Feature Calibration Network for Occluded Pedestrian Detection

Tianliang Zhang, *Student Member, IEEE*, Qixiang Ye, *Senior Member, IEEE*, Baochang Zhang, *Member, IEEE*, Jianzhuang Liu, *Senior Member, IEEE*, Xiaopeng Zhang, and Qi Tian, *Fellow, IEEE*

***Abstract*— Pedestrian detection in the wild remains a challenging problem especially for scenes containing serious occlusion. In this paper, we propose a novel feature learning method in the deep learning framework, referred to as Feature Calibration Network (FC-Net), to adaptively detect pedestrians under various occlusions. FC-Net is based on the observation that the visible parts of pedestrians are selective and decisive for detection, and is implemented as a self-paced feature learning framework with a self-activation (SA) module and a feature calibration (FC) module. In a new self-activated manner, FC-Net learns features which highlight the visible parts and suppress the occluded parts of pedestrians. The SA module estimates pedestrian activation maps by reusing classifier weights, without any additional parameter involved, therefore resulting in an extremely parsimony model to reinforce the semantics of features, while the FC module calibrates the convolutional features for adaptive pedestrian representation in both pixel-wise and region-based ways. Experiments on CityPersons and Caltech datasets demonstrate that FC-Net improves detection performance on occluded pedestrians up to 10% while maintaining excellent performance on non-occluded instances.**

***Index Terms*— Pedestrian detection, occlusion handling, feature calibration, feature learning, self-paced learning.**

I. Introduction

PEDESTRIAN detection is an important research topic in the computer vision area, driven by many real-world applications including autonomous driving [1], video surveillance [2], and robotics [3]–[5]. With the rise of deep learning, pedestrian detection has achieved unprecedented performance in simple scenes. However, the performance for detecting heavily occluded pedestrians in complex scenes remains far from being satisfactory [6]–[11]. For example, when the occlusion rate is higher than 35% (Caltech pedestrian dataset), state-of-the-art methods [9] report miss rates larger than 50% at 0.1 False Positive Per Image (FPPI). This seriously hinders the deployment of pedestrian detection in real-world scenarios.

To address the occlusion issue, one commonly used method is the part-based model [12], which leverages a divide-and-conquer strategy to handle visible and occluded parts. However, such a method suffers deficiency in handling complex occlusions due to a limited number of parts and the fixed part partition strategy. The other commonly used method is the attention model [13], which replaces "hard" object parts with "soft" attention regions by introducing feature enforcement and/or sampling modules [9], [14]. Nevertheless, the attention model usually operates in parallel with the detector learning procedure, ignoring the class-specific semantic information produced by the detectors. This could mix the attentive regions of negatives and positives and make the feature enforcement dubiously oriented.

In this paper, we propose self-activation (SA) and feature calibration (FC) modules, and target at adaptating the convolutional features to pedestrians of various occlusions. The SA module defines the corresponding relationship between pedestrians and convolutional feature channels, without any additional parameter involved. Such relationship is reflected by a classifier weight vector, which is constructed during the learning of the detection network. By multiplying such a weight vector with the feature maps in a channel-wise manner, the visual patterns across channels are collected and a pedestrian activation map is calculated, as shown in Fig. 1. The activation map is further fed to the FC module to reinforce or suppress the convolutional features in both pixel-wise and region-based manners.

Integrating the SA and FC modules with a deep detection network leads to our feature calibration network (FC-Net). In each learning iteration, FC-Net updates the classifier weights, which are reused to calibrate the features, iteratively. The key idea of calibration is leveraging the pedestrian activation map as an indicator to reinforce the features in visible pedestrian parts while depressing the features in occluded pedestrian regions. With multiple iterations of feature calibration, FC-Net attentively learns discriminative features for pedestrian representation in a self-paced manner.

The contributions of this work include:

(1) We propose a self-activation approach, and provide a simple yet effective way to estimate pedestrian activation maps by reusing the classier weights of the detection network.

Manuscript received January 15, 2020; revised June 7, 2020 and October 11, 2020; accepted November 23, 2020. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61836012 and 61771447 and Strategic Priority Research Programme of Chinese Academy of Science under Grant XDA27010303. The Associate Editor for this article was K. Wang. *(Corresponding author: Qixiang Ye.)*

Tianliang Zhang was with Huawei Noah's Ark Lab, Shenzhen 518129, China. He is now with the School of Electronics, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: zhangtianliang17@mails.ucas.ac.cn).

Qixiang Ye is with the School of Electronics, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qxye@ucas.ac.cn).

Baochang Zhang is with the Institute of Deep Learning, Baidu Research, Beijing 100191, China (e-mail: bczhang@139.com).

Jianzhuang Liu is with Huawei Noah's Ark Lab, Shenzhen 518129, China (e-mail: liu.jianzhuang@huawei.com).

Xiaopeng Zhang and Qi Tian are with Cloud & AI, Huawei Technologies, Shenzhen 518129, China (e-mail: zhangxiaopeng12@huawei.com; tian.qi1@huawei.com).

Digital Object Identifier 10.1109/TITS.2020.3041679

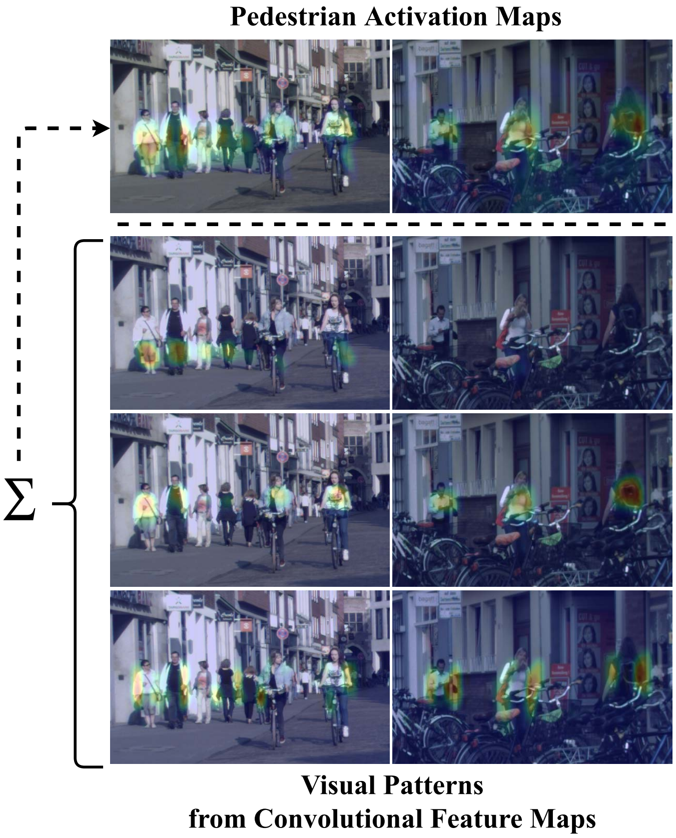


Fig. 1. Pedestrian activation maps (upper) and visual patterns (lower). (Best viewed in color).

(2) We design a feature calibration module, and upgrade a deep detection network to the feature calibration network (FC-Net), which can highlight the visible parts and suppress the occluded parts of pedestrians.

(3) We apply FC-Net on commonly used pedestrian detection benchmarks, and achieve state-of-the-art detection performance with slight computational cost overhead. We also validate the applicability of FC-Net to general object detection.

The remainder of the paper is organized as follows. In Section II, related work about pedestrian detection and occlusion handling is described. In Section III, the implementation details of the SA and FC modules are presented. In Section IV, the learning procedure of FC-Net for pedestrian detection is described. We show the experiments in Section V and conclude the paper in Section VI.

II. Related Work

There is a long history of pedestrian detection research, and various feature representations have been proposed including Histogram of Gradients (HOG) [15], [16], Local Binary Patterns (LBP) [17], Integral Channel Feature (ICF) [18], [19], and informed Haar-like features [20]–[22]. Various sensors including 3-D Range Sensors [23], Near-Infrared Cameras [24], Stereo Cameras [25], CCD Cameras [26] and a combination of them [27] have been employed. In what follows, we mainly review approaches with convolutional neural networks (CNNs) and models about occlusion handing.

A. Pedestrian Detection

With the rise of deep learning, pedestrian detection methods move from hand-crafted features to CNN-based features. Early approaches focus on exploring effective network architectures and hyper-parameters for feature learning [4]. Since 2014, RCNN [28], which integrates high-quality region proposals with deep feature representation, has been leading the object detection area. In the following years, Fast R-CNN [29] and Faster R-CNN [30] were proposed to aggregate feature representation and improve the detection efficiency. By using deep learning features [31]–[34] for general object detection, these approaches have achieved unprecedented good performance. In [8], Zhang *et al.* borrowed the Faster R-CNN framework for pedestrian detection, by increasing the resolution of feature maps and adding hard negative mining modules.

Despite of the effectiveness of these approaches on general object detection, detecting heavily occluded pedestrians remains an open and challenging problem, as indicated by the low performance of existing state-of-the-art approaches (the miss rate is often higher than 50% when the false positive rate per image is 0.1 [9]). The primary reason for the low performance lies in that the occluded parts of pedestrians generate random features which can significantly decrease the representation capability of convolutional features. The problem about how to suppress the features from occluded regions while reinforcing those from visible parts of pedestrians requires to be further investigated.

B. Occlusion Handling

1) Part-Based Models: One major line of methods for occluded pedestrian detection resorts to the part-based model [12], [35]–[38], which leverages a divide-and-conquer strategy, *i.e.*, using different part detectors, to handle pedestrians with different occlusions.

In [39], the Franken-classifiers learns a set of detectors, where each detector accounts for a specific type of occlusion. Zhou *et al.* [40] proposed using multi-label classifiers, implemented by two fully connected layers, to localize the full body and the visible parts of a pedestrian, respectively. Zhang *et al.* [41] proposed CircleNet to implement reciprocating feature adaptation and used an instance decomposition training strategy. In [42], a joint deep learning framework was proposed and multi-level part detection maps were used for estimating occluded patterns. In [43], an occlusion-aware R-CNN (OR-CNN) was presented, with an aggregation loss and a part occlusion-aware region of interest (PORoI) pooling. The authors enforced proposals to be close to the corresponding objects, while integrating the prior structure of human body to predict visible parts.

Although effective, part-based models suffer complex occlusions as the limited number of parts experiences difficulty in covering various occlusion situations. Increasing the number of parts could alleviate such a problem but will increase the model complexity and the computational cost significantly.

2) Attention Models: The other line of methods involves attention-based models [13], which replace "hard" object

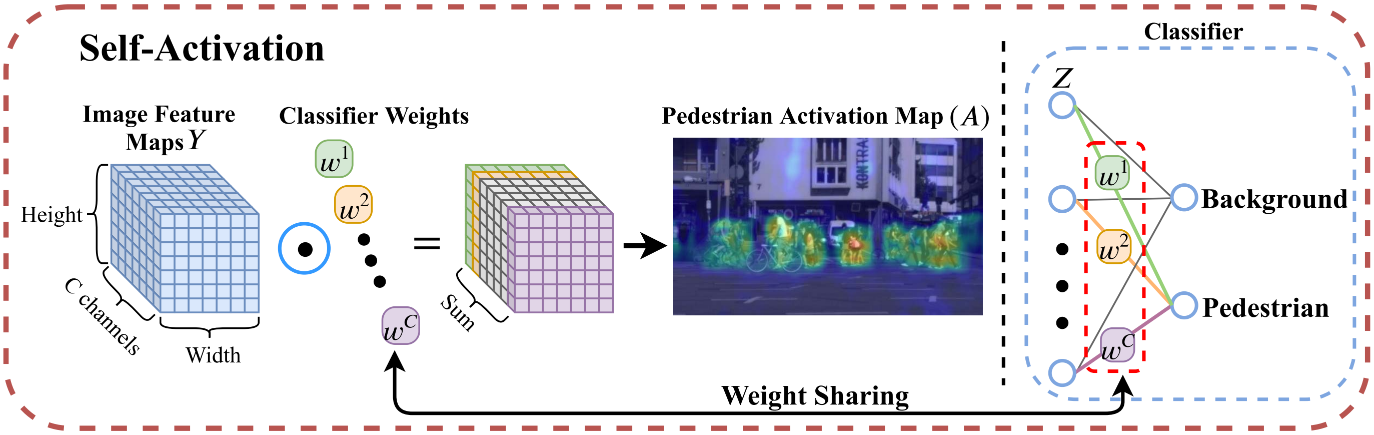


Fig. 2. Self-activation module. (Best viewed in color).

parts with “soft” attention regions by introducing attention or saliency modules [6], [43].

In [13], the Faster R-CNN with attention guidance (FasterRCNN-ATT) was proposed to detect occluded instances. Assuming that each occlusion pattern can be formulated as a combination of body parts, a part attention mechanism was proposed to represent various occlusion patterns by squeezing the features from multiple channels. In [44], the feature learning procedure for pedestrian detection was reinforced with pixel-wise contextual attention based on a saliency network. In [45], a scale-aware pedestrian attention module was proposed to guide the detector to focus on pedestrian regions. The scale-aware attention module targets at exploiting fine-grained details in proper scales into deep convolutional features for pedestrian representation. Thermal images are used to detect pedestrians at night, but not suitable for applications in daytime. Ghose *et al.* [44] used saliency maps to augment thermal images, which is an attention mechanism for pedestrian detectors especially during daytime.

The introduction of attention/saliency has boosted the performance of pedestrian detection. Nevertheless, most existing approaches ignore the class-specific confidence produced by the detection network, and therefore experience difficulty in discriminating the attention regions of positives from those of negatives. In [6], a repulsion loss (RepLoss) approach was designed to enforce pedestrian localization in crowded scenes. With RepLoss, each proposal is forced to be attentive to its designated targets, while kept away from other ground-truth objects. Nevertheless, the discriminative capacity of features is not reinforced despite that the spatial localization is aggregated.

3) Generative Models: Generative methods [46]–[48] have been explored to produce training samples and solve the occlusion problem. The Cycle GAN [46] method transforms synthetic images to real-world scenes for data augmentation. Pedestrian-Synthesis-GAN [48] generates labeled pedestrian data and adopts such data to enforce the performance of pedestrian detectors. Meanwhile structural context descriptor [47] is used to characterize the structural properties of individuals in crowd scenes. A weighted majority voting method [49] inspired by domain adaptation is used to generate labels for other visual tasks.

In this paper, we propose the self-activation approach to explore the class-specific confidence predicted by the detection network. Our approach not only discriminates occluded regions from visible pedestrian parts, but also couples with the detection network to reinforce feature learning in a self-paced manner. The self-activation approach is inspired by class-activation maps (CAMs) [50], a kind of top-down feature activation approach. However, it is essentially different from CAMs as the activation is performed during the feature learning procedure, while that of CAM is performed after the network training is completed. Our work is also related to the squeeze-and-excitation (SE) network [14], which adaptively recalibrates channel-wise feature responses by explicitly modelling inter-dependencies between channels. The difference lies in that our approach leverages the semantic information “squeezed” in the classifier and therefore enforces the discriminative capacity of features more effectively.

III. Feature Calibration

The core of our Feature Calibrating Network (FC-Net) is a self-activation (SA) module, as shown in Fig. 2, which estimates a pedestrian activation map by reusing classifier weights, without any additional parameter involved. The pedestrian activation map is used to manipulate the network with a feature calibration (FC) module in pixel-wised and region-based manners, as shown in Fig. 3 and Fig. 4 respectively. The SA and FC modules are iteratively called during the network training to enforce visible parts while suppressing occluded regions.

A. Self-Activation (SA)

In the Faster-RCNN framework, the pedestrian classifier output, $y(Z) = f(W^T Z + b)$, is made up of a linear model and a nonlinear function. For the binary classification problem, the network has two weight vectors in the fully-connected layer of the classifier, one for the pedestrian and the other for the background. The weight vector for the pedestrian, denoted as $W = (w^1, w^2, \ldots, w^C)^T \in \mathbb{R}^C$, where C is the

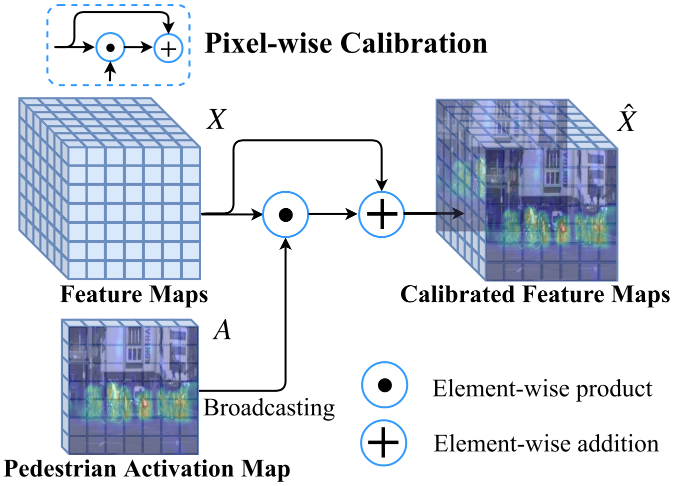


Fig. 3. Pixel-wise calibration. (Best viewed in color).

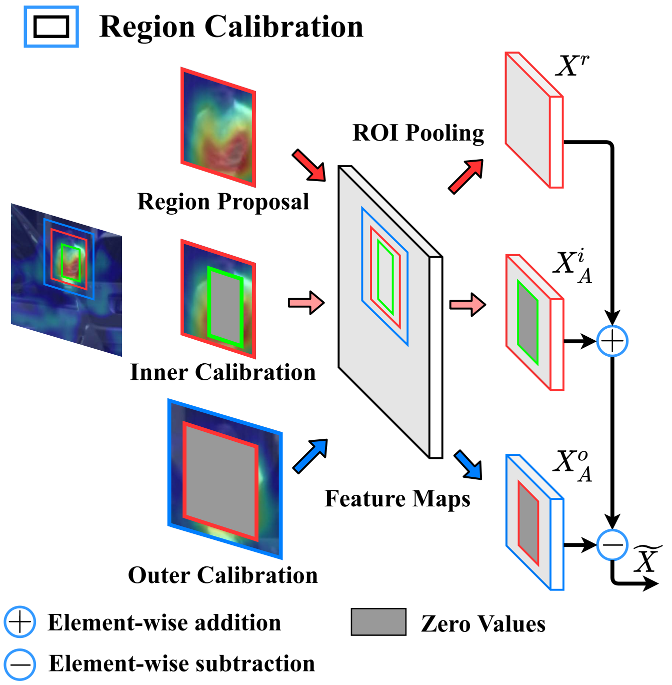


Fig. 4. Region calibration.

number of feature channels, as shown in Fig. 2. Different feature channels represented by the feature Z detect different pedestrian parts, as shown in Fig. 1. To reflect the detected parts in the output $y(Z)$, their corresponding weights must be large, and if some channels only detect background parts, their corresponding weights should be small. This means that the weights actually "squeeze" the channel-wise semantic information for pedestrian representation.

The self-activation module (Fig. 2) reuses the semantic information squeezed in the classifier weights to construct a pedestrian activation map. This procedure is implemented by weighting and summing all of the convolutional feature channels. Specifically, let $Y \in \mathbb{R}^{M\times N\times C}$ be the convolutional feature maps of an image, where M and N respectively denote the width and height of the feature maps. An element $A_{m,n}$ on the pedestrian activation map, $A \in \mathbb{R}^{M\times N}$, is calculated as

$$A_{m,n} = \sum_{c=1}^{C} w^c \cdot Y^c_{m,n}, \tag{1}$$

where m and n denote the 2D coordinates over the feature maps, and c is the index of the feature channels.

The baseline detector is the Faster RCNN equipped with ResNet, which has a global average pooling (GAP) layer after Conv5, as shown in Fig. 5. The GAP layer converts the multiple values of a feature map (channel) to a single value. As a result, multiple feature maps are converted to a vector, which has the same element number with the classifier.

For a pedestrian, different feature channels are sensitive to different parts as the convolutional filters are learned for different visual patterns ($w^c \cdot Y^c$). Benefiting from the fact that the RoI pooling (see Fig. 5 later for the detail) does not change the order of feature channels, the learning procedure constructs a statistical relationship between the feature channels and the weight vector. The larger is a weight element, the more informative is the corresponding feature channel. With Eq. 1, we can aggregate visual patterns into a pedestrian activation map, which indicates the statistical importance of pixels for pedestrian representation. With the pedestrian activation map, we can enforce the features from visible pedestrian parts, as well as depressing occluded regions when the values of either the corresponding feature channels or the weights are small.

B. Feature Calibration (FC)

To make use of the information incorporated into the pedestrian activation map, we follow it with a feature calibration step which aims to aggregate the convolutional features. Towards this goal, such calibration is expected to handle occlusion effectively. First, it should be adaptive to spatial occlusion (in particular, it must be capable of suppressing the channels which output high feature values on occluded regions), and second, it should incorporate the context information so that when an important part of pedestrians is occluded, the region features can still be used for detection.

To meet these requirements, we design pixel-wise calibration and region-based calibration. The former enforces the feature maps to focusing on visible and discriminative parts of pedestrians, while the latter leverages the pedestrian activation map to select the most discriminative regions via introducing multi-level context information.

The pixel-wise calibration reinforces or suppresses the convolutional features in the learning procedure according to the pedestrian activation map. When an important part of pedestrians was occluded, the context regions were validated to provide discriminate information from the perspective of concurrence. For example, pedestrians often stay on the sidewalk or bicycles, but seldom in the air. The region calibration module can leverage the features of context regions for better detection.

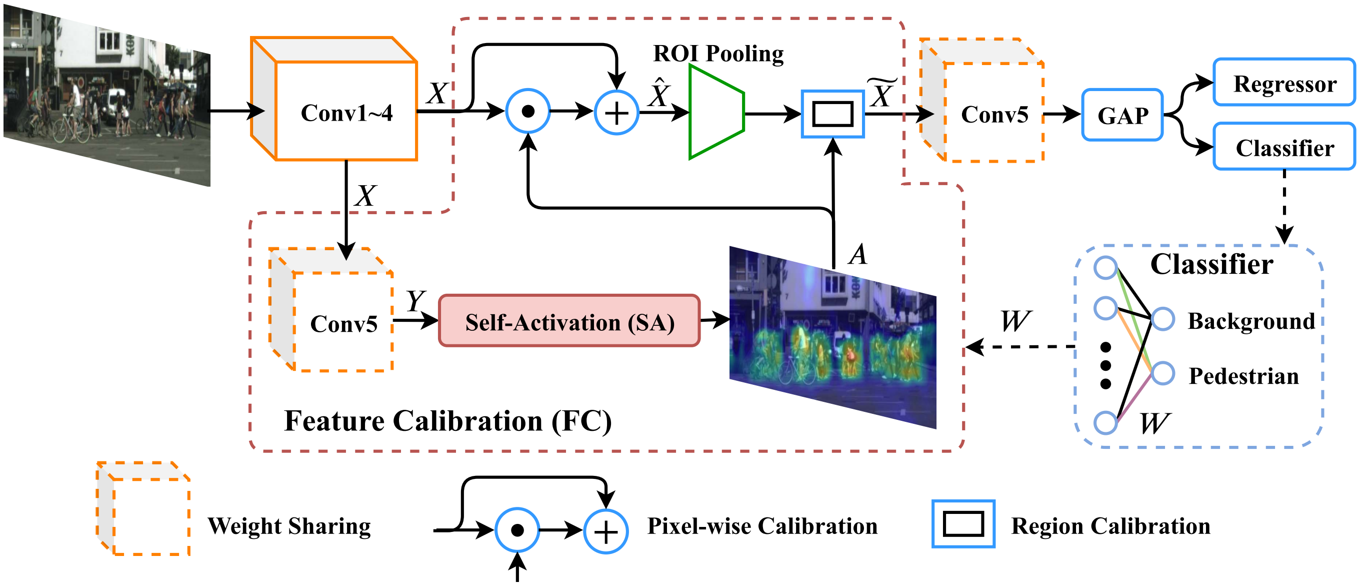


Fig. 5. The architecture of the feature calibration network (FC-Net), which is made up of a deep detection network, a self-activation (SA) module, and a feature calibration (FC) module. After each feed-forward procedure, the detection network learns the classifier weights, which are reused by the SA module for the calibration of the convolutional features. With multiple iterations of the feature calibration, FC-Net learns features which highlight the visible parts and suppress the occluded regions of pedestrians. In the network, GAP stands for global average pooling.

1) Pixel-Wise Calibration: The pixel-wise feature calibration, as shown in Fig. 3, is performed with a pixel-wise product operation and an addition operation. Denoting the feature maps before and after the calibration as $X = \{X^c\}$ and $\hat{X} = \{\hat{X}^c\}$, with c being the channel index, the pixel-wise calibration operation is performed as

$$\hat{X}^c = A \odot X^c + X^c, \ \ c = 1, \ldots, C, \tag{2}$$

where $\odot$ denotes the element-wise product. The calibrated features are then inserted into the network for other feature computation.

Note that the pixel-wise calibration is performed with a product and an addition operations. The product operation converts the occlusion and non-occlusion confidence reflected by the pedestrian activation map to each feature channel. Nevertheless, given pedestrians of various appearances and clutter backgrounds, the pedestrian activation map is not necessarily accurate. The usage of the addition operation keeps the original features and thus smooths the effect of pixel-wise calibration. As the pedestrian activation map (PAM) is calculated by weighting and summing all of the convolutional features, it combines the discriminative information from both classifier weights and features to highlight visible pedestrian parts, in a self-activation fashion. The classifier weights themselves, however, can not indicate visible or occluded parts of pedestrians.

2) Region Calibration: Given the pedestrian activation map, an adaptive context module can be further developed to enhance the feature representation towards good detection.

As shown in Fig. 4, we first define an inner calibration region and an outer calibration region for each region proposal. The outer calibration region is defined as a rectangle similar to the region proposal but with height $= h \times r$ and width $= w \times r$, where h and w are the height and width of the region proposal, respectively, and $r > 1$ is a hyper-parameter. Similarly, the inner calibration region is defined as a rectangle with its height $= h/r$ and width $= w/r$. The inner calibration region is inside the region proposal and covers the area with the largest sum of pixel values on the pedestrian activation map. The outer calibration region covers the region proposal and is with the same center as the inner calibration region. The coordinates of the calibration regions are determined with an exhaustive search around the region proposal.

From the above definitions, we know that the locations of the two calibration rectangles are determined by the pedestrian activation map A. Let X^r, X^i_A, and X^o_A be three features of the same size after RoI pooling. As shown in Fig. 4, both X^r and X^i_A are from the region proposal, but the features of the inner calibration rectangle inside X^i_A are set to 0; and X^o_A is from the outer calibration rectangle, but the features of the region proposal inside X^o_A are set to 0. Then the calibrated feature $\widetilde{X}$ of the region proposal is calculated as

$$\widetilde{X} = X^r + X^i_A - X^o_A. \tag{3}$$

In three cases, we describe the effects of Eq. 3 below. (i) When the region proposal perfectly detects a pedestrian, overall, the feature values of both X^r and X^i_A are large, while those of X^o_A are small. Then $\widetilde{X}$ is enhanced significantly. (ii) When the region proposal only covers part of a pedestrian, overall, all the feature values of X^r, X^i_A, and X^o_A are relatively large, which results in little enhanced/depressed features. (iii) When a pedestrian takes only a small part of the region proposal, overall, all the feature values of X^r, X^i_A, and X^o_A are relatively small, which results in no enhancement of the features. From these effects, we can see that with the guidance of the pedestrian activation map A, the features are calibrated towards good detection.

There are two common factors which cause missing detection of occluded pedestrians. First, the occluded parts introduce significant noises to features, which could confuse the detector towards miss-classification. Second, the features of the visible parts could be not discriminative enough to detect occluded pedestrians, particularly when the background is complex. Leveraging the pedestrian activation map as an indicator, we use the pixel-wise and region calibration modules to enhance features of the visible parts and suppress those of the occluded parts, improving the opportunity to detect occluded pedestrians. When an important part of pedestrians was occluded, the context regions were validated to provide discriminate information from the perspective of concurrence. For example, pedestrians often stay on the sidewalk or bicycles, but seldom or the air. The region calibration module can leverage the features of the context regions for better detection.

The region calibration is fulfilled by a spatial pooling function, which aggregates the features in the context areas. With region calibration, the proposal region features, inner calibration features and outer calibration features are fused. In the procedure, the information within the context region is not removed but fused according to a negative weight, so that the outer region does not cover any pedestrian part. This facilitates improving pedestrian localization accuracy.

IV. Network Structure

Based on the Faster-RCNN framework and the proposed SA and FC modules, we construct the pedestrian object detector, FC-Net, as shown in Fig. 5. Following the last convolutional layer (Conv5) in the detection network, the convolutional features of each region proposal are spatially pooled into a C-dimensional feature vector with a global average pooling (GAP) layer, where C denotes the number of the feature channels. Such a feature vector is then converted into the confidence for a class of object (pedestrian or background), by multiplying it and the weight vector of the fully connected layer with a soft-max operation.

With the feature calibration and network learning, FC-Net works in this way: $X \rightarrow W \rightarrow A \rightarrow \hat{X} \cdots$. During the learning procedure of the network, the features of various pedestrian instances are aggregated to the classifier weight vector W. With the SA and FC modules, the weight vector is employed to generate the activation map A, which is further used to reinforce or depress the features X. With multiple iterations of learning, FC-Net actually implements a special kind of self-paced feature learning. The SA and FC modules are stacked together to form a new architecture, which is universal for deep learning-based object detection.

The proposed SA and FC modules are extremely compressed without additional parameters, involving channel-wise feature calibration, *i.e.*, loosely speaking, $\hat{X} = A \odot X$ and $A = W \cdot X$, where W is borrowed from the detection network. W involves in the forward process and improves the feature learning process of FC-Net, by fully investigating the high-level semantic information "squeezed" in the classifier. During the feature calibration procedure, the semantic information is excited to activate the feature maps so that they can focus on visible pedestrian parts while suppressing occluded regions.

V. Experiments

In this section, we first describe the experimental settings about datasets, evaluation metrics, and implementation details. We then evaluate the effectiveness of the proposed SA and FC modules on the benchmark datasets. Finally, the performance of FC-Net and the comparisons with state-of-the-art pedestrian detectors are presented.

A. Experimental Settings

1) Datasets: Two common datasets, Caltech [51] and CityPersons [10], are used to evaluate FC-Net. The Caltech dataset contains approximately 10 hours of street-view videos taken with a camera mounted on a vehicle. The most challenging aspect of the dataset is the large number of low-resolution and occluded pedestrians. We sample 42,782 images from set00 to set05 for training and 4,024 images from set06 to set10 for testing. The CityPersons dataset is built upon the semantic segmentation dataset Cityscapes [52]. It contains 18 different cities in Germany in three different seasons and various weather conditions. There are 5,000 images, 2,975 for training, 500 for validation, and 1,525 for testing. This dataset is much more "crowded" than Caltech, and the most challenging aspect of the pedestrian objects is heavy occlusion.

2) Evaluation Metric: To demonstrate the effectiveness of FC-Net under various occlusion levels, we follow the strategy in [6] and [43] to define three subsets from the validation set in CityPersons: (i) *Reasonable* (occlusion $< 35\%$ and height > 50 pixels), (ii) *Partial* ($10\% <$ occlusion $< 35\%$ and height > 50 pixels), and (iii) *Heavy* (occlusion $> 35\%$ and height > 50 pixels). The commonly used average-log miss rate $\mathbf{MR}^{-2}$ computed in the False Positive Per Image (FPPI) range of $[10^{-2}, 10^{0}]$ [10] is used as the performance metric.

3) Implementation Details: The baseline detection network is the commonly used Faster R-CNN [30]. It is specified for pedestrian detection by following the settings in [10]. ResNet-50 [53] is used as the backbone network as it is faster and lighter than VGG-16. By using Faster R-CNN as the baseline detection network, we achieve 15.18% $\mathbf{MR}^{-2}$ on the CityPersons validation set, which is sightly better than the reported result, 15.4% $\mathbf{MR}^{-2}$, in [10].

The implementation details of FC-Net are consistent with that of the maskrcnn-benchmark project [54]. We train the network for 6k iterations, with the base learning rate set to 0.008 and decreased by a factor of 10 after 5k iterations on CityPersons. The Stochastic Gradient Descent (SGD) solver is adopted to optimize the network on 8 Nvidia V100 GPUs. A mini-batch involves 1 image per GPU. The weight decay and momentum are set to 0.0001 and 0.9, respectively. We only use single-scale training and testing samples ($\times 1$ or $\times 1.3$) for fair comparisons with other approaches.

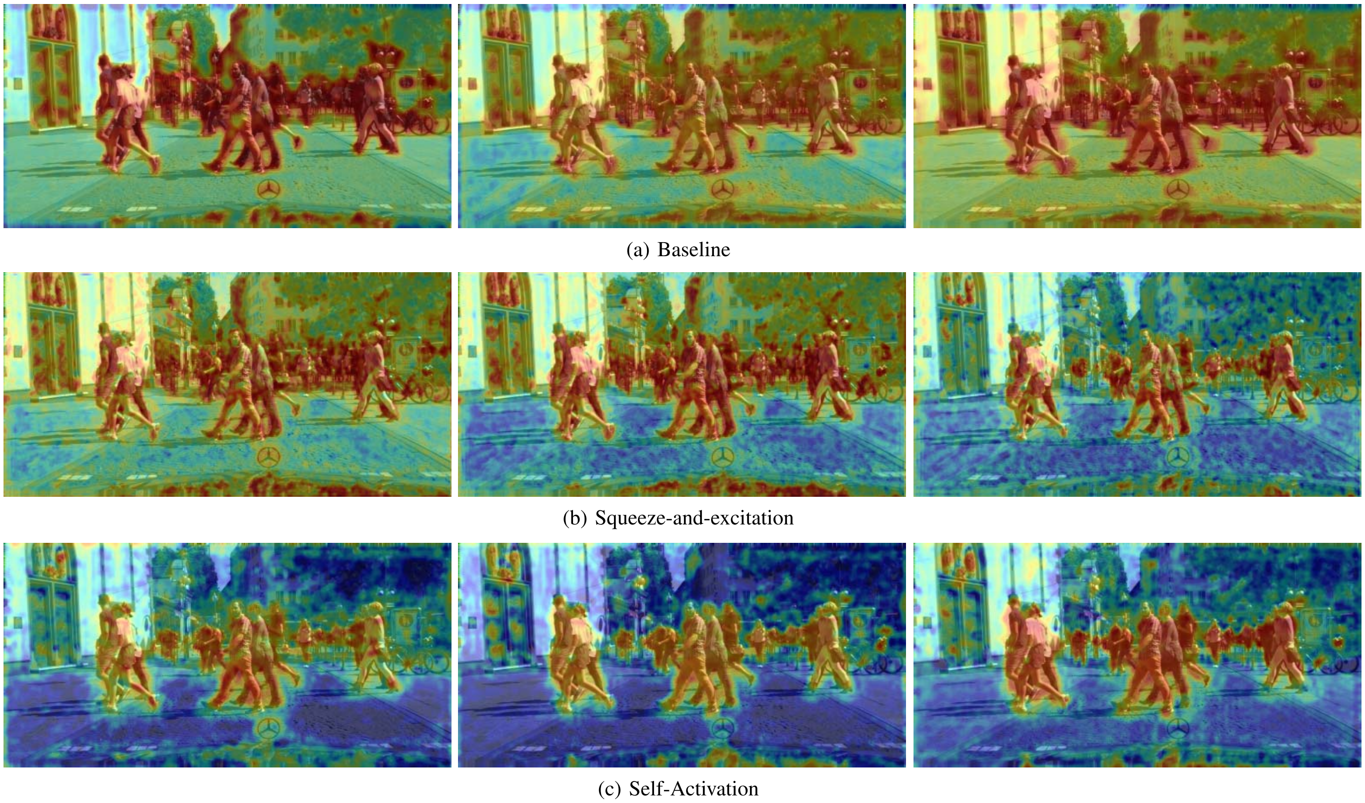


Fig. 6. Comparison of pedestrian activation maps. The proposed SA module can more effectively highlight the visible parts of the pedestrians than the squeeze-and-excitation (SE) network [14]. (Best viewed in color).

B. Self-Activation (SA)

For feature activation, squeeze-and-excitation (SE) is one of the most related works based on the self-attention mechanism to calibrate convolutional features [14]. Our SA module clearly differs from it, without any additional parameter added, and only reuses the weight vector of the classifier to enhance feature learning. As shown in Fig. 6, our SA module can more effectively highlight the visible parts of the pedestrians than the SE network [14]. This shows that the high-level semantic information is crucial to suppress the backgrounds and enhance the foregrounds.

Fig. 6c compares the results of different methods from top to bottom rows. From left to right, we show the results of the same methods at different iterations. Compared with the baseline Faster RCNN [30] (first row) and the SE network [14] (second row), the proposed SA module effectively highlights the pedestrian regions while depressing the background.

As the weights squeeze the statistical importance of pedestrian parts and feature channels, they are used to aggregate the parts/channels into an activation map, which enforces visible parts while depressing occluded parts. Fig. 7 shows the pedestrian activation maps of some non-occluded instances and occluded instances. It can be seen that our approach can adaptively suppress various occluded regions and enforce visible parts of pedestrian objects.

C. Feature Calibration (FC)

1) Pixel-Wise Calibration: In pedestrian detection, *background* is an important factor causing detection errors [4], [6].

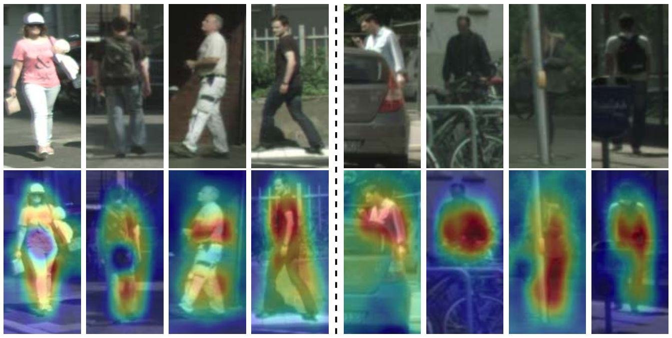


Fig. 7. Examples of pedestrians and pedestrian activation maps. Left: non-occluded instances. Right: occluded instances.

We therefore propose using background error to validate the effect of the FC module. One background error is defined as the case when the intersection over union (IoU) between a detection result and the ground truth is less than 0.2. Fig. 8 compares the error from background before and after using feature calibration. It can be seen that our pixel-wise calibration effectively reduces missed detections and false positives caused by background noise. In Fig. 8, the blue curve shows the background errors of the baseline are very significant, i.e., larger than 70% from $FPPI = 0.056$ to $FPPI = 0.316$. By using our pixel-wise calibration module, the background errors are significantly reduced (black curve), especially from $FPPI = 0.316$ to $FPPI = 1.0$. At $FPPI = 1.0$, FC-Net with the pixel-wise calibration reduces background errors from

TABLE I
ABLATION STUDY OF THE PROPOSED FEATURE CALIBRATION (FC) MODULE ON THE CITYPERSONS VALIDATION DATASET WITH MR^{-2}. SMALLER NUMBERS INDICATES BETTER PERFORMANCE

Method			Scale	*Reasonable*		*Heavy*		*Reasonable+Heavy*	
	Pixel-wise calibration	Region calibration		MR	ΔMR	MR	ΔMR	MR	ΔMR
FC-Net			×1	15.18	-	51.05	-	31.05	-
	√	√	×1	13.93	+1.25	46.79	+4.26	29.64	+1.41
			×1.3	13.21	-	46.75	-	29.45	-
	√		×1.3	12.43	+0.78	43.60	+3.15	27.80	+1.65
		√	×1.3	11.71	+1.50	**41.36**	+5.39	26.70	+2.75
	√	√	×1.3	**11.63**	+1.58	42.77	+3.98	**26.21**	+3.24

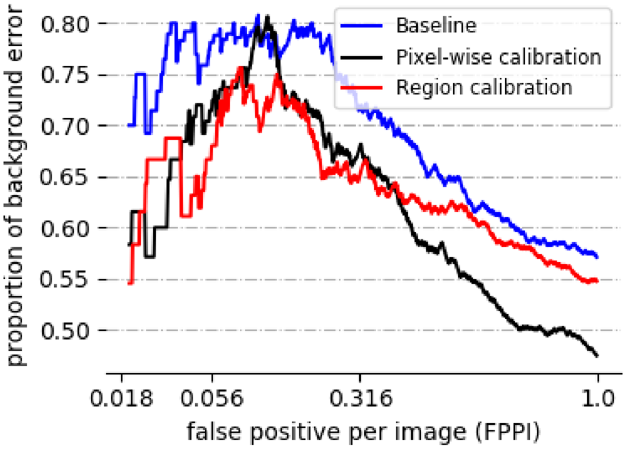


Fig. 8. By applying the proposed feature calibration (FC) module, the proportion of false positives caused by background is significantly reduced.

58% to 48%. This shows that with the pixel-wise feature calibration the background errors are significantly suppressed.

In Table I, we quantitatively evaluate the effect of the pixel-wise feature calibration. Compared with the baseline, FC-Net with the pixel-wise feature calibration reduces 0.78% MR^{-2} at ×1.3 scale on the *Reasonable* subset, 3.15% MR^{-2} on the *Heavy* subset, and 1.65% MR^{-2} on the *Reasonable+Heavy* subset.

2) Region Calibration: By using the region calibration module, the background errors are significantly reduced (red curve in Fig. 8). At FPPI=0.056, FC-Net with the region calibration reduces the proportion of the background errors from 76% to 66%.

In Table I, FC-Net with the region calibration reduces 1.50% MR^{-2} at ×1.3 scale on the *Reasonable* subset, 5.39% MR^{-2} on the *Heavy* subset, and 2.75% MR^{-2} on the *Reasonable+Heavy* subset,

The positions of the inner and outer calibration rectangles are determined by searching the regions of highest value sum on the activation map. What we need to determine is the ratio parameter r empirically. As shown in Table II, by searching in the range [1.0, 2.0], we observe that the best performance is achieved at $r = 1.8$ for height. With the best ratio for height, we further observe that the best width ratio[1] is 1.0 as shown in Table III. The reason of the width ratio smaller than the height ratio is that other pedestrians in horizontal directions may exist in the surrounding regions of a proposal, which may confuse the detector. Note that the pixel-wise calibration does not rely on any context information and therefore is effective even in crowded scenes.

TABLE II
WITH THE WIDTH RATIO = 1, MR^{-2} UNDER DIFFERENT RATIOS FOR HEIGHT BETWEEN THE OUTER RECTANGLE AND THE REGION PROPOSAL, AND BETWEEN THE REGION PROPOSAL AND THE INNER RECTANGLE (SEE FIG. 4)

Ratio (Height)	1.0	1.4	1.6	1.8	2.0
Reasonable	13.21	13.00	12.40	**11.71**	12.78
Heavy	46.75	43.34	42.10	**41.36**	42.09
Reasonable+Heavy	29.45	27.75	27.13	**26.70**	27.11

TABLE III
WITH THE HEIGHT RATIO = 1.8, MR^{-2} UNDER DIFFERENT RATIOS FOR WIDTH BETWEEN THE OUTER RECTANGLE AND THE REGION PROPOSAL, AND BETWEEN THE REGION PROPOSAL AND THE INNER RECTANGLE (SEE FIG. 4)

Ratio (Width)	1.0	1.4	1.6	1.8
Reasonable	**11.71**	12.09	12.90	12.87
Heavy	**41.36**	42.51	43.29	42.98
Reasonable+Heavy	**26.70**	27.53	28.23	28.81

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CITYPERSONS VALIDATION SET WITH MR^{-2}

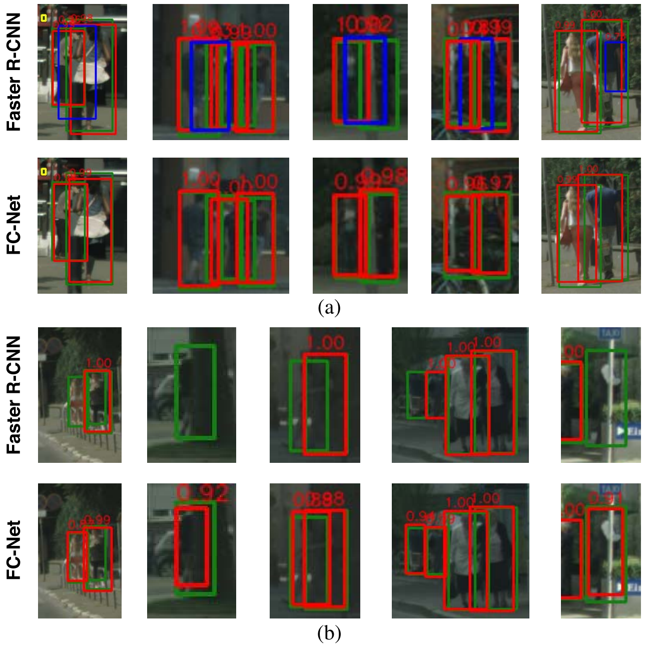
Method	Scale	*Reasonable*	*Heavy*	*Partial*
Adapted FasterRCNN [10]	×1	15.4	-	-
	×1.3	12.8	-	-
Repulsion Loss [6]	×1	13.2	56.9	16.8
	×1.3	11.6	55.3	14.8
OR-CNN [43]	×1	12.8	55.7	15.3
	×1.3	**11.0**	51.3	13.7
CircleNet [41]	×1	-	-	-
	×1.3	11.8	50.2	12.2
AEMS-RPN [55]	×1	13.7	-	-
	×1.3	12.2	-	-
FC-Net (Ours)	×1	13.5	44.3	14.0
	×1.3	11.6	**42.8**	**11.9**

[1] Note that the ratio for width may be different from that for height.

TABLE V

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CITYPERSONS TEST DATASET WITH MR^{-2}. THE RESULTS OF OUR APPROACH ARE EVALUATED BY THE AUTHORS OF CITYPERSONS AND THE COMPARED RESULTS ARE FROM THE OFFICIAL WEBSITE OF CITYPERSONS[2]

Method	Scale	*All*	*Reasonable*	*Reasonable_small*	*Heavy*
Adapted FasterRCNN [10]	×1.3	43.86	12.97	37.24	50.47
Repulsion Loss [6]	×1.5	**39.17**	11.48	15.67	52.59
OR-CNN [43]	×1.3	40.19	**11.32**	**14.19**	51.43
FC-Net (Ours)	×1.3	39.26	12.24	16.67	**41.14**



(a)

(b)

Fig. 9. Comparison of Faster R-CNN and FC-Net on occluded pedestrians. The red bounding boxes indicate correctly detected pedestrians. The blue boxes indicate false positives and the green boxes the ground-truth. (a) FC-Net produces fewer false positives. (b) FC-Net detects more occluded pedestrians than Faster R-CNN.

The width ratio and the height ratio are two hyper-parameters for region calibration. The ablation experiments in Table II and Table III show that the context information in the vertical direction is more important than that in the horizontal direction. The reason could be that there is more concurrence information between pedestrians and backgrounds in the vertical direction. When there are multiple pedestrians in horizontal directions, the context information could be interfered.

D. Occlusion Handling

To show the effectiveness of the proposed SA and FC modules for occlusion handling, we evaluate the detection performance on the validation set of CityPersons where there exist significant person-to-person and car-to-person occlusions.

In Fig. 9, we compare the detection results of Faster R-CNN and FC-Net on occluded samples. It can be seen that FC-Net produces fewer false positives and detects more pedestrians than Faster R-CNN.

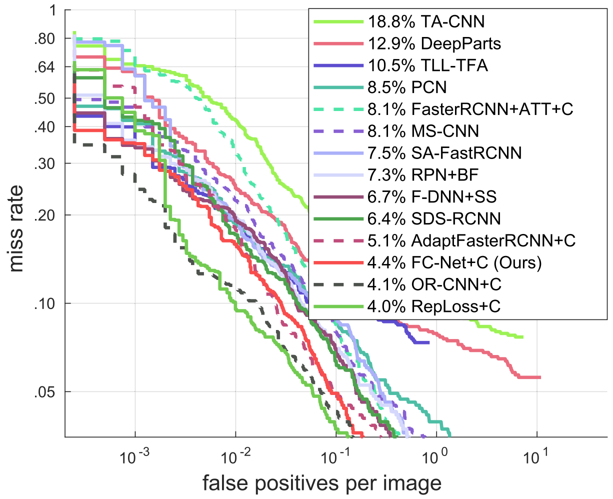


Fig. 10. Comparison with state-of-the-art approaches on the Caltech dataset. "C" indicates models pre-trained on CityPersons. FC-Net achieves 4.4% MR^{-2} and stays on the performance leading board.

E. Performance and Comparison

1) Citypersons Dataset: We compare FC-Net with state-of-the-art approaches including Adapted FasterRCNN [10], Repulsion Loss [6], and OR-CNN [43] on the validation and test sets of CityPersons.

In Table IV, with the ×1.3 scale of the input image, our approach achieves 8.5% and 1.8% lower MR^{-2} than OR-CNN on the *Heavy* subset and the *Partial* subset, respectively, while maintaining a comparable performance on the *Reasonable* subset. With the ×1 scale of the input image, it outperforms OR-CNN by 8.9% and 1.4% on the *Heavy* subset and the *Partial* subset, respectively.

As shown in the last column of Table V, FC-Net outperforms OR-CNN up to **10.29**% (41.14% vs. 51.43%) MR^{-2} on the *Heavy* subset. On the *All* subset, it produces comparable performance to other approaches.

In Table VI, we compare FC-Net with an attention guided approach, FasterRCNN+ATT [13], which is a state-of-the-art approach specified for occluded pedestrian detection. Surprisingly, FC-Net outperforms FasterRCNN+ATT by **9.87%** on the *Heavy* subset and **9.59%** on the *Reasonable+Heavy* subset. It also outperforms FasterRCNN+ATT on the *Reasonable* subset. We implement the attention module of FasterRCNN+ATT in the FC-Net framework (denoted as FC-Net+ATT), and find that FC-Net (ours) also outperforms FC-Net+ATT.

[2] https://bitbucket.org/shanshanzhang/citypersons.

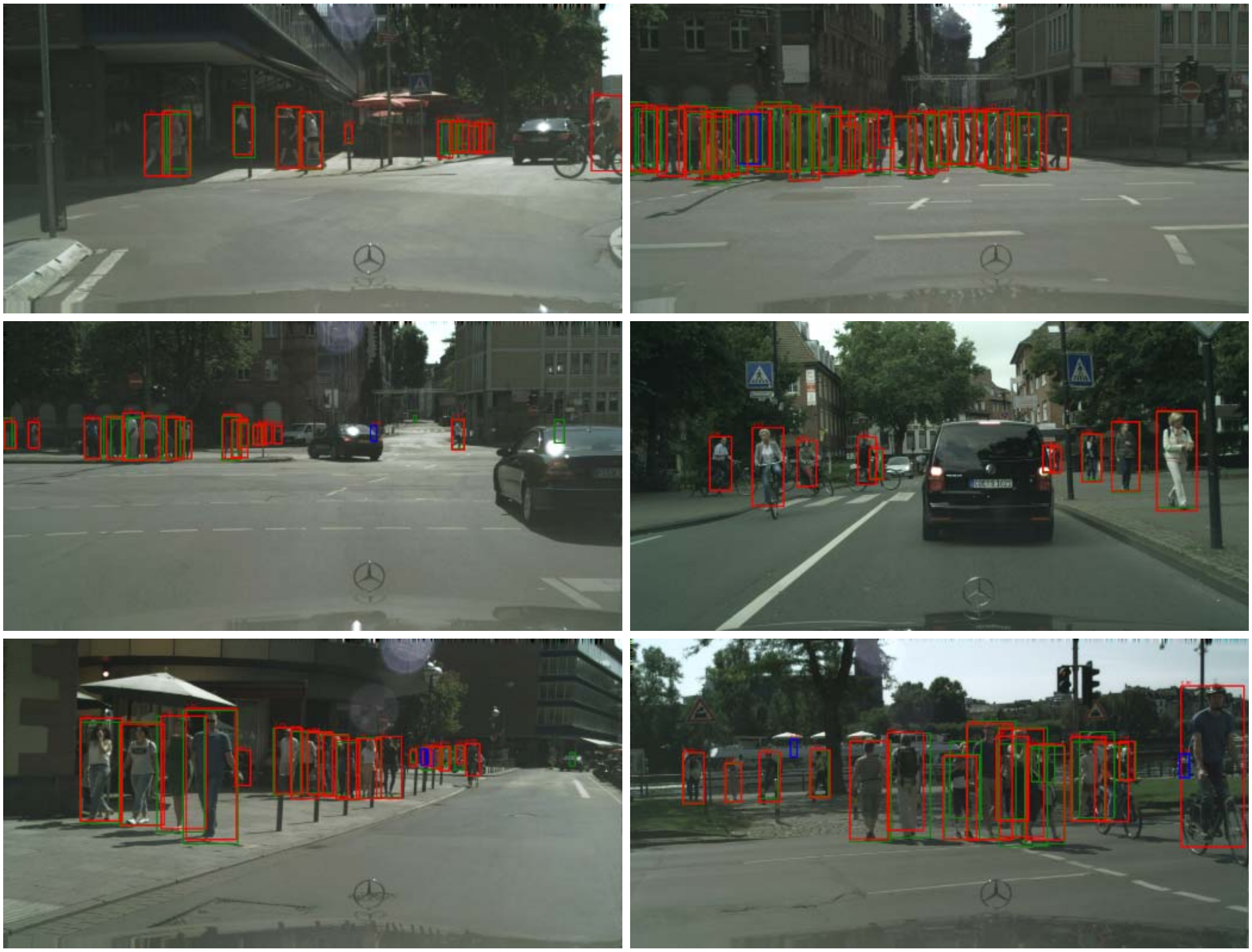


Fig. 11. Examples on the CityPersons. The red, blue and green boxes indicate correctly detected pedestrians, false positives, and ground-truth, respectively.

TABLE VI

COMPARISON WITH THE STATE-OF-THE-ART FASTERRCNN+ATT [13] ON THE CITYPERSONS VALIDATION SET WITH MR^{-2}, WHICH IS AN ATTENTION GUIDED APPROACH SPECIFIED FOR OCCLUDED PEDESTRIAN DETECTION

Method	*Reasonable*	*Heavy*	*Reasonable+Heavy*
FasterRCNN+ATT [13]	15.96	56.66	38.23
FC-Net+ATT	14.82	49.02	31.01
FC-Net (Ours)	**13.93**	**46.79**	**29.64**

TABLE VII

COMPARISON OF CONTEXT MODULES. FOR A FAIR COMPARISON, ALL THE METHODS USE SINGLE SCALE FEATURES

Method	*Reasonable*	*Heavy*	*Reasonable+Heavy*
MS-CNN [56]	13.22	45.27	28.62
MultiPath [57]	12.19	44.04	27.42
FC-Net (Ours)	**11.63**	**42.77**	**26.21**

In Table VII, the proposed region-calibration module is compared with the context model in MS-CNN [56] and MultiPath [57]. It can be seen that the proposed module outperforms them. The reason lies in that our region-calibration, under the guidance of the pedestrian activation map, can adaptively produce inner and outer regions according to the pedestrian activation map. In contrast, those in MS-CNN and MultiPath are not adaptive as they use pre-defined regions.

In Fig. 11, some detection examples on the CityPersons dataset are shown. We find that FC-Net can precisely detect the pedestrians with heavy occlusions. Nevertheless, we also observe some false detections from low-resolution and/or occluded regions. The false detections could be caused by the detector's over-fitting to the few hard positives during training.

2) Caltech Dateset: On this dataset, we use the high quality annotations provided by [4]. Following the commonly used evaluation metric [43], the log-average miss rate over 9 points ranging from 10^{-2} to 10^{0} FPPI is used to evaluate the performance of the detectors. We pre-train FC-Net on CityPersons, and then fine-tune it on the training set of Caltech. We evaluate FC-Net on the *Reasonable* subset of the Caltech dataset, and compare it to other state-of-the-art methods (*e.g.*, [6]–[9], [12], [13], [43], [56], [58]). As shown in Fig. 10, FC-Net, achieving 4.4% MR^{-2}, is on the performance leading board.

TABLE VIII
GENERAL OBJECT DETECTION RESULTS EVALUATED ON PASCAL VOC 2007

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Faster R-CNN [30]	75.8	**82.3**	72.9	56.0	61.2	80.6	85.4	83.7	**55.2**	**82.7**	63.2	**82.0**	84.0	80.9	82.6	**46.5**	73.5	66.8	75.8	**74.2**	73.3
FC-Net	**82.1**	79.1	**74.9**	**61.9**	**63.1**	**81.4**	**85.8**	**85.7**	55.1	78.9	**65.6**	**82.0**	**85.6**	**81.0**	**82.8**	46.4	**75.0**	**72.7**	**80.1**	73.4	**74.6**

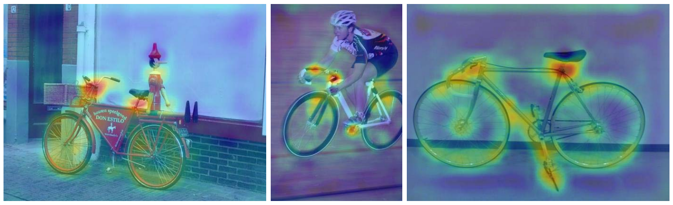


Fig. 12. The activation maps of bikes.

TABLE IX
COMPARISON OF DETECTION SPEEDS ON CITYPERSONS

	Inference Time (second/image)
Faster R-CNN [30]	0.248
FC-Net with SA	0.283
FC-Net with FC	0.284
FC-Net with SA+FC	0.290

F. General Object Detection

In addition to pedestrian detection, the proposed FC-Net is generally applicable to other object detection. To validate it, we test FC-Net on the PASCAL VOC 2007 dataset, which consists of 20 object categories. To implement FC-Net on PASCAL VOC, we still use the Faster R-CNN framework as a baseline, and the RseNet-50 pre-trained on ImageNet as the backbone network. The model is fine-tuned on the training and validation subsets of PASCAL VOC, and is evaluated on the test subset of it. In Table VIII, FC-Net outperforms the baseline by 1.3% mAP. Particularly, it improves the mAPs of "aero", "boats", "sofa" and "train" by 6.3%, 5.9%, 5.9%, and 4.3% respectively, which are significant improvements for the challenging object detection task. The activation maps of some categories, *e.g.*, "bike", have many "holes", as there exist many background pixels within the object regions. This could cause a false enforcement of background features which decrease the detection performance.

G. Detection Efficiency

In Table IX, we compare the test efficiency of FC-Net with the Faster R-CNN baseline. With the superior performance on detecting occluded pedestrians, FC-Net has only a slight computational cost overhead. The SA and FC modules are called once in each training iteration. Therefore, their training iteration number is equal to that of the network. During inference, the SA and FC modules are called once with only an increment of 0.042 second per image.

VI. CONCLUSION

Existing pedestrian detection approaches are unable to adapt to occluded instances while maintaining good performance on non-occluded ones. In this paper, we propose a novel feature learning method, referred to as feature calibration network (FC-Net), to adaptively detect pedestrians with heavy occlusion. FC-Net is made up of a self-activation (SA) module and a feature calibration (FC) module. The SA module estimates a pedestrian activation map by reusing the classifier weights, and the FC module calibrates the convolutional features for adaptive pedestrian representation. With the SA and FC modules, FC-Net improves the performance of occluded pedestrian detection, in striking contrast with state-of-the-art approaches. It is also applicable to general object detection tasks with significant performance gain. The underlying nature behind FC-Net is that it implements a special kind of self-paced feature learning, which can reinforce the features in visible object parts while suppressing those in occluded regions. This provides a fresh insight for pedestrian detection and other general object detection with occlusions.

REFERENCES

[1] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.

[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[3] Q. Ye *et al.*, "Self-learning scene-specific pedestrian detectors using a progressive latent model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2057–2066.

[4] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.

[5] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. 15th Eur. Conf. Comput. Vis.*, Oct. 2018, pp. 764–781.

[6] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[7] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.

[8] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection," in *Proc. 14th Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 443–457.

[9] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4960–4969.

[10] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4457–4465.

[11] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6034–6043.

[12] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1904–1912.

[13] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[18] P. Dollár, Z. Tu, P. Perona, and S. J. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

[19] W. Ke, Y. Zhang, P. Wei, Q. Ye, and J. Jiao, "Pedestrian detection via PCA filters based convolutional channel features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1394–1398.

[20] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 947–954.

[21] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.

[22] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[23] K. Li, X. Wang, Y. Xu, and J. Wang, "Density enhancement-based long-range pedestrian detection using 3-D range data," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 5, pp. 1368–1380, May 2016.

[24] Y.-S. Lee, Y.-M. Chan, L.-C. Fu, and P.-Y. Hsiao, "Near-infrared-based nighttime pedestrian detection using grouped part models," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1929–1940, Aug. 2015.

[25] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 380–391, Sep. 2009.

[26] X.-B. Cao, H. Qiao, and J. Keane, "A low-cost pedestrian-detection system with a single optical camera," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 58–67, Mar. 2008.

[27] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 4, pp. 619–629, Dec. 2007.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[29] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[30] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[31] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[32] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[33] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2395–2409, Oct. 2019.

[34] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-MIL: Continuation multiple instance learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2199–2208.

[35] A. Prioletti, A. Mogelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. B. Moeslund, "Part-based pedestrian detection and feature-based tracking for driver assistance: Real-time, robust algorithms, and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1346–1359, Sep. 2013.

[36] J. Xu, D. Vazquez, A. M. Lopez, J. Marin, and D. Ponsa, "Learning a part-based pedestrian detector in a virtual world," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2121–2131, Oct. 2014.

[37] M. Pedersoli, J. Gonzalez, X. Hu, and X. Roca, "Toward real-time pedestrian detection based on a deformable template model," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 355–364, Feb. 2014.

[38] W. Liu, B. Yu, C. Duan, L. Chai, H. Yuan, and H. Zhao, "A pedestrian-detection method based on heterogeneous features and ensemble of multi-view–pose parts," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 813–824, Apr. 2015.

[39] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1505–1512.

[40] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3506–3515.

[41] T. Zhang, Z. Han, H. Xu, B. Zhang, and Q. Ye, "CircleNet: Reciprocating feature adaptation for robust pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4593–4604, Nov. 2020.

[42] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063.

[43] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. 15th Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 657–674.

[44] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 988–997.

[45] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. 15th Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 745–761.

[46] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8198–8207.

[47] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4179–4188.

[48] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, and P. Zhou, "Pedestrian-synthesis-GAN: Generating pedestrian data in real scene and beyond," 2018, *arXiv:1804.02047*. [Online]. Available: http://arxiv.org/abs/1804.02047

[49] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, "Domain-weighted majority voting for crowdsourcing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 163–174, Jan. 2019.

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[51] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[52] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] F. Massa and R. Girshick. (2018). *Maskrcnn-Benchmark: Fast, Modular Reference Implementation of Instance Segmentation and Object Detection Algorithms in PyTorch*. [Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark

[55] H. Wang, Y. Li, and S. Wang, "Fast pedestrian detection with attention-enhanced multi-scale RPN and soft-cascaded decision trees," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 12, pp. 5086–5093, Dec. 2019.

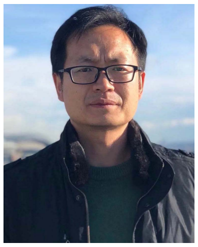
[56] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. 14th Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 354–370.

[57] S. Zagoruyko *et al.*, "A MultiPath network for object detection," in *Proc. Procedings Brit. Mach. Vis. Conf.*, Aug. 2016.

[58] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5079–5087.

Tianliang Zhang (Student Member, IEEE) received the B.S. degree in electronic information engineering from the Wuhan University of Technology (WUT) in 2013, and the M.S. degree in industrial engineering from the University of Chinese Academy of Sciences in 2017, where he is currently pursuing the Ph.D. degree with the School of Electronic, Electrical, and Communication Engineering. His research interests include visual object detection and deep learning.

Qixiang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees from the Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. He has been a Professor with the University of Chinese Academy of Sciences since 2016, and was a Visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, until 2013. His research interests include image processing, visual object detection, and machine learning. He has published more than 50 papers in refereed conferences and journals, including the IEEE CVPR, ICCV, ECCV, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is on the Editorial board of *The Visual Computer Journal* (Springer).



Baochang Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of the Technology, Harbin, China, in 1999, 2001, and 2006, respectively. From 2006 to 2008, he was a Research Fellow with The Chinese University of Hong Kong, Hong Kong, and with Griffith University, Brisbane, Australia. He is currently an Academic Advisor at the Institute of Deep Learning, Baidu Research. He has published more than 50 papers in refereed conferences and journals. His research interests include pattern recognition, machine learning, face recognition, and wavelets.

Jianzhuang Liu (Senior Member, IEEE) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong, Hong Kong, in 1997. From 1998 to 2000, he was a Research Fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a Post-Doctoral Fellow, an Assistant Professor, and an Adjunct Associate Professor with The Chinese University of Hong Kong. In 2011, he joined the Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, as a Professor. He is currently a Principal Researcher with Huawei Technologies Company Ltd., Shenzhen. He has authored over 150 articles. His research interests include computer vision, image processing, machine learning, multimedia, and graphics.

Xiaopeng Zhang received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University in 2017, under the supervision of Prof. H. Xiong and Prof. Q. Tian. He is currently a Senior Researcher with Cloud & AI, Huawei Technologies. Before that, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, from 2017 to 2019, and a member of the Learning and Vision Lab, under the supervision of Jiashi Feng and Shuicheng Yan.

Qi Tian (Fellow, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, the M.S. degree in ECE from Drexel University, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC). He is currently a Chief Scientist in Artificial Intelligence at Cloud BU, Huawei. From 2018 to 2020, he was the Chief Scientist in Computer Vision at Huawei Noah's Ark Lab. He was also a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA), from 2002 to 2019. From 2008 to 2009, he took one-year faculty leave at Microsoft Research Asia (MSRA). His research interests include computer vision, multimedia information retrieval, and machine learning, and published more than 590 refereed journal and conference papers. His Google citation is over 24100 with H-index 76. He was the coauthor of best papers, including IEEE ICME 2019, ACM CIKM 2018, ACM ICMR 2015, PCM 2013, MMM 2013, ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, a Student Contest Paper in ICASSP 2006, and coauthor of a Best Paper/Student Paper Candidate in ACM Multimedia 2019, ICME 2015, and PCM 2007. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He received the 2017 UTSA President's Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Awards from the College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, ACM TOMM, and MMSJ, and in the Editorial Board of *Journal of Multimedia* (JMM) and *Journal of Machine Vision and Applications*. He is a Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, and *Journal of Computer Vision and Image Understanding*.