i-VisionGroup@Tsinghua

Multi-Grained Deep Feature Learning for Pedestrian Detection

Chunze Lin, Jiwen Lu, Jie Zhou

Tsinghua University, China

Outline

① Introduction

② Related Work

③ Proposed Approach

④ Experimental Results

What is Pedestrian Detection?

Input image

machine

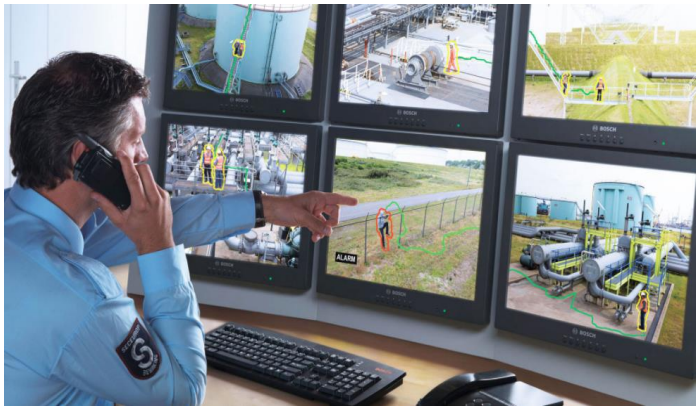
Output detection

Complex problem for machine

- **Recognition**: What is a pedestrian?
- **Localization**: Where are pedestrians?

Why Pedestrian Detection?

Autonomous driving

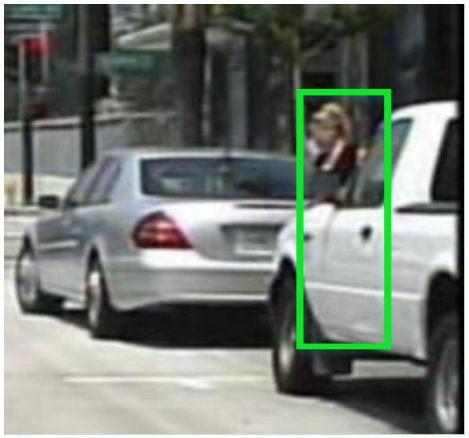


Intelligent surveillance

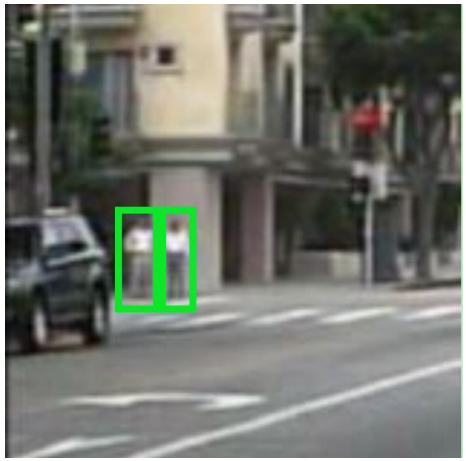
Robotics

Challenges

Large variances of scales



Occlusions



Blurry representation

Noisy background

Outline

① Introduction

② Related Work

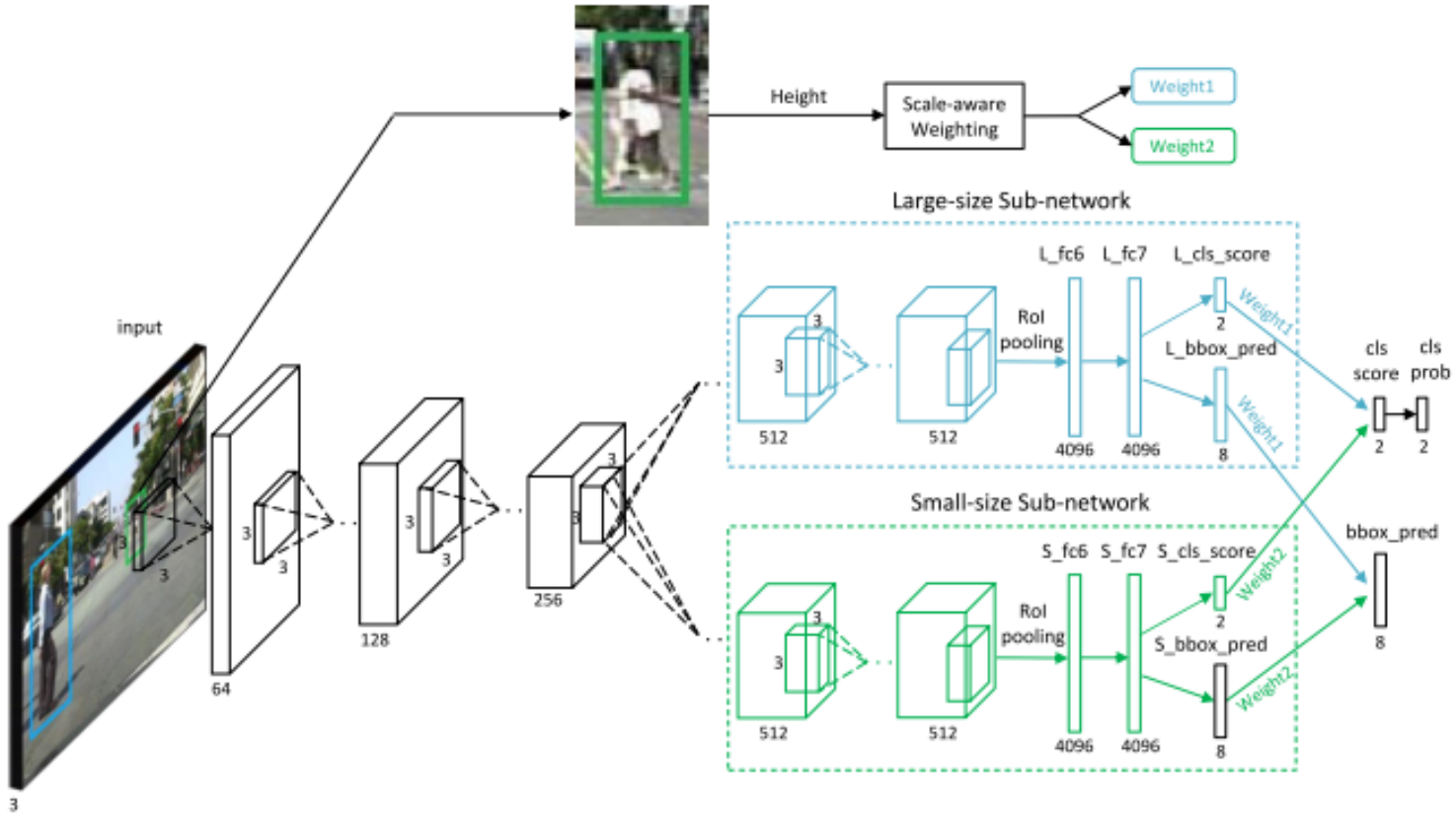
③ Proposed Approach

④ Experimental Results

Possible Solutions

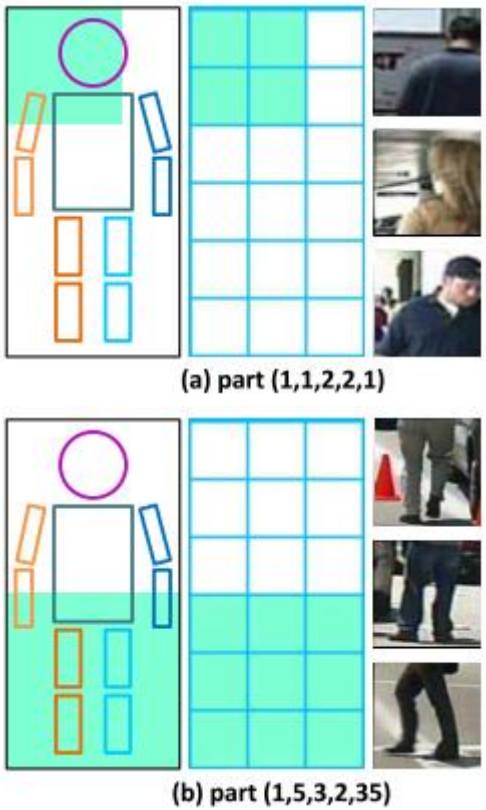
- Learn multiple models for different scales
- Human parts detection
- Hard negative samples mining

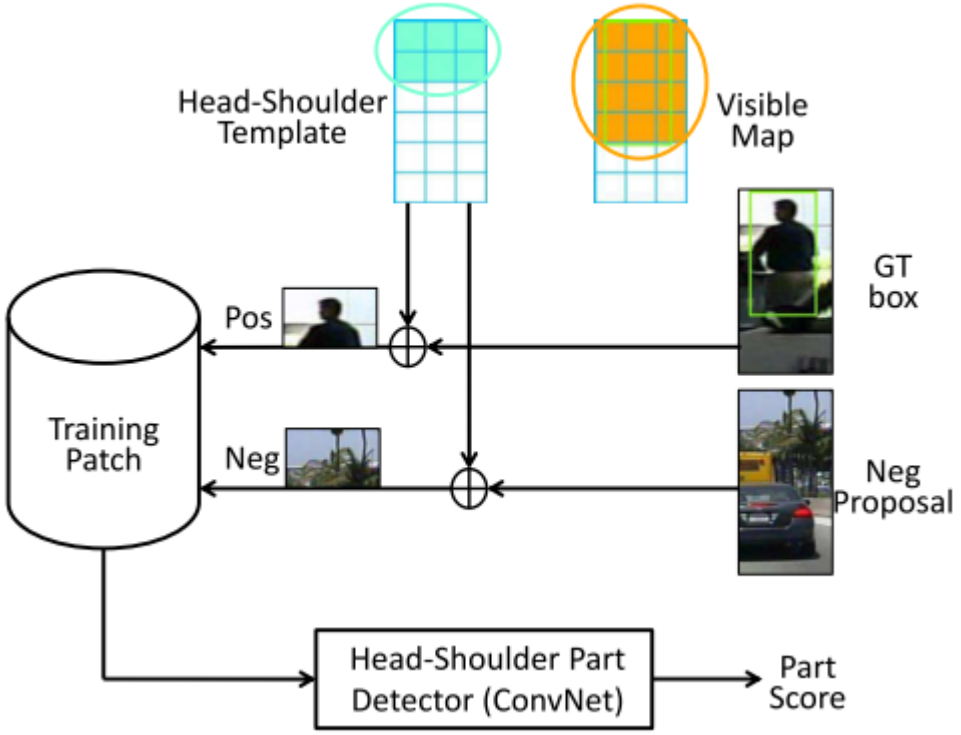
Scale-aware Fast R-CNN for Pedestrian Detection[1]



[1] Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2018). Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, *20*(4), 985-996.

Deep learning strong parts for pedestrian detection[2]

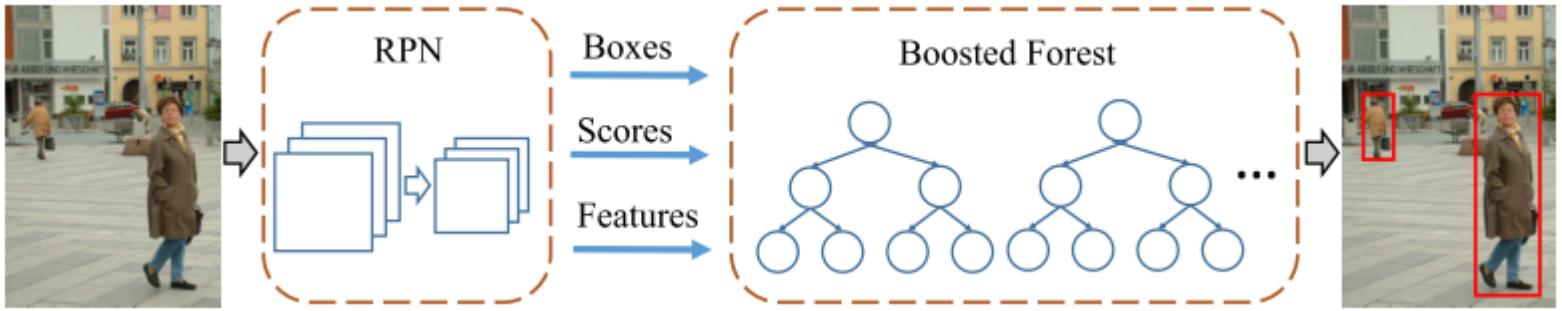




[2] Tian, Y., Luo, P., Wang, X., & Tang, X. (2015). Deep learning strong parts for pedestrian detection. In Proceedings of the IEEE international conference on computer vision (pp. 1904-1912).

Is faster R-CNN doing well for pedestrian detection?[3]

(a) Small positive instances (b) Hard negatives



[3] Zhang, L., Lin, L., Liang, X., & He, K. (2016, October). Is faster R-CNN doing well for pedestrian detection?. In *European Conference on Computer Vision* (pp. 443-457).

Outline

① Introduction

② Related Work

③ Proposed Approach

④ Experimental Results

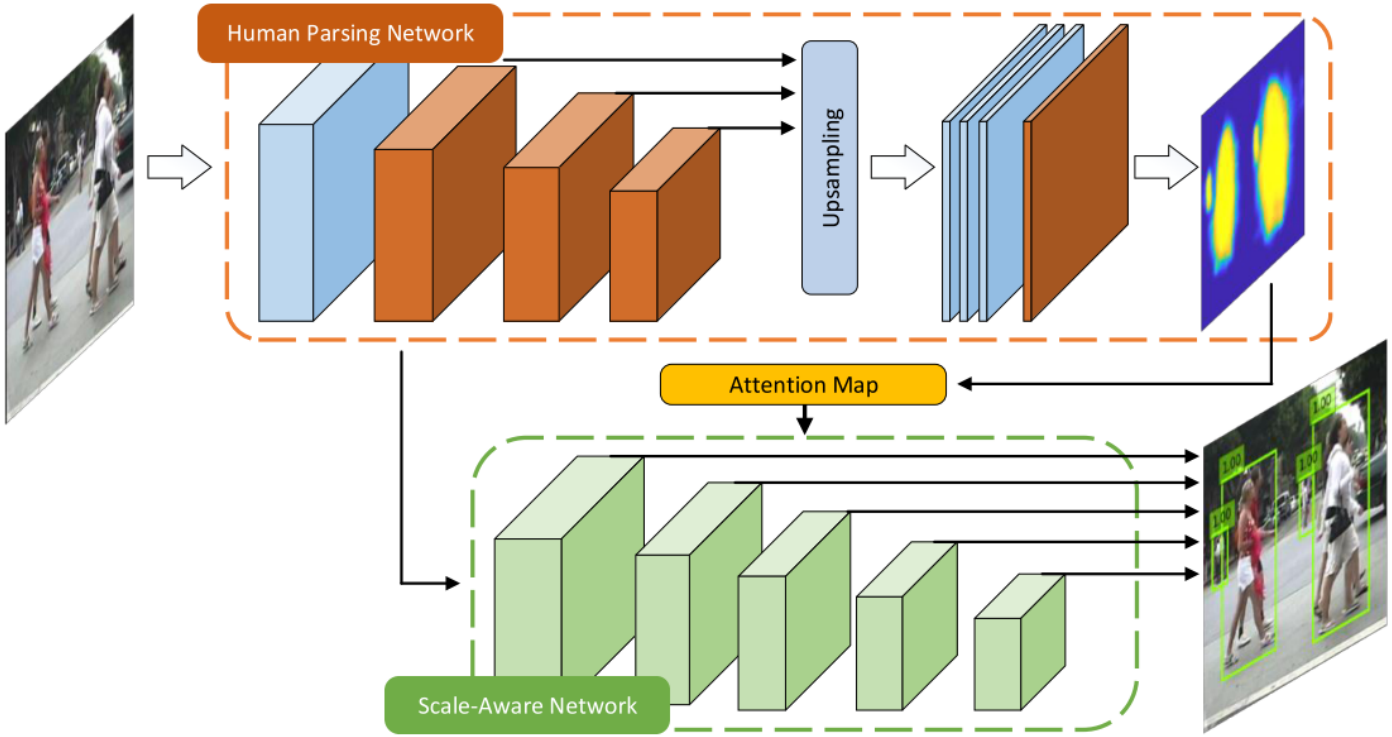
Motivation & Contributions

- Full-body and part-based methods: too coarse to localize small and occluded pedestrians

- **Fine-grained information** with pixel-wise classification to help detection

- **Multiple feature maps** of different resolutions to deal with scale variances issue

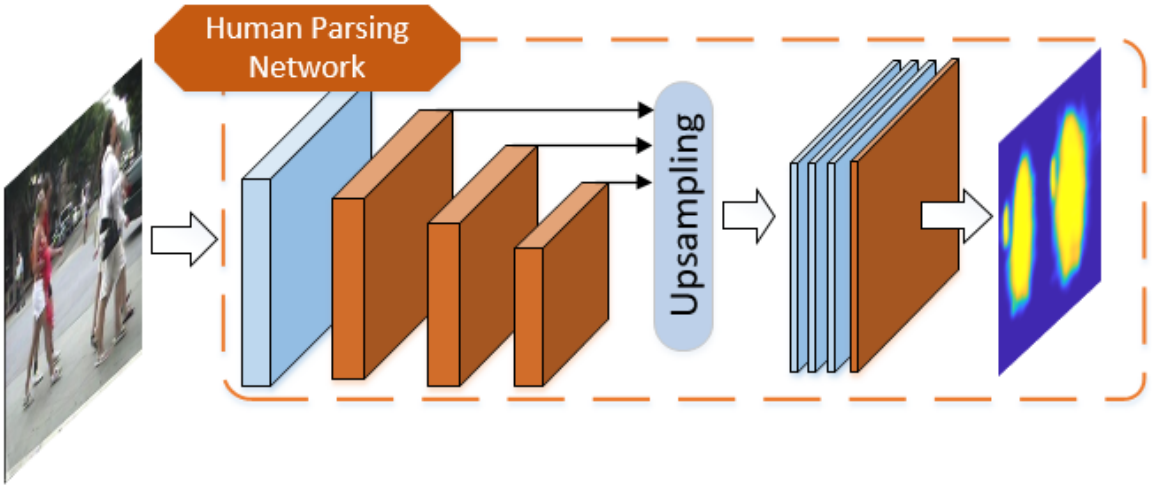
Flowchart



- Human Parsing Network
  - Generate human parsing mask and convert it into attention map
- Scale-Aware Network
  - Exploit intermediate feature maps for multiple scales detection
  - Attention map guides the detection to focus on pedestrians

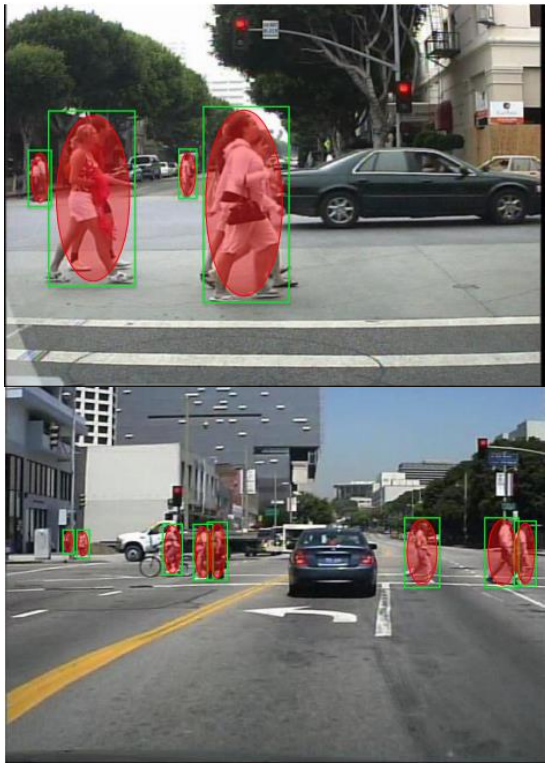
Human Parsing Network

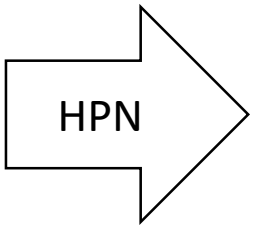
- Network Architecture
  - Truncated VGG16 with ‘atrous’ convolution
  - Deconvolution to up-sample to image size
  - Concatenate multiple layers to form hierarchical feature maps

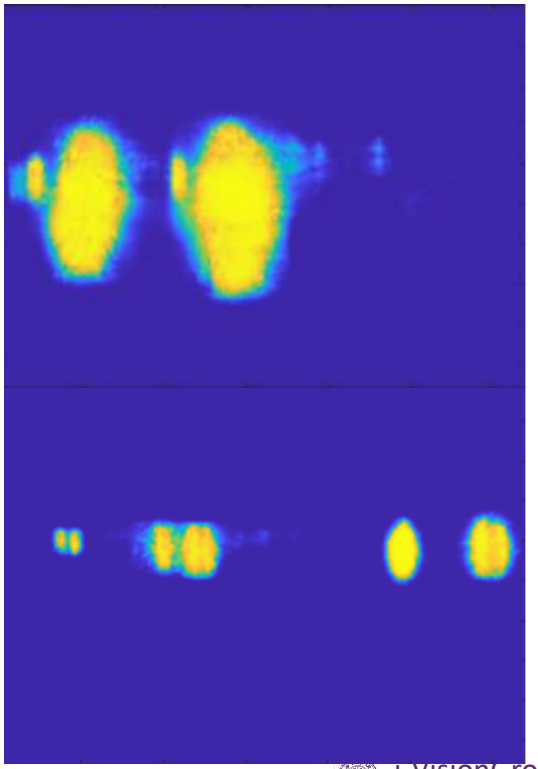


Human Parsing Network

- Weakly Supervised Training
  - Only bounding box annotations available
  - Consider 80% pixels at the center area of the bounding box as foreground
    - → Eliminate background noise
    - → Focus on main parts of human

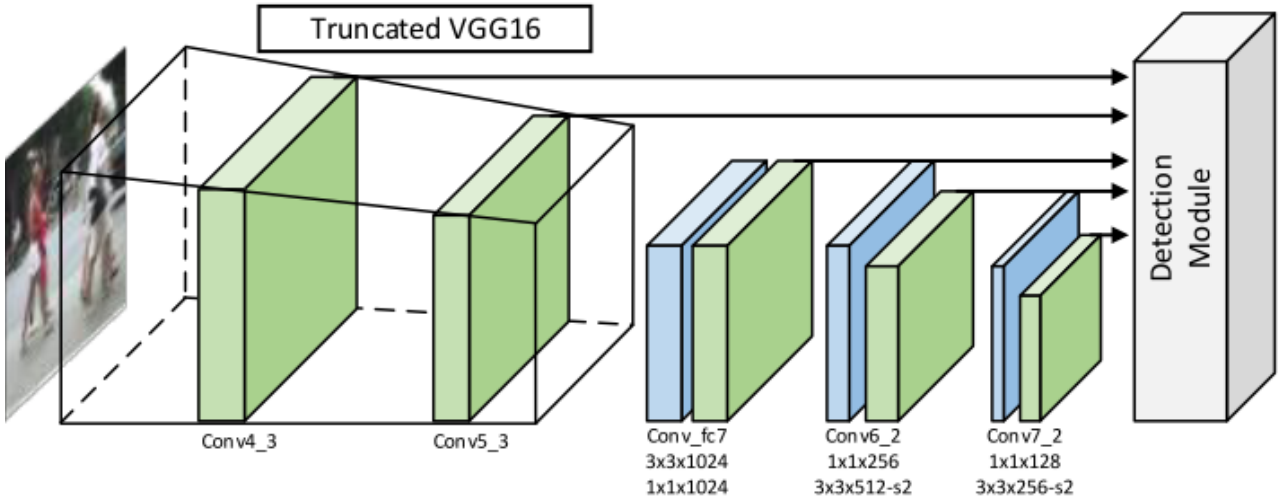






Scale-Aware Network

- Network Structure
  - Truncated VGG16 + extra convolutional layers
- Multiple scale detection
  - High resolution feature maps (shallower layers) for small targets detection
  - High-level semantic feature maps (deeper layers) for large pedestrians detection
  - Each detection layer followed by a detection module

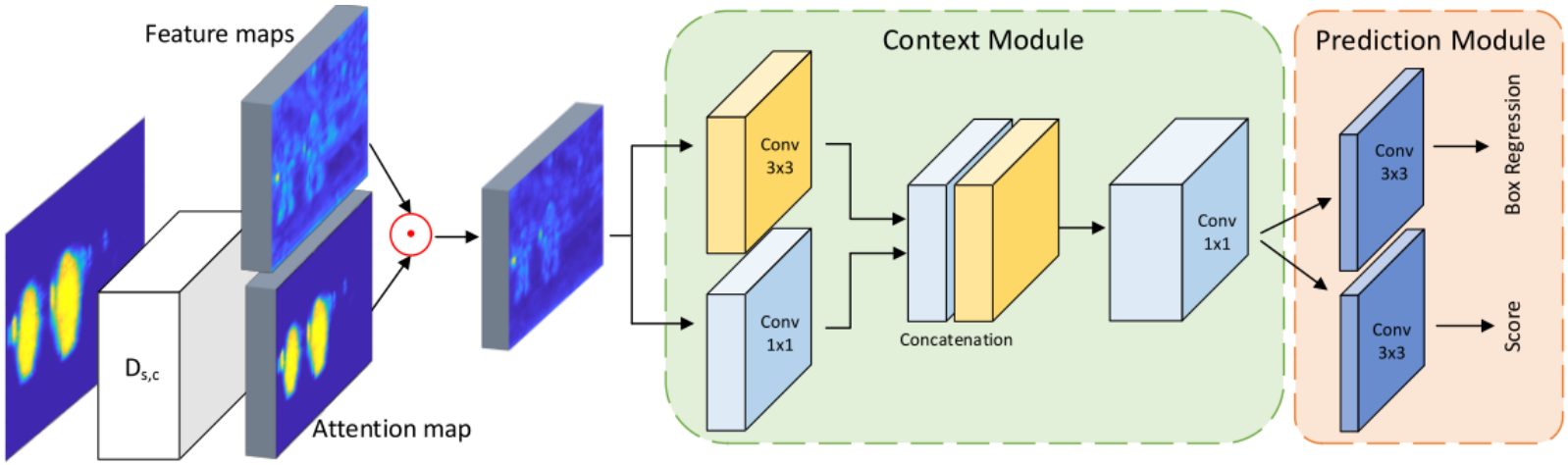


Scale-Aware Network

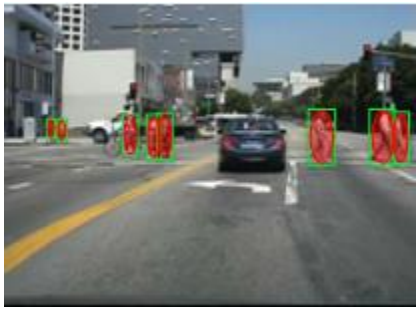
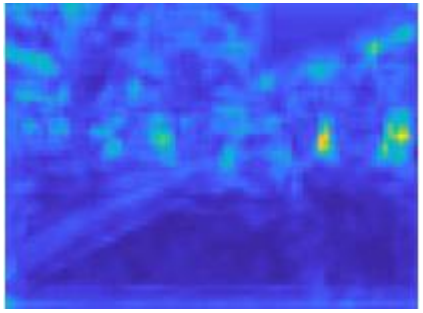
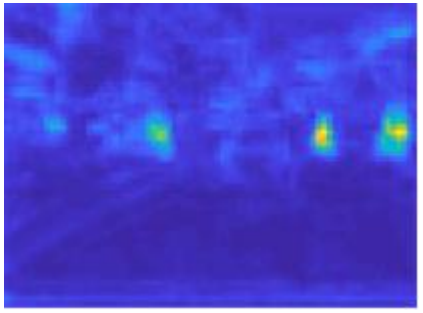
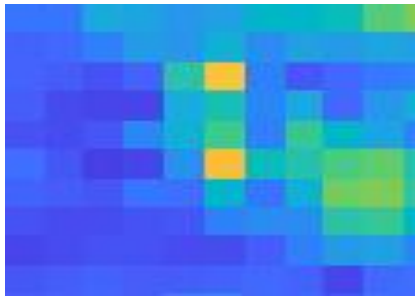
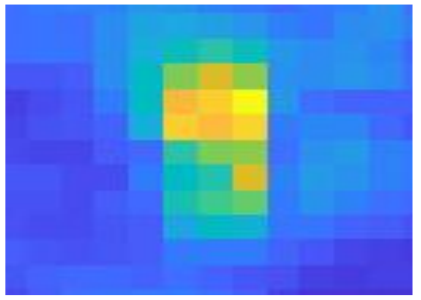
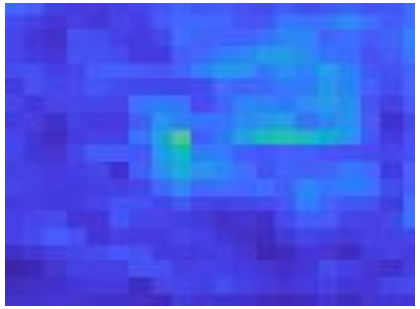
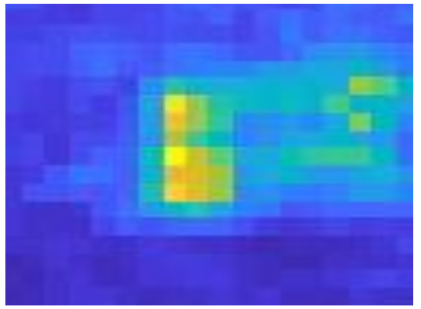
- Detection Module
  - Encode attention map into feature maps

$$\mathrm{A}_{s,c} = D_{s,c}(\mathrm{M}) \odot \mathrm{F}_{s,c}$$

  - Context module: concatenate 2 layers of different receptive fields
    → Incorporate more context information
  - Prediction module outputs the detection results



Visualization of feature maps

Image/Patch	Initial feature maps	Feature maps with attention
		
		
		

Optimization

- Implementation details
  1. Separately train Scale-Aware Network and Human Parsing Network
  2. Jointly optimize both networks

  → Facilitate the convergence

- Multi-task loss

$$L = \underbrace{L_{\text{box}} + \lambda_c L_{\text{conf}}}_{\text{SAN}} + \underbrace{\lambda_s L_{\text{seg}}}_{\text{HPN}}$$

Outline

① Introduction

② Related Work

③ Proposed Approach

④ Experimental Results

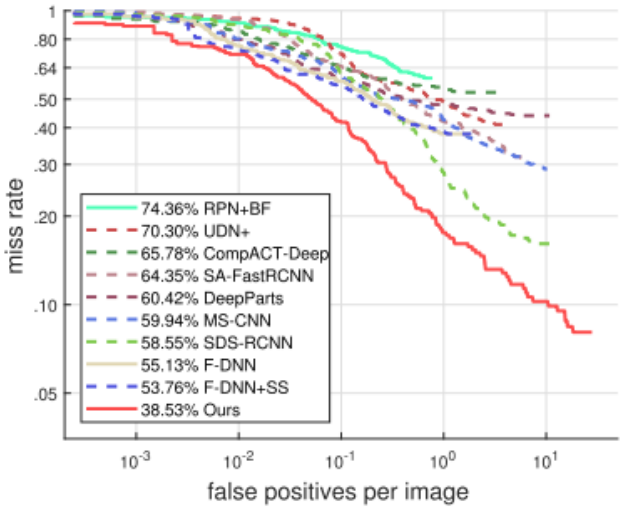
Experimental Results

- Caltech Pedestrian
  - 42,782 training images
  - 4,024 test images
  - Evaluation metric: average miss rate

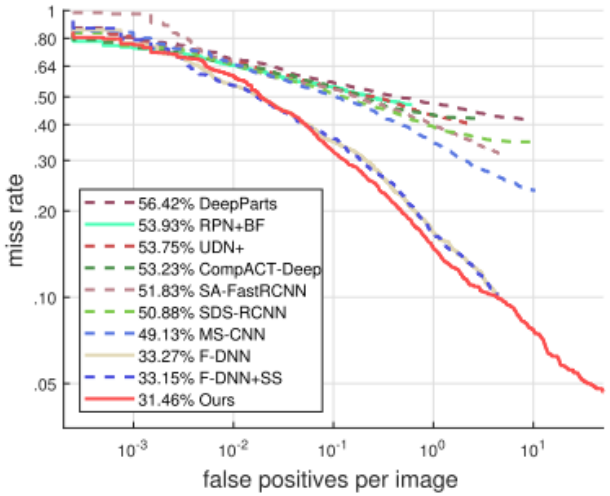
- KITTI
  - 7,481 training images
  - 7,518 test images
  - Evaluation metric: mean average precision (AP)

Experimental Results

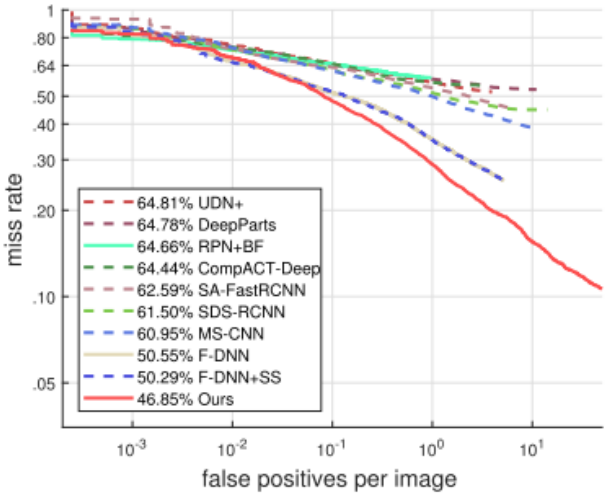
- Caltech Pedestrian
  - **Heavy occluded**: taller than 50 pixels, visibility $\in [0.36, 0.80]$
  - **Medium**: pedestrian height $\in [30, 80]$ pixels, reasonable visibility
  - **Overall**: all pedestrian taller than 20 pixels, with or non occlusion



(a) Heavy Occluded



(b) Medium



(c) Overall

Experimental Results

- KITTI
  - Moderate setting: pedestrian taller than 25 pixels with or non occlusion

- Computing Time
  - Real time pedestrian detector
  - Our method is at least 2x faster
  - Great trade-off of performance and runtime

Method	Caltech	KITTI	Runtime
RPN+BF [4]	74.36	61.29	0.5s
SA-FastRCNN [25]	64.35	65.01	0.5s
DeepParts [16]	60.42	58.67	1s
MS-CNN [5]	59.94	73.70	0.14s
SDS-RCNN [11]	58.55	63.05	0.21s
F-DNN[13]	55.13	-	0.3s
F-DNN+SS [13]	53.76	-	2.48s
JL-Tops [10]	49.20	-	0.6s
Ours	38.53	66.32	0.07s

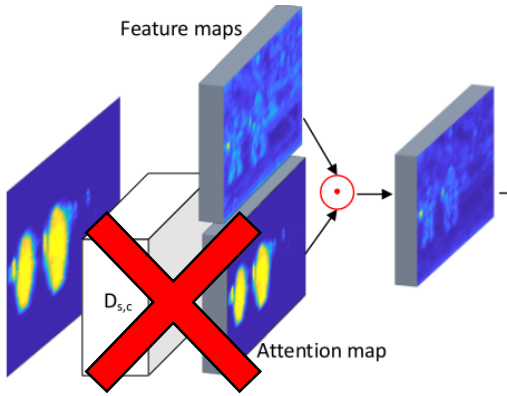
Experimental Results

- Ablation Studies

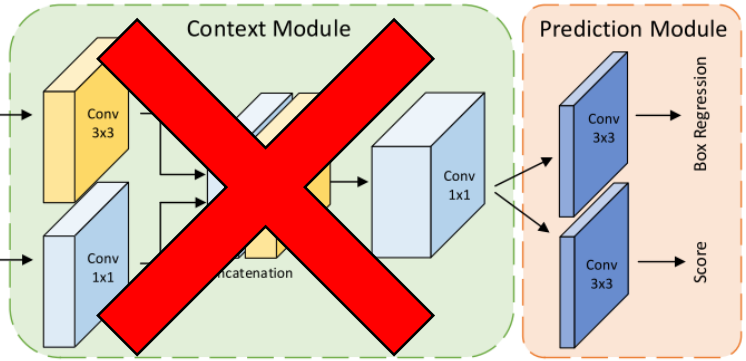
Disable main components successively

- Segmentation mask: performance drops by ~2%
- Context module: performance drops by ~2.5%

Component Disabled	Medium	Heavy	Overall
Context module	35.31	44.37	49.99
Segmentation mask	33.27	40.27	47.83
Our-MDFL	31.46	38.53	46.85



1. Segmentation mask



2. Context module

Conclusion and Future Works

- Fine Grained Attention map
  - Guide the detector to focus on pedestrians
  - Eliminate background interference

- Future works
  - Implement the proposed method into video based detector
  - Exploit temporal information

Thanks!