香港中文大學
The Chinese University of Hong Kong

DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection

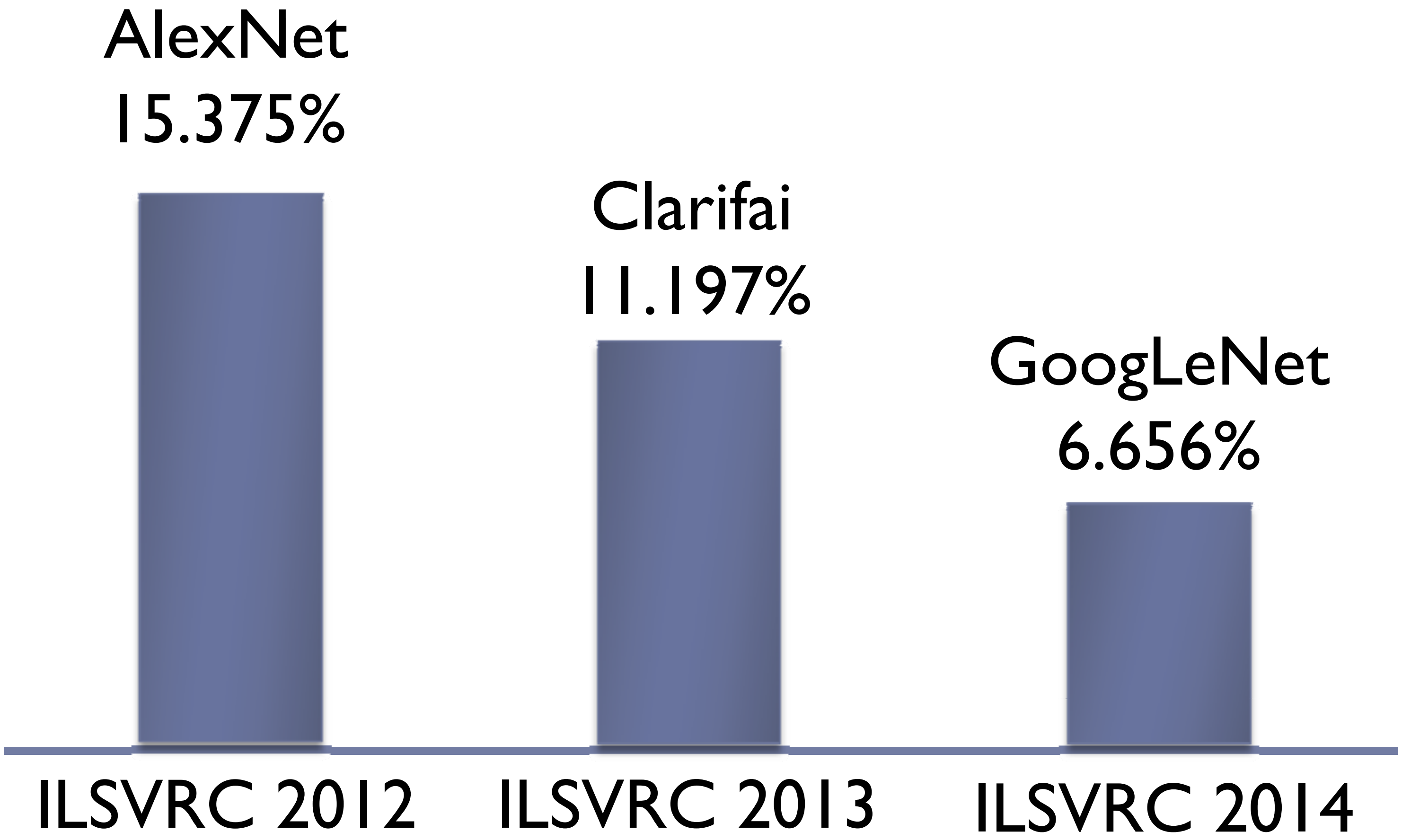
Xiaogang Wang
Department of Electronic Engineering, Chinese University of Hong Kong

ImageNet Image Classification Challenge 2012

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

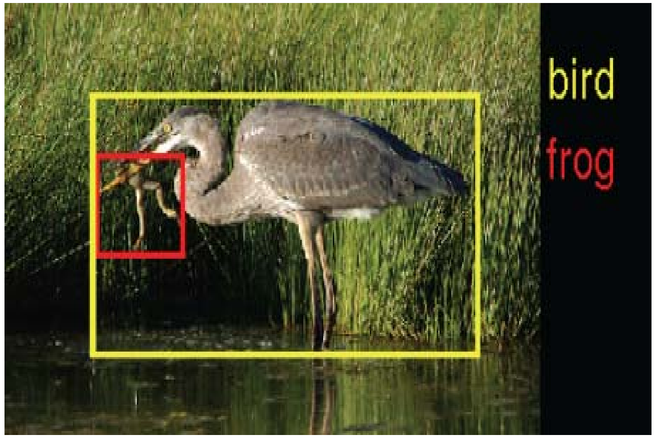
Krizhevsky, Sutskever, Hinton, NIPS'12

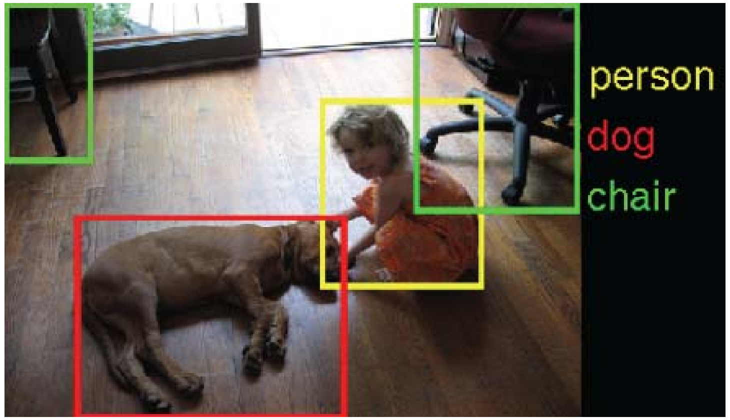
Top5 Image Classification Error on ImageNet

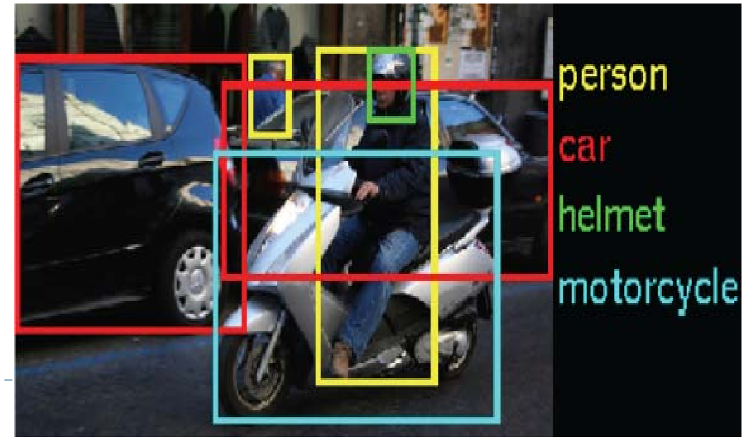


ImageNet Object Detection Task (2013)

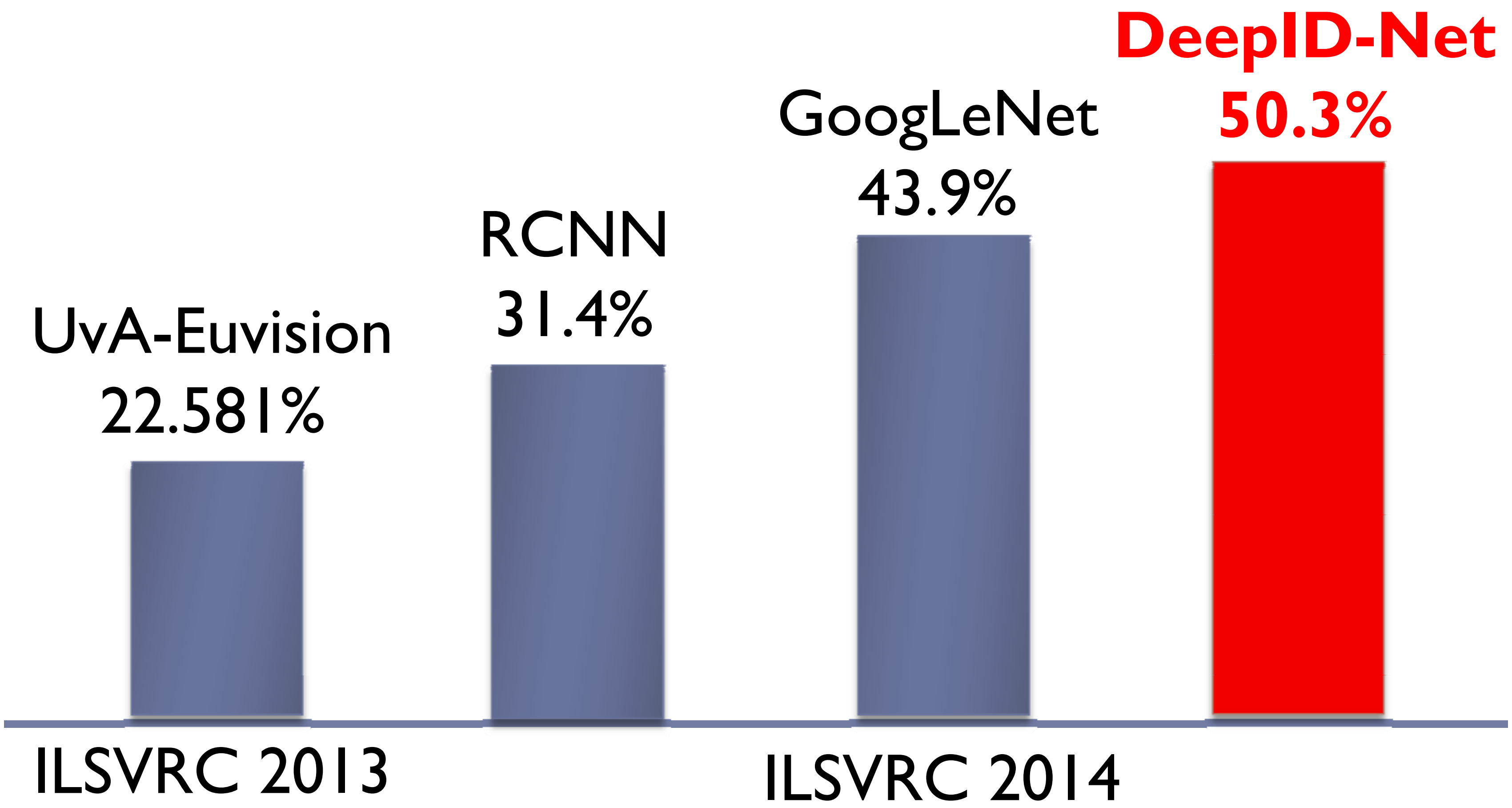
- 200 object classes
- 40,000 test images





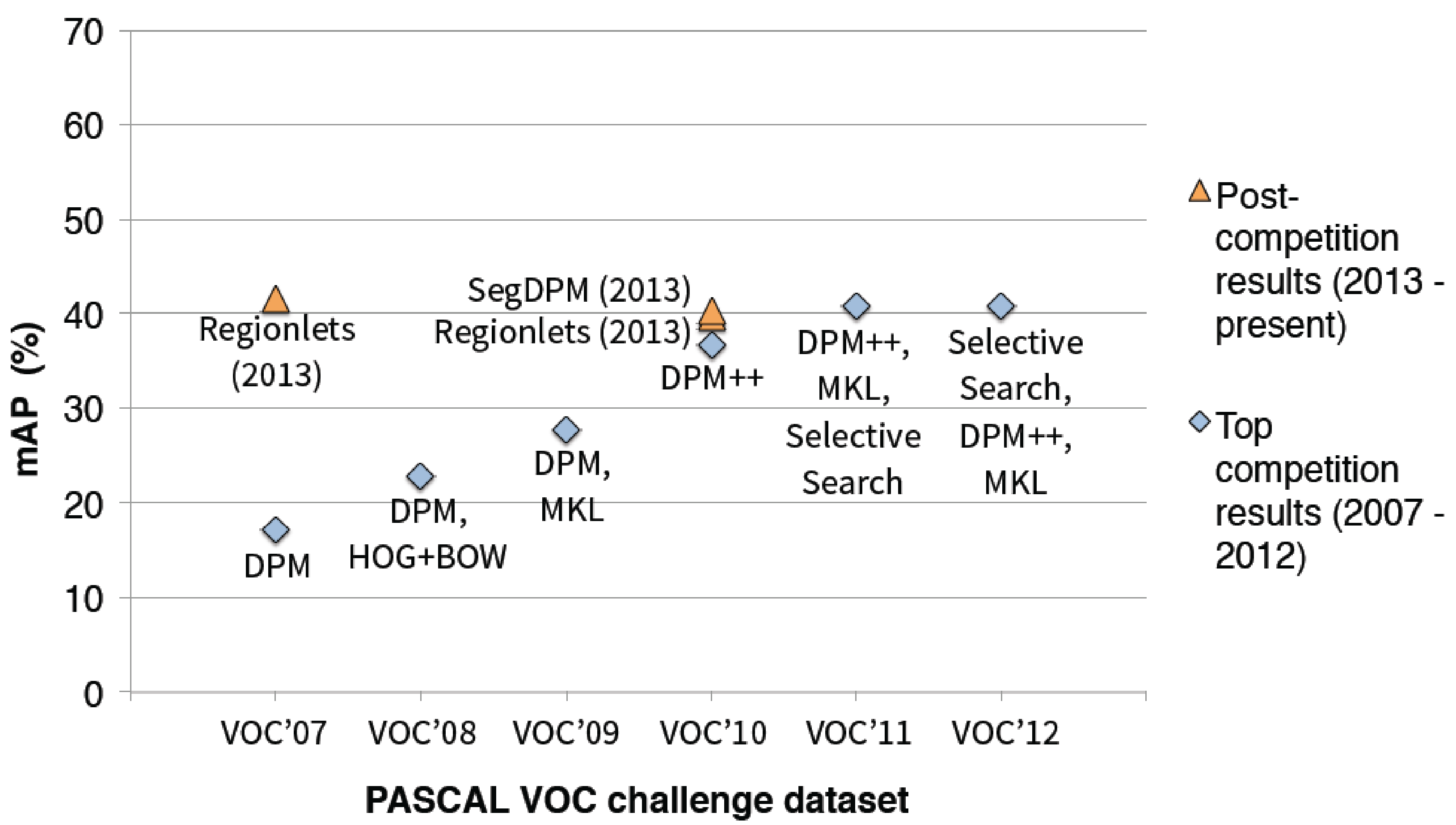


Mean Average Precision (mAP)

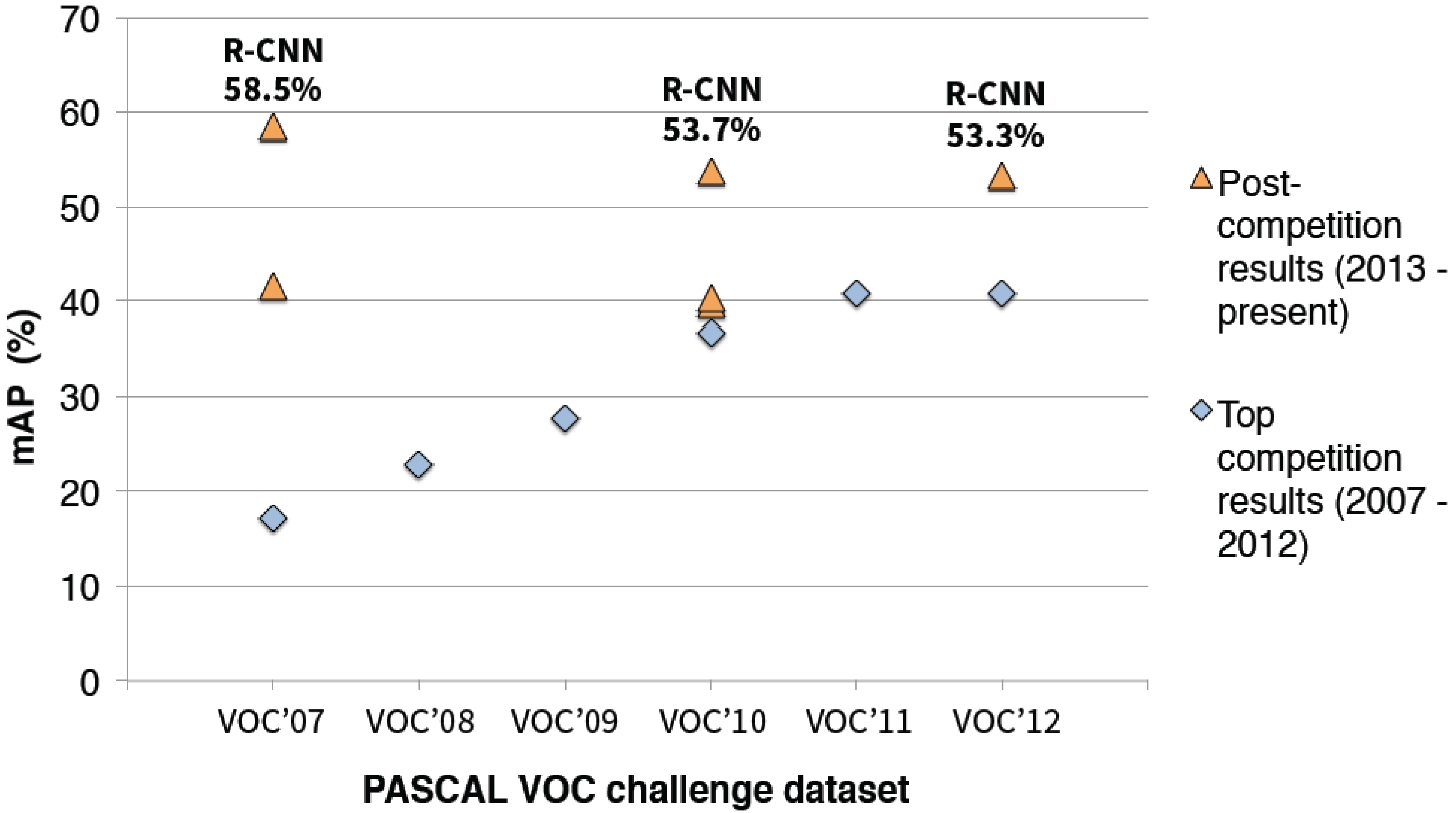


W. Ouyang and X. Wang, et al. "DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection," CVPR 2015

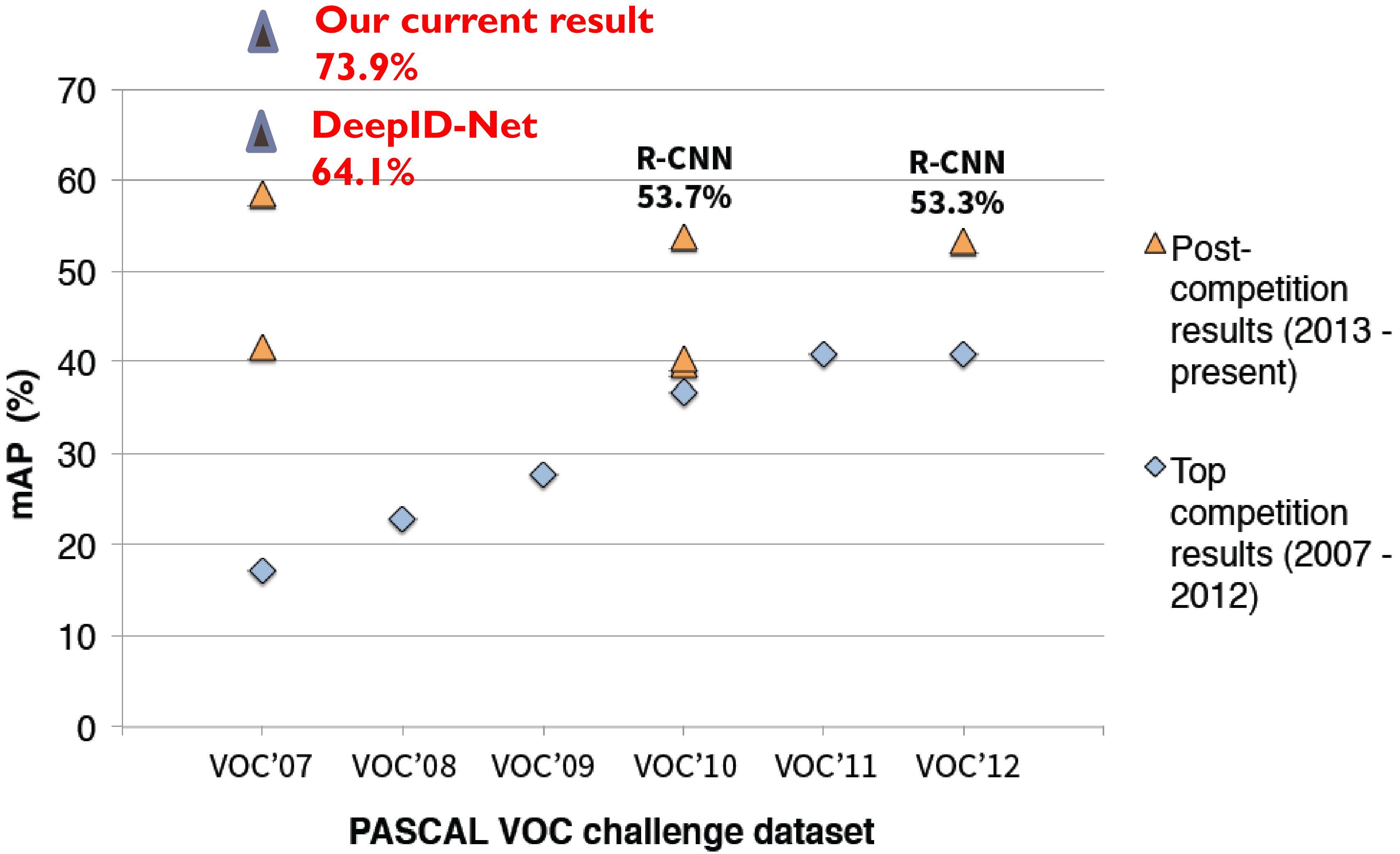
PASCAL VOC (SIFT, HOG, DPM…)



PSCAL VOC (CNN features)

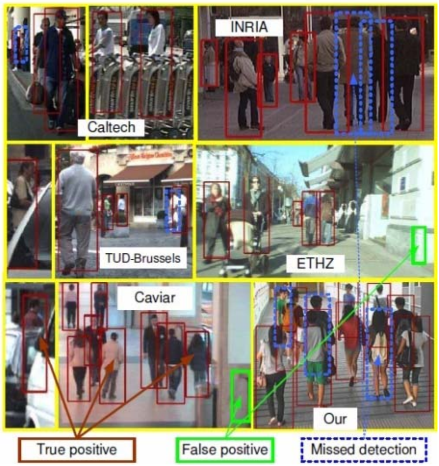


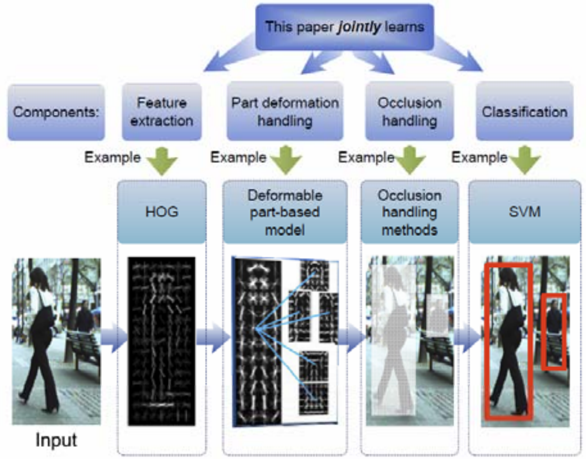
PSCAL VOL (CNN features)



Pedestrian Detection

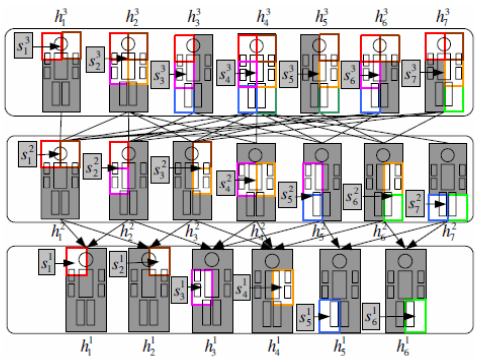
Improve state-of-the-art average miss detection rate on the largest Caltech dataset from 63% to 17%



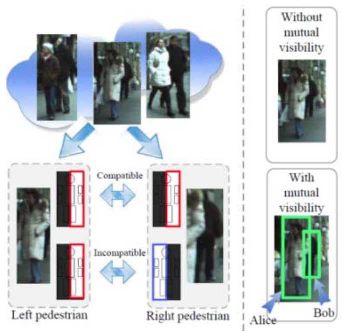


ICCV'13

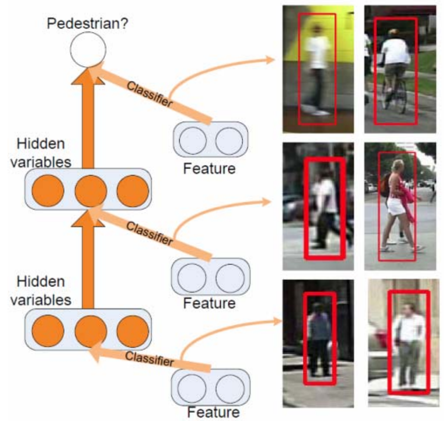
CVPR'14



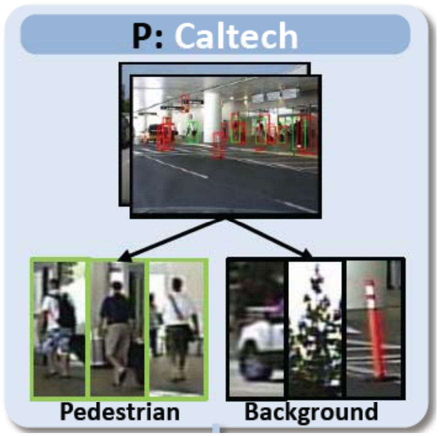
CVPR'12



CVPR'13

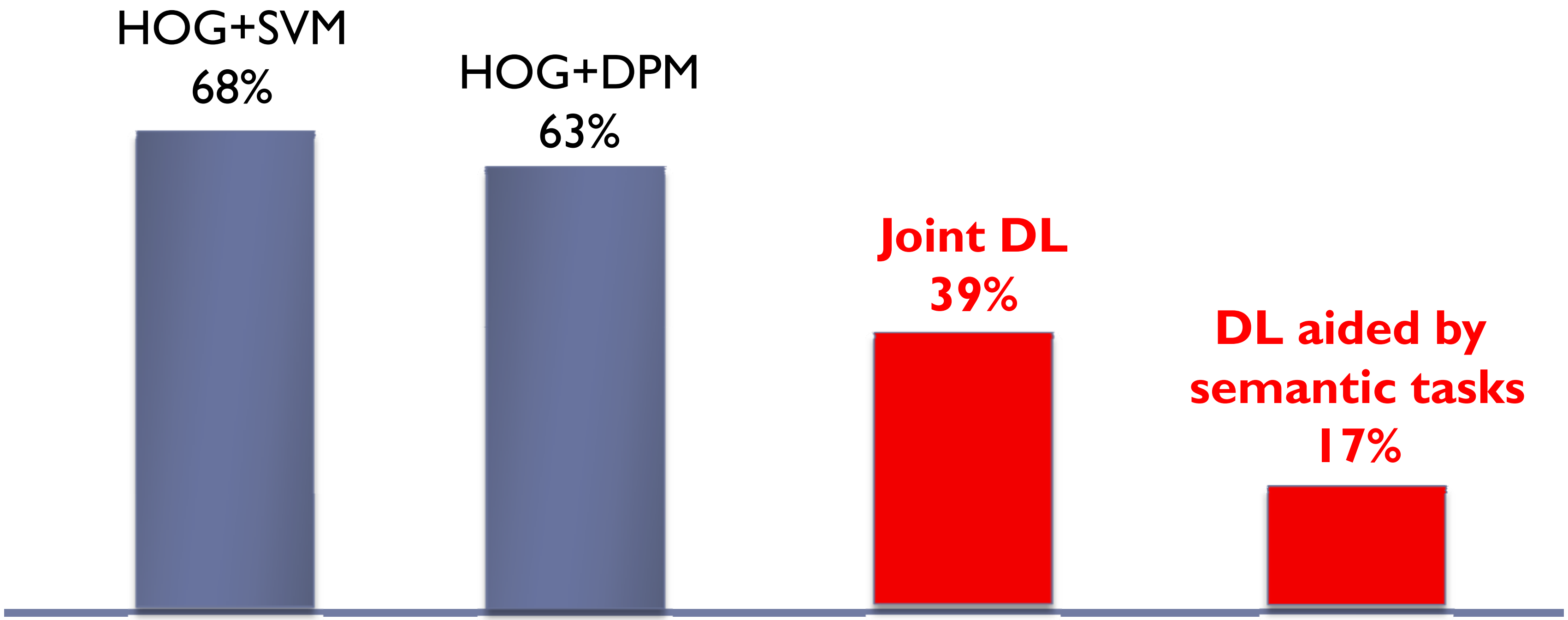


ICCV'13



CVPR'15

Pedestrian Detection on Caltech (average miss detection rates)



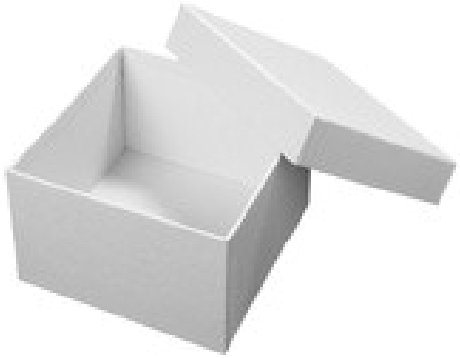
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," ICCV 2013.

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

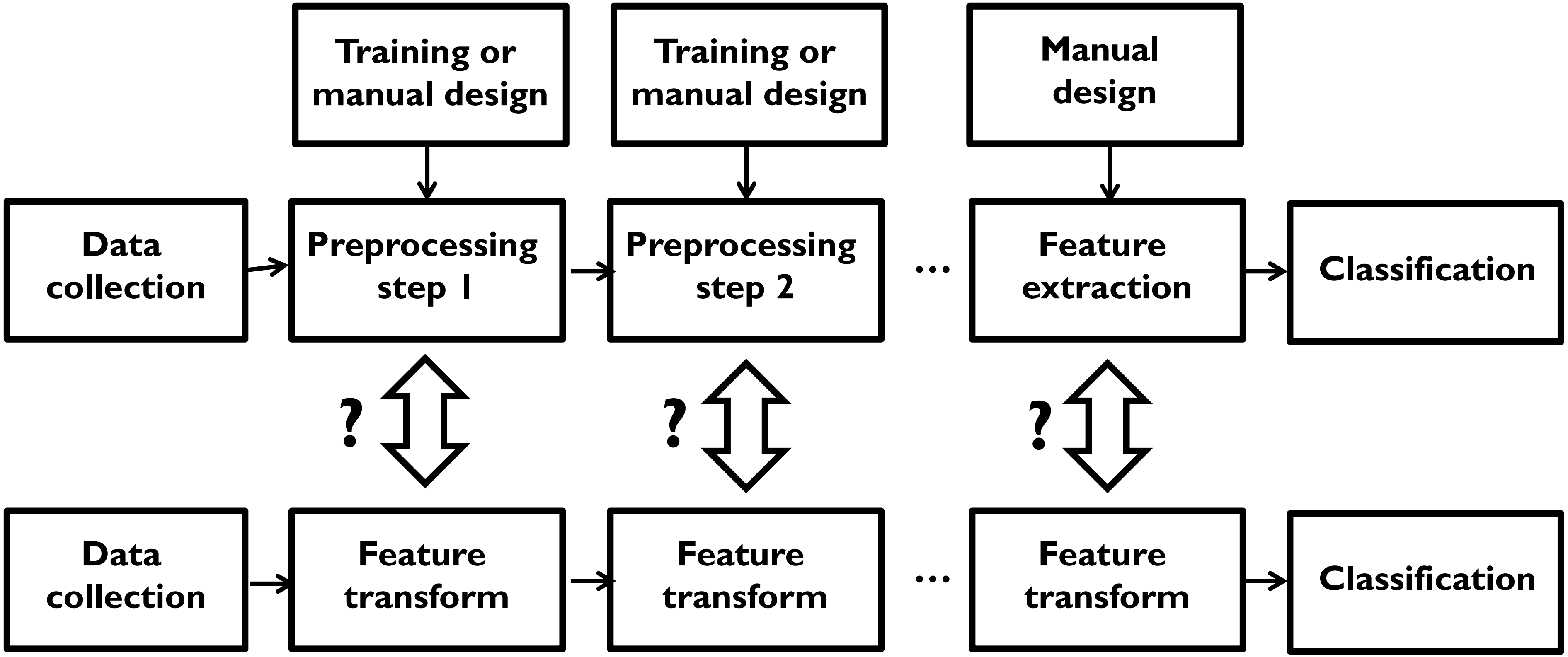
Outline

- **Joint deep learning: pedestrian detection**
- DeepID-Net: general object detection on ImageNet
- Conclusions

Is deep model a black box?



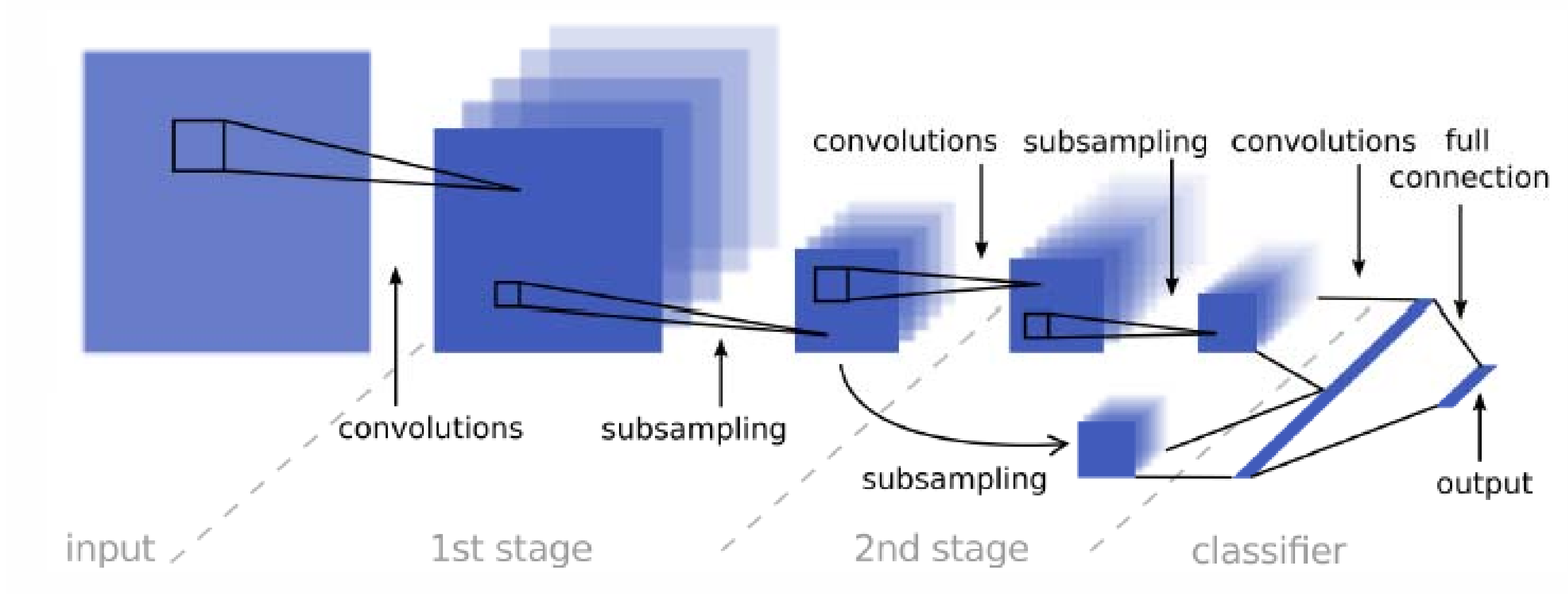
Joint Learning vs Separate Learning



End-to-end learning

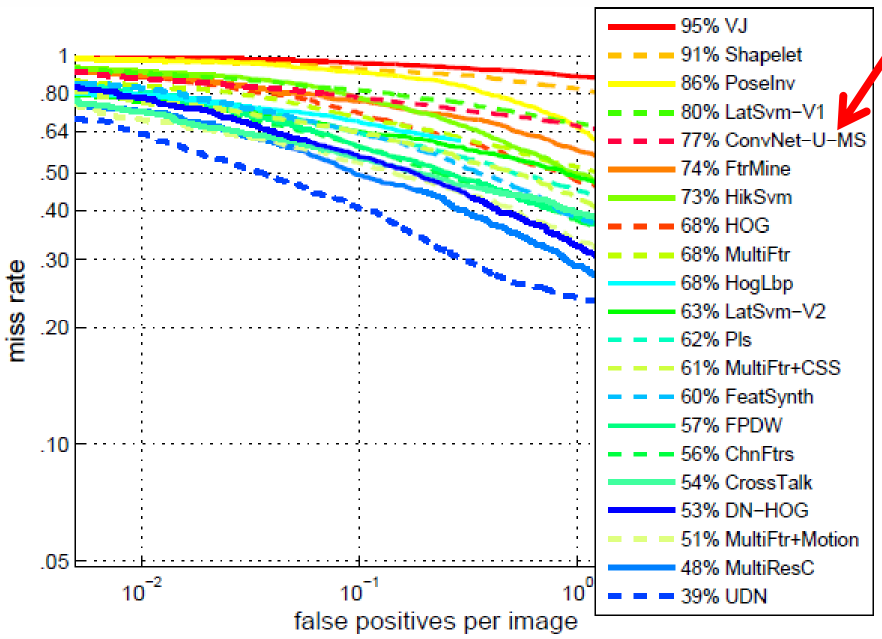
Deep learning is a framework/language but not a black-box model

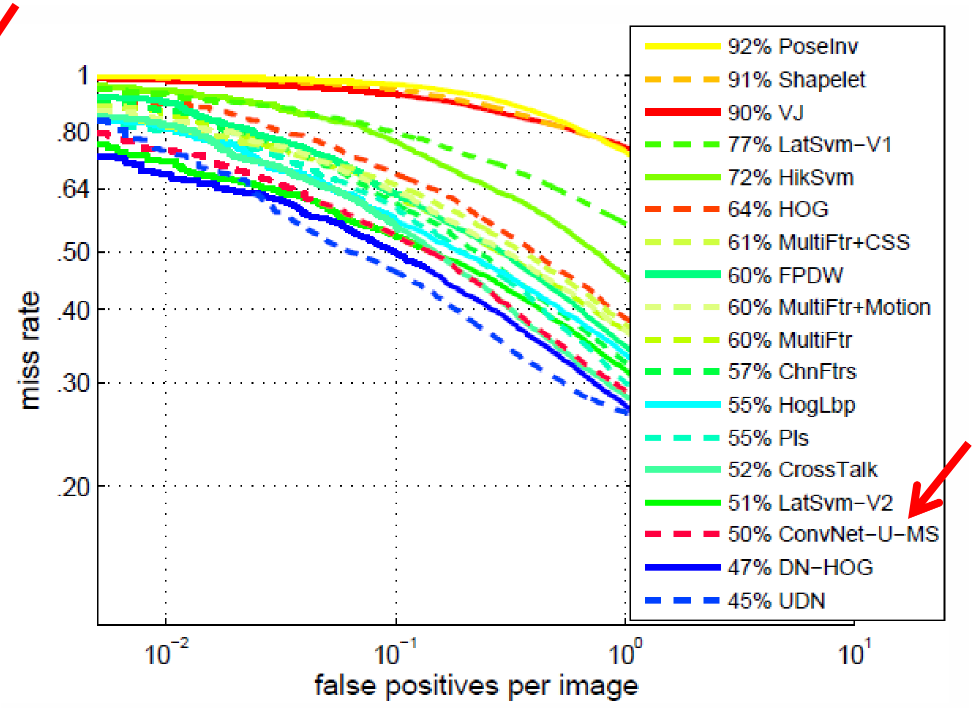
Its power comes from joint optimization and increasing the capacity of the learner



ConvNet−U−MS

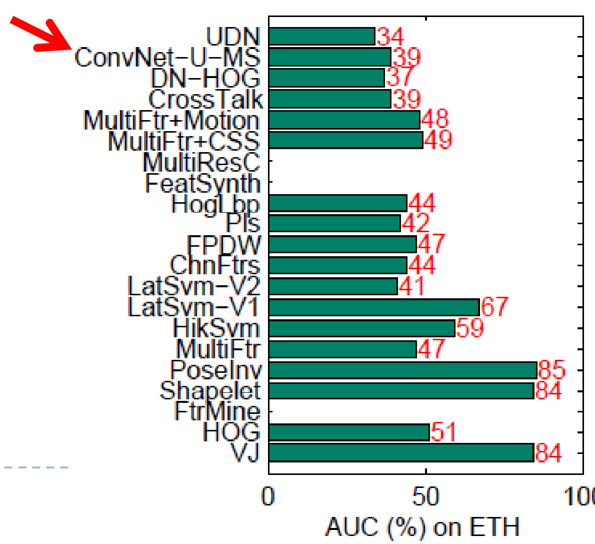
– Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, "Pedestrian Detection with Unsupervised Multi-Stage Feature Learning," CVPR 2013.

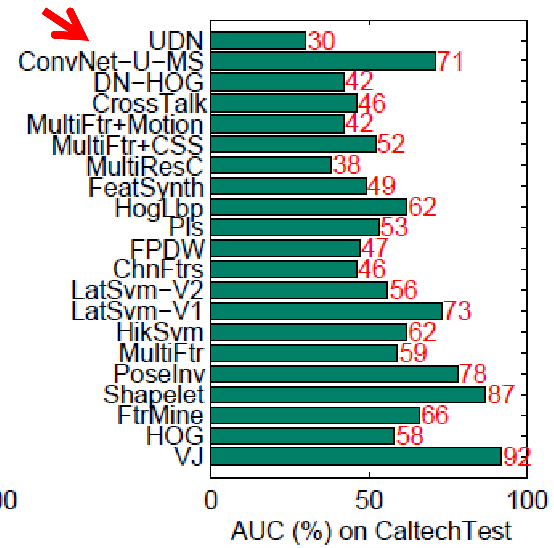


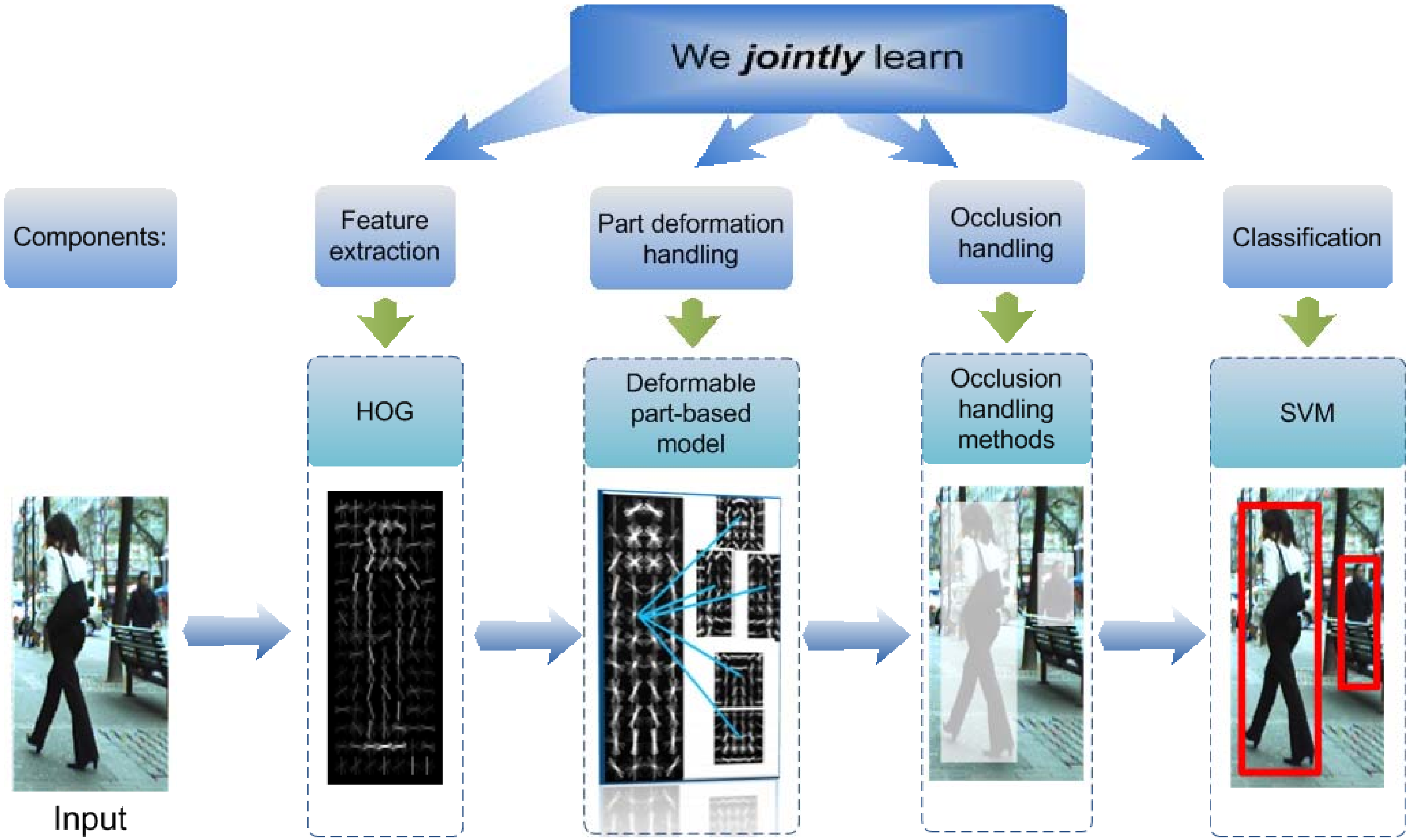


Results on Caltech Test

Results on ETHZ

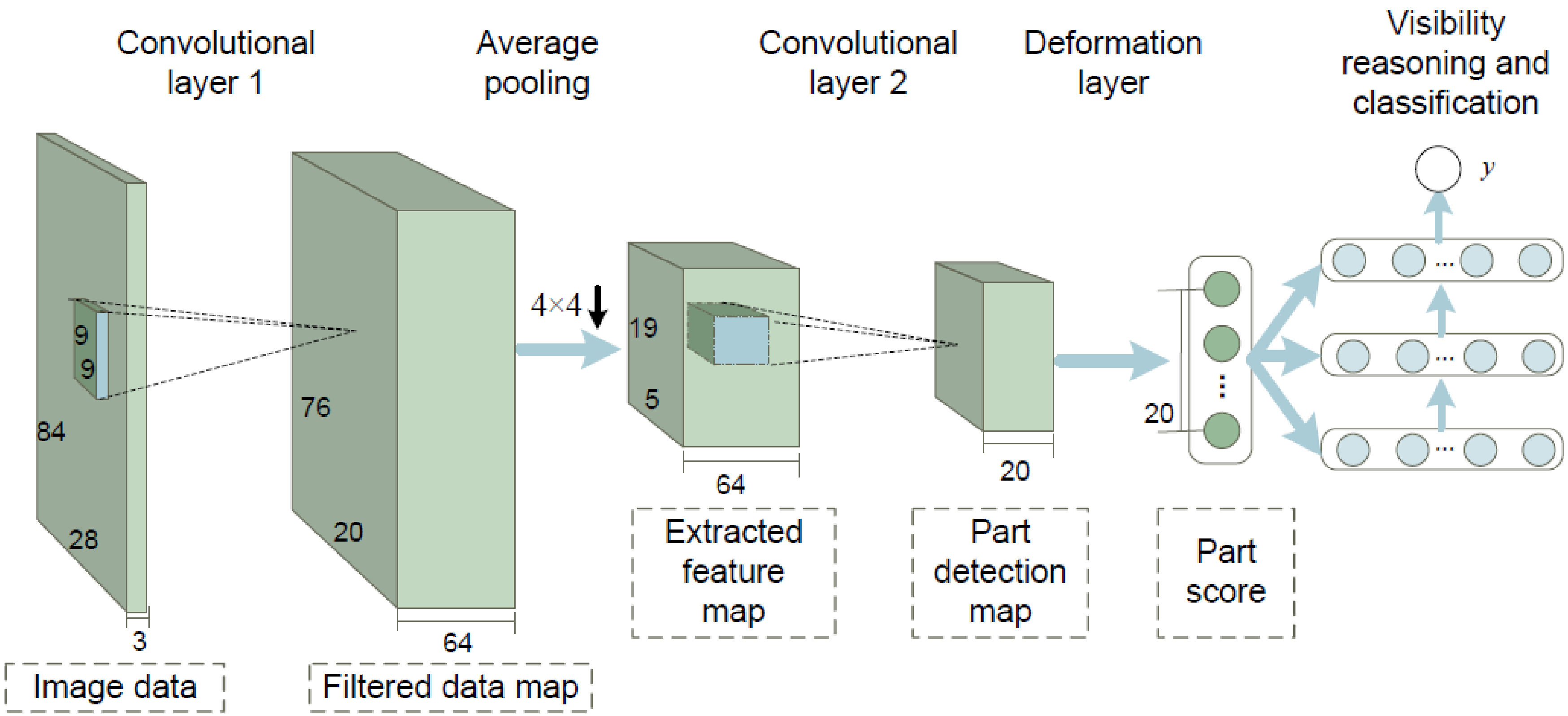






- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)
- P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)
- W. Ouyang and X. Wang. A Discriminative Deep Model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

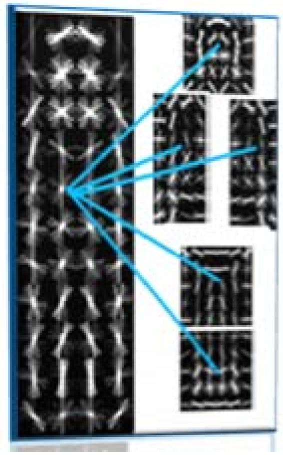
Our Joint Deep Learning Model



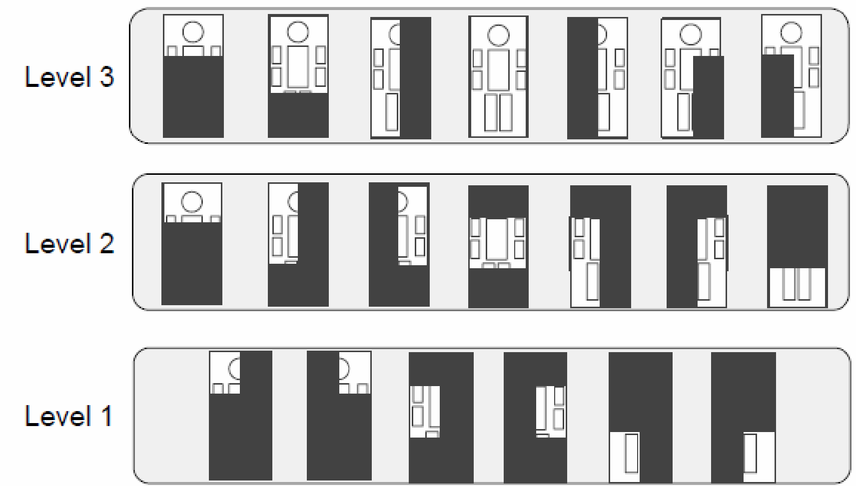
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling Part Detectors

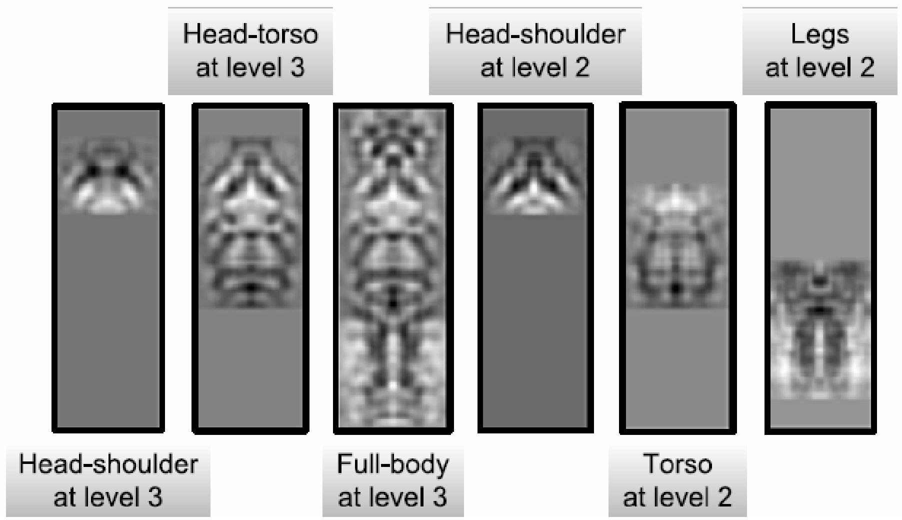
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

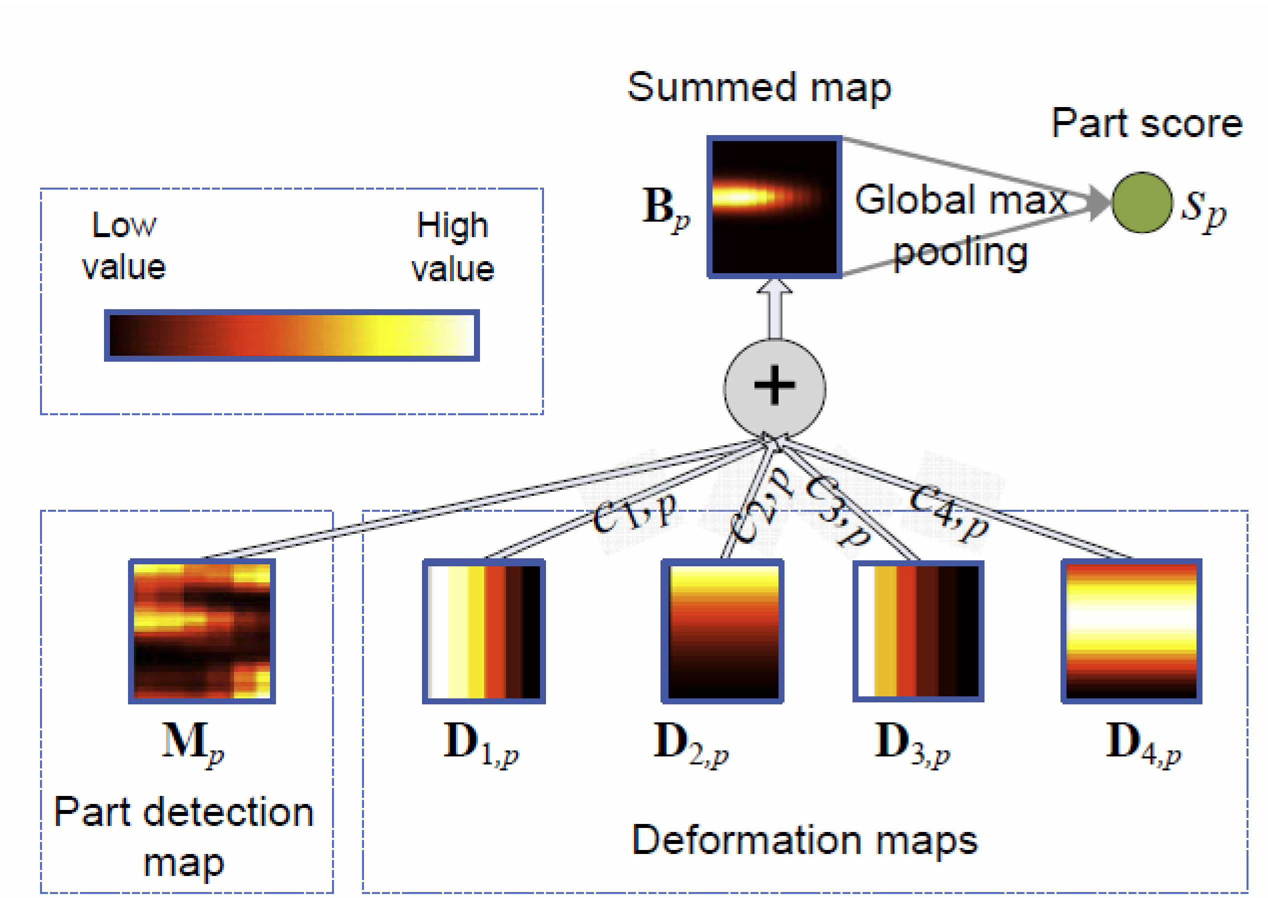


Part models

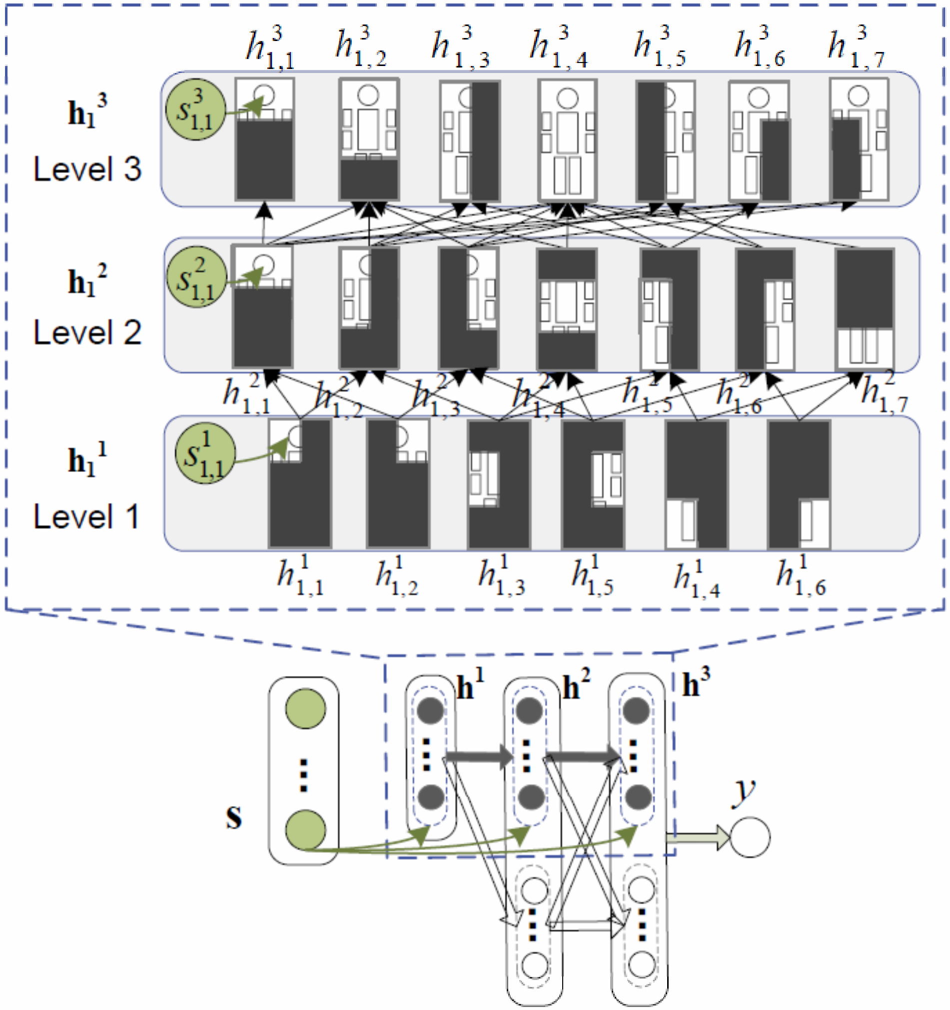


Learned filtered at the second convolutional layer

Deformation Layer



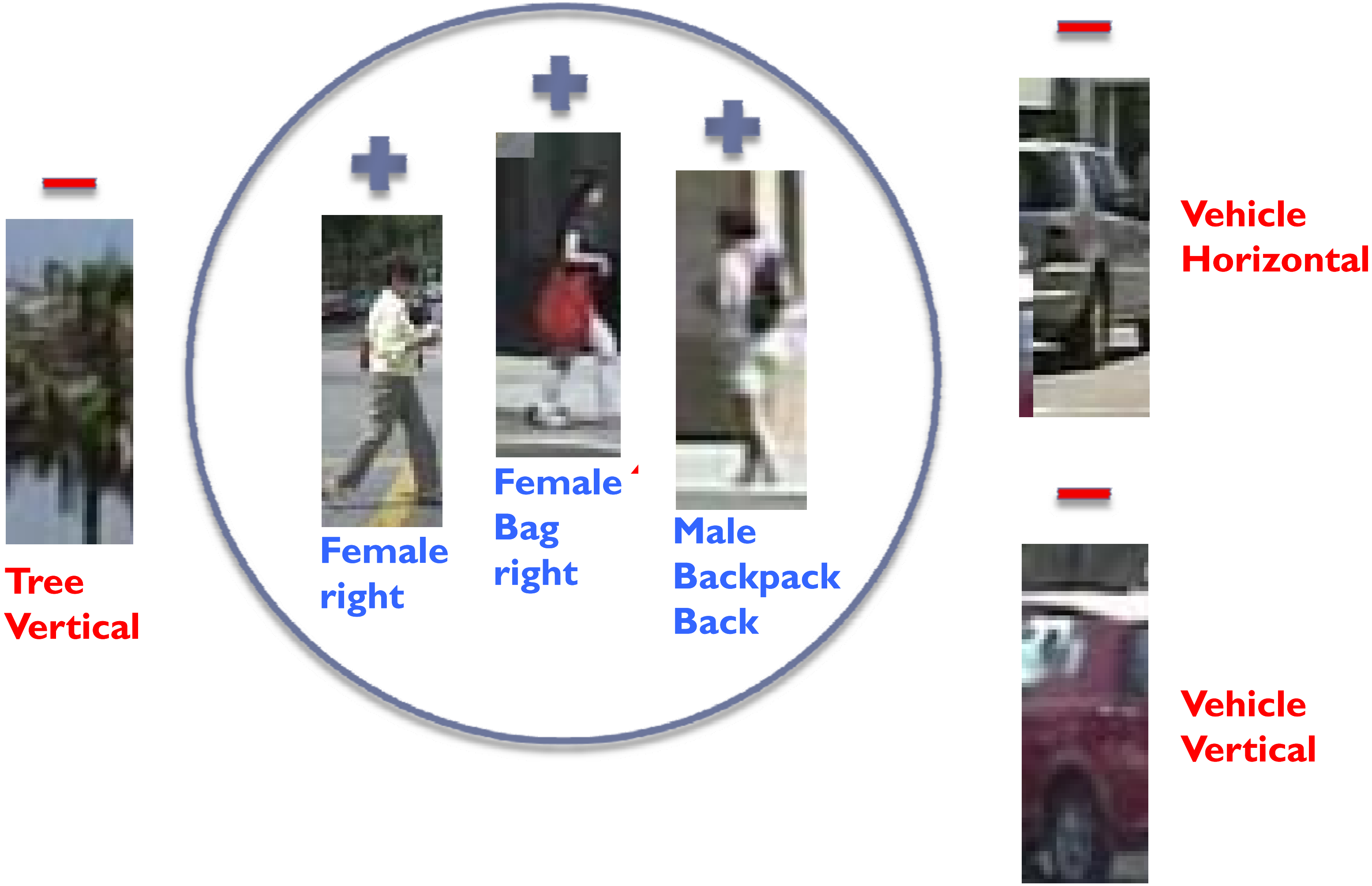
Visibility Reasoning with Deep Belief Net



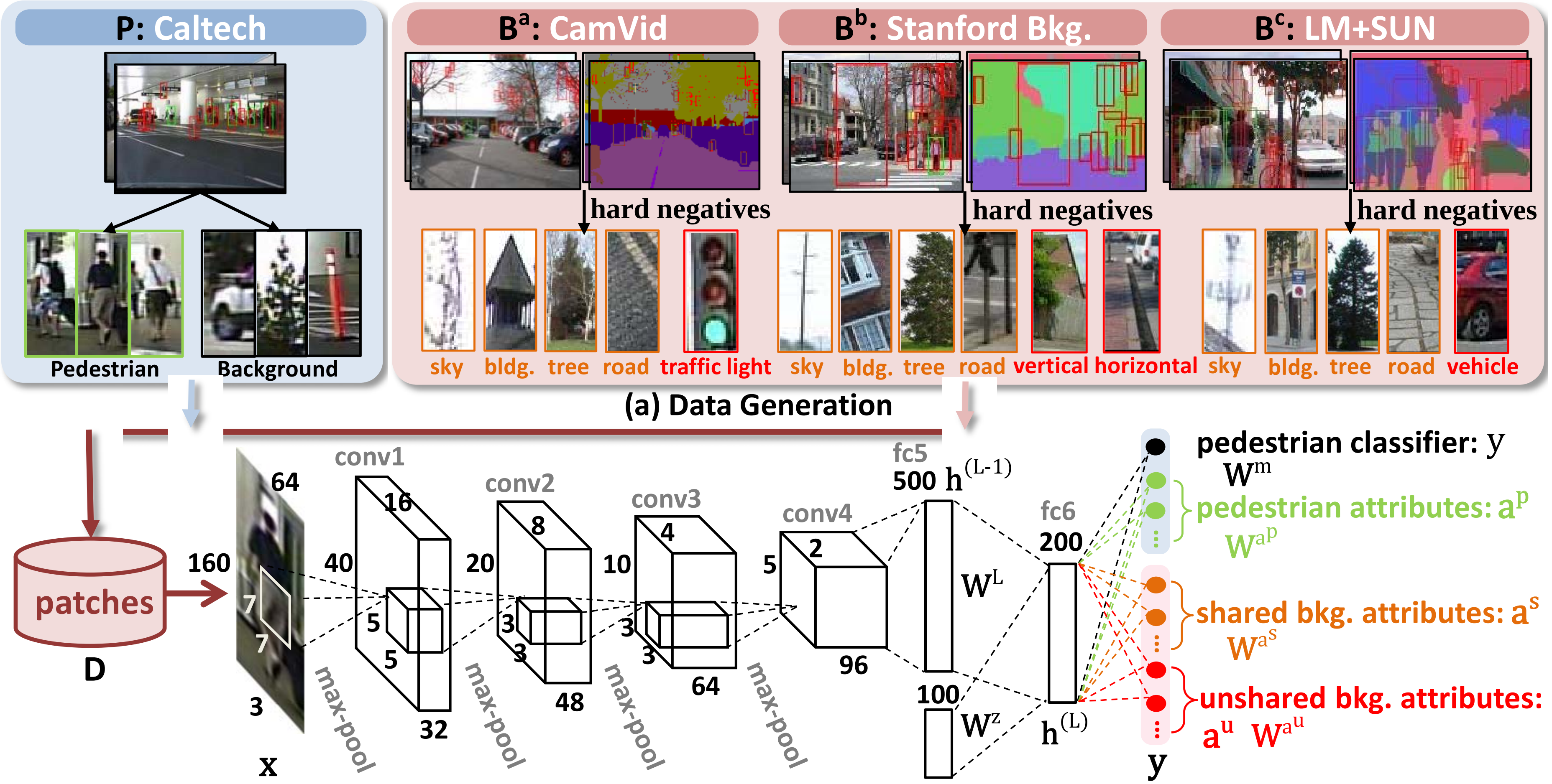
$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}^l_{*,j} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

Pedestrian Detection aided by Deep Learning Semantic Tasks



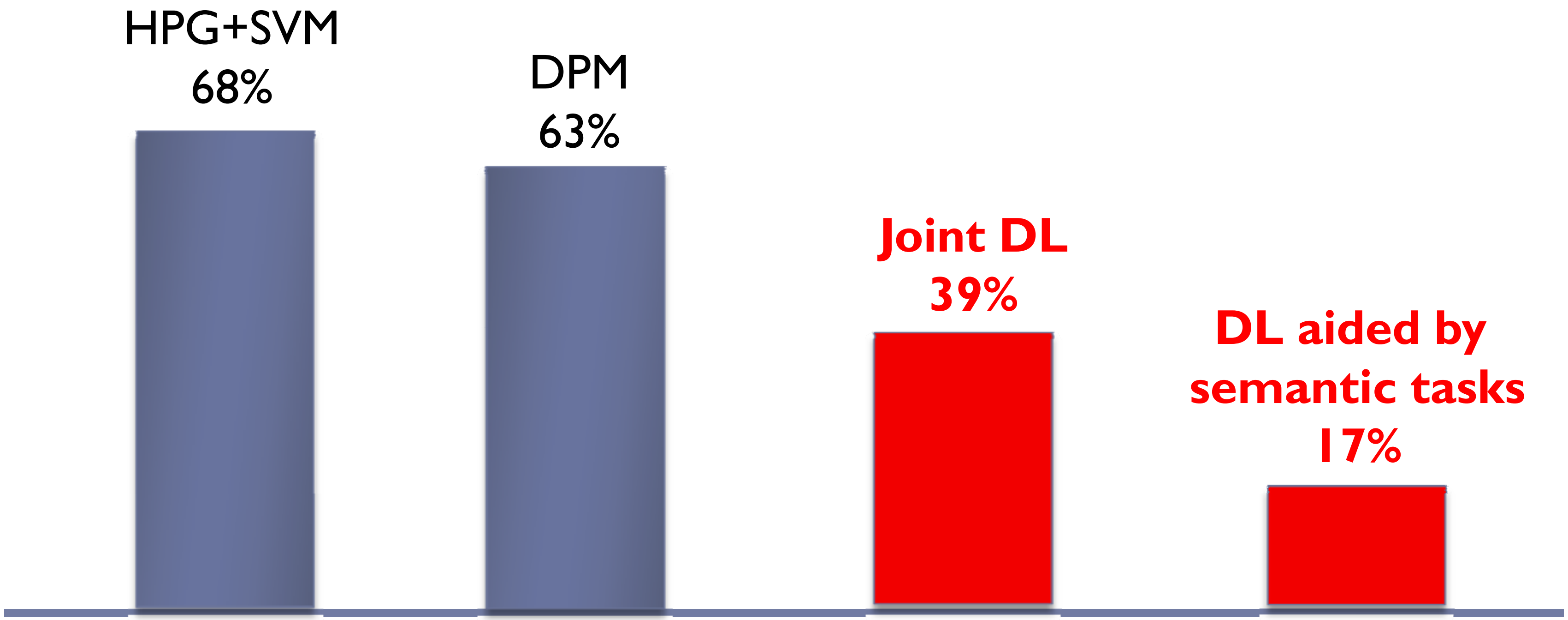
- Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," CVPR 2015



(a) Data Generation

(b) TA-CNN

Pedestrian Detection on Caltech (average miss detection rates)



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

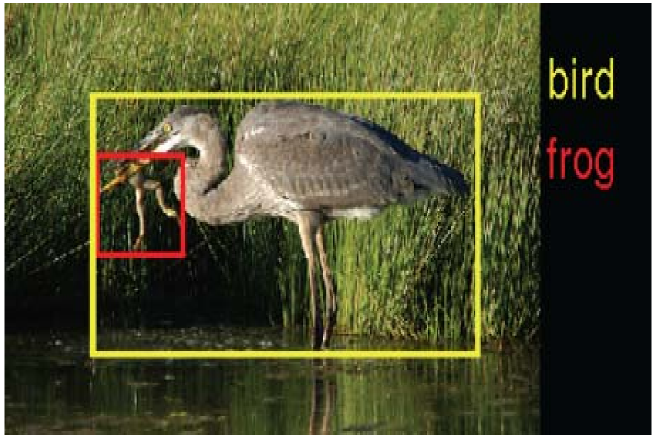
Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," CVPR 2015.

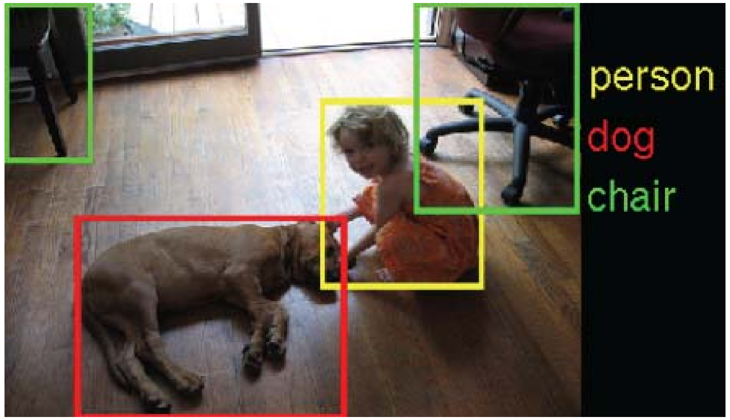
Outline

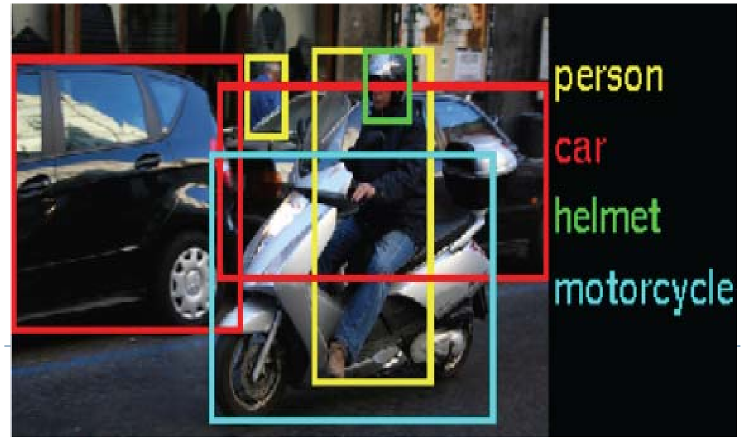
- Joint deep learning: pedestrian detection
- **DeepID-Net: general object detection on ImageNet**
- Conclusions

Challenges of Object Detection

- Huge number of classes
- Appearance variation in different classes



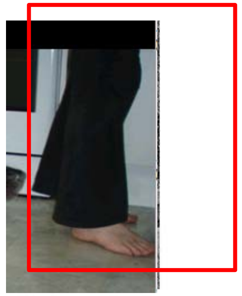




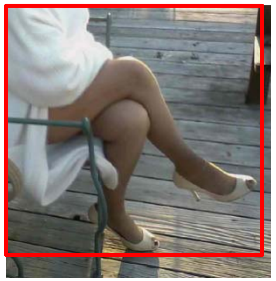
Challenges -- person

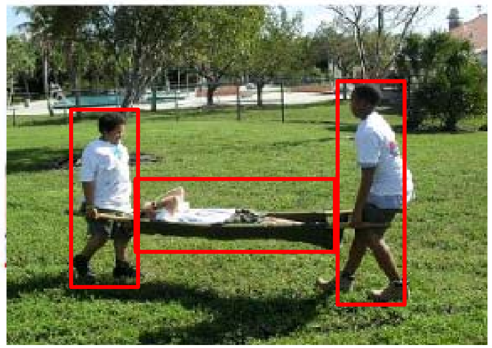
- Intra-class variation
  - Part existence










Challenges -- person

- Intra-class variation
  - Part existence
  - Color

Challenges -- person

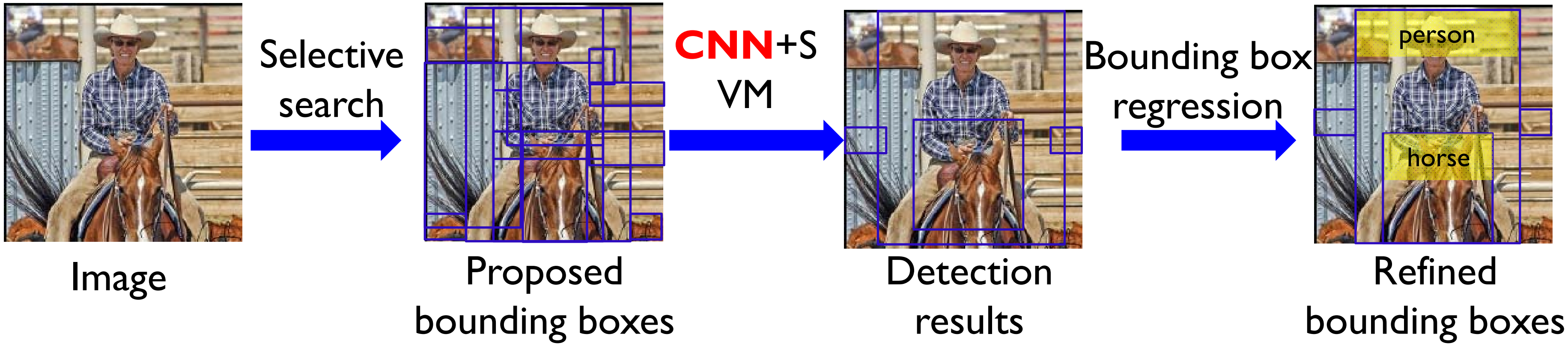
- Intra-class variation
  - Part existence
  - Color
  - Occlusion

Challenges -- person

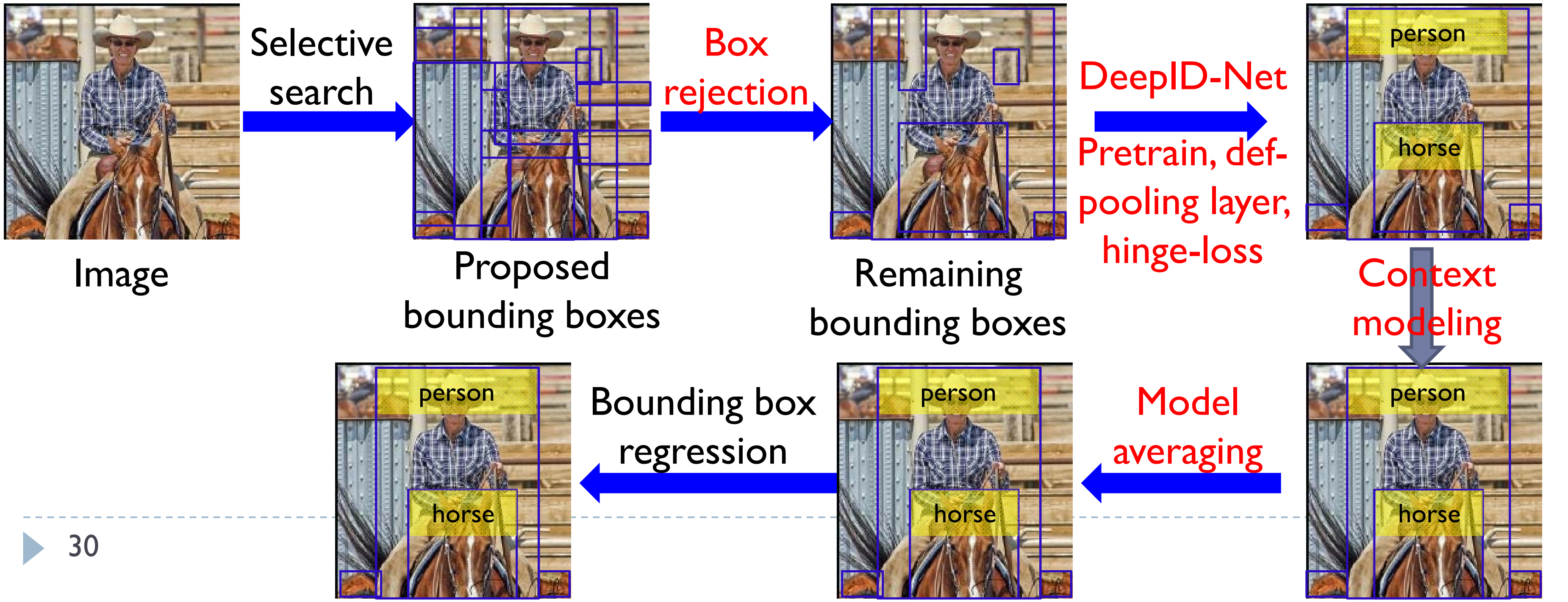
- Intra-class variation
  - Part existence
  - Color
  - Occlusion
  - Deformation

Object Detection on ImageNet

RCNN (**mean average precision: 31.4%**)



DeepID-Net (**mean average precision: 50.3%**)

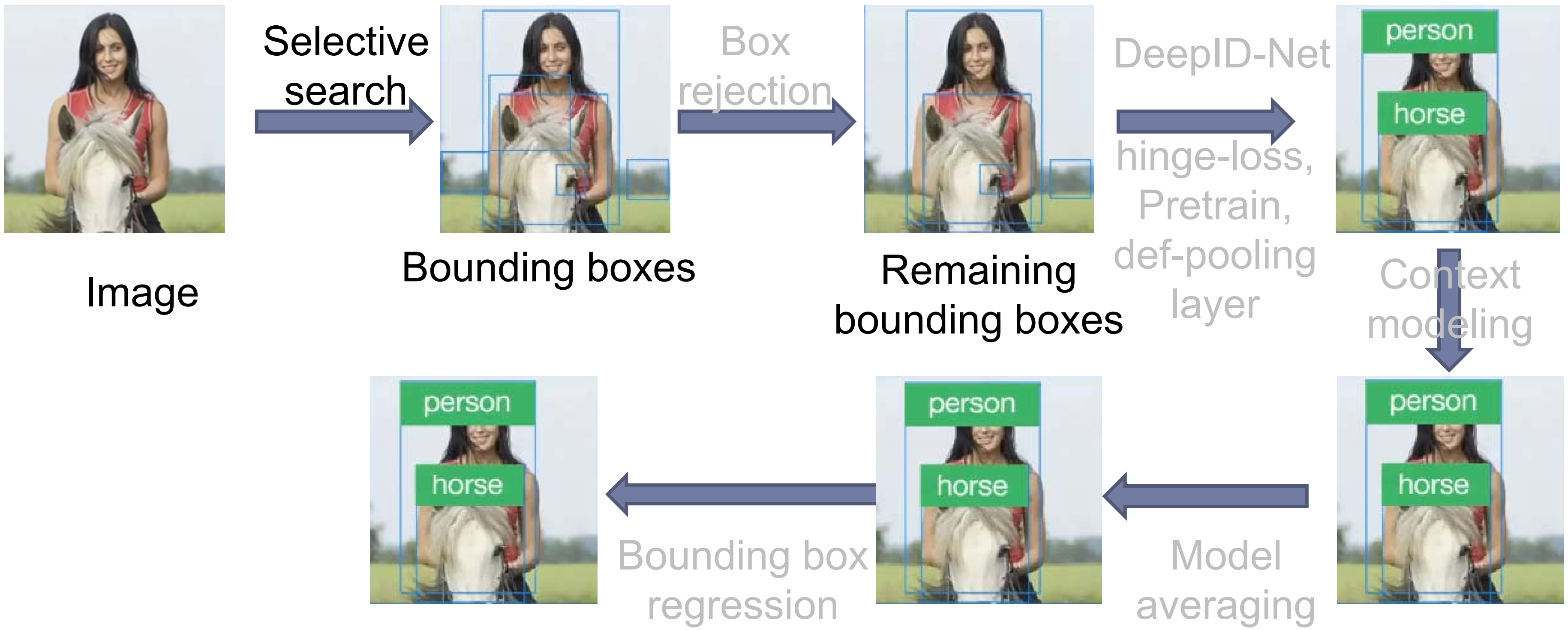


Consideration for deep learning based general object detection

- Time
  - Test
  - Training
- Accuracy
  - Learning discriminative and invariant features
  - Capture complex deformation and occlusion of parts
  - Rich contextual information

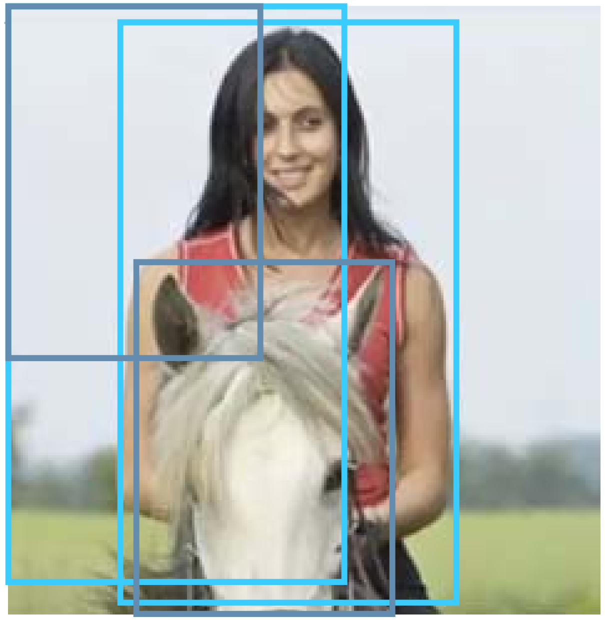
mAP 31 → to 50.3

Our pipeline



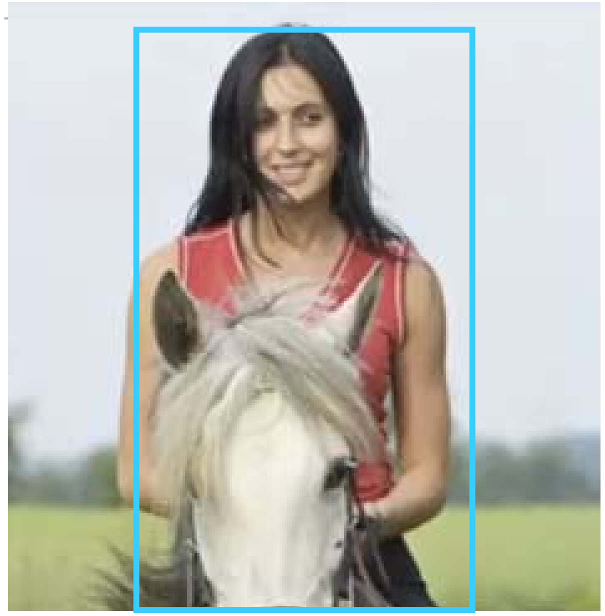
Object detection – old framework

- **Sliding window**
- Feature extraction
- Classification



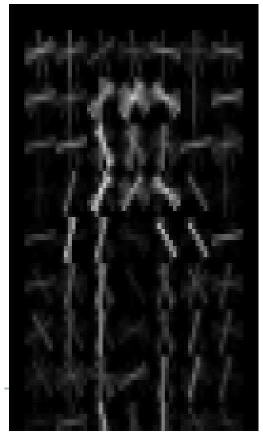
For each window size
For each window
1. Feature extraction
2. Classification
End;
End;

Object detection – the framework



Sliding window

Feature exaction



Feature vector:

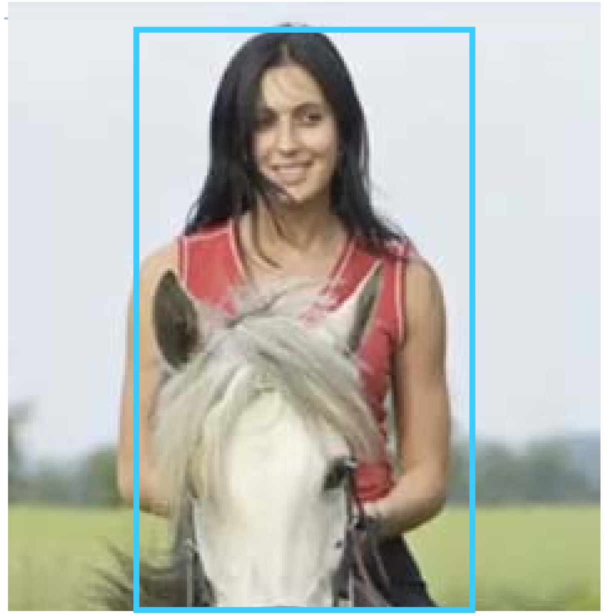
$\vec{\mathbf{x}} = [x_1\ x_2\ x_3\ x_4\ \ldots]$

- Sliding window
- Feature extraction
- Classification

For each window size
    For each window
        1. Feature extraction
        2. Classification
    End;
End;

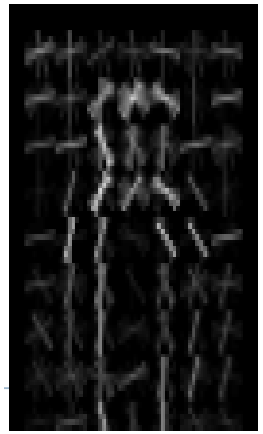
Object detection – the framework

- Sliding window
- Feature extraction
- **Classification**



Sliding window

Feature exaction



Feature vector:

$\vec{\mathbf{x}} = [x_1\ x_2\ x_3\ x_4\ \ldots]$

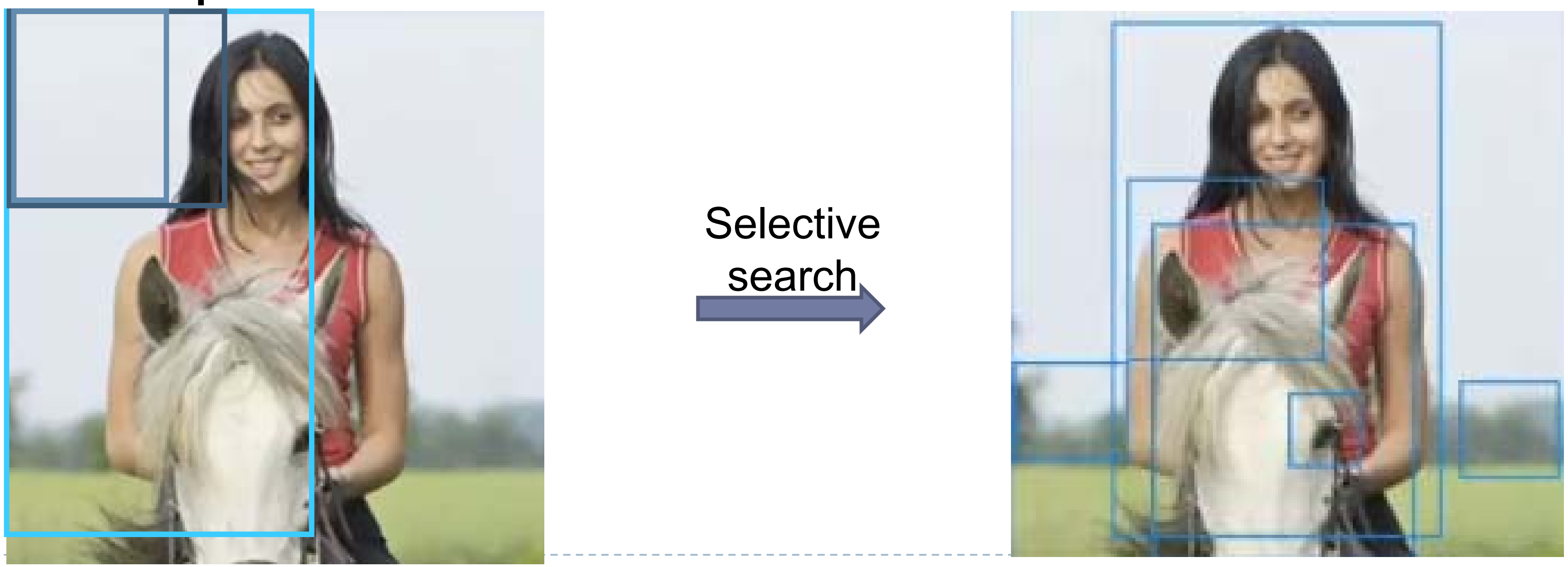
Classification

Object or not?

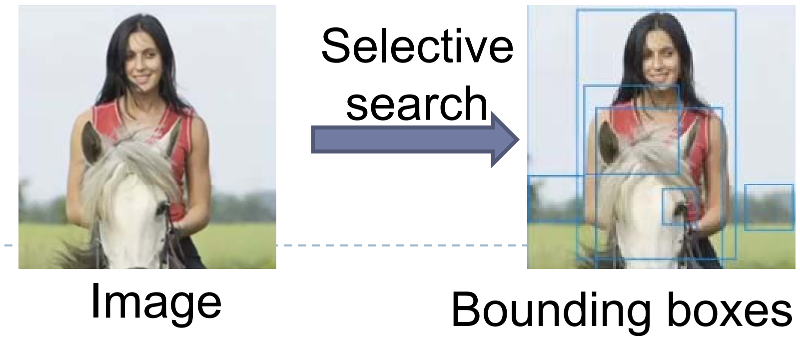
For each window size
  For each window
    1. Feature extraction
    2. Classification
  End;
End;

Problem of sliding windows

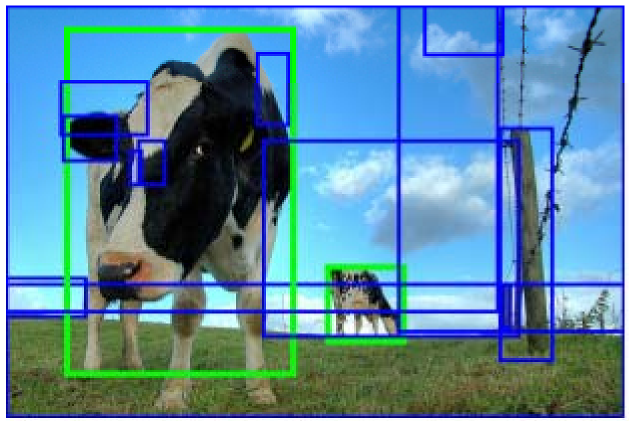
- Single-scale detection: 10k to 100k windows per image
- Multi-scale detection: 100k to 1m windows per image
- Multiple aspect ratio: 10m to 100m windows per image
- Selective search: 2k windows per image of multiple scales and aspect ratios



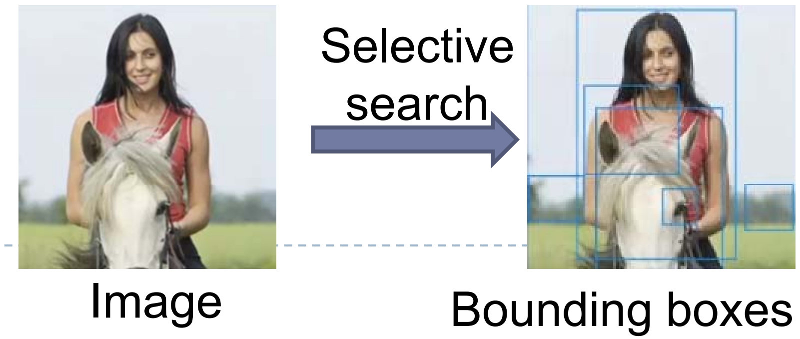
Selective search



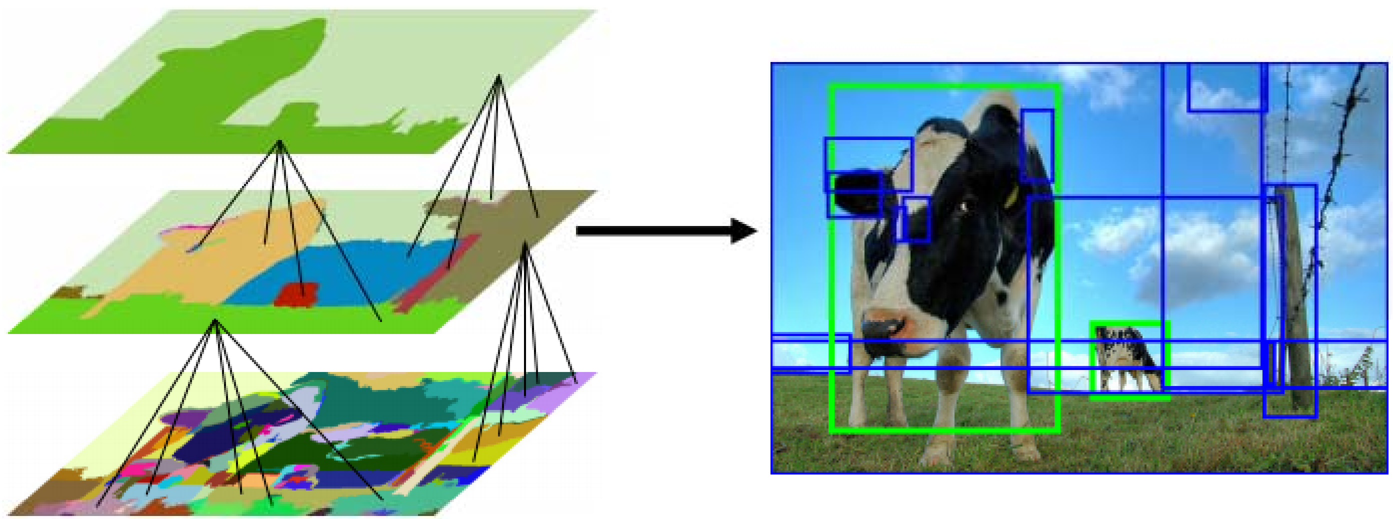
- Initial segments from over-segmentation [Felzenszwalb2004]



Selective search



- Initial segments from over-segmentation [Felzenszwalb2004]
- Based on hierarchical grouping
- Group adjacent regions on region-level similarity
- Consider all scales of the hierarchy

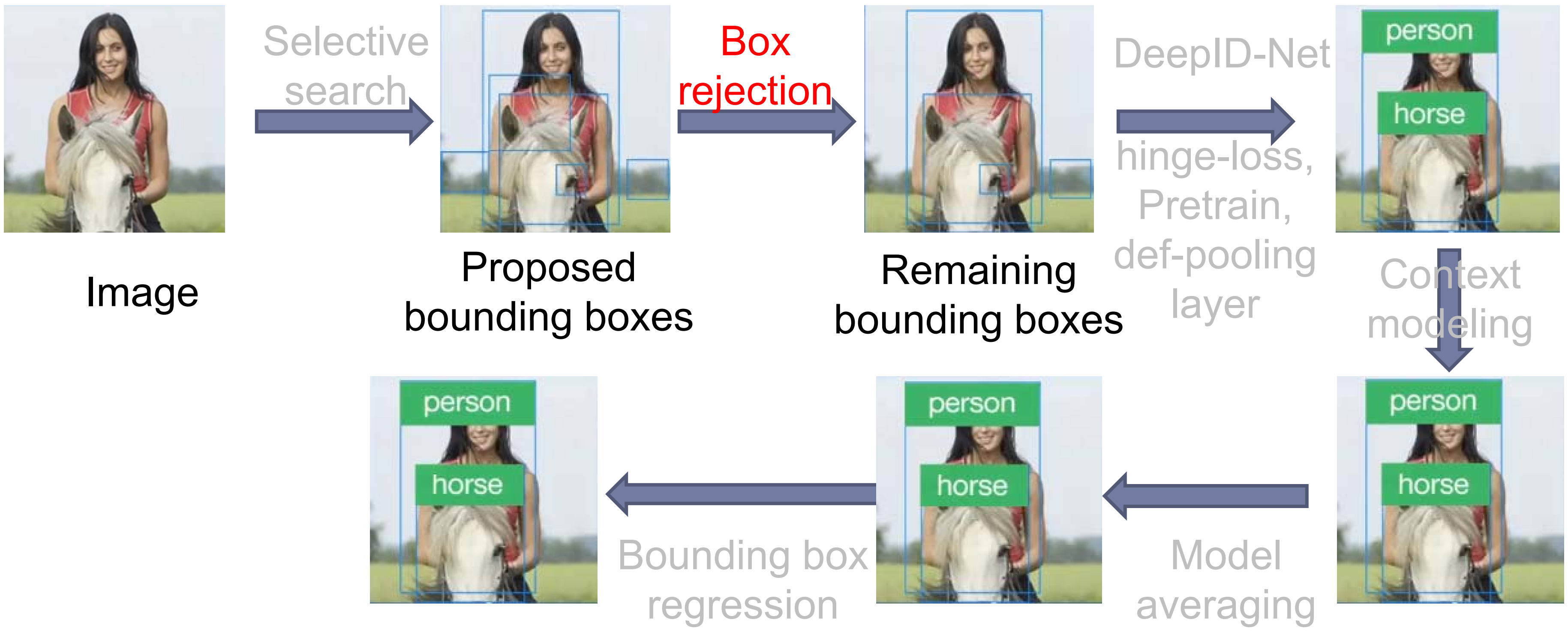


Our investigation

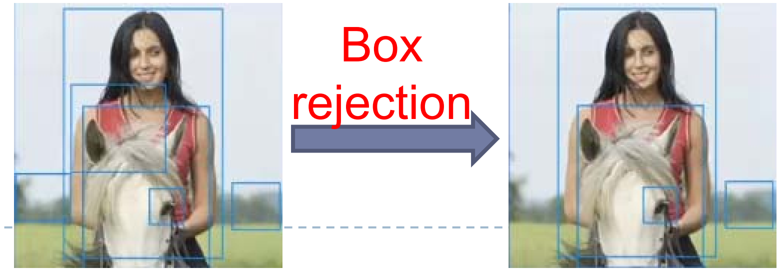
- Speed-up the pipeline
- Effectively learn the deep model
- Make use of domain knowledge from computer vision
  - Deformation pooling
  - Context modelling

mAP 31 → to 50.57 on val2

Our approach



Bounding box rejection

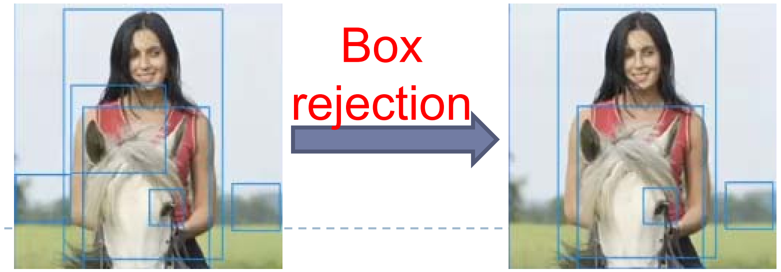


- Motivation
  - Selective search: ~ 2400 bounding boxes per image
  - Feature extraction using AlexNet
    - ILSVRC val: ~20,000 images, ~2.4 days
    - ILSVRC test: ~40,000 images, ~4.7days
- Bounding box rejection by RCNN:
  - For each box, RCNN has 200 scores $S_{1\ldots200}$ for 200 classes
  - If $\max(S_{1\ldots200}) < -1.1$, reject. 6% remaining bounding boxes

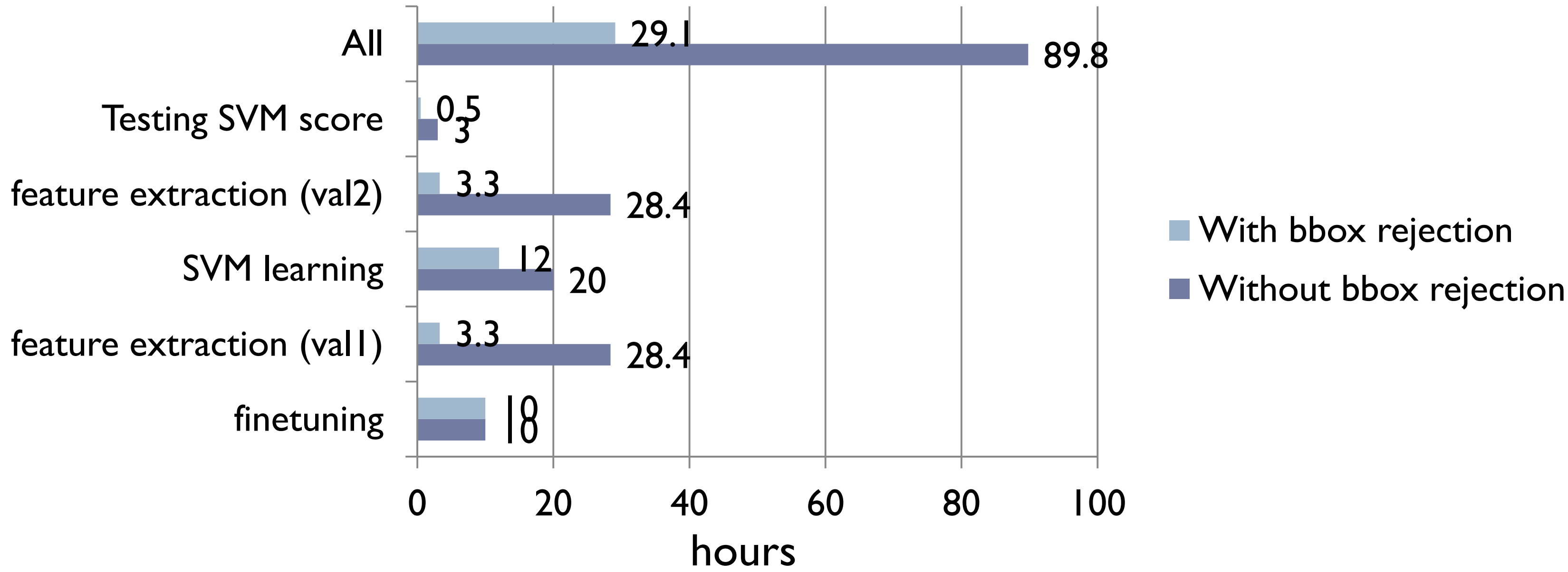
Remaining window	100%	20%	6%
Recall (val_1)	92.2%	89.0%	84.4%
Feature extraction time (seconds per image)	10.24	2.88	1.18

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *CVPR*, 2014

Bounding box rejection



- Speed up the pipeline
  - Save the feature extraction time by about 10 times.
- Improve mean AP by 1%

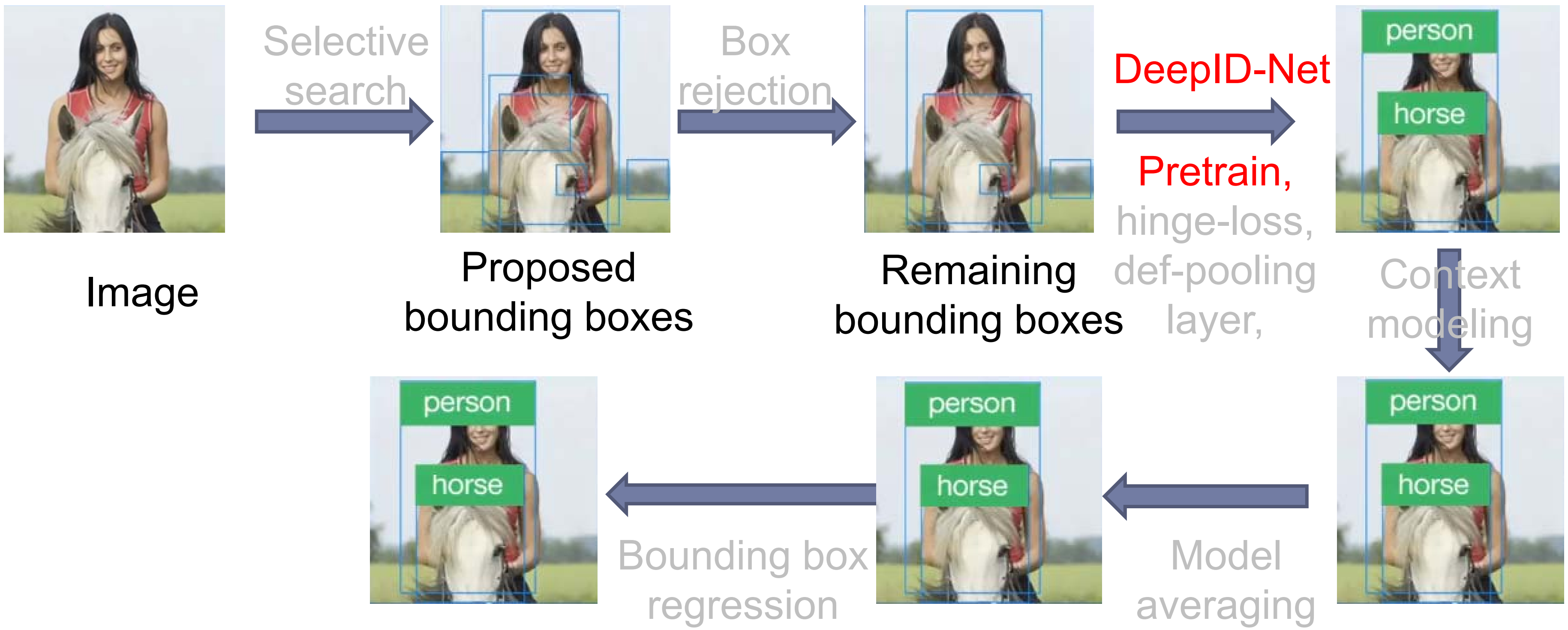


Remaining window	100%	20%	6%
Recall (val_1)	92.2%	89.0%	84.4%
Feature extraction time (seconds per image)	10.24	2.88	1.18

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *CVPR*, 2014

mAP 31 → to 50.57

Our pipeline

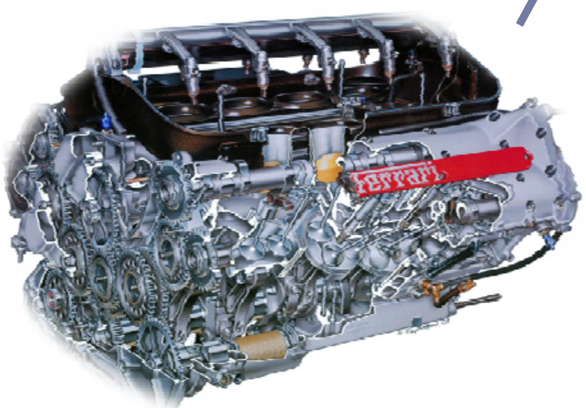


Deep learning is feature learning

Image classification

Object detection

Tracking

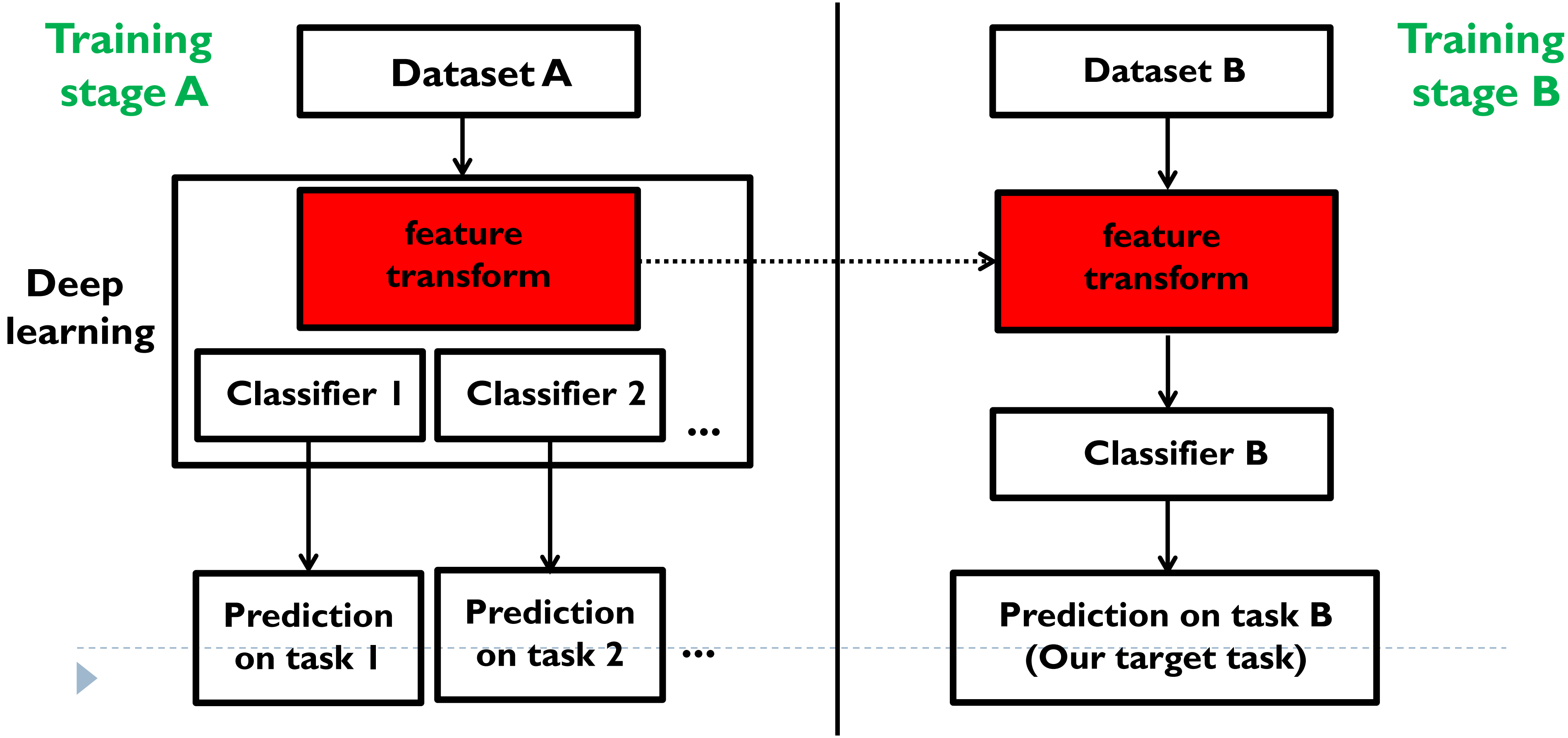


Segmentation

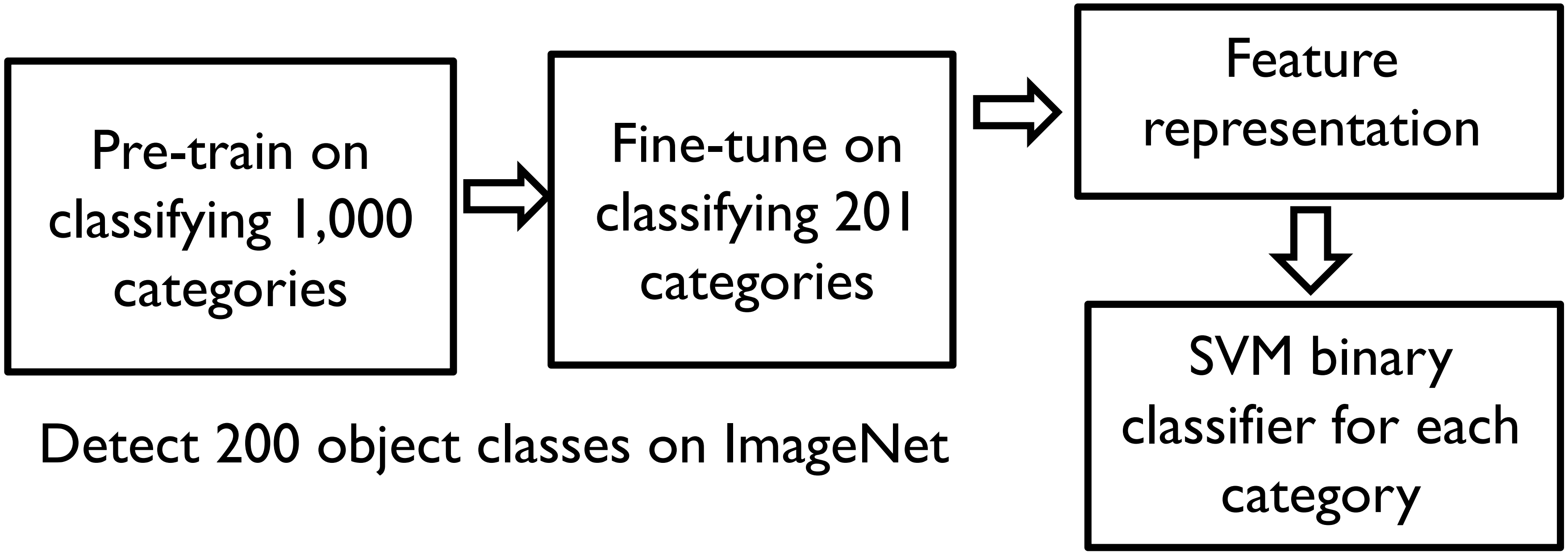
- **Features learned on ImageNet**

Learning features and classifiers separately

- How to effectively learn features?
  - With challenging tasks
  - Predict high-dimensional vectors



Directly training 200 binary classifiers with CNNs are not good



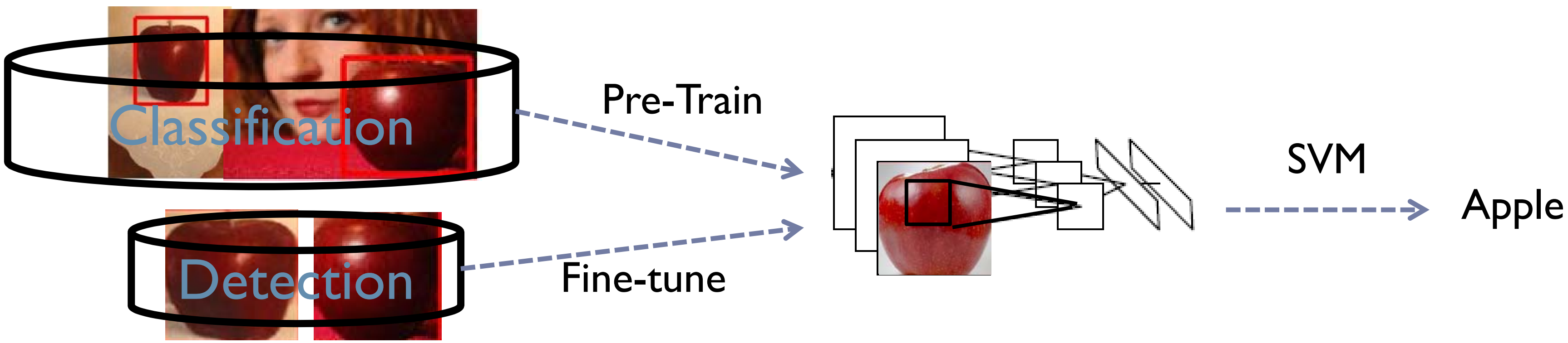
Girshick, Ross, et al. *CVPR*, 2014

Why need pre-training with many classes?

- Each sample carries much more information
- One big negative class with many types of objects confuses CNN on feature learning
- Make the training task challenging, not easy to overfit

Feature learning

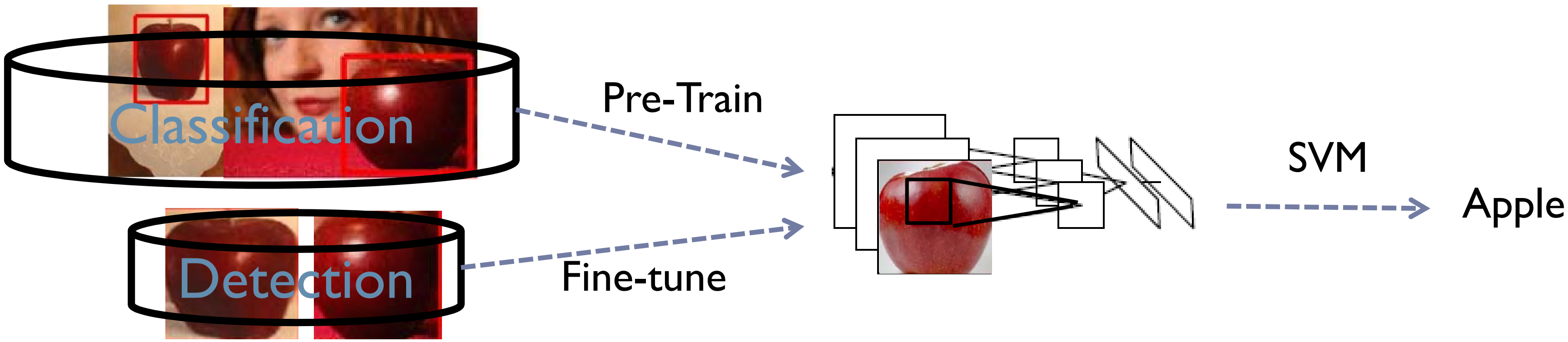
- Pretrain for *image-classification* with 1000 classes
- Finetune for *object-detection* with 200+1 classes
  - Transfer the representation learned from ILSVRC Classification to PASCAL (or ImageNet) detection
- Use the fine-tuned features for learning SVM



Girshick, Ross, et al. *CVPR*, 2014

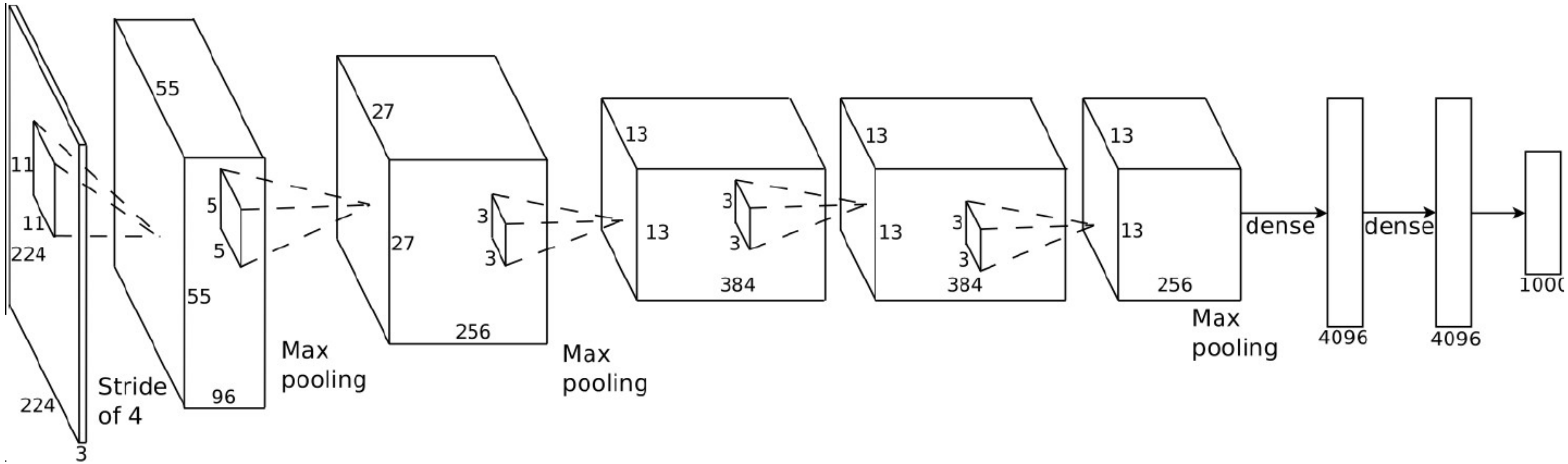
Feature learning

- Pretrain for *image-classification* with 1000 classes
- Finetune for *object-detection* with 200+1 classes
- Use the fine-tuned features for learning SVM
- Existing approaches mainly investigate on network structure
  - Number of layers/channels, filter size, dropout



Deep model design

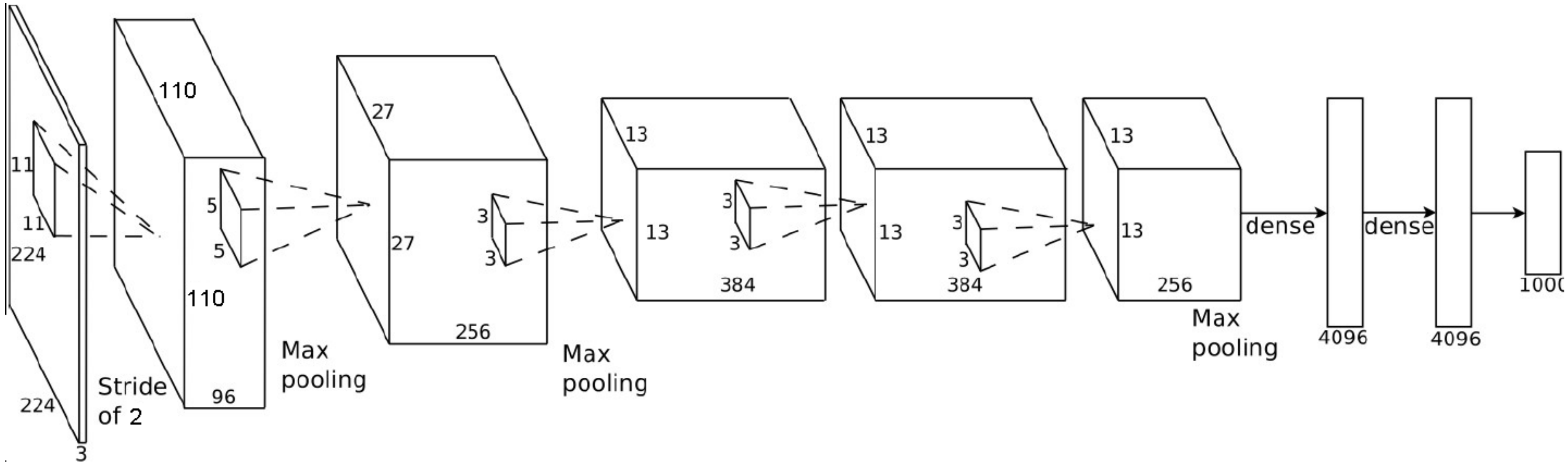
- Network structure



Net structure	AlexNet	AlexNet
Annotation level	Image	Image
Bbox rejection	n	y
mAP (%)	29.9	30.9

Deep model design

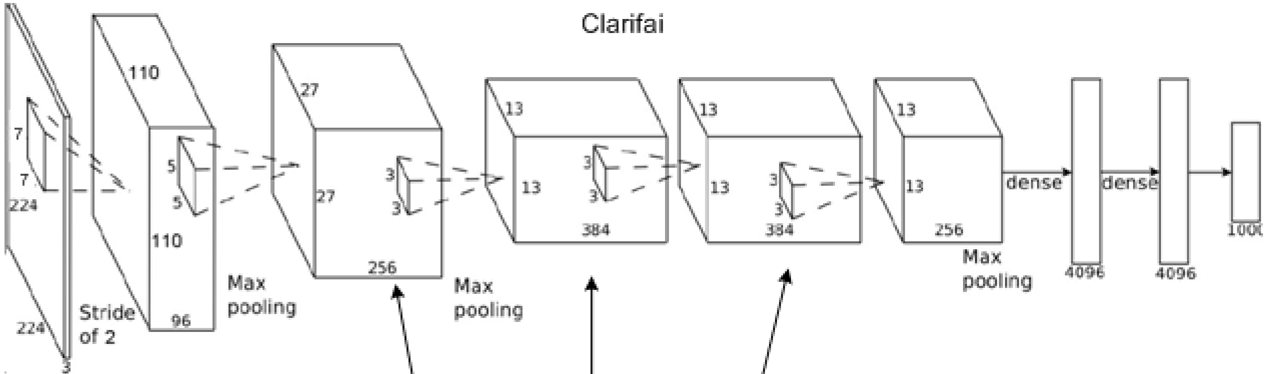
- Network structure



Net structure	AlexNet	AlexNet	Clarifai
Annotation level	Image	Image	Image
Bbox rejection	n	y	y
mAP (%)	29.9	30.9	31.8

Deep model design

- Network structure



Layer	1	2	3	4	5	6	7	8	Output 9
Stage	conv + max	conv + max	conv	conv	conv	conv + max	full	full	full
# channels	96	256	512	512	1024	1024	4096	4096	1000
Filter size	7x7	7x7	3x3	3x3	3x3	3x3	-	-	-
Conv. stride	2x2	1x1	1x1	1x1	1x1	1x1	-	-	-
Pooling size	3x3	2x2	-	-	-	3x3	-	-	-
Pooling stride	3x3	2x2	-	-	-	3x3	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	221x221	36x36	15x15	15x15	15x15	15x15	5x5	1x1	1x1

Net structure	AlexNet	AlexNet	Clarifai	Overfeat
Annotation level	Image	Image	Image	Image
Bbox rejection	n	y	y	y
mAP (%)	29.9	30.9	31.8	36.6

Deep model design

- Network structure

Net structure	AlexNet	AlexNet	Clarifai	Overfeat	GoogleNet
Annotation level	Image	Image	Image	Image	Image
Bbox rejection	n	y	y	y	y
mAP (%)	29.9	30.9	31.8	36.6	37.8

Feature learning – pretrain

- Classification
  - Pretrain for *image-classification* with 1000 classes
  - Finetune for *object detection* with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200

Image classification

Object detection

Feature learning – pretrain

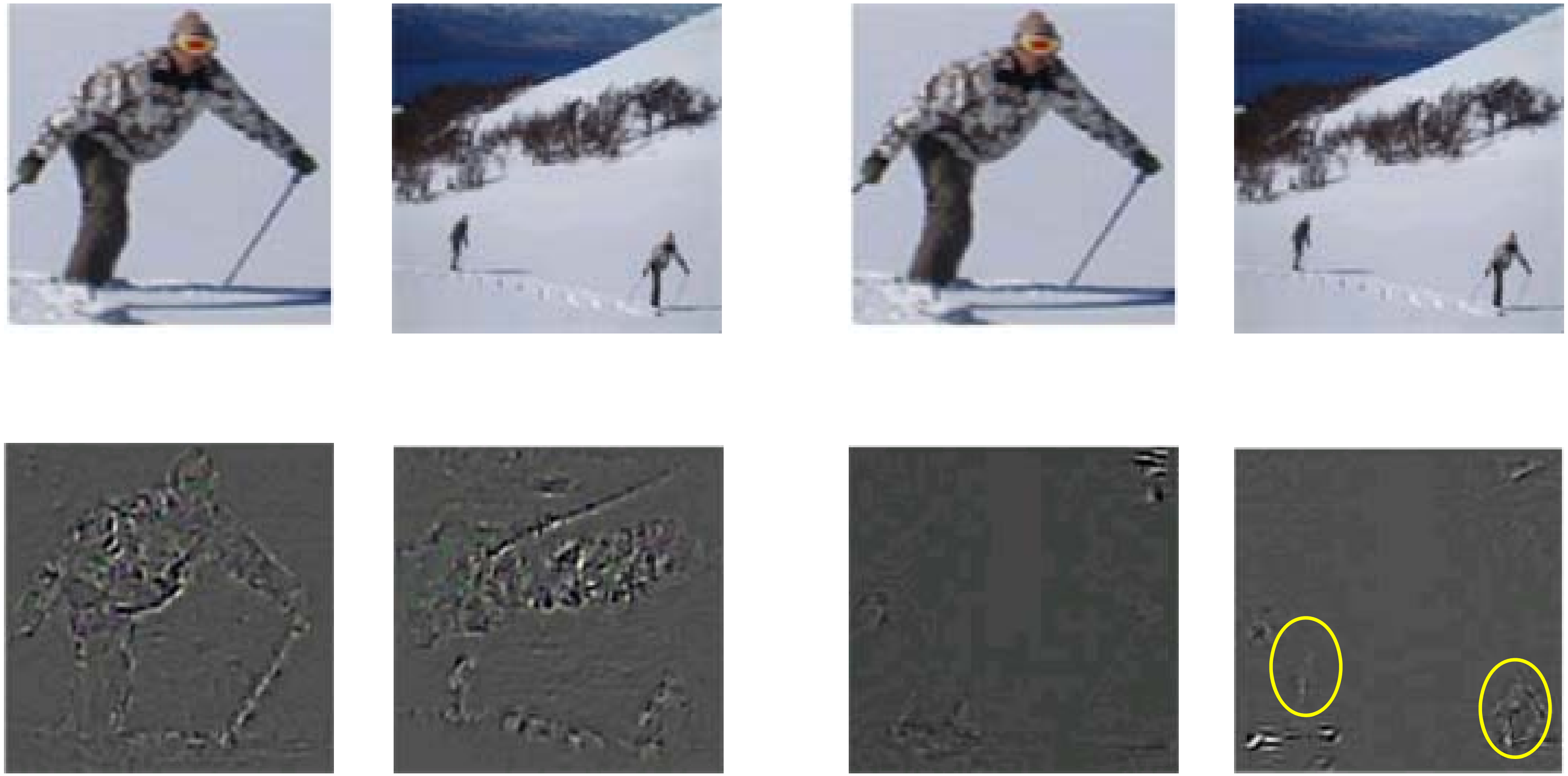
- Classification
  - Pretrain for *image-classification* with 1000 classes
  - Finetune for *object detection* with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200

Image classification

Object detection

Feature learning – pretrain

- Classification



Pretrained on object-level annoation

Pretrained on image-level annotation

Feature learning – pretrain

- Classification (Cls)
  - Pretrain for *image-classification* with 1000 classes
  - Gap: classification vs. detection, 1000 vs. 200
- Detection (Loc)
  - Pretrain for *object-detection* with 1000 classes

Pretraining scheme	Cls	Cls	Loc
Net structure	AlexNet	Clarifai	Clarifai
mAP (%) on val2	29.9	31.8	36.0

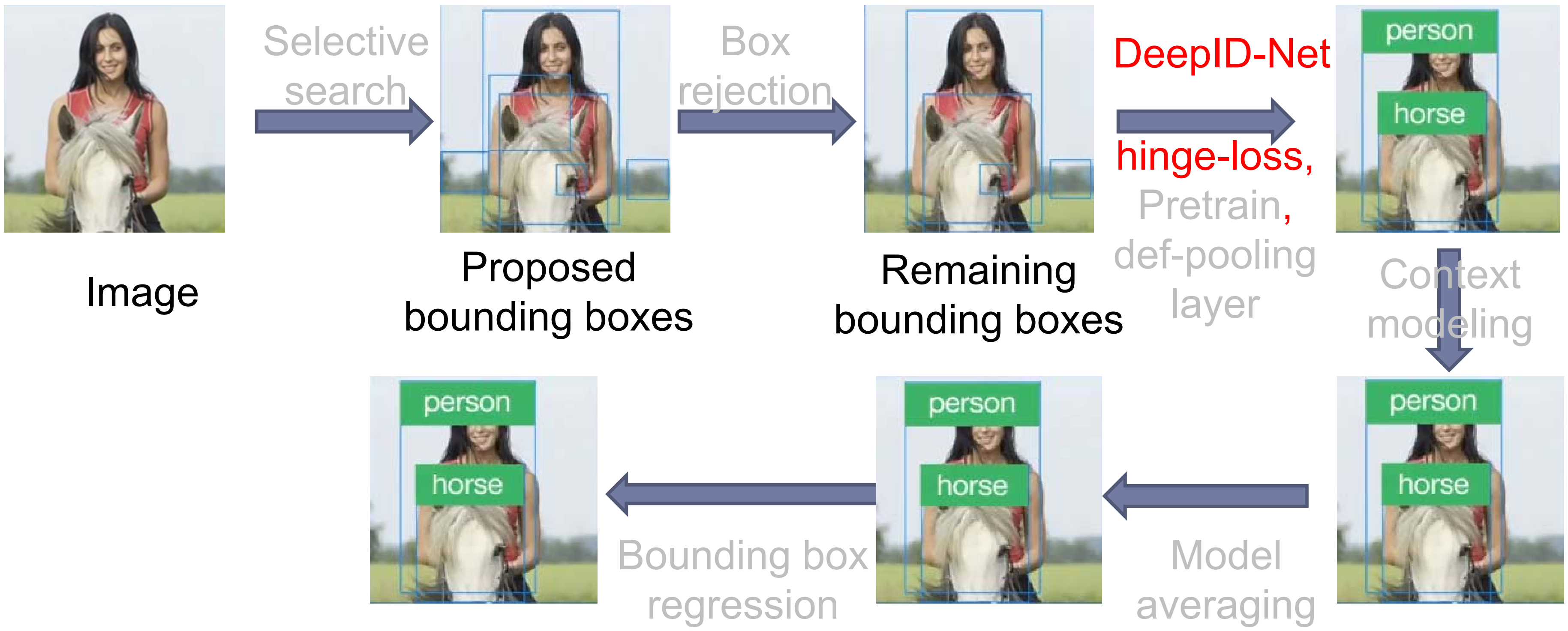
Result and discussion

- RCNN (Cls+Det),
- Our investigation
  - Better pretraining on 1000 classes
  - Object-level annotation is more suitable for pretraining

AlexNet	Image annotation	Object annotation
200 classes (Det)	20.7	32
1000 classes (Cls-Loc)	31.8	36

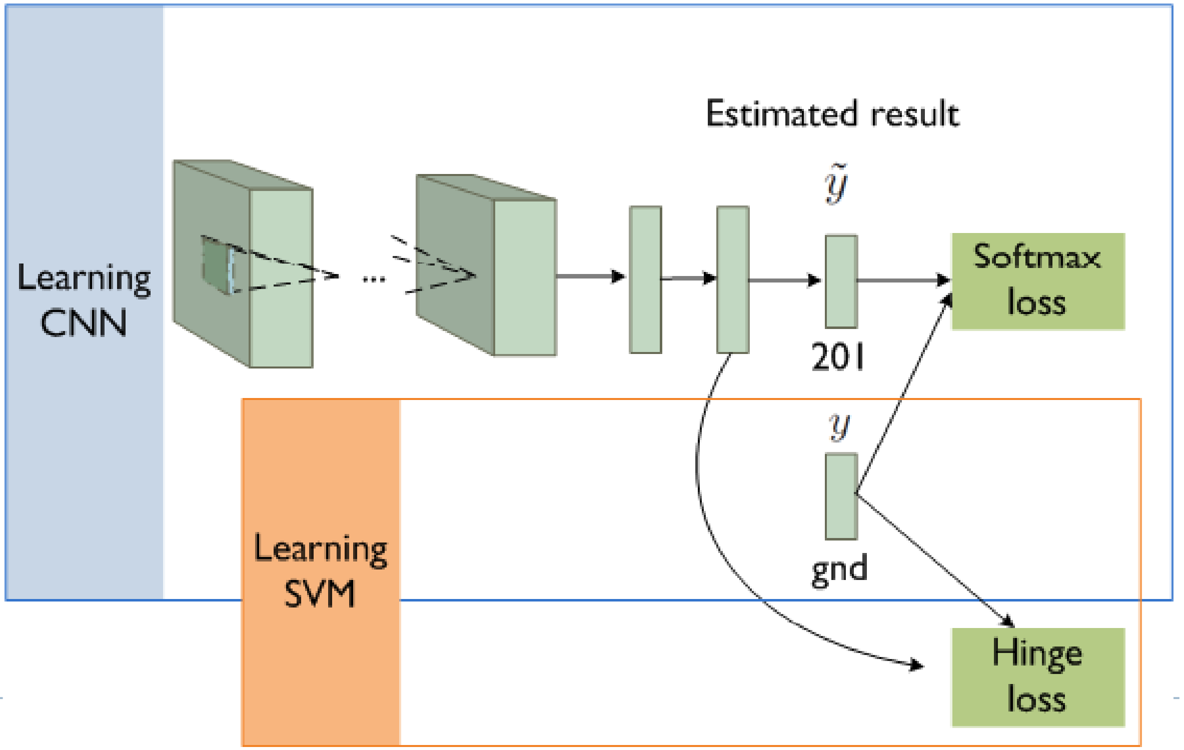
mAP 31 → to 50.57 on val2

Our approach



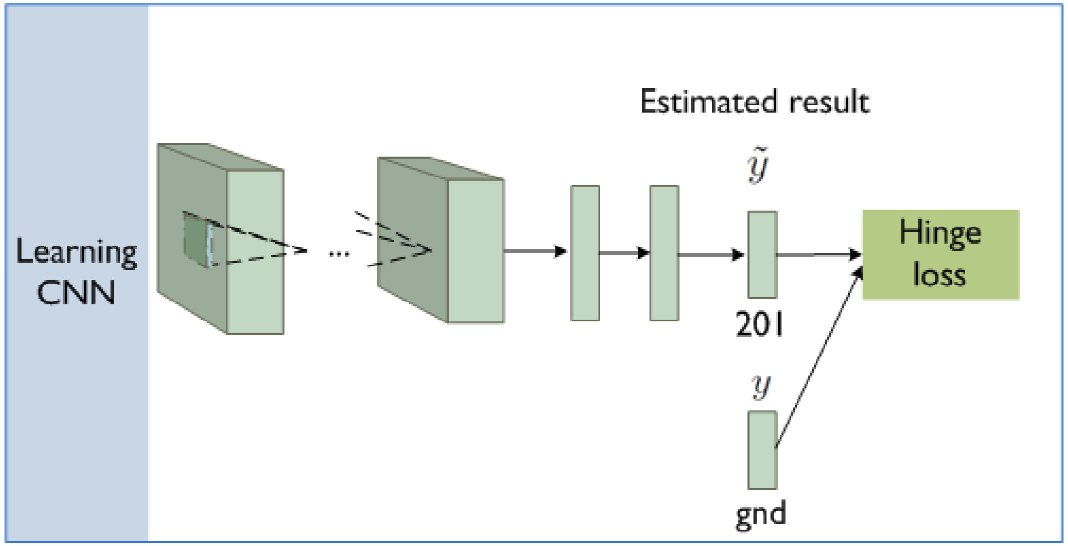
Feature learning – SVM-net

- Existing approach
  - Learn features using soft-max loss (Softmax-Net)
  - Train SVM with the learned features



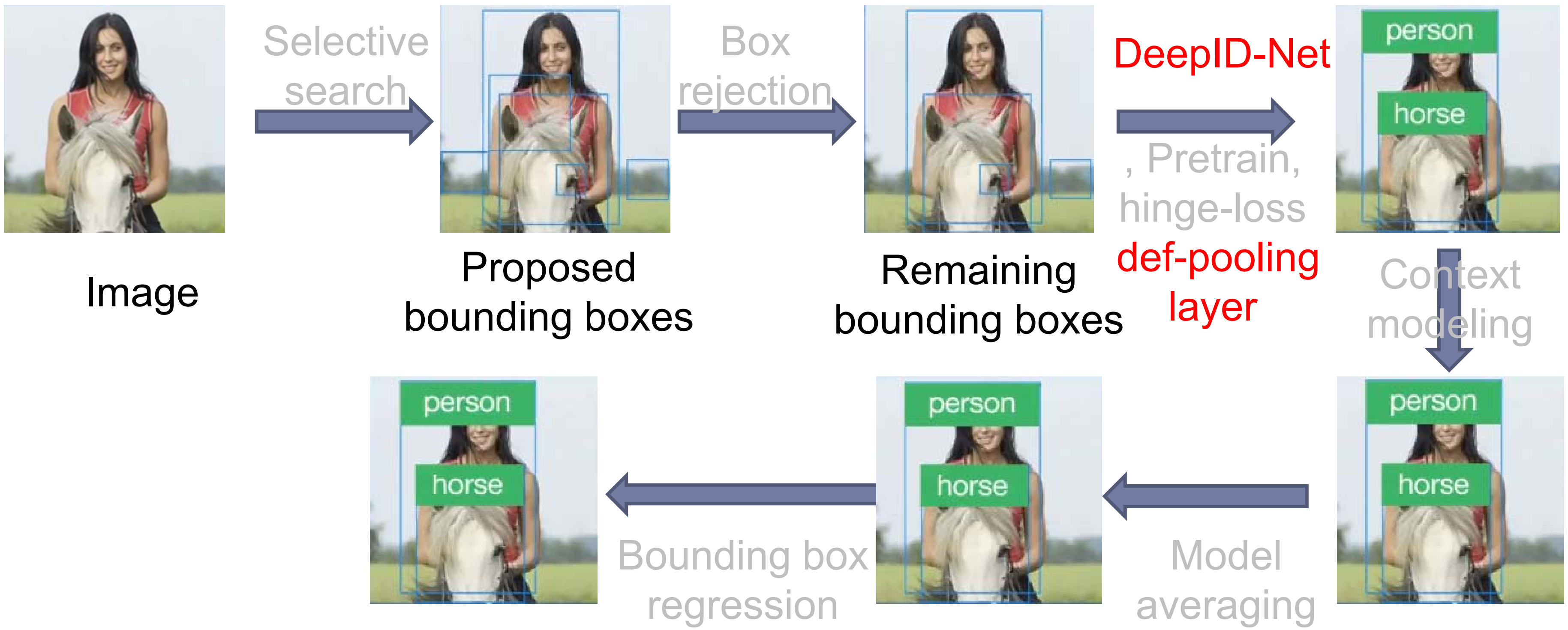
Feature learning – SVM-net

- Existing approach
  - Learn features using soft-max loss (Softmax-Net)
  - Train SVM with the learned features
- Replace Soft-max loss by Hinge loss when fine-tuning (SVM-Net)
  - Merge the two steps of RCNN into one
  - Require no feature extraction from training data (~60 hours)



mAP 31 → to 50.3

Our pipeline



Deep model training – def-pooling layer

- RCNN (ImageNet Cls+Det)
  - Pretrain on image-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes
  - Gap: classification vs. detection, 1000 vs. 200
- Our approach (ImageNet Loc+Det)
  - Pretrain on object-level annotation with 1000 classes
  - Finetune on object-level annotation with 200 classes with def-pooling layers

Net structure	Without Def Layer	With Def layer
mAP (%) on val2	36.0	38.5

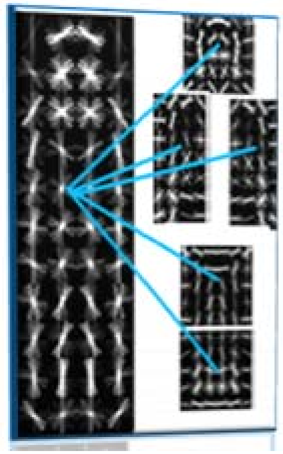
Deformation

- Learning deformation [a] is effective in computer vision society.
- Missing in deep model.
- We propose a new deformation constrained pooling layer.

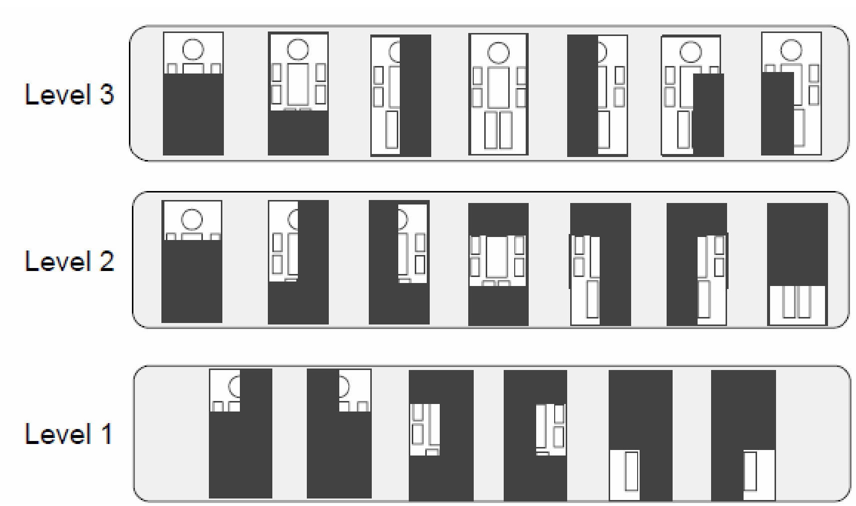
[a] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Trans. PAMI, 32:1627–1645, 2010.

Modeling Part Detectors

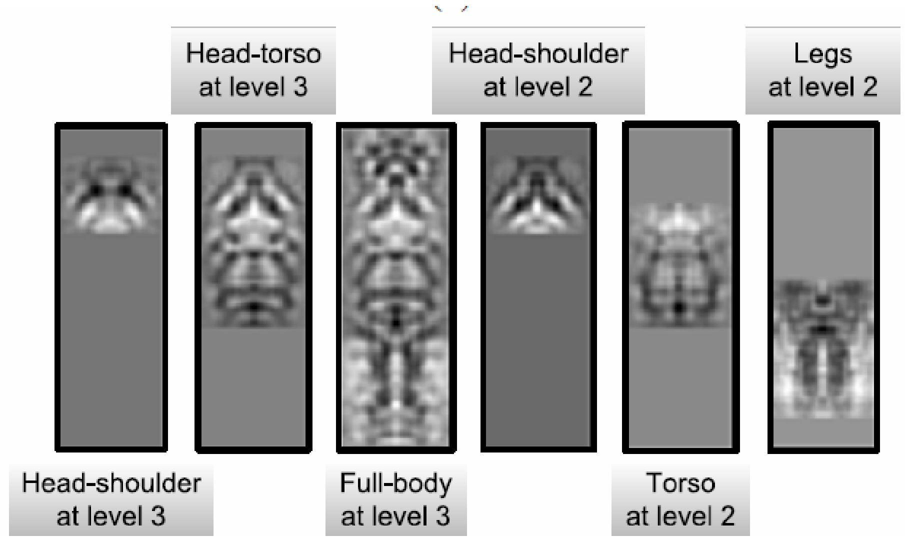
- Different parts have different sizes
- Design the filters with variable sizes



Part models learned from HOG



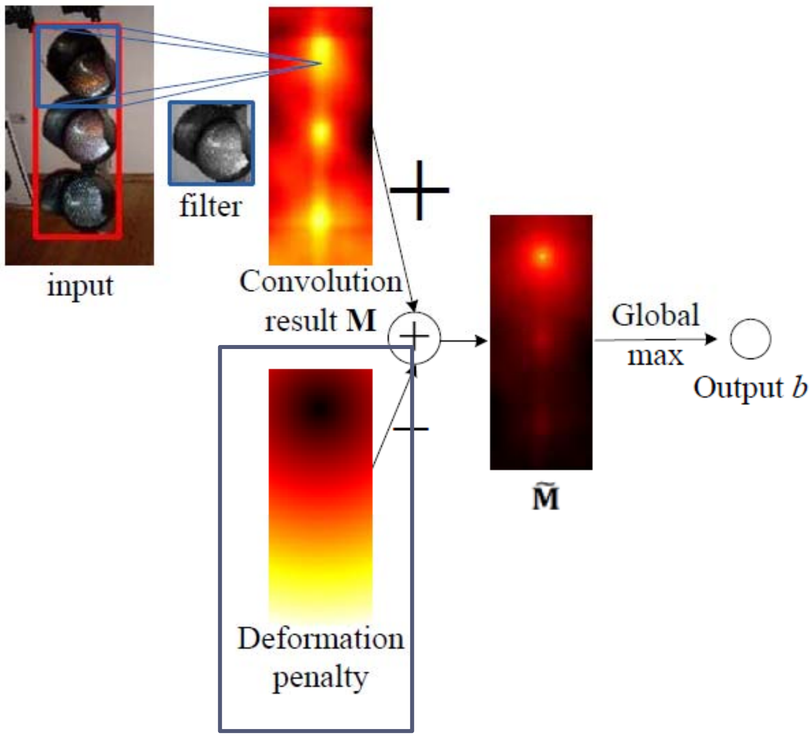
Part models



Learned filtered at the second convolutional layer

Deformation Layer [b]

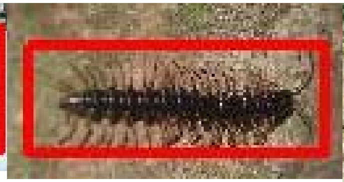
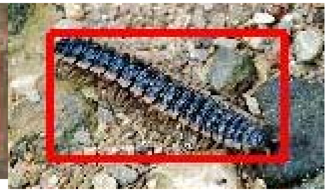
$$\mathbf{B}_p = \mathbf{M}_p + \boxed{\sum_{n=1}^{N} c_{n,p} \mathbf{D}_{n,p}} \qquad s_p = \max_{(x,y)} b_p^{(x,y)}$$



[b] Wanli Ouyang, Xiaogang Wang, "Joint Deep Learning for Pedestrian Detection ", ICCV 2013.

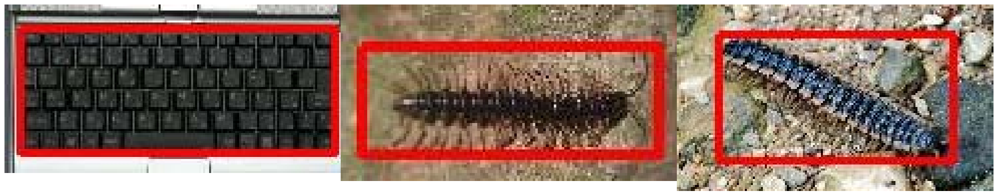
Deformation layer for repeated patterns

Pedestrian detection	General object detection
Assume no repeated pattern	Repeated patterns


Deformation layer for repeated patterns

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



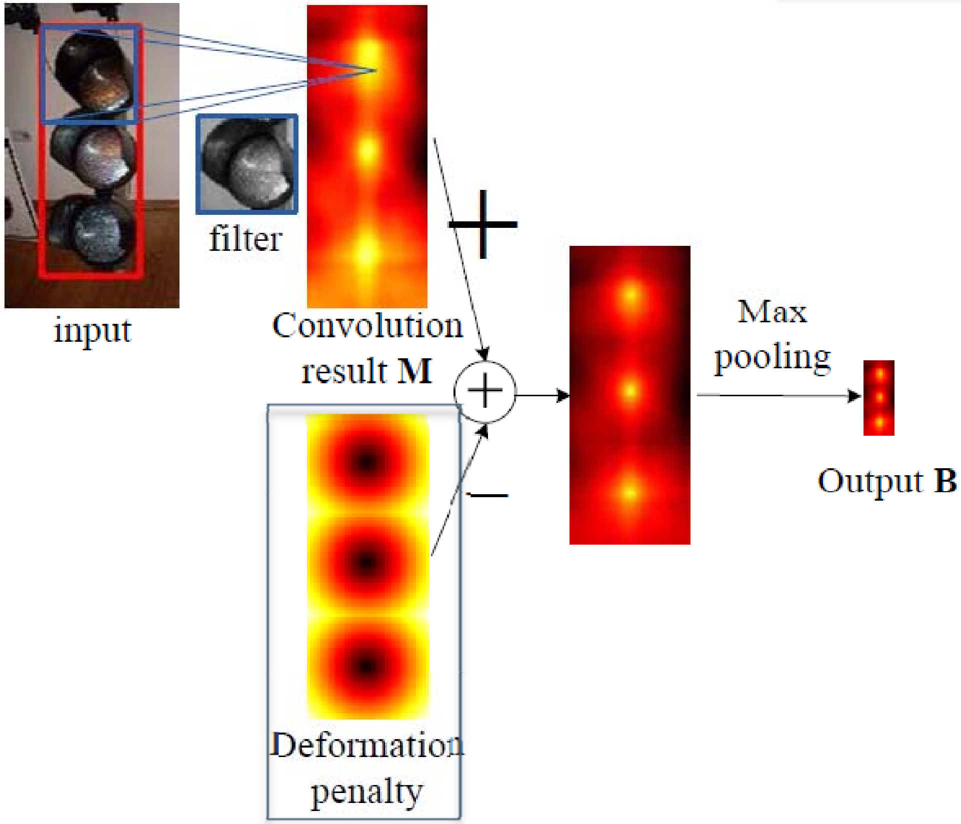


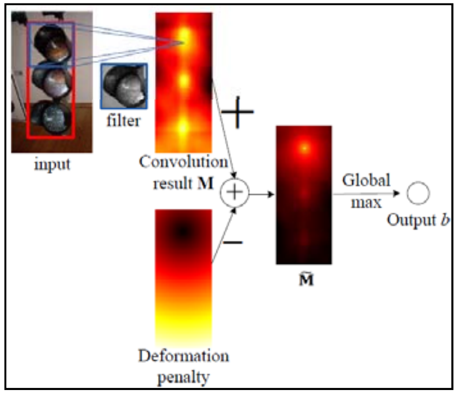



Deformation constrained pooling layer

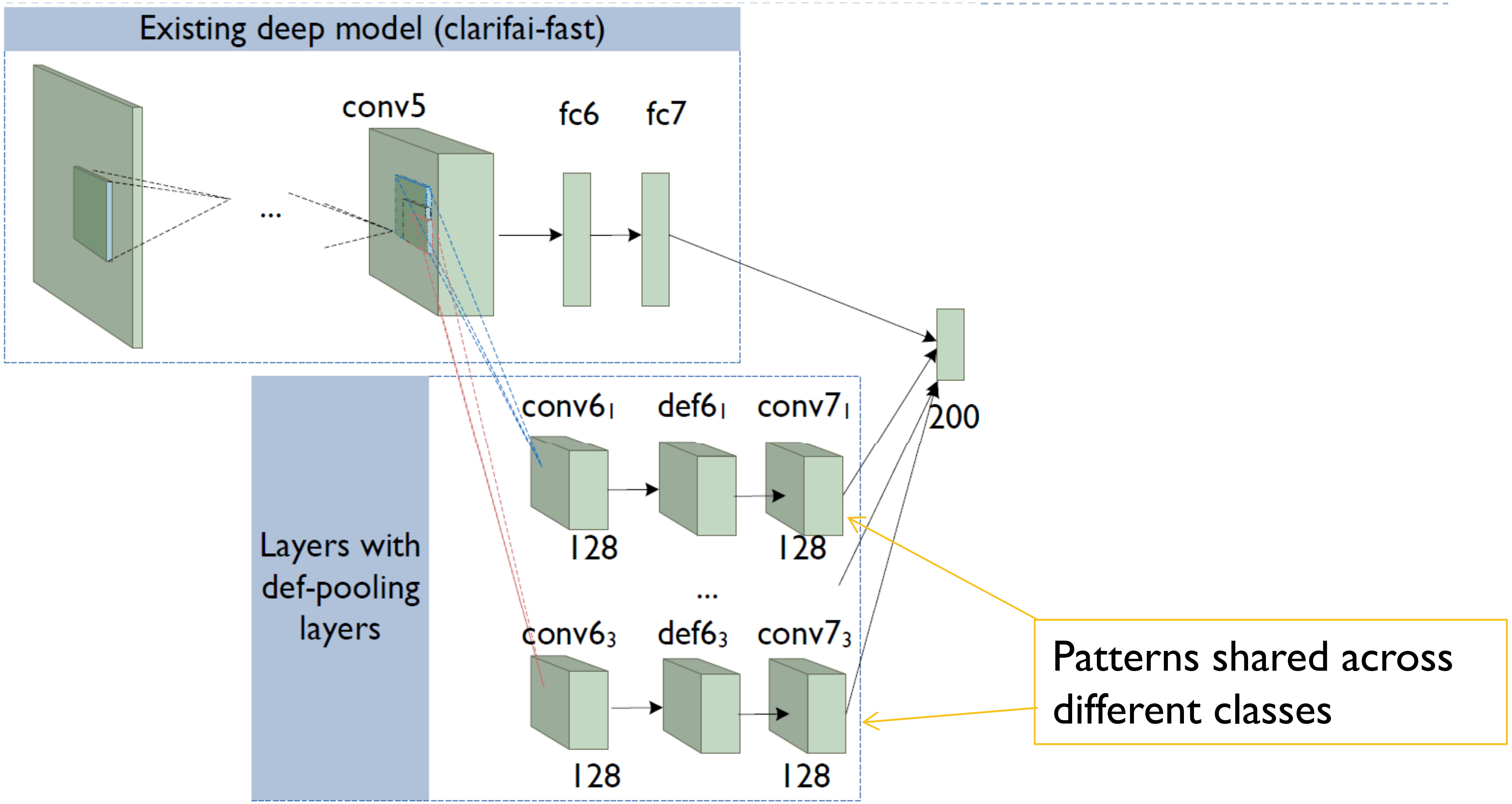
Can capture multiple patterns simultaneously

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





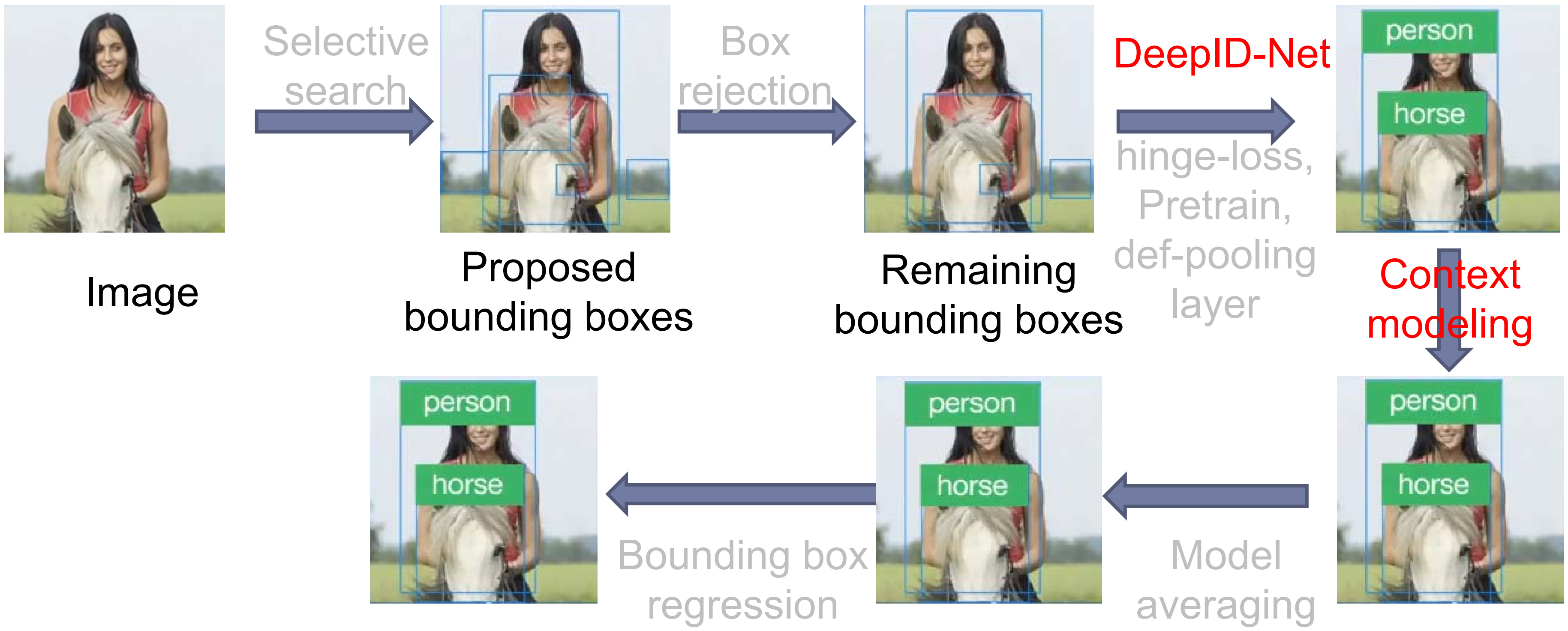
Our deep model with deformation layer



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

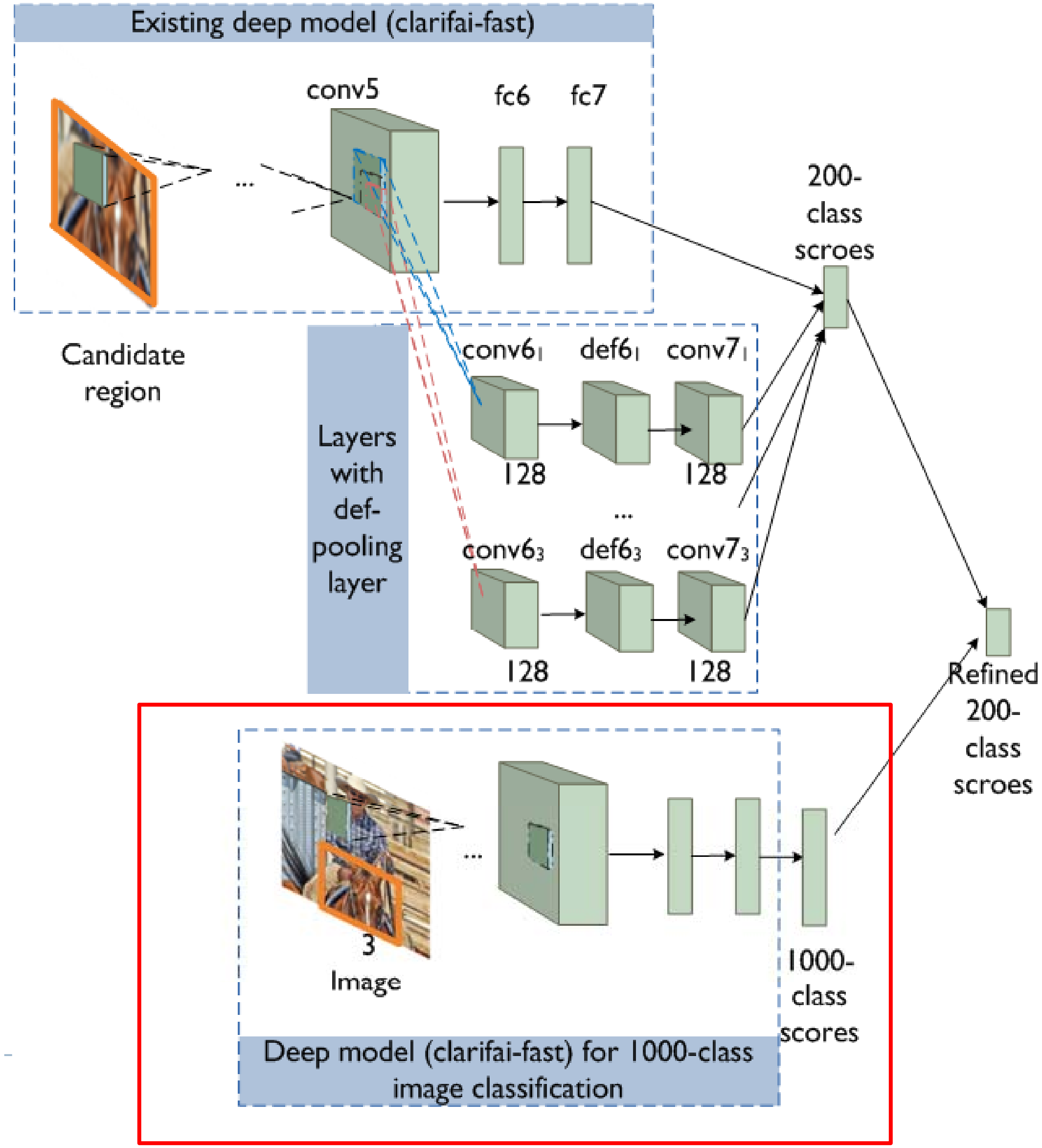
mAP 31 → to 50.57 on val2

Our approach



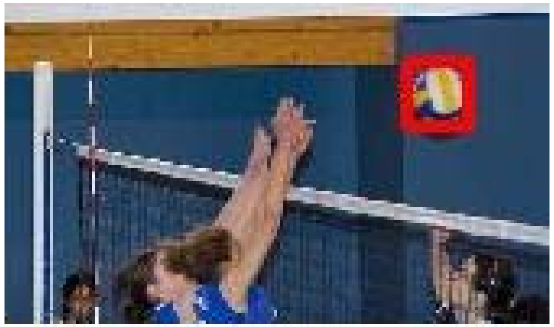
Context modeling

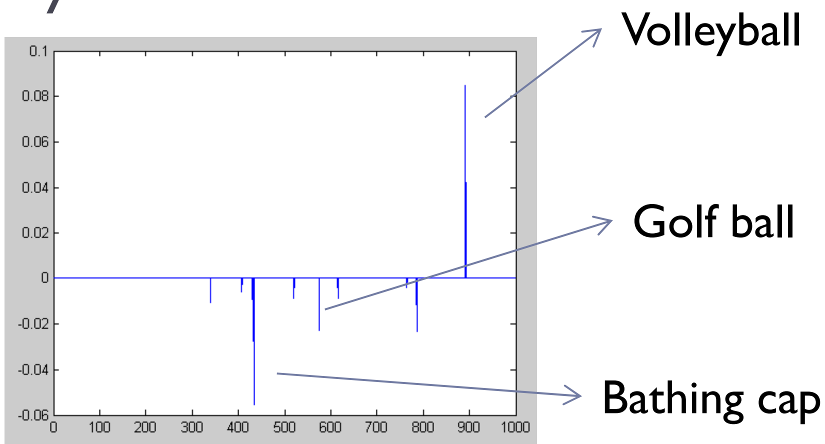
- Use the 1000 class Image classification score.
- ~1% mAP improvement.

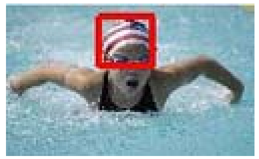


Context modeling

- Use the 1000-class Image classification score.
  - ~1% mAP improvement.
  - Volleyball: improve ap by 8.4% on val2.

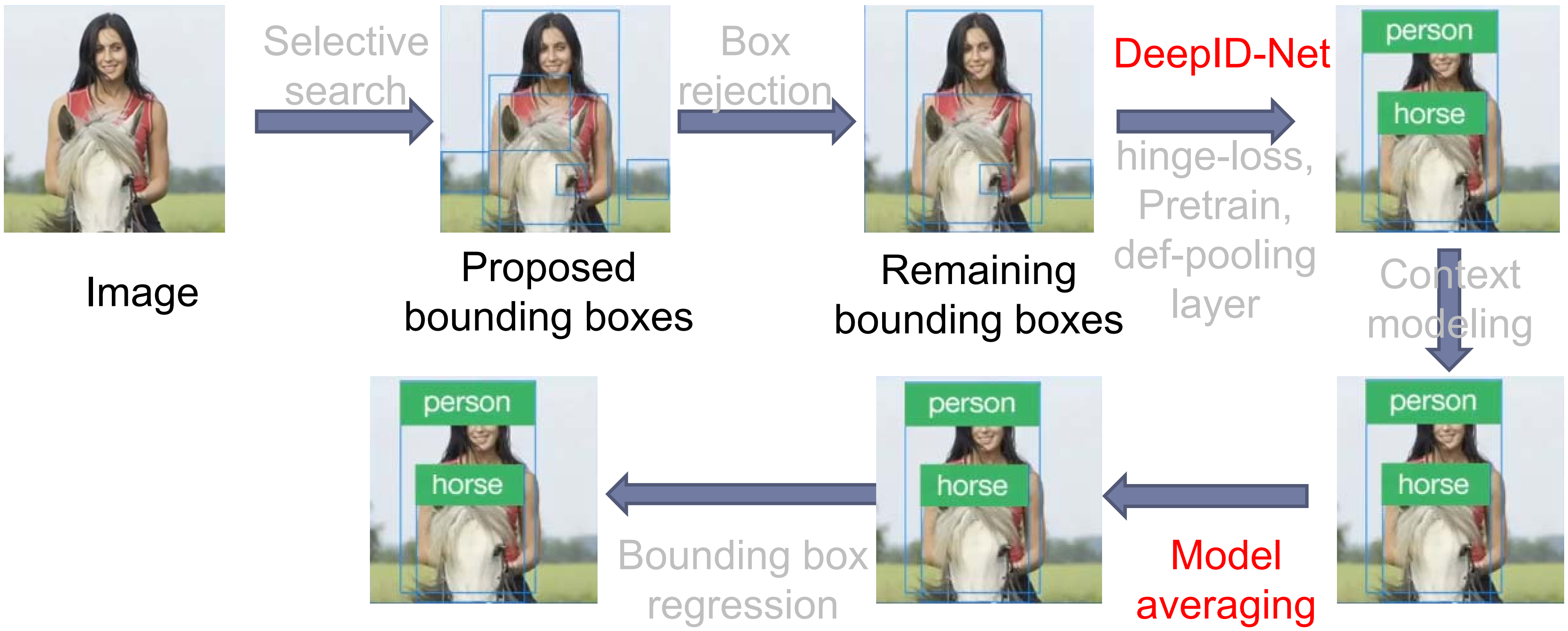






mAP 31 → to 50.57 on val2

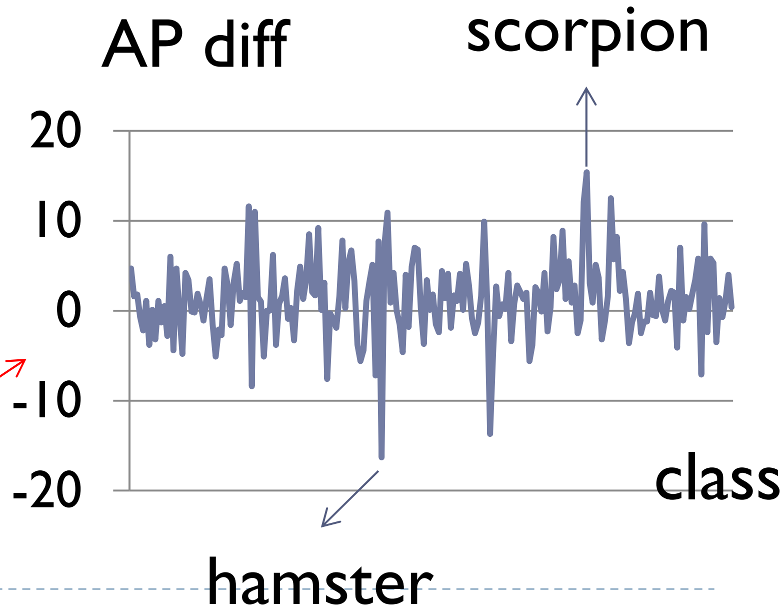
Our approach



Model averaging

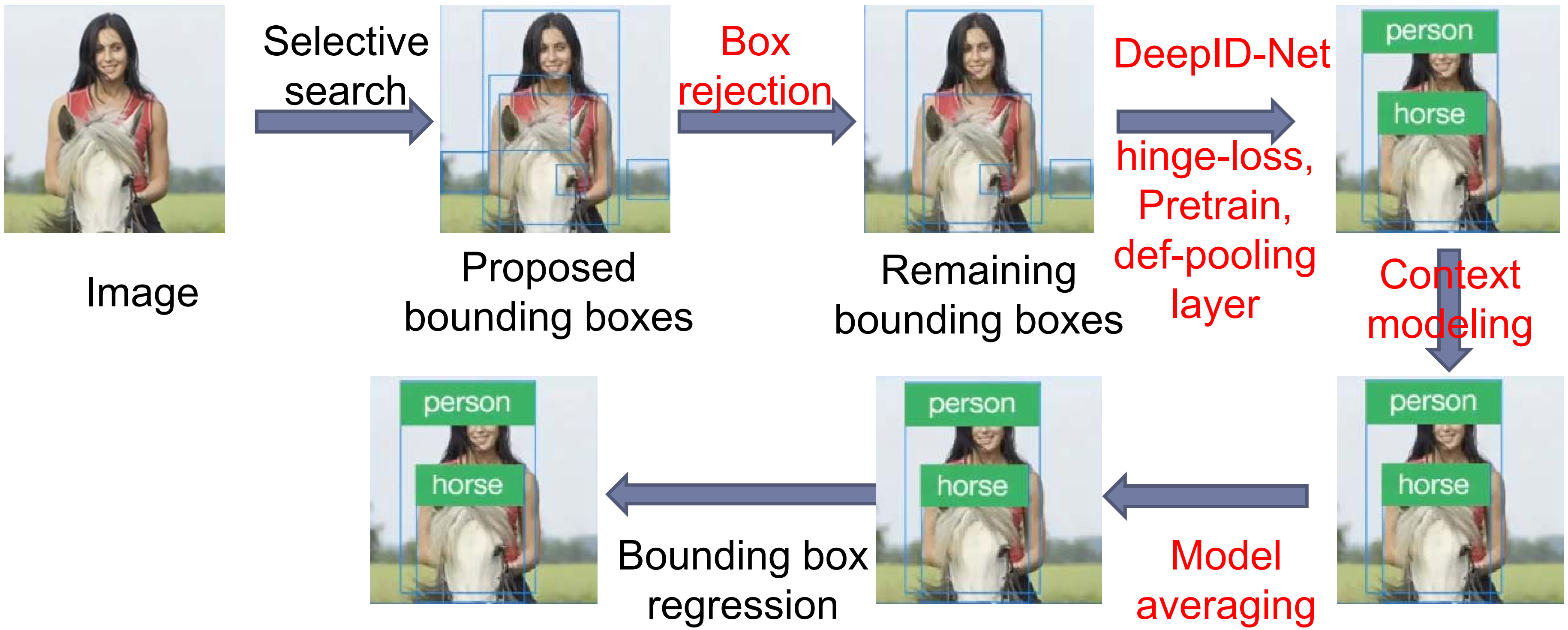
- Models of different structures are complementary on different classes.

Net structure	AlexNet	AlexNet	Clarifai
Annotation level	Image	Object	Object
Bbox rejection	n	n	n
mAP （%）	29.9	34.3	35.6



mAP 31 → to 50.57 on val2

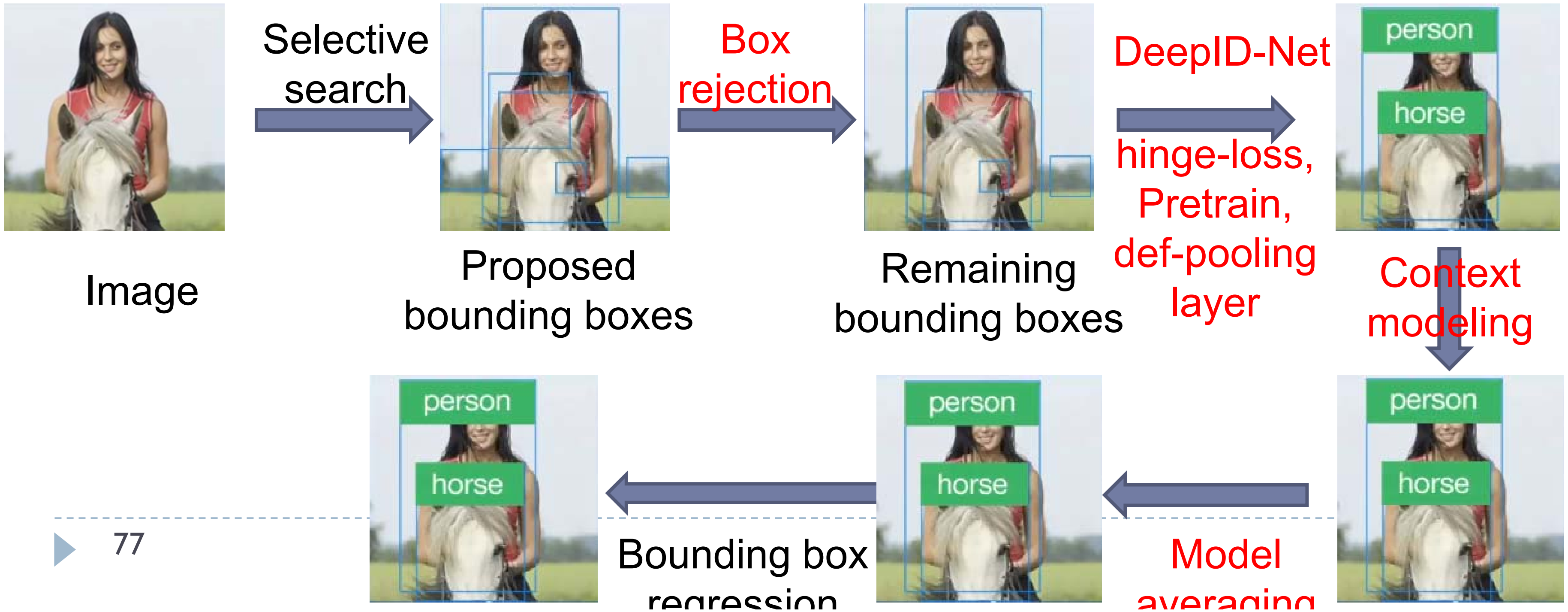
Our approach



Comparison with state-of-the-art

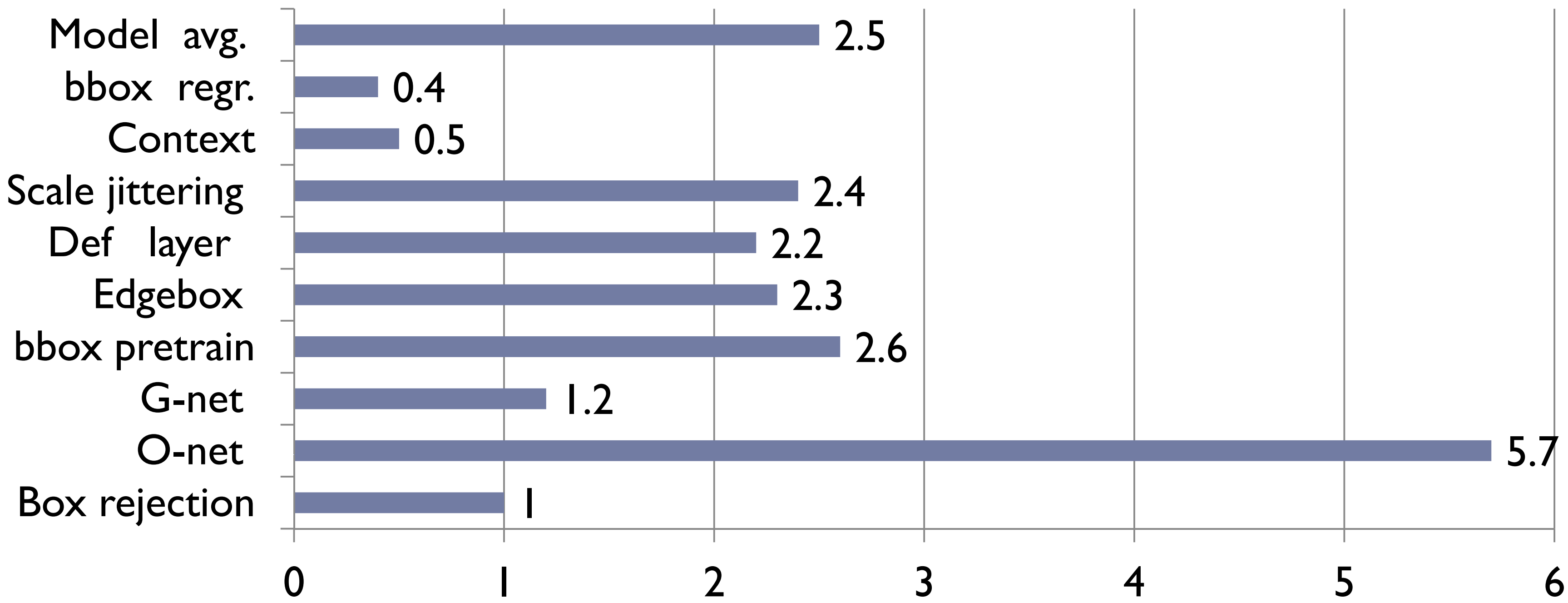
Detection Pipeline	Flair	RCNN	Berkeley Vision	UvA-Euvision	DeepInsight	GoogLeNet	Ours
mAP on val2 (avg)	n/a	n/a	n/a	n/a	42	44.5	50.7
mAP on val2 (sgl)	n/a	31.0	33.4	n/a	40.1	38.8	48.2
mAP on test (avg)	22.6	n/a	n/a	n/a	40.5	43.9	50.3
mAP on test (sgl)	n/a	31.4	34.5	35.4	40.2	38.0	47.9

Our approach



Component analysis

Detection Pipeline	RCNN	Box rejection	O-net	G-net	+bbox pretrain	+Edge box	+Def layer	Scale jittering	+ctx	+bbox regr.	Model avg.
mAP on val2	29.9	30.9	36.6	37.8	40.4	42.7	44.9	47.3	47.8	48.2	50.7
mAP on test										47.9	50.3



Summary

- Speed-up the pipeline:
  - Bounding rejection. Save feature extraction by about 10 times, slightly improve mAP (~1%).
  - Hinge loss. Save feature computation time (~60 h).
- Improve the accuracy
  - Pre-training with object-level annotation, more classes. 4.2% mAP
  - Def-pooling layer. 2.5% mAP
  - Context. 0.5-1% mAP
  - Model averaging. Different model designs and training schemes lead to high diversity

Conclusions

- Jointly optimize vision components (joint deep learning)
- Propose new layers based on domain knowledge (def-pooling layer)
- Carefully design the strategies of learning feature representations
  - Feature learned aided by semantic tasks
  - Pre-training with challenging tasks and rich predictions
  - The chosen training tasks help to achieved desired feature invariance and discriminative power
  - Adapted to specific tasks in test

Multimedia Laboratory

The Chinese University of Hong Kong

Wanli Ouyang

Xiaogang Wang

Xiaoou Tang

Chen Change Loy

Ping Luo

Hongsheng Li

Xingyu Zeng

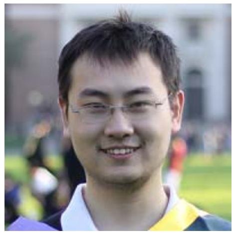
Shi Qiu

Yongkong Tian

Zhenyao Zhu

Zhe Wang

Shuo Yang



Chen Qian

Yuanjun Xiong

Ruohui Wang