IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 20, 2020, accepted August 7, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018306

DA-Net: Pedestrian Detection Using Dense Connected Block and Attention Modules

RUIHONG YIN[1], RUFEI ZHANG[2], WEI ZHAO[1], AND FENG JIANG[2]
[1]School of Electronic and Information Engineering, Beihang University, Beijing 100191, China
[2]Beijing Institute of Control and Electronics Technology, Beijing 100038, China

Corresponding author: Wei Zhao (zhaowei203@buaa.edu.cn)

ABSTRACT Pedestrian detection plays an important role in some areas such as autonomous driving, but due to heavy occlusion and various scales, it is still challenging. In this article, we propose an improved pedestrian detection method called DA-Net based on the two-stage detector Feature Pyramid Network (FPN). DA-Net adds Dense Connected Block (DCB), a combination of channel-wise attention module (CWAM) and global attention module (GAM) to the network. FPN can produce features with various scales and semantic information, which is good for the detection of pedestrians on various scales. Due to many small-scale targets in pedestrian detection, we only regard the low layers with enough details of targets in FPN as prediction layers. After several DCBs to deepen the network, prediction layers in our network can encode richer semantic information of targets, which can make the location of a target more precisely. In order to highlight visible parts of occluded pedestrians and ignore occluded parts, CWAM weights each channel of features with different importance. GAM aggregates global information and long-range dependencies for small-scale and occluded targets. Thus, the combination of CWAM and GAM is not only beneficial for coping with occlusion problem in pedestrian detection, but also for gaining environmental information for small-scale targets. Evaluation results on CUHK and CityPersons datasets show that our proposed method achieves improved performance with log-average miss rate reduction of 9.6% on the CUHK dataset and 6.1% on the Heavy subset of CityPersons dataset compared with FPN.

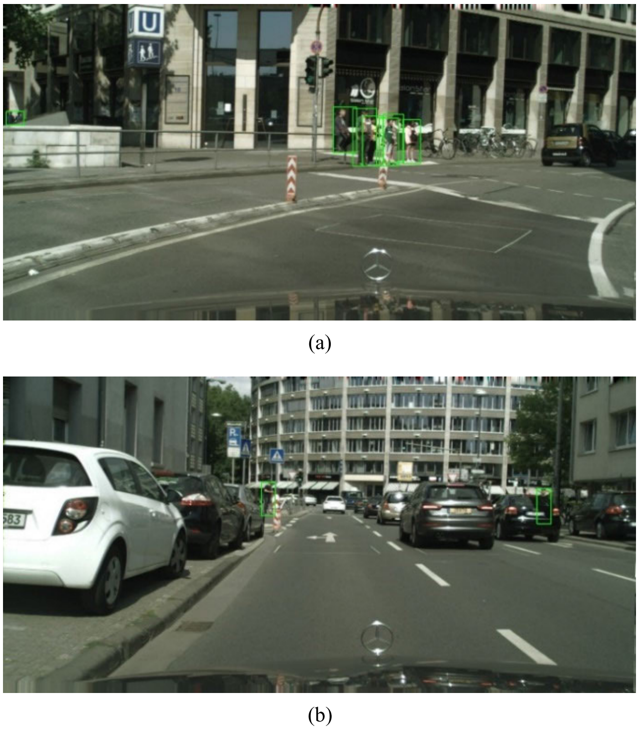
INDEX TERMS Attention module, dense connected block, feature pyramid network, pedestrian detection.

I. INTRODUCTION

Object detection is one of the essential research fields in computer vision, whose task is to find all objects in an image. At present, it is widely utilized in areas such as military, medicine, and intelligent transportation. Object detection includes two processes, i.e. location and classification. In the location stage, models output the coordinates of objects. Models recognize the categories of targets in the classification stage. The existing object detection methods can be mainly divided into two categories: traditional algorithms based on hand-crafted features [1]–[4] and algorithms based on features adopted by convolutional neural networks(CNNs) [5]–[7]. CNN-based detection algorithms can learn the features of target adaptively with strong generalization and feature expression ability. CNN-based detection methods also fall into two categories: one-stage algorithms [7]–[10] and two-stage algorithms [5], [6], [11], [12]. One-stage algorithms are based on regression, like SSD [8] and YOLO [7]. Two-stage algorithms are based on region proposal network, including RCNN [5], Fast RCNN [11], Faster RCNN [6], FPN [12], and so on. Since one-stage algorithms do not have the region proposal network, their detection speed is fast but detection accuracy is usually low. Two-stage algorithms extract the region proposals first and then locate the targets, so detection accuracy is high at the expense of time.

Pedestrian detection is an important branch of object detection. It has attracted a lot of attention in recent years [3], [13]–[16] and plays an important role in some areas such as autonomous driving. Different from general object detection, pedestrian detection has two main characters. First, there are occlusions between pedestrian and pedestrian or pedestrian and background like cars (seen in Fig. 1). Occlusion may lead to missing targets at some times. Second, from Fig. 2, we can see that there is a large variety of scales in the pedestrian dataset CityPersons and many targets have small scales. Small-scale targets have very little information.

The associate editor coordinating the review of this manuscript and approving it for publication was Long Cheng.


(a)

(b)

FIGURE 1. Images with pedestrian targets. From the photos, we can see some targets with heavy occlusion, various and small scales.

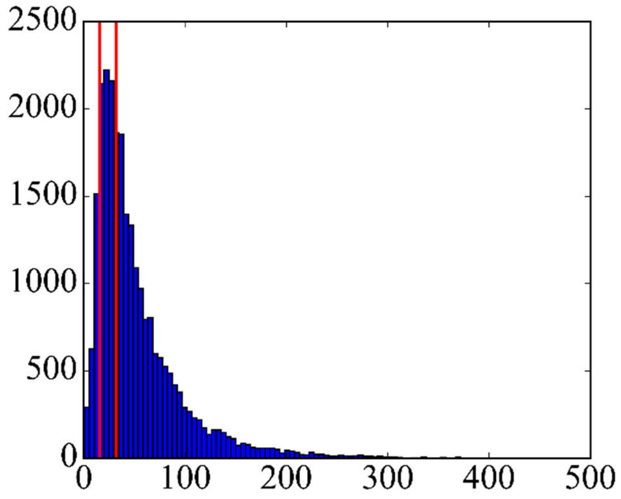


FIGURE 2. Distribution of target scales on CityPersons dataset. The short size of images is resized to 800. The x-axis, y-axis are the scale and number of targets respectively. The two red lines represent scale=16 and 32 respectively.

Therefore, it is not totally appropriate to use common object detection algorithms directly into pedestrian detection. New and effective algorithms are needed. In the last few years, lots of algorithms have been proposed to deal with these two problems. Our method also aims at solving occlusion and various scales in pedestrian detection.

Attention mechanism in CNNs [17] is modeled on the characteristics of human attention, expressed as giving different weights to different parts of a target. Many CNNs with attention mechanism are applied to computer vision tasks such as object detection [18], [19] and pose estimation [20]. In squeeze-and-excitation networks (SENet) [21], the SE block contains a lightweight gating mechanism that applies the network to weight each channel according to their importance. In pedestrian detection, Faster RCNN+ATT [19] also exploits channel attention module for heavy occlusion task. In addition, there are some attention modules used to extract environmental information. For instance, non-local network (NLNet) [22] learns to gain long-range dependencies rather than local information, which is good for detecting targets with little information. However, NLNet is too complex and increases too much computation to the structure.

Another challenging problem in pedestrian detection is that there is a large variation of pedestrian scales and the number of small-scale targets is especially high. Image pyramid and feature pyramid structure are common tools to tackle this problem. SNIP [23] is an example of image pyramid, which uses a scale normalization method to adapt each resolution during multi-scale training. However, SNIP is at the expense of inference time. Feature maps from different layers have variant resolutions and receptive fields, so features with different strides of CNN can be used to cope with scale variation. Besides, feature pyramid can compensate for time expenses in the image pyramid. SSD [8] and MSCNN [24] perform detection with feature pyramid but without fusion between low-level features and high-level features. FPN [12] designs a top-down architecture with lateral connections for building feature maps at multiple scales. It can compensate for missing information in the down-sampling process of CNNs.

Although attention mechanisms and pyramid structures are proposed to overcome obstacles in pedestrian detection, the separate structure is limited to solve the problems in pedestrian detection. There is a need for a model to detect pedestrians with various scales and severe occlusion well. In this article, we propose a network (DA-Net) using Dense Connected Block and attention modules for pedestrian detection. We adopt Feature Pyramid Network (FPN) with ResNet50 [25] to face with various scales in pedestrian detection and only use the low layers for prediction. Deep CNN can extract rich semantic information that is beneficial for the task of location. Thus, our method applies Dense Connected Block (DCB) to increase the depth of prediction layers to gain sufficient information. To reduce the influence of occluded parts during detection, channel-wise attention module (CWAM) is introduced to learn channel-wise weights for visible and occluded parts. Global attention module (GAM) is designed for detecting small-scale and occluded targets because these targets have little information. GAM can capture contextual information for them. In summary, the contributions of this article are as follows,

1) FPN with baseline ResNet50 is used for classification and location. The high layers for prediction are removed and only several low layers remain, which is good for detecting small-scale targets and reducing computation.

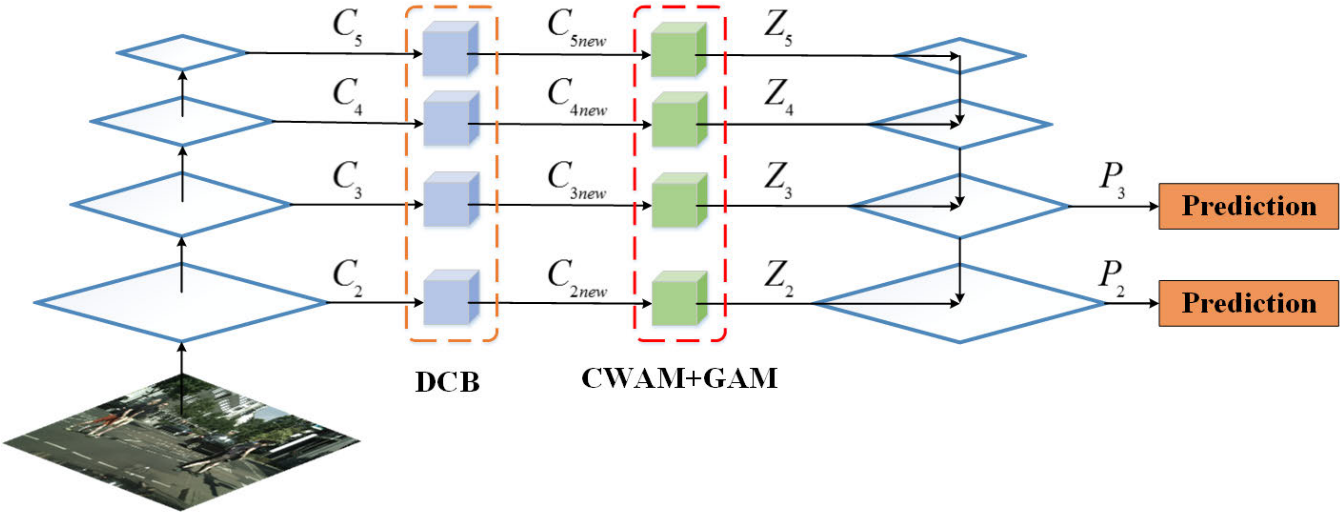


FIGURE 3. Our proposed structure. $\{C_2, C_3, C_4, C_5\}$ are outputs of ResNet50 from the second stage to the fifth stage. DCB is Dense Connected Block. CWAM+GAM represents the combination of CWAM and GAM. $\{P_2, P_3\}$ are used for location and classification.

2) DCB is utilized to deepen CNN and gain richer semantic information of targets, which is beneficial for location task. According to experiment results, we find that richer semantic information in feature maps is important for pedestrian detection.
3) We apply a combination of CWAM and GAM to detect small-scale and occluded targets accurately. By weighting each channel of a feature map, CWAM enables the network to stress the importance of features in visible parts of pedestrians. GAM can extract long-range dependencies and environmental information for small-scale and occluded targets.

The rest of this article is organized as follows. Section II describes our proposed method to tackle problems in pedestrian detection. We discuss the experiment results and explore the effectiveness of each component in our network in Section III. In Section IV, we draw conclusions on this work.

II. THE PROPOSED METHOD

In this section, we will introduce our proposed structure in Fig. 3. First, we introduce the whole structure in Section II-A. Then, we explain each part of the structure, FPN in Section II-B, DCB in Section II-C, a combination of CWAM and GAM in Section II-D.

A. FRAMEWORK

Taking speed and accuracy into account, we extend FPN with ResNet50 for the variety of target scales in pedestrian detection. The proposed structure is illustrated in Fig. 3. After the outputs of ResNet50 from the second stage to the fifth stage, we add DCB to increase the depth of each branch and obtain richer semantic information for accurate detection. Then, a combination of CWAM and GAM is employed in our model. CWAM is proposed to gain weights for each part of pedestrians. It can enable the network to attach importance to visible parts of pedestrians by learned weights. We apply GAM to encode global information and enrich the features of small-scale and occluded targets. After DCB and the combination of CWAM and GAM, the high layers are up-sampled and then added with previous layers. Because there are a large number of small-scale targets and the low layers have high resolution and keep more details for them, we only use low-level feature maps $\{P_2, P_3\}$ for classification and location.

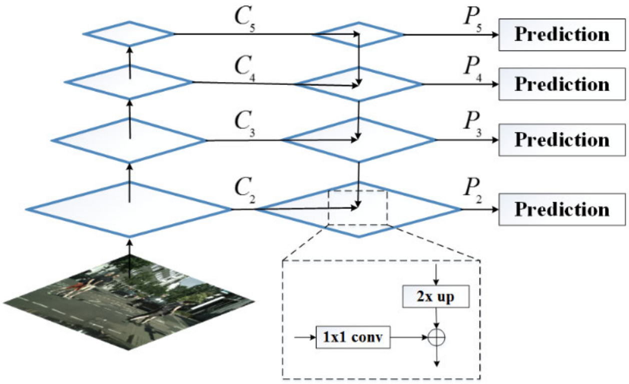


FIGURE 4. Feature Pyramid Network.

B. FEATURE PYRAMID NETWORK

In our framework, we use FPN structure with ResNet50 [25]. FPN builds a feature pyramid structure, which takes a single-scale image as input and outputs feature maps at multiple levels. Adjacent feature maps in FPN have a stride of 2. FPN in Fig. 4 consists of three parts, i.e., a bottom-up pathway, a top-down pathway, and lateral connections. The bottom-up pathway is the feed-forward computation of ResNet50, during which it produces feature maps at multiple scales by down-sampling layers. The feature maps at low layers have large resolutions and more details of targets but are lack of global information. On the contrary, the feature maps at high layers have small resolutions but large receptive fields and rich semantic information. Global information is of great importance for the complicated detection task. Therefore, in the top-down pathway, the high-level features are up-sampled by bilinear interpolation and then enhanced with features from the bottom-up pathway. In this process, lateral connections adopting convolutional layers with kernel size 1 are proposed to make fused feature maps at the same dimensions. During prediction, all prediction layers share classifiers and regressors. Due to the multi-scale prediction layers, the ability to detect targets with small and various scales is significantly improved. In Fig. 2, we know there are plenty of small-scale targets on CityPersons dataset, so we only apply low layers $\{P_2, P_3\}$ for prediction.

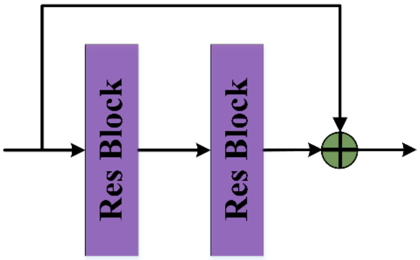


FIGURE 5. Dense Connected Block. Res Block denotes residual block.

C. DENSE CONNECTED BLOCK

In object detection, the task of location needs richer semantic information. The simplest way to gain richer semantic information is to increase the depth of prediction layers. Therefore, we design a structure called Dense Connected Block (DCB) to achieve this goal and improve information flow between blocks. DCB not only contains some convolutional layers to increase the depth but also has skip connections with former blocks. The skip connection has two advantages. One is to avoid the problem of vanishing gradients during back-propagation. Another advantage is to propagate information between blocks. We show the layout of DCB in Fig. 5. The structure can be expressed as follows,

$$B(C_i) = R(R(C_i)) \tag{1}$$

$$C_{inew} = B(C_i) + C_i \tag{2}$$

We denote the outputs of baseline ResNet50 from the second stage to the fifth stage as $\{C_2, C_3, C_4, C_5\}$. $\{C_2, C_3, C_4, C_5\}$ have strides of $\{4, 8, 16, 32\}$ of the input image respectively. C_i is one of the feature maps $\{C_2, C_3, C_4, C_5\}$ in Fig. 3. C_i goes through DCB and the output is C_{inew}. B represents the entirety of two residual blocks. R is the residual block shown in Fig. 6, which is composed of batch normalization, ReLU, and convolutional layers with kernel size 1 or 3. The structure shows dense skip connections between convolutional layers, so we call it as Dense Connected Block. In Section III, we will analyze the influence of adding DCB after different layers.

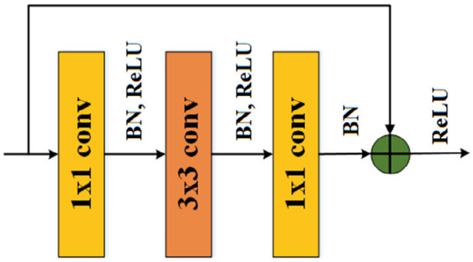


FIGURE 6. Residual Block. 1 × 1 conv, 3 × 3 conv, BN, ReLU represent a convolutional layer with kernel size 1, a convolutional layer with kernel size 3, batch normalization [26], Rectified Linear Unit [27] respectively.

D. A COMBINATION OF CWAM AND GAM

In our structure, we propose a combination of CWAM and GAM in Fig. 7 to solve the problems of small-scale and occluded targets at the same time. In this model, we design a cascade structure of CWAM and GAM. In addition, there is a skip connection between input and output to propagate information.

$$Z_i = H_G(H_C(C_{inew})) + C_{inew} \tag{3}$$

where H_C, H_G denote CWAM and GAM respectively. CWAM can acquire channel-wise information for occluded targets. GAM is adopted to extract long-range dependencies for small-scale and occluded targets.

1) CHANNEL-WISE ATTENTION MODULE

As we know, each channel of a feature map is related to different features of targets. Taking an image with pedestrians as input, we visualize some channels of an intermediate feature map in Fig. 8(c)-(h) to explain this. We can see that different channels of the feature map are associated with different body parts of a pedestrian, such as the head in Fig. 8(c)(d)(h), the upper body in Fig. 8(d), and the lower body in Fig. 8(e)(g) respectively. There is much more serious occlusion in the body regions that Fig. 8(c)-(e) correspond to, which makes the features of neighboring targets confusing. On the contrary, the features in Fig. 8(f)-(h) are more obvious and distinguishable for each pedestrian, which plays a vital role in detecting them. However, standard convolutional blocks treat channel-wise features equally, so they have difficulty in dealing with occlusion problem in the task of pedestrian detection. This encourages us to explore a channel-wise attention module (CWAM) to weight each channel of a feature map. Our CWAM in Fig. 9 can extract inter-channel relationships to guide the network to concentrate more on the visible body regions in Fig. 8(f)-(h) and less on the occluded regions in Fig. 8(c)-(e).

The process of CWAM can be divided into two steps. The first step is to yield a channel-wise vector. In the second step,

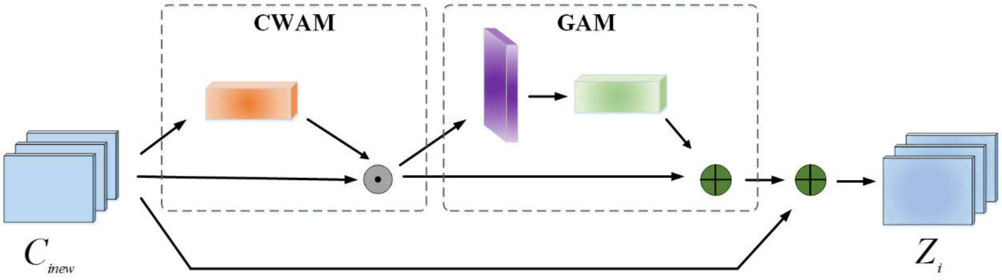


FIGURE 7. A combination of CWAM and GAM.

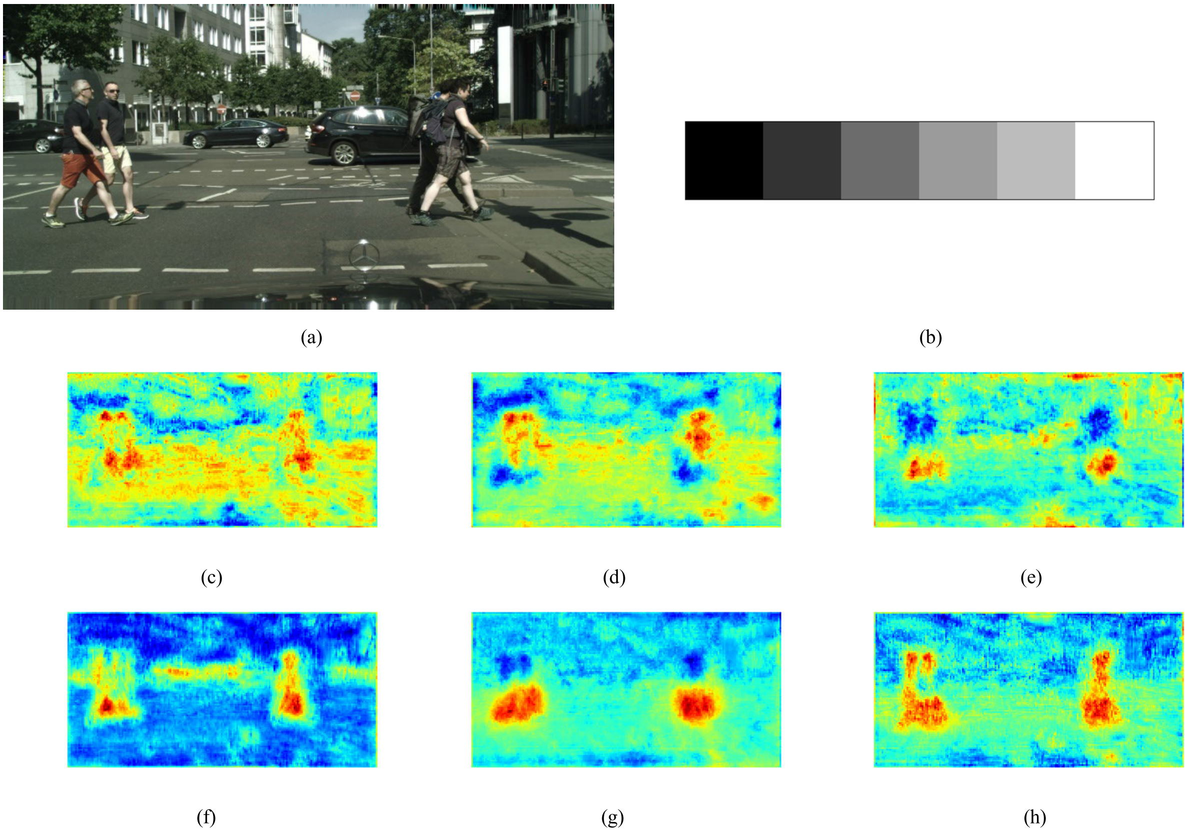


FIGURE 8. Visualization of CWAM. (a) is an input of the network. (c)-(h) are different channels of an intermediate feature map. Each grid in (b) is the weight of channels (c)-(h). The value increases from left to right in (b).

this vector is utilized to weight each channel of a feature map. These can be stated in detail as follows.

In the first step, the feature map $C_{inew} \in \mathbb{R}^{h \times w \times c_1}$ passes through a convolutional layer with kernel size 3 and a ReLU layer to extract the feature $F_1 \in \mathbb{R}^{h \times w \times c_1}$.

$$F_1 = D(C_{inew}) \tag{4}$$

Then, in order to aggregate global information of each channel and gain channel-wise feature maps, for simplicity, global average pooling g is applied on F_1. For the cth channel of F_1, the operation is as follows,

$$\begin{aligned} F_{2c} &= g(F_{1c}) \\ &= \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} F_{1c}(i,j) \end{aligned} \tag{5}$$

where $F_{1c}(i,j)$ is the pixel in the spatial position (i,j) of the cth channel in F_1. F_{2c} is the value of the cth channel in F_2. After that, we develop two fully connected layers W_1, W_2 to gain the relationship between different channels. There is a ReLU δ between fully connected layers to learn the nonlinear connection of channels. As we know, fully connected layers generate plenty of parameters for the network, so we cut down the output dimension of W_1 by 16 times to reduce parameters.

$$F_3 = W_2(\delta(W_1(F_2))) \tag{6}$$

Then, a softmax layer is adopted to normalize the range of F_3 to [0, 1]. The output V is the channel-wise weight vector of CWAM.

$$V_c = \frac{\exp(F_{3c})}{\sum_{c=1}^{c_1} \exp(F_{3c})} \tag{7}$$

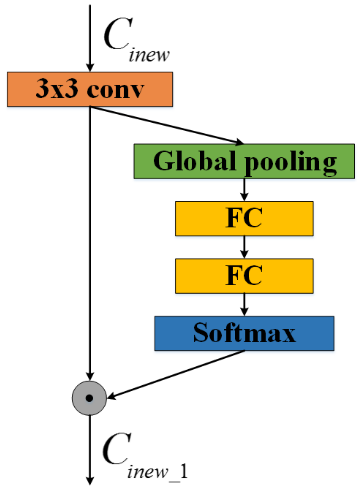


FIGURE 9. Channel-Wise Attention Module.

where F_{3c}, V_c are values of the cth channel in F_3 and V respectively. We select and visualize six weight values of V in Fig. 8(b), each grid of which gradually increases from left to right and is corresponding to the importance of channels in Fig. 8(c)-(h) respectively. Due to occlusion, the channels with features of the front foot and front body in Fig. 8(c)-(e) have lower weights. On the contrary, the channels extracting features of regions without occlusion, like the back foot and lower body in Fig. 8(f)-(h), get relatively higher weights.

In the second step, we use the learned weights V to rescale the original feature map F_1. In this way, features in F_1 have different importance.

$$C_{inew\_1} = V \odot F_1 \tag{8}$$

where $\odot$ represents element-wise multiplication. The parameters of operations in CWAM are shown in Table 1.

TABLE 1. Parameters of each part in the CWAM. The height, width, and channel are h, w, c_1 respectively. $r = c_1/16$.

Operation	Kernel size	Stride	Filter number	Output size
D	3	1	c_1	$h\times w\times c_1$
g	-	-	-	$1\times 1\times c_1$
W_1	-	-	r	$1\times 1\times r$
W_2	-	-	c_1	$1\times 1\times c_1$
softmax	-	-	-	$1\times 1\times c_1$

2) GLOBAL ATTENTION MODULE

As we know, small-scale pedestrians have fewer pixels than large-scale pedestrians. Compared with pedestrians without occlusion, occluded pedestrians lose some information of body regions. Thus, both small-scale and occluded targets are lack of information, which makes it difficult to recognize them. Global and environmental information is needed to help extract information for them. In our model, we attempt to adopt a global attention module (GAM) to acquire environmental features and enhance inter-spatial relationship for small-scale and occluded targets (seen in Fig. 10). The process is illustrated as follows.

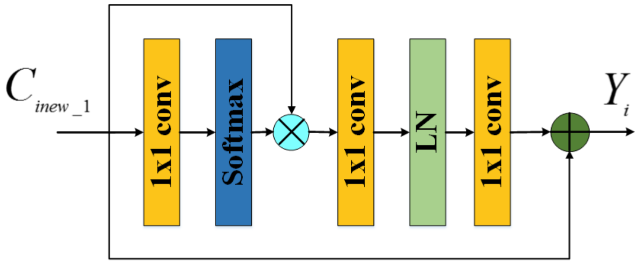


FIGURE 10. Global attention module.

First, we feed the feature map $C_{inew\_1} \in \mathbb{R}^{h\times w\times c_1}$ into a convolutional layer W_3 to gain the spatial feature $E_1 \in \mathbb{R}^{h\times w\times 1}$. Then, we reshape E_1 to the size of $1 \times 1 \times N$, $N = h \times w$, and a softmax layer is applied to it, which can convert the values to weights between 0 and 1.

$$E_1 = W_3(C_{inew\_1}) \tag{9}$$

$$E_{2j} = \frac{\exp(E_{1j})}{\sum_{k=1}^{N} \exp(E_{1k})} \tag{10}$$

where E_{1j} and E_{2j} are values in the jth channel of E_1 and E_2 respectively. After that, C_{inew_1} is reshaped to $N \times c_1$ and multiplied with E_2 by matrix multiplication to weight each spatial pixel in the feature map C_{inew_1}.

$$E_3 = E_2 \times C_{inew\_1} \tag{11}$$

Next, E_3 passes through convolutional layers, a layer normalization, and a ReLU to transform features. And we get the weight vector E_4.

$$E_4 = W_5(\delta(LN(W_4(E_3)))) \tag{12}$$

where W_4 represents a channel-downscaling convolutional layer to reduce the number of channels by 16. W_5 is a channel-upscaling convolutional layer to convert the number of channels to the original size. LN is layer normalization. δ denotes ReLU. Finally, E_4 is combined with the feature map C_{inew_1} by element-wise summation. In this way, C_{inew_1} is reinforced by contextual information E_4.

$$Y_i = E_4 + C_{inew\_1} \tag{13}$$

The parameters of operations in GAM are shown in Table 2.

III. EXPERIMENTS AND DISCUSSIONS

A. DATASET AND EVALUATION METRICS

CUHK dataset [28] is from Chinese University of Hong Kong and can be used for pedestrian detection. It contains 9 sets

TABLE 2. Parameters of each part in the GAM. The height, width, and channel are h, w, c_1 respectively. $r = c_1/16$, $N = h \times w$.

Operation	Kernel size	Stride	Filter number	Output size
W_3	1	1	1	$h \times w \times 1$
softmax	-	-	-	$1 \times 1 \times N$
W_4	1	1	r	$1 \times 1 \times r$
W_5	1	1	c_1	$1 \times 1 \times c_1$

and 1063 images. We divide the dataset into training set and test set. The training set contains the first 7 sets which have 807 images and the rest 2 sets consisting of 256 images constitute the test set. The target size of the pedestrians varies greatly, and the occlusion problem is severe.

CityPersons dataset [29] is a widely used benchmark in the task of pedestrian detection. It was built on the Cityscapes dataset [30], which was collected from multiple cities and countries across Europe. In the CityPersons dataset, occlusion between pedestrians is very serious, which causes difficulties for detection and makes it an ideal dataset for evaluating the performance of methods. Besides, the CityPersons dataset has a large variation on pedestrians' scales and includes many small pedestrians. We only use the original training and validation datasets that consist of 2975 and 500 images respectively.

We use the metric log-average Miss Rate (denoted as MR^{-2}) [31] in all of our experiments. MR^{-2} is computed by averaging miss rates at 9 false positives per image (FPPI) points evenly spaced in the range of $[10^{-2}, 10^{0}]$ in log space. And the lower the metric is, the better the algorithm is. We also use mean Average Precision (mAP) that is an actual metric to evaluate detection. mAP is an auxiliary metric in our experiments. And the higher mAP is, the better the algorithm is. For CityPersons dataset, we evaluate our methods on the Heavy subset like [31]. The Heavy subset is widely used in pedestrian detection to evaluate the effectiveness and robustness of detectors. On the Heavy subset, pedestrians used for evaluation are at least 50 pixels tall and have the occlusion range of 35-80 percent. For the CUHK dataset, due to a lack of visual annotation information of pedestrians, we evaluate our method on all the test sets.

B. IMPLEMENTATION DETAILS

We adopt FPN structure with ResNet50 as baseline and fine-tune the model ResNet50 pre-trained on the ImageNet dataset [32]. We apply stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0005.

During training, the ratio of foreground and background is 1:3. The RoIs which have IoU (Intersection over Union) overlap with ground truth bounding box at least 0.5 are regarded as positive proposals. The remaining RoIs are negative proposals. We use RoIAlign [33] to remove the harsh quantization of RoIPool [11] in our experiments. For CUHK dataset, the short side of input images is 600 pixels and the large side is 1000 pixels at most. The anchor scales are $\{64^2, 128^2\}$. For CityPersons dataset, the short side of input images is 800 pixels and the large side is 1600 pixels at most. We adopt 2 anchor scales of $\{16^2, 32^2\}$ on CityPersons dataset. On both datasets, we choose 3 anchor ratios of $\{0.33, 0.4, 0.5\}$. In addition to flipping images randomly, no other dataset augmentation is used on the input. In the test stage, we adopt non-maximum suppression (NMS) [34] to remove redundant boxes with IoU threshold 0.5. We train the network with learning rate 0.001 on CUHK dataset and 0.0005 on CityPersons dataset for 10 epochs, and then the learning rate is reduced by 10 times for another 5 epochs.

C. RESULTS AND ANALYSIS

In this section, we carry out ablation experiments on CityPersons dataset to show the effectiveness of our structure. Besides, we compare our proposed method with state-of-the-art methods.

1) EFFECTIVENESS OF DENSE CONNECTED BLOCK

In this part, we will analyze the influence of DCB. We set up a series of experiments to find the best DCB arrangement for pedestrian detection. In these experiments, we add DCB after $\{C_2\}$, $\{C_2, C_3\}$, $\{C_2, C_3, C_4\}$, $\{C_2, C_3, C_4, C_5\}$ respectively. For the low-level feature map C_2, due to few convolutional layers it passes through, we increase 2 DCBs after it. For high-level feature maps C_3, C_4, C_5, we increase one DCB after them.

We compare the results of different settings on CityPersons dataset. As the results in Table 3 show, original FPN with only $\{P_2, P_3\}$ for prediction and without DCB gets MR^{-2} 57.7% on Heavy subset. It reaches the best result of 54.5% when we add DCB after $\{C_2, C_3, C_4, C_5\}$. The experiment results can show that adding DCB after $\{C_2, C_3, C_4, C_5\}$ is very useful for extracting rich semantic information for classification and location. To verify our results, we visualize the results with and without DCB after $\{C_2, C_3, C_4, C_5\}$ in Fig. 11. We can see that our network with DCB in Fig. 11(c)(d) can detect pedestrians more accurately than that without DCB in Fig. 11(a)(b). Thus, we choose this DCB arrangement in our network as shown in Fig. 3 in the following experiments.

2) INFLUENCE OF THE COMBINATION OF CWAM AND GAM

After adding DCB for richer semantic information, we add a combination of CWAM and GAM after $\{C_{2new}, C_{3new}, C_{4new}, C_{5new}\}$ for getting channel-wise and environmental information. In this part, we will analyze the influence of this combination on the Heavy subset of CityPersons dataset. From Table 4, we can see that if we add attention modules after $\{C_{2new}, C_{3new}, C_{4new}, C_{5new}\}$ all, MR^{-2} comes to the top 51.6%, 2.9% better than the result (54.5%) without the combination. Therefore, our attention modules with CWAM and GAM plays a vital role in pedestrian detection, which is effective to extract enough contextual information and solve the occluded problem at the same time.

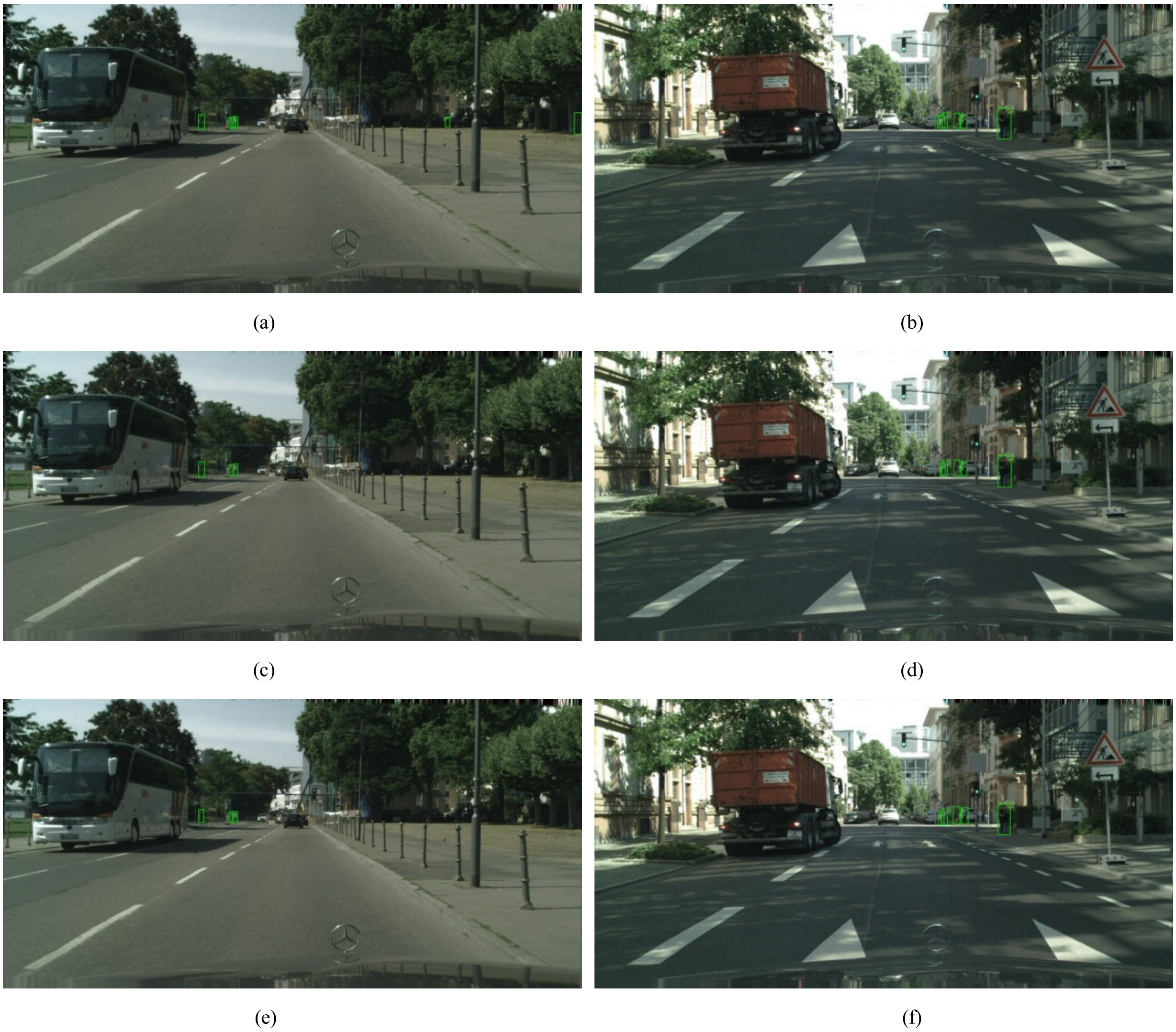


FIGURE 11. Comparisons between our network with DCB and without DCB on CityPersons dataset. The (a)(b) are the detection results of FPN with 2 prediction layers without DCB. The (c)(d) are the detection results of FPN with DCB. The (e)(f) are the ground truths.

TABLE 3. Comparisons of DCB added after different feature maps on Heavy subset of CityPersons dataset. The bold number indicates the best result. √ represents that DCB is added after that layer.

Dense Connected Block				CityPersons dataset
C_2	C_3	C_4	C_5	MR^{-2}
				57.7%
√				55.0%
√	√			55.4%
√	√	√		54.7%
√	√	√	√	**54.5%**

3) INFLUENCE OF DIFFERENT PREDICTION LAYERS

In this section, we explore the effectiveness of different prediction layers on the Heavy subset of CityPersons dataset. The results are shown in Table 5. In the second row, we only use $\{P_2\}$ for prediction, the MR^{-2} is very high and the detection result is especially bad. This is because the $\{P_2\}$ does not have deep features and its scale is single. This cannot cope with the variation of target scales. We get the best MR^{-2} 51.6% with prediction layers $\{P_2, P_3\}$. The results with prediction layers $\{P_2, P_3, P_4\}$ and $\{P_2, P_3, P_4, P_5\}$ are 53.5%

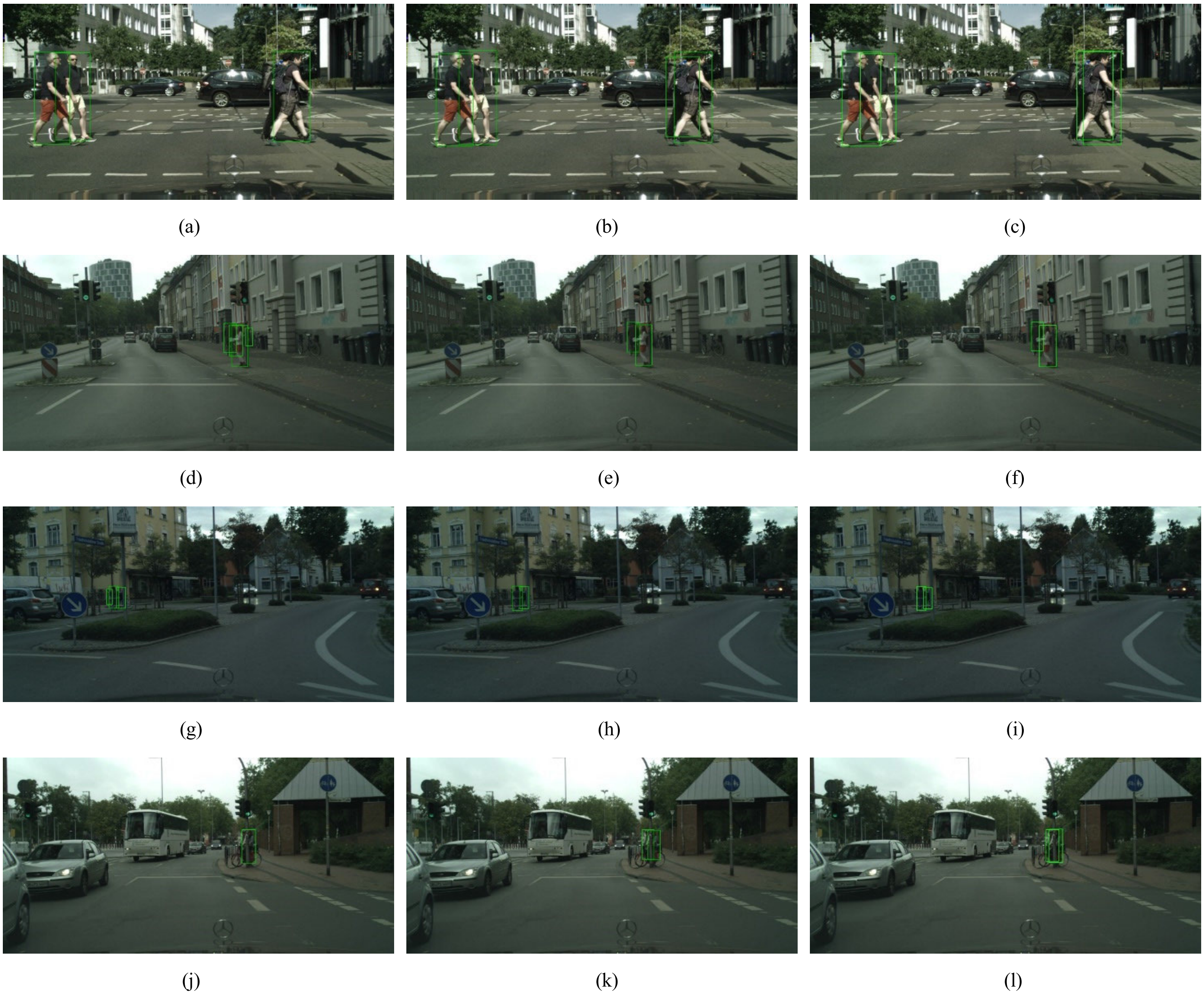


FIGURE 12. Visualization on CityPersons dataset. The images of each row from left to right are the detection results of FPN with 2 prediction layers, the detection results of our DA-Net, and the ground truths.

TABLE 4. Comparisons of the combination of CWAM and GAM on different feature maps on Heavy subset of CityPersons dataset. The bold number indicates the best result. √ represents that attention modules are added after that layer.

Attention modules				CityPersons dataset
C_{2new}	C_{3new}	C_{4new}	C_{5new}	MR^{-2}
				54.5%
√				53.1%
√	√			54.0%
√	√	√		54.6%
√	√	√	√	**51.6%**

and 54.3% respectively, a bit worse than the result of $\{P_2, P_3\}$. As we can see from Fig. 2, there are lots of small-scale targets on the CityPersons dataset. The high-level features are good for the detection of large-scale targets but may lead to some redundancy for small-scale targets.

4) COMPARISONS WITH STATE-OF-THE-ART METHODS

We compare our method with some state-of-the-art methods on the validation dataset of CityPersons dataset in Table 6. Our method gets MR^{-2} 51.6%, 13.2% lower than Faster RCNN [6] on the Heavy subset. Our method

TABLE 5. Comparisons of different prediction layers on Heavy subset of CityPersons dataset. The bold number indicates the best result.

Prediction layers	MR^{-2}
$\{P_2\}$	58.6%
$\{P_2, P_3\}$	**51.6%**
$\{P_2, P_3, P_4\}$	53.5%
$\{P_2, P_3, P_4, P_5\}$	54.3%

overpasses Faster RCNN+ATT [19] with self attention, visible-box attention, and part attention (6.7%, 5.7%, 5.1% lower respectively) on Heavy subset. This can explain that our attention modules with both CWAM and GAM is more effective. Besides, compared with RepLoss [35], we also get better MR^{-2} on Heavy subset, 5.9%, 7.5%, 5.3% lower than RepLoss with RepGT Loss, RepBox Loss, and RepGT+RepBox Loss respectively. In addition, we use the one-stage algorithm YOLOv3 [36] in the CityPersons dataset and compare it with our DA-Net. YOLOv3 is similar to FPN, which has several prediction layers with different resolutions. Although YOLOv3 is faster than ours in CityPersons dataset, MR^{-2} of ours is 6.4% lower than that of YOLOv3. We also compare our algorithm with the recent state-of-the-art methods like TLL [37], R2NMS [38], ALFNet [39], CSANet [40], CSP [16]. Our method outperforms TLL, R2NMS, TLL+MRF, ALFNet by 2%, 1.7%, 0.4%, 0.3% in MR^{-2}. Besides, our method runs faster than ALFNet when testing. Although CSANet and CSP have slightly better results on Heavy subset than ours, the testing speed of our method is twice faster than them. In Table 6, we can see that our method gets good results in both MR^{-2} and test time, which shows our structure is beneficial for pedestrian detection.

5) EVALUATION ON CUHK DATASET

To verify the generalization capacity of our proposed method, we train and evaluate our DA-Net on CUHK dataset. From Table 7, DA-Net on CUHK dataset gets MR^{-2} 37.2% (9.6% lower than FPN with two prediction layers), mAP 81.9%(4.0% higher than FPN with two prediction layers). We also show the results without DCB and the combination of CWAM and GAM to prove their effectiveness. When we add DCB on the original FPN, MR^{-2} is 4.3% lower (42.5% vs 46.8%) and mAP is 3.0% higher (80.9% vs 77.9%). Comparing DA-Net with the network without the combination, we find that our attention modules can improve the MR^{-2} by 5.3% and mAP by 1.0%. The results show that our DCB and the combination of CWAM and GAM are both valid for pedestrian detection.

TABLE 6. Comparisons with state-of-the-art methods on Heavy subset of CityPersons dataset. We bold the Top-2 results. Our method has good results on Heavy subset in both MR^{-2} and speed. FPN* represents FPN structure with prediction layers $\{P_2, P_3\}$.

Detector		MR^{-2}	Test Time
Faster RCNN[6]	-	64.8%	-
YOLOv3[36]	-	58.0%	**0.095s**
FPN* [12]	-	57.7%	-
	self attention	58.3%	-
Faster RCNN +ATT[19]	visible-box attention	57.3%	-
	part detection	56.7%	-
	RepGT Loss	57.5%	-
RepLoss[35]	RepBox Loss	59.1%	-
	RepGT + RepBox Loss	56.9%	-
OR-CNN[41]	-	55.7%	-
TLL[37]	-	53.6%	-
R2NMS[38]	-	53.3%	-
TLL+MRF[37]	-	52.0%	-
ALFNet[39]	-	51.9%	0.270s
CSANet[40]	-	**51.3%**	0.320s
CSP[16]	-	**49.9%**	0.330s
ours	**-**	51.6%	**0.150s**

TABLE 7. Results on CUHK dataset. FPN* represents FPN structure with prediction layers $\{P_2, P_3\}$. *Combination* represents the combination of CWAM and GAM. The bold number indicates the best result.

Setting	CUHK dataset	
	MR^{-2}	mAP
FPN*	46.8%	77.9%
FPN* + DCB	42.5%	80.9%
FPN* + DCB+ Combination	**37.2%**	**81.9%**

6) RESULTS VISUALIZATION ON CHALLENGING SCENARIOS

We visualize the detection results generated by our method and FPN structure with prediction layers $\{P_2, P_3\}$. We choose pictures in challenging scenarios with various-scale and occluded pedestrians. We show the detection results on CityPersons dataset in Fig. 12. From the visualization results, we know that our proposed model is more accurate and robust in detecting various-scale and occluded pedestrians.

IV. CONCLUSION

In this article, we add Dense Connected Block, a combination of channel-wise attention module and global attention module after the original FPN structure. We only use the low-level layers of original FPN for prediction to deal with the detection with lots of small-scale and various-scale targets. We apply Dense Connected Block to enrich semantic information. Channel-wise attention module can weight each channel with different importance, which can guide the network to emphasize visible parts of pedestrians to solve the occlusion problem. Global attention module is used to acquire long-range dependencies and environmental information for small-scale and occluded targets. The combination of those two modules can keep our network having those two advantages just mentioned at the same time. We conduct a series of ablation experiments on CityPersons dataset to verify their effectiveness. Compared with state-of-the-art methods, our proposed method overpasses most of these methods in both detection accuracy and speed. Our method gets MR^{-2} 51.6% on the Heavy subset of CityPersons dataset and MR^{-2} 37.2% on CUHK dataset, which shows its effectiveness and robustness for occluded and various-scale pedestrians.

REFERENCES

[1] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

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

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

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

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

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

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

[10] L. Zheng, C. Fu, and Y. Zhao, "Extend the shallow part of single shot multibox detector via convolutional neural network," in *Proc. 10th Int. Conf. Digit. Image Process. (ICDIP)*, Aug. 2018, Art. no. 1080613.

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

[12] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

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

[14] Q. Hu, P. Wang, C. Shen, A. van den Hengel, and F. Porikli, "Pushing the limits of deep cnns for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1358–1368, Jan. 2017.

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

[16] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

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

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

[20] Y. Zhang, J. Liu, and K. Huang, "Dilated hourglass networks for human pose estimation," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 483–499.

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

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[23] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

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

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

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[28] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3258–3265.

[29] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3213–3221.

[30] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[31] 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.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[34] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 850–855.

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

[36] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[37] 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. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 536–551.

[38] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10750–10759.

[39] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 618–634.

[40] Y. Zhang, P. Yi, D. Zhou, X. Yang, D. Yang, Q. Zhang, and X. Wei, "CSANet: Channel and spatial mixed attention CNN for pedestrian detection," *IEEE Access*, vol. 8, pp. 76243–76252, 2020.

[41] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 637–653.

RUIHONG YIN received the B.S. degree from the School of Electronic and Information Engineering, Beihang University, China, where she is currently pursuing the master's degree. Her research interest includes computer vision.

RUFEI ZHANG was born in 1981. He received the Ph.D. degree from Northwestern Polytechnical University, China. He is currently a Senior Engineer. His research interests include guidance, navigation, and control.

WEI ZHAO received the B.S., M.S., and Ph.D. degrees from the School of Automatic Control, Northwestern Polytechnical University, Xi'an, China. She did postdoctoral research at Beihang University, where she is currently a Full Professor. Her main research interests include digital image processing, automatic target recognition, signal processing in wireless sensor networks, and information fusion.

FENG JIANG was born in 1995. He received the M.S. degree from the Beijing Institute of Technology, China. He is currently an Engineer. His research interests include intelligent perception and collaborative control.