Illuminating Pedestrians via Simultaneous Detection and Segmentation

Garrick Brazil, Xi Yin, Xiaoming Liu
Computer Vision Lab, Michigan State University
{brazilga, yinxi1, liuxm}@msu.edu

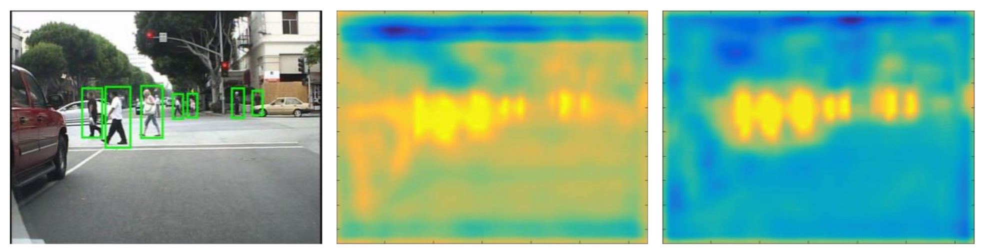
Problem & Contributions

Problem: Pedestrian detection from a single image

Semantic segmentation can be used to boost accuracy and efficiency of pedestrian detection without demanding additional data.

The main contributions are:

- Multi-task infusion framework for supervision on pedestrian detection and semantic segmentation, meant to infuse semantic features into shared layers
- Cascaded two-stage specialized networks
- Stage-wise fusion of classification scores



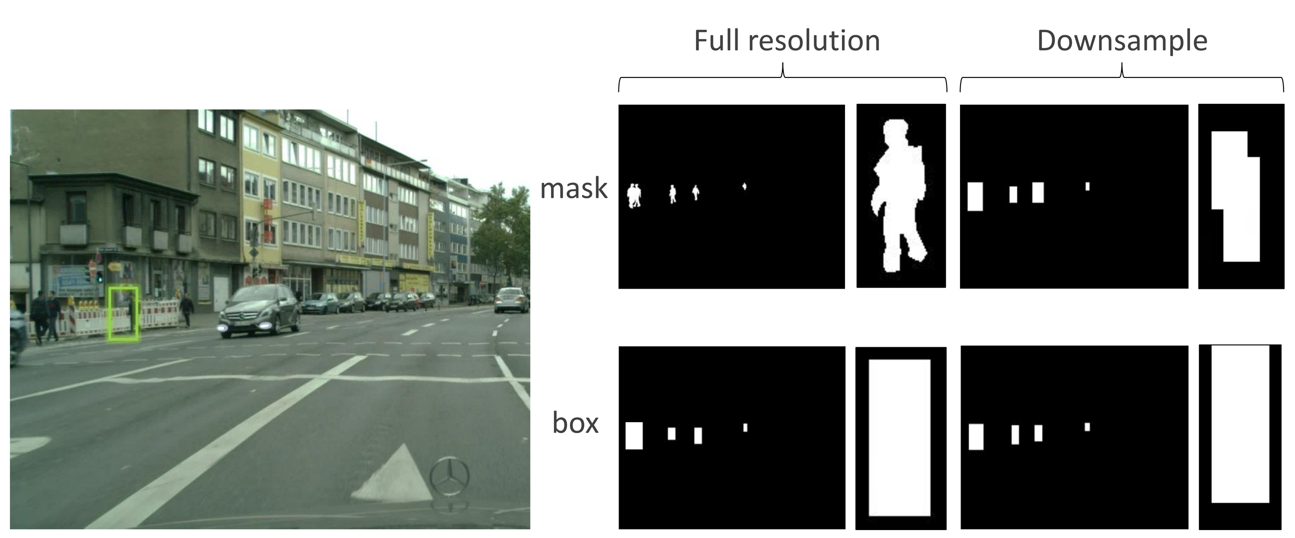
Test Image | Baseline Feature Map | SDS R-CNN Feature Map

Proposed Method

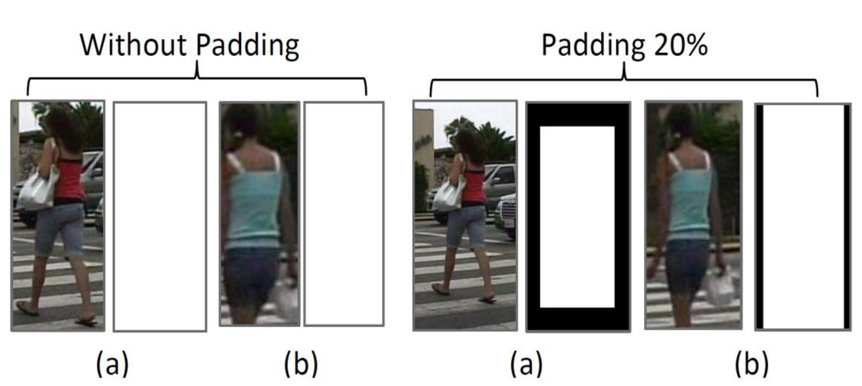
The method contains 2 primary stages trained separately. Each stage has an additional layer to supervise semantic segmentation using weakly annotated boxes.

- **Semantic Segmentation Layer**
  - Infuses semantic features of pedestrians into the shared layers (conv1-5).
  - Trained using weakly annotated boxes only.
- **Stage 1: Region Proposal Network (RPN)**
  - Sliding window detector across 9 anchors.
  - Labeling policy uses **lenient** IoU > 0.5 for high recall.
- **Stage 2: Binary Classification Network (BCN)**
  - Crops, pads, and warps RGB proposals to fixed size for a final classification.
  - Labeling policy differentiates from RPN by enforcing a **stricter** IoU > 0.7, thereby suppressing poorly localized boxes and achieving higher precision.

Weak Semantic Segmentation

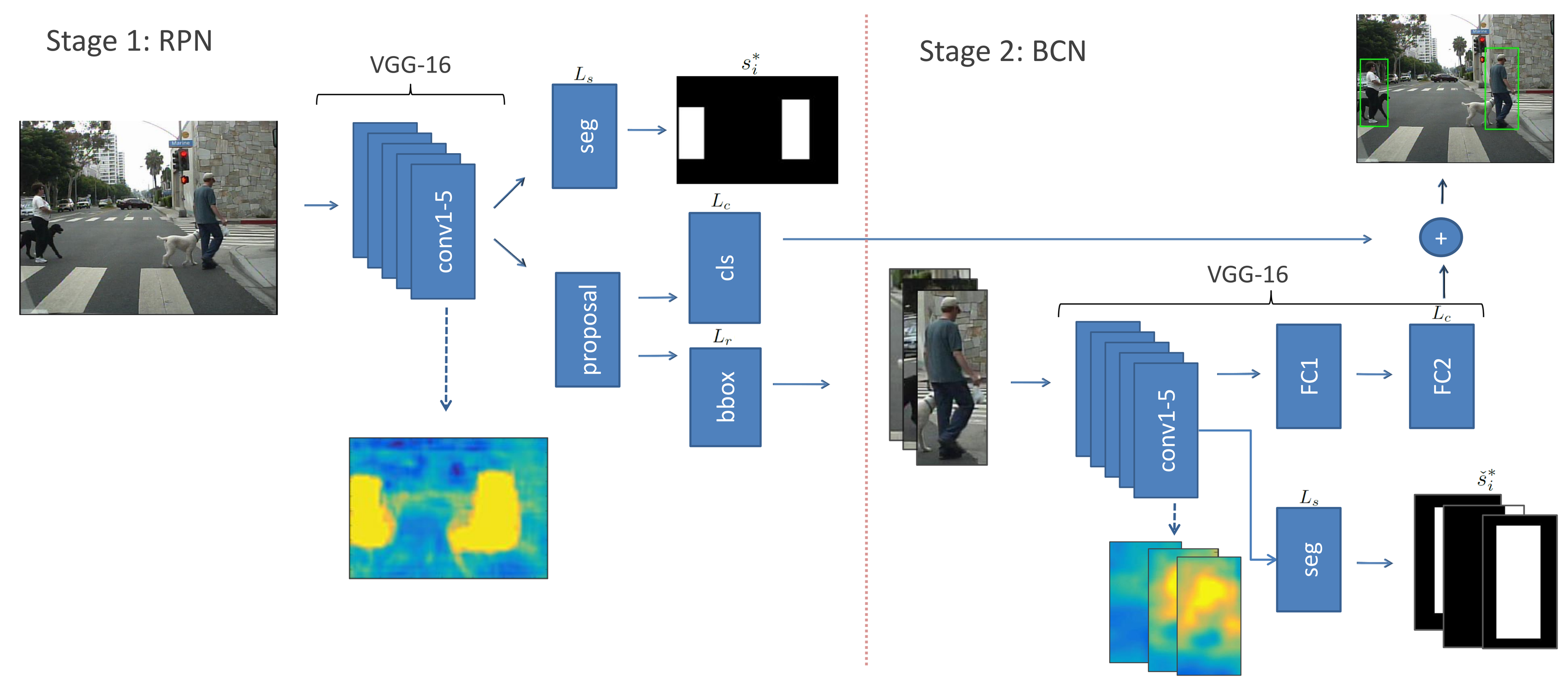


Visualization of the similarity between pixel-wise masks (Cityscapes dataset) and weak box annotations when down-sampled in our framework. By pooling, the differences between pixel-wise annotations and box annotations are negligible.



Example ground truth masks for the BCN with and without padding. Without padding there is no discernible difference between the ground truth masks of a well-localized proposal (a) and a poorly localized proposal (b).

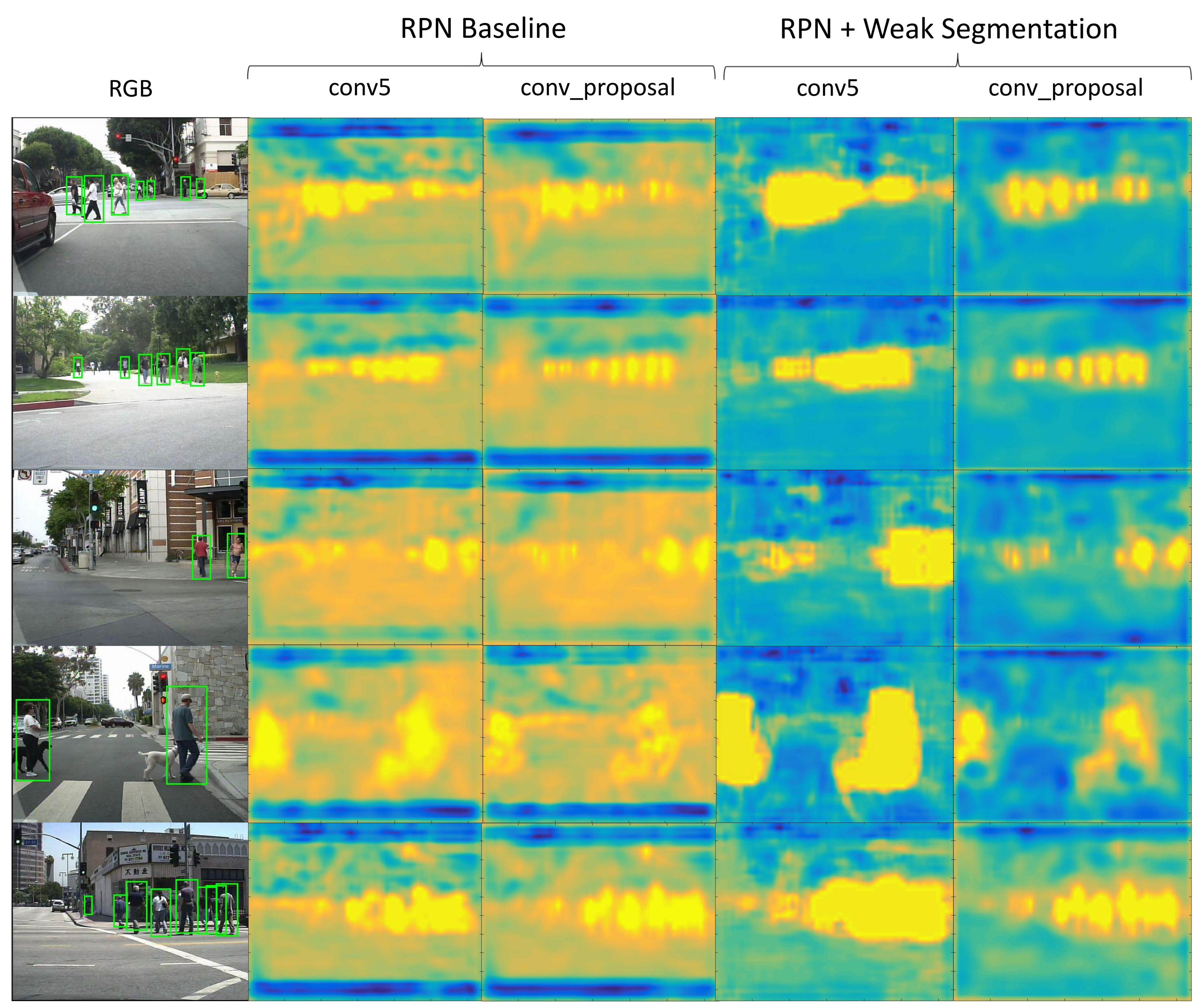
Network Architecture



$$L = \frac{1}{N_c}\sum_i L_c(p_i, p_i^*) + \frac{\lambda_1}{N_r}\sum_i L_r(t_i, t_i^*) + \frac{\lambda_2}{N_s}\sum_i L_s(s_i, s_i^*)$$

$$L = \frac{1}{N_c}\sum_i w_i L_c(\check{p}_i, \check{p}_i^*) + \frac{\lambda_1}{N_s}\sum_i w_i L_s(\check{s}_i, \check{s}_i^*)$$

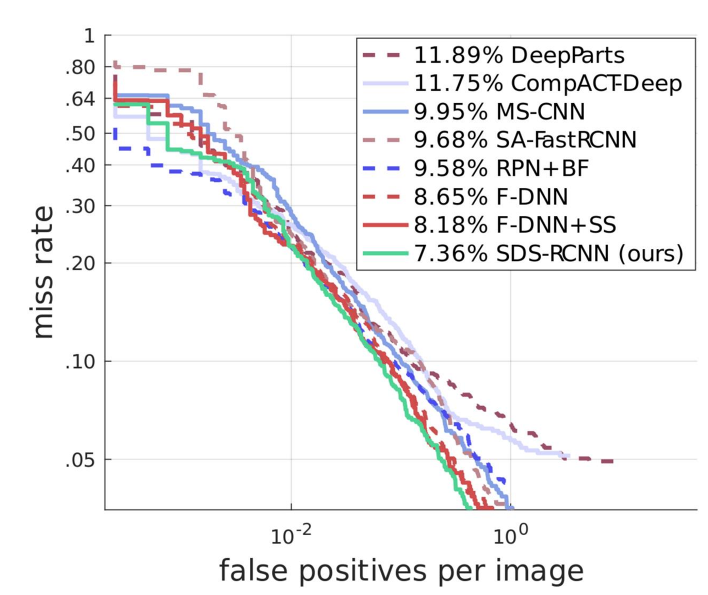
Experimental Results



Feature map visualizations of conv5 and the proposal layer for the baseline RPN (left) and the RPN infused with weak segmentation supervision (right). The weak semantic segmentation supervision helps *illuminate* pedestrians in the shared feature maps.

Method	Caltech	KITTI	Runtime
DeepParts	11.89	58.67	1s
CompACT-Deep	11.75	58.74	1s
MS-CNN	9.95	73.70	0.4s
SA-FastRCNN	9.68	65.01	0.59s
RPN+BF	9.58	61.29	0.60s
F-DNN	8.65	-	0.30s
F-DNN+SS	8.18	-	2.48s
SDS-RPN (ours)	9.63	-	0.13s
SDS-RCNN (ours)	7.36	63.05	0.21s

Comprehensive comparison with other state-of-the-art methods showing the Caltech miss rate, KITTI mAP score, and runtime.



Comparison of SDS-RCNN with the state-of-the-art methods on the Caltech dataset using the reasonable setting.

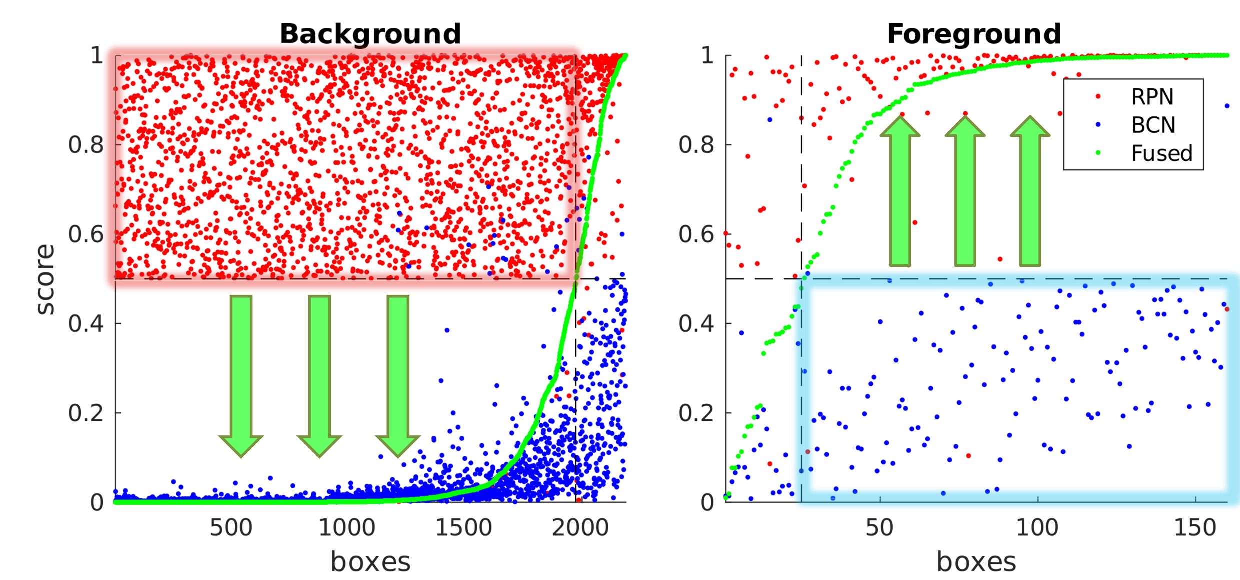
Ablation Study

Example error sources which are corrected by infusing semantic segmentation. The majority cases corrected are occlusion (48%), unusual poses (28%), misc (24%).

Component Disabled	RPN	BCN	Fusion
proposal padding	10.67	13.09	7.69
cost-sensitive	9.63	14.87	7.89
strict supervision	10.67	17.41	8.71
weak segmentation	13.84	18.76	10.41
SDS-RCNN	10.67	10.98	7.36

Ablation experiments evaluated using the Caltech test set. Each ablation experiment reports the miss rate for the RPN, BCN, and fused score with one component disabled.

Stage-wise Fusion



Visualization of the diversification between the RPN and BCN classification. We plot only boxes which the RPN and BCN of SDS-RCNN disagree on using a threshold of 0:5. The BCN drastically reduces false positives of the RPN, while the RPN corrects many missed detections by the BCN.

Shared Layer	BCN MR	Fused MR	Runtime
conv5	16.24	10.87	0.15s
conv4	15.53	10.42	0.16s
conv3	14.28	8.66	0.18s
conv2	13.71	8.33	0.21s
conv1	14.02	8.28	0.25s
RGB	10.98	7.36	0.21s

Stage-wise sharing experiments which demonstrate the trade-off of runtime efficiency and accuracy, using the Caltech dataset. As sharing is increased from RGB (no sharing) to conv5, both the BCN and Fused miss rate (MR) become less effective.

Conclusions

- Multi-task infusion of semantic segmentation features helps to illuminate pedestrians in shared feature maps, making downstream classification easier and more robust to pose and occlusion.
- Collectively semantic segmentation supervision, stage-wise fusion, and stricter supervision result in a **23% relative reduction of error** and executes **2x faster** than competitive methods.

project website + source code