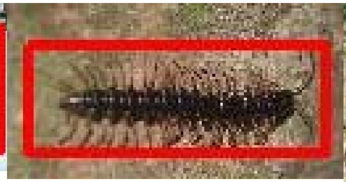
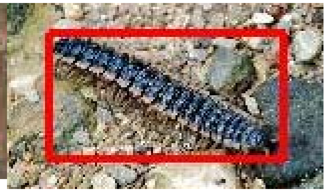
Learned filtered at the second convolutional layer

Deformation layer for repeated patterns

Pedestrian detection	General object detection
Assume no repeated pattern	Repeated patterns

Deformation layer for repeated patterns

Pedestrian detection	General object detection
Assume no repeated pattern	Repeated patterns
Only consider one object class	Patterns shared across different object classes



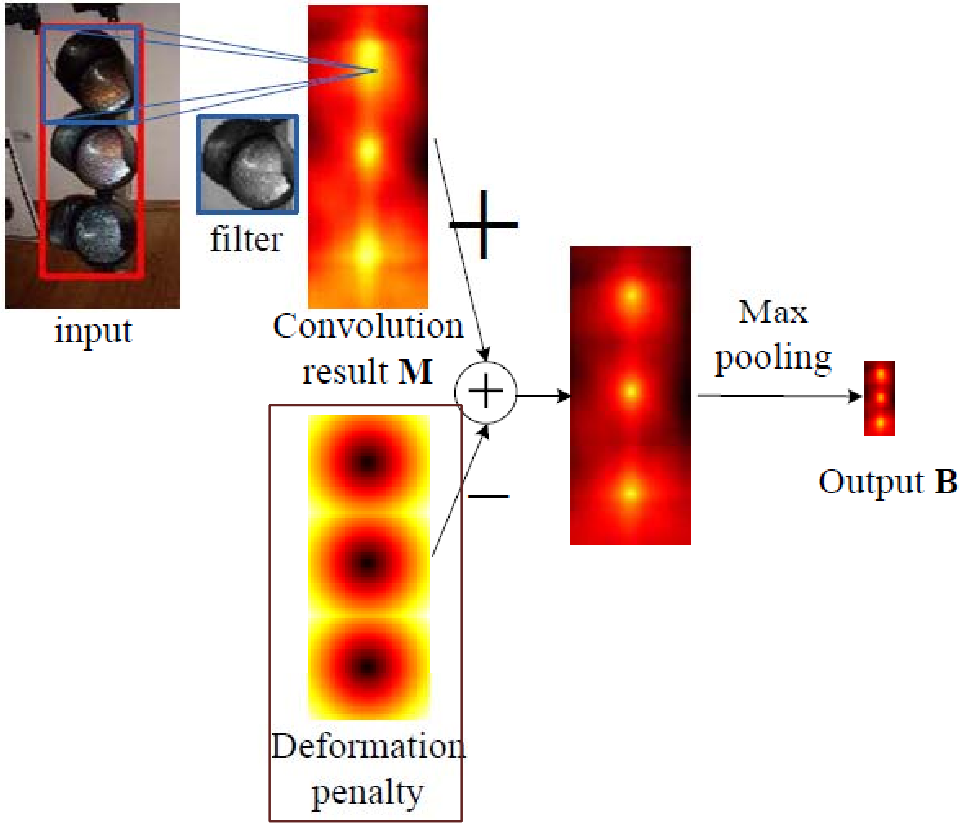


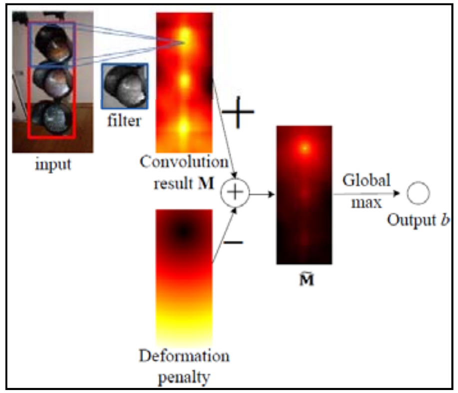


Deformation constrained pooling layer

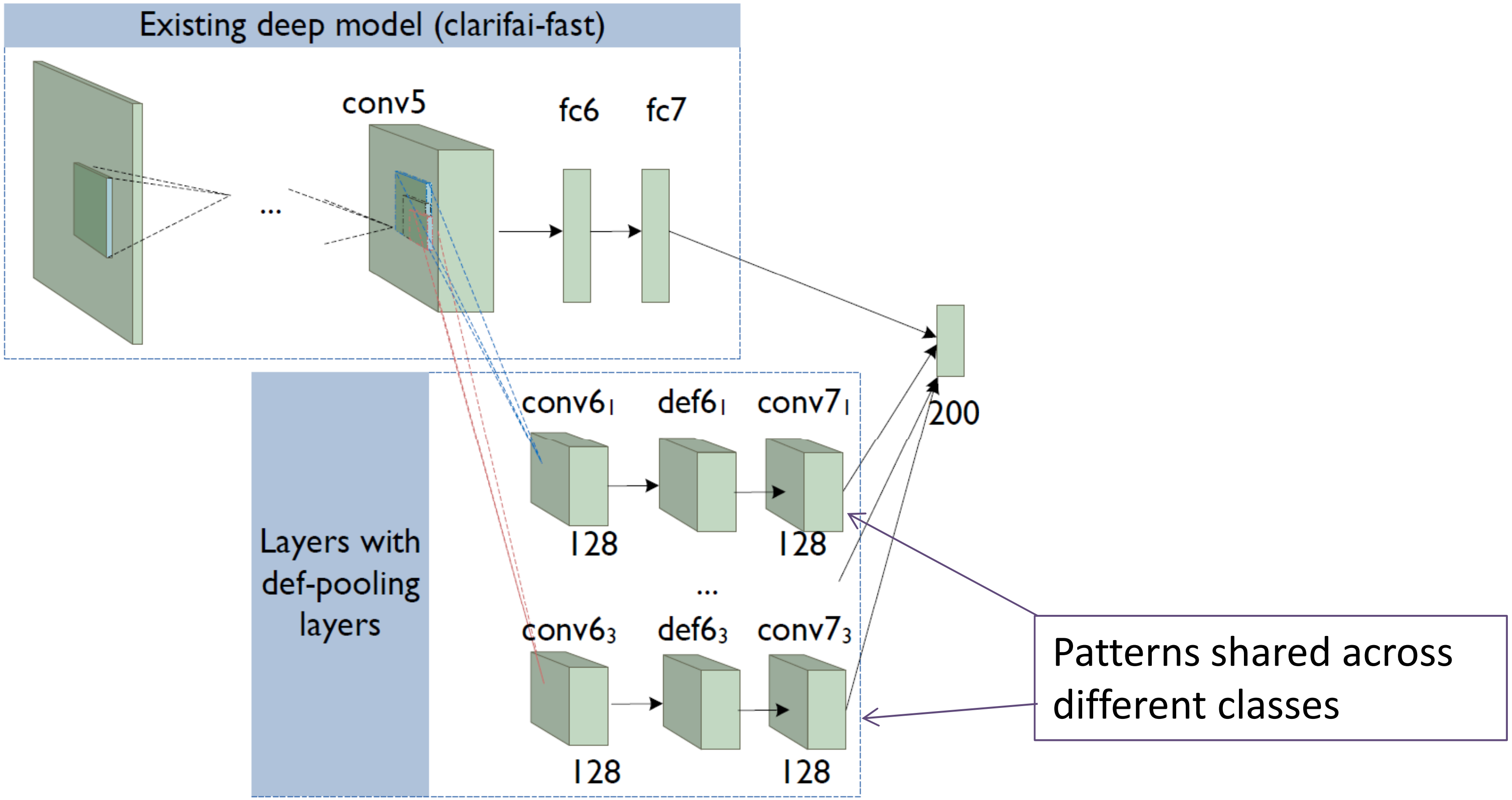
Can capture multiple patterns simultaneously

$$b^{(x,y)} = \max_{i,j \in \{-R,\cdots,R\}} \{m^{(k_x \cdot x+i, k_y \cdot y+j)} - \boxed{\sum_{n=1}^{N} c_n d_n^{i,j}}\},$$





Our deep model with deformation layer



Training scheme	Cls+Det	Loc+Det	Loc+Det
Net structure	AlexNet	Clarifai	Clarifai+Def layer
Mean AP on val2	0.299	0.360	0.385

- ImageNet 2014 – object detection challenge

	GoogLeNet (Google)	DeepID-Net (CUHK)	DeepInsight	UvA-Euvision	Berkley Vision	RCNN
Model average	0.439	**0.439**	0.405	n/a	n/a	n/a
Single model	0.380	**0.427**	0.402	0.354	0.345	0.314

W. Ouyang et al. "DeepID-Net: deformable deep convolutional neural networks for object detection", CVPR, 2015

Our understanding of deep learning

- Most two important operations (filtering and pooling) have been widely used in computer vision
- Expect other domain knowledge can inspire new layers such as deformation-pooling

Many important ideas in object detection can be generalized to deep learning…

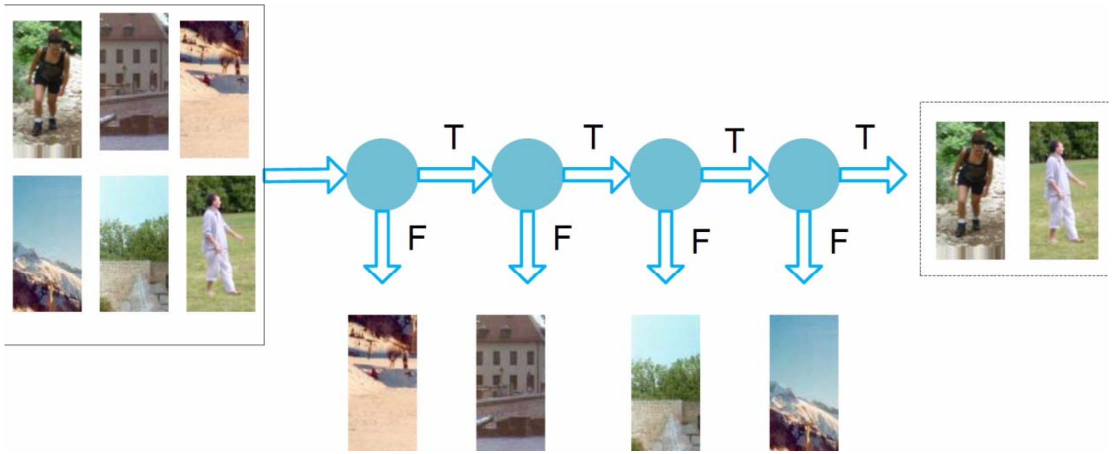
Multi-Stage Contextual Deep Learning:

- ✧ **Simulate cascaded detector and contextual boost**
- ✧ **Train different detectors for different types of samples**
- ✧ **Model contextual information**
- ✧ **Stage-by-stage pretraining strategies**

X. Zeng, W. Ouyang and X. Wang, "Multi-Stage Contextual Deep Learning for Pedestrian Detection," ICCV 2013

Cascaded Classifiers

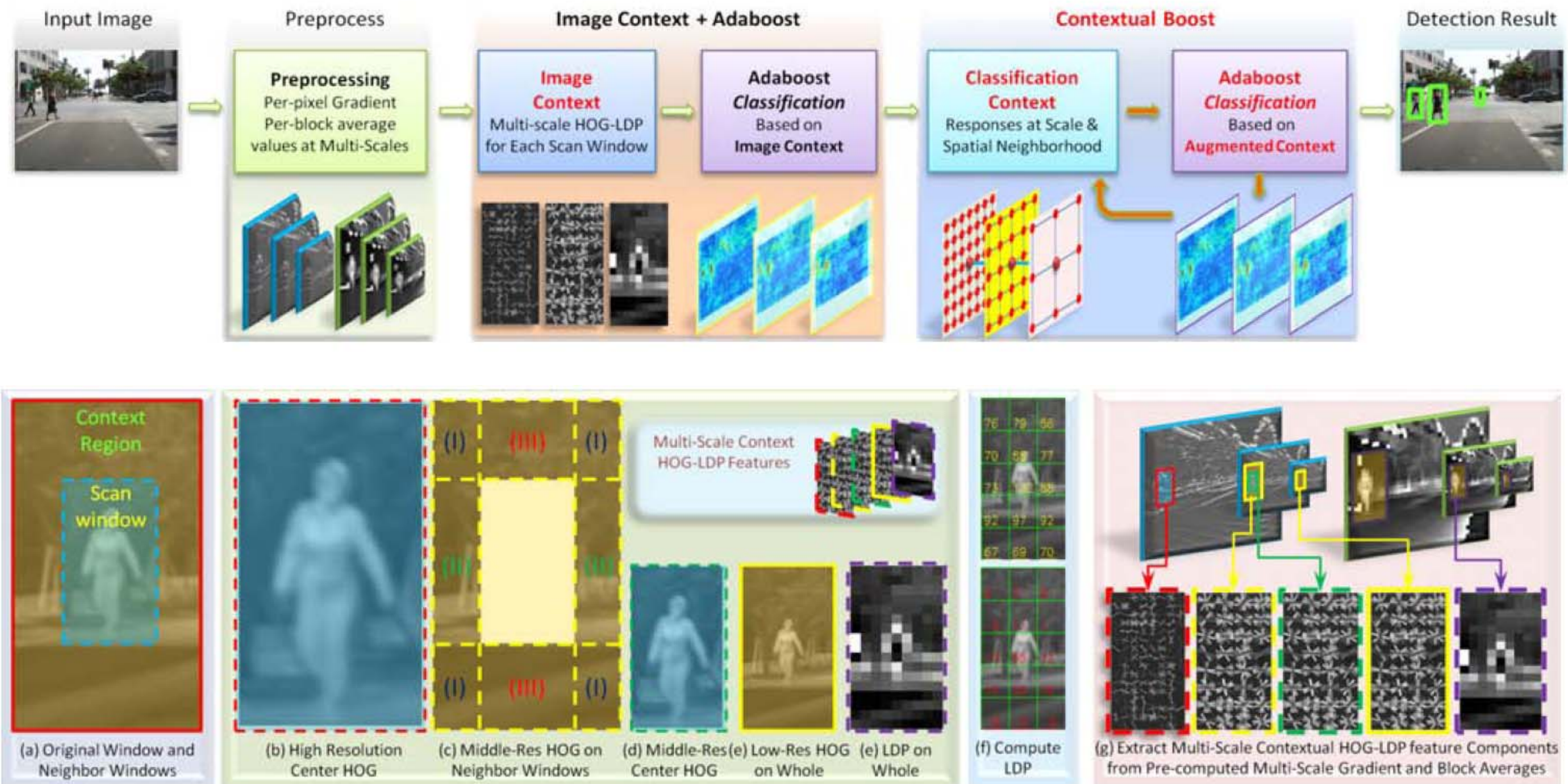
- The classifier of each stage deals with a specific set of samples
- The score map output by one classifier can serve as contextual information for the next classifier



Conventional cascaded classifiers for detection

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

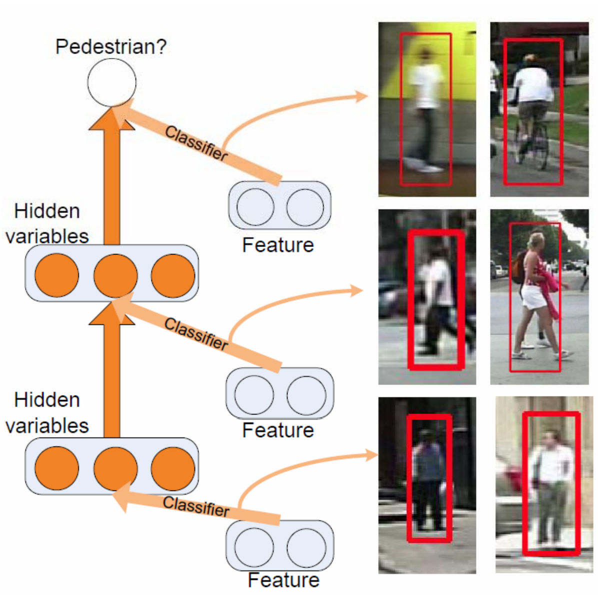
Contextual Boost

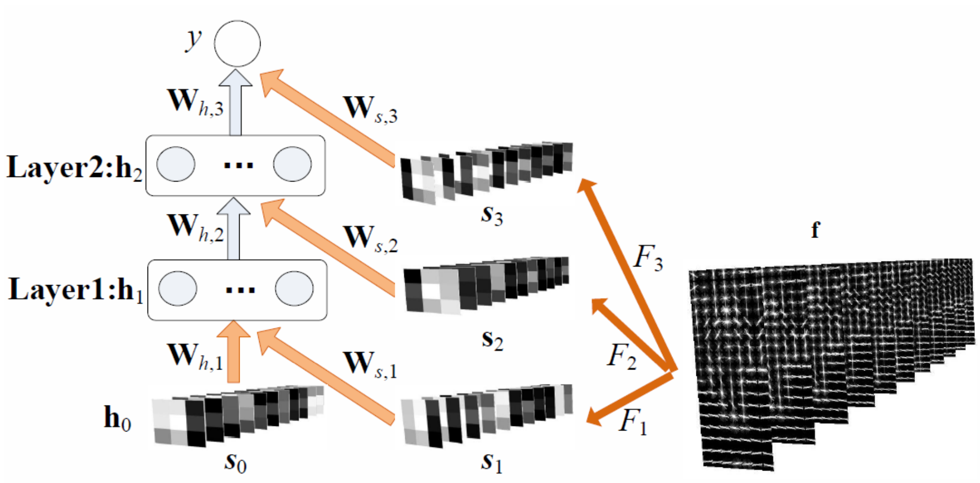


Y. Ding and J. Xiao, "Contextual Boost for Pedestrian Detection," CVPR 2012

Multi-stage deep learning

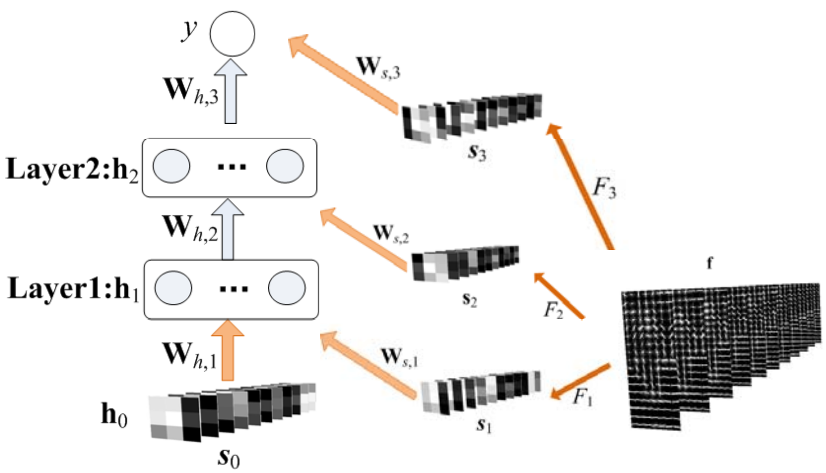
- Simulate the cascaded classifiers by mining hard samples to train the network stage-by-stage
- Cascaded classifiers are jointly optimized instead of being trained sequentially
- The deep model keeps the score map output by the current classifier and it serves as contextual information to support the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





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 sample 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



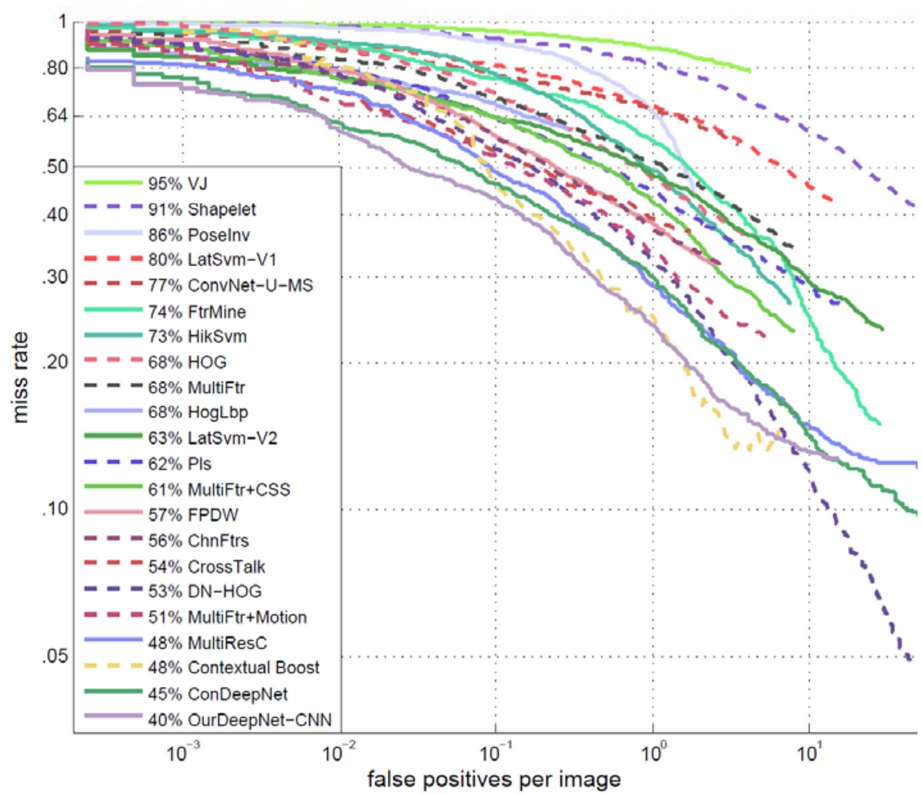
Log error function:

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

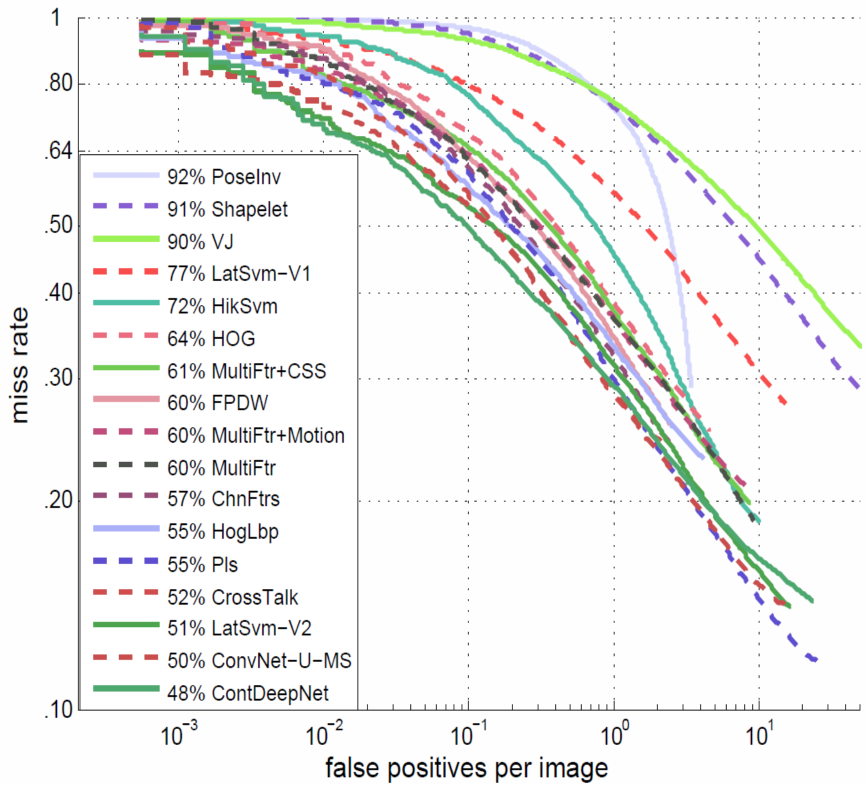
Gradients for updating parameters:

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

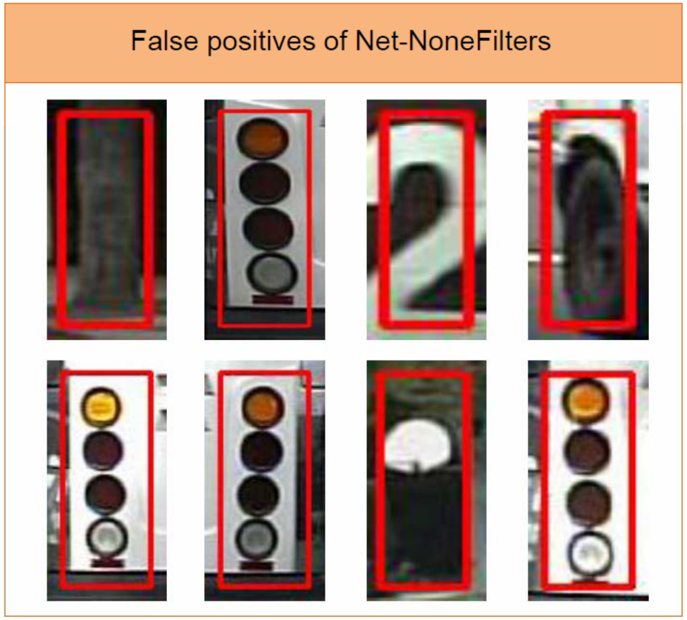
Experimental Results

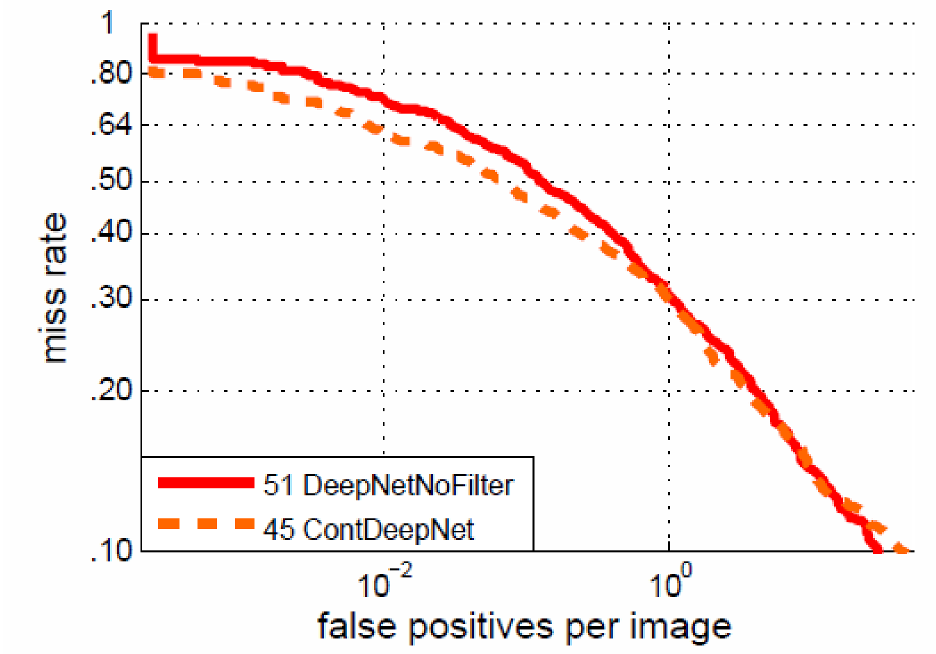


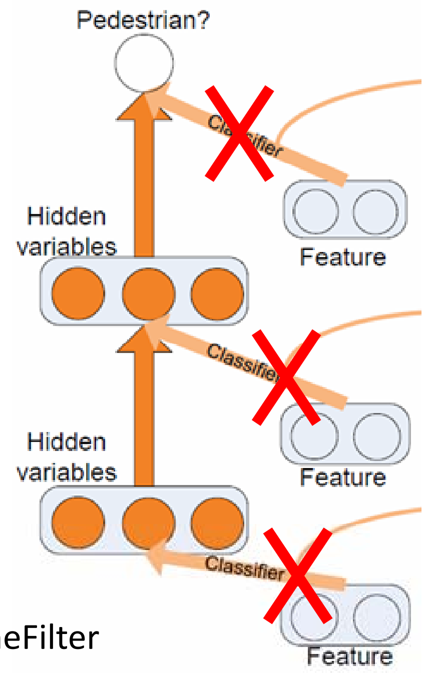
Caltech



ETHZ

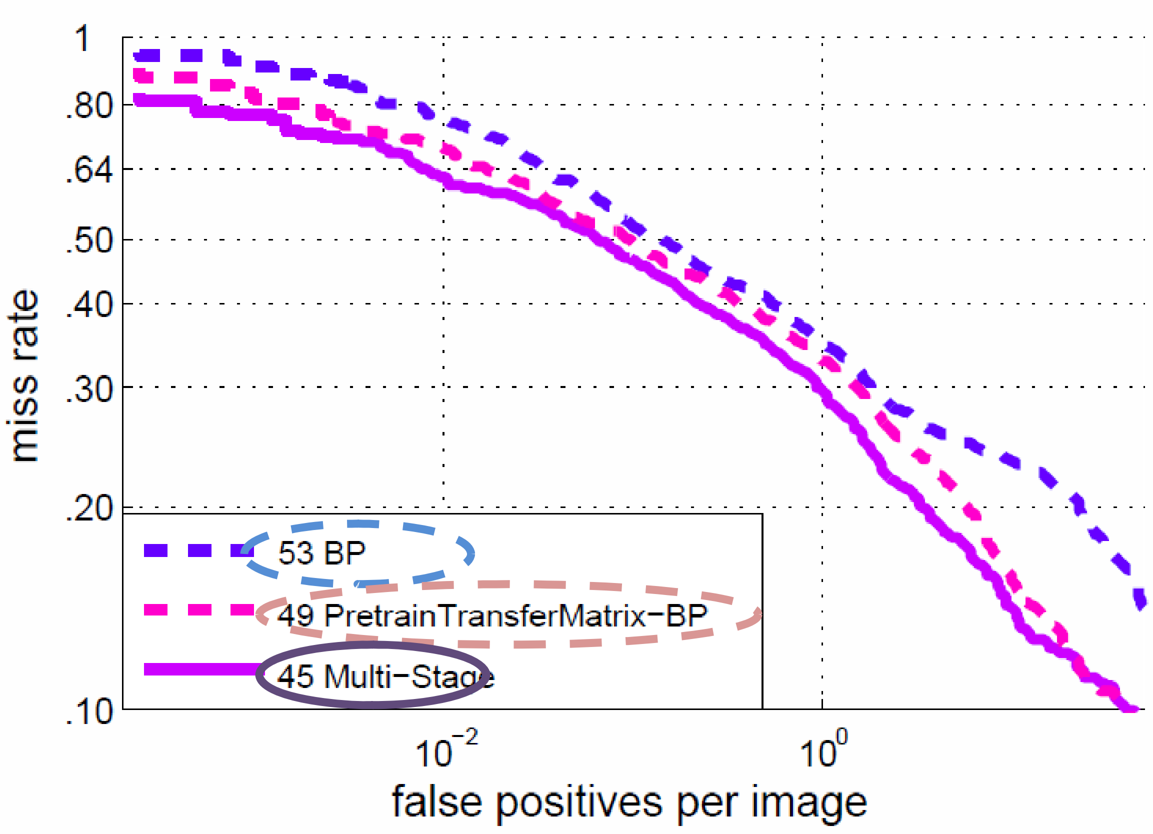






DeepNetNoneFilter

Comparison of Different Training Strategies



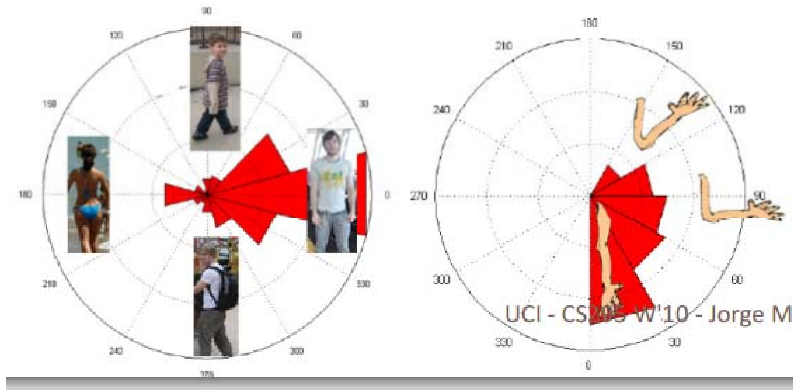
Network-BP: use back propagation to update all the parameters without pre-training
PretrainTransferMatrix-BP: the transfer matrices are unsupervised pertrained, and then all the parameters are fine-tuned
Multi-stage: our multi-stage training strategy

Switchable Deep Network

- ✧ **Use mixture components to model complex variations of body parts**
- ✧ **Use salience maps to depress background clutters**
- ✧ **Help detection with segmentation information**

P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable Deep Network for Pedestrian Detection", CVPR 2014

Poselet: modeling mixture components of body parts



L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations,” ICCV 2009

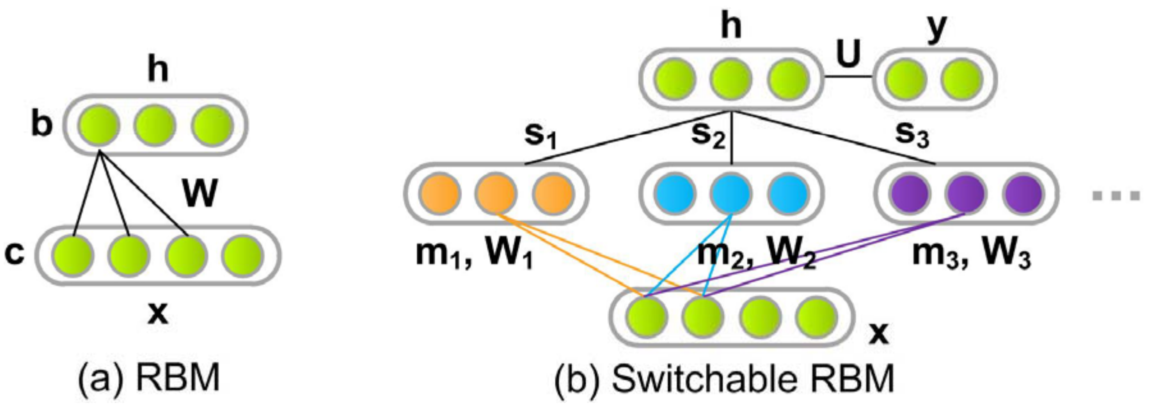
Switchable Deep Network for Pedestrian Detection

- *Background clutter* and large variations of pedestrian appearance.
- **Proposed Solution.** A Switchable Deep Network (SDN) for learning the foreground map and removing the effect background clutter.

Switchable Deep Network for Pedestrian Detection

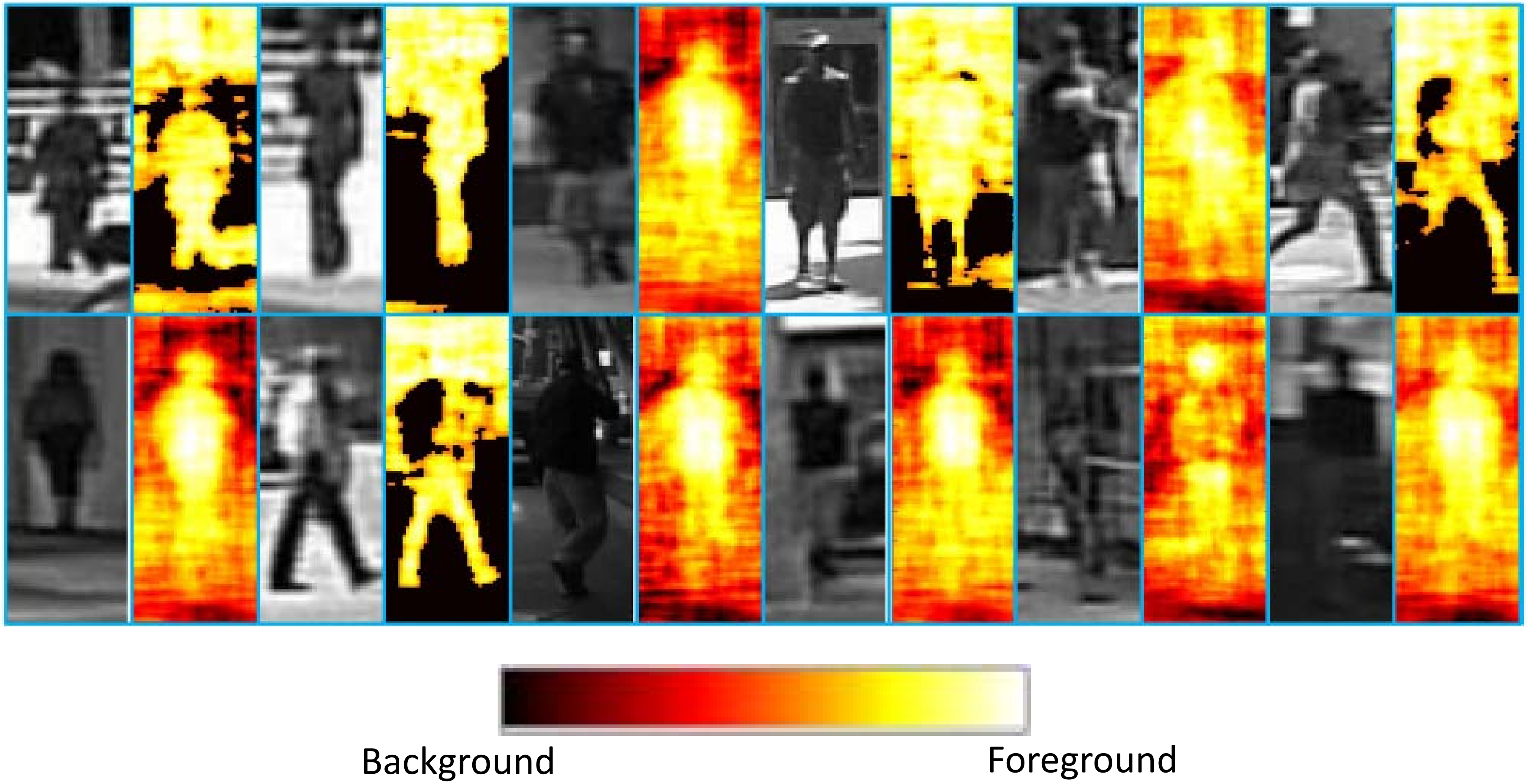
- Switchable Restricted Boltzmann Machine

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}, \mathbf{s}, \mathbf{m}; \Theta) = -\sum_{k=1}^{K} s_k \mathbf{h}_k^T (\mathbf{W}_k (\mathbf{x} \circ \mathbf{m}_k) + \mathbf{b}_k) - \sum_{k=1}^{K} s_k \mathbf{c}_k^T (\mathbf{x} \circ \mathbf{m}_k) - \mathbf{y}^T \mathbf{U} \sum_{k=1}^{K} s_k \mathbf{h}_k - \mathbf{d}^T \mathbf{y},$$

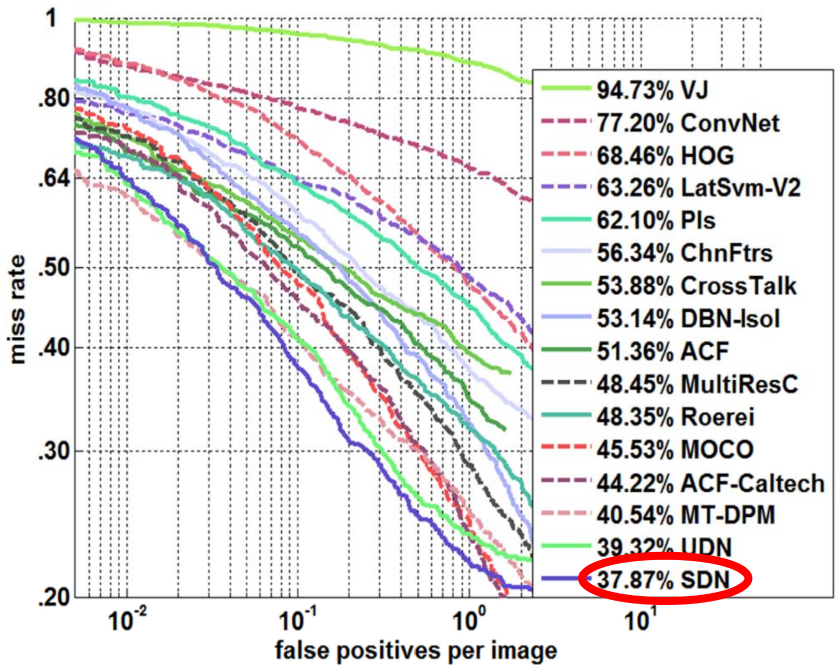


Switchable Deep Network for Pedestrian Detection

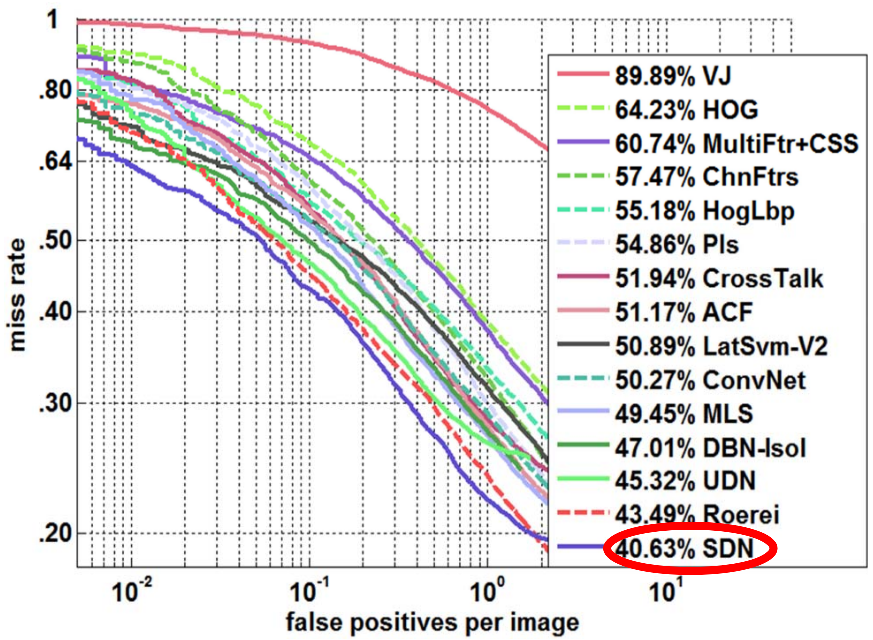
- Switchable Restricted Boltzmann Machine



Switchable Deep Network for Pedestrian Detection



(a) Performance on Caltech Test



(b) Performance on ETH

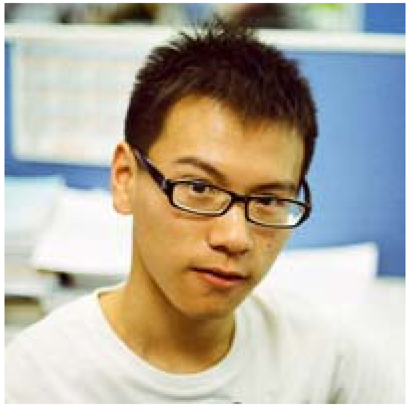
Deep Learning for Face Recognition

The projects started from December of 2012

DeepID

Yi Sun

MVP



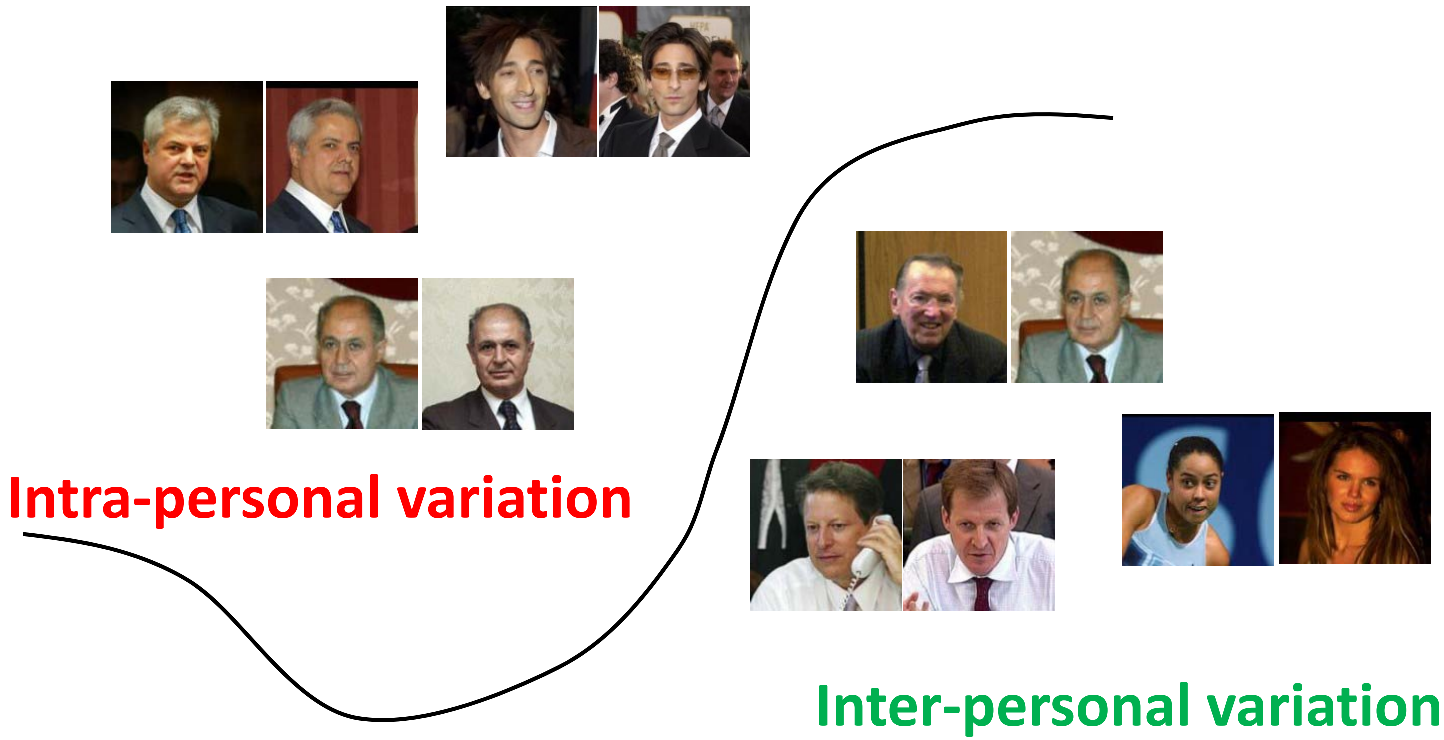
Zhenyao Zhu

Ping Luo

We started the research on face recognition since 2012

- X. Wang and X. Tang, “Unified Subspace Analysis for Face Recognition,” ICCV 2013.
- X. Wang and X. Tang, “A Unified Framework for Subspace Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 26, No.9, pp. 1222-1228, 2004.

Eternal Topic on Face Recognition



How to separate the two types of variations?

Go Back to the Starting Point

- Linear discriminant analysis (LDA) (PAMI’97)
- Bayesian face recognition (PR’00)
- Unified subspace analysis (PAMI’04)

Linear Discriminate Analysis (PAMI'97)

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^t \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_w \mathbf{W}|}$$

$$\mathbf{S}_b = \sum n_k (\bar{\mathbf{x}}_k - \bar{x})(\bar{\mathbf{x}}_k - \bar{x})^t \propto \sum (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{k'})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{k'})^t$$

$$\mathbf{S}_w = \sum_k \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t \propto \sum_{(i,j) \in \Omega} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t$$

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} |\mathbf{W}^t \mathbf{S}_b \mathbf{W}| \quad s.t. \quad |\mathbf{W}^t \mathbf{S}_w \mathbf{W}| = 1$$

LDA seeks for linear feature mapping which maximizes the distance between class centers under the constraint what the intrapersonal variation is constant

$$\mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{W}^t \mathbf{x}_i$$

$$f^* = \arg\max_{f^*} \sum_{k,k'} |f(\bar{\mathbf{x}}_k) - f(\bar{\mathbf{x}}_{k'})|^2$$

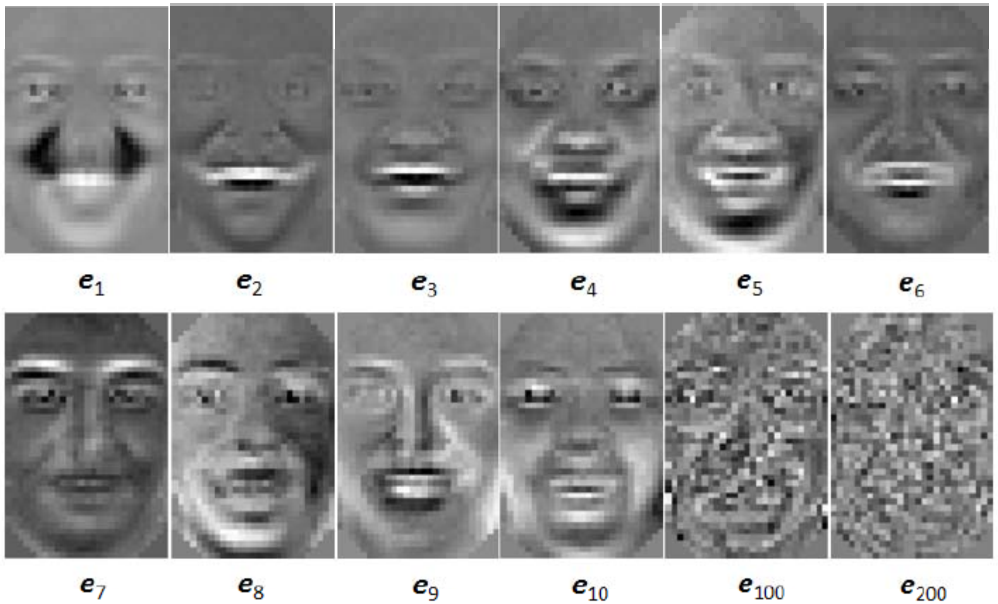
$$s.t. \sum_{(i,j)\in\Omega_I} |f(\mathbf{x}_i) - f(\mathbf{x}_j)|^2 = 1$$

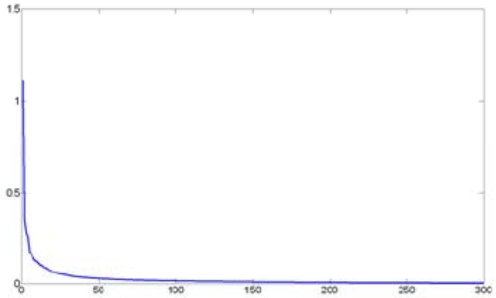
Bayesian Face Recognition (PR'00)

...

Training images

$\boldsymbol{\Delta} = \mathbf{X}_1 - \mathbf{X}_2$





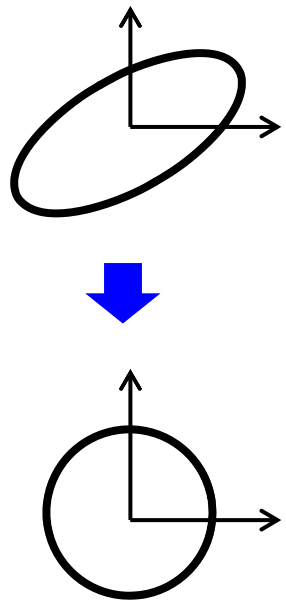
Eigenvalues

Intrapersonal subspace

$$\Delta_k = \mathbf{x}_{new} - \bar{\mathbf{x}}_k$$

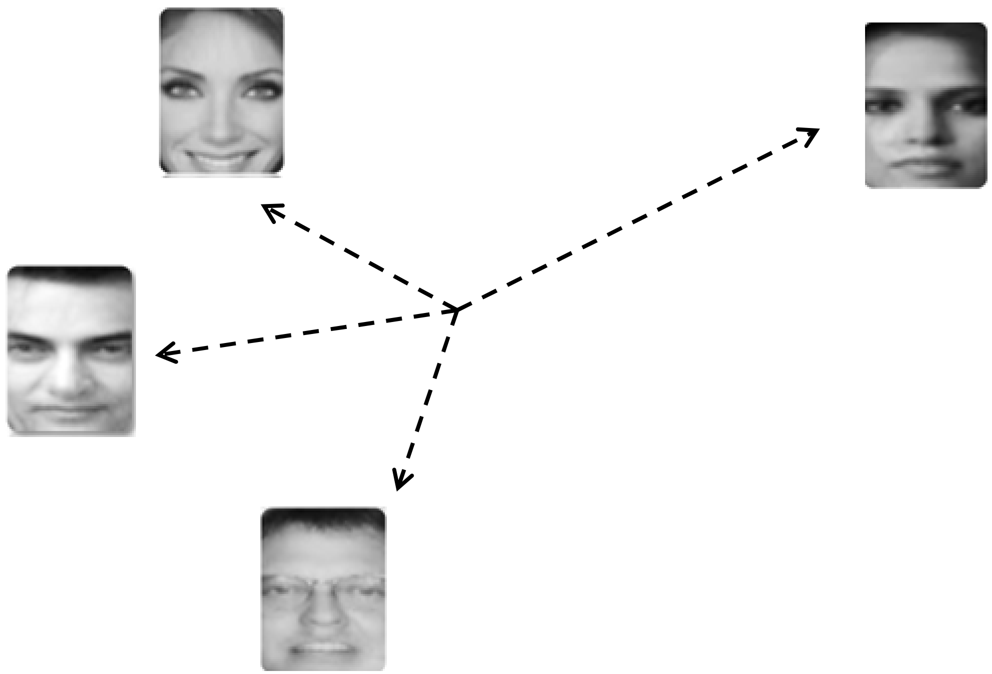
$$y_{ki} = \mathbf{e}_i^t(\mathbf{x}_{new} - \bar{\mathbf{x}}_k)$$

$$r^2(\Delta_k) = \sum_{i=1}^{d'} y_{ki}^2 / \lambda_i$$



Scatter Class Centers

- Further do PCA on class centers after reducing intrapersonal variation with whitening

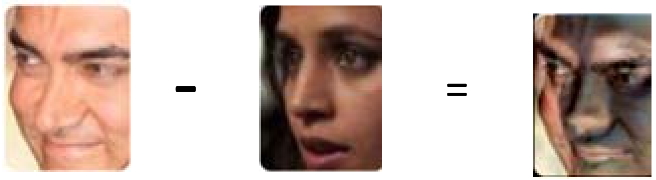


Unified Subspace Analysis (PAMI'04)

- Eigenface: PCA on images to reduce dimensionality and remove noise (when later steps increase intrapersonal difference, some noise could be magnified in wrong directions)
- Bayesianface: PCA on intrapersonal difference vectors to extract the patterns of intrapersonal variations, and depress them by dividing eigenvalues
- Fisherface: PCA on class centers to make them as far as possible and extract identity information

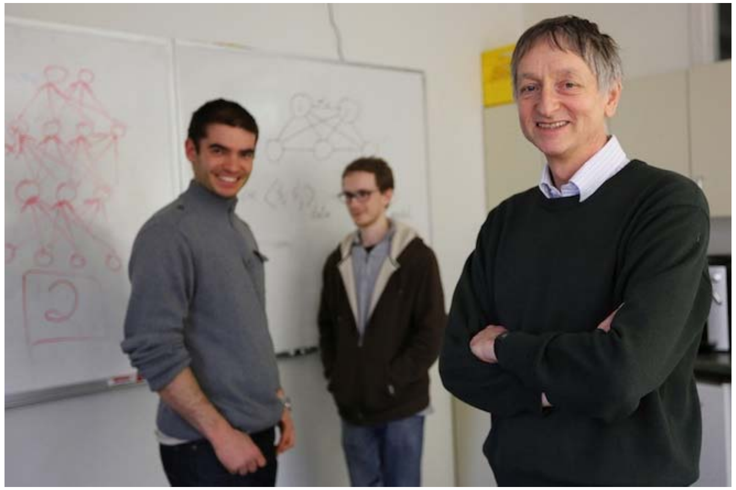
Limitations of Existing Approaches

- A lot of information has been lost when calculating the difference $\Delta = X_1 - X_2$



- Linear models with shallow structures cannot separate intra- and inter-personal variations, which are complex, nonlinear, and in high-dimensional image space

Deep Learning Won ImageNet Image Classification Challenge 2012

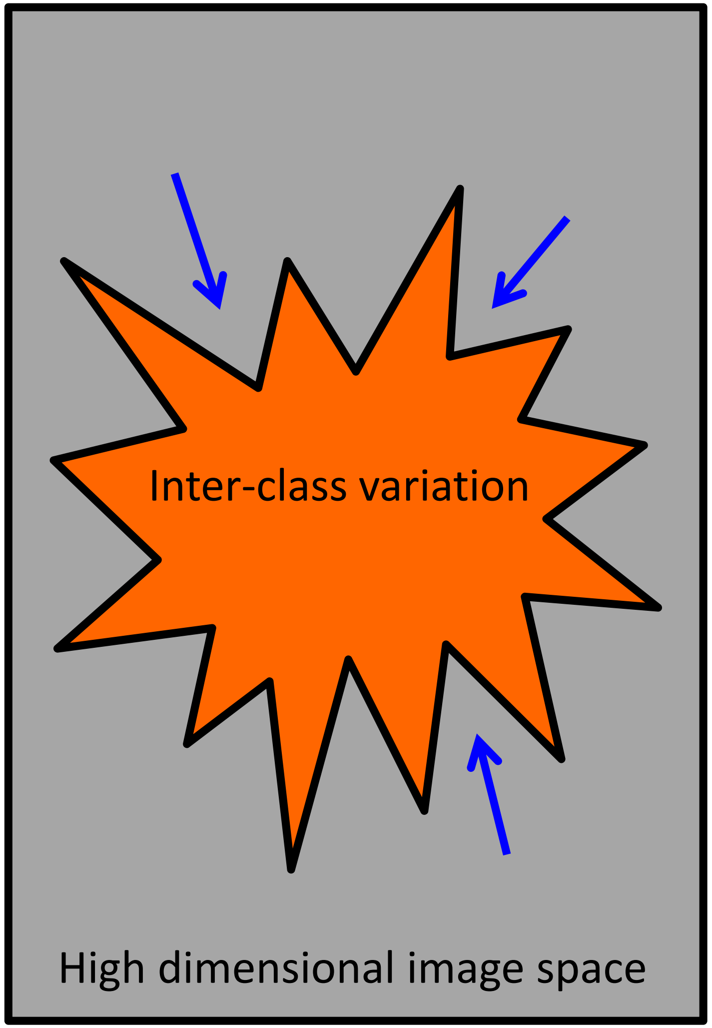


Motivated us to feed an image pair (I_1 , I_2) to CNN and train a powerful nonlinear classifier

$$S(I_1, I_2) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)} \xrightarrow{?} \textbf{CNN} (\mathbf{I_1} , \mathbf{I_2})$$

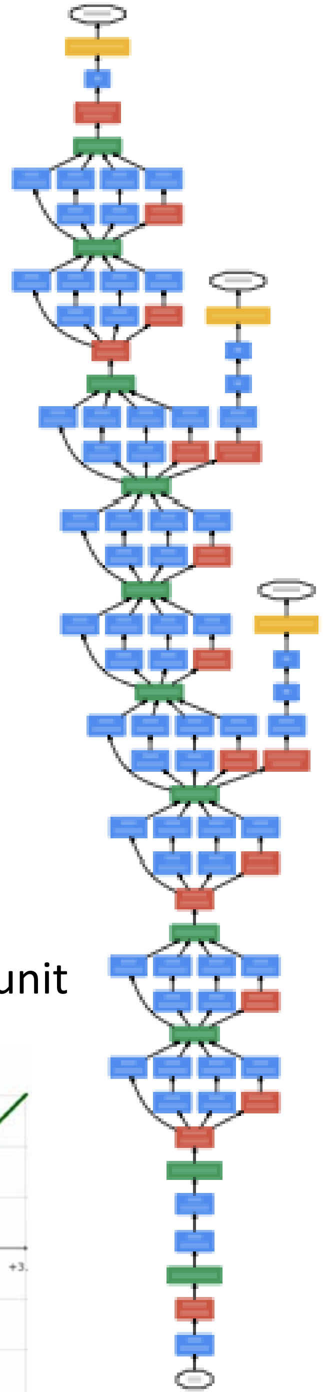
Deep Learning for Face Recognition

- Extract identity preserving features through hierarchical nonlinear mappings
- Model complex intra- and inter-personal variations with large learning capacity

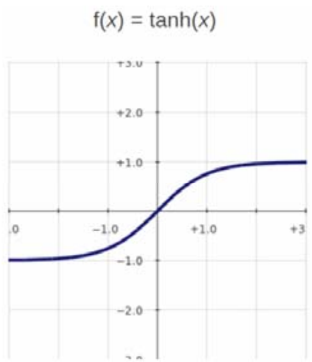


- **Linear transform**
- **Pooling**
- **Nonlinear mapping**

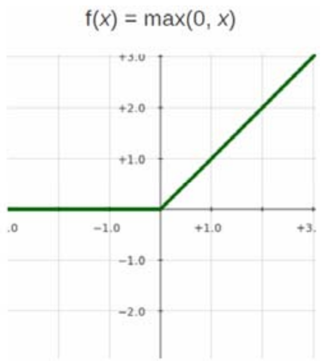
GoogleNet



Sigmoid



Rectified linear unit

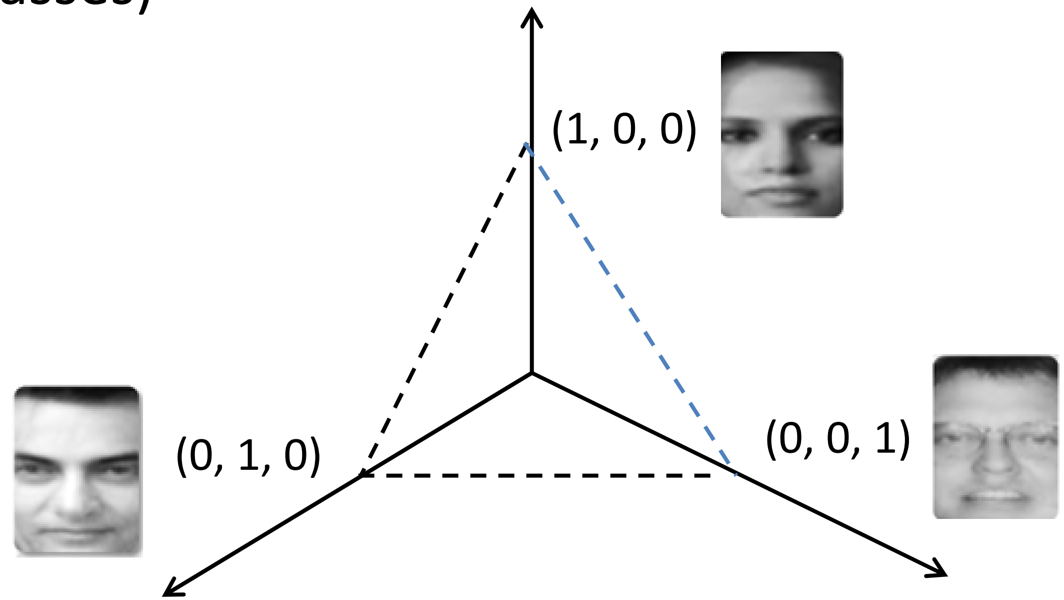


Learn Identity Features from Different Supervisory Tasks

- Face identification: classify an image into one of N identity classes
  – multi-class classification problem
- Face verification: verify whether a pair of images belong to the same identity or not
  – binary classification problem

$$S(I_1, I_2) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)} \xrightarrow{?} \textbf{CNN } (I_1 , I_2)$$

Minimize the intra-personal variation under the constraint that the distance between classes is constant (i.e. contracting the volume of the image space without reducing the distance between classes)



$$\mathbf{y} = f(\mathbf{x}); \quad g = \mathrm{softmax}()$$

$$f^* = \arg\min_f \sum_{(i,j)\in\Omega_I} ||f(\mathbf{x}_i) - f(\mathbf{x}_j)||^2$$

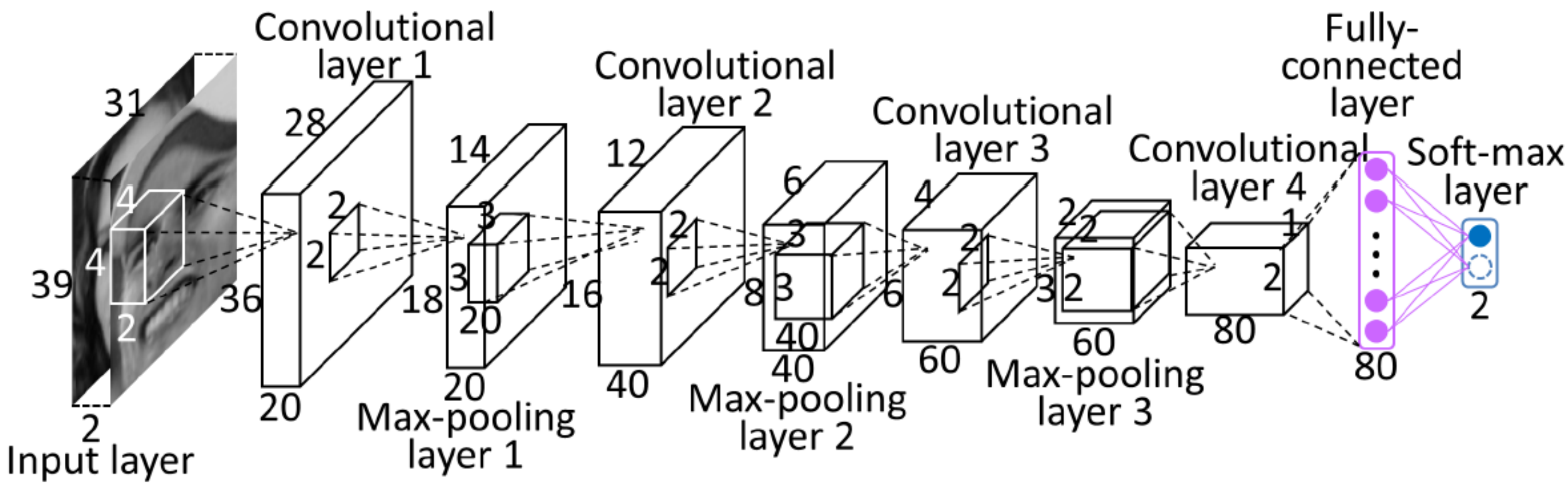
$$s.t. \ \ |g(f(\mathbf{x}_i)) - g(f(\mathbf{x}_j))| = 1, \ \ label(\mathbf{x}_i) \neq label(\mathbf{x}_j)$$

Learn Identity Features with Verification Signal

- Extract relational features with learned filter pairs

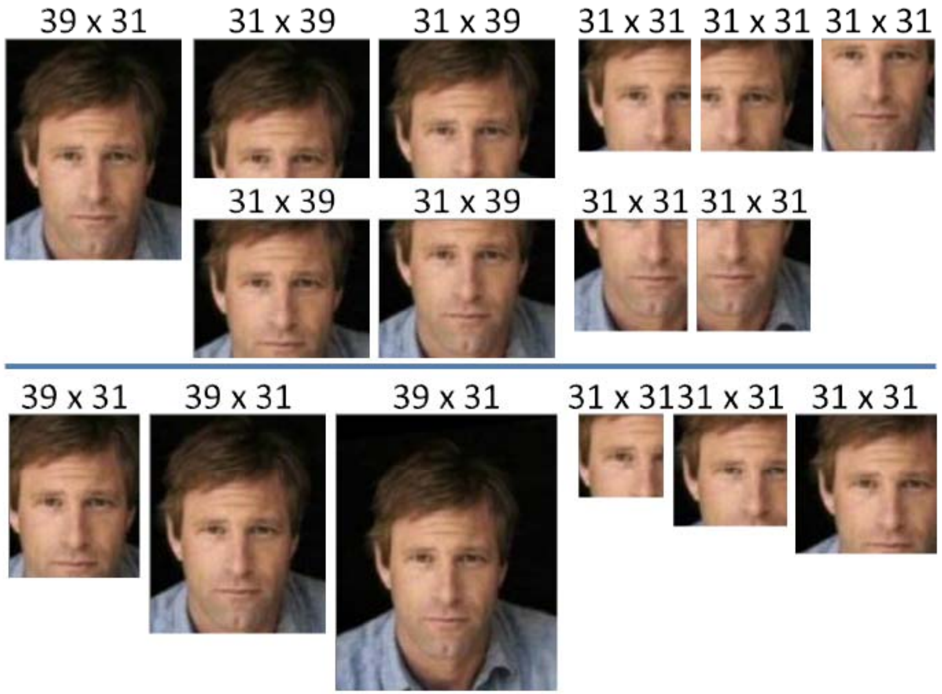
$$y^j = f\left(b^j + k^{1j} * x^1 + k^{2j} * x^2\right)$$

- These relational features are further processed through multiple layers to extract global features
- The fully connected layer can be used as features to combine with multiple ConvNets

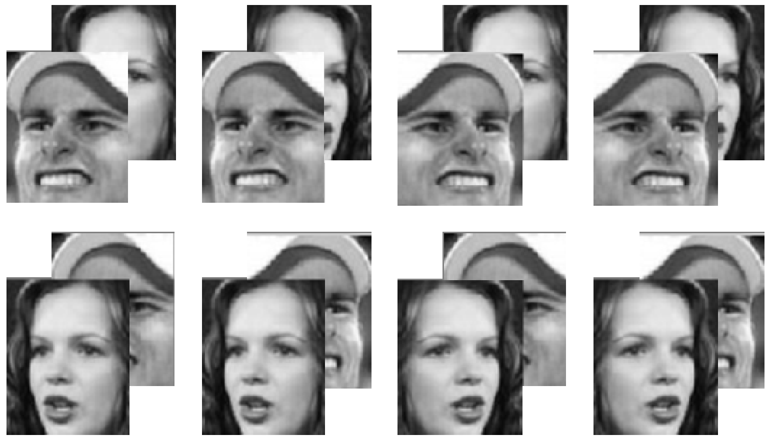


Generate Multiple CNNs

- 10 face regions, 3 scales, color/gray and 8 modes
- Base on three-point alignment

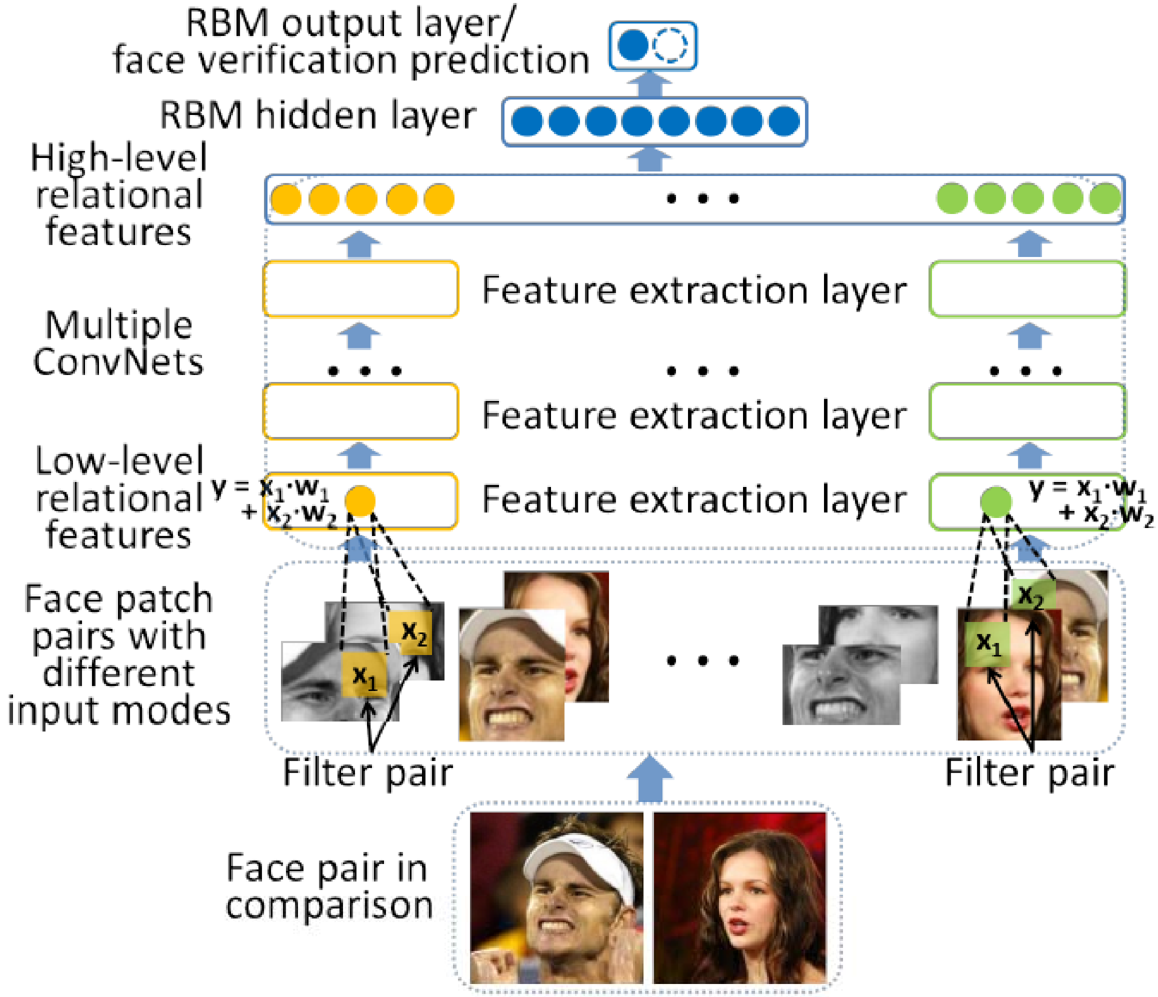


Regions and scales



modes

RBM Combines Features Extracted by Multiple ConvNets



Results on LFW

- Outside training data: the CelebFaces dataset has 87,628 face images of 5,436 celebrities. Its identities have no overlap with LFW

	hid	hid+out	out
dimension	38,400	38,880	480
each dim (%)	60.25	60.58	86.63
PCA+LDA (%)	94.55	94.42	93.41
SVM linear (%)	95.12	95.04	93.45
SVM rbf (%)	94.95	94.89	94.00
classRBM (%)	**95.56**	95.32	93.79

Taking the last hidden layer (hid) as features for combination is more effective than using the output of CNNs (out)

Results on LFW

- Fine tuning RBM and ConvNets improves the performance
- Averaging 5 RBMs (each is trained with a randomly generated training set) can improves performance

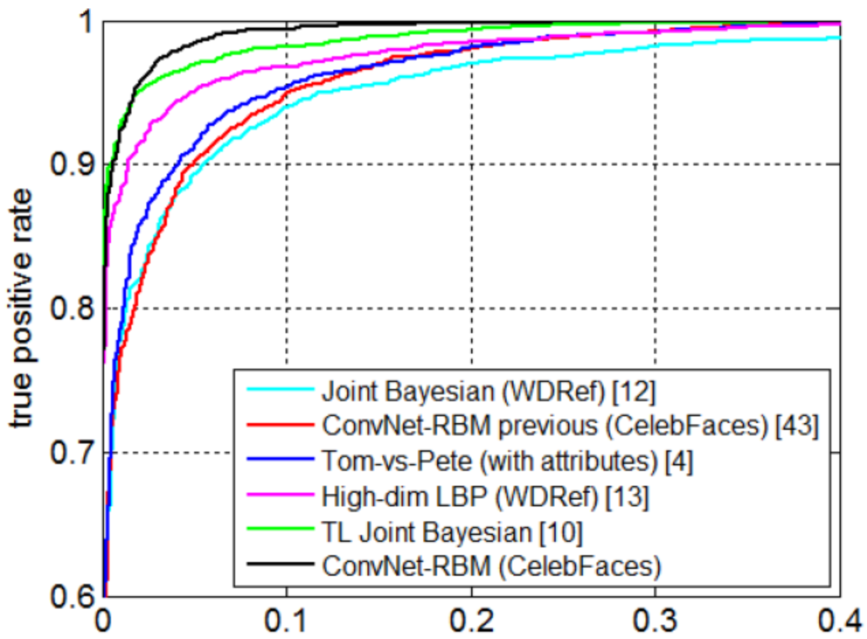
	LFW (%)	CelebFaces (%)
Single ConvNet	85.05	88.46
RBM	93.45	95.56
Fine-tuning	93.58	96.60
Model averaging	**93.83**	**97.08**

LFW: only using training images from LFW with unrestricted protocol
CelebFaces: using CelebFaces as training set without training images from LFW

Results on LFW

- Unrestricted protocol using outside training data

Method	Accuracy (%)
Joint Bayesian [12]	92.42 ± 1.08
ConvNet-RBM previous [43]	92.52 ± 0.38
Tom-vs-Pete (with attributes) [4]	93.30 ± 1.28
High-dim LBP [13]	95.17 ± 1.13
TL Joint Bayesian [10]	96.33 ± 1.08
ConvNet-RBM	$\mathbf{97.08 \pm 0.28}$



Summary of Results

- Use the last hidden layer instead of the output of CNNs as features
- Fusion of features from more face regions (CNNs) improves the performance
- Fine tuning RBM and CNNs improves performance
- Averaging the outputs of multiple RBMs improves the performance
- Drawbacks: computational cost is high and features cannot be computed offline

Features learned from a large number of classes from ImageNet has good generalization capability

The key of deep learning is to learn feature representations instead of classifiers

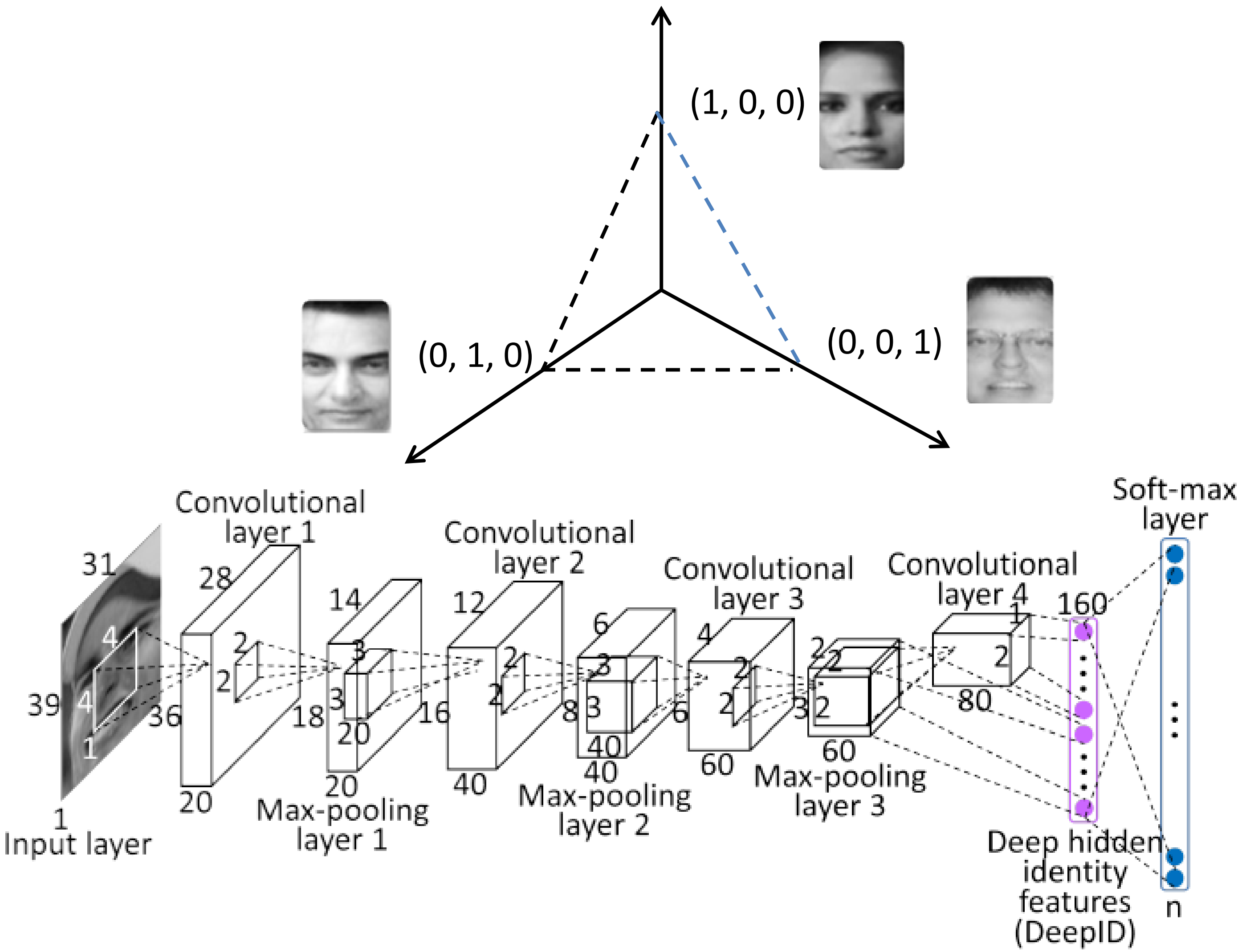
⇩

Can this idea be generalized to face recognition?

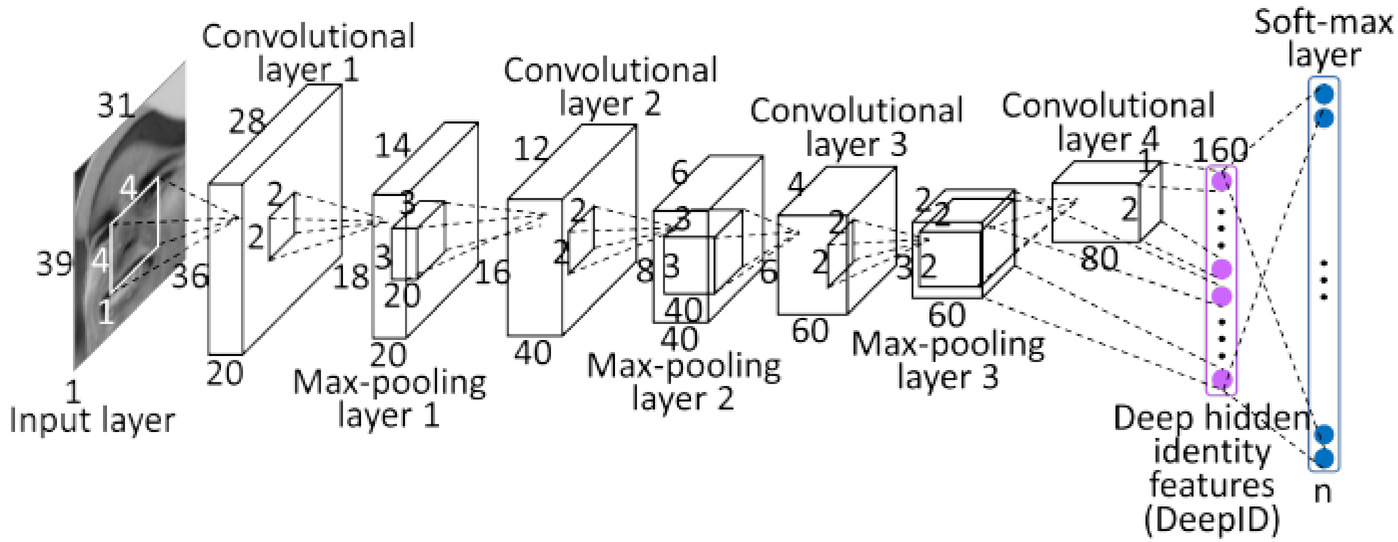
Our understanding of deep learning

- **Deeply learned features can be well generalized to other datasets and recognition tasks**
- **The generalization power increases when the supervision task is more challenging**

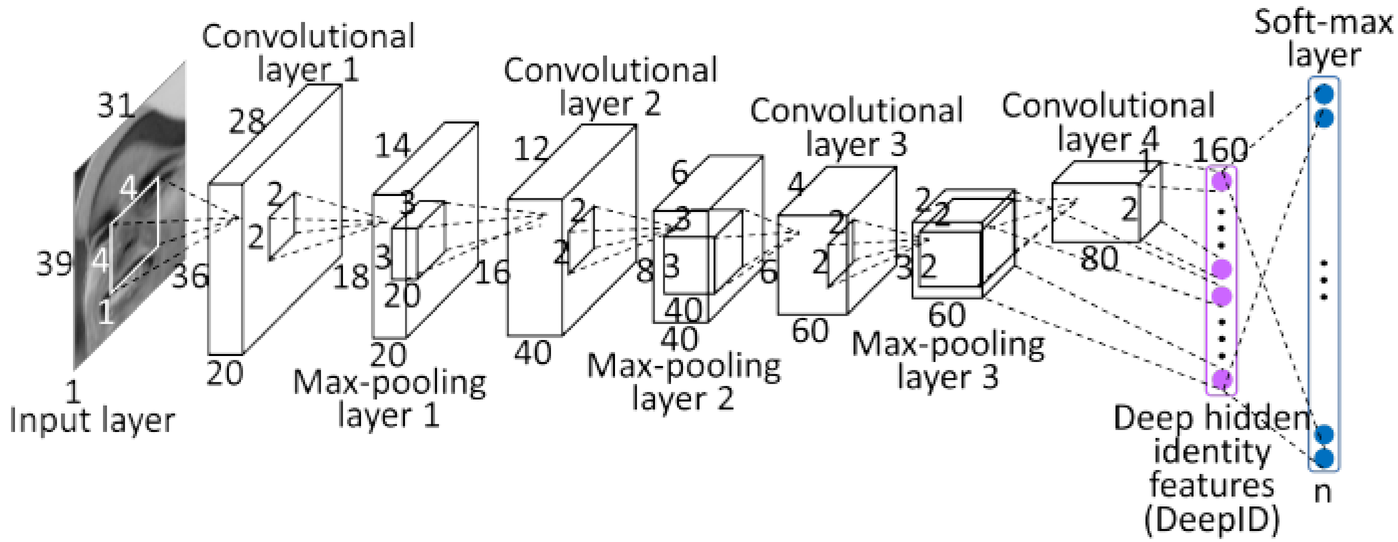
Learn Identity Features with Identification Signal



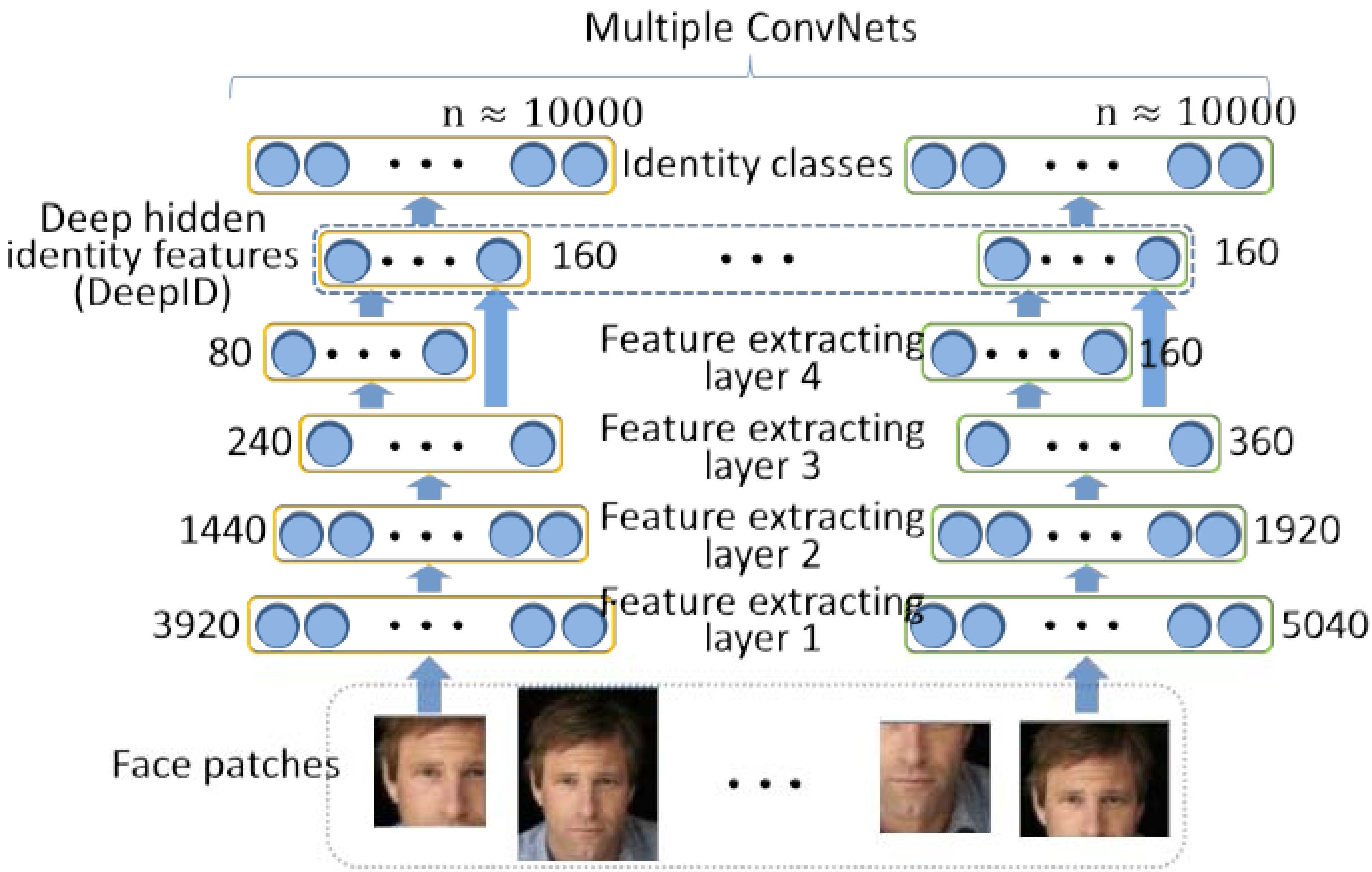
- During training, each image is classified into 10,000 identities with 160 identity features in the top layer
- These features keep rich inter-personal variations
- Features from the last two convolutional layers are effective
- The hidden identity features can be well generalized to other tasks (e.g. verification) and identities outside the training set



- High-dimensional prediction is more challenging, but also adds stronger supervision to the network
- As adding the number of classes to be predicted, the generalization power of the learned features also improves



Extract Features from Multiple ConvNets



Learn Identity Features with Identification Signal

- After combining hidden identity features from multiple CovNets and further reducing dimensionality with PCA, each face image has 150-dimenional features as signature
- These features can be further processed by other classifiers in face verification. Interestingly, we find Joint Bayesian is more effective than cascading another neural network to classify these features

Result on LFW

- We enlarge CelebFaces dataset to CelebFaces+, which include 202,599 images of 10,117 celebrities. CelebFaces+ has no overlap with LFW on identities

Method	Accuracy (%)	No. of points	No. of images	Feature dimension
Joint Bayesian [8]	92.42 (o)	5	99,773	2000×4
ConvNet-RBM [31]	92.52 (o)	3	87,628	N/A
CMD+SLBP [17]	92.58 (u)	3	N/A	2302
Fisher vector faces [29]	93.03 (u)	9	N/A	128×2
Tom-vs-Pete classifiers [2]	93.30 (o+r)	95	20,639	5000
High-dim LBP [9]	95.17 (o)	27	99,773	2000
TL Joint Bayesian [6]	96.33 (o+u)	27	99,773	2000
DeepFace [32]	97.25 (o+u)	6 + 67	4,400,000 + 3,000,000	4096×4
DeepID on CelebFaces	**96.05** (o)	5	87,628	150
DeepID on CelebFaces+	**97.05** (o)	5	202,599	150
DeepID on CelebFaces+ with transfer	**97.45** (o+u)	5	202,599	150

“o” denotes using outside training data, however, without using training data from LFW

“o+u” denotes using outside training data and LFW data in the unrestricted protocol for training

Joint Identification-Verification Signals

- Every two feature vectors extracted from the same identity should are close to each other

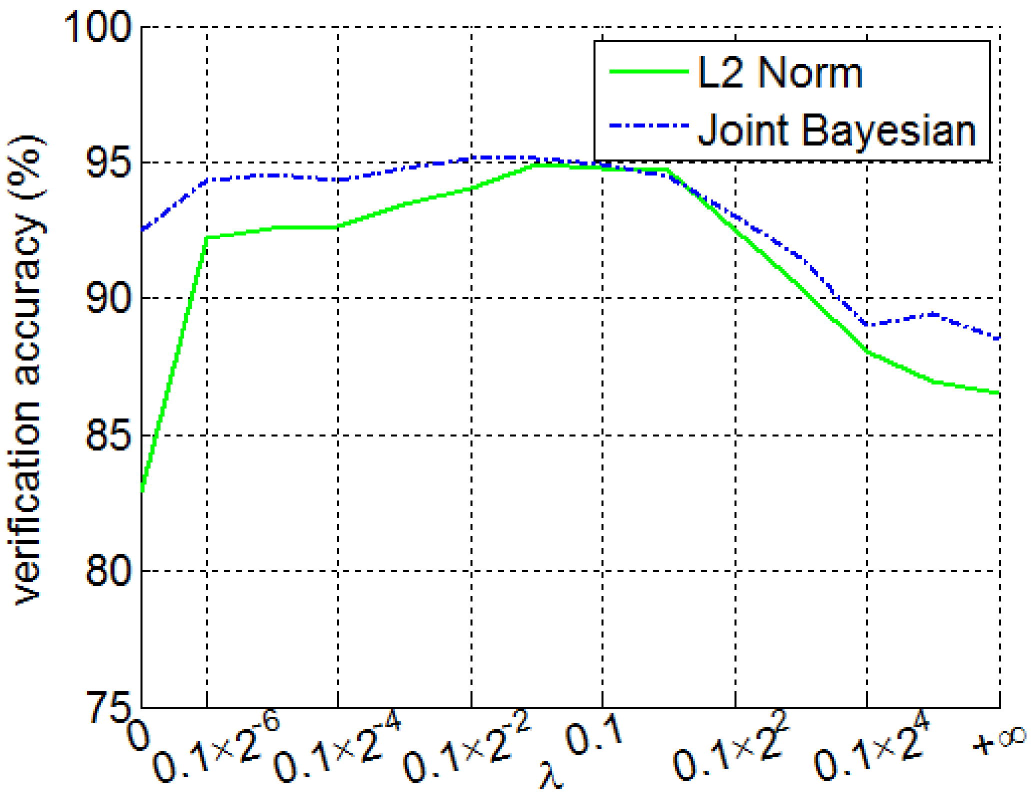
$$\mathrm{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\left\|f_i - f_j\right\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \left\|f_i - f_j\right\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

f_i and f_j are feature vectors extracted from two face images in comparison

y_{ij} = 1 means they are from the same identity; y_{ij} = -1means different identities

m is a margin to be learned

Balancing Identification and Verification Signals with Parameter λ

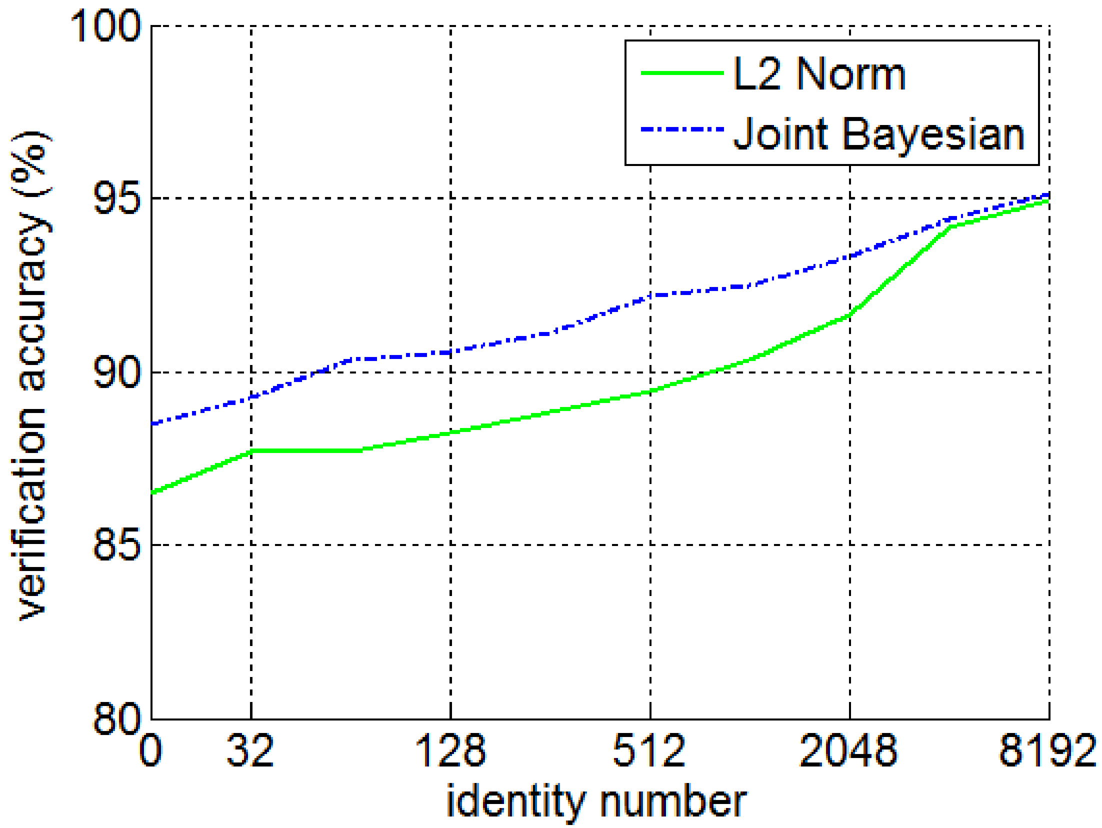


λ = 0: only identification signal

λ = +∞: only verification signal

Rich Identity Information Improves Feature Learning

- Face verification accuracies with the number of training identities

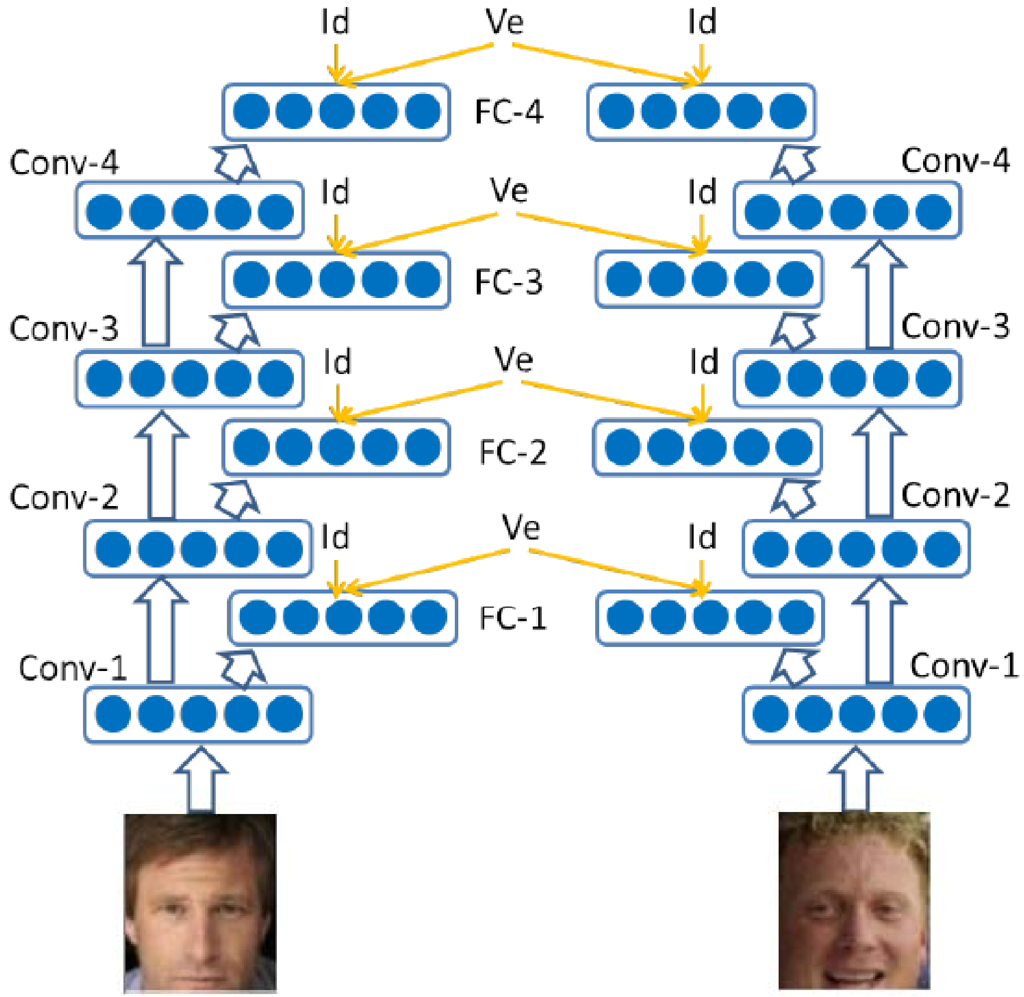


Summary of DeepID2

- 25 face regions at different scales and locations around landmarks are selected to build 25 neural networks
- All the 160 X 25 hidden identity features are further compressed into a 180-dimensional feature vector with PCA as a signature for each image
- With a single Titan GPU, the feature extraction process takes 35ms per image

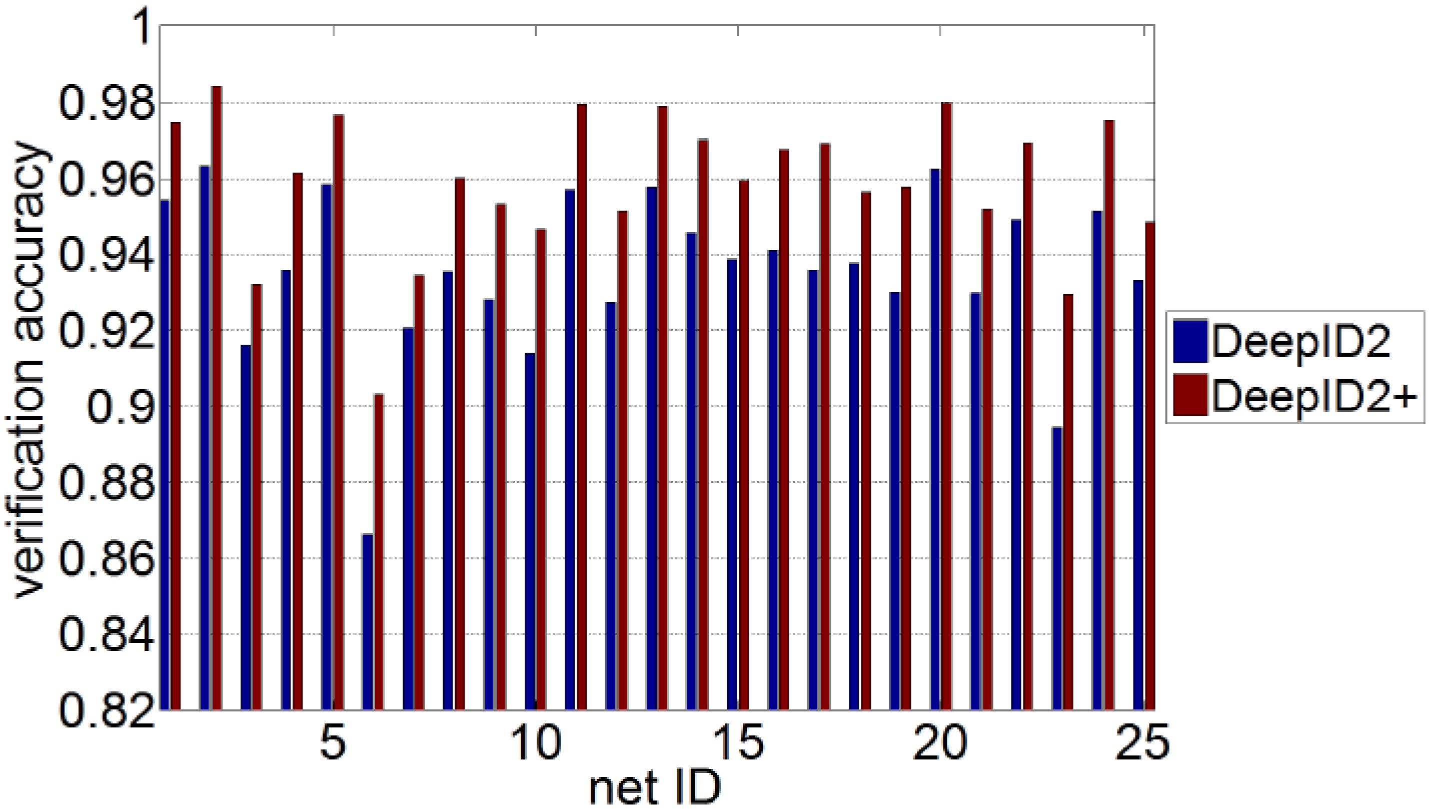
DeepID2+

- Larger net work structures
- Larger training data
- Adding supervisory signals at every layer



Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. arXiv:1412.1265, 2014.

Compare DeepID2 and DeepID2+ on LFW



Comparison of face verification accuracies on LFW with ConvNets trained on 25 face regions given in DeepID2

Best single model is improved from 96.72% to 98.70%

Final Result on LFW

Methods	High-dim LBP [1]	TL Joint Bayesian [2]	DeepFace [3]	DeepID [4]	DeepID2 [5]	DeepID2+ [6]
Accuracy (%)	95.17	96.33	97.35	97.45	99.15	99.47

[1] Chen, Cao, Wen, and Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. CVPR, 2013.

[2] Cao, Wipf, Wen, Duan, and Sun. A practical transfer learning algorithm for face verification. ICCV, 2013.

[3] Taigman, Yang, Ranzato, and Wolf. DeepFace: Closing the gap to human-level performance in face verification. *CVPR,* 2014.

[4] Sun, Wang, and Tang. Deep learning face representation from predicting 10,000 classes. *CVPR,* 2014.

[5] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. NIPS, 2014.

[6] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. arXiv:1412.1265, 2014.

Closed- and open-set face identification on LFW

Method	Rank-1 (%)	DIR @ 1% FAR (%)
COST-S1 [1]	56.7	25
COST-S1+s2 [1]	66.5	35
DeepFace [2]	64.9	44.5
DeepFace+ [3]	82.5	61.9
DeepID2	91.1	61.6
DeepID2+	**95.0**	**80.7**

[1] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *TR MSU-CSE-14-1,* 2014.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verifica- tion. In *Proc. CVPR,* 2014.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web- scale training for face identification. Technical report, arXiv:1406.5266, 2014.

Face Verification on YouTube Faces

Methods	Accuracy (%)
LM3L [1]	81.3 ± 1.2
DDML (LBP) [2]	81.3 ± 1.6
DDML (combined) [2]	82.3 ± 1.5
EigenPEP [3]	84.8 ± 1.4
DeepFace [4]	91.4 ± 1.1
DeepID2+	93.2 ± 0.2

[1] J. Hu, J. Lu, J. Yuan, and Y. P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," ACCV 2014

[2] J. Hu, J. Lu, and Y. P. Tan, "Discriminative deep metric learning for face verification in the wild," CVPR 2014

[3] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt, "Eigen-pep for video face recognition," ACCV 2014

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," CVPR 2014.

Unified subspace analysis

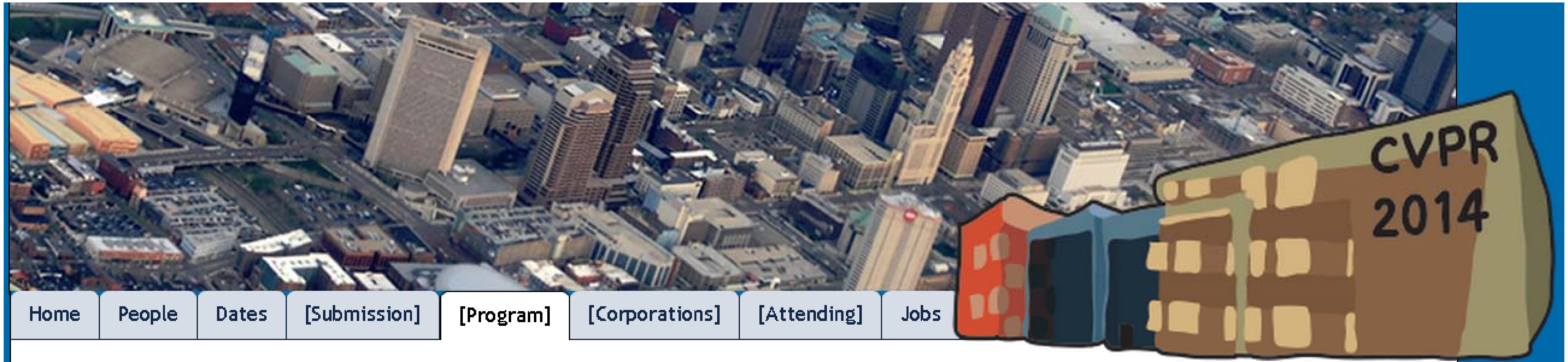
- Identification signal is in S_b; verification signal is in S_w
- Maximize distance between classes under constraint that intrapersonal variation is constant
- Linear feature mapping

Joint deep learning

- Learn features by joint identification-verification
- Minimize intra-personal variation under constraint that the distance between classes is constant
- Hierarchical nonlinear feature extraction
- Generalization power increases with more training identities

- Need to be careful when magnifying the inter-personal difference; Unsupervised learning many be a good choice to remove noise

We still do not know limit of deep learning yet



CVPR 2014 Plenary Speakers

Neural mechanisms for face processing

Professor Doris Tsao, California Institute of Technology (Caltech)

How the brain distills a representation of meaningful objects from retinal input is one of the central challenges of systems neuroscience. Functional imaging experiments in the macaque reveal that one ecologically important class of objects, faces, is represented by a system of six discrete, strongly interconnected regions. Electrophysiological recordings show that these 'face patches' have unique functional profiles. By studying the distinct visual representations maintained in these six face patches, the sequence of information flow between them, and the role each plays in face perception, we are gaining new insights into hierarchical information processing in the brain.

What has been learned by DeepID2+?

Properties owned by neurons?

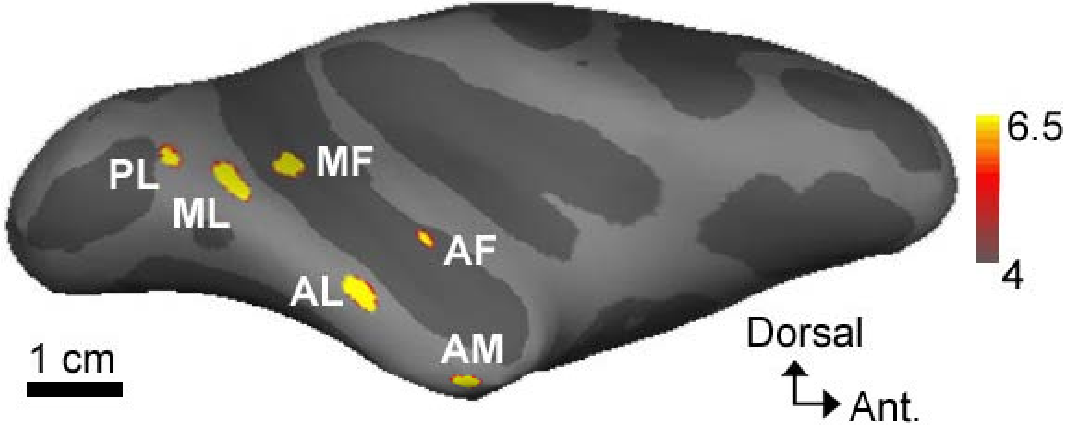
Moderate sparse

Selective to identities and attributes

Robust to data corruption

These properties are naturally owned by DeepID2+ through large-scale training, without explicitly adding regularization terms to the model

Biological Motivation

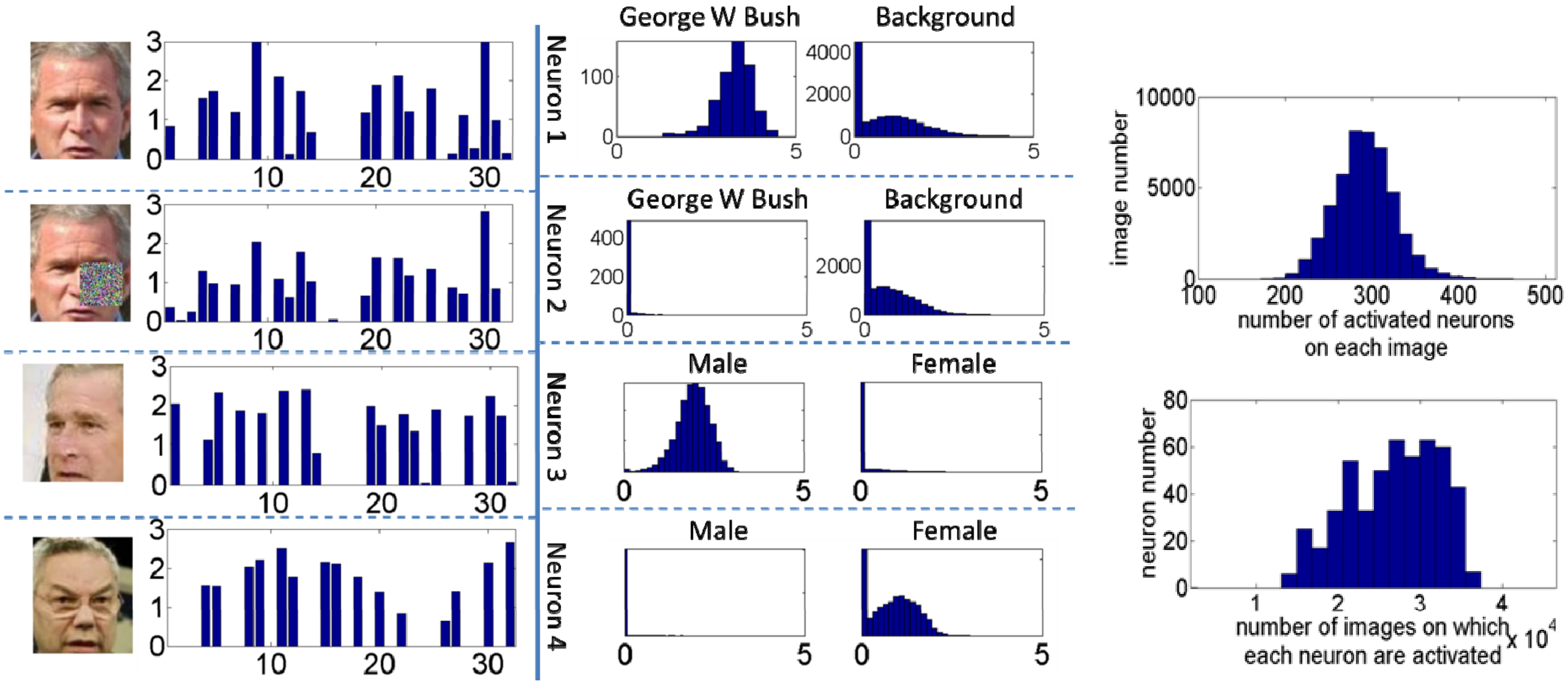


- Monkey has a face-processing network that is made of six interconnected face-selective regions
- Neurons in some of these regions were view-specific, while some others were tuned to identity across views
- View could be generalized to other factors, e.g. expressions?

Winrich A. Freiwald and Doris Y. Tsao, “Functional compartmentalization and viewpoint generalization within the macaque face-processing system,” *Science,* 330(6005):845–851, 2010.

Deeply learned features are moderately space

- For an input image, about half of the neurons are activated
- An neuron has response on about half of the images



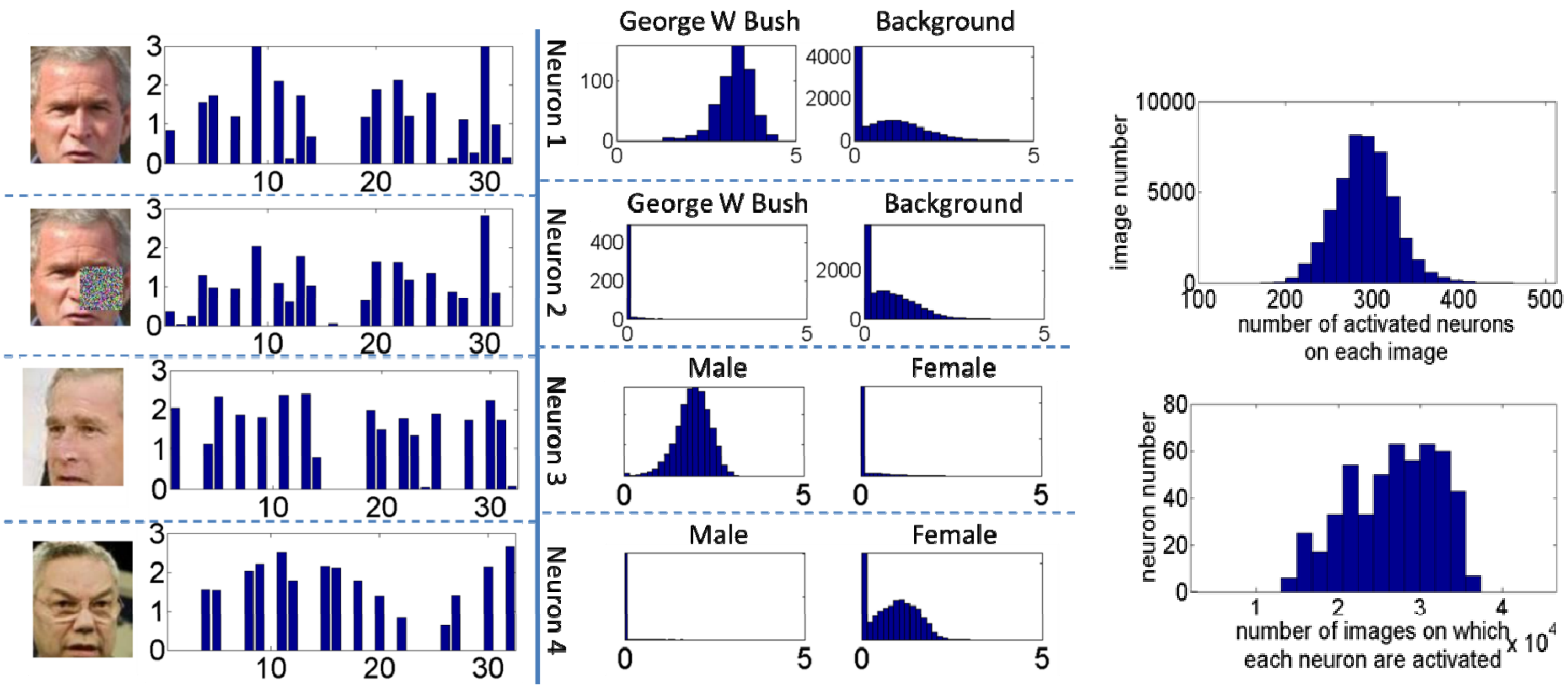
Deeply learned features are moderately space

- The binary codes on activation patterns of neurons are very effective on face recognition
- Activation patterns are more important than activation magnitudes in face recognition

	Joint Bayesian (%)	**Hamming distance (%)**
Single model (real values)	98.70	n/a
Single model (binary code)	97.67	96.46
Combined model (real values)	99.47	n/a
Combined model (binary code)	99.12	97.47

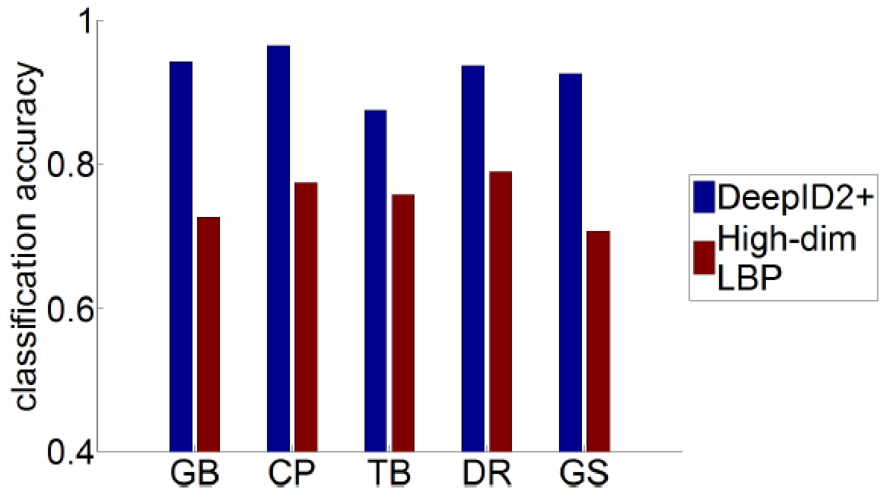
Deeply learned features are selective to identities and attributes

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

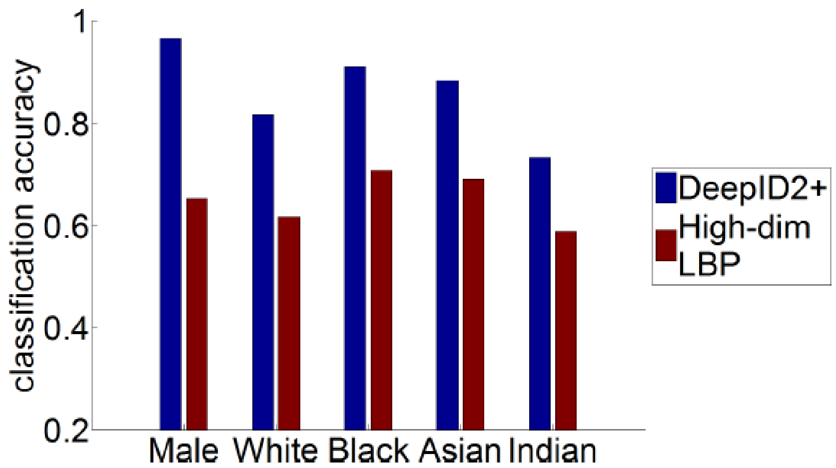


Deeply learned features are selective to identities and attributes

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



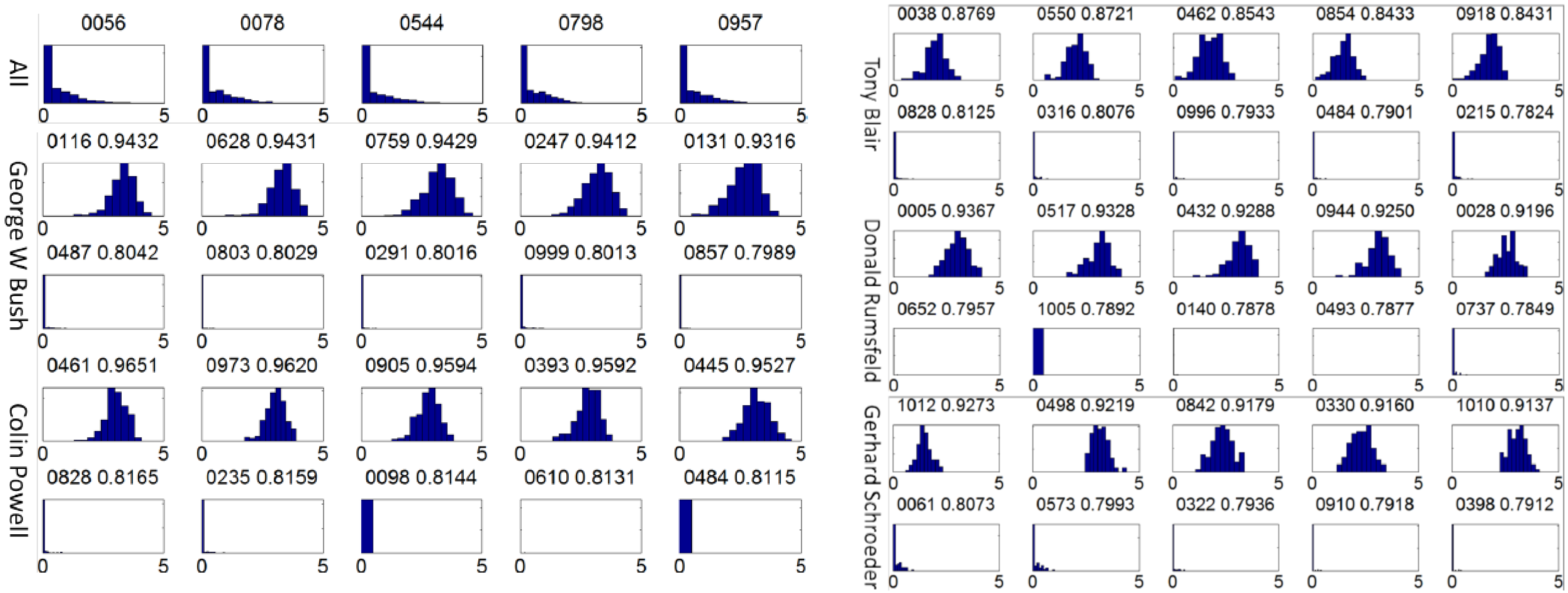
Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. GB, CP, TB, DR, and GS are five celebrities with the most images in LFW.



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

Deeply learned features are selective to identities and attributes

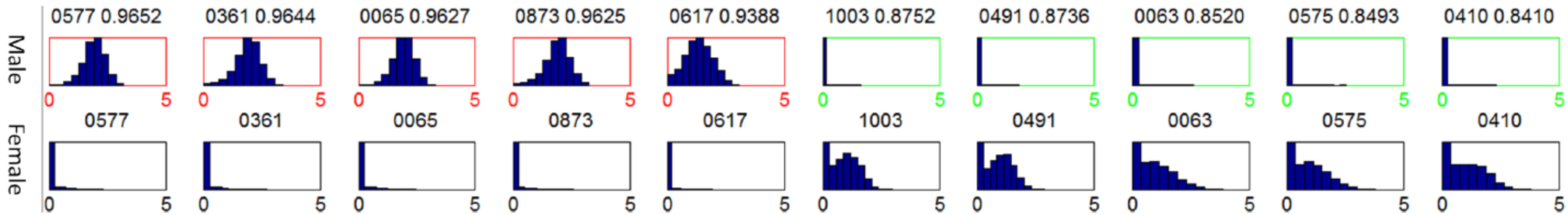
- Excitatory and inhibitory neurons



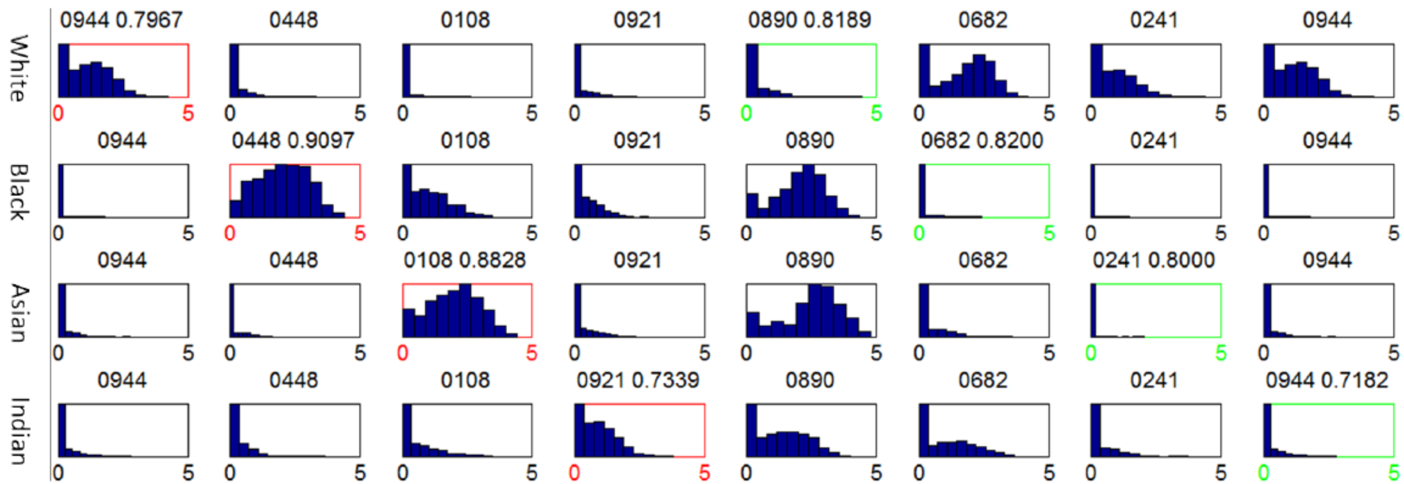
Histograms of neural activations over identities with the most images in LFW

Deeply learned features are selective to identities and attributes

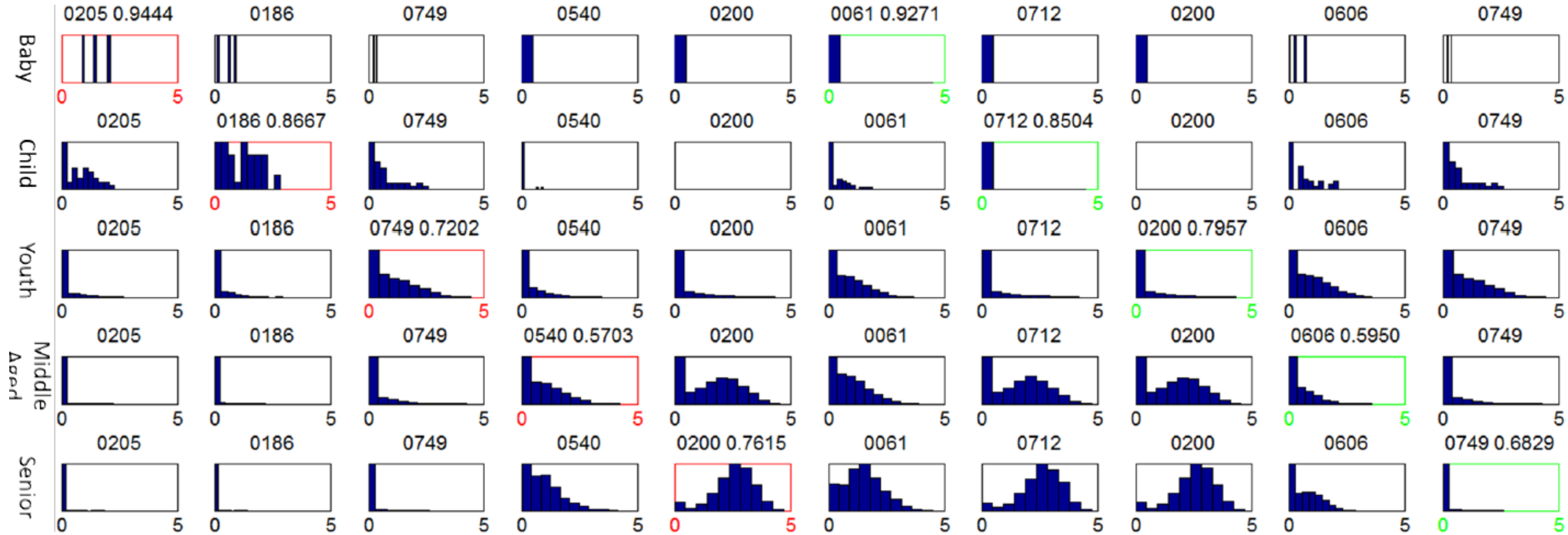
- Excitatory and inhibitory neurons



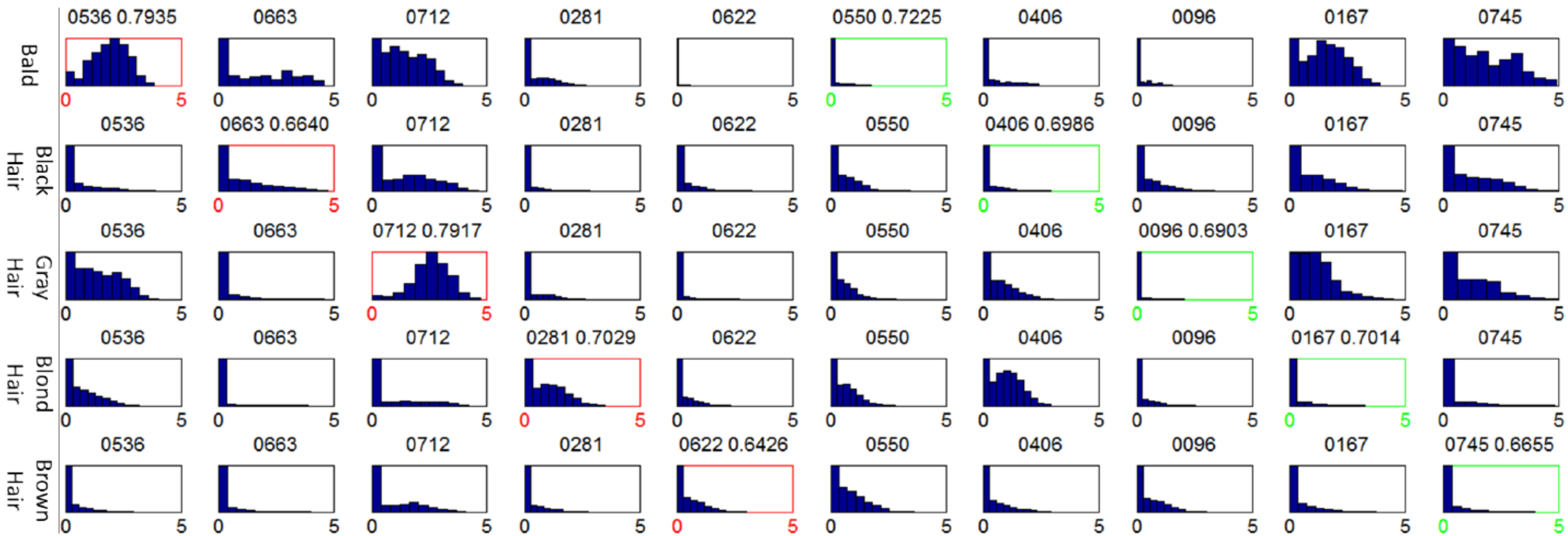
Histograms of neural activations over gender-related attributes (Male and Female)



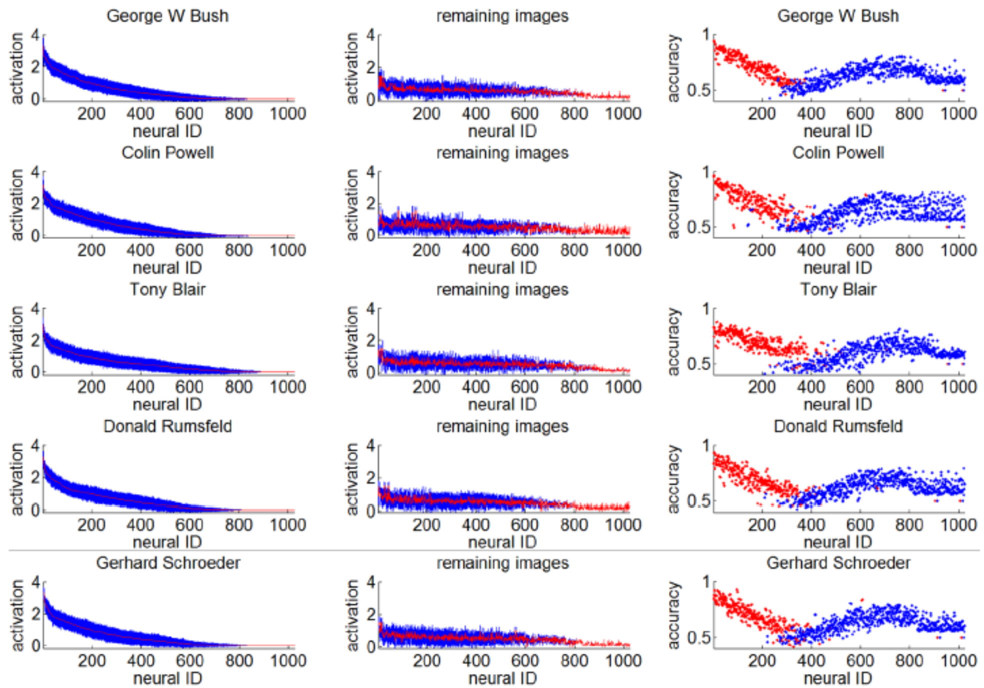
Histograms of neural activations over race-related attributes (White, Black, Asian and India)



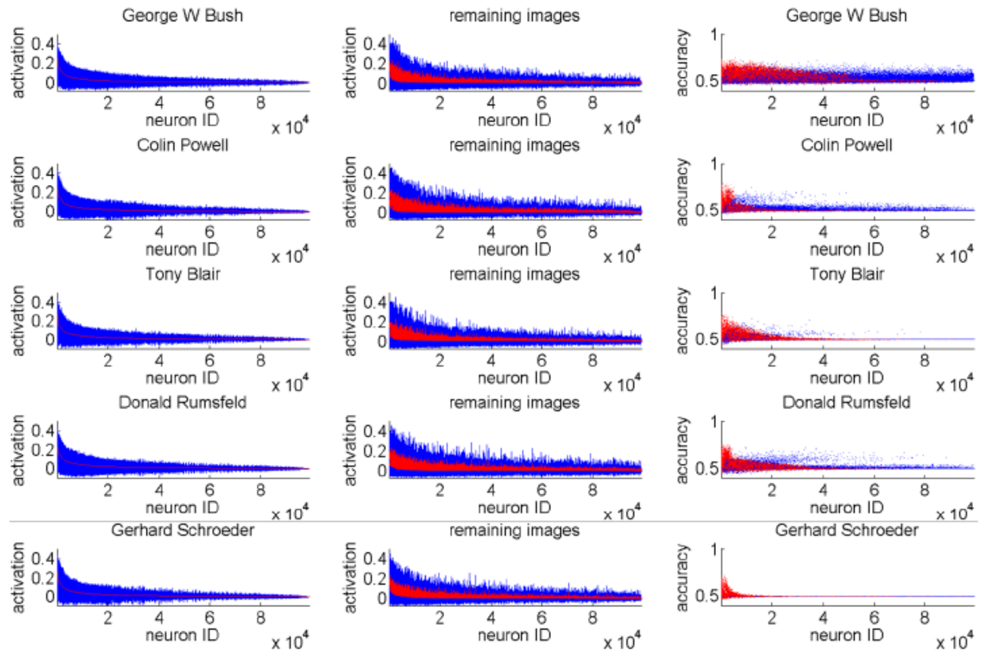
Histogram of neural activations over age-related attributes (Baby, Child, Youth, Middle Aged, and Senior)



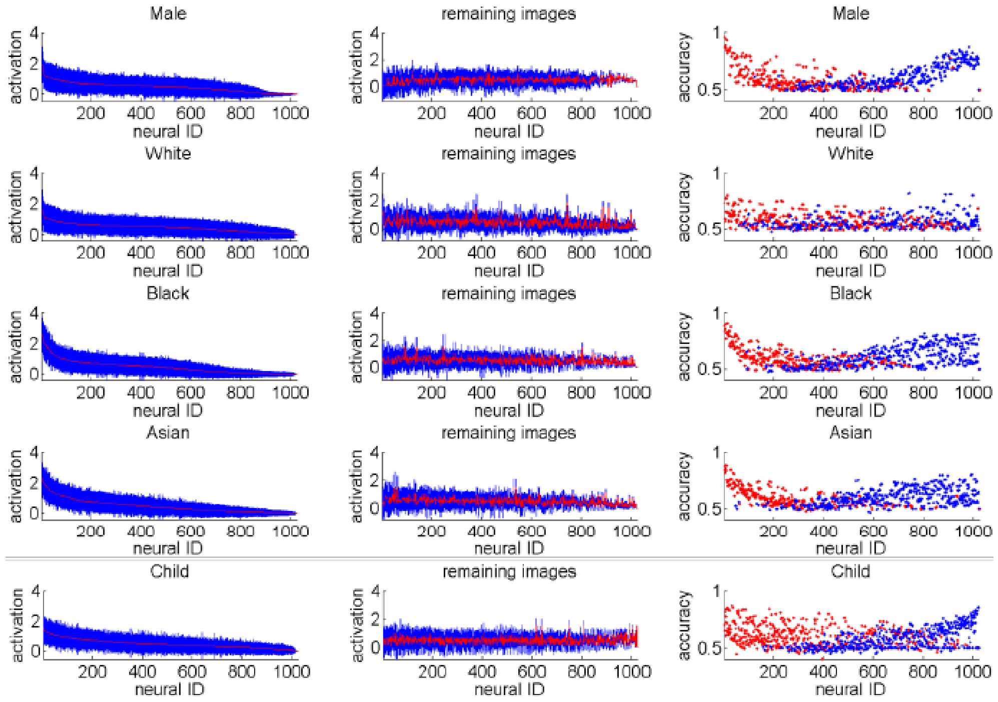
Histogram of neural activations over hair-related attributes (Bald, Black Hair, Gray Hair, Blond Hair, and Brown Hair.



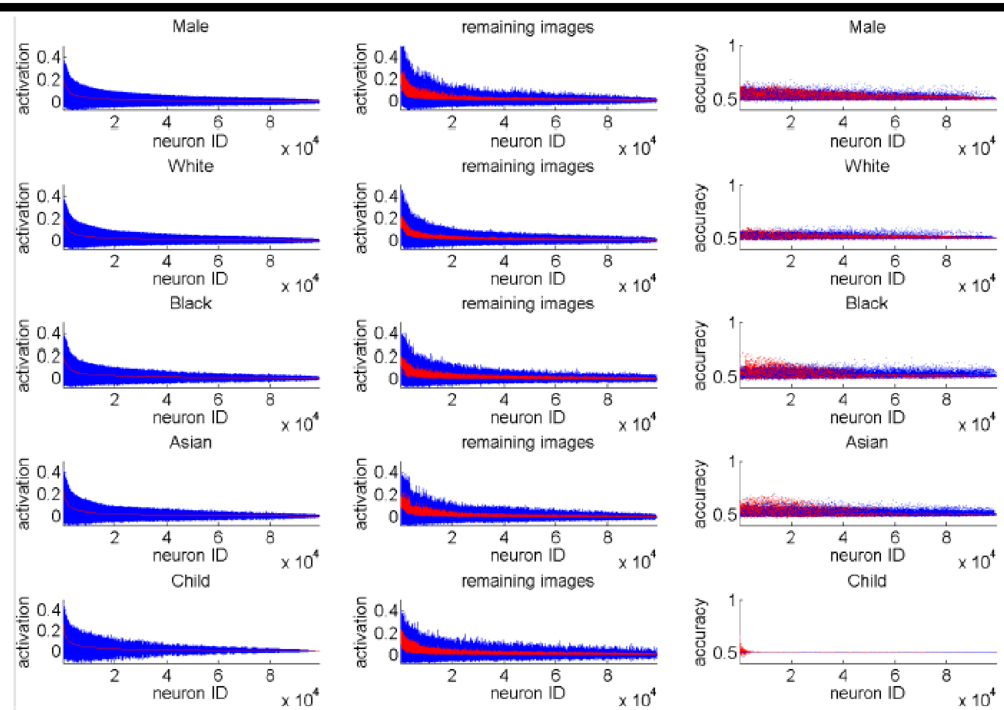
DeepID2+



High-dim LBP



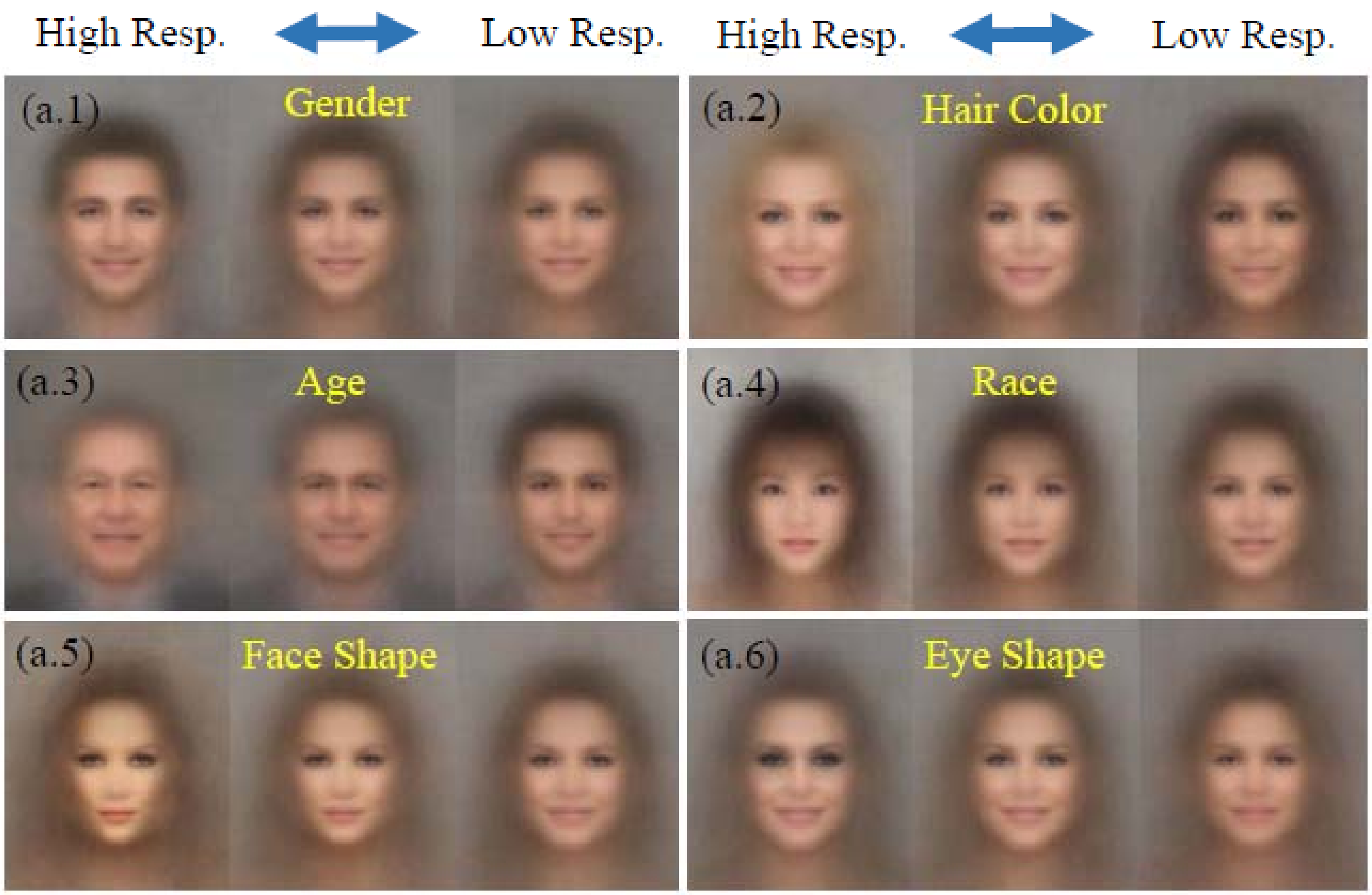
DeepID2+



High-dim LBP

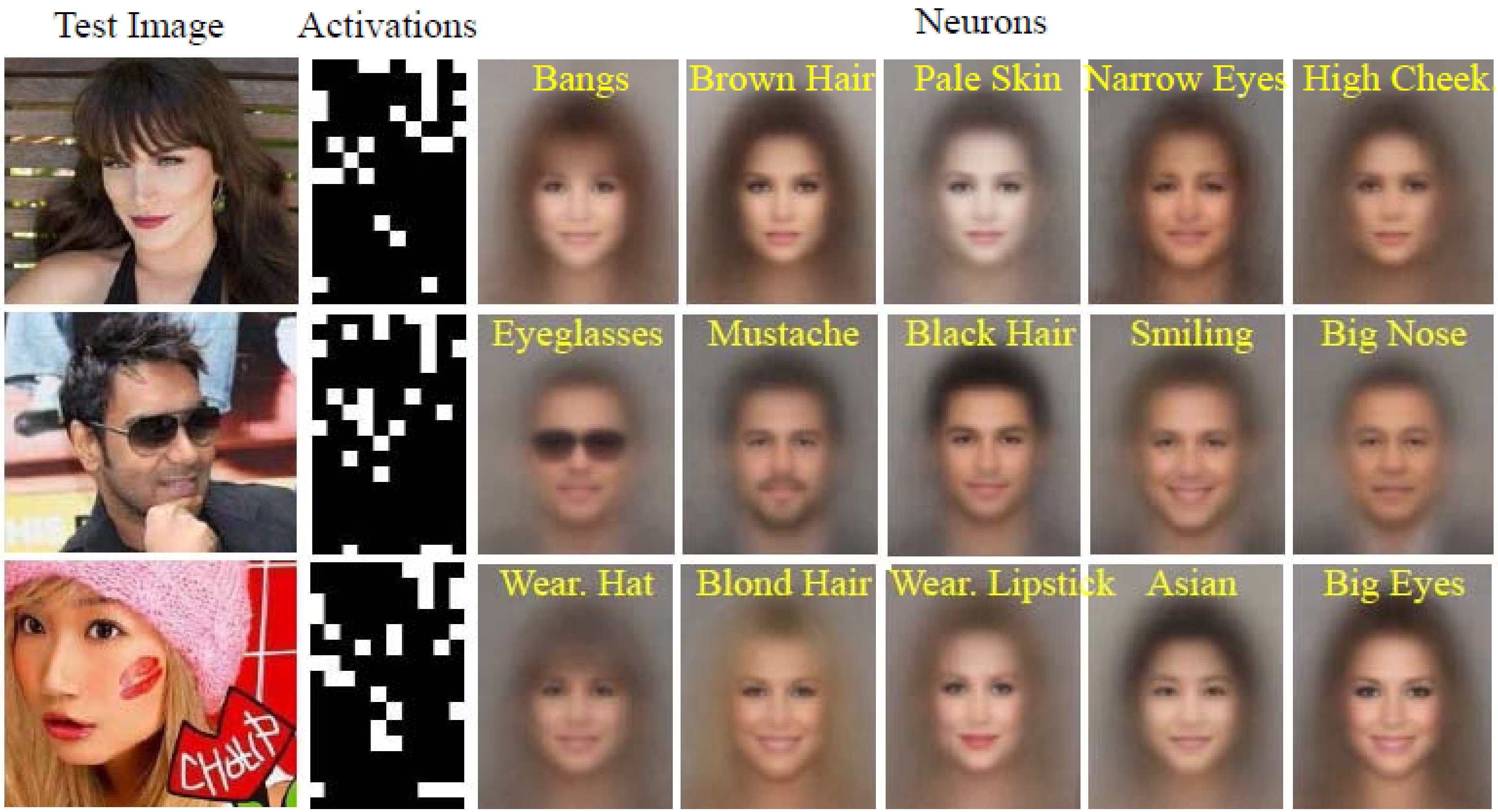
Deeply learned features are selective to identities and attributes

- Visualize the semantic meaning of each neuron



Deeply learned features are selective to identities and attributes

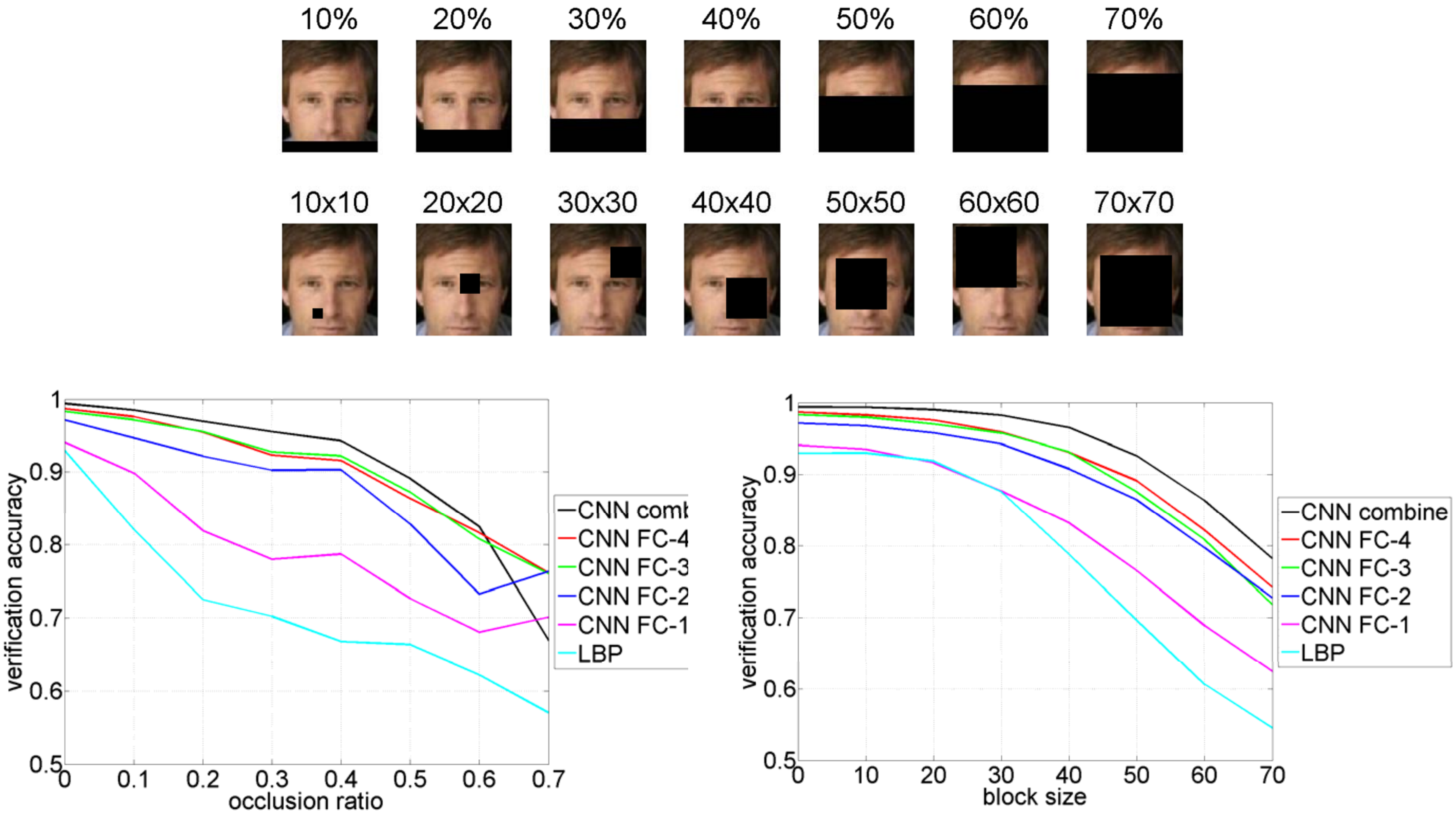
- Visualize the semantic meaning of each neuron

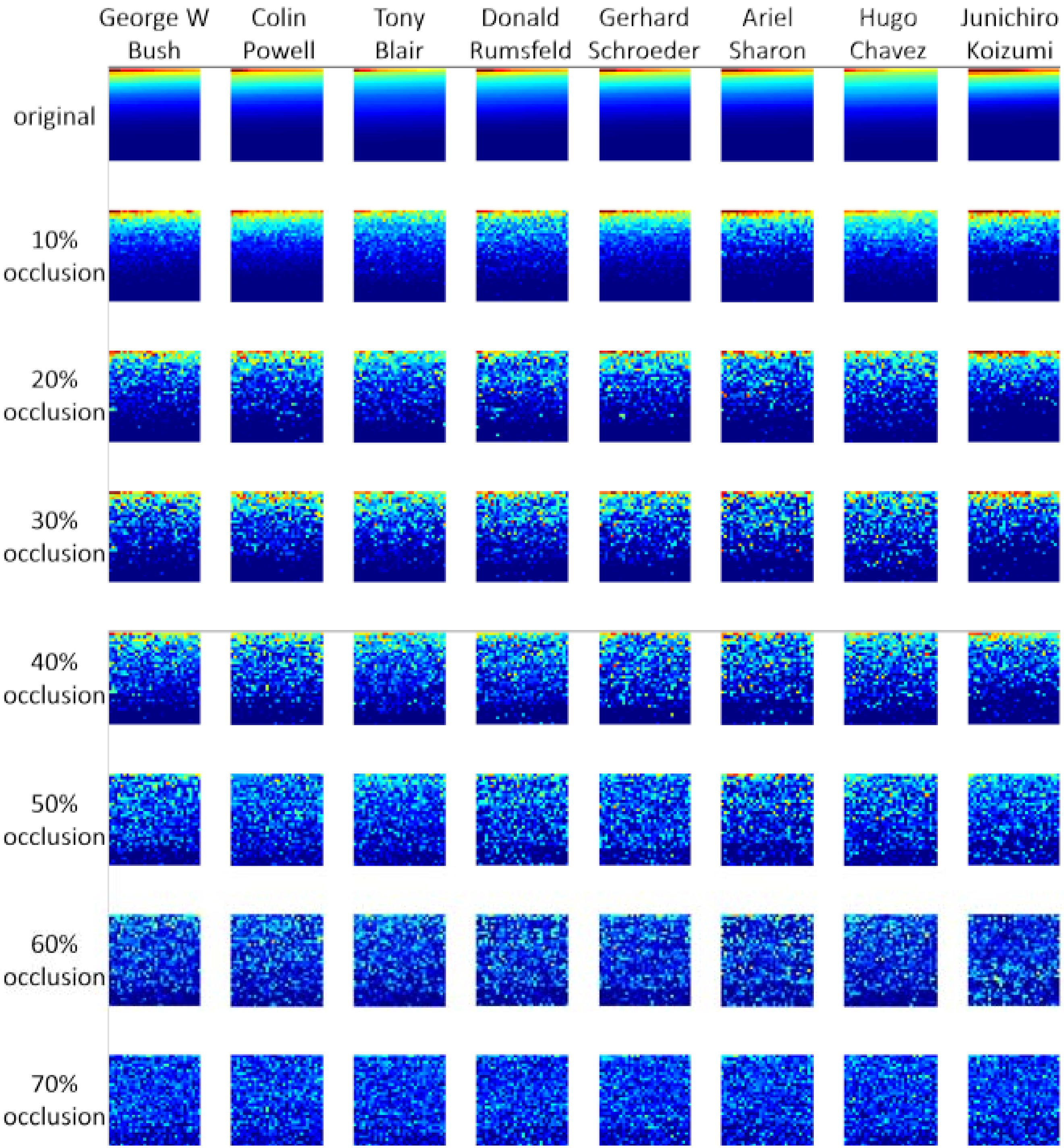


Neurons are ranked by their responses in descending order with respect to test images

Deeply learned features are robust to occlusions

- Global features are more robust to occlusions



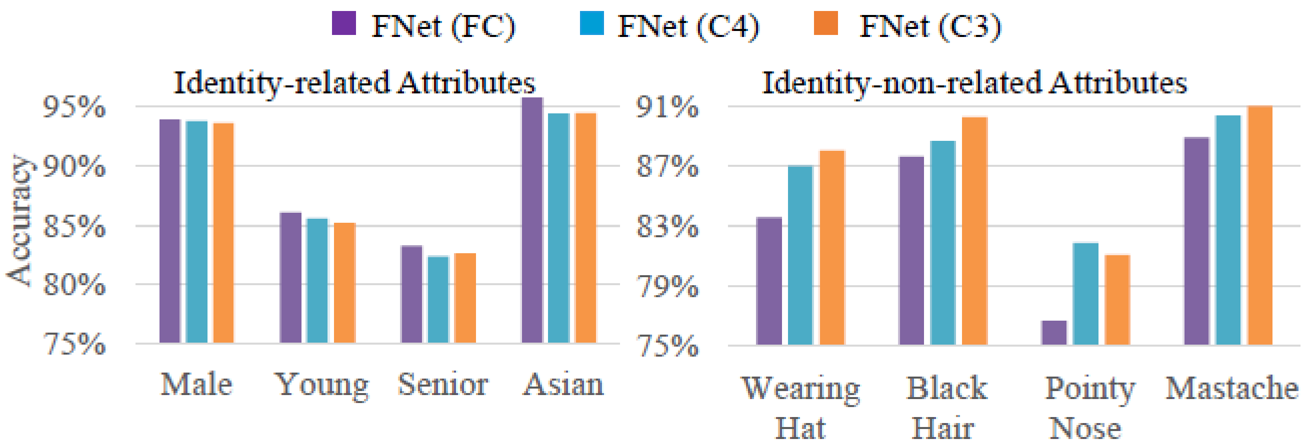

George W Bush
Colin Powell
Tony Blair
Donald Rumsfeld
Gerhard Schroeder
Ariel Sharon
Hugo Chavez
Junichiro Koizumi
original
10% occlusion
20% occlusion
30% occlusion
40% occlusion
50% occlusion
60% occlusion
70% occlusion

Can features learned by DeepID be effectively applied to other face related tasks, such as face localization and face attribute recognition?

Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild”, arXiv: 1411.7766, 2014

DeepID2 features for attribute recognition

- Features at top layers are more effective on recognizing identity related attributes
- Features at lowers layers are more effective on identity-non-related attributes



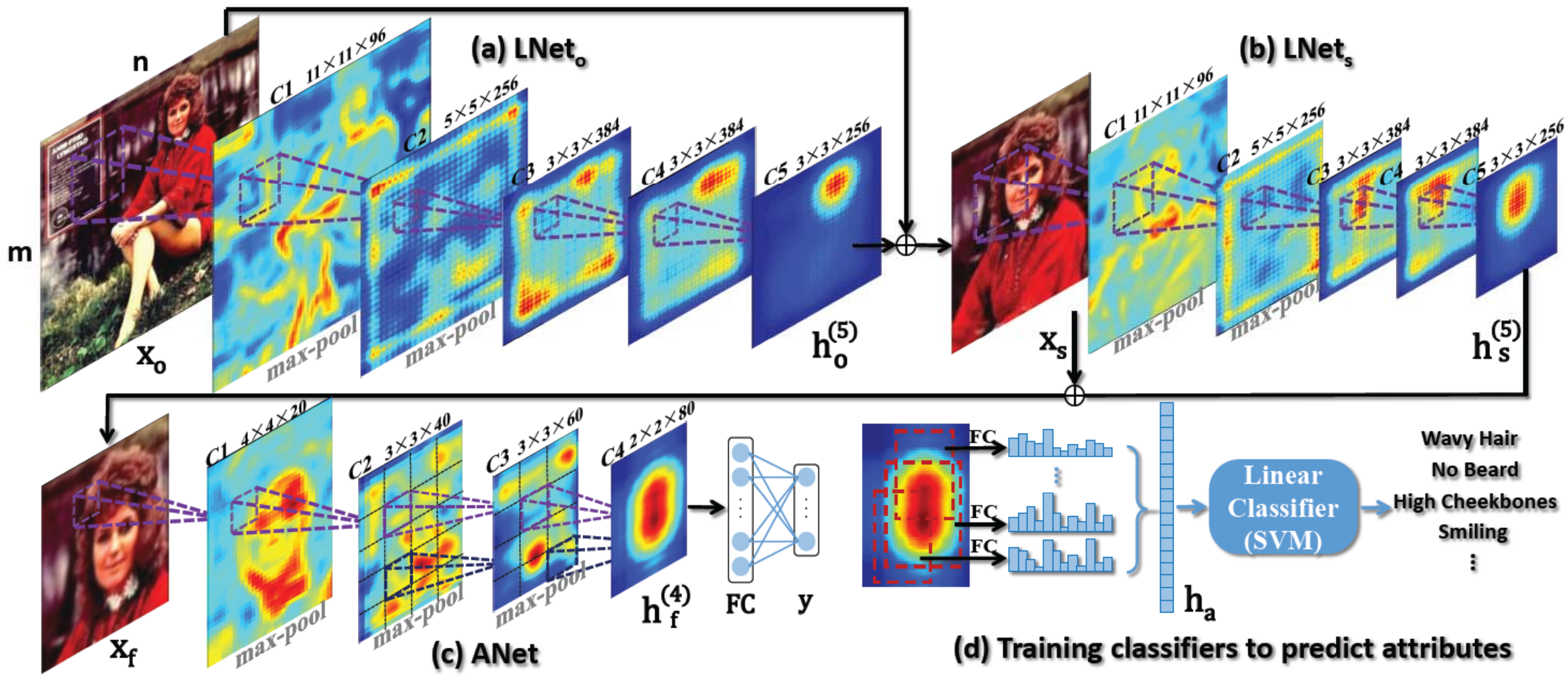
DeepID2 features for attribute recognition

- DeepID2 features can be directly used for attribute recognition
- Use DeeID2 features as initialization (pre-trained result), and then fine tune on attribute recognition
- Average accuracy on 40 attributes on CelebA and LFWA datasets

	CelebA	LFWA
FaceTracer [1] (HOG+SVM)	81	74
PANDA-W [2] (Parts are automatically detected)	79	71
PANDA-L [2] (Parts are given by ground truth)	85	81
DeepID2	**84**	**82**
Fine-tune (w/o DeepID2)	83	79
DeepID2 + fine-tune	**87**	**84**

Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” arXiv:1411.7766, 2014.

Features learned by DeepID and attribute recognition are effective on face localization



Lnets are pre-trained with ImageNet

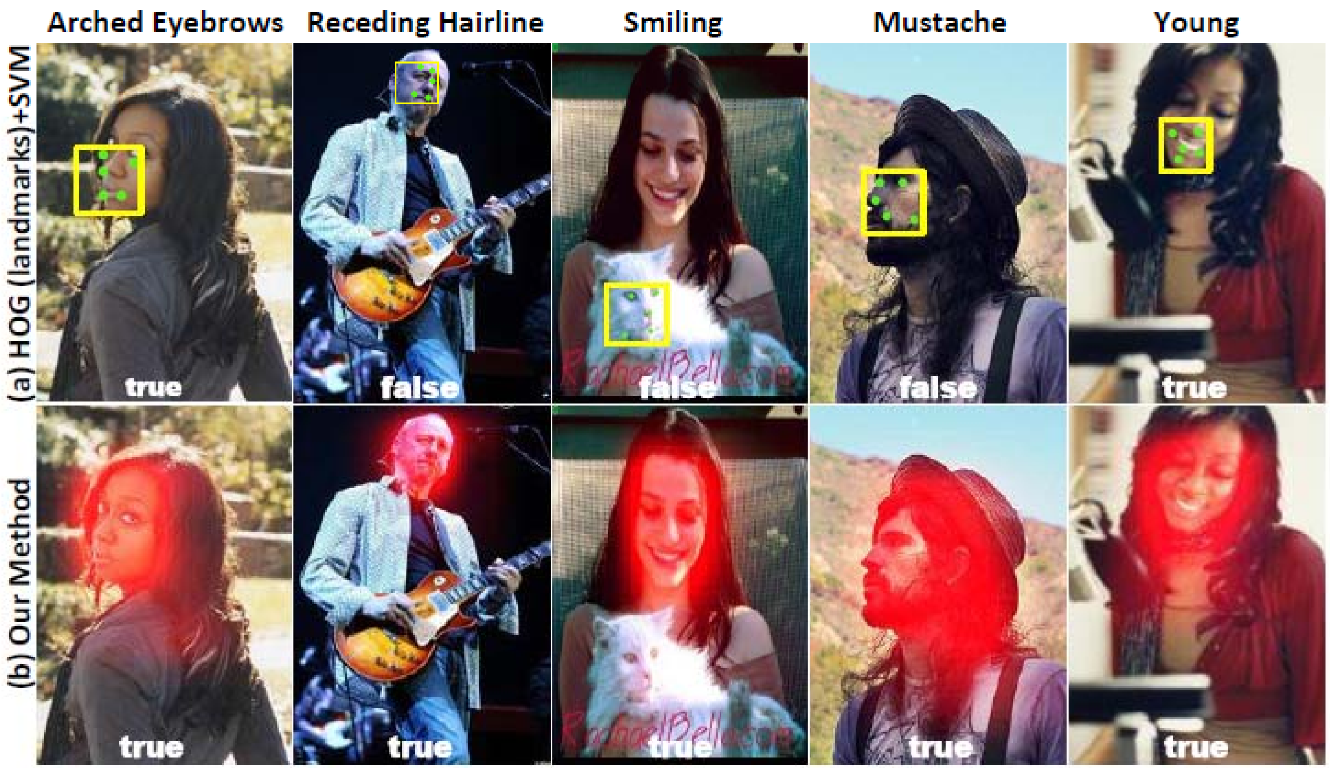
ANet is pre-trained with DeepID

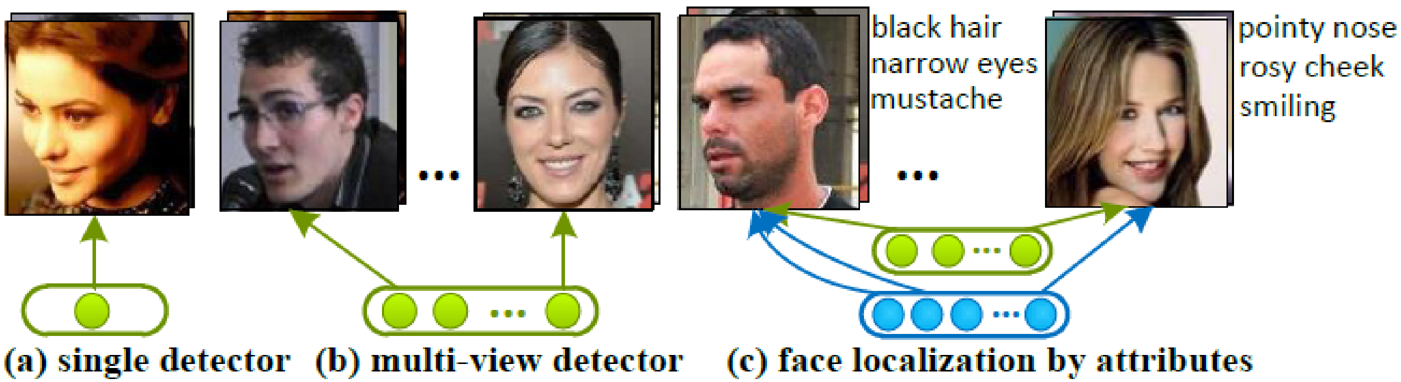
Both are fine-trained with face attributes

Lnet$_o$ calculates a response map which indicates the region of head-shoulder

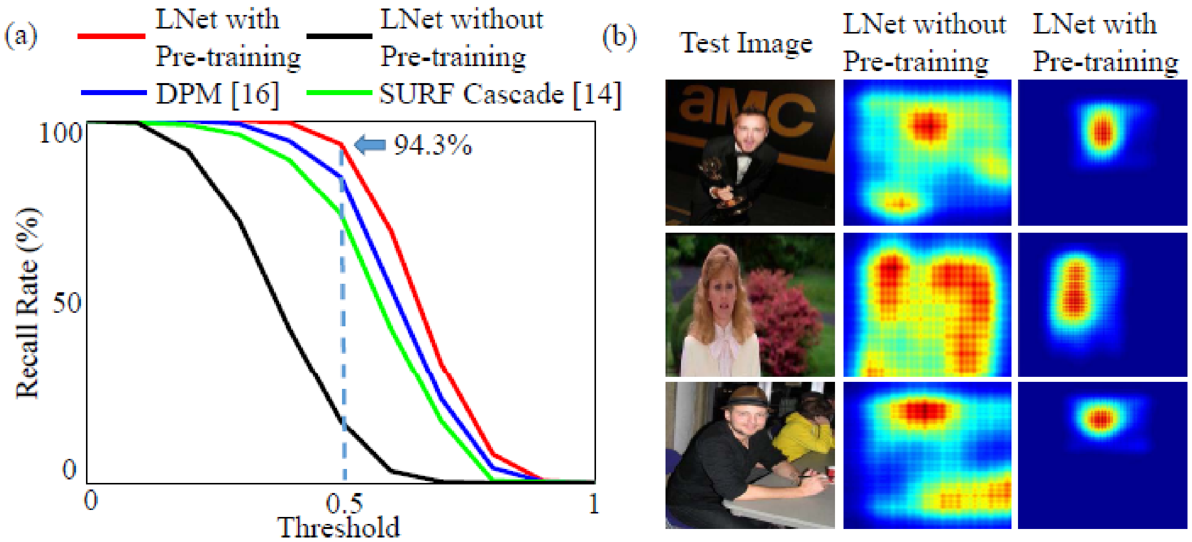
Lnet$_s$ refines the location of face

Anet extracts features to recognize attributes





Each neuron learned from face attribute recognition servers as a face detector, and it extends the idea of multi-view face detector to an extreme case



(a)
LNet with Pre-training
LNet without Pre-training
DPM [16]
SURF Cascade [14]
94.3%
Recall Rate (%)
100
50
0
0
0.5
1
Threshold
(b)
Test Image
LNet without Pre-training
LNet with Pre-training

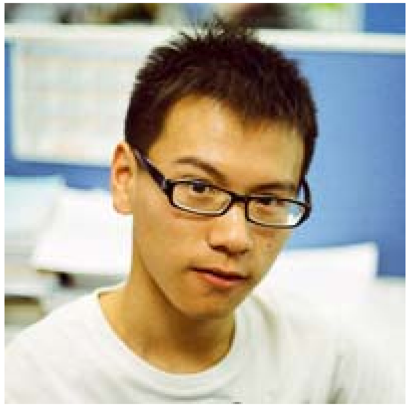
Deep Learning for Face Recognition

The projects started from December of 2012

DeepID

Yi Sun

MVP



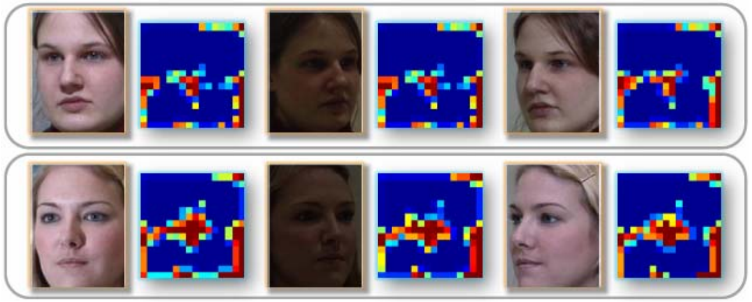
Zhenyao Zhu

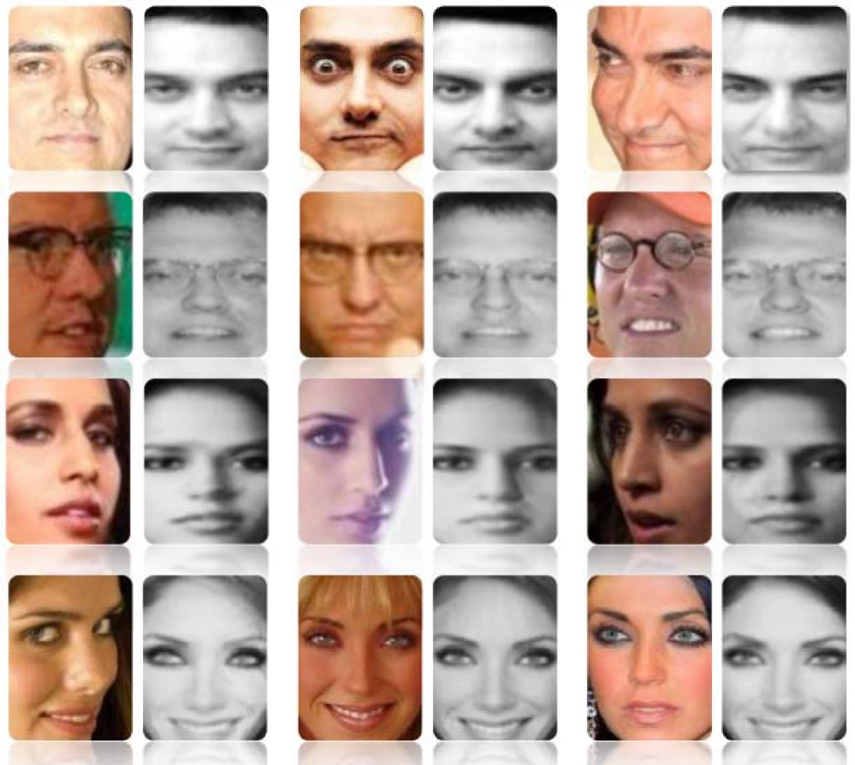
Ping Luo

Our understanding of deep learning

- Deep models can disentangle hidden factors with different neurons
- Deep models can be a combination of random and determinant neurons
- Image reconstruction is a stronger supervision task and can be used to learn features

Example 2: deep learning face identity features by recovering canonical-view face images

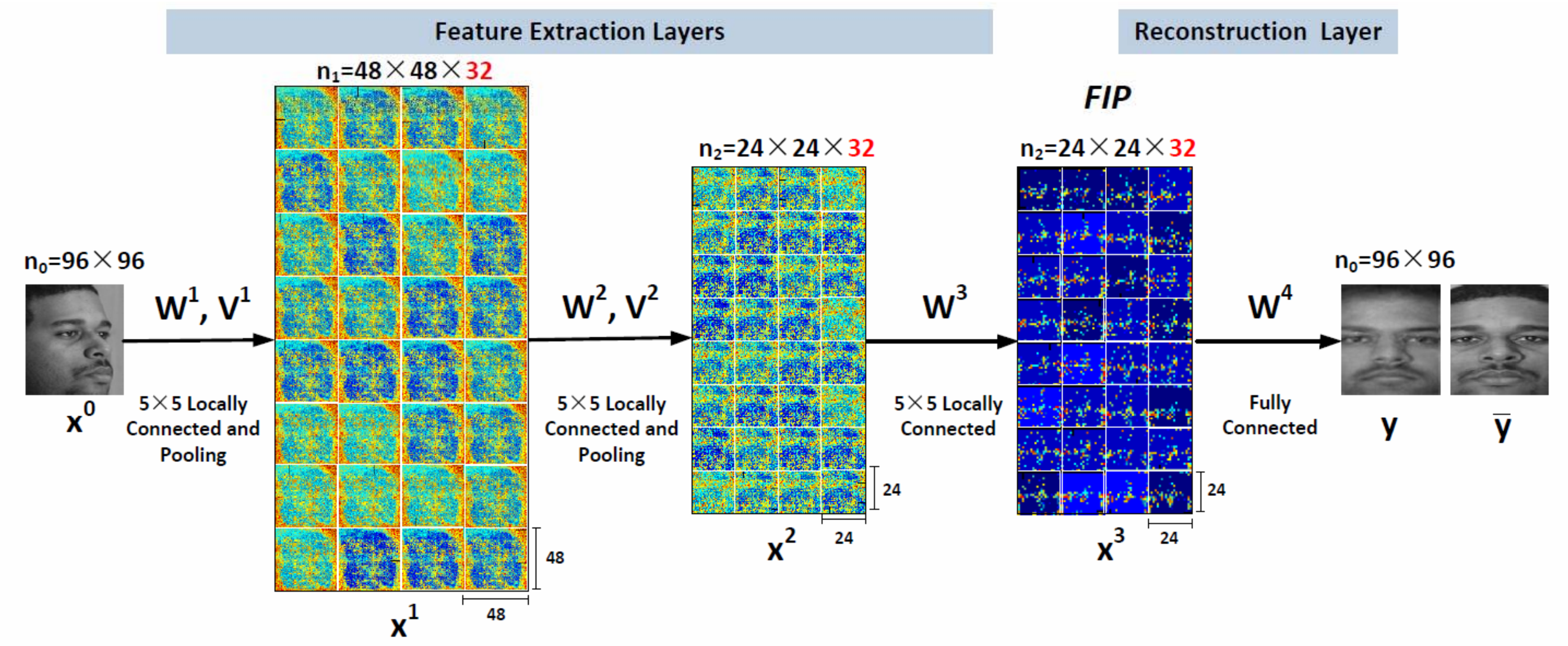




Reconstruction examples from LFW

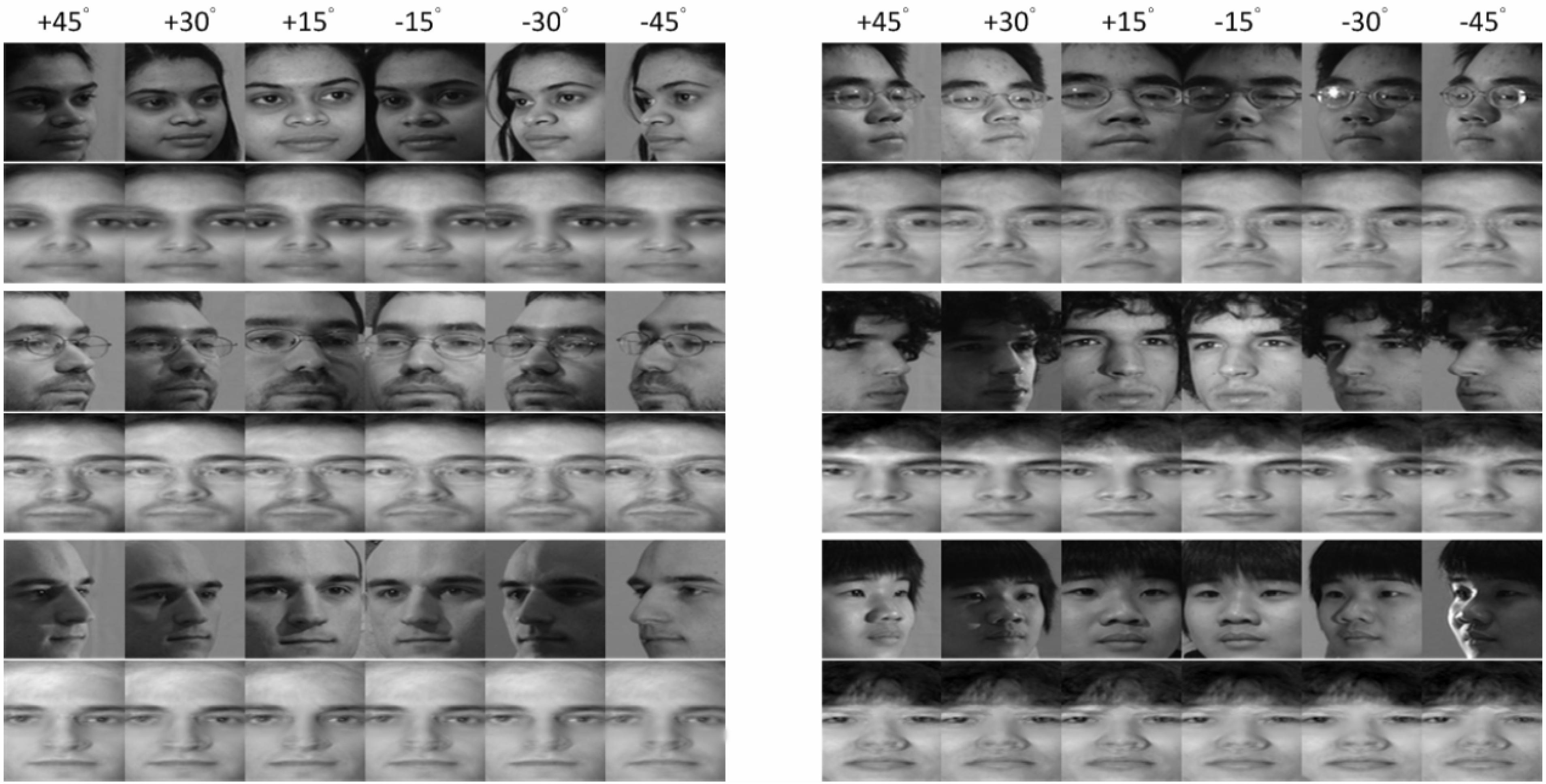
Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep Learning Identity Preserving Face Space," ICCV 2013.

- Deep model can disentangle hidden 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 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

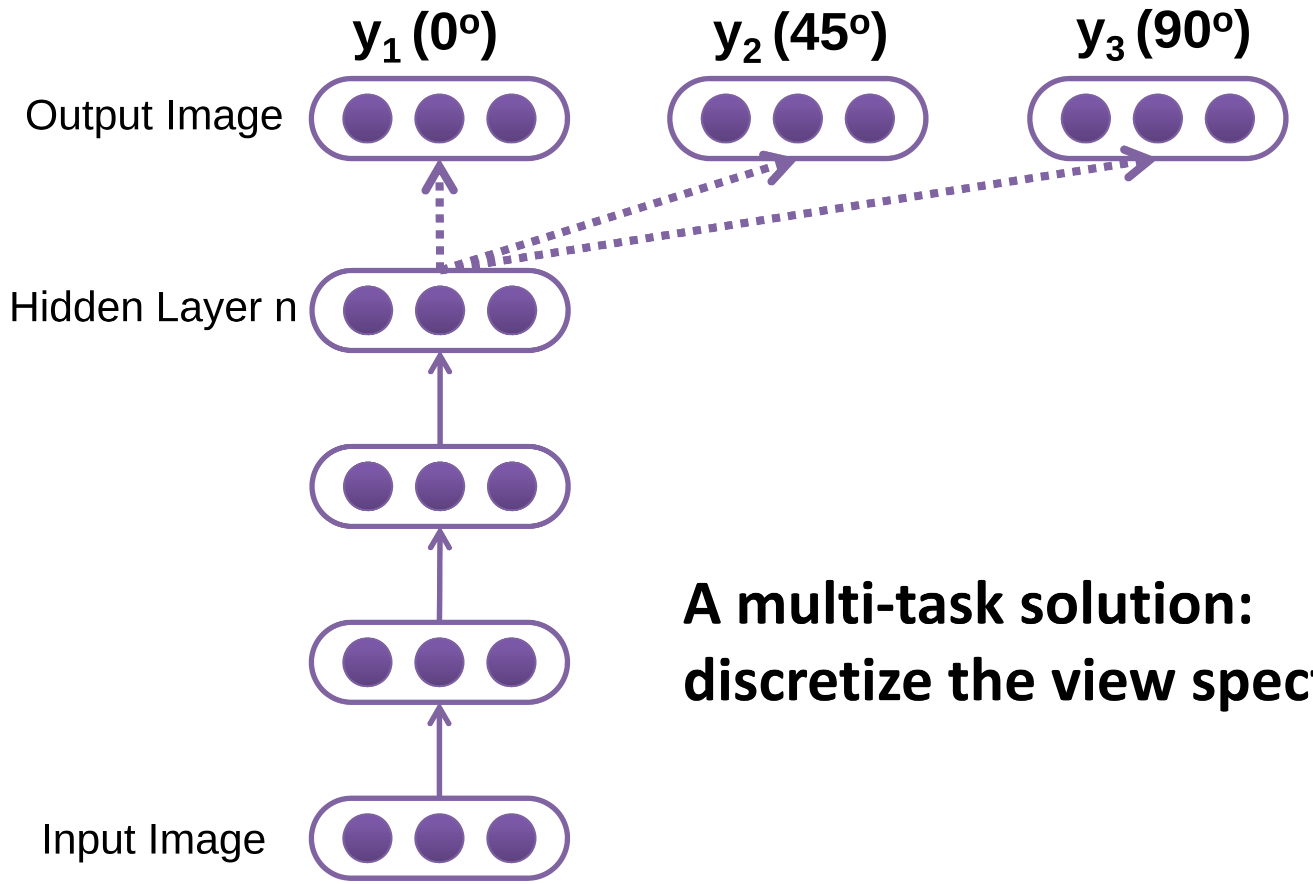
[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012. 3, 6

[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

It is still not a 3D representation yet

Can we reconstruct all the views?

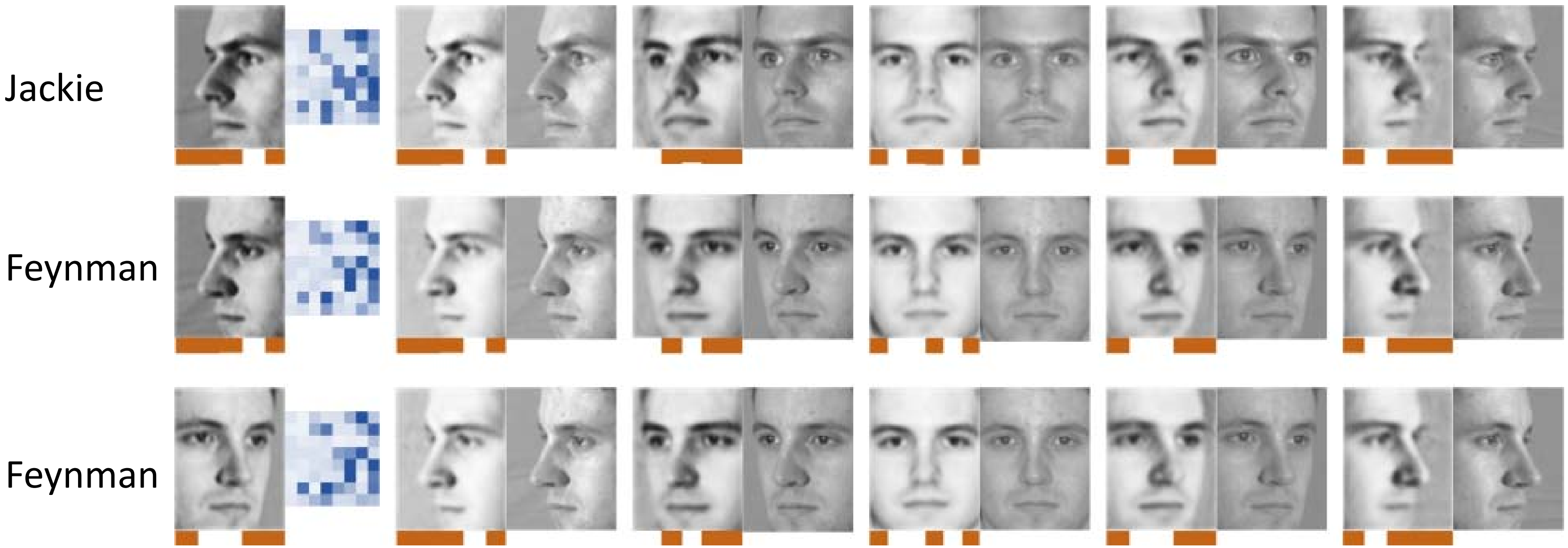


A multi-task solution: discretize the view spectrum

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

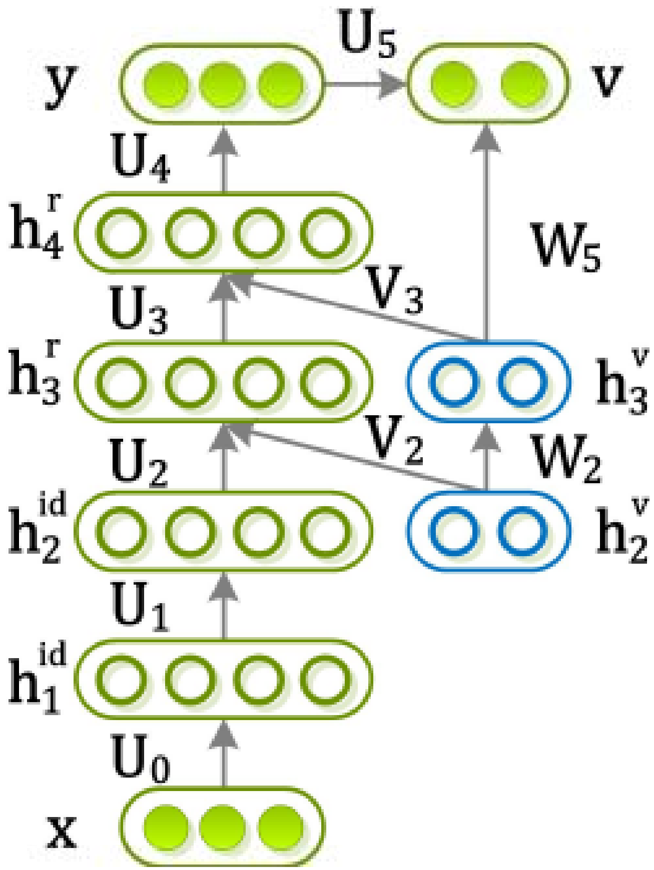
Deep Learning Multi-view Representation from 2D Images

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



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

Deep Learning Multi-view Representation from 2D Images



x and y are input and ouput 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
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
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**
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**
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.

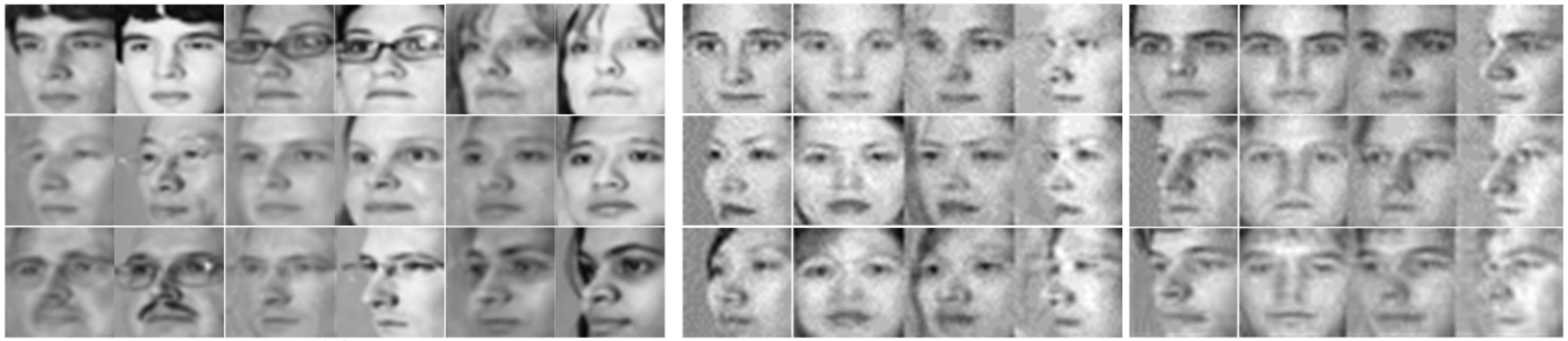
[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28:2037–2041, 2006.

[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

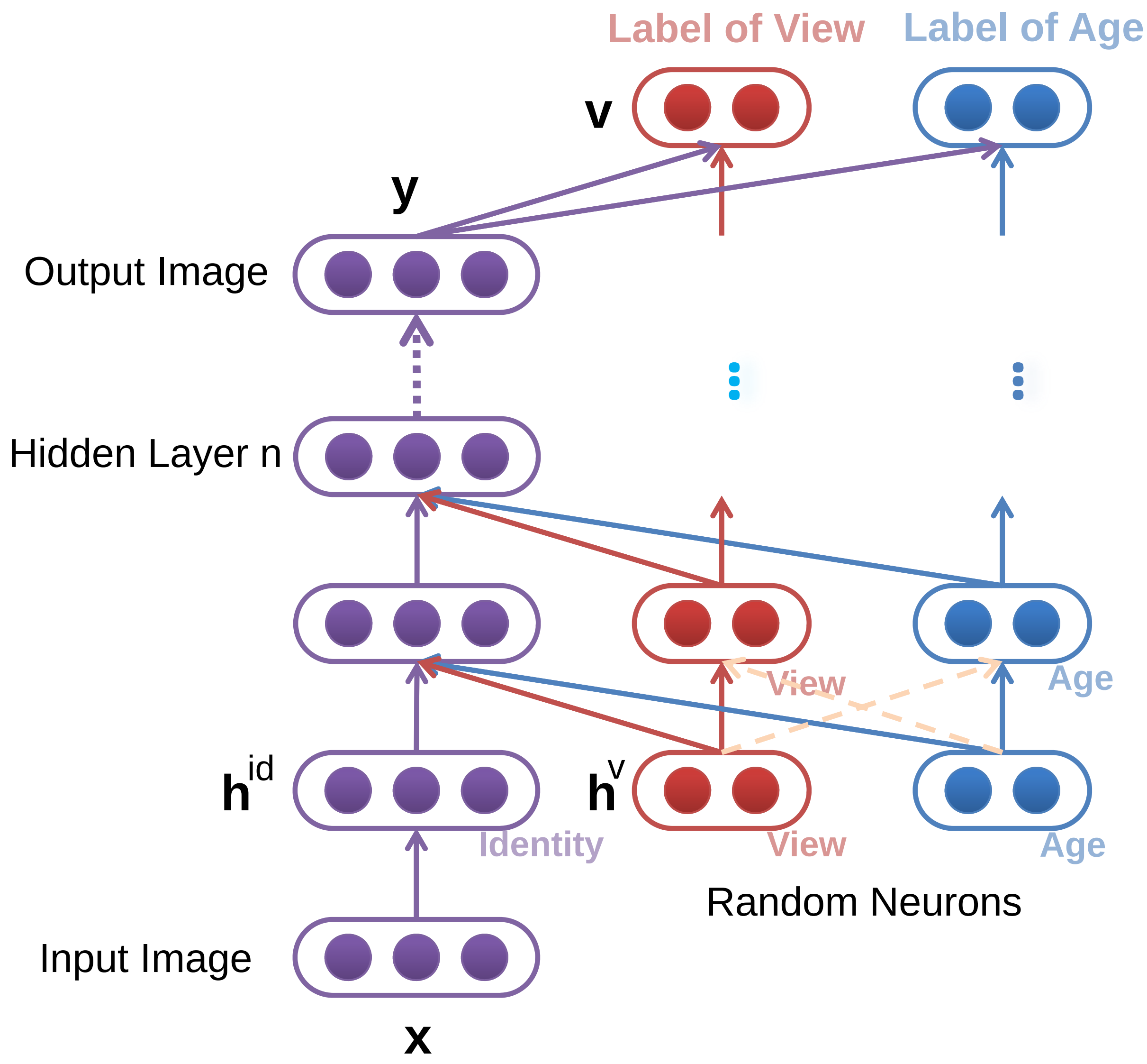
- Interpolate and predict images under viewpoints unobserved in the training set



(a) (b)

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



Tips

- Apply deep learning to new applications
- Bridge the connection between conventional pattern recognition systems and deep models, and get ideas from domain applications to propose new deep models and training strategies
- Understand why deep learning works, get insights and generalize those insights – have your own philosophy on deep learning
- Many neural networks were proposed in 1980s and 1990s and they can be revisited

Tips

- Many machine learning models were motivated by computer vision applications. However, computer vision did not have close interaction with neural networks in the past 15 years. We expect fast development of deep learning driven by applications.
- The most successful deep model in computer vision is CNN. The two most important operations in CNN, i.e. filtering and pooling, were also widely used in vision systems. We expect other effective domain knowledge, such more advanced pooling operations which are also robust to rotation and scaling, can be incoporated into deep models.

Tips

- Study the properties of neurons, which may provide the directions of theoretical studies on deep learning. Study the difference and similarity between the mechanisms of neural networks and human brains