[ICIP 2017]

Direct Multi-Scale Dual-Stream Network for Pedestrian Detection

Sang-Il Jung and Ki-Sang Hong

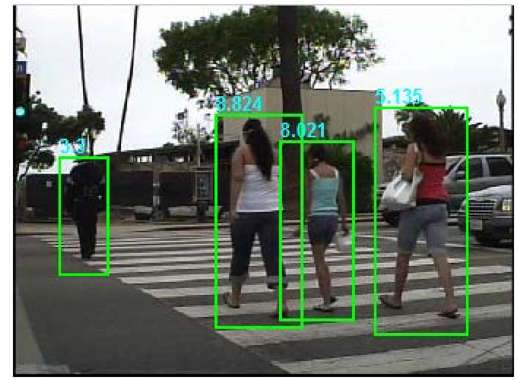
Image Information Processing Lab., POSTECH

Pedestrian Detection

- **Goal**
  - To draw bounding boxes that tightly enclose pedestrians given an image

Input: an Image $\mathbf{I}$

Pedestrian Detector



Output: a set of bounding boxes and scores $\{(B_i, s_i)\}_{i=1,\ldots,N}$

Classification

$s_i \in \mathbb{R}$

Is it a pedestrian or not?
How confident?

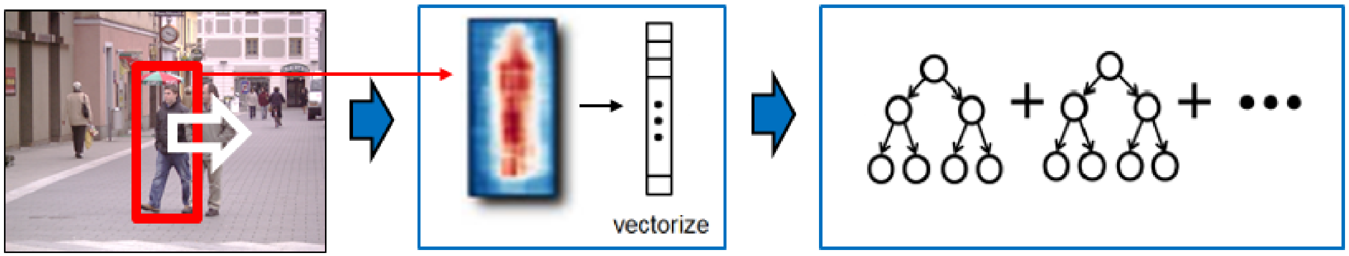
Localization

$B_i = (x_i, y_i, w_i, h_i) \in \mathbb{R}^4$

Where is the box located? (x_i, y_i)
What is the size of the box? (w_i, h_i)

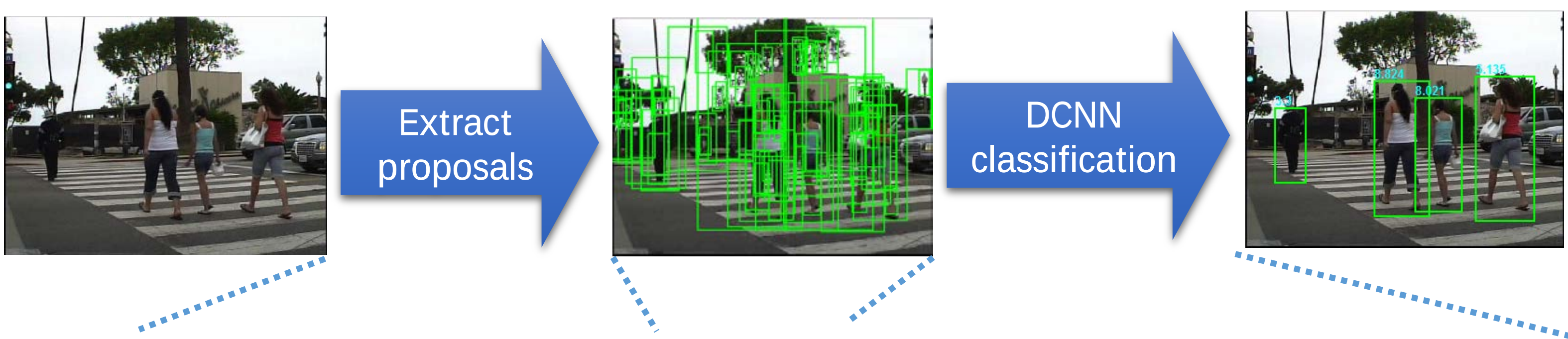
Introduction

- **Pedestrian detection methods**
  - AdaBoost based methods: sliding window, hand-designed feature



Sliding window — Generate feature — Classify using boosted forest

  - (Region) DCNN based methods: region proposal, DCNN classification



- AdaBoost based methods
- Region proposal network

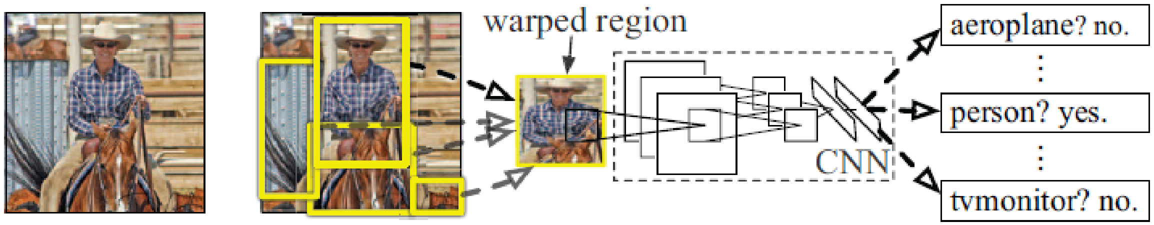
- Deep convolutional neural network
- Fine-tuning for pedestrian detection
- Bounding box regression

Introduction

- **Object detection (Region, DCNN)**

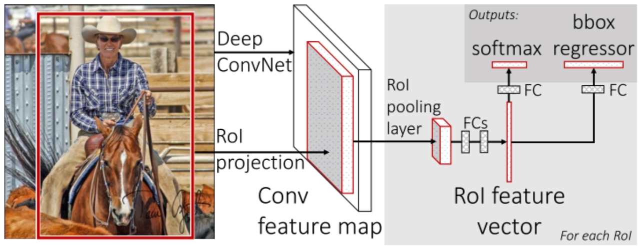
  - **(R-CNN) Region-based CNN**

    R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014.



  - **(Fast R-CNN)**

    R. Girshick. Fast r-cnn, ICCV 2015.



+ share convolutional features

  - **(Faster R-CNN)**

    S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, NIPS 2015.

+ integrate region proposal network (RPN)

Introduction

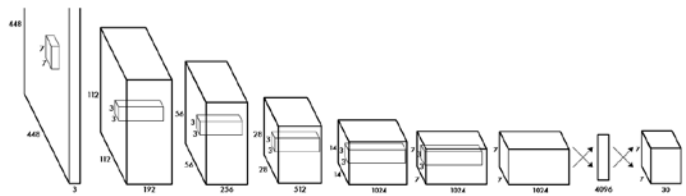
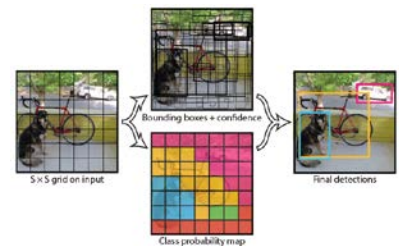
- **Pedestrian detection (Region, DCNN)**

Input: An image	Extract proposals	Classification	
(**SCF+AlexNet**) J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians, CVPR 2015.	**External proposal methods** : AdaBoost-based pedestrian detector SCF, ACF ⋮	**Feature:** DCNN **Classification:** softmax	➢ **R-CNN** Pipeline
(**SAFastRCNN**) J. Li, X. Liang, S. Shen, T. Xu, and S. Yan. Scale-aware fast r-cnn for pedestrian detection, arXiv 2015.		Joint **classification** and **bounding box regression** using DCNN	➢ **Fast R-CNN** Pipeline (share conv. features) ➢ Scale-aware: two sub-networks
(**MS-CNN**) Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection, ECCV 2016.	**Region Proposal Network (RPN)**		➢ **Faster R-CNN** Pipeline (+RPN) ➢ Branching multi-scale network
(**RPN+BF**) L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection, ECCV 2016.		**Feature:** DCNN **Classifier:** BF	➢ DCNN proposal + BF classifier

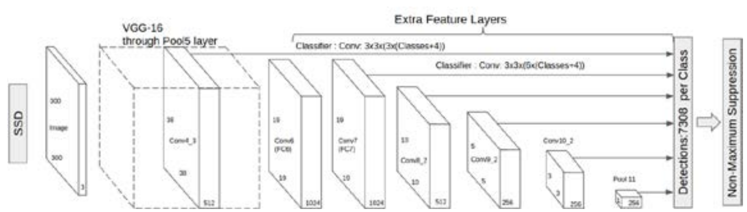
Introduction

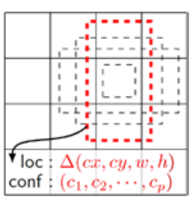
- **Direct detector without extracting proposals**
  - Object detection

    (**YOLO**) J. Redmon, S. Divvala, R. Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016.

    (**SSD**) W. Liu, D. Angueelov, D. Erhan, C. Szegedy, S. Reed, C-Y. Fu, A. C. Berg. SSD: Single Shot MultiBox Detector, ECCV 2016.

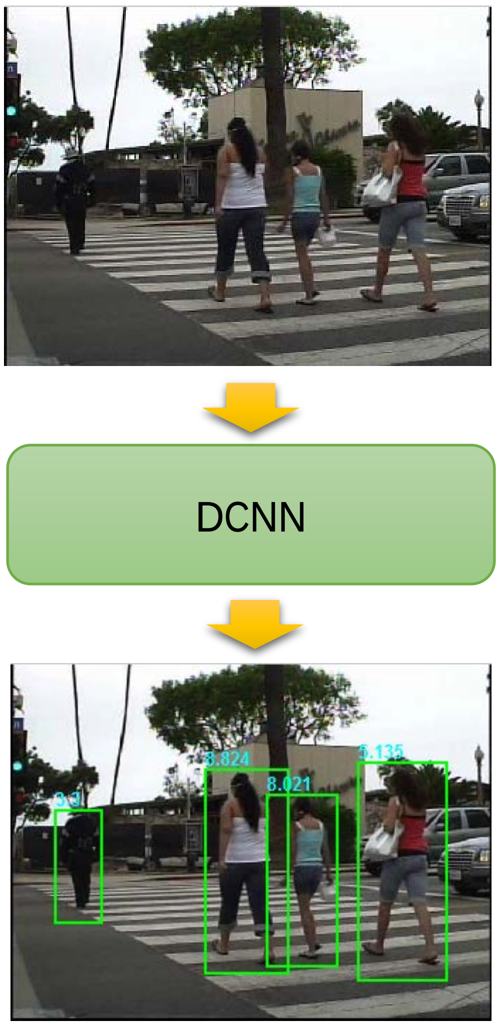




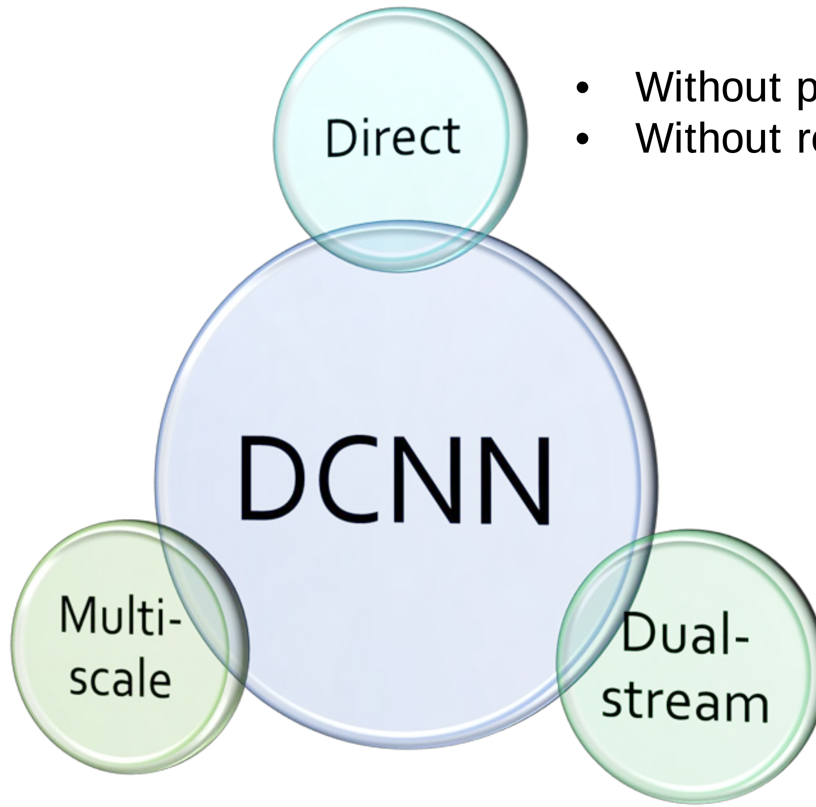
  - Pedestrian detection

Direct Multi-scale Dual-stream network (DMDnet)	"Direct" + "Multi-scale" + "Dual-stream"	**Our work!**

Overview (ours)



A single DCNN that takes a full-size image as an input and outputs pedestrians of various sizes



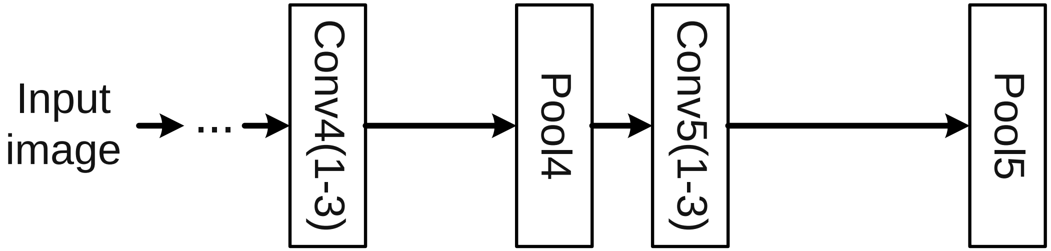
- Without proposal
- Without resampling

- Detect Various-sized pedestrians

- Local + contextual feature

Network architecture

- Base feature extraction: convolutional layers of VGG-16 network

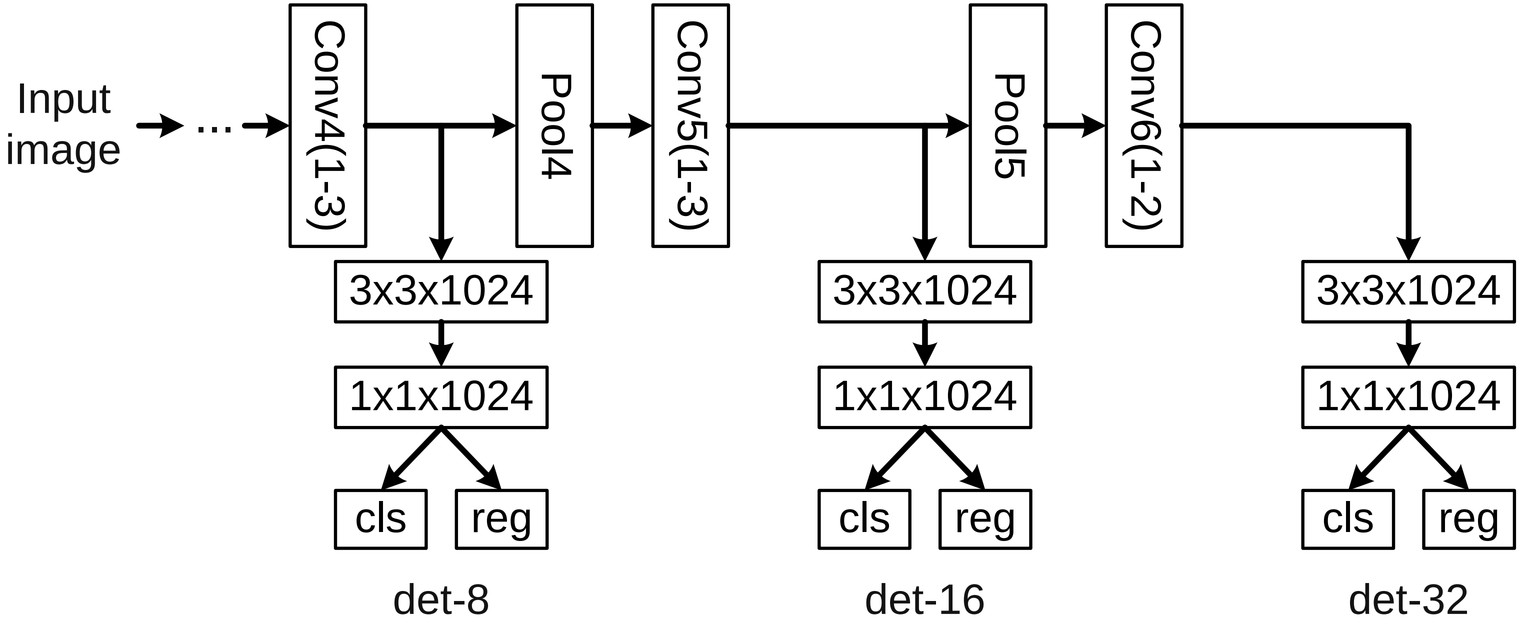


(**VGG-16 network**) K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, ICLR 2015.

Network architecture

(**MS-CNN**) Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection, ECCV 2016.

- **DMnet:** resemble the RPNs of MS-CNN

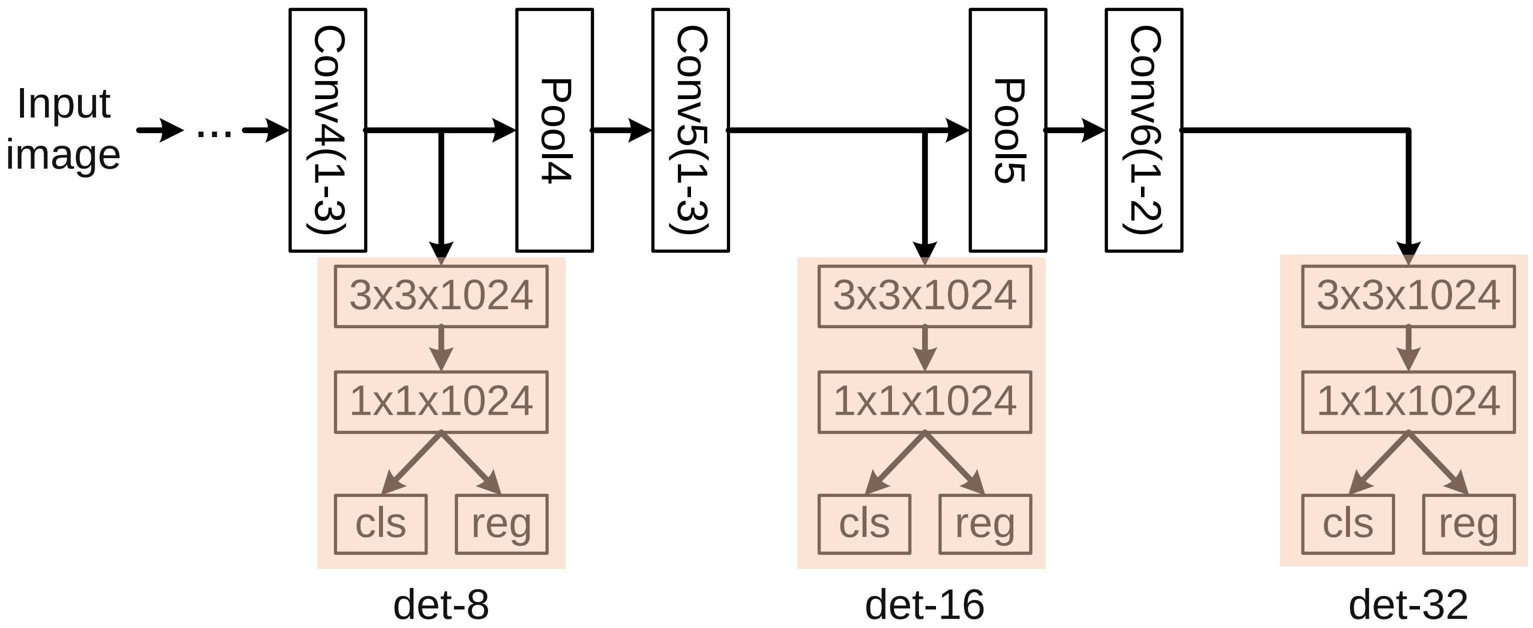


Direct + Multi-scale Network

Network architecture

(**MS-CNN**) Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection, ECCV 2016.

- **DMnet:** resemble the RPNs of MS-CNN

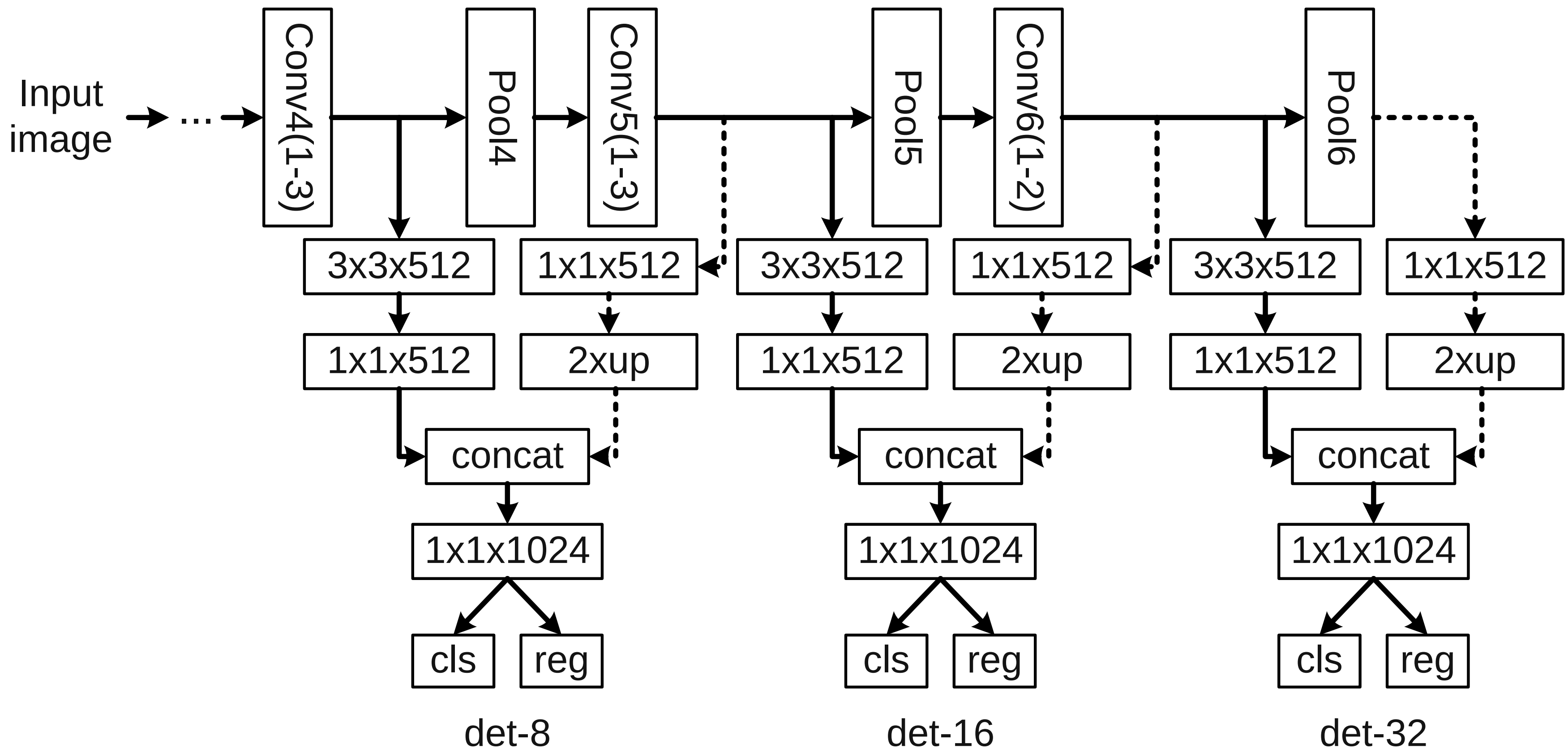


Each detector sees the region of the proper receptive fields according to the sizes of pedestrians (anchor boxes).

Direct + Multi-scale Network

Network architecture

- **DMDnet**

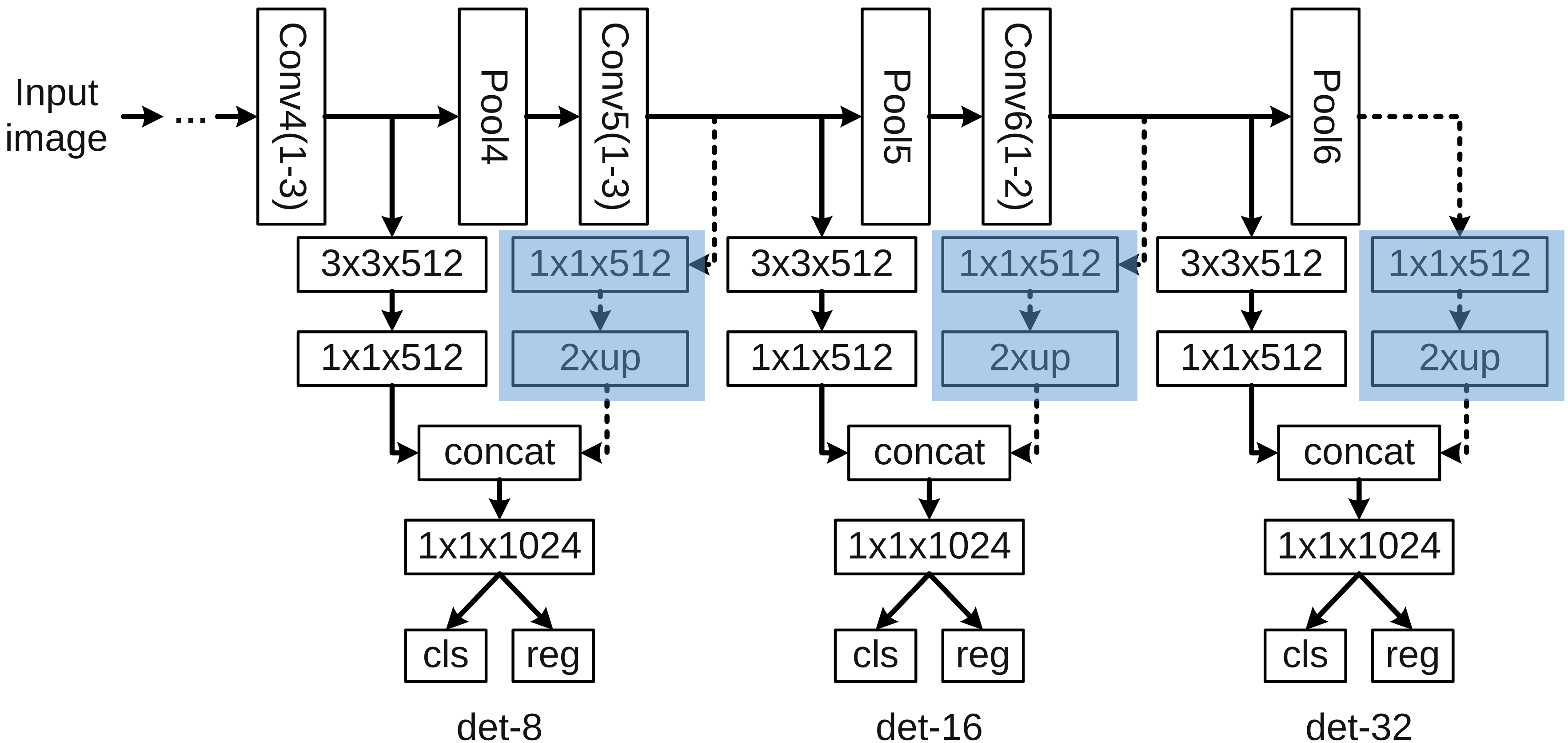


Direct + Multi-scale + Dual-stream Network

Network architecture

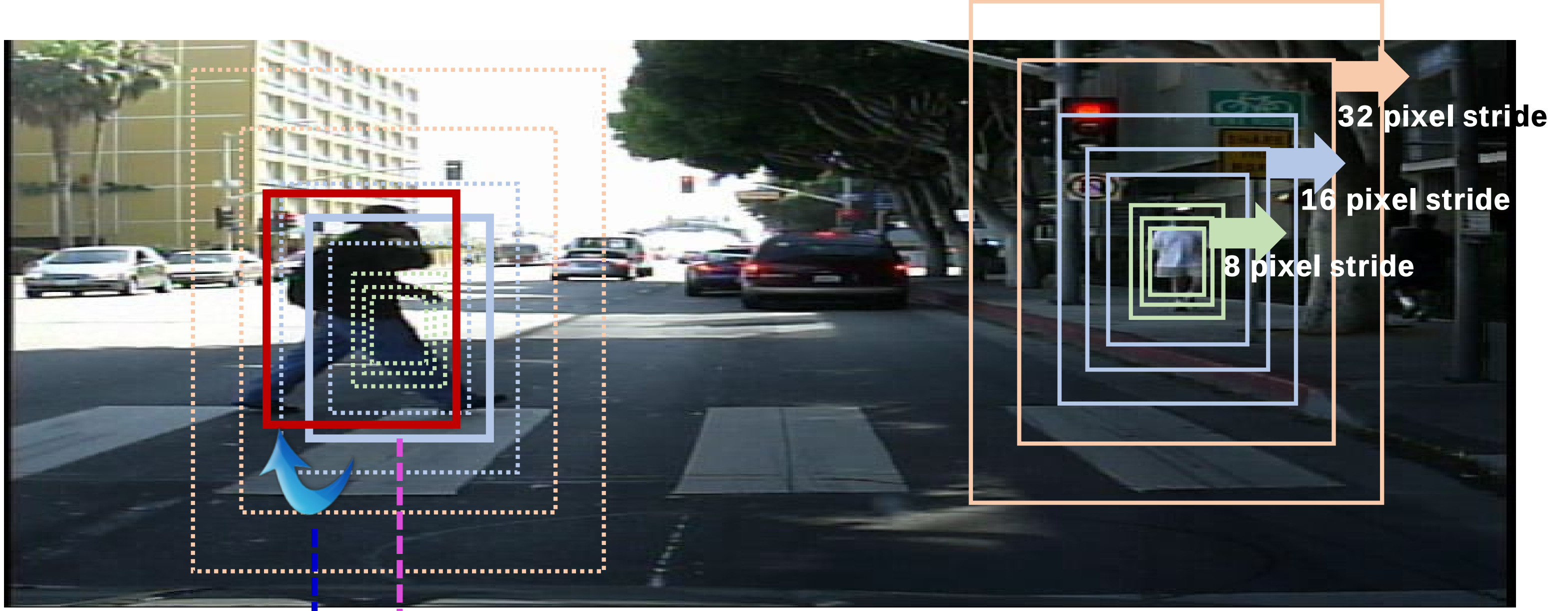
- **DMDnet**

The second stream networks encodes contextual information!



Direct + Multi-scale + Dual-stream Network

Anchor boxes



Classification: does the anchor box contain a pedestrian?

Localization: compute box transformation offset

- Generating anchor boxes of eight different scales with scale stride 1.3
- Assignment

Detection networks	det-8	det-16	det-32
Branching layers	Conv4-3	Conv5-3	Conv6-2
Receptive fields	92×92	196×196	340×340
Heights of Anchor boxes	$\{50, 65, 84.5\}$	$\{109.8, 142.8, 185.6\}$	$\{241.3, 313.7\}$

Loss

- **Combined loss**

R. Girshick. Fast r-cnn, ICCV 2015.

$$L = \sum_{k=1}^{K} \sum_{\mathbf{b}_i \in B^k} \left(\boxed{l_{cls}^k(\hat{p}_i, y_i)} + \gamma \boxed{l_{loc}^k(\hat{\mathbf{t}}_i, \mathbf{t}_i)} \right)$$

Localization loss: smooth$_{L1}$ loss

Classification loss: 2-way softmax log loss

B^k : the boxes that belongs to the k-th anchor box type

$\mathbf{b}_i$: the i-th box

$\hat{p}_i$: the estimated probability

y_i : the ground-truth class label

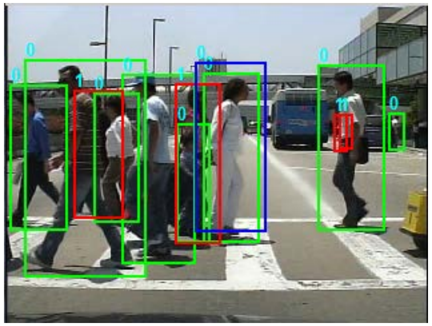
$\hat{\mathbf{t}}_i$: the estimated bounding box regression offsets

$\mathbf{t}_i$: the target bounding box regression offset

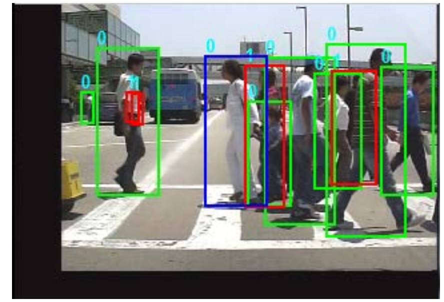
γ : balancing hyper-parameter

Target generation

Training image



Scale / Flip

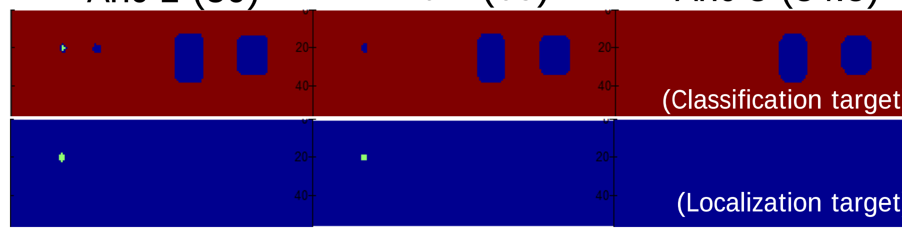


Width 2x

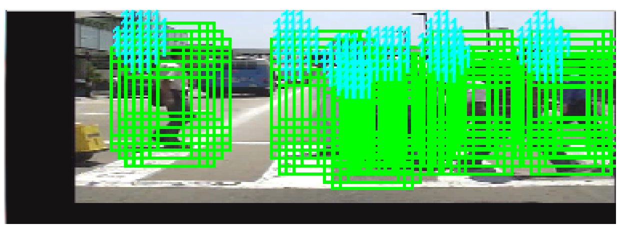
Anchor box

det-8

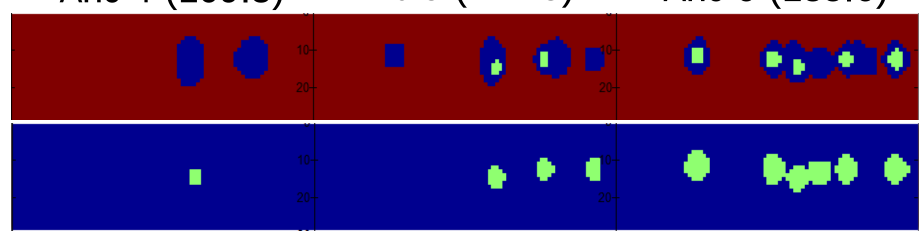
Anc 1 (50) Anc 2 (65) Anc 3 (84.5)



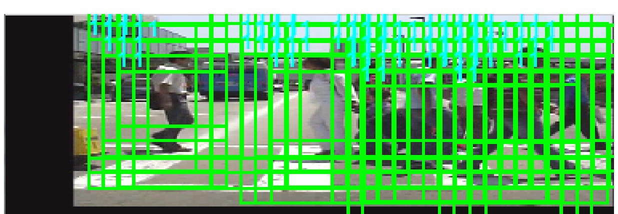
det-16



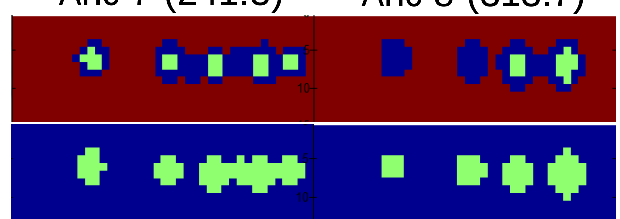
Anc 4 (109.8) Anc 5 (142.8) Anc 6 (185.6)



det-32



Anc 7 (241.3) Anc 8 (313.7)



: ignore

: positive (cls/loc)

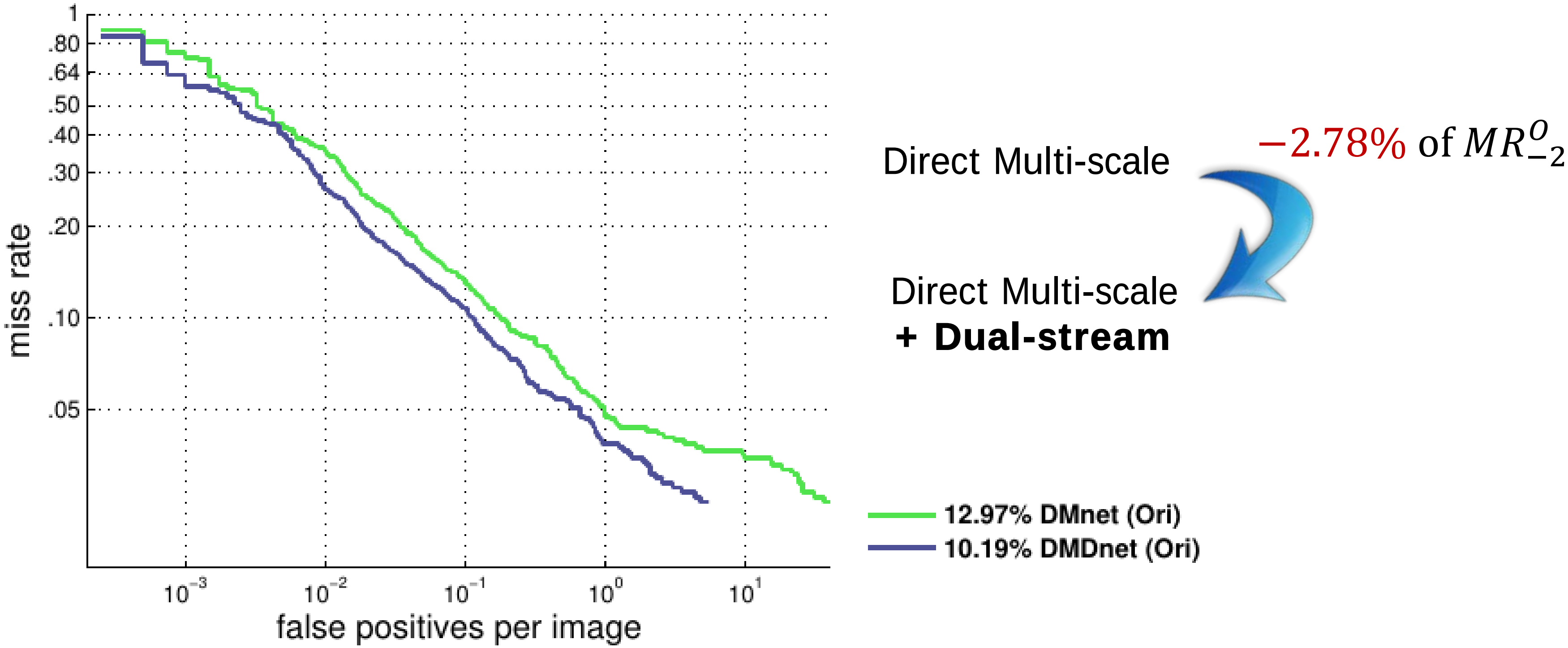
: negative (cls)

Target generation

- **Sampling**
  - Purpose: to balance the number of samples (pos/neg ,cls/loc) when the combined loss is computed during training
  - Sampling criteria (for a mini-batch of 2 images)
    - 1) Positive sample: select $n_p(\leq 32)$ with IOU $\geq$ 0.6
    - 2) Negative sample: select $n_n(= \min(32, 3n_p))$ with IOU $\leq$ 0.4
    - 3) Localization sample: select $n_l(\leq 32)$ with IOU $\geq$ 0.45
  - Sampling method
    - Random sampling: select the samples according to uniform distribution
    - Bootstrapping sampling: select the samples that have top-$n_{\{p,n,l\}}$ largest loss
  - Sampling strategy
    - Random sampling for the first 12,000 iterations and Bootstrapping sampling for the rest of iterations
    - If the Bootstrapping sampling is applied from the beginning, then the loss does not converge

Experimental Results

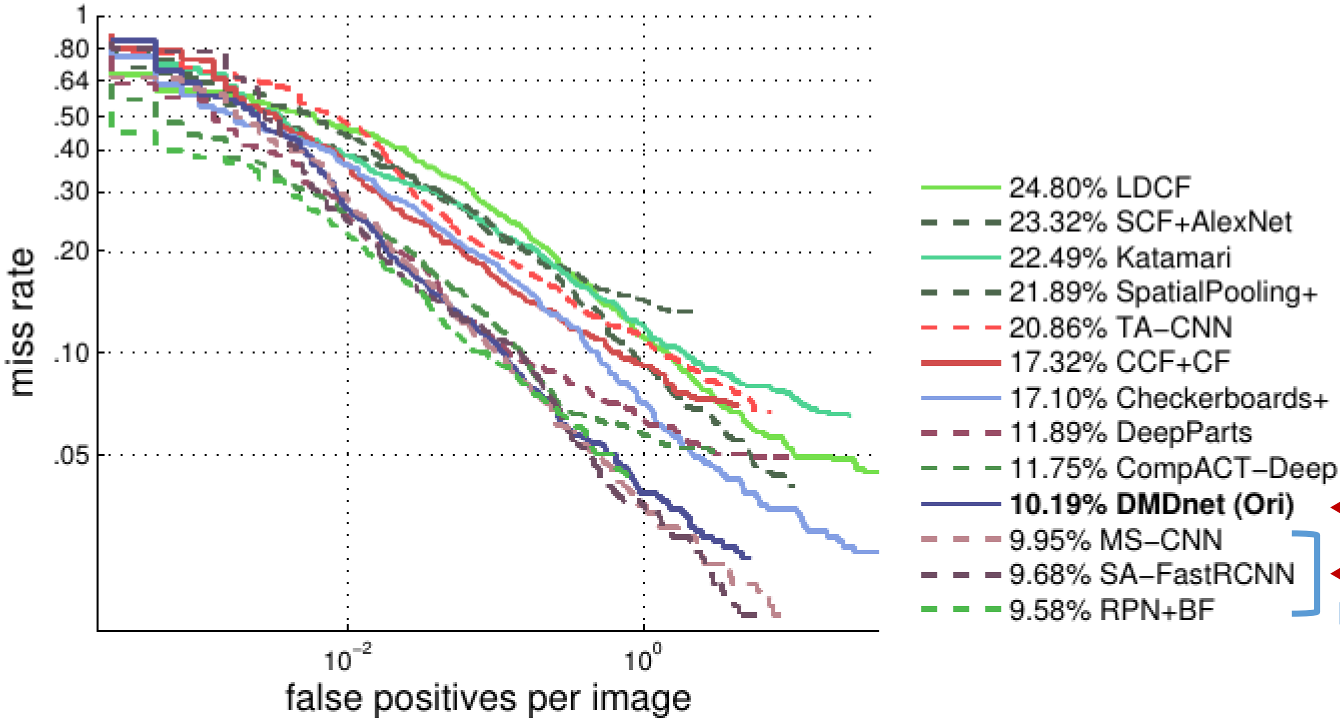
- **DMnet vs. DMDnet**



The miss rates over FPPI of DMnet and DMDnet. The evaluation was conducted on the **Caltech** testing set with original annotation (MR^{O}_{-2}). The `(Ori)' indicates that the networks were trained with **original** annotation.

Experimental Results

- **Comparison with the state-of-the-art methods**



Comparable !

Region-based methods

Evaluation on the **Caltech** testing set with **original** annotation (MR^{O}_{-2}).

Experimental Results

Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection. CVPR 2016.

- **Top-40 false positives (FPPI 0.01)**

: ground-truth
: false positive

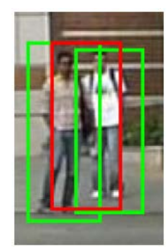
Missing annotation (11)

Bad annotations (10)

Background (12)

Confusing (3)

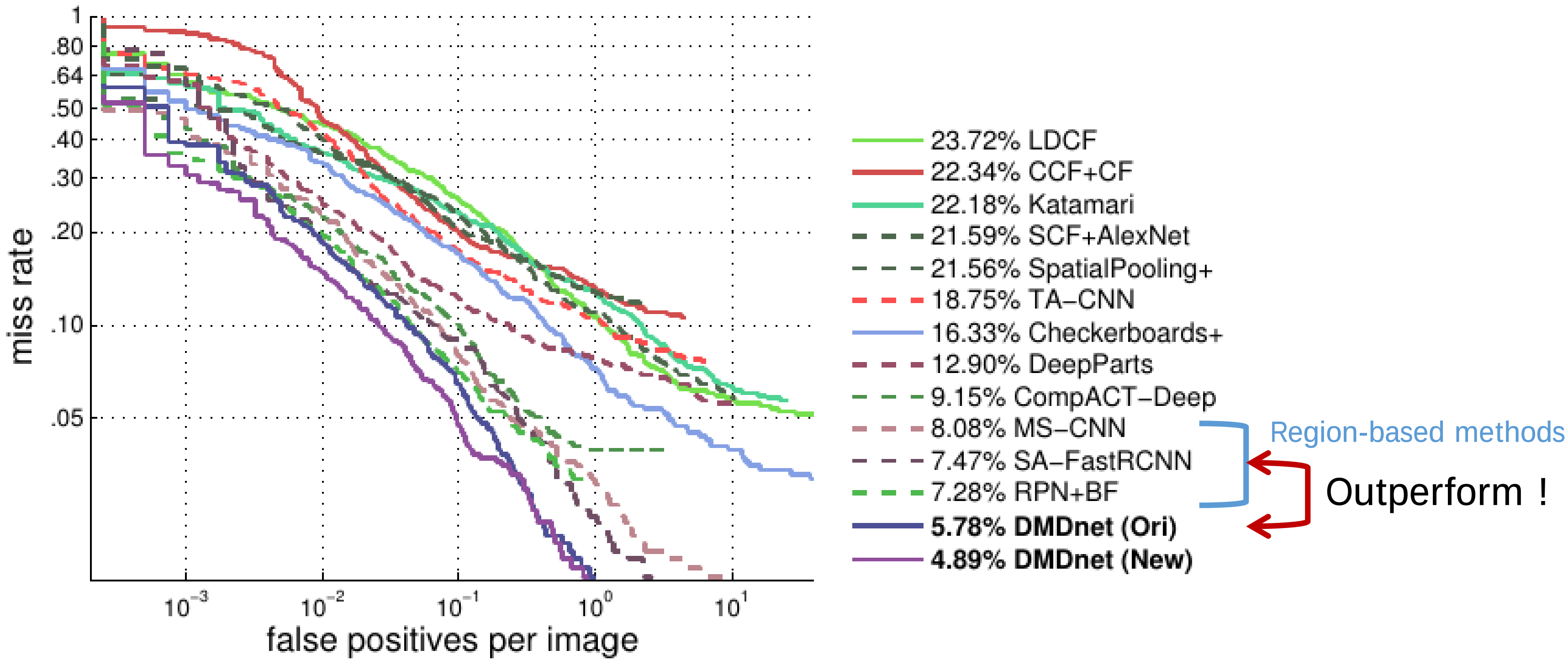
Bad localization (3)



Double detection (1)

Experimental Results

- **Comparison with the state-of-the-art methods**



Evaluation on the **Caltech** testing set with **new** annotation (MR^{N}_{-2}). All other methods except `DMDnet (New)' were trained with the original annotation.

Experimental Results

- **Detection speed**

Detection accuracy and speed. The detection speeds of other methods were obtained from the original papers. Due to the differences in hardware and implementation details, direct comparison is invalid, but this roughly shows that DMDnet is computationally efficient.

Methods	MR^{O}_{-2}	MR^{N}_{-2}	Detection speed
MS-CNN	9.95%	8.08%	8 im/s
SA-FastRCNN	9.68%	7.47%	2.7 im/s
RPN+BF	**9.58**%	7.28%	2 im/s
DMDnet*	10.19%	**5.78**%	**8.4** im/s

* Intel i7 3.60-GHz CPU, TitanX GPU, MatConvNet library

(**MatConvNet**) A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab, ACM, 2015.
(**MS-CNN**) Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection, ECCV 2016.
(**RPN+BF**) Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection?, ECCV 2016.
(**SA-FastRCNN**) Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. arXiv 2015.

Concluding Remarks

- **Summary**
  - A direct DCNN for pedestrian detection
  - Direct detector → No extracting proposals and resampling
  - Multi-scale → Branching networks depending on the size of pedestrians
  - Dual-stream → Concatenating two types of features
  - Outperformed detection accuracy on the Caltech dataset with new annotation
  - Fast processing time: ~8.4 images/s

Thank you !