This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Illuminating Pedestrians via Simultaneous Detection & Segmentation

Garrick Brazil, Xi Yin, Xiaoming Liu
Michigan State University, East Lansing, MI 48824
{brazilga, yinxi1, liuxm}@msu.edu

Abstract

Pedestrian detection is a critical problem in computer vision with significant impact on safety in urban autonomous driving. In this work, we explore how semantic segmentation can be used to boost pedestrian detection accuracy while having little to no impact on network efficiency. We propose a segmentation infusion network to enable joint supervision on semantic segmentation and pedestrian detection. When placed properly, the additional supervision helps guide features in shared layers to become more sophisticated and helpful for the downstream pedestrian detector. Using this approach, we find weakly annotated boxes to be sufficient for considerable performance gains. We provide an in-depth analysis to demonstrate how shared layers are shaped by the segmentation supervision. In doing so, we show that the resulting feature maps become more semantically meaningful and robust to shape and occlusion. Overall, our simultaneous detection and segmentation framework achieves a considerable gain over the state-of-the-art on the Caltech pedestrian dataset, competitive performance on KITTI, and executes $2\times$ *faster than competitive methods.*

1. Introduction

Pedestrian detection from an image is a core capability of computer vision, due to its applications such as autonomous driving and robotics [14]. It is also a long-standing vision problem because of its distinct challenges including low resolution, occlusion, cloth variations, etc [30]. There are two central approaches for detecting pedestrians: object detection [2, 29] and semantic segmentation [4, 5]. The two approaches are highly related by nature but have their own strengths and weaknesses. For instance, object detection is designed to perform well at localizing distinct objects but typically provides little information on object boundaries. In contrast, semantic segmentation does well at distinguishing pixel-wise boundaries among classes but struggles to separate objects within the same class.

Intuitively, we expect that knowledge from either task will make the other substantially easier. This has been demonstrated for generic object detection, since having segmentation masks of objects would clearly facilitate detection. For example, Fidler et al. [13] utilize predicted segmentation masks to boost object detection performance via a deformable part-based model. Hariharan et al. [18] show how segmentation masks generated from MCG [1] can be used to mask background regions and thus simplify detection. Dai et al. [6] utilize the two tasks in a 3-stage cascaded network consisting of box regression, foreground segmentation, and classification. Their architecture allows each task to share features and feed into one another.

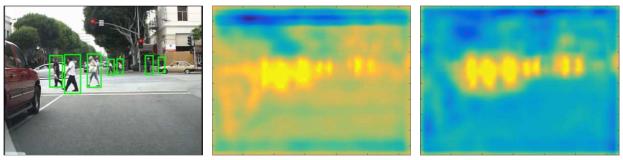

Figure 1. Detection results on the Caltech test set (left), feature map visualization from the RPN of conventional Faster R-CNN (middle), and feature map visualization of SDS-RCNN (right). Notice that our feature map substantially illuminates the pedestrian shape while suppressing the background region, both of which make positive impact to downstream pedestrian detection.

In contrast, the pairing of these two tasks is rarely studied in pedestrian detection, despite the recent advances [2, 21, 29]. This is due in part to the lack of pixel-wise annotations available in classic pedestrian datasets such as Caltech [8] and KITTI [14], unlike the detailed segmentation labels in the COCO [22] dataset for generic object detection. With the release of Cityscapes [5], a high quality dataset for urban semantic segmentation, it is expected that substantial research efforts will be on *how to leverage semantic segmentation to boost the performance of pedestrian detection*, which is the core problem to be studied in this paper.

Given this objective, we start by presenting a competitive two-stage baseline framework of pedestrian detection deriving from RPN+BF [29] and Faster R-CNN [23]. We contribute a number of key changes to enable the second-stage classifier to specialize in stricter supervision and additionally fuse the refined scores with the first stage RPN. These changes alone lead to state-of-the-art performance on

the Caltech benchmark. We further present a simple, but surprisingly powerful, scheme to utilize multi-task learning on pedestrian detection and semantic segmentation. Specifically, we infuse the semantic segmentation mask into shared layers using a *segmentation infusion layer* in both stages of our network. We term our approach as "simultaneous detection and segmentation R-CNN (SDS-RCNN)". We provide an in-depth analysis on the effects of joint training by examining the shared feature maps, e.g., Fig. 1. Through infusion, the shared feature maps begin to illuminate pedestrian regions. Further, since we infuse the semantic features during training only, the network efficiency at inference is unaffected. We demonstrate the effectiveness of SDS-RCNN by reporting considerable improvement (23% relative reduction of the error) over the published state-of-the-art on Caltech [8], competitive performance on KITTI [14], and a runtime roughly $2\times$ faster than competitive methods.

In summary our contributions are as follows:

- ◇ Improved baseline derived from [23, 29] by enforcing stricter supervision in the second-stage classification network, and further fusing scores between stages.
- ◇ A multi-task infusion framework for joint supervision on pedestrian detection and semantic segmentation, with the goal of *illuminating* pedestrians in shared feature maps and easing downstream classification.
- ◇ We achieve the new state-of-the-art performance on Caltech pedestrian dataset, competitive performance on KITTI, and obtain $2\times$ faster runtime.

2. Prior work

Object Detection: Deep convolution neural networks have had extensive success in the domain of object detection. Notably, derivations of Fast [16] and Faster R-CNN [23] are widely used in both generic object detection [2, 15, 28] and pedestrian detection [21, 26, 29]. Faster R-CNN consists of two key components: a region proposal network (RPN) and a classification sub-network. The RPN works as a sliding window detector by determining the *objectness* across a set of predefined anchors (box shapes defined by aspect ratio and scale) at each spatial location of an image. After object proposals are generated, the second stage classifier determines the precise class each object belongs to. Faster R-CNN has been shown to reach state-of-the-art performance on the PASCAL VOC 2012 [12] dataset for generic object detection and continues to serve as a frequent baseline framework for a variety of related problems [15, 18, 19, 30].

Pedestrian Detection: Pedestrian detection is one of the most extensively studied problems in object detection due to its real-world significance. The most notable challenges are caused by small scale, pose variations, cyclists, and occlusion [30]. For instance, in the Caltech pedestrian dataset [8] 70% of pedestrians are occluded in at least one frame.

The top performing approaches on the Caltech pedestrian benchmark are variations of Fast or Faster R-CNN. SA-FastRCNN [16] and MS-CNN [2] reach competitive performance by directly addressing the scale problem using specialized multi-scale networks integrated into Fast and Faster R-CNN respectively. Furthermore, RPN+BF [29] shows that the RPN of Faster R-CNN performs well as a standalone detector while the downstream classifier degrades performance due to collapsing bins of small-scale pedestrians. By using higher resolution features and replacing the downstream classifier with a boosted forest, RPN+BF is able to alleviate the problem and achieve 9.58% miss rate on the Caltech reasonable [9] setting. F-DNN [10] also uses a derivation of the Faster R-CNN framework. Rather then using a single downstream classifier, F-DNN fuses multiple parallel classifiers including ResNet [19] and GoogLeNet [25] using soft-reject and further incorporates multiple training datasets to achieve 8.65% miss rate on the Caltech reasonable setting. The majority of top performing approaches utilize some form of a RPN, whose scores are typically discarded after selecting the proposals. In contrast, our work shows that fusing the score with the second stage network can lead to substantial performance improvement.

Simultaneous Detection & Segmentation: There are two lines of research on simultaneous detection and segmentation. The first aims to improve the performance of both tasks, and formulates a problem commonly known as *instance-aware semantic segmentation* [5]. Hariharan et al. [18] predict segmentation masks using MCG [1] then get object instances using "slow" R-CNN [17] on masked image proposals. Dai et al. [6] achieve high performance on instance segmentation using an extension of Faster R-CNN in a 3-stage cascaded network including mask supervision.

The second aims to explicitly improve object detection by using segmentation as a strong cue. Early work on the topic by Fidler et al. [13] demonstrates how semantic segmentation masks can be used to extract strong features for improved object detection via a deformable part-based model. Du et al. [10] use segmentation as a strong cue in their F-DNN+SS framework. Given the segmentation mask predicted by a third parallel network, their ensemble network uses the mask in a post-processing manner to suppress background proposals, and pushes performance on the Caltech pedestrian dataset from 8.65% to 8.18% miss rate. However, the segmentation network degrades the efficiency of F-DNN+SS from 0.30 to 2.48 seconds per image, and requires multiple GPUs at inference. In contrast, our novel framework infuses the semantic segmentation masks into shared feature maps and thus does not require a separate segmentation network, which outperforms [10] in both accuracy and network efficiency. Furthermore, our use of weak box-based segmentation masks addresses the issue of lacking pixel-wise segmentation annotations in [8, 14].

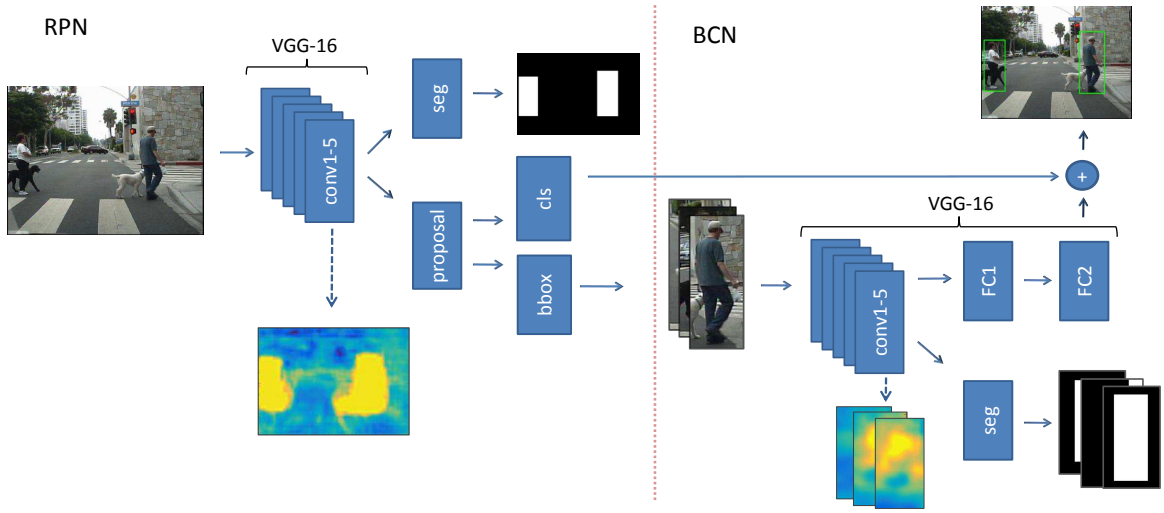

Figure 2. Overview of the proposed SDS-RCNN framework. The segmentation layer infuses semantic features into shared conv1-5 layers of each stage, thus *illuminating* pedestrians and easing downstream pedestrian detection (proposal layers in RPN, and FC1-2 in BCN).

3. Proposed method

Our proposed architecture consists of two key stages: a region proposal network (RPN) to generate candidate bounding boxes and corresponding scores, and a binary classification network (BCN) to refine their scores. In both stages, we propose a *semantic segmentation infusion layer* with the objective of making downstream classification a substantially easier task. The infusion layer aims to encode semantic masks into shared feature maps which naturally serve as strong cues for pedestrian classification. Due to the impressive performance of the RPN as a standalone detector, we elect to fuse the scores between stages rather than discarding them as done in prior work [2, 10, 27, 29]. An overview of the SDS-RCNN framework is depicted in Fig. 2

3.1. Region Proposal Network

The RPN aims to propose a set of bounding boxes with associated confidence scores around potential pedestrians. We adopt the RPN of Faster R-CNN [23] following the settings in [29]. We tailor the RPN for pedestrain detection by configuring $N_a = 9$ anchors with a fixed aspect ratio of 0.41 and spanning a scale range from $25 - 350$ pixels, corresponding to the pedestrain statistics of Caltech [8]. Since each anchor box acts as a sliding window detector across a pooled image space, there are $N_p = N_a \times \frac{W}{f_s} \times \frac{H}{f_s}$ total pedestrian proposals, where f_s corresponds to the feature stride of the network. Hence, each proposal box i corresponds to an anchor and a spatial location of image $\mathbf{I}$.

The RPN architecture uses conv1-5 from VGG-16 [24] as the backbone. Following [23], we attach a proposal feature extraction layer to the end of the network with two sibling output layers for box classification (*cls*) and bounding box regression (*bbox*). We further add a segmentation infusion layer to conv5 as detailed in Sec. 3.3.

For every proposal box i, the RPN aims to minimize the following joint loss function with three terms:

$$L = \lambda_c \sum_i L_c(c_i, \hat{c}_i) + \lambda_r \sum_i L_r(t_i, \hat{t}_i) + \lambda_s L_s. \quad (1)$$

The first term is the classification loss L_c, which is a softmax logistic loss over two classes (pedestrian vs. background). We use the standard labeling policy which considers a proposal box at location i to be pedestrian ($c_i = 1$) if it has at least 0.5 Intersection over Union (IoU) with a ground truth pedestrian box, and otherwise background ($c_i = 0$). The second term seeks to improve localization via bounding box regression, which learns a transformation for each proposal box to the nearest pedestrian ground truth. Specifically, we use $L_r(t_i, \hat{t}_i) = R(t_i - \hat{t}_i)$ where R is the robust (smooth L_1) loss defined in [16]. The bounding box transformation is defined as a 4-tuple consisting of shifts in x, y and scales in w, h denoted as $t = [t_x, t_y, t_w, t_h]$. The third term L_s is the segmentation loss presented in Sec. 3.3.

In order to reduce multiple detections of the same pedestrian, we apply non-maximum suppression (NMS) greedily to all pairs of proposals after the transformations have been applied. We use an IoU threshold of 0.5 for NMS.

We train the RPN in the Caffe [20] framework using SGD with a learning rate of 0.001, momentum of 0.9, and mini-batch of 1 full-image. During training, we randomly sample 120 proposals per image at a ratio of 1:5 for pedestrian and background proposals to help alleviate the class imbalance. All other proposals are treated as ignore. We initialize conv1-5 from a VGG-16 model pretrained on ImageNet [7], and all remaining layers randomly. Our network has four max-pooling layers (within conv1-5), hence $f_s = 16$. In our experiments, we regularize our multi-task loss terms by setting $\lambda_c = \lambda_s = 1, \lambda_r = 5$.

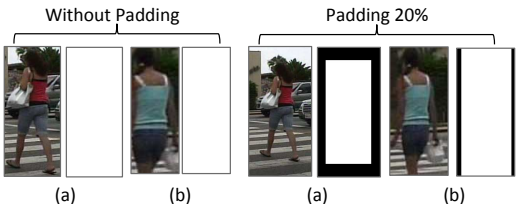

Figure 3. Example proposal masks with and without padding. There is no discernible difference between the non-padded masks of well-localized (a) and poorly localized (b) proposals.

3.2. Binary Classification Network

The BCN aims to perform pedestrian classification over the proposals of the RPN. For generic object detection, the BCN usually uses the downstream classifier of Faster R-CNN by sharing conv1-5 with the RPN, but was shown by [29] to degrade pedestrian detection accuracy. Thus, we choose to construct a separate network using VGG-16. The primary advantage of a separate network is to allow the BCN freedom to specialize in the types of "harder" samples left over from the RPN. While sharing computation is highly desirable for the sake of efficiency, the shared networks are more predestined to predict similar scores which are redundant when fused. Therefore, rather than cropping and warping a shared feature space, our BCN directly crops the top N_b proposals from the RGB input image.

For each proposal image i, the BCN aims to minimize the following joint loss function with two terms:

$$L = \lambda_c \sum_i w_i L_c(c_i, \hat{c}_i) + \lambda_s L_s. \tag{2}$$

Similar to RPN, the first term is the classification loss L_c where c_i is the class label for the ith proposal. A cost-sensitive weight w_i is used to give precedence to detect large pedestrians over small pedestrians. There are two key motivations for this weighting policy. First, large pedestrians typically imply close proximity and are thus significantly more important to detect. Secondly, we presume that features of large pedestrians may be more helpful for detecting small pedestrians. We define the weighting function given the ith proposal with height h_i and a pre-computed mean height $\bar{h}$ as $w_i = 1 + \frac{h_i}{\bar{h}}$. The second term is the segmentation loss presented in Sec. 3.3.

We make a number of significant contributions to the BCN. First, we change the labeling policy to encourage higher precision and further diversification from the RPN. We enforce a *stricter* labeling policy, requiring a proposal to have IoU > 0.7 with a ground truth pedestrian box to be considered pedestrian ($c_i = 1$), and otherwise background ($c_i = 0$). This encourages the network to suppress poorly localized proposals and reduces false positives in the form of double detections. Secondly, we choose to fuse the scores of the BCN with the confidence scores of the RPN at test time. Since our design explicitly encourages the two stages

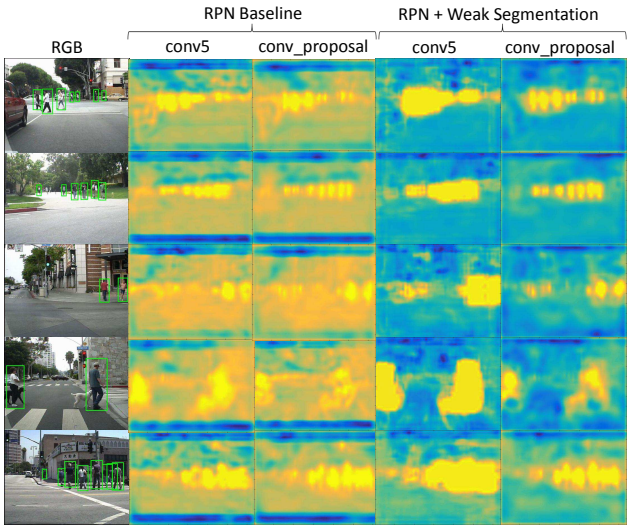

Figure 4. Feature map visualizations of conv5 and the proposal layer for the baseline RPN (left) and the RPN infused with weak segmentation supervision (right).

to diversify, we expect the classification characteristics of each network to be complementary when fused. We fuse the scores at the feature level prior to softmax. Formally, the fused score for the ith proposal, given the predicted 2-class scores from the RPN $= \{\hat{c}^r_{i0}, \hat{c}^r_{i1}\}$ and BCN $= \{\hat{c}^b_{i0}, \hat{c}^b_{i1}\}$ is computed via the following softmax function:

$$\hat{c}_i = \frac{e^{(\hat{c}^r_{i1} + \hat{c}^b_{i1})}}{e^{(\hat{c}^r_{i1} + \hat{c}^b_{i1})} + e^{(\hat{c}^r_{i0} + \hat{c}^b_{i0})}}. \tag{3}$$

In effect, the fused scores become more confident when the stages agree, and otherwise lean towards the dominant score. Thus, it is ideal for each network to diversify in its classification capabilities such that at least one network may be *very* confident for each proposal.

For a modest improvement to efficiency, we remove the pool5 layer from the VGG-16 architecture then adjust the input size to 112×112 to keep the fully-connected layers intact. This is a fair trade-off since most pedestrian heights fall in the range of $30 - 80$ pixels [8]. Hence, small pedestrian proposals are upscaled by a factor of $\sim 2\times$, allowing space for finer discrimination. We further propose to pad each proposal by 20% on all sides to provide background context and avoid partial detections, as shown in Fig. 3.

We train the BCN in the Caffe [20] framework using the same settings as the RPN. We initialize conv1-5 from the trained RPN model, and all remaining layers randomly. During training, we set $N_b = 20$. During inference, we set $N_b = 15$ for a moderate improvement to efficiency. We regularize the multi-task loss by setting $\lambda_c = \lambda_s = 1$.

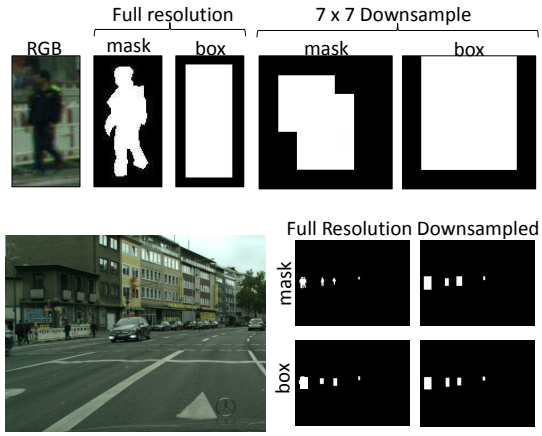

Figure 5. Visualization of the similarity between pixel-wise segmentation masks (from Cityscapes [5]) and weak box-based masks when downsampled in both the BCN (top) and RPN (bottom).

3.3. Simultaneous Detection & Segmentation

We approach simultaneous detection and segmentation with the motivation to make our downstream pedestrian detection task easier. We propose a segmentation infusion layer trained on weakly annotated pedestrian boxes which *illuminate* pedestrians in the shared feature maps preceding the classification layers. We integrate the infusion layer into both stages of our SDS-RCNN framework.

Segmentation Infusion Layer: The segmentation infusion layer aims to output two masks indicating the likelihood of residing on pedestrian or background segments. We choose to use only a single layer and a 1×1 kernel so the impact on the shared layers will be as high as possible. This forces the network to directly infuse semantic features into shared feature maps, as visualized in Fig. 4. A deeper network could achieve higher segmentation accuracy but will infer less from shared layers and diminish the overall impact on the downstream pedestrian classification. Further, we choose to attach the infusion layer to conv5 since it is the deepest layer which precedes both the proposal layers of the RPN and the fully connected layers of the BCN.

Formally, the final loss term L_s of both the RPN and BCN is a softmax logistic loss over two classes (pedestrian vs. background), applied to each location i, where w_i is the cost-sensitive weight introduced in 3.2:

$$\lambda_s \sum_i w_i L_s(\mathbf{S}_i, \hat{\mathbf{S}}_i). \tag{4}$$

We choose to levereage the abundance of bounding box annotations available in popular pedestrian datasets (e.g., Caltech [8], KITTI [14]) by forming weak segmentation ground truth masks. Each mask $\mathbf{S} \in \mathbb{R}^{\mathbb{W} \times \mathbb{H}}$ is generated by labeling all pedestrian box regions as $\mathbf{S}_i = 1$, and otherwise background $\mathbf{S}_i = 0$. In most cases, box-based annotations would be considered too noisy for semantic segmentation. However, since we place the infusion layer at conv5, which has been pooled significantly, the differences between box-based annotations and pixel-wise annotations diminish rapidly w.r.t. the pedestrian height (Fig. 5). For example, in the Caltech dataset 68% of pedestrians are less than 80 pixels tall, which corresponds to 3×5 pixels at conv5 of the RPN. Further, each of the BCN proposals are pooled to 7×7 at conv5. Hence, pixel-wise annotations may not offer a significant advantage over boxes at the high levels of pooling our networks undertake.

Benefits Over Detection: A significant advantage of segmentation supervision over detection is its simplicity. For detection, sensitive hyperparamters must be set, such as anchor selection and IoU thresholds used for labeling and NMS. If the chosen anchor scales are too sparse or the IoU threshold is too high, certain ground truths that fall near the midpoint of two anchors could be missed or receive low supervision. In contrast, semantic segmentation treats all ground truths indiscriminate of how well the pedestrian's shape or occlusion-level matches the chosen set of anchors. In theory, the incorporation of semantic segmentation infusion may help reduce the *sensitivity* of conv1-5 to such hyperparamters. Furthermore, the segmentation supervision is especially beneficial for the second stage BCN, which on its own would only know *if* a pedestrian is present. The infusion of semantic segmentation features inform the BCN *where* the pedestrian is, which is critical for differentiating poorly vs. well-localized proposals.

4. Experiments

We evaluate our proposed SDS-RCNN on popular datasets including Caltech [8] and KITTI [14]. We perform comprehensive analysis and ablation experiments using the Caltech dataset. We refer to our collective method as SDS-RCNN and our region proposal network as SDS-RPN. We show the performance curves compared to the state-of-the-art pedestrian detectors on Caltech in Fig. 6. We further report a comprehensive overview across datasets in Table 1.

4.1. Benchmark Comparison

Caltech: The Caltech dataset [8] contains $\sim$350K pedestrian bounding box annotations across 10 hours of urban driving. The log average miss rate sampled against a false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ is used for measuring performance. A minimum IoU threshold of 0.5 is required for a detected box to match with a ground truth box. For training, we sample from the standard training set according to Caltech10$\times$[31], which contains $42{,}782$ training images. We evaluate on the standard $4{,}024$ images in the Caltech $1\times$ test set using the *reasonable* [9] setting, which only considers pedestrians with at least 50 pixels in height and with less than 35% occlusion.

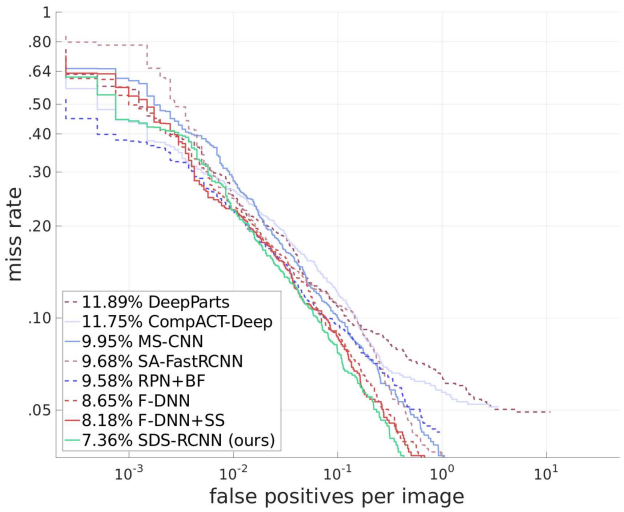

Figure 6. Comparison of SDS-RCNN with the state-of-the-art methods on the Caltech dataset using the *reasonable* setting.

SDS-RCNN achieves an impressive 7.36% miss rate. The performance gain is a relative improvement of 23% compared to the best published method RPN+BF (9.58%). In Fig. 6, we show the ROC plot of miss rate against FPPI for the current top performing methods reported on Caltech.

We further report our performance using just SDS-RPN (without cost-sensitive weighting, Sec. 4.2) on Caltech as shown in Table 1. The RPN performs quite well by itself, reaching 9.63% miss rate while processing images at roughly 3× the speed of competitive methods. Our RPN is already on par with other top detectors, which themselves contain a RPN. Moreover, the network significantly outperforms other standalone RPNs such as in [29] (14.9%). Hence, the RPN can be leveraged by other researchers to build better detectors in the future.

KITTI: The KITTI dataset [14] contains ∼80K annotations of cars, pedestrians, and cyclists. Since our focus is on pedestrian detection, we continue to use only the pedestrian class for training and evaluation. The mean Average Precision (mAP) [11] sampled across a recall range of $[0, 1]$ is used to measure performance. We use the standard training set of 7,481 images and evaluate on the designated test set of 7,518 images. Our method reaches a score of 63.05 mAP on the moderate setting for the pedestrian class. Surprisingly, we observe that many models which perform well on Caltech do not generalize well to KITTI, as detailed in Table 1. We expect this is due to both sensitivity to hyperparameters and the smaller training set of KITTI (∼6× smaller than Caltech10×). MS-CNN [2] is the current top performing method for pedestrian detection on KITTI. Aside from the novelty as a multi-scale object detector, MS-CNN augments the KITTI dataset by random cropping and scaling. Thus, incorporating data augmentation could alleviate the smaller training set and lead to better generalization across datasets. Furthermore, as described in the ablation study of Sec. 4.2, our weak segmentation supervision primarily improves the detection of unusual shapes and poses (e.g., cyclists, people sitting, bent over). However, in the KITTI evaluation, the person sitting class is ignored and cyclists are counted as false positives, hence such advantages are less helpful.

Method	Caltech	KITTI	Runtime
DeepParts [26]	11.89	58.67	1s
CompACT-Deep [3]	11.75	58.74	1s
MS-CNN [2]	9.95	73.70	0.4s
SA-FastRCNN [21]	9.68	65.01	0.59s
RPN+BF [29]	9.58	61.29	0.60s
F-DNN [10]	8.65	-	0.30s
F-DNN+SS [10]	8.18	-	2.48s
SDS-RPN (ours)	9.63	-	0.13s
SDS-RCNN (ours)	7.36	63.05	0.21s

Table 1. Comprehensive comparison of SDS-RCNN with other state-of-the-art methods showing the Caltech miss rate, KITTI mAP score, and runtime performance.

Efficiency: The runtime performance of SDS-RCNN takes ∼0.21s/image. We use images of size 720 × 960 pixels and a single Titan X GPU for computation. The efficiency of SDS-RCNN surpasses the current state-of-the-art methods for pedestrian detection, often by a factor of 2×. Compared to F-DNN+SS [10], which also utilizes segmentation cues, our method executes ∼10× faster. The next fastest runtime is F-DNN, which takes 0.30s/image with the caveat of requiring multiple GPUs to process networks in parallel. Further, our SDS-RPN method achieves very competitive accuracy while only taking 0.13s/image (frequently ∼3× faster than competitive methods using a single GPU).

4.2. Ablation Study

In this section, we evaluate how each significant component of our network contributes to performance using the reasonable set of Caltech [8]. First, we examine the impact of four components: weak segmentation supervision, proposal padding, cost-sensitive weighting, and stricter supervision. For each experiment, we start with SDS-RCNN and disable one component at a time as summarized in Table 2. For simplicity, we disable components globally when applicable. Then we provide detailed discussion on the benefits of stage-wise fusion and comprehensively report the RPN, BCN, and fused performances for all experiments. Finally, since our BCN is designed to not share features with the RPN, we closely examine how sharing weights between stages impacts network diversification and efficiency.

Weak Segmentation: The infusion of semantic features into shared layers is the most critical component of SDS-RCNN. The fused miss rate degrades by a full 3.05% when

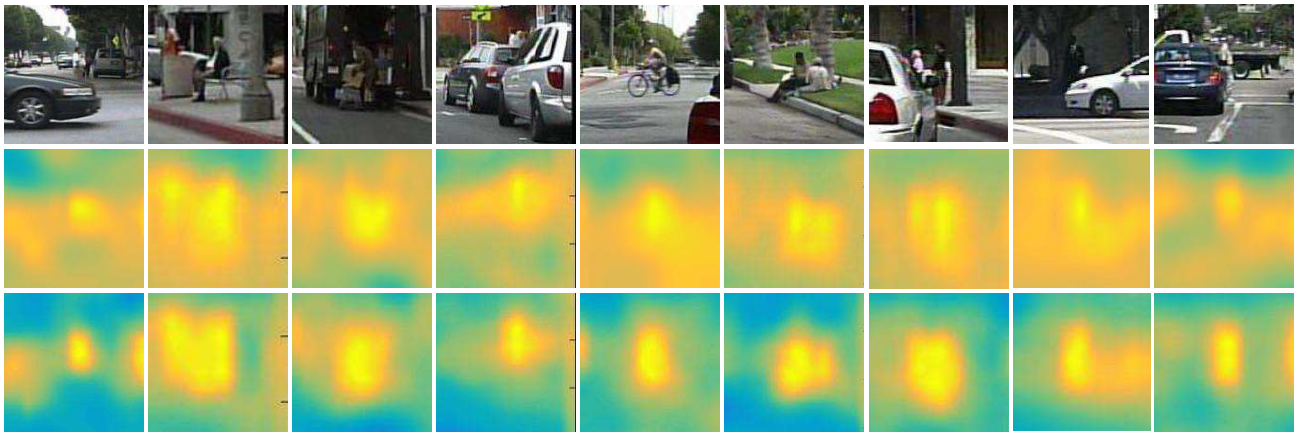

Figure 7. Example error sources which are corrected by infusing semantic segmentation into shared layers. Row 1 shows the test images from Caltech1×. Row 2 shows a visualization of the RPN proposal layer using the baseline network which fails on these examples. Row 3 shows a visualization of the proposal layer from SDS-RCNN, which corrects the errors. Collectively, occlusion and *unusual* poses of pedestrians (sitting, cyclist, bent over) make up for 75% of the corrections, suggesting that the the segmentation supervision naturally informs the shared features on robust pedestrian parts and shape information.

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

Table 2. 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 at a time.

the segmentation supervision is disabled, while both individual stages degrade similarly. To better understand the types of improvements gained by weak segmentation, we perform a failure analysis between SDS-RCNN and the "baseline" (non-weak segmentation) network. For analysis, we examine the 43 pedestrian cases which are missed when weak segmentation is disabled, but corrected otherwise. Example error corrections are shown in Fig. 7. We find that ∼48% of corrected pedestrians are at least partially occluded. Further, we find that ∼28% are pedestrians in *unusual* poses (e.g., sitting, cycling, or bent over). Hence, the feature maps infused with semantic features become more robust to atypical pedestrian shapes. These benefits are likely gained by semantic segmentation having indiscriminant coverage of all pedestrians, unlike object detection which requires specific alignment between pedestrians and anchor shapes. A similar advantage could be gained for object detection by expanding the coverage of anchors, but at the cost of computational complexity.

Proposal Padding: While padding proposals is an intuitive design choice to provide background context (Fig. 3), the benefit in practice is minor. Specifically, when proposal padding is disabled, the fused performance only worsens from 7.36% to 7.69% miss rate. Interestingly, proposal padding remains critical for the individual BCN performance, which degrades heavily from 10.98% to 13.09% without padding. The low sensitivty of the fused score to padding suggests that the RPN is already capable of localizing and differentiating between partial and full-pedestrians, thus improving the BCN in this respect is less significant.

Cost-sensitive: The cost-sensitive weighting scheme used to regularize the importance of large pedestrians over small pedestrians has an interesting effect on SDS-RCNN. When the cost-sensitive weighting is disabled, the RPN performance actually improves to an impressive 9.63% miss rate. In contrast, without cost-sensitive weighting the BCN degrades heavily, while the fused score degrades mildly. A logical explanation is that imposing a precedence on a single scale is counter-intuitive to the RPN achieving high recall across *all* scales. Further, the RPN has the freedom to learn scale-dependent features, unlike the BCN which warps to a fixed size for every proposal. Hence, the BCN can gain significant boost when encouraged to focus on large pedestrian features, which may be more scale-independent than features of small pedestrians.

Strict Supervision: Using a stricter labeling policy while training the BCN has a substantial impact on the performance of both the BCN and fused scores. Recall that the strict labeling policy requires a box to have IoU > 0.7 to be considered foreground, while the standard policy requires IoU > 0.5. When the stricter labeling policy is reduced to the standard policy, the fused performance degrades by 1.35%. Further, the individual BCN degrades by 6.43%, which is on par with the degradation observed when weak

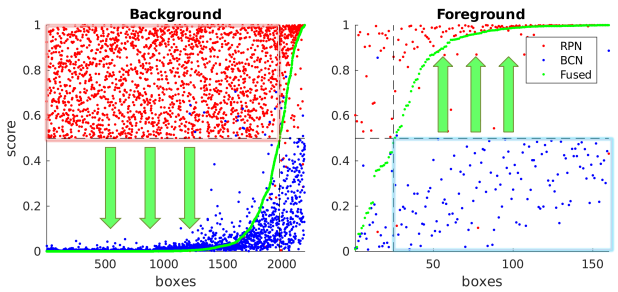

Figure 8. Visualization of the diversification between the RPN and BCN classification characteristics. We plot only boxes which the RPN and BCN of SDS-RCNN disagree on using a threshold of 0.5. The BCN drastically reduces false positives of the RPN, while the RPN corrects many missed detections by the BCN.

segmentation is disabled. We examine the failure cases of the strict versus non-strict BCN and observe that the false positives caused by double detections reduce by $\sim$22%. Hence, the stricter policy enables more aggressive suppression of poorly localized boxes and therefore reduces double detections produced as localization errors of the RPN.

Stage Fusion: The power of stage-wise fusion relies on the assumption that the each network will diversify in their classification characteristics. Our design explicitly encourages this diversification by using separate labeling policies and training distributions for the RPN and BCN. Table 2 shows that although fusion is useful in every case, it is difficult to anticipate how well any two stages will perform when fused without examining their specific strengths and weaknesses.

To better understand this effect, we visualize how fusion behaves when the RPN and BCN disagree (Fig. 8). We consider only boxes for which the RPN and BCN disagree using a decision threshold of 0.5. We notice that both networks agree on the majority of boxes ($\sim$80K), but observe an interesting trend when they disagree. The visualization clearly shows that the RPN tends to predict a significant amount of background proposals with high scores, which are corrected after being fused with the BCN scores. The inverse is true for disagreements among the foreground, where fusion is able to correct the majority of pedestrians boxes given low scores by the BCN. It is clear that whenever the two networks disagree, the fused result tends toward the true score for more than $\sim$80% of the conflicts.

Sharing Features: Since we choose to train a separate RPN and BCN, without sharing features, we conduct comprehensive experiments using different levels of stage-wise sharing in order to understand the value of diversification as a trade-off to efficiency. We adopt the Faster R-CNN feature sharing scheme with five variations differing at the point of sharing (conv1-5) as detailed in Table 3. In each experiment, we keep all layers of the BCN except those before and including the shared layer. Doing so keeps the effective depth of the BCN unchanged. For example, if the shared layer is conv4 then we replace conv1-4 of the BCN with a RoIPooling layer connected to conv4 of the RPN. We configure the RoIPooling layer to pool to the resolution of the BCN at the shared layer (e.g., conv4 $\rightarrow 14 \times 14$, conv5$\rightarrow 7 \times 7$).

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

Table 3. 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.

We observe that as the amount of sharing is increased, the overall fused performance degrades quickly. Overall, the results suggest that forcing the networks to share feature maps lowers their freedom to diversify and complement in fusion. In other words, the more the networks share the more susceptible they become to redundancies. Further, sharing features up to conv1 becomes slower than no stage-wise sharing (e.g., RGB). This is caused by the increased number of channels and higher resolution feature map of conv1 (e.g., $720 \times 960 \times 64$), which need to be cropped and warped. Compared to sharing feature maps with conv3, using no sharing results in a very minor slow down of 0.03 seconds while providing a 1.30% improvement to miss rate. Hence, our network design favors maximum precision for a reasonable trade-off in efficiency, and obtains speeds generally $2\times$ faster than competitive methods.

5. Conclusion

We present a multi-task infusion framework for joint supervision on pedestrian detection and semantic segmentation. The segmentation infusion layer results in more sophisticated shared feature maps which tend to *illuminate* pedestrians and make downstream pedestrian detection easier. We analyze how infusing segmentation masks into feature maps helps correct pedestrian detection errors. In doing so, we observe that the network becomes more robust to pedestrian poses and occlusion compared to without. We further demonstrate the effectiveness of fusing stage-wise scores and encouraging network diversification between stages, such that the second stage classifier can learn a stricter filter to suppress background proposals and become more robust to poorly localized boxes. In our SDS-RCNN framework, we report new state-of-the-art performance on the Caltech pedestrian dataset (23% relative reduction in error), achieve competitive results on the KITTI dataset, and obtain an impressive runtime approximately $2\times$ faster than competitive methods.

References

[1] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 328–335, 2014. 1, 2

[2] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016. 1, 2, 3, 6

[3] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3361–3369, 2015. 6

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 1

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 2, 5

[6] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016. 1, 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009. 1, 2, 3, 4, 5, 6

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012. 2, 5

[10] X. Du, M. El-Khamy, J. Lee, and L. S. Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. *arXiv preprint arXiv:1610.03466*, 2016. 2, 3, 6

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html. 6

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2

[13] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3294–3301, 2013. 1, 2

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 5, 6

[15] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142, 2015. 2

[16] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 2, 3

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[18] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 1, 2

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3, 4

[21] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *arXiv preprint arXiv:1510.08160*, 2015. 1, 2, 6

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 3

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 2

[26] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1904–1912, 2015. 2, 6

[27] S. Tsogkas, I. Kokkinos, G. Papandreou, and A. Vedaldi. Deep learning for semantic part segmentation with high-level guidance. *arXiv preprint arXiv:1505.02438*, 2015. 3

[28] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 2129–2137, 2016. 2

[29] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? *arXiv preprint arXiv:1607.07032*, 2016. 1, 2, 3, 4, 6

[30] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1267, 2016. 1, 2

[31] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1751–1760. IEEE, 2015. 5