CircleNet: Reciprocating Feature Adaptation for Robust Pedestrian Detection

Tianliang Zhang, *Student Member, IEEE*, Zhenjun Han, *Member, IEEE*, Huijuan Xu, *Member, IEEE*, Baochang Zhang, *Member, IEEE*, and Qixiang Ye, *Senior Member, IEEE*

***Abstract*— Pedestrian detection in the wild remains a challenging problem especially when the scene contains significant occlusion and/or low resolution of the pedestrians to be detected. Existing methods are unable to adapt to these difficult cases while maintaining acceptable performance. In this paper we propose a novel feature learning model, referred to as CircleNet, to achieve feature adaptation by mimicking the process humans looking at low resolution and occluded objects: focusing on it again, at a finer scale, if the object can not be identified clearly for the first time. CircleNet is implemented as a set of feature pyramids and uses weight sharing path augmentation for better feature fusion. It targets at reciprocating feature adaptation and iterative object detection using multiple top-down and bottom-up pathways. To take full advantage of the feature adaptation capability in CircleNet, we design an instance decomposition training strategy to focus on detecting pedestrian instances of various resolutions and different occlusion levels in each cycle. Specifically, CircleNet implements feature ensemble with the idea of hard negative boosting in an end-to-end manner. Experiments on two pedestrian detection datasets, Caltech and CityPersons, show that CircleNet improves the performance of occluded and low-resolution pedestrians with significant margins while maintaining good performance on normal instances.**

***Index Terms*— CircleNet, feature learning, pedestrian detection, traffic scenes.**

I. Introduction

PEDESTRIAN detection is an important problem in intelligent transportation with many real-world applications in intelligent surveillance systems platforms, driver assistant systems, and autonomous vehicles [1]–[6]. Although extensively investigated, robust pedestrian detection at a long distance with a single low-cost camera remains unsolved.

Deep learning methods have achieved unprecedented success on visual object detection; nevertheless, they have trouble with adapting difficult pedestrian instances without sacrificing performance on normal instances [7]–[12]. Take traffic scenes for example, systems have difficulty with instances that are occluded and/or have low-resolution objects, and with scenes in cluttered backgrounds [7], [13].

To detect these occluded and/or low-resolution pedestrians, it is natural to consider taking higher-resolution features in the lower-level feature pyramid to assist detection using deep convolutional neural network (CNN) [14], [15]. The Feature Pyramid Network (FPN) [14], for example, introduces an additional network branch from top to bottom to increase the feature representation capacity at lower layers (Fig. 1a). The Path Aggregation Network (PANet) [15] adds a path augmentation from bottom to top on the basis of FPN to further boost the information flow for feature representation (Fig. 1b). These approaches improve the performance of low-resolution instances; however, instances with heavy occlusion remain challenging.

As humans, we may not be able to recognize the hard instances with only one glance, but need to focus on the specific area of the image in question. Inspired by this, we propose a feature learning framework, CircleNet, shown in Fig. 1c, which reciprocates the feature adaptation by formatting multiple top-down and bottom-up feature fusion pathways to enhance the feature representation for occluded and low-resolution objects. When more information pathways incorporated, we need to pay attention to the model capacity for these additional layers due to the fact that more parameters increase the chance of over-fitting. We propose using a weight sharing policy for these additional layers and experimentally validate that sharing the network parameters of the top-down/bottom-up pathways across repeating steps can maintain a healthy balance between model capacity and the generalization ability.

CircleNet is a general network architecture with FPN and PANet as its special cases. It naturally mimics the functionality of the cognitive phenomenon that the long-latency responses of the neurons contain many levels of forward and backward pipelines in the information processing of the early visual cortex [16]. CircleNet achieves feature adaptation by mimicking the process humans looking at low resolution and occluded objects: focusing on it again, if the object cannot be identified clearly at the first time. More specifically, CircleNet uses a circling structure to implement feature adaptation, which means that pedestrians of various appearance are handled in different circles. The circling structure is constructed by a top-down branch and a bottom-up branch, which fuses features in a reciprocating manner and facilitates learning adaptive feature representation.

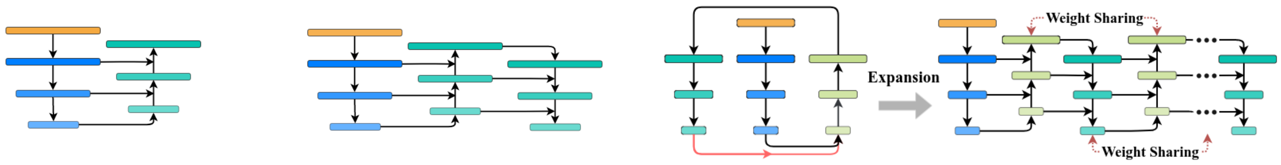
Manuscript received January 13, 2019; revised May 22, 2019 and September 9, 2019; accepted September 12, 2019. This work was supported in part by the NSFC under Grant 61836012, Grant 61671427, and Grant 61771447 and in part by the Beijing Municipal Science and Technology Commission under Grant Z181100008918014. The Associate Editor for this article was W. Lin. *(Corresponding author: Qixiang Ye.)*

T. Zhang, Z. Han, and Q. Ye are with the School of Electronics, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 108408, China (e-mail: zhangtianliang17@mails.ucas.ac.cn; hanzhj@ucas.ac.cn; qxye@ucas.ac.cn).

H. Xu is with the Department of Computer Science, University of California at Berkeley, Berkeley, CA 94720 USA.

B. Zhang is with the Department of Automation, Beihang University, Beijing 100191, China.

Digital Object Identifier 10.1109/TITS.2019.2942045



(a) FPN

(b) PANet

(c) CircleNet (Ours)

Fig. 1. From feature pyramid network to CircleNet. (a) Feature pyramid network (FPN) [14]. (b) Path Aggregation Network (PANet) [15]. (c) Our proposed CircleNet is implemented as a set of feature pyramids using weight sharing path augmentation. (Lateral connection paths are illustrated in Fig. 2.)

The most relevant approach, Path Aggregation Network (PANet) [15] adds one path augmentation from bottom to top on the basis of FPN to boost the information flow for feature representation. However with a single bottom-up path and without parameter sharing, PANet is still limited in learning adaptive features. Our CircleNet includes more cycles with multiple information flows, where useful information can flow on parallel circles or vertical paths in an adaptive manner, as shown in Fig. 1c.

The other relevant work, Cascade R-CNN [17], consists of a sequence of detectors trained with increasing IoU thresholds, to be sequentially more selective against close false positives. Nevertheless, Cascade R-CNN requires more learnable parameters, which take the risk of over-fitting given limited training samples. In contrast, the proposed CircleNet shares the network parameters across repeating steps, and thereby maintains a healthy balance between model capacity and the generalization ability.

To take advantage of the unique feature representation in each cycle and boost the feature adaptation capability, we further propose using instance decomposition during the training and test phases. This idea is motivated by the preliminary observations that shallow layers benefit small-scale object detection while deep cycles benefit occluded object detection. In the instance decomposition procedure, training samples are partitioned into sub-sets according to their resolution and occlusion rates, and are assigned to various circles and feature maps. As cycling goes on, Fig. 1c, we obtain adaptive feature representations corresponding to the sample distributions of various instance decomposition groups, and the network can be viewed as an ensemble classifier. Each branch will focus on detecting objects of a specific scale or occlusion ratio, and the CircleNet shares the backbone network and operates in an ensemble mode with a boosting strategy.

The contributions of the paper are as follows:

1. A novel feature learning architecture, referred to as CircleNet, which fuses the features in a reciprocating manner and improves the capability and adaptability of feature representation.

2. An instance decomposition strategy designed for CircleNet, which makes full use of adapted and unique feature representations at each cycle for low-resolution and occluded pedestrian detection.

3. The demonstration of state-of-the-art performance for occluded and/or low-resolution pedestrians on two standard benchmarks.

The remainder of the paper is organized as follows. Related works are reviewed in Section II. The CircleNet architecture is described in Section III and pedestrian detection with CircleNet is presented in Section IV. A discussion about CircleNet is given in Section V. Experiments are presented in Section VI and conclusions are given in Section VII.

II. Related Work

In the section, we first review the commonly used pedestrian detection methods. We then mainly review CNN-based pedestrian detection approaches and analyzed recent methods dealing with low-resolution and occluded issues.

A. Pedestrian Detection

As one of the most import object sensing tasks, pedestrian detection has been extensively investigated during the past decades. Various sensors including 3-D Range Sensors [18], Near-Infrared Cameras [19], Stereo Cameras [20], CCD Cameras [21] and a combination of them [22] have been employed. For visual pedestrian detection, various hard-crafted and learning-based visual features including Histogram of Gradients (HOG) [23], [24], Local Binary Patterns (LBP) [25], Aggregated Channel Features [26], [27], Informed Haar-like features [28], [29], Rectangle Feature [30], and Convolutional Neural Network (CNN) [31] have been been explored. With the visual feature presentations, SVM [32], Random Forest [33], and cascaded models [21], [34] were used as classifiers.

As a major branch of visual pedestrian detection method, part-based model was widely explored by first dividing a pedestrian object into parts and then. training part-based models to detect pedestrians of various postures [35]–[38]. Integrated with deep learning features, part-based models have shown great advantages to detect mult-view and multi-posture pedestrians, but remained challenging for pedestrians objects with low-resolution and heavy occlusion.

Another branch of pedestrian detection method rooted in feature/classifier ensemble. With tree classifier ensemble, Xu *et al.* [39] targeted at achieving not only high detection accuracy but also high detection speed. With error correcting output code classification of manifold sub-classes, Ye *et al.* [32] can robustly detect multi-view and multi-posture pedestrians. In [40], the cascade implementation of the additive KSVM (AKSVM) was proposed for the application of pedestrian detection. AKSVM avoided kernel expansion by using look-up tables, and it was implemented in cascade form, thereby speeding up pedestrian detection.

B. CNN-Based Pedestrian Detection

Early CNN-based pedestrian detection [2], [41] was primarily based on the RCNN structure [31], [42], which relies on high-quality object proposals to achieve pedestrian localization and detection. More recently the Faster R-CNN [43] architecture has become popular as it integrates region proposals with object classification for end-to-end learning. In [44] and [45], Tiny-DSOD and AP-loss approaches were proposed to reduce the computational resource while maintaining the accuracy for detection. AP-Loss can also alleviate the extreme foreground-background class imbalance issue caused by the large number of anchors.

By borrowing general object detection frameworks to tackle pedestrian detection, these approaches have already achieved unprecedented performance. Nevertheless, detecting low-resolution and 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 20% when false positive rate per image is 0.01 [13]).

C. Low-Resolution Pedestrian Detection

To handle low-resolution objects, a number of approaches have explored using hierarchical features in CNNs, such as SSD [46], MS-CNN [47], SA-FastRCNN [48], and FPN [14]. These methods leveraged fused hierarchical features to aggregate the receptive fields and representative capability from different layers.

Semantic segmentation [10], temporal features, and depth information [12] have also been explored to address the problem of low-resolution. Adaptive Faster R-CNN [11] was used to process low-resolution instances by optimizing the scales and aspect ratios of region proposals and the resolution of the image. A cascaded Boosted Forest (BF) was applied on up-sampled feature maps to detect low-resolution pedestrians [9]. In [49], fine-grained information was incorporated into features to make them more discriminative for human body parts. In [50], topological line localization (TLL) and temporal feature aggregation were used to detect low-resolution pedestrians.

D. Occlusion Handling

One effect of occlusion is that it significantly aggravates the appearance of pedestrian instances. To address this issue, one simple strategy is to use a classifier ensemble or detector [51], [52]. In [17], the Cascade R-CNN was proposed to address these problems. It consists of a sequence of detectors trained with increasing the Intersection over Union (IoU) thresholds, to sequentially select difficult instances.

Another effect of occlusion is that it causes the loss of certain pedestrian parts and significantly increases the difficulty of spatial localization. To address the problem, a pedestrian object was decomposed into parts, which can be used to model the partial occlusion of objects and the attention of classifiers and features. Tian *et al.* [52] proposed the DeepPart model where each part detector is a strong detector that can detect pedestrian by observing only a part of a proposal. In [7] a repulsion loss (RepLoss) approach was designed to enforce spatial localization in crowd scenes. With RepLoss, each proposal is forced to be close to its designated target, and keep it away from other ground-truth objects. 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 used to represent various occlusion patterns in one single model.

Existing approaches have included effective strategies to process low-resolution and occluded instances, but often require empirically defined strategies. Some of them achieve higher performance on difficult instances but get lower performance on normal instances. In this paper we propose to mimic the human cognition processes to handle these difficult instances without sacrificing the performance on normal instances. The proposed CircleNet leverages a recurrent structure with parameter sharing to realize feature adaptation. The top-down and bottom-up branches in multiple circles facilitate fusing features in a cognitively plausible way. In contrast, PANet [15] adds only one bottom-up path and do not use parameter sharing.

E. Hard Instance Handling

OHEM [53] is a widely used to handle the hard instances. It mines hard examples according to the training loss of instances and makes the learned models more discriminative. Our instance decomposition strategy is different from OHEM. It is based on the heuristics that the normal instances and hard instances have different feature representations. An instance decomposition strategy is used with the deep circles paying more attention to hard examples while the shallow circles paying more attention to normal instances. In contrast, OHEM uses the same features for all hard and normal instances.

With larger receptive fields, deep circles are easy to detect occluded pedestrians than shallow ones. Existing works [54]–[56] have explored context information for object detection, and find that appropriate context is helpful for detecting objects with occlusion. We take advantage of more context information to learn features that are adaptive to hard pedestrian examples. Considering the property of context information and the circling architecture, we argue that the proposed instance decomposition strategy is effective in the CircleNet structure.

III. CircleNet

In this section, we first introduce the architecture of CircleNet. We then describe the reciprocating feature adaptation and instance decomposition performed with CircleNet.

A. Architecture

CircleNet is made up of a backbone network and a circling architecture (Fig. 2). The top-down branches enforce semantics by up-sampling and concatenating deep-layer features. The bottom-up branches enlarge the receptive field and incorporate context information by down-sampling and concatenating features from shallow layers. The top-down and bottom-up branches construct a circling network that fuses features in a reciprocating manner. The backbone network is a commonly

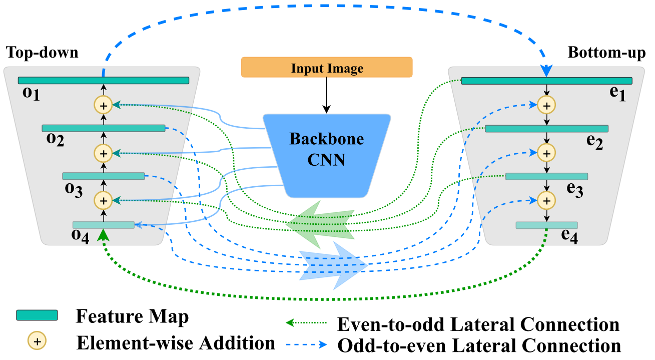


Fig. 2. CircleNet architecture. CircleNet is made up of a backbone network and top-down and bottom-up branches. The top-down branches enforce semantics by up-sampling and concatenating deep-layer features. The bottom-up branches enlarge the receptive field and incorporate context information by down-sampling and concatenating features from shallow layers. The top-down and bottom-up branches construct a circling network that fuses features in a reciprocating manner.

used fully convolutional network, which computes a feature map hierarchy consisting several scales with a scaling factor of 2.

For the top-down (odd) pathway, let $o_n, n = 1, \dots, N$, denote the n^{th} feature map, and o_{n+1} and e_n denote the input and the side-output, respectively. e_n stores the information from the current layer and o_{n+1} brings more high level information from the deep layers. Similarly, for the bottom-up (even) pathway, let e_{n+1} denote the $(n+1)^{th}$ feature map, and e_n and o_{n+1} denote the input and the side-output, respectively. o_{n+1} stores the information from the current layer and e_n is used to generate new features. Cascading the multiple top-down and bottom-up pathways together is implemented as a cascaded feature pyramid network.

To upgrade the cascading architecture to a circling one, we share weights for all top-down pathways and for all bottom-up pathways from different circles, as shown in Fig. 1c. In this way, we keep a minimal number of learnable parameters and learn features that are adaptive to different instances. This means we only need to learn weights for an odd branch and an even pathway $\{w^e, w^o\}$. By weight sharing, we update the cascading architecture to a circling architecture, which learns different features in each circle t as:

$$\mathbf{x}_t = F(\mathbf{x}_{t-1}, w^e, w^o), \tag{1}$$

where $\mathbf{x}_t = \{o_n^t, e_n^t\}$, $w^e = \{w_n^e\}$, and $w^o = \{w_n^o\}$, $n = 1, \dots, N, t = 1, \dots, T$. n is the index of the feature maps and t is the index of circles (see Section III-B for details). $F(\cdot)$ denotes the feature extraction function.

B. Reciprocating Feature Adaptation

The circling architecture implements feature fusion and adaptation in a reciprocating manner by concatenating multiple top-down and bottom-up pathways. Each pathway uses lateral connections to fuse features which come from the backbone network or a pathway. The top-down pathway follows the implementation of FPN [14]. In the feed-forward procedure of any circle, the feature of the n^{th} top-down layer at t^{th} circle, o_n^t, is generated by fusing the features from the backbone CNN and the top-down layer o_{n+1}, or features from the bottom-up layer e_n^t and o_{n+1}, as

$$\begin{aligned} o_n^t &= F_n(e_n^t, o_{n+1}^t, w_n^e) \\ &= w_n^{e11} * (w_n^{e33} * e_n^t) + \uparrow o_{n+1}^t, \end{aligned} \tag{2}$$

where $w_n^e = \{w_n^{e11}, w_n^{e33}\}$ are 1×1 and 3×3 convolutional filters to fuse features and generate o_n^t. $\uparrow$ denotes the up-sampling operation.

Similarly, the features of the $(n+1)^{th}$ bottom-up layer, e_{n+1}^t, are generated by fusing the features from the bottom-up layer o_{n+1}^{t-1} and its previous bottom-up layer e_n^t, as

$$\begin{aligned} e_{n+1}^t &= F_n(e_n^t, o_{n+1}^{t-1}, w_{n+1}^o) \\ &= w_{n+1}^{o11} * (w_{n+1}^{o33} * o_{n+1}^{t-1}) + \downarrow e_n^t, \end{aligned} \tag{3}$$

where $w_n^o = \{w_n^{o11}, w_n^{o33}\}$ are 1×1 and 3×3 convolutional filters to fuse features and generate e_{n+1}. $\downarrow$ denotes the down-sampling operation.

The up-sampling operation is implemented by the nearest neighbor interpolation [57]. Nearest-neighbor interpolation (also known as proximal interpolation or, in some contexts, point sampling) is a simple method of multivariate interpolation in one or more dimensions. The nearest neighbor interpolation selects the value of the nearest point which does not consider the values of neighboring points, yielding a piecewise-constant interpolant. The down-sampling operation is implemented by a strided convolution.

In the feed-forward procedure, the top-down pathways hallucinate higher resolution features by up-sampling spatially coarser but semantically stronger feature maps from higher pyramid levels. These features are then enhanced with features from either the backbone network or the bottom-up pathway, via lateral connections. Each lateral connection merges feature maps of the same spatial size from the bottom-up pathway and the top-down pathway. The top-down pathways further fuse features by reducing their resolution but enforcing their semantics. Reducing the resolution of feature maps can enlarge the receptive field and collect context information, which is crucial to detect low-resolution and occluded instances.

By using multiple top-down and bottom-up pathways as shown in Eqs. 2 and 3, the CircleNet implements reciprocating feature fusion that provides a higher probability of producing features with both strong semantics and context information.

C. Instance Decomposition

With reciprocating feature fusion, the circling architecture has the potential to learn features that are adaptive to pedestrian instances of various appearance. This motivates us to empirically decompose the instances into different feature maps on the circles, so that normal and hard instances are handled with proper features. In experiments, we observed that deep circles pay more attention to samples with occlusion, while for samples with no occlusion the detectors perform well in the shallow circles. Thereby, an instance decomposition strategy is proposed so that difficult instances are decomposed into deep circles according to the training loss or the occlusion ratio. Low-resolution instances are decomposed into shallow feature maps.

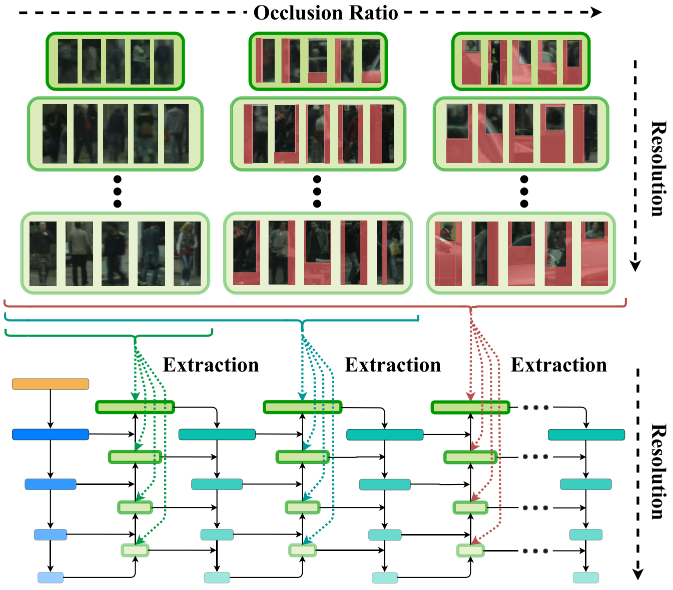


Fig. 3. An illustration of instance decomposition during the training phase. (Best viewed in color and zoomed in.)

Empirically, we use circle-based and layer-based instance decomposition and partition the training set D into $N \times T$ sub-sets, where N and T are the number of feature maps and circles, as shown in Fig. 3.

Along the circles, D is decomposed into sub-sets with some instance overlapping, as given by

$$D = \bigcup_t D^t, D^t \cap D^{(t+1)} \neq \emptyset, \tag{4}$$

where $D^t = \sum_n D^t_n$. D^t is a sub-set of easy instances, while $D^{(t+1)}$ is a larger sub-set which includes additional instances that are harder than D^t. The harder instances are identified by their occlusion rates or training loss.

Along the feature maps in each circle, D^t is decomposed into sub-sets without instance overlap, as

$$D^t = \bigcup_n D^t_n, D^t_n \cap D^t_{n+1} = \emptyset, \tag{5}$$

where D^t_{n+1} is a sub-set of instances whose resolutions are higher than D^t_n. We assign high resolution instances to deep layers and low resolution instances to shallow layers.

IV. Pedestrian Detection With CircleNet

Using the CircleNet as a backbone network, we implemented pedestrian detection by using the region proposal network (RPN) to generate object proposals at each feature layer. The CircleNet is optimized using stochastic gradient descent (SGD) in an end-to-end manner.

A. Implementation

CircleNet contains a backbone CNN for basic feature extraction and a circling module for feature adaptation. The basic feature maps for ResNets [58] are the outputs of last residual blocks for conv2, conv3, conv4, and conv5. These basic feature maps are fed to the circling module to generate adapted feature maps $\{o^t_n, e^t_n\}, n = 1, \ldots, 4, t = 1, \ldots, T$, which are used for region proposal generation, object classification, and bounding box regression.

An RPN is a sliding-window class-agnostic object detector for generating region proposals, and a predictor classifies each proposal and refines the proposal for object localization. We use the RPN and the predictor on each circle and note that the parameters of the RPNs and the predictor heads are shared across all circles. The procedure of detection involves several steps. First the input images go through the backbone CNN and the circling module, and the RPNs are used to generate proposals. Second the features of each proposal are generated by cropping from the adapted features, *e.g.*, $\{o^t_n\}$. More specifically, the proposal features are extracted from the n^{th} layer of feature pyramid according to the proposal' size in the current circle ($T = t$), where $n = \lfloor k_0 + log_2(area/224) \times \theta \rfloor$, θ is a hyper-parameter relevant to the dataset. "area" denotes the area of the proposal. In experiment, we set $k_0 = 4$ and $\theta = log_2(2)$, which means that the features of the small proposals are extracted from high-resolution pyramid layers while the features of the large proposals are extracted from low-resolution pyramid layers. Third, a Region-of-Interest (RoI) pooling layer produces the same length of feature for each proposal. Finally, these features are fed to the predictors to determine the detection.

B. Learning

We train the network by optimizing both the classification and regression tasks for the RPNs and the predictors. The softmax loss function is used for classification, and the smooth L1 [59] loss function is used for regression. The layers on a top-down pathway focus on aggregating the semantic information from deep layers. We therefore add a pseudo segmentation loss[1] on the top-down pathways. By introducing a segmentation layer which drives the features to focus on pedestrian regions and to suppress the negative samples from cluttered backgrounds. The loss function, defined as

$$\begin{aligned}\mathcal{L} = \sum_t \mathcal{L}^t = \sum_t \Big(\sum_n (\mathcal{L}{rpn\_cls}^t_n + \mathcal{L}{rpn\_reg}^t_n \\ + \mathcal{L}{cls}^t_n + \mathcal{L}{reg}^t_n) + \mathcal{L}{seg}^t\Big),\end{aligned} \tag{6}$$

is applied to each RPN, predictor and segmentation layer. The softmax loss $\mathcal{L}{rpn\_cls}$ and $\mathcal{L}{cls}$ optimize the classes for anchors and proposals. The regression loss $\mathcal{L}{rpn\_reg}$ optimizes the relative displacement between anchors and ground truth, and $\mathcal{L}{reg}$ optimizes the relative displacement between proposals and ground truth. The forward and backward propagation procedures are shown in Fig. 4, which shows that the feature/gradient can be directly propagated along circles.

V. Understanding CircleNet

From the perspective of learning, CircleNet implements a special kind of classifier ensemble. For multiple feature layers on the circles, we have multiple classifiers, each of

[1] A pseudo-mask is generated for each image by setting the pixels in the pedestrian bounding boxes to 1 and in the background to 0.

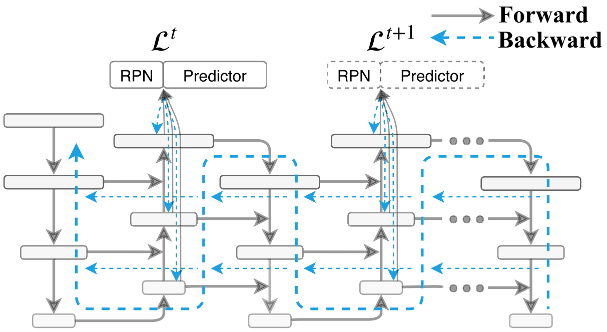


Fig. 4. On each of the feature layer of the top-down pathway, an RPN and a predictor is implemented for pedestrian detection. During the learning procedure, forward connections (solid lines with arrows) along the CircleNet are for feature extraction and object score prediction. Backward connections (dashed lines with arrows) are for gradient propagation.

which is responsible for a sub-set of instances. In the deep learning framework, the hard instances have large training loss, which means that they have large weight during the learning procedure. This is actually a boosting-like learning strategy. In the training phrase, the pedestrian instances of different appearance, *e.g.*, resolution and occlusion rates, are processed with multiple base classifiers. In the test phase, the maximum classification scores are used to determine final detections.

By sharing the backbone based in a feature reciprocating procedure, we can actually boost the base classifiers learned on different sub-sets. As a result, the feature (base classifier) achieved on difficult samples can also be adapted to normal samples. This not only benefits pedestrian detection but also leads to a new ensemble method in the deep learning framework.

VI. Experiments

In this section, the experimental protocols are introduced and the effects of the CircleNet are analyzed. The performance of pedestrian detection with CircleNet and comparisons with other state-of-the-art detectors are presented. Ablation experiments are carried out to validate the effectiveness of CircleNet backbone for pedestrian detection in challenging traffic scenes.

A. Experimental Protocols

1) Datasets: Two of the most popular datasets for pedestrian detection, **Caltech** [60] and **CityPersons** [11], are used to evaluate CircleNet. The Caltech dataset contains approximately 10 hours of street-view video taken with a camera mounted on a vehicle. The most challenging aspect of the dataset is the large number of low-resolution pedestrians. We sample 42,782 images from set00 to set05 for training and 4,024 images from set06-set10 for testing.

The CityPersons dataset is built on the semantic segmentation dataset Cityscapes [61]. This dataset contains 5,000 images (2,975 for training, 500 for validation, and 1,525 for testing) captured from 18 cities in Germany at three different seasons and various weather conditions. The scene in CityPersons dataset is much more “crowded” than Caltech dataset, and the challenging aspects of the pedestrian objects include complex backgrounds, low resolution, and heavy occlusion.

2) Evaluation Metrics: The experiments are conducted on instances across different occlusion levels, including: (1) **Reasonable:** visibility $\in [0.65, \infty]$ & Height$\geq$50 (pixels); (2) **None:** visibility $\in [1.0, \infty]$; (3) **Partial:** visibility $\in [0.65, 1.0)$; (4) **Heavy:** visibility $\in [0.2, 0.65)$; and normal and low-resolution instances, as: (1) **Height$\geq$50** (pixels); (2) **Height$\geq$20** (pixels).

In most previous pedestrian detection works, the miss rate (MR) over false positive per image (FPPI) is commonly used as the evaluation protocol [11], [60]. This is preferred to precision recall curves for certain tasks, *e.g.*, automotive applications, as typically there is an upper limit on the acceptable false positives per image rate independent of pedestrian density. Such a protocol produces miss rates against FPPI by varying the threshold on detection scores. To emphasize the importance of missing detections, a log-average miss rate is used to summarize pedestrian detector performance, by averaging miss rates at nine FPPI rates in the range of $[10^{-1}, 10^{0}]$.

A detected box is classified as a true positive, if its classification score is larger than a threshold when the box is matched with a ground-truth box, *i.e.*, their IoU exceeds 50 percent. Each bounding box and ground truth can be matched at most once. Detections with the highest confidence are matched first; if a detected bounding box matches multiple ground-truth bounding boxes, the match with the highest overlap is used. Unmatched detected bounding boxes are classified as false positives.

3) Implementation Details: FPN [14] with the ResNet50 backbone [58] is used as the baseline detector. For a fair comparison with other state-of-the-art methods, we up-sample the image resolution to 900 × 1200, fine-tune the network trained on CityPersons. On the CityPersons, we follow the settings in [62] and up-sample the images with a 1.3× ratio and fine-tune the network pre-trained on ImageNet.[2]

B. Ablation Studies

1) Circling Architecture: We define the information through a bottom-up pathway and a top-down pathway as one circle, and use the FPN, which has no circling architecture, as a baseline model. The half-circling architecture (CircleNet-1/2) that has an additional bottom-up pathway to FPN is a PANet [15]. The CircleNets with one, two, and three circles are labeled as CircleNet-1 ($T = 1$), CircleNet-2 ($T = 2$), and CircleNet-3 ($T = 3$), respectively.

In Table I, it can be seen that CircleNet-1 outperforms FPN and CircleNet-1/2, while CircleNet-2 outperforms CircleNet-1. CircleNet-3 has slightly lower performance than CircleNet-2 on the “All” (height$\geq$50) and (height$\geq$20) sub-set. The reason could be that by increasing the number of circles, the training difficulty increases. Considering that CircleNet-3 can provide sufficient feature layers to represent

[2] Some recent works fine-tune the network pre-trained on Citypersons to detection pedestrians on Caltech. This can aggregates the performance at the cost of training complexity.

TABLE I

DETECTION PERFORMANCE OF FPN AND CIRCLENET ON THE CALTECH TEST SET. MR^{-2} IS USED TO COMPARE THE PERFORMANCE. LOWER SCORE INDICATES BETTER PERFORMANCE

Model	Description	Height≥50					Height≥20			
		All	None	Partial	Heavy	Reasonable	All	None	Partial	Heavy
FPN [14]	Baseline	31.21	12.72	37.40	82.96	15.79	59.85	48.39	69.41	89.71
CircleNet-1/2	PANet structure	30.02	12.76	37.82	79.20	15.80	57.26	45.48	68.98	88.72
CircleNet-1	One circle ($T = 1$)	28.74	**10.85**	**34.28**	79.87	**13.75**	**57.00**	**45.21**	68.29	88.86
CircleNet-2	Two circles ($T = 2$)	**28.12**	11.58	36.49	75.08	14.63	57.25	45.62	69.27	**88.19**
CircleNet-3	Three circles ($T = 3$)	28.69	11.66	36.32	**75.01**	14.72	58.71	48.89	**68.18**	88.65
FPN [14] (trained with occluded instances)	Baseline	25.88	12.58	32.03	64.01	14.85	57.02	48.51	64.33	79.33
CircleNet-2 (trained with occluded instances)	Two circles	**24.14**	12.84	**28.74**	**54.66**	15.02	**55.36**	**46.97**	65.77	**75.05**

TABLE II

COMPARISON OF THE DETECTION SPEEDS ON CALTECH DATASET

Method	**FPN**	**CircleNet-1**	**CircleNet-2**	**CircleNet-3**
Inference Time (ms/image)	51	61	74	87

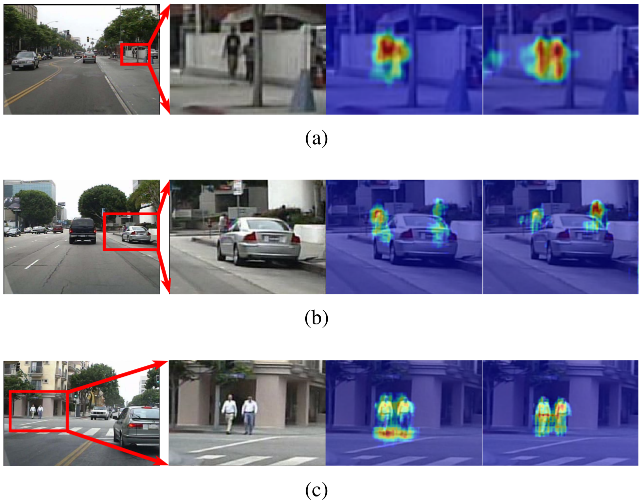
pedestrians of various normal and hard instances, we do not test more circles in following experiments.

Table I shows that the CircleNets outperform the baseline FPN and PANet (CircleNet-1/2) by significant margins. On the “All” and “Reasonable” sub-set (height≥50), CircleNet-2 outperforms FPN by 3.09% (28.12% vs. 31.21%) and 1.16% (14.63% vs. 15.79%). On the “None” and “Partial” (Height≥50) occlusion sub-set, CircleNet-2 outperforms FPN by 1.14% (11.58% vs. 12.72%) and 0.91% (36.49% vs. 37.40%), respectively. On the “Heavy” occlusion sub-set (height≥50), CircleNet outperforms FPN by 7.84% (75.08% vs. 82.96%). On “All” sub-set (Height≥20), CircleNet-2 noticeably outperforms FPN by 2.60% (57.25% vs. 59.85%). CircleNet-2 outperforms PANet (CircleNet-1/2) by 1.9% (28.12% vs. 30.02%) on the “All” subset of Caltech dataset. For a fair comparison, the compared PANet does not use adaptive feature pooling. When training the networks with occluded instances as shown in the last two rows of Table I, CircleNet still outperforms the baseline approach, particularly for low-resolution instances.

The proposed CircleNet also boosts the performance of the region proposal network (RPN) under two commonly used metrics. As shown in Fig. 7, the first metric evaluates the recall rate under different IoU thresholds. The second metric evaluates the variation of recall rate under different numbers of region proposals. CircleNet outperforms the baseline FPN by significant margins in terms of both metrics.

As is known, down-stream features bring not only the semantic information but also noise. To alleviate the noise, we do not use the down-steam feature directly. Instead, we use a convolution layer to encode the fused features from top-down and bottom-up pathways to filter the noises from down-stream features.

In Table II, we compare the inference speeds of the FPN and CircleNet. FPN takes 51ms to process an image while CircleNet-1, CircleNet-2 and CircleNet-3 take 61ms, 74ms, and 87ms, respectively. Circle-Net can significantly improve the detection accuracy with moderate computational cost overhead. With a Telsla 1080TI GPU, the best CircleNet-2 runs at 13.5 FPS, which can feed the requirement of many real-world applications.


(a)

(b)

(c)

Fig. 5. Visualization of feature adaptation. The first column shows the original image. A cropped patch with a pedestrian is presented in the second column. The visualization of the feature map that from Circle-1 can be seen in the third column and the fourth column shows the Circle-2. (Best view in color.)

2) Feature Adaptation: Feature adaptation is implemented in a reciprocating manner by concatenating multiple top-down and bottom-up pathways. Fig. 5 presents the visualization of the features from different circles (Circle-1 and Circle-2) in CircleNet-2.

In Fig. 5a, a pedestrian is almost ignored by the first circle (Circle-1) as the occlusion. With feature adaptation, the second circle (Circle-2) learns representative features of the pedestrian. Fig. 5b shows two circles learn different features that are adaptive to different occlusion patterns. In Fig. 5c the learned features of the first circle activate some background areas, while those of the second circle activate the pedestrian boundary.

Fig. 6 further shows the effect of feature adaption contributed by CircleNet. In the bottom layer, Circle-2 better activates pedestrian regions than Circle-1, while in the top layer, Circle-1 performs better on pedestrians with no occlusion. For pedestrians with heavy occlusion, Circle-2 is able to activate more details. The feature layer with the best activation regions (red dotted ellipses) corresponds to the detection results.

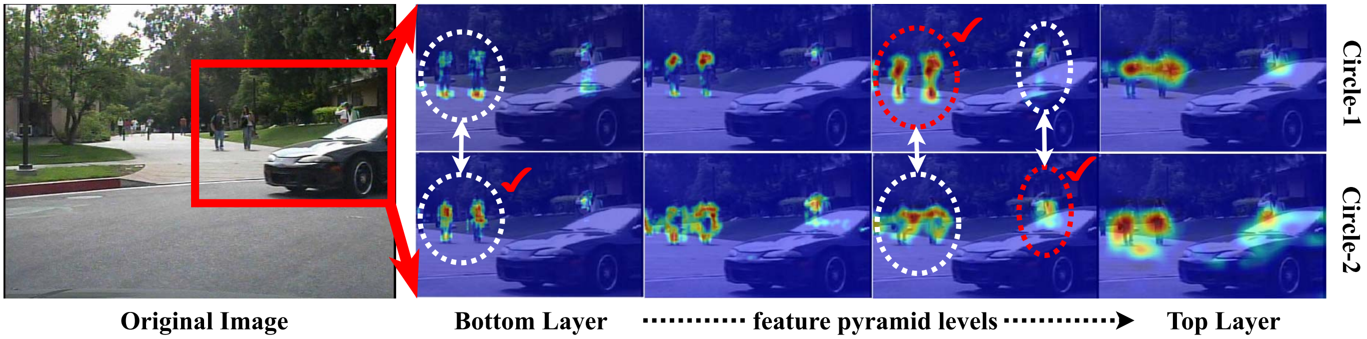
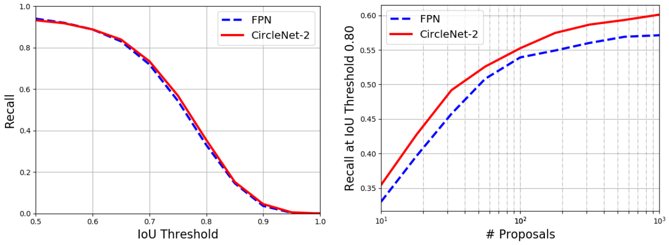


Fig. 6. Visualization of feature adaptation across circles. (Best view in color.)



(a) (b)

Fig. 7. Evaluation of the effect of CircleNet for region proposal network (RPN). (a) Recall rates under different IoU thresholds. (b) Recall rates under different numbers of region proposals.

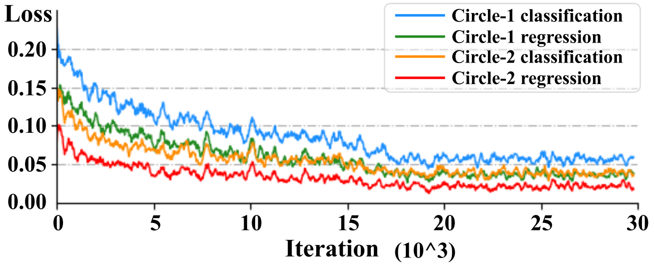


Fig. 8. Loss comparison of Circle-1 and Circle-2. (Best view in color.)

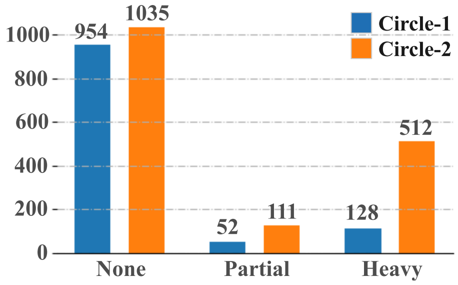
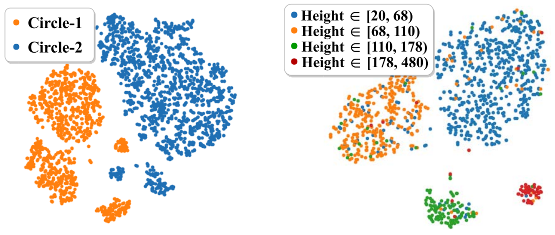


Fig. 9. Detection result statistics. Circle-1 and Circle-2 detect a comparable number of objects with occlusion ratio "None" and "Partial", but Circle-2 detects significantly more "Heavy" occlusion objects.



(a) (b)

Fig. 10. The t-SNE [63], [64] visualization of different feature embedding. (a) Instances from different circles. (b) Instances with different resolution. (Best view in color.)

Fig. 8 shows the classification and regression loss of Circle-1 and Circle-2 on the predictors. We can see that the loss of Circle-2 is larger than that of Circle-1. The reason lies in that during the feature adaptation procedure, Circle-2 focuses on learning the features for harder instances and the features of Circle-2 inherit the Circle-1, as analyzed below.

3) Instance Decomposition: We study the effect of different instance decomposition strategies on the circles. "None" means that Circle-1 and Circle-2 are trained with all instances. "By loss" means the instances are decomposed according to the loss, which requires Circle-2 to learn the hard instances with larger loss in the training batch. The RoIs of larger training loss in Circle-2 are defined as hard instances. We normalize the loss of instances in a training batch and scale them to a specific range. The final weights are obtained by adding them to a basic value as $w = \frac{l - lmin}{lmax - lmin} \times (1 - \alpha) + \alpha$, where l means the loss of the training batch, and we empirically set $\alpha = 0.7$. This strategy improves the performance on high-resolution sub-set (Height $\geq$ 50) by 0.49% (22.37% vs. 22.86%), while maintaining the performance on the sub-set (Height $\geq$20) (Table III).

Pedestrians with an occlusion ratio between 65% and 80% are defined as hard instances. "All-to-hard" means all samples are used in Circle-1, and "hard" samples are learned in Circle-2. "Easy-to-hard" means easy samples are decomposed in the first circle, and "Easy+hard" samples are decomposed to the second one. As shown in Table III, "Easy-to-hard" is the best instance decomposition strategy, which reduces the error rate MR^{-2} by 1.85% (52.69% vs. 54.54%) on the "All" (Height$\geq$20) sub-set.

In the test phase, we randomly select 2792 pedestrian samples from the test dataset, and plot the detection result statistics according to occlusion ratio (None, Partial and Heavy) in Fig. 9. It can be seen that both Circle-1 and Circle-2 have detected a comparable number of objects with occlusion ratio "None" and "Partial", but Circle-2 has detected significantly more "Heavy" occlusion objects.

TABLE III

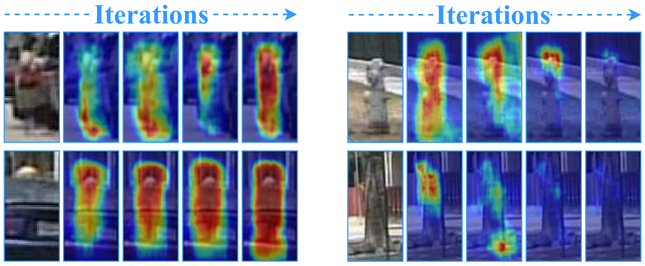
DETECTION PERFORMANCE OF CIRCLENET WITH INSTANCE DECOMPOSITION AND MULTIPLE SUPERVISIONS ON THE CALTECH TEST SET

Model	Instance decomposition & supervision	Height≥50					Height≥20			
		All	None	Partial	Heavy	Reasonable	All	None	Partial	Heavy
CircleNet	w/o	24.14	12.84	28.74	54.66	15.02	55.36	46.97	65.77	75.05
CircleNet+ID1	None	22.86	12.33	27.68	50.4	14.54	54.54	47.42	66.11	69.42
CircleNet+ID2	By loss (OHEM) [53]	22.37	12.92	28.45	48.35	14.83	54.55	47.15	65.31	72.25
CircleNet+ID3	All-to-hard	23.84	12.25	28.00	55.19	14.48	55.05	46.40	65.65	76.48
CircleNet+ID4	Easy-to-hard	21.62	11.83	26.45	48.70	13.78	52.69	44.29	64.83	73.09
CircleNet+MS	Multiple supervision	21.57	12.44	27.50	45.47	14.38	54.62	47.12	66.29	71.55
CircleNet+	Multiple supervision + decomposition	18.05	8.42	20.27	44.53	10.21	46.42	37.48	59.26	66.28

TABLE IV

COMPARISON OF CIRCLENET WITH OTHER STATE-OF-THE-ART METHODS ON THE CALTECH TEST SET

Model	Height≥50					Height≥20			
	All	None	Partial	Heavy	Reasonable	All	None	Partial	Heavy
DeepParts [52]	22.79	10.64	19.93	60.42	11.89	64.78	58.43	70.39	81.81
MS-CNN [47]	21.53	8.15	19.24	59.94	9.95	60.95	53.67	67.16	79.51
RPN+BF [9]	24.01	7.68	24.23	74.36	9.58	64.66	56.38	72.55	87.48
AdaptFasterRCNN [11]	20.03	7.01	26.55	57.58	9.18	60.11	52.67	68.50	79.58
SDS-RCNN [10]	19.72	**5.95**	**14.86**	58.55	**7.36**	61.50	54.45	66.46	78.78
FasterRCNN+ATT [13]	18.21	8.46	22.29	45.18	10.33	54.51	47.54	64.47	71.02
CircleNet (Ours)	**18.05**	8.42	20.27	**44.53**	10.21	**46.42**	**37.48**	**59.26**	**66.28**



(a) Positive examples (b) Negative examples

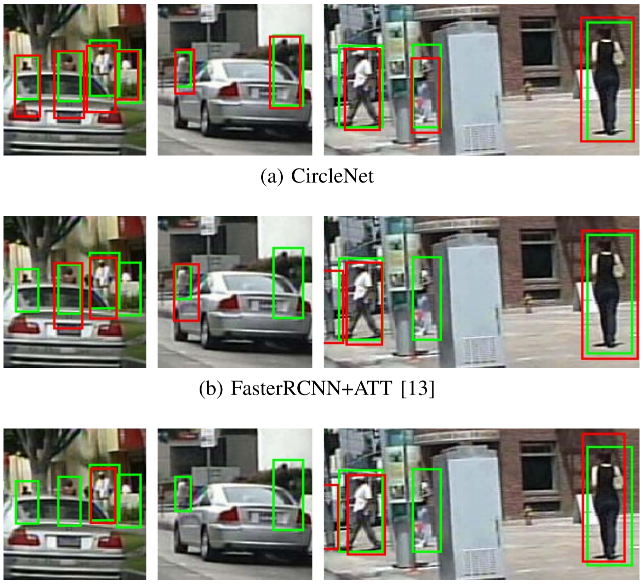
Fig. 11. With multiple supervisions, the feature representation for the positive examples are enforced while that for the negative examples are depressed. (Best viewed in color.)

We further visualize the feature embedding of these pedestrian samples using t-SNE [63], [64], Fig. 10. It can be seen that the samples from Circle-1 and Circle-2 form clusters, which correspond to pedestrian instances of different appearances in Fig. 10a. This figure shows the feature differences between adapted circles. Fig. 10b shows the clusters of samples in terms of resolution. CircleNet, as a classifier ensemble, can process these clusters well.

4) Multiple Supervision: With object detection loss and pseudo-segmentation loss, we target at activating features on the object border while further explore the feature adaptability of CircleNet. Fig. 11a illustrates that when training, the activation mask[3] gradually fills the bounding boxes. As a pixel-wise classification task, semantic segmentation with pseudo-mask supervision helps suppressing the negative samples from cluttered backgrounds, Fig. 11b.

C. Performance and Comparison

In Table IV, we compare the performance of CircleNet with state-of-the-art approaches on Caltech. MS-CNN [47], RPN+BF [9], AdaptFaster-RCNN [11], and SDS-RCNN [10] achieve top results on the "Reasonable" sub-set, but do not perform well on heavily occluded instance or low-resolution instances. CircleNet beats all compared approaches on low-resolution and occluded cases, while reporting acceptable performance on "Reasonable" sub-set.



(a) CircleNet

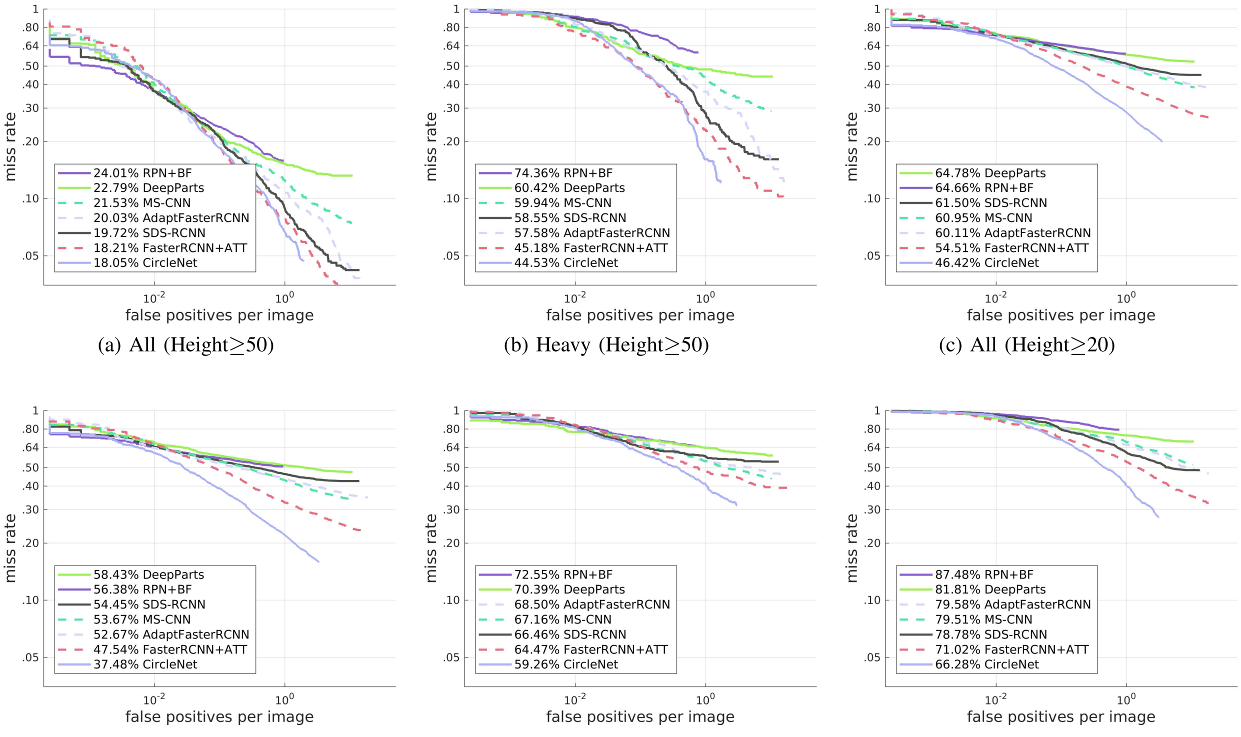
(b) FasterRCNN+ATT [13]

(c) MS-CNN [47]

Fig. 12. Qualitative detection results of cropped image patches at FPPI=0.1 on the Caltech test set. The green solid boxes indicate ground truth; the red boxes denote detection results. (Best viewed in color.)

As shown in Table IV, CircleNet outperforms the state-of-the-art methods on the Caltech test set. It outperforms FastRCNN-ATT by 2.02% (20.27% vs. 22.29%) on the "Partial" sub-set (Height ≥ 50) and 8% (46.42% vs. 54.51%) on the "all" sub-set (Height ≥ 20). CircleNet also outperforms FastRCNN-ATT on the "Partial" and "Heavy" occlusion sub-set with significant margins.

[3] The activation mask is obtained by a segmentation layer (3×3 and 1×1 convolutional operations) and a sigmoid function.



(a) All (Height≥50) (b) Heavy (Height≥50) (c) All (Height≥20)

(d) None (Height≥20) (e) Partial (Height≥20) (f) Heavy (Height≥20)

Fig. 13. Performance comparison on the Caltech dataset. Lower curves indicate better performance.

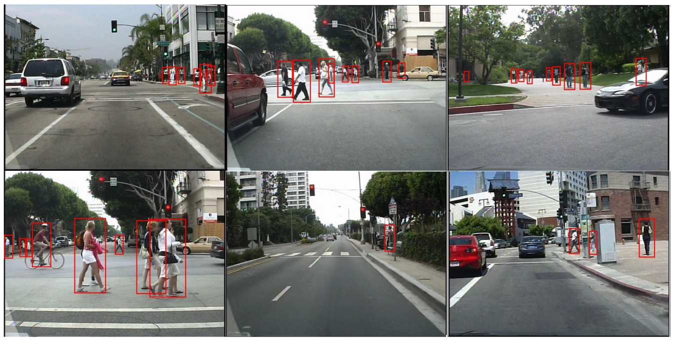


Fig. 14. Detection examples from the Caltech dataset. Red bounding boxes are predicted pedestrians with threshold 0.7.

TABLE V

COMPARISON OF CIRCLENET WITH OTHER STATE-OF-THE-ART METHODS ON THE CITYPERSONS VALIDATION SET. (*DENOTES THE EXPERIMENTAL PROTOCOLS USED IN [7], [62])

Method	*Reasonable*	*Heavy**	*Partial**	*Bare**
Adapted Faster RCNN [11]	12.8	-	-	-
Repulsion Loss [7]	11.6	55.3	14.8	7.0
OR-CNN [62]	11.0	51.3	13.7	5.9
CircleNet (Ours)	11.77	**50.22**	**12.21**	7.14

Fig. 12 shows qualitative results which indicates that CircleNet produces robust detections, which are well aligned with the ground-truth on various occlusion patterns. In contrast, the other two detectors (FasterRCNN+ATT [13] and MS-CNN [47]) missed some of the objects. In Fig. 13, the miss rate and FPPI curves show that the proposed CircleNet outperforms state-of-the-art approaches with significant margins.

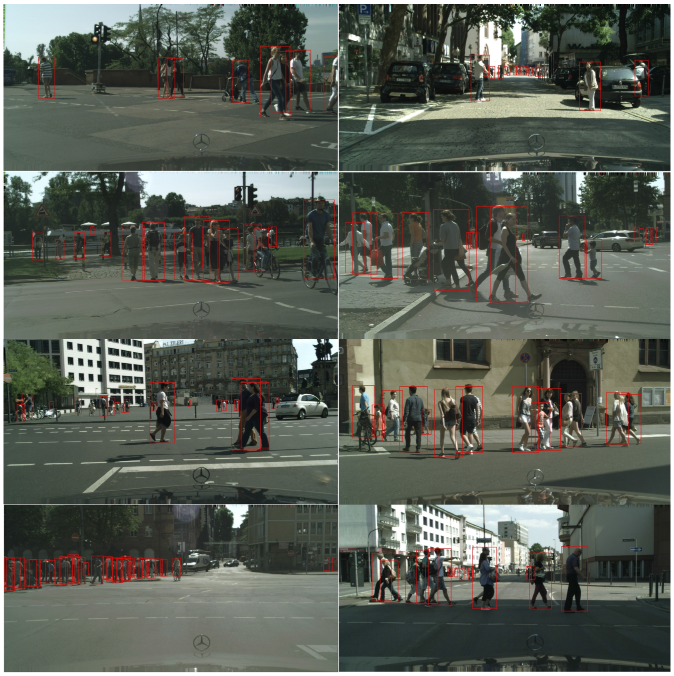


Fig. 15. Detection examples from the CityPersons dataset. Red bounding boxes are predicted pedestrians with threshold 0.7. (Best viewed in color and with zoom.)

In Table V, we compare CircleNet with state-of-the-art detectors on the CityPersons validation set, using the performance reported by the authors [7], [62]. It can be seen that CircleNet consistently outperforms the state-of-the-art OR-CNN and FasterRCNN+ATT on this validation set.

In Fig. 14 and Fig. 15, the results show that CircleNet can effectively detect pedestrians of low-resolution, occlusion, and clutter backgrounds.

VII. Conclusion

Pedestrian detection in the wild remains a challenging problem for the "hard" instances with heavy occlusion and/or low resolution. In this paper, we developed a new feature learning model, referred to as CircleNet, which reciprocates the feature adaptation by formatting deep-to-shallow and shallow-to-deep feature fusion pathways. These cycling loops not only improve the representative capability and adaptability of convolutional features for objects of various appearance, but also naturally mimic the process which we humans attempt to detect and recognize small and highly occluded objects. We also propose hard instance decomposition strategies to assign instances along feature layers and circles to fully utilize the feature adaptation capability introduced by the circling architecture. Significant performance improvement over the baseline FPN approach demonstrates that CircleNet is simple yet effective. Its shows great potential for single-camera based pedestrian detection systems and provides fresh insight to occluded and low-resolution object detection problems.

Acknowledgment

The authors would like to express their sincere appreciation to the reviewers for their constructive comments.

References

[1] Q. Ye *et al.*, "Self-learning scene-specific pedestrian detectors using a progressive latent model," in *Proc. CVPR*, 2017, pp. 509–518.

[2] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. CVPR*, 2016, pp. 1259–1267.

[3] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. ECCV*, 2018, pp. 734–750.

[4] C. Flores, P. Merdrignac, R. de Charette, F. Navas, V. Milanés, and F. Nashashibi, "A cooperative car-following/emergency braking system with prediction-based pedestrian avoidance capabilities," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1837–1846, May 2019.

[5] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. G. Jung, "A new approach to urban pedestrian detection for automatic braking," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 594–605, Dec. 2009.

[6] X. Li *et al.*, "A unified framework for concurrent pedestrian and cyclist detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 2, pp. 269–281, Feb. 2017.

[7] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. CVPR*, 2017, pp. 7774–7783.

[8] X. Du, M. El-Khamy, J. Lee, and L. S. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. WACV*, 2017, pp. 953–961.

[9] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. ECCV*, 2016, pp. 443–457.

[10] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proc. ICCV*, 2017, pp. 4950–4959.

[11] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. CVPR*, 2017, pp. 3213–3221.

[12] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. CVPR*, 2017, pp. 3127–3136.

[13] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. CVPR*, 2018, pp. 6995–7003.

[14] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017, pp. 2117–2125.

[15] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. CVPR*, 2018, pp. 8759–8768.

[16] T. S. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *J. Opt. Soc. Amer. A, Opt. Image Sci. Vis.*, vol. 20, no. 7, pp. 1434–1448, 2003.

[17] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. CVPR*, 2018, pp. 6154–6162.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[24] P. F. Felzenszwalb, R. B. Girshick, D. A. 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.

[25] T. Ojala, M. Pietikäinen, and T. Mäenpää, "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.

[26] P. Dollár, 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.

[27] W. Ke, Y. Zhang, P. Wei, Q. Ye, and J. Jiao, "Pedestrian detection via PCA filters based convolutional channel features," in *Proc. ICASSP*, 2015, pp. 1394–1398.

[28] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. CVPR*, 2014, pp. 947–954.

[29] S. Zhang, R. Benenson, M. Omran, J. H. 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.

[30] S. Zhang, C. Bauckhage, and A. B. Cremers, "Efficient pedestrian detection via rectangular features based on a statistical shape model," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 763–775, Apr. 2015.

[31] X. Chen, P. Wei, W. Ke, Q. Ye, and J. Jiao, "Pedestrian detection with deep convolutional neural network," in *Proc. ACCV Workshops*, 2014, pp. 354–365.

[32] Q. Ye, Z. Han, J. Jiao, and J. Liu, "Human detection in images via piecewise linear support vector machines," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 778–789, Feb. 2013.

[33] J. Marín, D. Vázquez, A. M. López, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, Dec. 2013, pp. 2592–2599.

[34] J.-Y. Kwak, B. C. Ko, and J. Y. Nam, "Pedestrian tracking using online boosted random ferns learning in far-infrared imagery for safe driving at night," *IEEE Trans. Intell. Transp. Syst.*, vol. 18., no. 1, pp. 69–81, Jan. 2017.

[35] A. Prioletti, A. Møgelmose, 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. Vázquez, A. M. López, J. Marín, 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. Gonzàlez, 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] Y. Xu, X. Cao, and H. Qiao, "An efficient tree classifier ensemble-based approach for pedestrian detection," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 107–117, Feb. 2011.

[40] J. Baek, J. Kim, and E. Kim, "Fast and efficient pedestrian detection via the cascade implementation of an additive kernel support vector machine," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 902–916, Apr. 2017.

[41] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. CVPR*, 2015, pp. 4073–4082.

[42] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2014, pp. 580–587.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[44] Y. Li, J. Li, W. Lin, and J. Li, "Tiny-DSOD: Lightweight object detection for resource-restricted usages," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 59–63.

[45] K. Chen *et al.*, "Towards accurate one-stage object detection with AP-loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5119–5127.

[46] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[47] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. ECCV*, 2016, pp. 354–370.

[48] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[49] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. ECCV*, 2018, pp. 732–747.

[50] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. ECCV*, 2018, pp. 536–551.

[51] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. CVPR*, 2012, pp. 3258–3265.

[52] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. ICCV*, 2015, pp. 1904–1912.

[53] A. Shrivastava, A. Gupta, and R. B. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 761–769.

[54] S. Wang, J. Cheng, H. Liu, and M. Tang, "PCN: Part and context information for pedestrian detection with CNNs," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, London, U.K., Sep. 2017, pp. 1–13.

[55] S. Zagoruyko *et al.*, "A multipath network for object detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, York, U.K., Sep. 2016, pp. 1–13.

[56] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1134–1142.

[57] K. S. Ni and T. Q. Nguyen, "Adaptable K-nearest neighbor for image interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar./Apr. 2008, pp. 1297–1300.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[59] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015, pp. 1440–1448.

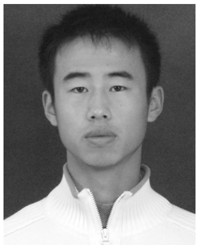
[60] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. CVPR*, 2009, pp. 304–311.

[61] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. CVPR*, 2016, pp. 3213–3223.

[62] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. ECCV*, 2018, pp. 637–653.

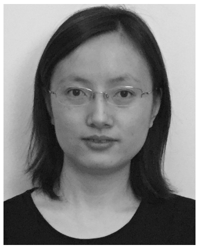
[63] L. van der Maaten and G. Hinton, "Visualizing non-metric similarities in multiple maps," *Mach. Learn.*, vol. 87, no. 1, pp. 33–55, 2012.

[64] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.

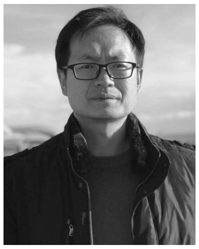


Tianliang Zhang received the B.S. degree in electronic information engineering from the Wuhan University of Technology (WUT) in 2013 and the M.S. degree in industrial engineering from the University of Chinese Academy of Sciences in 2017, where he is currently pursuing the Ph.D. degree with the School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

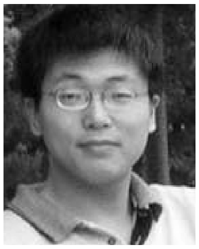
Zhenjun Han (M'17) received the B.S. degree in software engineering from Tianjin University, Tianjin, China, in 2006, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2011. He has been an Associate Professor with the University of Chinese Academy of Sciences since 2014. He has published more than 20 articles in refereed conferences and journals, including the IEEE CVPR and the IEEE TRANSACTIONS ON CSVT. His current research interests include image processing and visual surveillance.



Huijuan Xu (M'10) received the bachelor's degree from the Hefei University of Technology in 2009, the master's degree from the University of Chinese Academy of Sciences in 2012, and the Ph.D. degree from the Computer Science Department, Boston University, in 2018. She currently holds post-doctoral position at UC Berkeley. Her research focuses on deep learning, computer vision, and natural language processing, particularly in the areas of visual question answering, video language description, and activity detection.



Baochang Zhang (M'12) 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 Griffith University, Brisbane, Australia. He is currently a tenured Associate Professor with the Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. He has published more than 50 articles in refereed conferences and journals. His research interests include pattern recognition, machine learning, face recognition, and wavelets.



Qixiang Ye (M'10–SM'15) 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 was a Visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, until 2013. He has been a Professor with the University of Chinese Academy of Sciences since 2016. He has published more than 50 articles in refereed conferences and journals, including the IEEE CVPR, ICCV, ECCV, and IEEE TRANSACTIONS ON CSVT, TIP, ITS, and PAMI. His research interests include image processing, visual object detection, and machine learning. He is on the Editorial Board of *Visual Computer* (Springer).