香港中文大學
The Chinese University of Hong Kong

Introduction to Deep Learning

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

Outline

- Historical review of deep learning
- Introduction to classical deep models
- Why does deep learning work?
- Properties of deep feature representations

Machine Learning

$$x \longrightarrow \boxed{F(x)} \longrightarrow y \begin{cases} \text{Class label (Classification)} \\ \\ \text{Vector (Estimation)} \end{cases}$$

Object recognition $\longrightarrow$ {dog, cat, horse, flower, …}

Low-resolution image

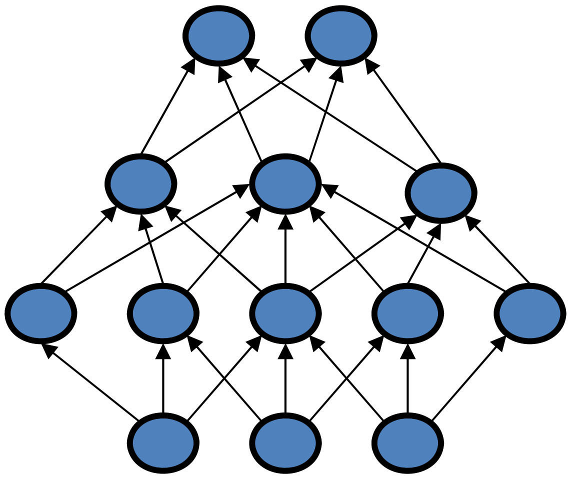
Super resolution $\longrightarrow$

High-resolution image

Neural network
Back propagation

Nature

1986

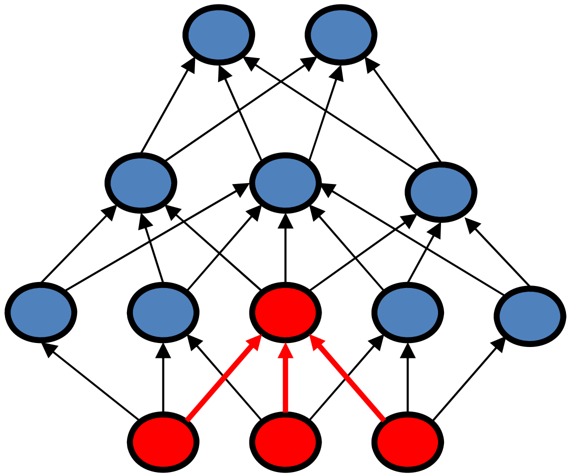


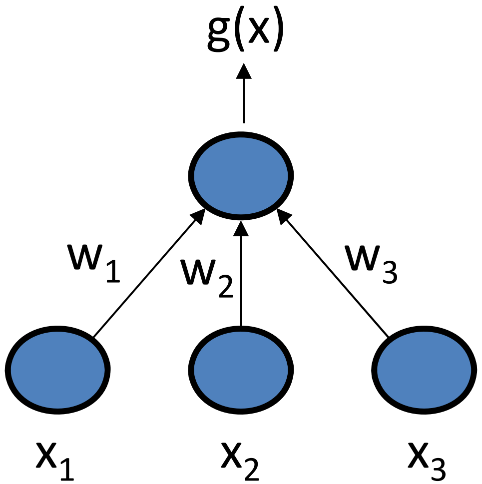
- Solve general learning problems
- Tied with biological system

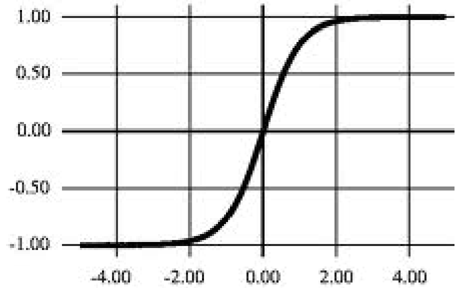
Neural network
Back propagation

Nature

1986

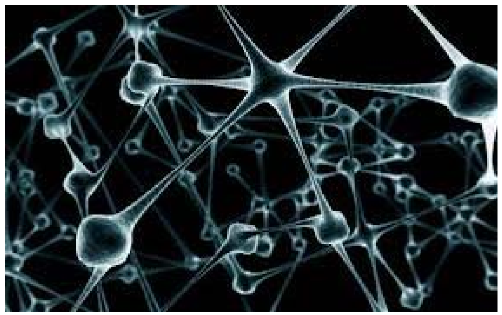






f(*net*)

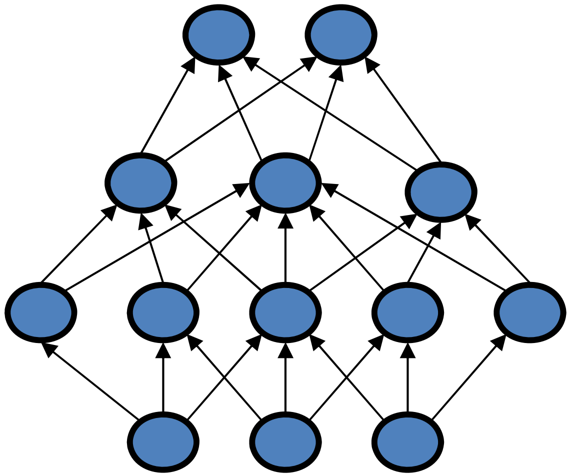
$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t\mathbf{x})$$



Neural network
Back propagation

Nature

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

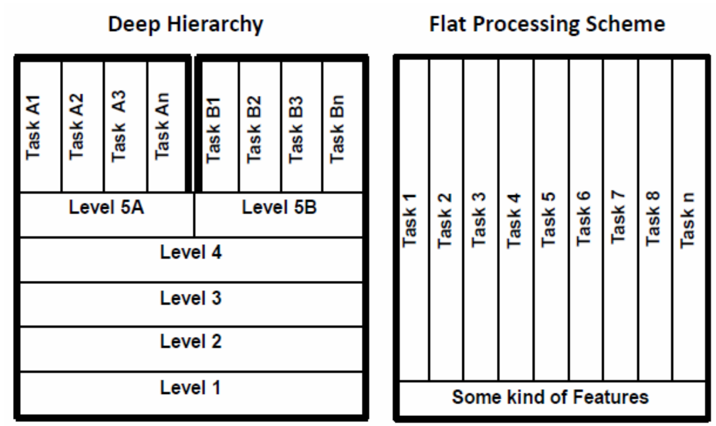
Nature

1986

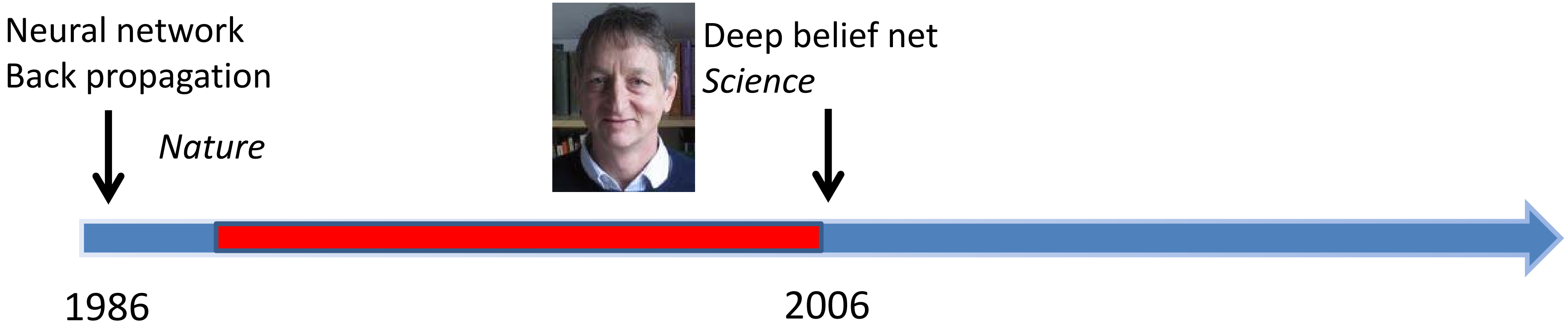
2006

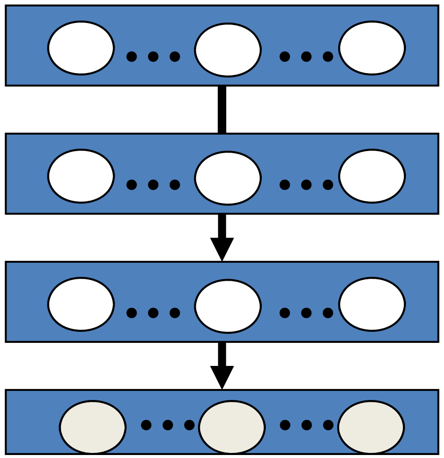
- SVM
- Boosting
- Decision tree
- KNN
- …

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



Kruger et al. TPAMI'13



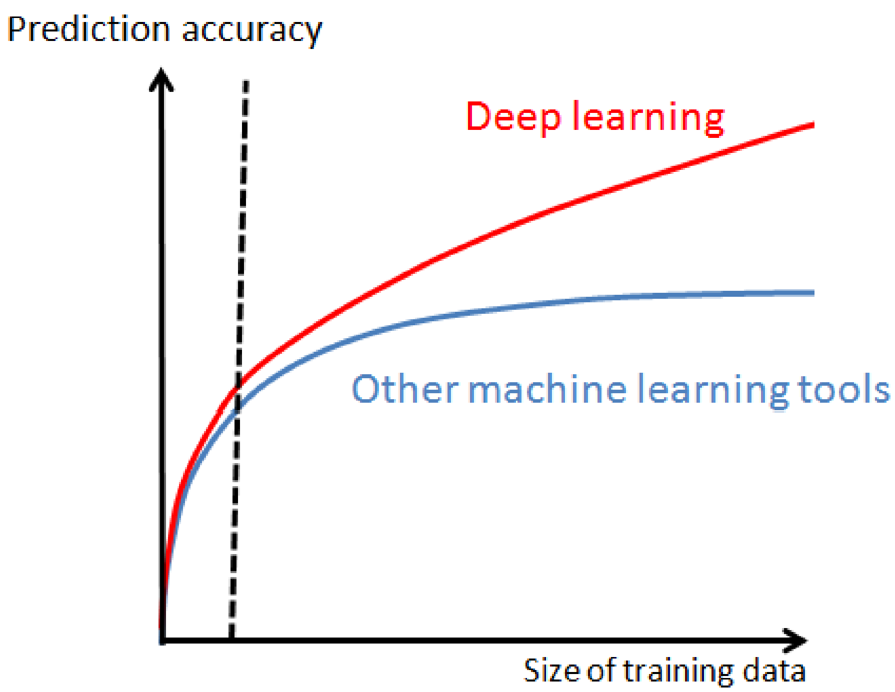


- Unsupervised & Layer-wised pre-training
- Better designs for modeling and training (normalization, nonlinearity, dropout)
- New development of computer architectures
  - GPU
  - Multi-core computer systems
- Large scale databases

Big Data !

Machine Learning with Big Data

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



How to increase model capacity?

Curse of dimensionality

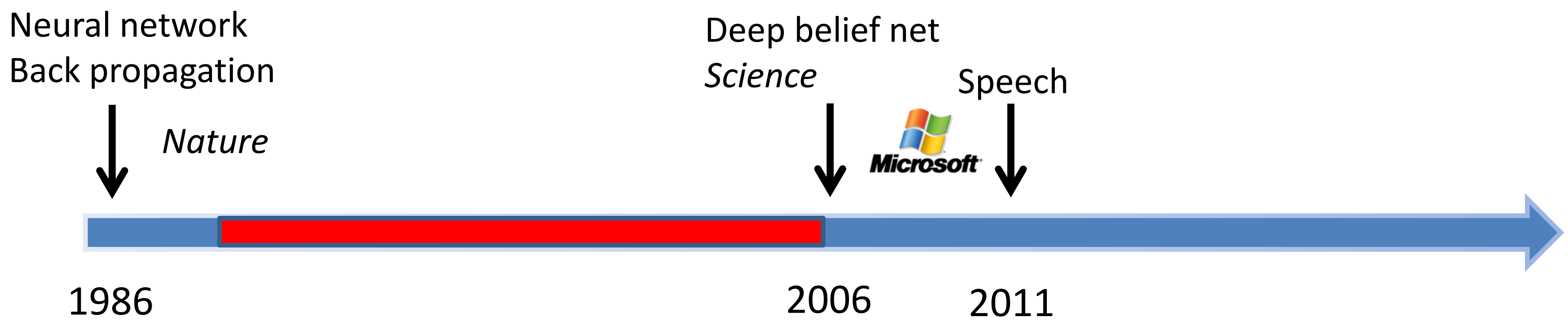
⇩

Blessing of dimensionality

⇩

Learning hierarchical feature transforms
(Learning features with deep structures)

D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: Highdimensional feature and its efficient compression for face verification. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2013.

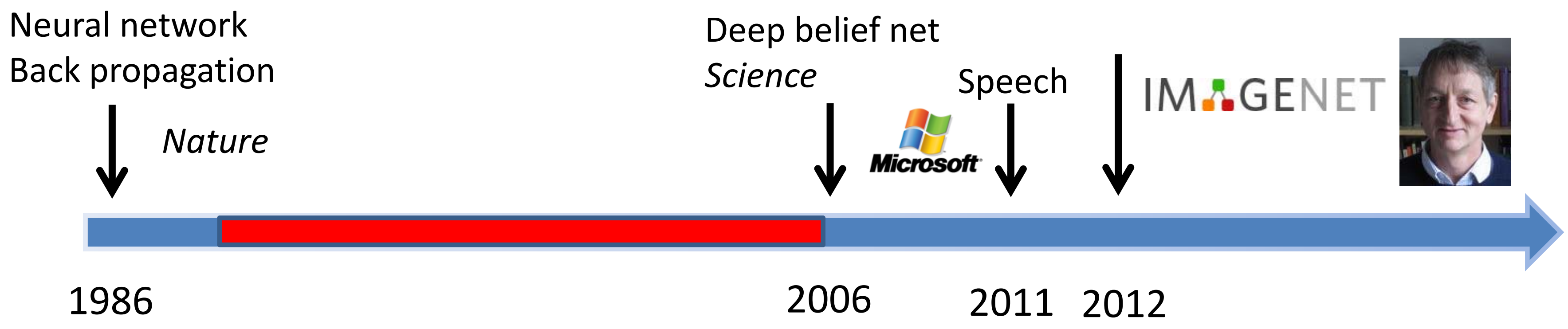


deep learning results

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.

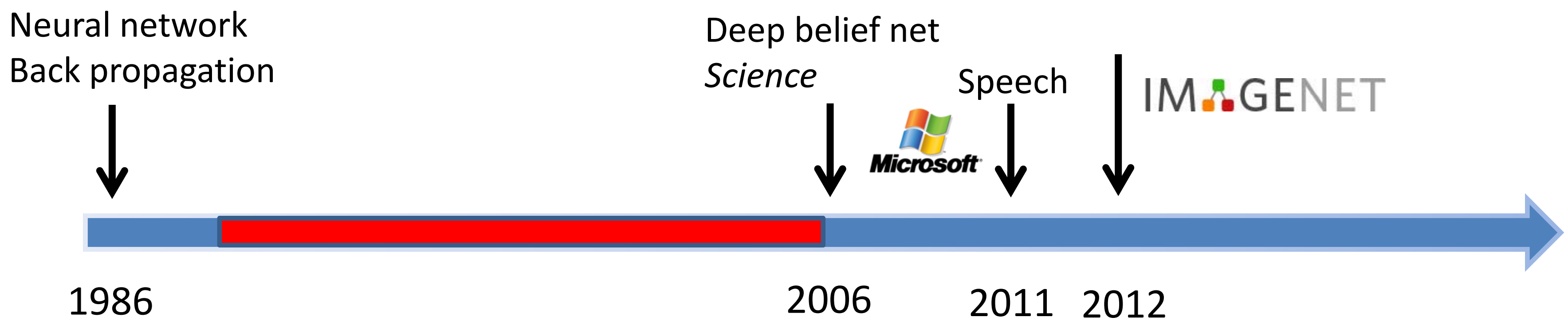


Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

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

A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.

Examples from ImageNet



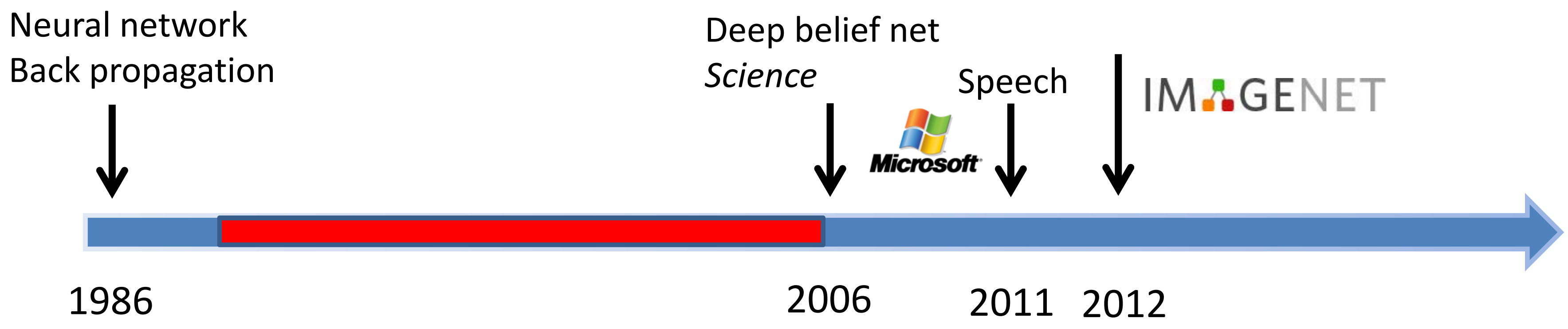
- ImageNet 2013 – image classification challenge

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

- ImageNet 2013 – object detection challenge

Rank	Name	Mean Average Precision	Description
1	UvA-Euvision	0.22581	Hand-crafted features
2	NEC-MU	0.20895	Hand-crafted features
3	NYU	0.19400	Deep learning

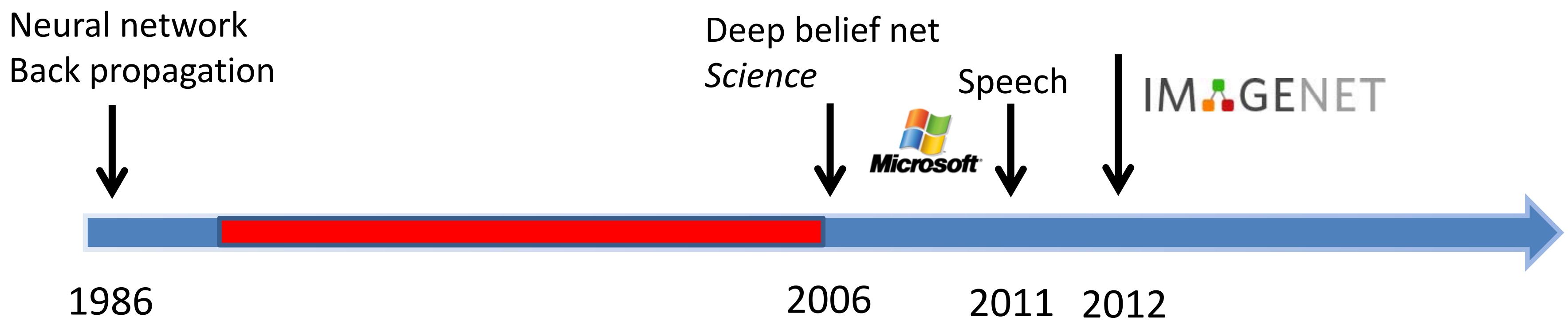


- ImageNet 2014 – Image classification challenge

Rank	Name	Error rate	Description
1	Google	0.06656	Deep learning
2	Oxford	0.07325	Deep learning
3	MSRA	0.08062	Deep learning

- ImageNet 2014 – object detection challenge

Rank	Name	Mean Average Precision	Description
1	Google	0.43933	Deep learning
2	CUHK	0.40656	Deep learning
3	DeepInsight	0.40452	Deep learning
4	UvA-Euvision	0.35421	Deep learning
5	Berkley Vision	0.34521	Deep learning

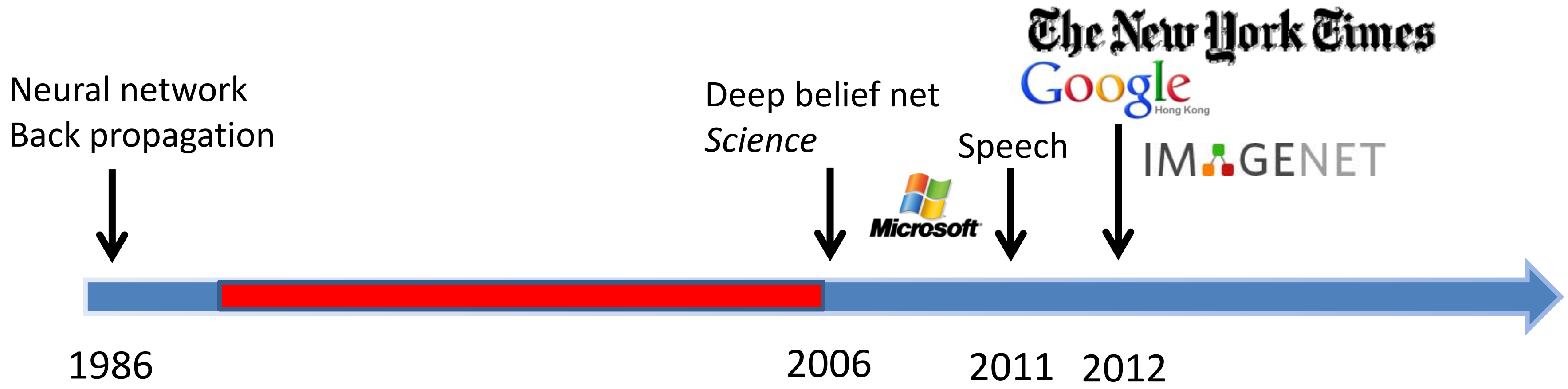


- ImageNet 2014 – object detection challenge

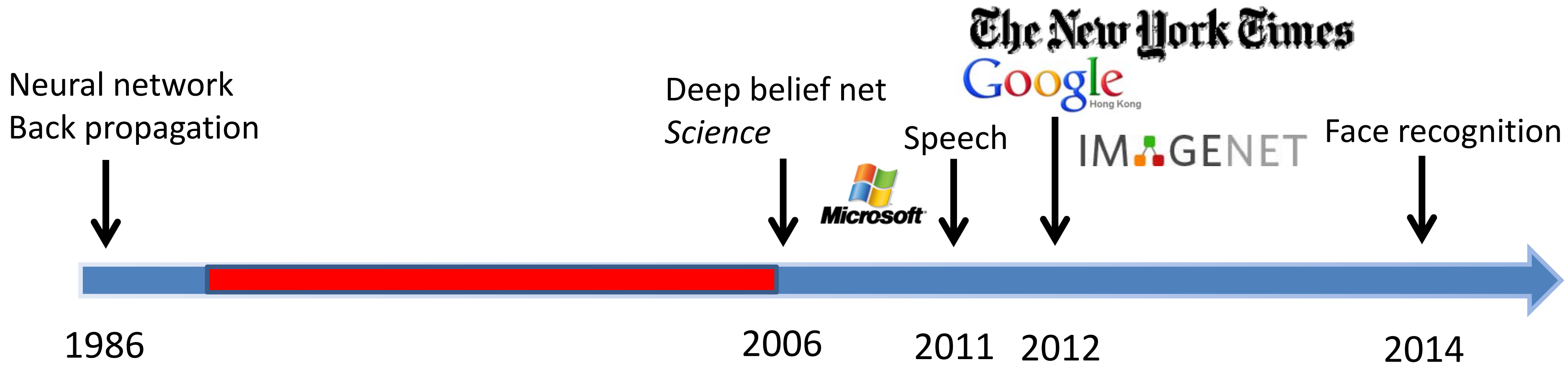
	RCNN (Berkley)	Berkley vision	UvA-Euvision	DeepInsight	GooLeNet (Google)	DeepID-Net (CUHK)
Model average	n/a	n/a	n/a	40.5	43.9	**50.3**
Single model	31.4	34.5	35.4	40.2	38.0	**47.9**

Wanli Ouyang

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



- Google and Baidu announced their deep learning based visual search engines (2013)
  - [Google]()
    - "on our test set we saw **double the average precision** when compared to other approaches we had tried. We acquired the rights to the technology and went full speed ahead adapting it to run at large scale on Google's computers. We took cutting edge research straight out of an academic research lab and launched it, in just a little over six months."
  - [Baidu]()



- Deep learning achieves 99.47% face verification accuracy on Labeled Faces in the Wild (LFW), higher than human performance

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

Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CVPR, 2015.

Labeled Faces in the Wild (2007)

Random guess (50%)

Eigenface (60%)

Best results
without deep learning

MSRA TL Joint Bayesian (96.33%)

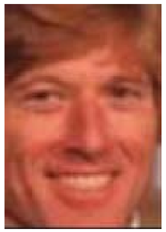
Human cropped (97.53%)

Human funneled (99.20%)

CUHK deep learning result (99.53%)

Google deep learning result (99.6%)

Baidu deep learning result (99.8%)



Unrestricted, Labeled Outside Data Results

Attribute classifiers[11]	0.8525 ± 0.0060
Simile classifiers[11]	0.8414 ± 0.0041
Attribute and Simile classifiers[11]	0.8554 ± 0.0035
Multiple LE + comp[14]	0.8445 ± 0.0046
Associate-Predict[18]	0.9057 ± 0.0056
Tom-vs-Pete[23]	0.9310 ± 0.0135
Tom-vs-Pete + Attribute[23]	0.9330 ± 0.0128
combined Joint Bayesian[26]	0.9242 ± 0.0108
high-dim LBP[27]	0.9517 ± 0.0113
DFD[33]	0.8402 ± 0.0044
TL Joint Bayesian[34]	0.9633 ± 0.0108
face.com r2011b[19]	0.9130 ± 0.0030
Face++[40]	0.9727 ± 0.0065
DeepFace-ensemble[41]	0.9735 ± 0.0025
ConvNet-RBM[42]	0.9252 ± 0.0038
POOF-gradhist[44]	0.9313 ± 0.0040
POOF-HOG[44]	0.9280 ± 0.0047
FR+FCN[45]	0.9645 ± 0.0025
DeepID[46]	0.9745 ± 0.0026
GaussianFace[47]	0.9852 ± 0.0066
DeepID2[48]	0.9915 ± 0.0013

Table 6: Mean classification accuracy $\hat{u}$ and standard error of the mean S_E.

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. →

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous. →

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child? →

Additive Manufacturing

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts. →

Baxter: The Blue-Collar Robot

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people. →

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss.

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket.

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible.

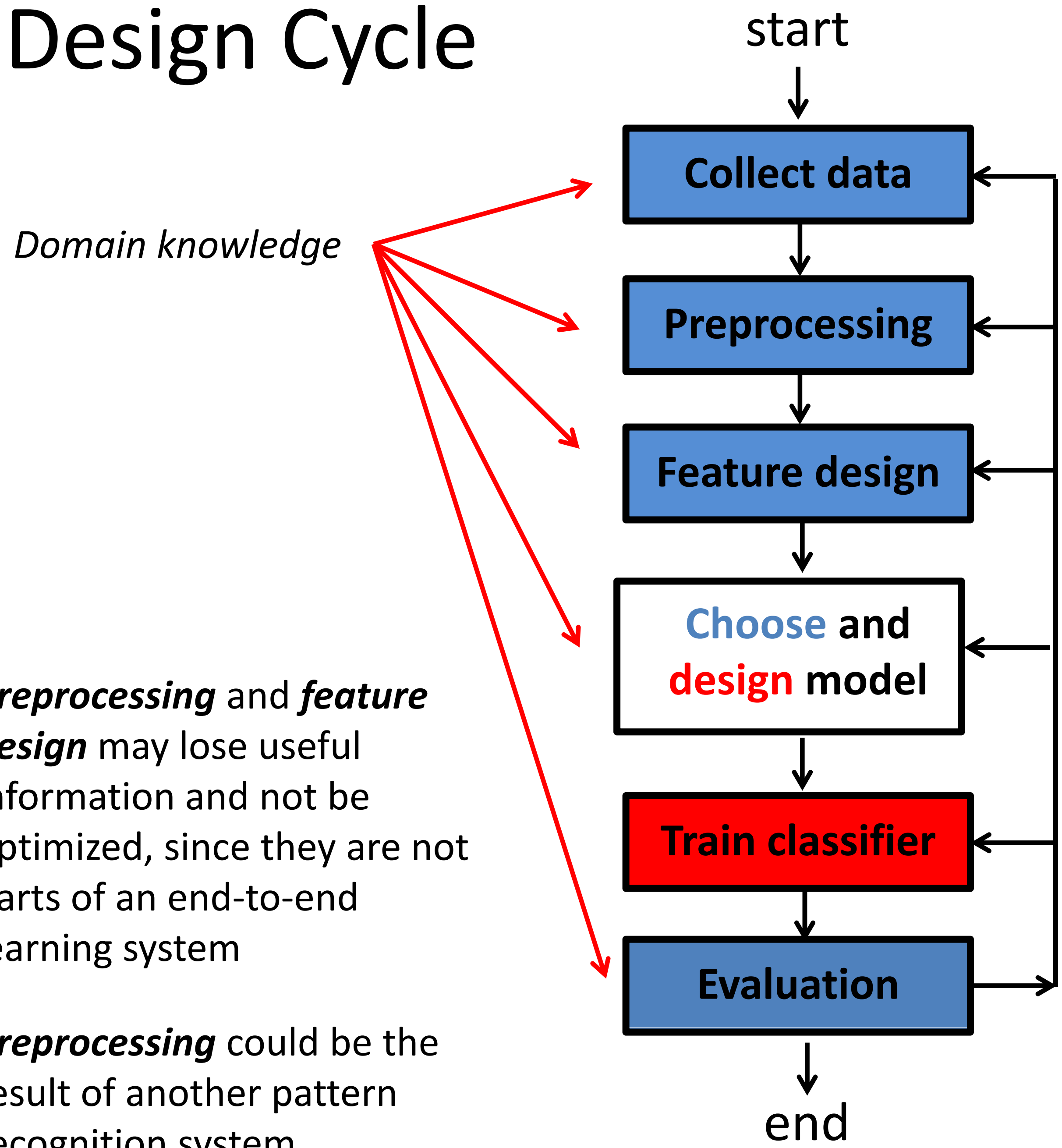
Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave – and even help us understand the spread of diseases.

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical.

Design Cycle



Interest of people working on **computer vision, speech recognition, medical image processing,…**

Interest of people working on **machine learning**

Interest of people working on **machine learning** and **computer vision, speech recognition, medical image processing,…**

Preprocessing and ***feature design*** may lose useful information and not be optimized, since they are not parts of an end-to-end learning system

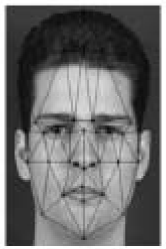
Preprocessing could be the result of another pattern recognition system

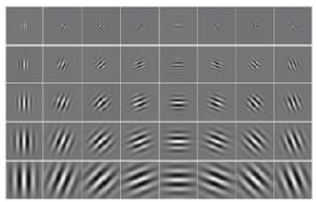
Person re-identification pipeline

Pedestrian detection → Pose estimation → Body parts segmentation → Photometric & geometric transform → Feature extraction → Classification

Face recognition pipeline

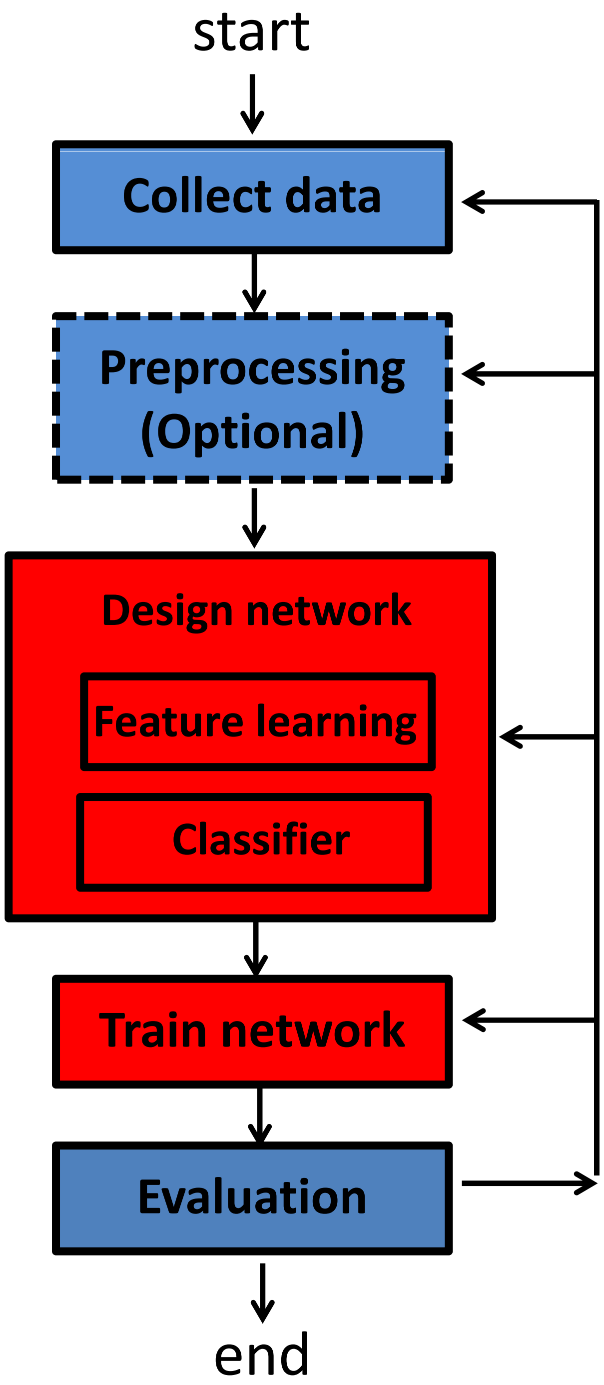
Face alignment → Geometric rectification → Photometric rectification → Feature extraction → Classification



Design Cycle with Deep Learning

- Learning plays a bigger role in the design circle
- Feature learning becomes part of the end-to-end learning system
- Preprocessing becomes optional means that several pattern recognition steps can be merged into one end-to-end learning system
- Feature learning makes the key difference
- We underestimated the importance of data collection and evaluation



What makes deep learning successful in computer vision?

Li Fei-Fei

IMAGENET

Geoffrey Hinton

Data collection

One million images with labels

Evaluation task

Predict 1,000 image categories

Deep learning

CNN is not new

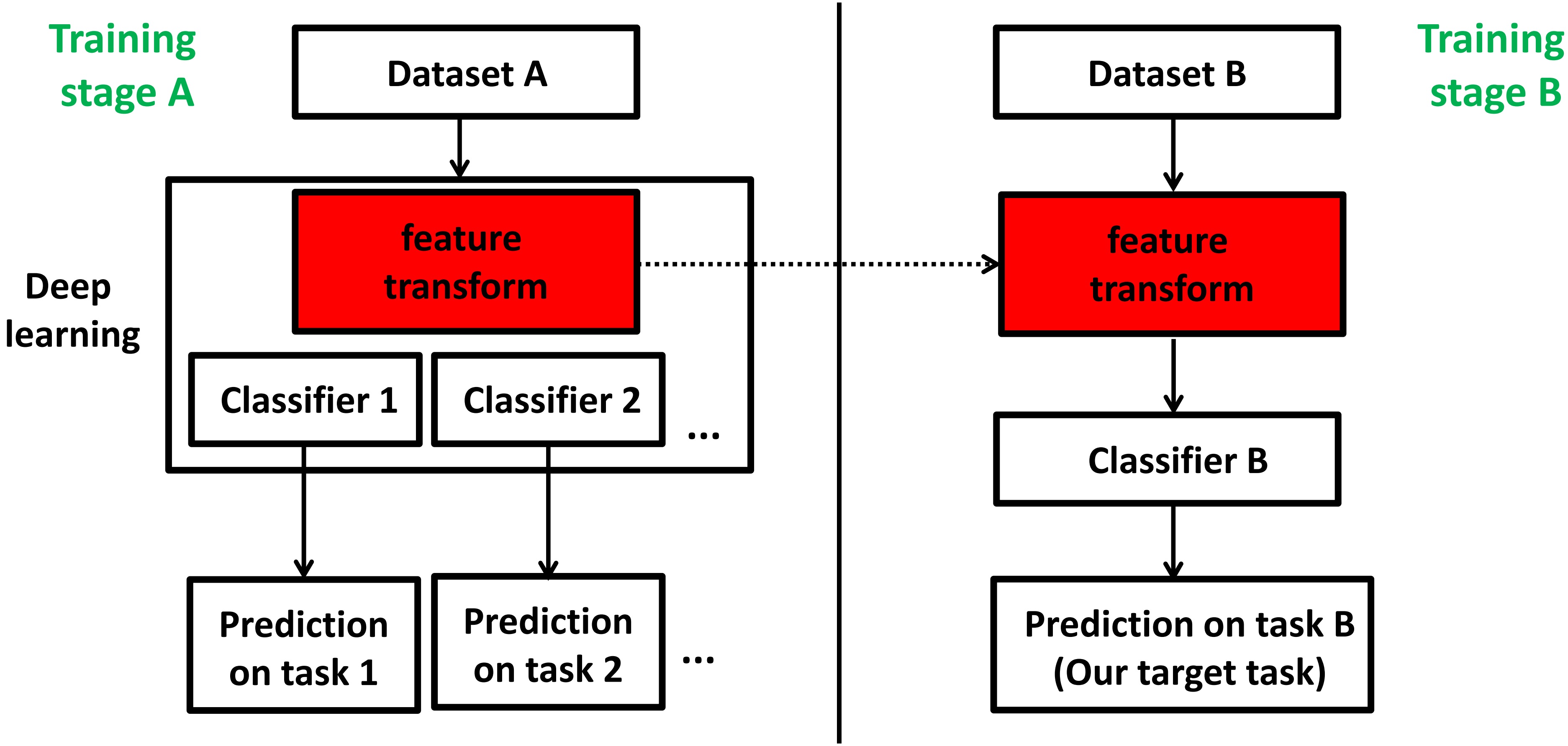
Design network structure

New training strategies

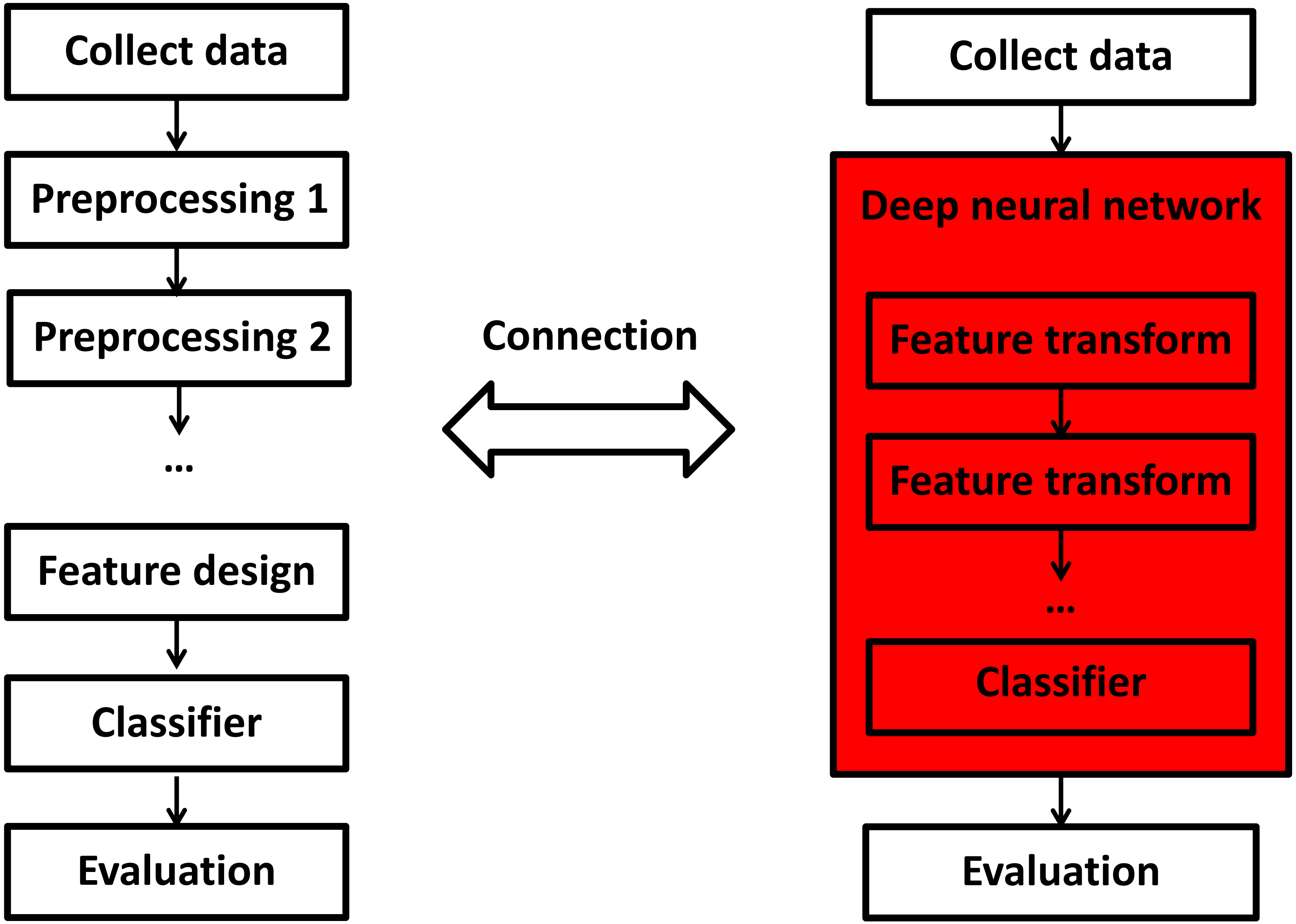
Feature learned from ImageNet can be well generalized to other tasks and datasets!

Learning features and classifiers separately

- Not all the datasets and prediction tasks are suitable for learning features with deep models



Deep learning can be treated as a language to described the world with great flexibility



Introduction to Deep Learning

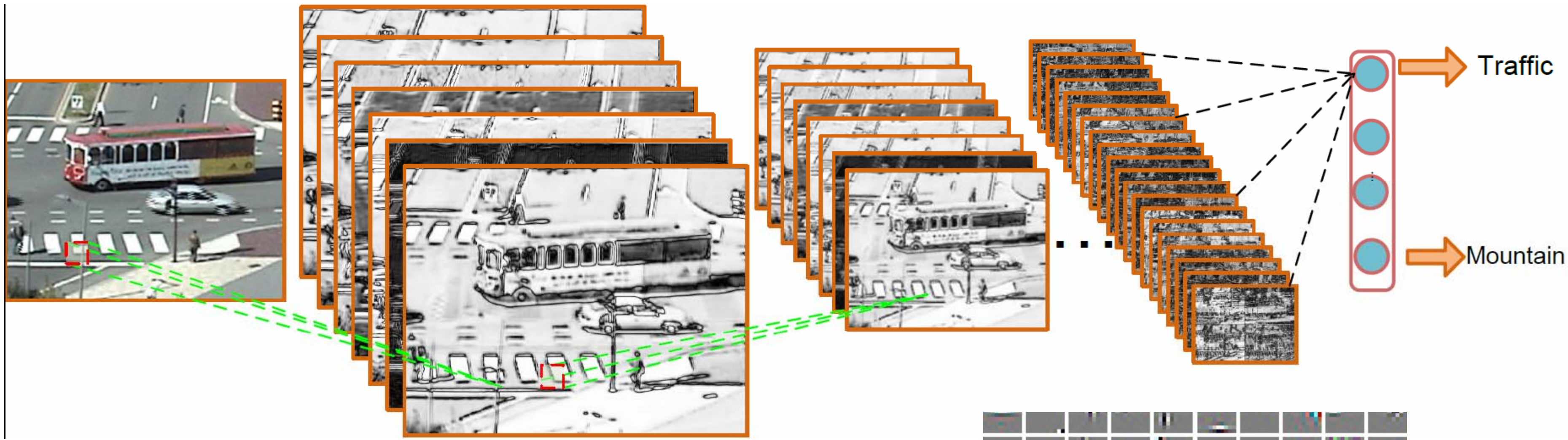
- Historical review of deep learning
- **Introduction to classical deep models**
- Why does deep learning work?
- Properties of deep feature representations

Introduction on Classical Deep Models

- Convolutional Neural Networks (CNN)
  - Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based Learning Applied to Document Recognition,” Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.
- Deep Belief Net (DBN)
  - G. E. Hinton, S. Osindero, and Y. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” Neural Computation, Vol. 18, pp. 1527-1544, 2006.
- Auto-encoder
  - G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” Science, Vol. 313, pp. 504-507, July 2006.

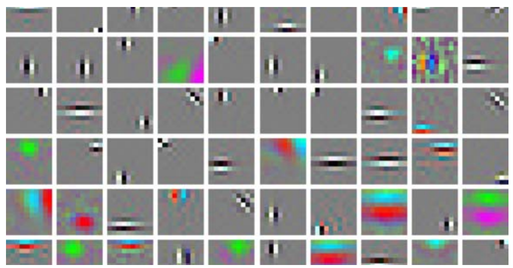
Classical Deep Models

- Convolutional Neural Networks (CNN)
  - First proposed by Fukushima in 1980
  - Improved by LeCun, Bottou, Bengio and Haffner in 1998



Convolution

Pooling

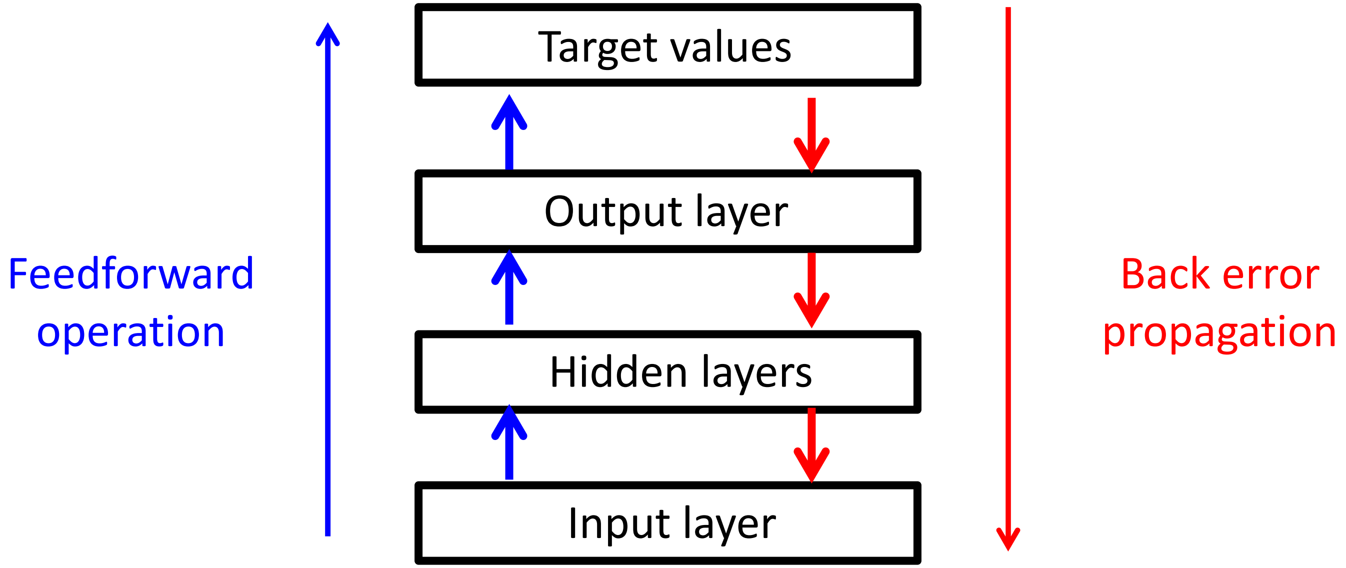


Learned filters

Backpropagation

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \bigtriangledown J(\mathbf{W})$$

W is the parameter of the network; *J* is the objective function

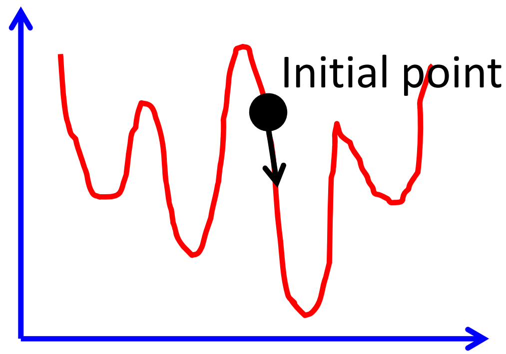


D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-propagation Errors," Nature, Vol. 323, pp. 533-536, 1986.

Classical Deep Models

- Deep belief net
  - Hinton'06

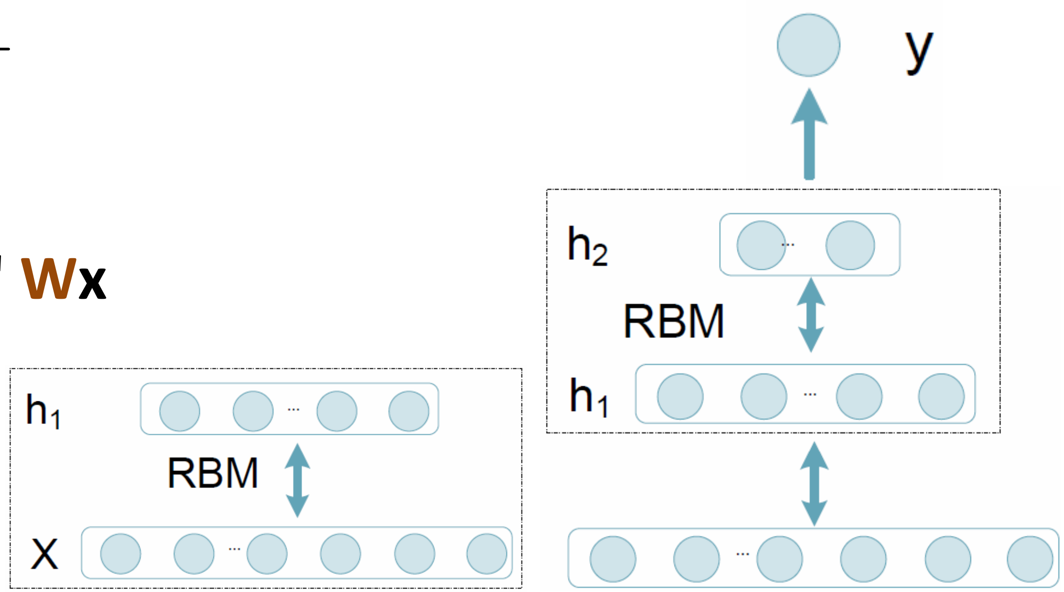
Pre-training:
- Good initialization point
- Make use of unlabeled data



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

$$P(\mathbf{x},\mathbf{h_1}) = \frac{e^{-E(\mathbf{x},\mathbf{h}_1)}}{\sum_{\mathbf{x},\mathbf{h}_1} e^{-E(\mathbf{x},\mathbf{h}_1)}}$$

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



Classical Deep Models

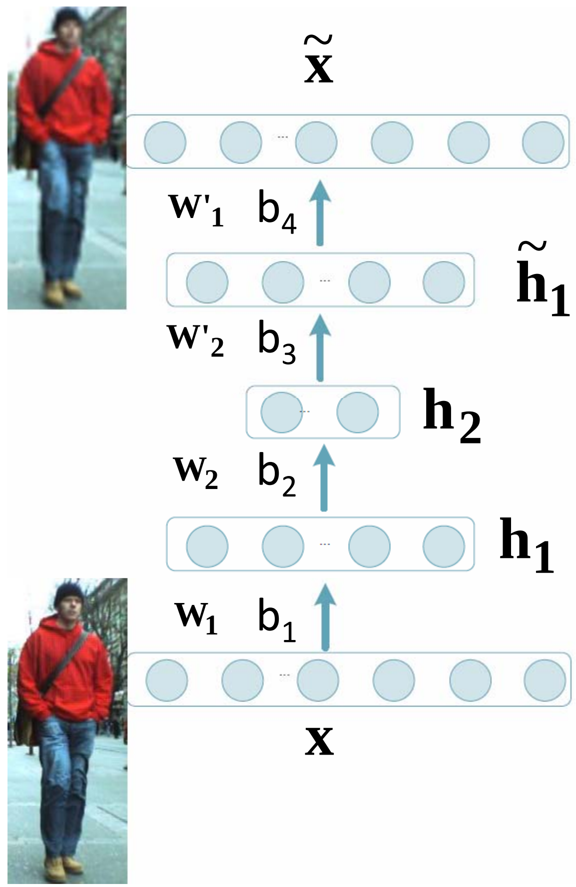
- Auto-encoder
  - Hinton and Salakhutdinov 2006

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

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



Introduction to Deep Learning

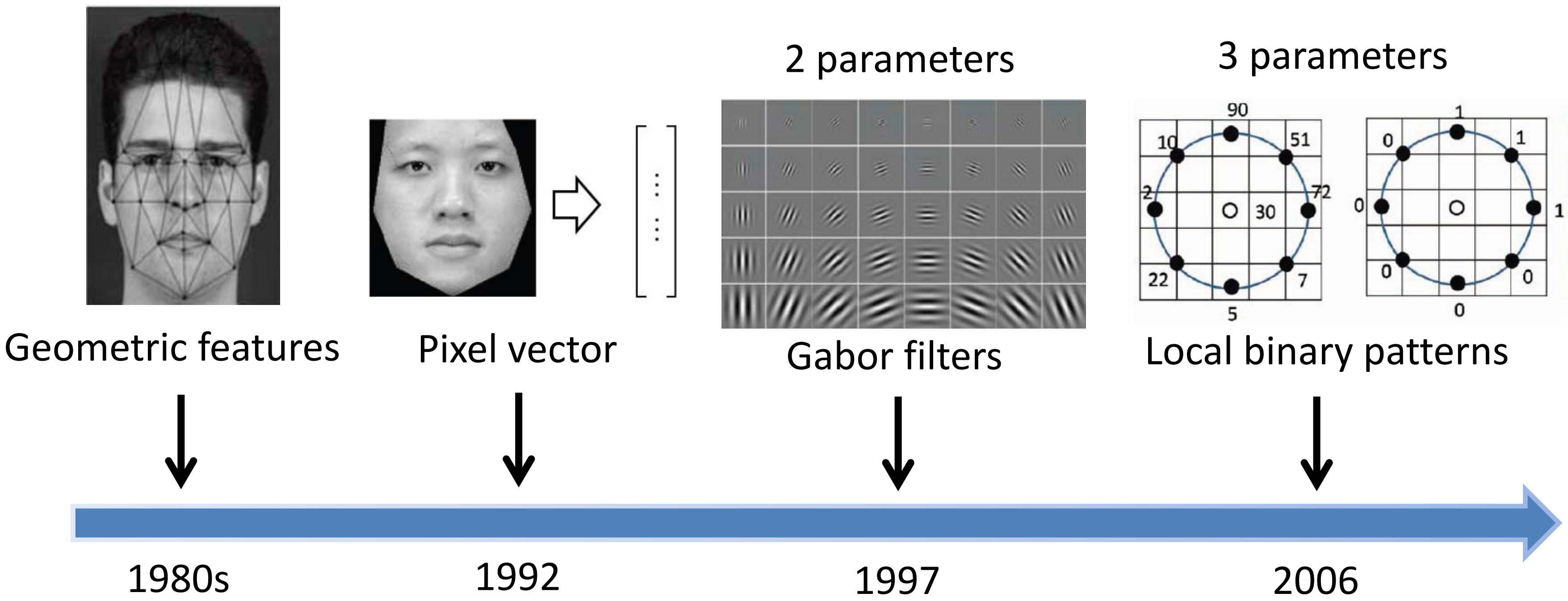
- Historical review of deep learning
- Introduction to classical deep models
- **Why does deep learning work?**
- Properties of deep feature representations

Feature Learning vs Feature Engineering

Feature Engineering

- The performance of a pattern recognition system heavily depends on feature representations
- Manually designed features dominate the applications of image and video understanding in the past
  – Reply on human domain knowledge much more than data
  – Feature design is separate from training the classifier
  – If handcrafted features have multiple parameters, it is hard to manually tune them
  – Developing effective features for new applications is slow

Handcrafted Features for Face Recognition



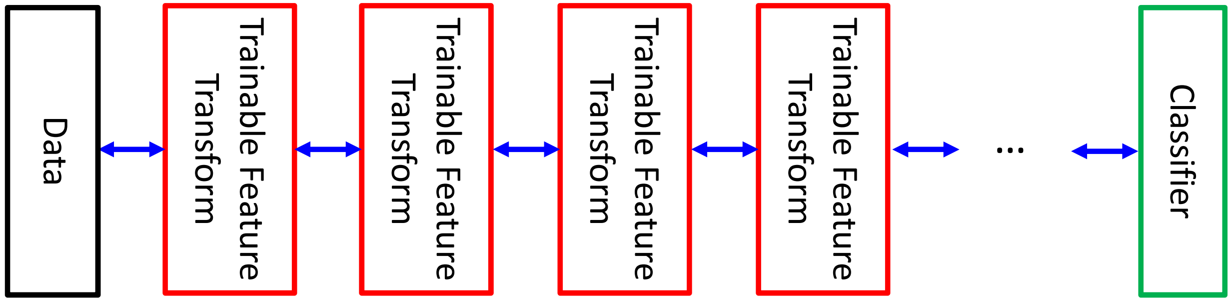
Feature Learning

- Learning transformations of the data that make it easier to extract useful information when building classifiers or predictors
  - Jointly learning feature transformations and classifiers makes their integration optimal
  - Learn the values of a huge number of parameters in feature representations
  - Faster to get feature representations for new applications
  - Make better use of big data

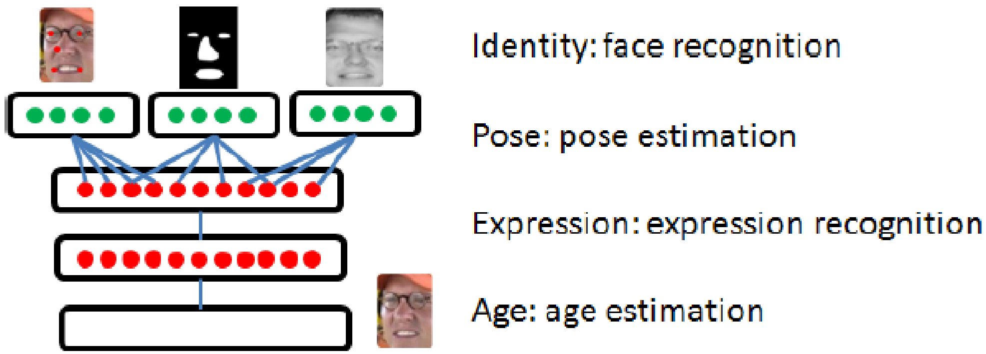
Deep Learning Means Feature Learning

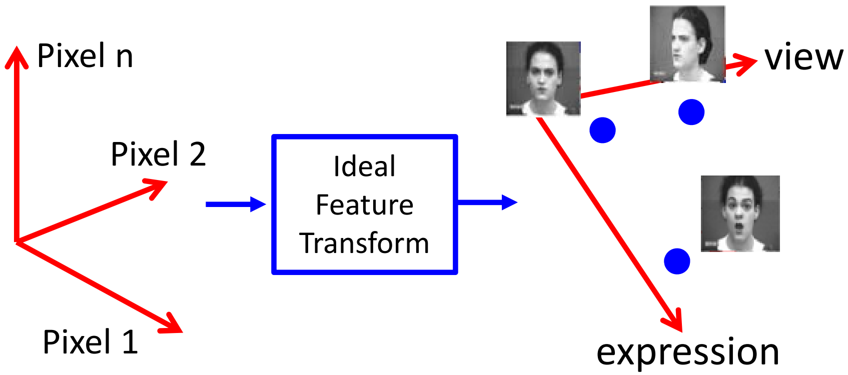
- Deep learning is about learning hierarchical feature representations

$$\mathbf{y} = F(\mathbf{W}^k \cdot F(\mathbf{W}^{k-1} \cdot F(\ldots F(\mathbf{W}^0 \cdot \mathbf{x}))$$



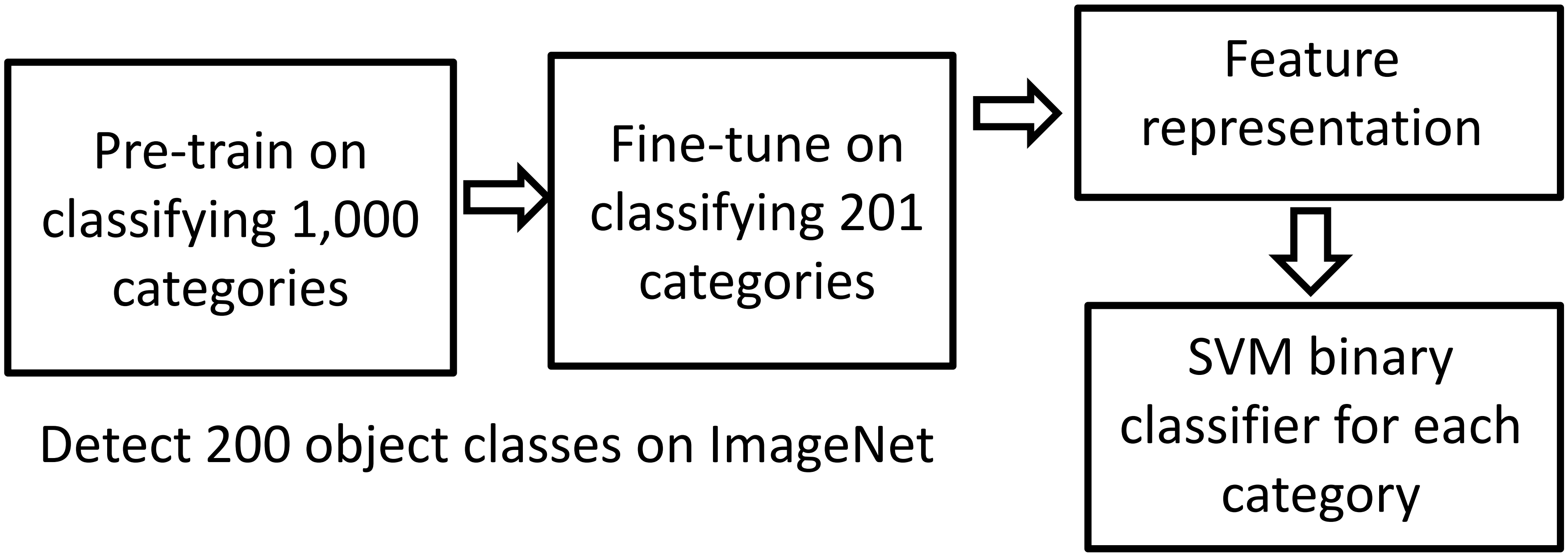
- Good feature representations should be able to disentangle multiple factors coupled in the data





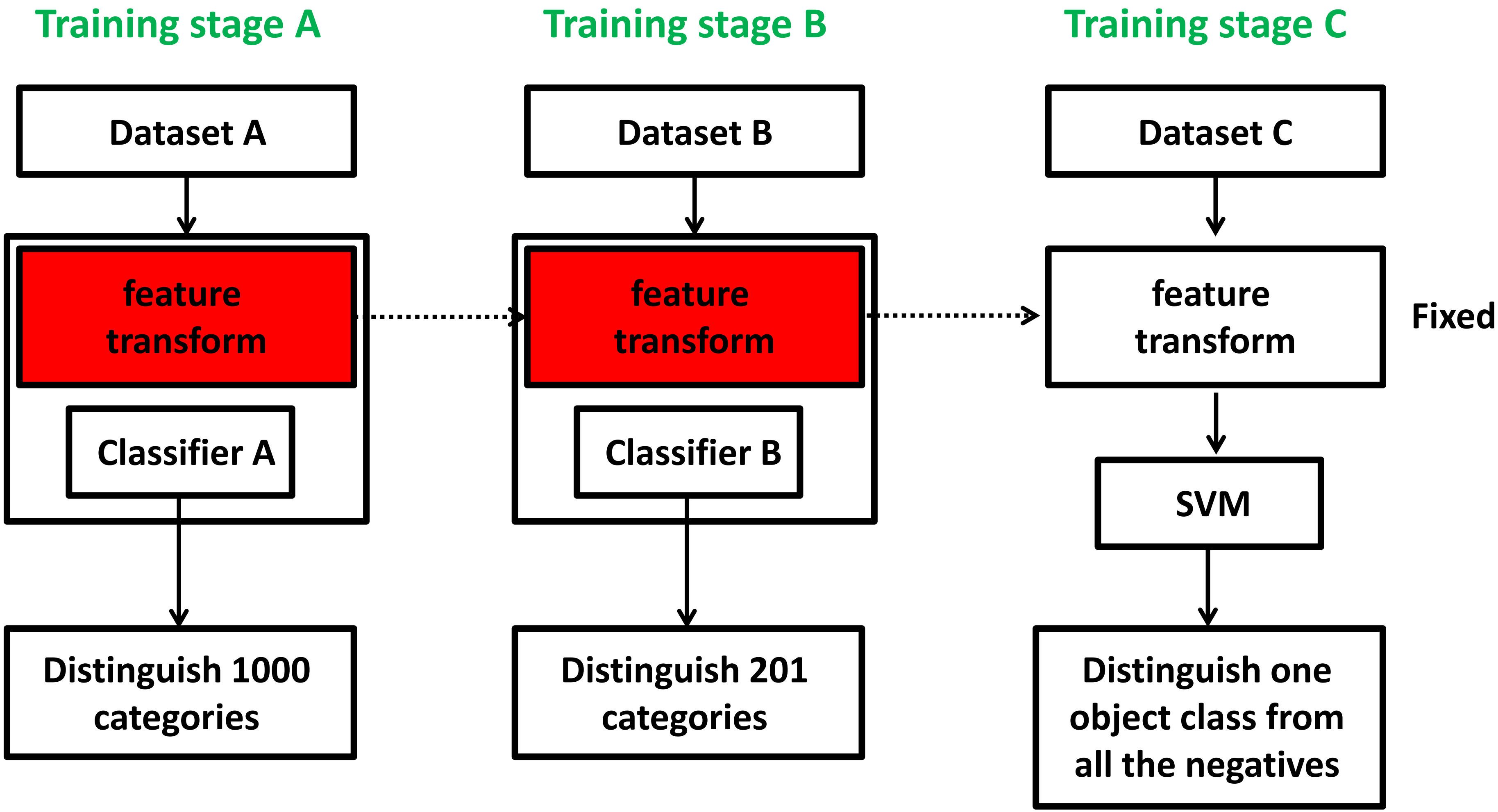
Deep Learning Means Feature Learning

- How to effectively learn features with deep models
  - With challenging tasks
  - Predict high-dimensional vectors



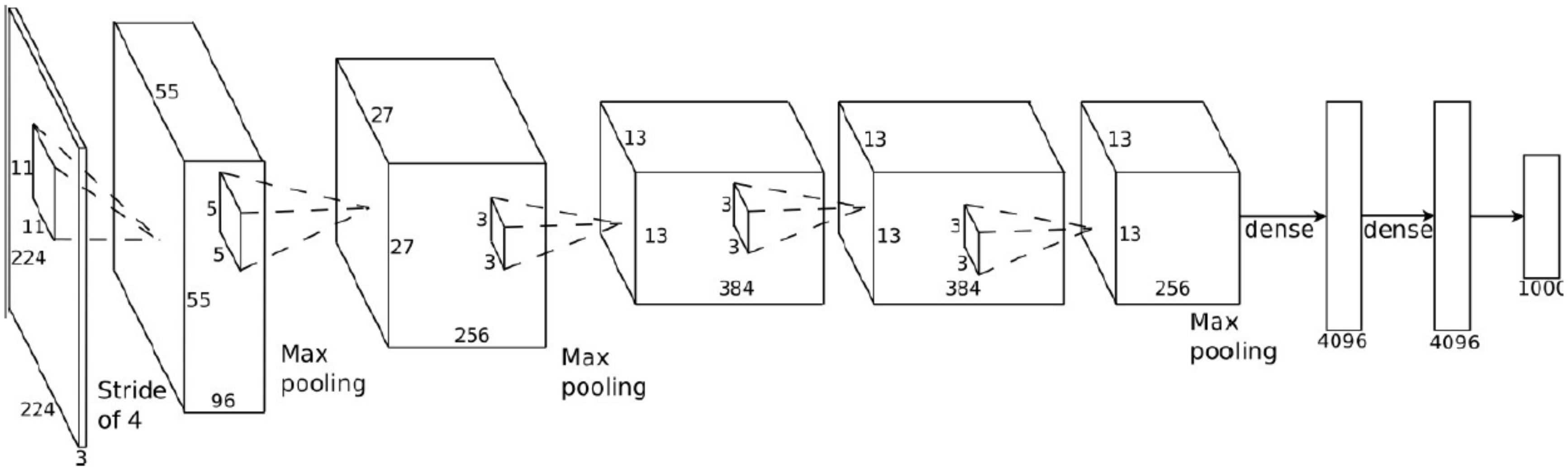
Detect 200 object classes on ImageNet

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


Training stage A
Training stage B
Training stage C
Dataset A
Dataset B
Dataset C
feature transform
feature transform
feature transform
Fixed
Classifier A
Classifier B
SVM
Distinguish 1000 categories
Distinguish 201 categories
Distinguish one object class from all the negatives

Example 1: deep learning generic image features

- Hinton group's groundbreaking work on ImageNet
  - They did not have much experience on general image classification on ImageNet
  - It took one week to train the network with 60 Million parameters
  - The learned feature representations are effective on other datasets (e.g. Pascal VOC) and other tasks (object detection, segmentation, tracking, and image retrieval)



96 learned low-level filters

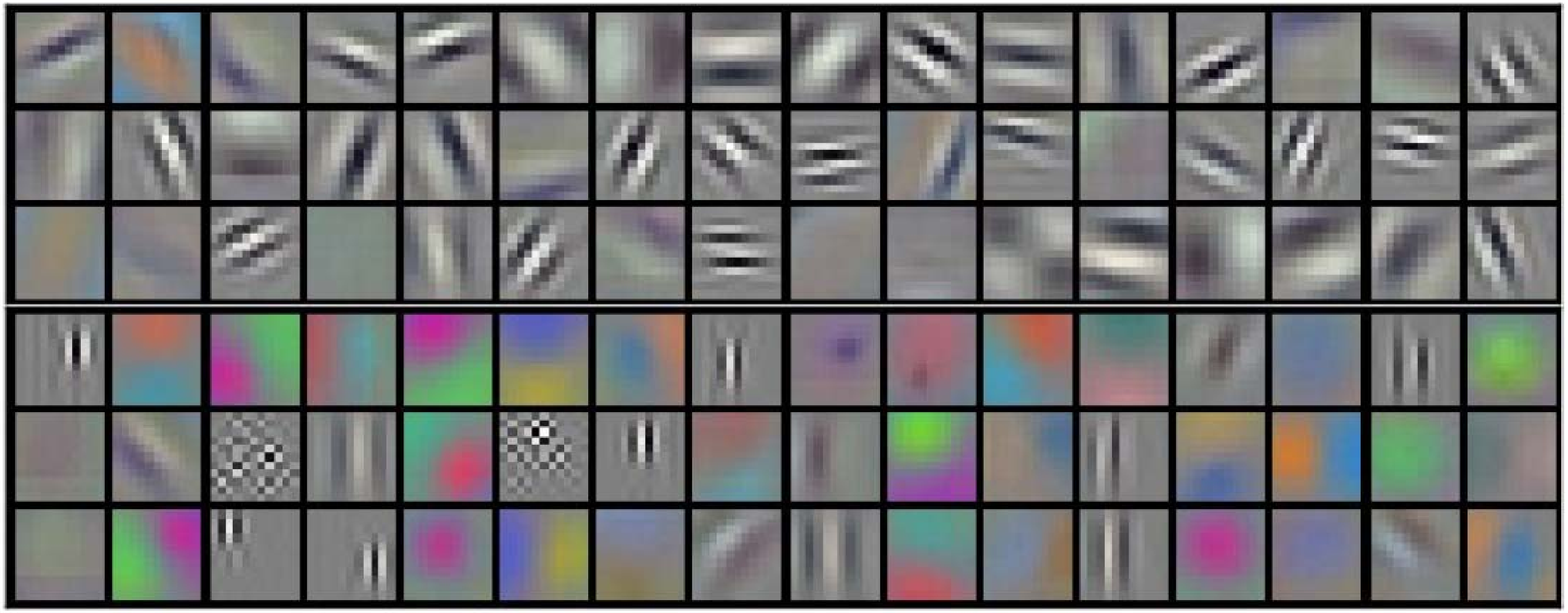
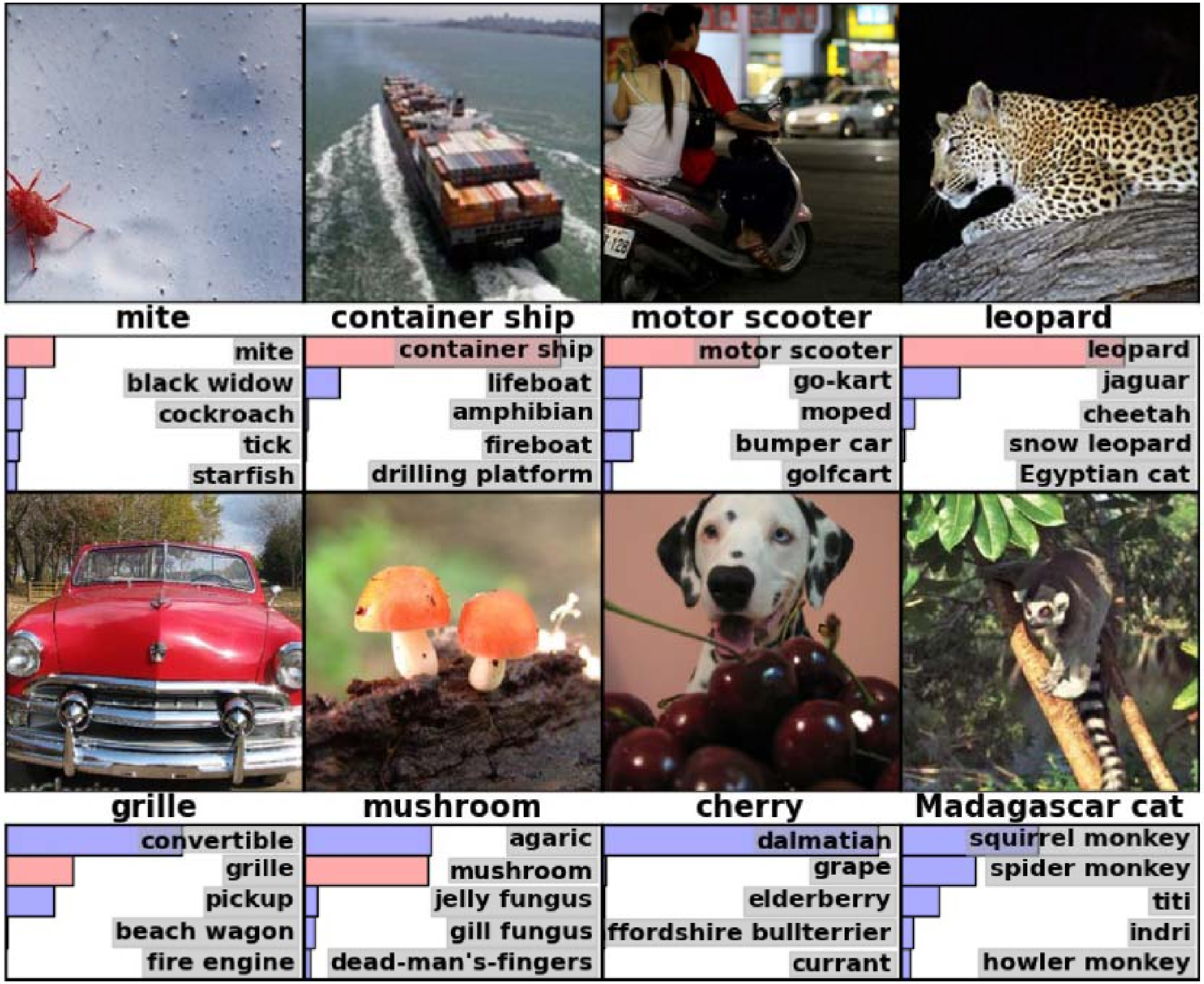
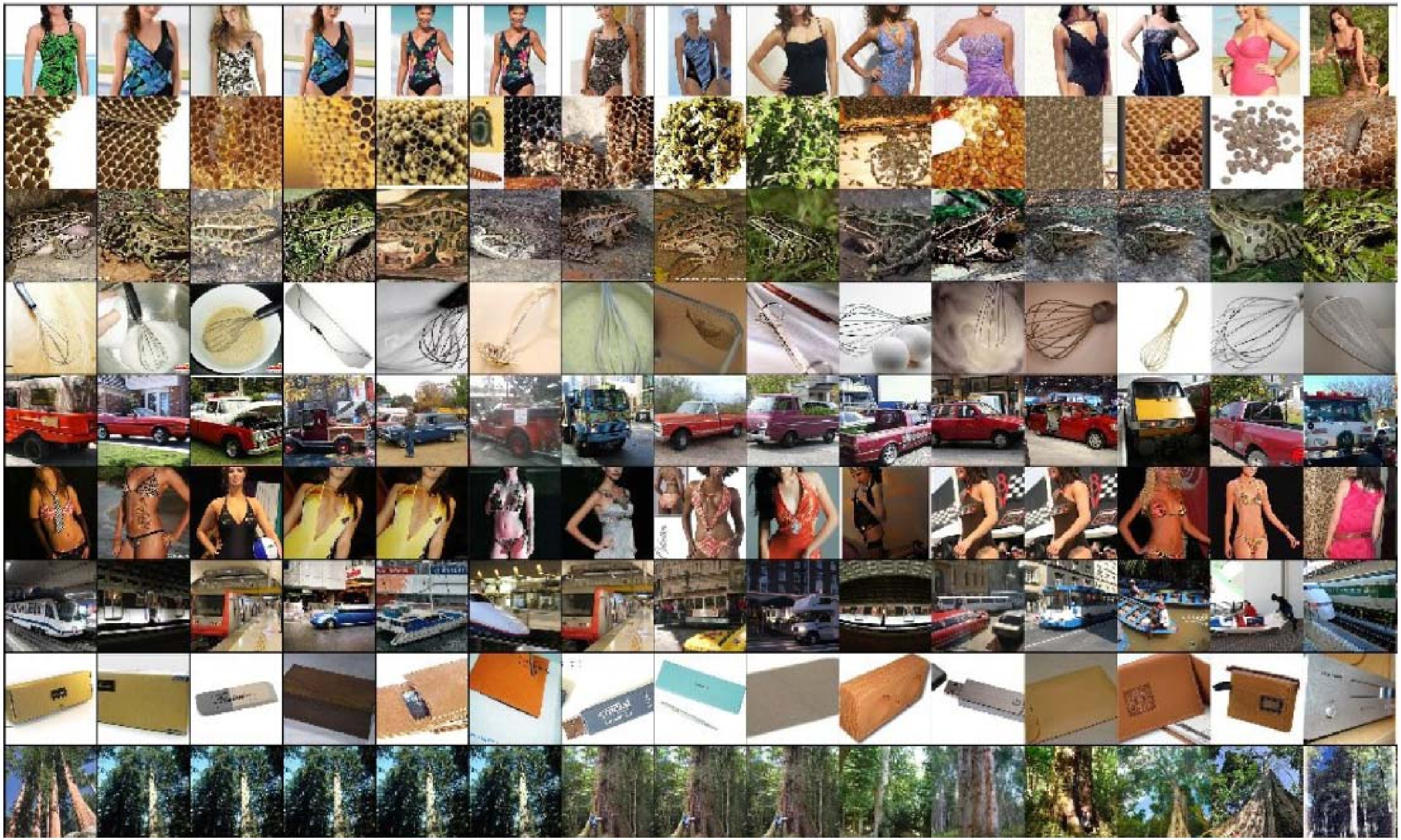


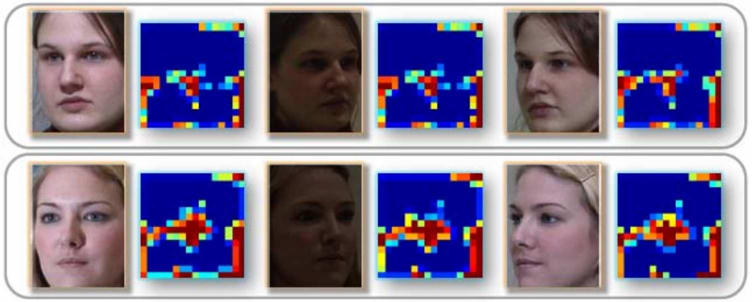
Image classification result

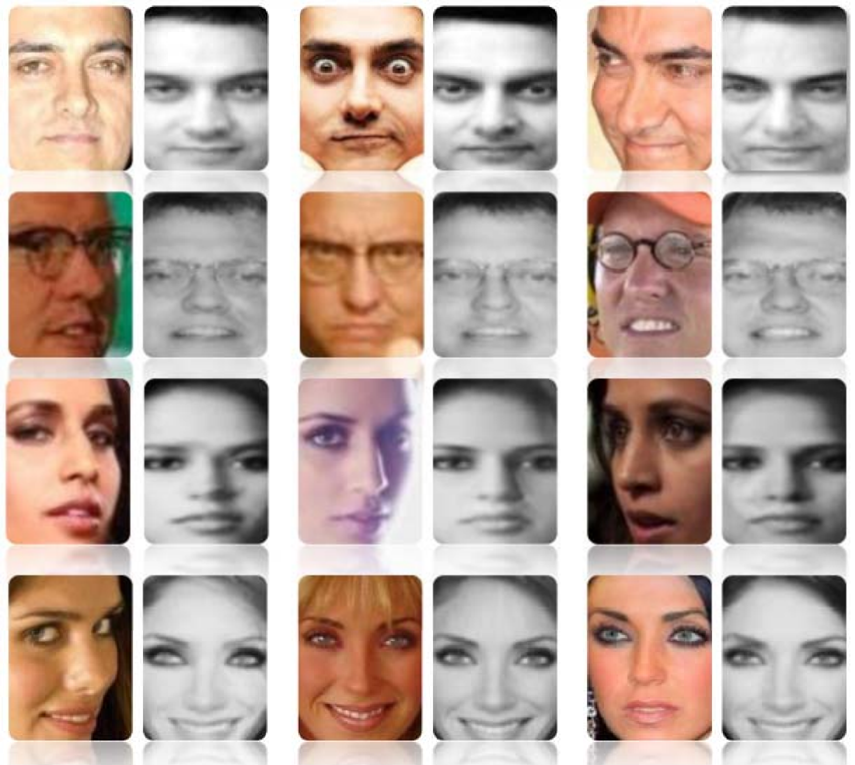


Top hidden layer can be used as feature for retrieval



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

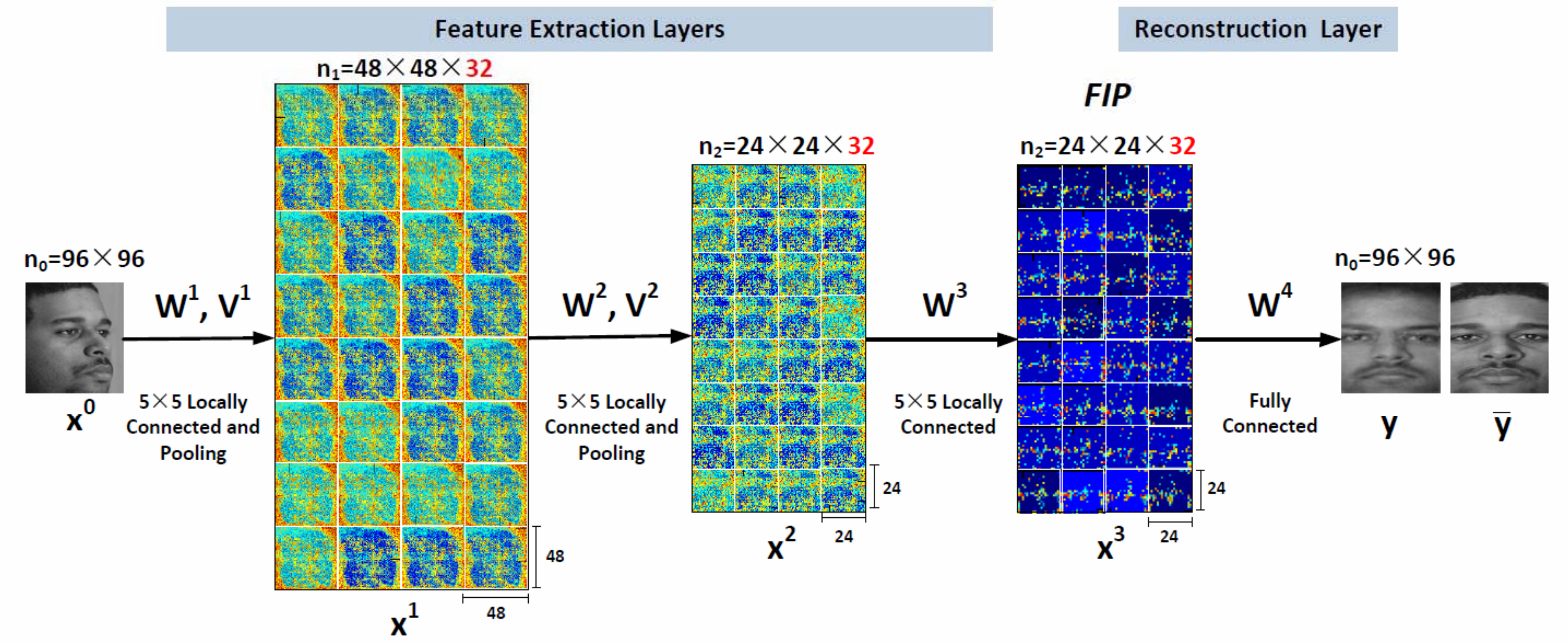




Reconstruction examples from LFW

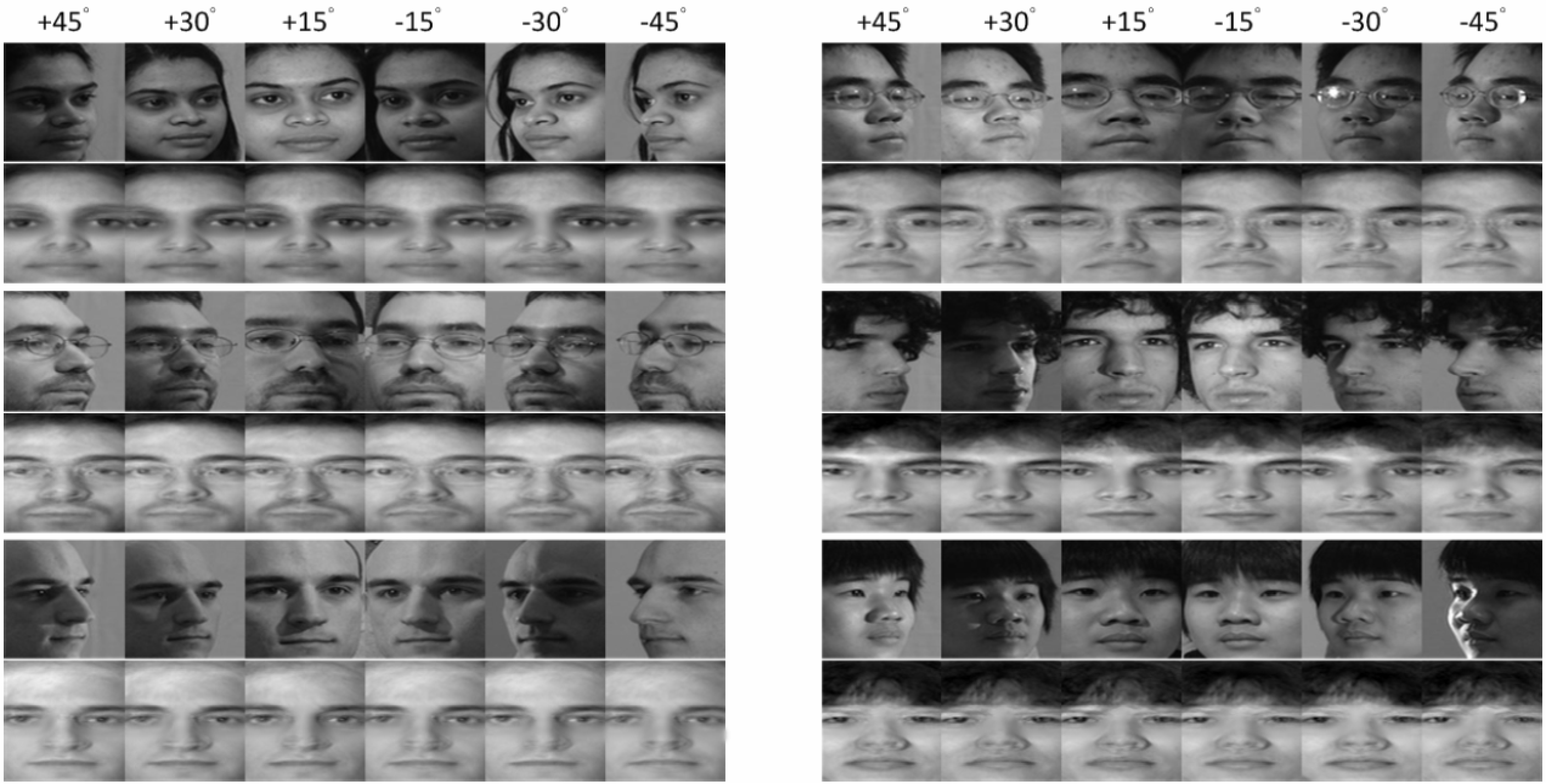
Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Identity Preserving Face Space,” ICCV 2013.

- 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

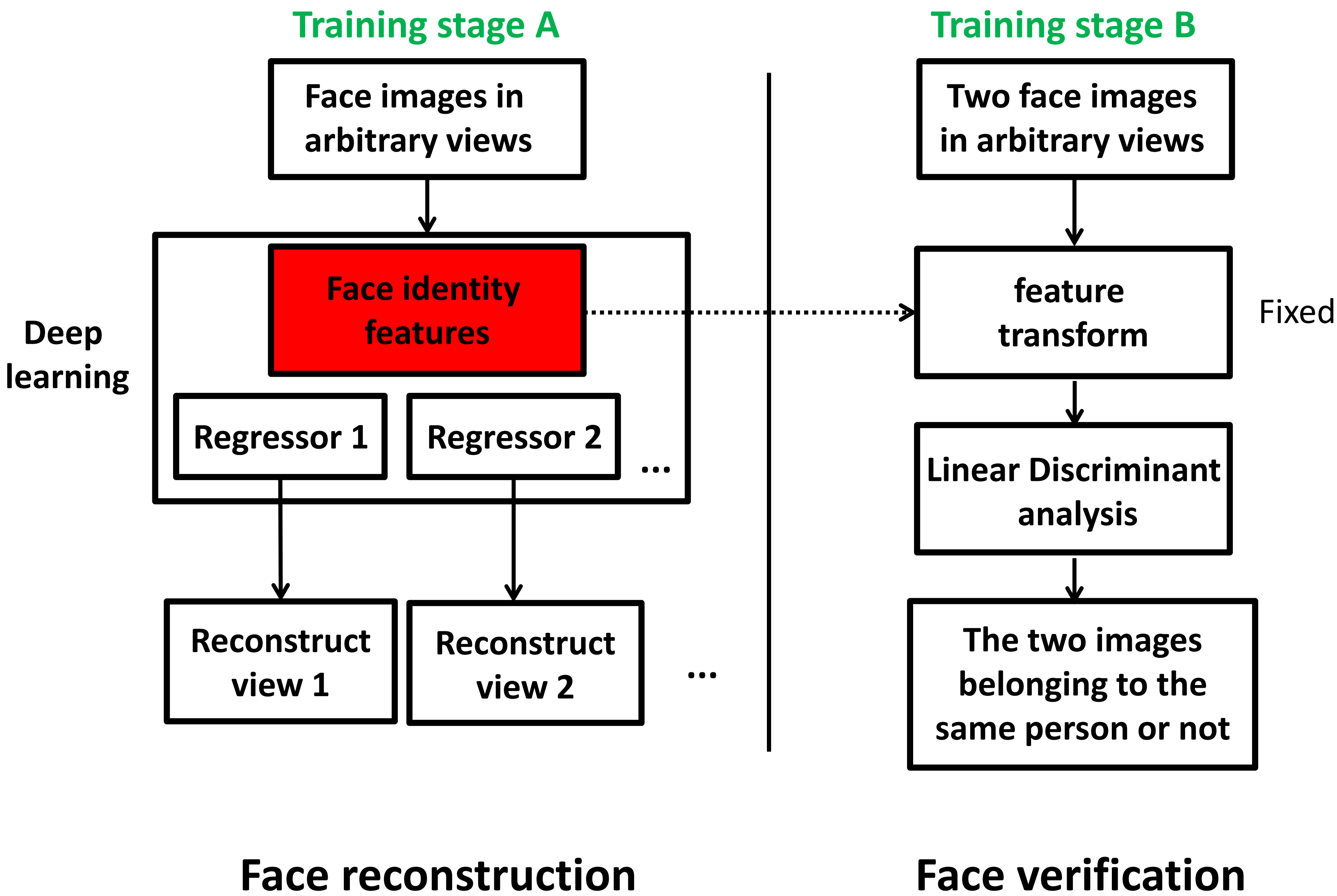
[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012. 3, 6

[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

Deep learning 3D model from 2D images, mimicking human brain activities

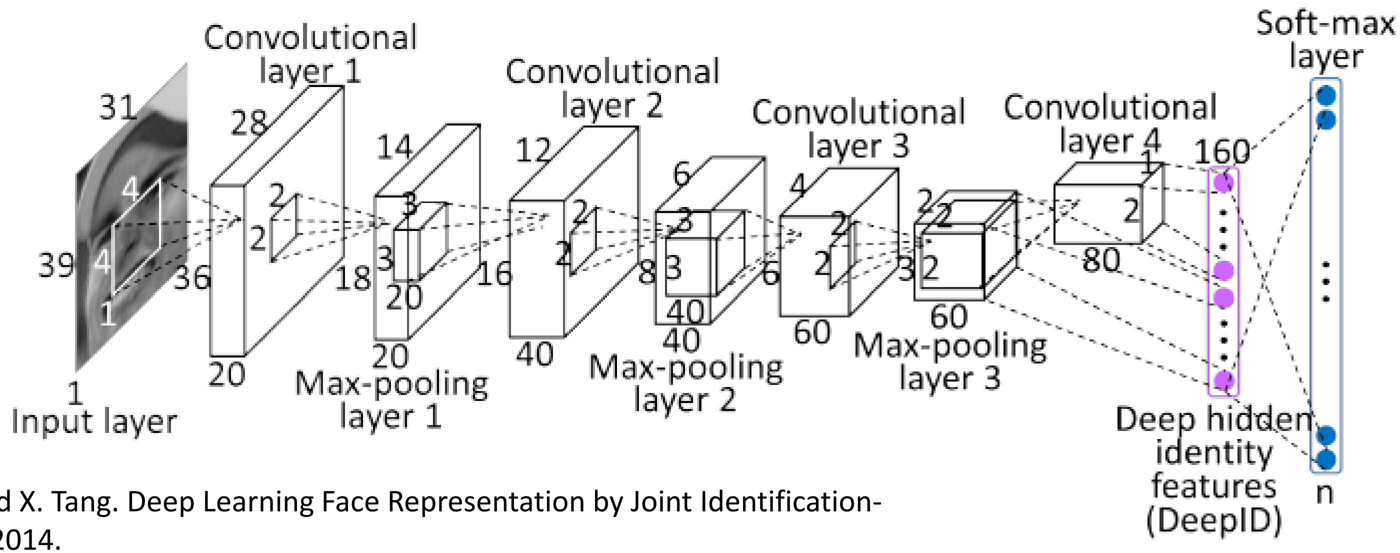
Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning and Disentangling Face Representation by Multi-View Perception,” NIPS 2014.



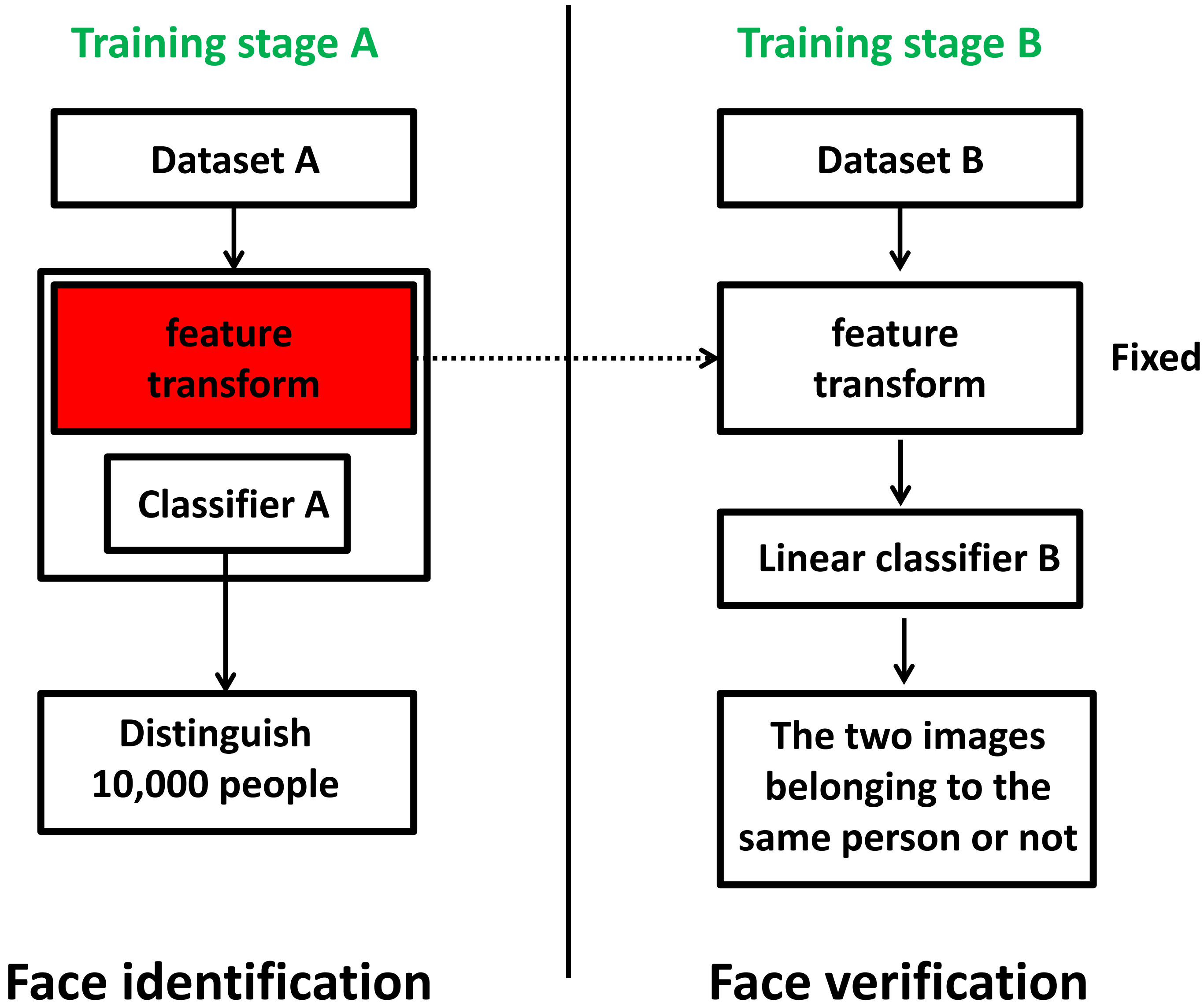
Training stage A
Face images in arbitrary views
Face identity features
Deep learning
Regressor 1
Regressor 2
...
Reconstruct view 1
Reconstruct view 2
...
Training stage B
Two face images in arbitrary views
feature transform
Fixed
Linear Discriminant analysis
The two images belonging to the same person or not
Face reconstruction
Face verification

Example 3: deep learning face identity features from predicting 10,000 classes

- At training stage, each input image is classified into 10,000 identities with 160 hidden identity features in the top layer
- The hidden identity features can be well generalized to other tasks (e.g. verification) and identities outside the training set
- As adding the number of classes to be predicted, the generalization power of the learned features also improves



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

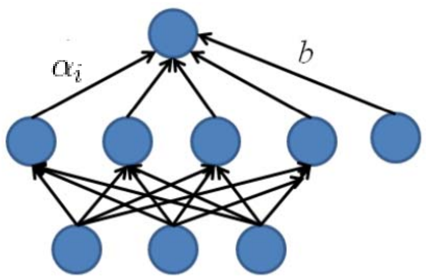

Training stage A
Dataset A
feature transform
Classifier A
Distinguish 10,000 people
Face identification
Training stage B
Dataset B
feature transform
Fixed
Linear classifier B
The two images belonging to the same person or not
Face verification

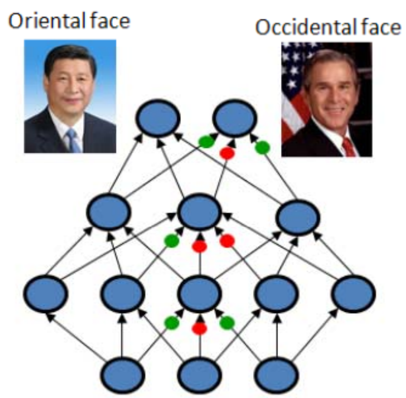
Deep Structures vs Shallow Structures
(Why deep?)

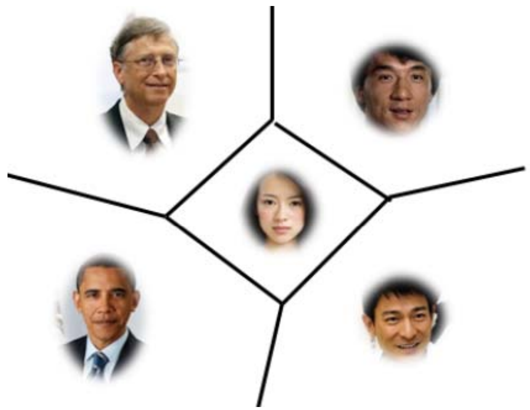
Shallow Structures

- A three-layer neural network (with one hidden layer) can approximate any classification function
- Most machine learning tools (such as SVM, boosting, and KNN) can be approximated as neural networks with one or two hidden layers
- Shallow models divide the feature space into regions and match templates in local regions. O(N) parameters are needed to represent N regions

SVM $g(x) = b + \sum_i \alpha_i K(x, x_i)$







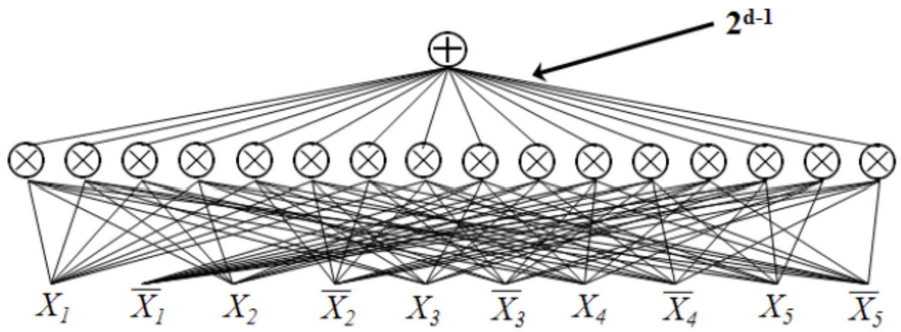
Deep Machines are More Efficient for Representing Certain Classes of Functions

- Theoretical results show that an architecture with insufficient depth can require many more computational elements, potentially exponentially more (with respect to input size), than architectures whose **depth is matched to the task** (Hastad 1986, Hastad and Goldmann 1991)
- It also means many more parameters to learn

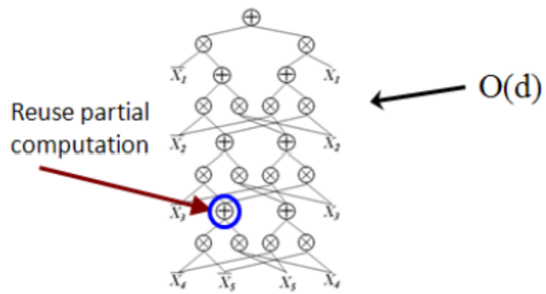
- Take the d-bit parity function as an example

$$(X_1, \ldots, X_d) \in \{0,1\}^d \longmapsto \begin{cases} 1, & \text{if } \sum_{i=1}^{d} X_i \text{ is even} \\ -1, & \text{otherwise} \end{cases}$$

- d-bit logical parity circuits of depth 2 have exponential size (Andrew Yao, 1985)



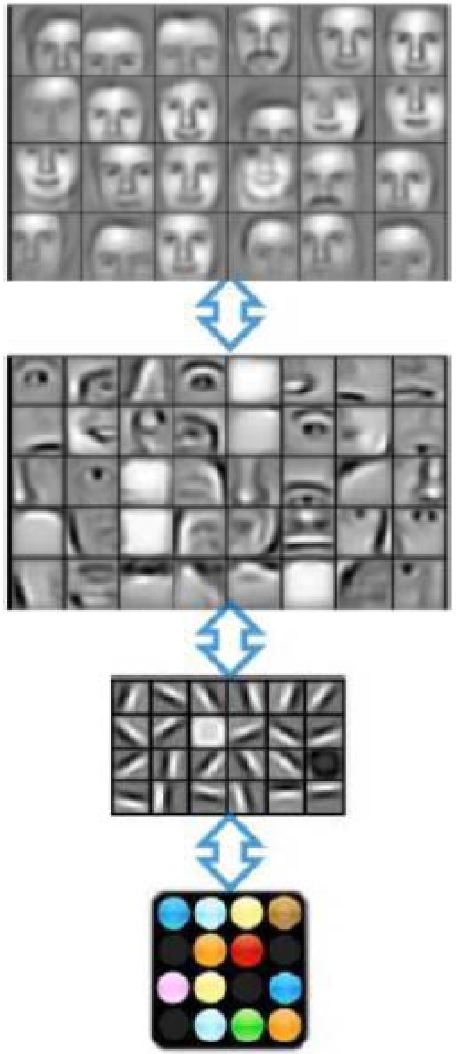
Shallow structure



Deep structure

- There are functions computable with a polynomial-size logic gates circuits of depth k that require exponential size when restricted to depth k -1 (Hastad, 1986)

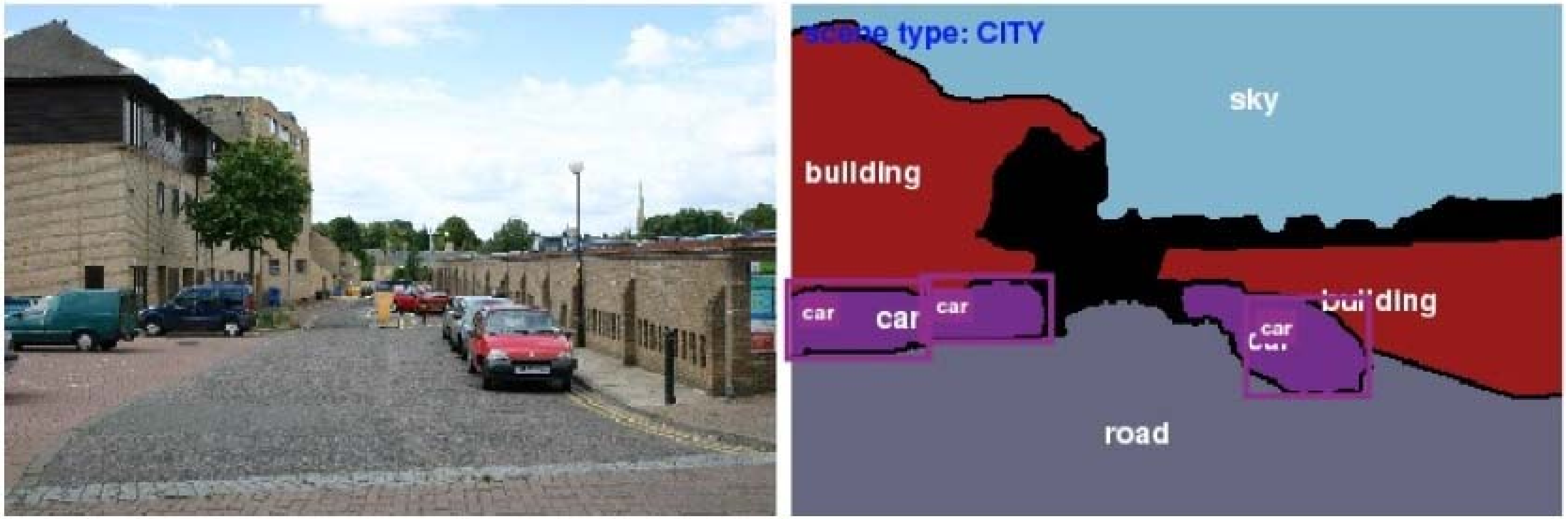
- Architectures with multiple levels naturally provide sharing and re-use of components



Honglak Lee, NIPS'10

Humans Understand the World through Multiple Levels of Abstractions

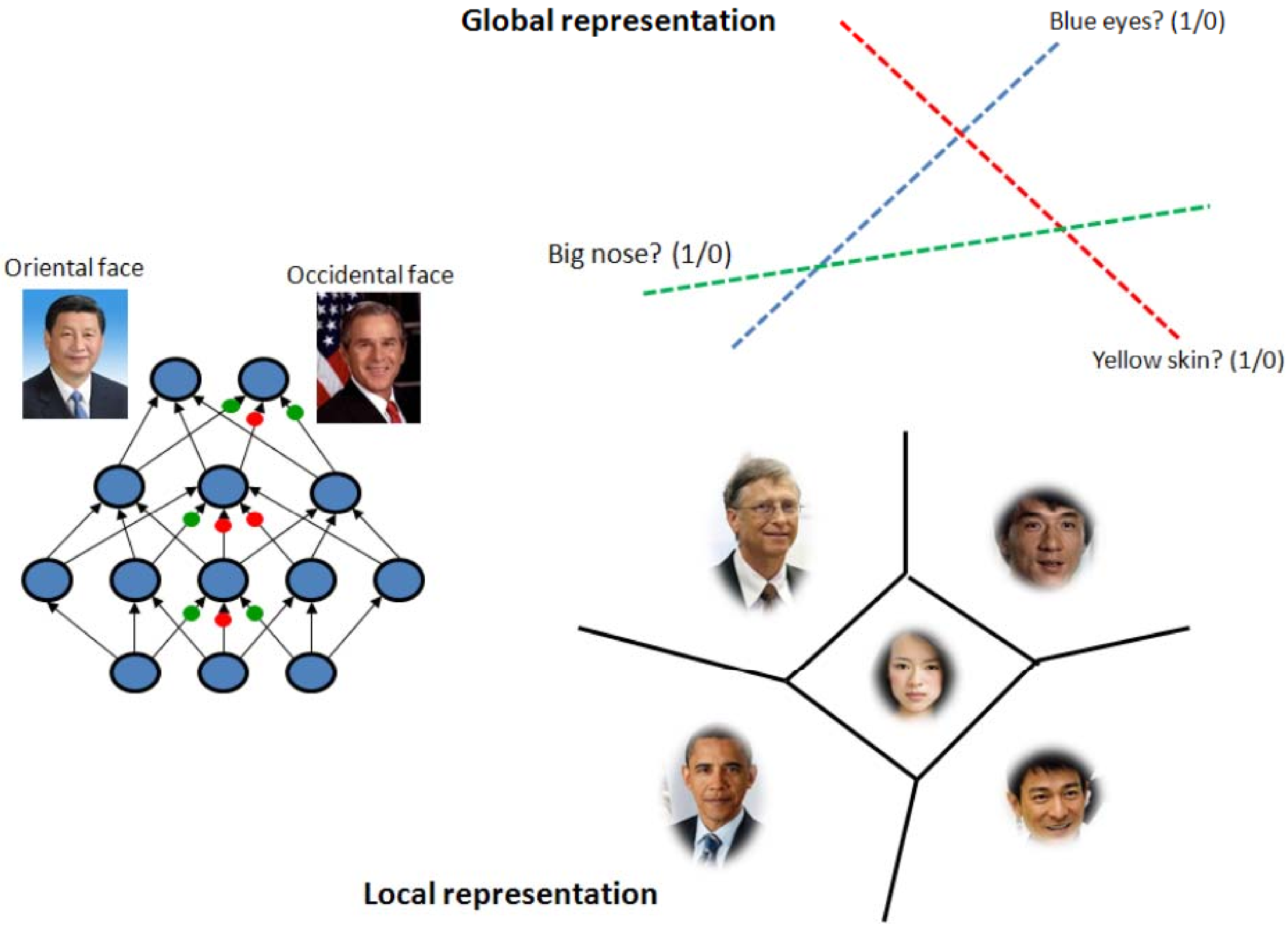
- We do not interpret a scene image with pixels
  - Objects (sky, cars, roads, buildings, pedestrians) -> parts (wheels, doors, heads) -> texture -> edges -> pixels
  - Attributes: blue sky, red car
- It is natural for humans to decompose a complex problem into sub-problems through multiple levels of representations



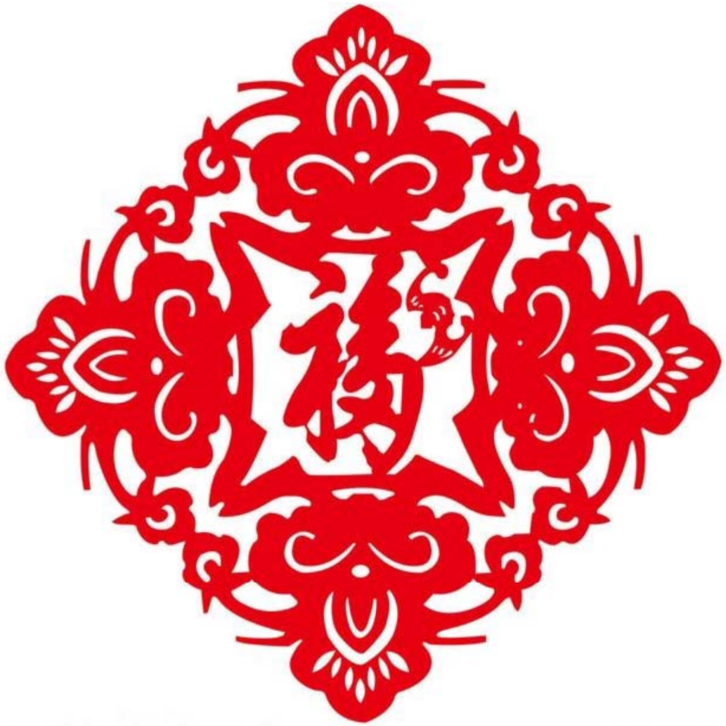
Humans Understand the World through Multiple Levels of Abstractions

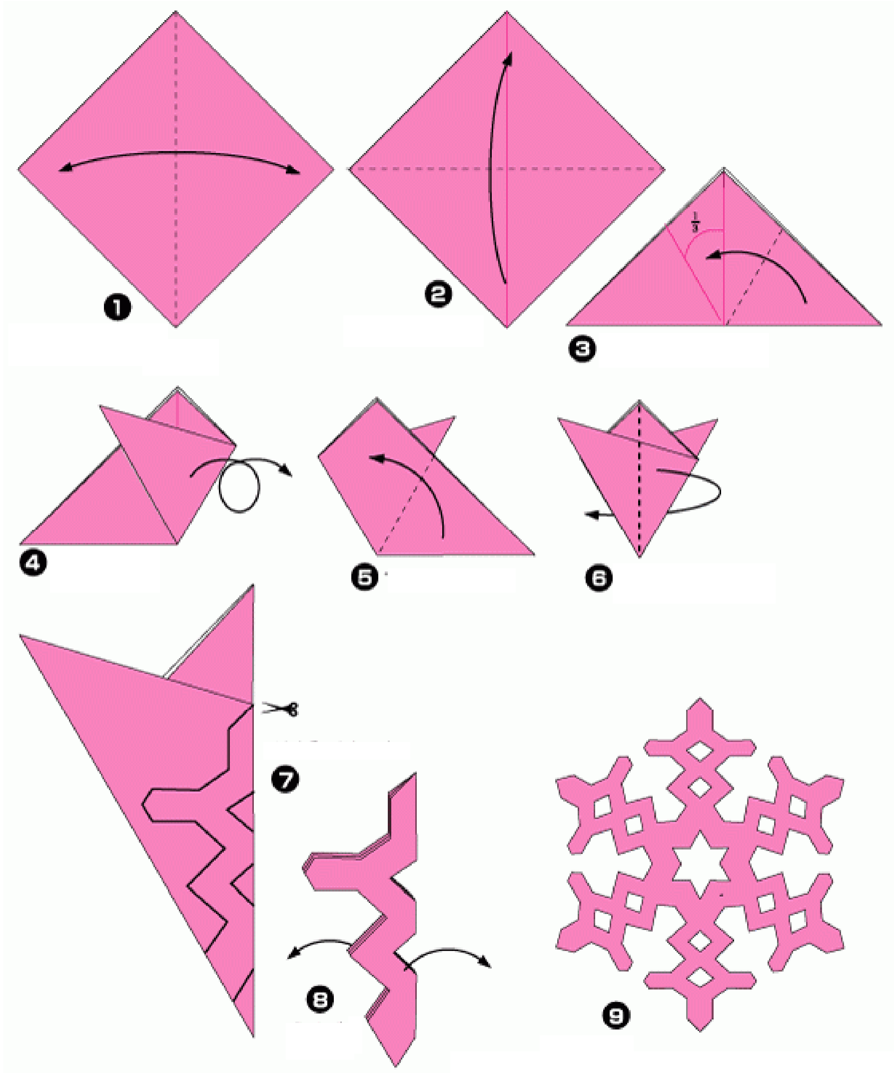
- Humans learn abstract concepts on top of less abstract ones
- Humans can imagine new pictures by re-configuring these abstractions at multiple levels. Thus our brain has good generalization can recognize things never seen before.
  - Our brain can estimate shape, lighting and pose from a face image and generate new images under various lightings and poses. That's why we have good face recognition capability.

Local and Global Representations



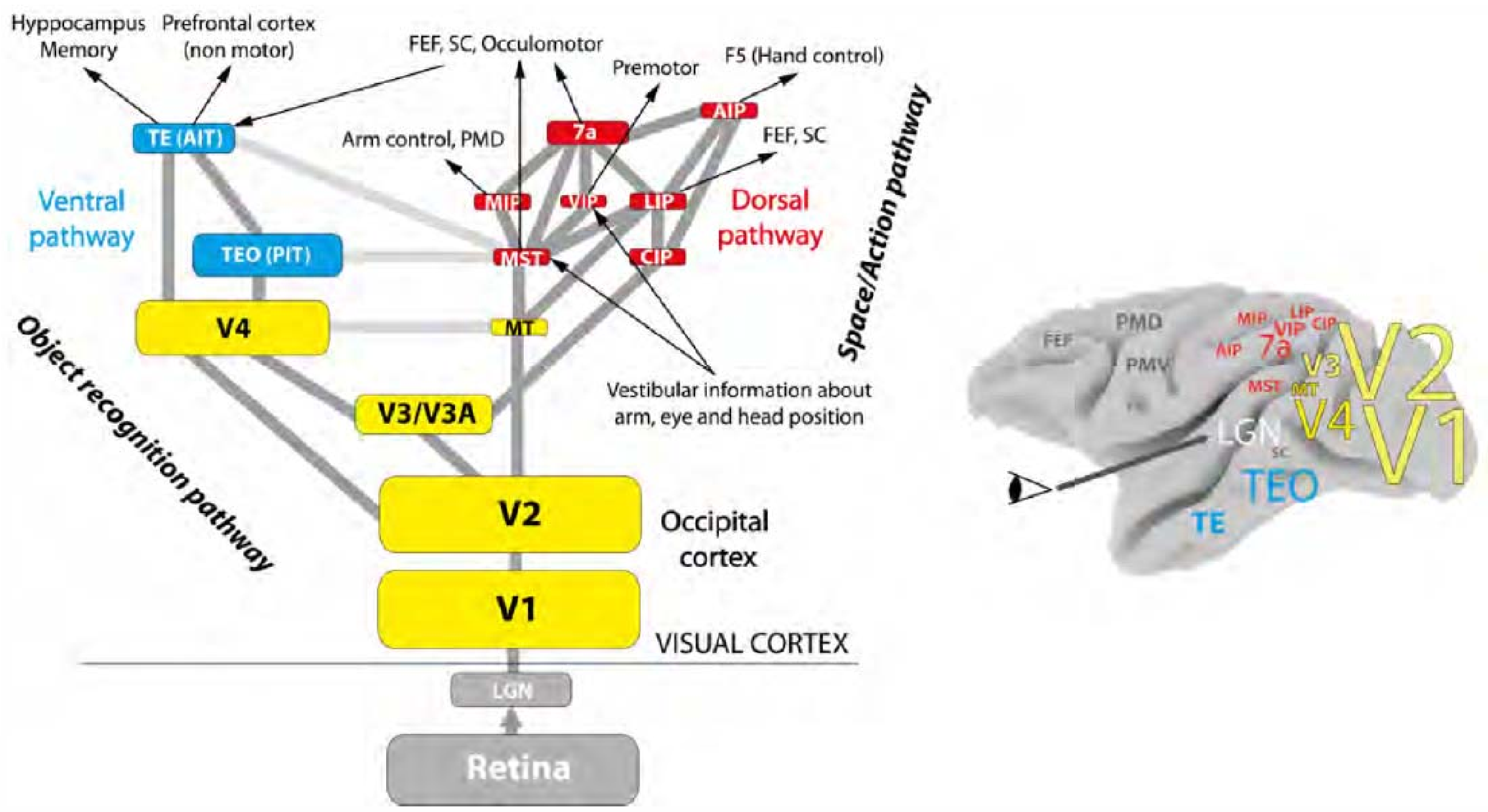
- The way these regions carve the input space still depends on few parameters: this huge number of regions are not placed independently of each other
- We can thus represent a function that looks complicated but actually has (global) structures
- The assumption is that one can learn about each feature without having to see the examples for all the configurations of all the other features, i.e. these features correspond to underlying factor explaining the data


福

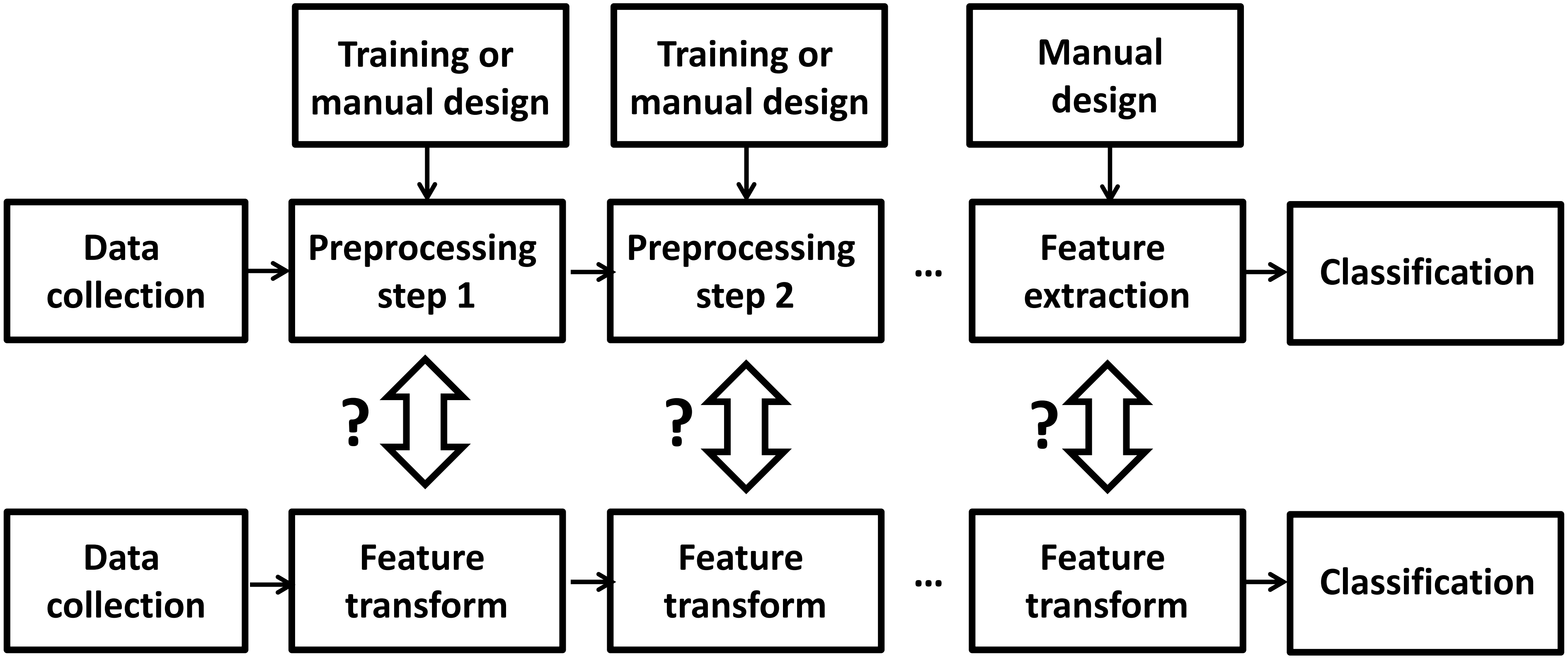

1
2
3
4
5
6
7
8
9

Human Brains Process Visual Signals through Multiple Layers

- A visual cortical area consists of six layers (Kruger et al. 2013)



Joint Learning vs Separate Learning

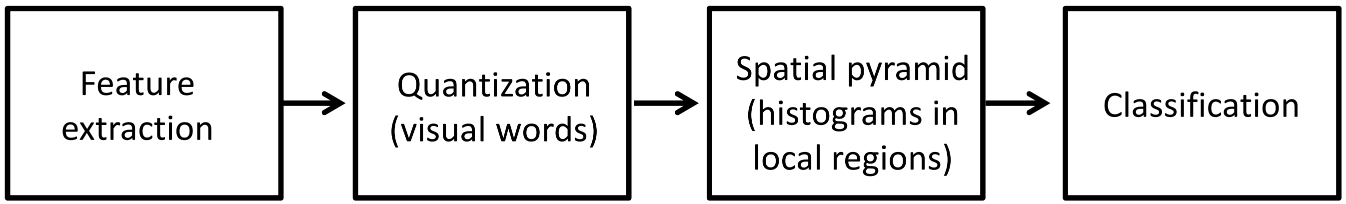


End-to-end learning

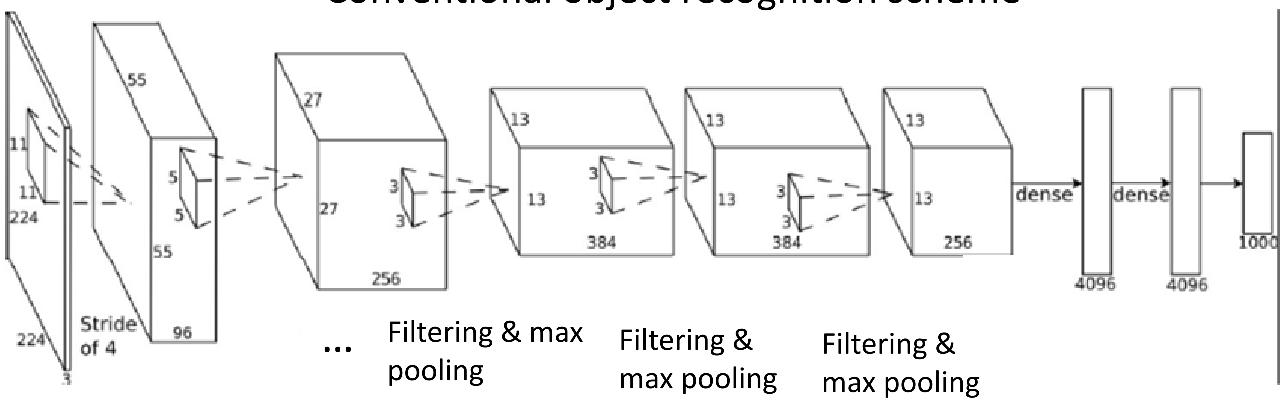
Deep learning is a framework/language but not a black-box model

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

- Domain knowledge could be helpful for designing new deep models and training strategies
- 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)



Conventional object recognition scheme



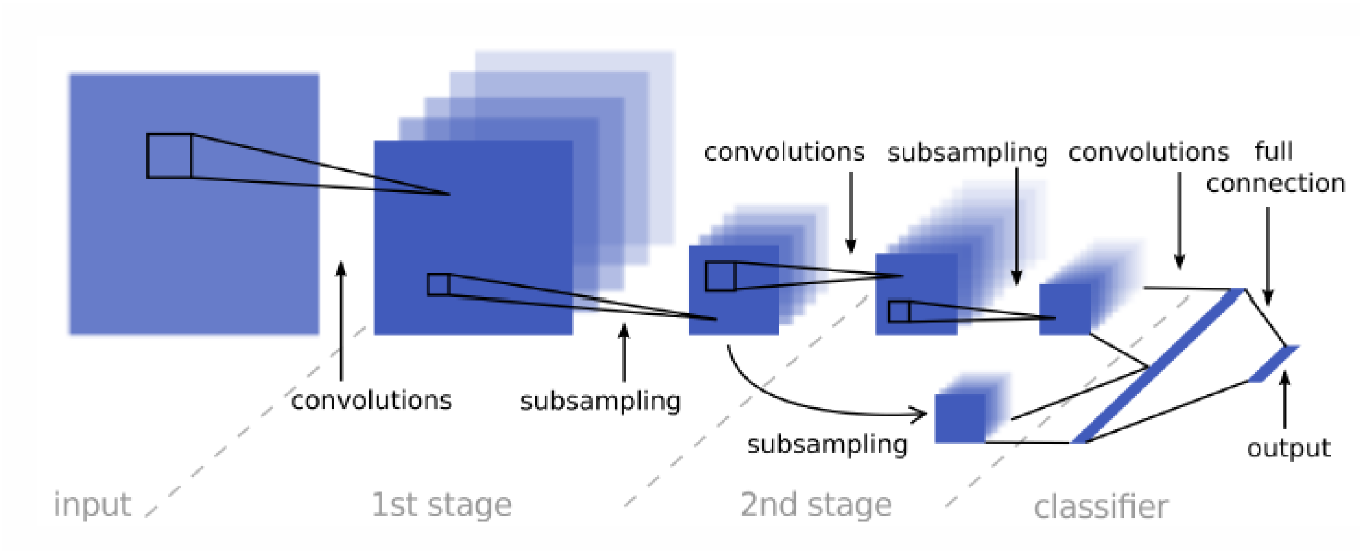
Feature extraction ↔ filtering

Quantization ↔ filtering

Spatial pyramid ↔ multi-level pooling

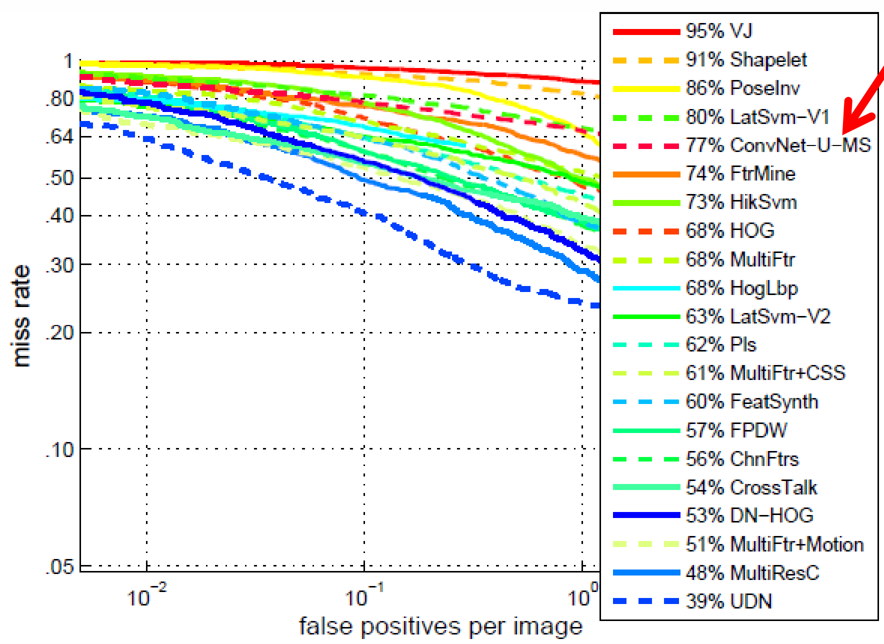
Krizhevsky NIPS'12

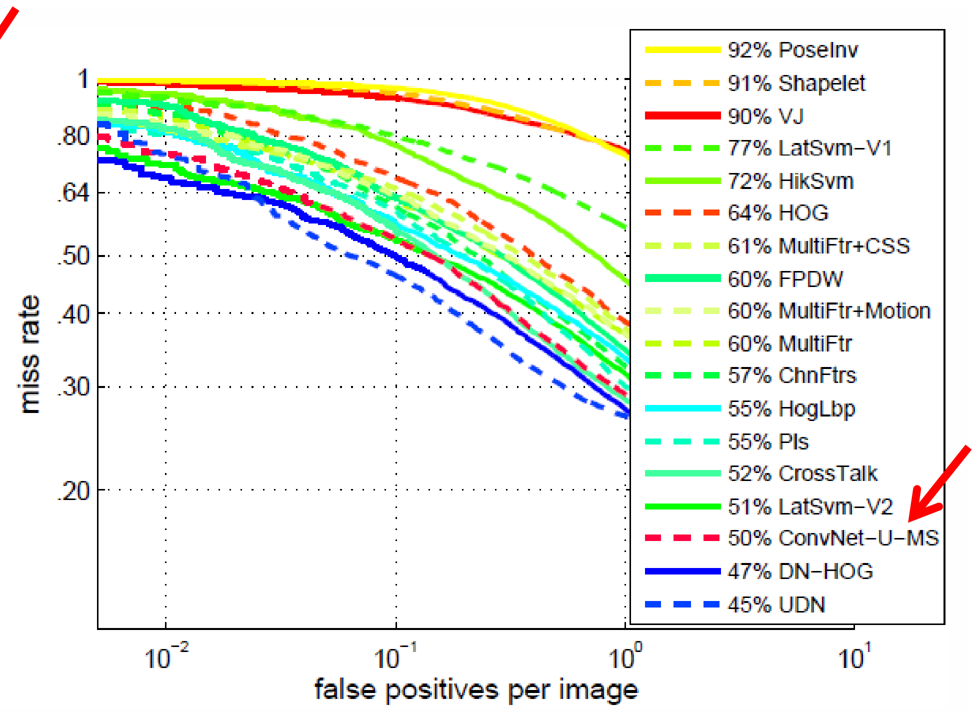
What if we treat an existing deep model as a black box in pedestrian detection?



ConvNet−U−MS

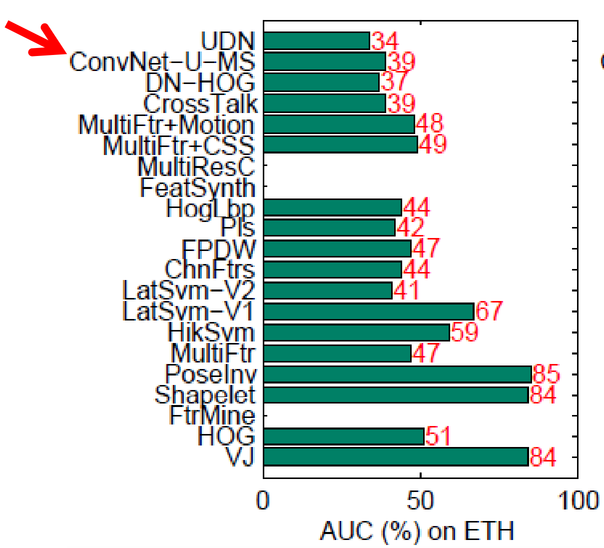
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” CVPR 2013.

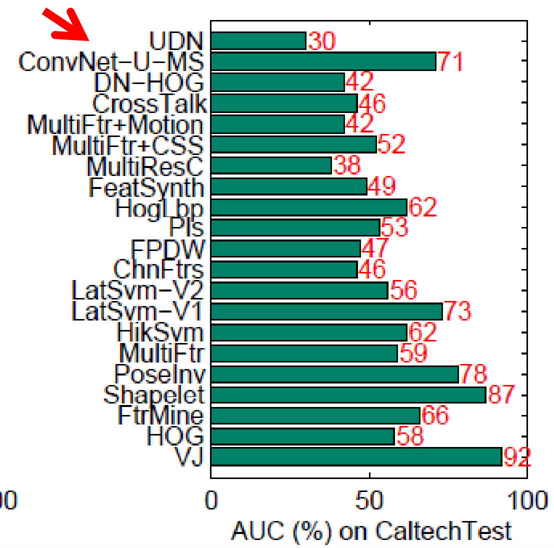


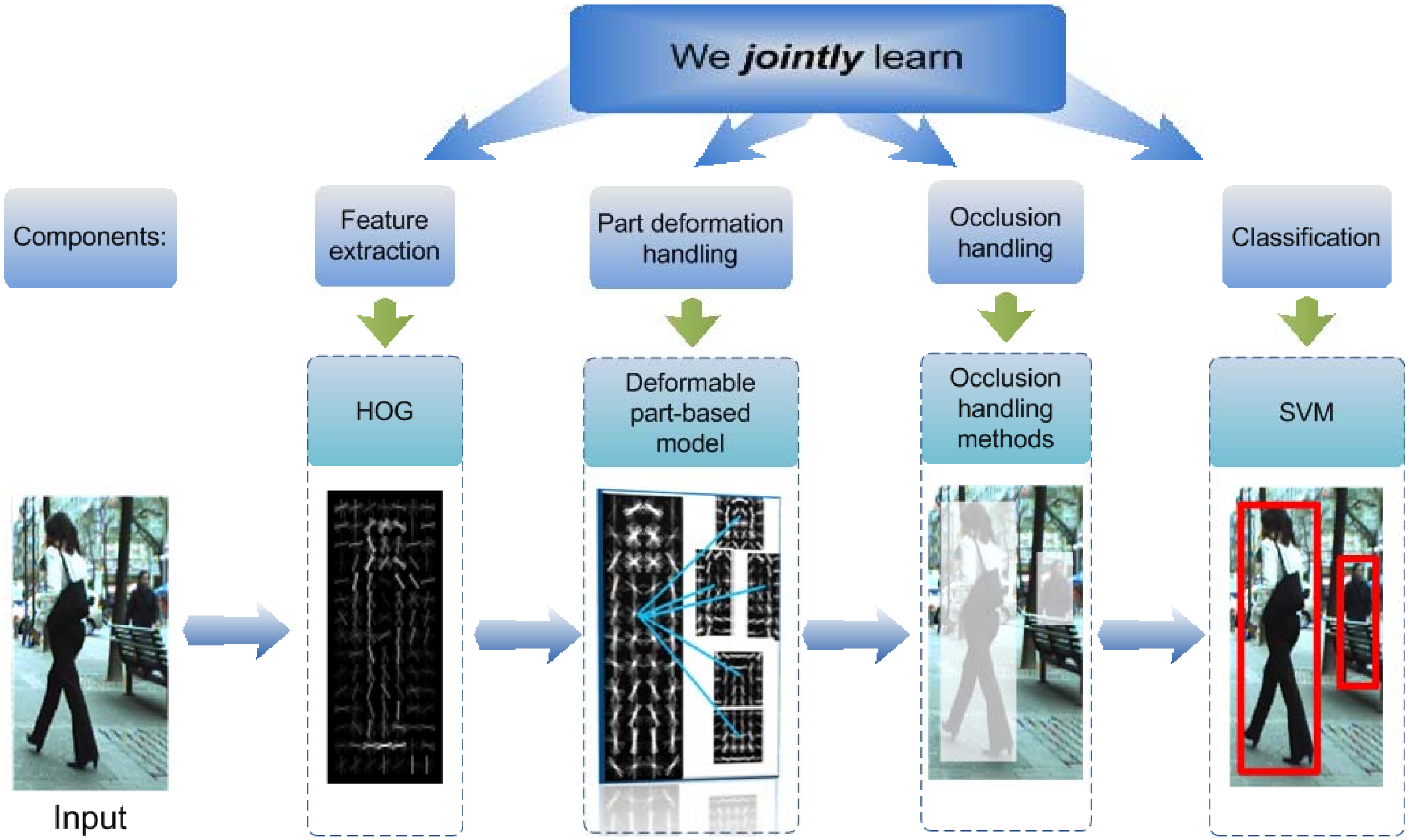


Results on Caltech Test

Results on ETHZ

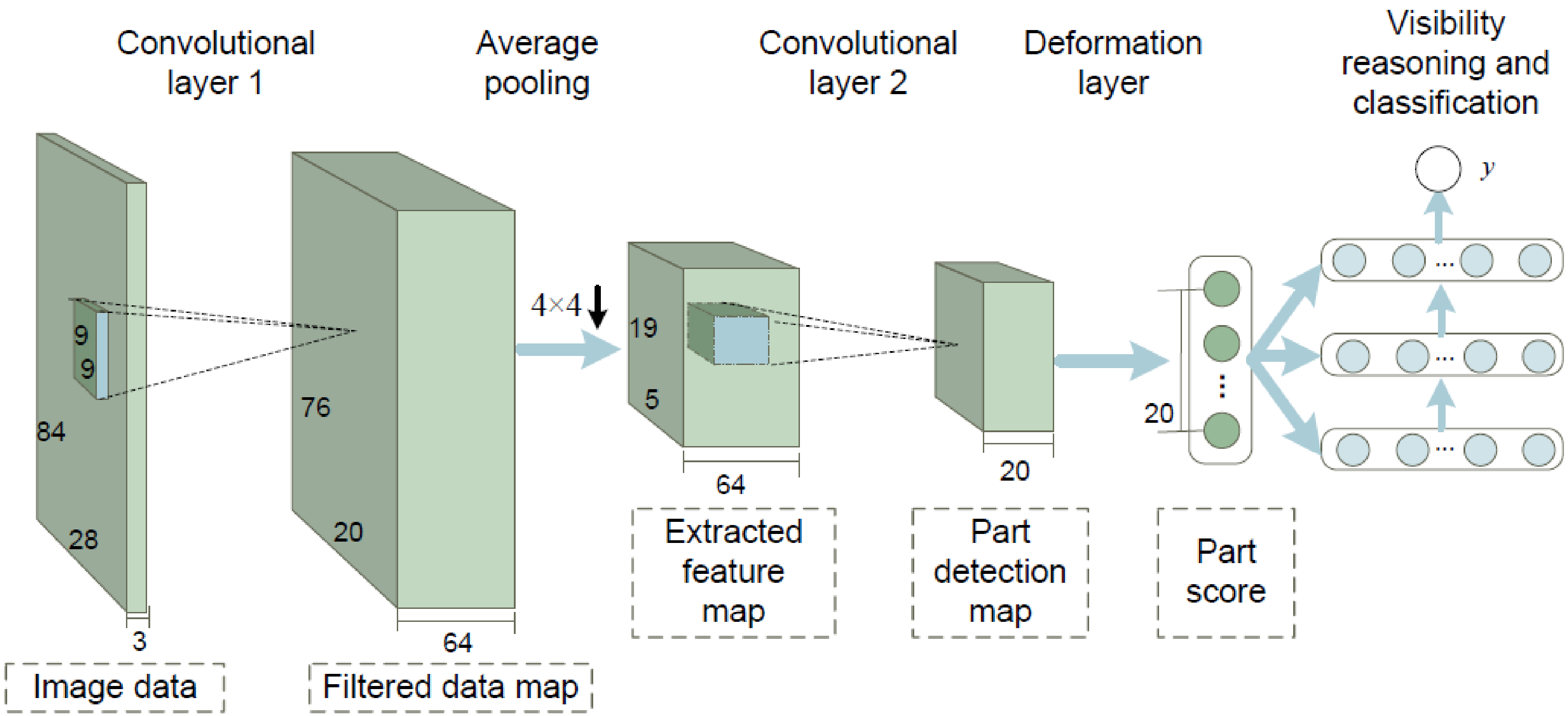






- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)
- P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)
- W. Ouyang and X. Wang. A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling. CVPR, 2012.

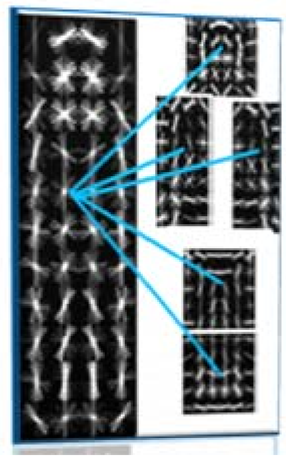
Our Joint Deep Learning Model



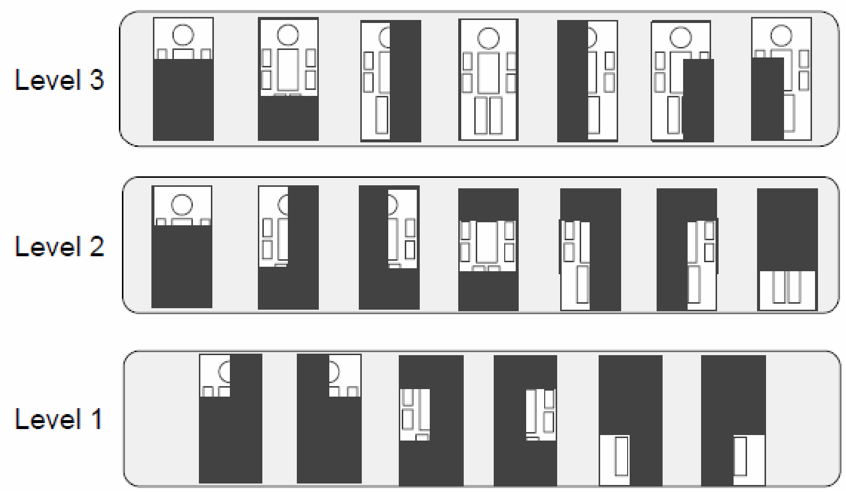
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling Part Detectors

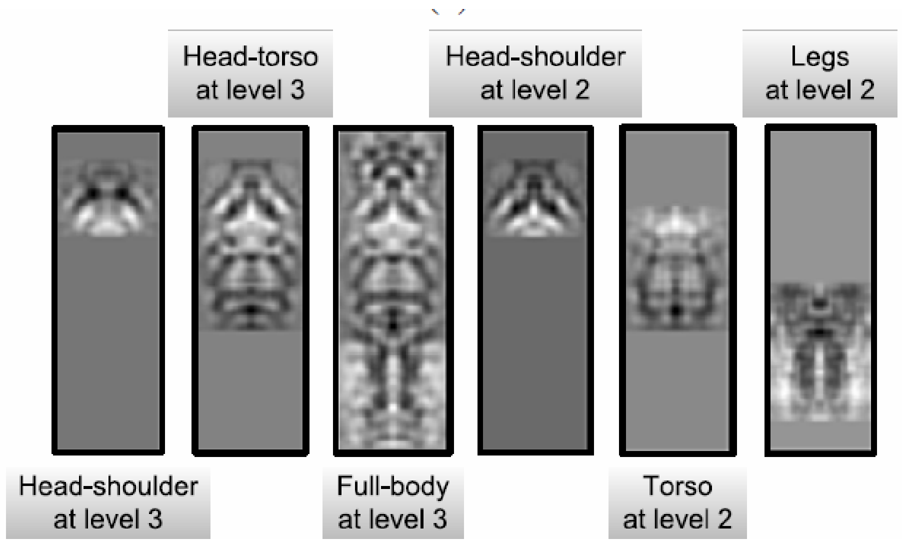
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

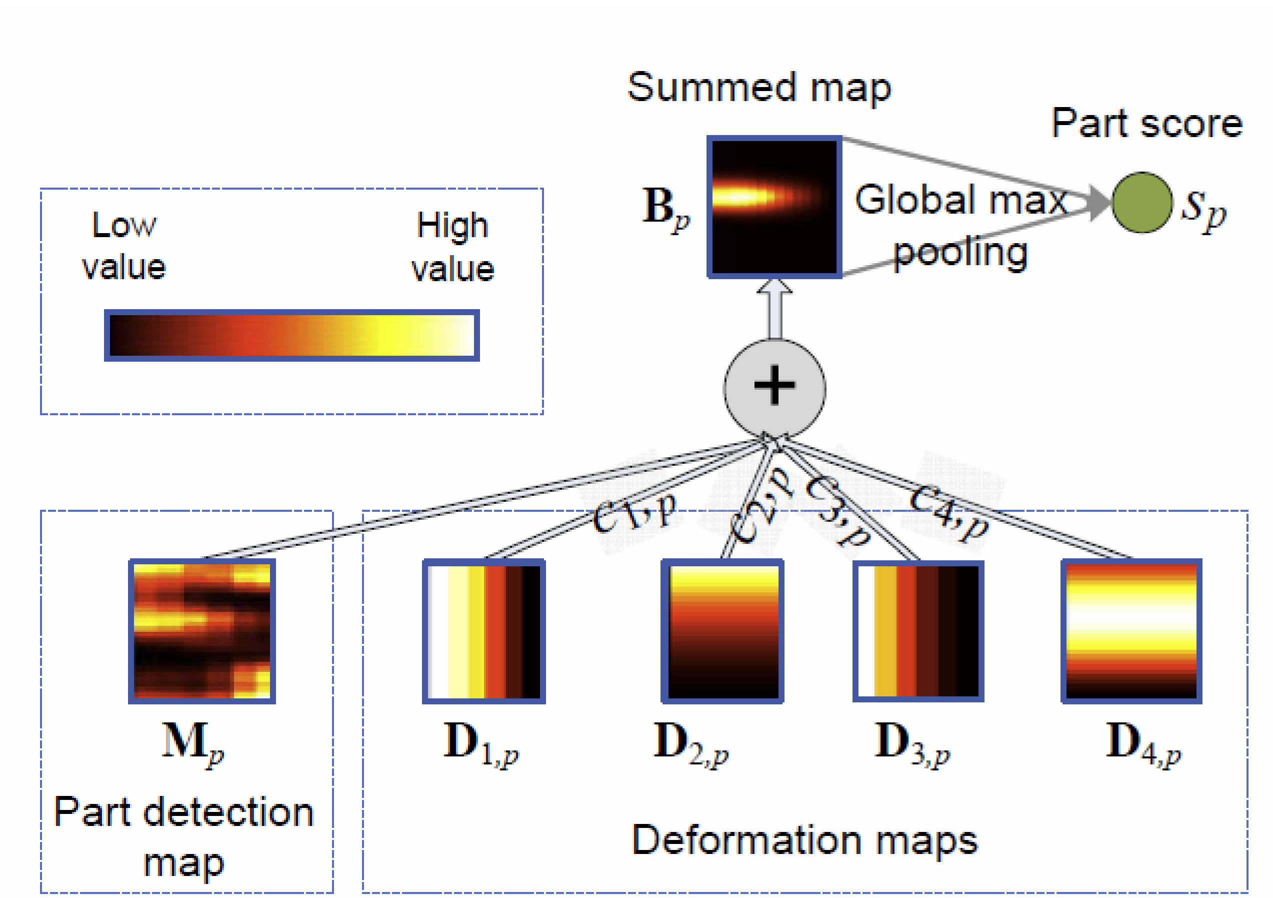


Part models

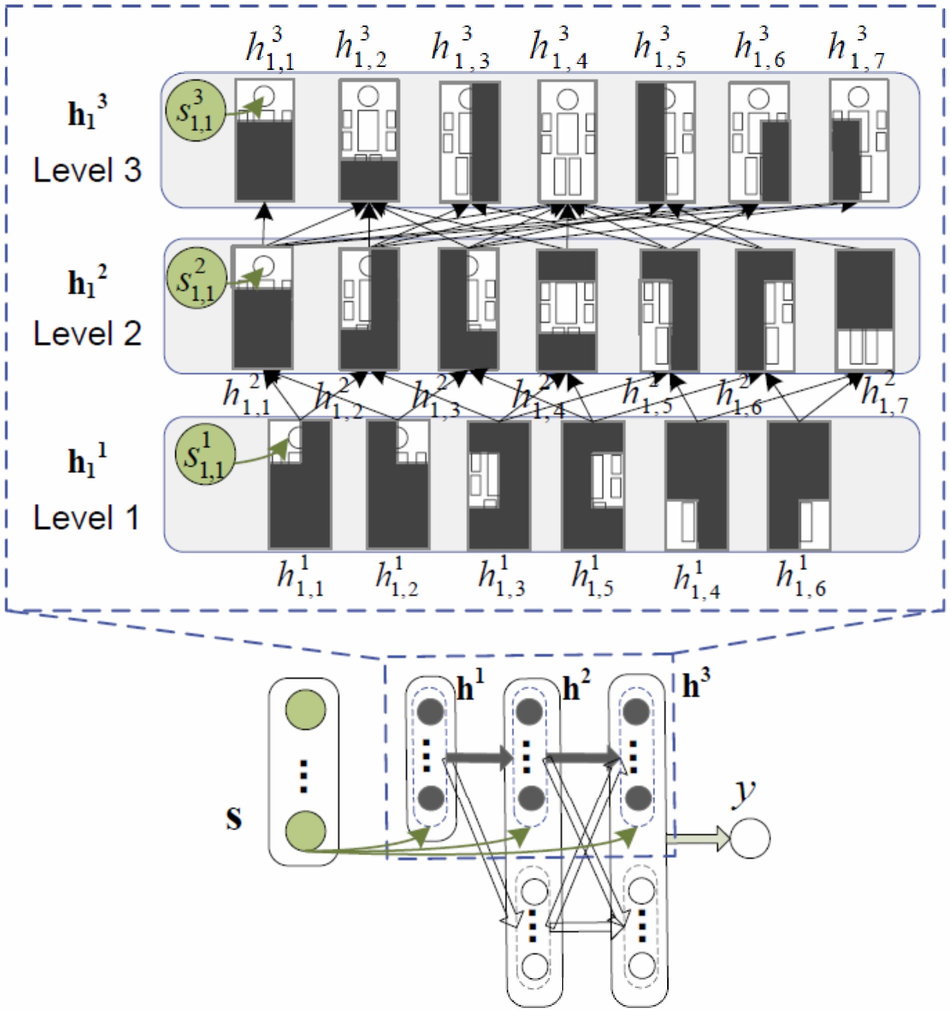


Learned filtered at the second convolutional layer

Deformation Layer



Visibility Reasoning with Deep Belief Net

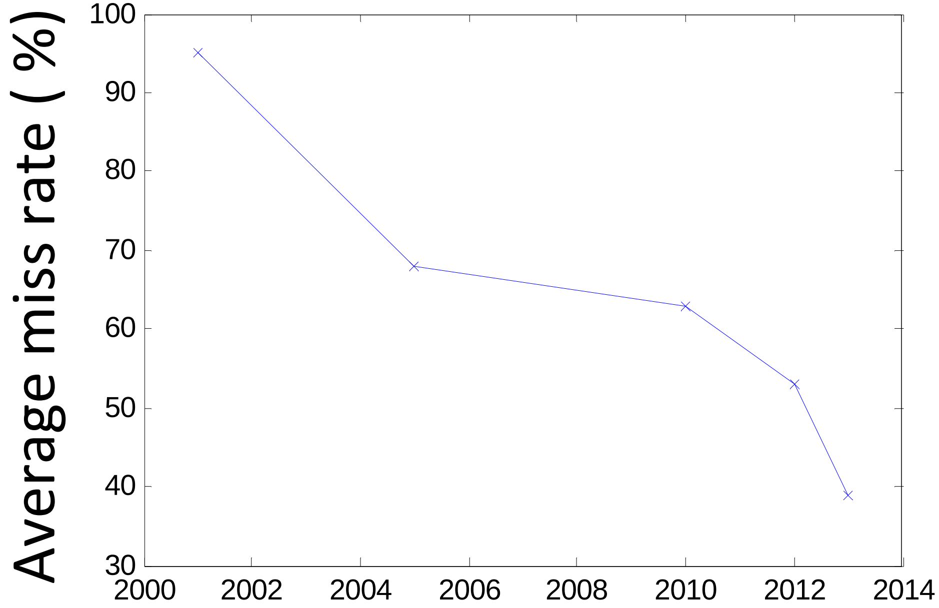


$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

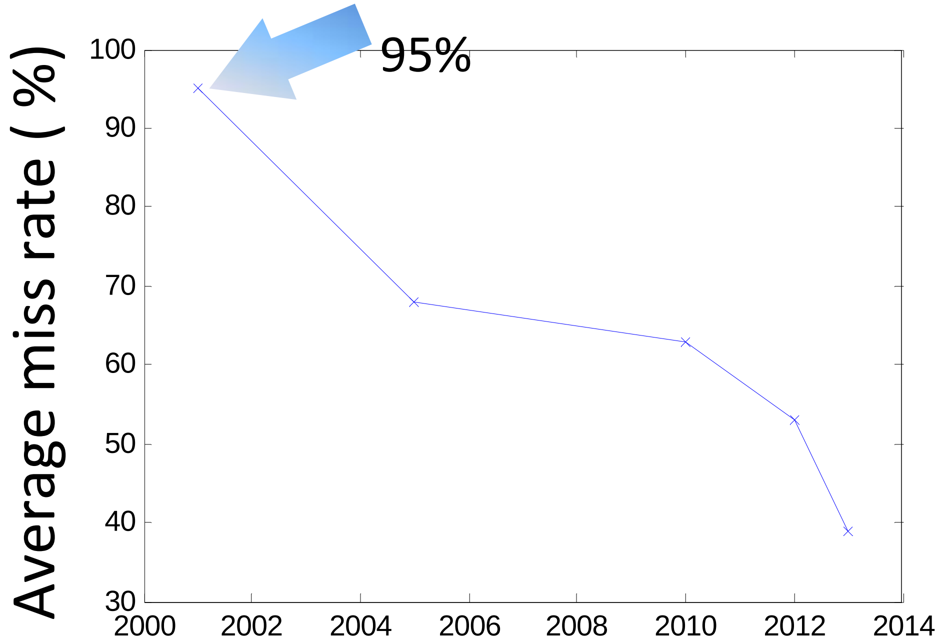
Experimental Results

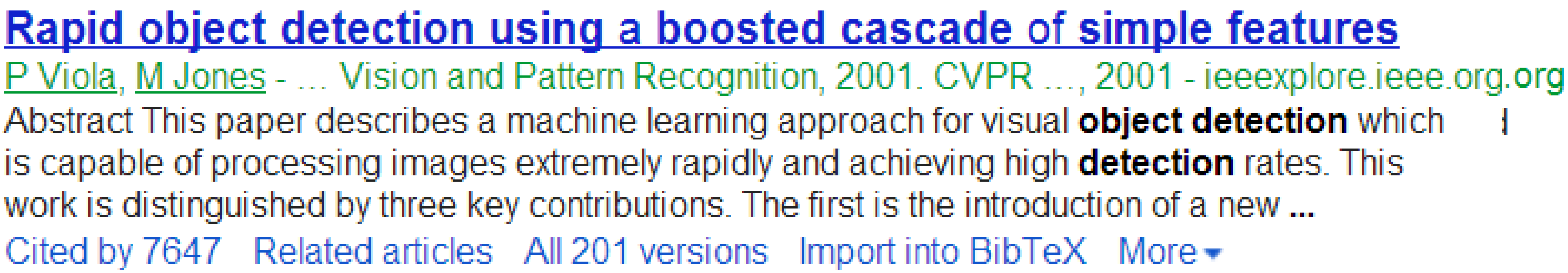
- Caltech – Test dataset (largest, most widely used)



Experimental Results

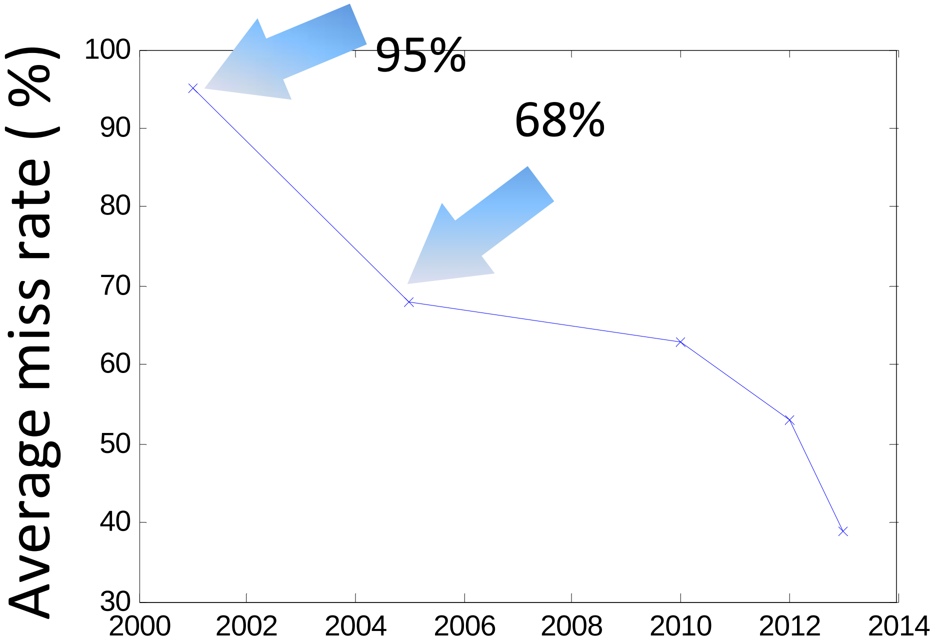
- Caltech – Test dataset (largest, most widely used)





Experimental Results

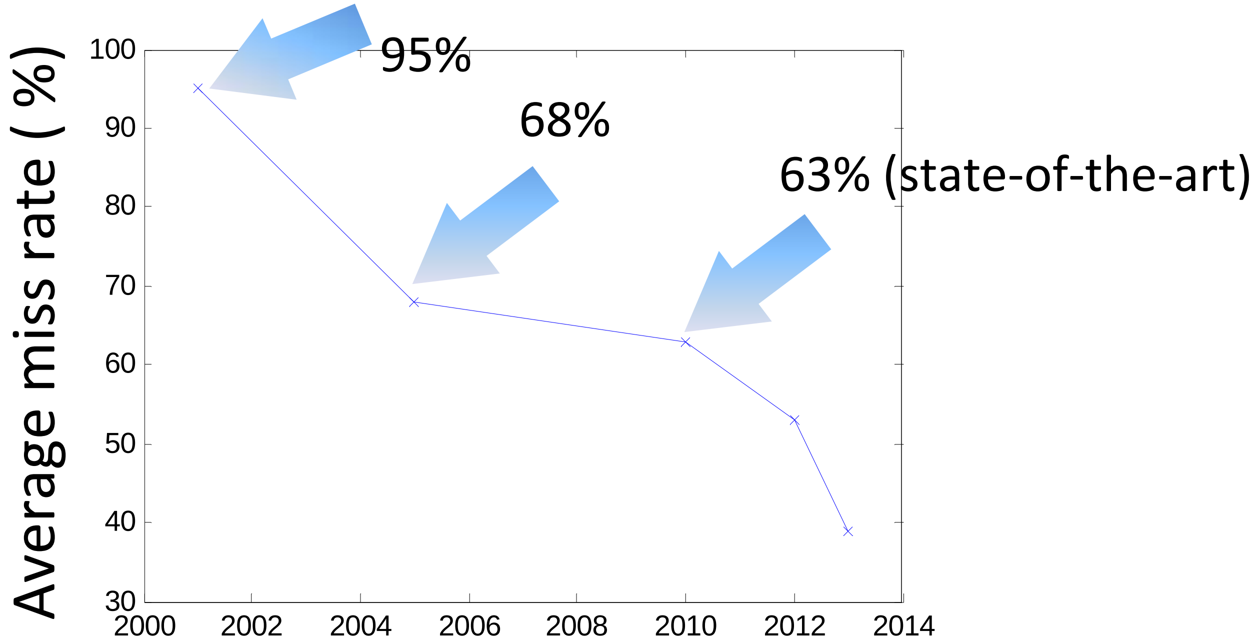
- Caltech – Test dataset (largest, most widely used)



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

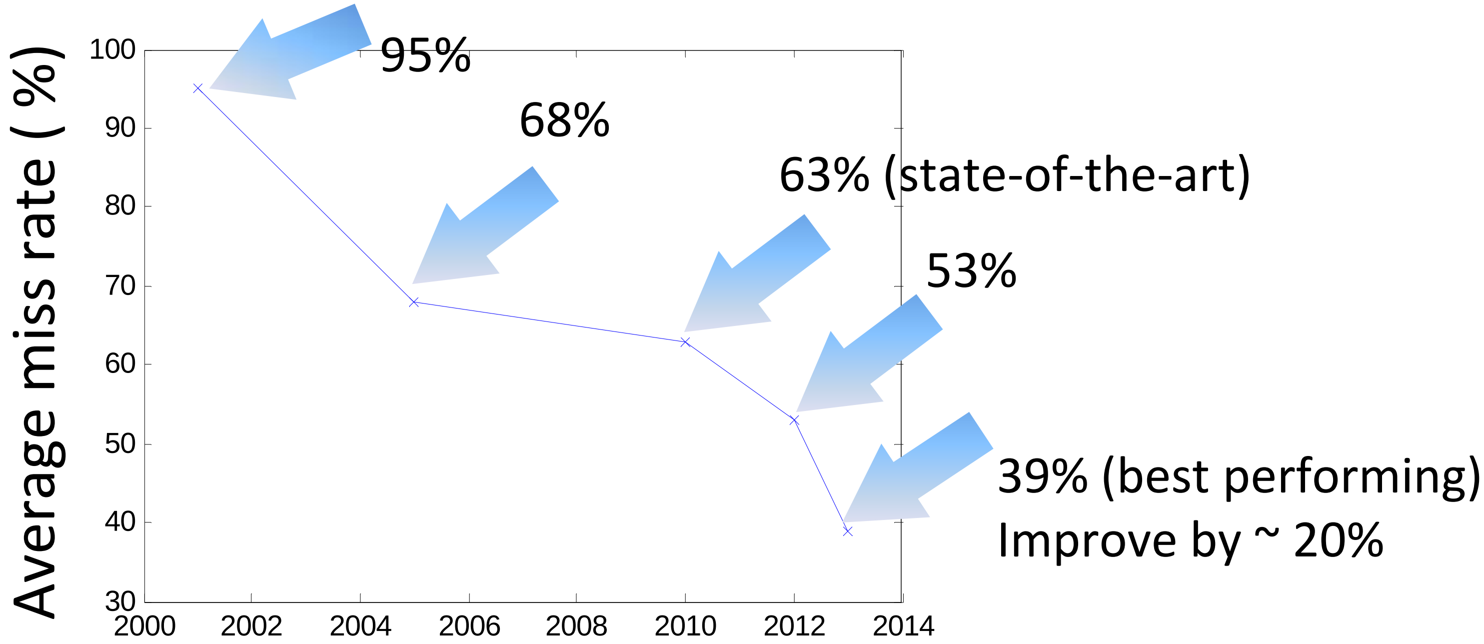
- Caltech – Test dataset (largest, most widely used)



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

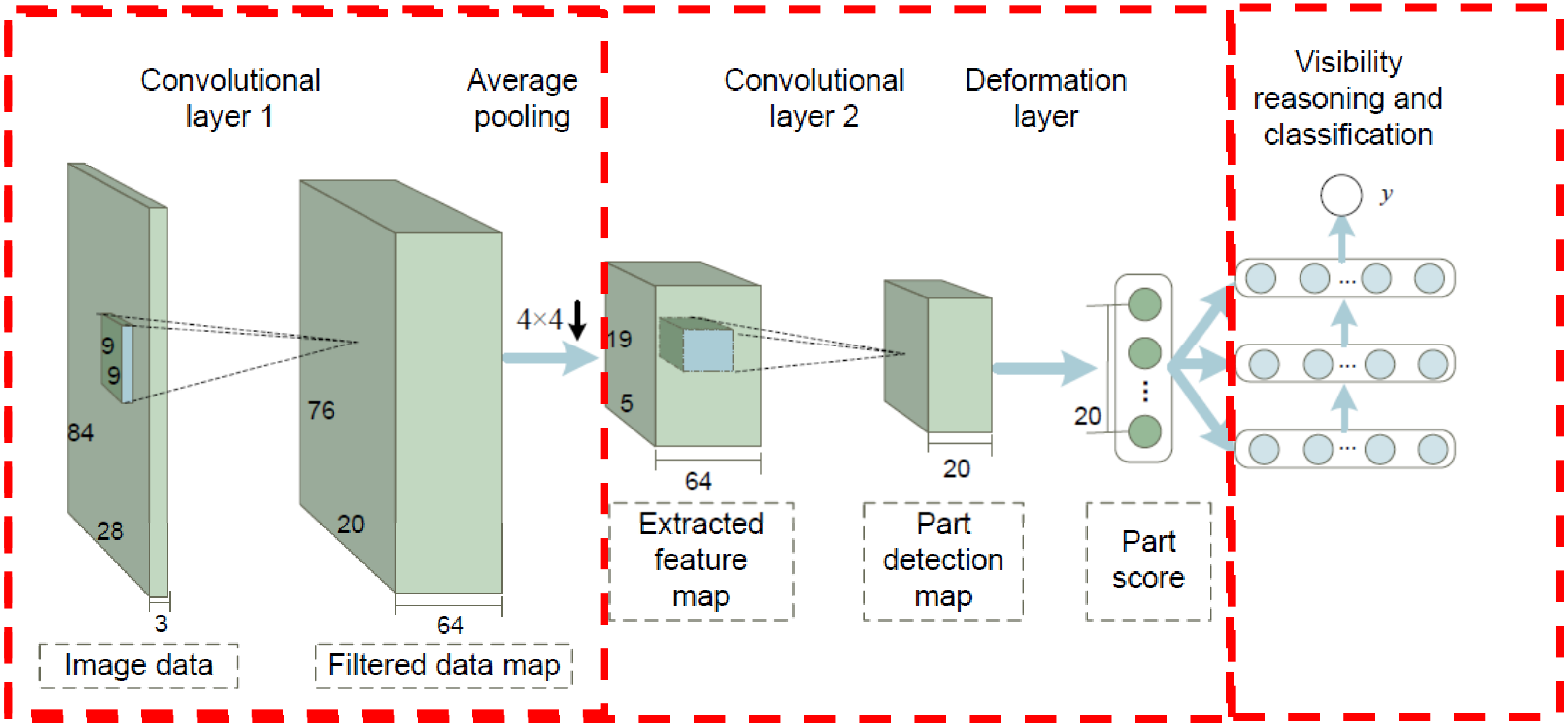
Experimental Results

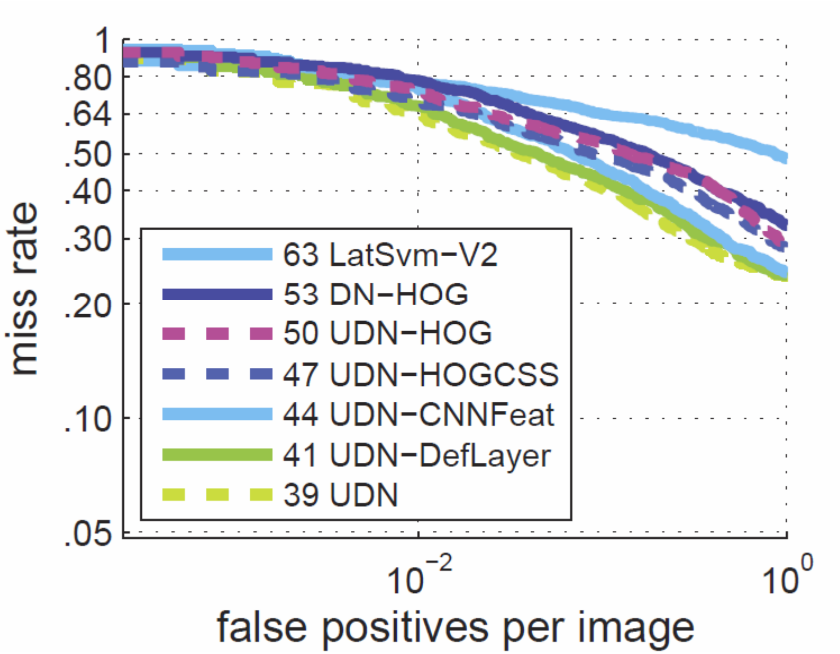
- Caltech – Test dataset (largest, most widely used)



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.

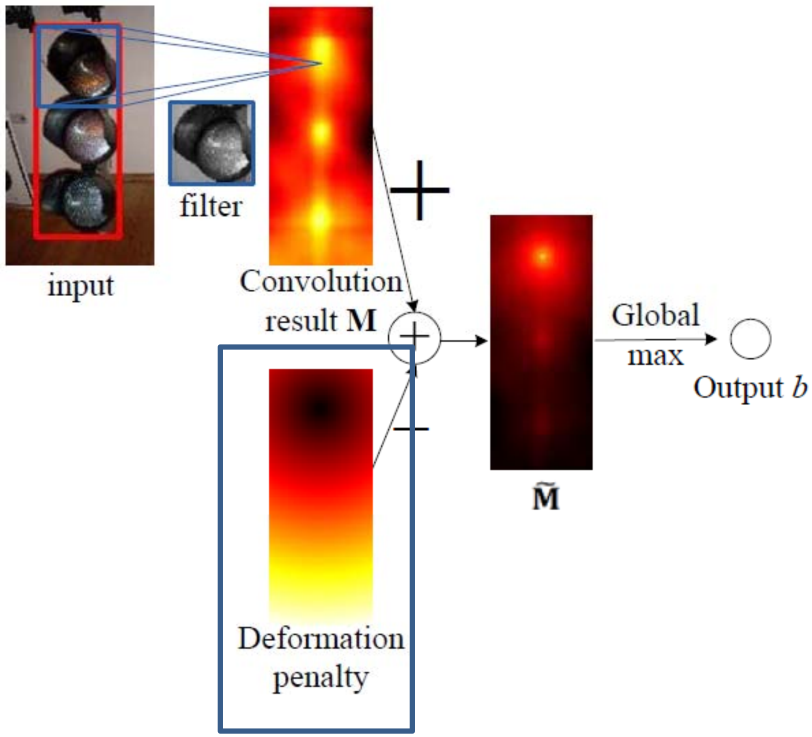




DN-HOG
UDN-HOG
UDN-HOGCSS
UDN-CNNFeat
UDN-DefLayer

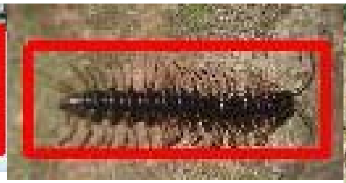
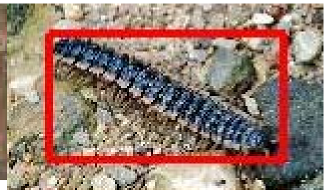
Deformation layer for general object detection

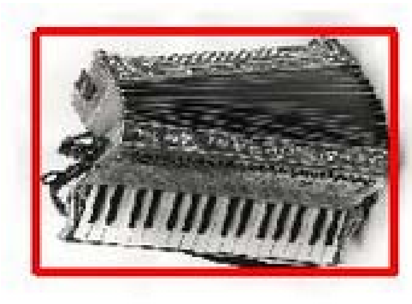
$$\mathbf{B}_p = \mathbf{M}_p + \boxed{\sum_{n=1}^{N} c_{n,p}\mathbf{D}_{n,p}} \qquad s_p = \max_{(x,y)} b_p^{(x,y)}$$



Deformation layer for repeated patterns

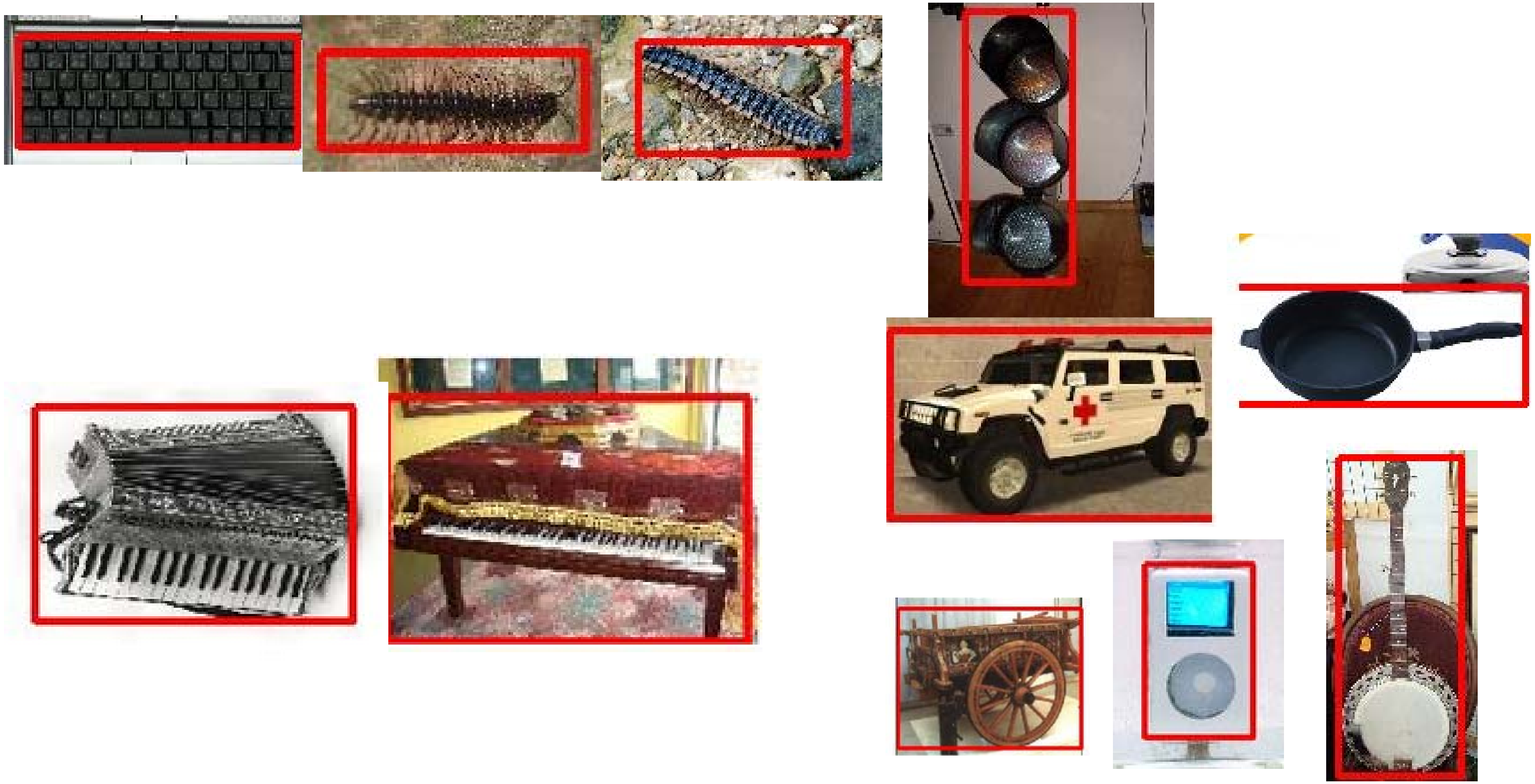
Pedestrian detection	General object detection
Assume no repeated pattern	Repeated patterns


Deformation layer for repeated patterns

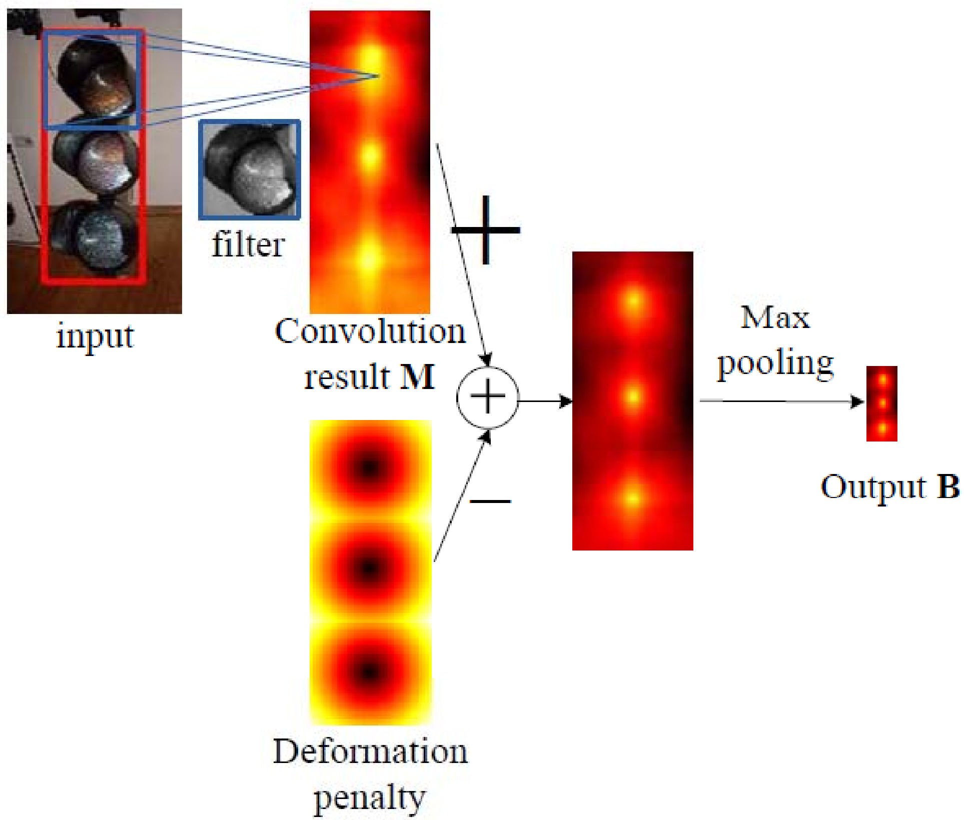
Pedestrian detection	General object detection
Assume no repeated pattern	Repeated patterns
Only consider one object class	Patterns shared across different object classes

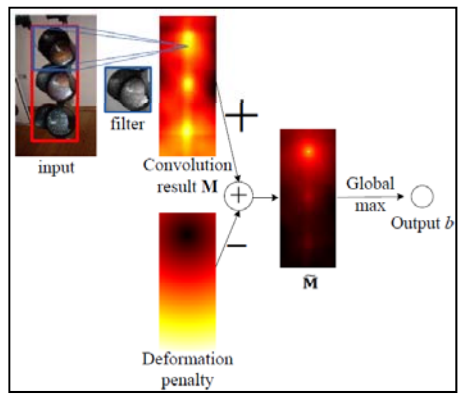


Deformation constrained pooling layer

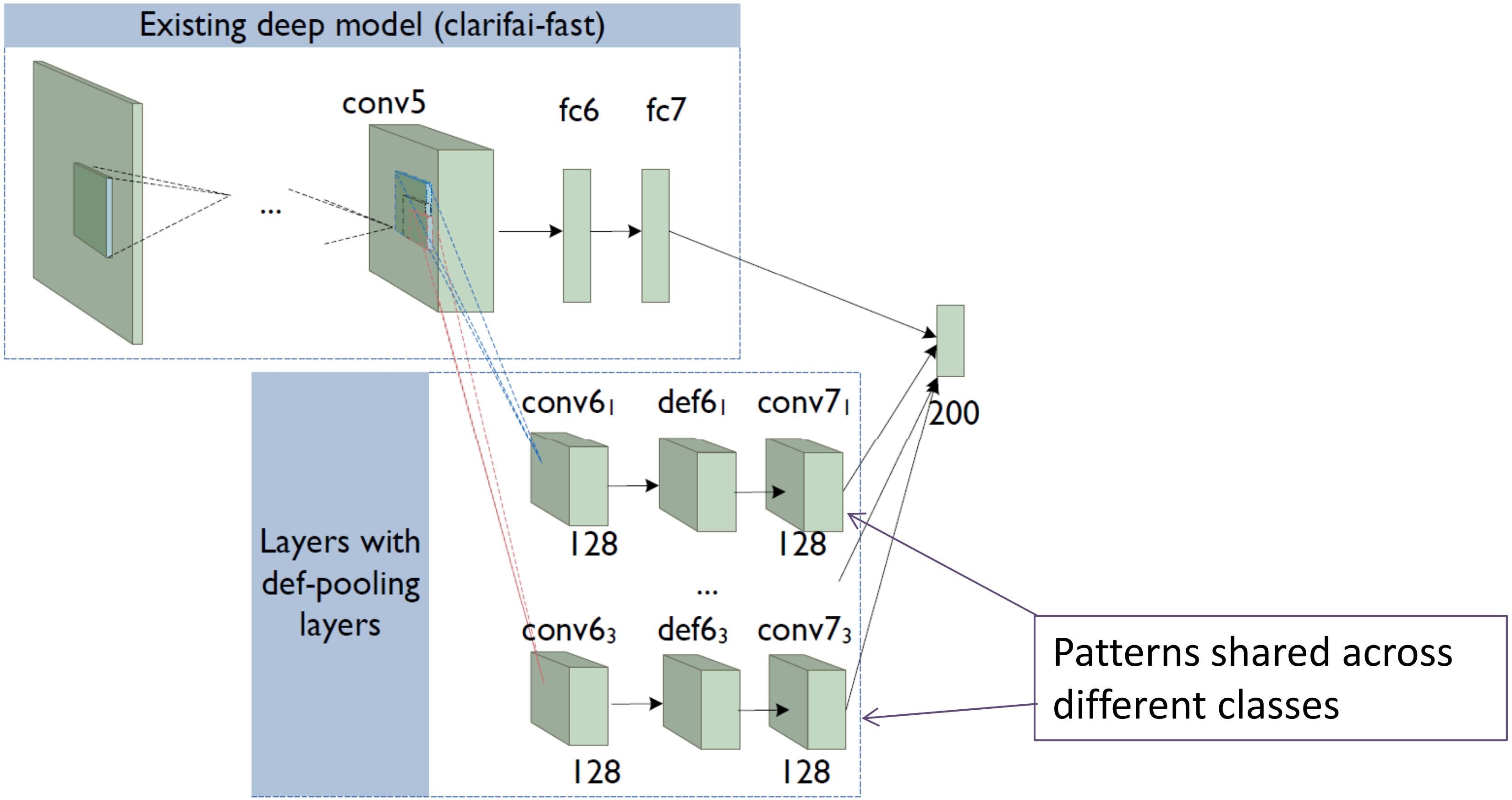
Can capture multiple patterns simultaneously

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





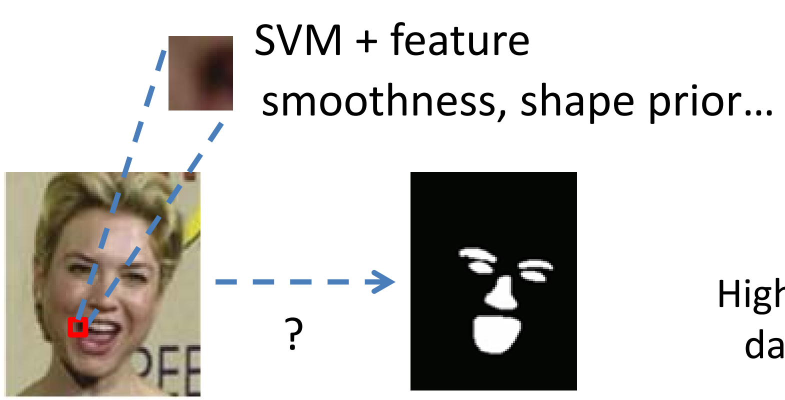
Deep model with deformation layer

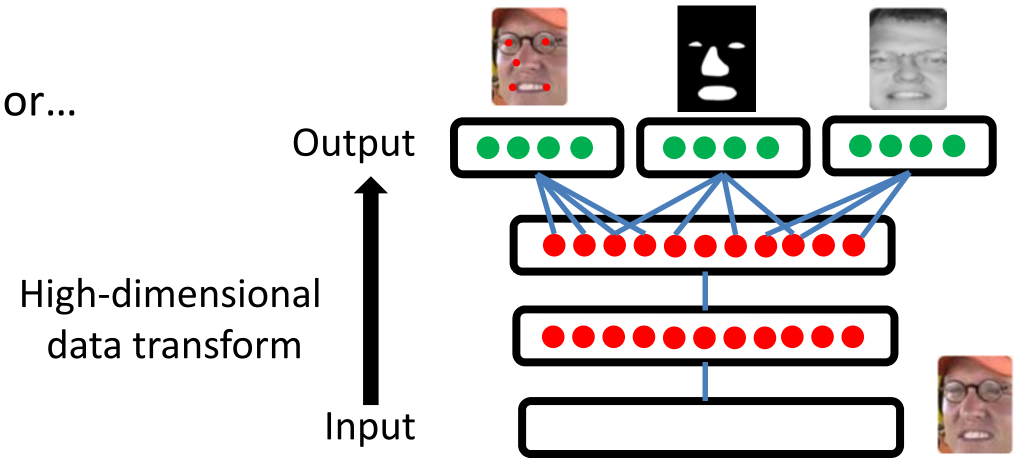


Training scheme	Cls+Det	Loc+Det	Loc+Det
Net structure	AlexNet	Clarifai	Clarifai+Def layer
Mean AP on val2	0.299	0.360	0.385

Large learning capacity makes high dimensional data transforms possible, and makes better use of contextual information

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





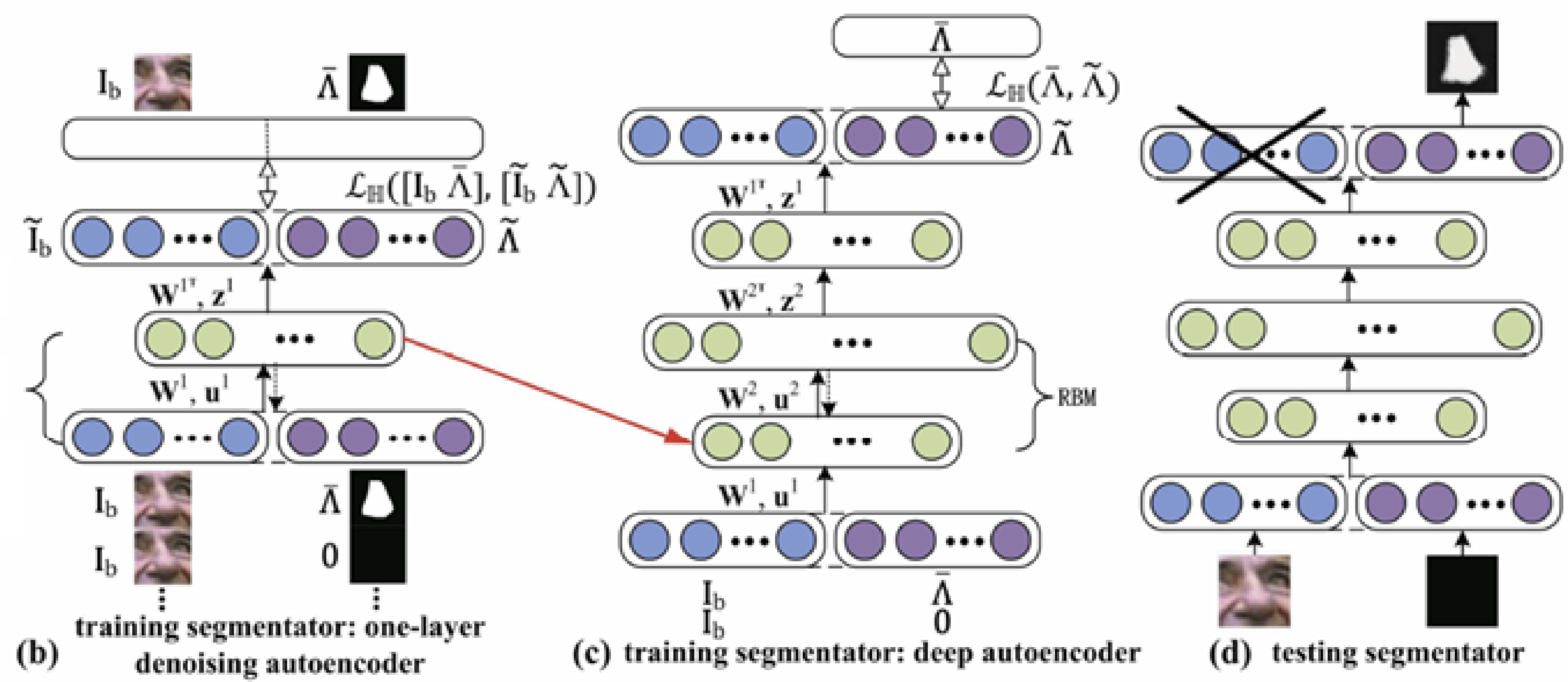
Face Parsing

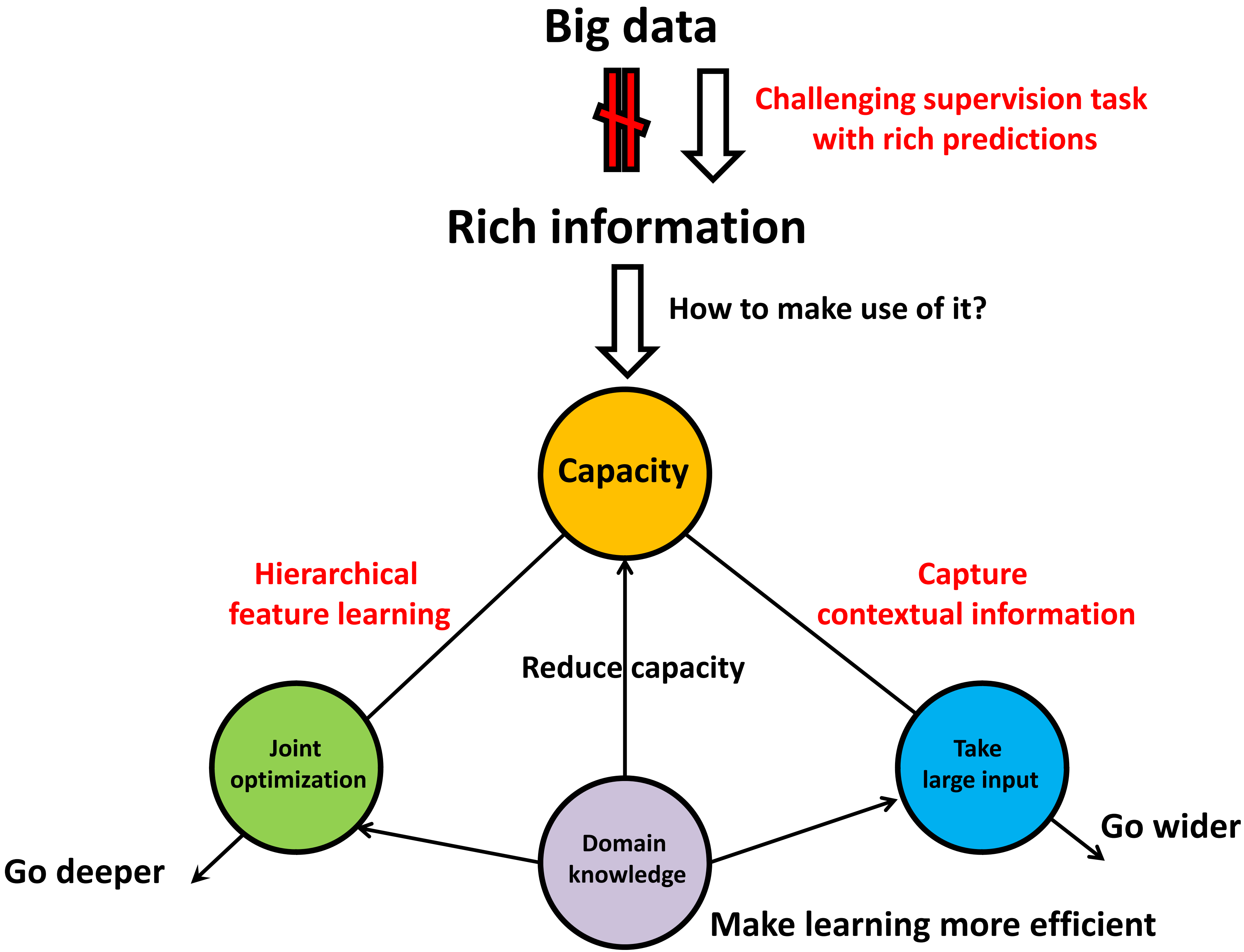
- P. Luo, X. Wang and X. Tang, “Hierarchical Face Parsing via Deep Learning,” CVPR 2012

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

Training Segmentators

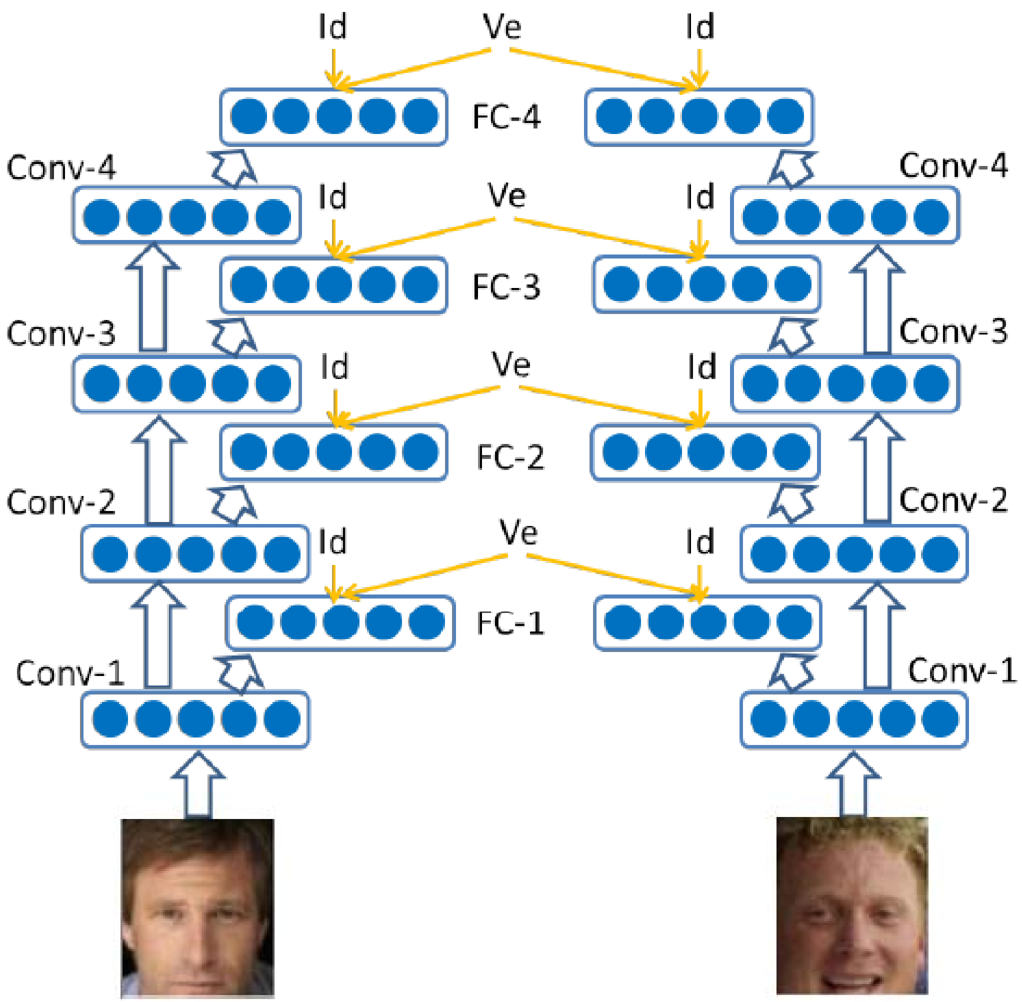



Big data
Challenging supervision task
with rich predictions
Rich information
How to make use of it?
Capacity
Hierarchical
feature learning
Capture
contextual information
Reduce capacity
Joint
optimization
Take
large input
Go wider
Go deeper
Domain
knowledge
Make learning more efficient

Introduction to Deep Learning

- Historical review of deep learning
- Introduction to classical deep models
- Why does deep learning work?
- **Properties of deep feature representations**

Example of DeepID2+ for Face Recognition



Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CVPR, 2015.

What has been learned by DeepID2+?

Properties owned by neurons?

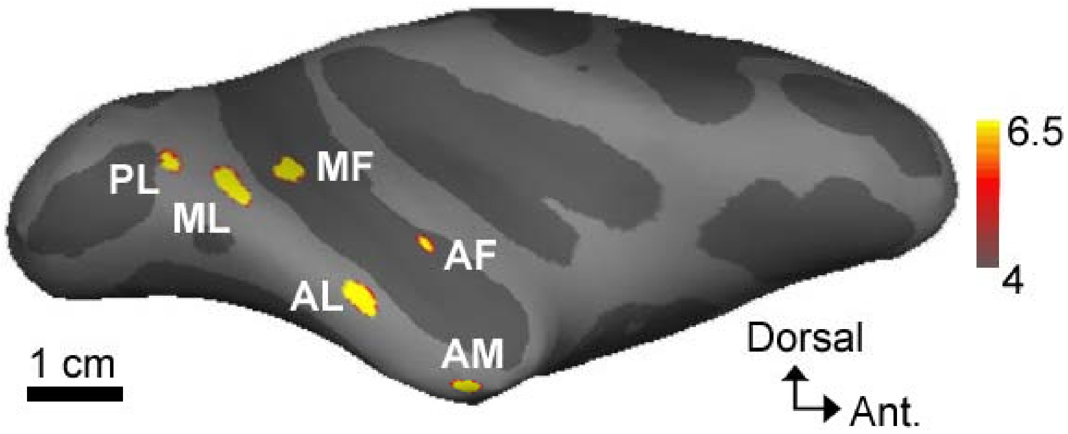
Moderate sparse

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

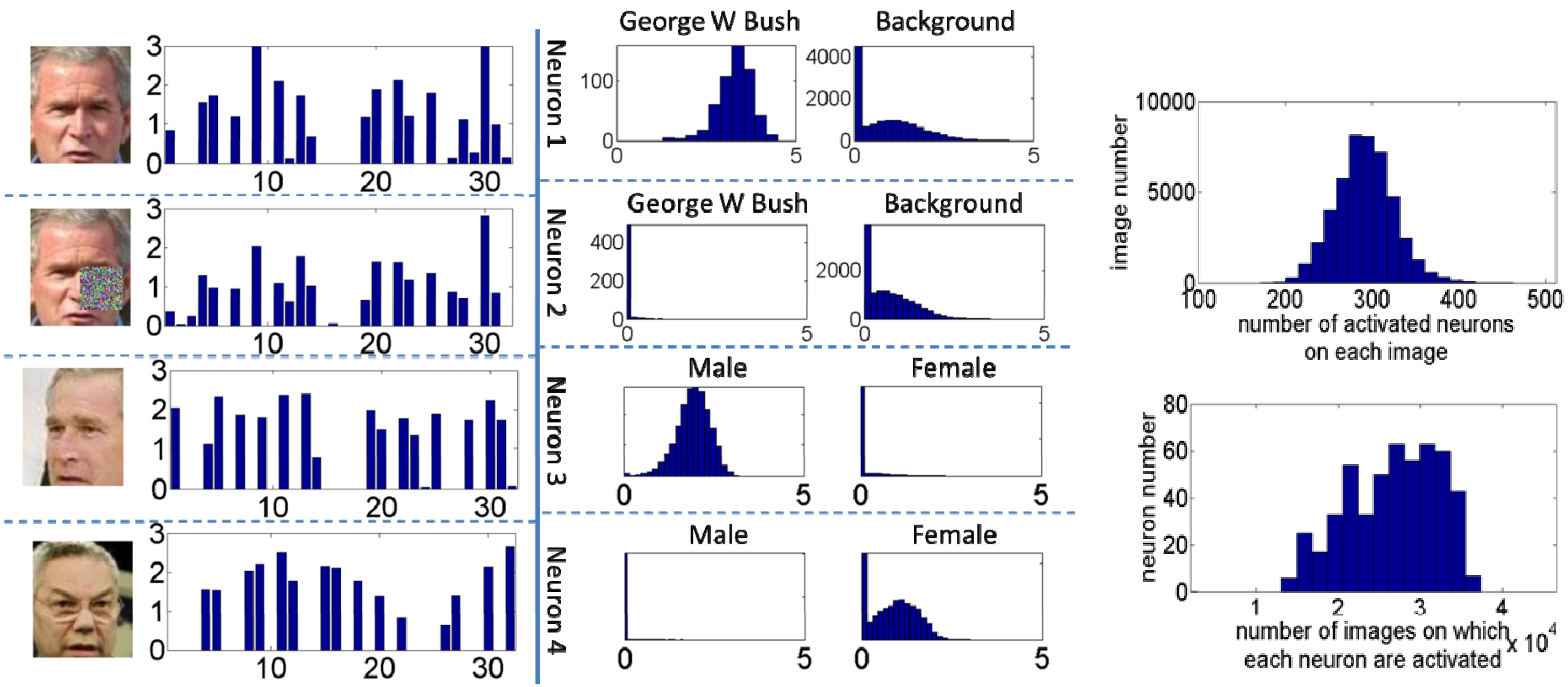


- Monkey has a face-processing network that is made of six interconnected face-selective regions
- Neurons in some of these regions were view-specific, while some others were tuned to identity across views
- View could be generalized to other factors, e.g. expressions?

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



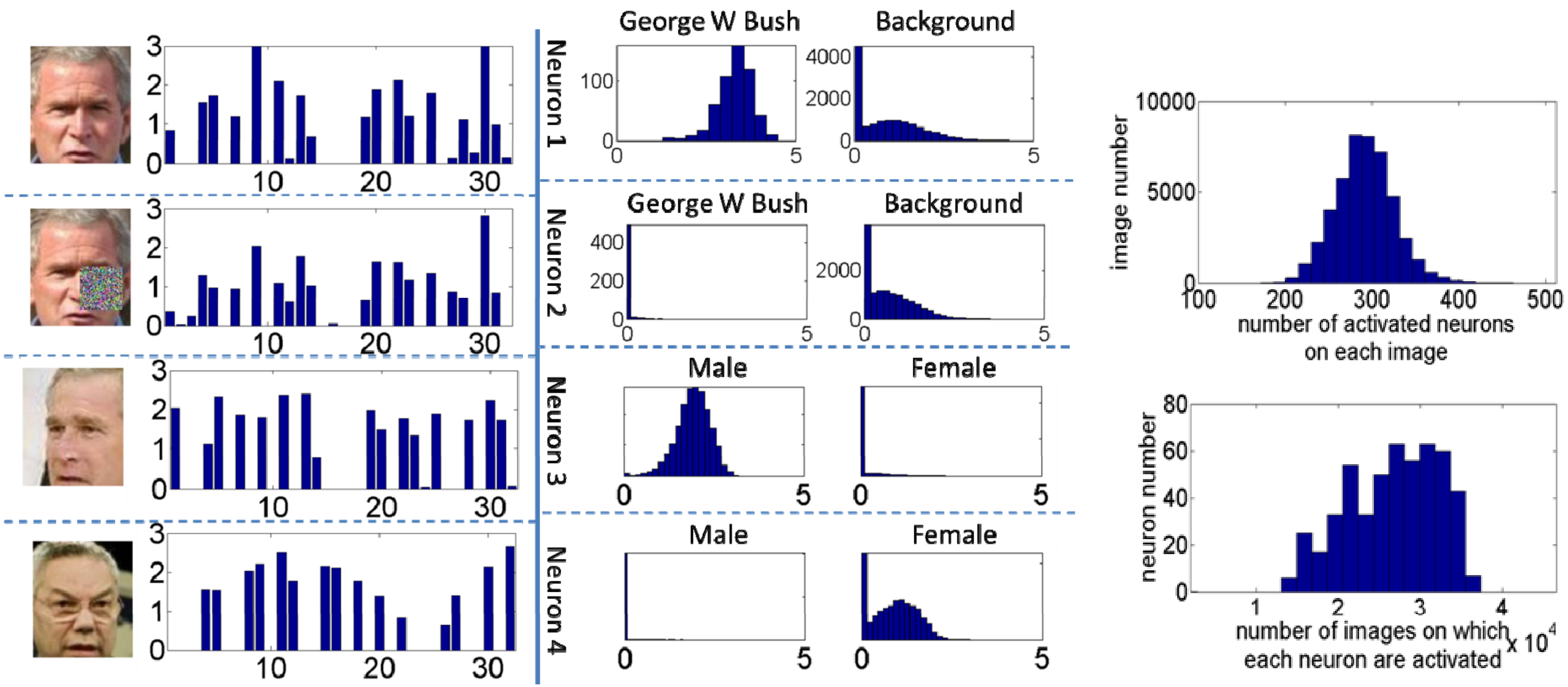
Deeply learned features are moderately space

- 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

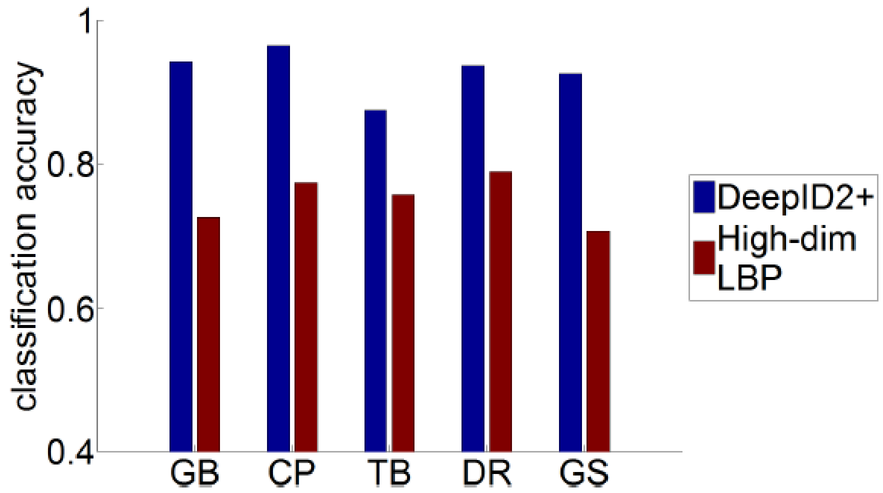
Deeply learned features are selective to identities and attributes

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

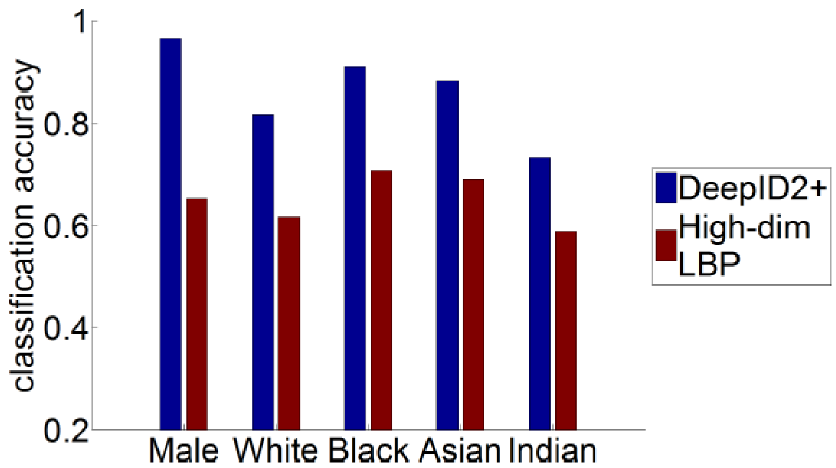


Deeply learned features are selective to identities and attributes

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



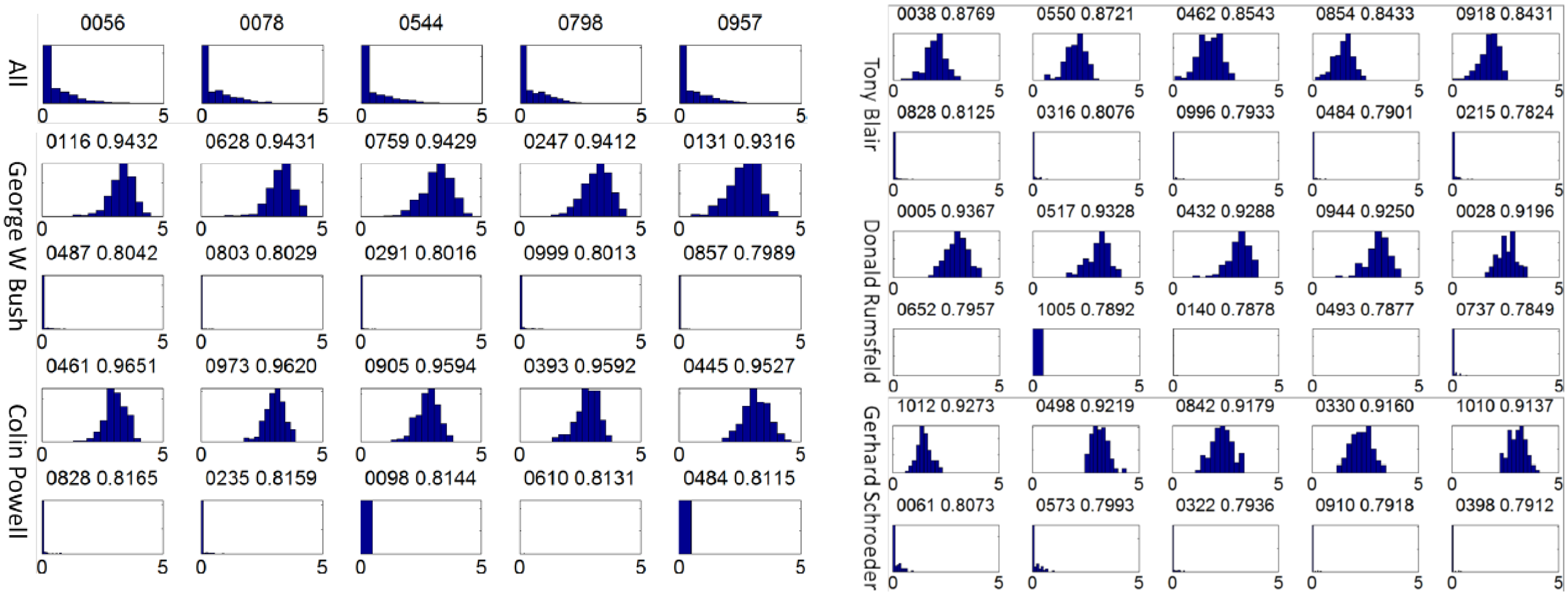
Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. GB, CP, TB, DR, and GS are five celebrities with the most images in LFW.



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

Deeply learned features are selective to identities and attributes

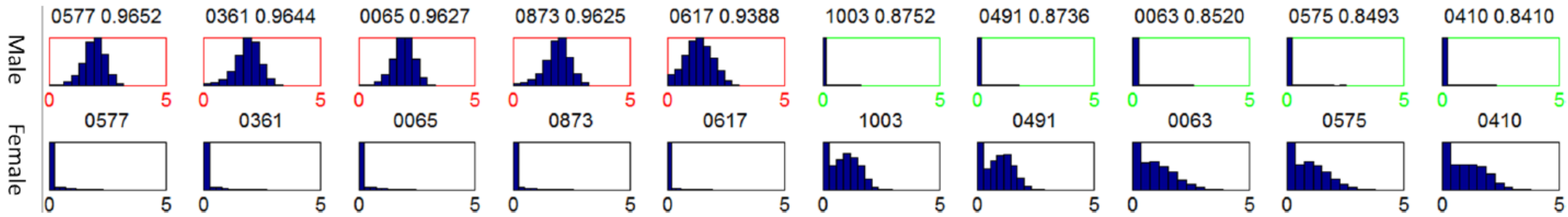
- Excitatory and inhibitory neurons



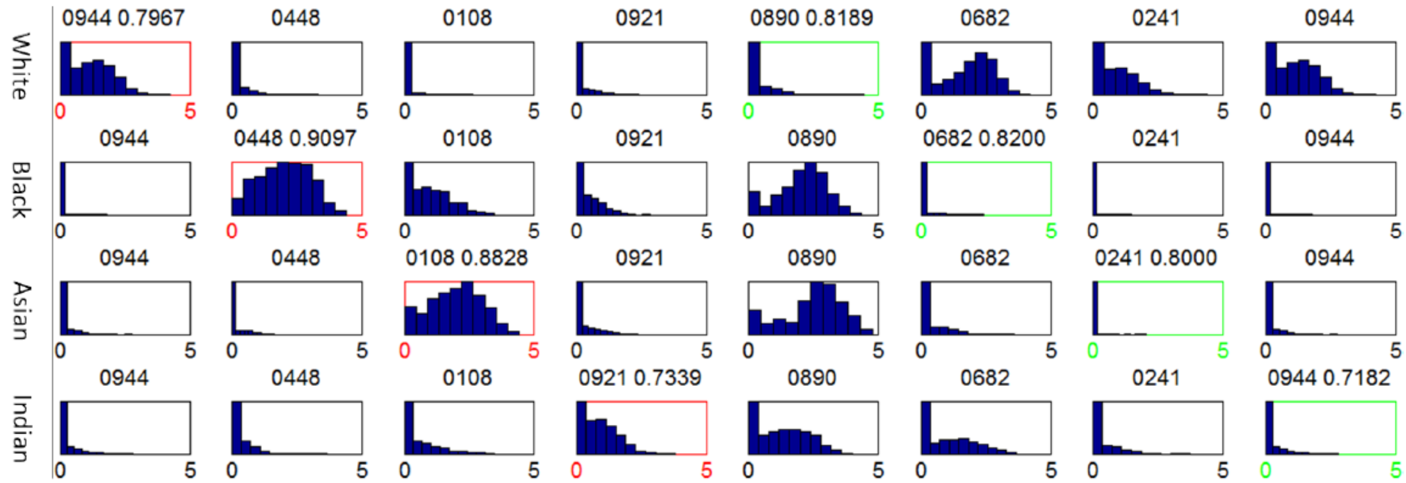
Histograms of neural activations over identities with the most images in LFW

Deeply learned features are selective to identities and attributes

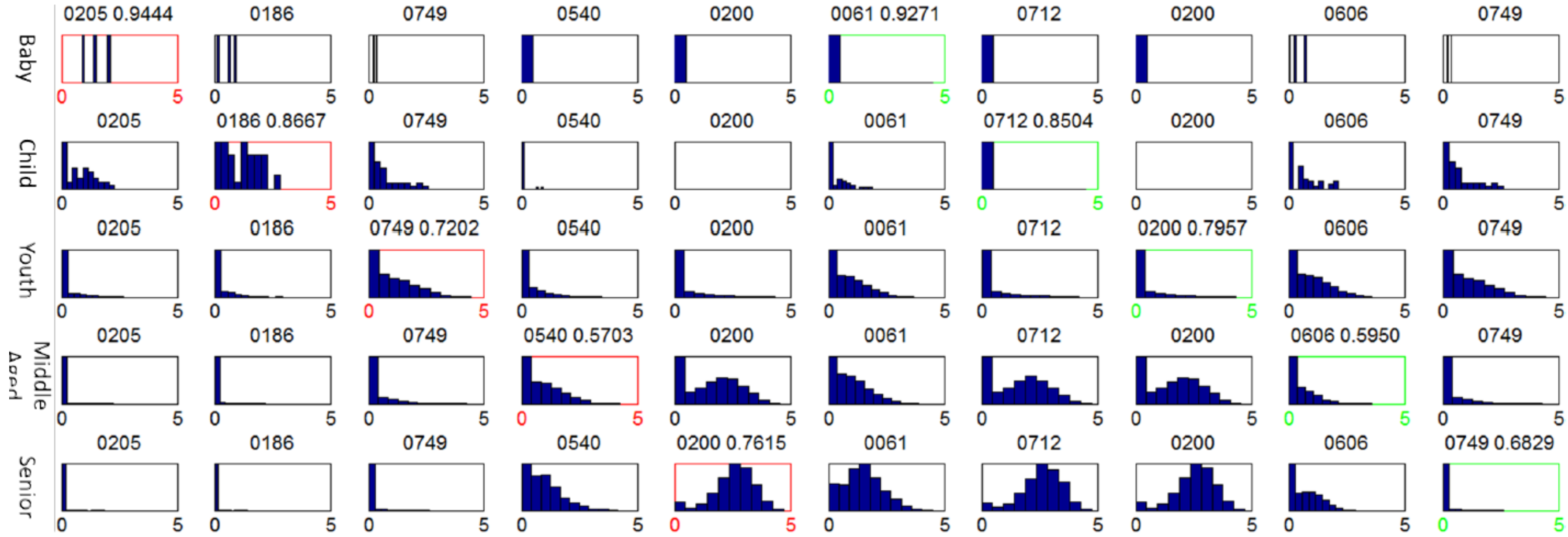
- Excitatory and inhibitory neurons



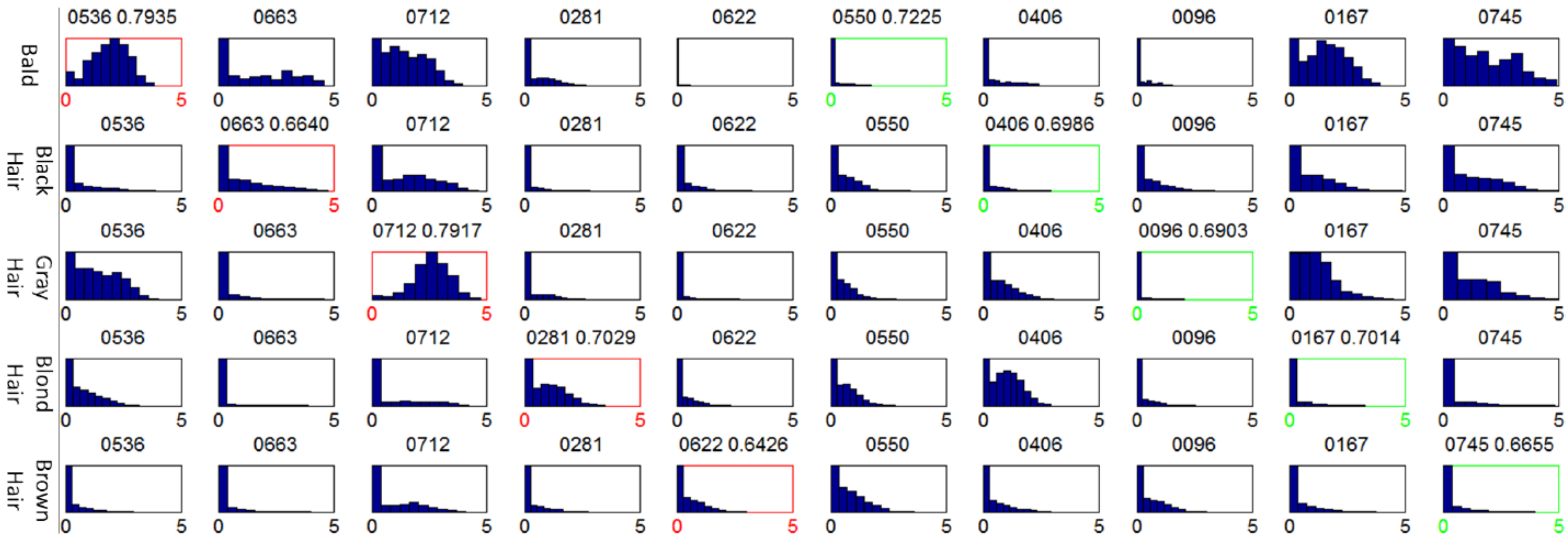
Histograms of neural activations over gender-related attributes (Male and Female)



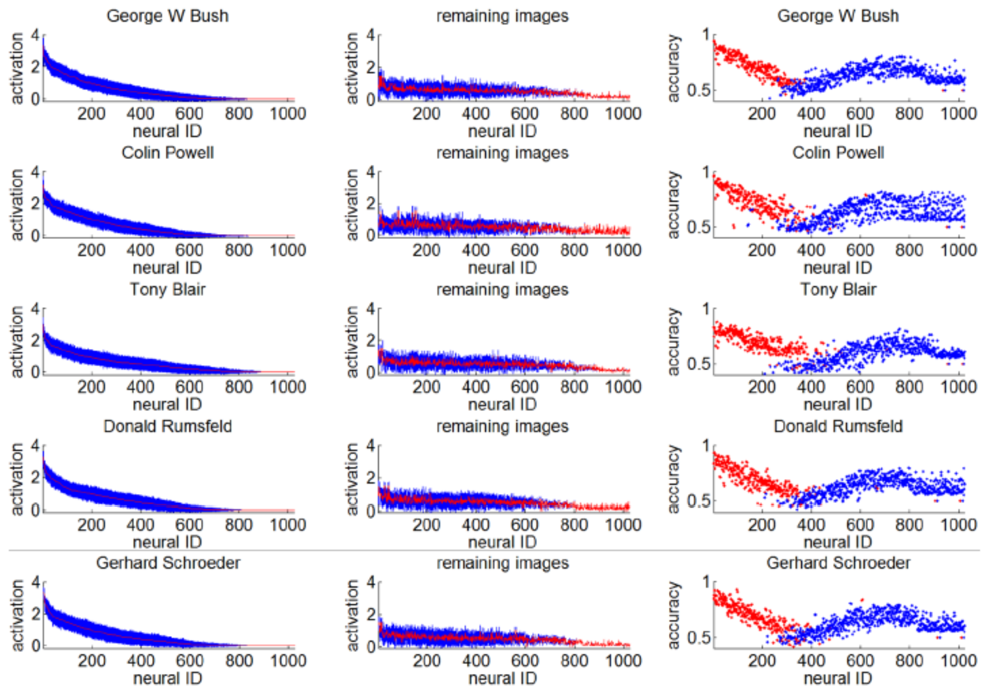
Histograms of neural activations over race-related attributes (White, Black, Asian and India)



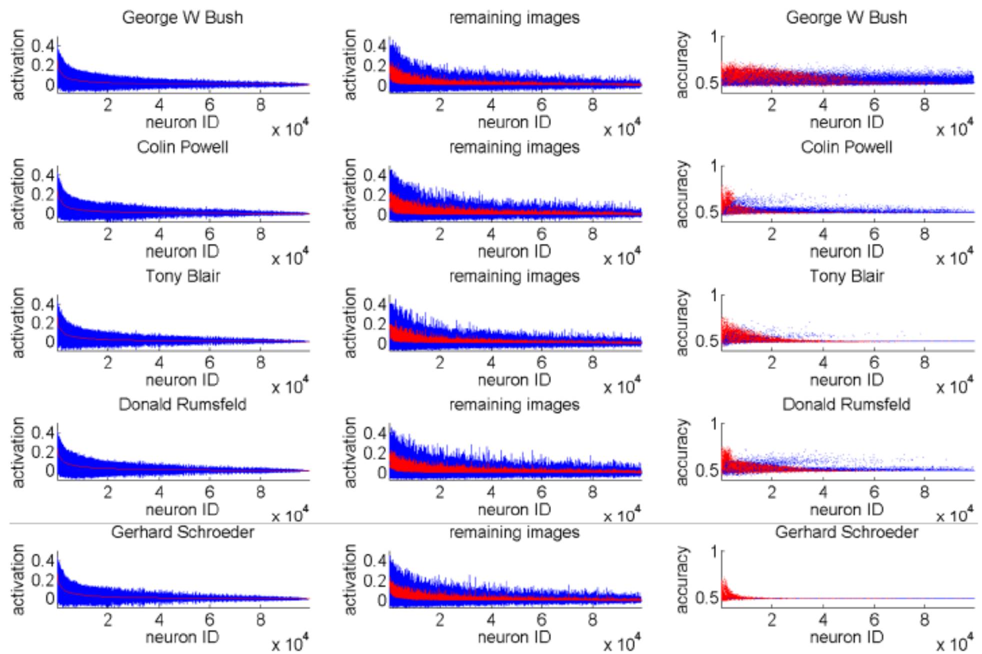
Histogram of neural activations over age-related attributes (Baby, Child, Youth, Middle Aged, and Senior)



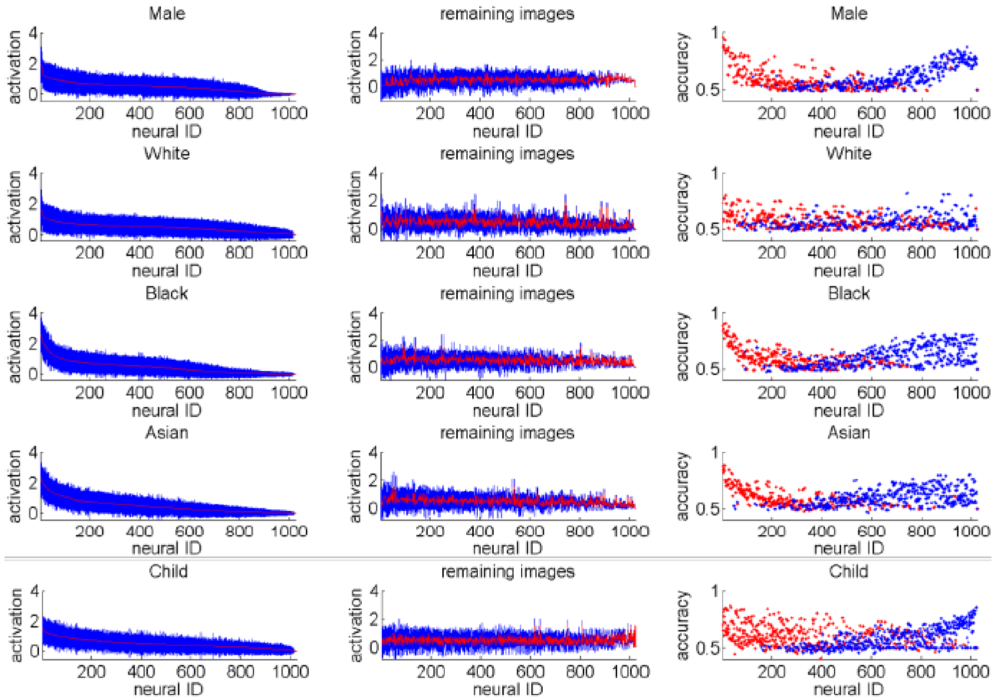
Histogram of neural activations over hair-related attributes (Bald, Black Hair, Gray Hair, Blond Hair, and Brown Hair.



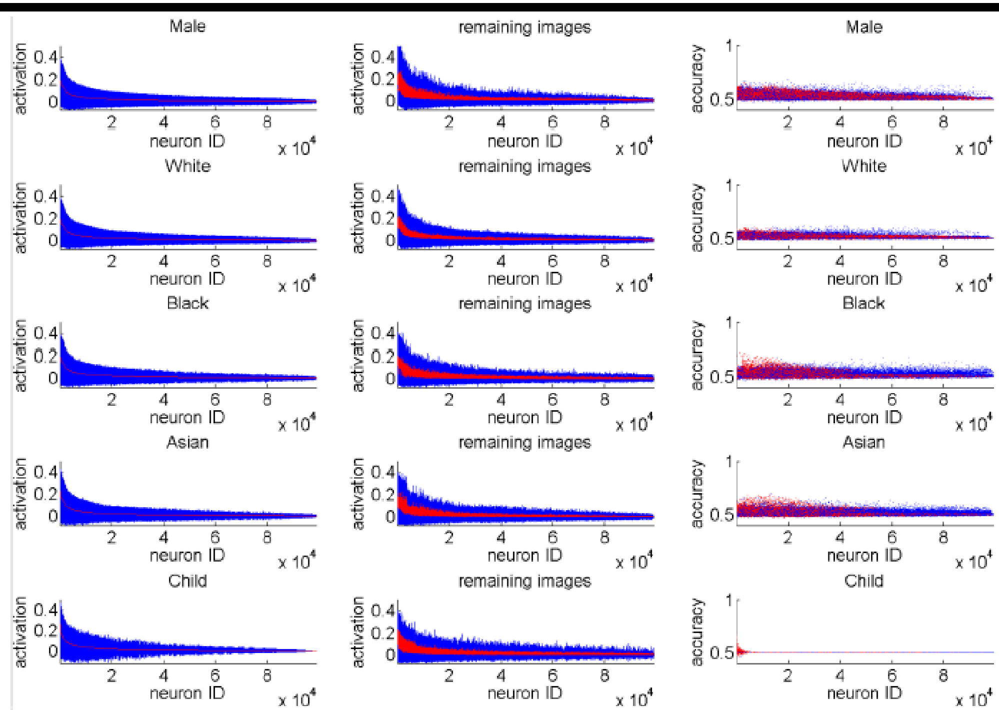
DeepID2+



High-dim LBP



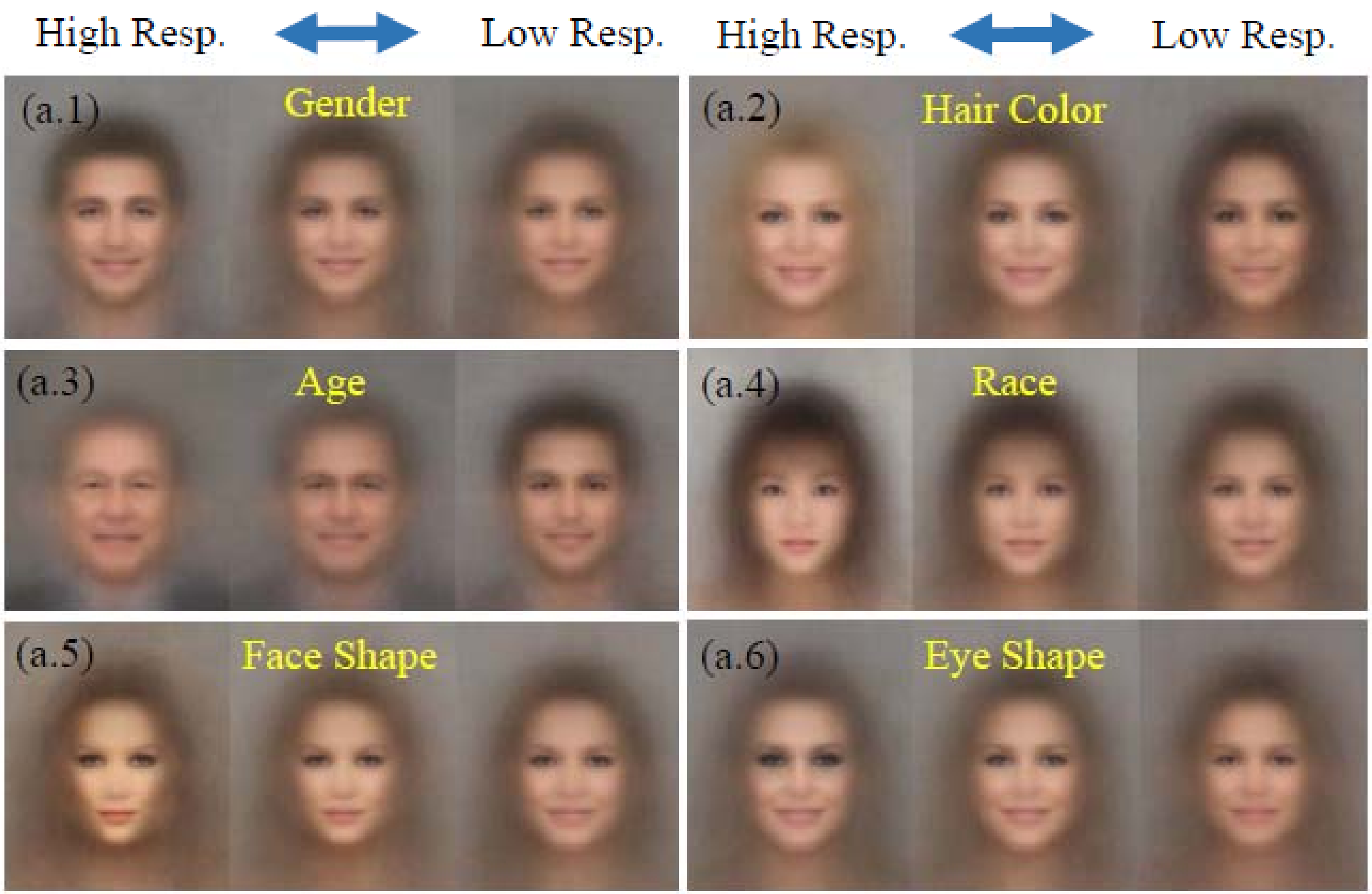
DeepID2+



High-dim LBP

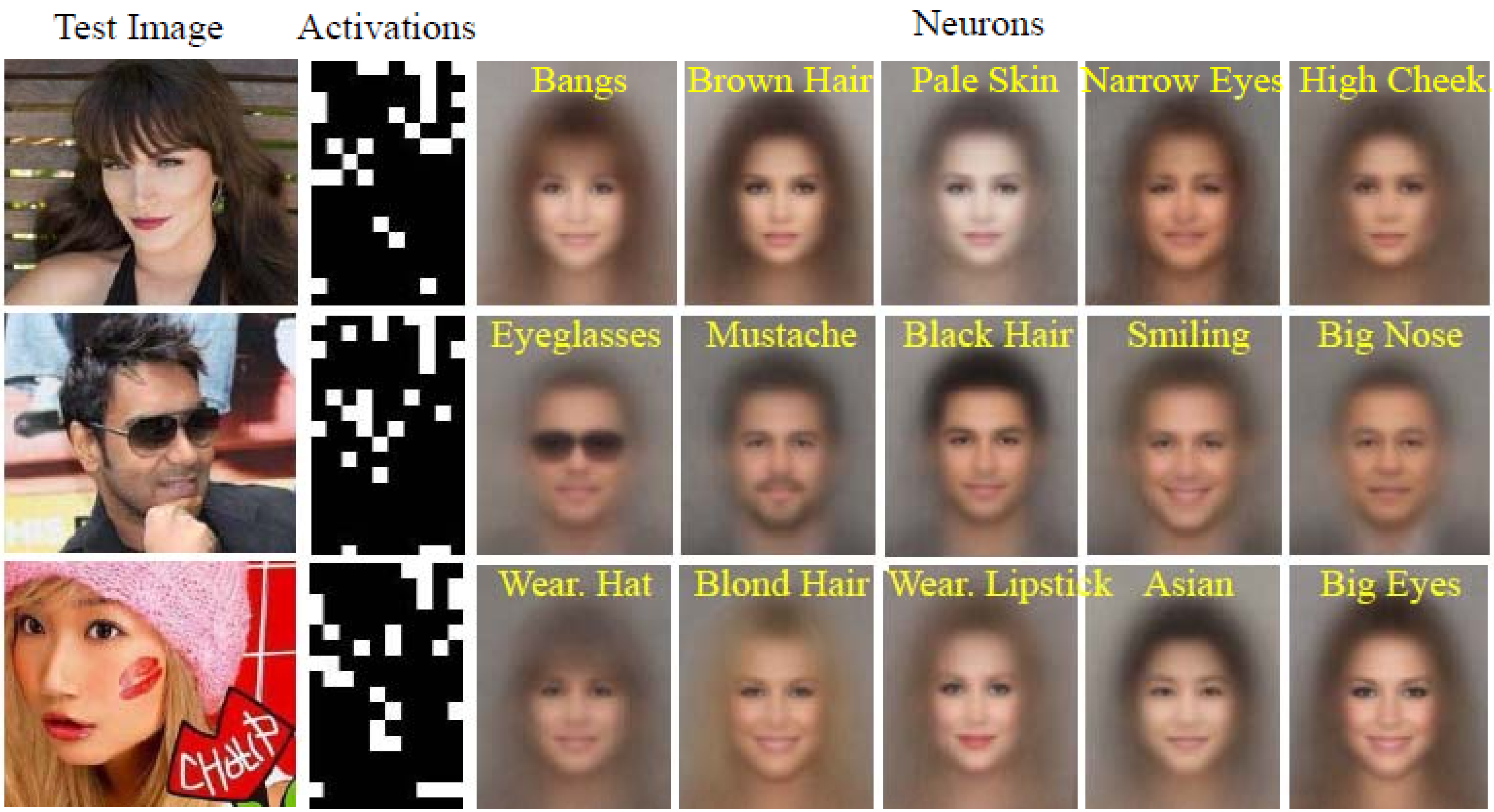
Deeply learned features are selective to identities and attributes

- Visualize the semantic meaning of each neuron



Deeply learned features are selective to identities and attributes

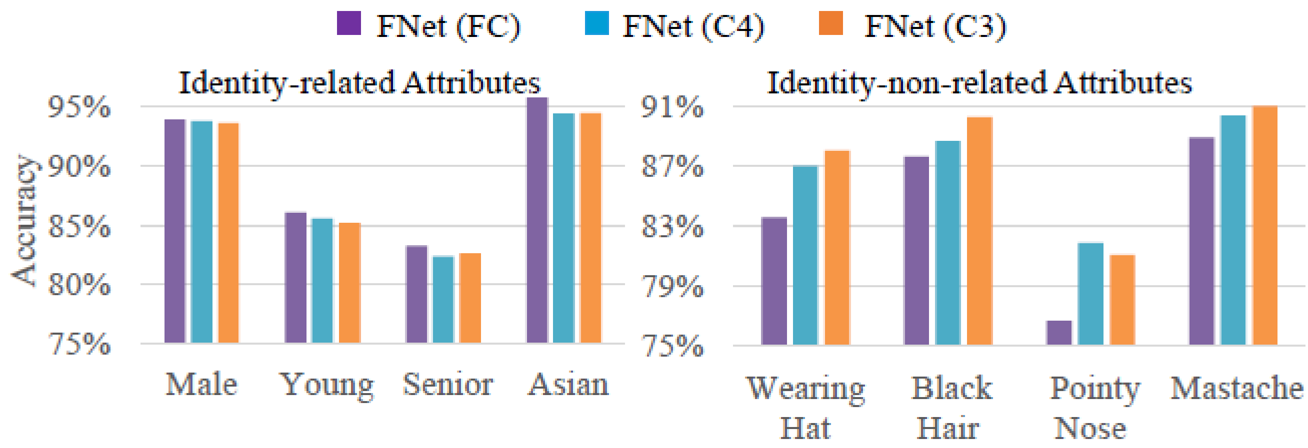
- Visualize the semantic meaning of each neuron



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

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

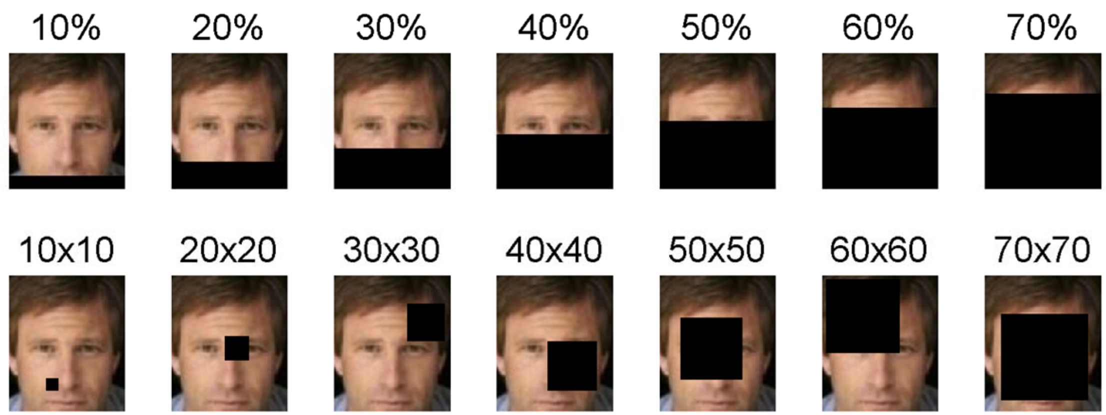
- DeepID2 features can be directly used for attribute recognition
- Use DeeID2 features as initialization (pre-trained result), and then fine tune on attribute recognition
- Average accuracy on 40 attributes on CelebA and LFWA datasets

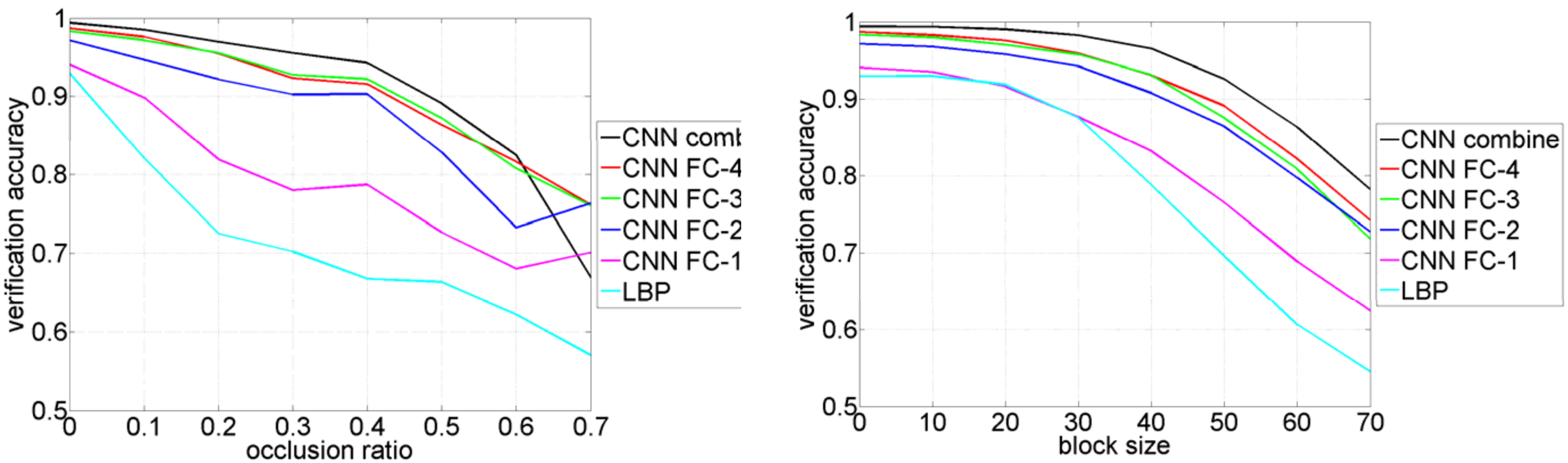
	CelebA	LFWA
FaceTracer [1] (HOG+SVM)	81	74
PANDA-W [2] (Parts are automatically detected)	79	71
PANDA-L [2] (Parts are given by ground truth)	85	81
DeepID2	**84**	**82**
Fine-tune (w/o DeepID2)	83	79
DeepID2 + fine-tune	**87**	**84**

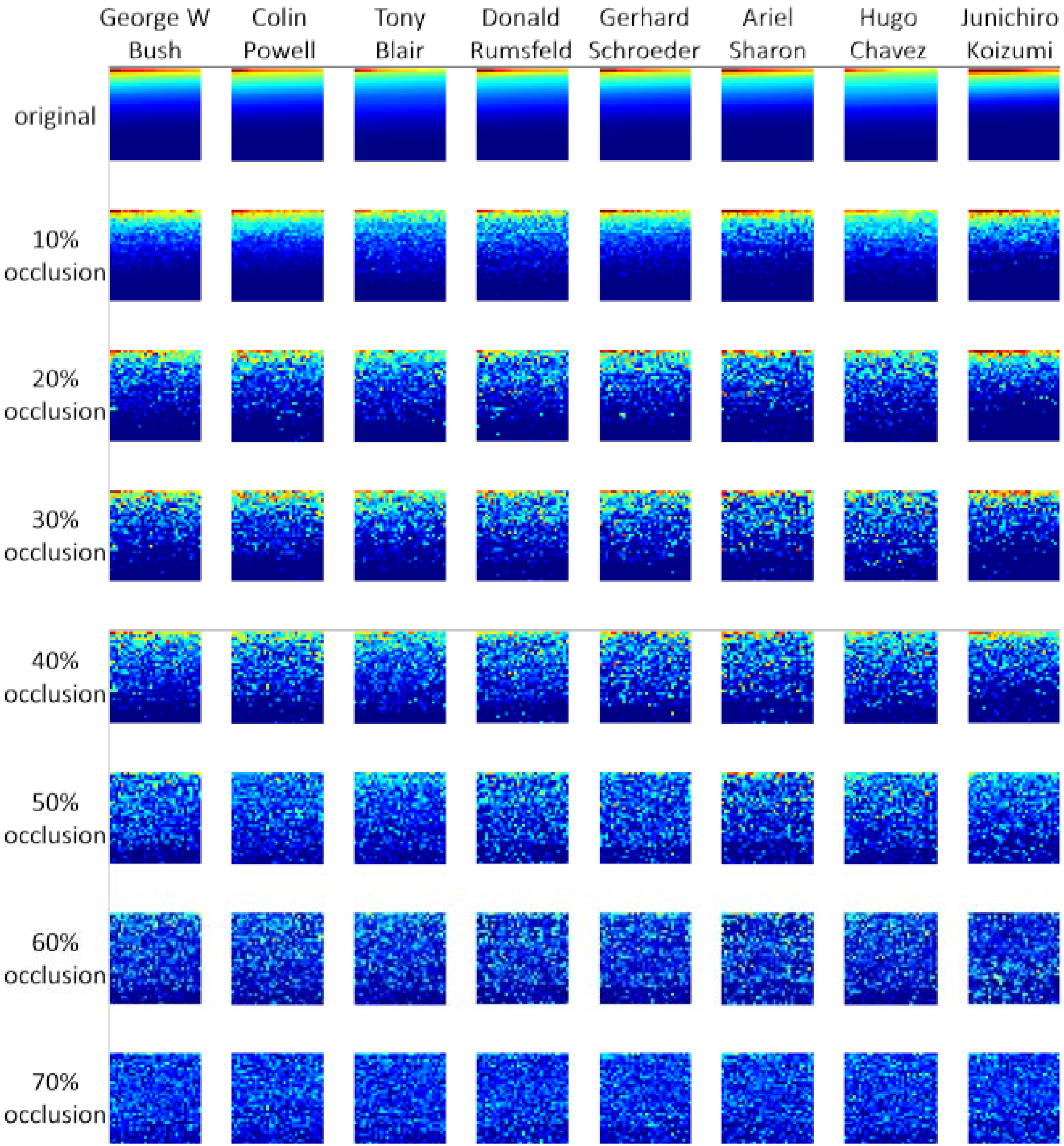
Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," arXiv:1411.7766, 2014.

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

Summary

- Automatically learns hierarchical feature representations from data and disentangles hidden factors of input data through multi-level nonlinear mappings
- For some tasks, the expressive power of deep models increases exponentially as their architectures go deep
- Jointly optimize all the components in a vision and crate synergy through close interactions among them
- Benefitting the large learning capacity of deep models, we also recast some classical computer vision challenges as high-dimensional data transform problems and solve them from new perspectives
- It is more effective to train deep models with challenging tasks and rich predictions

Summary

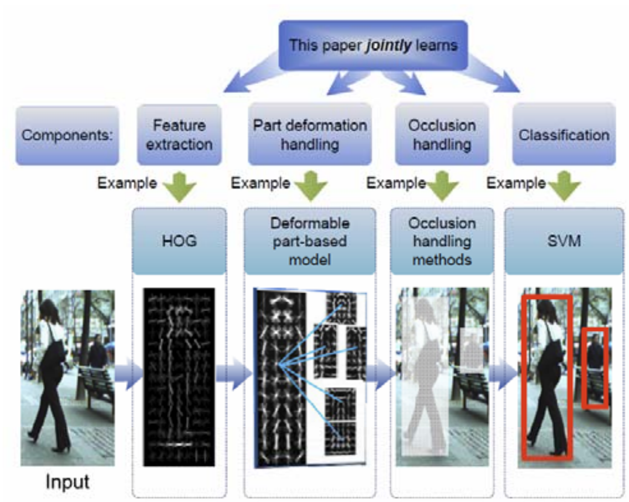
- Deeply learned features are moderately sparse, identity and attribute selective, and robust to data corruption
- Binary neuron activation patterns are effective for face recognition than activation magnitudes
- Neurons in the higher layers are more robust to occlusions and more effective on recognizing identity related attributes; while neurons in the lower layers are more effective on the remaining attributes
- These properties are naturally learned by DeepID2+ through large-scale training

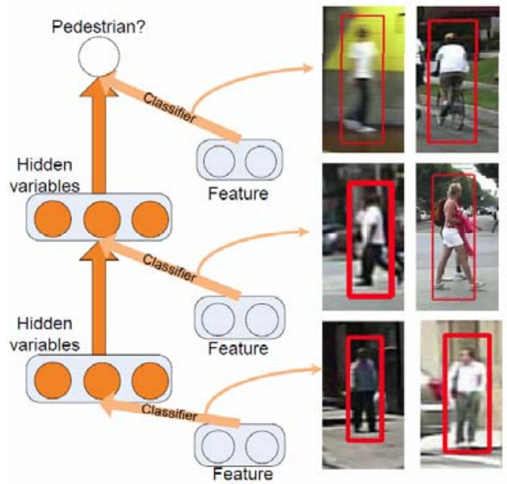
References

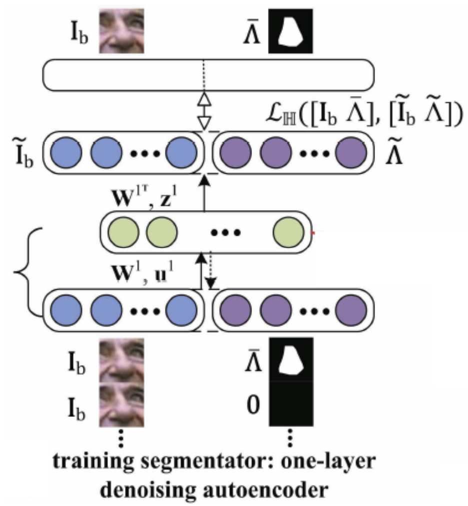
- D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Representations by Back-propagation Errors," Nature, Vol. 323, pp. 533-536, 1986.
- N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, L. Wiskott, "Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?" IEEE Trans. PAMI, Vol. 35, pp. 1847-1871, 2013.
- A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Proc. NIPS, 2012.
- Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation by Joint Identification-Verification," NIPS, 2014.
- K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," Biological Cybernetics, Vol. 36, pp. 193-202, 1980.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.
- G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, Vol. 18, pp. 1527-1544, 2006.

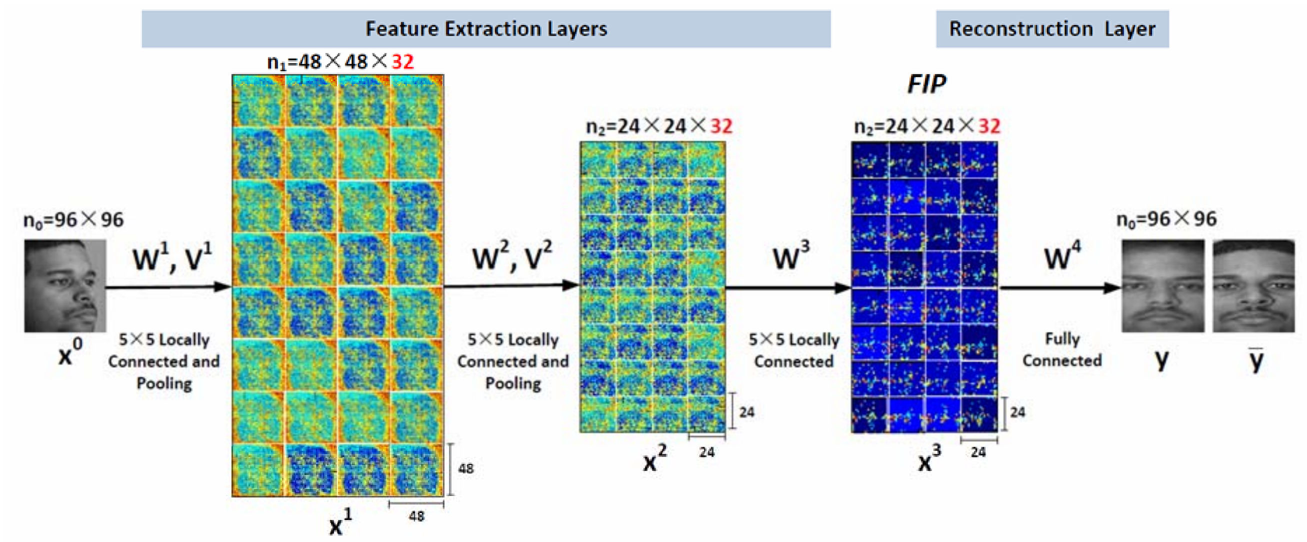
- G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” Science, Vol. 313, pp. 504-507, July 2006.
- Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Identity Face Space,” Proc. ICCV, 2013.
- Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning and Disentangling Face Representation by Multi-View Perception,” NIPS 2014.
- Y. Sun, X. Wang, and X. Tang, “Deep Learning Face Representation from Predicting 10,000 classes,” Proc. CVPR, 2014.
- J. Hastad, “Almost Optimal Lower Bounds for Small Depth Circuits,” Proc. ACM Symposium on Theory of Computing, 1986.
- J. Hastad and M. Goldmann, “On the Power of Small-Depth Threshold Circuits,” Computational Complexity, Vol. 1, pp. 113-129, 1991.
- A. Yao, “Separating the Polynomial-time Hierarchy by Oracles,” Proc. IEEE Symposium on Foundations of Computer Science, 1985.
- Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” CVPR 2013.
- W. Ouyang and X. Wang, “Joint Deep Learning for Pedestrian Detection,” Proc. ICCV, 2013.
- P. Luo, X. Wang and X. Tang, “Hierarchical Face Parsing via Deep Learning,” Proc. CVPR, 2012.
- Honglak Lee, “Tutorial on Deep Learning and Applications,” NIPS 2010.

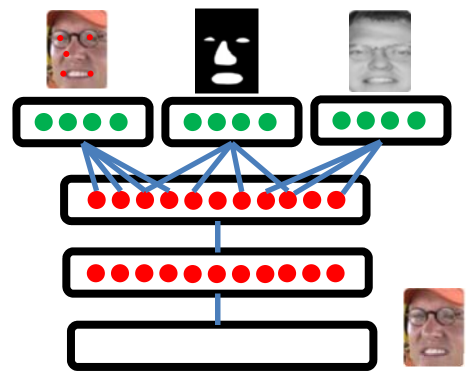
Thank you!











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