www.ccfcv.org

中国计算机学会•计算机视觉专业组

欢迎关注

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

模式识别国家重点实验室
National Lab of Pattern Recognition

中国科学院自动化研究所
Institute of Automation Chinese Academy of Sciences

视觉大数据的深度计算

王　亮

智能感知与计算研究中心/模式识别国家重点实验室
中国科学院自动化研究所

2015年11月

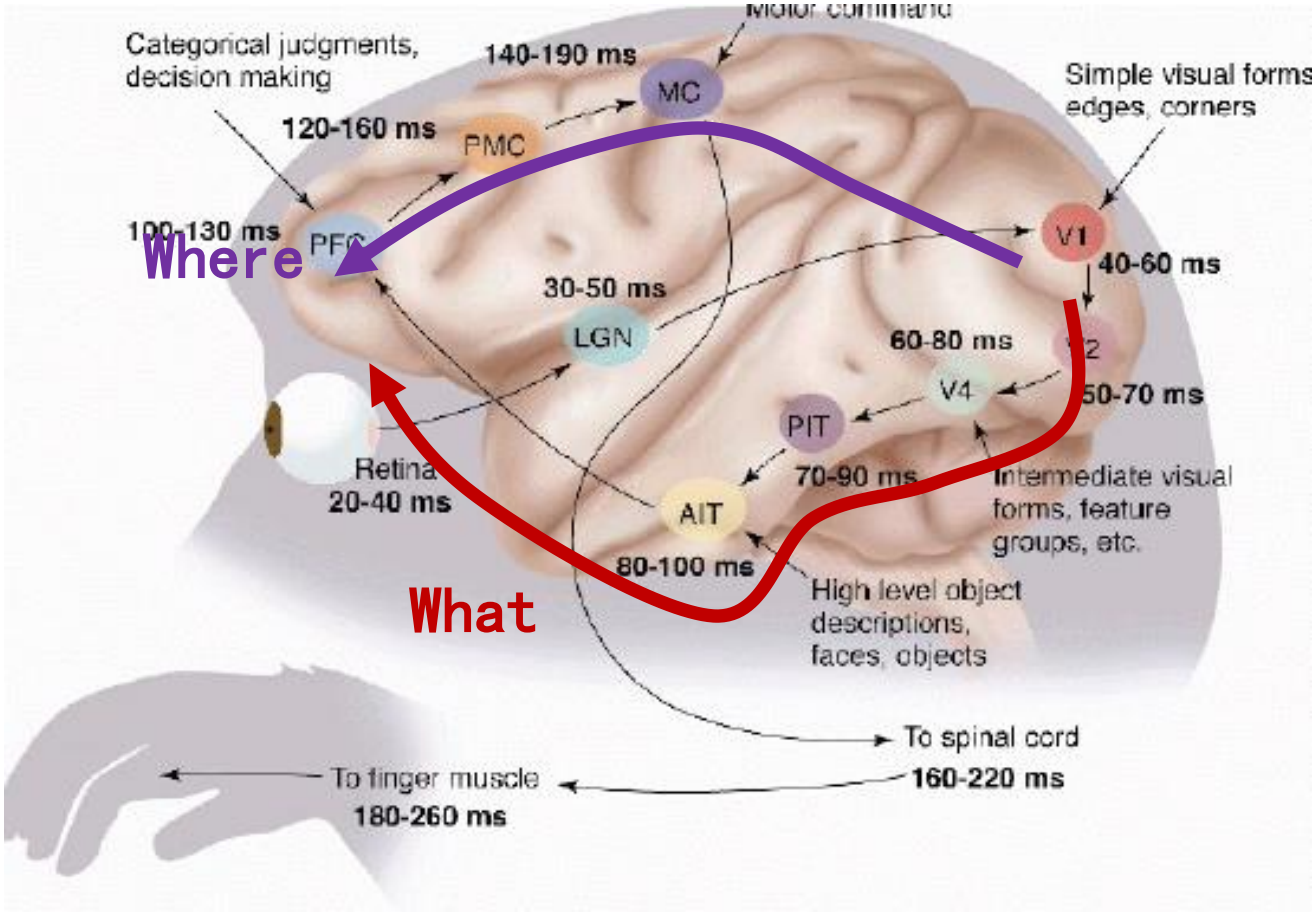
报告提纲

- 视觉大数据
- 大规模视觉计算
- 我们的工作
- 未来方向

报告提纲

- 视觉大数据
- 大规模视觉计算
- 我们的工作
- 未来方向

人类至少有70%以上的外界信息是由视觉系统所接收、处理和感知的



视觉大数据：记录客观世界

随着采集设备的普及，
图像视频大数据记录了人们
生活的方方面面，数据量以前所
未有的速度在积累和增加

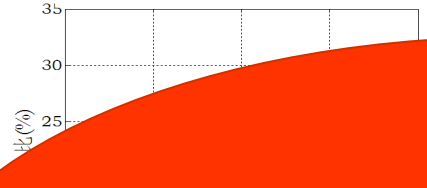
视觉大数据的重要性

互联网

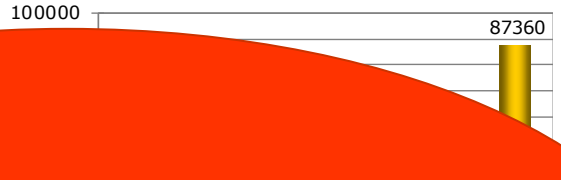
Facebook用户量已经超过8亿，每天上传图片超过3亿张，视频超过300万个

移动互联网

社交网络数据增长率



视频监控数据量增长趋势（PB）



视觉数据→视觉红利

Visual Data → Visual Dividend

Video → Value

广电网

视联网

视觉大数据是模式识别的前沿方向，是人工智能的突破口，是信息产业新的增长点

图像视频大数据的主要应用场所

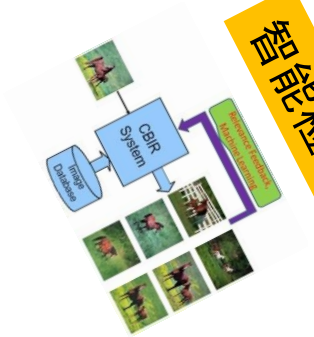
休闲娱乐

智能监控

无人机/车

考勤安检

网络信息过滤



智能检索

公安刑侦

机器人

遥感测绘

报告提纲

- 视觉大数据
- 大规模视觉计算
- 我们的工作
- 未来方向

大规模视觉计算的概念

- 视觉计算
  - 对视觉信息（或数据）的分析与处理
- 大规模视觉计算
  - 对大规模视觉信息（或数据）的分析与处理
    = 视觉大数据计算
  - 大规模视觉信息：规模大、类别多、来源广

大规模视觉计算的挑战

与传统视觉计算相比，大规模视觉计算面临的挑战：

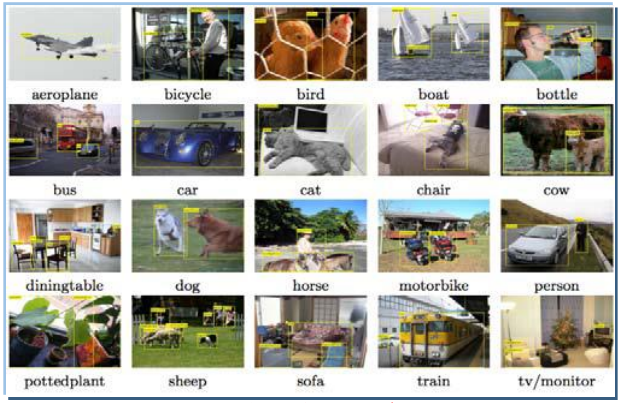
视频、论坛、博客、微博、短信等

跨景跨媒

海量庞杂

多源异质

VOC：20类目标，<2万图片



Top 1分类精度：>90%

ImageNet：1000类目标，130万图片

传统算法在ImageNet上<50%

计算机视觉系统&人眼视觉系统

最先进的机器人依然不能
仅靠视觉系统穿越十字路口

鲁棒性远逊色于人眼视觉系统

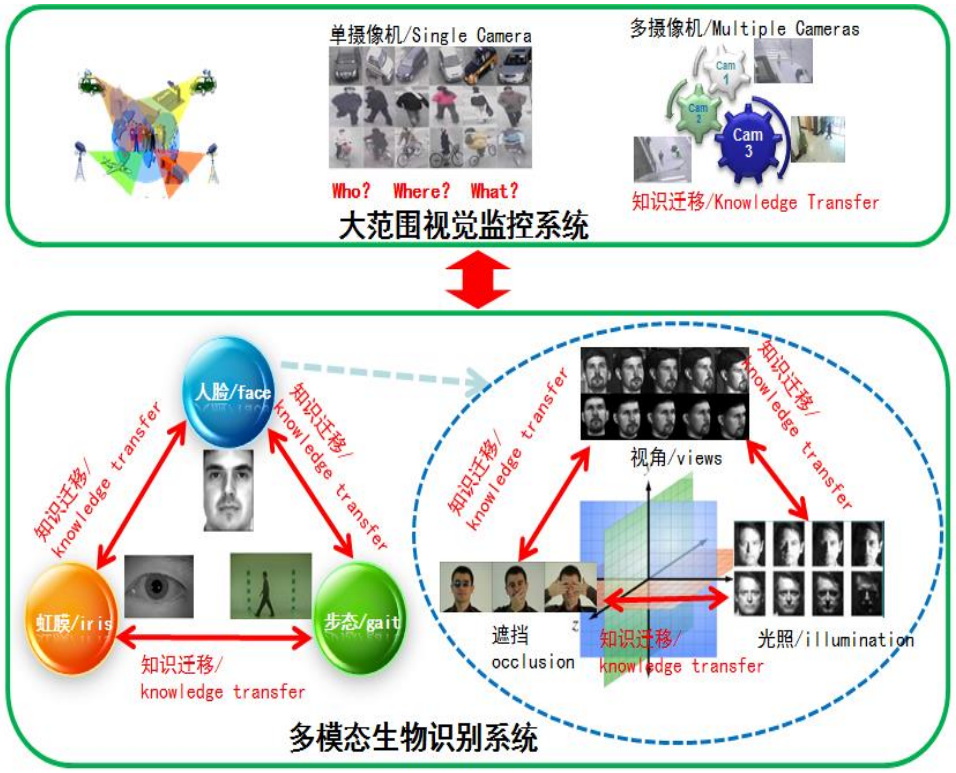
学术界最先进的算法在1000类目标数据库上得到约30%的检测精度，而人眼视觉系统可以在复杂场景下轻松识别超过2万类目标！

我们希望通过研究大规模视觉计算来逼近人眼视觉系统，有效地处理规模大、类别多、来源广的视觉大数据！

VOC2010检测 计算机:35%，人眼:95%

大规模视觉计算的关键问题

人工智能顶级会议NIPS2012设立了子会：
Large Scale Visual Recognition and Retrieval
总结了大规模视觉计算蕴含的几个关键科学问题



算法层

- 大规模特征表达
- 大规模模型学习
- 大规模知识迁移

系统层

- 大规模数据库构建
- 大规模数据处理平台

关键问题1：大规模特征表达

视频、论坛、博客、微博、短信等

跨景跨媒

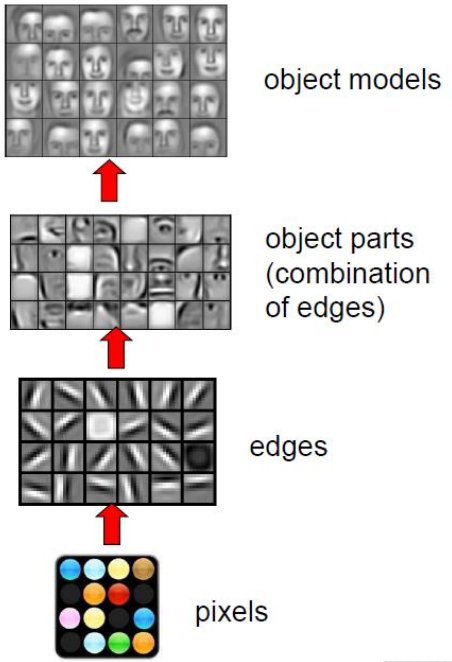
多源异质

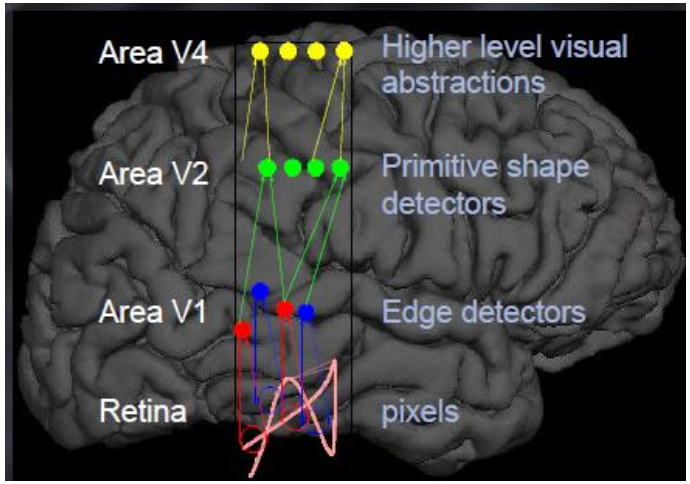
如何在跨景跨媒、多源异质的视觉大数据中找到具有较好泛化性和不变性的表达？

鲁棒特征表达

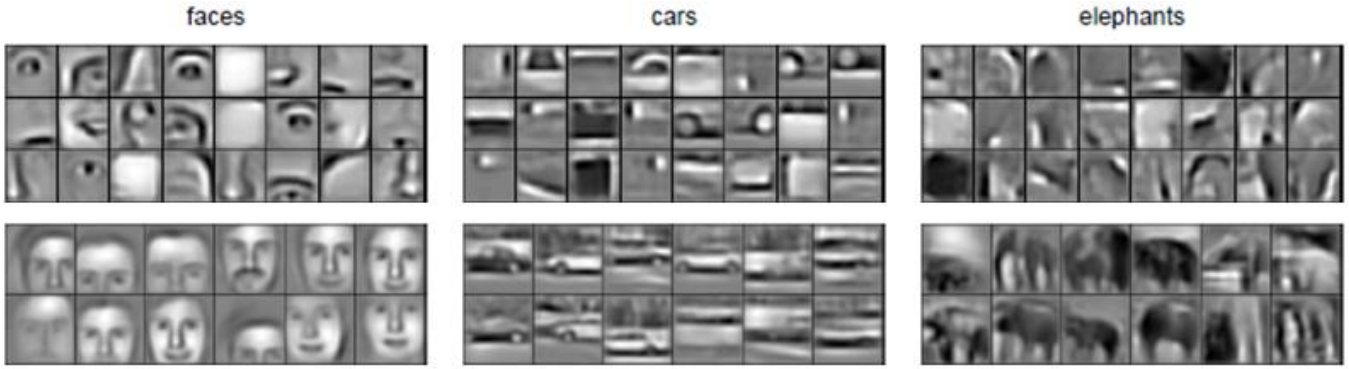
基于先验知识的鲁棒特征表达

先验知识：像素->边->部件->物体；稀疏，低秩；小世界网络



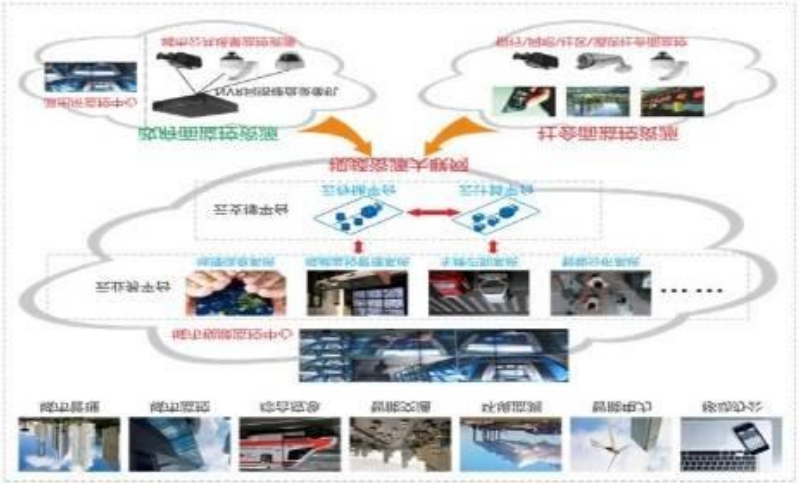


从大规模样本中自动学习出鲁棒特征表达



Slide credit: Kai Yu

关键问题2：大规模模型学习



海量庞杂

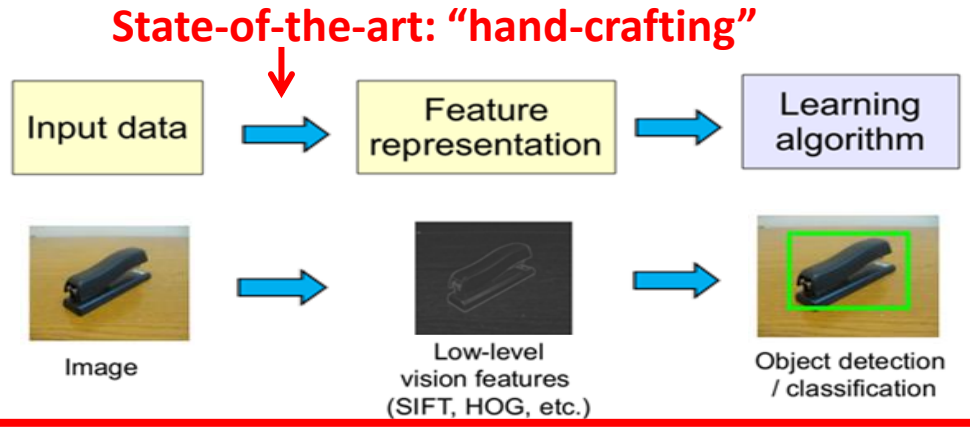
ImageNet：1000类目标，130万图片

种类繁多

面对海量庞杂、种类繁多的视觉大数据，人为设计的特征不一定适用于大规模模型学习

直接从海量数据中学习模型

视觉大数据时代模型学习的突破：端到端深度学习



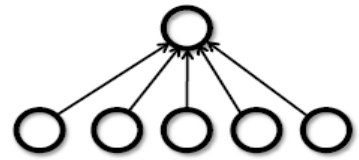
- 需要经验知识手工设置视觉特征提取算法
- 缺少与环境的信息交互以及知识库的决策支持

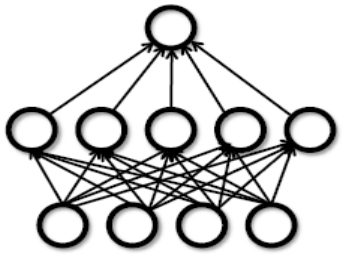
- 不再区分特征提取和模式分类
- 通过深度神经网络非线性模拟从pixel到label的映射关系

从数据直接到概念要素是当前对大数据语义理解的变革性思路——“语义就在大数据中”

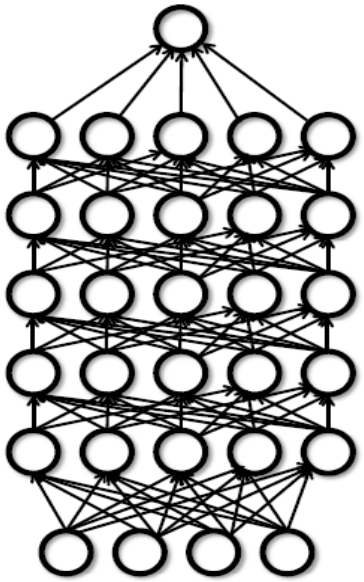
深度学习兴起的一个重要契机就是大数据时代的到来！

浅层模型



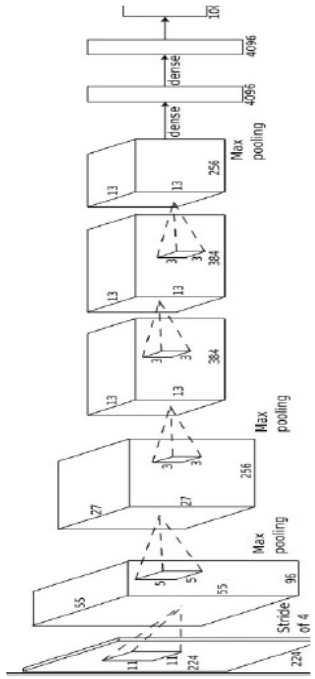


深度模型



Won the 2012 ImageNet LSVRC. 60 Million parameters, 832M MAC ops

4M	FULL CONNECT	4Mflop
16M	FULL 4096/ReLU	16M
37M	FULL 4096/ReLU	37M
	MAX POOLING	
442K	CONV 3x3/ReLU 256fm	74M
1.3M	CONV 3x3ReLU 384fm	224M
884K	CONV 3x3/ReLU 384fm	149M
	MAX POOLING 2x2sub	
	LOCAL CONTRAST NORM	
307K	CONV 11x11/ReLU 256fm	223M
	MAX POOL 2x2sub	
	LOCAL CONTRAST NORM	
35K	CONV 11x11/ReLU 96fm	105M



Slide Courtesy: LeCun et al.

深度神经网络DNN旨在建立模拟人脑进行分析学习的神经网络，它模仿人脑的神经网络机制来解释数据，例如图像，声音和文本。

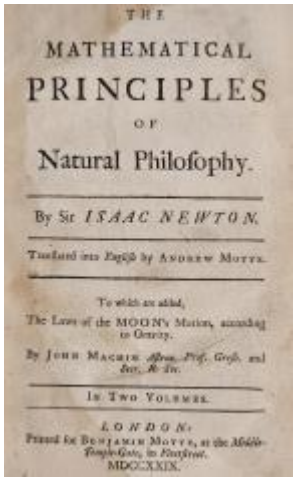
深度模型学习是大数据时代下视觉计算研究的一个重要突破，推动了视觉计算众多领域的飞速发展！

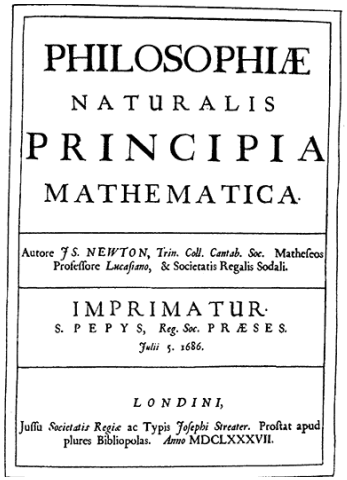
关键问题3：大规模知识迁移

心理学依据

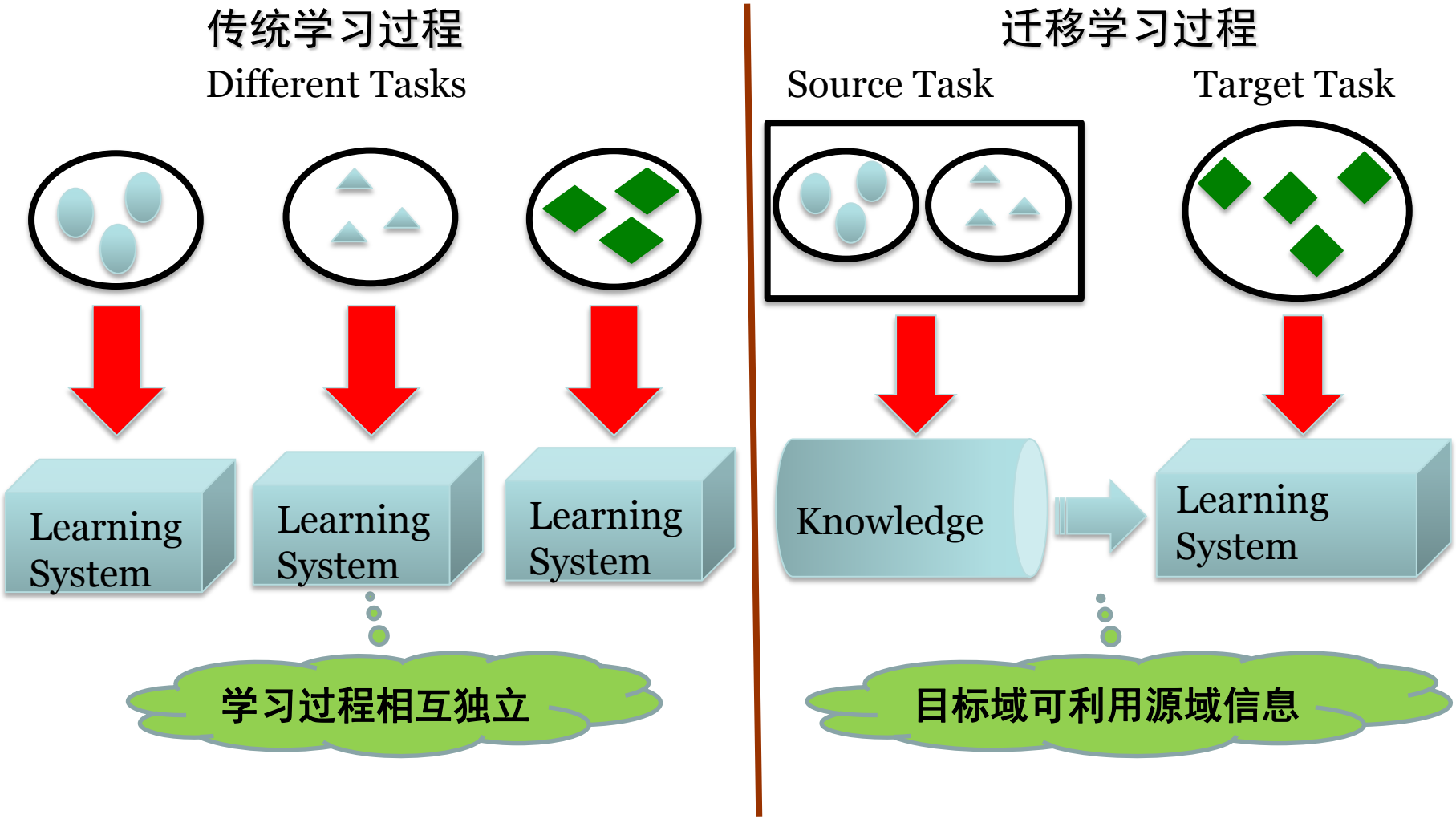
人类能够将某种知识或技能迁移到另一种相似的领域中

——[Thorndike and Woodworth, *Psychological Review* (1901)]

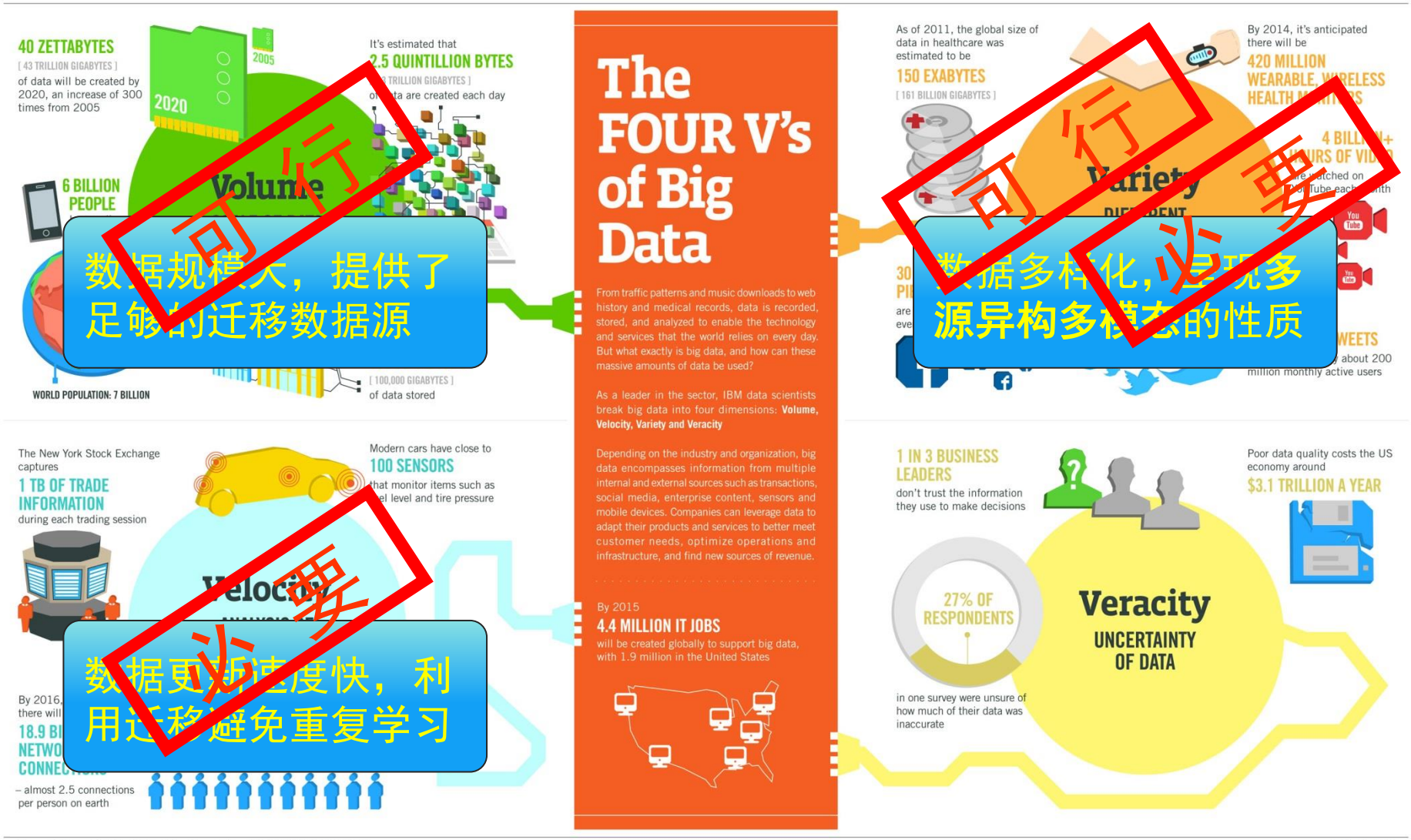




传统学习 vs. 迁移学习


传统学习过程
Different Tasks
Learning System
Learning System
Learning System
学习过程相互独立
迁移学习过程
Source Task
Target Task
Knowledge
Learning System
目标域可利用源域信息

大数据背景下的知识迁移



可行

数据规模大，提供了足够的迁移数据源

可行

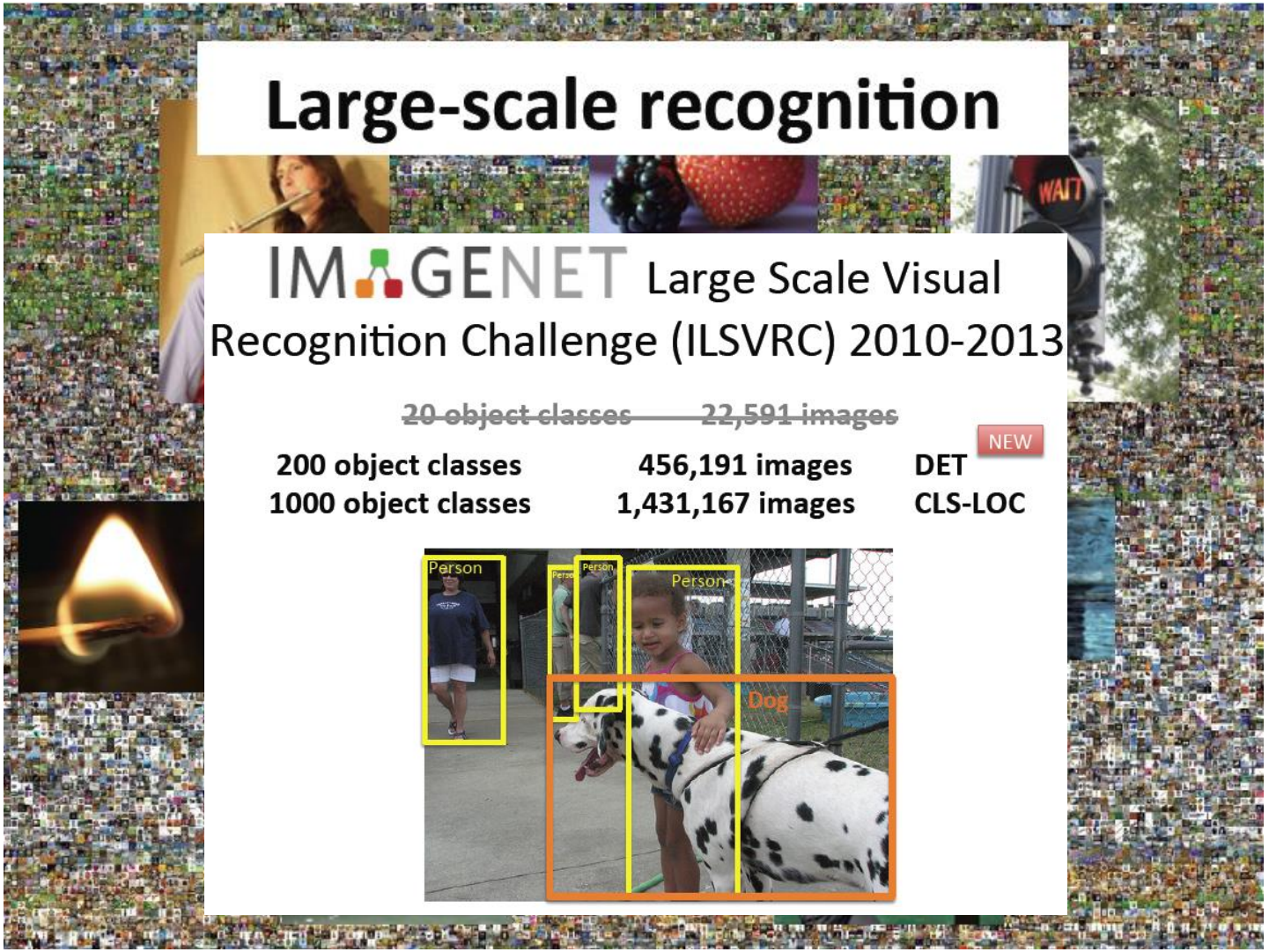
必要

数据多样化，呈现多源异构多模态的性质

必要

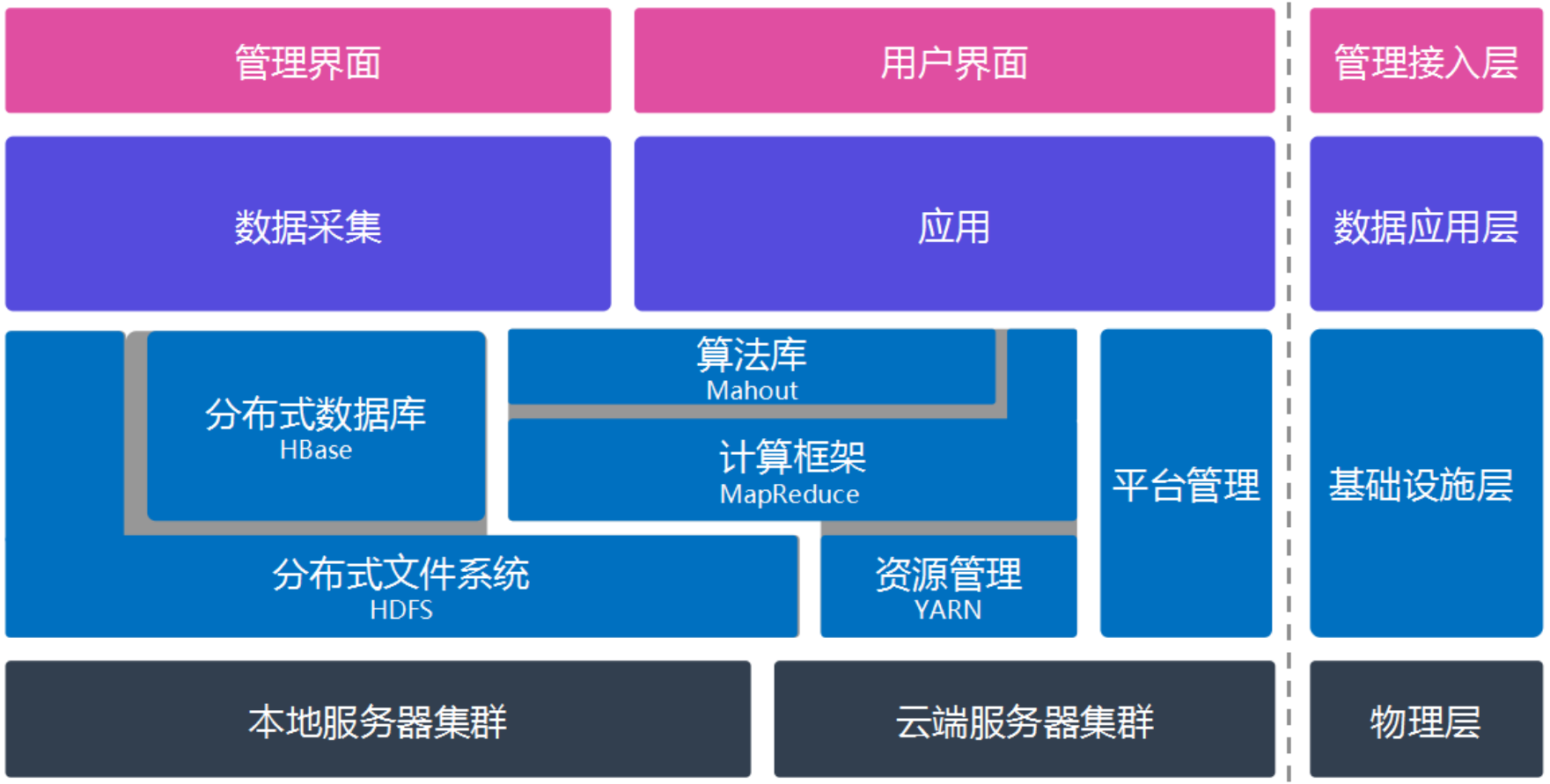
数据更新速度快，利用迁移避免重复学习

问题4：大规模视觉数据库



Slide credit: Fei-Fei Li et al.

问题5：大规模视觉数据处理平台



报告提纲

- 视觉大数据
- 大规模视觉计算
- 我们的工作
- 未来方向

深度学习背景

深度学习的前身

- **神经网络**
  - 1962 – neurobiological inspiration through simple/complex cell, *Hubel and Wiesel*
  - 1970 – efficient error backpropagation, *Linnainmaa*
  - 1979 – deep neocognitron, convolution, *Fukushima*
  - 1987 – autoencoder, *Ballard*
  - 1989 – backpropagation for CNN, *Lecun*
  - 1991 – fundamental deep learning problem, *Hochreiter*
  - 1991 – deep recurrent neural network, *Schmidhuber*
  - 1997 – supervised LSTM RNN, *Schmidhuber*

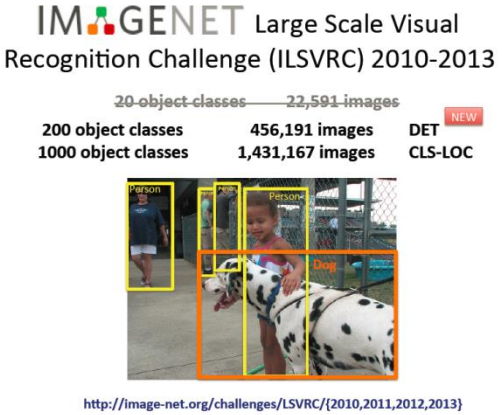
Deep Learning in Neural Networks: An Overview. Jurgen Schmidhuber, Technical Report 2014

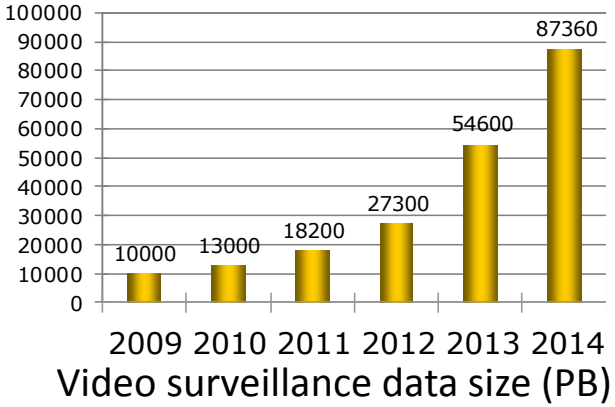
深度学习的前身

- 神经网络通过模拟大脑认知的机理解决各种机器学习问题
- 神经网络的缺陷：
  1. 包含大量参数，导致计算复杂度高
  2. 需要大训练集，容易导致过拟合问题
  3. 与其他模型相比，在识别准确率上没有明显优势
- 因此，多数学者转而选择：
  1. SIFT和LBP等手工设计的特征
  2. SVM和Boosting等浅层分类模型

深度学习兴起的契机

Big Data



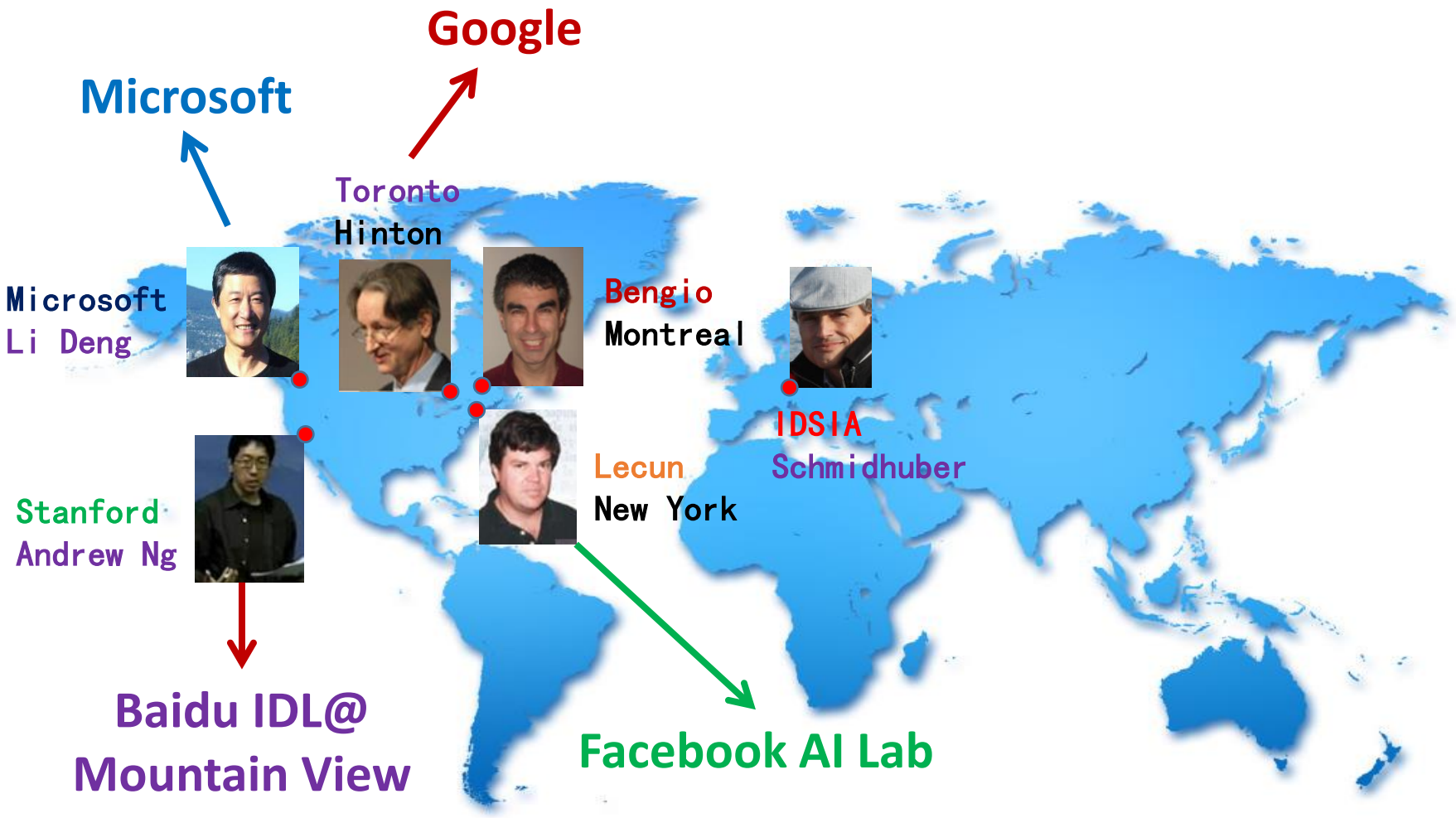


Cheap Computation

大规模数据和高性能计算的出现，使得拥有大量模型参数的神经网络可以被高效快速的拟合

point performance			
Memory bandwidth (ECC off)	288 GB/sec	250 GB/sec	208 GB/sec
Memory size (GDDR5)	12 GB	6 GB	5 GB
CUDA cores	2880	2688	2496

深度学习先锋者(2006-)



他们引导深度学习经历了三个主要阶段！

1: “RBM/AE”阶段(2006–)

RBM: restricted boltzmann machine; AE: auto-encoder

Breakthrough in 2006

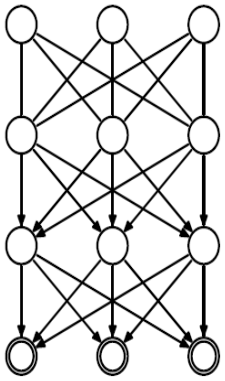
Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

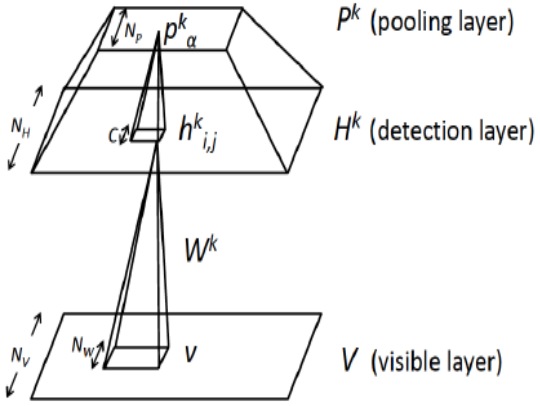
High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such “autoencoder” networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

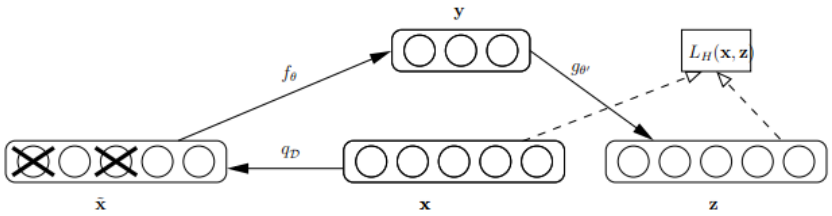
Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer “encoder” network

2006 VOL 313 SCIENCE www.sciencemag.org



Deep Belief Network

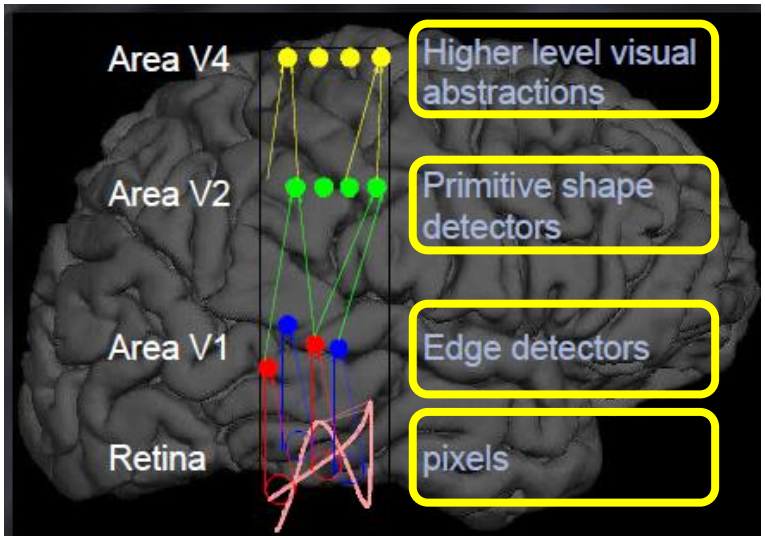


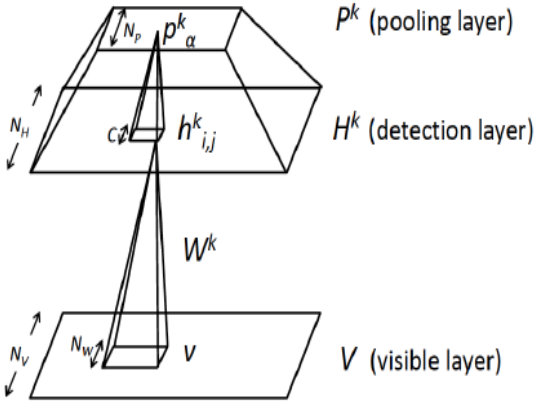


多种RBM和AE模型

- **模型特点：**生成式模型，中等规模数据集，较深层网络
- **热点问题：**替代传统手工设计特征，进行数据表示学习(representation learning)

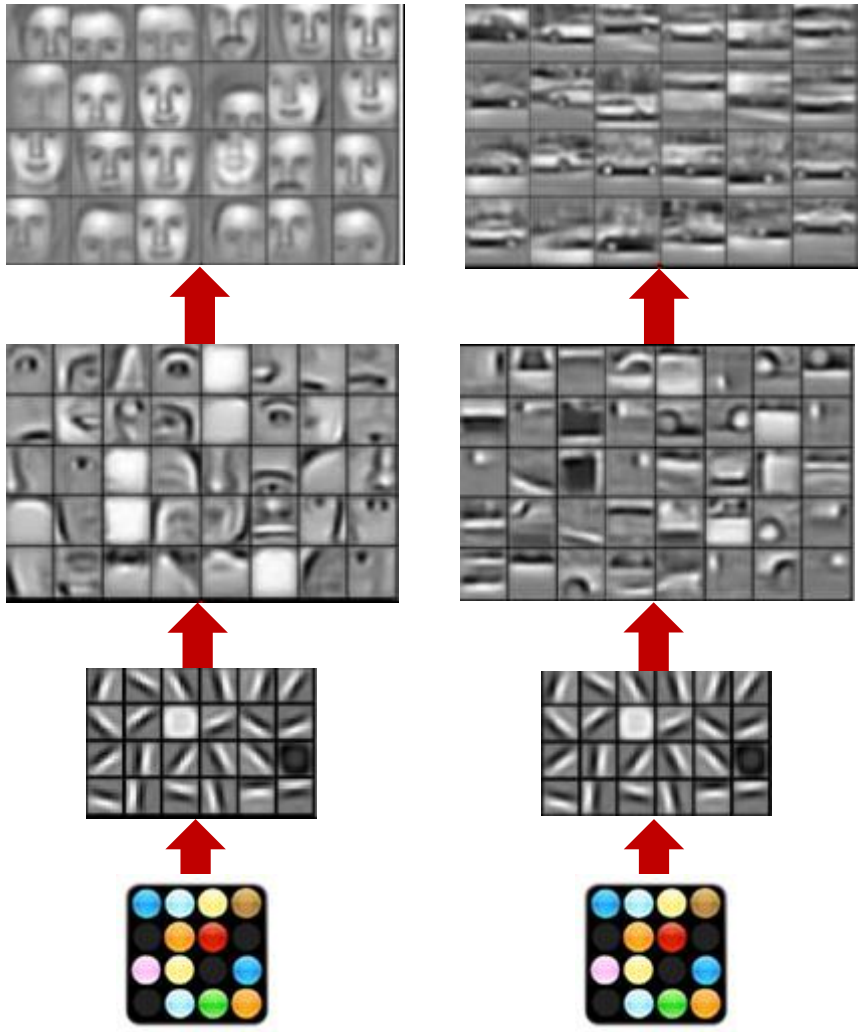
分层目标表示学习





Face 

Car 



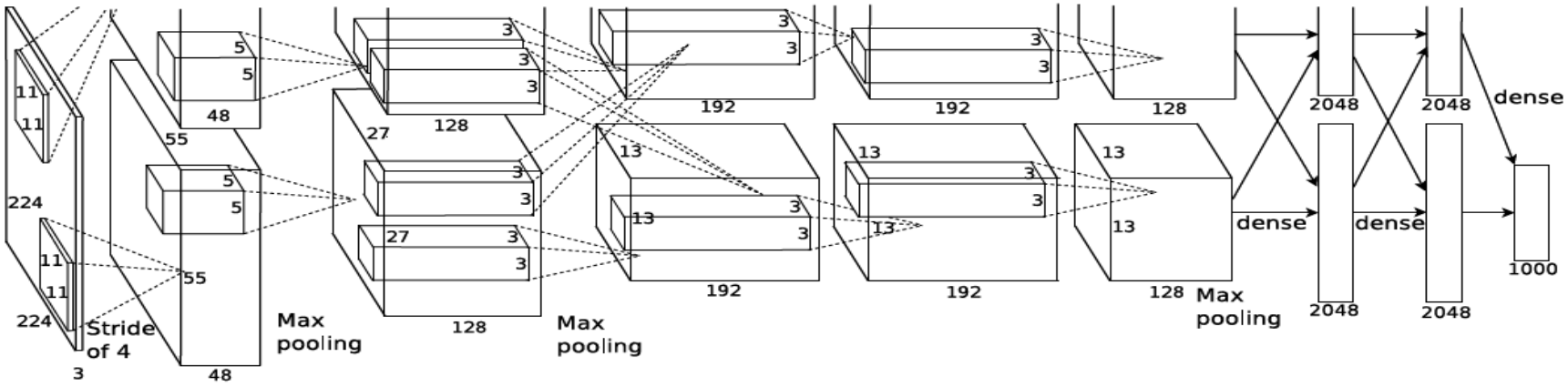
[Lee et al., ICML 2009]

2: “CNN” 阶段 (2012–)

CNN: convolutional neural network

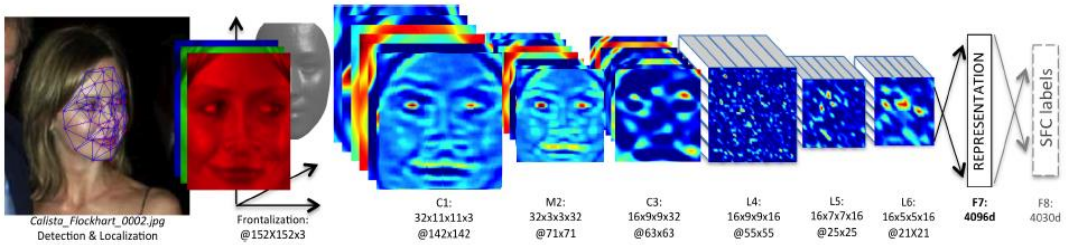
ImageNet竞赛

- Traditional model
  74% 2011
- CNN based deep model
  85% 2012
  89% 2013
  92% 2014

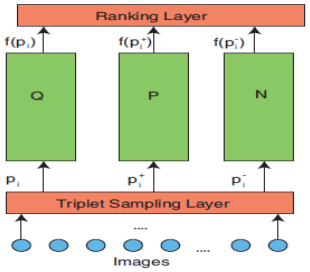


CNN→视觉应用

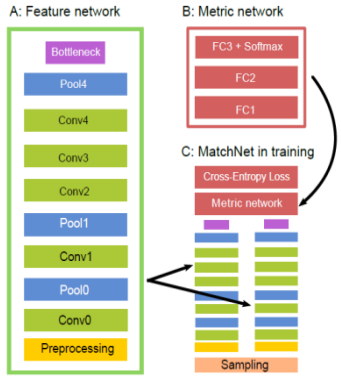
Learn discriminate, task-oriented features



DeepFace, CVPR2014

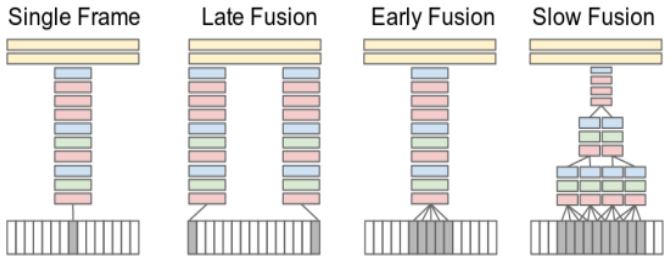


DeepRanking, CVPR2014

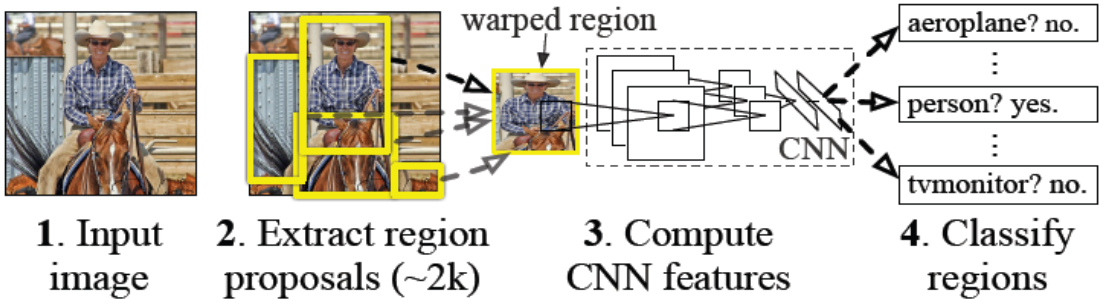


DeepMatching, CVPR2015

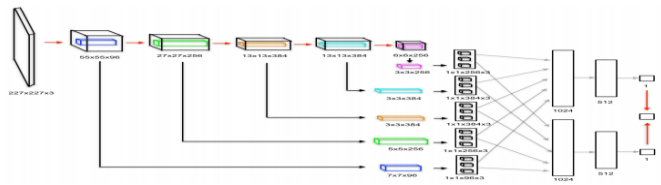
DeepPose, CVPR2014



DeepVideo, CVPR2014

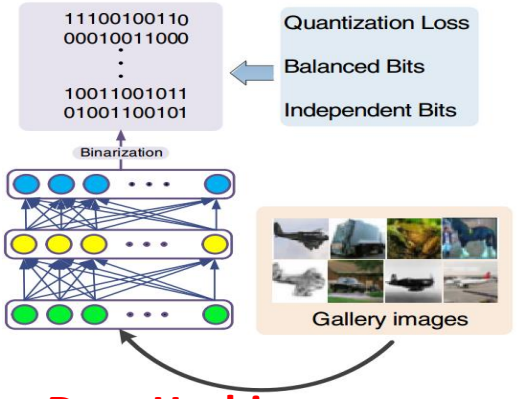


RCNN for detection, CVPR2014

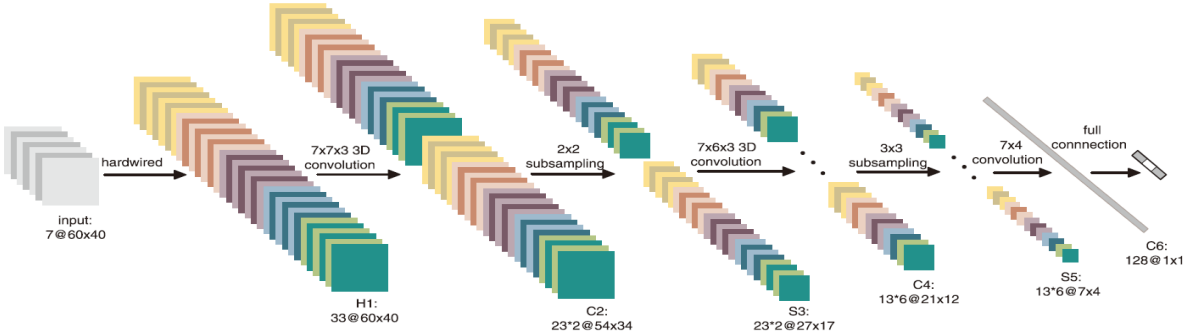


DeepEdge, CVPR2015

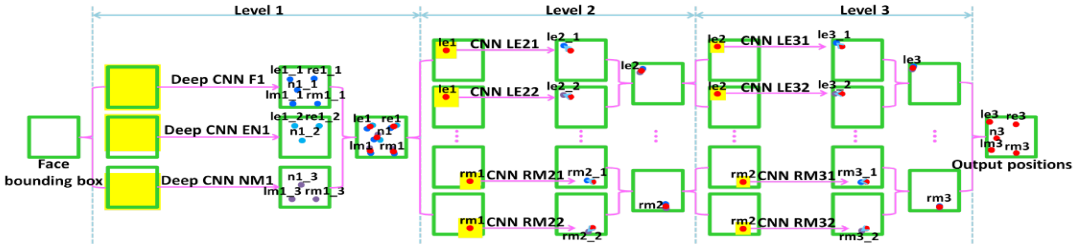
CNN→视觉应用



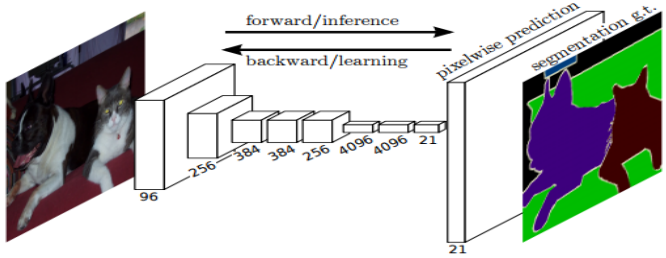
DeepHashing, CVPR2015



3D CNN for action recognition, TPAMI2013



DeepLandmark, CVPR2013



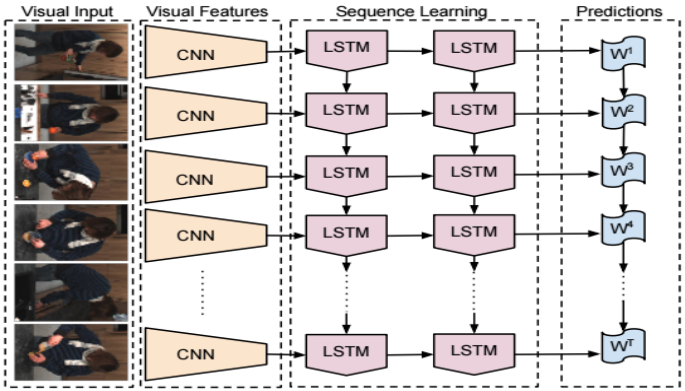
DeepSegmentation, CVPR2015

- **模型特点：**判别式模型，大规模数据集，深层网络，并行分布式计算
- **热点问题：**侧重于处理静态图像相关的各种任务，并刷新大量state-of-the-art结果

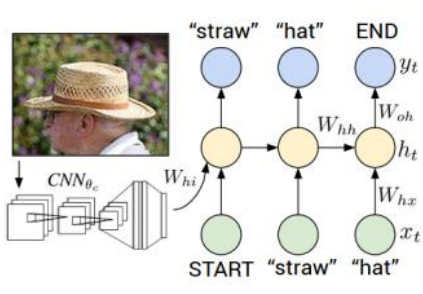
3：“RNN”阶段(2014–)

RNN: recurrent neural network

Model dependency relation in sequence data



Activity recognition, CVPR2015



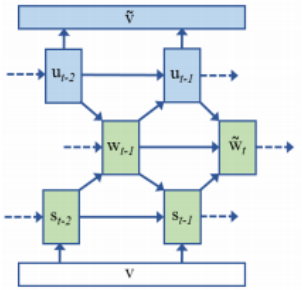
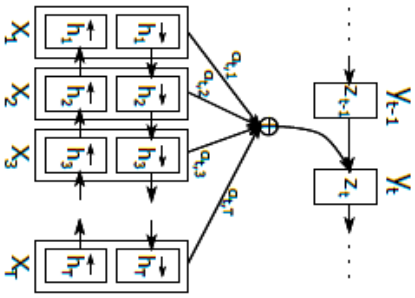
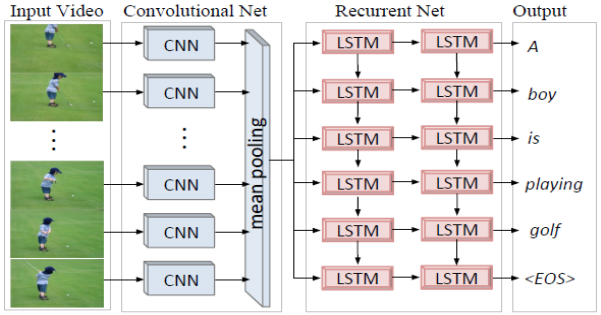
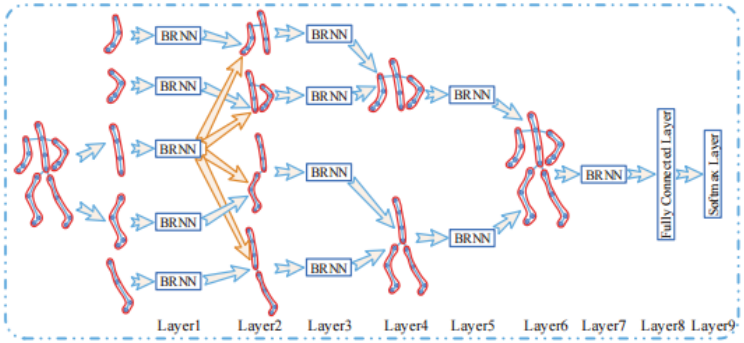


Image caption, CVPR2015&ICML2015

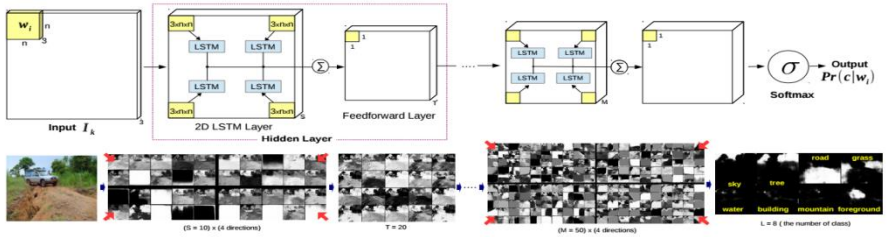




Video caption, CVPR2015

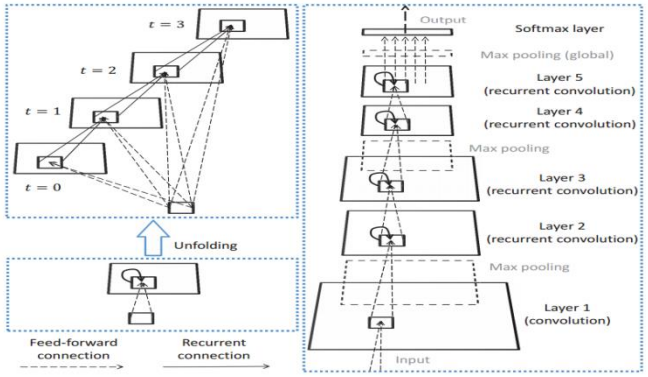


Action recognition, CVPR2015

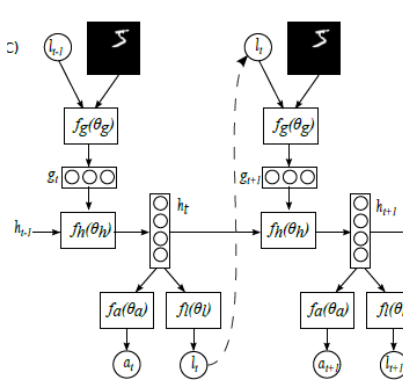


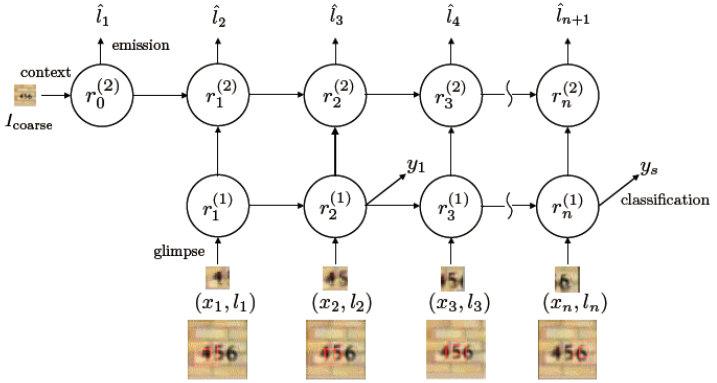
Scene labeling, CVPR2015

RNN→视觉应用

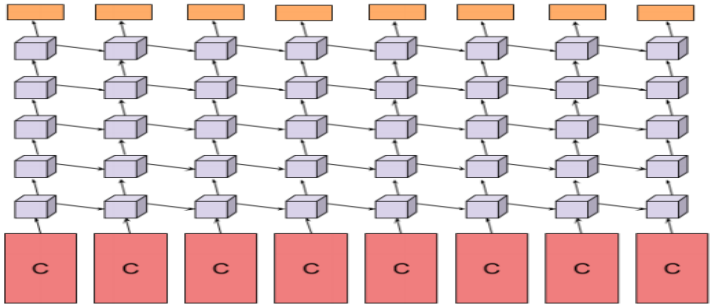


Object recognition, CVPR2015





Visual attention, NIPS2014&ICLR2015



Video classification, CVPR2015

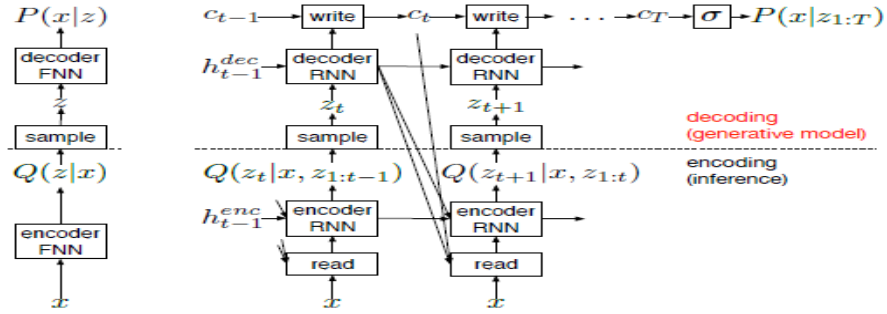


Image generation, ICML2015

- **模型特点：** 时序模型，模型结构设计，优化算法
- **热点问题：** 对序列数据中的（时间）相依关系进行建模

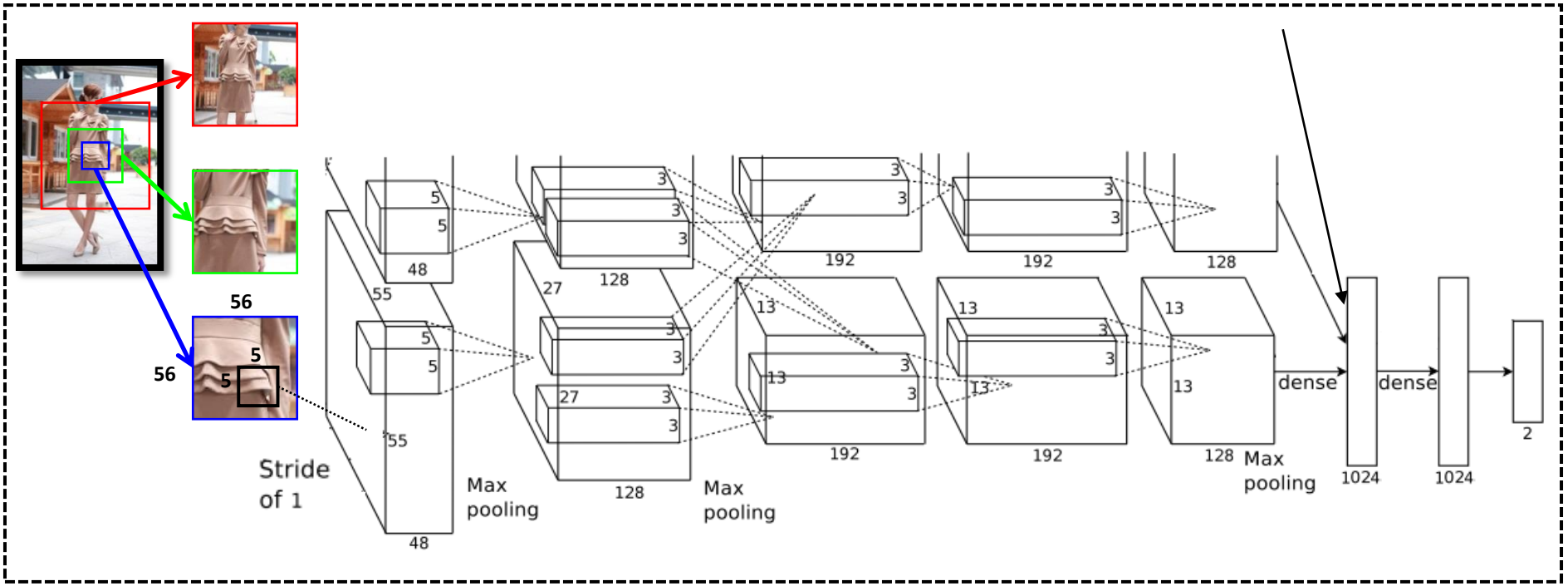
1. 物体分割、检测与识别

目标图像分割（百度竞赛冠军2013）

- 在图像或视频中，把用户指定的目标分割出来
- 展示：人形图像分割（也可以是任意目标）




解决方案



- Hierarchical contexts
  - context patches on different scales are bound together at the beginning

人形分割

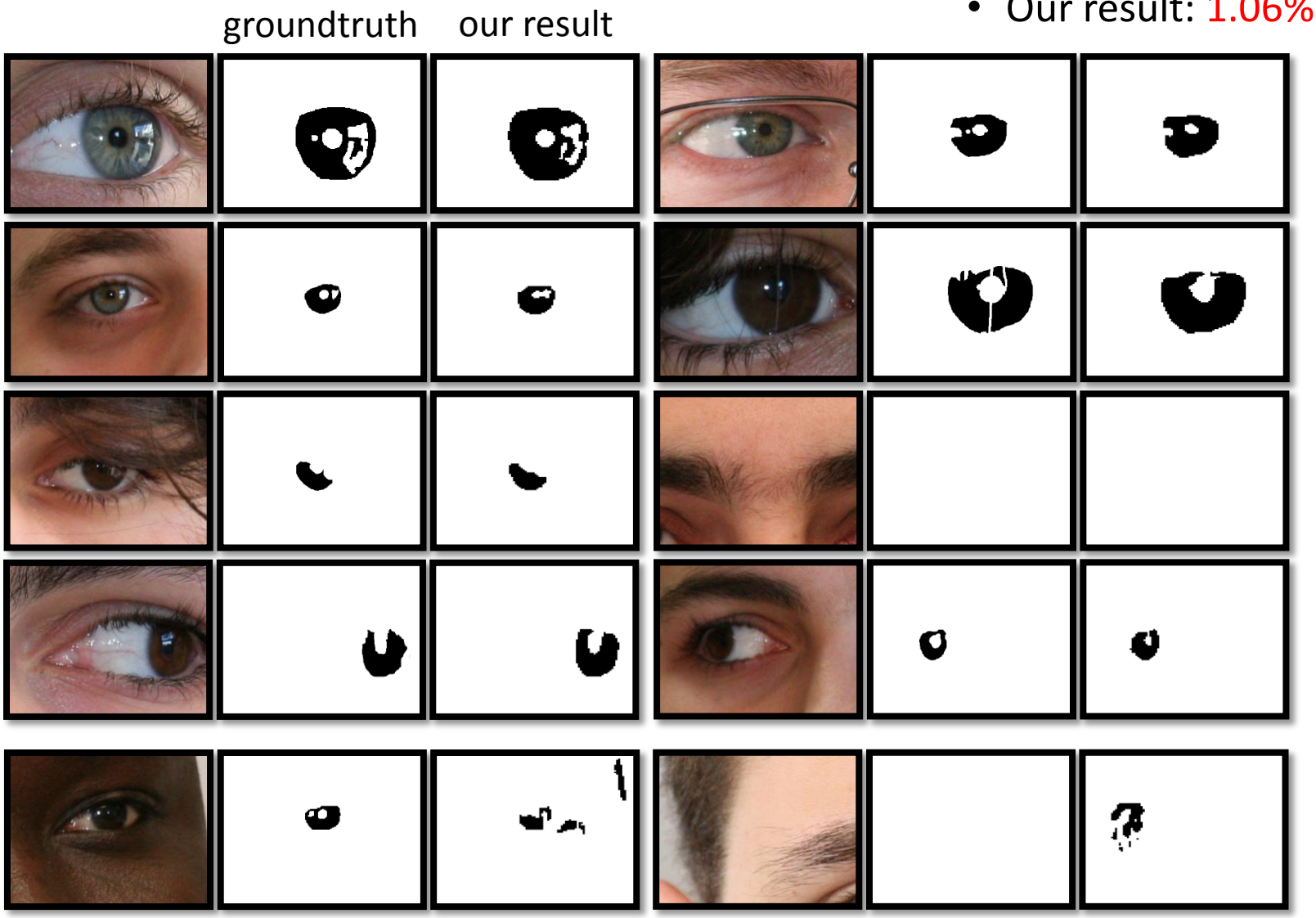
人形分割

- 以绝对优势获得中国云移动互联网人形图像分割大赛冠军，以及特别奖
- 国内最高的人形图像分割水平和国际先进技术

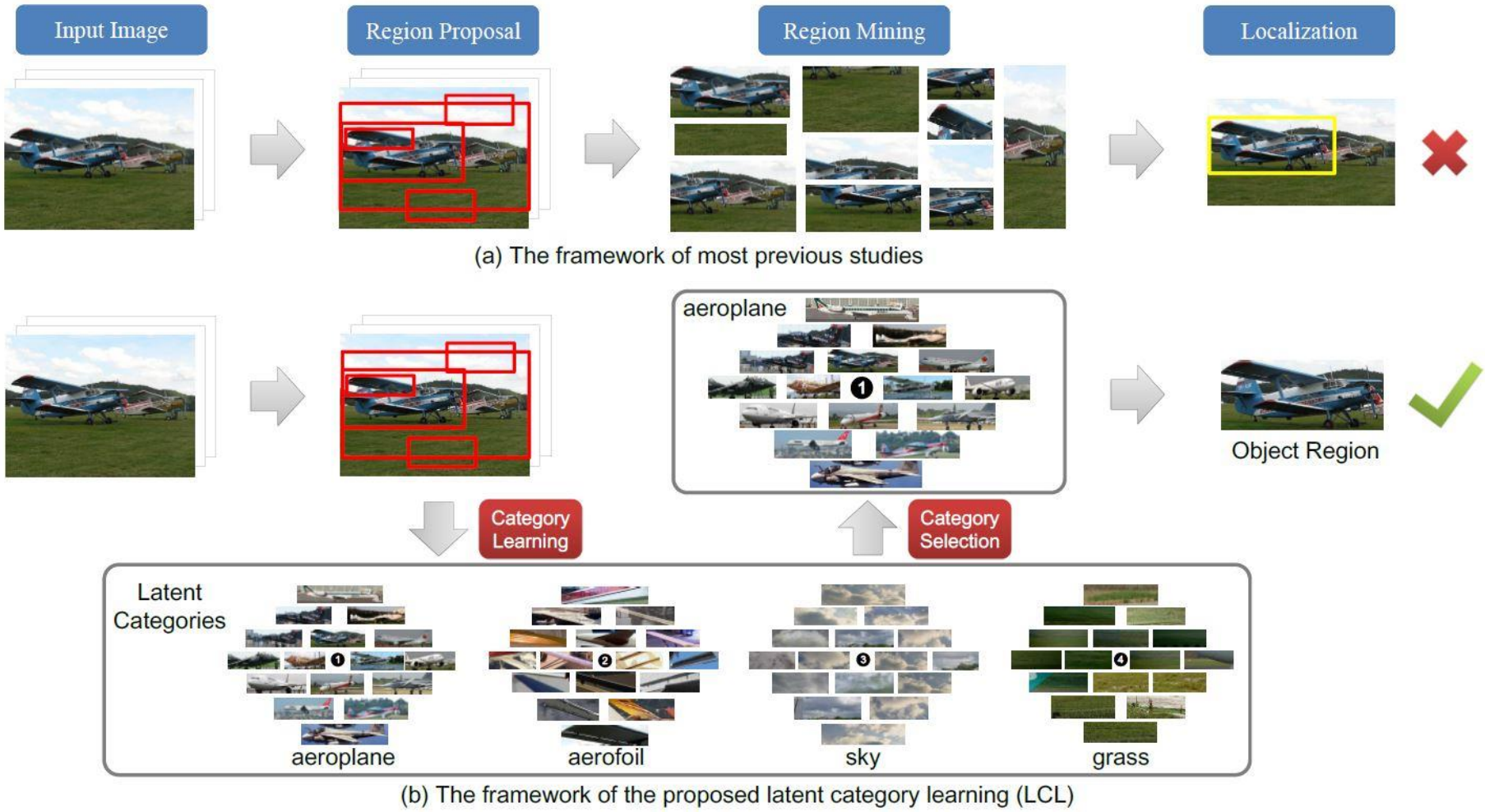
Rank	Name	Score	Publish Time
1	NLPR&CRIPAC	0.8683	Tue Oct 15 CST 2013
2	Freedom	0.7817	Tue Oct 15 CST 2013
3	WEESEE	0.7600	Tue Oct 15 CST 2013
4	CASIA_IGIT	0.7595	Tue Oct 15 CST 2013
5	DT_ppseg	0.7587	Tue Oct 15 CST 2013
6	FlyHigh	0.7328	Tue Oct 15 CST 2013
7	EagleEye	0.7213	Tue Oct 15 CST 2013
8	sysu_vision	0.7167	Tue Oct 15 CST 2013
9	DeepLearner	0.6111	Tue Oct 15 CST 2013
10	RandomForest	0.5117	Tue Oct 15 CST 2013

虹膜分割

- Dataset: NICE1
- State-of-the-art: 1.3%
- Our result: 1.06%



弱监督目标检测（ECCV2014）



Reduce background influence to obtain clean object region


(a) Training Images
aeroplane
bicycle
bird
cat
tvmonitor
(b) Candidate Regions
category learning on aeroplane
(c) Category Learning
aeroplane
aerofoil
sky
grass
negative regions
selection model of category1
(d) Category Selection
MAX Precision
positive images
BoW
negative images
BoW
Positive Training Set
(e) Discrimination Evaluation
classification model of category1

检测准确度

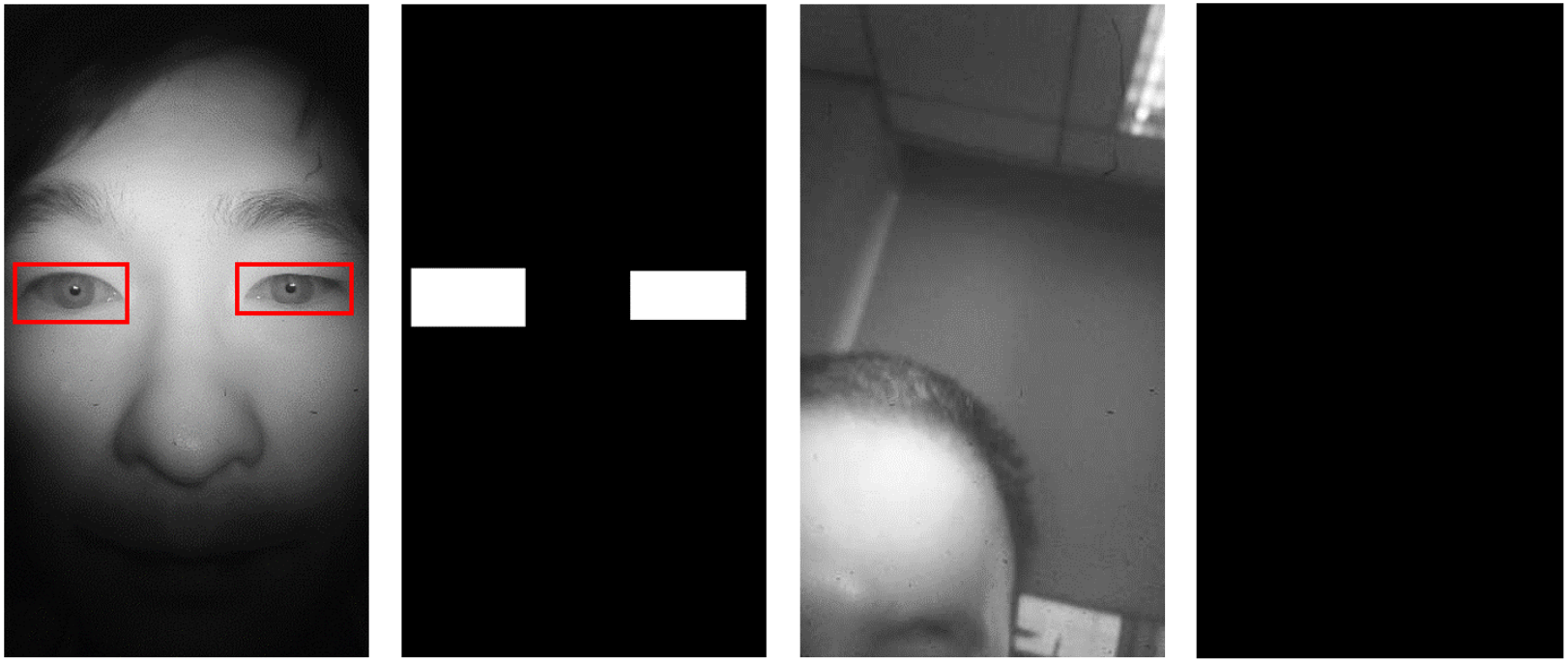
Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Drift-Detect [8]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3
Object-Centric [9]	-	-	-	-	-	-	-	-	-	-
Multifold MIL [26]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0
Latent SVM [27]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9
LCL-kmeans	41.5	29.7	24.9	12.0	10.7	30.3	40.9	31.8	10.5	21.8
LCL-pLSA	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6
DPM 5.0 [15]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1
CNN Supervise [38]	61.8	62.0	38.8	35.7	29.4	52.5	61.9	53.9	22.6	49.7

Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mAP**
Drift-Detect [8]	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
Object-Centric [9]	-	-	-	-	-	-	-	-	-	-	15.0
Multifold MIL [26]	4.9	14.1	31.9	41.9	19.3	11.1	27.6	12.1	31.0	40.6	22.4
Latent SVM [27]	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
LCL-kmeans	15.4	29.4	24.3	37.8	19.1	14.7	33.1	24.1	36.2	43.0	**26.6**
LCL-pLSA	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	**30.9**
DPM 5.0 [15]	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
CNN Supervise [38]	40.5	48.8	49.9	57.3	44.5	28.5	50.4	40.2	54.3	61.2	47.6

8% improvement on detection rate, and competitive with the DPM 5.0

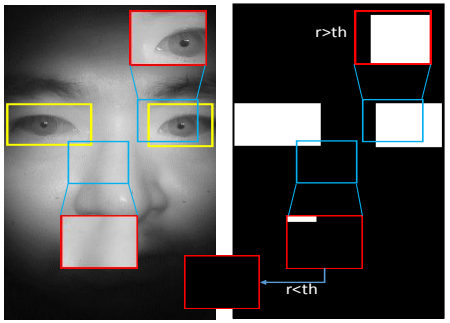
快速高精度人眼检测（CCCV2015最佳学生论文）

- 在人脸红外图像中，对人眼位置进行快速高精度快速定位
- 挑战：在普通PC下速度>100fps，精度>85%

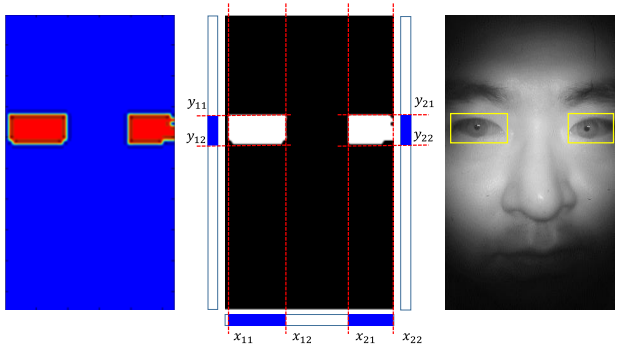


我们的模型

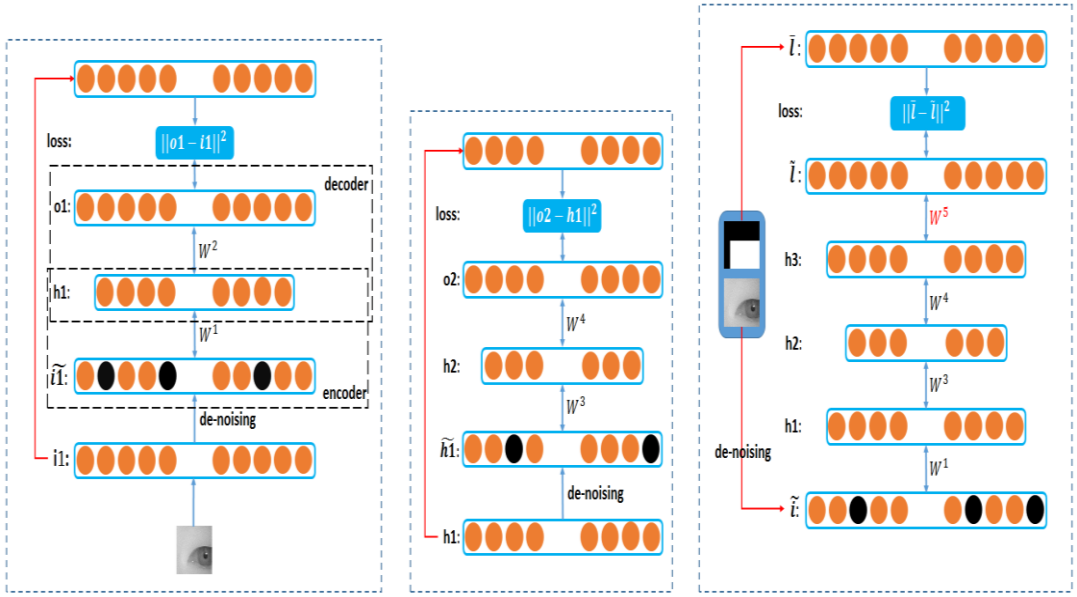
- 利用stacked auto-encoder学习输入图像到label map的映射
- 然后从label map中提取人眼的位置



Patch based label map

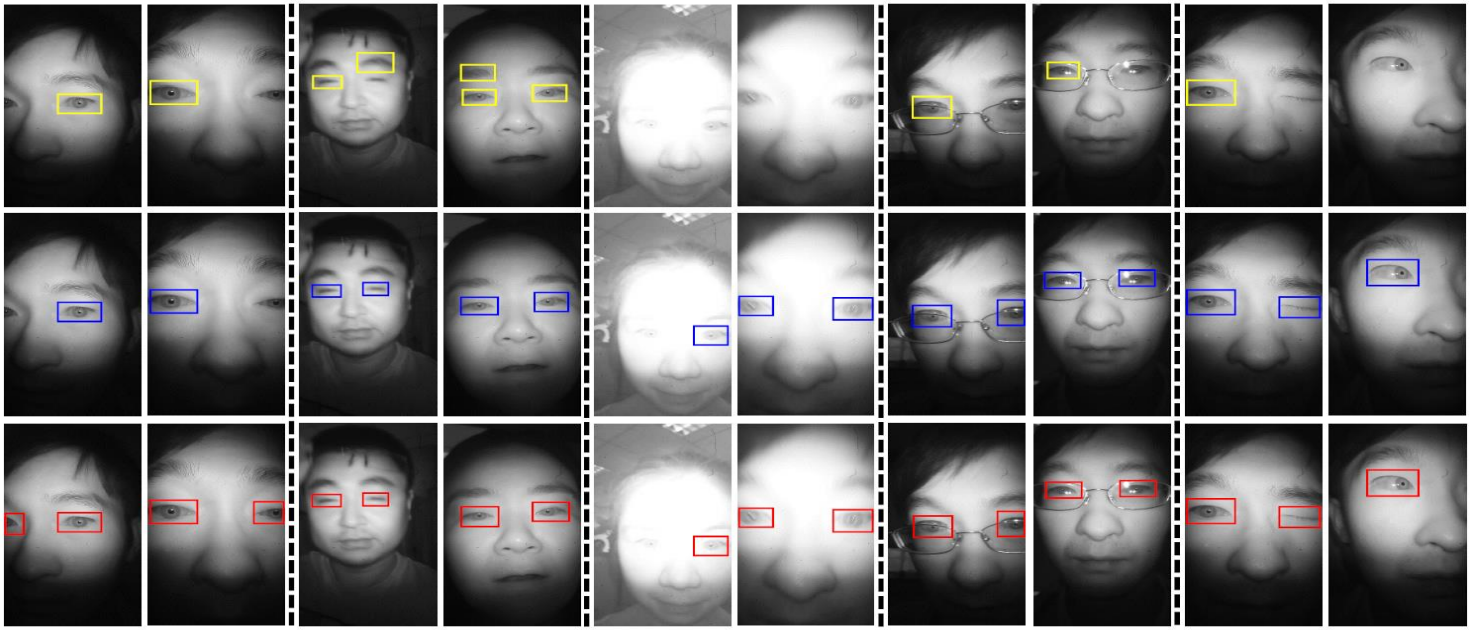


Bounding boxes acquisition



Layer wise pre-training and stacked auto-encoder fine-tuning

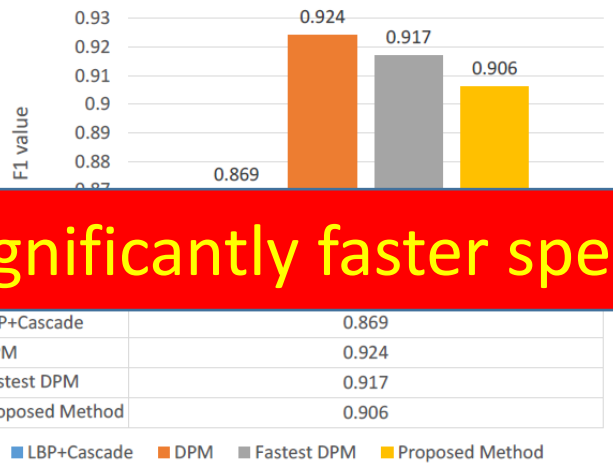
结果展示



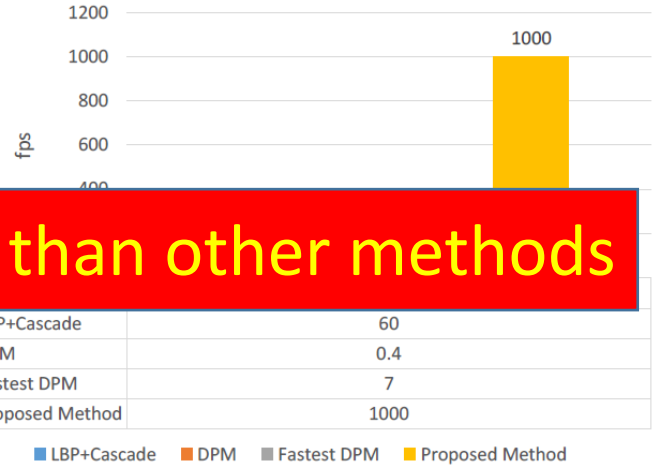
LBP+Cascade

Our

DPM



LBP+Cascade	0.869
DPM	0.924
Fastest DPM	0.917
Proposed Method	0.906



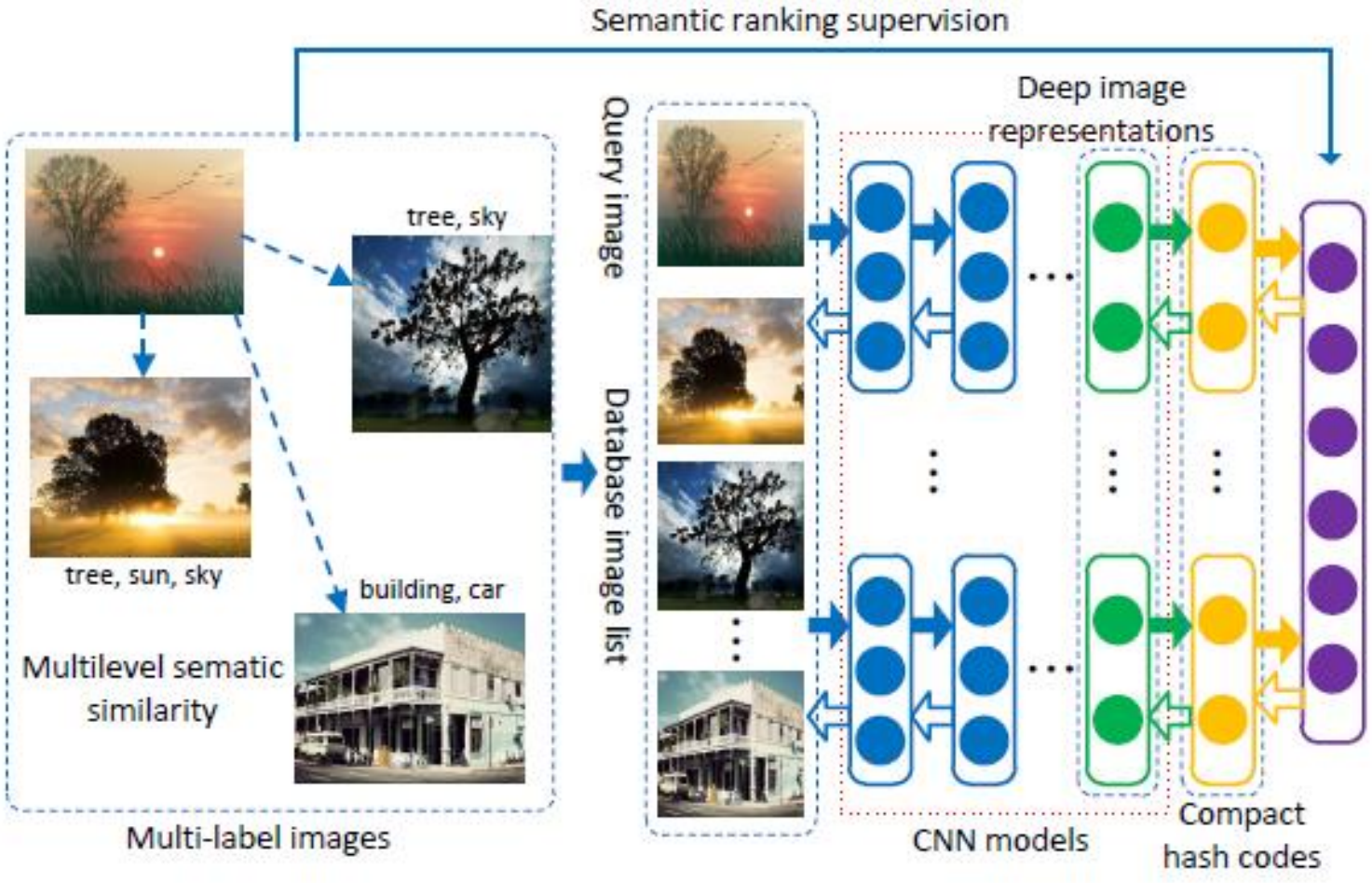
LBP+Cascade	60
DPM	0.4
Fastest DPM	7
Proposed Method	1000

Significantly faster speed than other methods

2. 多标签图像检索

深度语义检索（CVPR2015）

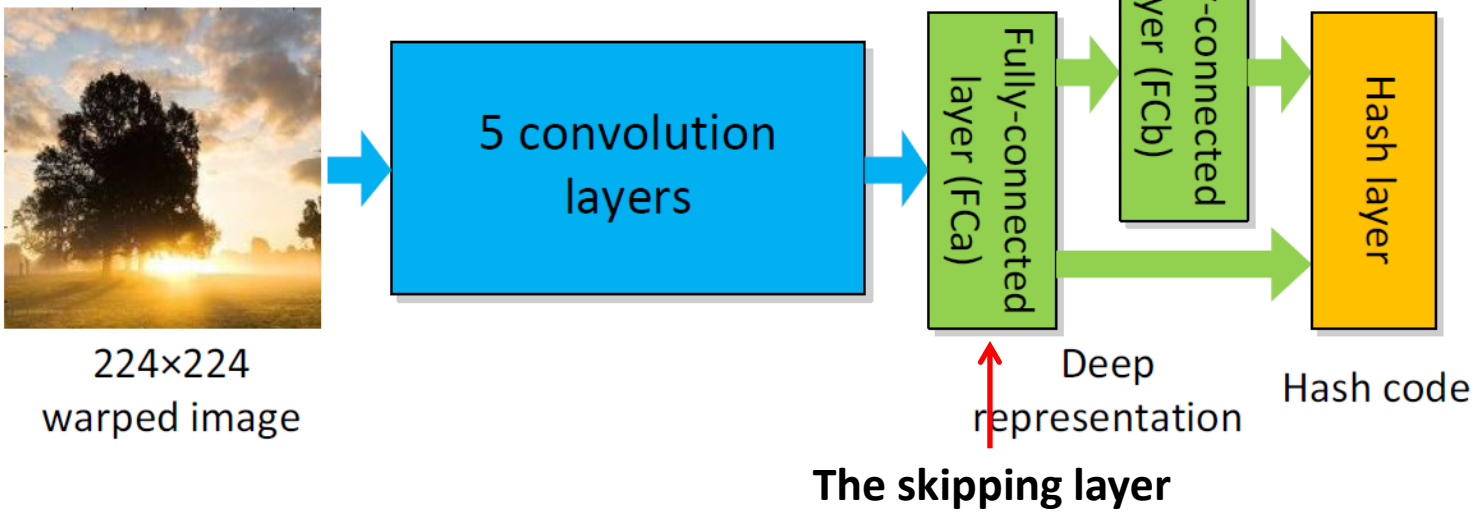
- Deep Semantic Ranking Hashing for Multi-Label Image Retrieval



目标函数

- Deep hash functions:

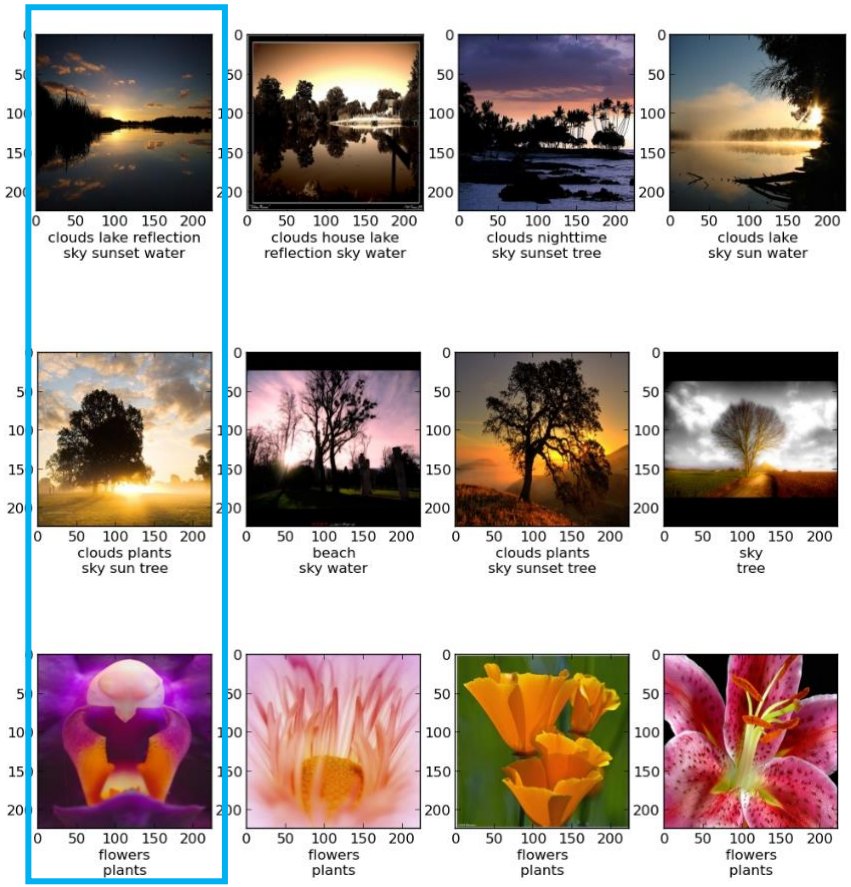
$$h(\mathbf{x};\mathbf{w}) = sign(\mathbf{w}^{\mathrm{T}}[\, f_a(\mathbf{x});\, f_b(\mathbf{x})])$$



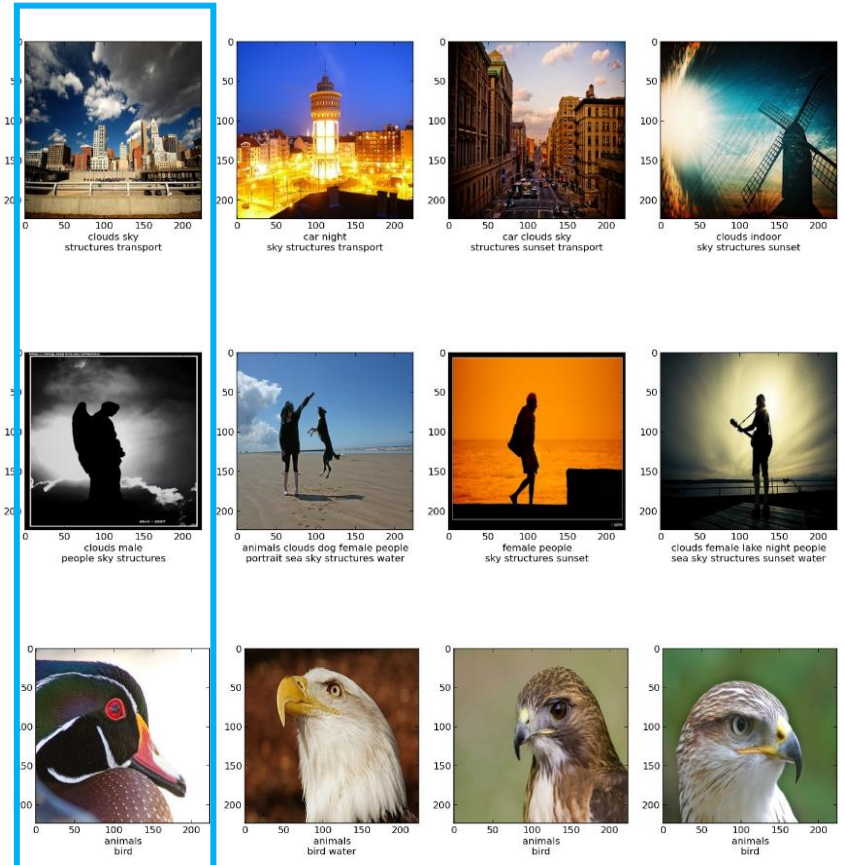
Utilizing more feature information biased toward visual appearance.

图片检索结果

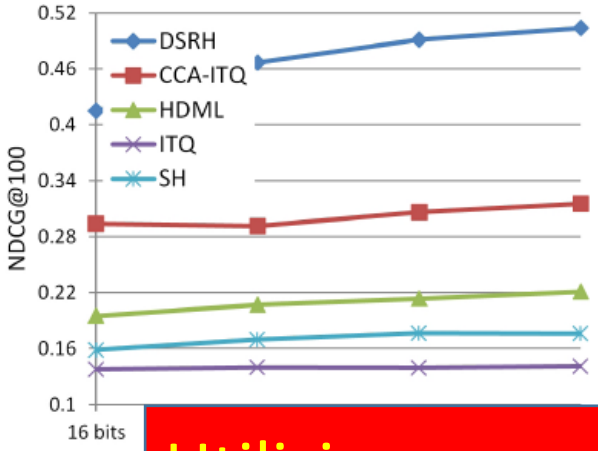
- 在MIRFLICKR-25K 和 NUS-WIDE 数据库上的检索样例

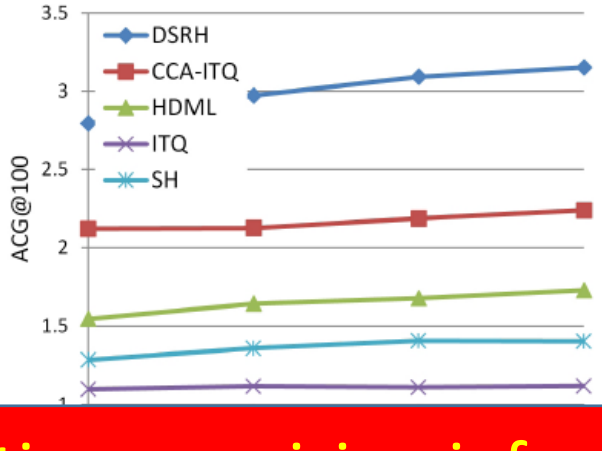


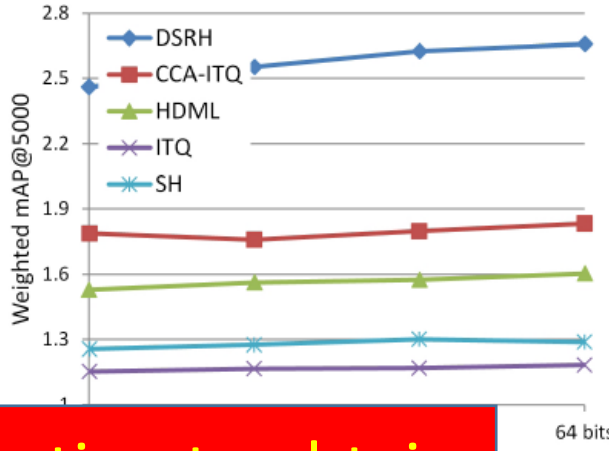
Query



Query

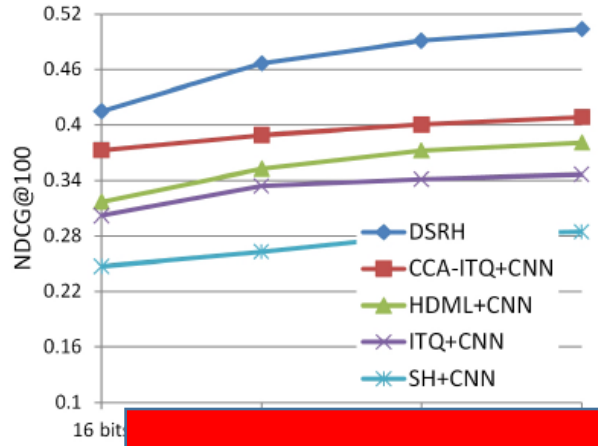


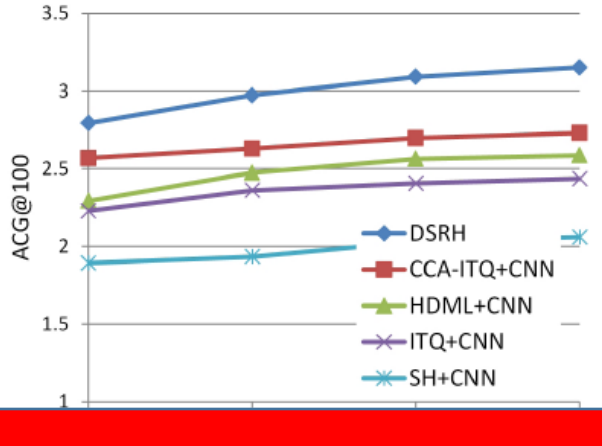


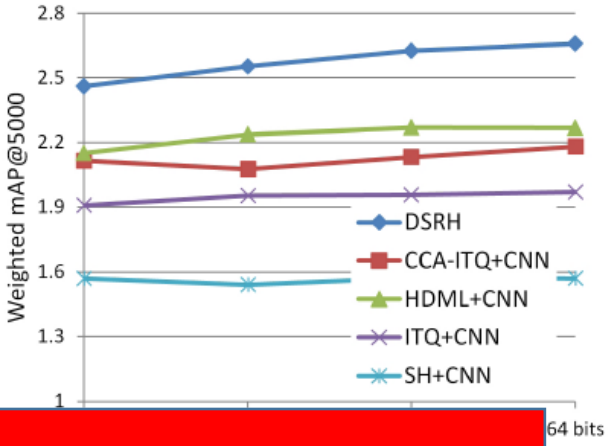


Utilizing semantic supervision information to obtain features can achieve better performance

Figure 5. R CKR-25K dataset usi





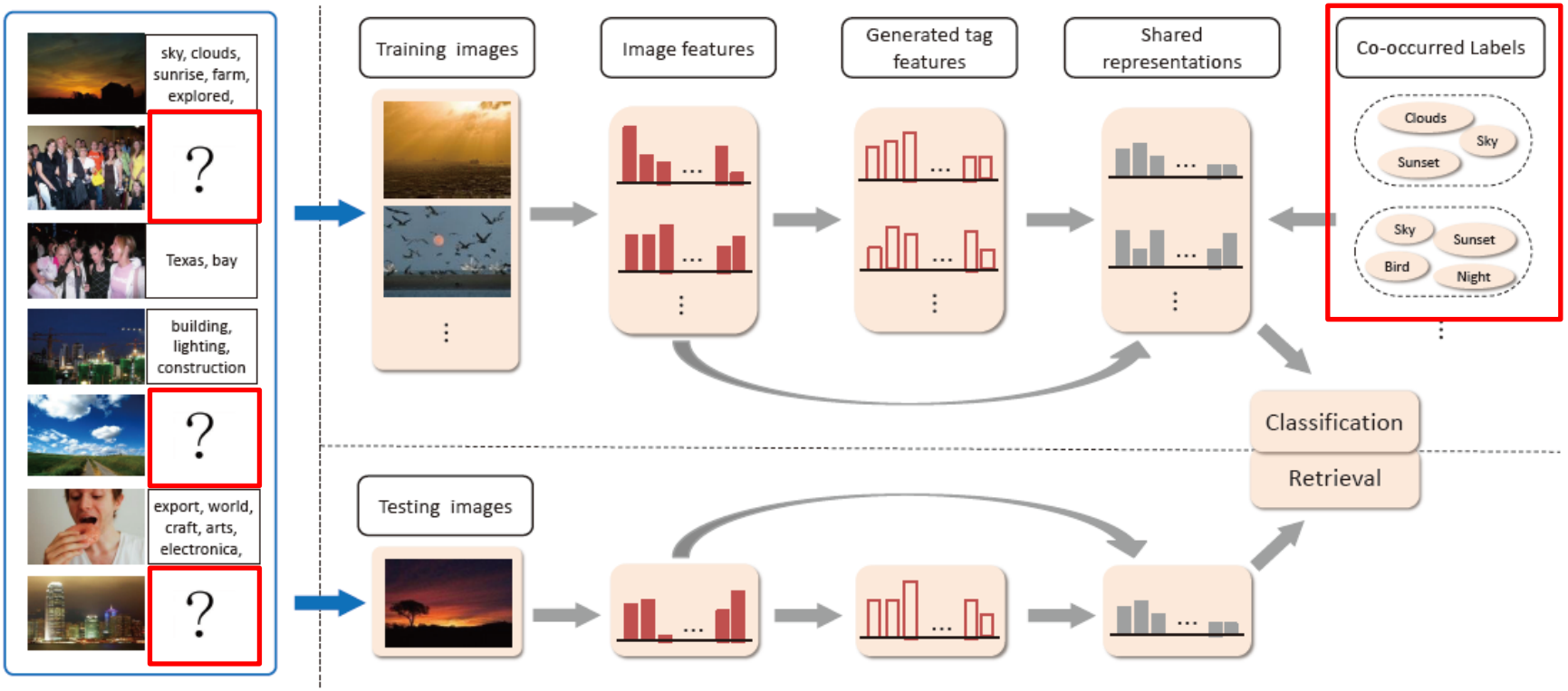


Multilevel semantic ranking supervision can better preserve the semantic structure of multi-label images

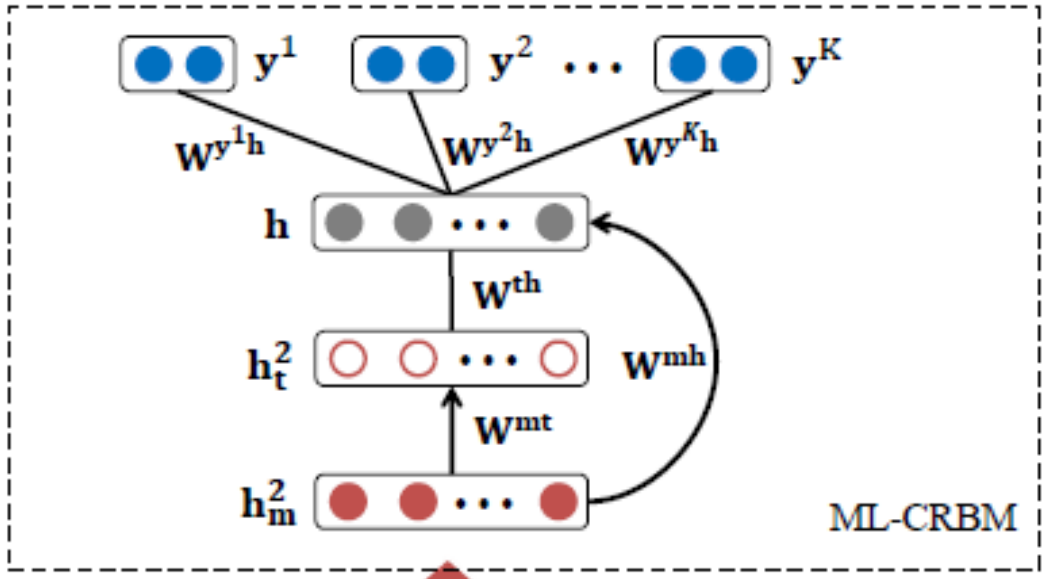
Figure 6. n the MIRFLI

多标签图像文本的分类与检索（TMM2015）

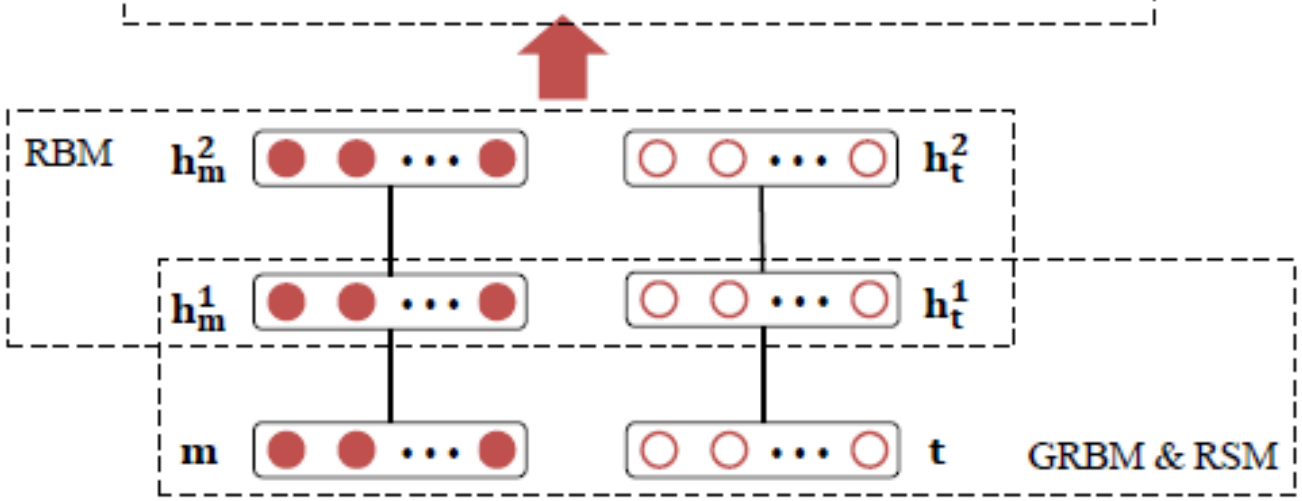
- 对多标签的多模态数据进行表示学习，用于分类检索任务
- 考虑了类别缺失问题和类别标签的共生关系



多标签条件玻尔兹曼机

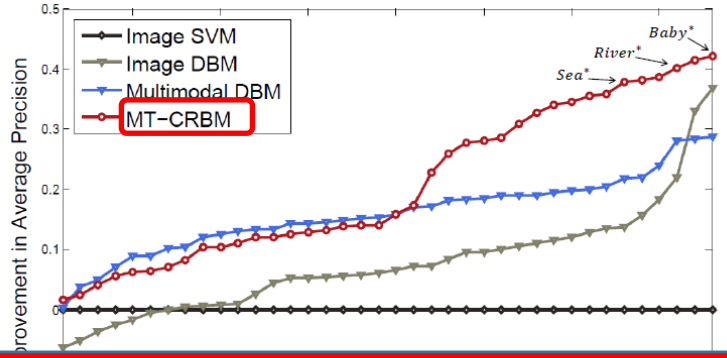


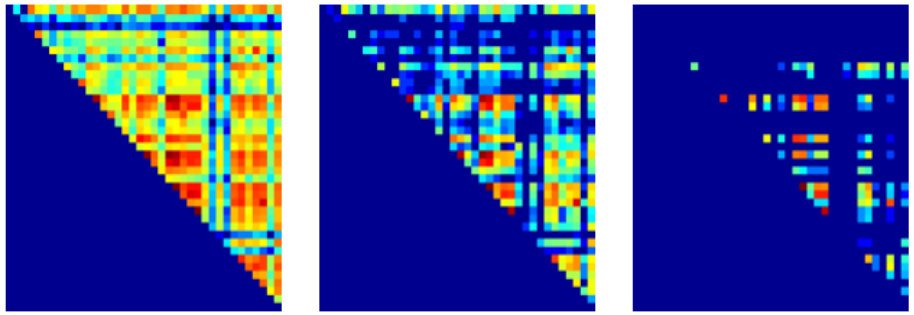
有监督的多模态表示学习



对于每个模态，提取相应的特征表示

检索与分类结果





Greatly promote the precisions of some classes which have fewer training samples

Make fewer mistakes on the prediction of the label co-occurrence

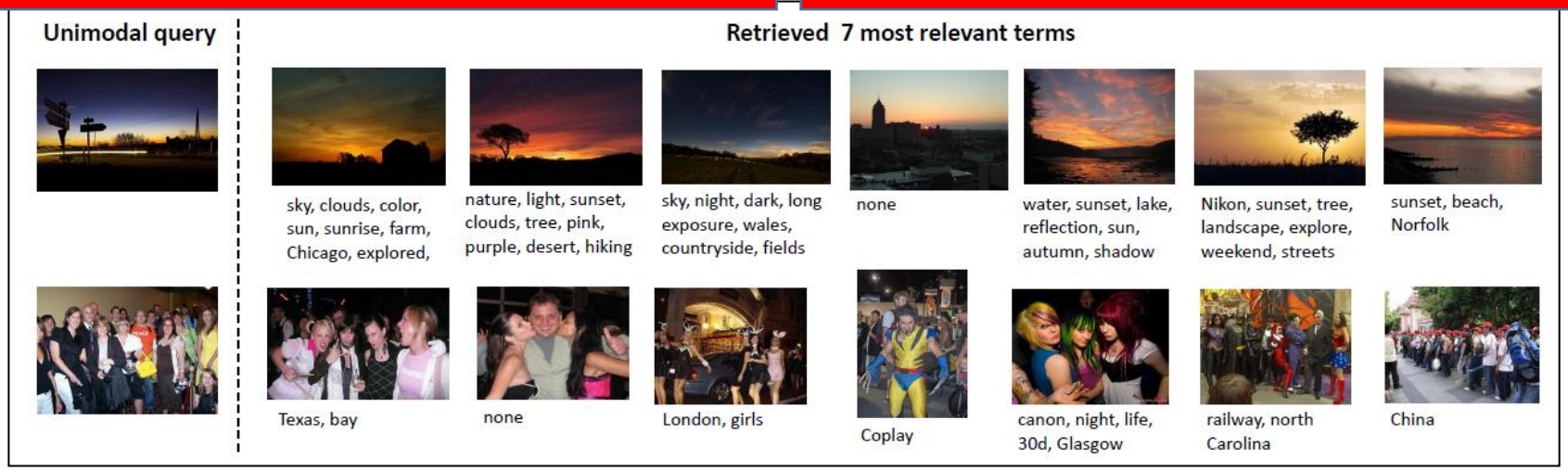
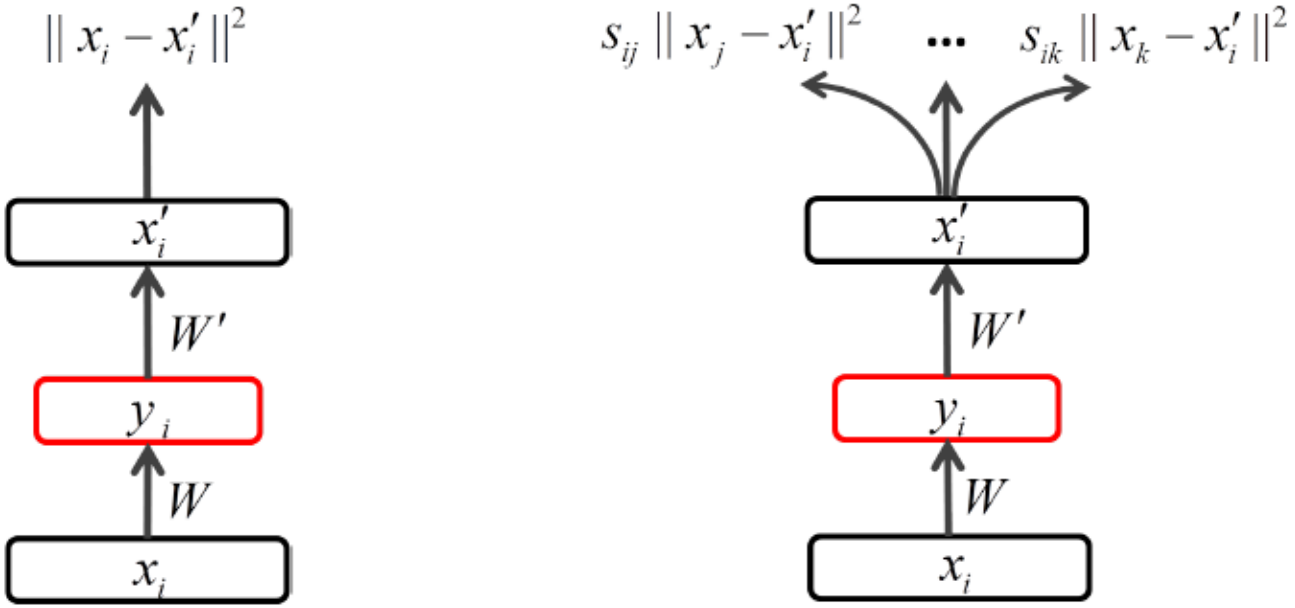


Figure 3. Unconstrained multimodal multi-label retrieval on the MIR Flickr dataset.

3. 数据关系学习

广义自编码机（DeepVision2014最佳论文奖）

为了保持学习到数据表示的局部结构关系，在重构过程中考虑了数据之间的相似性关系



Original Autoencoder　　Generalized Autoencoder

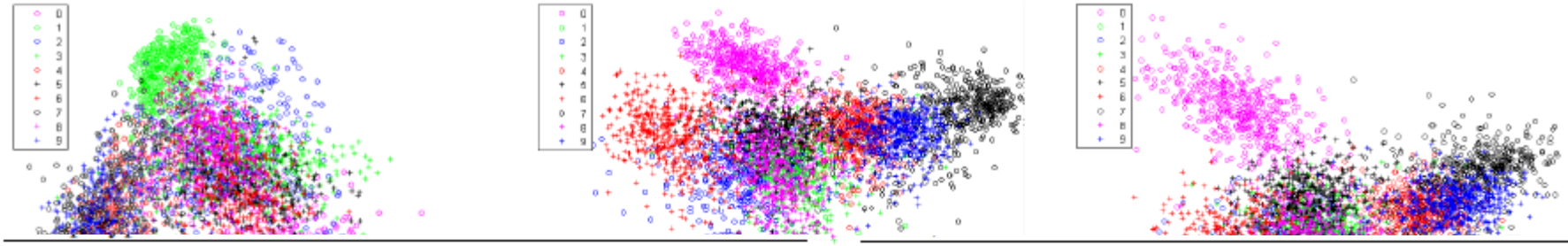
$$E(W, W') = \sum_{i=1}^{n} e_i(W, W') = \sum_{i=1}^{n} \sum_{j \in \Omega_i} s_{ij} L(x_j, x_i')$$

统一降维框架

	$\| x_i - x_i' \|^2$	$s_{ij} \| x_j - x_i' \|^2 \quad \cdots \quad s_{ik} \| x_k - x_i' \|^2$
Method	Reconstruction Set	Reconstruction Weight
GAE-PCA	$j = i$	$s_{ij} = 1$
GAE-LDA	$j \in \Omega_{c_i}$	$s_{ij} = \frac{1}{n_{c_i}}$
GAE-ISOMAP	$j : x_j \in X$	$s_{ij} \in S = -H\Lambda H/2$
GAE-LLE	$j \in N_k(i)$,	$s_{ij} = (M + M^T - M^T M)_{ij}$ if $i \neq j$;
	$j \in (N_k(m) \cup m), j \neq i$ if $\forall m, i \in N_k(m)$	0 otherwise
GAE-LE	$j \in N_k(i)$	$s_{ij} = exp\{-\|\|x_i - x_j\|\|^2/t\}$
GAE-MFA	$j \in \Omega_{k_1}(c_i)$,	$s_{ij} = 1$
	$j \in \Omega_{k_2}(\bar{c}_i)$	$s_{ij} = -1$

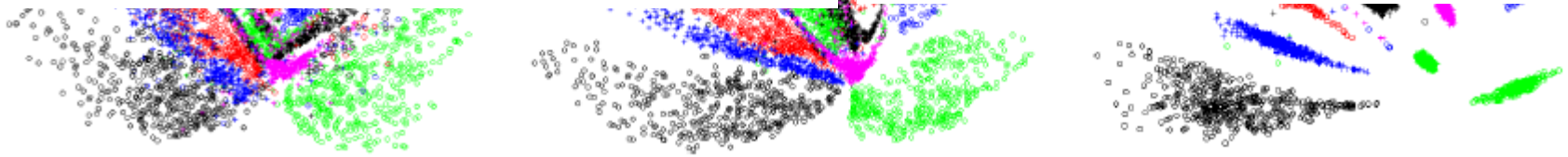
$$E(W, W') = \sum_{i=1}^{n} e_i(W, W') = \sum_{i=1}^{n} \sum_{j \in \Omega_i} s_{ij} L(x_j, x_i')$$

结果展示



Method	ER	Our Model	ER
PCA	20.6% (150)	dGAE-PCA	3.5%
Kernel PCA	8.1% (g)		
LDA	5.7% (67)	dGAE-LDA	1.2%
Kernel LDA	1.6% (pp)		
ISOMAP	–	dGAE-ISOMAP	2.5%
LLE	–	dGAE-LLE	3.6%
LPP	4.6%(110)	dGAE-LE	1.1%
Kernel LPP	1.7% (pp)		
MFA	2.6% (85)	dGAE-MFA	1.1%
Kernel MFA	2.1% (pp)		

Method	ER	Our Model	ER
PCA	6.2% (55)	dGAE-PCA	5.3%
Kernel PCA	8.5% (pp)		
LDA	16.1% (9)	dGAE-LDA	4.4%
Kernel LDA	4.6% (pp)		
ISOMAP	–	dGAE-ISOMAP	6.4%
LLE	–	dGAE-LLE	5.7%
LPP	7.9%(55)	dGAE-LE	4.3%
Kernel LPP	4.9% (pp)		
MFA	9.5% (45)	dGAE-MFA	3.9%
Kernel MFA	6.8% (pp)		



(d) dGAE-LE (e) dGAE-MFA (f) dGAE-LDA

深度聚类（ICPR2014最佳学生论文）

- 利用深度嵌入网络学习适用于聚类的表示

Objective function:

min

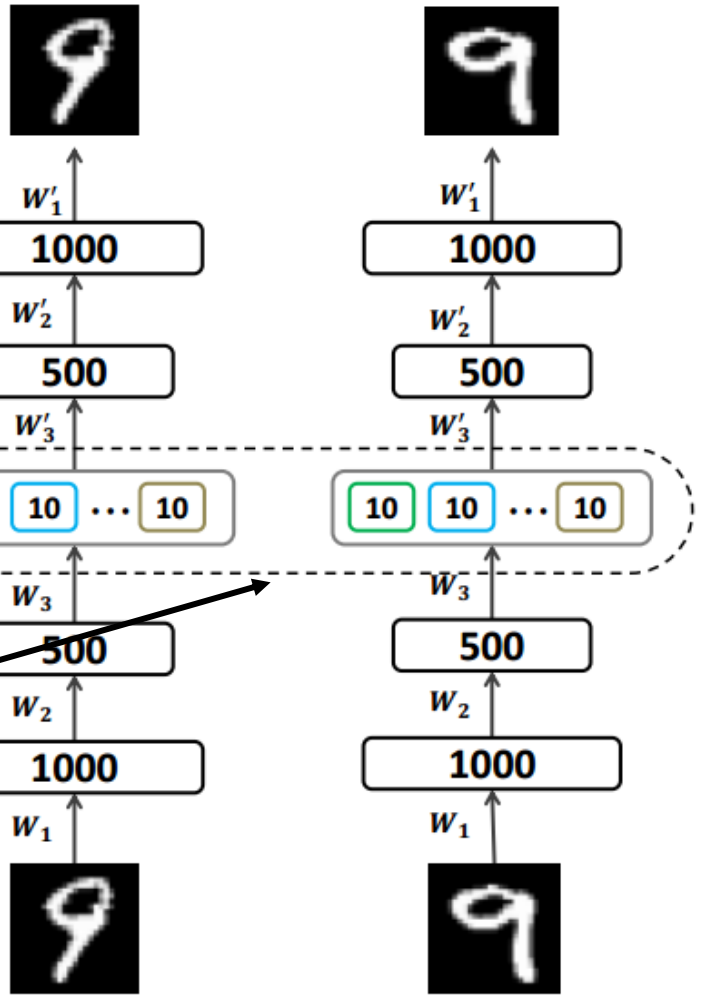
Reconstruction error

+

Group sparsity

+

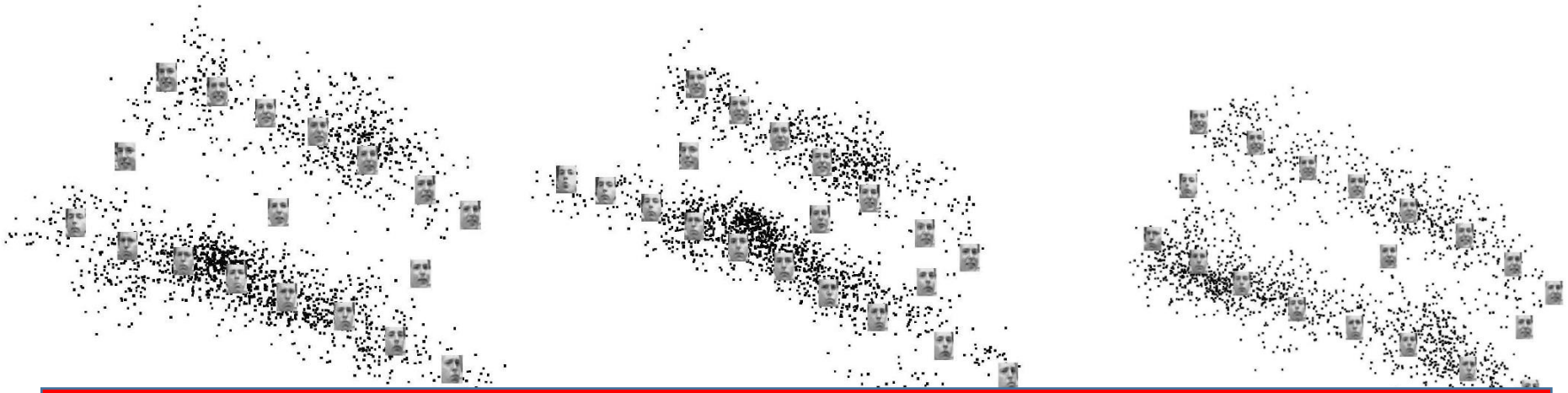
Locality-preserving



聚类结果

Table 1. Comparsion of clustering methods on three datasets.

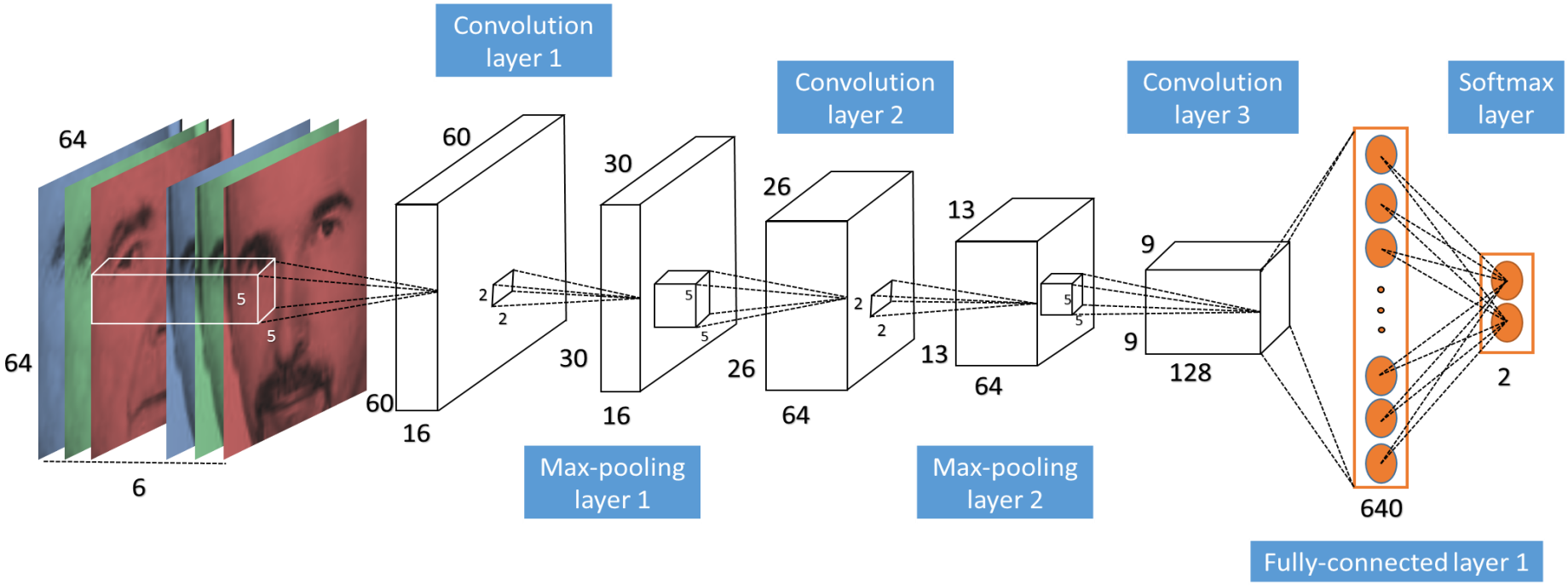
Method	COIL20		Yale-B		PIE	
	NMI	ACC (%)	NMI	ACC (%)	NMI	ACC (%)
K-means	0.76±0.02	59.02±4.40	0.70±0.04	65.80±6.18	0.19±0.00	8.36±0.40
Ncut	0.86±0.03	64.57±6.87	0.79±0.03	62.34±6.37	0.25±0.01	9.55±0.65
Proposed	**0.87±0.01**	**72.40±3.39**	**0.92±0.04**	**81.73±9.64**	**0.42±0.00**	**11.19±0.29**



Among each cluster, the facial expressions and viewpoints are varying smoothly along the tube-like manifold

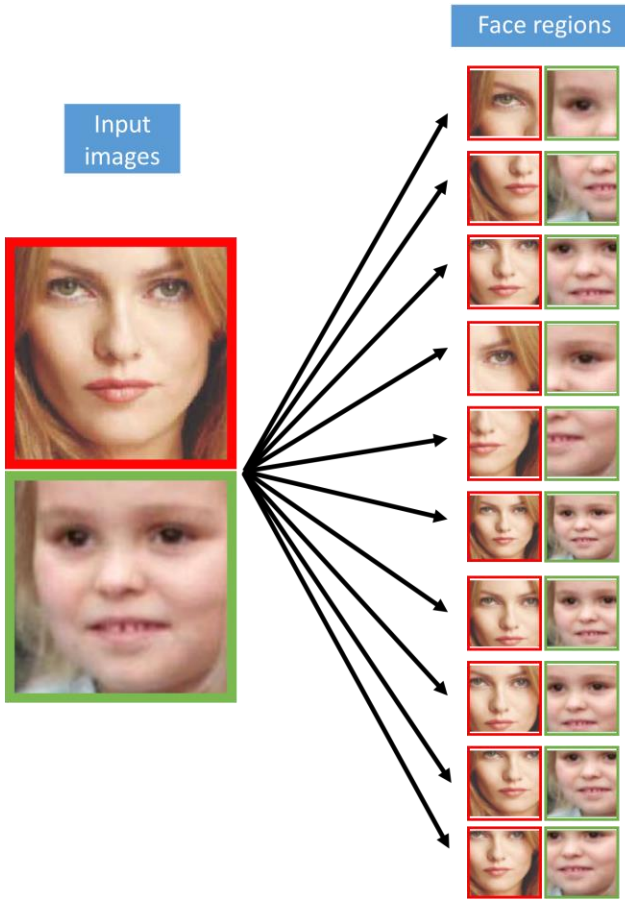
亲属关系识别（BMVC2015）

- 通过成对人脸图片来判断亲属关系
- 模型的输入为一对待判断某种关系的照片，输出对亲属关系的预测



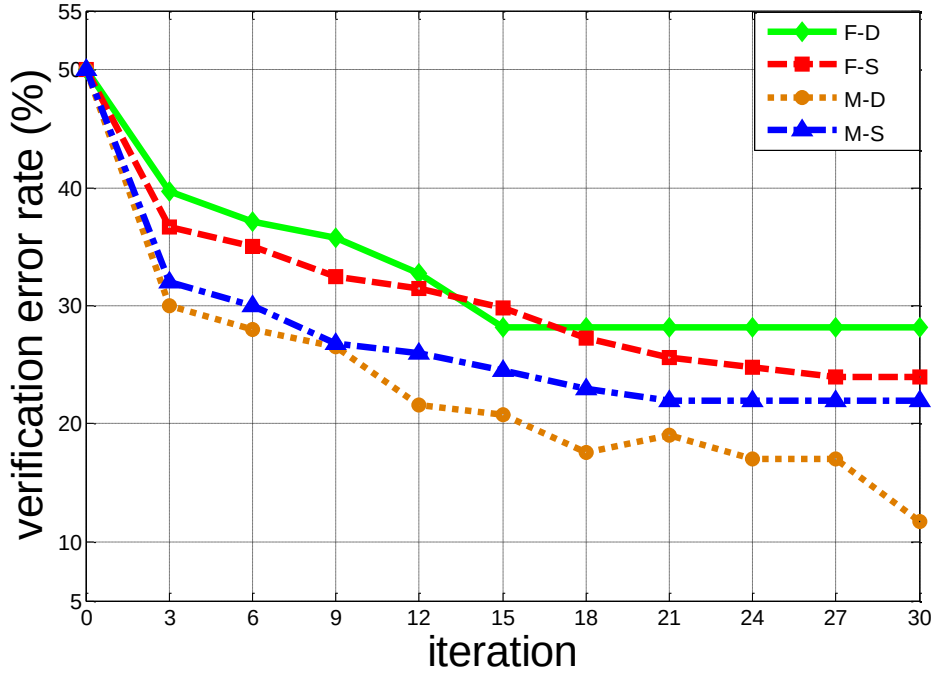
关键点特征提取

- 将输入照片按照关键点位置划分为10部分
- 对每一部分进行特征提取，然后进行融合

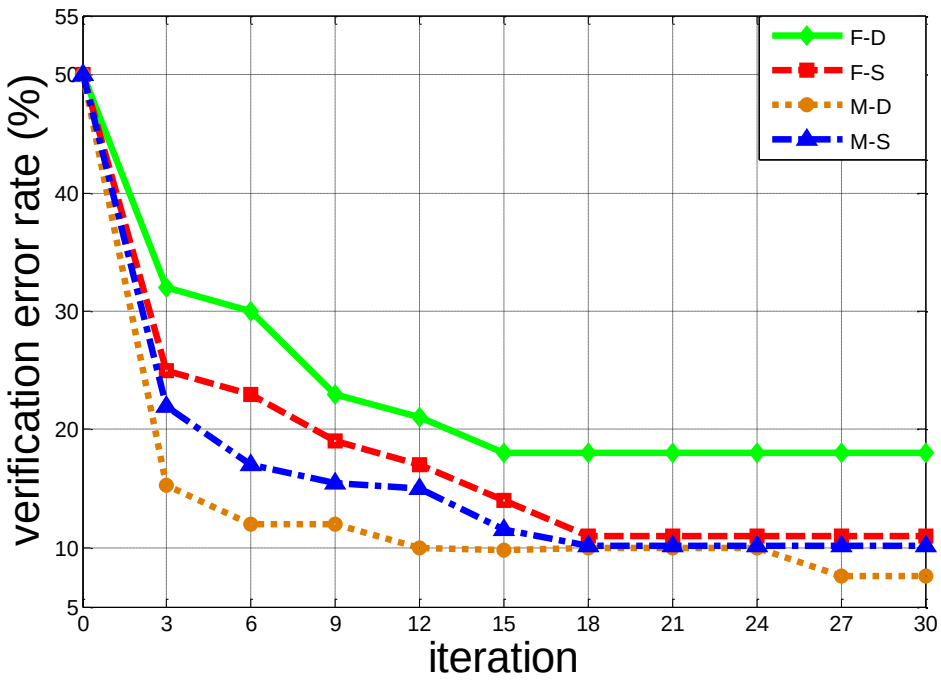


识别结果

- 在两个公开数据库 KFW-I和 KFW-II上，正确率分别提高到77.5%和88.4%



(a)

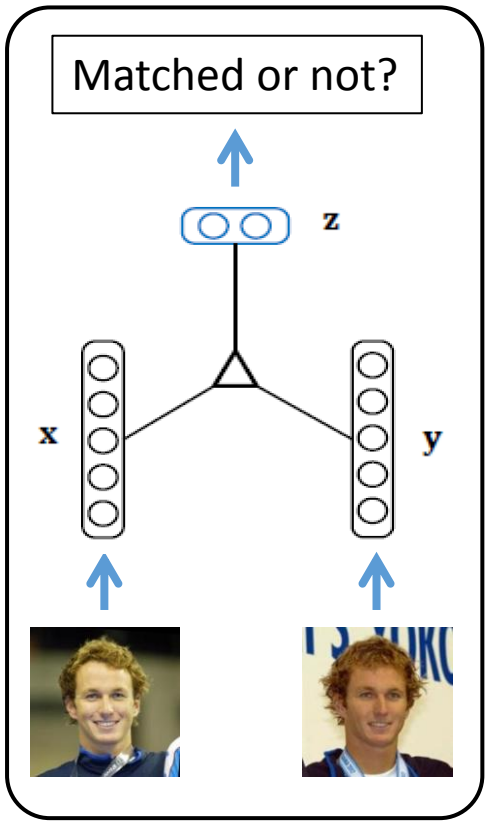


(b)

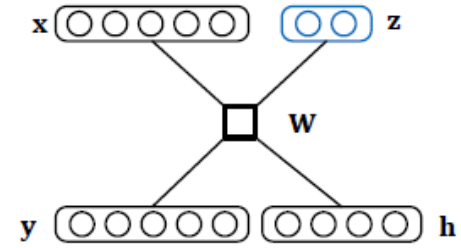
图1：4种亲属关系的识别错误率：(a) KFW-I和(b) KFW-II

深度关系学习（ICCV2015）

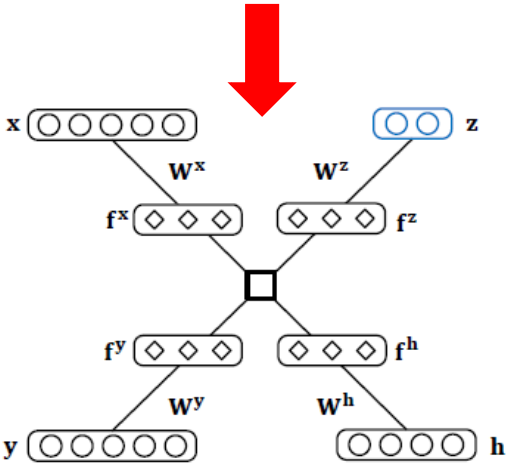
- Propose new supervised learning algorithms of High-order Boltzmann Machines
- Apply to relational learning tasks, e.g., face verification



(a) CHBM



(b) Gated CHBM



(c) Factored Gated CHBM

Untangle factors of variation, e.g., expression

Factorize the 4-order weight tensor into matrices

图像变换及流形可视化

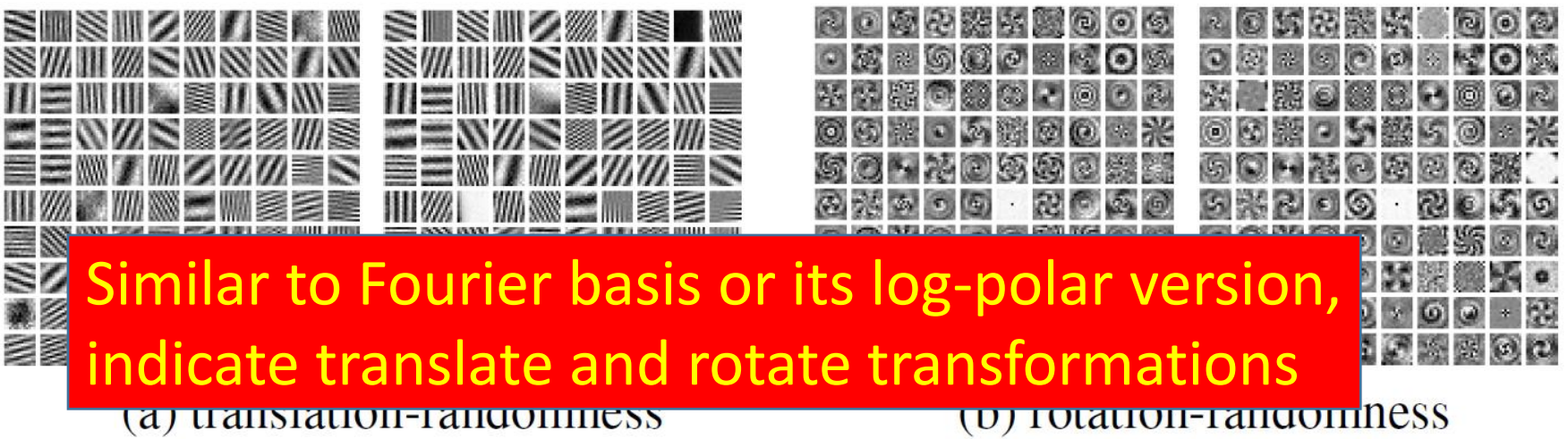


Figure 1. Visualization of learned pairwise filters on the translational and rotary pairs of images.

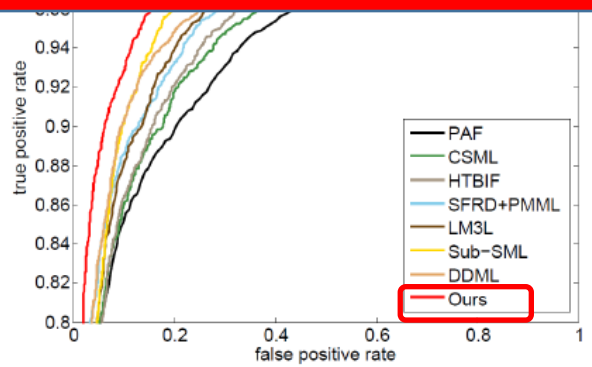
Figure 2. T-SNE visualization of predicted similarities on the MNIST-basic dataset.

人脸验证结果

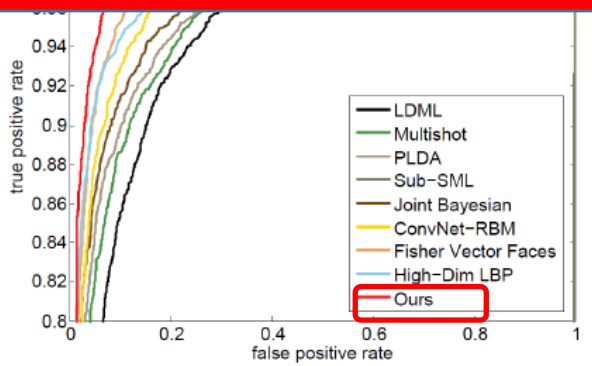
Method	Accuracy
PAF [41]	87.77 ± 0.51
Convolutional DBN [22]	87.77 ± 0.62
CSML [28]	88.00 ± 0.37
HTBIF [29]	88.13 ± 0.58
SFRD+PMML [8]	89.35 ± 0.50
LM3L [17]	89.57 ± 1.53
Sub-SML [4]	89.73 ± 0.38
DDML [16]	90.68 ± 1.41
VMRS [2]	91.10 ± 0.59

Method	Accuracy
LDML [11]	87.50 ± 0.40
Multishot [36]	89.50 ± 0.51
PLDA [23]	90.07 ± 0.51
Sub-SML [4]	90.75 ± 0.64
Joint Bayesian [6]	90.90 ± 1.48
ConvNet-RBM [33]	91.75 ± 0.48
VMRS [2]	92.05 ± 0.45
Fisher Vector Faces [32]	93.03 ± 1.05
High-Dim LBP [7]	93.18 ± 1.07

Achieve the current best results on the LFW dataset under two protocols, without using outside labeled training data



(a) Restricted protocol



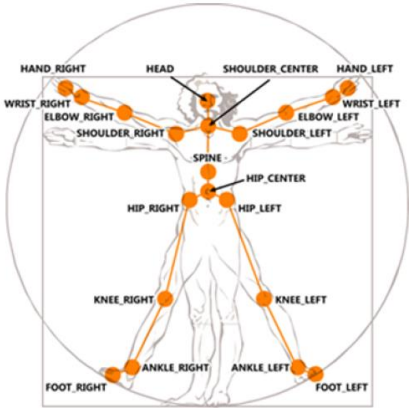
(b) Unrestricted protocol

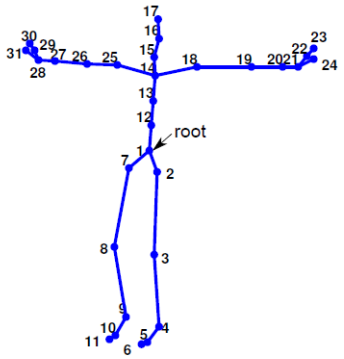
Figure 3. ROC curves by state-of-the-art methods on the LFW dataset, under restricted and unrestricted protocols.

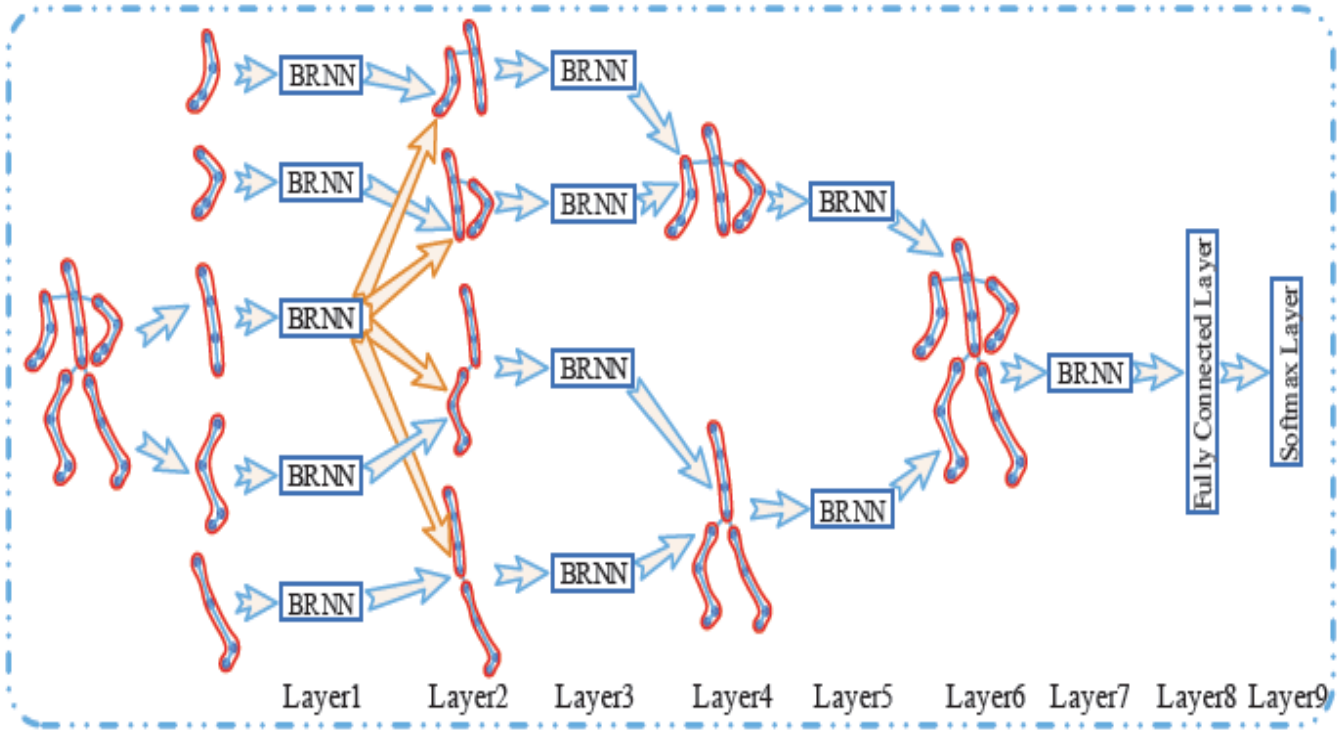
4. 视频分析

骨架行为识别（CVPR2015）

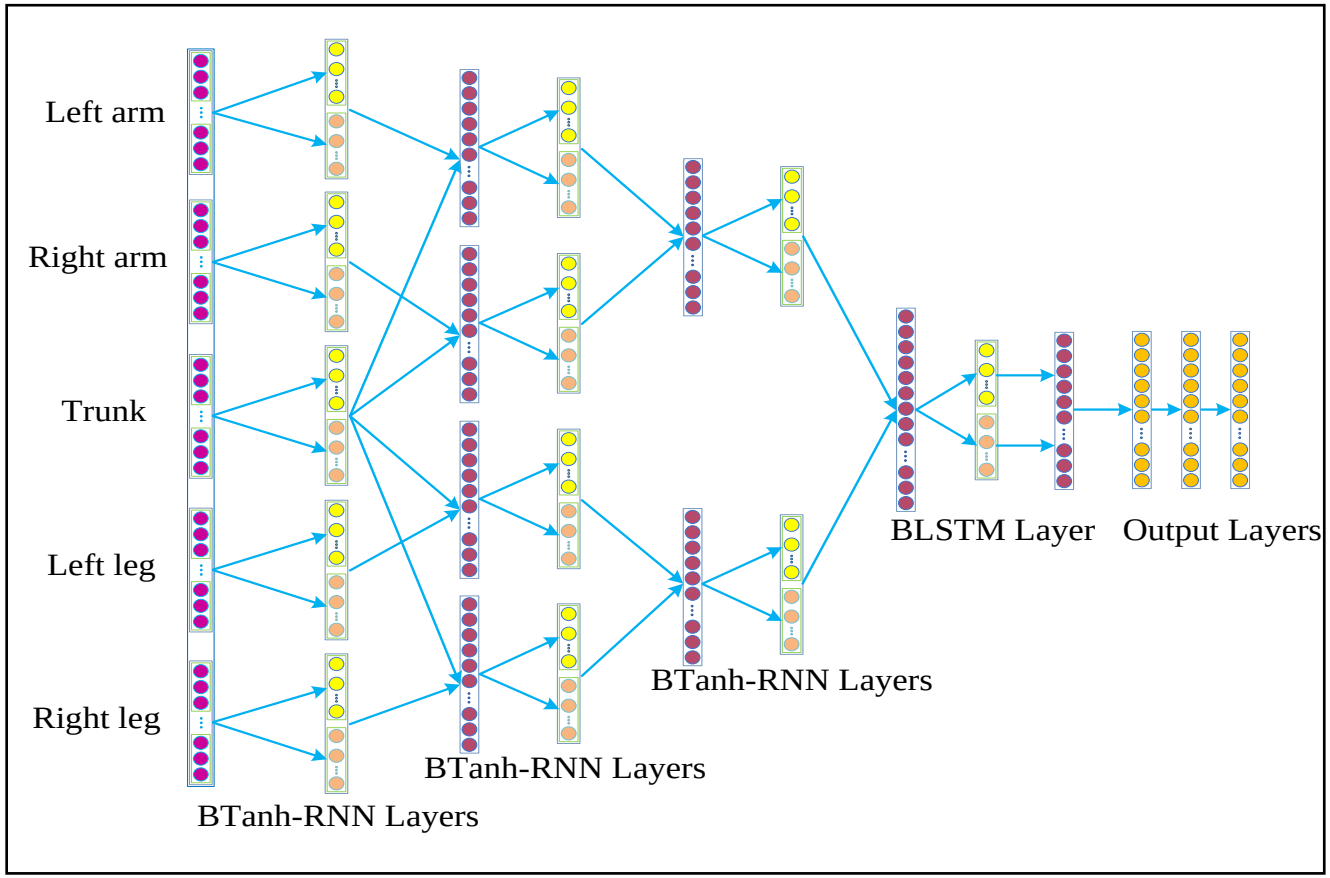
- Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition
- Use five divided skeletons but not a whole skeleton

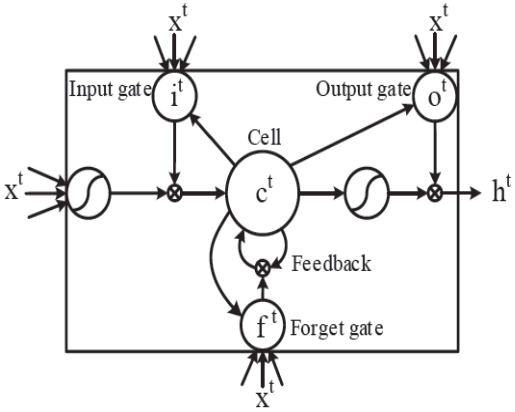


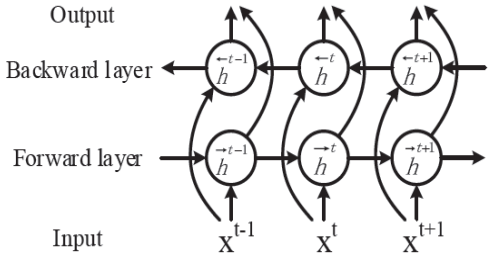




模型框架







The architecture of our proposed model

实验结果

Table 1. Experimental results on MSR Action3D.

Method	AS1(%)	AS2(%)	AS3(%)	Ave(%)
Li et al., 2010	72.9	71.9	79.2	74.7
Chen et al., 2013	96.2	83.3	92.0	90.47
Gowa				
Vemu				
HBR				

Table 2. Experimental results on the MHAD.

MethodAcc	Acc.(%)
Ofli et al., 2014	95.37
Vantigodi et al., 2014	97.58

Computational Efficiency:
52.46ms/sequence,
4
ence

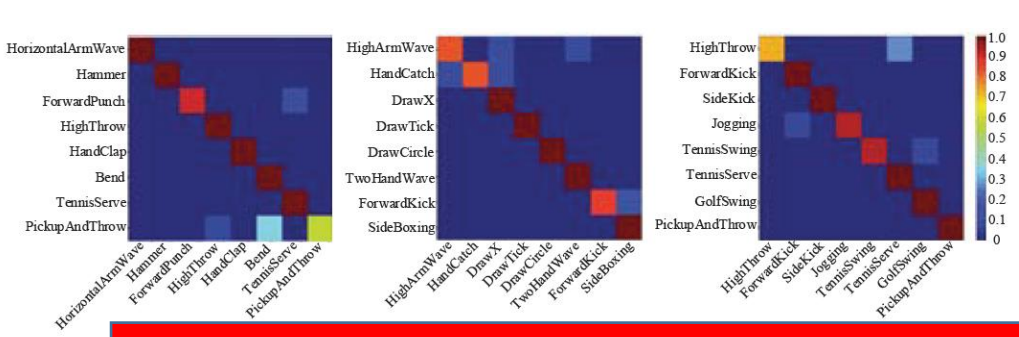
Verify the effectiveness of divided feature extraction, LSTM and bidirectional connection

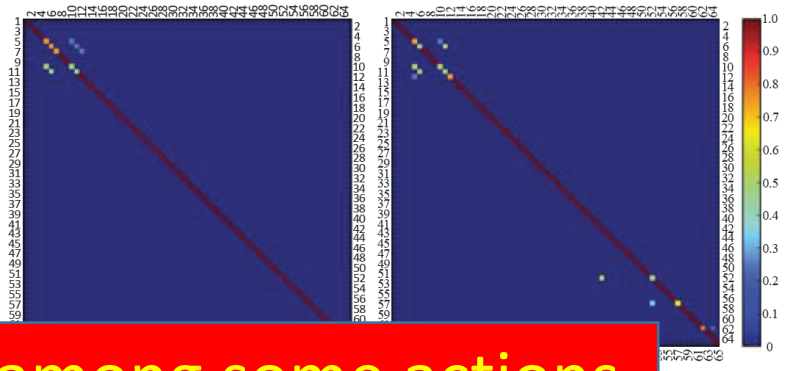
Table 3. Experimental results on HDM05.

Method	Ave. Acc. (%)	Std.
Cho and Chen, 2013	95.59	0.76
HBRNN-L	**96.92**	**0.50**

Table 4. Robustness tests.

HBRNN-L					
Noise Var.	0.5	1	2	4	8
Acc. (%)	**99.64**	**98.18**	**90.91**	**82.55**	**69.82**





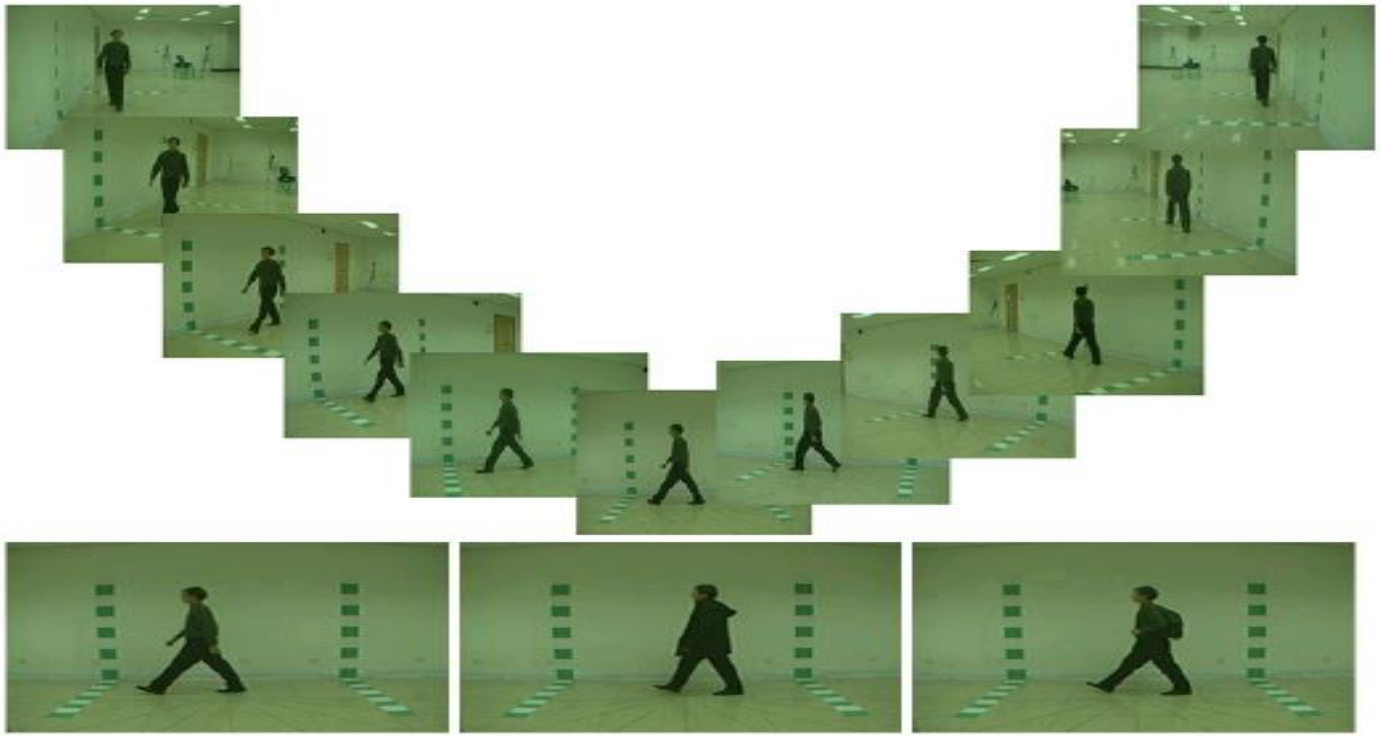
(a) A

Misclassifications mainly occur among some actions sharing the similar spatial and temporal variations

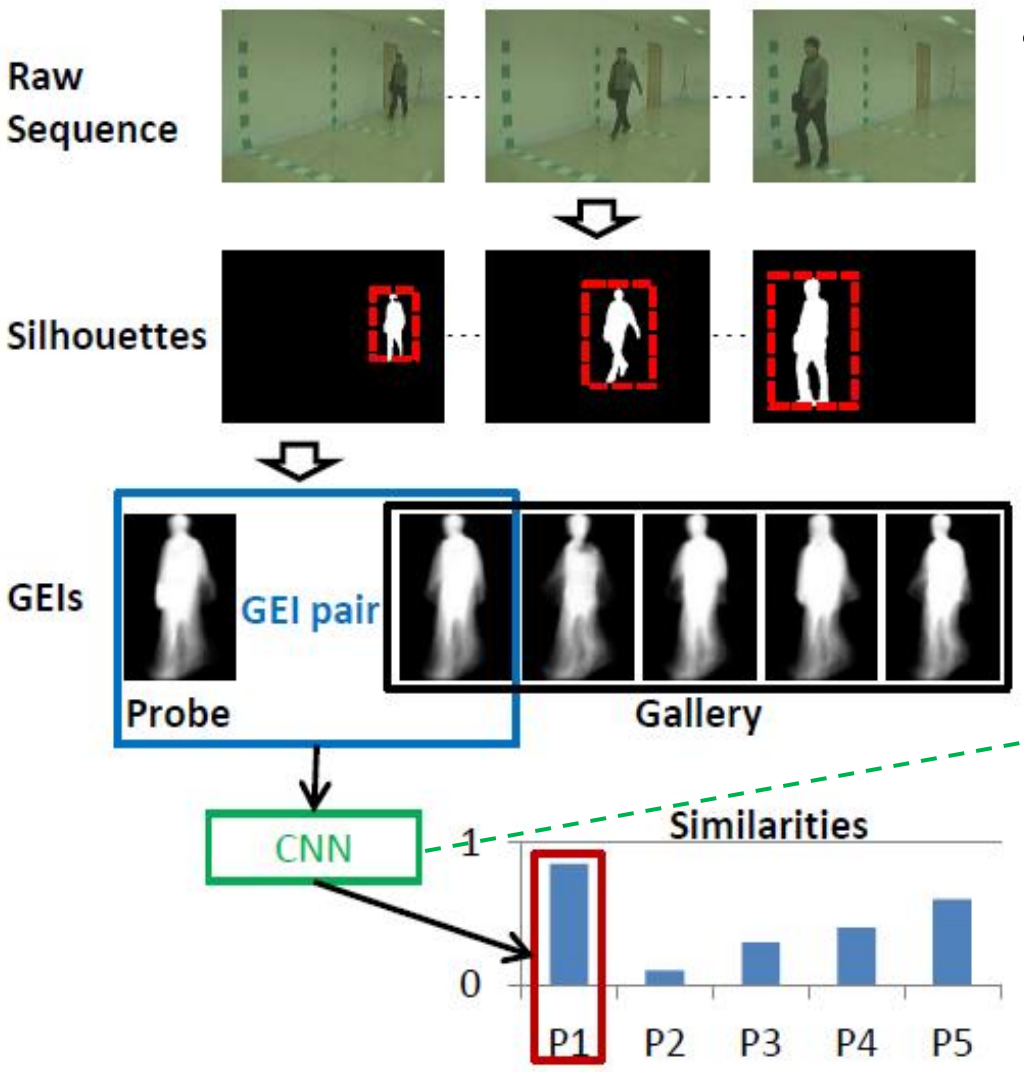
Figu ces on
Acti

跨视角步态识别（TPAMI，Minor）

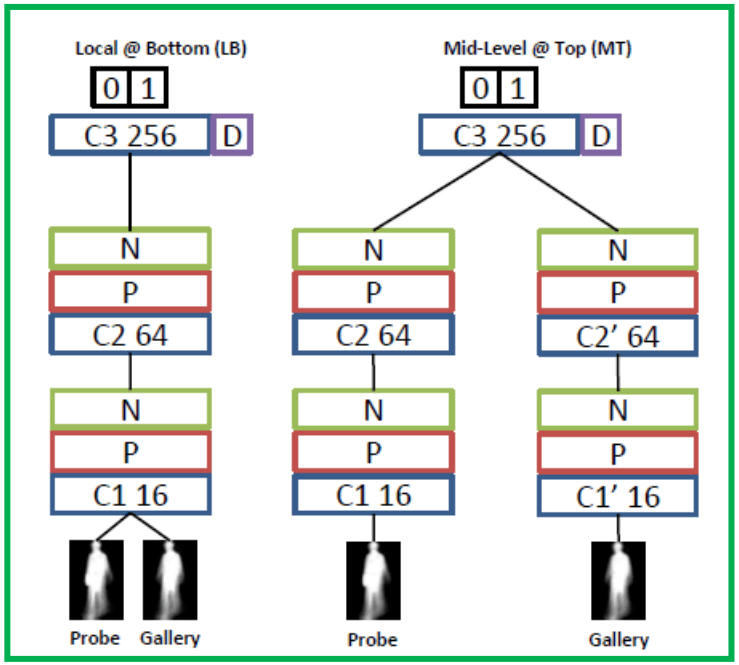
- 通过视频中人的行走方式（任意角度）来进行人的身份识别
- 不同视角下人的步态序列变化非常大，人眼都难以区分



解决方案



- 利用卷积神经网络来
  - 建模不同视角下步态序列的表观变化
  - 并度量两个序列的相似性



测试多种相似度度量网络

识别结果

- 最难的跨视角步态数据库CASIA-B
  - 我们算法的步态识别精度为90%，而之前国际最高水平仅为75%
- 国际最大的步态数据库OULP
  - 不跨视角测试下，我们算法的步态识别精度为98%，而之前国际最高水平仅为85%
  - 跨视角测试下，我们算法达到了91%，目前还没有其他算法在该数据库下测试跨视角情况

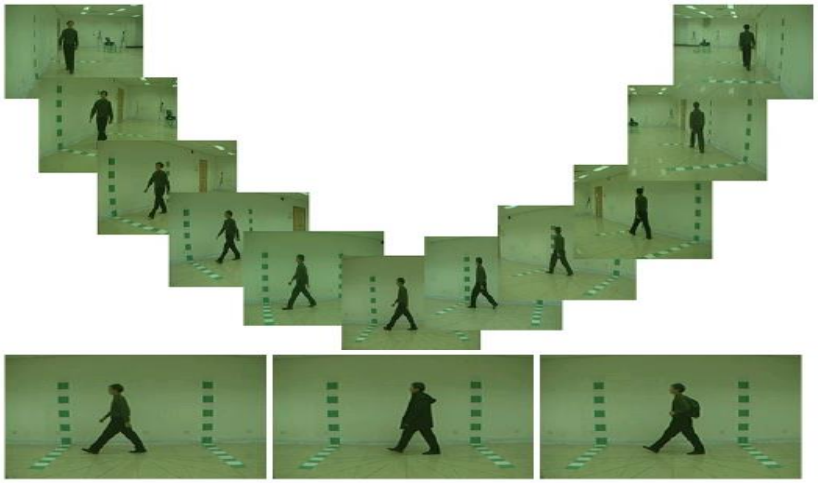
	Gallery	0°-180°				36°-144°		
	Probe	0°	54°	90°	126°	54°	90°	126°
24 train	GEI+ViDP [1]	-	59.1	50.2	57.5	83.5	76.7	80.7
	GEI+CMCC [2]	46.3	52.4	48.3	56.9	-	-	-
	GEI+CNN (ours)	**48.0**	**67.3**	**62.6**	**70.5**	**84.1**	**85.4**	**86.3**
74 train	GEI+ViDP [1]	-	64.2	60.4	65.0	87.0	87.7	89.3
	GEI+CNN (ours)	**76.3**	**90.8**	**83.0**	**90.4**	**98.5**	**98.5**	**98.5**

[1] View-invariant discriminative projection for multi-view gait-based human identification. M. Hu et al., TIFS, 2013.
[2] Rocognizing gaits across views through correlated motion co-clustering. W. Kusakunniran et al., TIP, 2014.

图像集的深度表示（TMM2015）

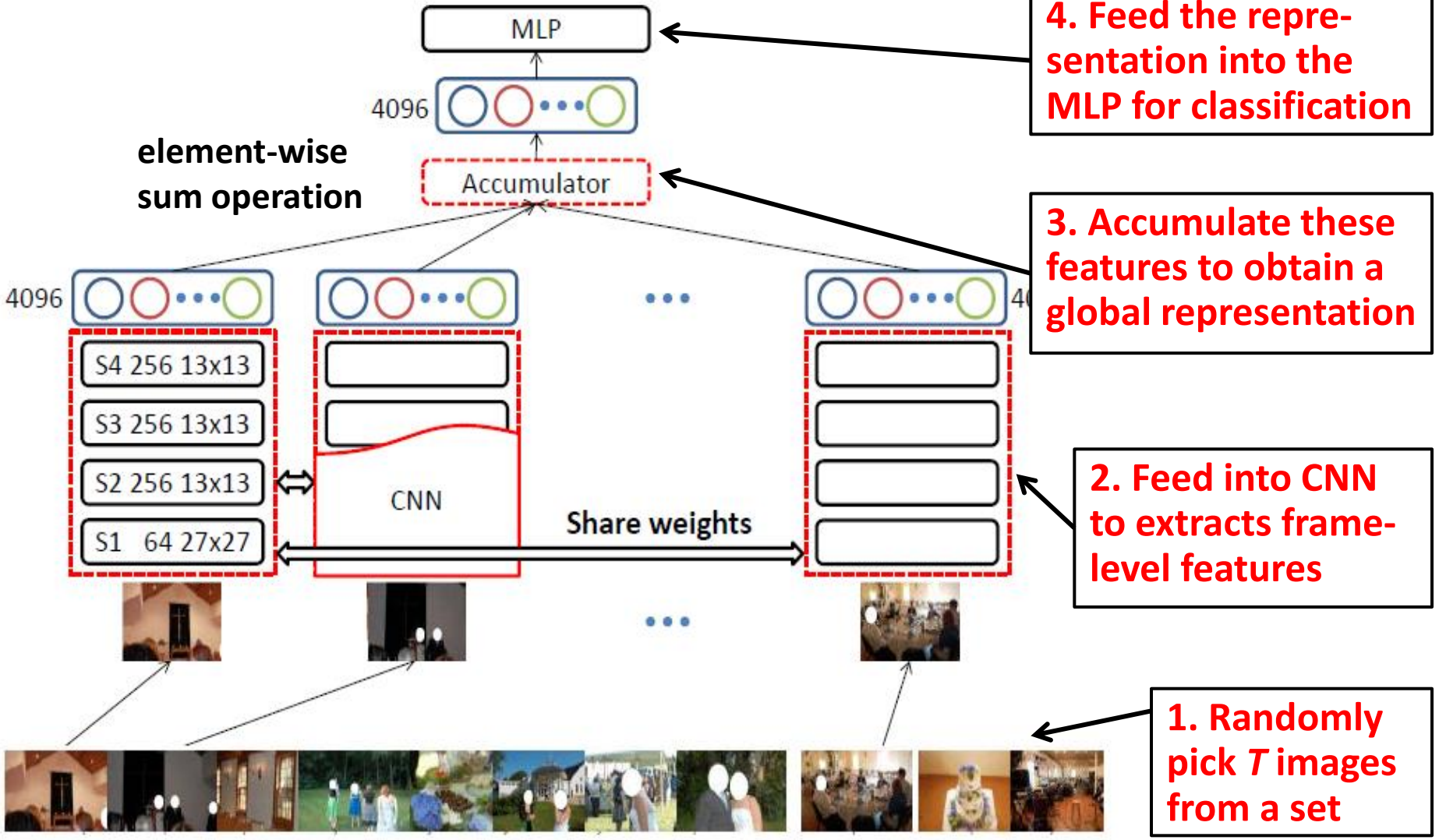
- To learn powerful features from sets of labeled raw images, e.g., video frames
- Focus on dealing with sets of images, no matter if the sets bear temporal structures

包含图片的相册
personal event classification



包含帧的视频
gait recognition

解决方案


MLP
4096
Accumulator
element-wise
sum operation
4096
S4 256 13x13
S3 256 13x13
S2 256 13x13
S1 64 27x27
CNN
Share weights
4. Feed the repre-sentation into the MLP for classification
3. Accumulate these features to obtain a global representation
2. Feed into CNN to extracts frame-level features
1. Randomly pick T images from a set

实验结果

Table 1. Comparison of our method with previous ones on the test set of the PEC dataset by class-wise and average accuracies.

Method	Birthday	Children's birthday	Christmas	Concert	Cruise	Easter	Exhibition	Mean (%)	Recall@2 (%)
No motion [5]	0	30	50	100	80	50	70	51.43	70.00
With motion [5]	10	30	70	100	50	50	70	55.71	72.86
T=1, M=N_{im}	0	30	100	100	80	10	40	56.43	77.14
T=1, M=64	4.0±4.9	28.0±6.0	95.0±5.0	98.0±4.0	67.0±7.8	16.0±4.9	36.0±4.9	55.71±1.53	74.71±1.97
T=16, M=8	0.0±0.0	56.0±6.6	77.0±6.4	89.0±3.0	68.0±7.5	27.0±9.0	79.0±3.0	**63.36**±1.50	77.07±2.01
T=1, M=64, A-net	6.0±6.6	49.0±3.0	96.0±4.9	99.0±3.0	75.0±8.1	44.0±4.9	69.0±10.4	71.71±1.29	85.36±0.86
T=16, M=8, A-net	12.0±7.5	57.0±4.6	89.0±3.0	100.0±0.0	82.0±4.0	44.0±4.9	75.0±5.0	**73.43**±1.27	**90.43**±1.25

Our method does not exploit any motion information at all

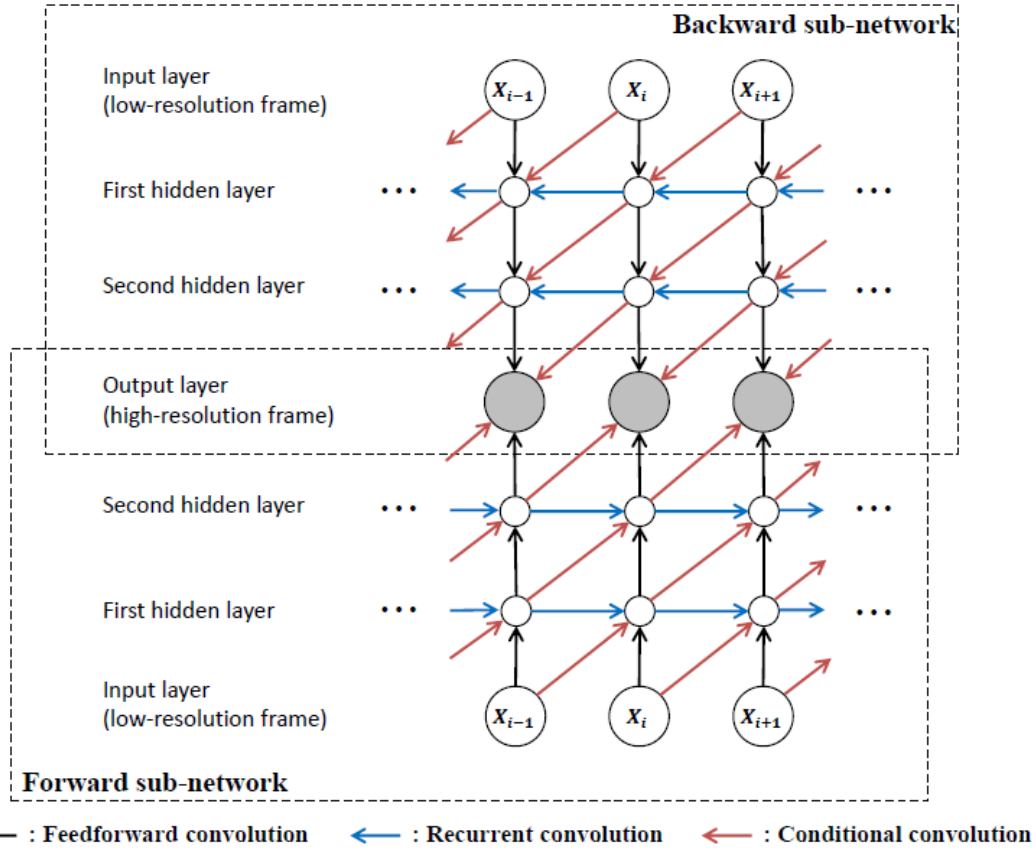
The obvious improvement in performance shows the effectiveness of our learned features

Table 2. Comparison of our method with previous ones on CASIA-B by average accuracies.

Gallery NM #1-4	0°-180°			36°-144°		
Probe NM #5-6	54°	90°	126°	54°	90°	126°
CCA [40]	–	–	–	66	66	67
ViDP [42]	64.2	60.4	65.0	87.0	87.7	89.3
Ours	**77.7**	59.9	**75.0**	**89.2**	81.5	**90.0**

多帧超分辨率（NIPS2015）

- Model **long-term temporal dependency** of video sequences
- Replace all full connections with weight-sharing convolutions



: Feedforward convolution

learn spatial dependency between a low-resolution frame and its high-resolution result

: Recurrent convolution

model long-term temporal dependency across video frames

: Conditional convolution

enhance visual-temporal dependency modelling

PSNR结果与运行时间

Table 1. The results of PSNR (dB) and test time (sec) on the test video sequences.

Video	Bicubic		SC [24]		K-SVD [25]		NE+NNLS [4]		ANR [22]	
	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time
Dancing	26.83	-	26.80	45.47	27.69	2.35	27.63	19.89	27.67	0.85
Flag	26.35	-	26.28	12.89	27.61	0.58	27.41	4.54	27.52	0.20
Fan	31.94	-	32.50	12.92	33.55	1.06	33.45	8.27	33.49	0.38
Treadmill	21.15	-	21.27	15.47	22.22	0.35	22.08	2.60	22.24	0.12
Turbine	25.09	-	25.77	16.49	27.00	0.51	26.88	3.67	27.04	0.18
Fan	33.46	1.76	31.91	5.96	31.91	323	32.14	-	**33.73**	0.64
Treadmill	22.22	0.57	21.15	4.01	22.32	127	21.20	-	**22.63**	0.20
Turbine	26.98	0.80	26.25	4.81	24.27	173	25.60	-	**27.71**	0.26
Average	27.52	1.66	26.48	6.76	26.64	418	26.52	-	**28.15**	0.61

Clearly surpass state-of-the-art methods in PSNR, due to the effective temporal dependency modelling

Table 2. Test variants of BRCN with different network architectures.

Vide					
Dan					
Flag					
Fan	31.91	33.63	33.65	33.65	33.73
Treadmill	21.15	22.59	22.56	22.59	22.63
Turbine	26.25	27.47	27.50	27.62	27.71
Average	26.48	27.99	28.02	28.09	28.15

Achieve orders of magnitude faster speed than other multi-frame SR methods

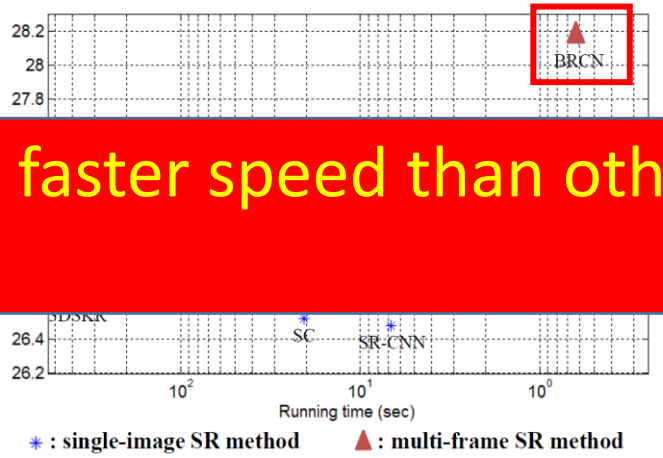
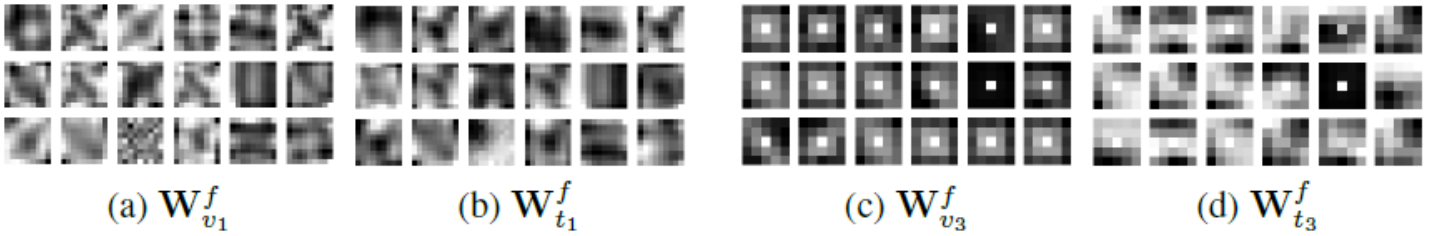


Figure 1. Running time vs. PSNR.

视频超分辨率结果



(a) $\mathbf{W}_{v_1}^f$ (b) $\mathbf{W}_{t_1}^f$ (c) $\mathbf{W}_{v_3}^f$ (d) $\mathbf{W}_{t_3}^f$

Figure 1. Visualization of learned filters in the first (a&b) and last (c&d) layers.

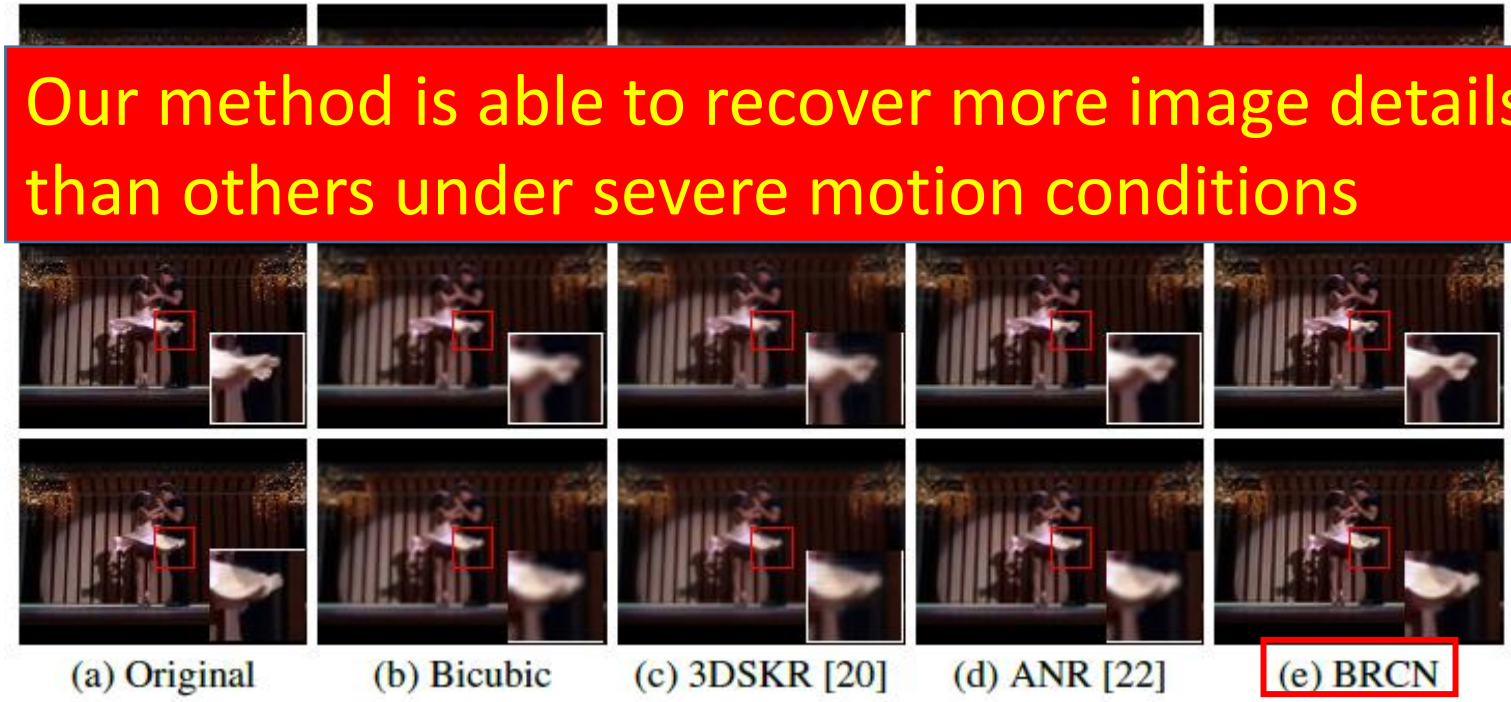
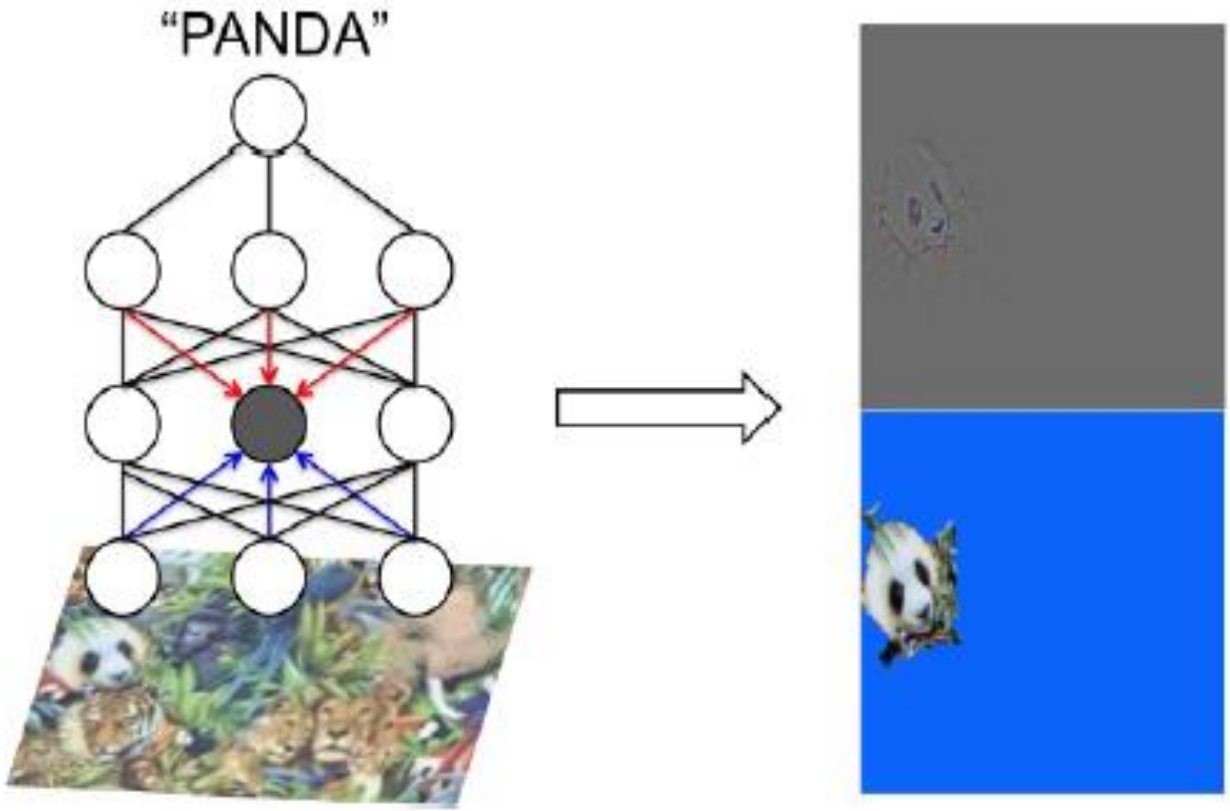


Figure 2. Comparison among original and super resolved results.

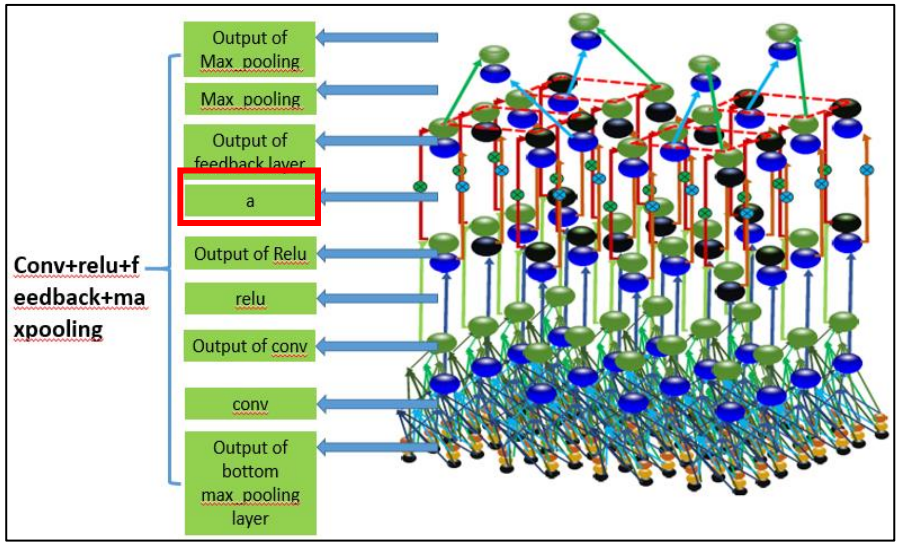
5. 神经网络可视化

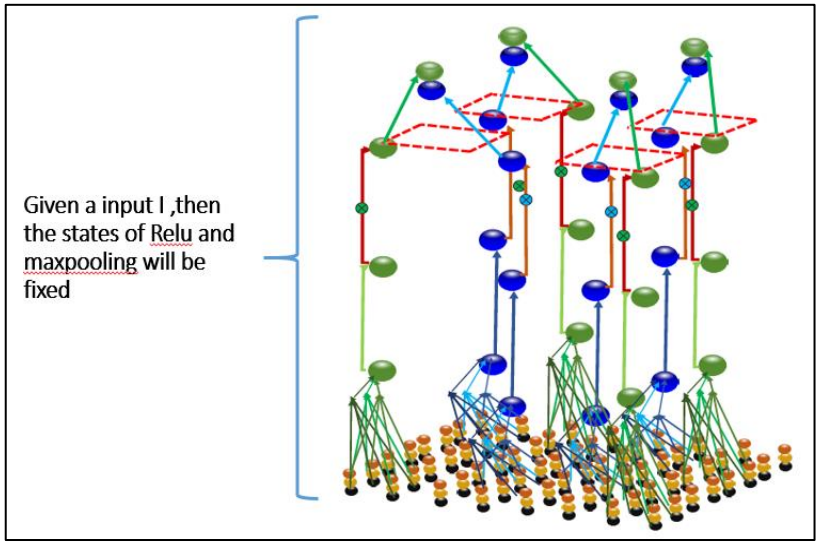
反馈卷积神经网络（ICCV2015）

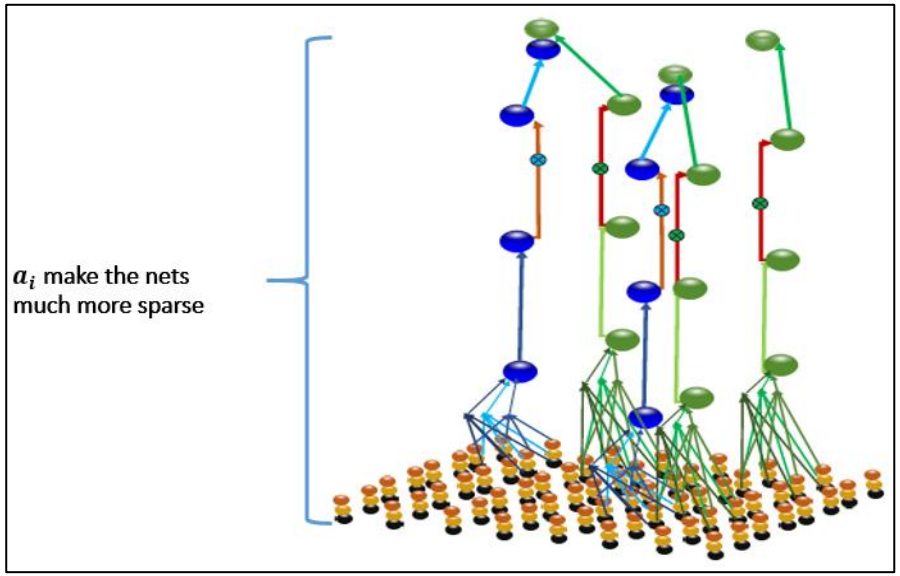
- Feedback for class specified visualization and classification
- In human's brain, visual attention is dominated by "goals" from our mind easily in a top-down manner



反馈卷积神经网络







Feedback Neural Networks

Mathematics

Given $s_0 = f(I_0)$, for $\forall$ I, find T, T_0

$$\min_{T,T_0} ||I * T + T_0 - \max_{a1.....a_n} f(I, a_1, a_2, a_3 \dots a_n)||$$

Mode:

1)hard: when a_i only can be 0 or 1

2)soft: when a_i can be any value in [0 ,1]

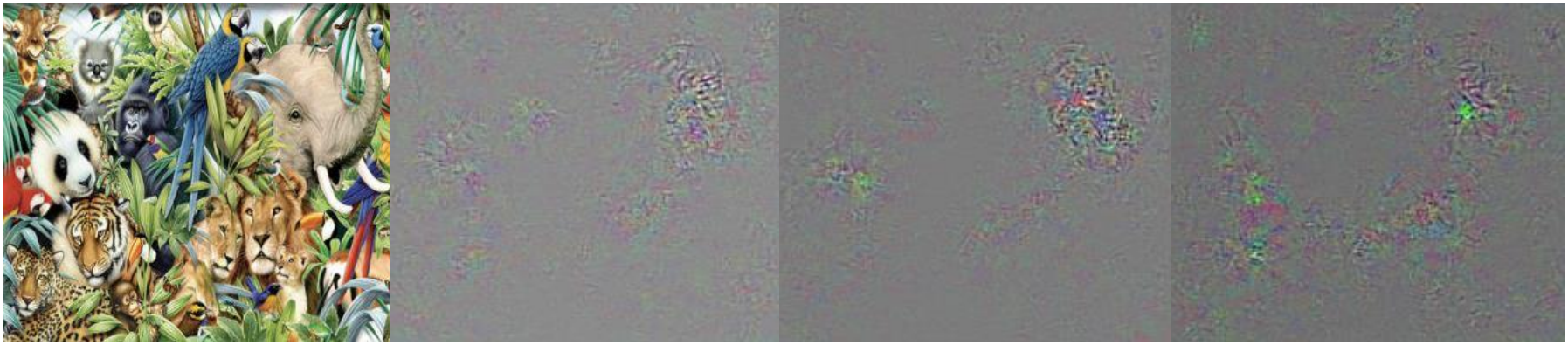
the solution of $\max_{a1.....a_n} f(I, a_1, a_2, a_3 \dots a_n)$:

1) Inital all the a_i with the state of relu layers.

2)Update a_i by $a_i = a_i + r * df/da_i$

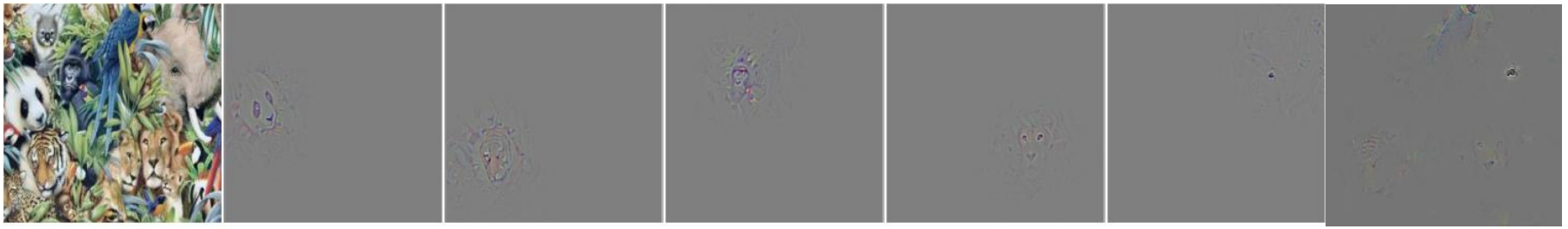
视觉注意机制的可视化

- 模型迭代过程：



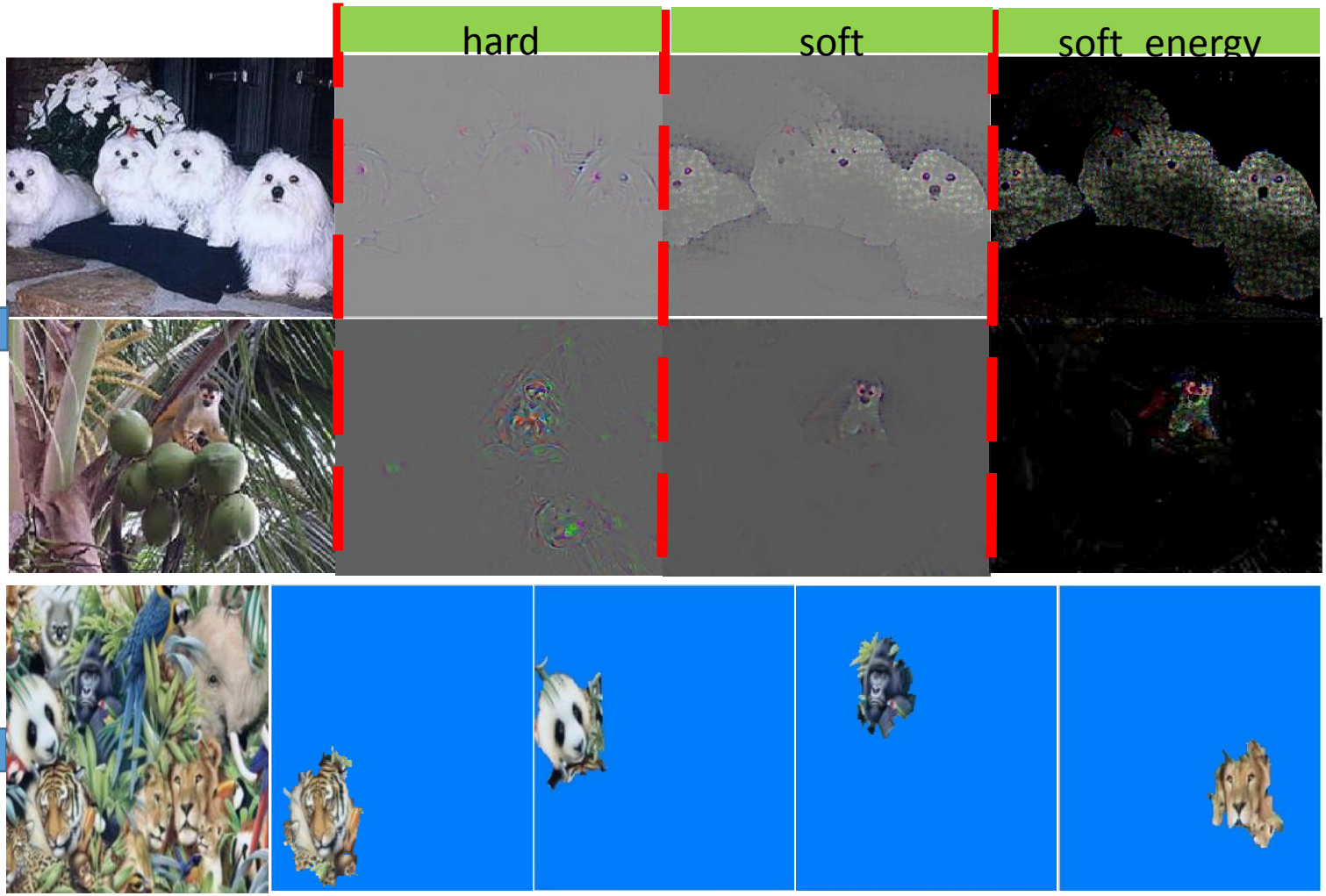
input　　Vgg_panda　　Vgg_elephant　　Vgg_lion

- Googlenet 的视觉注意可视化：



(a) Input Image　(b) Panda　(c) Tiger　(d) Gorilla　(e) Lion　(f) Elephant　(g)bird

可视化结果



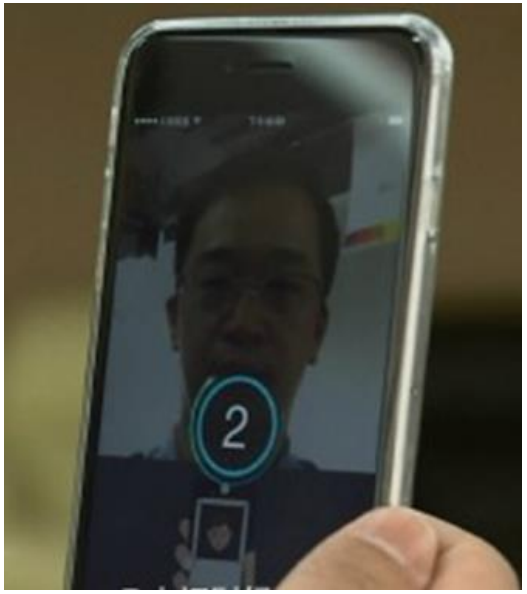
报告提纲

- 视觉大数据
- 大规模视觉计算
- 我们的工作
- 未来方向

未来方向(1/6)

- 深度图像分析
  - 进一步提升效果
  - 转化实际应用，例如移动端应用

年龄估计

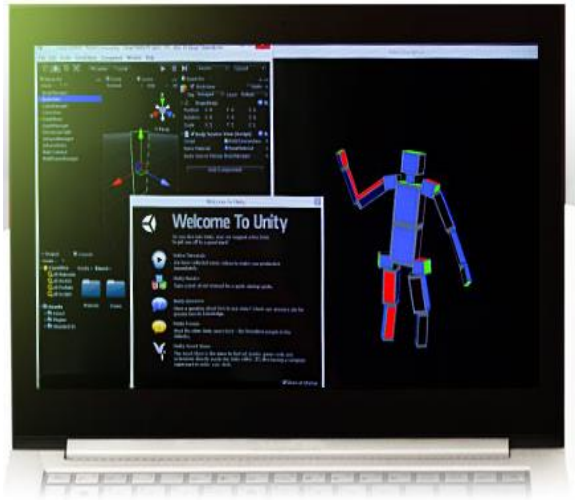


刷脸支付

物体标注

未来方向(2/6)

- 深度视频分析
  - 还处于起步阶段
  - 主要应用包括行为识别等



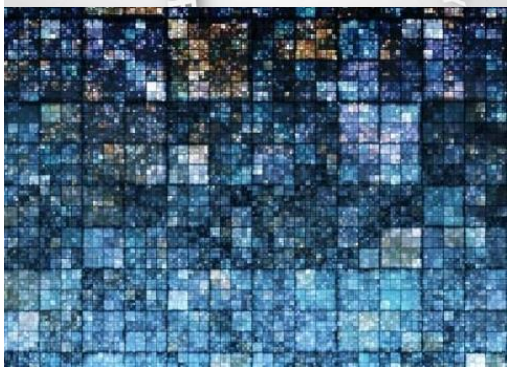
人机交互的行为识别

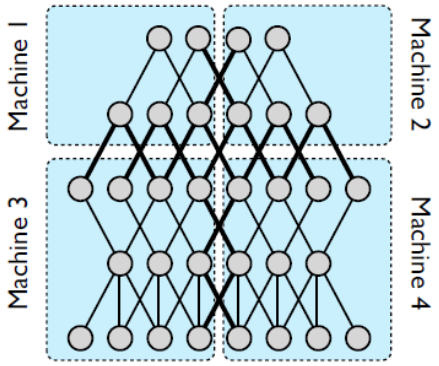
监控视频分析

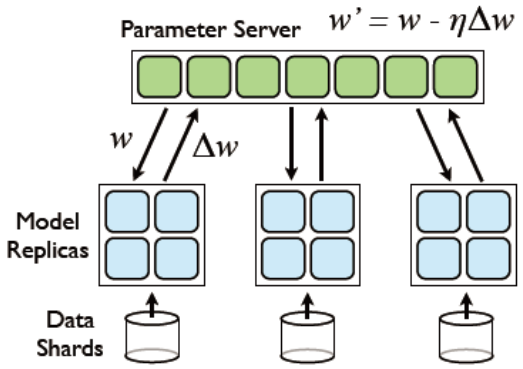
第一视角的视频分析

未来方向(3/6)

- 大规模深度学习
  - 处理更大规模的数据
  - 多GPU并行以及分布式处理



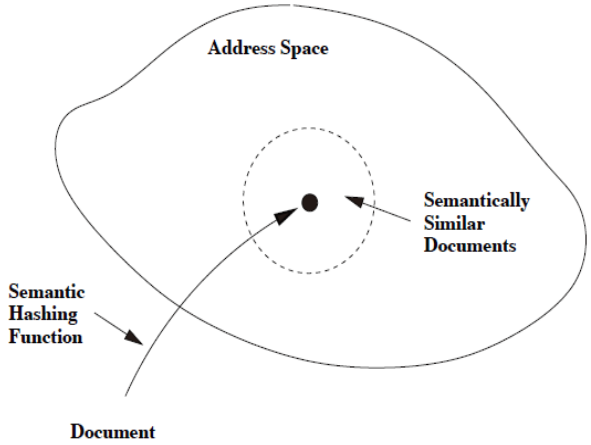




未来方向(4/6)

- 无监督(半监督)学习
  - 在实际应用中，监督信息常常缺失
  - 标注有监督信息代价太大

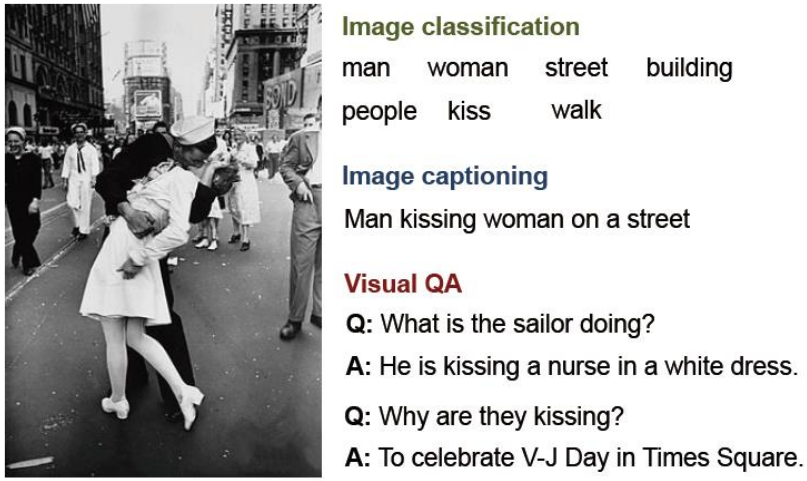
无监督学习得到的人脸和猫脸



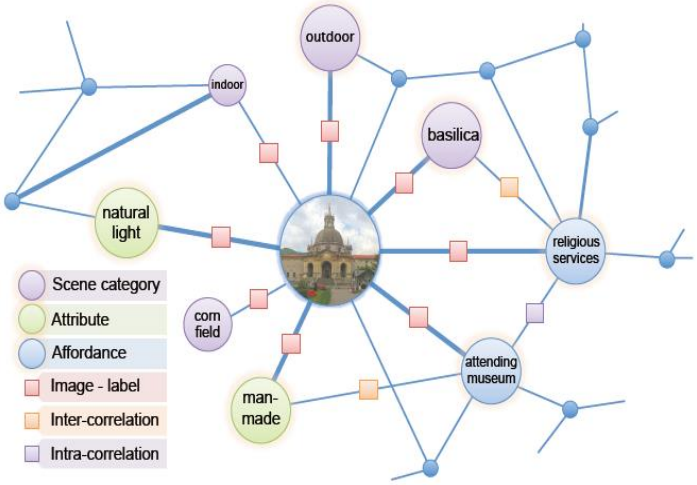
海量数据的流形假设

未来方向(5/6)

- 大规模多模态学习
  - 视觉信息的理解离不开其他数据模态，例如文本和语音
  - 重点对模态间的关联关系进行建模



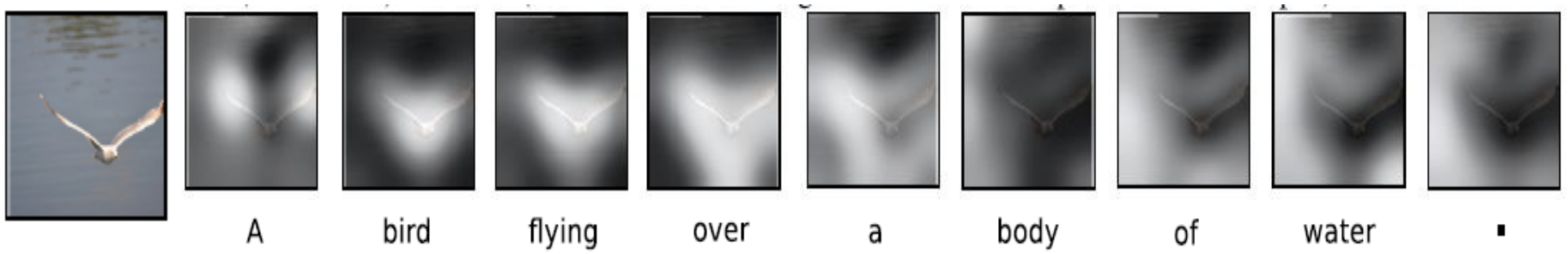
Visual question answering



Multimodal Knowledge Base

未来方向(6/6)

- 类脑智能研究
  - 早期神经网络是通过模拟大脑认知的机理解决各种机器学习问题
  - 如今部分生物机制已经被应用到深度学习中，例如视觉注意机制和神经元跨层连接机制
  - 未来的发展可以更多的借鉴脑神经科学的研究成果



基于深度视觉注意的图像描述

智能感知与计算中心

简介

团队介绍


谭铁牛
研究员


王亮
研究员


黄凯奇
研究员

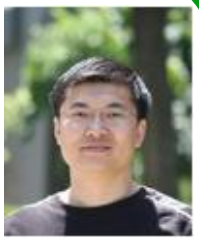

孙哲南
研究员


赫然
副研究员


董晶
副研究员


张彰
副研究员


黄永祯
副研究员


侯广琦
高工

4名研究员、5名副研/高工、9名助研、16名项目聘用人员、46名研究生

iSEE（Intelligent Situation Evaluation and Exploration）

SIR（Smart Identity Recognition）

DIG（Data Intelligence Gathering）

创新智能感知与计算技术，实现人类社会安全态势的透彻感知

iSEE平台

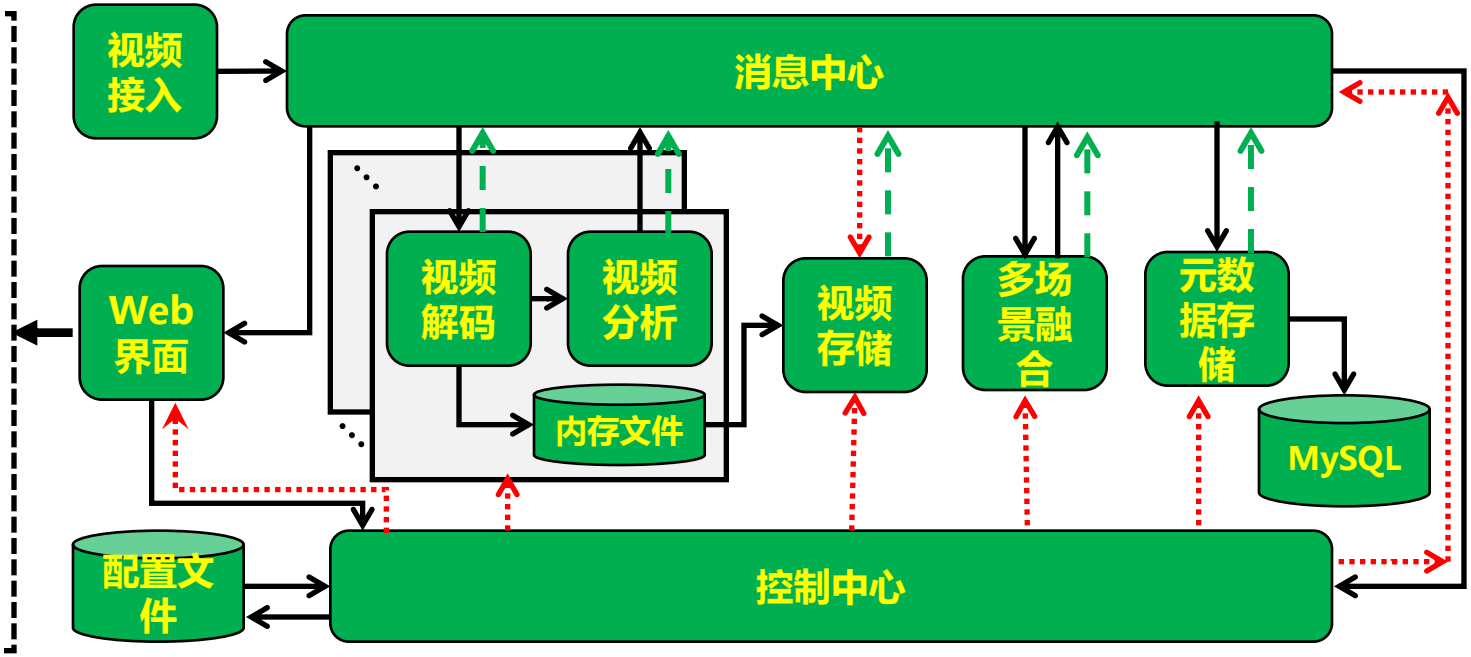
- 多摄像机、大范围、全天候、实时智能视频监控研发平台

●跨场景运动跟踪数据库（MCT）
●大规模行人属性和重识别数据库
●超大规模人群流量数据库

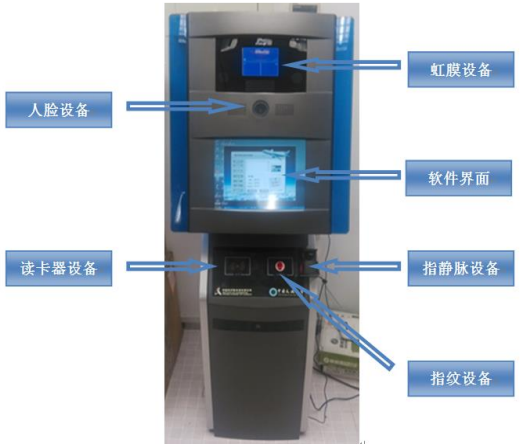
海量目标、事件检索

时空感兴趣模式挖掘

大范围场景态势评估

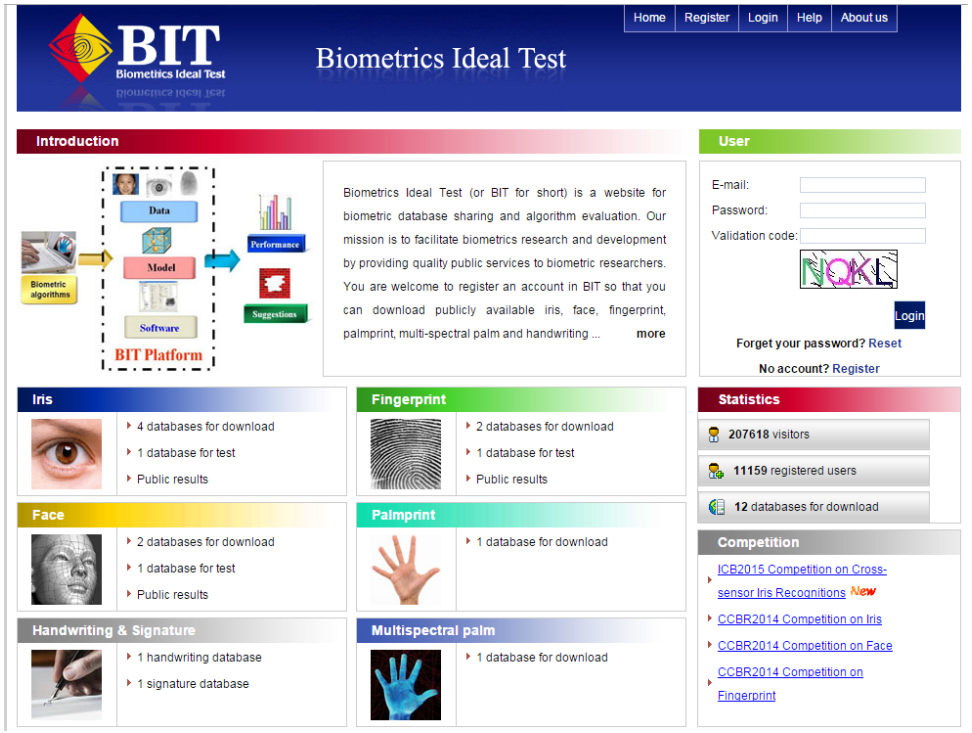


SIR平台



虹膜、人脸、指纹、掌纹、静脉图像数据采集与识别平台

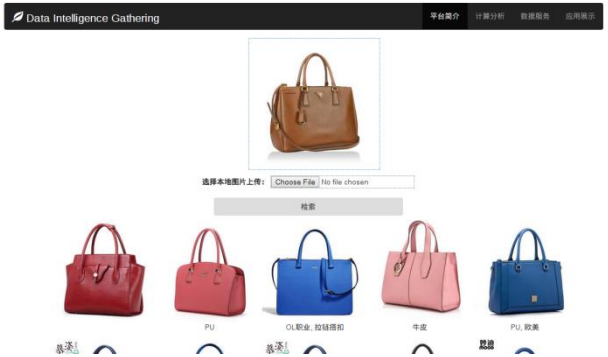
远距离虹膜人脸识别系统



多模态生物特征数据库共享平台BIT已有120多个国家的11159名用户，组织了多次国际国内生物识别算法竞赛

DIG平台

基于先进的分布式的大规模数据存储技术，集成最前沿的数据分析算法，提供完整的海量数据采集、存储、管理、计算、分析和应用的智能大数据分析和挖掘平台。



互联网图片检索示范应用

团队发展思路

打造智能识别科技的产业链和生态圈

Science

天津中科智慧科技研究院
中国英特尔物联技术研究院

Technology

IrisKing 中科虹霸　　istek 中科唯实　……

Product

生物识别
产业联盟

图像视频大数据
产业联盟

Ecosystem

谢谢

（Q&A）