Scale-Aware Hierarchical Detection Network for Pedestrian Detection

XIAOWEI ZHANG, SHUAI CAO, AND CHENGLIZHAO CHEN
Shandong Key Laboratory of Intelligent Information Processing, School of Computer Science and Technology, Qingdao University, Qingdao 266071, China
Corresponding authors: Xiaowei Zhang (xiaowei19870119@sina.com) and Chenglizhao Chen (cclz123@163.com)
This work was supported in part by the National Natural Science Foundation of China under Grant 6190070308, and in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2019BF028.

ABSTRACT Several or even dozens of times spatial scale variation is one of the major bottleneck for pedestrian detection. Although the Region-based Convolutional Neural Network (R-CNN) families have shown promising results for object detection, they are still limited to detect pedestrians with large scale variations due to the fixed receptive field sizes on a single convolutional output layer. In contrast to previous methods that simply combined pedestrian predictions on feature maps with different resolution, we propose a scale-aware hierarchical detection network for pedestrian detection under large scale variations. First, we introduce a cross-scale features aggregation module to accomplish feature augmentation for pedestrian representation through merging the lateral connection, the top-down path and bottom-up path. Specifically, the cross-scale features aggregation module can adaptively fuse hierarchical features to enhance feature pyramid representation for robust semantic and accurate localization. Further, we design a scale-aware hierarchical detection network to effectively integrate multiscale pedestrian detection into a unified framework by adaptively perceiving the augmented feature level for special-scale pedestrian detection. Experimentally, the proposed scale-aware hierarchical detection network forms a more robust and discriminative model for pedestrian instances with different scales on widely-used ETH and Caltech benchmarks. In particular, compared with the state-of-the-art method FasterRCNN+ATT, the log-average miss rate of pedestrian detection is reduced by 11.98% for medium scale pedestrians (between 30-80 pixels in height), and 14.12% for whole scale pedestrians (above 20 pixels in height) on Caltech benchmark.

INDEX TERMS Scale variation, feature aggregation, scale-aware weighting, hierarchical detection.

I. INTRODUCTION

Pedestrian detection stands out from the traditional object detection tasks, in view of its broad application prospects in computer vision, such as video surveillance, autonomous driving, robotics. Despite significant improvements have been made on pedestrian detection [4], [8], [24], [27], [41] over the years, the most existing efforts generally work very well for large scale pedestrian instances [17]–[19], [23], [34], [36], [51]. Compared with pedestrian detection under large scale, much less attention has been put toward medium and small scale ones, as similar observations in the literature [14], [54].

For autonomous driving system, detecting medium and small size pedestrians are an important topic because there may leave sufficient time to alert the driver. Assuming the vehicle traveling at an urban speed of 15m/s and a pedestrian of 1.8m tall, the person with 80 pixels in height is just 1.5 s away, while a person with 30 pixels is 4 s away. Take one latest effort AR-Ped [2] for example, it has been reported that empirically their detector is capable of achieving 6.45% log-average miss rate for pedestrians taller than 50 pixels on Caltech Pedestrian Benchmark [20], however the same error rate increases to 49.31% MR for pedestrians of 30-80 pixels in height. Fig. 1(a) shows several failed examples of pedestrian detection using state-of-the-art method AR-Ped [2] under large scale appearance variations on Caltech benchmark. As Fig. 1(b) illustrates the scale distribution of pedestrians in height on Caltech dataset, we group pedestrians by their image size (height in pixels) into three scales following [54]: near (80 or more pixels), medium (between 30-80 pixels), and far (between 20-30 pixels). Note that about 81.67% MR of the pedestrians lie in the medium scale on Caltech dataset. The degraded performance for pedestrian detection under large scale variations may be attributed to the following inherent challenges. First, small-size pedestrian instances often
convey smaller amount of information while having a greater proportion of noise with obscure appearance and blurred boundaries. It is in general difficult to distinguish them from the background clutters. Second, visual semantic concepts of an object can emerge in different spatial scales depending on the size of the target objects. For a pedestrian instance of interest, visual features are effective only at a proper scale where optimal response is obtained. This difference is more pronounced in complex scenes containing pedestrian instances of diverse scales.

To address the issue of pedestrian detection under large scale appearance variations, the Faster-RCNN [9] exploits a multiscale region proposal network (RPN), which achieves excellent object detection performance. However, multi-scale detection is generated by sliding a fixed set of filters over a fixed set of convolutional feature maps. This results in an inconsistency between the sizes of objects and filter receptive fields, as the scales of objects are variable, yet the sizes of filter receptive fields are fixed. Instead of using a fixed set of receptive fields, in most related works [1], [7], [15], [44], [49], [56], [57] that aim to detect multi-scale pedestrians, pedestrian detection is performed by redefining the perceptive fields of convolution based on object sizes at multiple output layers. However, in our view, these methods either simply selecting multiple output layers based on the sizes of receptive fields [11] [4] [5], or using feature fusion to expand the receptive field on a single output layer [44] [55] to obtain multi-scale receptive fields, are lack of enhancing entire feature hierarchy for multiscale pedestrian detection. This motivates us to construct an aggregation feature representation to enhance semantic information and localization signals for scale-aware pedestrian detection.

Motivated by above insight and analysis for the representation of feature hierarchy pyramid, we propose a scale-aware hierarchical detection network for pedestrian detection under large scale variations. First, we accomplish feature aggregation based on FPN [43] to enhance semantic information and localization signals for feature representation, by merging the lateral connection, the top-down path and the bottom-up path. Furthermore, in view of the feature differences for pedestrians under different scales, scale-aware hierarchical detection network is designed to learn adaptively perceive pedestrian instances within certain scale ranges, by probing the feature differences with different scales from augmented pyramid features.

To sum up, our work possesses the following contributions:

1) We introduce a cross-scale features aggregation module to enhance feature pyramid representation by fusing robust semantic and accurate localization for pedestrians with different scales, which accomplishes feature augmentation from the lateral connection, the top-down path and bottom-up path.

2) A novel scale perception strategy by normalized Gaussian fate function is designed to integrate multiple detection heads into a unified framework through adaptively perceiving the cross-scale features aggregation module for scale-aware hierarchical detection network.

3) Experimentally, compared with the state-of-the-art method FasterRCNN+ATT [44], the log-average miss rate of pedestrian detection is reduced by 11.98% for medium scale pedestrians (between 30-80 pixels in height), and 14.12% for...
whole scale pedestrians (above 20 pixels in height) on Caltech benchmark.

II. RELATED WORK

There has been lasting research activities on pedestrian detection with vast literature. Before the emergence of CNN, hand-crafted features have been widely used to obtain good performance for pedestrian detection, including HOG [10], Edgelets [38], ICF [21] and its variants ACF [28], LDCF [17], [22], SCF [23]. The most popular pedestrian detector is deformable part model (DPM) [50], which combines rigid root filter and deformable part filters based on HOG feature pyramid and latent SVM classifiers for detection.

Deep ConvNet due to its stronger feature representation ability exhibits obvious performance gains on pedestrian detection [28, 13, 29, 18, 2]. CCF [27] absorbs merits from filtered channel features and Convolutional Neural Networks (CNN), and transfers low-level features from pre-trained CNN models to feed the boosting forest model for pedestrian detection. ConvNet [41] uses an unsupervised method based on convolutional sparse coding to pre-train CNN for pedestrian detection. Deep Parts [13] consists of extensive part detectors, and each part detector is a strong detector that can detect pedestrian by observing only a part of a proposal. SDP [11] investigates scale-dependent pooling and layer-wise cascaded rejection classifiers from CNN to detect objects. CompACT-Deep [16] leverages both hand-crafted and CNN features to form complexity-aware cascaded detectors for an optimal trade-off between accuracy and speed. Especially, Faster-RCNN [9] has addressed a multiscale region proposal network that shares full-image convolutional features with the detection network, leading to an excellent performance for pedestrian detection.

However, spatial scale variation is one of main challenge for pedestrian detection due to the large variance of instance scales in a cross-scenario. To address the issue, an upsampling or dilated operations [5], [11] are employed to alleviate the decline which just adopts a fixed set of filter respective fields existed in Faster-RCNN [9]. MS-CNN [5] combines multiple output layers by feature upsampling of deconvolution to produce a strong multi-scale object detector. SA-FastRCNN [4] exploits multiple built-in subnetworks by a divide-and-conquer strategy to adaptively detect pedestrians across scales. RPN+BF [7] reuses the high-resolution convolutional features of RPN by cascaded boosted forests for multiscale pedestrian detection. ADM [49] executes sequences of coordinate transformation on multi-layer feature maps to deliver accurate pedestrian locations. TridentNet [53] constructs a parallel multi-branch architecture to expend receptive fields on the detection of different scale objects through dilated convolution. However, these methods have not effectively to fuse the robust semantic information of targets existed in high-level convolutional layers and the precise localization signals of the lower convolutional layers for multiscale pedestrian detection.

To exploit strong semantic for prediction, FPN [43] augments a top-down pathway and lateral connections to propagate high-level semantic information for reasonable classification capability. DSSD [47] adopts deconvolution layers to aggregate context and the high-level semantics for enhancing shallow features. M2Det [3] presents multi-level feature pyramid network to fuse multiscale features for detecting objects of different scales. On the other hand, many fine details and higher resolution existed in low-level feature maps are benefits for localization accuracy. PANet [32] builds a strong indicator to accurately localize instance segmentation by a pathway with clean lateral connections from the low level to top ones. DLA [48] augments standard architectures with deeper aggregation across layers to obtain stronger layer-wise multi-scale representation capability. STDN [29] is equipped with embedded super-resolution scale-transfer layers to explore the inter-scale consistency nature across multiple detection scales. Recently, NAS-FPN [52] consists of a series of merging cells to fuse features across scales by a combination of top-down and bottom-up connections. Res2Net [46] constructs hierarchical residual-like connections within one single residual block to capture multi-scale features at a granular level.

Inspired by these observations and analysis of feature fusion to multiscale detection, in this paper, we explore a scale-aware hierarchical detection network for multi-scale pedestrian detection, by aggregating the strong semantic information from high-level features and the accurate localization signals from low-level layer to enhance pyramidal feature representations.

III. APPROACH OVERVIEW

A high-level overview of our approach architecture is shown in Fig. 2. Our proposed approach consists of two main components: cross-scale features aggregation module and scale-aware hierarchical detection network. The cross-scale features aggregation module is built on Feature Pyramid Network (FPN) [43] to enhance representation ability of pyramid features. FPN shows significant improvement as a generic feature extractor for object recognition, significantly which propagates semantically strong features to enhance pyramid features with reasonable classification capability by the top-down path. Similarly, with many fine details and strong responses of local patterns are existed in low-level convolutional layers, which are benefits for high localization accuracy. For this reason, we design a cross-scale features aggregation module to adaptively aggregate features of pyramidal hierarchy to enhance the localization capability.

Further, the scale-aware hierarchical detection network based on the Fast R-CNN framework [6] combines the complementary detection branches on hierarchical pyramid feature maps from the cross-scale features aggregation module. And the detection heads in hierarchical detection network pretrained on ImageNet based on ResNet [33] are all sharing parameters for each proposal to learn scale-aware hierarchical weights by minimizing the error rate for
pedestrians with different scales, regardless of their feature levels.

A. CROSS-SCALE FEATURES AGGREGATION MODULE

Feature Pyramid Network (FPN) [43] shows significant improvement as a generic feature extractor for object recognition, which propagates semantically strong features to enhance pyramid features with reasonable classification capability. Followed by previous evidence on the benefits of the strategy of feature approximation [28], we denote the output of last residual blocks as \{C_1, C_2, C_3, C_4, C_5\} for conv1, conv2, conv3, conv4, and conv5 in ResNet. And given a list of multi-scale pyramid features \{P_1, P_2, P_3, P_4, P_5\} from FPN [43], where \(P_i\) represents the feature at pyramid level \(i\).

However, the feature fusion from FPN only directly builds on the lateral connection and the top-down pathway, ignoring the impact of bottom-up path augmentation to enhance feature representation for accurate localization signals existing in low-level convolutional layers.

Our goal is to find a transformation function \(f\) that can effectively aggregate multi-scale features and output a list of new features: \(X_{out} = f(X^{in})\), \(X^{in}\) may be \(C_i\), \(P_i\), or their union. Different from the feature augmentation generated by FPN, we propose a cross-scale features aggregation module (CFAM) to merge a bottom-up pathway to FPN. Specifically, we use \(\{H_1, H_2, H_3, H_4, H_5\}\) to denote augmented feature pyramid and in which the spatial resolution of feature maps is gradually upsampled with factor 2 from \(H_i\) to \(H_{i-1}\). As shown in Fig. 3(b), each feature aggregation module takes a convolutional feature map \(C_{i-1}\) with higher resolution, an identify mapping feature maps \(C_i\) and a coarser feature map \(H_{i+1}\) with stronger semantic to generate the augmented feature map \(H_i\). Note that we adopt an average pooling to downsample the spatially finer feature maps, which implies to directly propagate strong responses of local patterns from low-level pyramid levels for accurately localization by the bottom-up augmented pathway.

The key idea of CFAM is to adaptively aggregate multi-scale context information from feature maps of the convolutional layers at adjacent scales to generate more discriminative features. As shown in Fig. 3(b), each aggregating module merges a top-down path, lateral connections and a bottom-up augmented path by addition. The feature aggregating module takes a convolutional feature map \(C_{i-1}\) with higher resolution, an identify mapping feature map \(C_i\) and a coarser feature map \(H_{i+1}\) with stronger semantic to generate the fused feature map \(H_i\). This is an iterated process to build augmented feature pyramid until to the finest resolution map \(H_5\). At the beginning of iteration, we adopt a \(1 \times 1\) convolutional layer on \(C_5\) to produce the coarsest but semantically strongest resolution map \(H_5\). Then the lower-level feature map \(C_{i-1}\) goes through a \(2 \times 2\) average pooling layer with stride 2 to reduce the spatial size to generate the down-sampled feature map in the bottom-up augmented pathway. Each element of feature map \(H_{i+1}\), the down-sampled feature map and the identify mapping feature map \(C_i\) are added to generate fused feature map. Finally, we append a \(1 \times 1\) convolution on each merged map to generate the final augmented feature map \(H_i\) for following sub-networks, which is used to reduce the aliasing effect of upsampling and downsampling. In the feature aggregation module, these augmented feature maps are respectively
corresponding to \{C_3, C_4, C_5\} with the same spatial sizes, and we set 1024-channel outputs for each augmented feature pyramid \{H_3, H_4, H_5\} to scale-aware hierarchical detection network.

**B. SCALE-AWARE HIERARCHICAL DETECTION NETWORK**

The coverage of multiple scales is a critical problem to different scale ranges for pedestrian detection. Different from the multi-scale mechanism of the RPN [9], we divide the region proposals into three scales (near, medium, and far) from higher convolutional layers C_4, each scale is transported an augmented feature pyramid H_i to detect pedestrian instances within certain scale ranges as shown in Fig. 4. We hypothesize that pedestrian instances with different scales can be better modeled by hierarchical detection network with the valid range of filter receptive fields. Specifically, each pedestrian anchor scale needs to effectively match the receptive field size of the ROI pooling though different spatial pooling structure.

Let \( L^m(X, Y|W) \) represents multi-task loss function for each pedestrian proposal under specific feature level \( H_m \), and is given by:

\[
L^m(X, Y|W) = L^m_{\text{cls}}(p_i, \hat{p}_i) + \lambda \hat{p}_i L^m_{\text{loc}}(b_i, \hat{b}_i).
\]

where \( \hat{p}_i \) is 1 if the anchor is labeled positive, otherwise is 0. \( p_i \) is the predicted probability of the anchor being a proposal. \( \hat{b}_i = (\hat{b}_i^x, \hat{b}_i^y, \hat{b}_i^w, \hat{b}_i^h) \) represents the ground-truth box associated with a positive anchor, and \( b_i = (b_i^x, b_i^y, b_i^w, b_i^h) \) represents the parameterized coordinates of the predicted bounding box. The classification loss \( L^m_{\text{cls}} \) is the softmax loss from specific feature level \( H_m \). For the regression loss, we use \( L^m_{\text{loc}} = R(b_i - \hat{b}_i) \) where \( R \) is the robust loss function (smooth-L1) defined in [6]. The term \( p_i L^m_{\text{loc}} \) means the regression loss which is activated only for positive anchors \( \hat{p}_i = 1 \) and is disabled otherwise \( \hat{p}_i = 0 \).

To adaptively match valid feature level and anchor scale for multiscale pedestrian detection, SDP [11] adopts a hard isolation strategy by the pixels in height of an object proposal to detect multiscale objects. SA-FastRCNN [4] exploits a soft isolation strategy by Sigmoid gate function defined over the object proposal sizes to generate scale-aware weighting for multi-scale detection subnetworks. In this paper,
we design a novel scale perception strategy by normalized Gaussian gate function for scale-aware hierarchical detection network (SHDN) as shown in Fig. 4, and the model loss function is defined as:

\[
L(W) = \sum_{m=1}^{M} \sum_{i \in U} \omega_m L^m(X_i, Y_i | W) 
\]

where \(M\) is the number of hierarchical feature pyramid as mentioned in Section III A, \(U = \{ (X_i, Y_i) \}_{i=1}^{N} \) contains the training examples of multi-scale for pedestrian instances, and \(\omega_m\) is the normalized scale-aware weight to corresponded hierarchical loss \(L^m(X_i, Y_i | W)\), and is initialized by \( \omega_m = e^{\hat{\omega}_m} / \sum_{i=1}^{M} e^{\hat{\omega}_i}, \hat{\omega}_m = e^{-(s - \hat{s}_m)^2/2(\gamma_m)^2} \).

Here \( s = log_2(h) \) denotes the height scale of the pyramids which has already been normalized to a narrow range prior to detection, \(\hat{s}_m\) and \(\gamma_m\) is the average height scale and the scaling coefficient for specific feature level \(H_m\), respectively. Given a sliding window, the Gaussian function with lower \(\gamma_m\) tends to enlarge the gap between the weights for pedestrian instances from different scale ranges. Based on the ResNet structure, the output size of ROI pooling is \(7 \times 7\) with a stride chosen from the set of \(\{8, 16, 32\}\) to construct deep network \(\{C_3, C_4, C_5\}\), then the valid receptive fields for hierarchical feature pyramid \(\{H_3, H_4, H_5\}\) is \(\{56, 112, 224\}\) pixels for the height of bounding box, respectively. Consequently, we assign scale-aware parameters \((\hat{s}_m, \gamma_m)\) as \((5.8, 1.25), (6.8, 2), (7.8, 1.25)\) for hierarchical feature pyramid \(\{H_3, H_4, H_5\}\), respectively. Note that we optimize multi-task loss function being shunt to scale-aware hierarchical detection module by the scale-aware weights parameters \((\hat{s}_m, \gamma_m)\), and all hyper parameters after ROI pooling layers are shared for all levels of the hierarchical feature pyramid.

For efficient training the scale-aware hierarchical detection network, sampling is used to compensate for the imbalance from the distribution of positive samples \(U^+_m\) and negative samples \(U^-_m\). In this paper, we adopt random sampling and bootstrapped sampling to collect a final set of negative samples, such that \(U^-_m = \xi U^+_m\). We utilize random sampling to randomly select easy negative samples according to a uniform distribution. Because hard negatives mining has large influence on the detection accuracy, bootstrapping sampling is exploited to improve detection performance by ranking the negative samples according to their objectness scores. On the other hand, to avoid the heavily asymmetric of positive samples \(U^+_m\) and negative samples \(U^-_m\) resulting in for each specific detection layer, the cross-entropy terms of positives and negatives are weighted in formula (3), which guarantee that each detection layer have enough positive samples to cover a certain range of scales.

\[
L_{cls} = \frac{1}{1 + \xi} \sum_{i \in U^+_m} -\log p_i(X_i) \\
+ \frac{\xi}{1 + \xi} \sum_{i \in U^-_m} -\log p_0(X_i) 
\]

IV. EXPERIMENTS
A. EXPERIMENTS DETAILS
Following ResNet [33] pretrained on ImageNet, we fine-tunes the convolutional neural network to extract visual features from observed video frames on Caltech training dataset. The convolutional layers and max pooling layers of the ResNet network are used as the shared convolutional layers before the Region-of-Interest (RoI) pooling layer to produce feature maps from the entire input image. The last convolutional block in ResNet is 2048-d, and we employ a randomly initialized 1024-d \(1 \times 1\) convolutional layer for reducing dimension. And we use single-scale training in which the size of the input image is resized as 600 pixels on the shortest side. The scale-aware feature aggregation network is trained with Stochastic Gradient Descent (SGD) with momentum of 0.9, and weight decay of 0.0005. As [9], [30] demonstrate that mining from a larger set of candidates (e.g., 2000) has no benefit, we use 300 RoIs for both training and testing of this paper. We fine-tune scale-aware hierarchical detection network using a learning rate of 0.001 for 20k mini-batches. Each mini-batch consists of 128 randomly sampled object proposals in one randomly selected image, where in 32 positive object proposals and the rest 96 negative object proposals. A positive label of pedestrian is assigned when \(IOU \geq 0.5\) between the object proposal and the ground truth box, and the negative label to RoIs if their \(IOU \leq 0.3\) for all ground-truth boxes. The whole scale-aware hierarchical detection network is trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory.

B. ABLATION EXPERIMENTS
1) EVALUATING THE CROSS-SCALE FEATURES AGGREGATION MODULE
As mentioned in [7], the Region Proposal Network (RPN) in Faster R-CNN indeed performs well as a stand-alone detector, but the downstream classifier degrades the pedestrian detection performance. In this subsection, we investigate cross-scale features aggregation module in terms of detection quality, evaluated by the log-average miss rate of pedestrian detection under \(IOU = 0.5\) on Caltech dataset.

First of all, we evaluate high-level convolutional layer (from ResNet-50-C3 to ResNet-50-C5) of ResNet [33] to extract ROI features to detect pedestrian by using a set of anchor scales from RPN. As shown in Table 1(a)(b)(c), for illustrating the effects of high-level convolutional features in ResNet-50 to detect pedestrian instances, the higher convolutional layers (e.g., \(C_4, C_5\)) obviously perform better than lower-level convolutional layers (e.g., \(C_3\)) for pedestrian instances with near scale. This can be attributed to higher-level convolutional features with more robust semantic information than lower-levels.

Further, compared to adopt the simple high-level convolutional layer (e.g., \(C_3, C_4, C_5\)) to detect pedestrian, FPN (e.g., \(P_3, P_4, P_5\)) fused the semantically strong features from higher convolutional layer to enhance pyramid features.
for classification capability. Especially, $P_3$ gets 76.59% MR for pedestrian detection under far scale, and decreases by 10.16% MR for medium scale pedestrian instances to $C_4$ as shown in Table 1(d). However, FPN only directly builds on the lateral connection and the top-down pathway for feature fusion, ignoring the impact of bottom-up pathway which is benefit for accurate localization. Compared to the improved performance of $P_3$ under far scale and medium scale, $P_4$ has degraded the pedestrian detection performance shown in Table 1(e), which may be due to lack of accurate localization signals existing in lower convolutional layers. Therefore, we propose a cross-scale features aggregation module (CFAM) to fuse semantic information and localization signals by adding a bottom-up augmented pathway to FPN. As shown in Table 1(g), $H_3$ has achieved the best pedestrian detection performance under far-scale and medium-scale, up to 72.83% MR and 36.50% MR respectively. Note that $H_4$ achieves 43.69% MR for pedestrians with all scales.

2) THE ROLE OF SCALE-AWARE HIERARCHICAL DETECTION NETWORK

In this subsection, the contribution of proposed scale-aware hierarchical detection network is evaluated by log-average miss rate under $IOU = 0.5$ on Caltech testing dataset for pedestrian detection. We conduct comparison experiments to verify the effectiveness of the proposed method within a single output layer and multiple output layers for detection heads. As shown in Table 2 (a)(b)(c), we compare the single output layer $H_3$, $H_4$ and $H_5$ as detection head from proposed cross-scale features aggregation module for pedestrian detection under different scales. We found that the $H_3$ performs better than other single output layer on log-average miss rate for pedestrian detection under far and medium scales. For near scale, $H_4$ has achieved the best detection performance in a single output layer, up to 2.12% MR with a relative improvement of 14.23% over the competitor $H_3$.

However, detecting pedestrian only from a single output layer cannot effectively cover the multiscale pedestrians appeared large scale variations, due to lacking of the scale complementary from multiple feature layers with different sizes of filter receptive fields. To effectively combine multiple output layers from feature pyramid for pedestrian detection, we adopt the scale-aware parameters $(\delta_m, \gamma_m)$ to initialize learning hierarchical weights $\omega_m$ for optimizing multi-task loss function in formula 2. Specifically, we assign scale-aware parameters $(\delta_m, \gamma_m)$ as $(\{5.8, 1.25\}, \{6.8, 2\}, \{7.8, 1.25\})$ for hierarchical feature pyramid $\{H_3, H_4, H_5\}$ respectively. In Table 2(d), combining layers $\{H_3, H_4\}$ gains 42.58% MR for all scales on Caltech benchmark, improving by 1.11% compared to the single output layer $H_4$, and achieves best detection performance 65.86% MR for far scale pedestrian instances. Particularly, combining layers $\{H_4, H_5\}$ does not improve pedestrian detection performance on medium and far scales, but achieves better detection performance 1.25% MR on near scale shown in Table 2(e). The reason may be attributed to our proposed hierarchical scale-aware detection network that each detection branch is used to learn a proper pyramid feature layer to focus on pedestrian instances within certain scale ranges. Moreover, the log-average miss rate is reduced to 40.39% for all scales pedestrian detection, 28.77% for medium scale and 1.08% for near scale, by combining layers $\{H_3, H_4, H_5\}$ as shown in Table 2(f). Note that combining $\{H_3, H_4, H_5\}$ gets the best performance compared to $\{C_3, C_4, C_5\}$, $\{P_3, P_4, P_5\}$, and $\{P_2, P_3, P_4, P_5\}$ shown in Table 2(g-i). The experiments demonstrate that the proposed hierarchical scale-aware detection network is more flexible and is able to take advantage of different sizes of filter receptive fields from multiple level pyramid features for large variance in pedestrian instance scales.

C. COMPARISON WITH STATE-OF-THE-ARTS

In this section, the performance of proposed algorithm is fully evaluated to the state-of-the-art methods on Caltech [20] and ETH [18] datasets. As [54] proposed the evaluation criteria, the log-average miss rate is used to summarize the detector performance. The performance is computed by averaging miss rate at FPPI rates evenly spaced in log-space within the range of $10^{-3}$ to $10^3$. The experiments demonstrate that jointing cross-scale features aggregation module and scale-aware hierarchical detection network outperforms the state-of-the-art pedestrian detection algorithms, especially on pedestrian instances with small sizes.
TABLE 2. Comparisons of pedestrian detection results by log-average miss rate (MR) under IOU=0.5 on Caltech dataset.

| Detection Network               | Proposals | ROI features | $M_{R_{all}}$ | $M_{R_{f}}$ | $M_{R_{m}}$ | $M_{R_{n}}$ |
|--------------------------------|-----------|--------------|---------------|-------------|-------------|-------------|
| (a) Based on a single output layer | $C_4$     | $H_1$        | 46.52%        | 72.83%      | 36.50%      | 16.35%      |
| (b) Based on a single output layer | $C_4$     | $H_4$        | 43.69%        | 86.50%      | 38.08%      | 2.12%       |
| (c) Based on a single output layer | $C_4$     | $H_5$        | 84.84%        |             | 85.82%      | 2.51%       |
| (d) Based on multiple output layers | $C_4$    | $\{H_5, H_4\}$ | 42.58%        | 65.86%      | 30.53%      | 1.48%       |
| (e) Based on multiple output layers | $C_4$    | $\{H_4, H_5\}$ | 44.37%        | 90.56%      | 33.66%      | 1.25%       |
| (f) Based on multiple output layers | $C_4$    | $\{H_3, H_4, H_5\}$ | 40.39%        | 70.69%      | 28.77%      | 1.08%       |
| (g) Based on multiple output layers | $C_4$    | $\{C_3, C_4, C_5\}$ | 50.47%        | 78.67%      | 37.43%      | 2.22%       |
| (h) Based on multiple output layers | $C_4$    | $\{P_3, P_4, P_5\}$ | 46.09%        | 74.42%      | 31.16%      | 1.79%       |
| (i) Based on multiple output layers | $C_4$    | $\{P_2, P_3, P_4, P_5\}$ | 45.12%        | 71.54%      | 32.83%      | 1.46%       |

FIGURE 5. Quantitative results of comparisons on Caltech dataset.

1) COMPARISON WITH STATE-OF-THE-ART METHODS ON CALTECH DATASET

The Caltech pedestrian dataset consists of approximately 10 hours of 640*480 30Hz video taken from a vehicle driving through regular traffic in an urban environment, which includes about 250,000 frames with a total of 2300 unique pedestrians. Similar to other relevant publications previously [13], [16], [17], [24], we adopt the different spatial scale pedestrians to evaluate our method on the Caltech testing dataset, and choose the Caltech training dataset and the INRIA training dataset [10] as our training set. The experimental evaluations of our proposed method with the state-of-the-art methods are constructed on the Caltech testing dataset, including LDCF [22], ACF+SDt [34], RPN+BF [7], MS-CNN [5], CompACT-Deep [16], TA-CNN [24], SA-FastRCNN [4], FasterRCNN+ATT [44], and AR-Ped [7].

To evaluate the effectiveness of our proposed scale-aware hierarchical detection network, quantitative results of comparison are presented for different scale ranges of pedestrian instances on Caltech dataset. Fig. 5 shows the comparison results of the log-average miss rate for pedestrians under different scale ranges. It can be observed that our proposed method significantly outperforms other methods and achieves the lowest log-average miss rate 28.77% on Caltech dataset of the medium scale shown in Fig. 5(a), which is lower than the state-of-the-art approach FasterRCNN+ATT [44] by 11.98%. As the similar trend shown in Fig. 5(b), our approach achieves 7.41% log-average miss rate for pedestrian instances
taller than 50 pixels in height, second only to the state-of-the-art approach AR-Ped [2].

For pedestrian instances in far scale ranges, most methods exhibit dramatic performance drops as shown in Fig. 5(c). While our proposed method outperforms better than the available state-of-the-art competitors, it is difficulty to identify pedestrians reliably for small-size pedestrian instances under 30 pixels in height. In Fig. 5(c), the log-average miss rate is reduced to 70.69%, improved 20.25% compared to FasterRCNN + ATT [44]. This is similar to human performance that is also quite good in the large scales but degrades noticeably at the medium and far scales. Significantly, for pedestrian instances in whole scale span ranges, our approach achieves the log-average miss rate 40.39% for all pedestrian instances taller than 20 pixels in height, better than the current FasterRCNN + ATT [44] by 14.12% as shown in Fig. 5(d).

Fig. 6 shows the detection results of our proposed scale-aware hierarchical detection network on Caltech dataset. The green dotted bounding boxes represent true positive windows when the intersection over union (IoU) between the detected window and the ground truth (green solid bounding box) exceeds 50%. Otherwise, the bounding boxes denote false positive windows by the red dotted bounding box. As shown in Fig. 6, the most of pedestrian instances with different scale ranges can be detected by our proposed approach. Moreover, because of adaptively perceiving the augmented feature level with different resolutions for special-scale pedestrians, the medium-size and small-size pedestrian instances also can be detected in proposed scale-aware hierarchical detection network. The red dotted bounding box represents the positive pedestrians which are not marked by the ground truth as shown in Fig. 6. This experiment shows that the usage of jointing the cross-scale features aggregation module and scale-aware hierarchical detection network for pedestrian detection outperforms the state-of-the-art algorithms, especially for pedestrian instances from medium and small scale ranges.

2) COMPARISON WITH STATE-OF-THE-ART METHODS ON ETH DATASET

The ETH benchmark dataset consists of 3 testing video sequences with a resolution of 640*480, and a frame rate of 13FPS. Studies in [7], [37] report that state-of-the-art algorithms have the remarkable detection performance evaluated on ETH dataset, including ChnFtrs [21], MultiFtr+Motion [35], JointDeep [37], pAUCBoost [40], ConvNet [41], DBN-Mut [12], SpatialPooling [39], TACNN [24], and RPN+BF [7]. As most approaches are trained on the INRIA training dataset [10], our proposed method is also trained on the INIRA training dataset. As Fig. 7(a) the log-average miss rate of our proposed approach achieves 44.75% next to the state-of-the-art SpatialPooling [39] 43.36% for pedestrians under medium scale. Similar trend to what we have observed for pedestrian with near scale,
our approach achieves 20.49% log-average miss rate, second only to the best available competitor’s RPN+BF [7] as shown in Fig. 7(b). Significantly, for pedestrian instances above pixels taller than 80 pixels in height, our approach gets 16.84% log-average miss rate, improving 0.78% compared to the state-of-the-art RPN+BF [7] shown in Fig. 7(c). Moreover, for a more challenging with large variation of scale (above 20 pixels in height), the log-average miss rate of our approach reduces 3.98% over RPN+BF [7] on ETH dataset as shown in Fig. 7(d). The results demonstrate that our proposed method has a substantially better detection performance for the multiscale pedestrian instances appeared large scale variations in natural scenes.

The pedestrian detection results of our proposed method are shown in Fig. 8 on ETH dataset. As shown in Fig. 8, the green dotted boxes demonstrate the detection results of our approach. Our proposed approach adaptively perceives the augmented feature level to generate the final detection results for special-scale pedestrian detection by scale-aware hierarchical detection network. And the small-size pedestrian instances also can be detected, where the red dotted bounding box represents the positive pedestrians which are not marked by the ground truth as shown in Fig. 8. One can observe that our method can successfully detect most of the pedestrian instances, especially for pedestrians with large scale variations.

V. CONCLUSION

This study describes an effective approach to detect pedestrian instances with different scale ranges. The proposed cross-scale features aggregation module adaptively fuses hierarchical features to enhance feature pyramid representation by merging the lateral connection, the top-down path and bottom-up path. Moreover, probing the differences of local features with different sizes of receptive fields, the proposed scale-aware hierarchical detection network effectively integrates multiscale pedestrian detection into a unified framework through adaptively perceiving the augmented feature level for special-scale pedestrian detection. Experimentally, compared with the state-of-the-art FasterR-CNN+ATT [44], the log-average miss rate of pedestrian detection is reduced by 11.98% for medium scale pedestrians (between 30-80 pixels in height), and 14.12% for whole scale pedestrians (above 20 pixels in height) on the Caltech benchmark.

REFERENCES

[1] Z. Chen, L. Zhang, A. M. Khattak, W. Gao, and M. Wang, “Deep feature fusion by competitive attention for pedestrian detection,” IEEE Access, vol. 7, pp. 21981–21989, 2019.
[2] G. Brazil and X. Liu, “Pedestrian detection with autoregressive network phases,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 7231–7240.
[3] Q. Zhao, T. Sheng, and Y. Wang, “MD2Det: A single-shot object detector based on multi-level feature pyramid network,” in Proc. AAAI Conf. Artif. Intell., Honolulu, HI, USA, 2019, pp. 9259–9266.
[4] J. Li, X. Liang, and S. Shen, “Scale-aware fast R-CNN for pedestrian detection,” in Proc. CVPR, Honolulu, HI, USA, Jun. 2017, pp. 985–996.
[5] T. Cai, Q. Fan, R. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in Proc. Eur. Conf. Comput. Vis., Amsterdam, The Netherlands, 2016, pp. 354–370.
[6] R. Girshick, “Fast R-CNN,” in Proc. ICCV, Santiago, Chile, Dec. 2015, pp. 1440–1448.
[7] L. Zhang, L. Lin, and X. Liang, “Is faster R-CNN doing well for pedestrian detection?” in Proc. ECCV, Cham, Switzerland: Springer, 2016, pp. 443–457.
[8] C. Fei, B. Liu, Z. Chen, and N. Yu, “Learning pixel-level and instance-level context-aware features for pedestrian detection in crowds,” IEEE Access, vol. 7, pp. 94944–94953, 2019.
[9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. NIPS, Montreal, QC, Canada, 2015, pp. 1–9.
[10] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), San Francisco, CA, USA, Jun. 2005, pp. 886–893.
[11] F. Yang, W. Choi, and Y. Lin, “Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, CA, USA, Jun. 2016, pp. 2129–2137.
[12] W. Ouyang, X. Zeng, and X. Wang, “Modeling mutual visibility relationship in pedestrian detection,” in Proc. CVPR, Portland, OR, USA, Jun. 2013, pp. 3222–3229.
[13] Y. Tian, P. Luo, X. Wang, and X. Tang, “Deep learning strong parts for pedestrian detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1904–1912.
[14] D. Hoiem, Y. Chodpathumwan, and Q. Dai, “Diagnosing error in object detectors,” in Proc. Eur. Conf. Comput. Vis., Florence, Italy, 2012, pp. 340–353.
[15] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, “Robust multi-resolution pedestrian detection in traffic scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 3033–3040.
[16] Z. Cai, M. Saberian, and N. Vasconcelos, “Learning complexity-aware cascades for deep pedestrian detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 3361–3369.
[17] S. Zhang, R. Benenson, and B. Schiele, “Filtered channel features for pedestrian detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 1751–1760.
[18] P. Dollar, S. Belongie, and P. Perona, “The fastest pedestrian detector in the west,” in Proc. Brit. Mach. Vis. Conf., Aberystwyth, U.K., 2010, pp. 1–11.
[19] P. Dollar, R. Appel, and W. Kienzle, “Crosstalk cascades for frame-rate pedestrian detection,” in Proc. Eur. Conf. Comput. Vis., Florence, Italy, 2012, pp. 645–659.
[20] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, Jun. 2009, pp. 304–311.
[21] P. Dollar, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in Proc. Brit. Mach. Vis. Conf., London, U.K., 2009, pp. 1–11.
[22] W. Nam, P. Dollar, and J. Han, “Local decorrelation for improved pedestrian detection,” in Proc. NIPS, Montreal, QC, Canada, 2014, pp. 424–432.
[23] B. Schiele, R. Benenson, M. Omran, and J. Hosang, “A decade of understanding pedestrian detection,” in Proc. Eur. Conf. Comput. Vis., Vienna, Austria, Jun. 2014, pp. 613–627.
[24] 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), Boston, MA, USA, Jun. 2015, pp. 5079–5087.
[25] J. Hosang, M. Omran, R. Benenson, and B. Schiele, “Taking a deeper look at pedestrians,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 4073–4082.
[26] M. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Proc. Eur. Conf. Comput. Vis., Zurich, Switzerland, 2014, pp. 818–833.
[27] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Convolutional channel features,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 82–90.
[28] 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, Aug. 2014, pp. 1532–1545.
[29] P. Zhou, B. Ni, C. Kang, J. Hu, and Y. Xu, “Scale-transferable object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 528–537.
[30] J. Dai, Y. Li, and K. He, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. NIPS, Barcelona, Spain, 2016, pp. 379–387.
[31] W. Liu, D. Anguelov, and D. Erhan, “SSD: Single shot multibox detector,” in Proc. ECCV, Amsterdam, The Netherlands, 2016, pp. 21–37.
[32] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8759–8768.

[33] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[34] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollar, “Exploring weak stabilization for motion feature extraction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 2882–2889.

[35] S. Walk, N. Majer, K. Schindler, and B. Schiele, “New features and insights for pedestrian detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., San Francisco, CA, USA, Jun. 2010, pp. 1030–1037.

[36] G. Chen, Y. Ding, J. Xiao, and T. X. Han, “Detection evolution with multi-order contextual co-occurrence,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 1798–1805.

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

[38] B. Wu and R. Nevatia, “Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors,” Int. J. Comput. Vis., vol. 75, no. 2, pp. 247–266, Aug. 2007.

[39] S. Paiaktsiatrikou, C. Shen, and A. van den Hengel, “Strengthening the effectiveness of pedestrian detection with spatially pooled features,” in Proc. Eur. Conf. Comput. Vis., Zurich, Switzerland, 2014, pp. 546–561.

[40] S. Paiaktsiatrikou, C. Shen, and A. van den Hengel, “Efficient pedestrian detection by directly optimizing the partial area under the ROC curve,” in Proc. ICCV, Sydney, NSW, Australia, Dec. 2013, pp. 1057–1064.

[41] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, “Pedestrian detection with unsupervised multi-stage feature learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 3626–3633.

[42] 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), Las Vegas, NV, USA, Jun. 2016, pp. 2921–2929.

[43] 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), Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.

[44] S. Zhang, J. Yang, and B. Schiele, “Occluded pedestrian detection through guided attention in CNNs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 6995–7003.

[45] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, “Scale-aware face detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jul. 2017, pp. 6186–6195.

[46] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2Net: A new multi-scale backbone architecture,” 2019, arXiv:1904.01169. [Online]. Available: http://arxiv.org/abs/1904.01169

[47] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. Berg, “DSSD: Deconvolutional single shot detector,” 2017, arXiv:1701.06659. [Online]. Available: http://arxiv.org/abs/1701.06659

[48] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, “Deep layer aggregation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 2403–2412.

[49] X. Zhang, L. Cheng, and B. Li, “Too far to see? Not really!—Pedestrian detection with spatially pooled features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 7036–7045.

[50] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, “Scale-aware trident networks for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6054–6063.

[51] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 743–761, Apr. 2012.

[52] 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), Las Vegas, NV, USA, Jun. 2016, pp. 2874–2883.

[53] S. K. Choudhury, R. P. Padhy, A. K. Sangaiiah, P. K. Sa, K. Muhammad, and S. Bakshi, “Scale aware deep pedestrian detection,” Trans. Emerg. Telecommun. Technol., vol. 30, no. 9, pp. 1–14, Sep. 2019.

[54] T. Liu, M. Elmkavit, and T. Stathaki, “SAM-RCNN: Scale-aware multi-resolution multi-channel pedestrian detection,” in Proc. BMVC, Newcastle, U.K., 2018, pp. 1–13.

---

**XIAOWEI ZHANG** received the B.S. degree in computer science from Shandong Normal University, Linfen, China, in 2009, the M.S. degree in computer science from Shandong Normal University, Jinan, China, in 2013, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2018. He was a Visiting Student at the Bioinformatics Institute (BII), A*STAR, Singapore, from 2016 to 2017. He is currently an Assistant Professor of computer science and engineering at Qingdao University. His current research interests include image/video analysis and understanding, computer vision, and machine learning.

---

**SHUAI CAO** received the B.S. degree in computer science and technology from the Liaoning University of Technology, Jinhzhou, China, in 2018. He is currently pursuing the M.S. degree in computer science and engineering with Qingdao University, China. His current research interests include pedestrian detection and machine learning.

---

**CHENGLIZHAO CHEN** received the Ph.D. degree in computer science from Beihang University, in 2017. He is currently an Assistant Professor with Qingdao University. His research interests include computer vision, machine learning, and pattern recognition.