CCF-CV走进高校@郑州航空工业管理学院

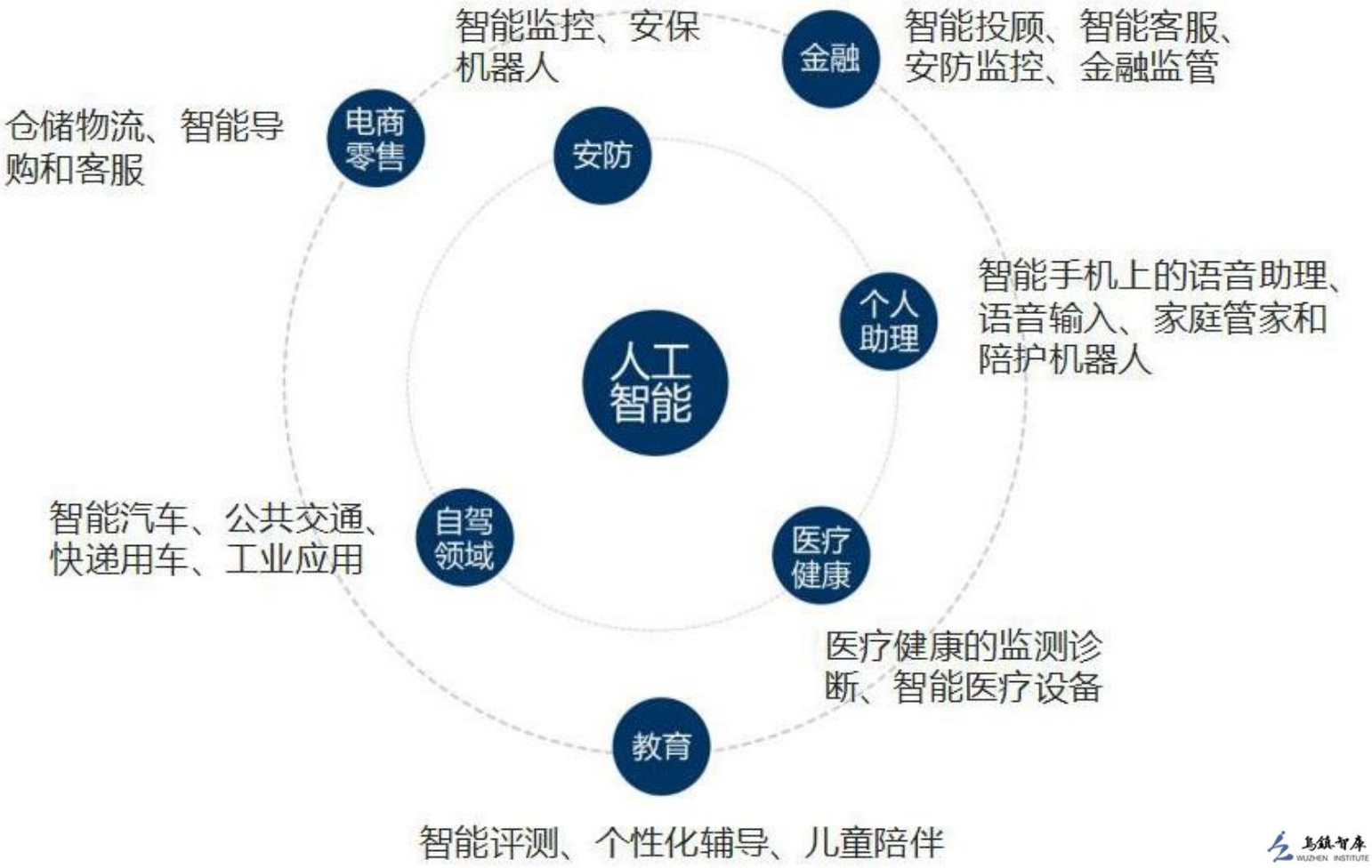
复杂视频的深度分析与理解方法

乔宇

中国科学院深圳先进技术研究院

2018年11月18日

计算机视觉是AI的核心领域



人类社会已经进入视觉信息的“大数据”时代

互联网
图像
视频

facebook

2500亿张照片，视频日播放**80亿**次

WeChat 微信

日上传图片10亿张，视频播放**20亿**次

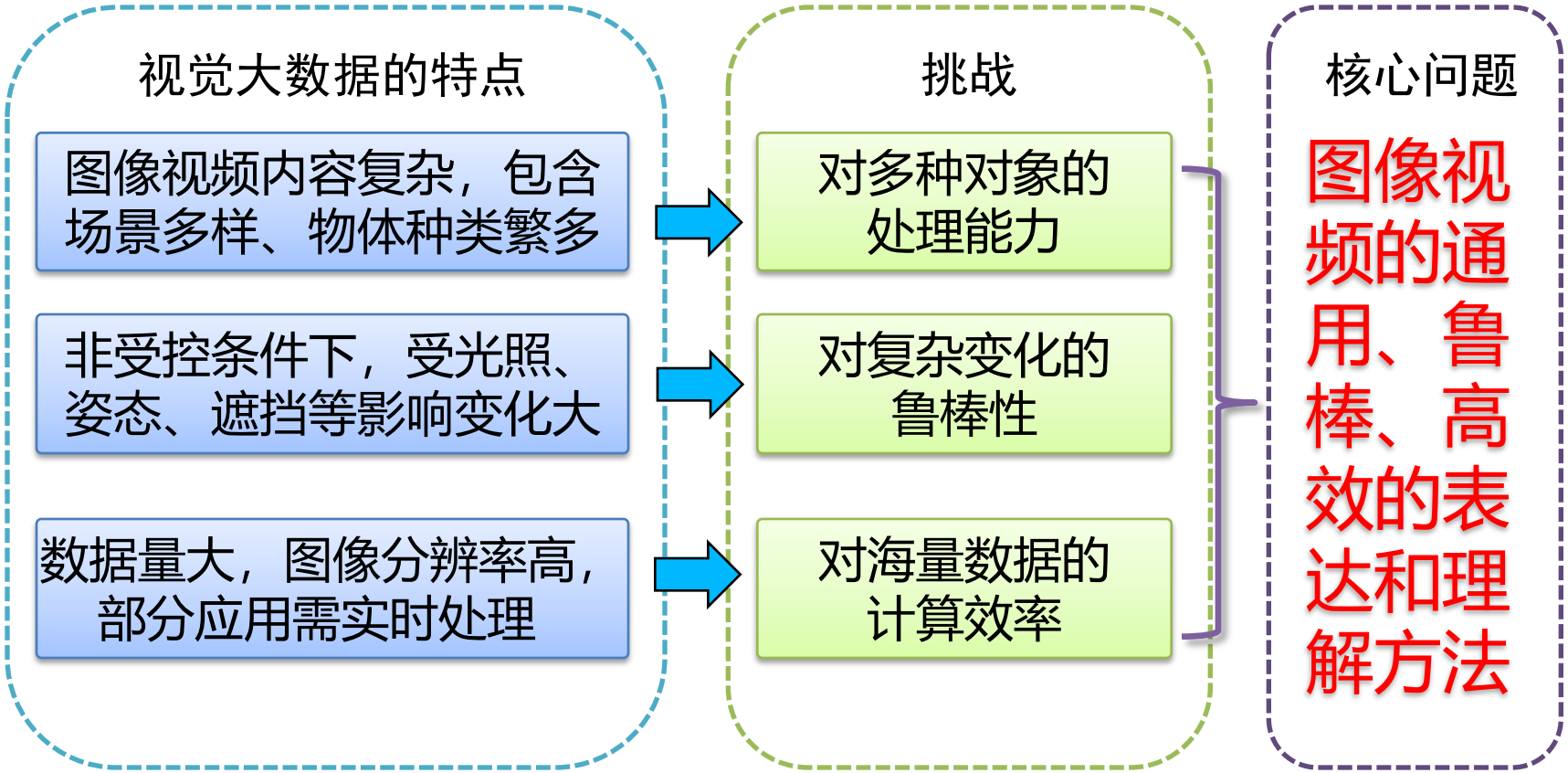
监控
视频

2017年,中国共装有1.76亿个监控摄像头。
深圳、广州等大型城市，市内监控探头总数超100万。

谭铁牛院士

“图像视频大数据是人工智能的突破口，是信息产业新的增长点。”

视觉大数据的挑战



深度学习方法的出现和发展，为解决上述问题和挑战提供了强有力的工具。

深度学习方法快速发展

深度神经网络已经在语音、视觉、自然语言处理等领域取得了很大成功，在学术界和工业界都引起了极大关注。

深度学习理论突破

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.

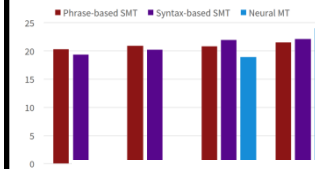
Dimensionality reduction facilitates the classification, visualization, communi- finds the directions of greatest variance in the data set and represents each data point by its n. We A that rework

深度置信网络

ImageNet竞赛：74% vs. 85%

1000类，1百万数据

机器翻译和文本生成



Machine Translation

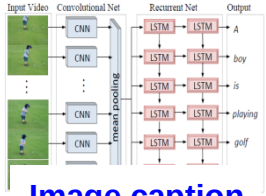
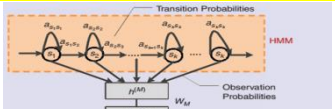


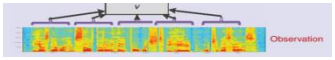
Image caption

2006 2011 2012 2013 2015 2016

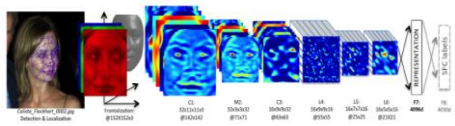
大规模语音识别



Switchboard：错误率降低8.9%



人脸识别



在LFW上识别率99%,超过人类

围棋

Alphago 4:1 李世乭

物体识别：ImageNet 数据库

ImageNet 包含100万图像，1000个类别。
Large Scale Visual Recognition Challenge from 2010

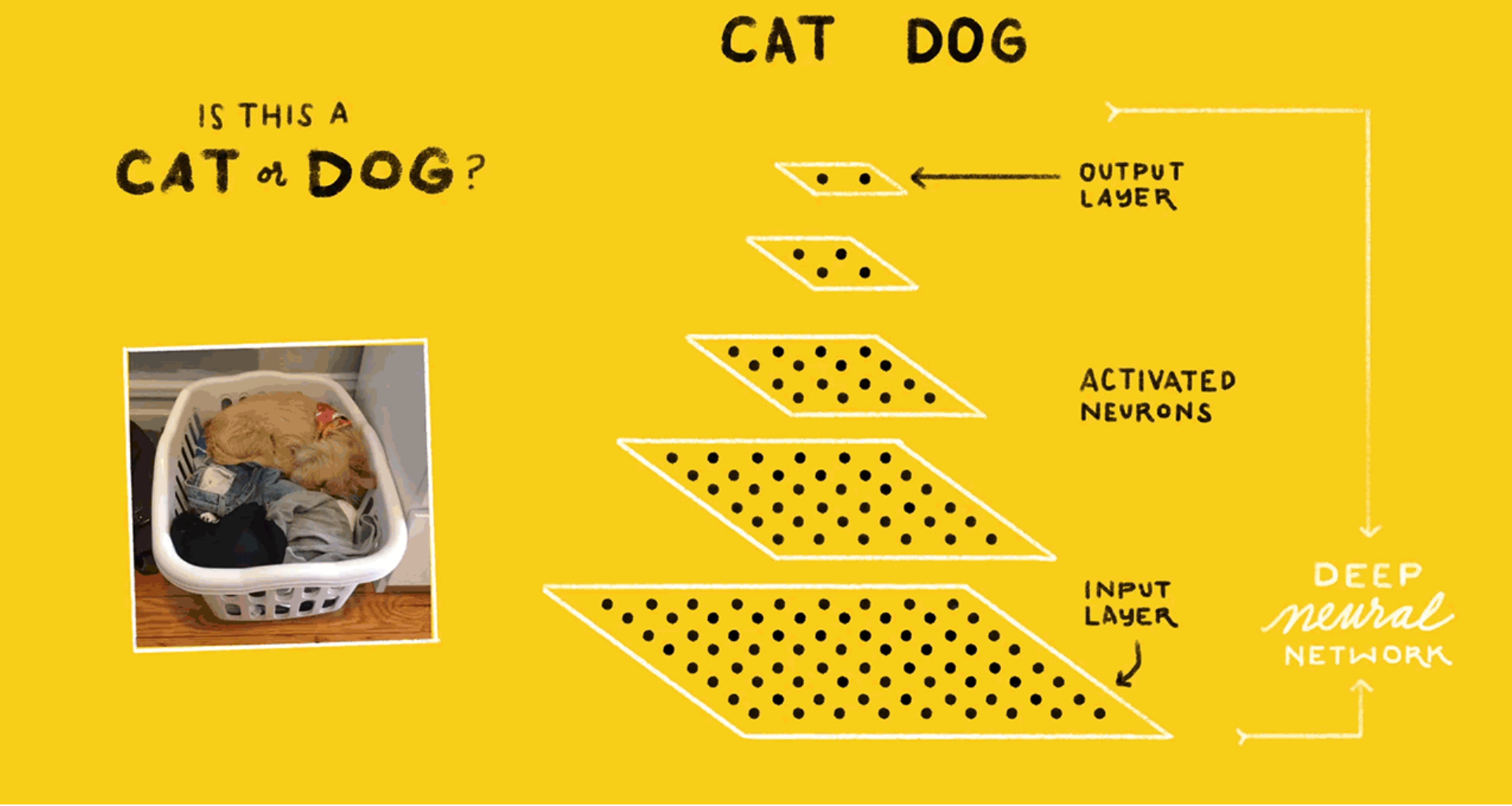
http://www.image-net.org/

ImageNet历年识别率

ImageNet分类任务Top5错误率(人的错误率是5.1%)

https://news.uc.cn/a_1803383491732101351 4/

卷积神经网络如何识别图片

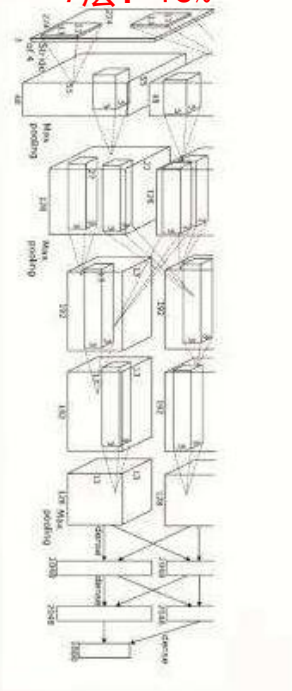


From J. Dean: Large-Scale Deep Learning for Intelligent Computer Systems. 2016

深度网络的结构演化

"AlexNet"

7层：16%



[Krizhevsky et al. NIPS 2012]

ReLU
DropOut
Data Argum.

"GoogLeNet"

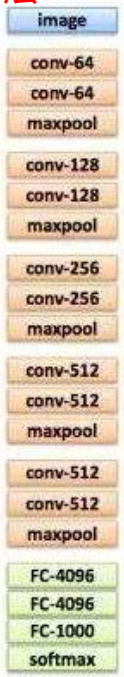
22层：6. 7%

[Szegedy et al. CVPR 2015]

Inception
Multiple branches
1x1 Conv
Bottleneck
Batch Normalization

"VGG Net"

19层：7. 3%

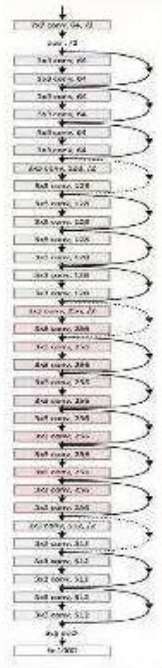


[Simonyan & Zisserman, ICLR 2015]

3x3 Conv.
Modularized design
Stage-wise training
- VGG-11 => VGG-13 …

"ResNet"

152层：3. 5%



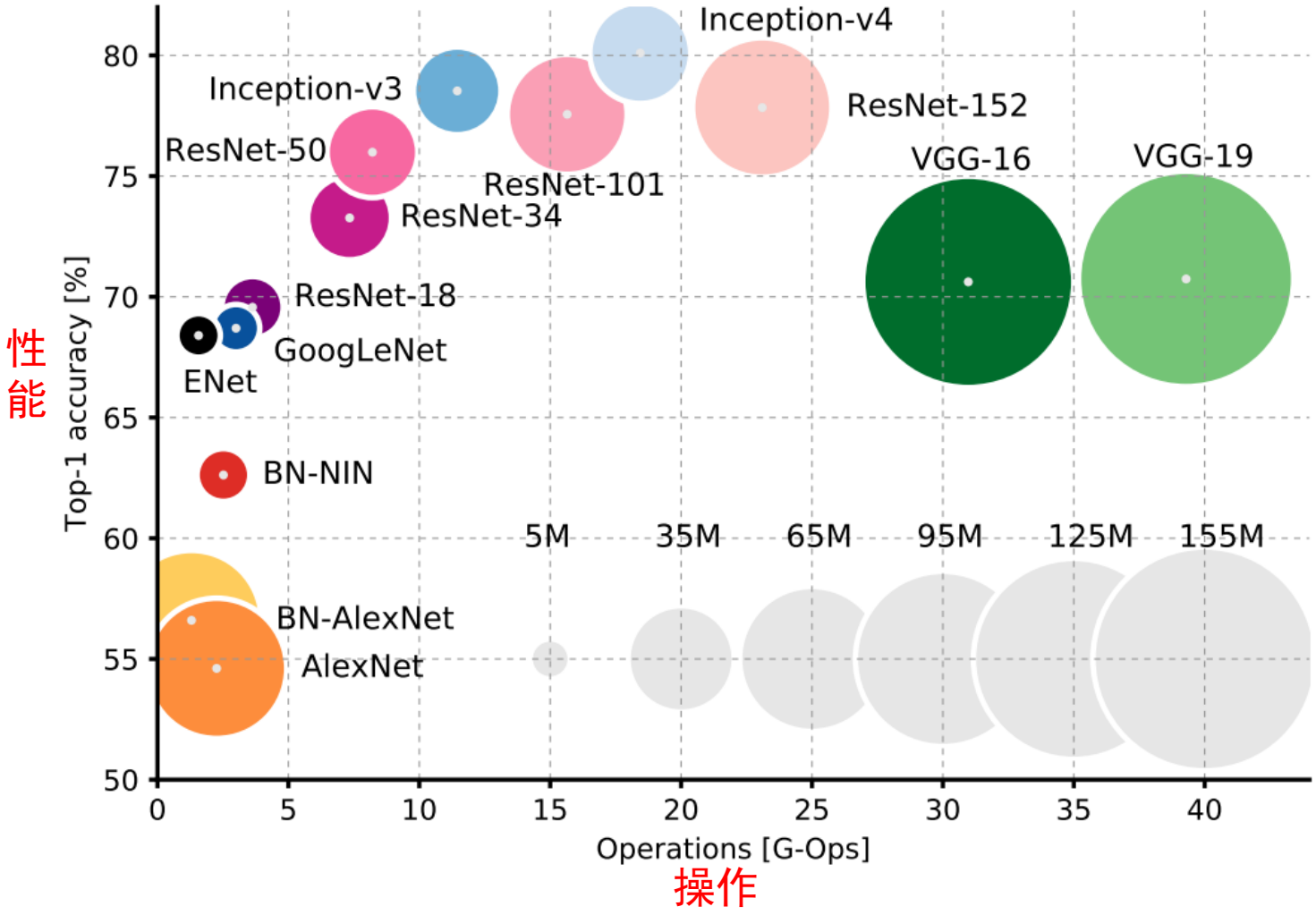
[He et al. CVPR 2016]

Identity Shout Cut
Better Grad. Prop.
Very Deep

深度网络的重要性

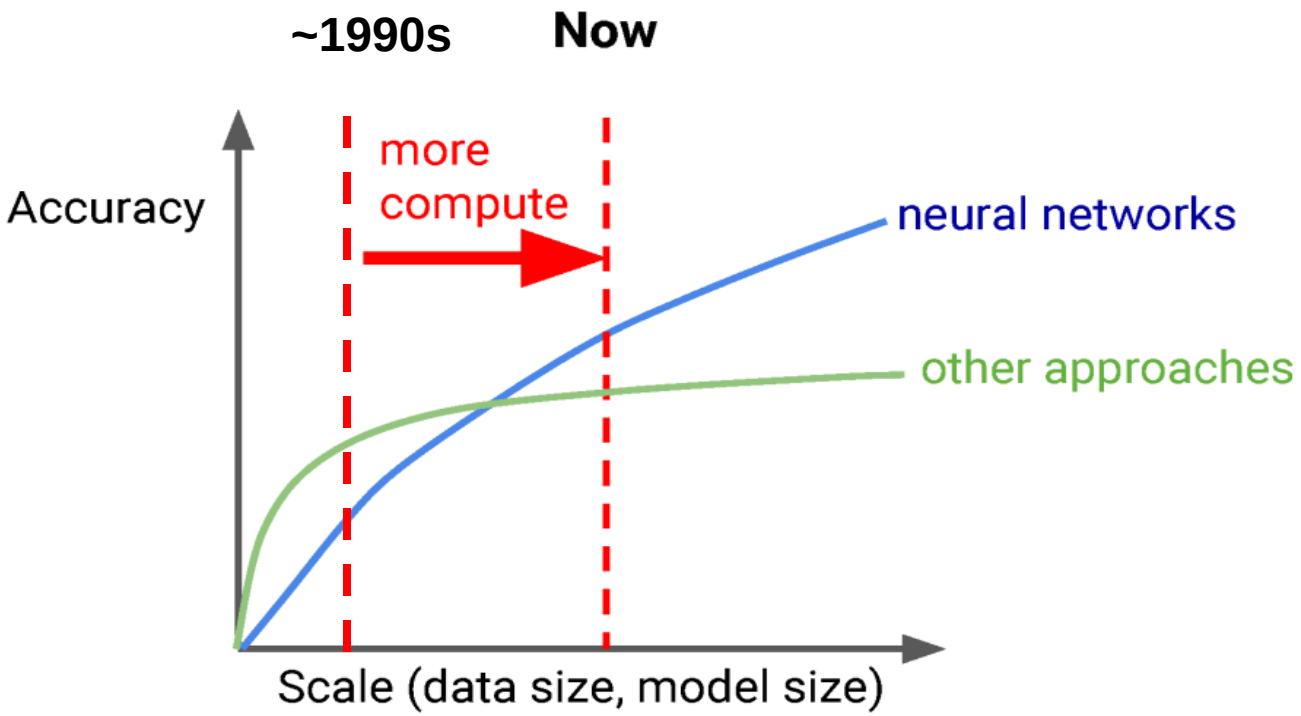
- Optimization (Gradient Propagation):
  - ReLU
  - Batch Normalization
  - Stage-wise training (or better initialization)
  - Identity Connection in ResNet
- Overfitting
  - Dropout
  - Data Argumentation
- Architecture Design (Modularized design)
  - 3x3 Layer in VGG
  - Inception Module
  - ResNet Block
- Light parameters
  - 3x3 Conv in VGG
  - 1x1 Conv + Bottleneck

深度网络的结构


Inception-v4
Inception-v3
ResNet-152
ResNet-50
VGG-16
VGG-19
ResNet-101
ResNet-34
ResNet-18
GoogLeNet
ENet
BN-NIN
5M
35M
65M
95M
125M
155M
BN-AlexNet
AlexNet
Top-1 accuracy [%]
80
75
70
65
60
55
50
0
5
10
15
20
25
30
35
40
Operations [G-Ops]
性能
操作

大数据和深度学习

- 小数据学习：过拟合，降低模型复杂度，加正则项
- 大数据学习：欠拟合、增加模型复杂度，优化，计算资源

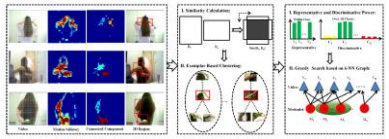


From J. Dean

我的视觉旅程

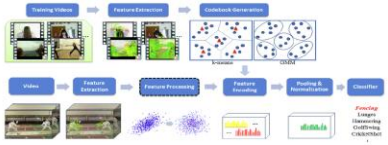
视频行为识别与理解

从视频行为理解和识别是计算机视觉的基本和热点问题，在监控、互联网等有着广泛的应用。在CVPR，ICCV，IJCV等顶级视觉会议和期刊发表了近30篇论文，其中2篇论文分别被ICCV和CVPR录用为Oral。



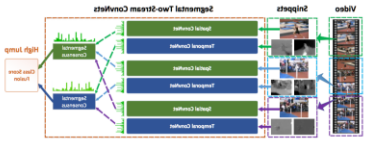
视频中层表示和结构模型

CVPR 2013，160引用
ICCV 2013，67引用
ECCV 2014，83引用
IJCV 2016



视频特征编码学习

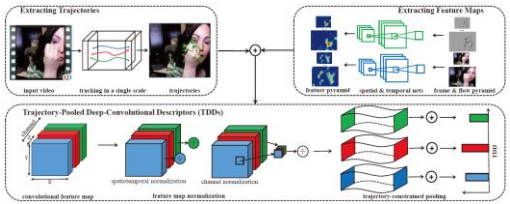
CVPR 2014 Oral， 110引用
ECCV 2014，247引用
CVIU 2019，360 引用
非深度学习领域UCF和HMDB性能最好方法



时序分割模型TSN

ECCV 2016, 485 引用
解决显存限制下，视频端到端的训练
ActivityNet 2016竞赛第一
被广泛应用

2013 2014 2015 2016 2017

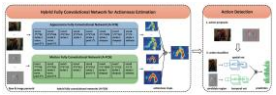


轨迹卷积特征TPD

CVPR 15， 547引用
第一个在UCF和HMDB数据库上全面超越传统方法的的深度模型
技术转移到华为

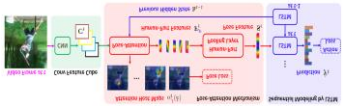
运动向量网络EMV-CNN

CVPR 2016，119引用
深度实时行为识别模型



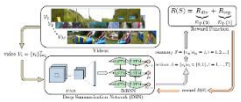
行为显著性估计

CVPR 2016
行为检测与显著性估计



姿态递归注意网络RPAN

ICCV2017 Oral
利用姿态估计引导动态行为序列建模



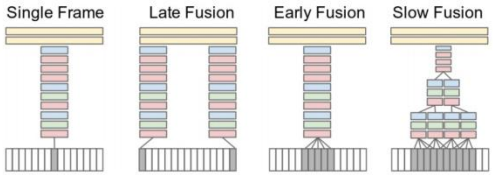
非监督强化视频概要

AAAI 2018 Oral

早期视频行为识别DL方法

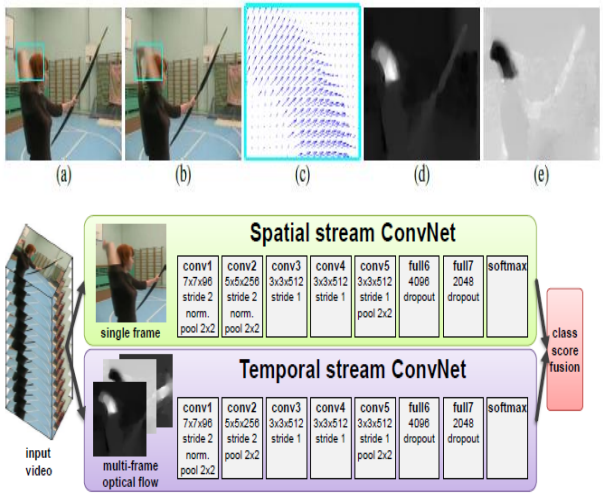
Spatial Temporal CNN

Sports-1M Dataset



[Karpathy et al., CVPR, 2014]

Two Stream-CNN



[Karen NIPS, 2014]

C3D: 3D VGGNet

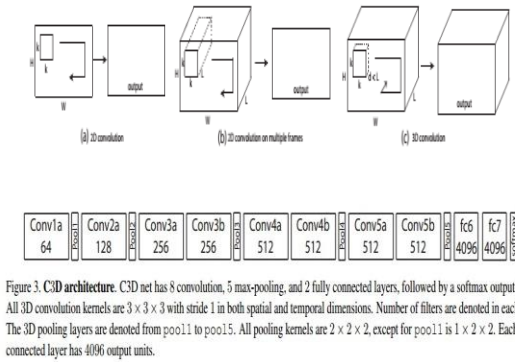
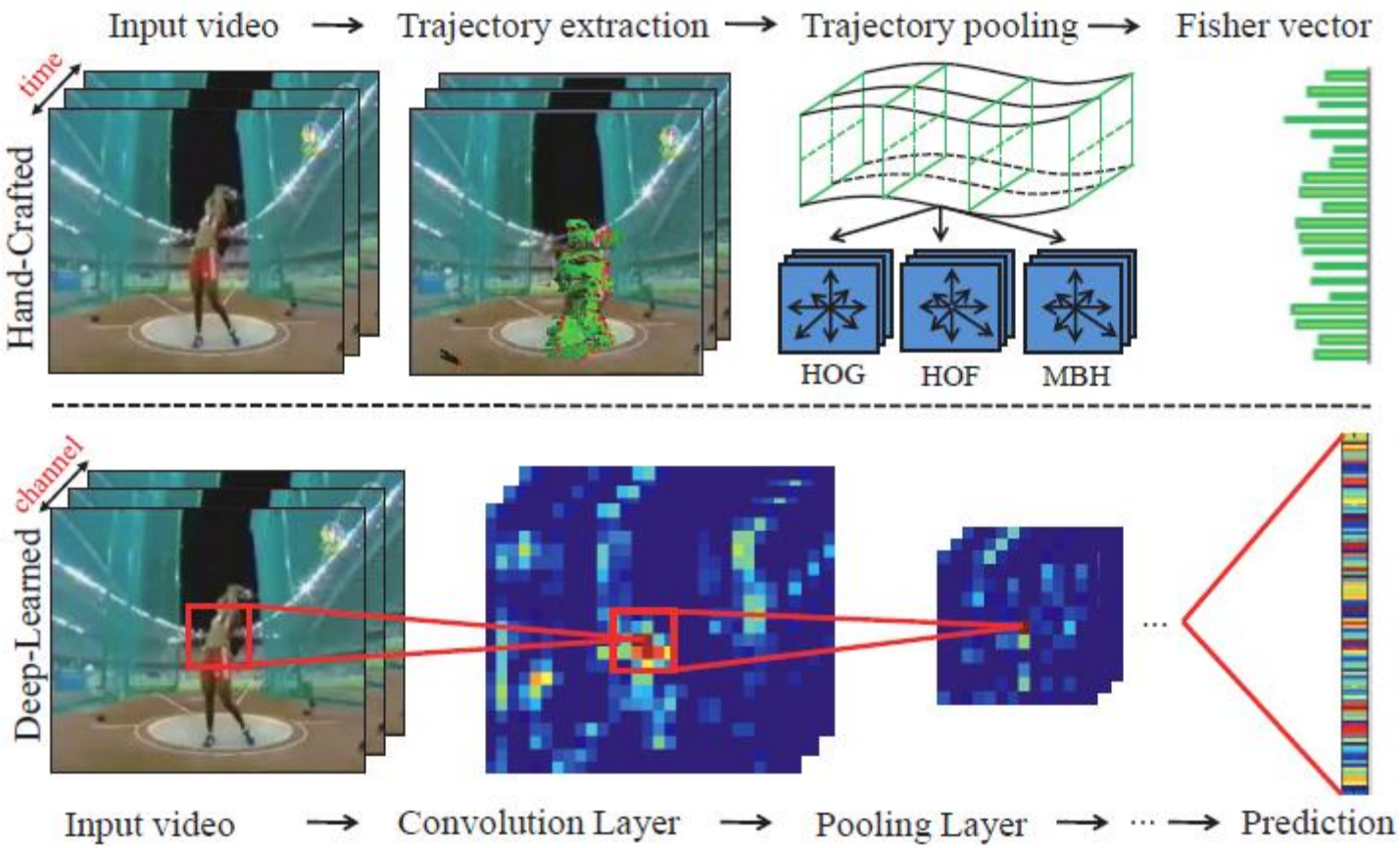


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

[Tran et al. CVPR 2015]

在UCF101的表现并没有明显好于非传统方法

工作1：轨迹池化卷积特征TDD(CVPR15)

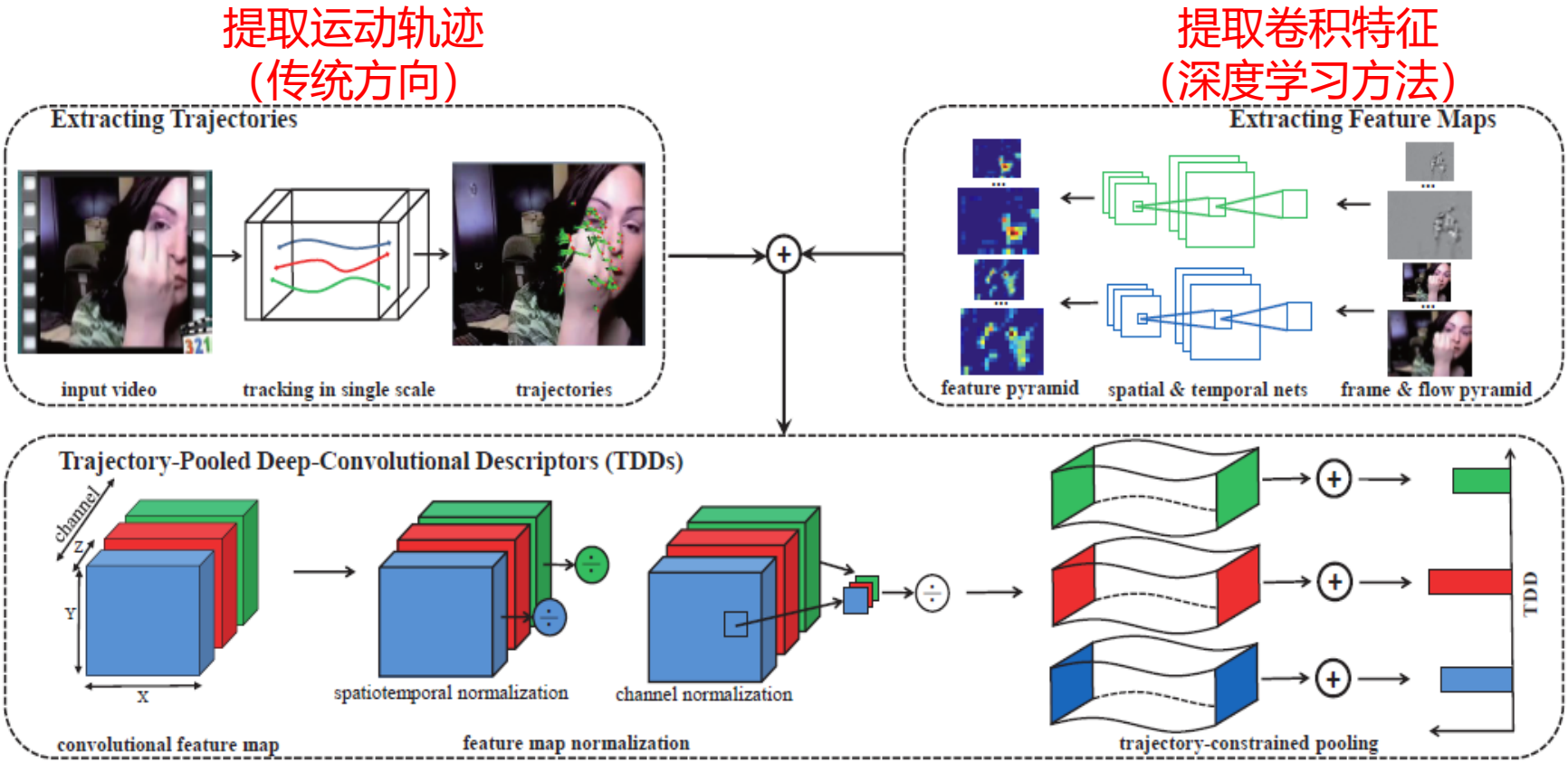


如何利用传统方法与深度学习方法的优势。

Limin Wang, Yu Qiao, Xiaoou Tang "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015

TDD的框架

Trajectory-pooled deep convolutional descriptor (TDD) 特征结合了传统方法的轨迹跟踪和深度学习方法的卷积特征提取。



沿着运动轨迹对特征进行编码

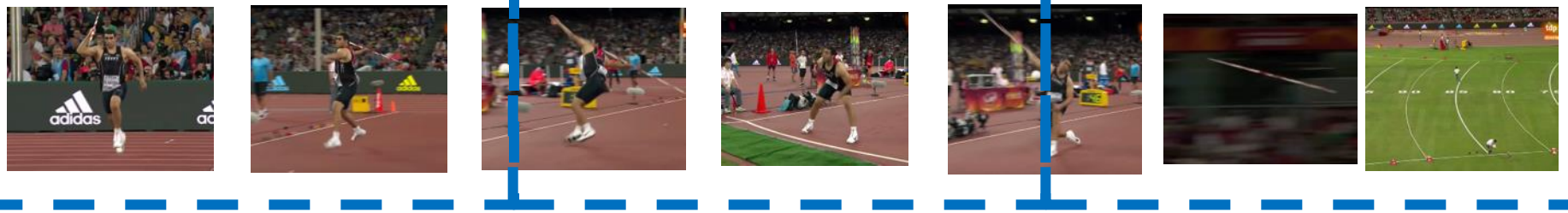
TDD的性能

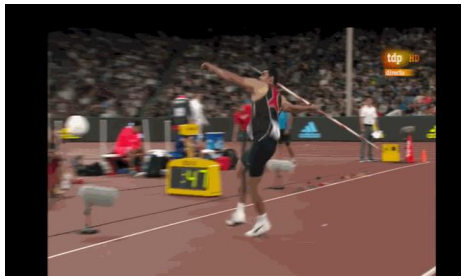
第一个在UCF和HMDB上全面超越传统浅层模型的深度学习方法。

Algorithm	HMDB51	UCF101
HOG [31, 32]	40.2%	72.4%
HOF [31, 32]	48.9%	76.0%
MBH [31, 32]	52.1%	80.8%
HOF+MBH [31, 32]	54.7%	82.2%
iDT [31, 32]	**57.2%**	**84.7%**
Spatial net [24]	40.5%	73.0%
Temporal net [24]	54.6%	83.7%
Two-stream ConvNets [24]	**59.4%**	**88.0%**
Spatial conv4	48.5%	81.9%
Spatial conv5	47.2%	80.9%
Spatial conv4 and conv5	**50.0%**	**82.8%**
Temporal conv3	54.5%	81.7%
Temporal conv4	51.2%	80.1%
Temporal conv3 and conv4	**54.9%**	**82.2%**
TDD	63.2%	90.3%
TDD and iDT	**65.9%**	**91.5%**

工作2：深度时序分割模型TSN (ECCV 16)

如何对视频序列进行建模和深度学习？



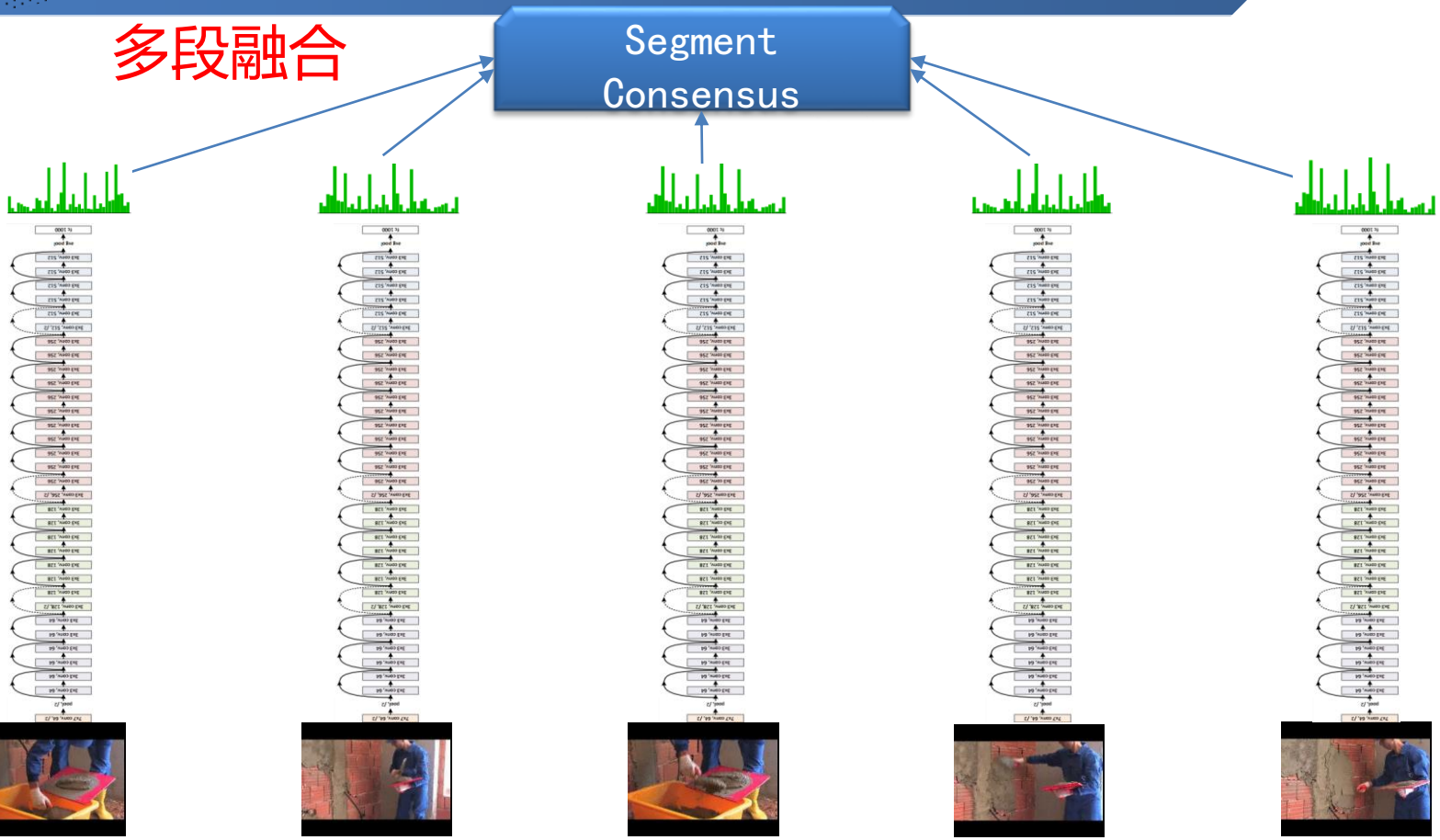


核心问题：视频的数据量大，特征维度很高，但深度学习的训练受制于显存和SGD算法。

1.Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang,Luc Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," Proc. European Conference Computer Vision (ECCV), 2016

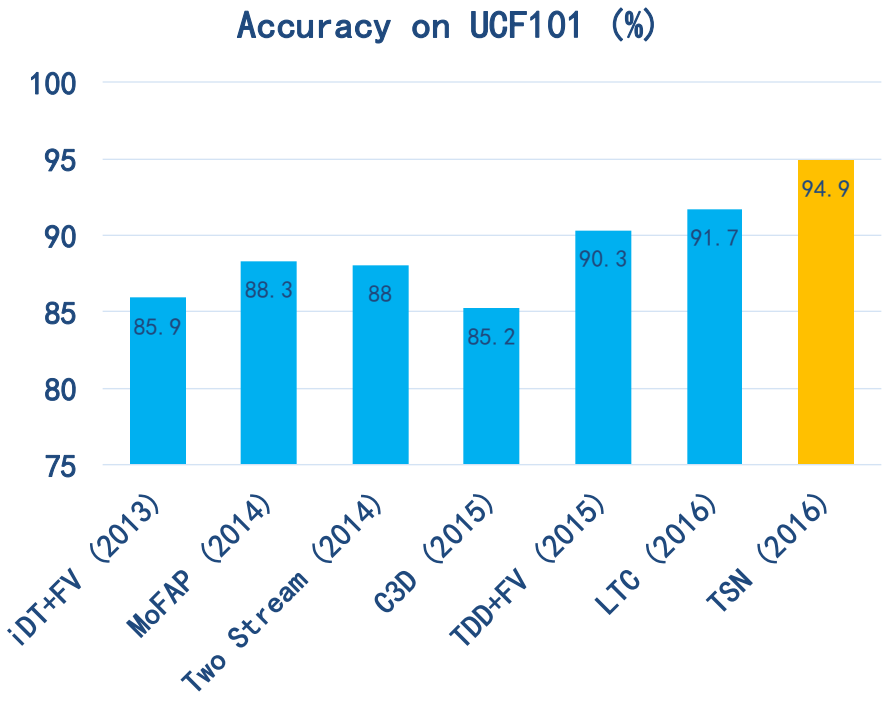
TSN框架

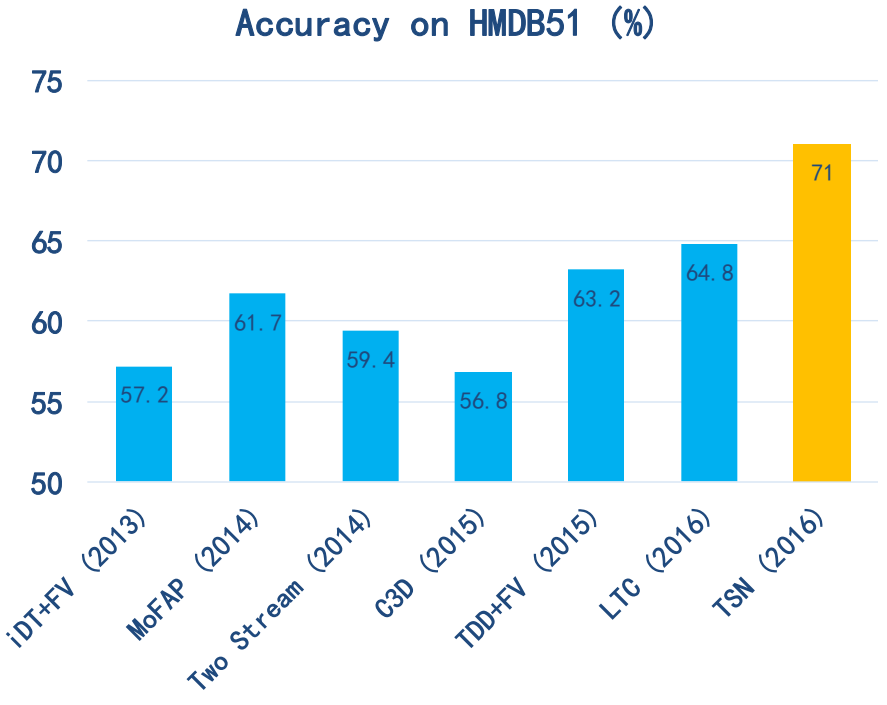
多段融合



分段
采样

TSN的性能评价


Accuracy on UCF101 (%)
100
95
90
85
80
75
85.9
88.3
88
85.2
90.3
91.7
94.9
iDT+FV (2013)
MoFAP (2014)
Two Stream (2014)
C3D (2015)
TDD+FV (2015)
LTC (2016)
TSN (2016)


Accuracy on HMDB51 (%)
75
70
65
60
55
50
57.2
61.7
59.4
56.8
63.2
64.8
71
iDT+FV (2013)
MoFAP (2014)
Two Stream (2014)
C3D (2015)
TDD+FV (2015)
LTC (2016)
TSN (2016)

ActivityNet 2016

ACTIVITYNET

200个类别，648小时视频，10k训练，5k测试

http://activity-net.org/challenges/2016/

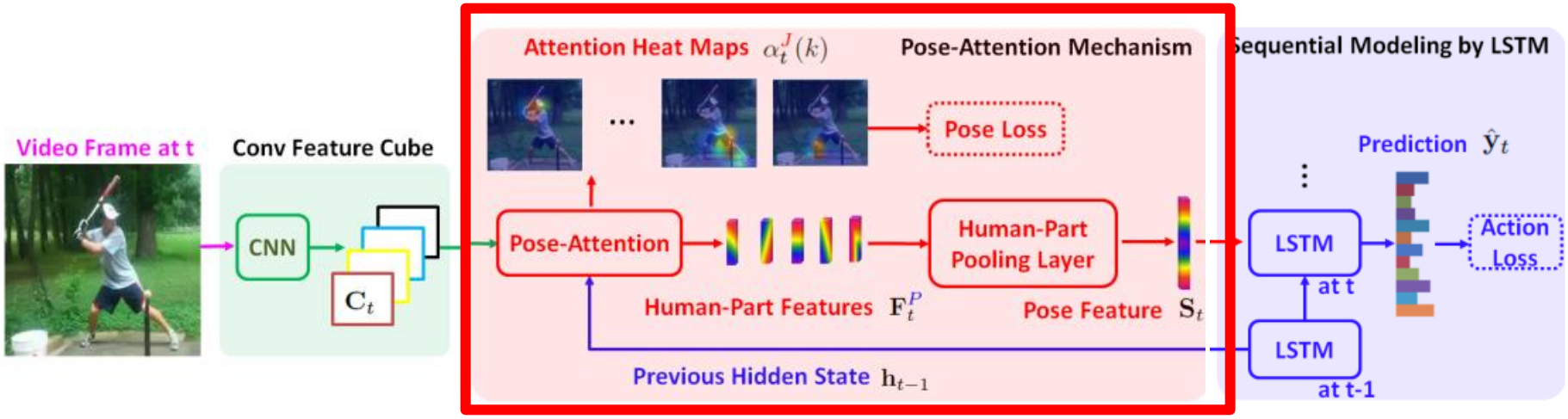
在24个队中排名第一。

Validation Set	mAp	Top-3 Acc.
Visual	90.4%	95.2%
Audio	15.2%	29.1%
Visual + Audio	90.9%	95.6%

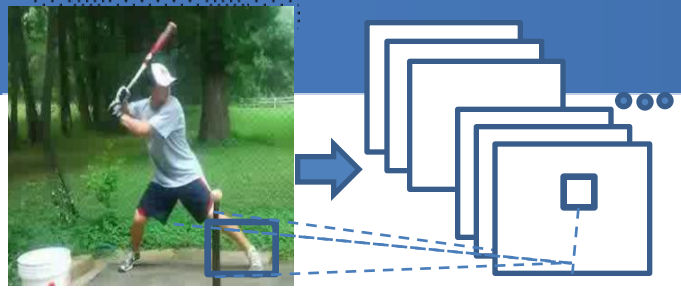
Testing Set	mAP	Top-3 Acc.
Visual CNN (Single)	91.2%	95.6%
Final Ensemble	93.2%	96.4%

工作3：递归姿态注意网络RPAN（ICCV17 Oral）

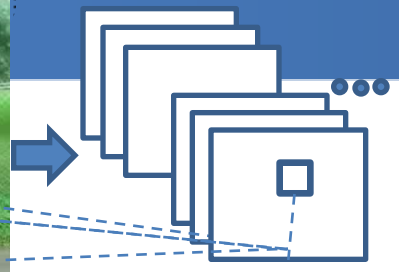
提出姿态注意机制RPAN对行为的动态过程进行建模。

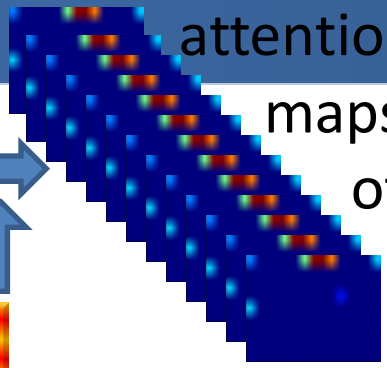


- 把行为识别和姿态估计两个任务进行结合。
- 利用姿态的变化，引导递归神经网络对行为的动态过程进行建模。



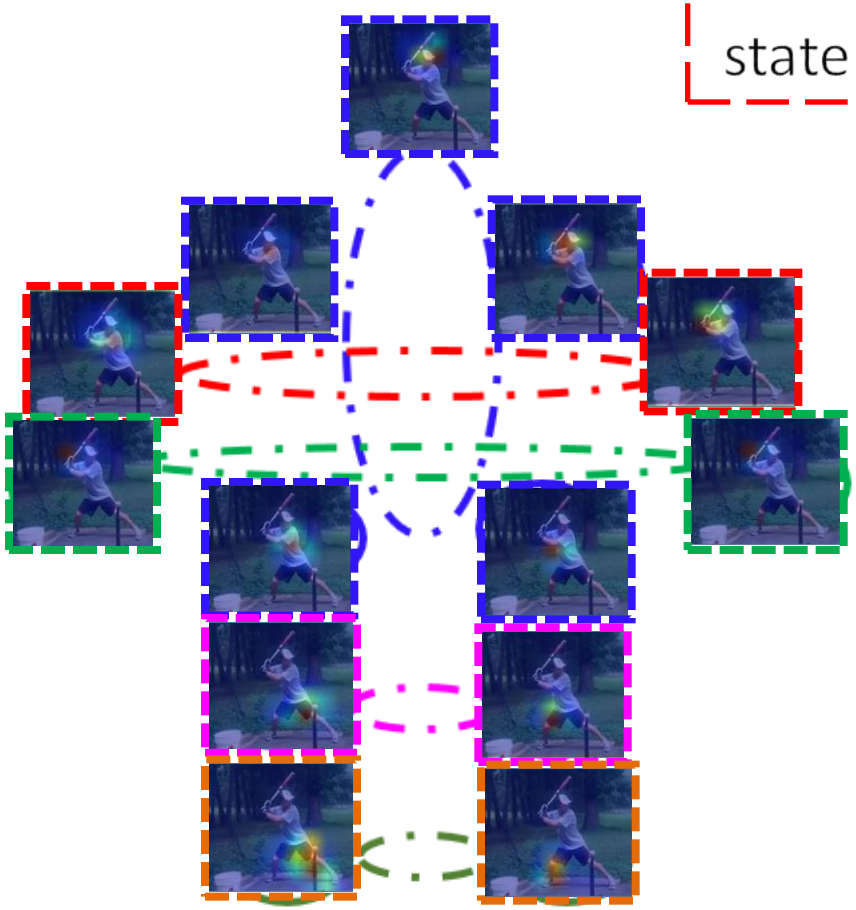
Pose Attention Mechanism





attention maps of k joints

LSTM hidden state at (t-1)



Torso

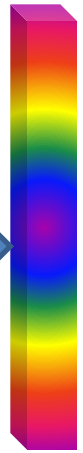
Elbow

Wrist

Knee

Ankle

mean/ max/ concat

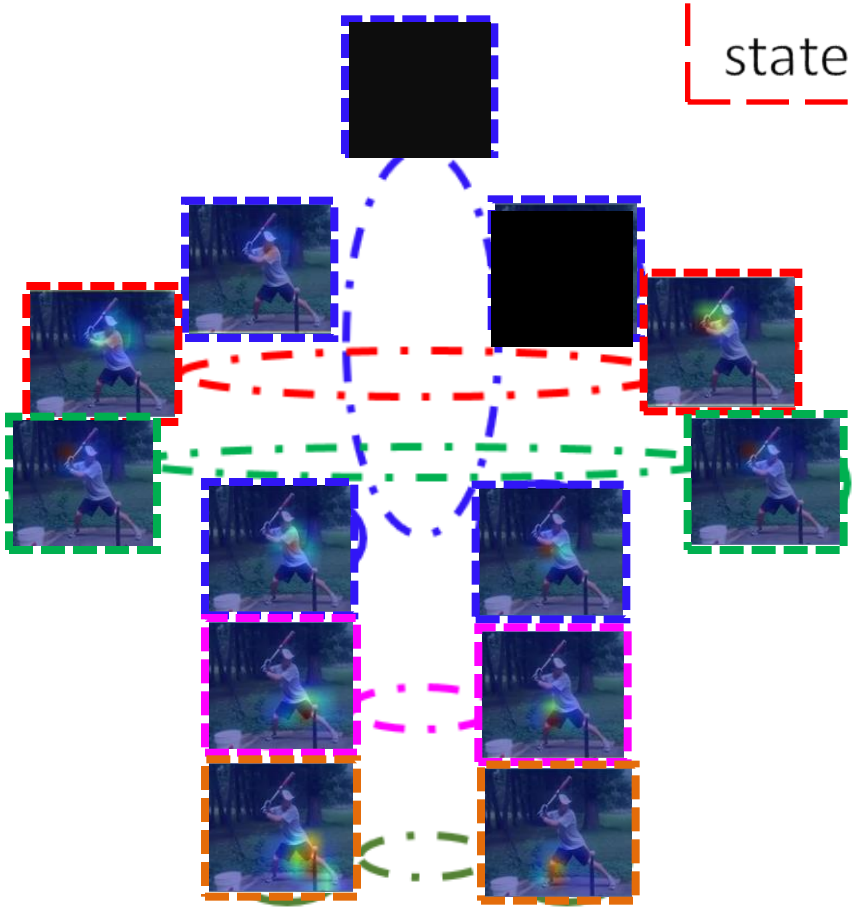


Pose featur

Pose Attention Mechanism

attention maps of k joints

LSTM hidden state at (t-1)



Torso

Elbow

Wrist

Knee

Ankle

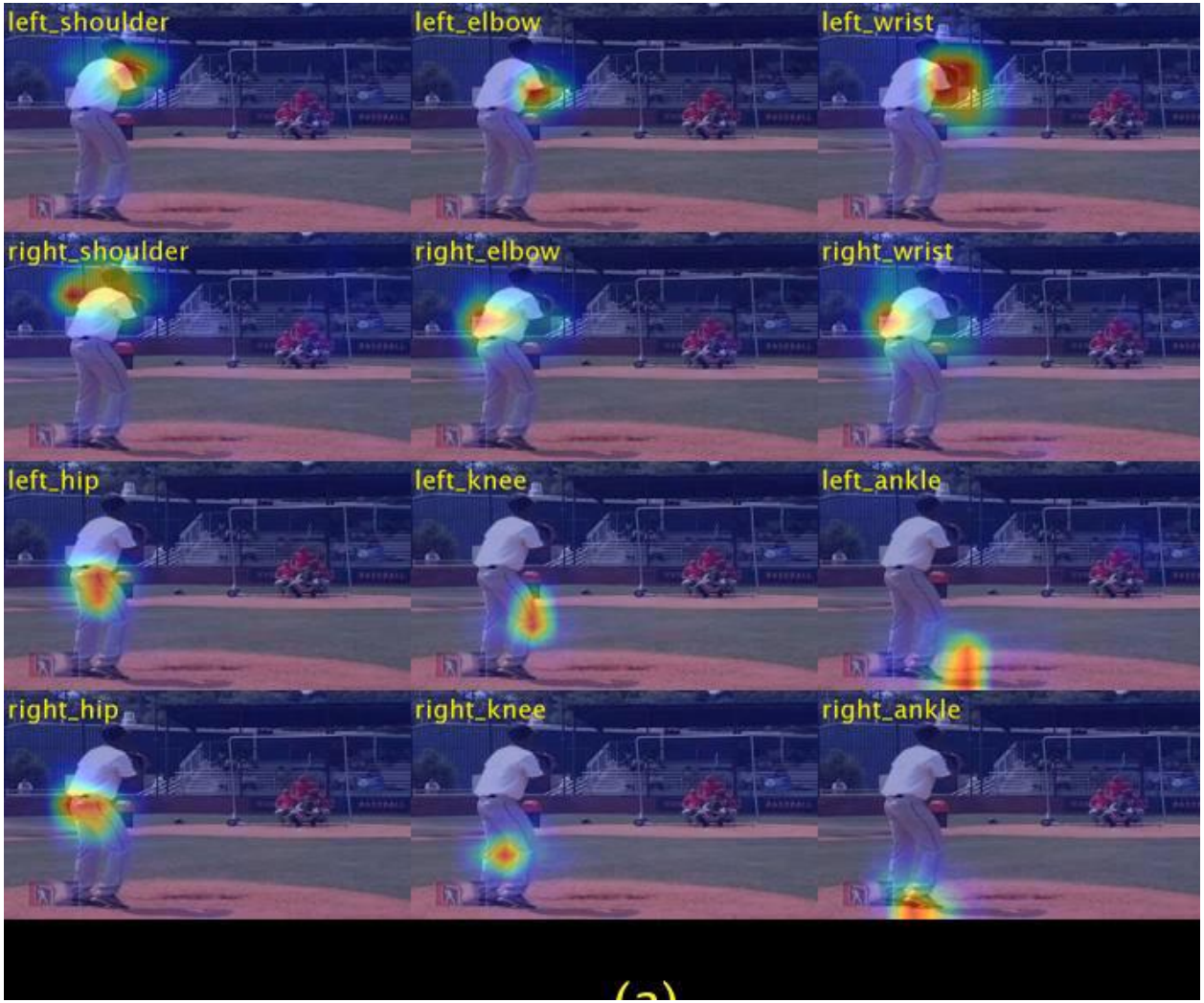
mean/ max/ concat

Pose feature

RPAN的性能

State-of-the-art	Authors	Year	Sub-JHMDB	PennAction
Dense+Pose	H. Jhuang, et al	2013	52.9	-
STIP	W. Zhang, et al	2013	-	82.9
Action Bank	W. Zhang, et al	2013	-	83.9
MST	J. Wang, et al	2014	45.3	74.0
AOG	B. X. Nie, et al	2015	61.2	85.5
P-CNN	G. Cheron et a	2015	66.8	-
Hierarchical	I. Lillo et al	2016	77.5	-
C3D	C. Cao, et al	2016	-	86.0
JDD	C. Cao, et al	2016	77.7	87.4
idt-fv	U. Iqbal et al	2017	60.9	92.0
Pose+ idt-fv	U. Iqbal et al	2017	74.6	92.9
Our RPAN			**78.6**	**97.4**

RPAN用于姿态估计


left_shoulder
left_elbow
left_wrist
right_shoulder
right_elbow
right_wrist
left_hip
left_knee
left_ankle
right_hip
right_knee
right_ankle
(a)

课题组部分论文

人脸分析与识别

共同判别分析	TIP 14
深度隐因子	CVPR 16
性别表情	CVPR 16
多任务级联	SPL 16
迁移性别	SPL 16
中心损失	ECCV 16
长尾识别	ICCV 17
层内过滤	ICCV 17

…

视频行为识别

运动词组	CVPR 13	轨迹卷积特征	CVPR 15
行为短语	ICCV 13	运动向量CNN	CVPR 16
隐层次模型	TIP 14	关键块挖掘	CVPR 16
动态姿态模型	ECCV 14	行为显著	CVPR 16
多视角编码	CVPR 14	时序分割模型	ECCV 16
堆栈FV编码	ECCV 14	递归姿态注意	ICCV 17
融合超向量	CVIU 16	注意序列模型	TIP 17
多层运动表示	IJCV 16	强化视频概要	AAAI 18
场景事件迁移	IJCV 18	动态运动迁移	CVPR 18

…

场景分类与理解

共生二值特征	PAMI 14
深度MSER	ECCV 14
递归系列识别	AAAI 16
连接序列候选	ECCV 16
知识引导学习	TIP 17
图块网络	TIP 17
局部监督网络	TIP 17
文本注意引导	ICCV 17
注意对齐框架	CVPR 18

…

模型和代码公开

场景理解与分类

•MR-CNNs (2nd in scene classification task ImageNet 2016, 1st in LSUN 2016)

•Weakly Supervised PatchNets (Top performance in MIT Indoor67 and SUN397)

行为识别和检测

•Temporal Segment Networks (NO1 in ActivityNet 2016)

•MV-CNNS(Speed:300帧/s)

•Trajectory-Pooled Deep-Convolutional Descriptors （Top performance in UCF101 and HMDB51)

人脸检测与识别

•MJ-CNN face detection (top performance in FDDB & WIDE)

•HFA-CNN face recognition (single model 99% in LFW)

场景文字检测与识别

•Connectionist Text Proposal Network for Scene Text Detection (Top performance in ICDAR)

下载地址

http://mmlab.siat.ac.cn/yuqiao/Codes.html