# SAM-RCNN: Scale-Aware Multi-Resolution Multi-Channel Pedestrian Detection

Tianrui Liu[1]
t.liu15@imperial.ac.uk

Mohamed Elmikaty[2]
melmikat@jaguarlandrover.com

Tania Stathaki[1]
t.stathaki@imperial.ac.uk

[1] Electrical and Electronic Engineering
Imperial College London
London, UK

[2] Jaguar Land Rover Research
Coventry, UK

## Abstract

Convolutional neural networks (CNN) have enabled significant improvements in pedestrian detection owing to the strong representation ability of the CNN features. Recently, aggregating features from multiple layers of a CNN has been considered as an effective approach, however, the same approach regarding feature representation is used for detecting pedestrians of varying scales. Consequently, it is not guaranteed that the feature representation for pedestrians of a particular scale is optimised. In this paper, we propose a Scale-Aware Multi-resolution (SAM) method for pedestrian detection which can adaptively select multi-resolution convolutional features according to pedestrian sizes. The proposed SAM method extracts the appropriate CNN features that have strong representation ability as well as sufficient feature resolution, given the size of the pedestrian candidate output from a region proposal network. Moreover, we propose an enhanced SAM method, termed as SAM+, which incorporates complementary features channels and achieves further performance improvement. Evaluations on the challenging Caltech and KITTI pedestrian benchmarks demonstrate the superiority of our proposed method.

## 1 Introduction

Pedestrian detection has wide applications in video surveillance, robotics automation and intelligent transportation. A robust pedestrian detector must be able to detect pedestrians of various poses and appearances and at different scales when they are placed in complex scenarios with cluttered backgrounds. A substantial number of methods have been developed in order to improve the detection accuracy [7, 10, 11, 24, 26, 33, 34, 35]. In particular, CNN based detectors [16, 27, 33] pushed the pedestrian detection performance to a new level. Compared to the traditional feature representation for pedestrian detection, CNN features have strong representation capacity that can largely handle the pose and appearance variations.

Current CNN-based detection methods use the same CNN feature representation to detect pedestrians of different sizes. Nonetheless, a sole feature representation does not always provide the best representation for objects at different sizes. As indicated in [21], the visual appearance and the feature representation of large-scaled and small-scaled pedestrians are
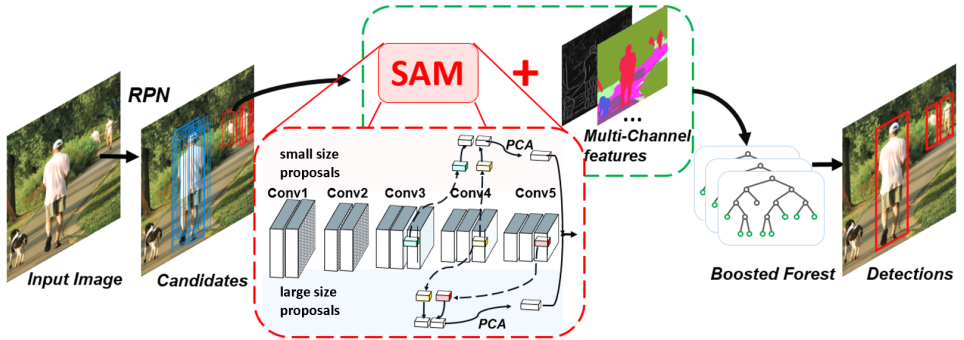
Figure 1: Overview of the proposed pedestrian detection framework. Feature extraction is performed differently using the proposed scale-aware multi-resolution (SAM) method according to the pedestrian candidate sizes from RPN (Section 3.2). For small pedestrian candidates, feature maps at shallow layers, such as *Conv*3 and *Conv*4, are used, while for large pedestrian candidates, feature maps at deeper layers are utilized because the feature resolution is sufficient large. SAM+ uses complementary feature channels with the CNN features (Section 3.3)

significantly different. This suggests that there is room for improvement if we could use different feature representation of objects of different sizes.

To solve this problem, we explicitly estimate the scale of the candidate pedestrians and propose to use a **Scale-Aware Multi-resolution (SAM)** strategy for pedestrian detection. Given the sizes of the candidate pedestrians, we can adaptively select suitable feature representations for pedestrians of different scales, rather than compromising on features that balance for pedestrians of all scales. Our intuition is that the best features (in terms of balancing feature abstraction level and feature resolution) for pedestrians of different sizes may come from different CNN layers. A large-size pedestrian should be represented by features from deep layers, whereas a small-size pedestrian should be represented by features from shallow layers which are of higher resolutions. With the proposed SAM strategy, the detector can choose appropriate CNN features which has the strongest representation ability and meanwhile retains sufficient feature resolution for pedestrians of a specific size.

Furthermore, we propose an enhanced version of the SAM detector, denoted as **SAM+**, by incorporating complementary feature channels with the CNN features for a better pedestrian detection. A novel **RoI histogram pooling** method is proposed to extract feature vectors from the additional feature maps for candidate regions of arbitrary size, leading to better performance than using RoI max pooling [7]. With the complementary features, our SAM+ detector can effectively eliminate hard negatives such as tree leaves and traffic lights. As pedestrians of different size may use different combinations of CNN layers and thus have features of different dimensions, we apply principal component analysis (PCA) to transform the feature vector into a fixed length vector such that it can be easily fed into the boosted forest (BF) classifier [28]. The flexibility of the BF classifier imposes no need for feature amplitudes normalization when combining multi-resolution CNN features and facilitates the integration of additional feature channels for the SAM+ detector.

The contributions of this work can be summarized as follows: 1. We proposed a scale-aware multi-resolution pedestrian detection framework which exploits multi-resolution fea-

tures from CNN and uses different combination of feature layers according to the size of the candidate pedestrians. 2. We propose an enhanced version of the SAM, termed as SAM+, which uses additional semantic and edge feature channels to obtain valuable complementary information for pedestrian detection. Experimental results show improvements on detection rates using additional cues. 3. The proposed SAM pedestrian detector achieves the state-of-the-art on the Caltech and KITTI pedestrian benchmarks. Along with the enforcement of the additional features channels, the proposed SAM+ detector outperforms the state-of-the-art.

## 2 Related Work

Traditional pedestrian detectors, such as ACF [10] and Checkerboards [34], are based on hand-engineered features which are usually descriptors of the gradients, edge and colors computed over a sliding window. These descriptors are used in conjunction with a classifier, such as boost forest, to perform pedestrian detection via classification. These methods were the dominant approaches for pedestrian detection before the emerging of the CNN based pedestrian detection methods.

Compared to the hand-engineered features, features generated by deep convolutional networks have demonstrated to have stronger feature representation capability in many detection tasks. Region-based convolutional neural networks (R-CNN) [15] method makes use of an "attention" mechanism which first proposes a small number of high potential candidate regions where classification is performed afterwards. R-CNN [15] and Fast R-CNN [14] apply selective search [30] for region proposal, and Faster R-CNN [27] replaces selective search with a built-in region proposal network (RPN) network that can effectively generate proposals. These general object detection methods when directly applied for pedestrian detection do not lead to satisfying results [33, 36] because Faster-RCNNs do not perform well on small size objects which dominate most pedestrian datasets [9, 13]. In [33], the Faster R-CNN is tailored to accommodate for pedestrian detection and achieves better results.

In an attempt to take advantage of the CNN features from multiple layers, the Inside-Outside Net [1] concatenates multiple layers of CNN and unifies the feature dimensions using a $1 \times 1$ convolution. As the convolutional features at each layer have very different amplitudes, they need to rely on $L_2$ normalization to normalize the features before the concatenation. In single shot multi-box detection (SSD) [25], several additional layers are built after $Conv5\_3$ of the VGG16 net [29]. SSD combines the CNN features starting from $Conv5\_3$ with the additional layers. However, SSD does not use the higher-resolution maps before $Conv5\_3$ that are crucial for detecting small objects. Feature pyramid network [23] combines low-resolution with high-resolution feature maps via a top-down pathway. The lower resolution feature maps are up-sampled by a factor of 2 and merged with the upper layer feature map by element-wise addition. The result is a feature pyramid that has high resolution and strong representation ability at all levels. Although these methods combine features from different layers of a deep convolutional network, they use the same features extraction for all candidates. Different from the existing methods, we focus on combining multi-layer features in a scale-aware manner to adaptively choose the most suitable feature representation for pedestrians of a certain scale.

## 3 Proposed Method

The overview of the proposed scale-aware multi-resolution (SAM) pedestrian detection framework is illustrated in Figure 1. Given an input image, RPN generates a pool of pedestrian

candidates with an estimated size and a confidence score that will be used as priors for the second stage classification. Pedestrian candidates are grouped according to their sizes, and each group performs feature extraction differently using our proposed SAM method. For small-scaled pedestrian candidates, CNN feature maps at shallow layers, such as *Conv*3 and *Conv*4, are used since CNN feature maps at even deeper layers are considered to be of too low resolution to provide useful information. For large-scaled pedestrian candidates, resolution of feature maps from deeper layers, e.g. *Conv*4 and *Conv*5, are of sufficient resolution and thus, can be utilised. Since pedestrians of different sizes adopt features from different layers of a CNN, their feature representations are of different sizes. PCA is applied to transform the feature representations of different sizes into a fixed length vector. In addition, complementary feature channels are integrated with the CNN features output from SAM and are fed into a boosted forest for classification. A detailed description of each processing block of the proposed framework is given below.

## 3.1 Region Proposal Network for Pedestrian Candidates Proposal and Scale Estimation

The scale of the candidate pedestrians is estimated using the RPN [27], which is a small network built on top of the last convolutional layer of the VGG-16 network. For general object detection [27], a three-scale three-ratio anchor is used to generate 9 proposals at each sliding position. For pedestrian candidates proposal, we use anchors of a single ratio of $\gamma = 0.41$ with 9 scales and fine-tune the RPN [33] on the Caltech pedestrian benchmark [9].

Each candidate window output from RPN is associated with information regarding window position, confidence score, and window size. The confidence score is passed to the boosted forest (BF) classifier as the initial detection score. The RPN can generate high quality pedestrian proposals. With 100 proposals per image, the RPN can achieve $> 99\%$ recall at an intersection over union (IoU) of 0.5, and $> 95\%$ recall at an IoU of 0.7. The candidate windows are sorted by their confidence scores in descending order. At test time, the top-ranked 100 candidates are passed to the BF classifier for classification, while for training the top-ranked $1,000$ proposals are used.

## 3.2 Scale-Aware Multi-Resolution Features

Features extracted from different CNN layers represent different levels of abstraction and can be all helpful for pedestrian detection. Features from a deeper CNN layer have stronger representation ability but lower resolution, whereas features from a shallower layer is of higher resolution but weaker representation ability. Using features from unsuitable CNN layers will bring difficulties to the classification process. For instance, when an image of $480 \times 360$ containing a small pedestrian of 50 pixels in height is forward to the VGG-16, it will pass through 4 pooling layers and the feature maps at Conv5 will be down-sampled by a factor of $2^4$. This leads to the size of active feature maps for the pedestrian being about only 3 pixels in height. With such low-resolution feature maps, classifiers can hardly discriminate between pedestrians and other irrelevant objects in the scene.

**CNN Features Extraction.** CNN features are extracted from the last layer of each convolutional block in the VGG-16 network, i.e., *Conv*1_2, *Conv*2_2, ..., *Conv*5_3. For simplicity, we refer to them as *Conv*1, *Conv*2, ..., *Conv*5. In addition, we also exploit the "*à* trous" version of CNN features. The "*à* trous" convolution technique is proposed in [4] which doubles

the feature resolution extracted from *Conv*4 to achieve better semantic segmentation performance. An *à* trous feature map is obtained by dilating the original filter size by a factor of 2 so that the stride of the original feature map can be reduced by 2. Using the *à* trous convolution enables a higher feature resolution while preserving the same feature representation ability. This is crucial for small object detection. Hence, we also perform experiments on the dilated version of *Conv*4 and *Conv*5 features and refer to them as *Conv*4*à* and *Conv*5*à* henceforth. RoI max pooling [27] is adopted in order to obtain fix-length CNN feature vectors for candidate region with varying sizes. Unlike in [27] where only *Conv*5 features are extracted, we combine multi-resolution feature maps from multiple layers of a CNN. In order to exploit suitable feature representation for pedestrians of different sizes, we have conducted extensive experiments on two subsets of pedestrians, i.e. small-size and large-size pedestrian sub-dataset, containing pedestrians of height in pixels belonging to the range of $[50, 80)$ and $[80, \infty)$, respectively. A comprehensive analysis of using different multi-resolution CNN features is given in Section 4.2.

**Multi-resolution Feature Combination using PCA.** Taking the advantage of multi-layer feature aggregation, we utilize multi-resolution CNN features wherein the feature combinations are determined according to the size of the candidate pedestrian. Since CNN features from different layers have different dimension, feature representation for candidates of different size could be of non-uniform length. The PCA algorithm is applied to project these features into fixed length vectors. For a feature combination, we collect a large number of training features from both pedestrian regions and background regions. The eigenvectors corresponding to the top $d$ eigenvalues have been reserved to form a projection matrix for dimension reduction. In this way, the feature representation of pedestrians with varying sizes can be transformed into the same size while preserving their representation ability.

## 3.3 SAM+: Enhance SAM using Additional Feature Channels

The enhanced SAM detector utilizes additional feature channels, i.e. semantic feature channel and edge feature channel to provide complementary information for pedestrian detection.

We propose an **RoI histogram pooling** method to extract feature vectors from the additional feature maps for candidate regions of arbitrary size. RoI histogram pooling works by dividing the $H \times W$ RoI window into an $m \times n$ grid of non-overlapping cells so that each cell is of approximate size $(H/m) \times (W/n)$. Differently from RoI max pooling [27] which pools the maximum value in each cell into the corresponding output grid cell, we pool out a normalized histogram of values $c \in C$ computed across the feature maps of this cell, i.e. $h_c = n_c/(m \times n)$ where $n_c$ is the number of pixels in this cell belonging to class $c$. Then histogram vectors $h_c$ of all cells in this RoI window are concatenated to generate the feature representation of this candidate region. For both RoI max pooling and RoI histogram pooling layer, we use $m = 12$ and $n = 5$ which is suitable for the pedestrian aspect ratio. We demonstrate in Section 4.2.2 that the features extracted using the proposed RoI histogram pooling perform better than those using RoI max pooling.

**Semantic feature channel.** We adopt the recent semantic segmentation method RefineNet [22] to perform pixel-wise semantic labeling and output valuable semantic information for pedestrian detection. RefineNet is trained on the Cityscapes dataset [5] to semantically label an image into $C = 20$ common classes (e.g. building, tree, sky, road, pedestrian, etc.). We encode the semantic segmentations with the object category index $c$ ($c = 1, ..., 20$) for each pixel as the semantic feature channel.

**Edge feature channel.** Our edge feature channel is encoded using the intensity values of the edge response from the HedNet [32]. Different from traditional edge detectors such as the Canny, the HedNet can generate semantically meaningful edge maps at object contour.

## 3.4 Boosted Forest for Integrated Multi-Resolution Multi-Channel Features

Boosted Forest (BF) has been widely used in computer vision tasks such as object recognition [12, 51], and super-resolution [17, 18, 20] as it can achieve fast and accurate classification. The flexibility of BF facilitates the combination of multi-layer CNN features without the need for feature amplitude normalization, and is also convenient for the integration of additional feature channels in the SAM+ detector. The confidence scores of the candidate pedestrians output from RPN are passed to BF as preliminary scores for classification.

We adopt the RealBoost algorithm [28] to perform bootstrapping for multi-stage hard negative samples mining. Our SAM detector performs 6 stages of bootstrapping passes in addition to the original training phase. The number of weak classifiers used in BF are $64, 128, 256, 512, 1024, 2048$ for stage $0, 1, ..., 5$ respectively. Initially, the training set consists of all positive examples and $30,000$ negative samples which are randomly sampled from background image regions. For each bootstrapping stage, $5,000$ hard negative samples are collected using the detector obtained from the previous stage.

We also built a smaller version of SAM, termed **SAM-Basic**, which allows us to exploit many more different settings in a short training time. SAM-Basic has 5 training stages. Each stage has $\{32, 64, 128, 256, 512\}$ weak classifiers, respectively. At the first stage, $10,000$ negative samples are randomly sampled and, the number of hard negative samples to be added at each bootstrapping pass is limited to $1,000$.

# 4 Experiments and Analysis

## 4.1 Datasets

**Caltech.** Caltech-USA dataset [9] and the improved annotations [55] are used for training and evaluation. As in [2, 21, 53], we use the Caltech10× training set which is obtained by extracting every $3^{rd}$ frame from the Caltech videos. The testing set contains 4024 images of size $480 \times 640$. Following the evaluation of Caltech benchmark [9], only bounding boxes restricted in the range of $x \in [5,635]$, $y \in [5,475]$ are evaluated. The "reasonable" evaluation setting is used which counts for pedestrians of height larger than 50 pixels and with less than 35% occlusion. Evaluations are measured using the log average miss rate (MR) of false positive per image (FFPI) ranging from $10^{-2}$ to $10^{0}$ ($MR_{-2}$). For SAM-Basic, the Caltech1× training set with 4250 images is used and the evaluation is performed on a subset of the Caltech testing set which contains 905 images.

**KITTI.** The KITTI dataset [13] is composed of 7,481 images for training and 7,581 images for testing. Since the ground-truth annotations of testing set is not publicly available, we use the training/validation set, as in [3, 26], for performance analysis. As in KITTI standard, we evaluate our detection methods at three levels of difficulty, i.e. "Easy", "Moderate"and "Hard", where the difficulty is measure by the minimal height, the occlusion and the truncation of an object. The mean Average Precision (mAP) with 0.5 overlap ratio is used to measure the pedestrian detection performance.
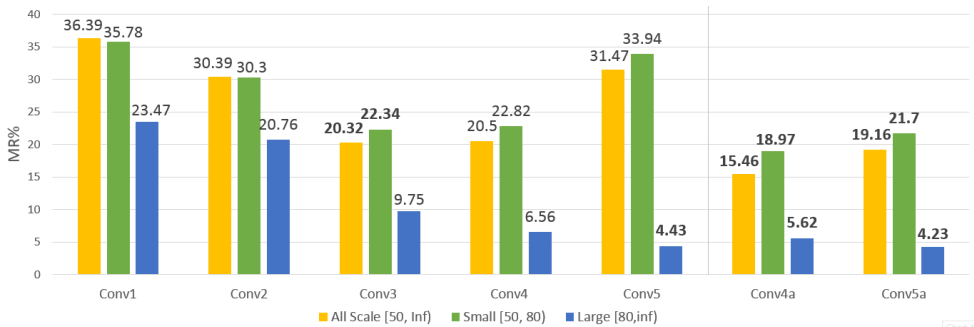
Figure 2: Comparison of SAM-Basic results ($MR_{-4}\%$) for "All Scale", "Small" and "Large" pedestrians using feature representation from different CNN layers, the lower the better.

Our implementation is based on the publicly available code for Faster-RCNN [8, 33] with Caffe [19]. All experiments were performed on a machine with a single GPU TITAN X and a CPU Intel Core i7 3.4 GHz.

## 4.2 Results and Discussions

### 4.2.1 Scale-Aware Multi-Resolution Convolutional Features

We analyse the performance of the SAM-Basic detector using different layers of CNN features and different feature combinations. The experiments were performed under the same parameter settings, except that the CNN features were extracted from different layers.

**Single Layer Convolutional Feature for Pedestrians of Difference Scales.** First, we demonstrate that the most suitable features for pedestrians of different scales are from different convolutional layers. We have trained and evaluated SAM-Basic detectors using the "Small", "Large" and All-scale" pedestrians and compared the results in Figure 2. At each time, features from a single convolutional layer, i.e. $Conv1, ..., Conv3, Conv4/Conv4à$ and $Conv5/Conv5à$, are used for training. The log miss rates (MR) of SAM-Basic detectors are averaged over the FPPI range $[10^{-4}, 10^0]$ ($MR_{-4}$) since $MR_{-2}$ for large-size pedestrians are nearly zero and is difficult to be compared.

From Figure 2, we can see that the SAM-Basic detectors perform differently when using features from different CNN layers. More importantly, the best feature representation for small pedestrians is different from that for large pedestrians. Excluding the à trous convolutional layers, the best performance for small-size pedestrians is achieved by using $Conv3$ features, while the best of large-size pedestrians is obtained using $Conv5$ features. This verifies that the optimized CNN features for pedestrians of different sizes are from different convolutional layers.

For large-size pedestrians, the lowest miss rates are achieved by using the $Conv5$ and $Conv5à$ features. The result indicates that using features from deeper layers which have stronger representation ability is indeed beneficial, as long as the feature resolution is proper for the object to be detected. This inference can be further confirmed by comparing the results between $Conv4$ (or $Conv4à$) and $Conv5$ (or $Conv5à$). For instance, features from $Conv4$ and $Conv4à$ are of the same representation ability, but the results using $Conv4à$ features are better because $Conv4à$ has doubled feature resolution. For small-size pedestrians, results

| Small | Conv34à | Conv4à5à | Conv34à5à |
|---|---|---|---|
| [50, 80) | **17.55** | 19.07 | 18.22 |

| Large | Conv4à5 | Conv4à5à | Conv34à5à |
|---|---|---|---|
| [80, ∞] | 7.49 | **5.88** | 6.69 |

Table 1: Comparison of SAM-Basic result ($MR_{-4}\%$) for small-size (upper) and large-size (lower) pedestrian detection using different combinations of multi-resolution CNN features.

using $Conv1$ and $Conv2$ features are poor because these shallow layers have weak feature representation capability. Performance using $Conv4$ and $Conv5$ features is not satisfied which can be explained by the in-sufficient feature resolution for detecting small pedestrians. Although features from deeper layers have stronger representation ability, these features have too low resolution to let the classifier make good classification. $Conv3$ yields the best result because it is the layer that can best balance feature abstraction level and feature resolution.

**Scale-Aware Multi-Resolution Convolutional Features.** According to the results on single layer convolutional features for pedestrians of difference scales, we conduct experiments using combinations of multi-layer CNN features. The features that lead to the best three results (bold in Figure 2) are used for combination. The performance of different combinations are shown in Table 1, where the multi-resolution features are shortened for concise (e.g., the combination of features from $Conv3$ and $Conv4à$ is denoted as $Conv34à$).

From Table 1, we can see that for large-size pedestrians, the feature combination of $Conv34à5à$ achieves the lowest miss rate; while for small-size pedestrians the best performance is yielded by $Conv34à$. Moreover, by comparing the results between $Conv4à5à$ and $Conv34à5à$ for both small and large-size pedestrians, we have an interesting observation: while the incorporation of $Conv3$ features facilitates the detection of small pedestrians, it is on the contrary harmful to the detection of large pedestrians. Therefore, it is not always better to combine more layers of CNN features. As $Conv3$ features are not of the most suitable resolution for large-size pedestrians, combining such features will even degrade the performance.

### 4.2.2 Comparison between SAM and SAM+

Given the results in Table 1, we evaluate our SAM-Basic detector using the best multi-resolution feature combinations for each scale range. That is, for small-size candidates, features are extracted from $Conv3$ and $Conv4à$, and for large-size pedestrians, features are extracted from $Conv4à$ and $Conv5à$. The dimension of $Conv34à$ feature is $512 + 256 = 786$, whereas the dimension of $Conv4à5à$ feature is $512 + 512 = 1024$. We apply PCA on $Conv4à5à$ features to reduce their dimension to 786, so that candidates of different sizes can have uniform feature length. To analyse the influence of feature dimension reduction using PCA, we evaluate the performance of SAM-Basic using the Conv45à features before and after dimension reduction. From Table 2, we can see that there is no performance deterioration in terms of $MR$ when the dimension of $Conv45à$ features is reduced from 1024 to 786.

The performance of the SAM+ detector using complementary feature channels is given in Table 3, where "max." and "hist." in the last column indicates that the complementary features are extracted using the RoI max pooling and our RoI histogram pooling method, respectively. By comparing the miss rate between "*max.*" and "*hist.*", we can see that the

| | Feature dim. | Energy% | *MR%* |
|---|---|---|---|
| Original | 1024 | 100 | 17.24 |
| Reduced | 786 | 98.4 | 17.15 |

Table 2: Comparison of SAM-Basic results (*MR*$_{-4}$%) using *Conv*45$\grave{a}$ features before and after PCA feature reduction.

| | All scale [50, ∞] | Large [80, ∞] | Small [50, 80) | Pooling |
|---|---|---|---|---|
| SAM-Basic | 16.16 | 7.00 | 18.79 | - |
| +Semantic | 15.77 | 7.15 | 17.98 | max. |
| | **15.14** | **6.63** | **16.55** | hist. |
| + Edge | 15.51 | 7.01 | 17.96 | max. |
| | **14.73** | **6.54** | **15.80** | hist. |

Table 3: Performance of SAM+ using semantic and edge feature channels (*MR*$_{-4}$%, the lower the better), "max." and "hist." indicates RoI max pooling and RoI histogram pooling, respectively.

proposed RoI histogram pooling performs better than RoI max pooling for semantic and edge feature extraction. We witness there is an overall improvement of 1.02% from the integration of the semantic feature channels. When we look at the fine-grained improvements for different scale ranges, we find that the improvement for small-size pedestrians (i.e., 2.24%) is larger than that for large-size pedestrians (i.e., 0.37%). This indicates that the semantic channel are more helpful for small pedestrians which are usually the hard cases in pedestrian detection. On the basis of using semantic feature channel, the integration of edge feature channel further improves the performance by 0.38%. Again, the detection rate of on small-size pedestrians benefits more than that of the large-size pedestrians.

### 4.2.3 Comparison with State-of-the-art Pedestrian Detection Methods

**Caltech.** In Figure 3, our SAM and SAM+ pedestrian detectors are compared with the state-of-the-art pedestrian detection methods, namely, Checkerboards [34], MRFC [6], CompACTDeep [2], SA-FastRCNN [21], MS-CNN [3] , RPN+BF [33], and HyperLearner [26]. Under the evaluation setting of IoU is 0.5 (Figure 3 (left)), the performance of SAM is on par with that of the latest HyperLearner method [26] even without using additional features. Our SAM+ detector has achieved a MR of 4.9% which outperforms the current state-of-the-art by 0.6%. Under a stricter evaluation condition of IoU is 0.7 (Figure 3 (right)), our proposed SAM and SAM+ method outperform all existing pedestrian detection methods with a larger margin and the performance of SAM+ is 0.4% better than that of SAM. This indicates that our proposed method not only achieves lower miss rate, but also obtains detection with more precise position.

**KITTI.** Table 4 shows the pedestrian detection results evaluated on the KITTI dataset. Again, our SAM method has competitive accuracy compared to the latest pedestrian detector [26] without using additional feature channels. Under the "hard" evaluation setting where pedestrians of 25 pixels in height and heavily occluded are counted, SAM outperforms HyperLearner [26] by 4.67%. The SAM+ method further improves the performance of SAM under all three evaluation settings.
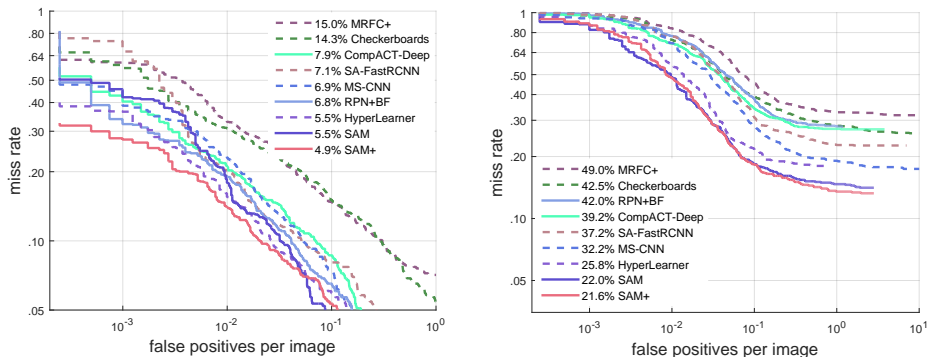
Figure 3: Comparison of results ($MR_{-2}$) on the Caltech test set evaluated using IoU 0.5 (left) and 0.7 (right), respectively.

| Method | Moderate | Easy | Hard |
|--------|----------|------|------|
| Faster-RCNN [33] | 71.05 | 76.00 | 62.08 |
| MS-CNN [3] | 72.26 | 76.38 | 64.08 |
| HyperNet [26] | 72.23 | 77.96 | 63.43 |
| HyperLearner [26] | 72.51 | 78.51 | 63.24 |
| SAM [Our] | **74.07** | **78.80** | **67.91** |
| SAM+ [Our] | **74.25** | **78.92** | **68.40** |

Table 4: Comparisons of pedestrian detection results (mAP%) on KITTI.

## 4.3 Conclusions

In this paper, we proposed a scale-aware multi-resolution (SAM) pedestrian detection framework which exploits different combination of multi-resolution CNN features for pedestrian candidates of different scales. Through extensive experiments, we found that for pedestrians of different scales, the features that can best balance feature abstraction level and resolution are from different convolutional layers. It is not always better to combine more layers of CNN features. Using features of unsuitable resolution will bring difficulties to the detector in the classification stage, and thus is harmful for detection accuracy. We also proposed the enhanced SAM+ detector which makes use of the additional feature channels as complementary information for pedestrian detection. Relying on the additional cues, some ambiguous pedestrian hypotheses that may be difficult to classify using the CNN features can be discriminated with the proposed method. Experiments indicate that the proposed detectors achieve state-of-the-art performance.

## 4.4 Acknowledgment

# References

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.

[2] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] Zhaowei Cai, Quanfu Fan, Rogerio Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

[6] A. D. Costea and S. Nedevschi. Semantic channels for fast pedestrian detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2360–2368, June 2016. doi: 10.1109/CVPR.2016.259.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*, volume 1, pages 886–893. IEEE. ISBN 0769523722.

[8] Piotr Dollar. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox.

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition (CVPR), 2009. IEEE Conference on*, pages 304–311. IEEE. ISBN 1424439922.

[10] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (8):1532–1545, 2014.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. ISSN 0162-8828.

[12] Juergen Gall and Victor Lempitsky. *Class-specific hough forests for object detection*, pages 143–157. Springer, 2013. ISBN 1447149289.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[14] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[17] J. J. Huang, W. C. Siu, and T. R. Liu. Fast image interpolation via random forests. *IEEE Transactions on Image Processing*, 24(10):3232–3245, Oct 2015. ISSN 1057-7149. doi: 10.1109/TIP.2015.2440751.

[18] Jun-Jie Huang and Wan-Chi Siu. Learning hierarchical decision trees for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 27 (5):937–950, 2017. doi: 10.1109/TCSVT.2015.2513661.

[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[20] Pier Luigi Dragotti Jun-Jie Huang, Tianrui Liu and Tania Stathaki. Srhrf+: Self-example enhanced single image super-resolution using hierarchical random forests. In *CVPR Workshop: New Trends in Image Restoration and Enhancement workshop*, 2017.

[21] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20 (4):985–996, 2018.

[22] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, July 2017.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Tianrui Liu and Tania Stathaki. Enhanced pedestrian detection using deep learning based semantic image segmentation. In *Digital Signal Processing (DSP), 2017 22nd International Conference on*, pages 1–5. IEEE, 2017.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[26] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[28] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[30] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

[31] Paul Wohlhart, Samuel Schulter, Martin K?stinger, Peter M Roth, and Horst Bischof. Discriminative hough forests for object detection. In *BMVC*, pages 1–11, 2012.

[32] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[33] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision*, pages 443–457. Springer, 2016.

[34] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Filtered channel features for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1751–1760. IEEE.

[35] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt 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.

[36] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.