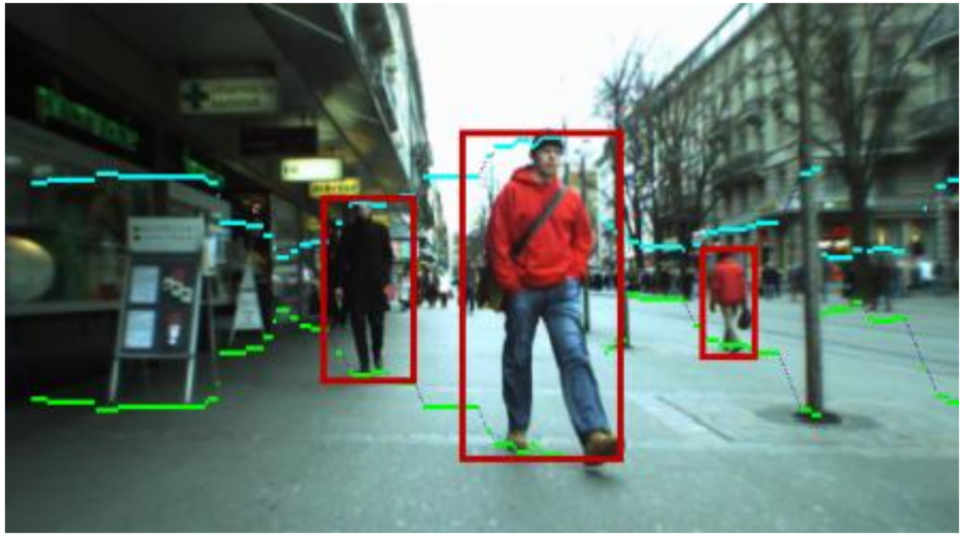
Part-Level Fully Convolutional Networks for Pedestrian Detection

Xinran Wang
School of Electronic Engineering
Xidian University, China

Pedestrian Detection

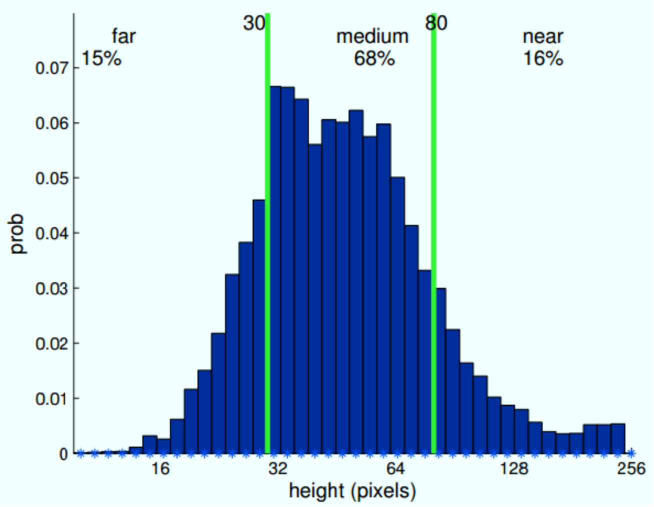
- Key problem for visual surveillance, automotive safety, and robotics applications
- Wide variety of appearances: Body pose, occlusions, clothing, lighting, and complex backgrounds



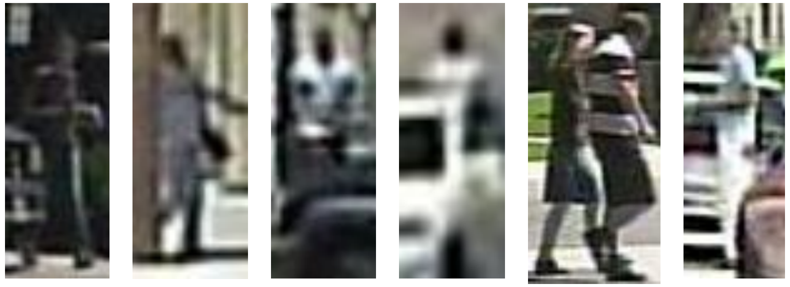
Pedestrian Statistics

The **Caltech Pedestrian Dataset** consists of approximately 10 hours of **640x480** 30Hz video taken from a vehicle driving through regular traffic in an urban environment. About 250,000 frames (in 137 approximately minute long segments) with a total of 350,000 bounding boxes and 2,300 unique pedestrians were annotated.

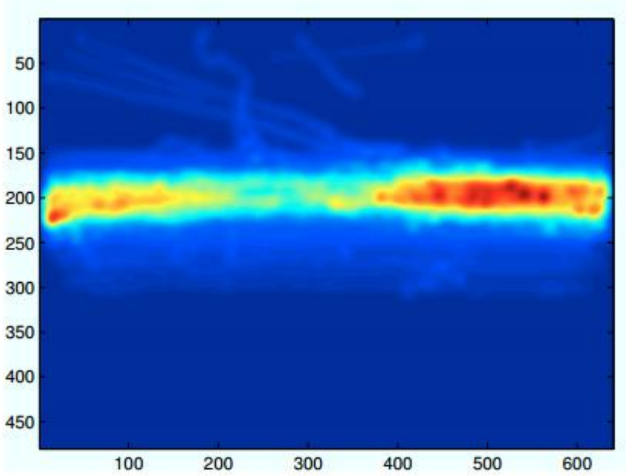
- **Various scales**: **10 ~ 250** in height (mainly 30~80)
- **Occlusion**: over 70% of pedestrians are occluded in at least one frame.
- **Distribution**: Narrow band running horizontally across the center of the image
- **Posture**: Stand still or walking



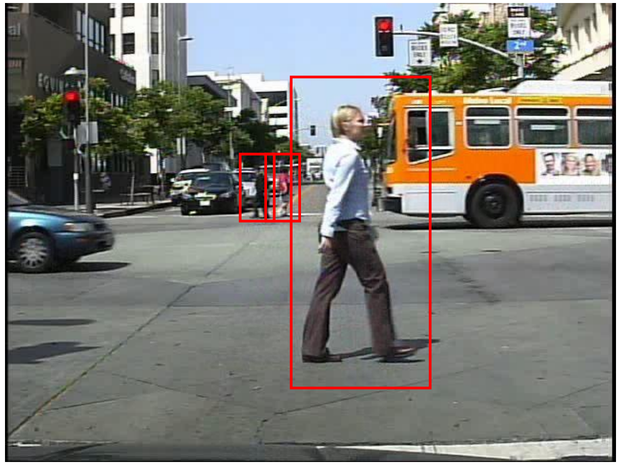
Height distribution



Occluded pedestrians



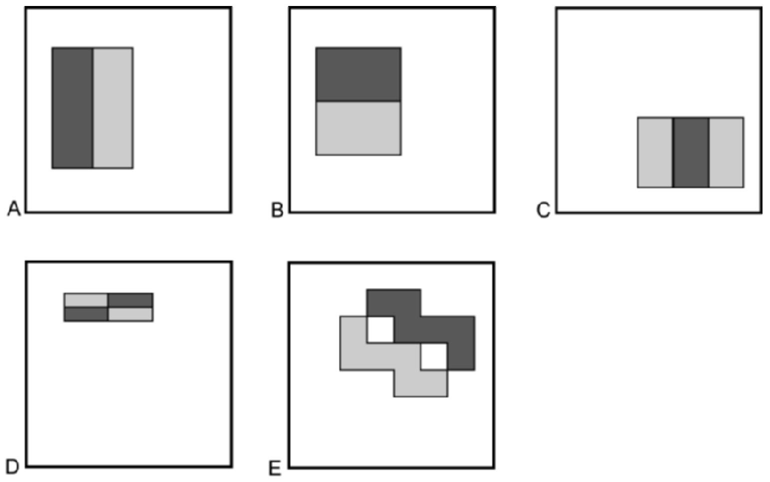
Distribution of pedestrians' position



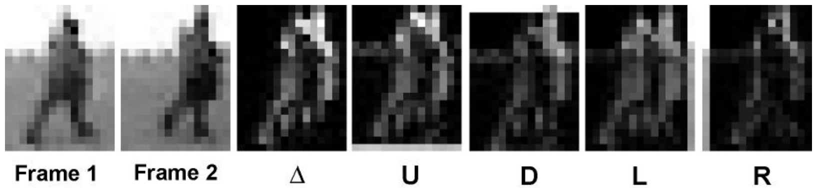
Sample frame

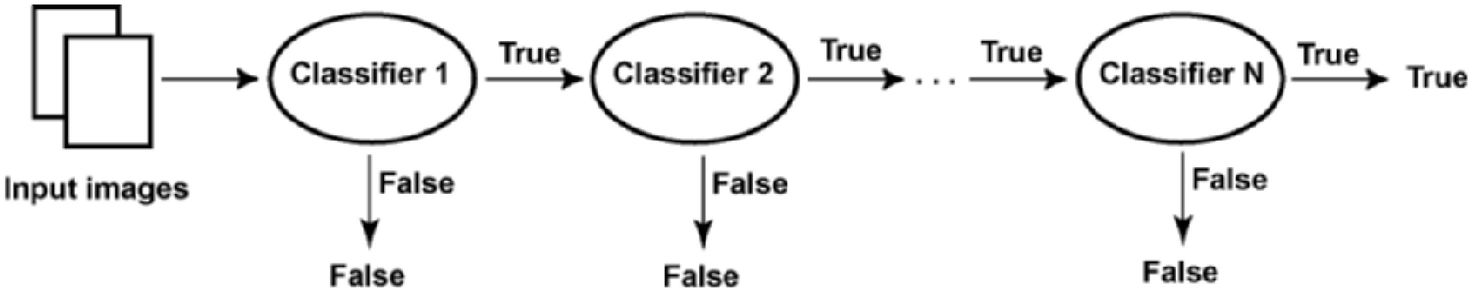
VJ Detector (ICCV 2003)

- Features of **motion and appearance** in integral images;
- Extension of the rectangle filters to the motion domain;
- Trained by AdaBoost algorithm;
- Very fast: 0.25 sec/image (360×240, 2.8 GHz P4 Processor)



Rectangle filters

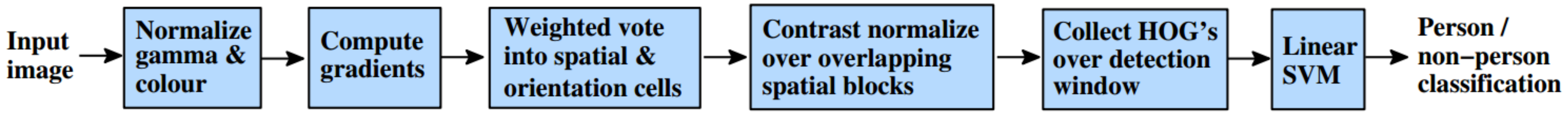


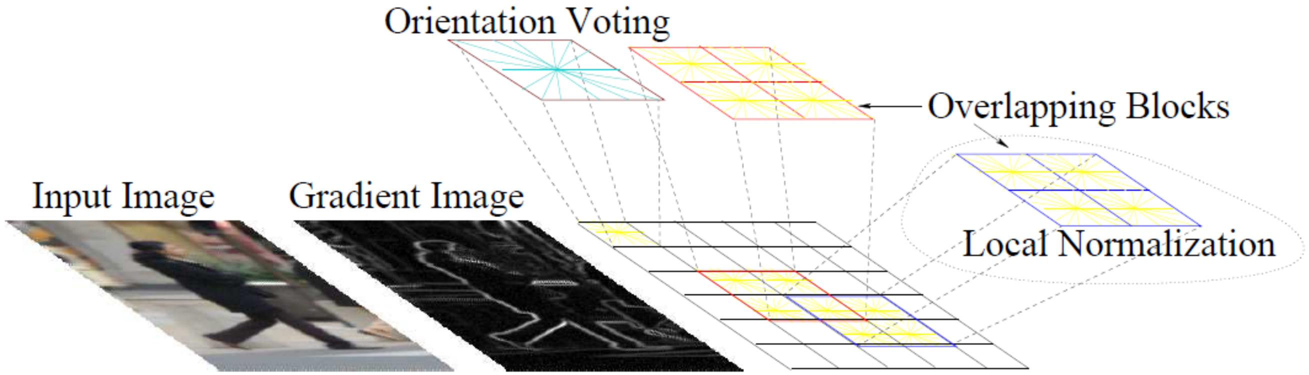


Cascade classifier

HOG Detector (CVPR 2005)

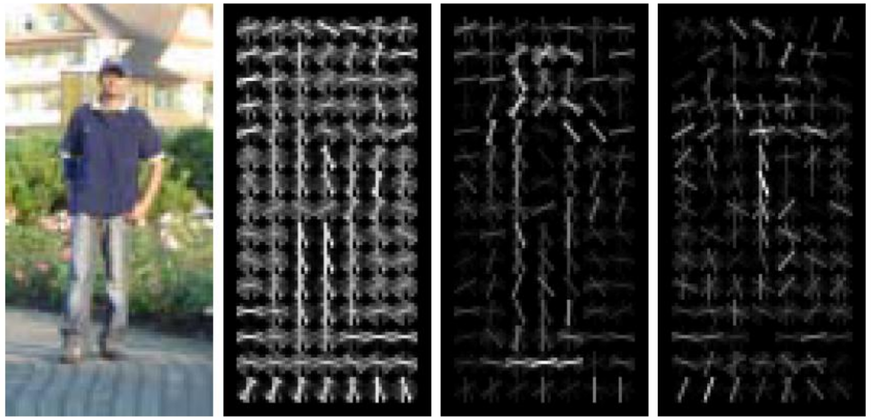
- Basic idea: Object appearance is characterized by the **distribution of local intensity gradients** or edge directions;
- Histogram of oriented gradient (HOG)+Linear SVM;





HOG

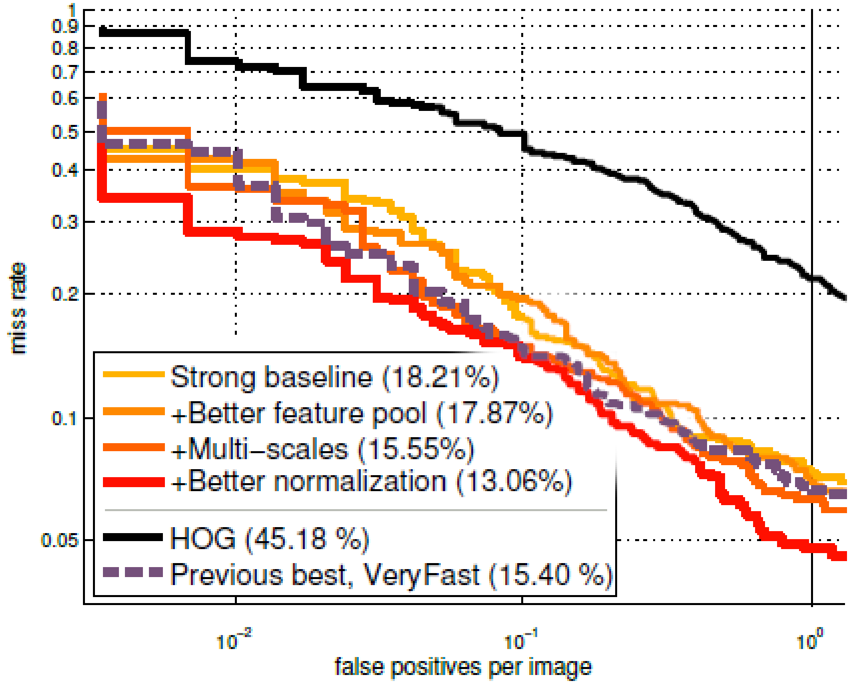
Positive Negative



R-HOG Weighted R-HOGs

SquaresChnFtrs (CVPR 2013)

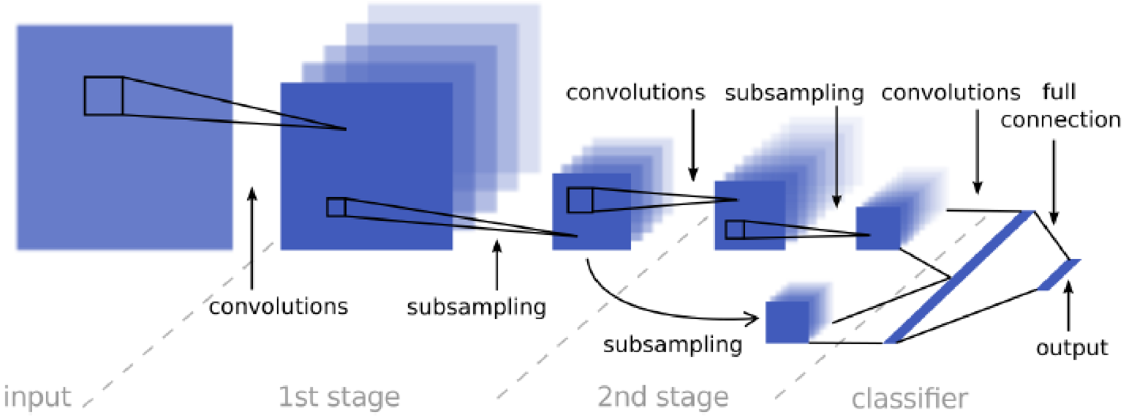
- Seek the strongest rigid detector;
- Best combination of features: **HOG+LUV** (with Nonlinear SVM);



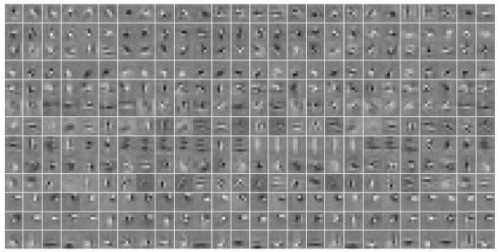
Detector aspect	Average miss-rate INRIA	Average miss-rate ETH
Strong baseline (§2)	18.21%	55.55%
+ AllFeatures (§4)	17.87%	55.50%
+ Multi-scales (§2)	15.55%	53.17%
+ GlobalNormalization (§5)	13.06%	43.90%
= Roerei detector	**13.06%**	**43.90%**
HOG+SVM	45.18%	65.03%
Previous best, VeryFast/MLS	15.40%	49.90%

ConvNet (CVPR 2013)

- Unsupervised method based on **convolutional sparse coding**;
- Two layers: Each layer initialized by convolutional sparse coding;
- 2nd stage: Extract a **global structure and local details**;



Multi-scale convolutional network



2nd layer filters

$$\tilde{z}_j = g_j \times \tanh\left(\sum_{i \in P_j} (x_i \otimes k_{j,i}) + b_j\right)$$

x: a set of feature maps
k: a set of 2D filters;

$$\mathbb{E}_{ConvSC} = \sum_i \|x_i - \sum_{j \in \bar{P}_i} \mathcal{D}_{i,j} \otimes z_j\|_2^2 + \lambda \|z\|_1$$

D: dictionary of filters

Problem Formulation

- **Proposal shifting problem** of pedestrian detectors:

  Poor localization quality of the detection proposals*

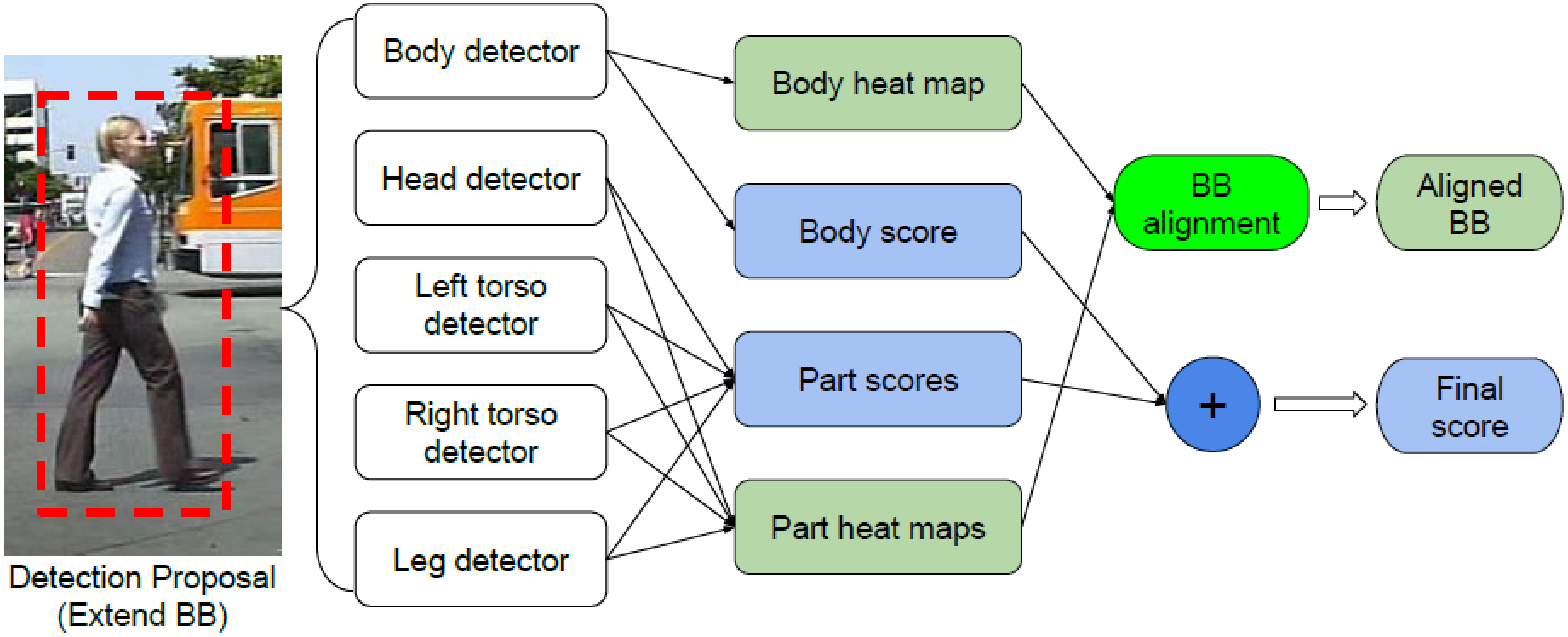
  - IoU ** = 0.5: Recall 93% of GT
  - IoU = 0.9: Recall only 10% of GT
- Detectors suffer from proposal shifting problem.
- Easily fail in body part detection:
  - Proposals lost some parts
  - Parts are not in the correct location
- **Part-based proposal alignment** is needed.

*Detection Proposal: Bounding box by pedestrian detection

**IoU: Intersection of Union

Examples of proposal shifting. Colored boxes are detection proposals, image regions with black boundaries are ground truths.

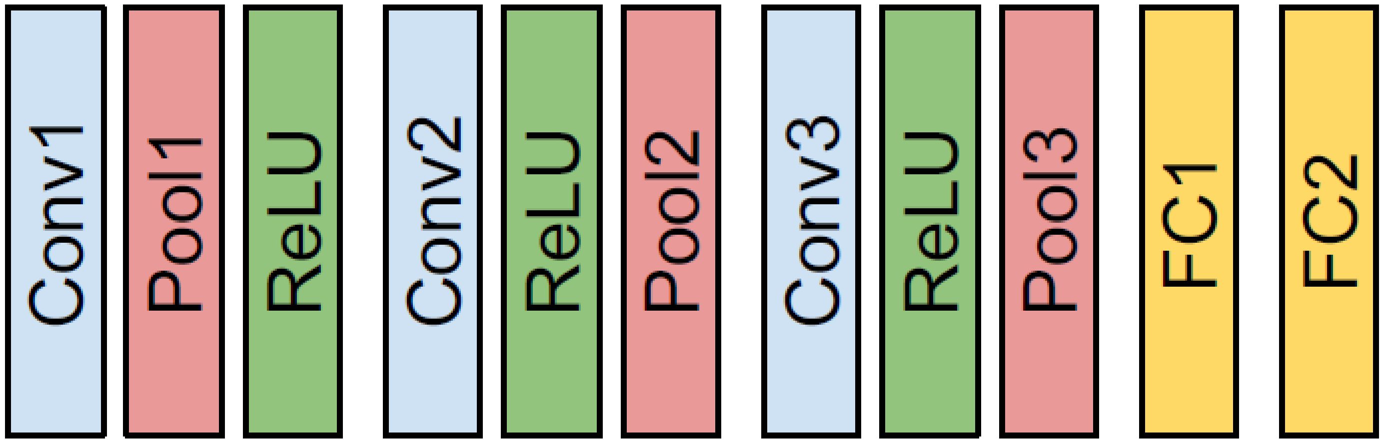
Proposed Method



We combine CNN and FCN* to generate the heat map and align the detection proposal.
We adopt part detection to recall the lost body parts.

*Shelhamer et al., “Fully convolutional networks for semantic segmentation,” Proc. IEEE CVPR 2015.

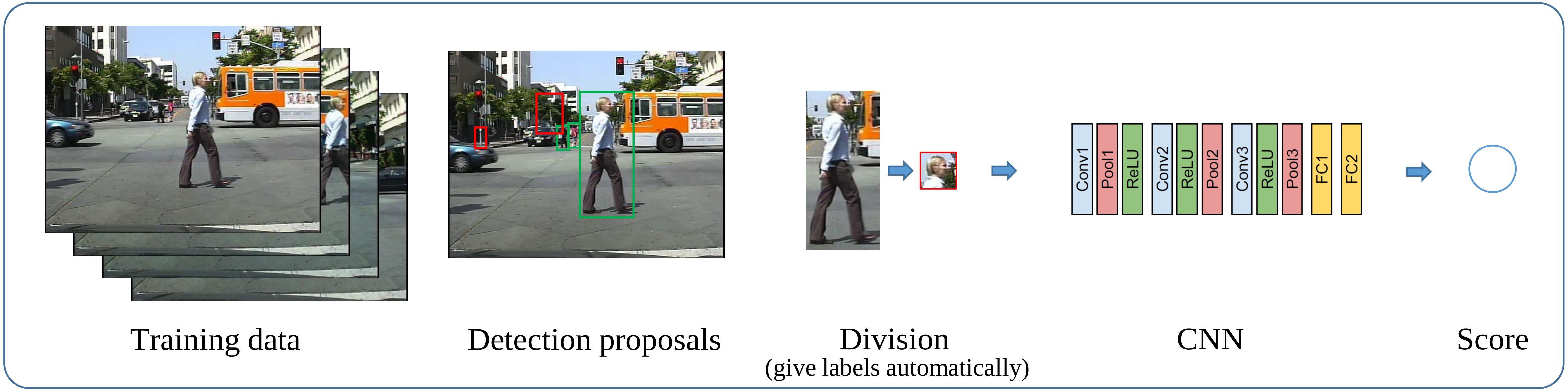
CNN Architecture: *CifarNet*



- Use `CifarNet`* for learning multiple layers of features (**caffe**)
- **3 convolutional layers, 3 pooling layers, 2 fully connected layers, softmax output**

*Hosang et al., "Taking a Deeper Look at Pedestrians," Proc. IEEE CVPR 2015.

Training: Part-Level CNNs

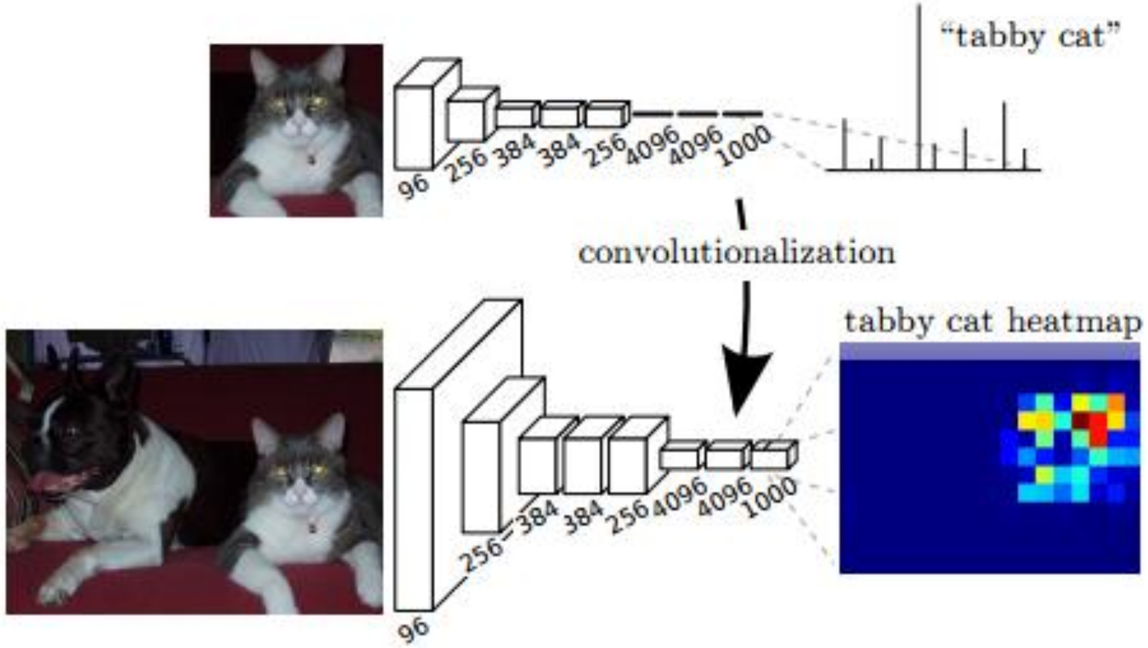


Example: Training the head detector

- **Training data**: Every 3^{th} frames in the training sequences (42782 frames; average 3 windows/frame);
- **Whole body**: Proposals resized into 128×64;
- **Part division**: Divide ground truth regions into 4 body parts (head, left torso & right torso: 32×32 , leg: 64×64)
- **Train** 5 CNNs for whole body, head, left torso, right torso and legs.
- 532 Mini-batch (128 patches) x 70 epochs to get the parameter set

FCN: Fully Convolutional Network

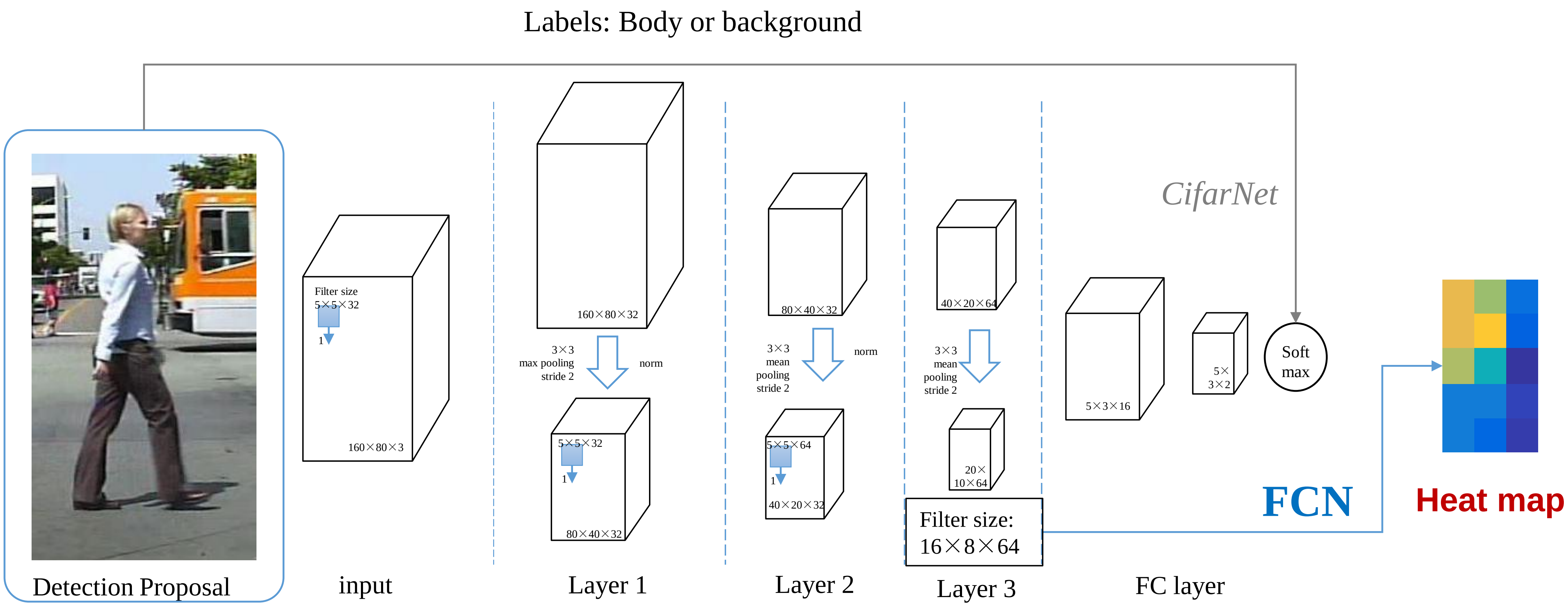
- Generate heat map for inference (**Semantics**);
- Transform fully connected layers into convolution layers;



$$f_{ks} \circ g_{k's'} = (f \circ g)_{k'+(k-1)s', ss'}$$

Convolutionalization

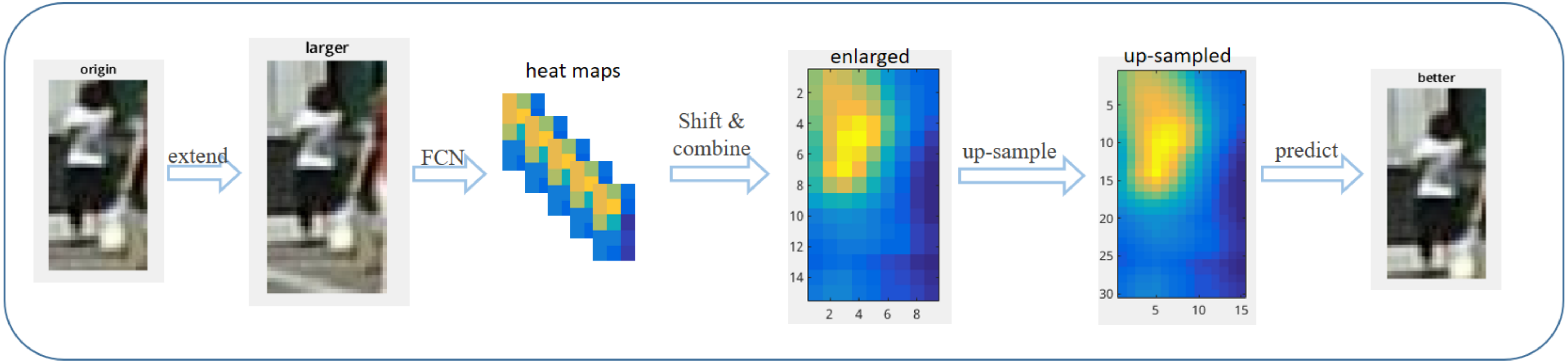
Testing: Part-Level FCN (CNN+FCN)



- Crop larger regions as proposals: 128×64→**160×80**
- **Detection proposals**: `SquaresChnFtrs`*(HOG+LUV);
- Output: **Heat map by FCN** for whole body or each part

• Benenson et al., “Seeking the strong rigid detector,” Proc. IEEE CVPR 2013.

BB Alignment

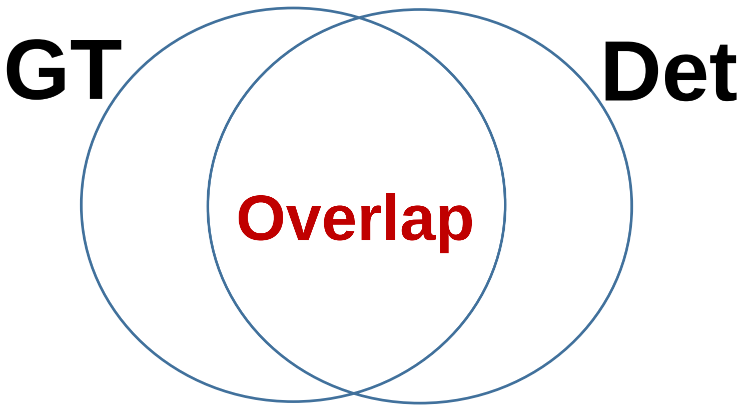


- **Origin**: Original BB, The person is located at the top left position;
- **Larger**: Enlarged BB;
- **Heat map**: Output of FCN (Coarse map);
- **Enlarged**: Shift each heat map by *f* (=3) pixels on 2 directions (dilation), and combine them;
- **Up-sampled**: Up-sampled heat map into a corresponding size;
- **Better**: Align BB with **the highest average score**;

Experimental Results

- Evaluation metrics:
  Miss rate-false positive per image (FPPI) curve;
  Log-average miss rate;

- Detection proposals:
  Generated by `SquaresChnFtrs` (Log-average miss rate: about 34.8%);



Overlap >50%

Experimental Results

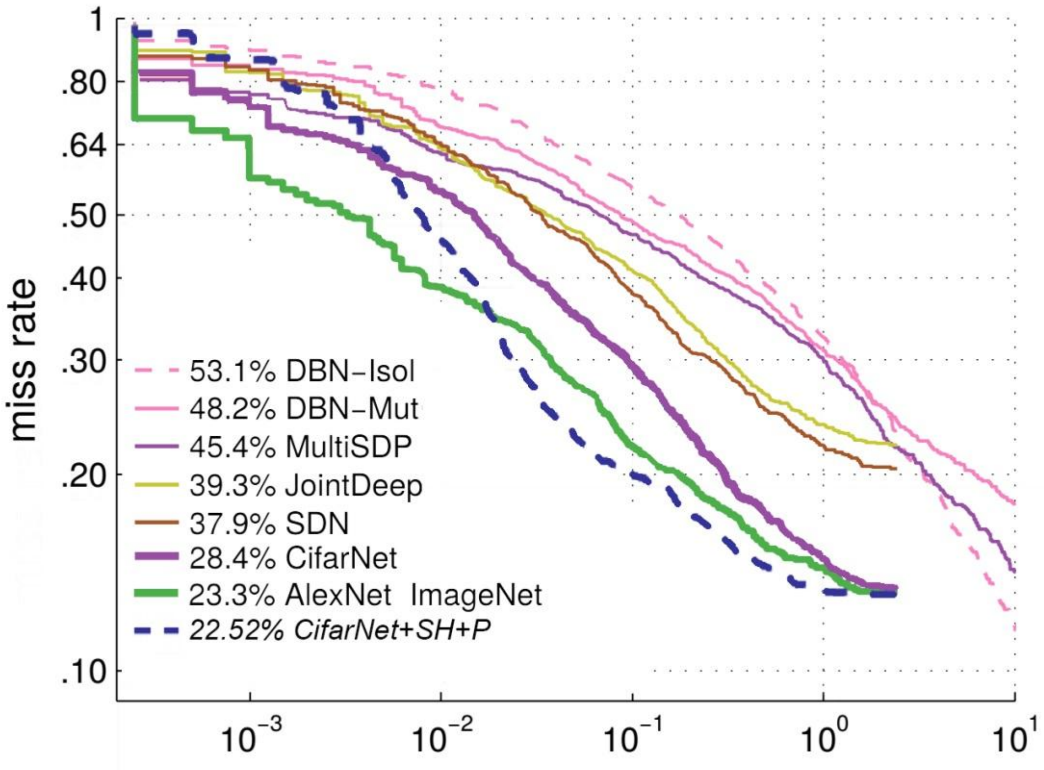
3 methods for performance comparison:

- `CifarNet`: CifarNet on pedestrian detection (CVPR 2015);
- `CifarNet+SH`: CifarNet with BB alignment;
- `CifarNet+SH+P`: Proposed (Part-level FCNs with BB alignment);

6.83% improvement in log-average miss rate over `CifarNet`

Method	Avg. miss rate (%)	Improvement (%)
CifarNet	29.35	----
CifarNet+SH	26.27	3.08
CifarNet+SH+P	*22.52*	3.75

Experimental Results



Comparison with other deep learning ones with a few convolution layers

	training data	miss rate (%)
ConvNet	INRIA	77.20
DBN-Isol	INRIA	53.14
DBN-Mut	INRIA/Caltech	48.22
JointDeep	INRIA/Caltech	39.32
SDN	INRIA/Caltech	37.87
LFOV	Caltech	35.85
DeepCascade	Caltech	31.11
CifarNet	Caltech	28.40
DeepCascade+	Caltech+	26.21
SCF+AlexNet	Caltech+ImageNet	23.32
Proposed	Caltech	*22.52*
TA-CNN	Caltech++	20.86
DeepParts	Caltech+ImageNet	11.89
SA-FastRCNN	Caltech+ImageNet	9.68

More layers

Experimental Results

- BB alignment results

Conclusions

- We have proposed part-level fully convolutional networks for pedestrian detection.

- We have handled **detection proposal shifting problem using deep learning.**

- Two main contributions to pedestrian detection:
  - Part-level detection to recall the lost body parts
  - CNN+FCN for BB alignment

- We have achieved **6.83% performance improvement** in log-average miss rate over `CifarNet`.

THANK YOU!