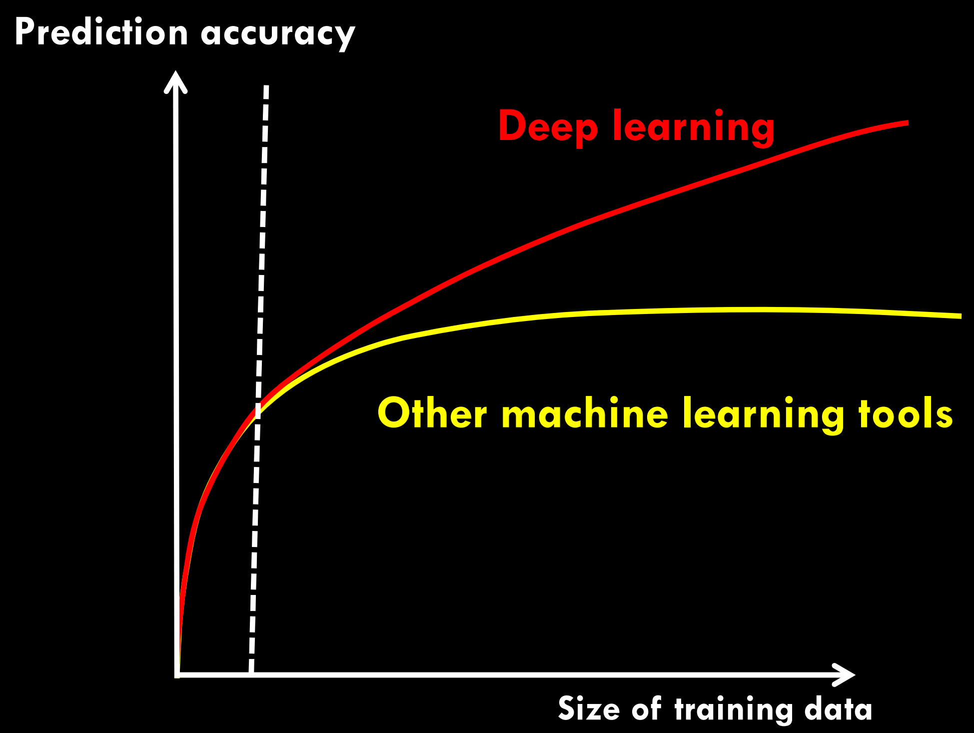
Deep Learning in Video Surveillance

Xiaogang Wang

The Chinese University of Hong Kong

Machine learning with big data

- Machine learning with small data: overfitting, reducing model complexity (capacity), adding regularization
- 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)

How to learn feature representation?

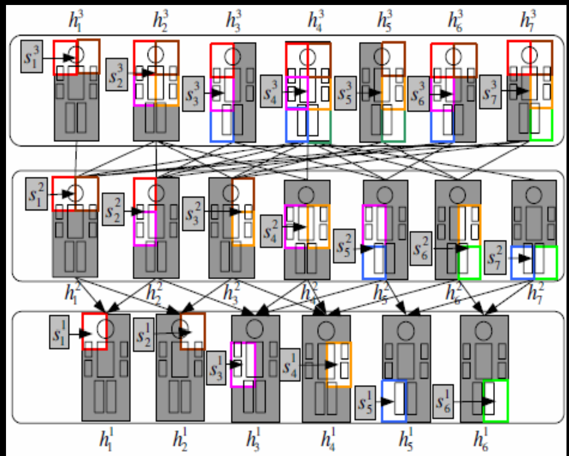
How to design network structures?

Outline

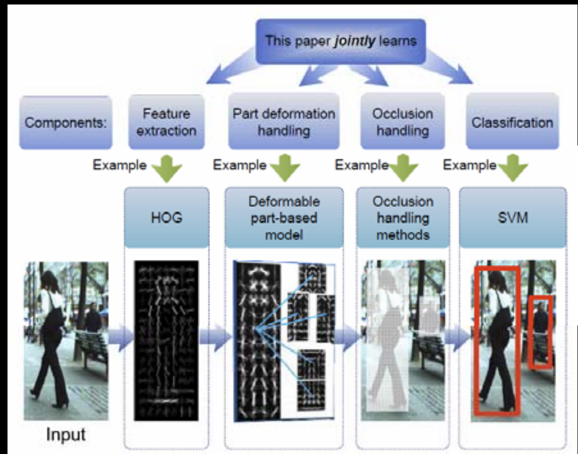
- Pedestrian detection
- Object tracking
- Crowd understanding

Pedestrian detection

Improve state-of-the-art average miss detection rate on the largest Caltech dataset from 63% to 11%

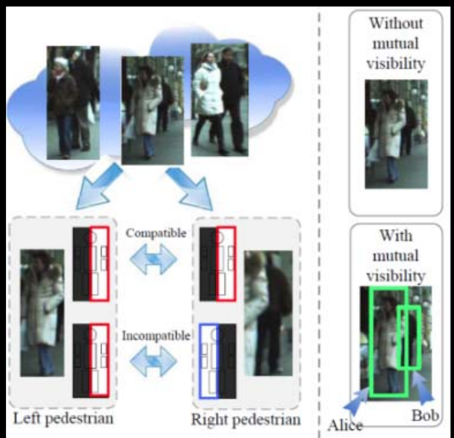


CVPR'12

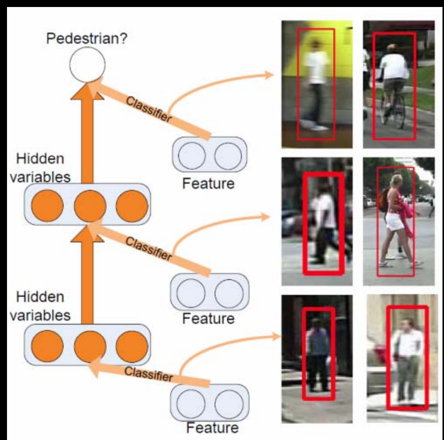


ICCV'13

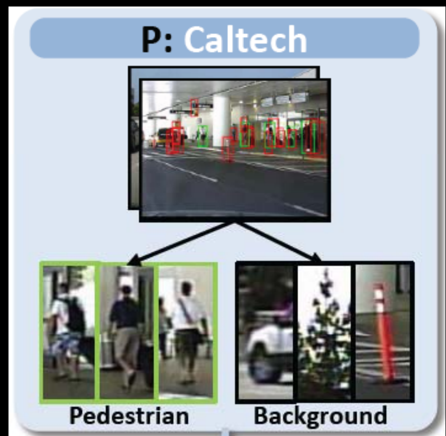
CVPR'14



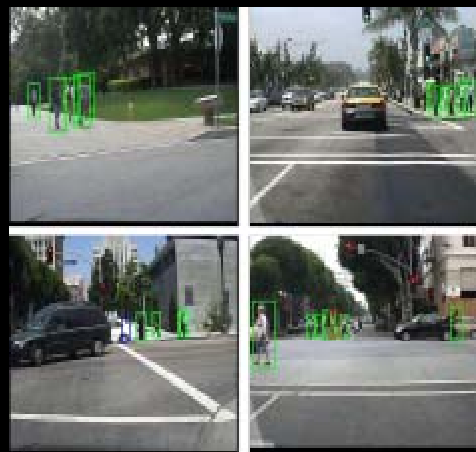
CVPR'13



ICCV'13

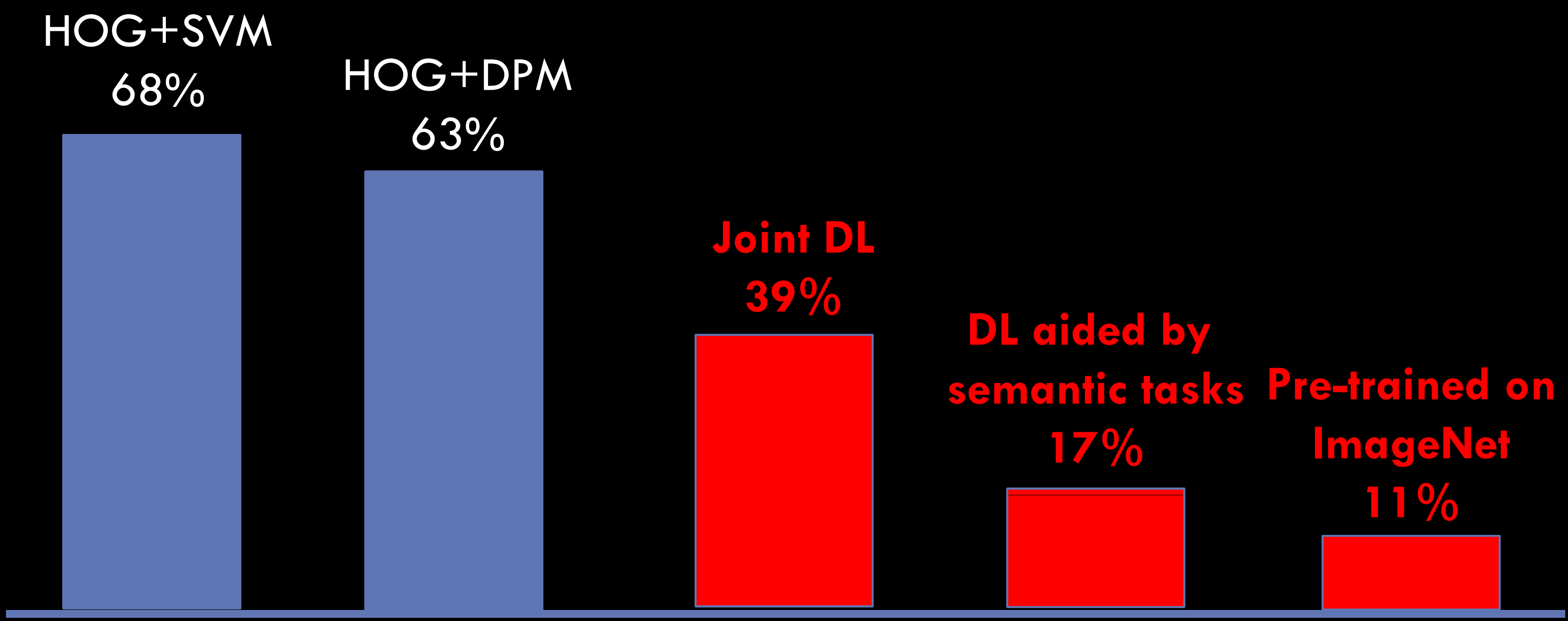


CVPR'15



ICCV'15

Pedestrian detection on Caltech (average miss detection rates)



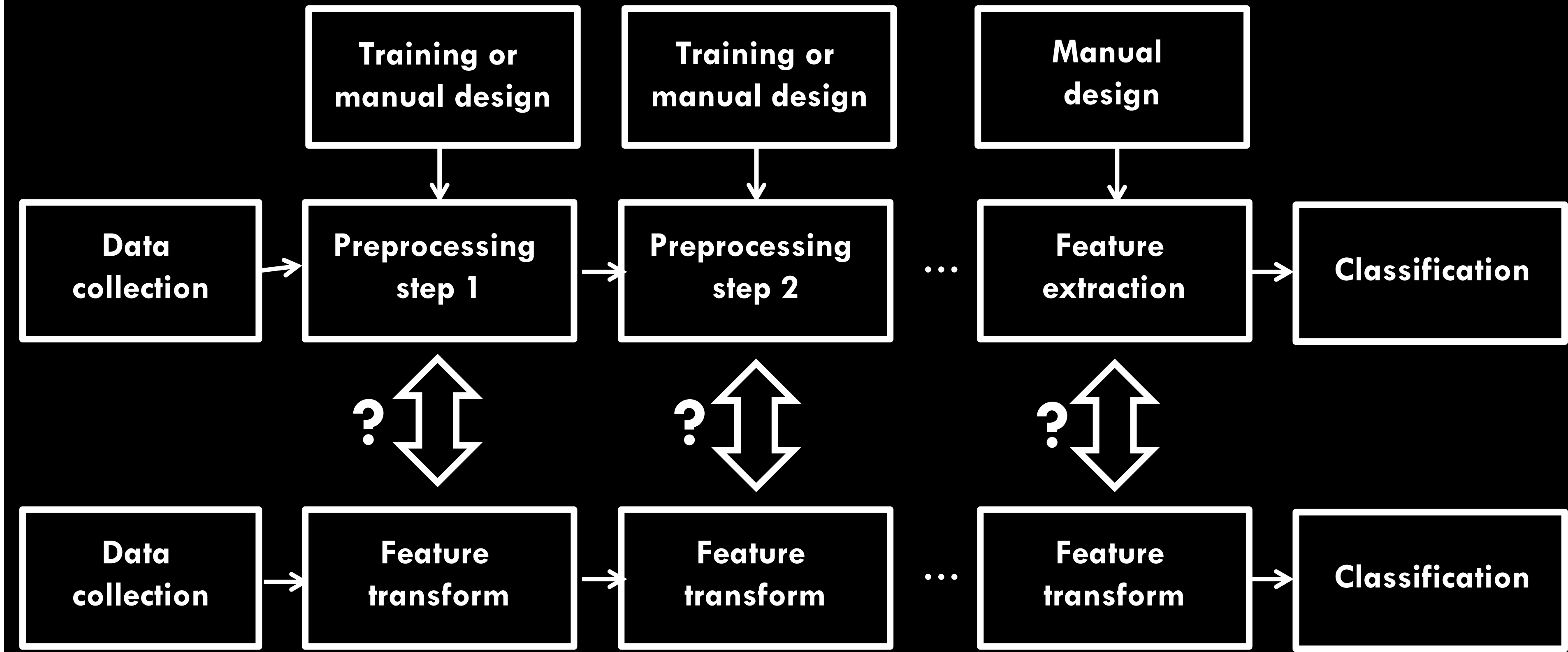
W. Ouyang and X. Wang, “Joint Deep Learning for Pedestrian Detection,” ICCV 2013.

Y. Tian, P. Luo, X. Wang, and X. Tang, “Pedestrian Detection aided by Deep Learning Semantic Tasks,” CVPR 2015.

Y. Tian, P. Luo, X. Wang, and X. Tang, “Deep Learning Strong Parts for Pedestrian Detection,” ICCV 2015.

Is deep model a black box?

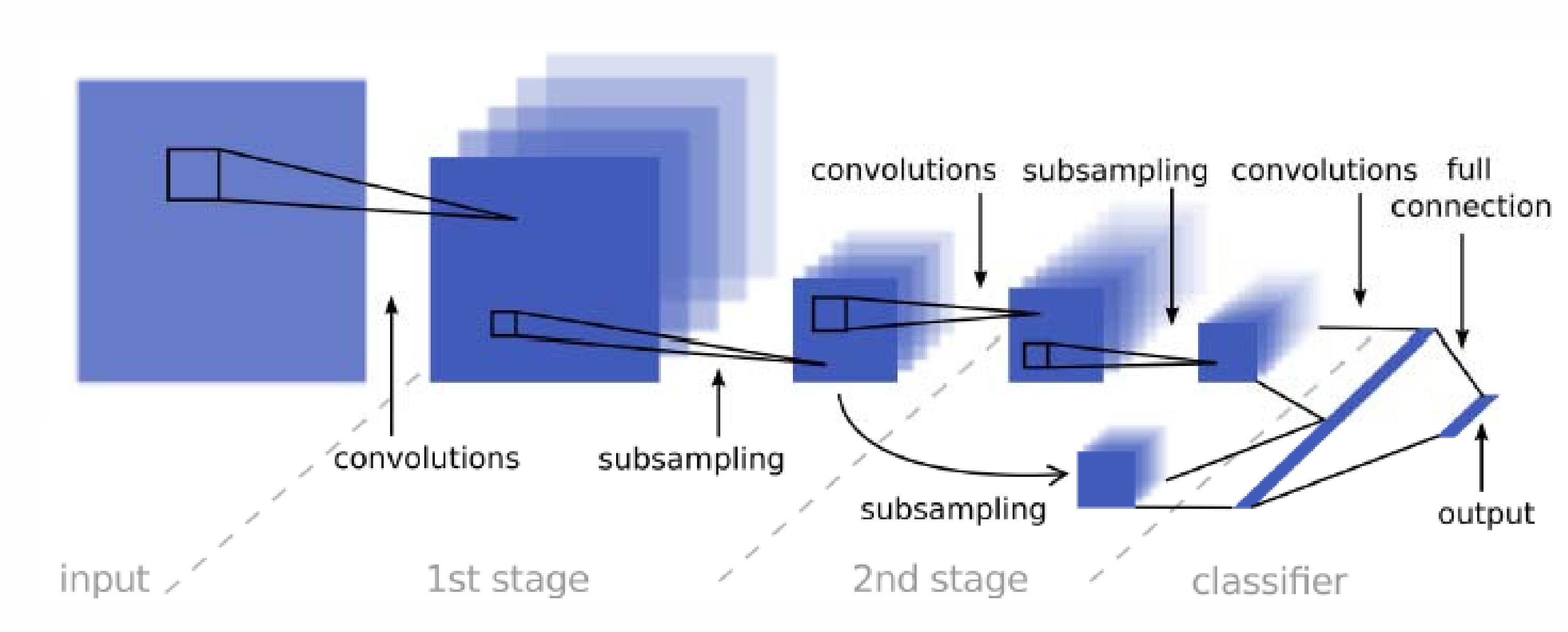
Joint learning vs separate learning



End-to-end learning

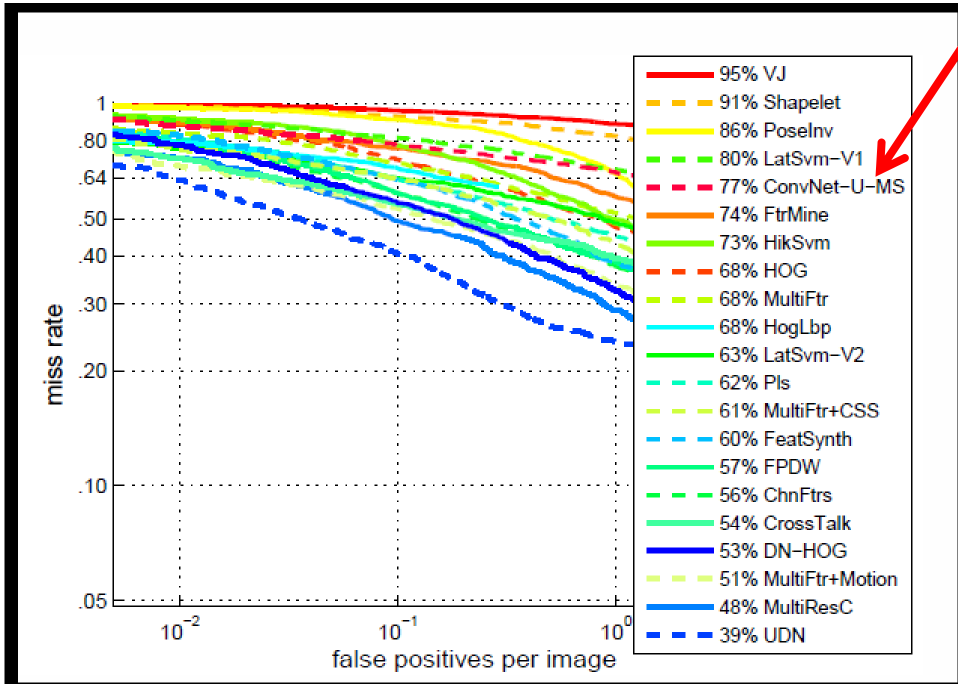
Deep learning is a framework/language but not a black-box model

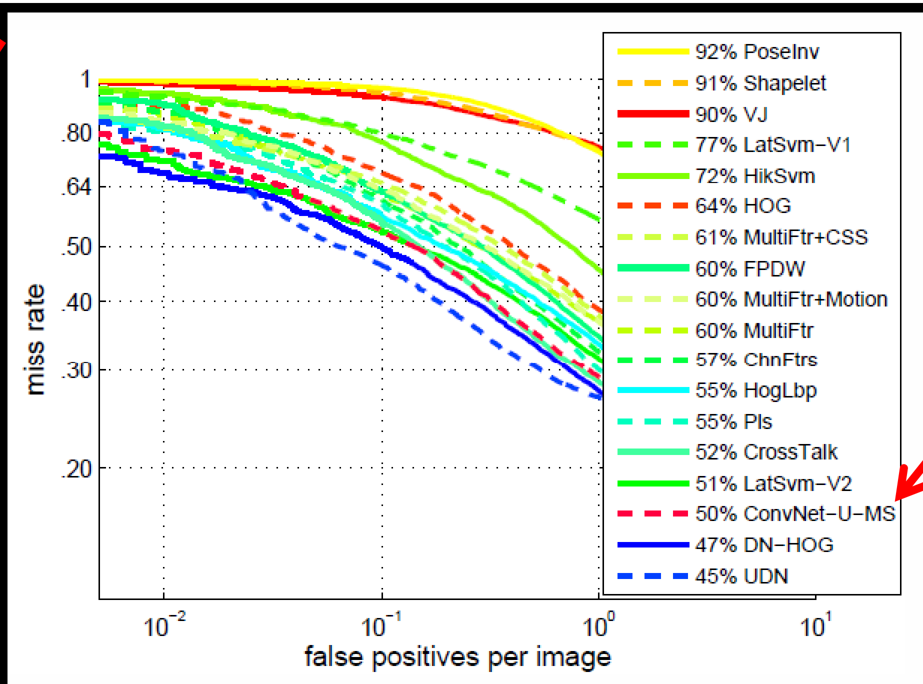
Its power comes from joint optimization and
increasing the capacity of the learner



ConvNet−U−MS

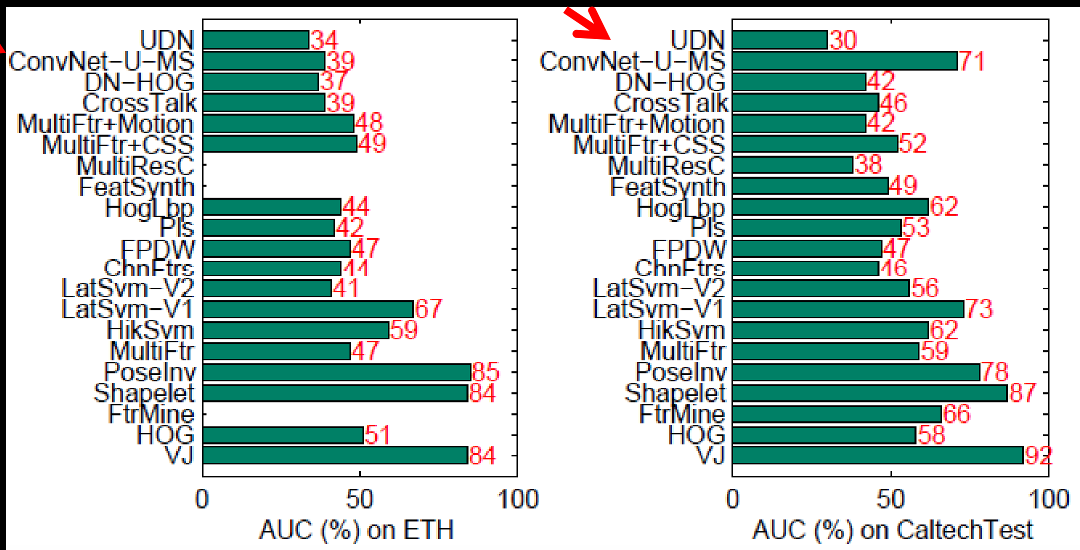
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” CVPR 2013.

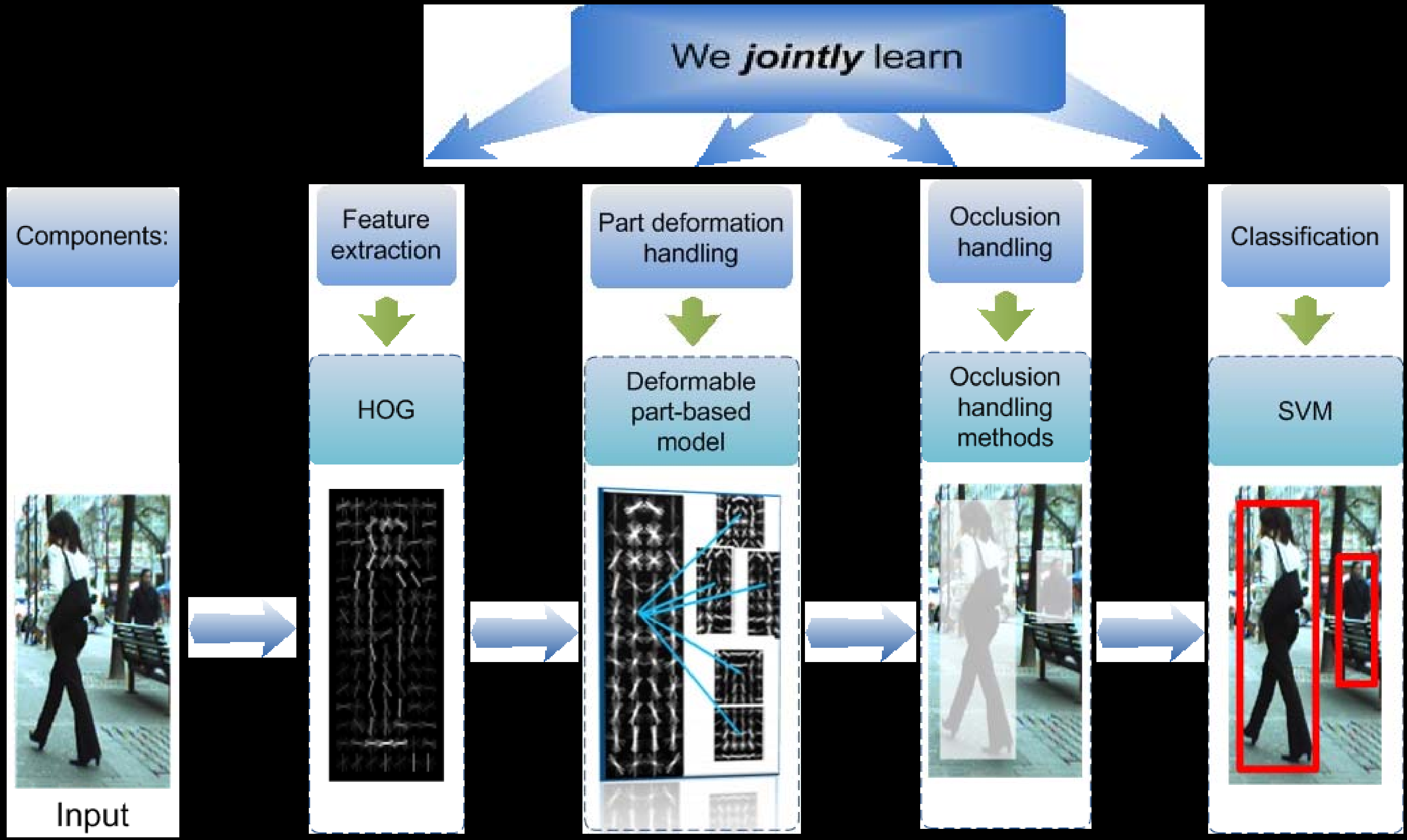




Results on Caltech Test

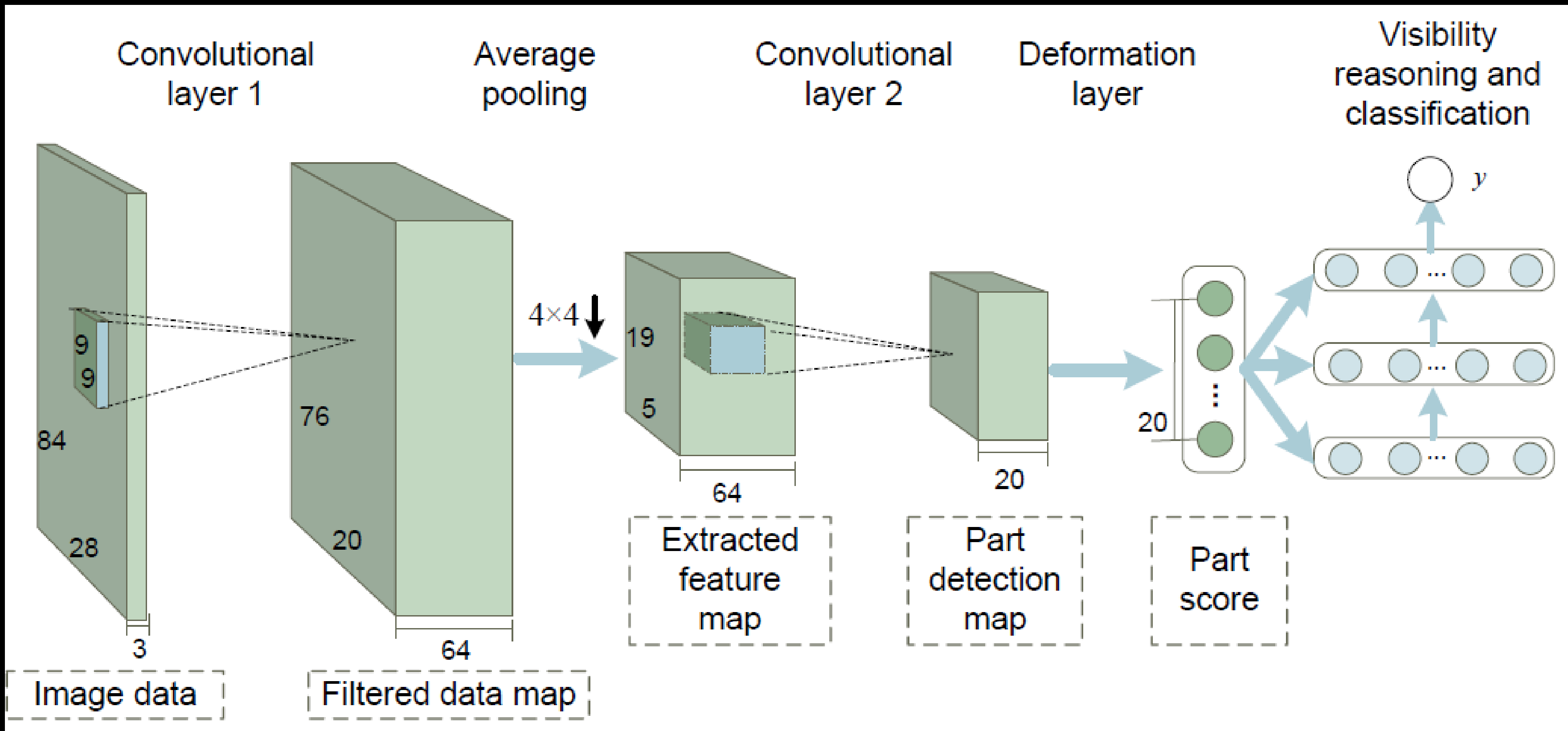
Results on ETHZ





- 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.

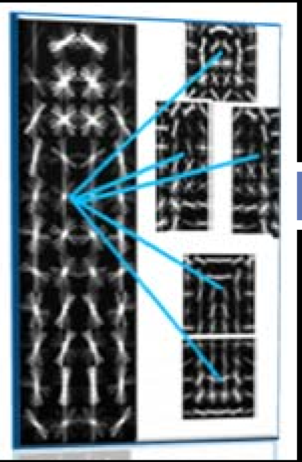
Our joint deep learning model



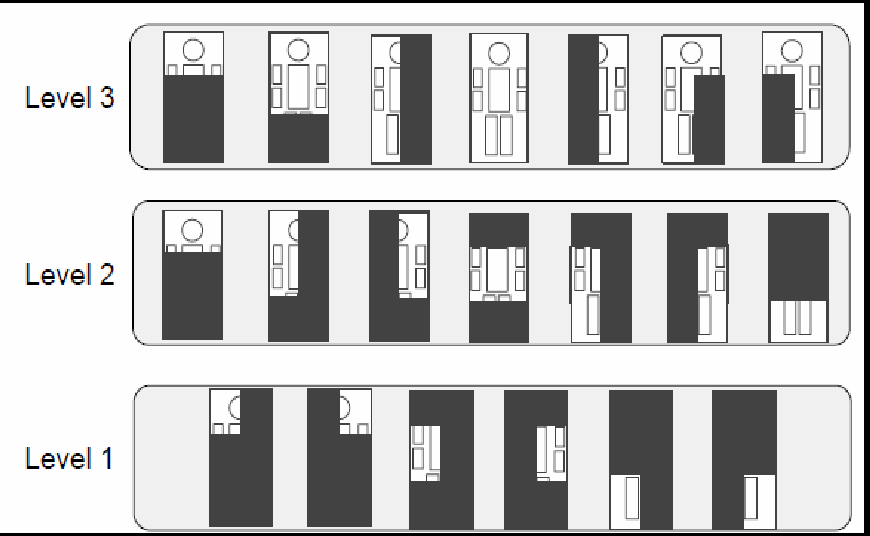
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling part detectors

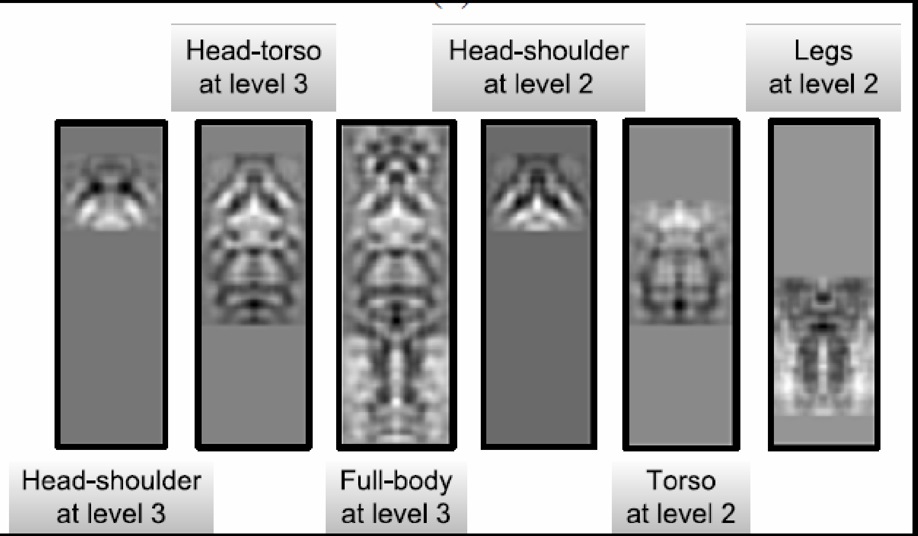
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

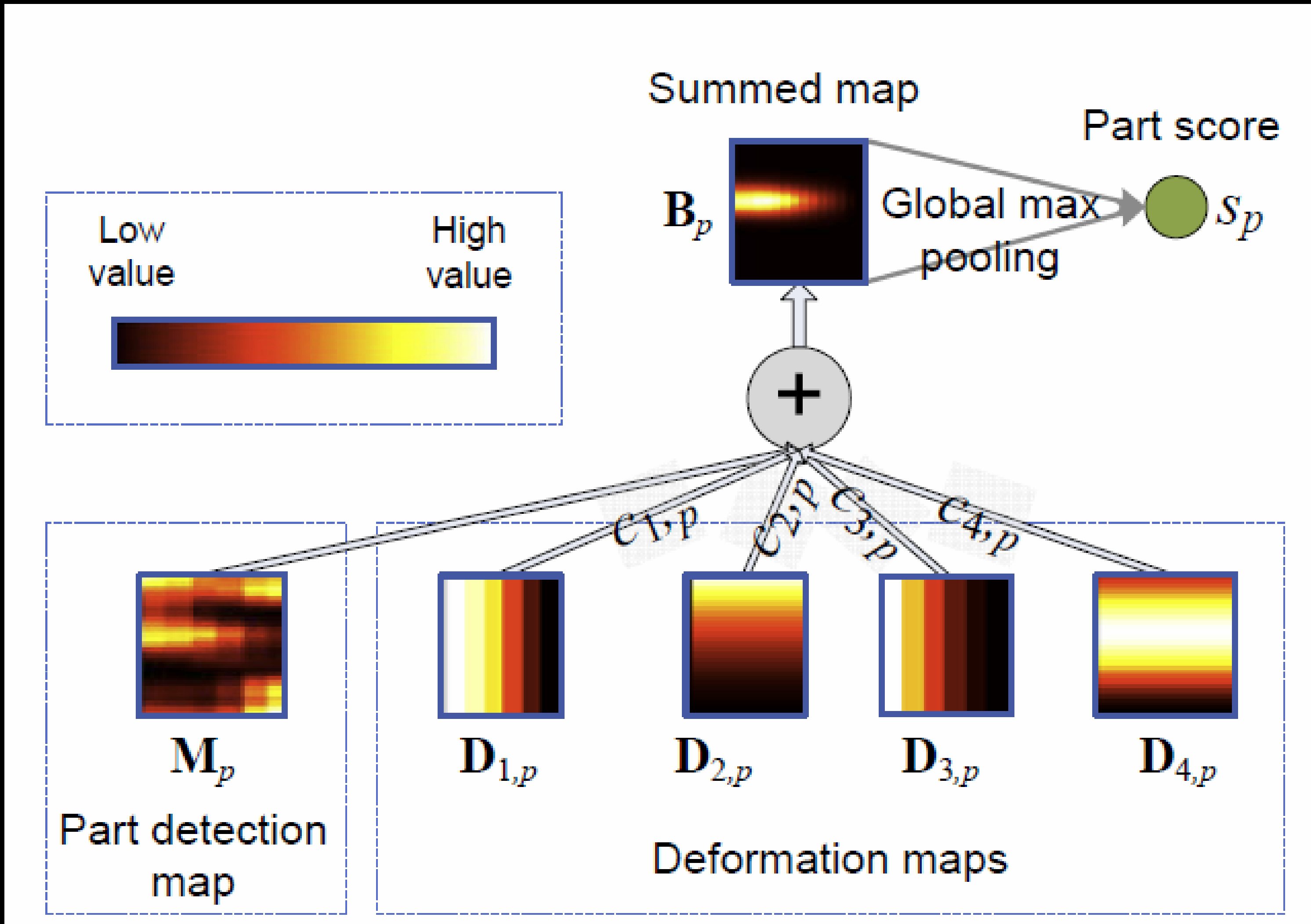


Part models

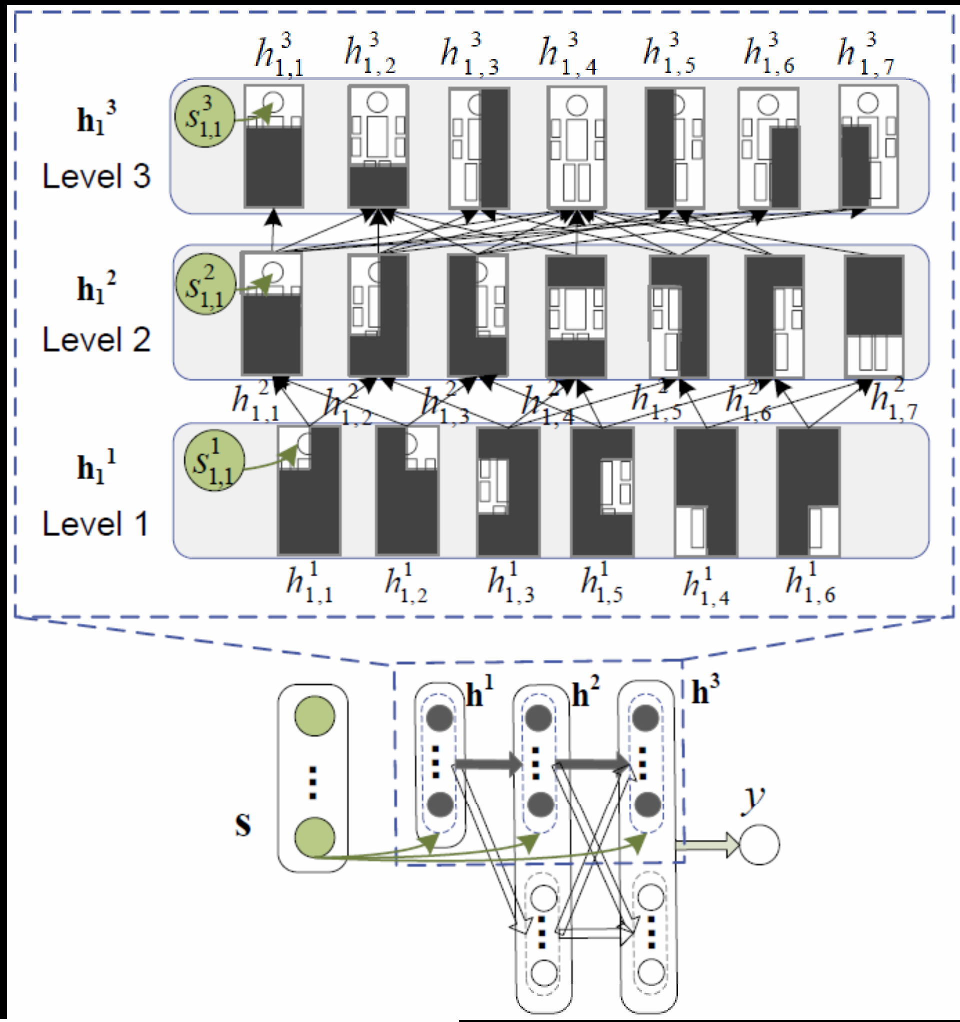


Learned filtered at the second convolutional layer

Deformation layer



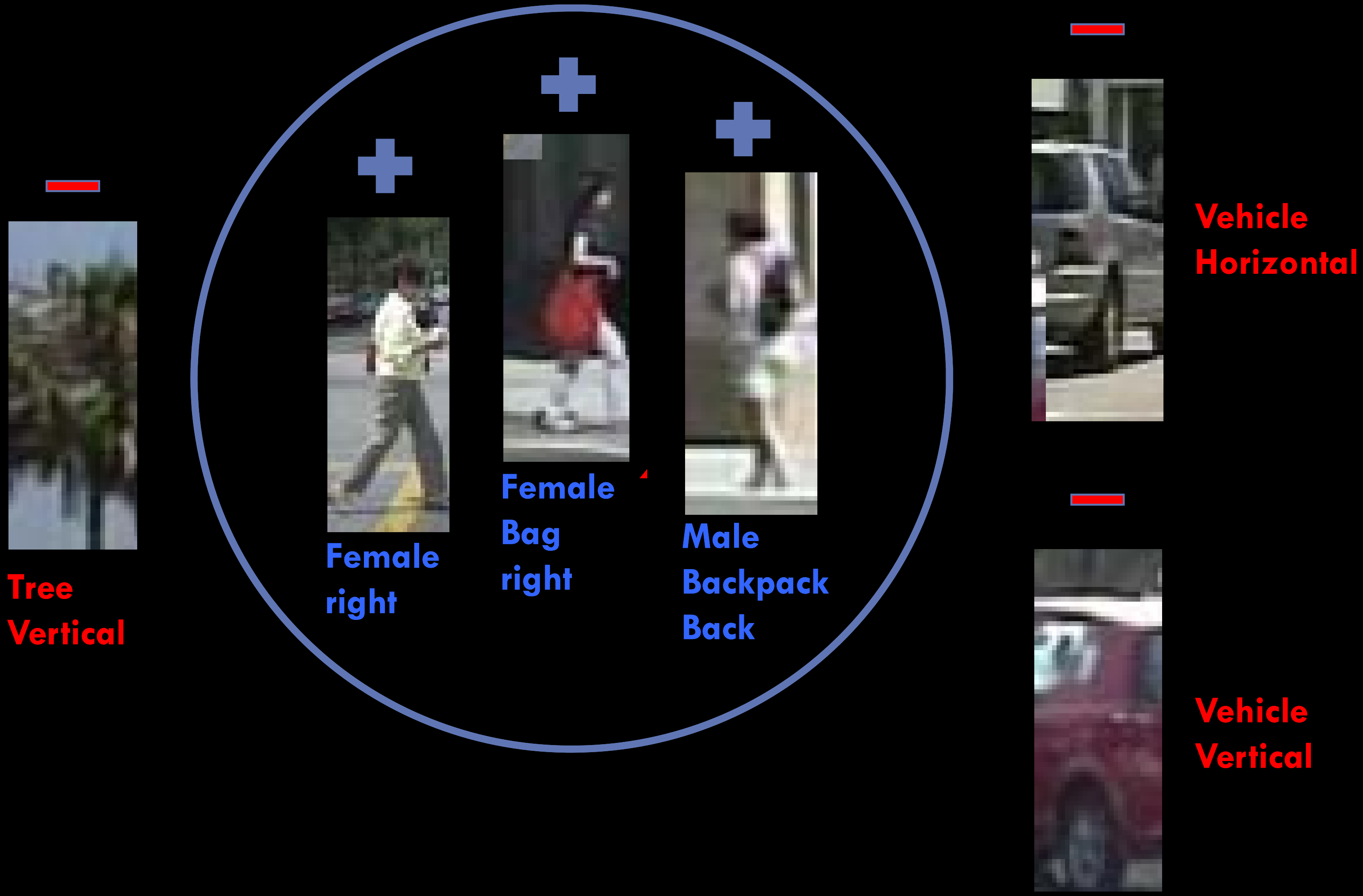
Visibility reasoning with deep belief net



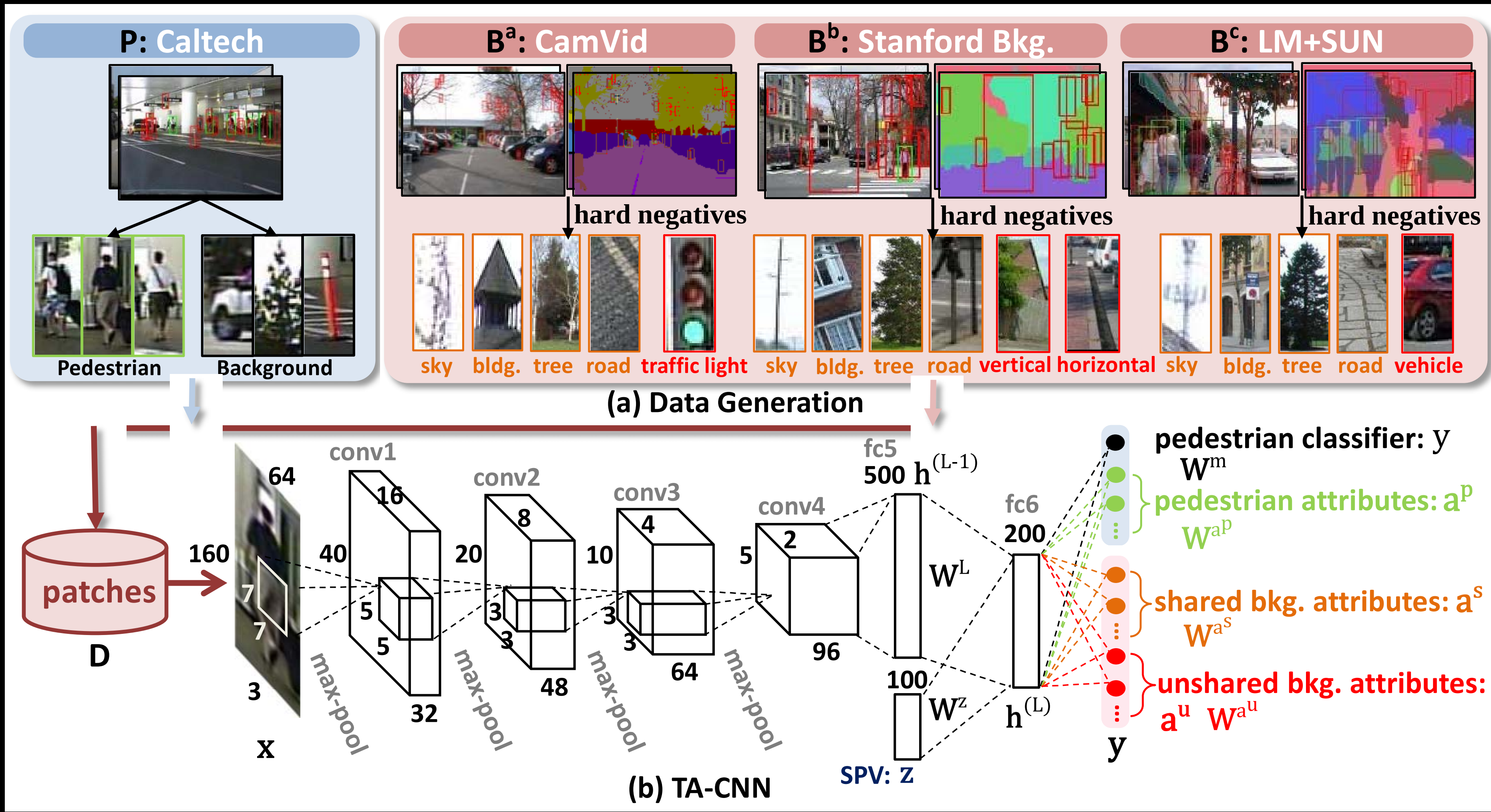
$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}_{*,j}^l + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

Pedestrian detection aided by deep learning semantic tasks

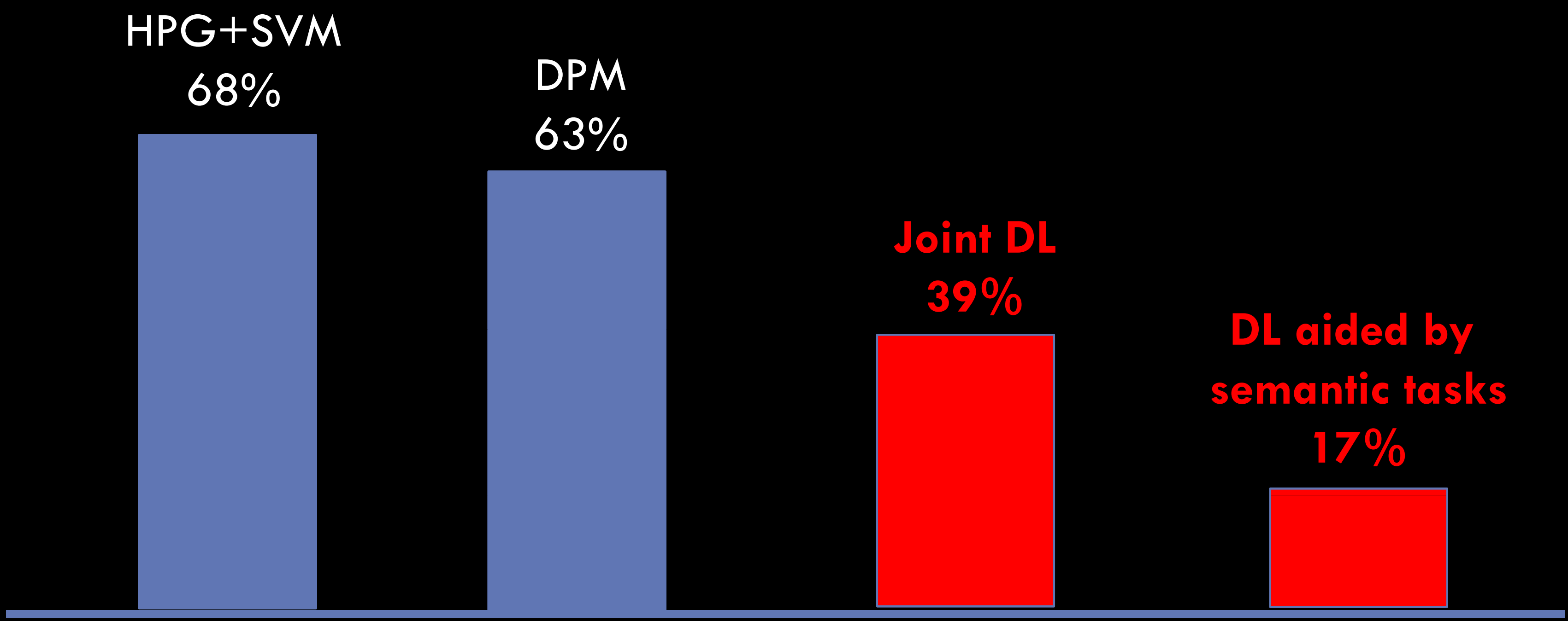


Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," CVPR 2015



P: Caltech
Pedestrian
Background
B^a: CamVid
B^b: Stanford Bkg.
B^c: LM+SUN
hard negatives
sky
bldg.
tree
road
traffic light
vertical
horizontal
vehicle
(a) Data Generation
patches
D
x
conv1
conv2
conv3
conv4
fc5
fc6
max-pool
500 h^(L-1)
W^L
W^z
100
SPV: z
200
h^(L)
pedestrian classifier: y
W^m
pedestrian attributes: a^p
W^ap
shared bkg. attributes: a^s
W^as
unshared bkg. attributes:
a^u W^au
y
(b) TA-CNN

Pedestrian Detection on Caltech (average miss detection rates)



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

Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," CVPR 2015.

Outline

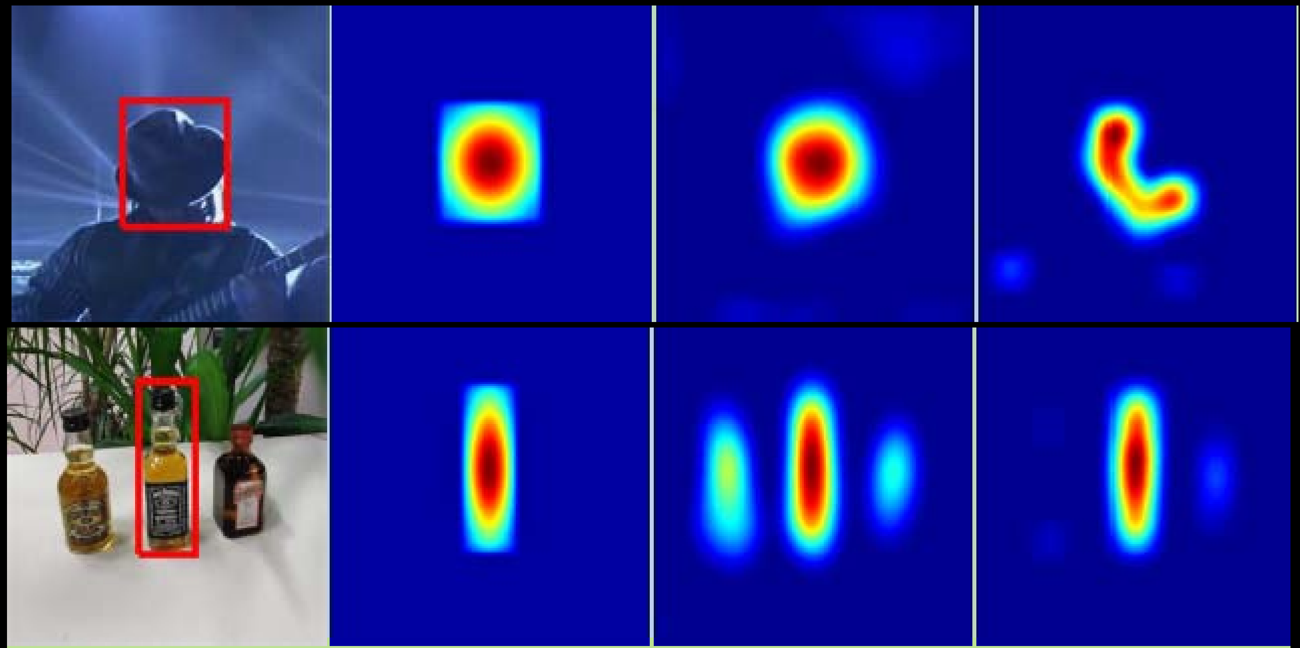
- Pedestrian detection
- **Object tracking**
- Crowd understanding

Motivations

- Explore the features pre-trained on massive data and classification task on ImageNet
- A top convolution layer encodes more semantic features and serves as a category detector
- A lower convolution layer carries more discriminative information and can better separate the target from distractors with similar appearance
- Both layers are jointly used with a switch mechanism during tracking
- A tracking target, only a subset of neurons are relevant

L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual Tracking with Fully Convolutional Networks," ICCV 2015.

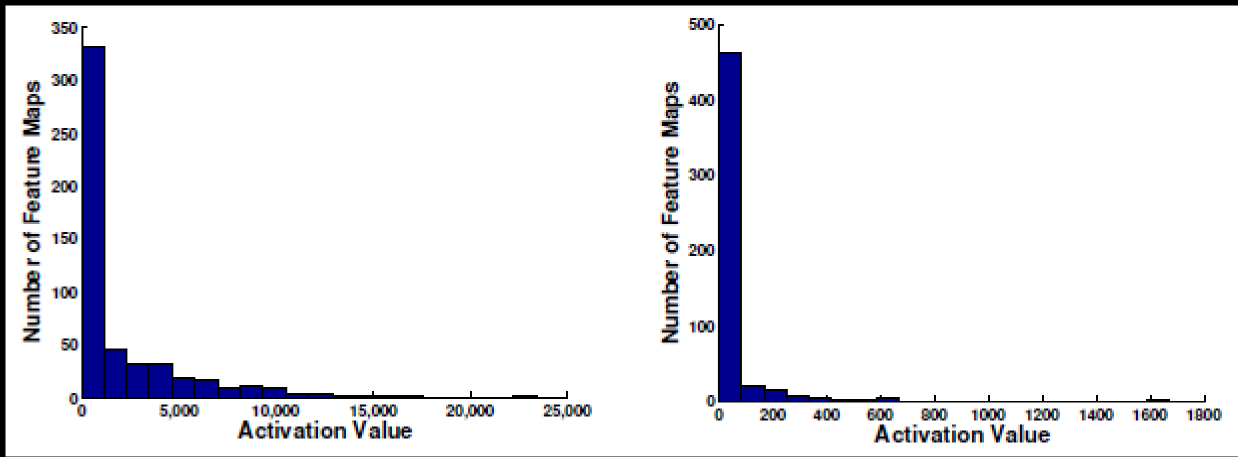
Observation 1: Different layers encode different types of features. Higher layers capture semantic concepts on object categories, whereas lower layers encode more discriminative features to capture intra class variations



(a) (b) (c)

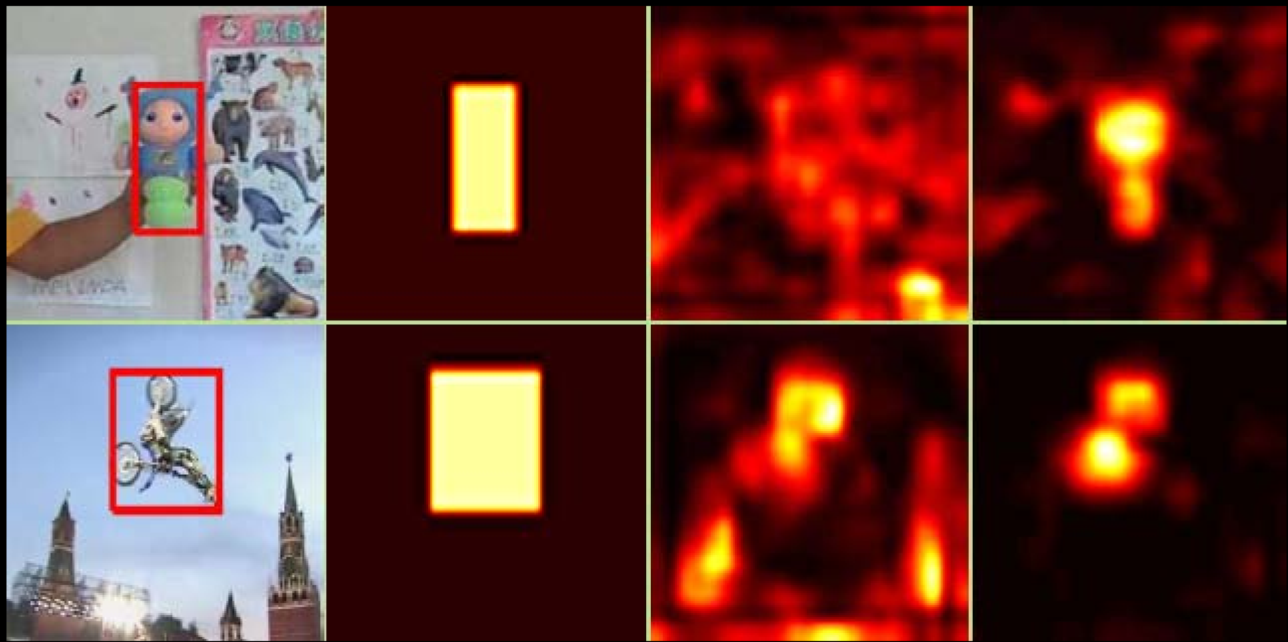
(a) Ground truth target heat map; (b) Predicted heat maps using feature maps of top convolution layers of VGG; (c) Predicted heat maps using feature maps of lower convolution layers of VGG

Observation 2: Although the receptive field of CNN feature maps is large, activated feature maps are sparse and localized. Activated regions are highly correlated to the regions of semantic objects



Activation value histograms of feature maps in top (left) and lower (right) layers

Observation 3: Many CNN feature maps are noisy or unrelated for the task of discriminating a particular target from its background

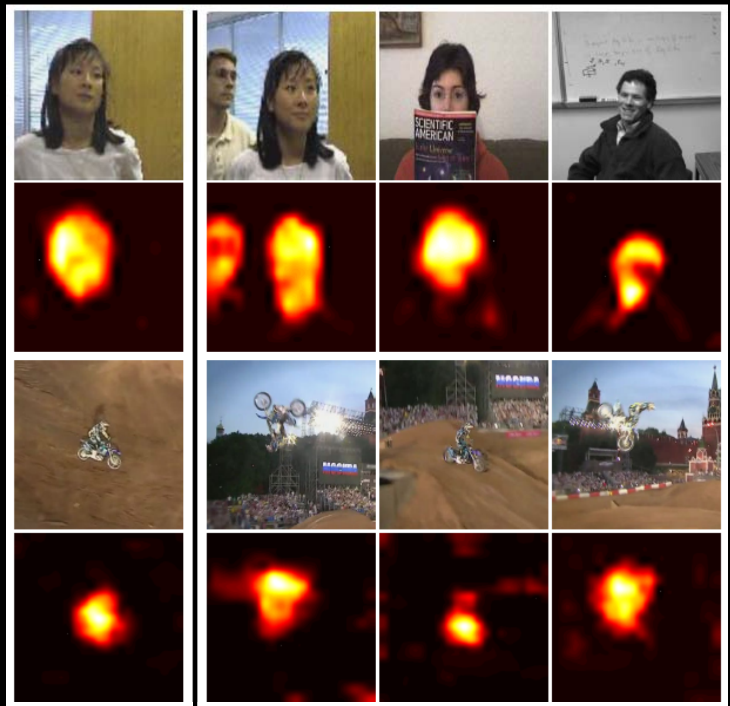


(a) (b) (c)

(a) Ground truth foreground mask, average feature maps of convolution layers; average selected feature maps of convolution layers

Selection of feature maps

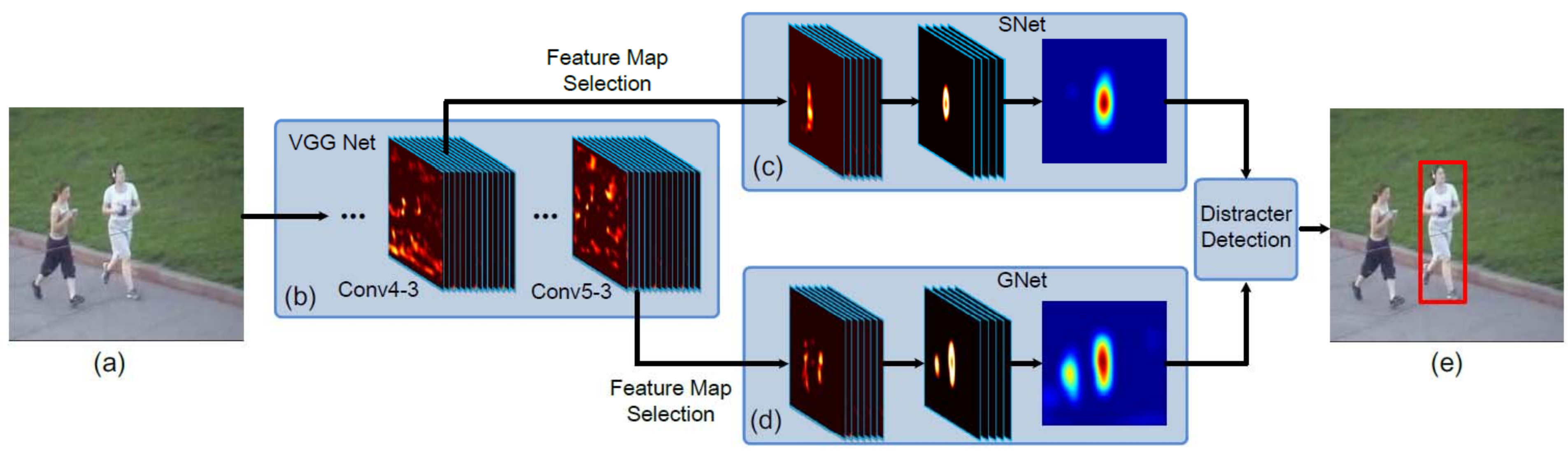
- Select feature maps by reconstructing foreground masks and their significance calculated with BP



The sparse coefficients are computed using the images in the first column and directly applied to the other columns without change

Fully convolutional network based tracker (FCN)

- GNet: capture the category information of the target and is built on the top layers of VGG
- SNet: discriminative the target from background with similar appearance and is built on the lower layers of VGG

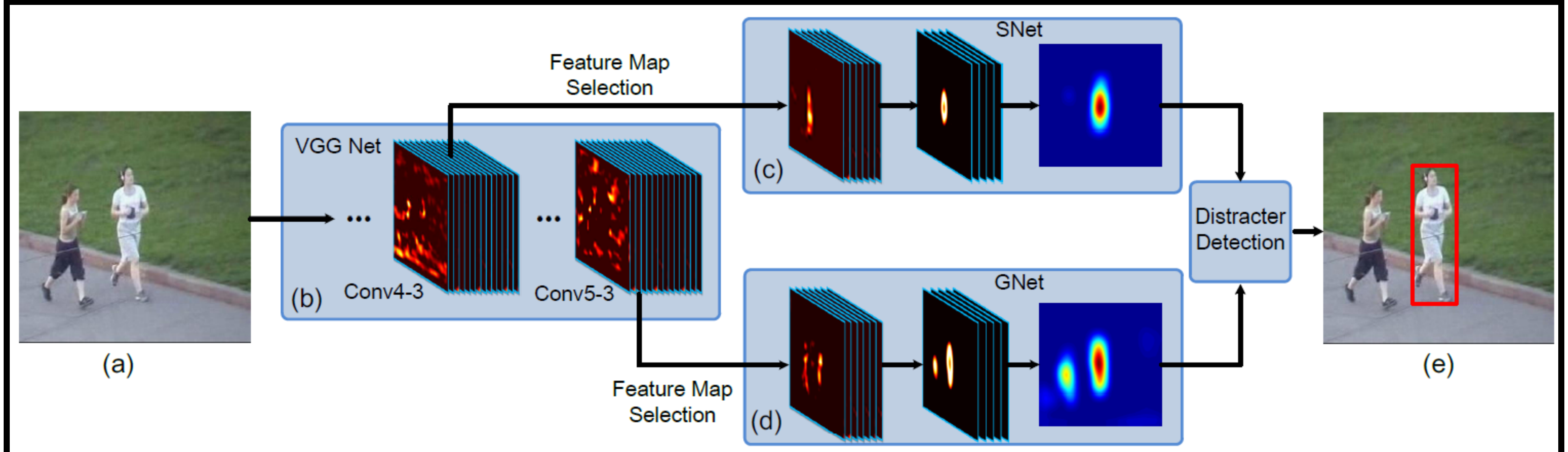


(b) VGG network; (c) SNet; (d) Gnet; (e) Tracking results

Both GNet and SNet are initialized in the first frame to perform foreground heat map regression for the target: GNet is fixed and SNet is updated every 200 frames

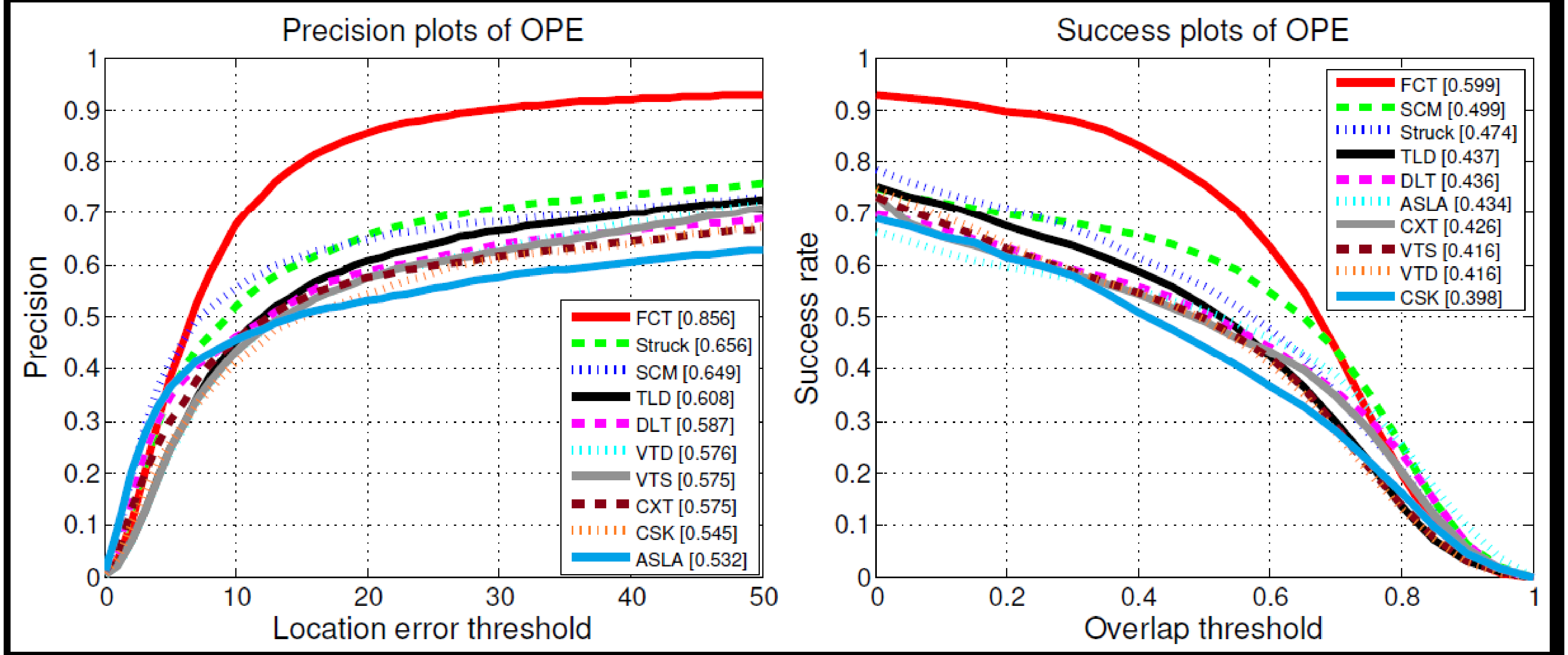
SNet is used if the background distractor is larger than a threshold; otherwise GNet is used

For a new frame, a region of interest (ROI) centered at the last target location containing both target and background context is cropped and propagated through the fully convolutional network



(b) VGG network; (c) SNet; (d) Gnet; (e) Tracking results

Precision plots and success plots of OPE for the top 10 trackers



[redacted]

Outline

- Pedestrian detection
- Object tracking
- **Crowd understanding**

Surveillance

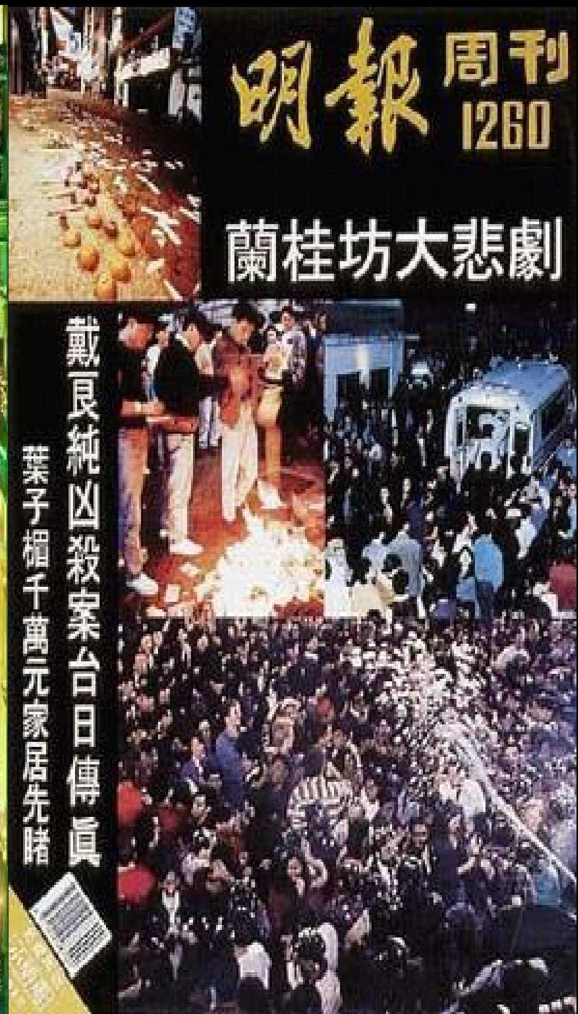
- **Crowd behavior analysis**
  - T. Hospedales, et al., CVPR’09
  - R. Mehran, et al., CVPR’09
  - V. Mahadevan, et al., CVPR’10
  - B. Zhou, et al., TPAMI’14
  - S. Yi, et al., CVPR’14
- **Crowd tracking**
  - S. Ali, et al., ECCV’08
  - M. Rodriguez, et al., ICCV’11
  - F. Zhu, et al., ECCV’14
- **Crowd segmentation**
  - S. Ali, et al., CVPR’07
  - A. B. Chan, et al., TPAMI’08

Surveillance
General Crowd Understanding

Movie/TV shows

Individual collection

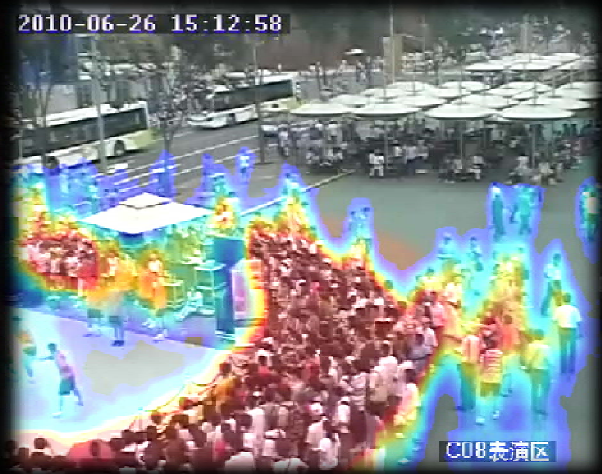
Crowd management to avoid disasters



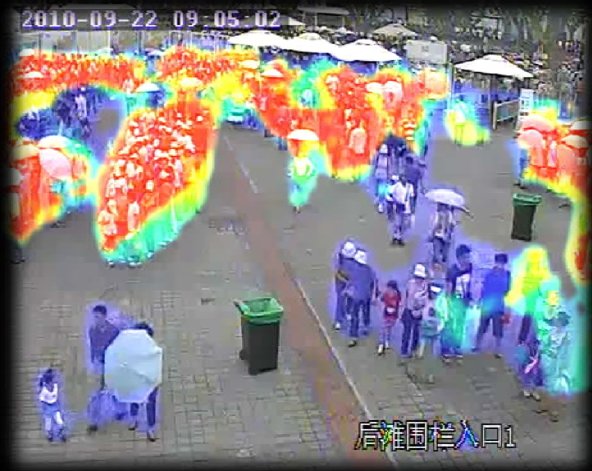
Shanghai **Hong Kong** **Mecca**

Crowd understanding

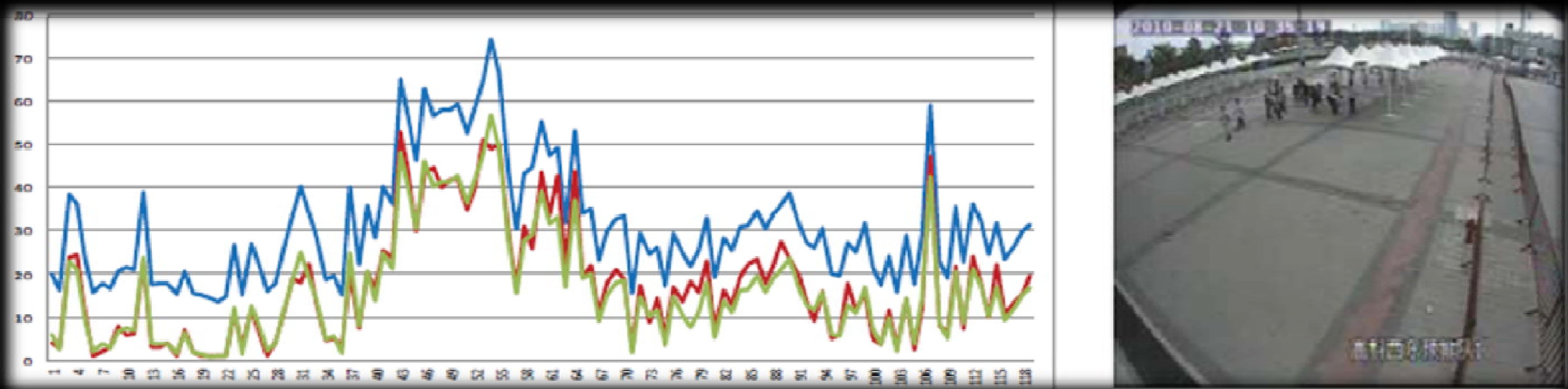
Crowd segmentation



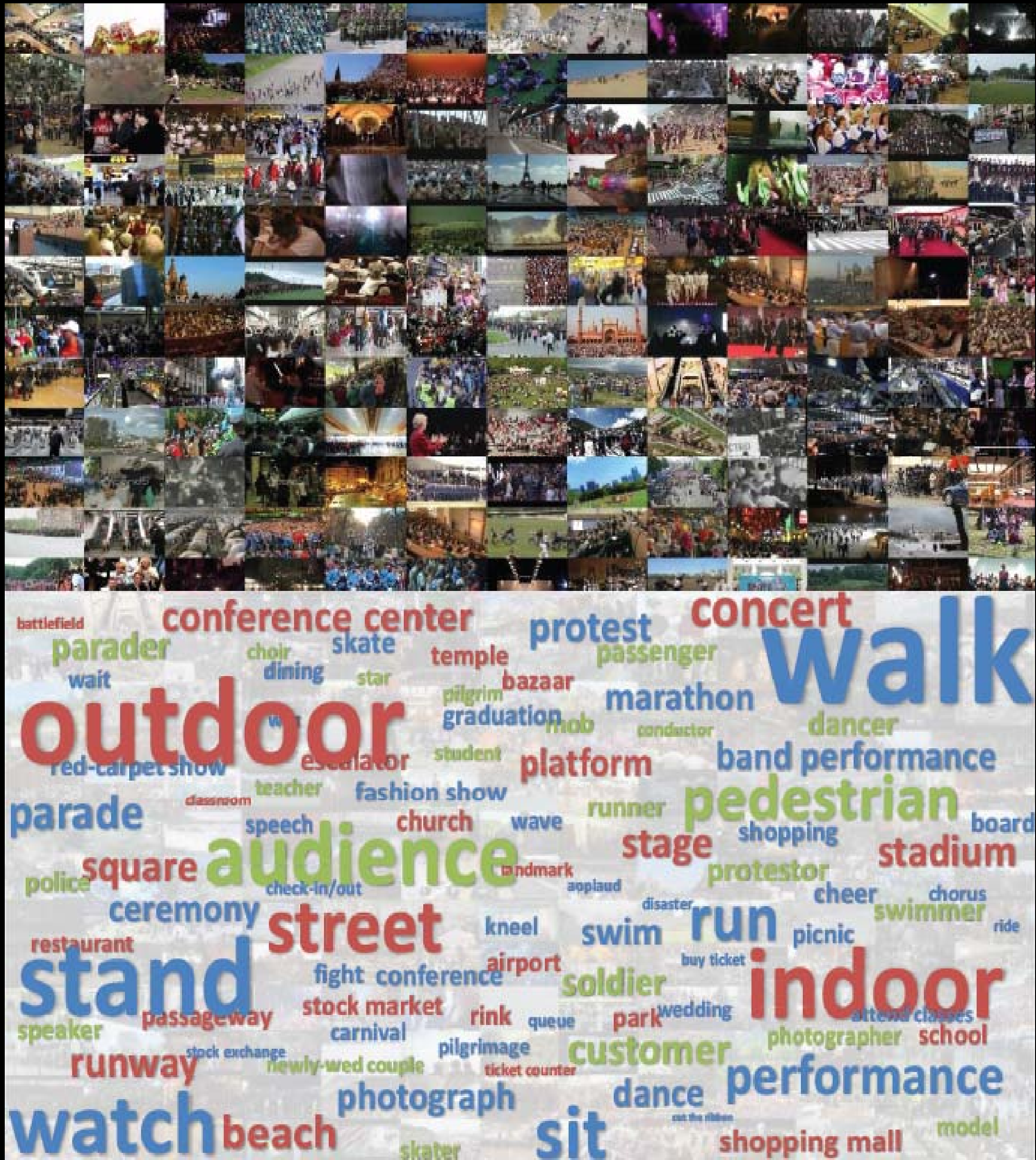
Density estimation



Stationary crowd detection



Crowd counting



outdoor | run
stand | marathon
runner | street

outdoor | rink
skater
skate

Crowd attribute recognition

Benchmark for cross-scene crowd understanding

WorldExpo'10 Crowd Dataset

1132 videos, from108 scenes

199932 annotated pedestrians

WWW Crowd Dataset

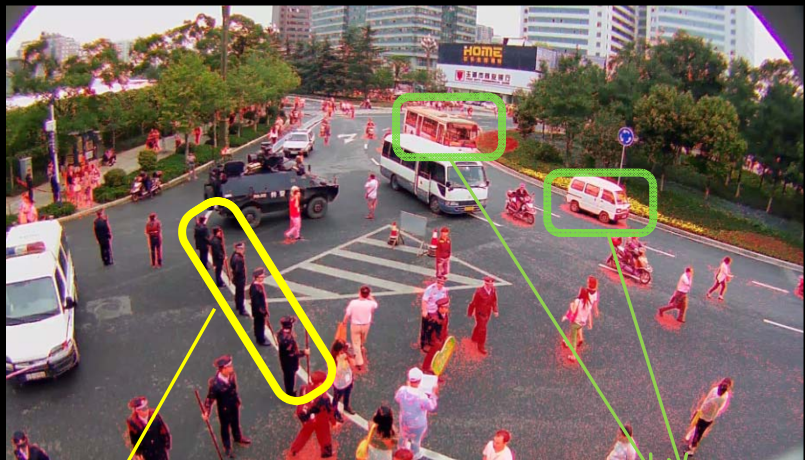
96 attributes

10,000 videos

8,257 crowded scenes

Crowd segmentation

- Traditional motion based approaches



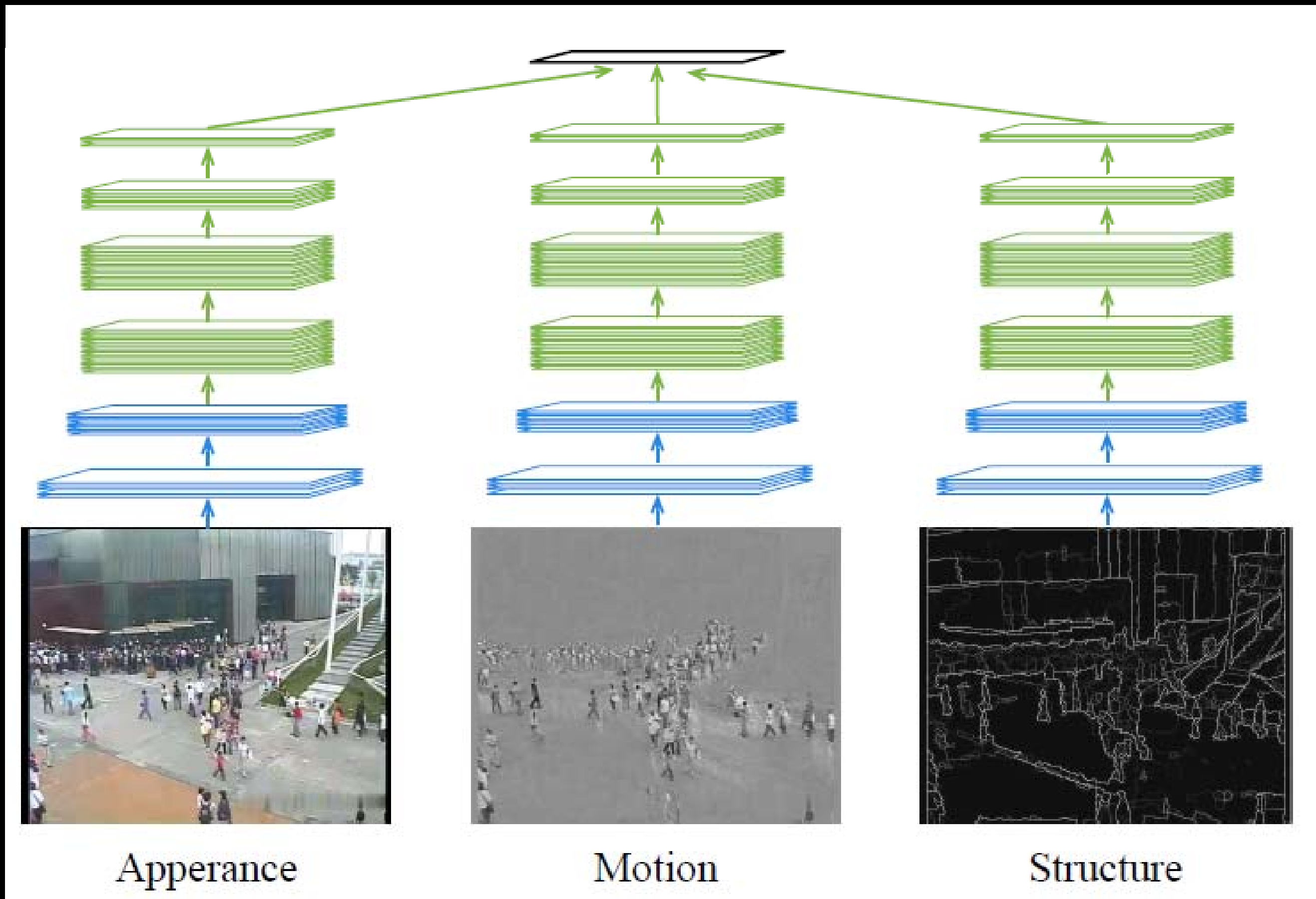
Still persons incomplete detection

Moving cars false detected as foreground

Deep learning

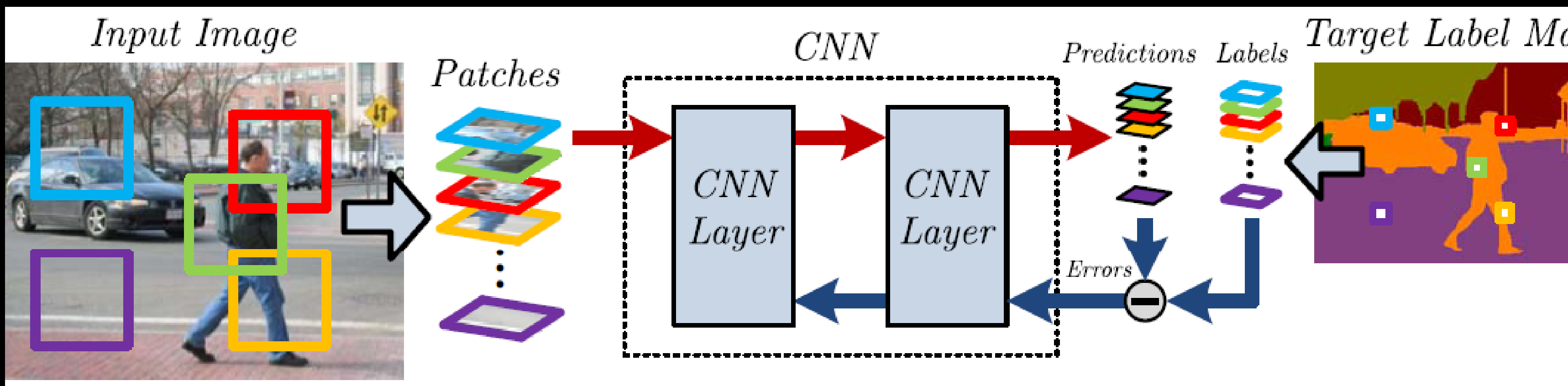
CNN based crowd segmentation

- Multi-stage fusion



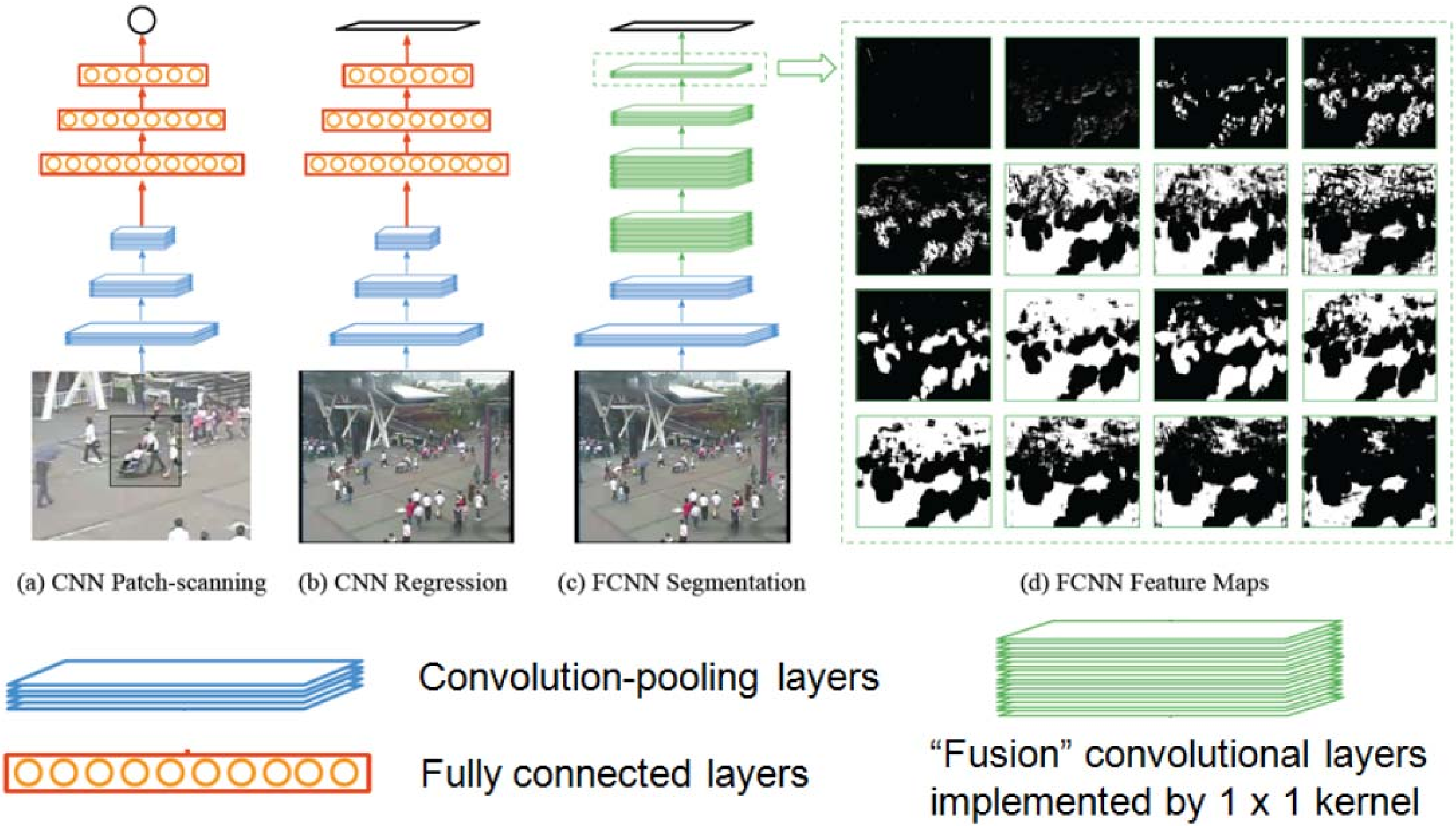
CNN for pixelwise classification

- CNN was proposed for whole image classification
- Pixelwise classificaion: predicting a label at very pxiel (e.g. segmentation, detection, and tracking)
- It is generally trained and tested in a patch-by-patch scanning manner, but involves much redundant computation



Fully convolutional neural network

- K. Kang and X. Wang, "Fully Convolutional Neural Networks for Crowd Segmentation," arXiv: 1411.4464, 2004.
- 2400 times speed up and take images of any size as input
- Replace the fully connected layers with 1 x 1 convolutional kernels


(a) CNN Patch-scanning
(b) CNN Regression
(c) FCNN Segmentation
(d) FCNN Feature Maps
Convolution-pooling layers
Fully connected layers
“Fusion” convolutional layers implemented by 1 x 1 kernel

Crowd segmentation

Crowd segmentation

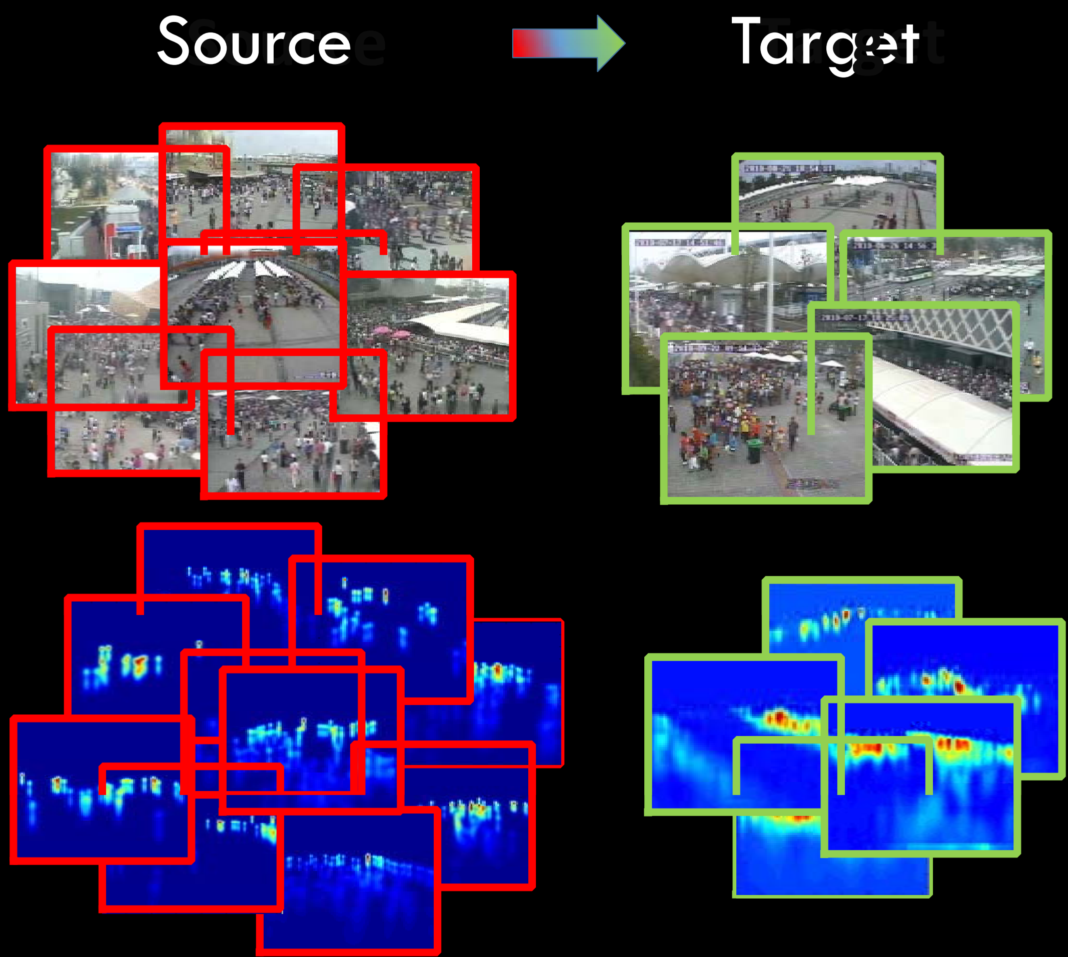
Stationary crowd detection

Crowd counting and density estimation

- Existing approaches are scene-dependent, i.e. requring training samples from the target scene
- Rely on motion-based crowd segmentation and use handcrafted features: LBP, HOG, area, perimeter

Cross-scene crowd counting via deep convolutional neural network

- C. Zhang, X. Wang, H. Li, and K. Yang, CVPR 15

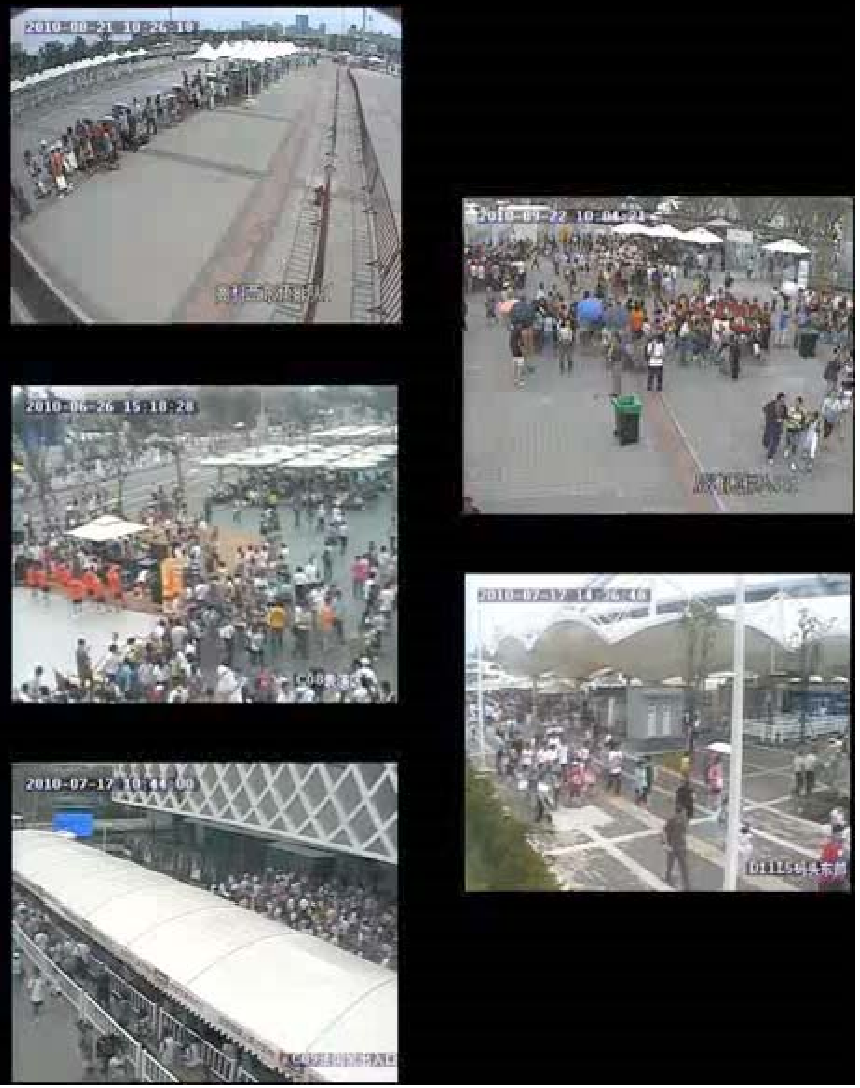


Source scenes

106 crowd scenes for training
1180 one-minute videoclips labeled

Target scenes

5 target scenes for testing
5 one-hour video clips labeled



A much larger dataset than before

Table 1. Statistics of three datasets: N_f is numbers of frames; N_c is numbers of scenes; R is the Resolution; FPS is frame per second; D is Density contained that minimum and maximum in the ROI; and T_p is total number of labeled pedestrian instances

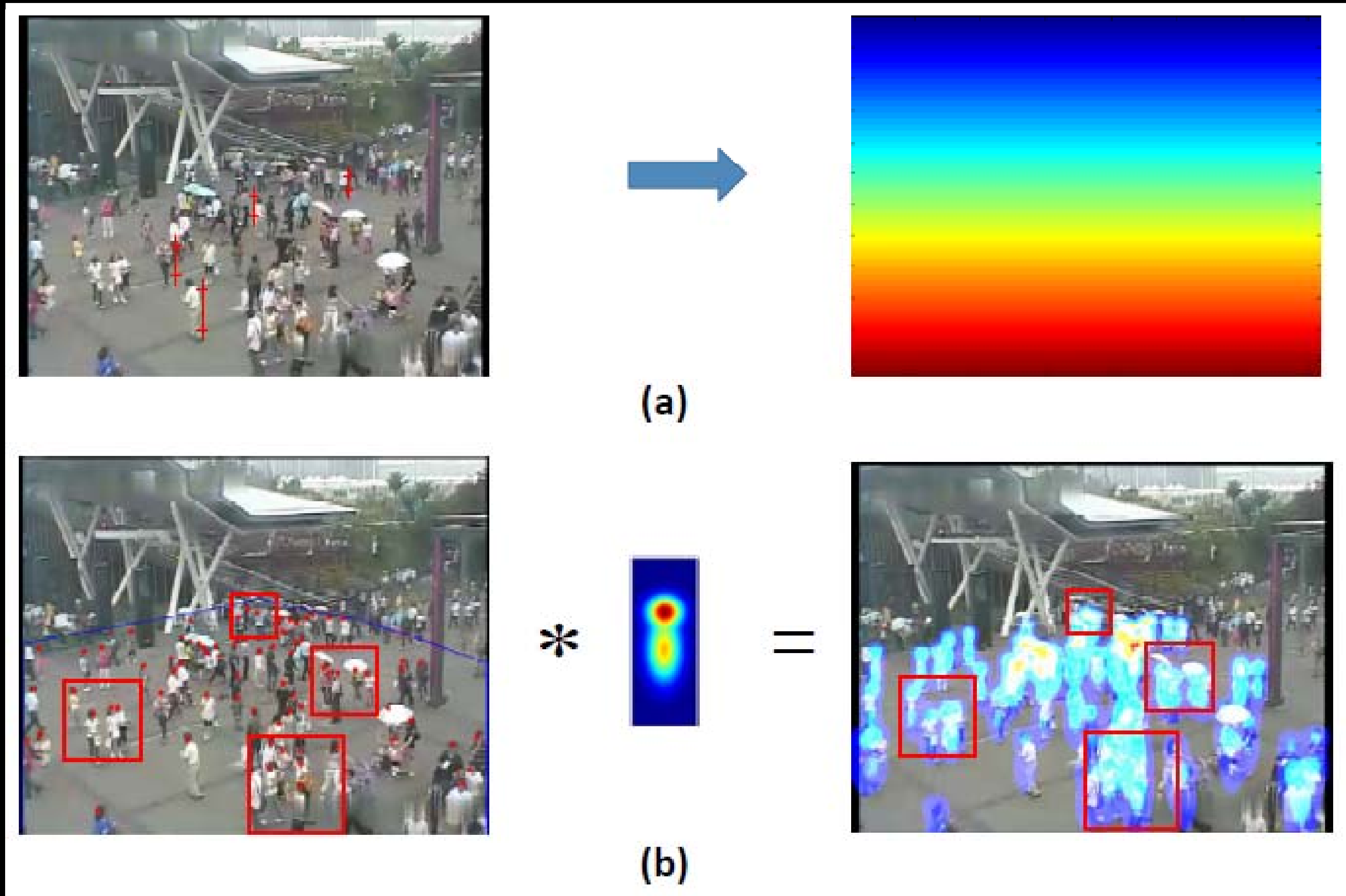
Dataset	N_f	N_c	R	FPS	D	T_p
UCSD	2000	1	158*238	10	11-46	49885
UCF_CC_50	50	50	–	image	94-4543	63974
WorldExpo	4.44 million	110	576*720	50	1-253	199923

(a)

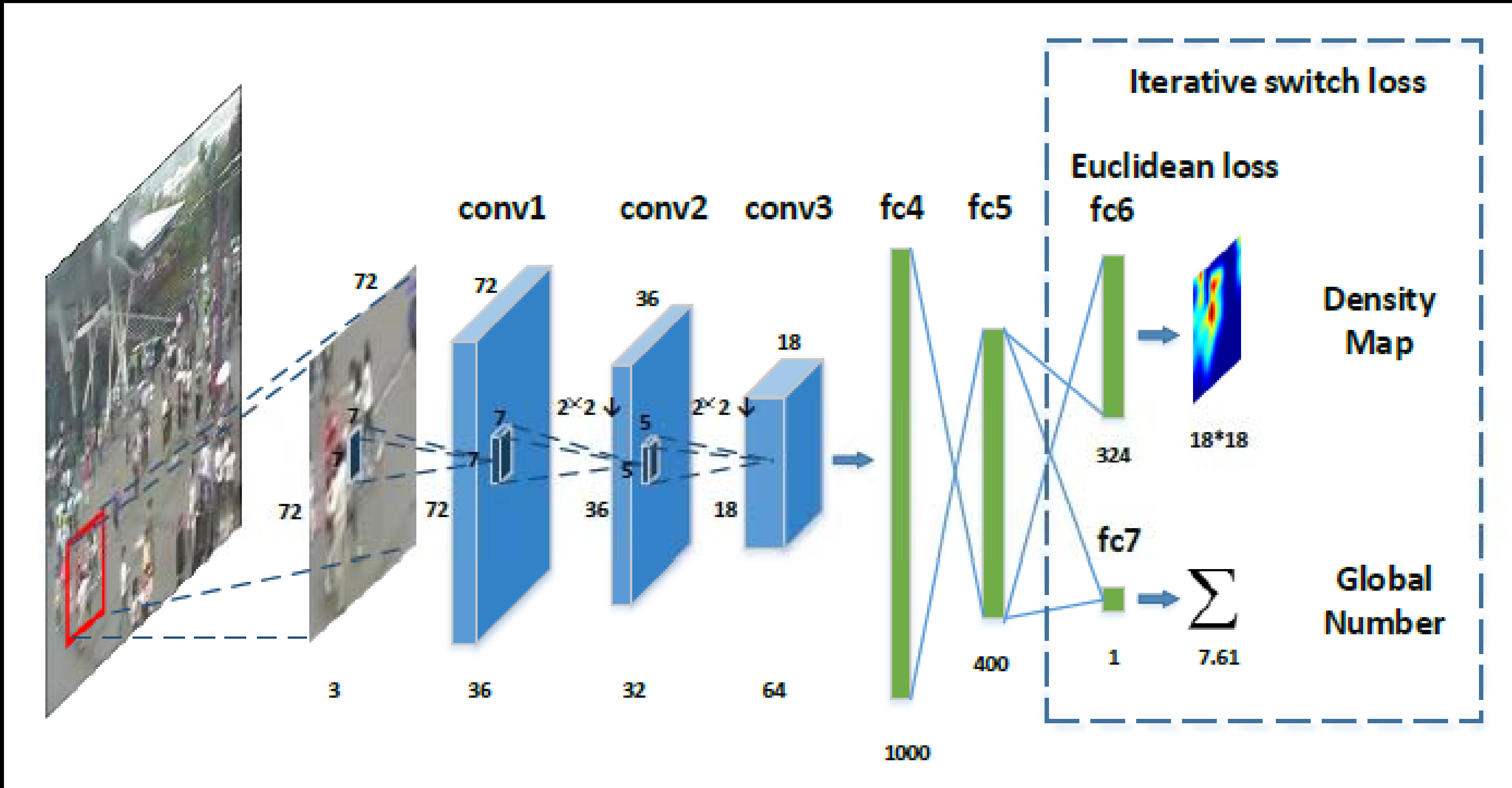
(b)

(c)

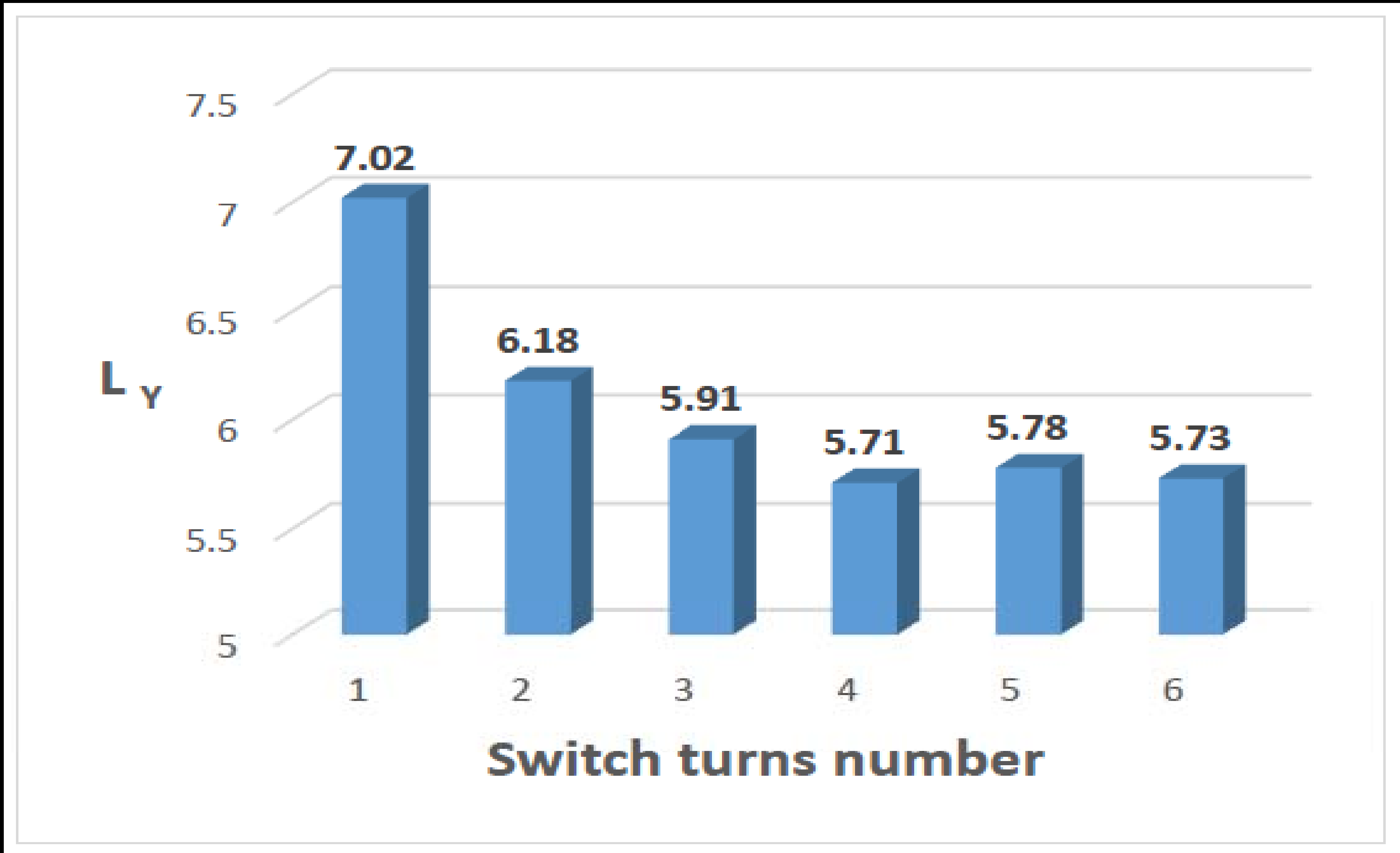
Joint crowd counting-density estimation



Deep convolutional neural network solution



Siwtching joint optimization helps to jump out from local minima



Crowd density estimation

Crowd counting

Crowd attribute recognition

How to categorize it ?

How to describe it ?

Or to categorize ?

One class label ?

orchestra performance ?
watch performance ?

orchestra performance ?
choir performance ?

orchestra performance ?
military marching ?

military marching?
watch performance ?

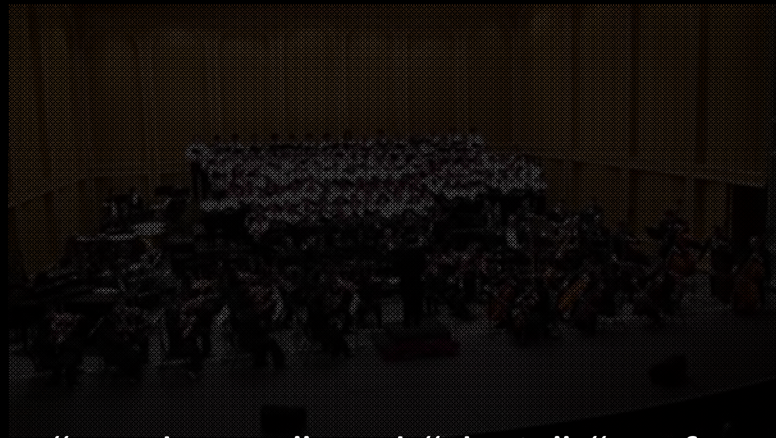
Attribute-based representation!

Or class label?

orchestra performance ?

watch performance ?

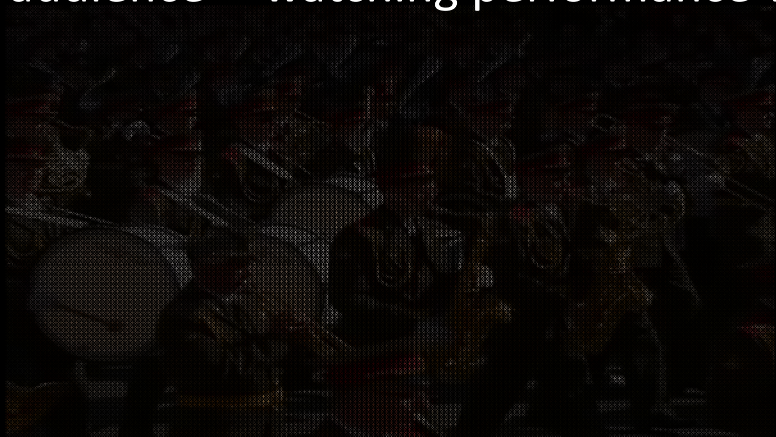
The "conductor" "conduct" "orchestra performance" in a "concert" with "audience" "watching performance".



orchestra performance ?

choir performance ?

The "conductor" and "choir" "perform/ chorus" on the "stage" with "orchestra performance" in an "indoor" "concert".



orchestra performance ?

military marching ?

The "military" "perform" "orchestra marching" on the "street".

military marching ?

watch performance ?

The "military" "march" on the "street" with "audience" "watching performance".

Attribute-based representation!

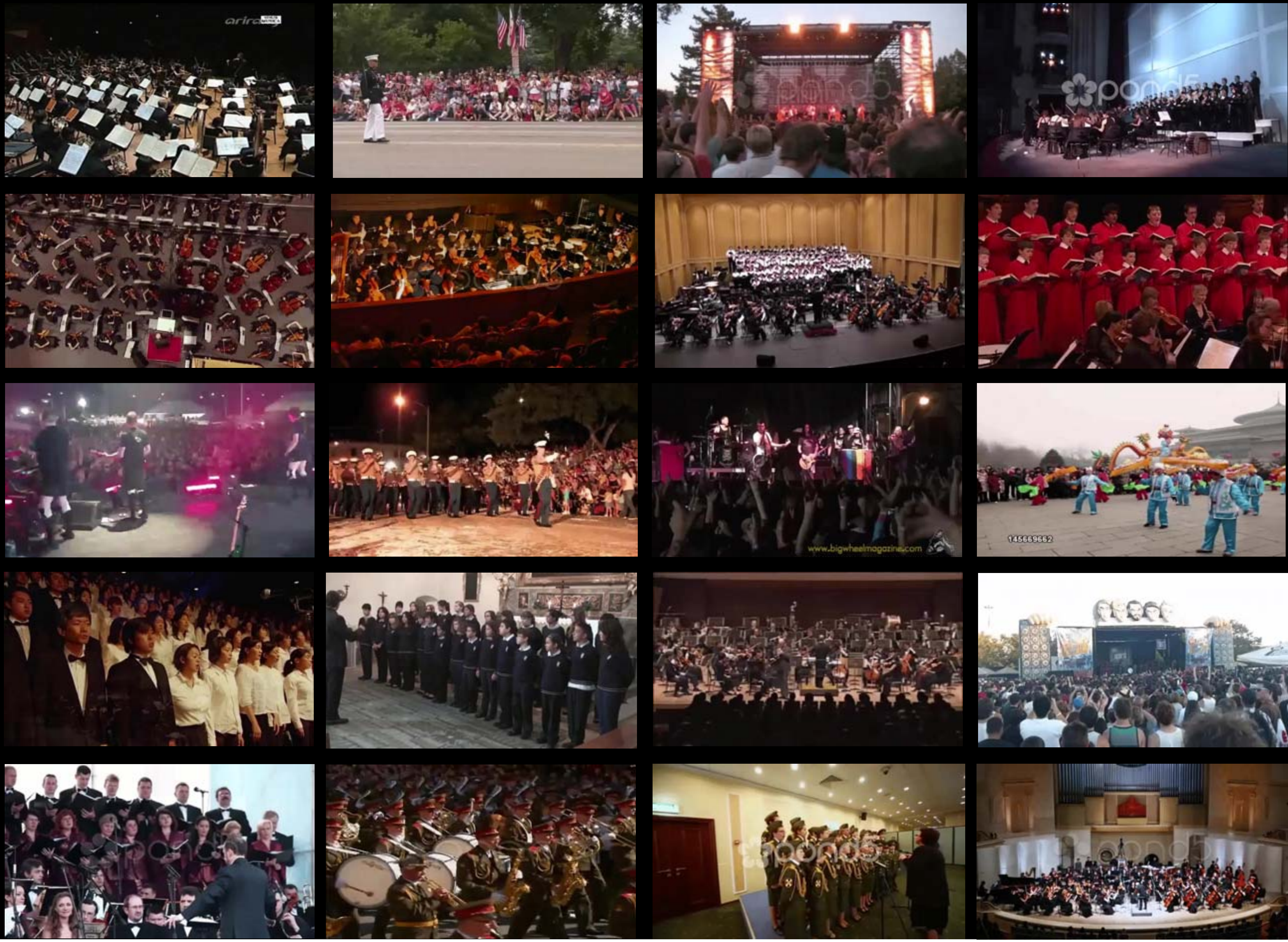
Where Who Why

The "military" "perform" "orchestra marching" on the "street".

The "military" "march" on the "street" with "audience" "watching performance".

Attribute-based representation!

performance stage conductor orchestra audience



Attribute-based representation!

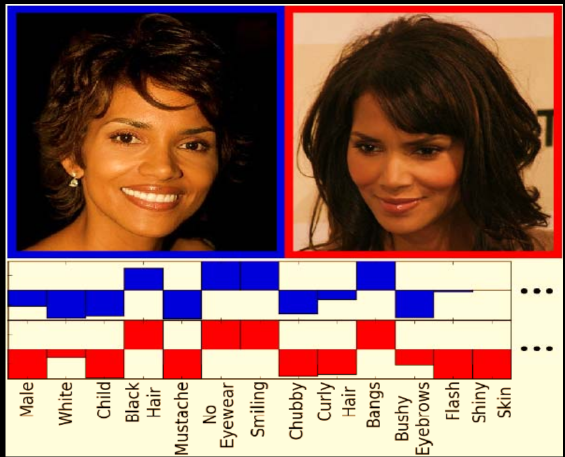
- Scene-independent
- More informative
- Natural for humans (i.e. *Who do What at someWhere*)

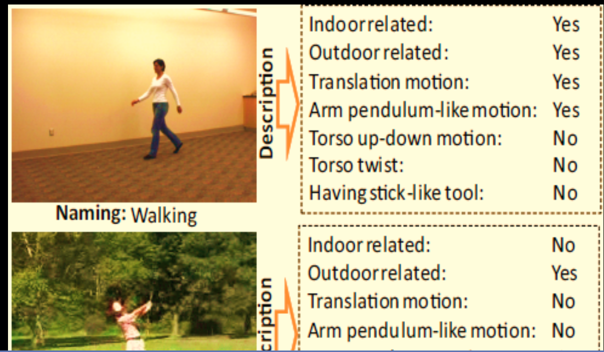
Attribute-based representation!

Face attribute

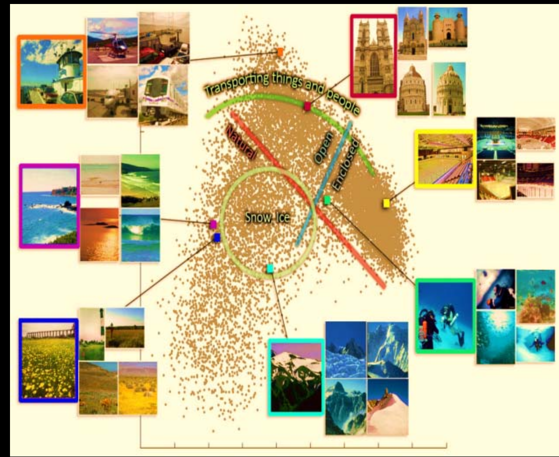
Action attribute

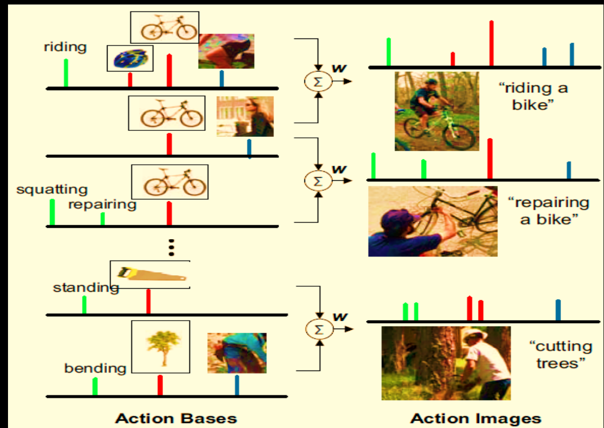
Scene attribute





Crowd attribute



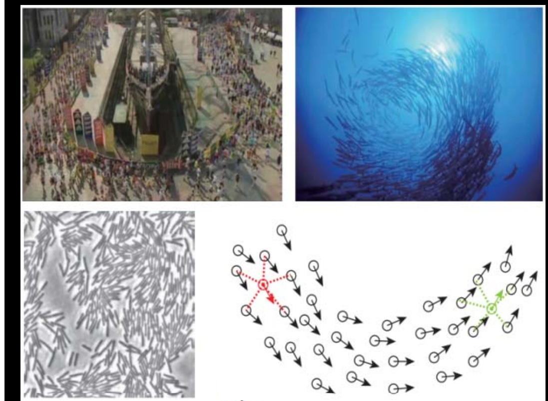


N. Kumar, A. C. Berg, et al., ICCV'09;
P. Luo, X. Wang, et al., ICCV'13.

J. Liu, B. Kuipers, et al., CVPR'11;
B. Yao, X. Jiang, et al., ICCV'11

G. Patterson and J. Hays, CVPR'12;
D. Parikh and K. Grauman, CVPR'11

The number of attributes is limited !



Crowd attribute

Collectiveness

Dataset: 413 videos, 62 scenes

[B. Zhou, X. Tang, et al.. TPAMI, 2014.]

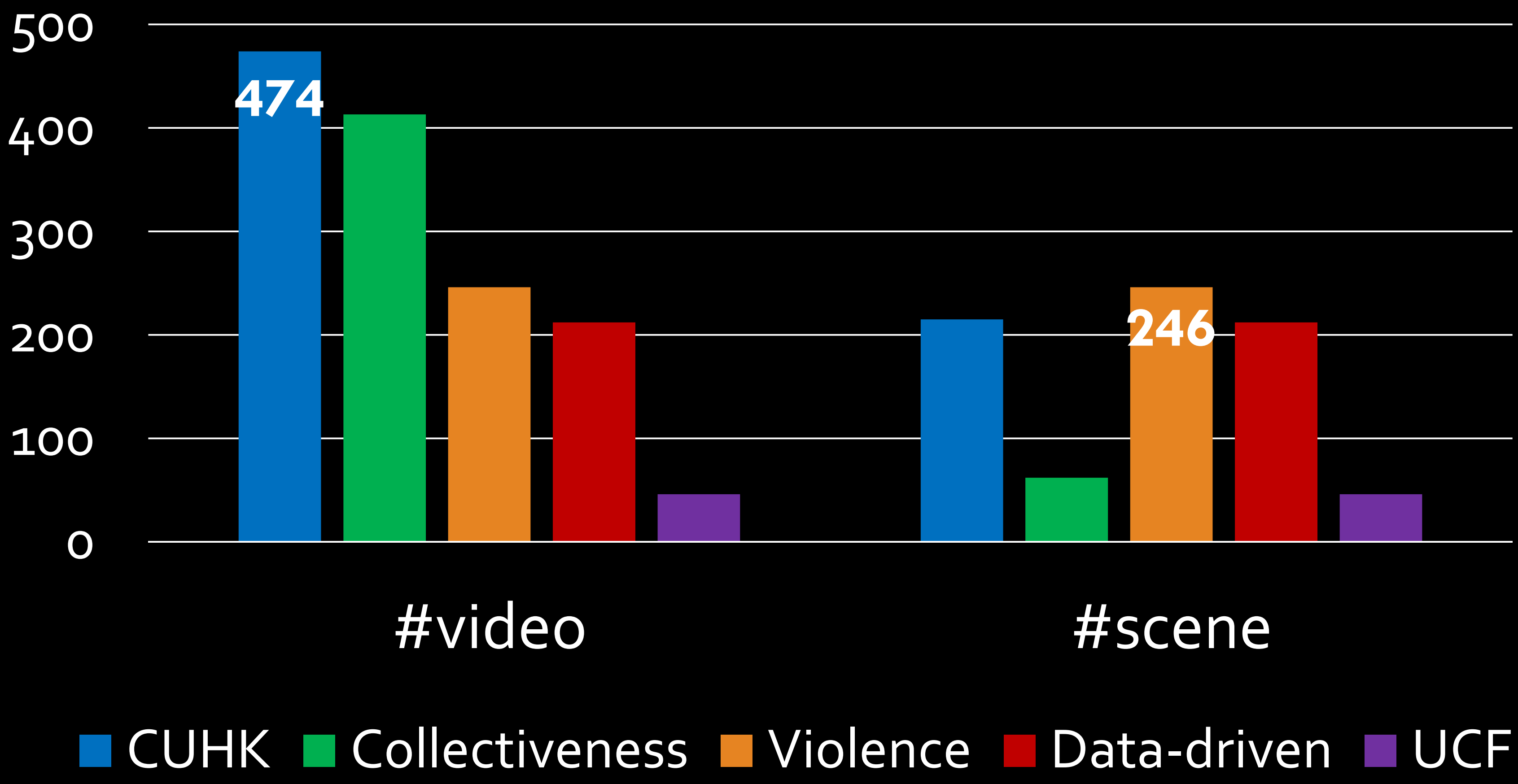
Collectiveness, Stability, Uniformity, and Conflict

Dataset: 474 videos, 215 scenes

[J. Shao, C. C. Loy, and X. Wang. CVPR, 2014.]

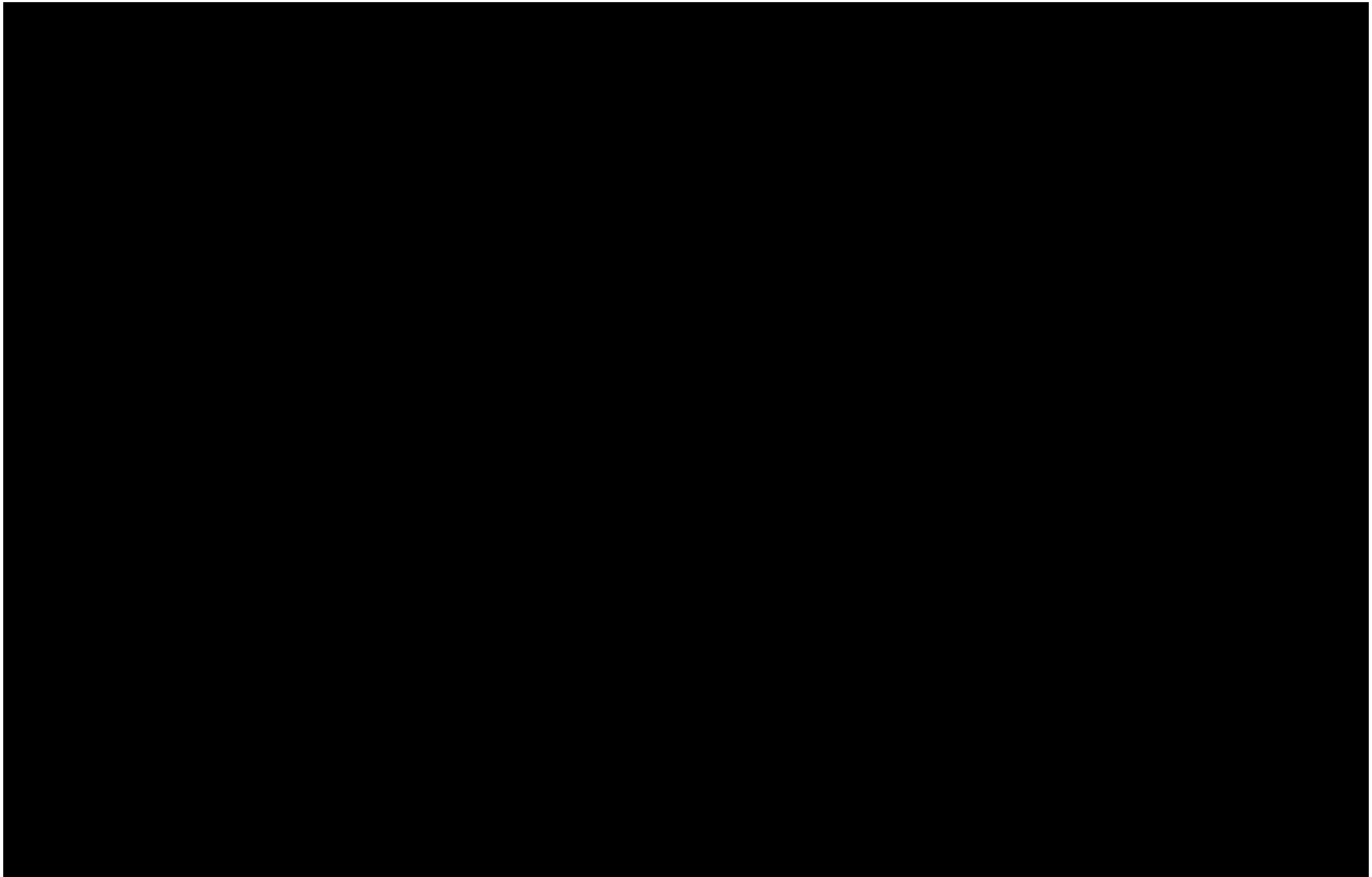
Existing Crowd Datasets

The datasets are small !



Our Goal

- ⌘ Construct a large-scale crowd video dataset
- ⌘ Study more crowd attributes



WWW Crowd Dataset

10000 videos, 8257 scenes, 8 million frames, 94 attributes

Crowd Attributes Collection

Partial raw tag wordle.

(The total number of retrieved tags is 7000+)

Crowd Attributes Collection

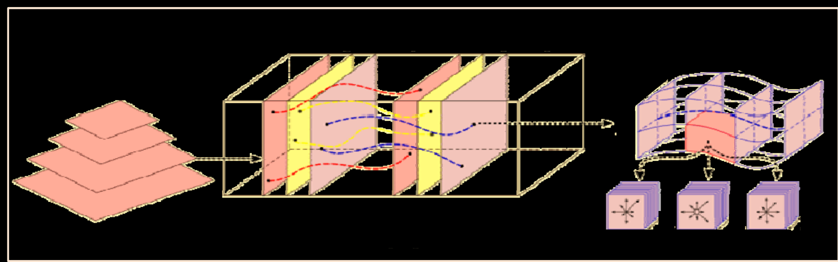
- We finally constructed an attribute set with 94 crowd-related attributes. It includes 3 types of attributes:
  - **Where** (e.g. street, temple, and classroom)
  - **Who** (e.g. star, protester, and skater)
  - **Why** (e.g. walk, board, and ceremony)

How to learn attributes from crowd videos?

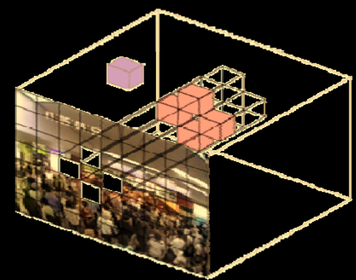
Hand-crafted features

- SIFT, HOG, GIST, SSIM, LBP, …
  - image classification and object detection

- Dense trajectory *[H. Wang et al. CVPR'11]*
  - action recognition



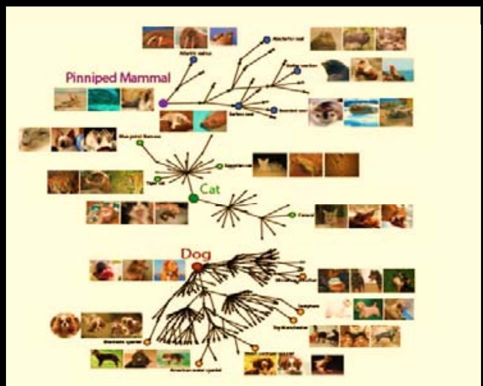
- Spatio-temporal motion patterns *[L. Kratz and K. Nishino, CVPR'09]*
  - anomaly detection

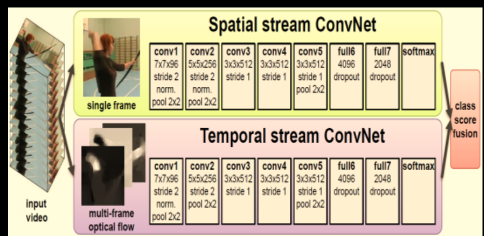


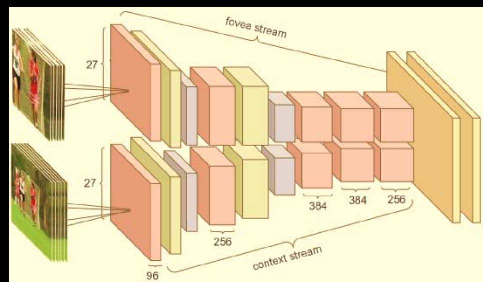
How to learn attributes from crowd videos?

Deeply learned features

- Convolutional neural networks (CNNs)
  - image classification
  - action recognition [*K. Simonyan, et al. CVPR'14*] and video classification [*A. Karpathy, et al. CVPR'14*]



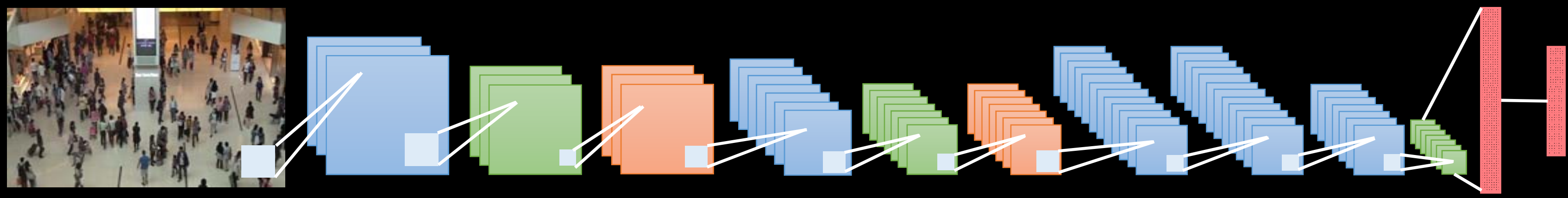




How to learn attributes from crowd videos?

A two-branch CNN model

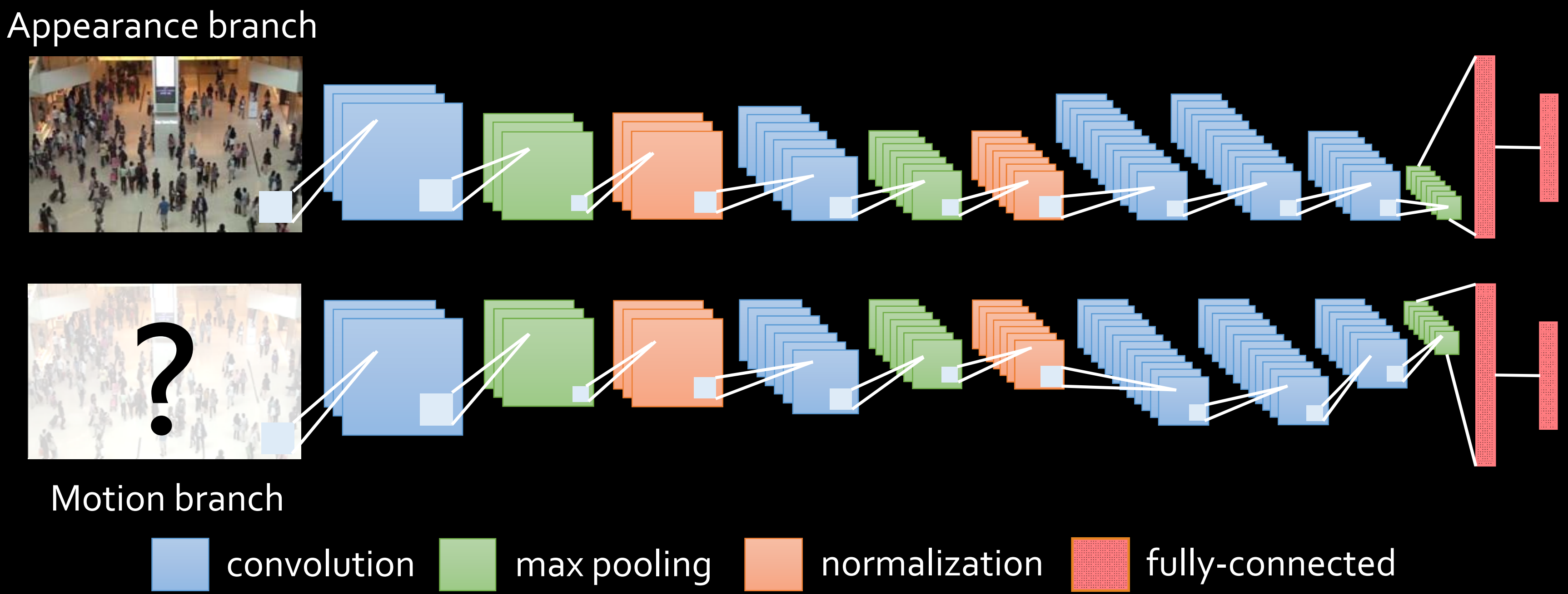
Appearance branch



convolution max pooling normalization fully-connected

How to learn attributes from crowd videos?

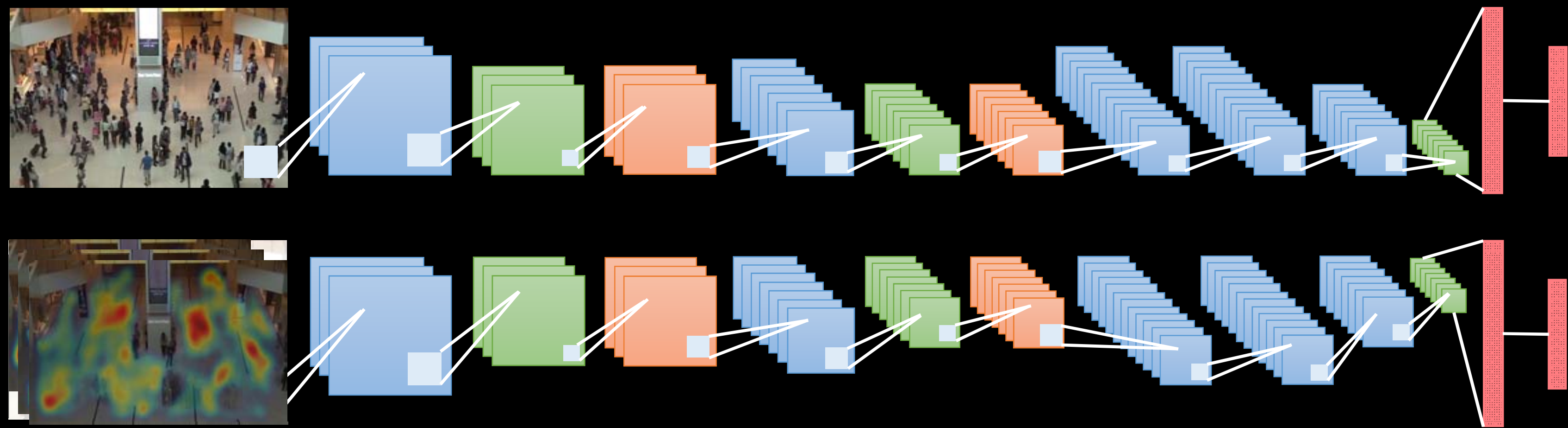
A two-branch CNN model



How to learn attributes from crowd videos?

A two-branch CNN model

Appearance branch

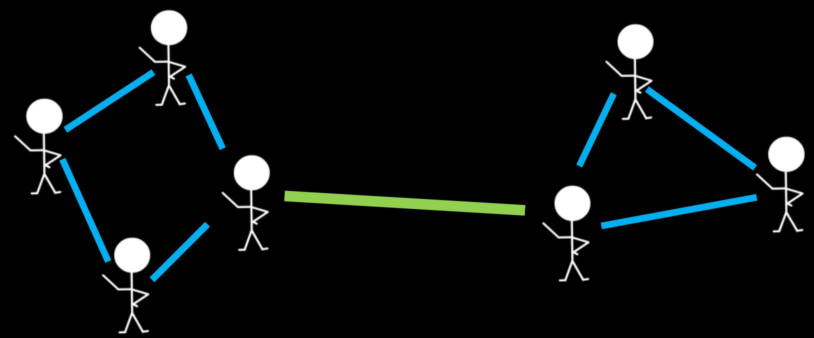


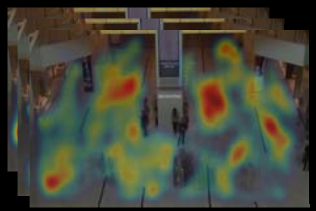
Motion branch

convolution max pooling normalization fully-connected

- multiple frames *[A. Karpathy, et al. CVPR'14]*
- optical flow *[K. Simonyan, et al. CVPR'14]*

How to learn attributes from crowd videos?





Motion channels *

Graph-driven crowd quantifications

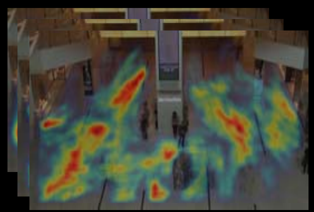
Geometric structure

Topological structure

Interaction

* J. Shao, C. C. Loy, and X. Wang. CVPR'14

How to learn attributes from crowd videos?

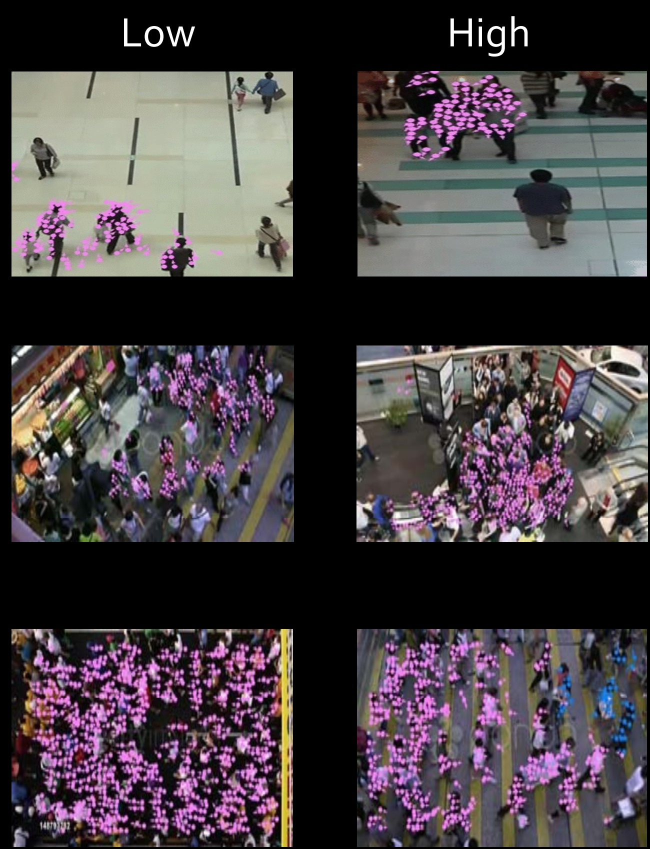


Motion channels *

stability

collectiveness

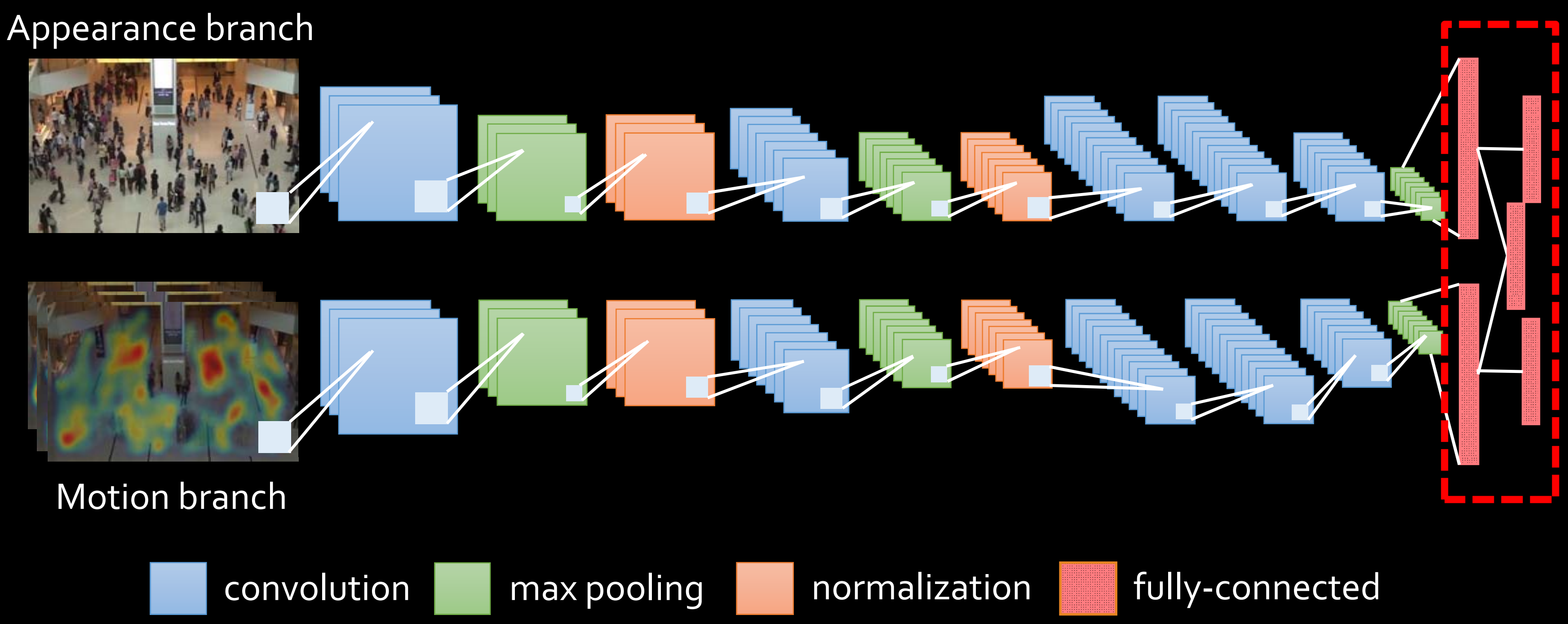
conflict



* J. Shao, C. C. Loy, and X. Wang. CVPR'14

How to learn attributes from crowd videos?

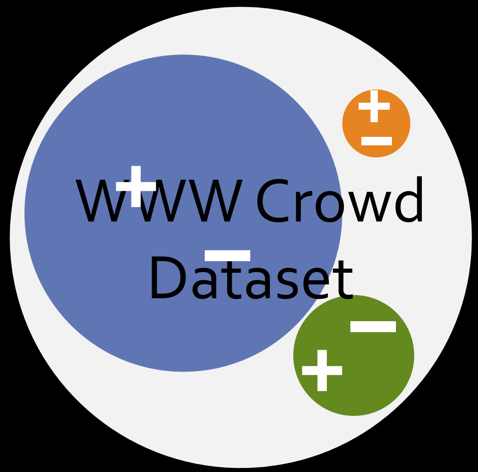
A two-branch CNN model




outdoor: 0.99668
walk: 0.93478
street: 0.81038
pedestrian: 0.70728
parade: 0.40149

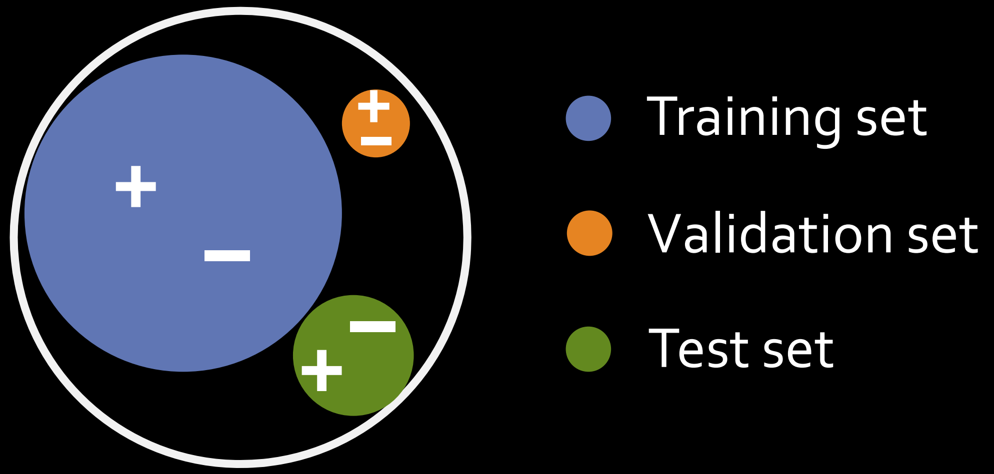

watch performance: 0.65004
audience: 0.58994
outdoor: 0.54932
stand: 0.18374
stadium: 0.15022
dance: 0.9968
outdoor: 0.98079
dancer: 0.97325
fight: 0.97323
mob: 0.96976

Experimental Settings



- Training set
- Validation set
- Test set

Experimental Settings



The proposed models

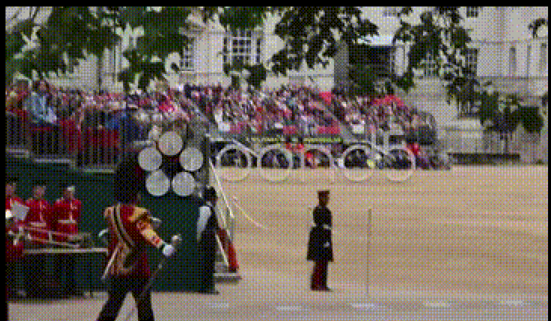
- Deeply Learned Static Features (**DLSF**)
- Deeply Learned Motion Features (**DLMF**)
- The model combining DLSF and DLMF (**DLSF+DLMF**)

DLSF vs. DLSF+DLMF

Correct prediction
Miss detection

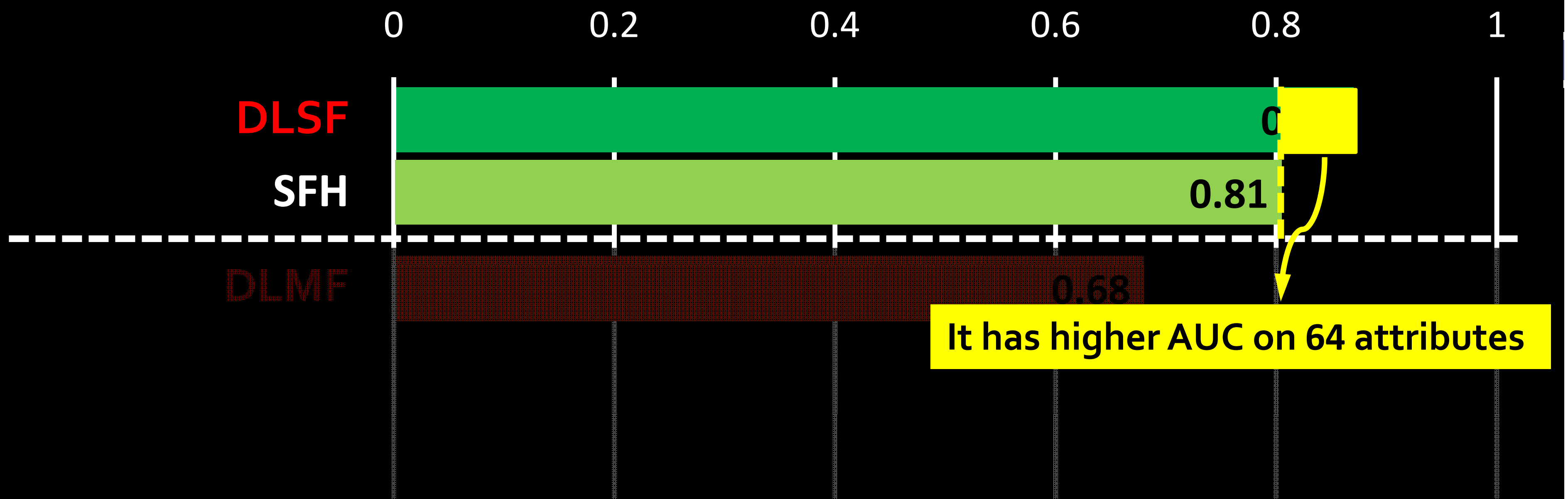
DLSF		DLSF+DLMF	
outdoor	run	outdoor	run
stand	marathon	stand	marathon
runner	street	runner	street

DLSF		DLSF+DLMF	
outdoor	rink	outdoor	rink
skater		skater	
skate		skate	



DLSF			DLSF+DLMF		
outdoor	walk	parader	outdoor	walk	parader
watch performance	performance	soldier	watch performance	performance	soldier
audience	parade		audience	parade	

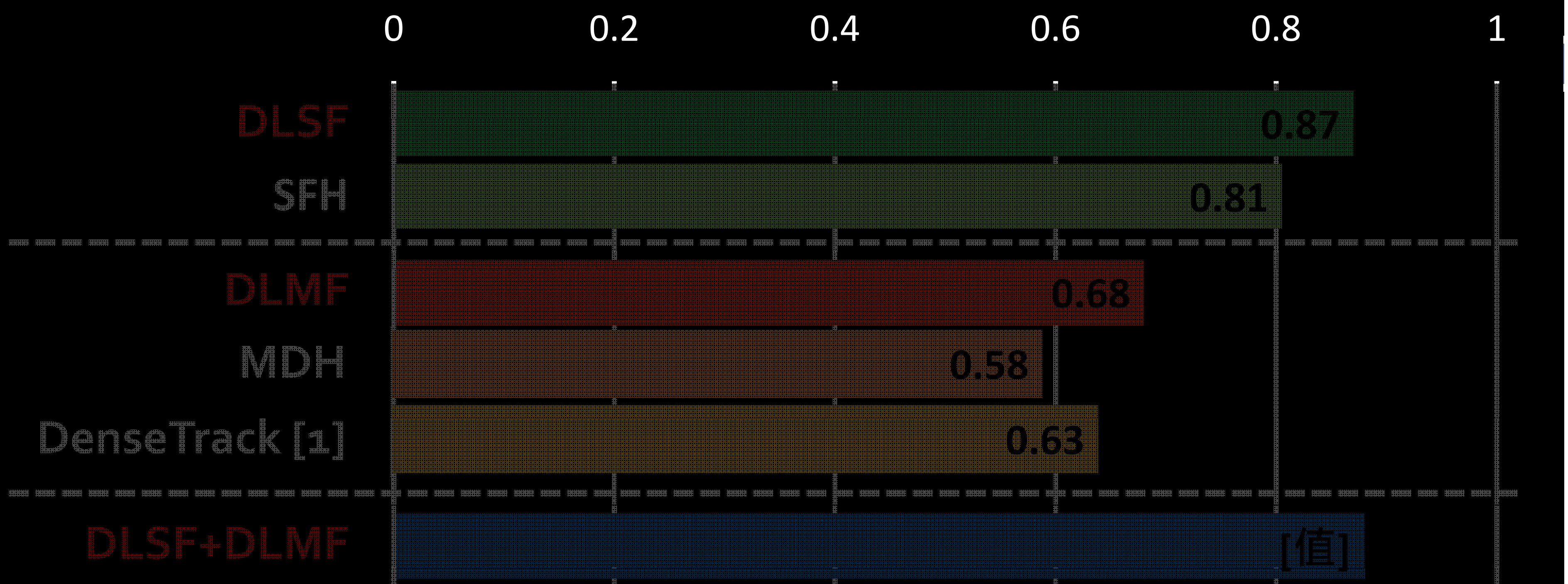
Quantitative Evaluation (AUC)



Static Feature Histogram (SFH)

- **{SIFT, GIST, HOG, Color histogram, SSIM, LBP} → Bag-of-words → SVM**

Quantitative Evaluation (AUC)

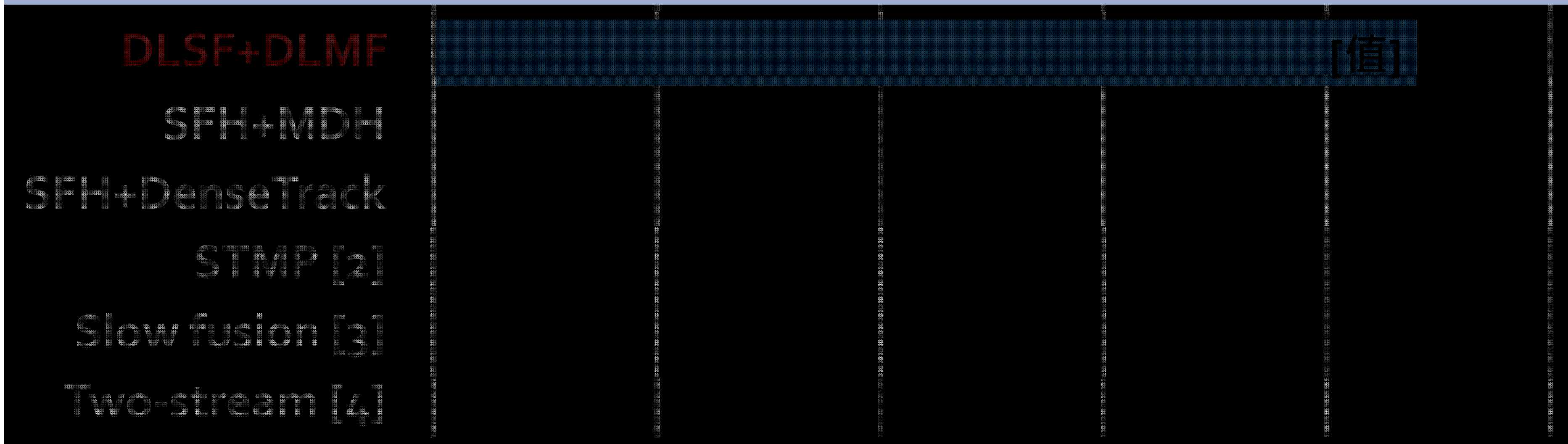


1. The proposed Motion Descriptor Histogram (**MDH**)
2. Dense Trajectory (**DenseTrack**)
   - State-of-the-art in action recognition

[1] H. Wang, et al. CVPR'11 [2] L. Kratz and K. Nishino. CVPR'09 [3] A. Kapathy, et al. CVPR'14 [4] K. Simonyan and A. Zisserman. NIPS'14

Quantitative Evaluation (AUC)

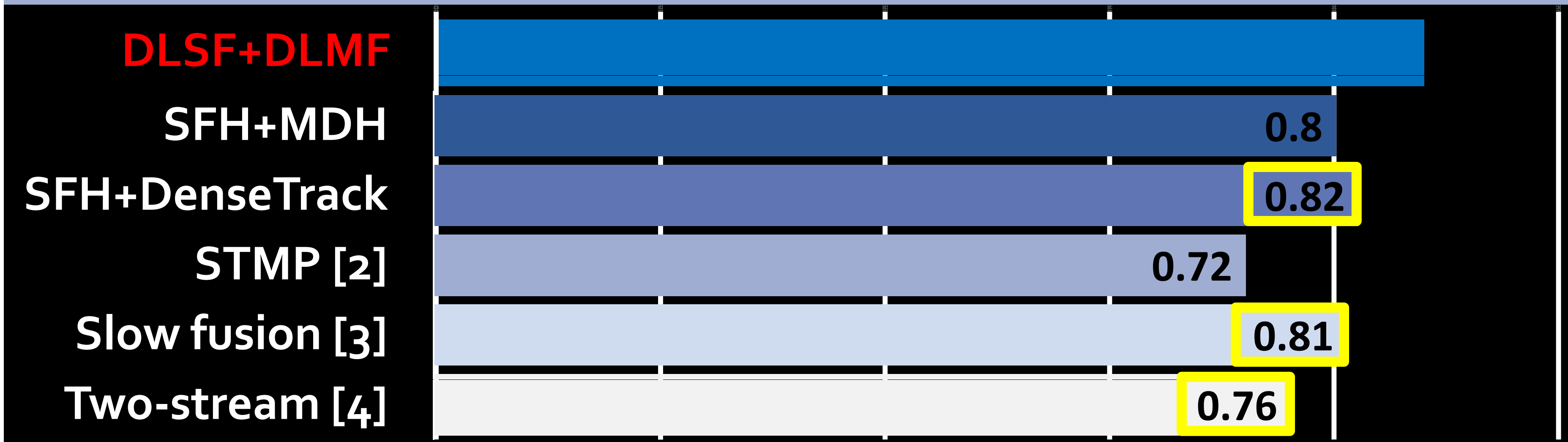
1. Static feature histogram + Motion descriptor histogram (**SFH+MDH**)
2. Static feature histogram + Dense trajectory (**SFH+DenseTrack**)
3. Spatio-temporal motion patterns (**STMP**)
4. Slow fusion scheme with multi-frames as input of CNN (**Slow Fusion**)
   - State-of-the-art deep learning method for (sports) video classification
5. Two-stream CNN with optical flow as input of motion stream (**Two-stream**)
   - State-of-the-art in action recognition



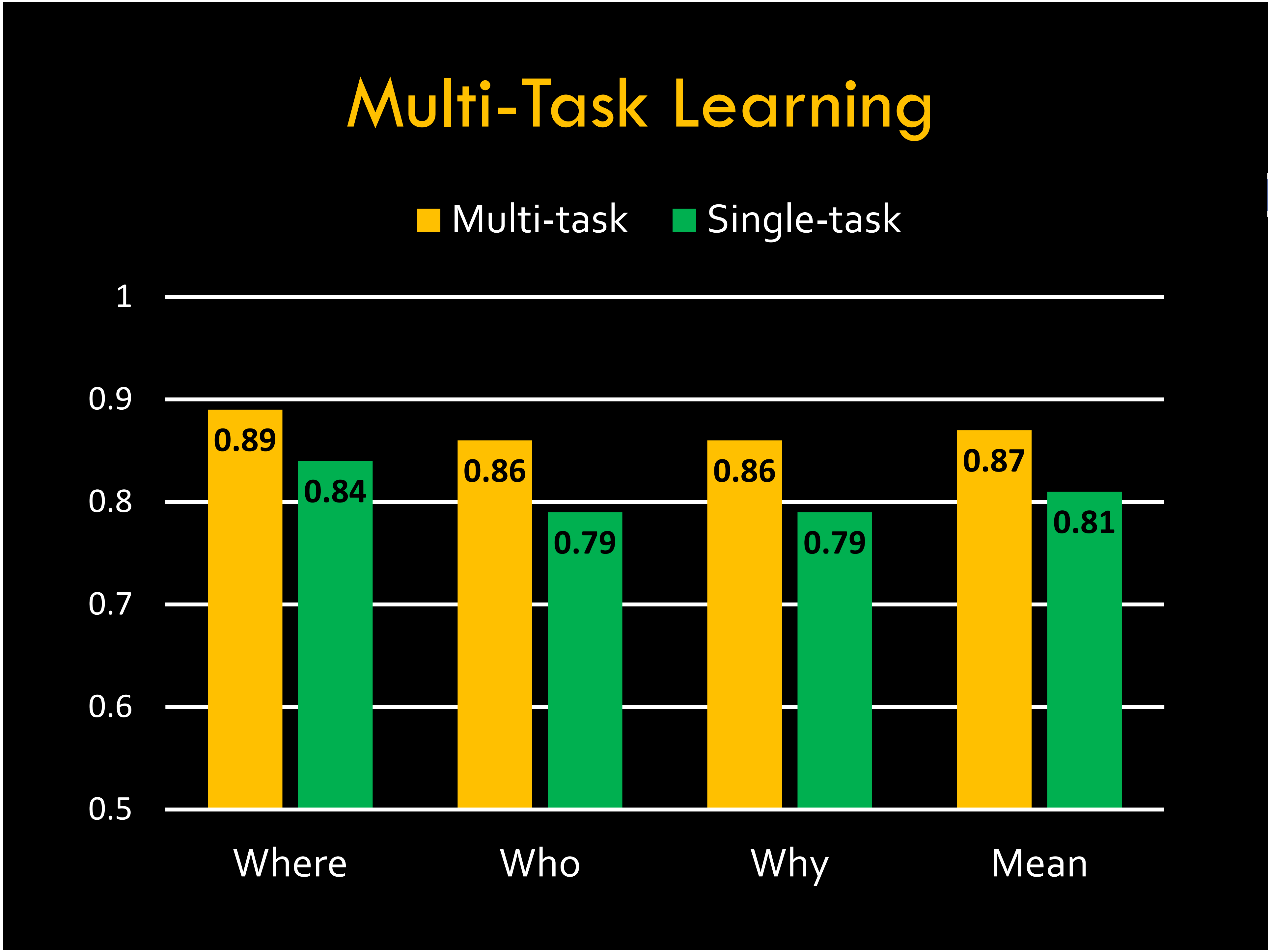
[1] H. Wang, et al. CVPR'11 [2] L. Kratz and K. Nishino. CVPR'09 [3] A. Kapathy, et al. CVPR'14 [4] K. Simonyan and A. Zisserman. NIPS'14

Quantitative Evaluation (AUC)

1. Static feature histogram + Motion descriptor histogram (SFH+MDH)
2. Static feature histogram + Dense trajectory (SFH+DenseTrack)
3. Spatio-temporal motion patterns (STMP)
4. Slow fusion scheme with multi-frames as input of CNN (Slow Fusion)
   - State-of-the-art deep learning method for (sports) video classification
5. Two-stream CNN with optical flow as input of motion stream (Two-stream)
   - State-of-the-art in action recognition



[1] H. Wang, et al. CVPR'11 [2] L. Kratz and K. Nishino. CVPR'09 [3] A. Kapathy, et al. CVPR'14 [4] K. Simonyan and A. Zisserman. NIPS'14


Multi-Task Learning
Multi-task
Single-task
1
0.9
0.8
0.7
0.6
0.5
0.89
0.84
0.86
0.79
0.86
0.79
0.87
0.81
Where
Who
Why
Mean

Conclusion

- Deep learning is driven by large scale training data
- Build diversified surveillance benchmarks, in order to scene-independent features representations
- Learn better feature representations from rich predictions
- Study the semantic meanings of the learned feature representations
- Build connections between deep models and conventional vision systems

Any Questions?