香港中文大學
The Chinese University of Hong Kong

ELEG 5040
Advanced Topics on Signal Processing
(Introduction to Deep Learning)

Xiaogang Wang

xgwang@ee.cuhk.edu.hk

Department of Electronic Engineering,
The Chinese University of Hong Kong

Course Information

- Course webpage

Signup link:

piazza.com/cuhk.edu.hk/spring2015/eleg5040

(If you want to take this course, please sign up with the link above)

Class link:

piazza.com/cuhk.edu.hk/spring2015/eleg5040/home

(If you are a listener, you still can access lecture notes from the link above without signup)

Course Information

- Instructor: Xiaogang Wang
  - SHB 415
  - Office hours: after Monday's class or by appointment
- Tutor: Kai Kang
  - SHB 304
  - kkang@ee.cuhk.edu.hk
  - Office hours: after the tutorial or by appointment

Course Information

- Lecture time & venue
  – Monday: 4:30pm – 5:15pm, ERB 404
  – Tuesday: 2:30pm – 4:15pm, ERB Lecture Theatre
- Options of tutorial time
  – Monday 17:30 - 18:15
  – Tuesday 17:30 - 18:15
  – Thursday 15:30 – 16:15
  – Friday 16:30 – 17:15

Course Information

- Homework (15%)
- Midterm (15%)
- Final exam (30%)
- Project (40%)
  - Applications of deep learning
  - Implementation of deep learning
  - Study deep learning algorithms
  - You should submit
    - A term paper of 4 pages (excluding figures) in maximum, double column, font size is equal or larger than 10.
    - Code and sample data
    - Project presentation
  - No survey
  - No collaboration

Course Information

- Examples of project topics
  - Implement CNN with GPU and compare its efficiency with Caffe
  - Modify Caffe to make it support multiple GPUs
  - Fast CPU implementation of CNN
  - We provide a baseline model of GoogLeNet on ImageNet, and you try improve it
  - Choose one of the deep learning related competitions (such as ImageNet), and compare your result with published ones
  - Propose a deep model to effectively learn dynamic features from videos
  - Deep learning for speech recognition
  - Deep learning for object detection
  - Will provide more later …
  - We you can discuss your topics with me at any time

Lectures

Week	Chapter	Content
1	Introduction to deep learning	Historical review of DL. DL makes difference. Classical deep models. Why DL works?
2	Machine learning basics	Classification, regression, capacity, underfitting, overfitting, generalization error, regularization, curse of dimensionality, bias, variance, supervised learning, unsupervised learning, descriminative model, generative model
3	Multilayer neural network	Feedforward operation, backpropagation, nonlinear activation functions, expressive power of three-layer neural network, nonlinear feature mapping, target functions, gradient descent, feature/representation learning
4	Convolutional neural network (CNN)	CNN, filters, pooling, efficient convolution algorithms, BP for CNN

Lectures

Week	Chapter	Content
5	Deep belief nets	Boltzmann machine (BM), Restricted Boltzmann machine (RBM), contrast divergence, deep Bolzmann machines, deep belief nets, convolutional Boltzmann machines, unsupervised pre-training, layerwise pre-training
6	Auto-encoder	Auto-encoder, denoising auto-encoder, stacked auto-encoder, unsupervised learning, manifold learning.
7	Optimization for training deep models	Stochastic gradient descent, pre-training, dropout, data augmentation, mini-batch, monitor the training process, learning rate, error surfaces, local minimum, normalization, data augmentation, initialization, training very deep networks
8	Midterm	
9	Large scale deep learning	Fast CPU implementation, GPU implementation, asynchronous parallel implementation

Lectures

Week	Chapter	Content
10	Multi-task and transfer deep learning	Multi-task learning, transfer learning, domain adaptation
11	Recurrent neural networks	Recurrent neural networks, Long-Short-Term-Memory network
12-13	Applications of deep learning	Image classification, object detection, , automatic image caption, face recognition, speech recognition
14	Understanding deep learning	Disentangling hidden factors, sparseness, robust to data corruption, distributed representation, contextual modeling, learning effective feature representation from rich predictions, exponential gain in representational efficiency from distributed representations, exponential gain in representation efficiency from depth, visualization of deep models

Tutorials

Week	Content
1	No tutorial (decide time and venue)
2	Python
3	Theano
4-5	CUDA, GPU programming (given by engineer from NVIDIA)
6-7	Deep learning toolbox
8-10	Caffe
11-12	Sharing research experience on deep learning
13-14	Review of lectures

6-hours Tutorial on CUDA/GPU Programming for DNN

- Instructor: Bin Zhou, PhD
  - NVIDIA CUDA Fellow, USTC Adjunct Research Prof
  - Chief Scientist and Director of Marine Remote Sensing & Information Processing Lab, SDIOI
- Prerequisites
  - Computer architecture basics
  - C programming language
  - Numerical methods and analysis
  - Neural network
- Video lecture

http://www.iqiyi.com/a_19rrhbvoe9.html#vfrm=2-3-0-1

Topic	Content
Basics of CUDA (1.5h)	1. CPU Architecture Review 2. Very Brief Review of Parallel Computing 3. Development Environment Configuration & Tools 4. GPU Architecture Review 5. GPU/CUDA Programming & Memory Model 6. CUDA Programming By Examples
Debugging, Profiling & Tools for CUDA/GPU (1 h, with lab works)	1. Programming, Compiling 2. Debugging under windows & Linux 3. Profiling for Performance 4. Library and Tools
DNN with GPU/CUDA (1.5h, with lab works)	1. Simple neural network with CUDA 2. cuDNN and cuda-convnet 3. Hands-on work for NN, cuDNN and cuda-convnet
CUDA Optimization for DNN (1h)	1. General Optimization Procedure & Consideration 2. Efficient CUDA Programming Skills 3. Memory Throughput Optimization 4. DNN Analytical Optimization

Topic	Content
Advanced Topics with Multi-GPU and more (0.5h)	1. Multi-GPU, Multi-Node 2. RDMA and GPUDirect 3. Hyper-Q 4. Dynamic Parallelization 5. Tegra K1
We're dealing with GPU/CUDA contents….	1. Programming Model? 2. Memory Model? 3. WARP? 4. Occupancy? 5. Optimization 6. Compute Bound or memory Bound? 7. CUDA-GDB 8. Parallel Nsight?

Introduction to Deep Learning

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

Machine Learning

$x \longrightarrow \boxed{F(x)} \longrightarrow y$

y:
- Class label (Classification)
- Vector (Estimation)

Object recognition → {dog, cat, horse, flower, ...}

Low-resolution image

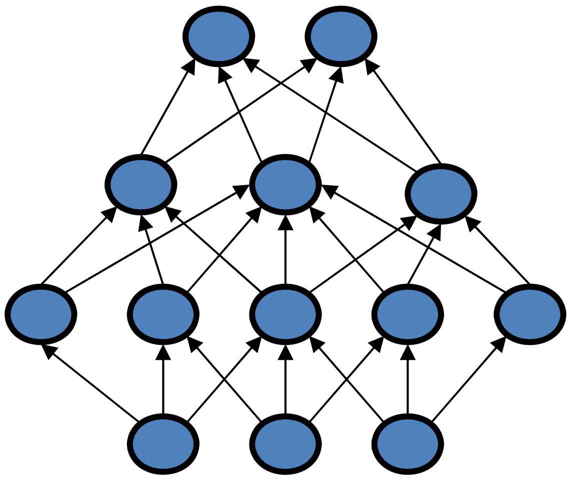
Super resolution →

High-resolution image

Neural network
Back propagation

Nature

1986

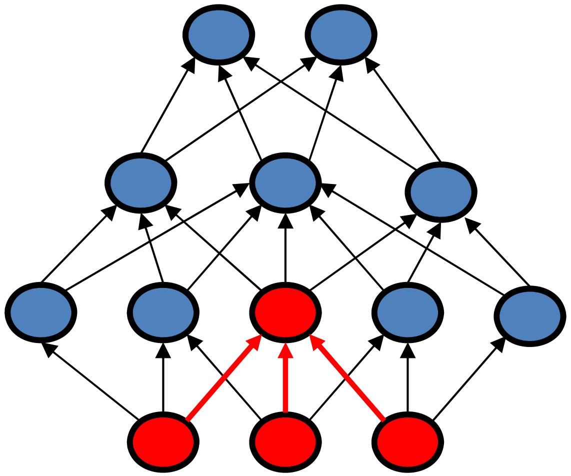


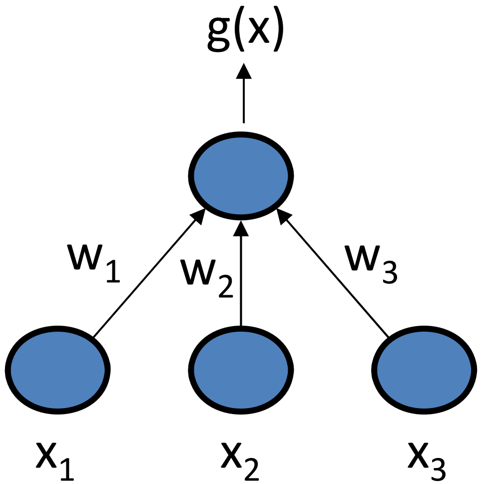
- Solve general learning problems
- Tied with biological system

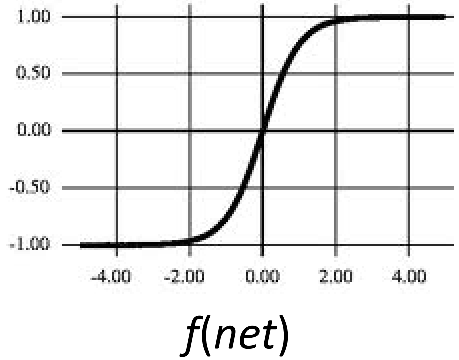
Neural network
Back propagation

Nature

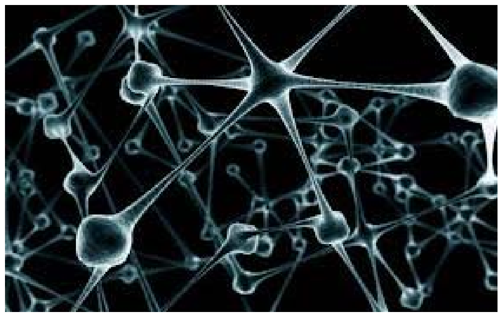
1986







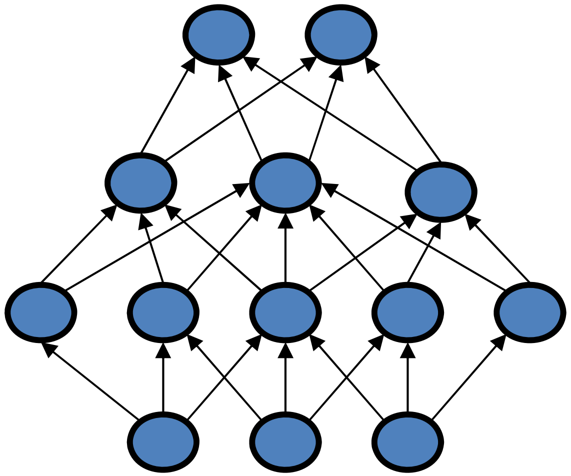
$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t \mathbf{x})$$



Neural network
Back propagation

Nature

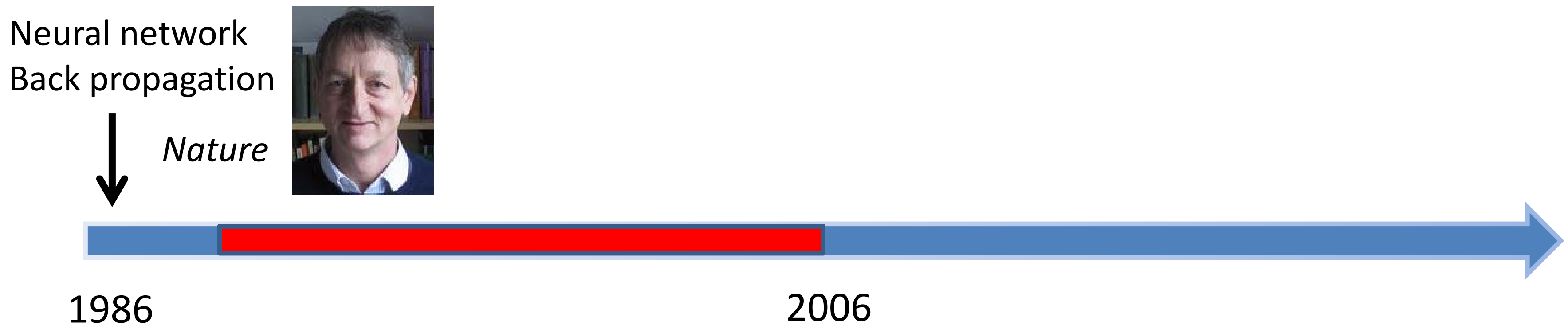
1986



- Solve general learning problems
- Tied with biological system

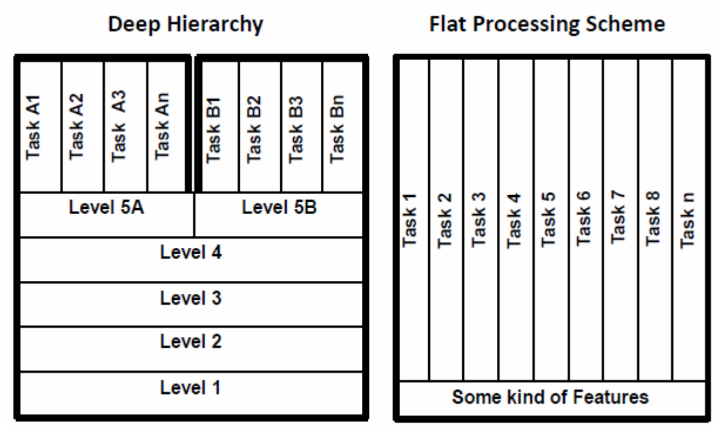
But it is given up…

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

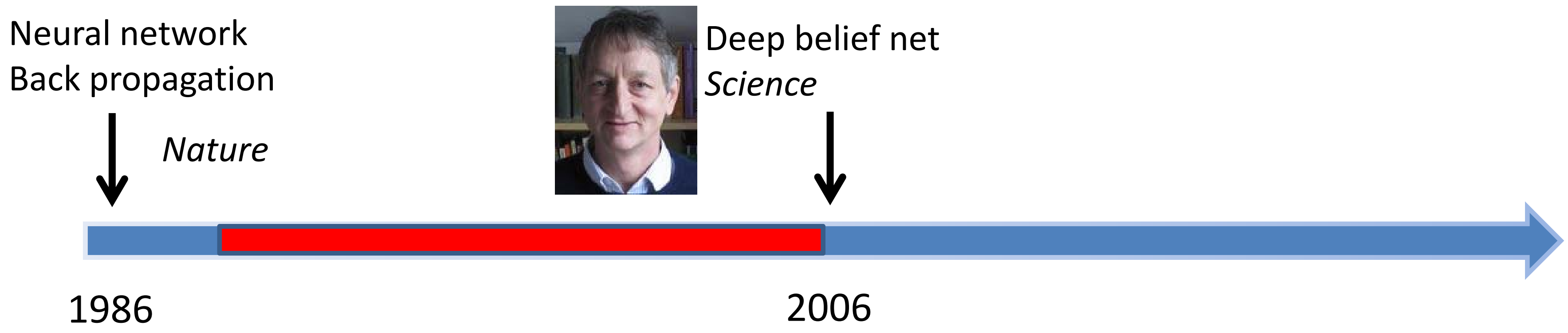


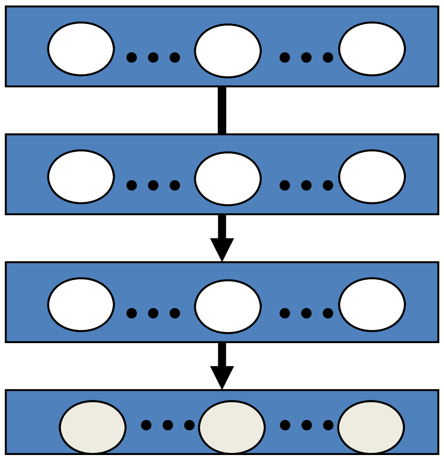
- SVM
- Boosting
- Decision tree
- KNN
- …

- 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



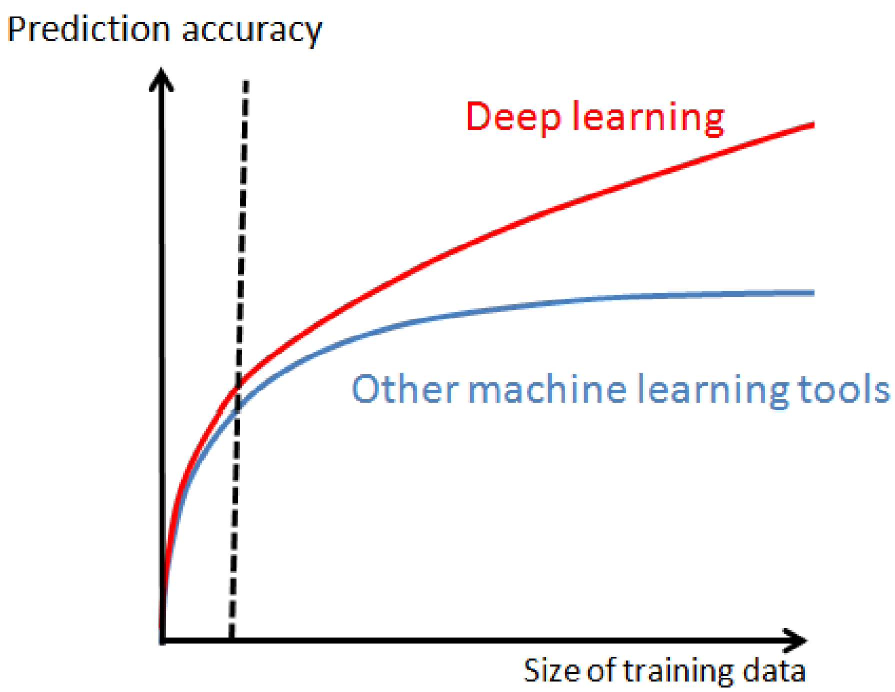


- Unsupervised & Layer-wised pre-training
- Better designs for modeling and training (normalization, nonlinearity, dropout)
- New development of computer architectures
  - GPU
  - Multi-core computer systems
- Large scale databases

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

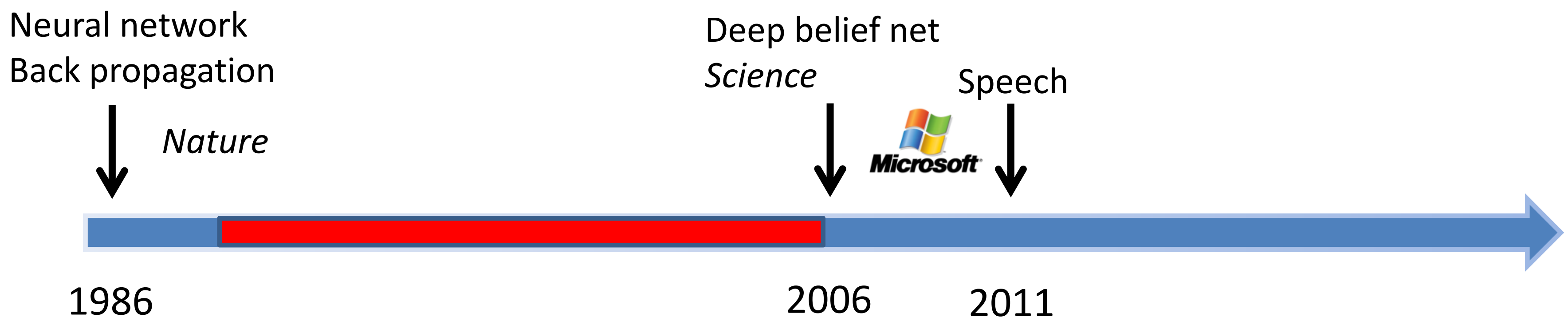
⇩

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.

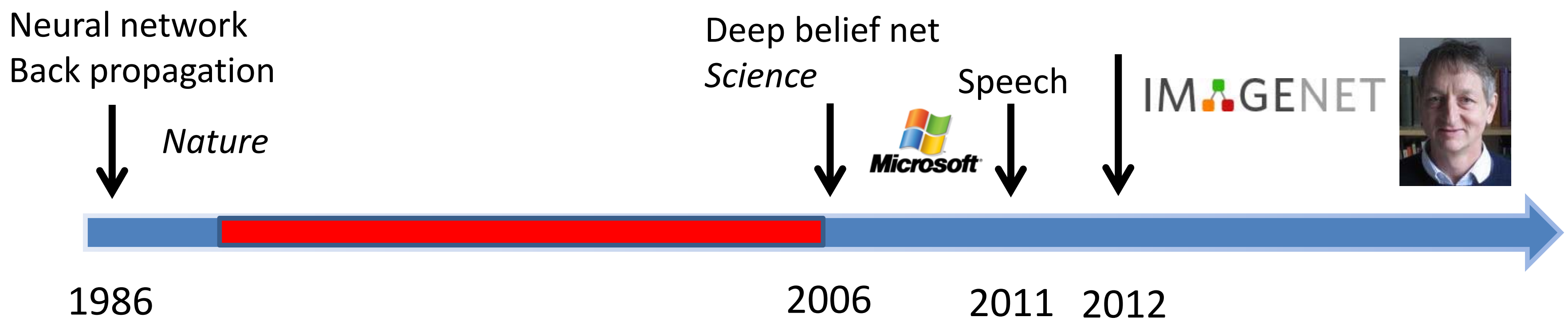


deep learning results

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.

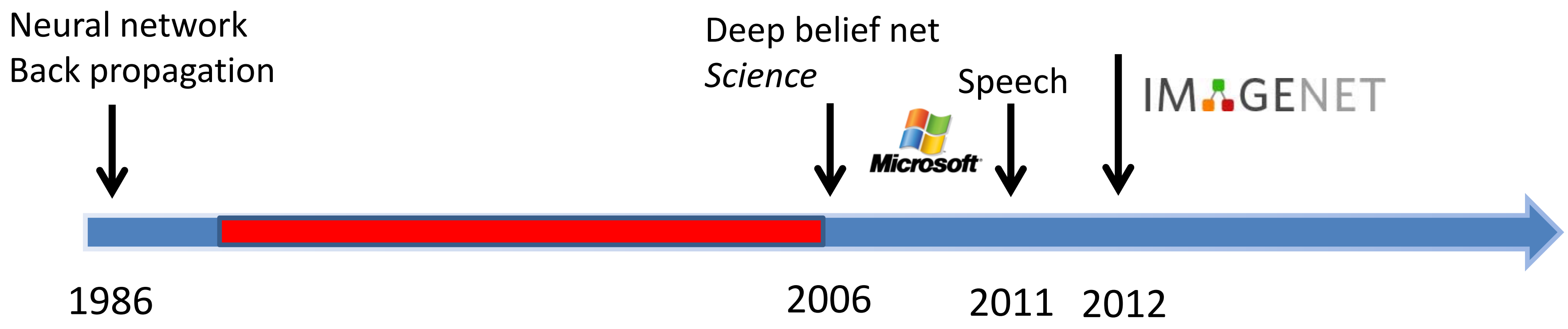


Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

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



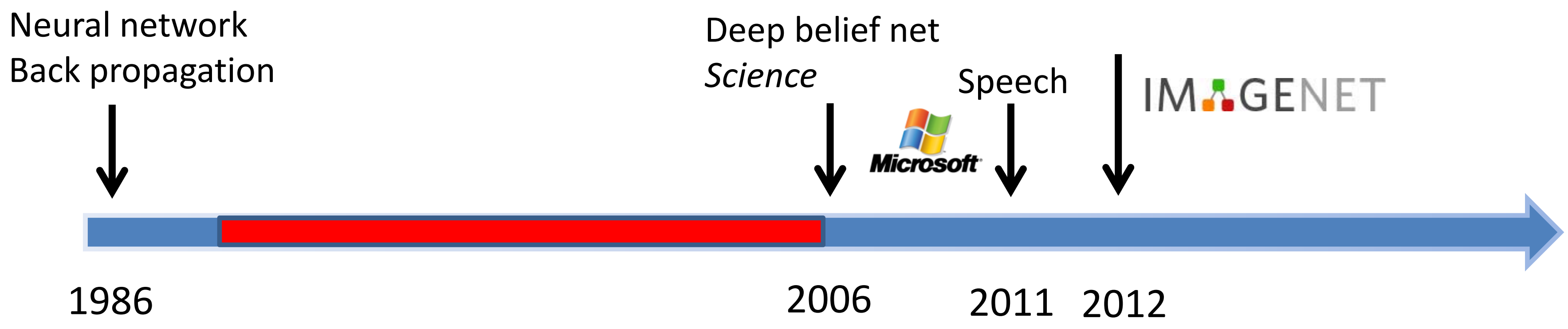
- ImageNet 2013 – image classification challenge

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

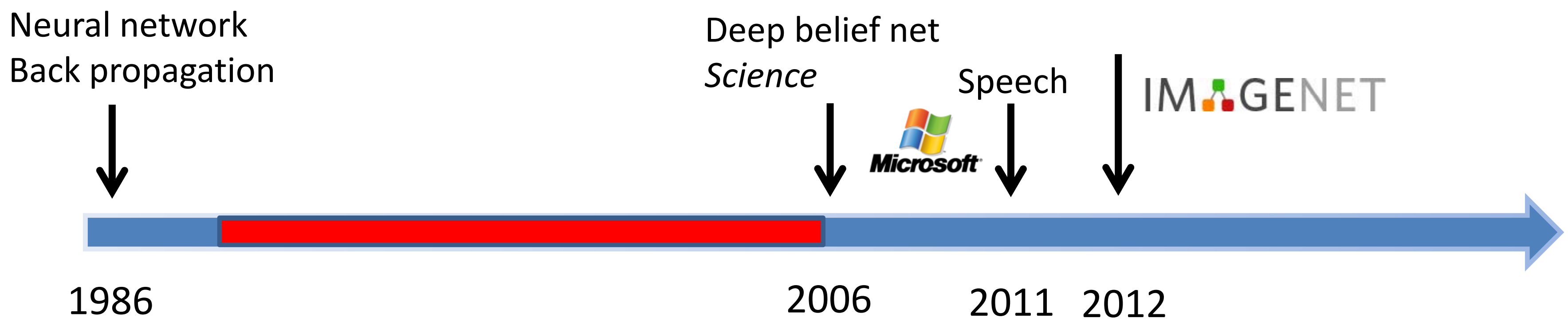


- ImageNet 2014 – Image classification challenge

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

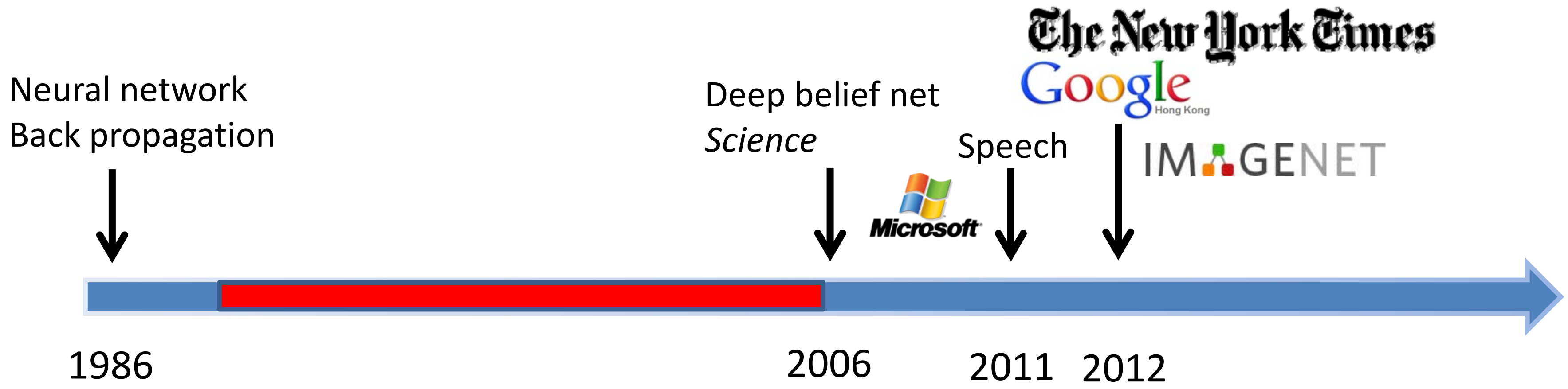
Rank	Name	Mean Average Precision	Description
1	Google	0.43933	Deep learning
2	CUHK	0.40656	Deep learning
3	DeepInsight	0.40452	Deep learning
4	UvA-Euvision	0.35421	Deep learning
5	Berkley Vision	0.34521	Deep learning



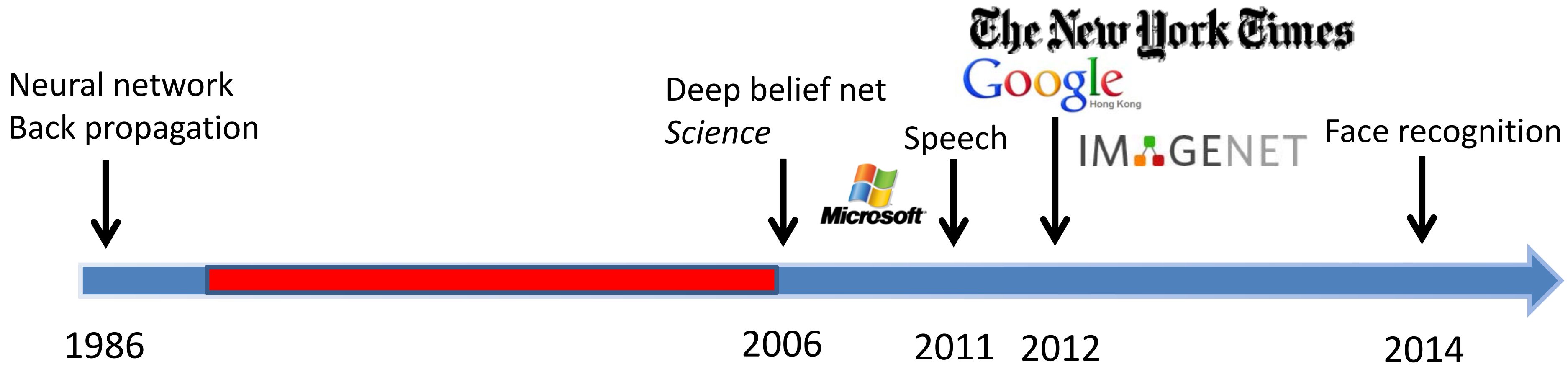
- ImageNet 2014 – object detection challenge

	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: multi-stage and deformable deep convolutional neural networks for object detection”, arXiv:1409.3505, 2014



- 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. arXiv:1412.1265, 2014.

Labeled Faces in the Wild (2007)

Best results without deep learning

Random guess (50%)

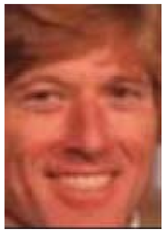
Eigenface (60%)

TL Joint Bayesian (96.33%), 2013

Human cropped (97.53%)

Human funneled (99.20%)

Our deep learning result (99.47%)



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

MIT Technology Review

BUSINESS NEWS

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

The acquisition, aimed at adding skilled experts rather than specific products, marks an acceleration in efforts by Google, Facebook, and other Internet firms to monpololize the biggest brains in artificial intelligence research.

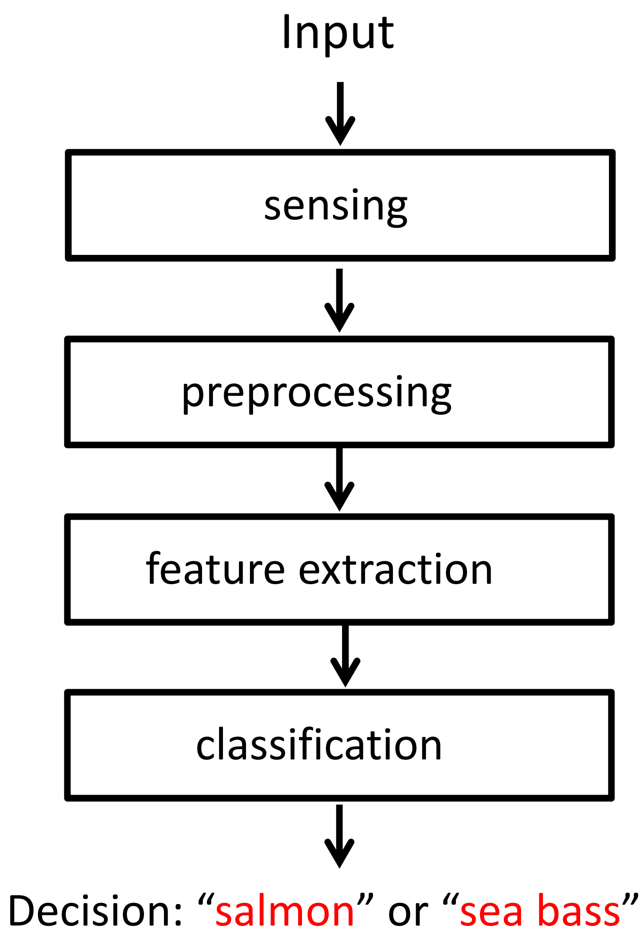
News on Deep Learning

Baidu established Institute of Deep Learning	2012
Hinton's group won ImageNet Contest	Oct. 2012
Hinton joined Google	March 2013
Google announced deep learning based visual search engine	March 2013
Baidu announced deep learning based visual search engine	June 2013
Yahoo acquired startup LookFlow working on deep learning	Oct. 2013
Facebook established a new AI lab in NewYork and recruited Yann LeCun	Dec. 2013
Google Acquires DeepMind for USD 400 Million	January 2014
Baidu established a new lab at Shenzhen, China	2014
Baidu established a new lab at silicon valley and Andrwe Ng is the director	May 2014
Deep learning reached human performance on face verification on LFW	June 2014

Deep Learning Changes the Design Cycle

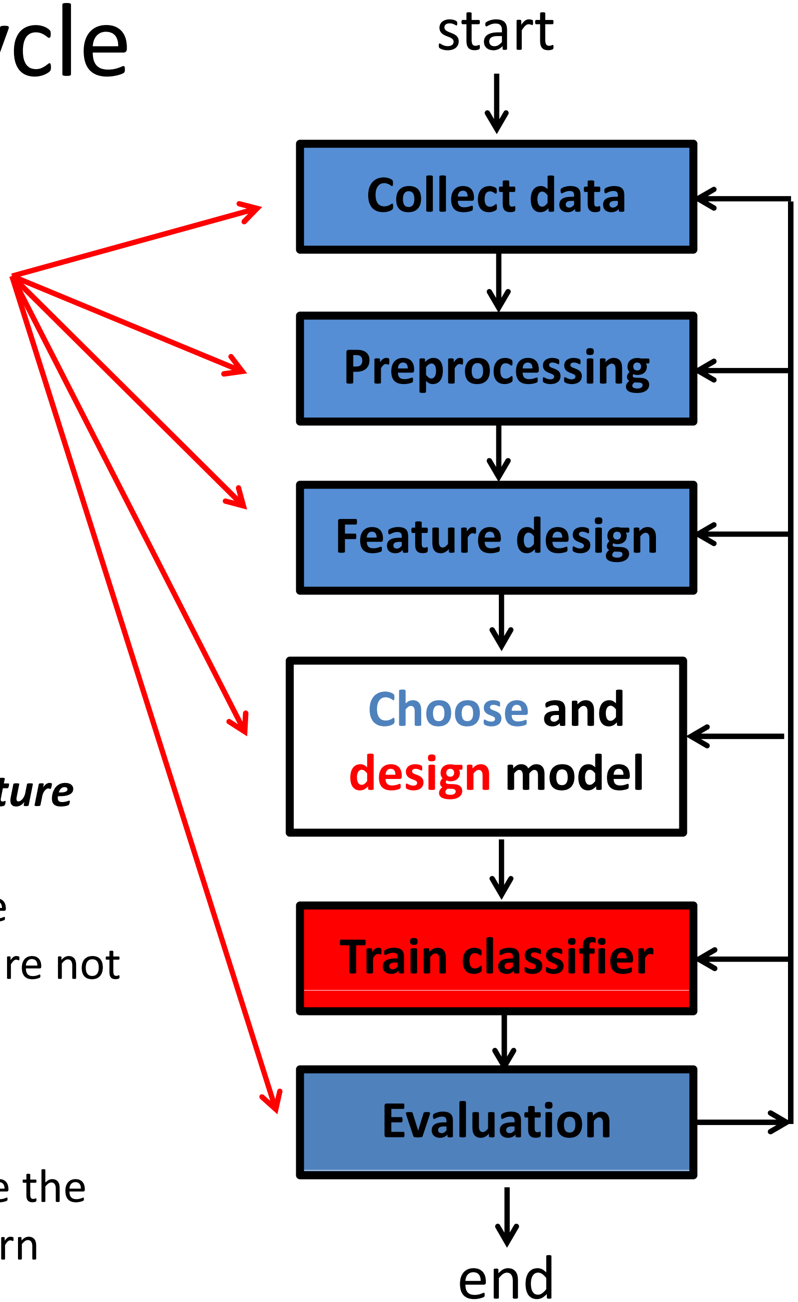
- How do machine learning, computer vision, and speech recognition people study pattern recognition problems in different ways?
- Why did some computer vision people have concern about their research life when deep learning emerged?
- How does deep learning change the design cycle of a pattern recognition system?

Pattern Recognition System



Design Cycle

Domain knowledge



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

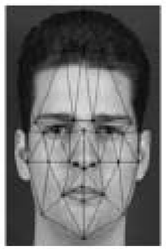
(a)

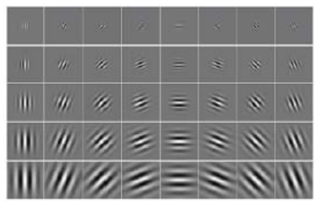
(b)

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

Face recognition pipeline

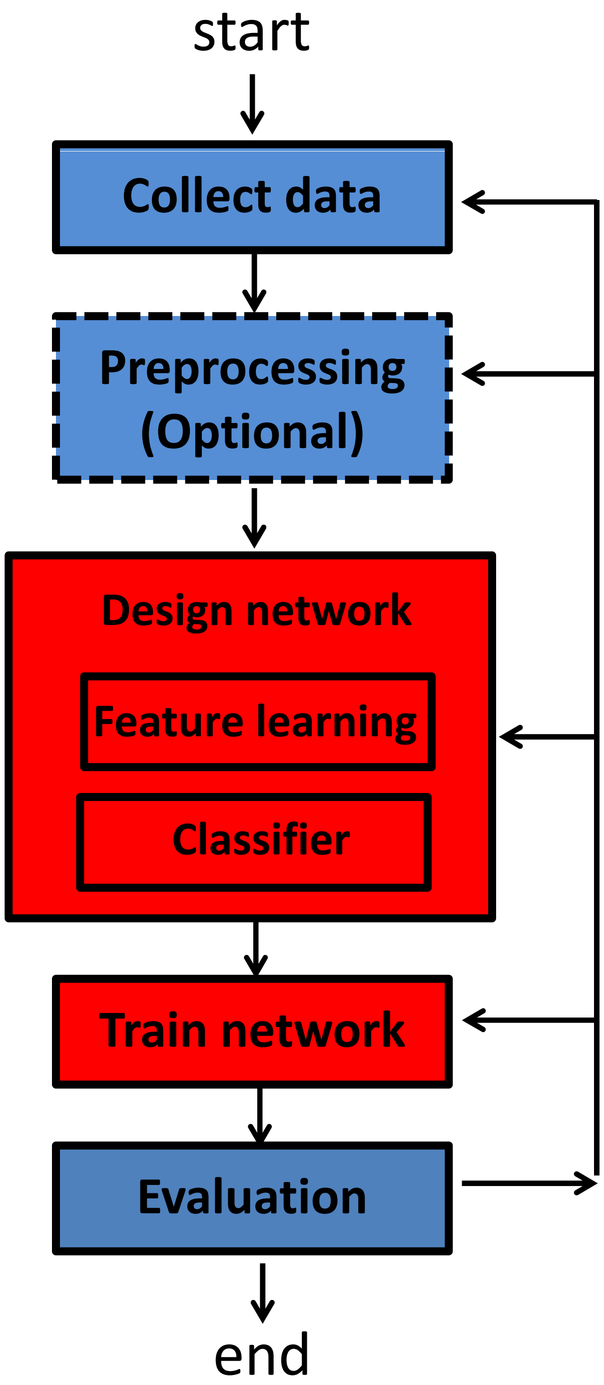
Face alignment → Geometric rectification → Photometric rectification → Feature extraction → Classification





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

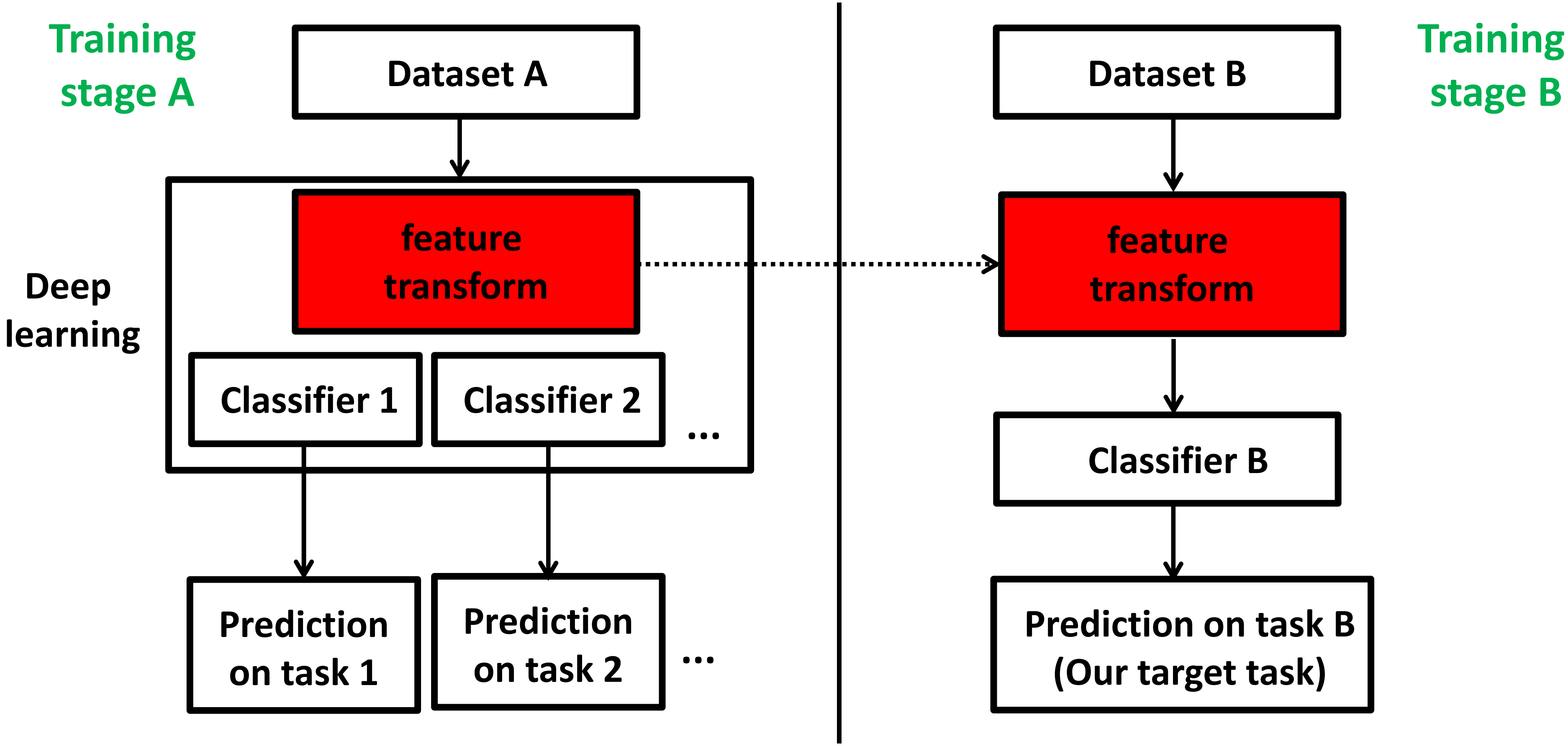
Design network structure

New training strategies

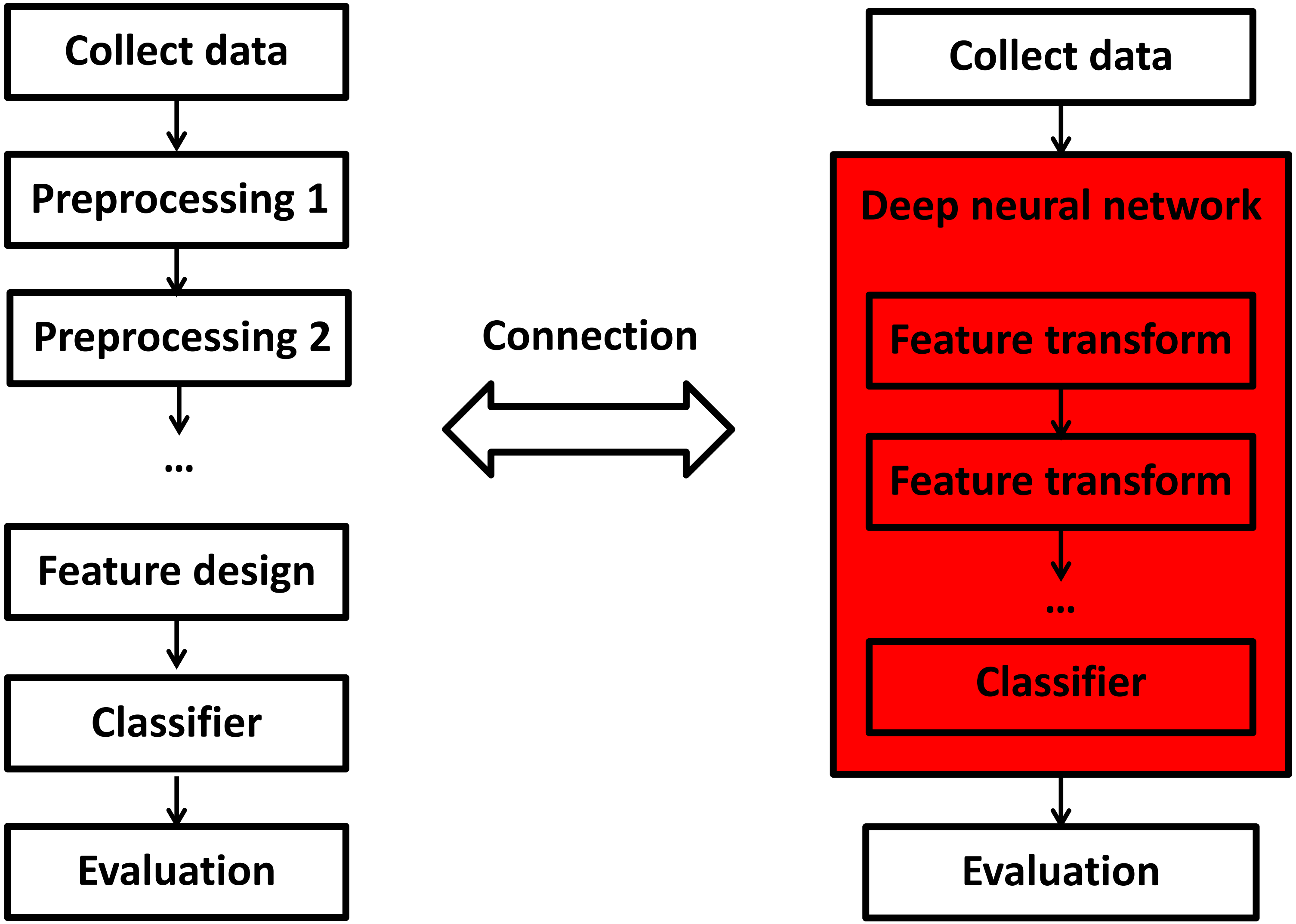
Feature learned from ImageNet can be well generalized to other tasks and datasets!

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

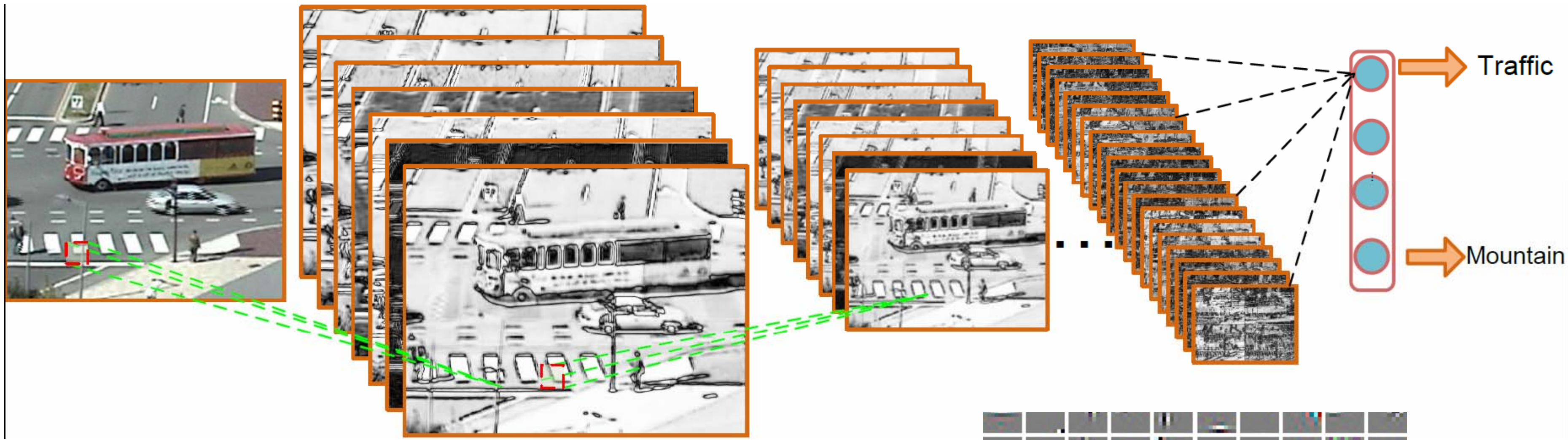
- Historical review of deep learning
- **Introduction to classical deep models**
- Why does deep learning work?

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.

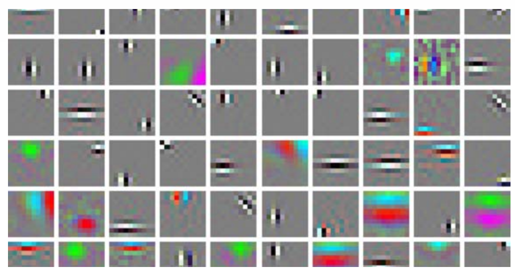
Classical Deep Models

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



Convolution

Pooling

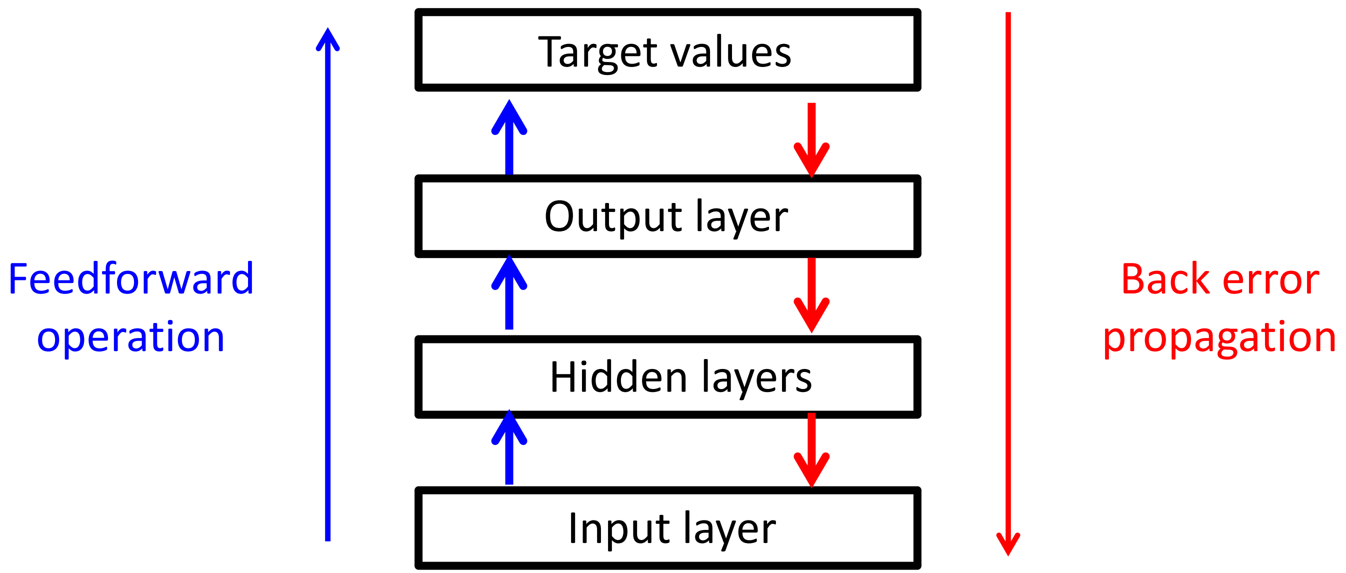


Learned filters

Backpropagation

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

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



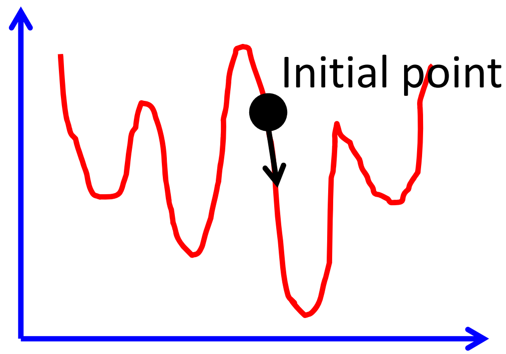
D. E. Rumelhart, G. E. Hinton, R. J. Williams, “Learning Representations by Back-propagation Errors,” Nature, Vol. 323, pp. 533-536, 1986.

Classical Deep Models

- Deep belief net
  - Hinton'06

Pre-training:

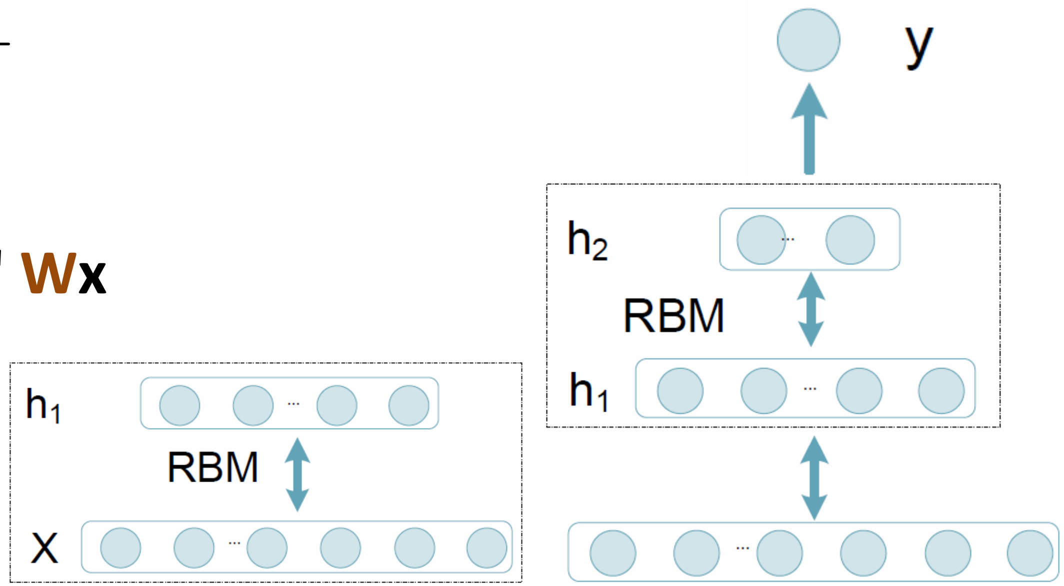
- Good initialization point
- Make use of unlabeled data



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

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

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



Classical Deep Models

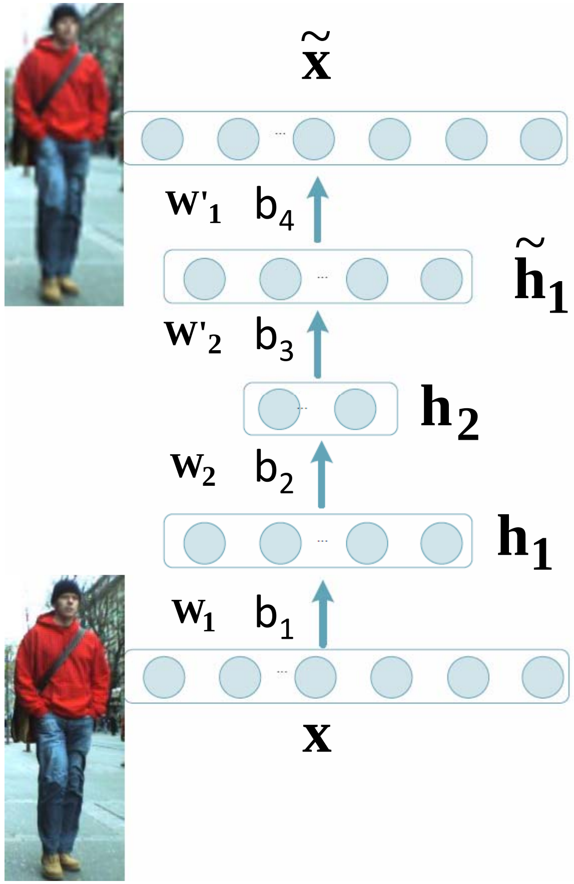
- Auto-encoder
  - Hinton and Salakhutdinov 2006

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

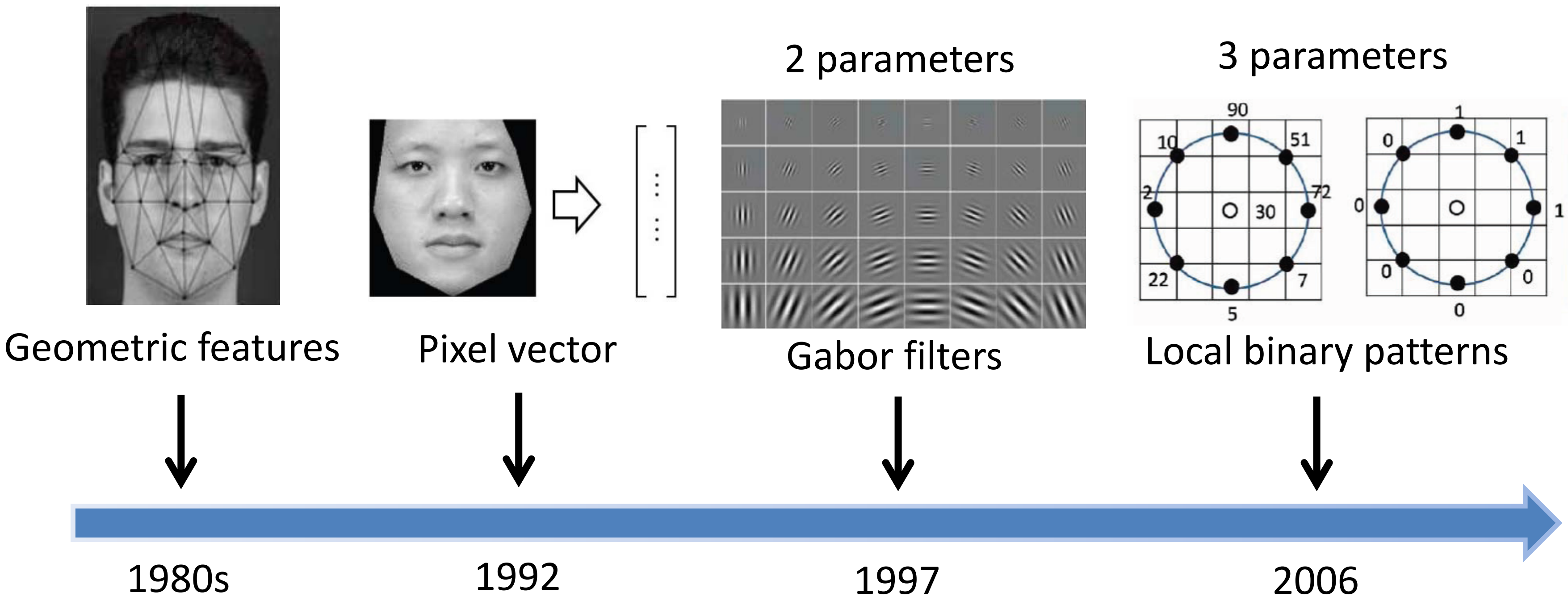
- Historical review of deep learning
- Introduction to classical deep models
- **Why does deep learning work?**

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



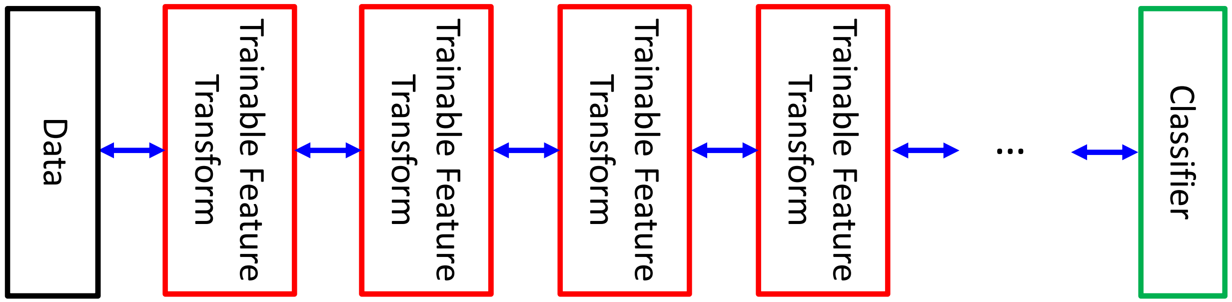
Feature Learning

- 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, **which dramatically increase the capacity of deep models**
  - Make better use of big data
  - Faster to get feature representations for new applications

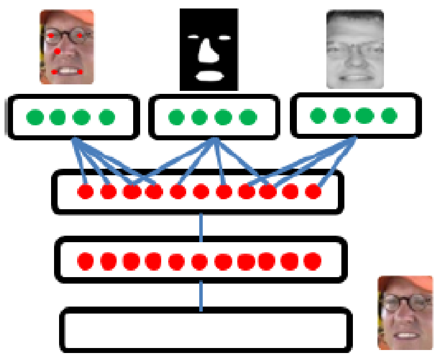
Deep Learning Means Feature Learning

- Deep learning is about learning hierarchical feature representations

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



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

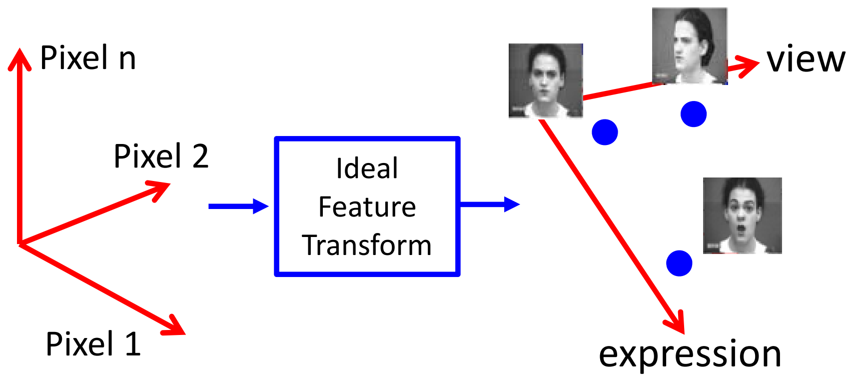


Identity: face recognition

Pose: pose estimation

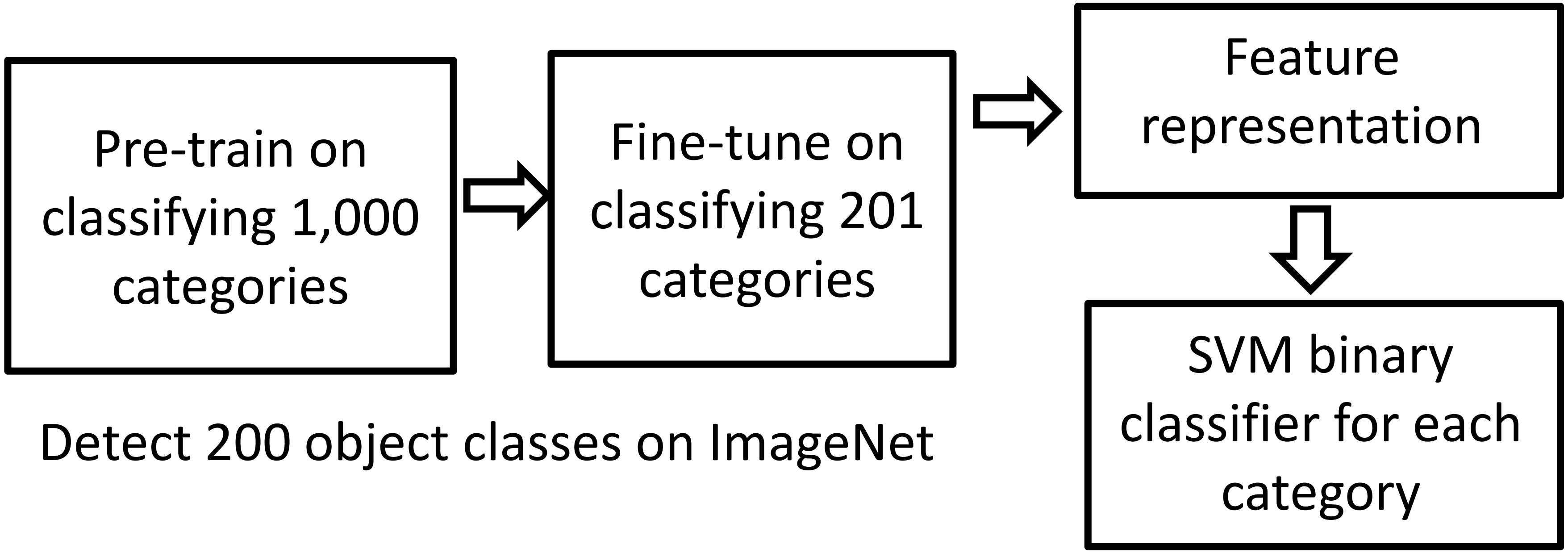
Expression: expression recognition

Age: age estimation



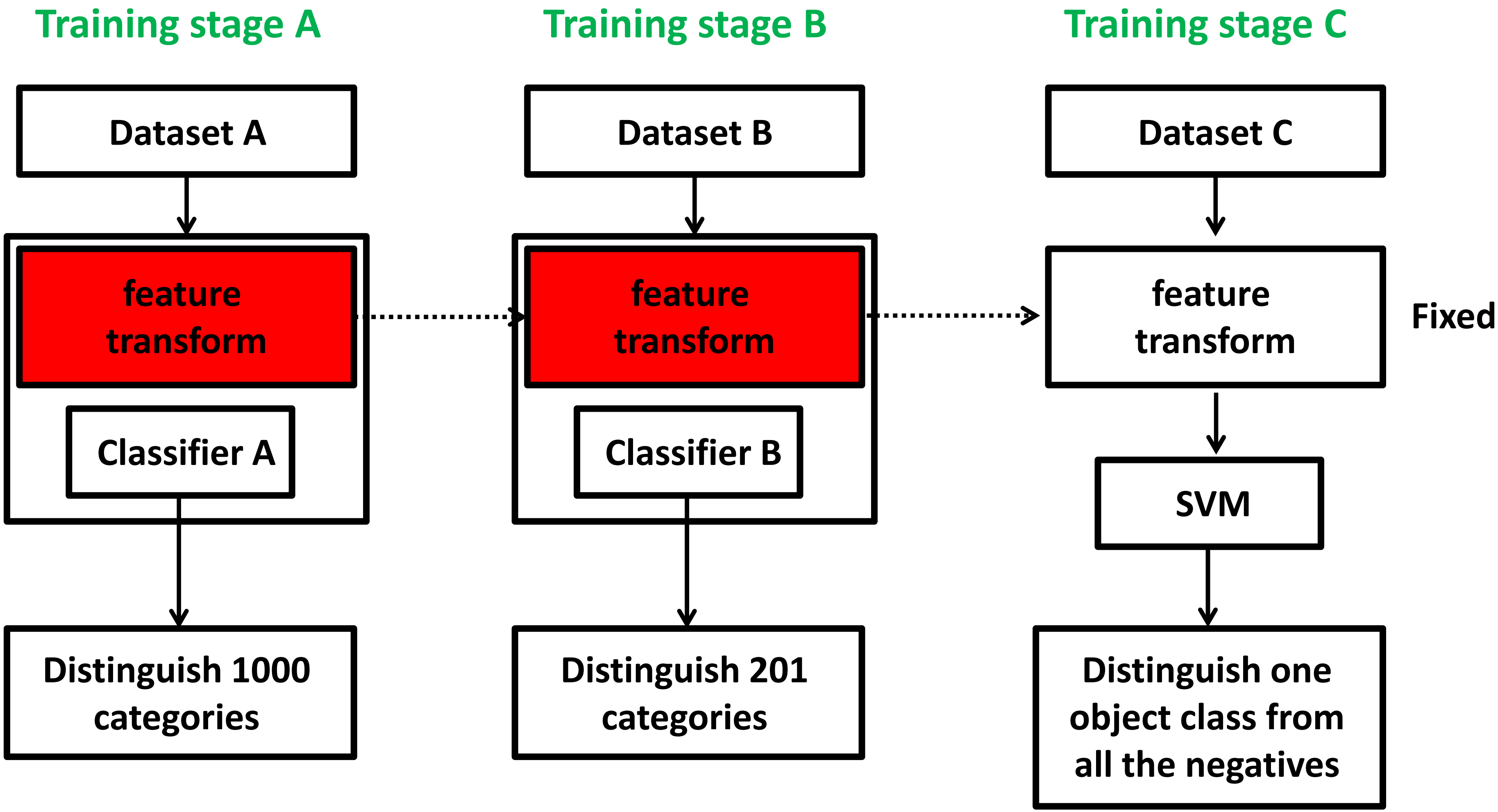
Deep Learning Means Feature Learning

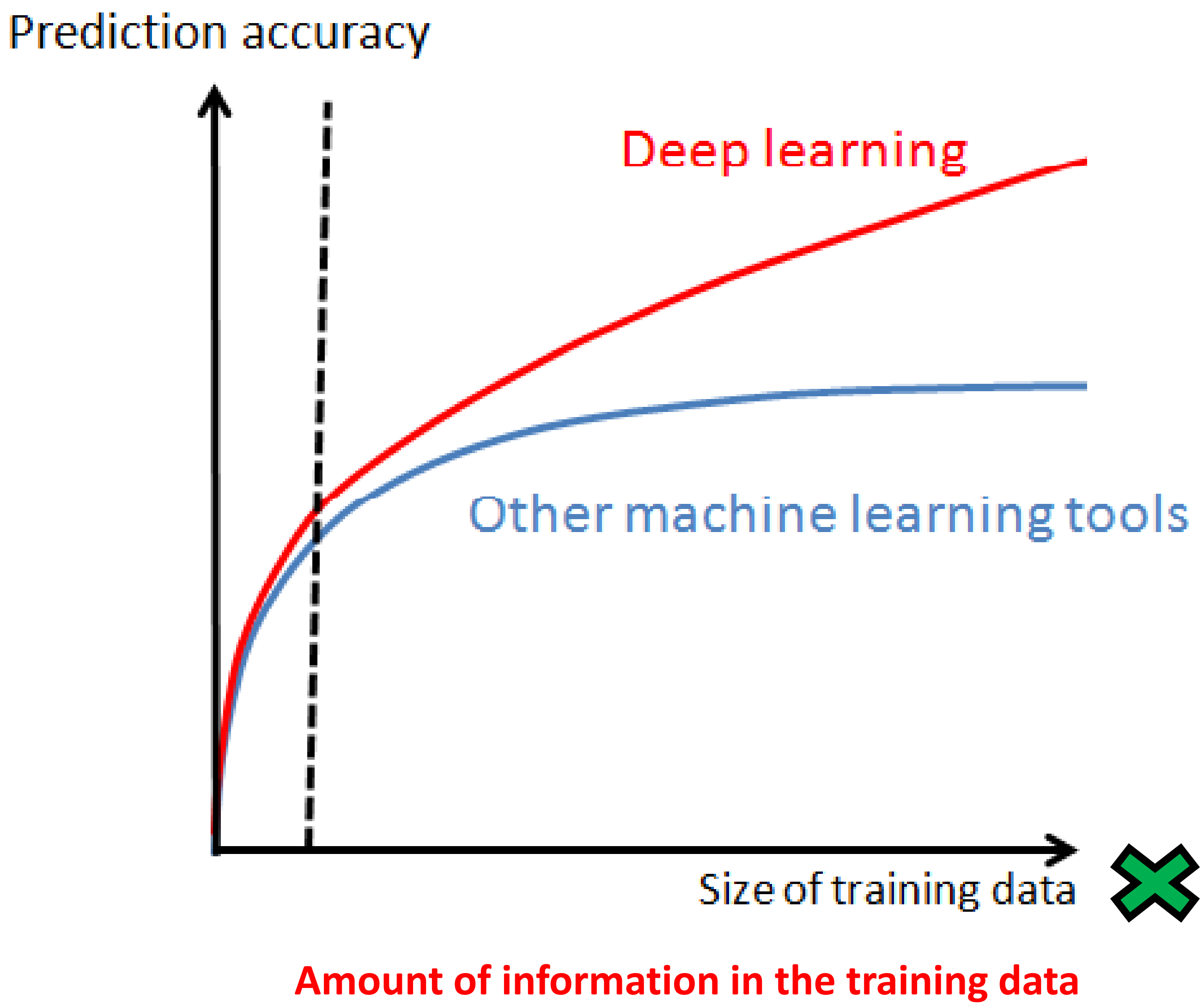
- How to effectively learn features with deep models
  - With challenging tasks
  - Predict high-dimensional vectors



Detect 200 object classes on ImageNet

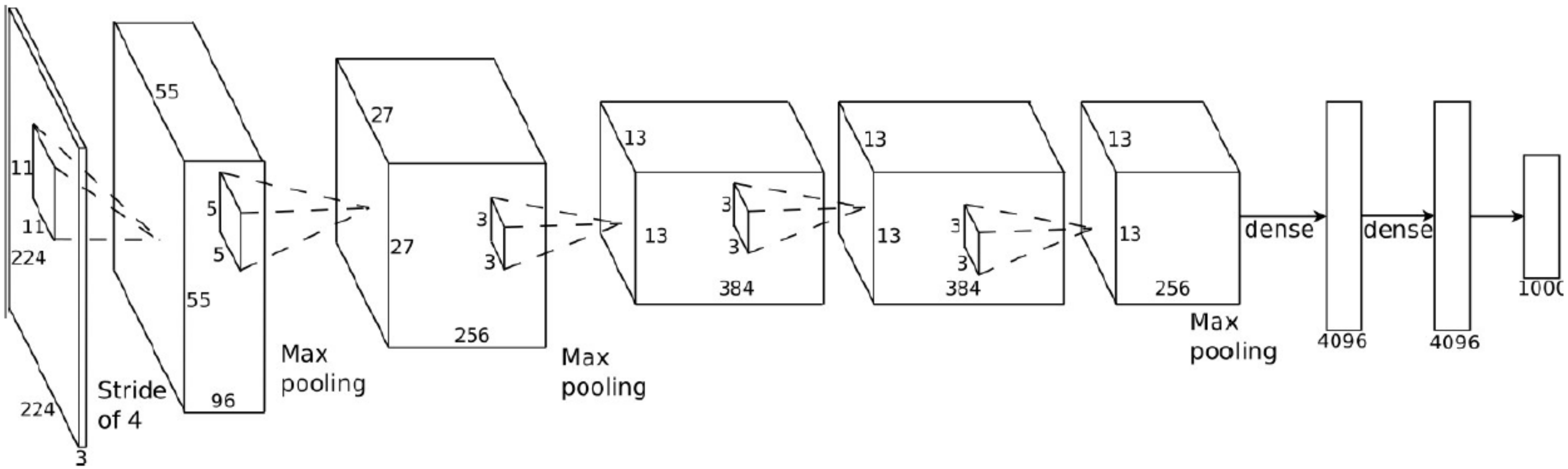
W. Ouyang et al. "DeepID-Net: multi-stage and deformable deep convolutional neural networks for object detection", arXiv:1409.3505, 2014


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


Prediction accuracy
Deep learning
Other machine learning tools
Size of training data
Amount of information in the training data

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

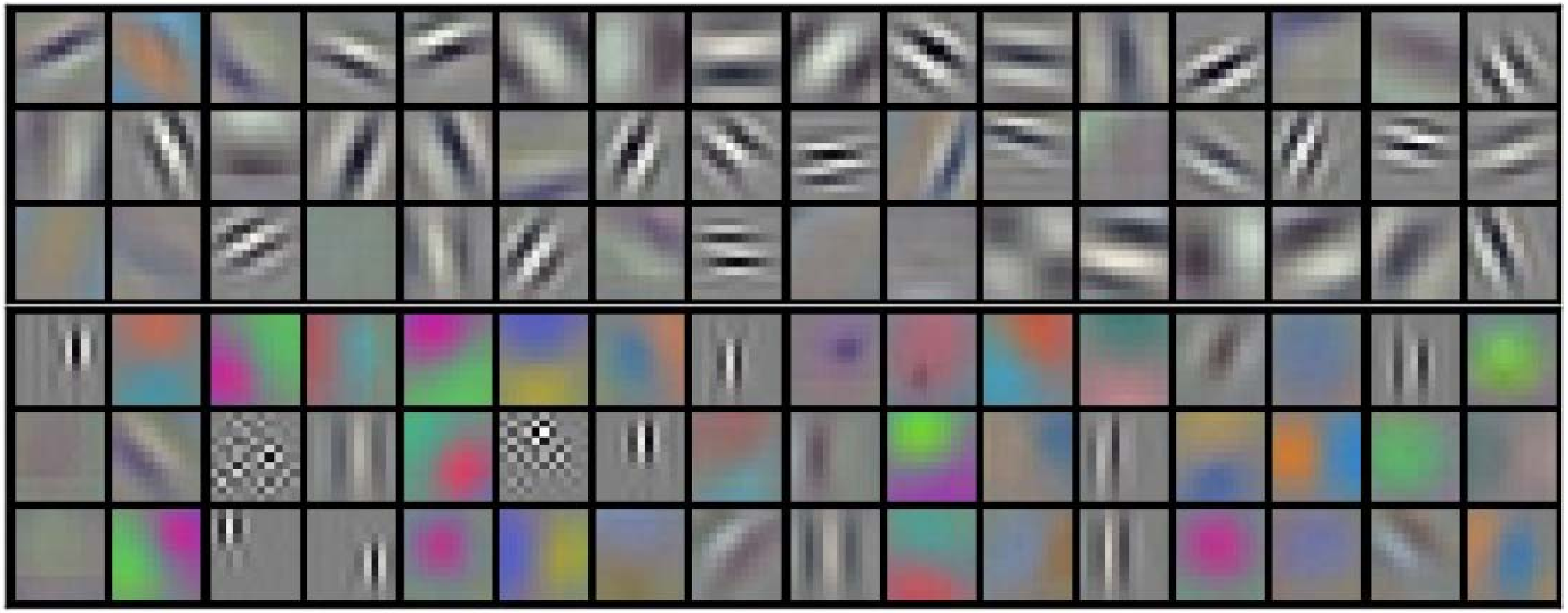
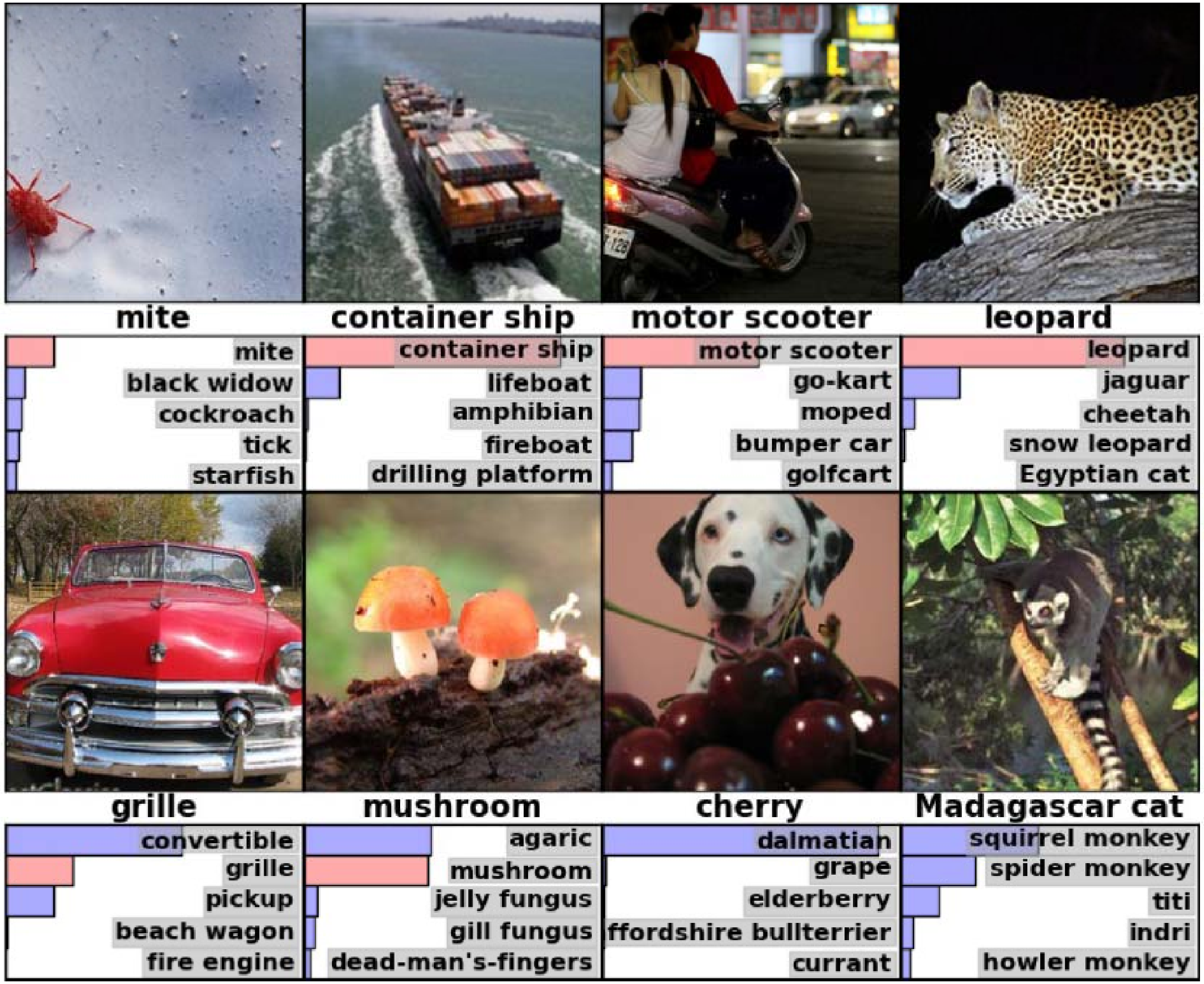
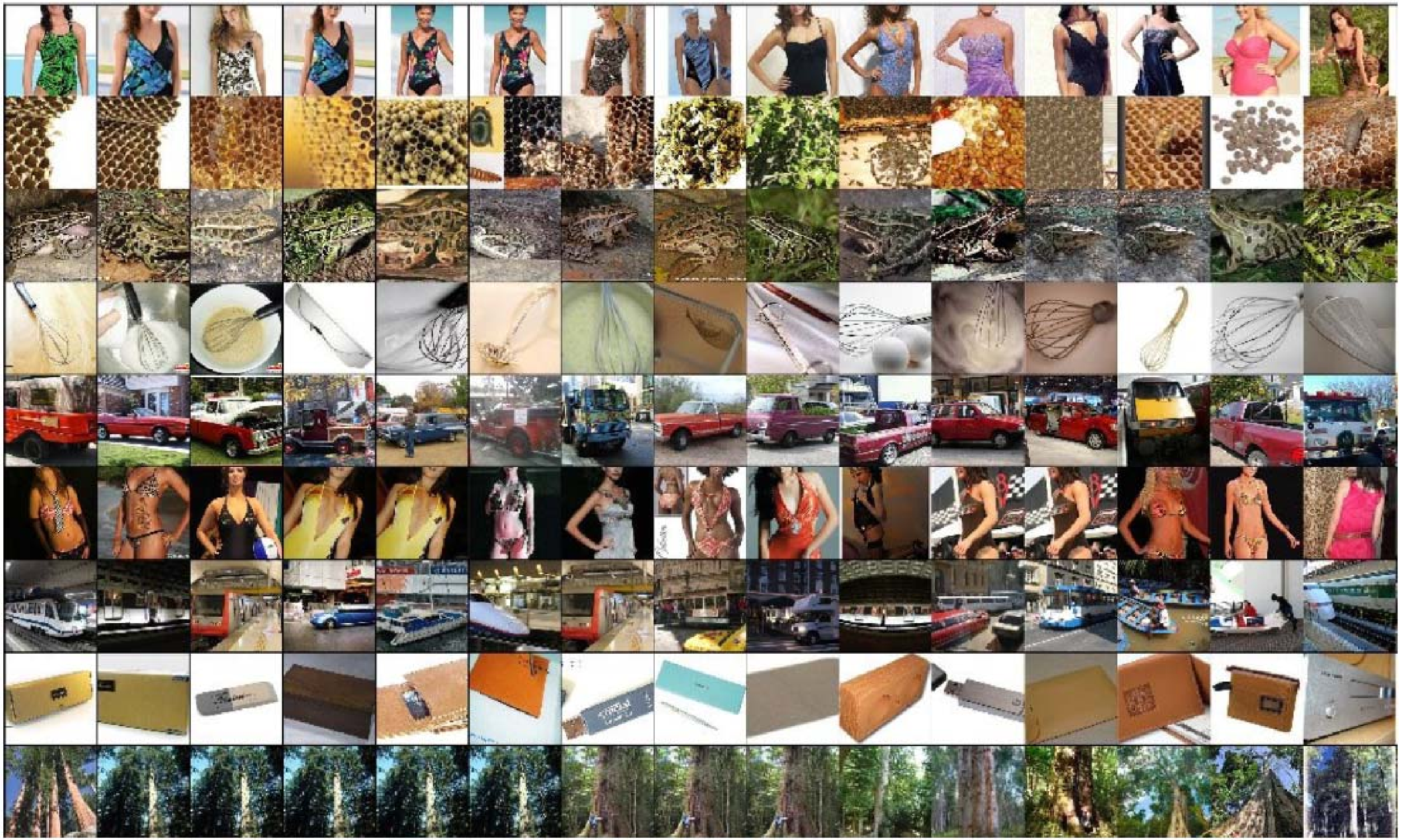


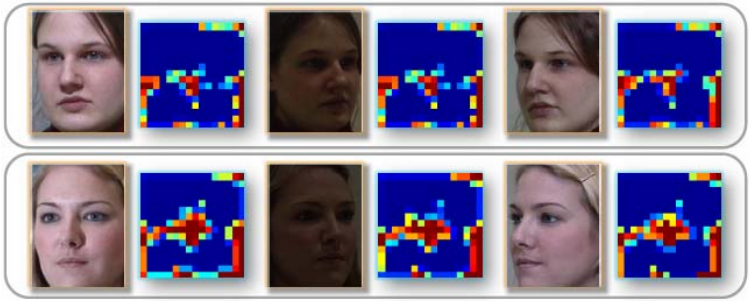
Image classification result

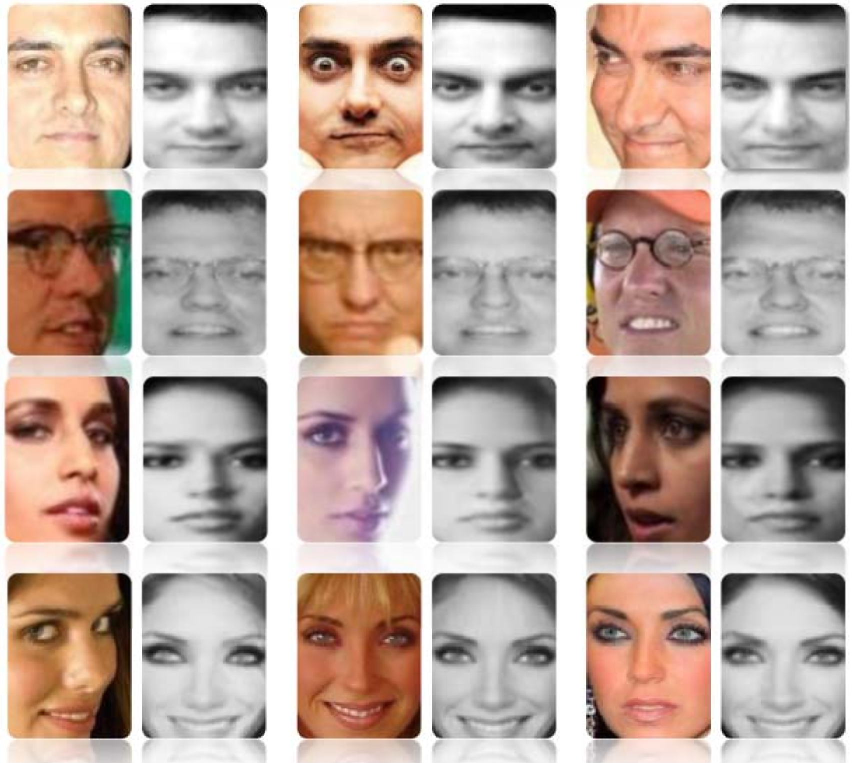


Top hidden layer can be used as feature for retrieval



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

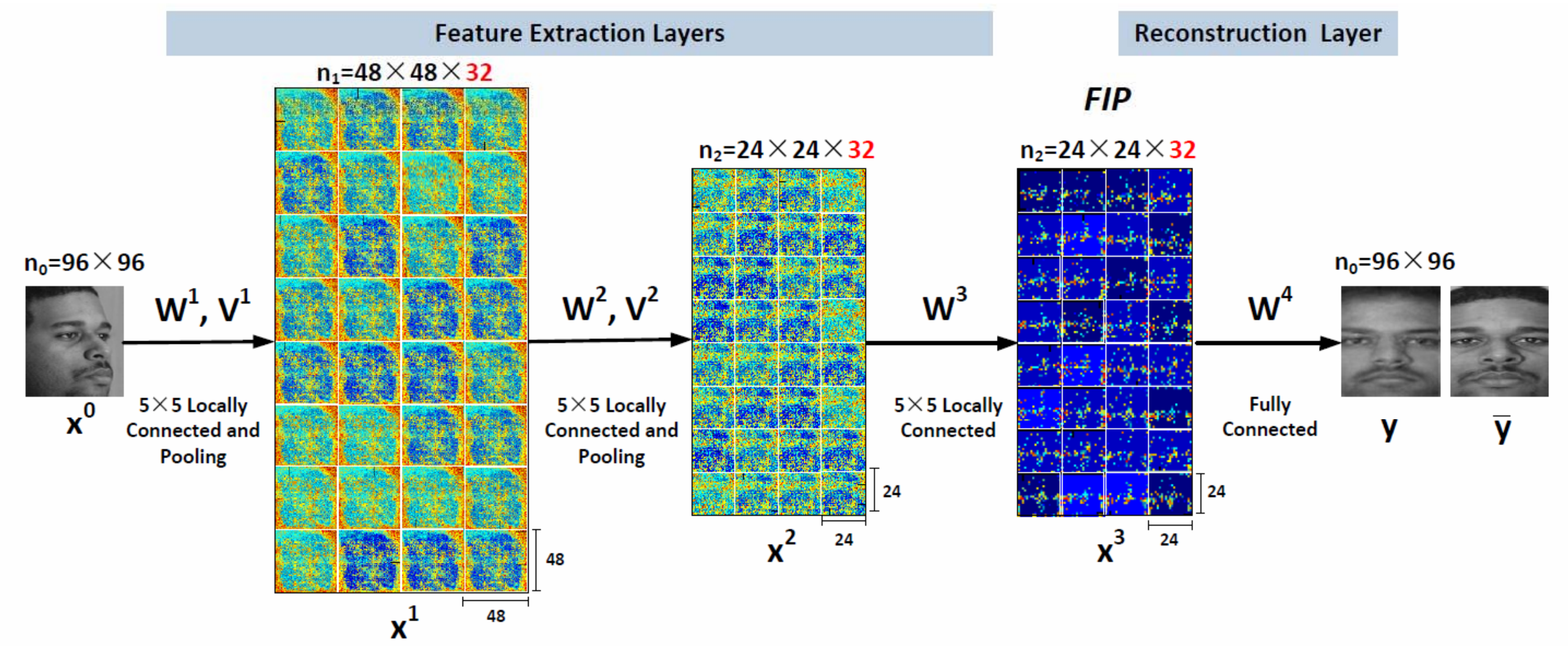




Reconstruction examples from LFW

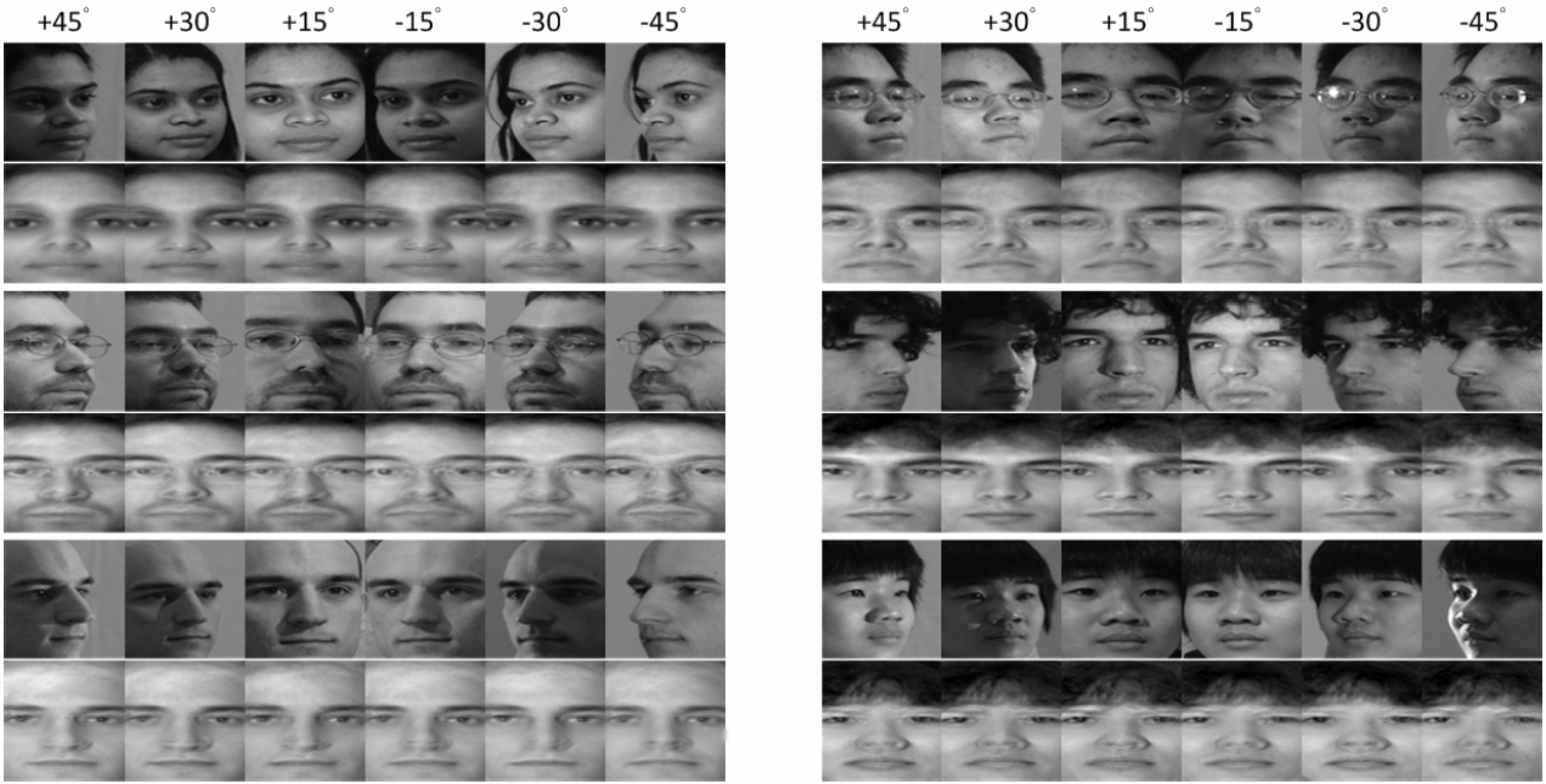
Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Identity Preserving Face Space,” ICCV 2013.

- 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

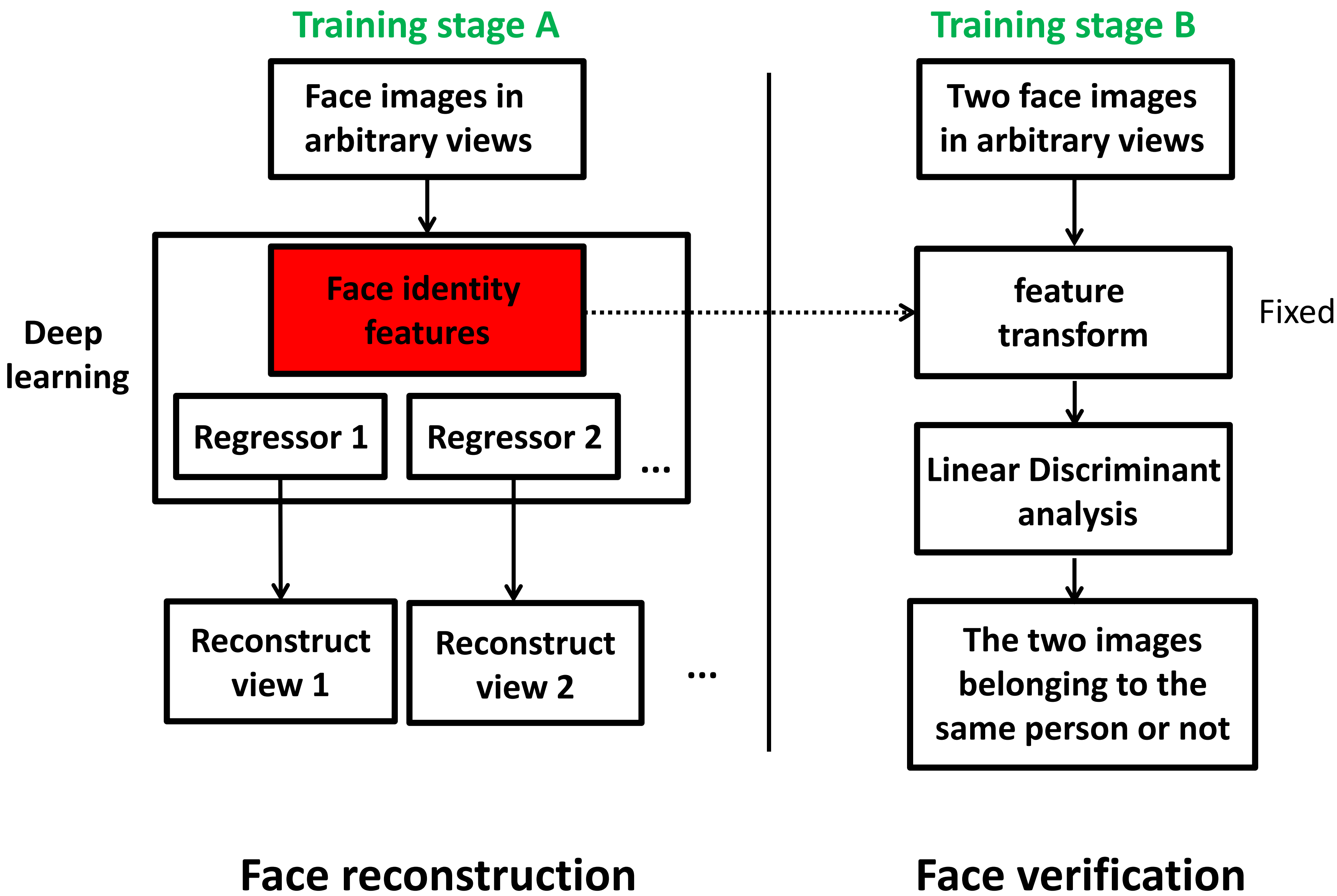
[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012. 3, 6

[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

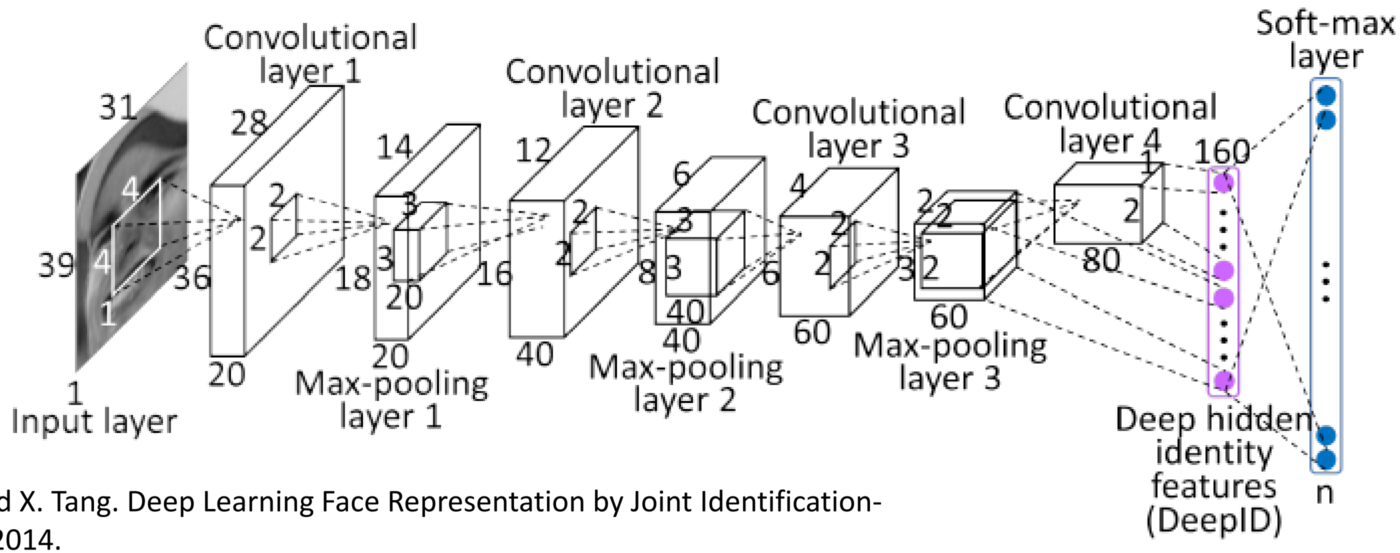
Deep learning 3D model from 2D images, mimicking human brain activities

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

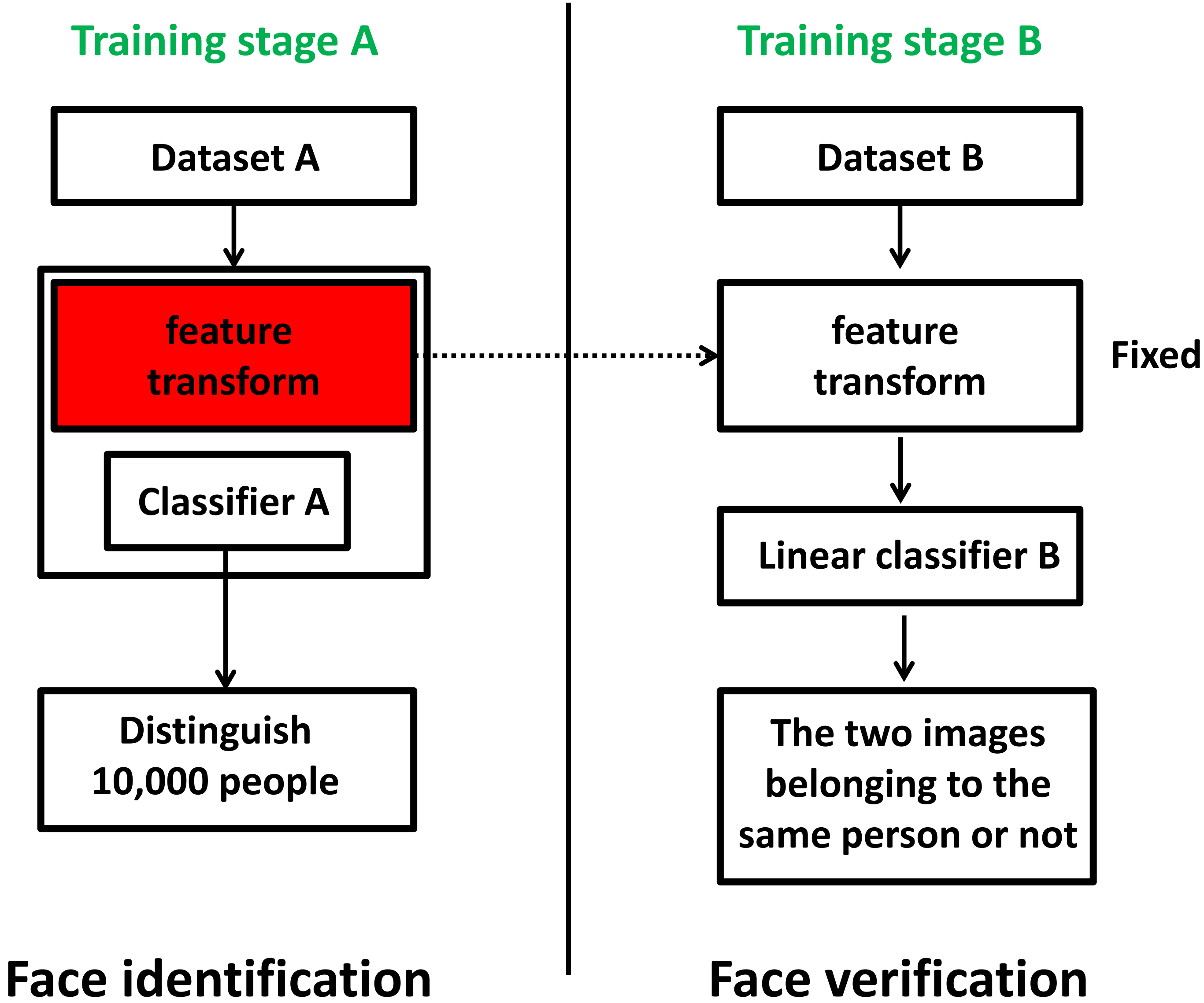

Training stage A
Face images in arbitrary views
Deep learning
Face identity features
Regressor 1
Regressor 2
...
Reconstruct view 1
Reconstruct view 2
...
Face reconstruction
Training stage B
Two face images in arbitrary views
feature transform
Fixed
Linear Discriminant analysis
The two images belonging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

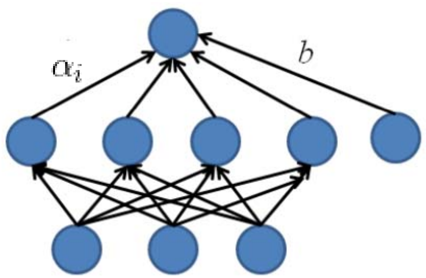

Training stage A
Dataset A
feature transform
Classifier A
Distinguish 10,000 people
Face identification
Training stage B
Dataset B
feature transform
Fixed
Linear classifier B
The two images belonging to the same person or not
Face verification

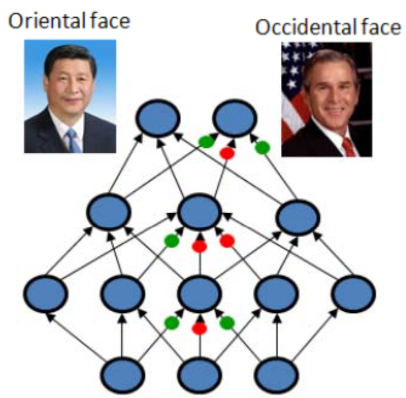
Deep Structures vs Shallow Structures
(Why deep?)

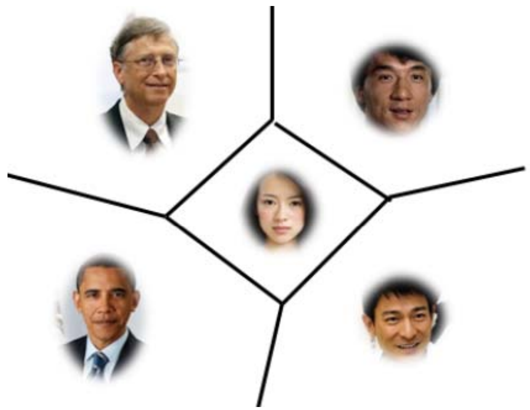
Shallow Structures

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







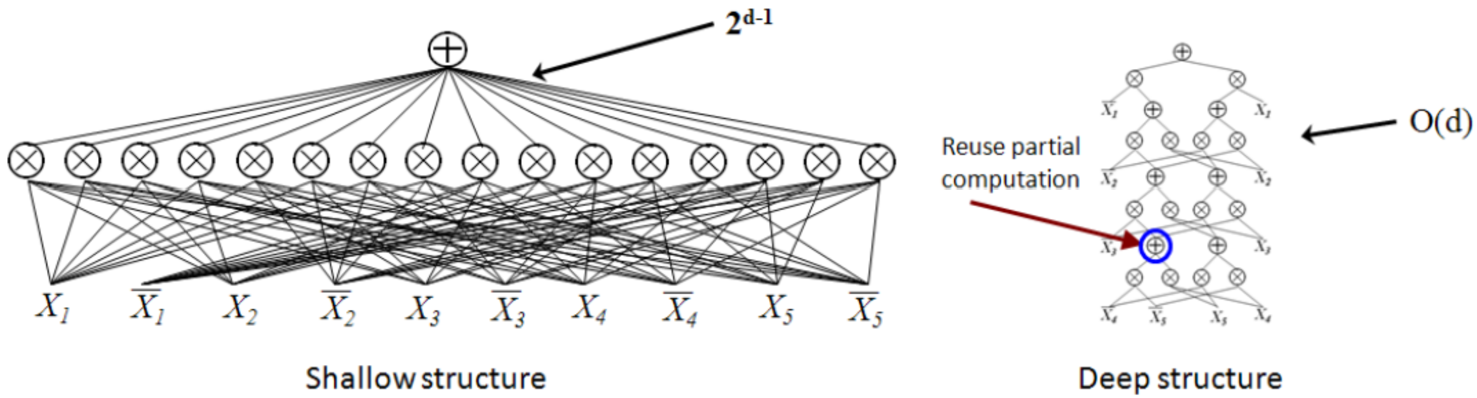
Deep Machines are More Efficient for Representing Certain Classes of Functions

- 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)

- Take the d-bit parity function as an example

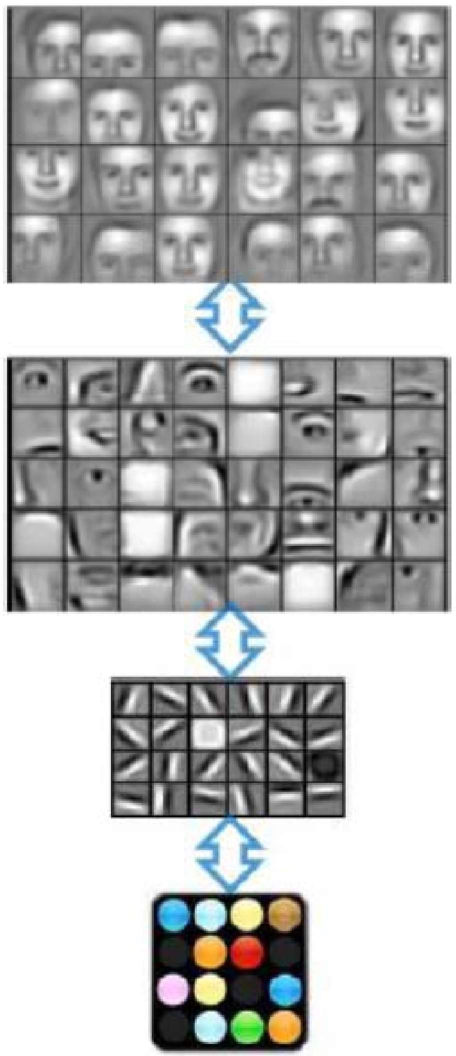
$$(X_{1,}\ldots, X_d) \in \{0,1\}^d \longmapsto \begin{cases} 1, & \text{if } \sum_{i=1}^{d} X_i \text{ is even} \\ -1, & \text{otherwise} \end{cases}$$

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



- 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

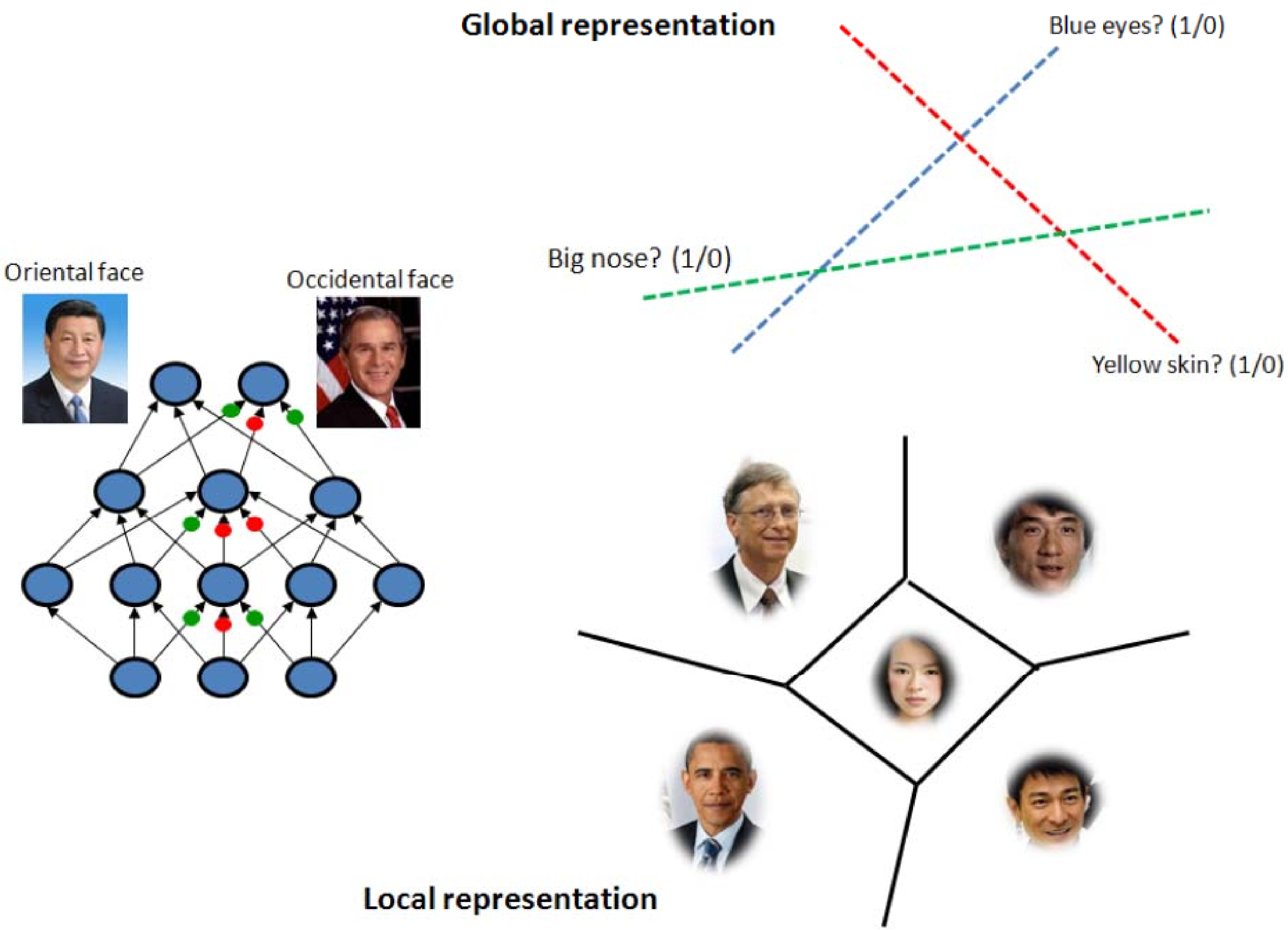
Humans Understand the World through Multiple Levels of Abstractions

- 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

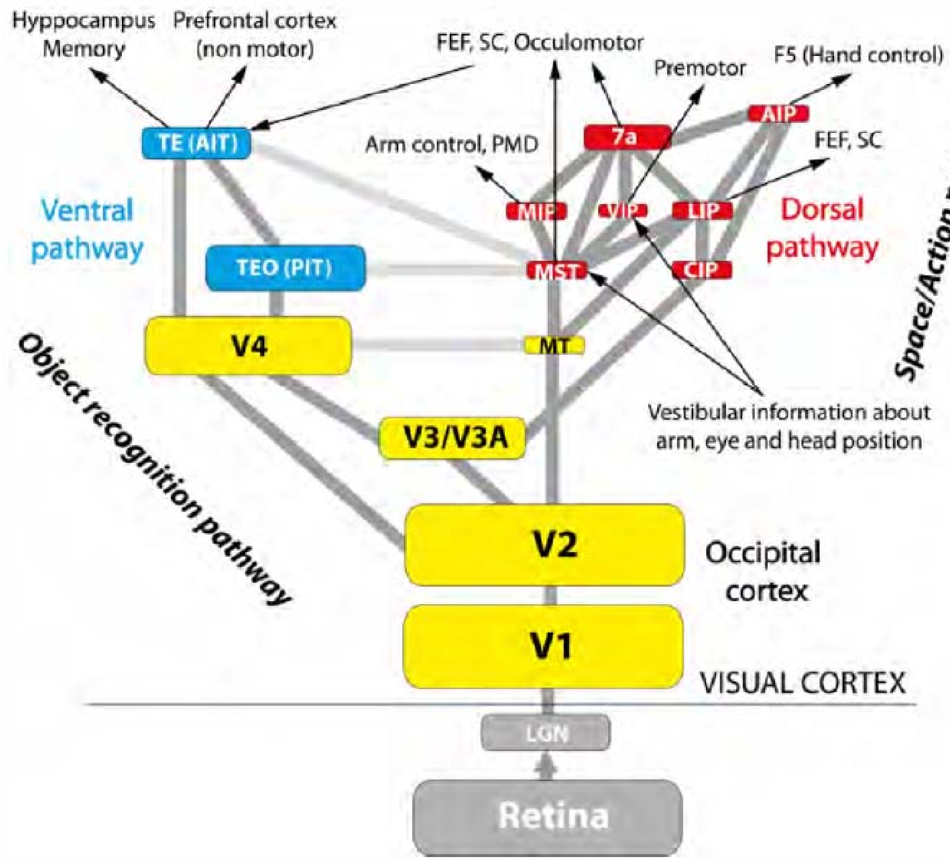
- Humans learn abstract concepts on top of less abstract ones
- Humans can imagine new pictures by re-configuring these abstractions at multiple levels. Thus our brain has good generalization can recognize things never seen before.
  - Our brain can estimate shape, lighting and pose from a face image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

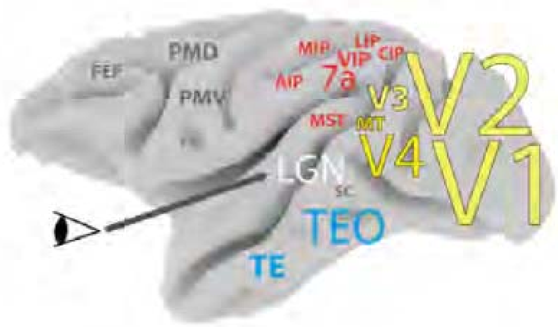
Local and Global Representations



Human Brains Process Visual Signals through Multiple Layers

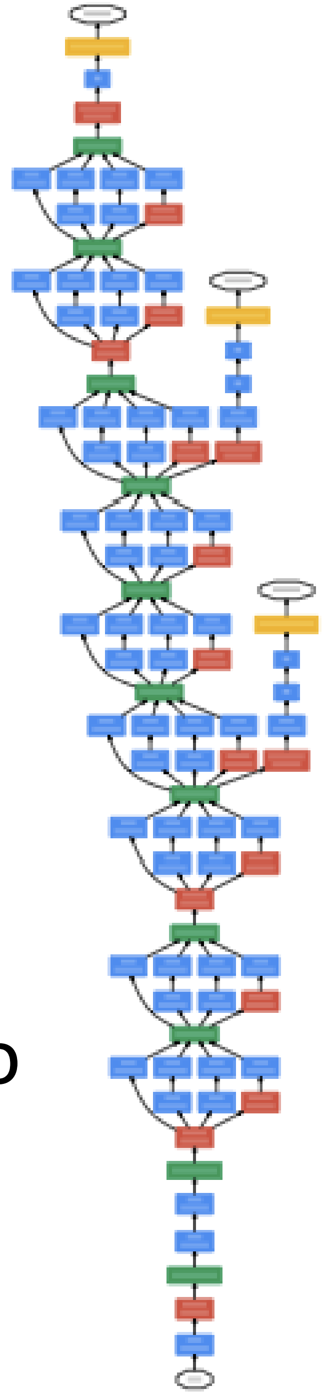
- A visual cortical area consists of six layers (Kruger et al. 2013)





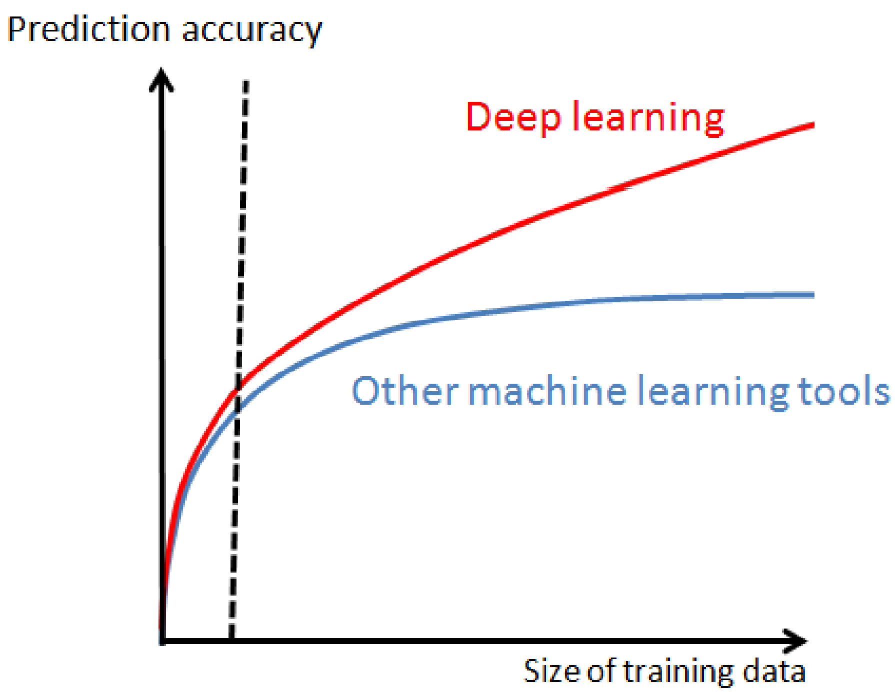
GoogLeNet
(Very Deep Neural Network)

- More than 20 layers
- The learning capacity is largely increased
- Add supervision at multiple layers
- The training error drops much more quickly
- The error rate is reduced from 15.3% to 6.6%



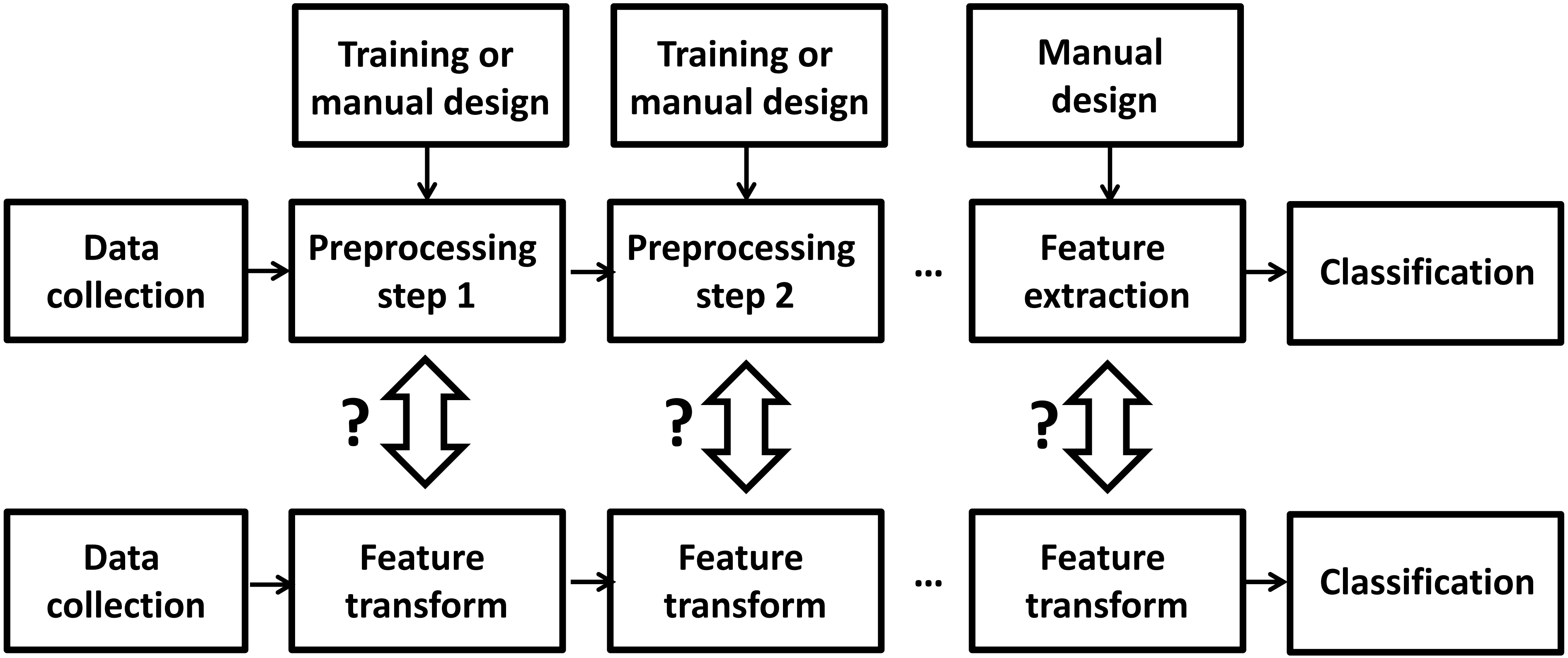
How do shallow models increase the model capacity?

- Typically increase the size of feature vectors



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.

Joint Learning vs Separate Learning

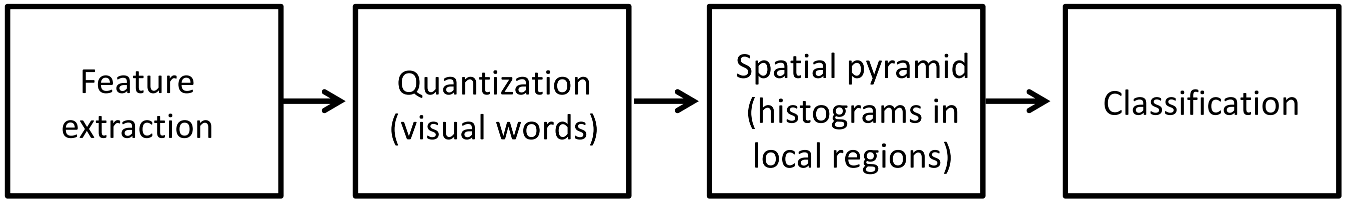


End-to-end learning

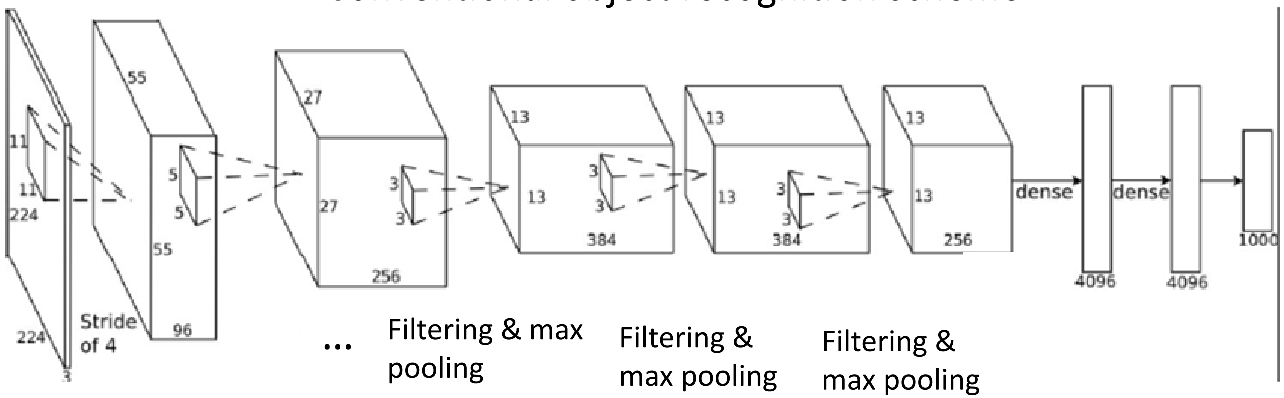
Deep learning is a framework/language but not a black-box model

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



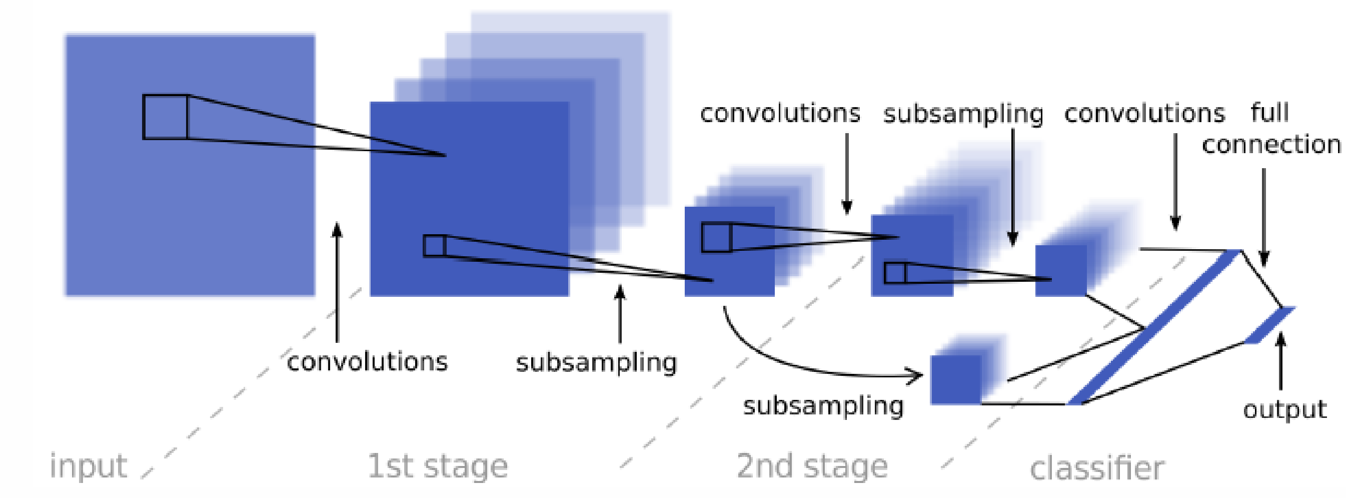
Feature extraction ↔ filtering

Quantization ↔ filtering

Spatial pyramid ↔ multi-level pooling

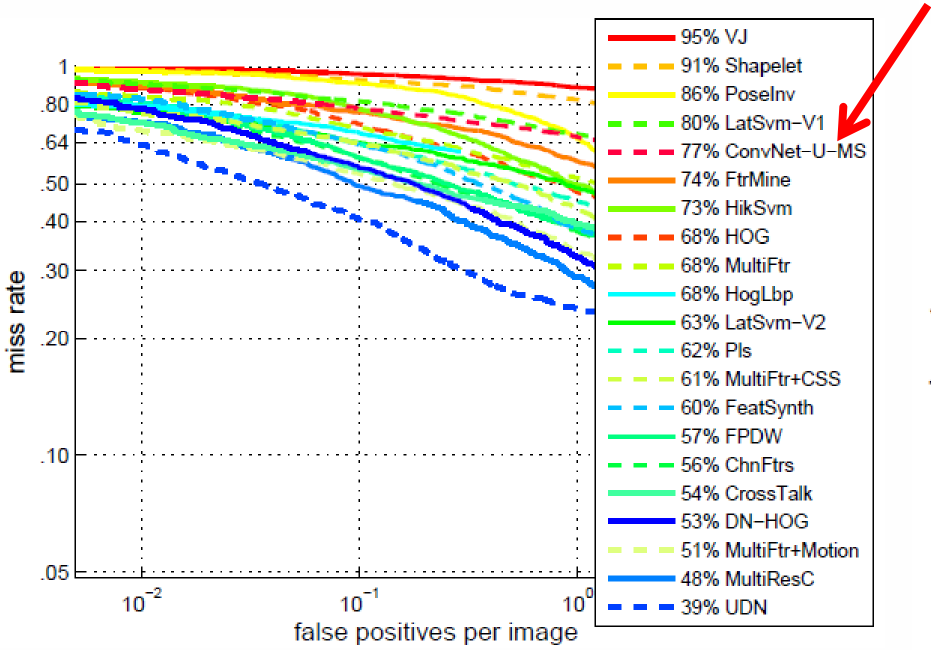
Krizhevsky NIPS'12

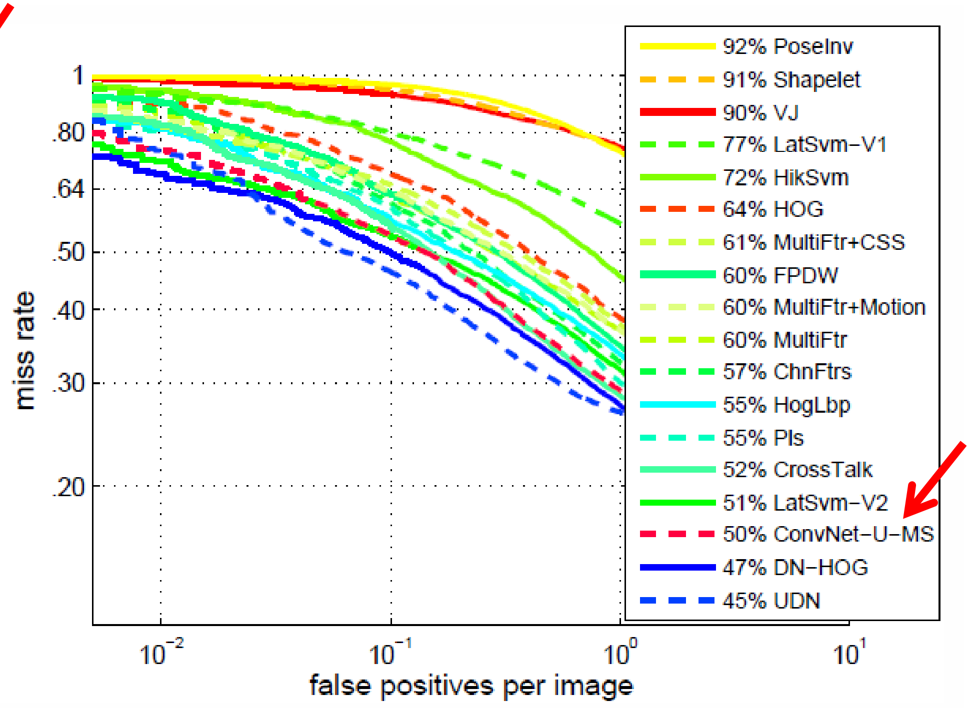
What if we treat an existing deep model as a black box in pedestrian detection?



ConvNet−U−MS

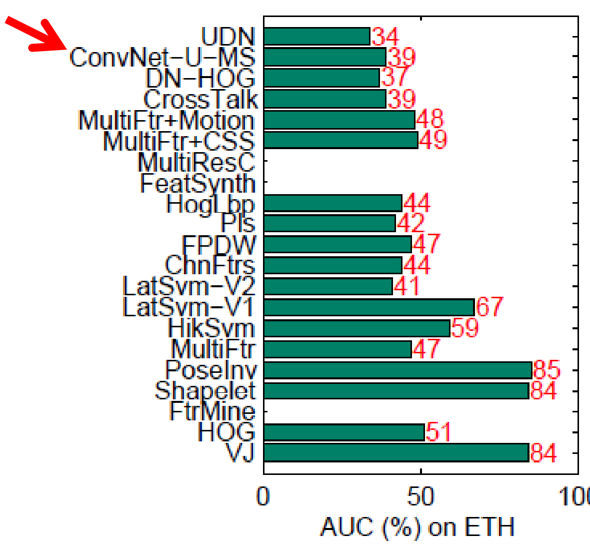
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” CVPR 2013.

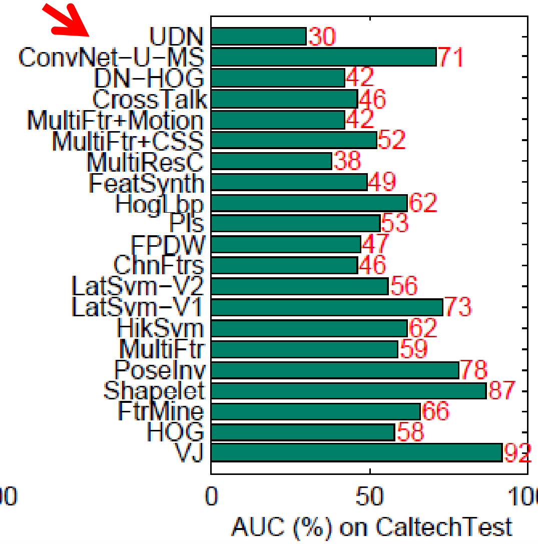


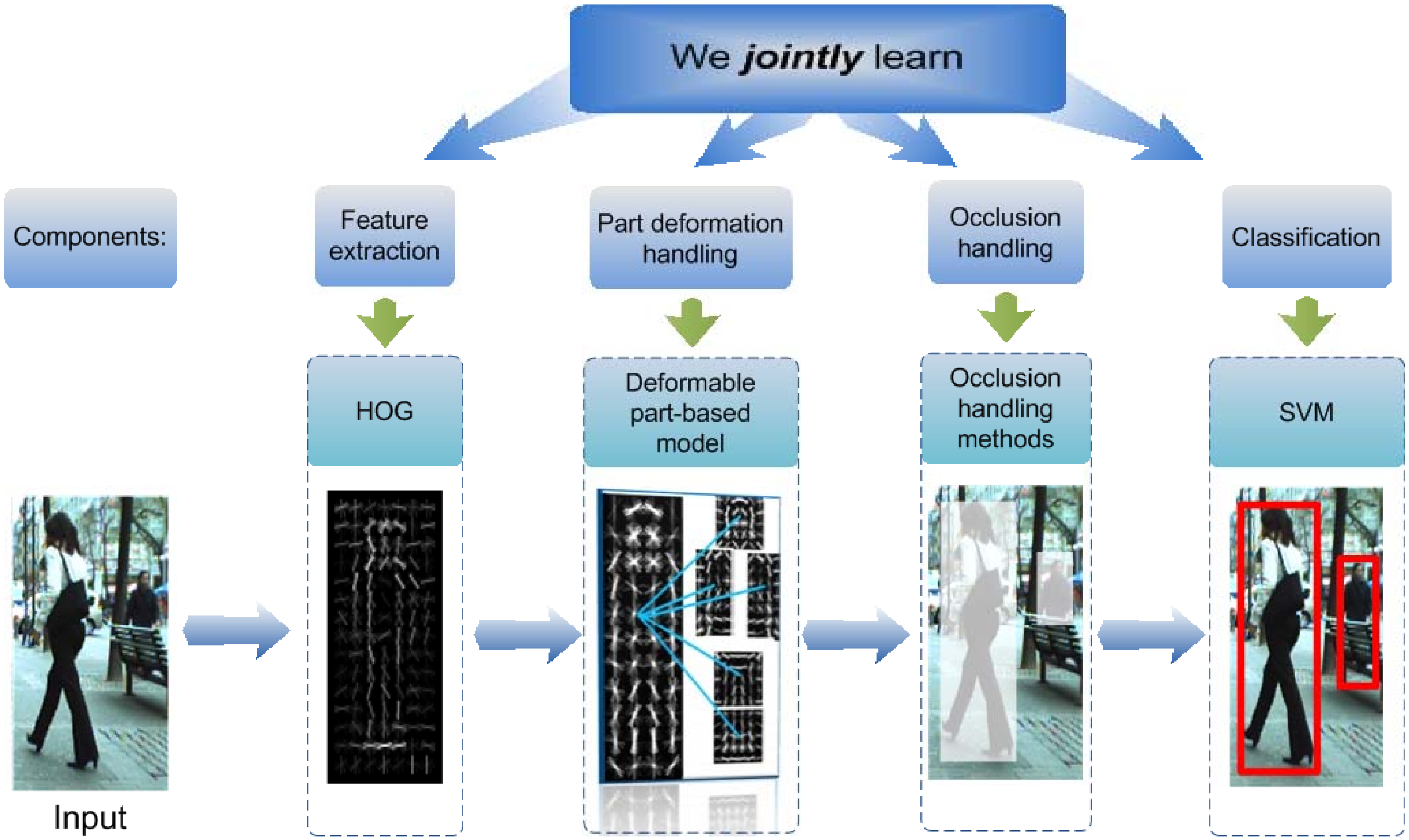


Results on Caltech Test

Results on ETHZ

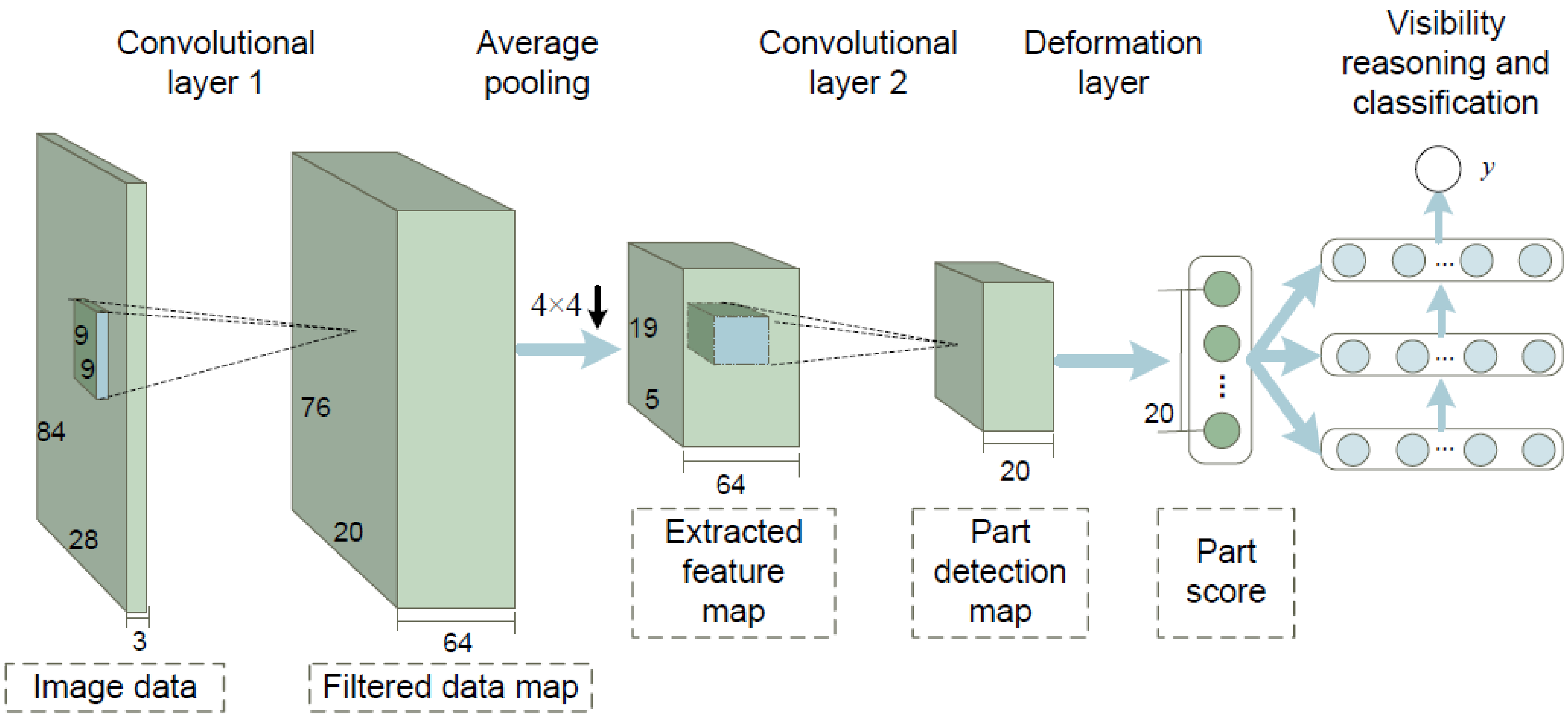






- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)
- P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)
- W. Ouyang and X. Wang. A Discriminative Deep Model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

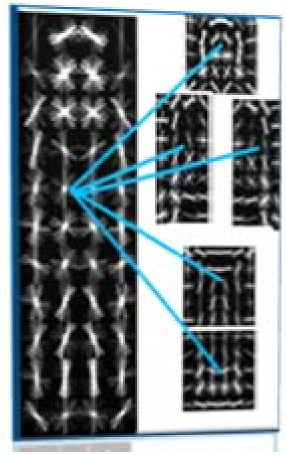
Our Joint Deep Learning Model



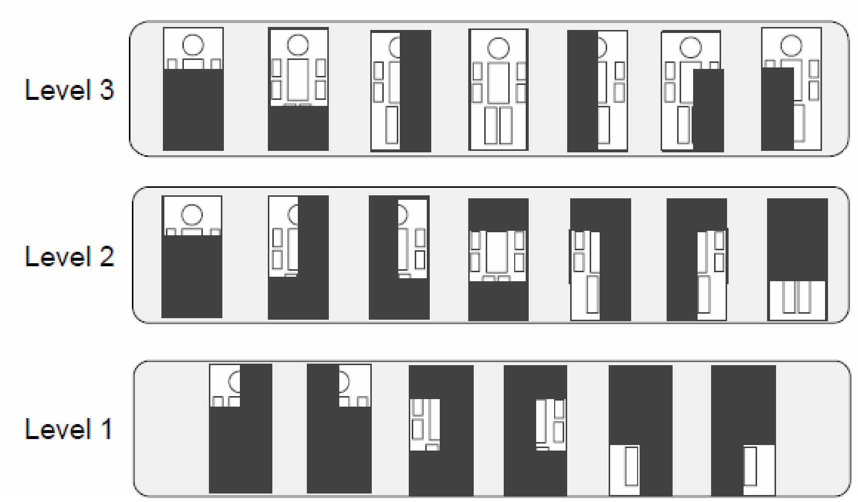
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling Part Detectors

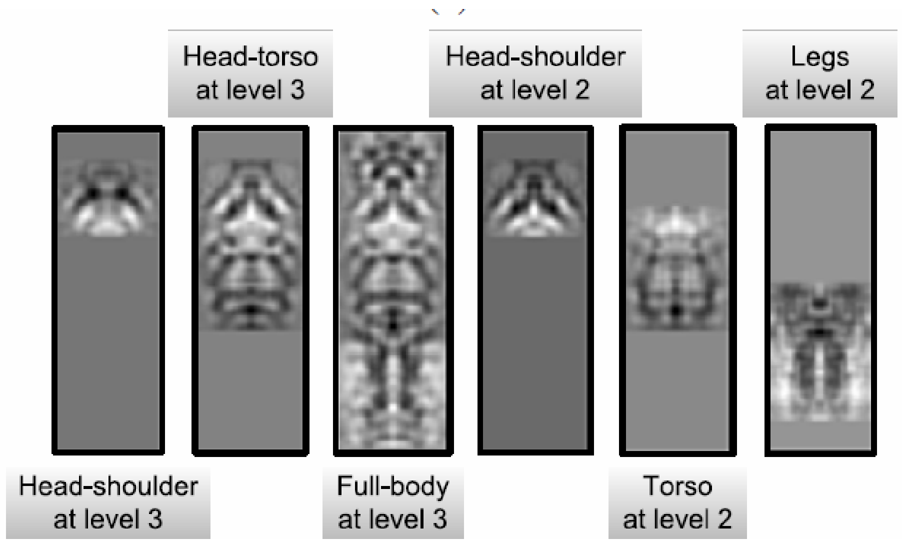
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

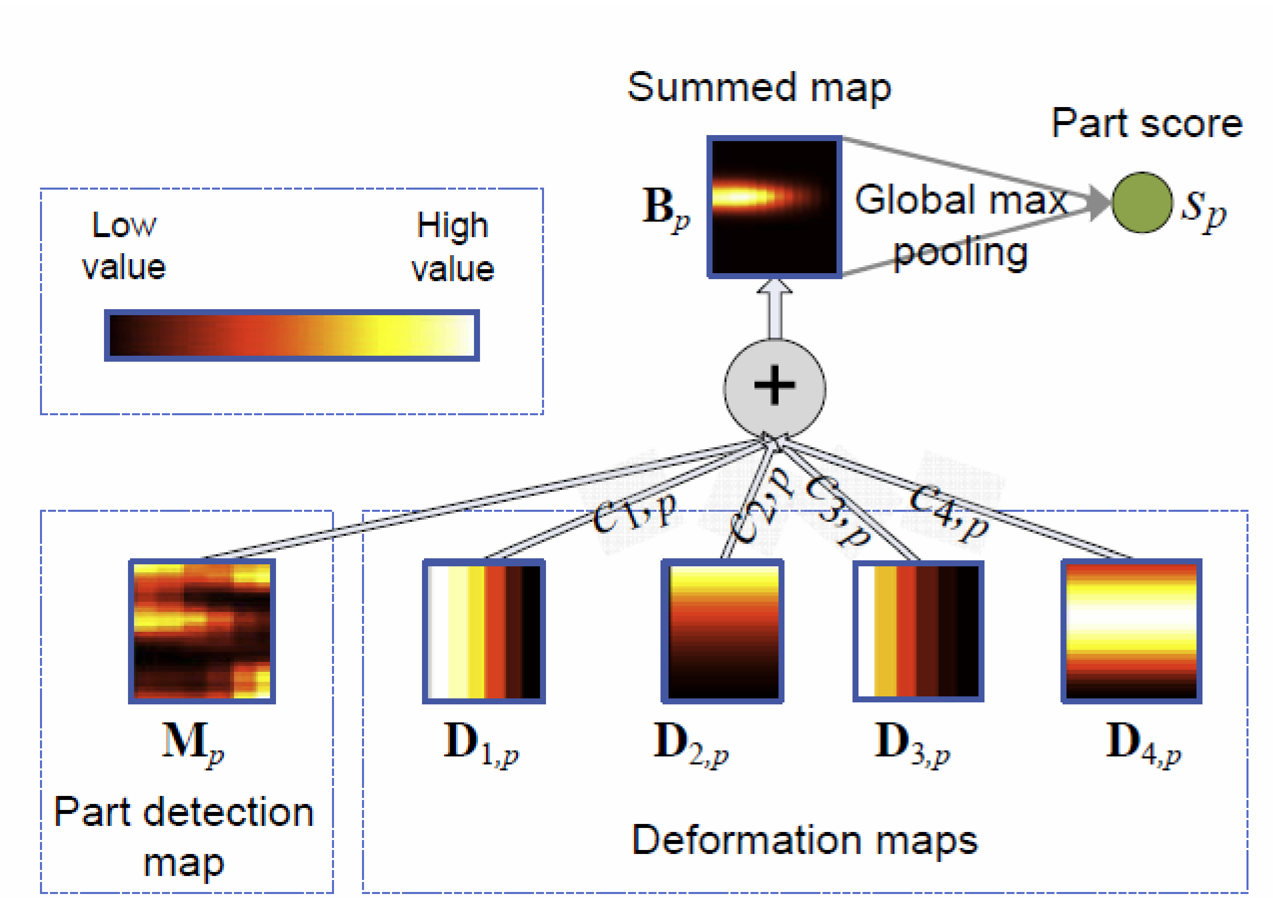


Part models

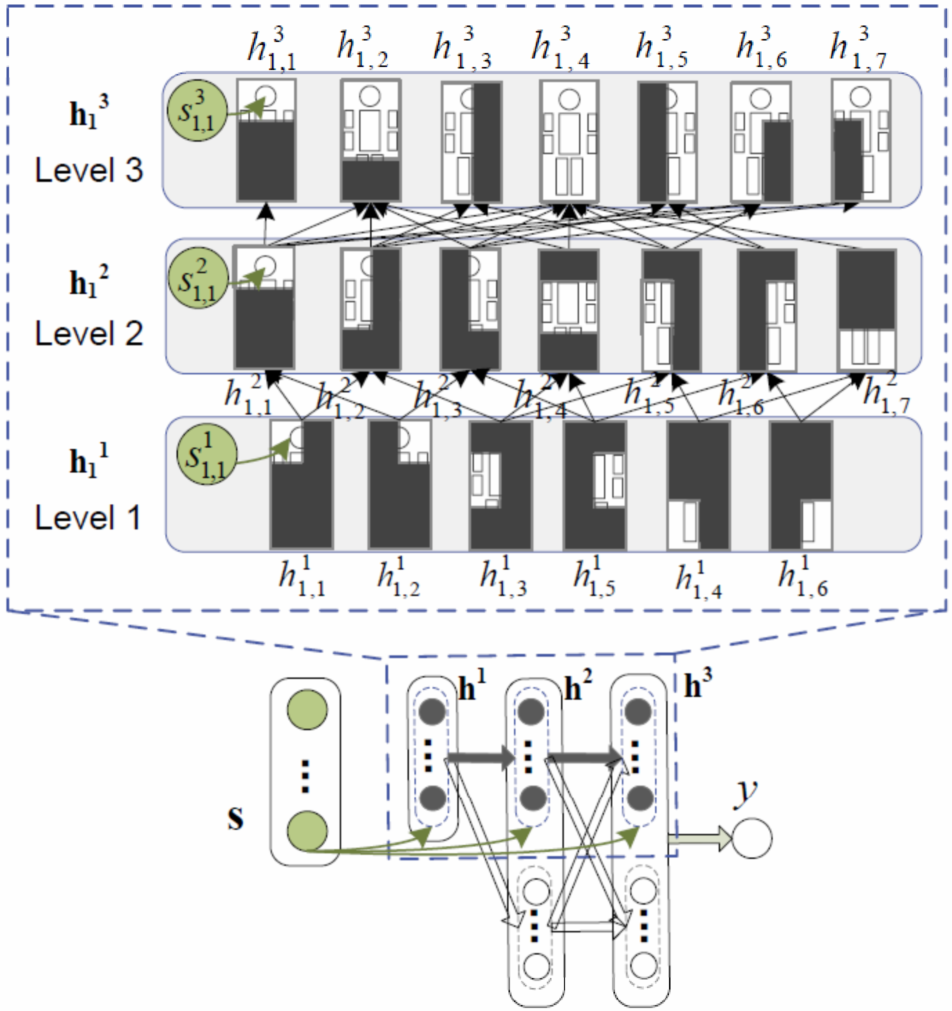


Learned filtered at the second convolutional layer

Deformation Layer



Visibility Reasoning with Deep Belief Net

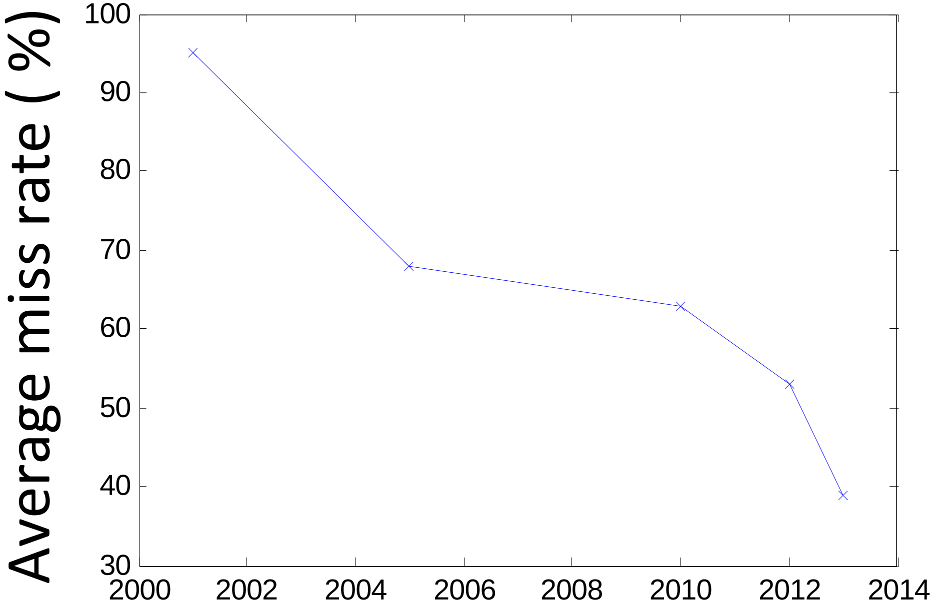


$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}}\mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

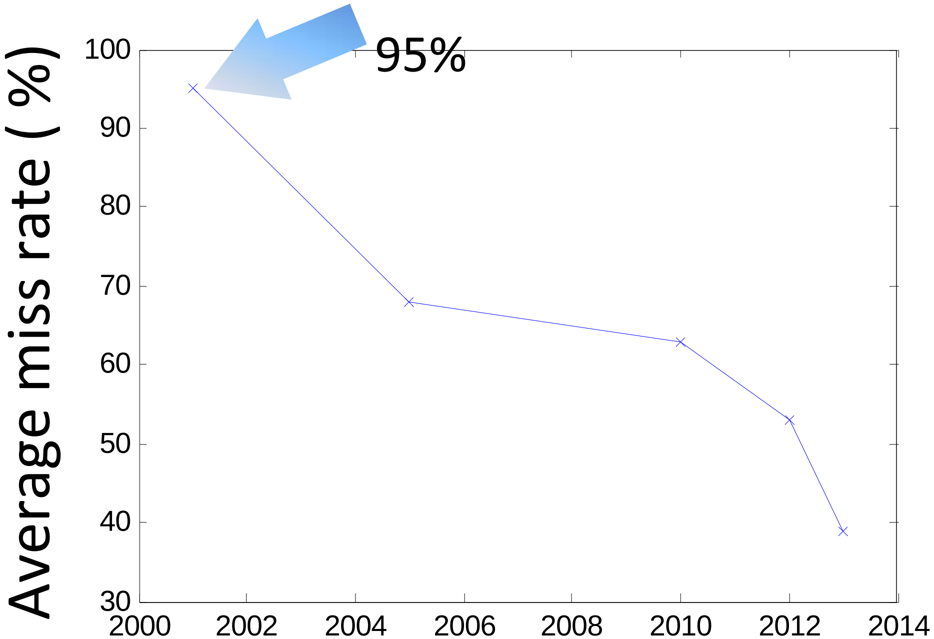
Experimental Results

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



Experimental Results

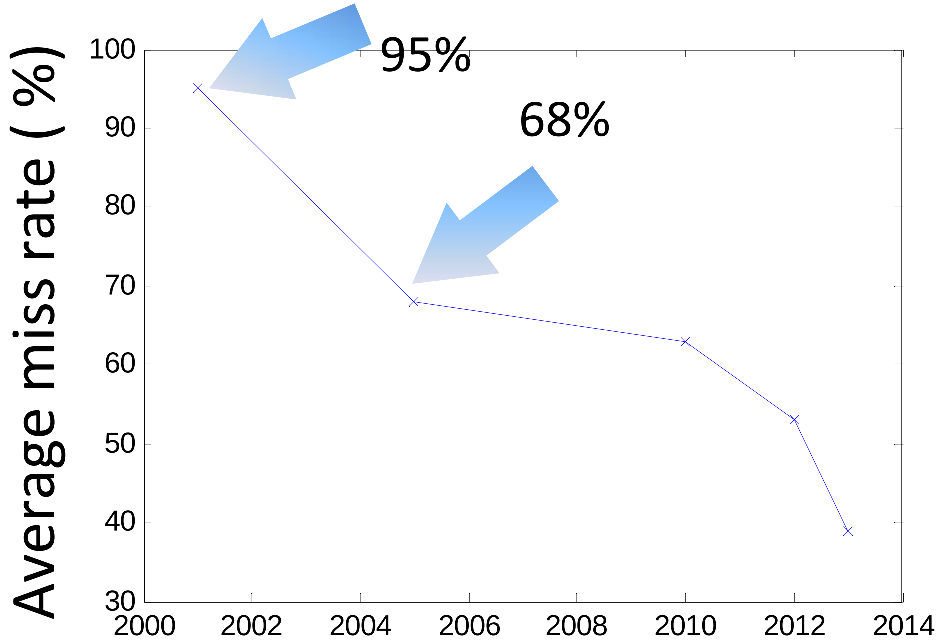
- Caltech – Test dataset (largest, most widely used)



Rapid object detection using a boosted cascade of simple features
P Viola, M Jones - … Vision and Pattern Recognition, 2001. CVPR …, 2001 - ieeexplore.ieee.org.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

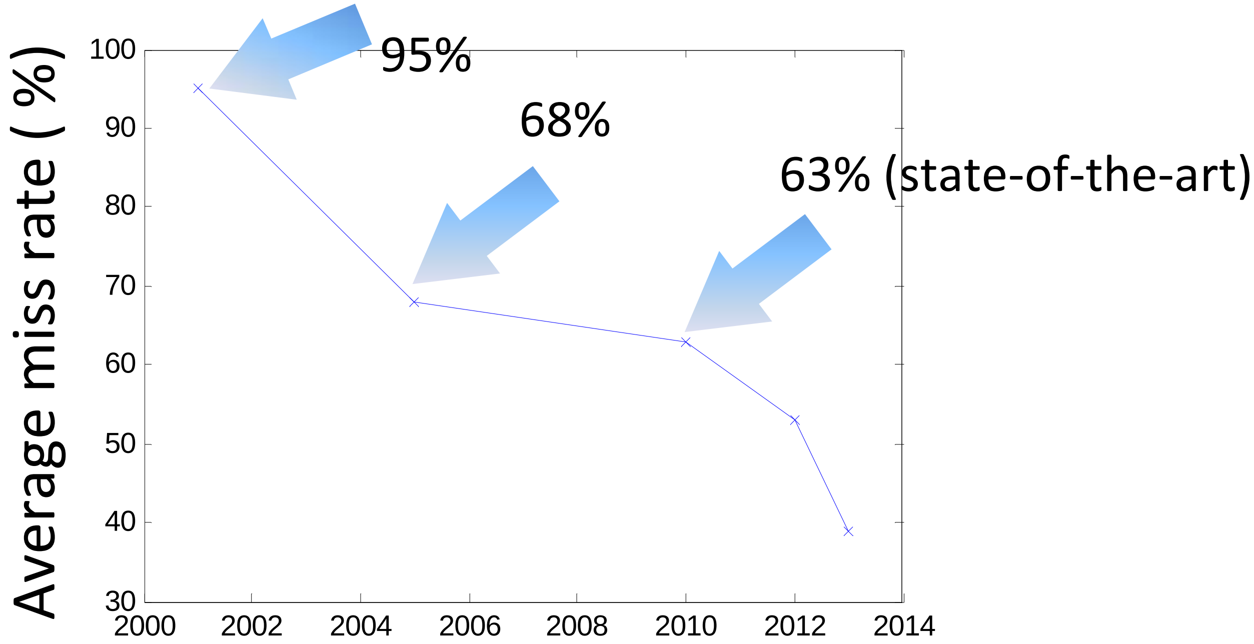
- Caltech – Test dataset (largest, most widely used)



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

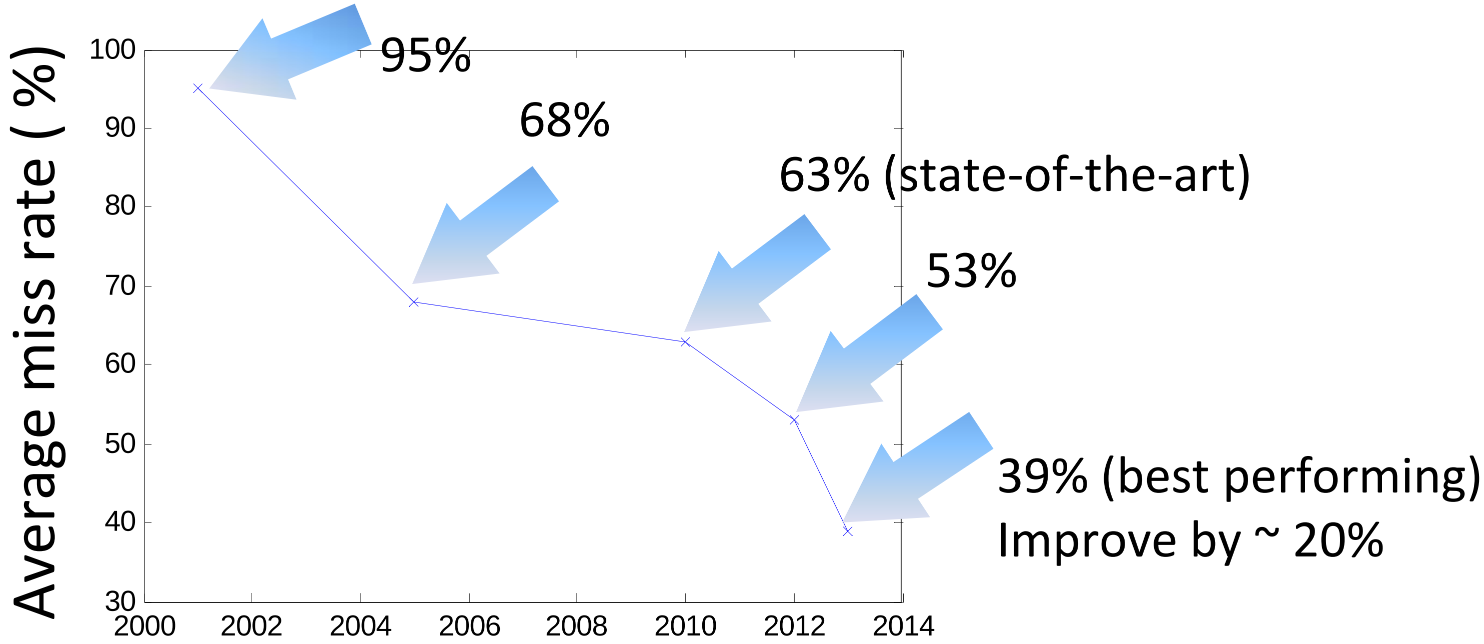
PF Felzenszwalb, RB Girshick… - Pattern Analysis and …, 2010 - ieeexplore.ieee.org

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

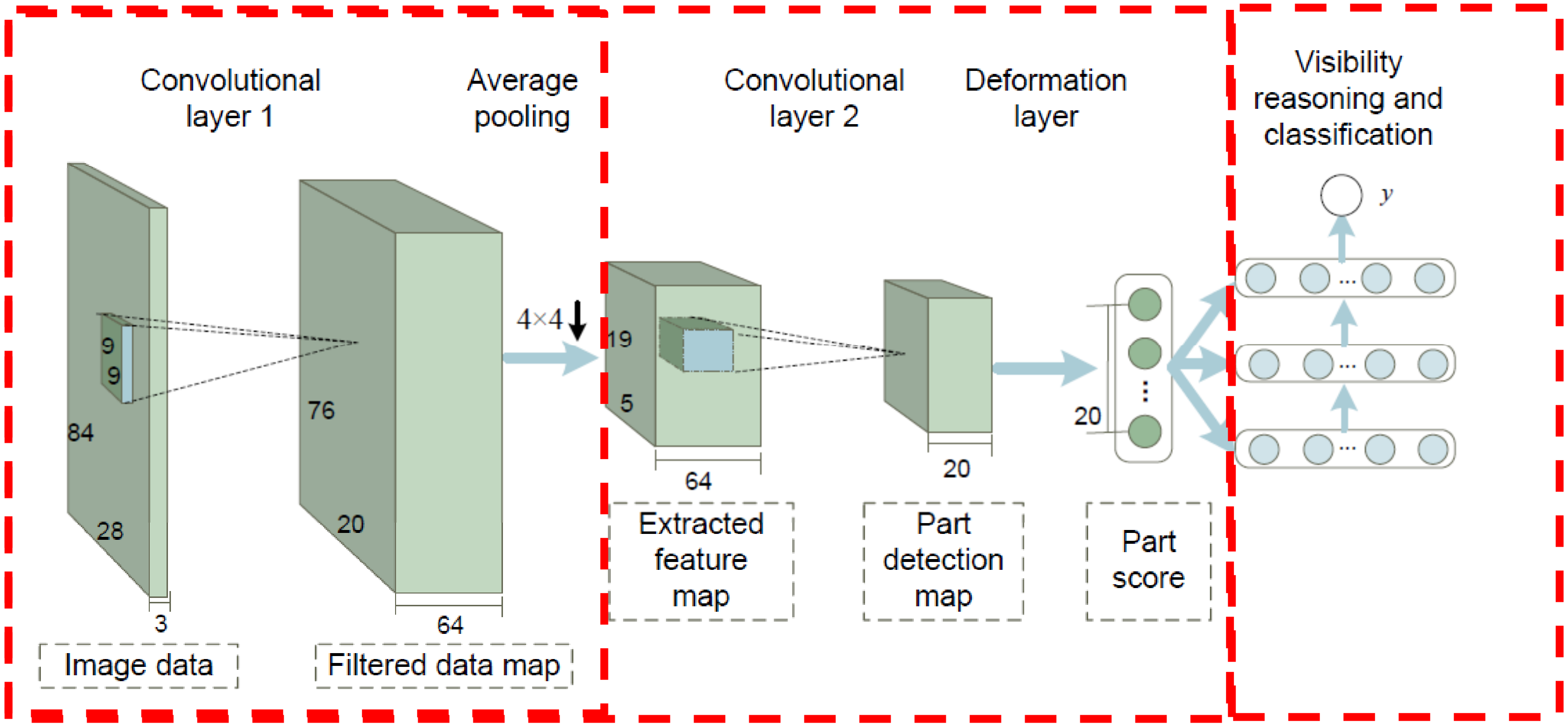
Experimental Results

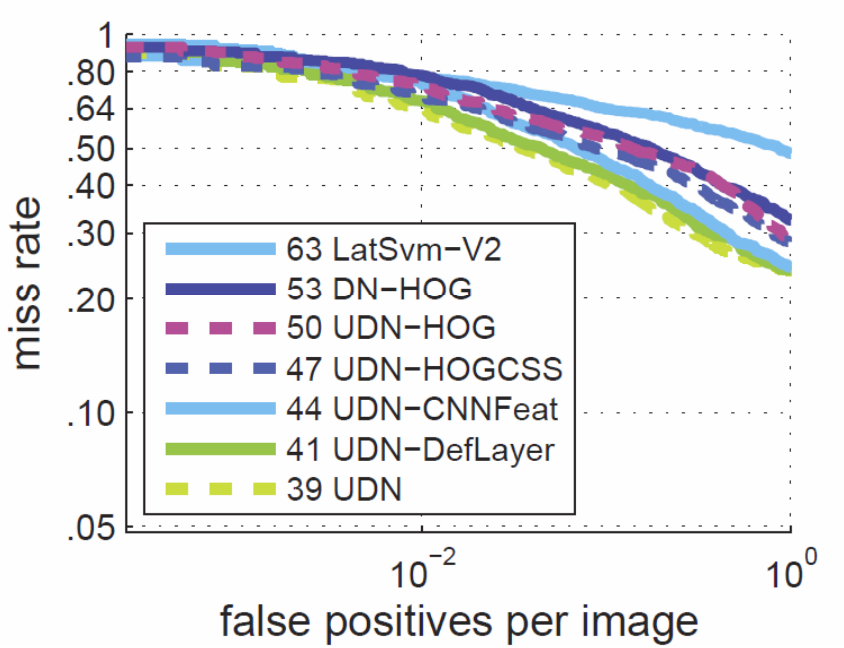
- Caltech – Test dataset (largest, most widely used)



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.

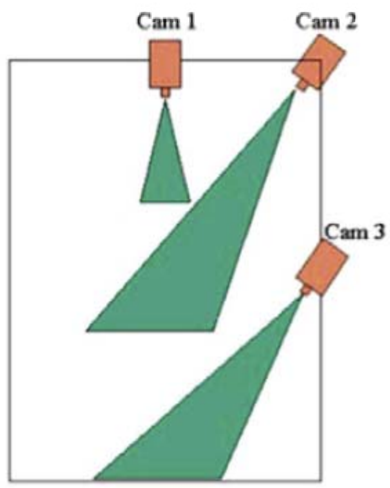




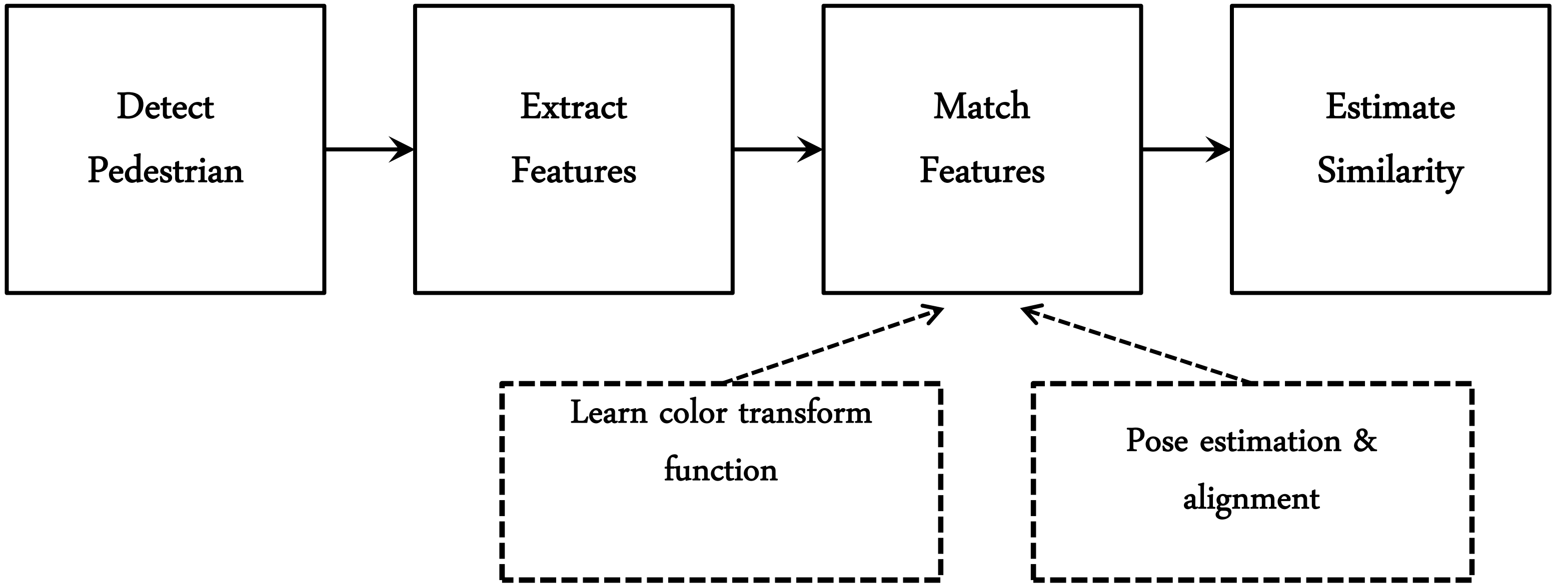
DN-HOG
UDN-HOG
UDN-HOGCSS
UDN-CNNFeat
UDN-DefLayer

Joint Deep Learning for Person Re-Identification

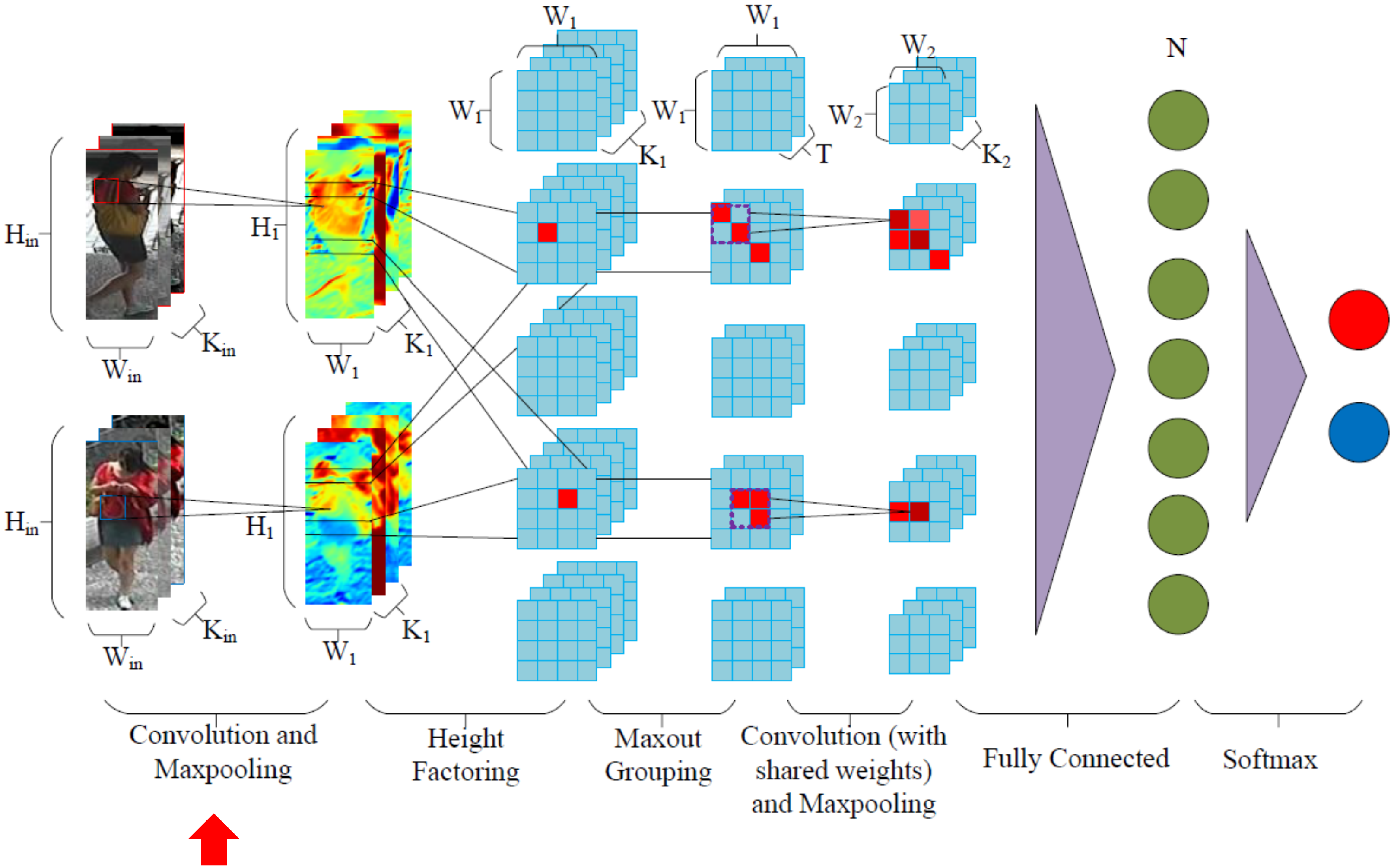
- Challenges
  - Photometric transforms (lighting, camera settings)
  - Geometric transforms (poses, views)
  - Occlusions
  - Background clutters



Pipeline of Existing Systems

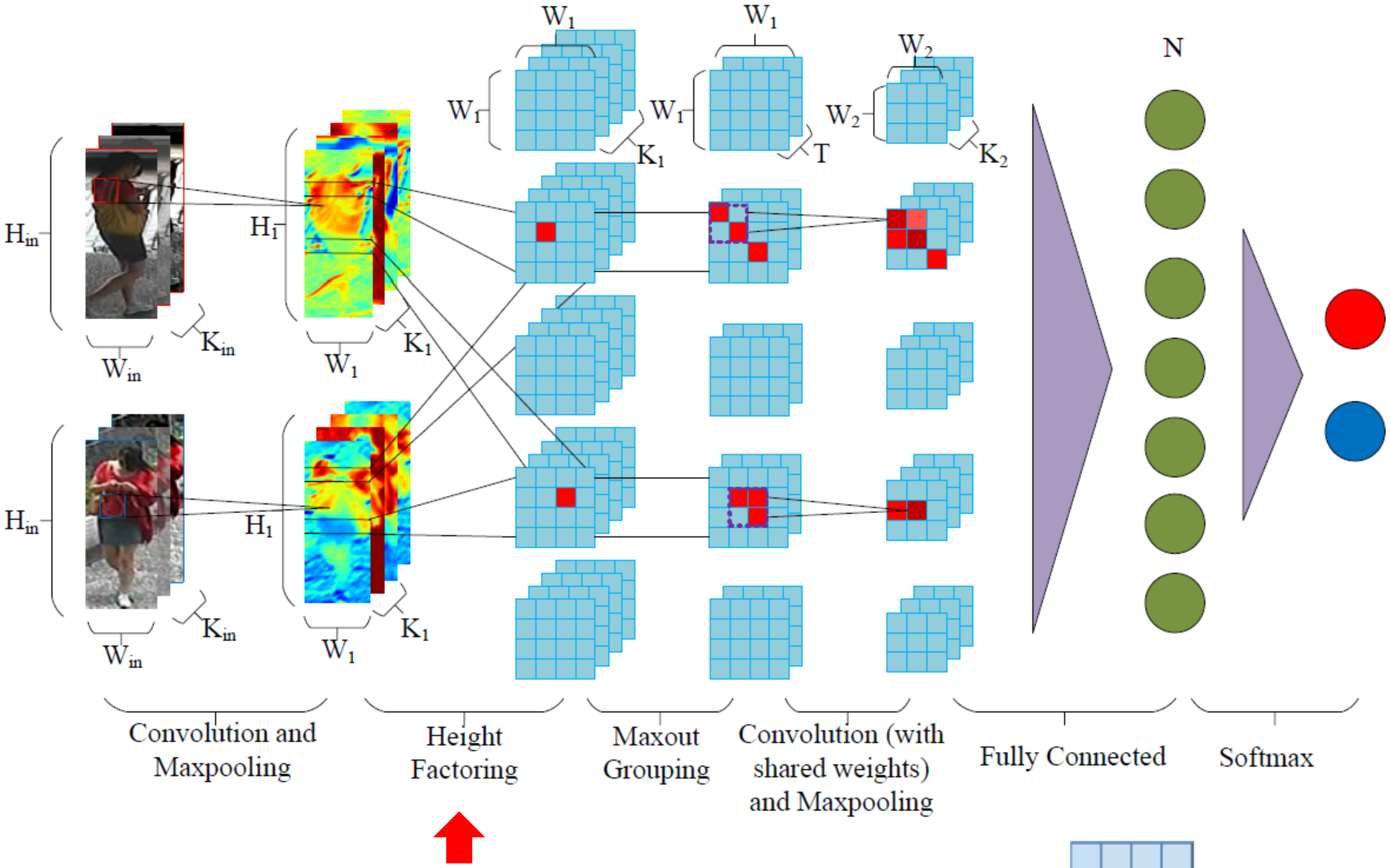


Optimize each module separately

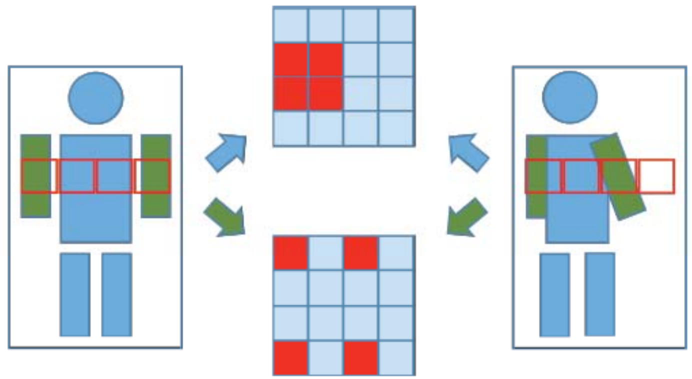


Filter pairing: filters ($\mathbf{W}_k$, $\mathbf{V}_k$) applied to different camera views are paired and their difference reflects the photometric transforms

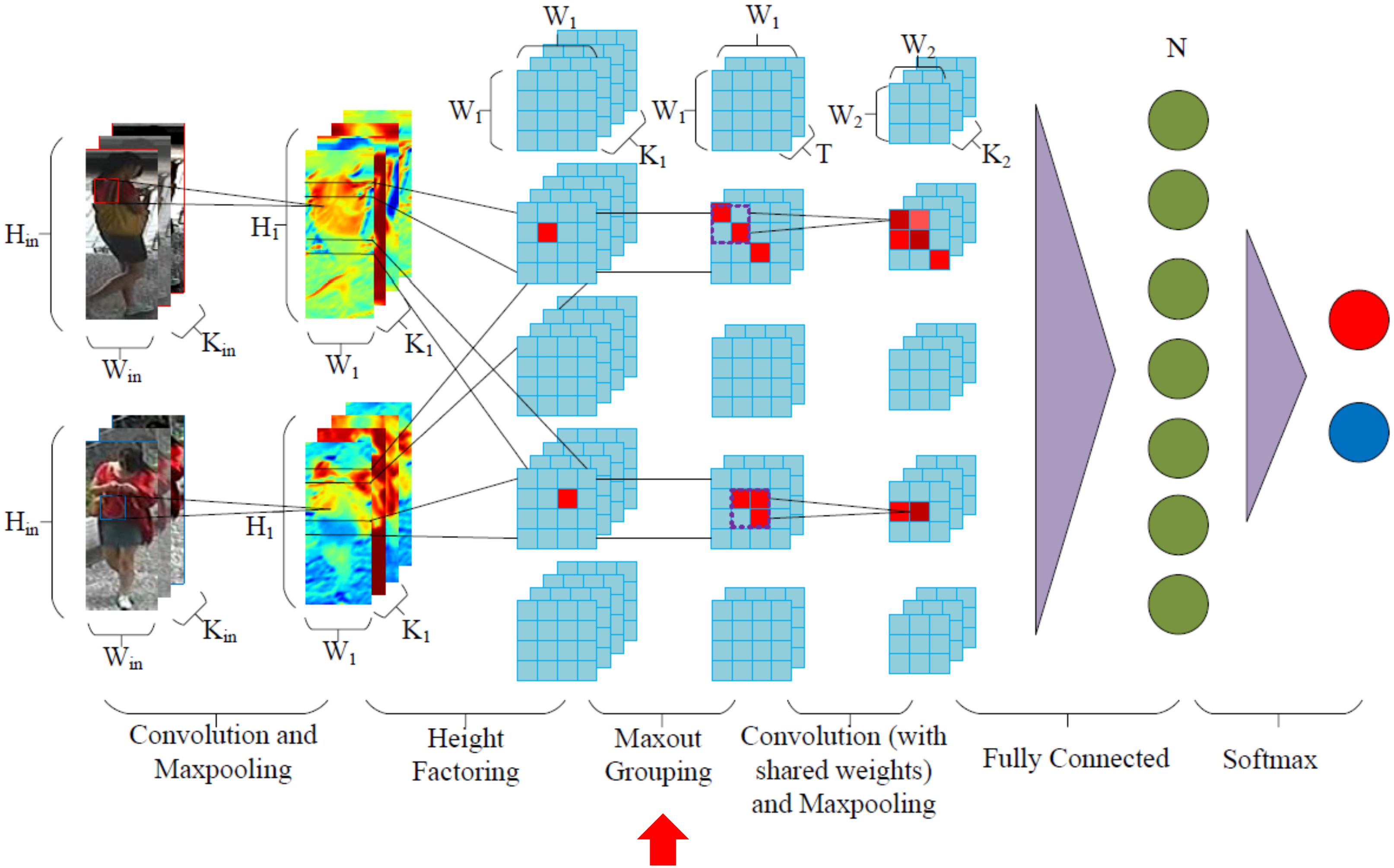
$$f_{ij}^k = \sigma((\mathbf{W}_k * \mathbf{I})_{ij} + b_k^I)$$
$$g_{ij}^k = \sigma((\mathbf{V}_k * \mathbf{J})_{ij} + b_k^J)$$



Patch matching layer:
match patches in the same horizontal stripe
K filters generate K displacement matrices for each stripe



$$S^k_{(i,j)(i',j')} = f^k_{ij} g^k_{i'j'}$$

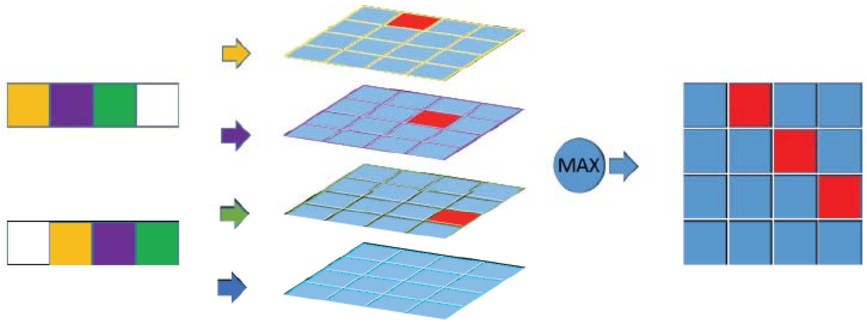


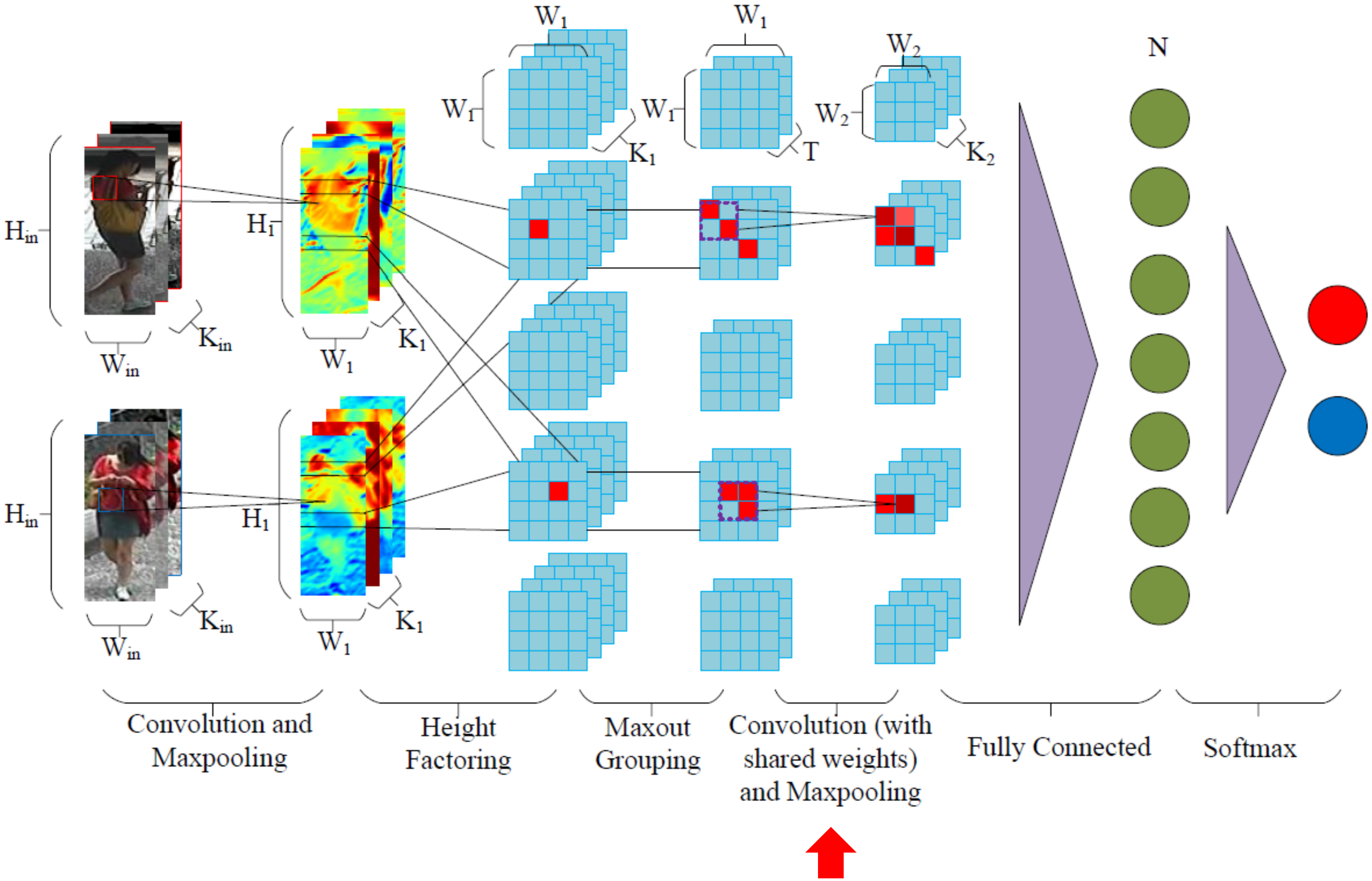
Maxout-grouping layer:

Each feature is represented by multiple channels

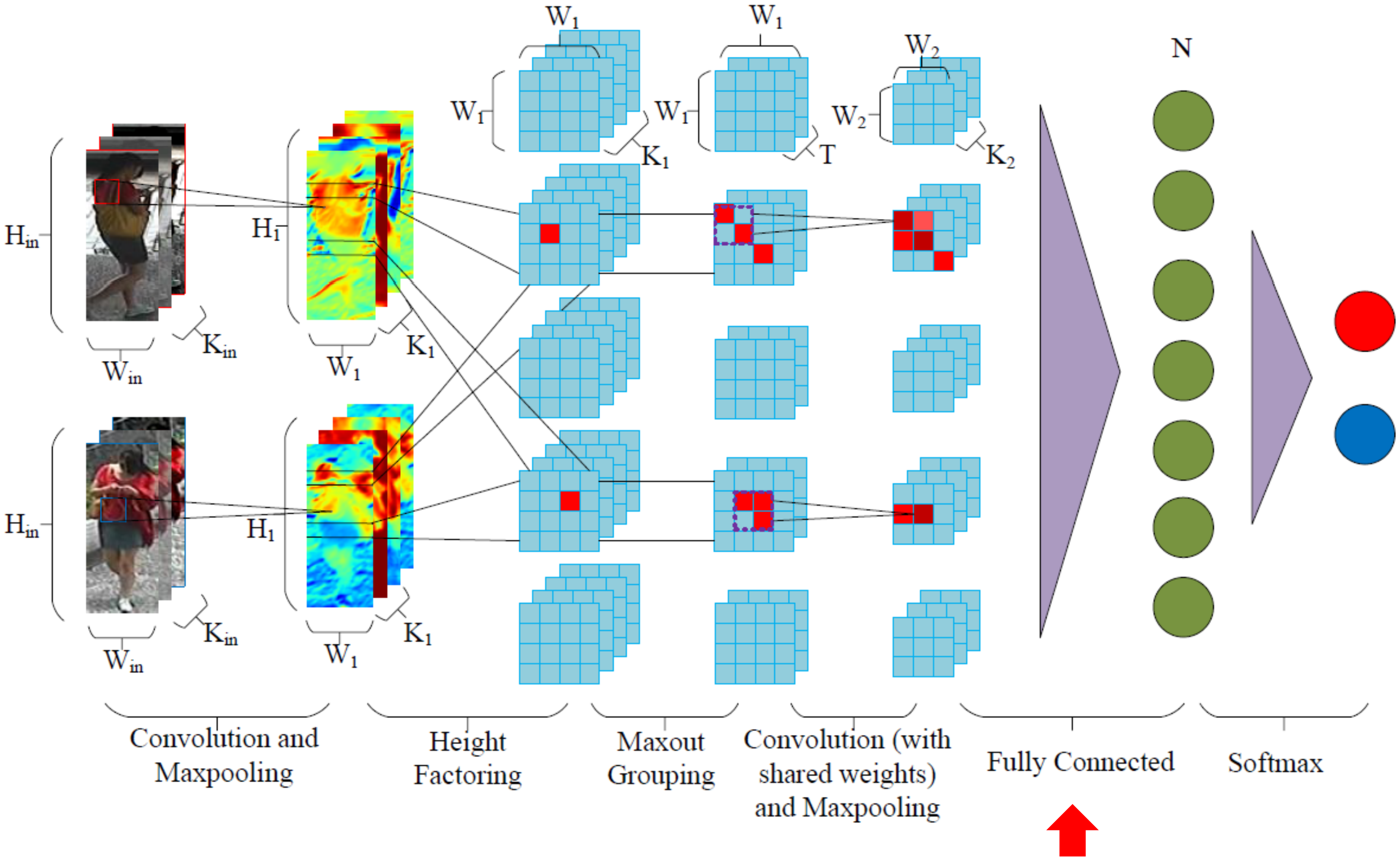
Model a mixture of photometric transforms

Robust to misdetection of patch matching





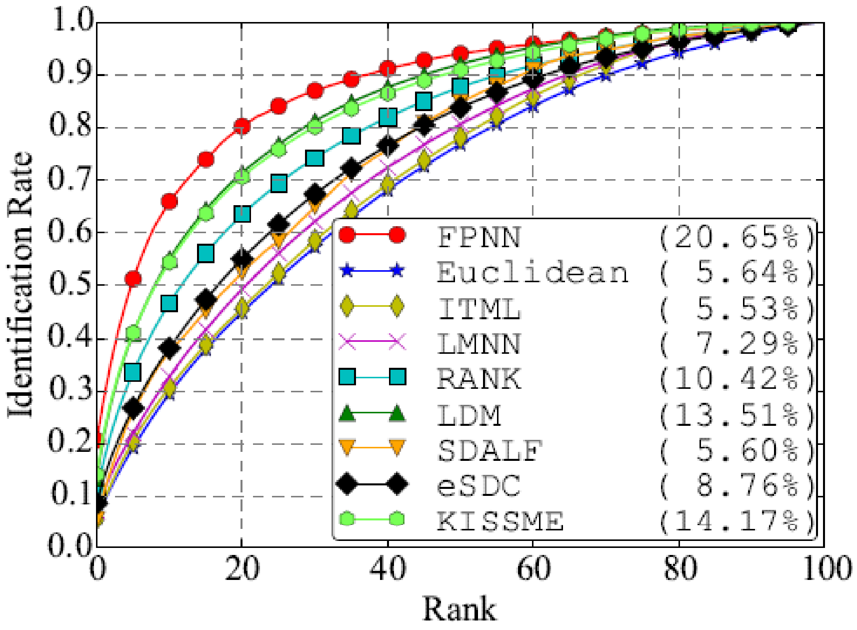
Through another **convolution and max-pooling layer**, the learned filters are applied to displacement matrices and capture local patterns of part displacements



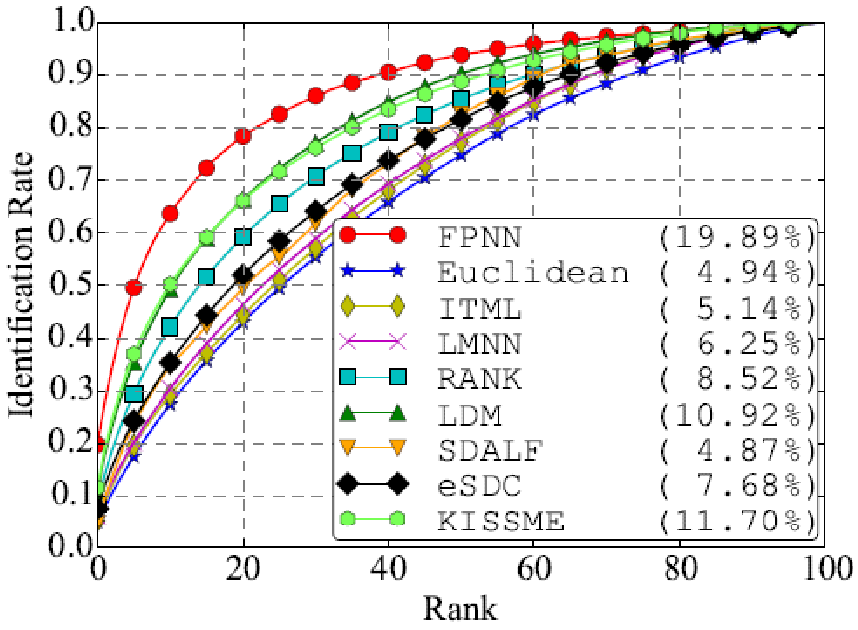
A fully connected layer captures global geometric transforms

Result

- Evaluated on 13,164 of 1,360 persons



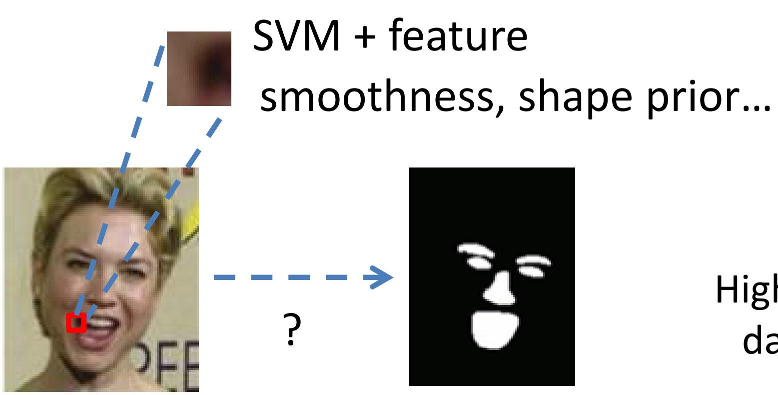
Manually labeled bounding boxes

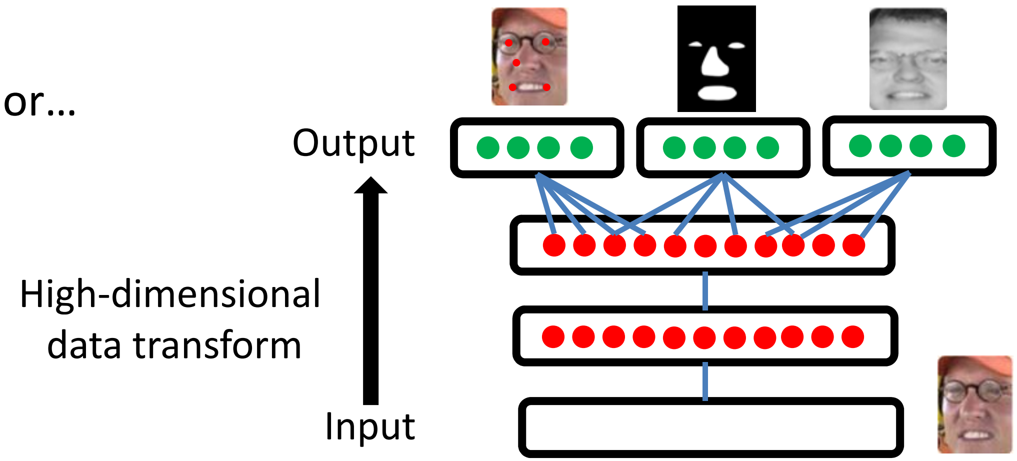


Automatically detected bounding boxes

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





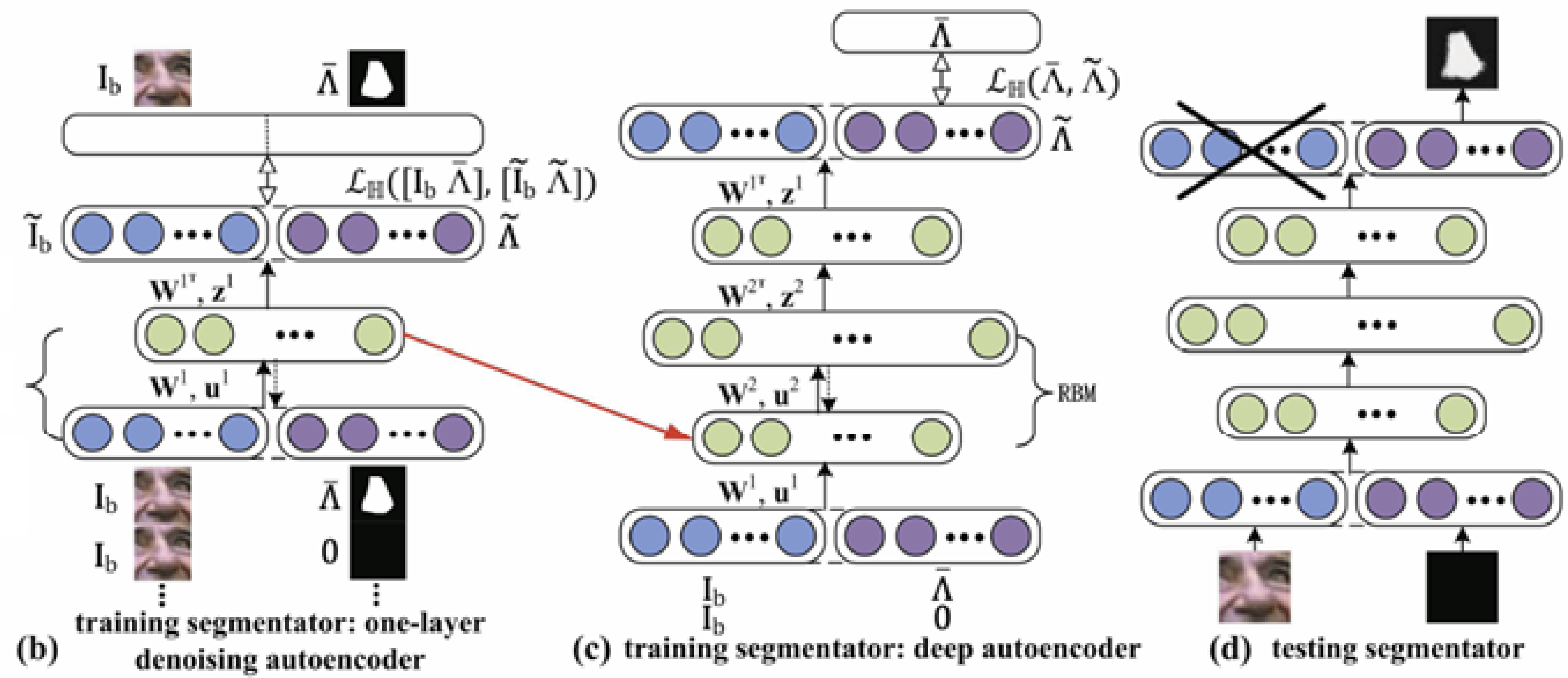
Face Parsing

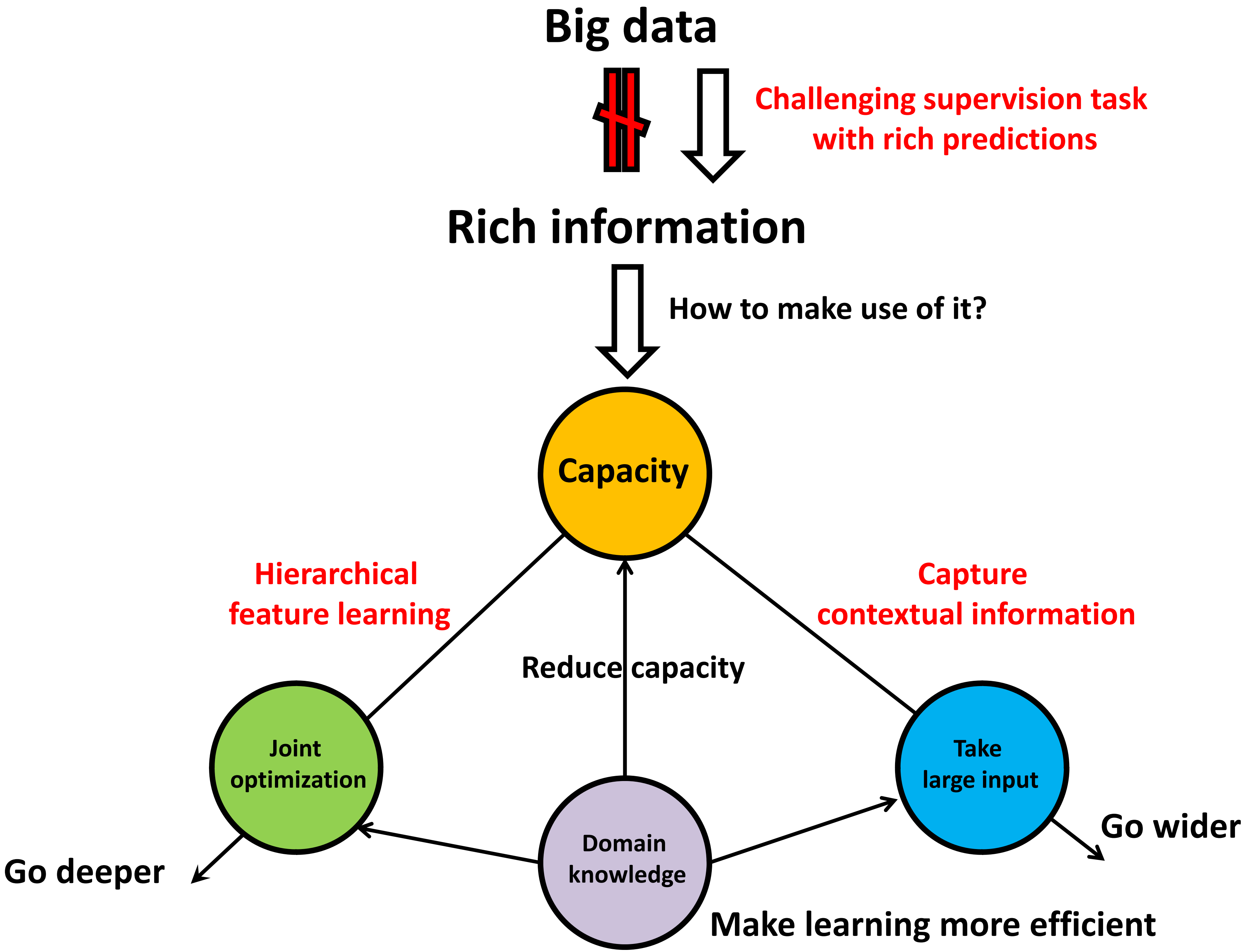
- P. Luo, X. Wang and X. Tang, “Hierarchical Face Parsing via Deep Learning,” CVPR 2012

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

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

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.