IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 21, 2020, accepted April 2, 2020, date of publication April 8, 2020, date of current version May 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986476

CSANet: Channel and Spatial Mixed Attention CNN for Pedestrian Detection

YUNBO ZHANG[1], PENGFEI YI[1], (Member, IEEE), DONGSHENG ZHOU[1,2], (Member, IEEE), XIN YANG[2], (Member, IEEE), DEYUN YANG[3], QIANG ZHANG[1,2], (Member, IEEE), AND XIAOPENG WEI[2], (Member, IEEE)

[1]Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education), School of Soft Engineering, Dalian University, Dalian 116622, China
[2]School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[3]School of Information Science and Technology, Taishan University, Tai'an 271000, China

Corresponding authors: Pengfei Yi (yipengfei@dlu.edu.cn) and Dongsheng Zhou (zhoudongsheng@dlu.edu.cn)

This work was supported in part by the State Key Program of National Natural Science Foundation of China under Grant U1908214, in part by the Program for the Liaoning Distinguished Professor, Program for Dalian High-level Talent Innovation Support, under Grant 2017RD11, and in part by the Science and Technology Innovation Fund of Dalian under Grant 2018J12GX036.

ABSTRACT Current mainstream pedestrian detectors tend to profit directly from convolutional neural networks (CNNs) that are designed for image classification. While requiring a large downsampling factor to produce high-level semantic features, CNNs cannot adaptively focus on the useful channels and regions of the feature maps, which limits the accuracy of pedestrian detection. To obtain a higher accuracy, we propose a single-stage pedestrian detector with channel and spatial attentions (CSANet), which can locate useful channels and regions automatically while extracting features. The backbone of CSANet is different from that of mainstream pedestrian detectors, which can effectively highlight the pedestrian-likely regions and suppress the background. Specifically, we model contextual dependencies from channel and spatial dimensions of the feature maps, respectively. The channel attention module can selectively promote CNNs to focus on key channels by integrating associated features. Meantime, the spatial attention module can illuminate semantic pixels by aggregating similar features of all channels. Eventually, the two modules are connected in series to further enhance the representation of feature maps. Experiment results show that CSANet achieves the state-of-the-art performance with MR^{-2} of 3.55% on Caltech dataset and obtains competitive performance on CityPersons dataset while maintaining a high computational efficiency.

INDEX TERMS Convolutional neural network, dual attention network, pedestrian detection.

I. INTRODUCTION

Pedestrian detection plays a critical role in computer vision tasks such as autonomous driving, robotics, and surveillance. In recent years, pedestrian detectors have made considerable progress with the revival of deep learning [1], [2]. However, current state-of-the-art pedestrian detectors still fall far short from the cognitive levels as fast and accurate as human [3].

Pedestrian detection can be traced back to traditional methods with low-level features, e.g. HOG [4]. The emergence of R-CNN [5] made the two-stage architectures of ''Region Proposal+CNN'' into an established method in object detection [6], [7]. Then Zhang *et al.* [8] discussed the effectiveness of Faster R-CNN framework in pedestrian detection tasks. The performance of small-scale pedestrian detection was optimized by improving the resolution of feature maps. Unfortunately, the lack of diverse pedestrian datasets limits the ability of Faster R-CNN to generalize pedestrian detection in various scenes. Zhang *et al.* [9] proposed a pedestrian benchmark called CityPersons, which contains rich and diverse images. They reported that AdaptFasterR-CNN [9], a model trained using the CityPersons, has stronger generalization capabilities. However, in the two-stage frameworks mentioned above, the cumbersome anchor boxes design and the generation of region proposals make it difficult to perform real-time detection.

In contrast to the two-stage detectors, single-stage detectors, e.g. YOLO [10], [11], are well known for their fast detection and considerable accuracy. Therefore, Noh *et al.* [12] proposed a single-stage detector to optimize real-time

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

indicators of pedestrian detection. Currently, researchers are keen on the design of simpler anchor-free detectors. In particular, H. Law *et al.* proposed an anchor-free object detector that detected the bounding boxes as a pair of keypoints [13], which led to a boom of anchor-free detectors. Subsequently, considerable anchor-free object detectors were designed [14]–[17]. Furthermore, researchers began to design anchor-free pedestrian detectors, such as [18], [19].

Notably, the pedestrians in traffic scenes have characteristics that are different from general objects, such as dynamic background and variable scales [8], [20]. Generally, researchers employ deep models to abstract higher-level semantics of object instances, which is helpful for identifying pedestrians in traffic scenarios. Unfortunately, this approach filters out a lot of small-scale pedestrians as well as the position information of large-scale pedestrians. Due to the inherent characteristics of CNNs, the key channels cannot be highlighted and the key spatial position cannot be illuminated. Convolution is a local operation, which obtains local information of an image by applying a convolution kernel to the local image. The local operation of CNNs results in its inability to capture images from a global view. Therefore, it is still a challenging task to design an efficient backbone for pedestrian detection.

In this paper, we designed a dual attention network to extenuate the limitations mentioned above for pedestrian detection. In detail, the dual attention network was constructed from two dimensions (channel and spatial) of feature maps. First, the global average pooling was used to aggregate the global information of feature channels. The original feature map was squeezed along the spatial axis to generate a 1D channel attention map. The channel attention map can model the correlations between channels. Then, the global pooling was used to obtain a 2D spatial attention map by compressing the original feature maps along the channel axis direction. Furthermore, the two modules constructed above were combined in proper order to refine the feature maps. In consideration of the superiority of the anchor-free detector, we designed an anchor-free pedestrian detector called CSANet based on the dual attention modules. In addition, the performance of CSANet on multi-scale pedestrians was enhanced by fusing multi-scale feature maps. In line with our expectations, CSANet achieved a significant performance on two pedestrian benchmarks, namely Caltech [21] and CityPersons [9].

Overall, our contributions were as follows:

1. We proposed a lightweight dual attention network, which not only models the relationship of each feature channel, but also improves the expression ability of feature maps at the pixel level.

2. We constructed a single-stage pedestrian detector based on the dual attention network and further analyzed the impact of several key components in CSANet on its performance.

3. CSANet achieved state-of-the-art performance on Caltech benchmark and competitive performance on CityPersons benchmark while maintaining computation efficiency.

II. RELATED WORKS

In this section, after reviewing cutting-edge technologies related to pedestrian detection, we finally found an approach to construct a lightweight, concise and effective pedestrian detector.

A. ANCHOR-BASED AND ANCHOR-FREE DETECTORS

Pedestrian detectors fall into two categories, the anchor-based and the anchor-free detectors. The former is dedicated to the improvement of accuracy, such as Faster R-CNN [7] and Mask R-CNN [22]. The latter focuses on the improvement of speed, such as YOLO [10], [11] and SSD [23]. Inspired by the methods above, the pedestrian detection tasks have made great progress [8], [24], [25]. Recently, CornerNet [13] has proposed an anchor-free method for detector based on keypoints detection. In some studies [18], [19], the idea of being anchor-free was applied to pedestrian detection, which fully explained the broad validity of the idea and opened up the perspective of pedestrian detection.

Our works fall into the category of anchor-free pedestrian detection based on keypoints detection, which has advantages in terms of speed and accuracy. In fact, keypoints belong to high-level semantic features. Therefore, anchor-free detectors rely strongly on the representation of feature maps. As we all know, some baseline methods often use deeper layer and larger downsampling factor to extract feature maps with greater abstraction. Different from the above methods, we used attention mechanism to enhance the expression ability of feature maps, which can increase the size of receptive fields. In addition, we set a small downsampling factor to maintain a high resolution of the feature maps.

B. ATTENTION MECHANISM IN IMAGE PROCESSING

Human visual attention mechanism inspires the development of attention mechanism in computer vision. Nowadays, the idea of attention was introduced into many computer vision tasks by researchers, such as image classification [27], [28], medical image segmentation [29], [30], image captioning [31], scene segmentation [32], remote sensing imagery analysis [33], etc.

Besides these outstanding works, there are more works on visual attention. Wang *et al.* [34] proposed a non-local neural network for capturing long-range dependencies. SENet [28] modeled the correlations among channels. Inspired by SENet and Iception [35], SKNet [36] made an improvement, which combined the channel attention module of SENet and the multi-branch convolutional layer of Iception. In addition, the spatial attention model is famous for STN [37] proposed by Google DeepMind, which can make up for the limitations of local convolution operation by aggregating the context information of the feature maps.

Chen *et al.* [38] integrated spatial, channel-wise and multi-layer visual attention in CNN for image captioning, and proposed two attention modules, Channel-Spatial and Spatial-Channel. Inspired by SCA-CNN [38], the network

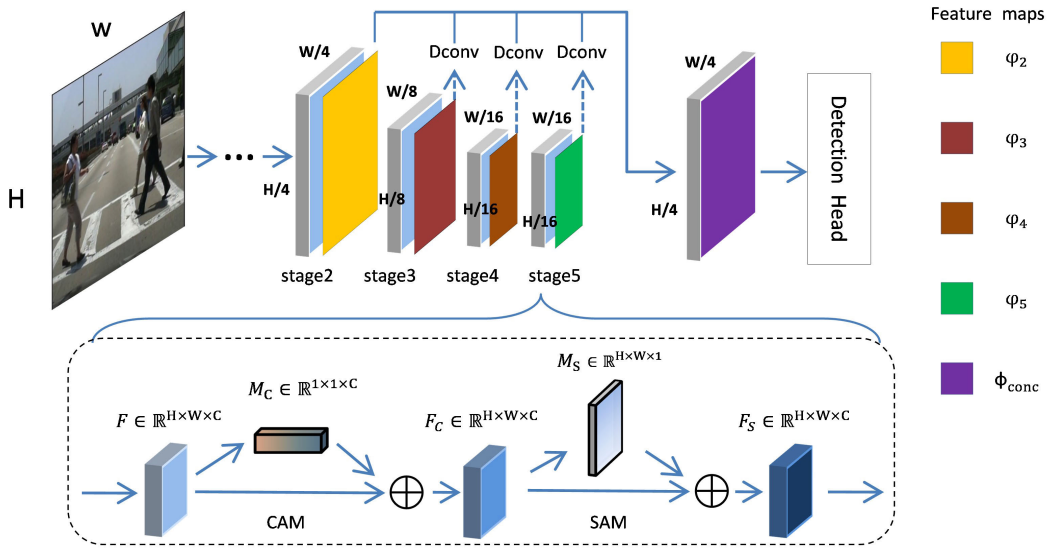


FIGURE 1. Overall architecture of CSANet. It mainly includes two components, the backbone and detection head. In dashed box, the channel attention model (CAM) and spatial attention model (SAM) are combined successively, which are embedded in the stage 5 of Resnet-50.

CBAM [39] integrated spatial and channel attention modules and achieved better results in image classification and object detection. In 2018, Guo *et al*. [40] started person re-identification task based on spatial and channel dual attention networks, which made great progress in person re-realization.

Inspired by [40], we carried out a pedestrian detection task in traffic scenes, which was similar to person re-realization. Our works fall into the design of a dual attention network for pedestrian detection, but they are different from all above attention networks. Firstly, we modeled the attention mechanism from channel and spatial dimensions of feature maps, which considered unequal roles of channel and spatial dimensions. In addition, inspired by GCNet [41], we used the addition operation to broadcast attention maps. In a word, our dual attention network can effectively improve the semantic abstraction of CNNs.

III. PROPOSED METHOD

In this section, we will firstly show the overall framework of CSANet and then introduce the mathematical modeling of channel attention module (CAM) and spatial attention module (SAM) separately. Finally, we will introduce the arrangement of the two attention modules.

A. OVERALL ARCHITECTURE

The overall framework of CSANet is shown in Fig. 1. The backbone network is ResNet-50 [42] with dual attention network embedded. Similarly to [18], the detection head module mainly includes three 1×1 convolutional layers, which predict the center position, scale, and offset, respectively. The ResNet-50 is divided into 5 stages. We define the output feature maps of 2 to 5 stages as $\varphi_2, \varphi_3, \varphi_4$ and φ_5, respectively. Following the practice [43], the input image is downsampled by 4, 8, 16, and 16, respectively. Among them, the low-level feature map can provide more accurate information on position, and the deeper feature map contains more semantic information. We merge the multi-scale feature maps of each stage in a simple way to get the final feature map ϕ_{conc}. Similarly to [18], [19] and [20], the resolution of the output feature maps at each stage is unified using the deconvolution operation before concatenation. The number of filters in three deconvolution layers is the same as that of the last convolution layer in stage 2. The number of filters in the deconvolution layer can be flexibly set. Usually, shallow features are known as universal, while the semantic information expressed by each channel of deep features is more category-specific. We compare the effectiveness of different types of dual attention network in ablation studies.

Taking the third residual block of stage 5 as an example, the dual attention network can be constructed as follows. Given the output of the residual block, we define it as the original feature map $F \in \mathbb{R}^{H\times W\times C}$. We feed it into the dual attention mechanism. CAM and SAM in turn derives a 1D channel attention map $M_C \in \mathbb{R}^{1\times 1\times C}$ and 2D spatial attention map $M_S \in \mathbb{R}^{H\times W\times 1}$. The original feature map F is sequentially refined by two attention maps. The calculation of two refined feature maps can be summarized as follows:

$$F_C = M_C(F) \oplus F, \tag{1}$$

$$F_S = M_S(F_C) \oplus F_C, \tag{2}$$

where $\oplus$ indicates element-wise addition, M_C is the channel attention map, M_S is the spatial attention map, F_C is the feature map refined by M_C, and F_S is the feature map refined by M_S.

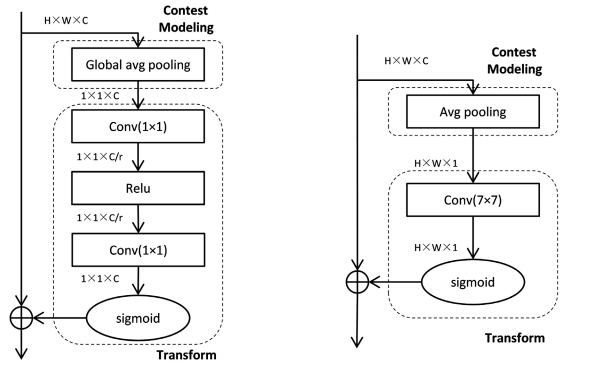
In what follows, we describe the modeling process of two modules in detail.

B. CHANNEL ATTENTION MODULE

In fact, each channel of high-level features can be regarded as a class-specific response and these channels have different responsiveness [28]. However, CNNs treat all channels equally. Therefore, we build a channel attention module to explicitly model interdependencies between channels. The attention model can emphasize interdependent feature maps and improve the feature representation of specific semantics.

In [30], [36], [38], [39], the channel attention module gives different weights to each channel to stress useful channels for image classification. However, such a channel relationship modeling method is not suitable for binary classification of pedestrian detection. The weighted operation mode excessively suppresses unimportant channel information and reduces the diversity of feature maps. In addition, while enhancing the information of a certain channel, the interference of noise in a complex background is also enhanced. We use addition operation to broadcast the channel map to the original feature map.

As shown in Fig. 2 (a), in order to calculate an effective channel attention map, the 2D feature map was compressed into a real number along the spatial axis of input.



(a) Channel Attention Module (CAM) (b) Spatial Attention Module (SAM)

FIGURE 2. Architecture of the two attention blocks. The feature maps are shown as feature dimensions, e.g. H×W×C denotes a feature map with height H, weight W and channel number C, and ⊕ denotes broadcasting of element-wise addition.

First, for the original feature map $F \in \mathbb{R}^{H\times W\times C}$ with channel number C, the global average pooling is used to aggregate the global information of each channel to obtain the channel attention map $U \in \mathbb{R}^{1\times 1\times C}$. The calculation process is:

$$u_c = GAP(f_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} f_c(i,j), \tag{3}$$

where GAP represents global average pooling operation, f_c represents c-th channel of size $H \times W$ in the original feature map F, and u_c is c-th real number of the channel attention map $U \in \mathbb{R}^{1\times 1\times C}$.

Then, we feed the channel feature map U into two fully connected layers to obtain a non-linear descriptor, and then the descriptor passes through the function of sigmoid to obtain the final channel attention map $M_C \in \mathbb{R}^{1\times 1\times C}$. The process is expressed as:

$$M_C = \sigma(W_2\delta(W_1 U)), \tag{4}$$

where the two fully connected layers are used to better fit the complex correlations between channels, δ represents the ReLU activation function, σ denotes the sigmoid function, W represents the scaling parameters in the fully connected layer, including $W_1 \in \mathbb{R}^{C/r\times C}$ and $W_2 \in \mathbb{R}^{C\times C/r}$, r is the reduction ratio, and r is set as 16.

Finally, we re-characterize the input feature map F with M_C. In the first step,M_C is broadcast into the same dimensions as $F \in \mathbb{R}^{H\times W\times C}$ and we donate it as $M'_C \in \mathbb{R}^{H\times W\times C}$. Here we use pixel-by-pixel addition operation to broadcast the channel attention map. The entire calculation process is expressed as:

$$f'_c = F_{add}(m'_c, f_c) = m'_c + f_c. \tag{5}$$

where F_{add} represents channel-wise addition, m'_c is the c-th channel of the broadcast attention map M'_C, and f_c is the c-th channel of the original feature map F, and f'_c is c-th channel of the refined feature map $F_C \in \mathbb{R}^{H\times W\times C}$. To some extent, m'_c represents the global information of the c-th channel of M'_C. The attention map M'_C integrates the representation capabilities of global semantic information.

C. SPATIAL ATTENTION MODULE

In pedestrian detection, the detailed position of the pedestrian plays a critical role, which is beneficial for locating the position of pedestrian. The discriminant feature representations are critical for pedestrian detection, which could be obtained by capturing long-range contextual dependencies between each pixel. In order to make up for the shortcomings of channel attention module, we further design the spatial attention module to refine the feature map from the pixel level to improve the feature representation of the feature map. Similar to the channel attention map, the spatial attention map is also broadcast to the original feature map by pixel-wise addition operation.

In order to calculate an effective spatial attention map, the 3D feature map is compressed into a 2D feature channel along the channel axis of the feature map F_C.

First, given the feature map $F_C \in \mathbb{R}^{H\times W\times C}$ after re-calibration, the average pooling operation is performed on all feature channels along the channel axis to obtain a feature map $V \in \mathbb{R}^{H\times W\times 1}$. The calculation process is:

$$v_{ij} = AP(f_{ij}) = \frac{1}{C}\sum_{k=1}^{C} f_{ij}(k), \tag{6}$$

where AP represents the average pooling operation, f_{ij} represents the pixel value of ij point in f'_c, C is the number of feature channels, and v_{ij} represents the pixel value of ij point in the feature map V.

Then, a convolution layer with step size of 1 and convolution kernel size of 7×7 is used to perform a convolution

operation on the feature map V. The activation function sigmoid is used in the process to obtain the attention map $M_S \in \mathbb{R}^{H\times W\times 1}$. The process is as follows:

$$M_S = \sigma(f^{7\times 7}(V)), \tag{7}$$

where σ denotes the sigmoid function and $f^{7\times 7}$ represents a convolution operation with the filter size of 7×7.

Finally, we re-characterize the input feature map F_C with the spatial attention map M_S. M_S is broadcast into the same dimensions as $F_C \in \mathbb{R}^{H\times W\times C}$ and we denote it as $M'_S \in \mathbb{R}^{H\times W\times C}$. As shown in formula (5), we use pixel-by-pixel addition operation. The whole calculation process is:

$$f'_s = F_{add}(m'_s, f_s) = m'_s + f_s. \tag{8}$$

where F_{add} represents spatial-wise addition, m'_s is the s-th channel of the broadcast attention map M'_S, f_s is the s-th channel of the refined feature map F_C by M_C, and f'_s is the s-th channel of the refined feature map $F_S \in \mathbb{R}^{H\times W\times C}$ by M_S.

Therefore, spatial attention map has a global contextual view and selectively aggregates contexts. In addition, the attention map focuses on the global information of each channel, increases the size of receptive fields, and enables CNNs to capture the image from a global perspective.

D. ARRANGEMENT OF TWO ATTENTION MODULES

In the works of [30], [38], [40], a multi-attention module organization method is proposed. Inspired by this, we connect CAM and CSM in series. The two attention modules can be embedded in ResNet-50 in a parallel or sequential manner. The channel attention module focuses on important channels, while the spatial attention module focuses on important regions of feature maps. Proper combination of two attention modules can maximize the effectiveness of the attention mechanism. Therefore, we discuss experimental results in ablation studies.

IV. EXPERIMENTS

In this section, we firstly introduce two pedestrian detection benchmarks, evaluation metric and implementation details separately. Then, we report the results of ablation studies on Caltech dataset. In addition, to fully prove the effectiveness of CSANet, we implement a visualization experiment. Finally, we show the results of comparisons among state-of-the-art pedestrian detectors on Caltech and CityPersons datasets.

A. EXPERIMENTS DETAILS

1) DATASETS

We evaluate the effectiveness of CSANet on two challenging benchmarks, Caltech [21] and CityPersons [9]. For the Caltech dataset, we follow the approach in [18], where the training data are augmented by extracting one of every 3 frames. There are 42,782 images with resolution 640×480 in the training set. The testing set has 4,024 official images. The evaluations are conducted based on the new annotations provided by [1].

The CityPersons dataset is derived from Cityscapes [44], and has pedestrian annotations with multiple occlusion levels. Experiments use a training set that contains 2,975 images and an official validation set that contains 500 images.

The evaluation metric follows the Caltech evaluation standard [21], which is log-average Miss Rate over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$ (expressed as MR^{-2}). Smaller value of MR^{-2} indicates better performance.

2) IMPLEMENTATION DETAILS

Our method is implemented in the Keras framework. The training and testing are performed on single NVIDIA GTX 1080Ti GPU. The backbone network is ResNet-50 pre-trained on ImageNet [45]. For the Caltech dataset, one mini-batch contains 16 images, and the learning rate is 10^{-4}, and the training is stopped after 120 epochs. For CityPersons dataset, we set a mini-batch that contains 2 images, and the learning rate is set to 2×10^{-4}, and the training is stopped after 150 epochs. Following [9], [18], for the Caltech dataset experiment, it also includes further optimization experiment using model initialized from CityPersons, which is trained with the learning rate 2×10^{-5}. The CityPersons dataset contains a large number of images under a variety of conditions.

B. ABLATION STUDY

In this subsection, we conduct an ablative analysis on the Caltech dataset to show the effectiveness of four main components of the proposed method. The ablation studies are mainly divided into four parts: 1) Which feature extraction method is more effective? 2) How important is the feature fusion? 3) Which is more efficient, addition or multiplication? 4) How to connect CAM with SAM?

1) WHICH FEATURE EXTRACTION METHOD IS MORE EFFECTIVE?

Our dual attention network can be easily embedded in each residual block of ResNet-50. Multi-layer visual attention in CNNs has proven to be more effective for image captioning [38]. In this part, we embed the dual attention network in multi-layer of the ResNet-50 to obtain the feature maps with different expressive abilities. We compare five methods of the extraction of feature maps in detail. The experiments are implemented under IoU =0.5 and IoU =0.75. The IoU reflects the overlap area between the prediction boxes and the ground truth boxes.

As shown in Table 1, the embedding method of stage 3-5 achieves the best performance with MR^{-2} of 3.88% under IoU = 0.5. Under the stricter IoU = 0.75, the embedding method of stage 2-5 improves the performance by about 36% compared with the embedding method of stage 5. Notably, under the IoU = 0.5, the models stage 2-4 and stage 5 have comparable performance with MR^{-2} of 4.28% and 4.27%, respectively. However, there is a large performance gap with ΔMR^{-2} of 4.77% between them under the threshold of IoU = 0.75. This comparison shows that dual attention

TABLE 1. Comparisons of the extraction methods of feature map.

Extraction methods	#Para (MB)	C	Test Time (ms/img)	MR^{-2} (%)	
				IoU=0.5	IoU=0.75
stage 2-5	41.6	2048	58.0	4.09	25.41
stage 2-4	40.0	2048	58.9	4.28	26.79
stage 3-5	41.5	2048	57.2	**3.88**	**20.05**
stage 4-5	41.4	2048	58.5	4.60	29.22
stage 5	40.6	2048	**52.2**	4.27	31.56

Comparisons of different extraction methods of the feature map, where stage 2-5 indicates the feature map extraction method, meaning that the dual attention network is embedded in the stage 2, 3, 4 and 5 of ResNet-50. The column of C shows the number of channel of the feature maps output by the last stage. #Para is the number of parameters of the model of CSANet. Test time is the runtime of one image with size of 640×480 pixels. Bold numbers indicate the best results.

network is more conducive to the detection of high-quality bounding boxes.

2) HOW IMPORTANT IS THE FEATURE FUSION?

In this part, we compare different combinations of multi-scale feature maps based on dual attention network under IoU =0.5 and IoU =0.75. Multi-scale feature fusion is essential for pedestrian detection [18], [20]. The shallower feature maps (e.g.φ_2) maintain a higher resolution, which is conducive to the detection of small-scale pedestrians. The deeper feature maps (e.g.φ_5) have a lower resolution and a larger receptive field, which is conducive to the detection of large-scale pedestrians. The proper feature fusion is beneficial to pedestrian detection of various scales.

It can be seen from Table 2 that the model combining low-level features such as φ_2 and φ_3 has a poor accuracy, but has smaller number of parameters and faster detection speed.

TABLE 2. Comparisons of different fusion methods.

Feature maps				C	#Para (MB)	Test Time (ms/img)	MR^{-2} (%)	
φ_2	φ_3	φ_4	φ_5		-		IoU=0.5	IoU=0.75
√	√			512	4.76	**33.9**	7.36	35.20
	√	√		1024	16.58	41.5	5.71	29.90
		√	√	2048	38.9	50.1	5.29	26.14
√	√	√		1024	17.2	44.2	5.43	29.57
	√	√	√	2048	41.5	57.2	**3.88**	**20.05**

Comparisons of different fusion strategies of the multi-scale feature maps. φ_2 , φ_3 , φ_4 and φ_5 represent the output of stage 2, 3, 4 and 5 of ResNet-50, respectively. The column of C shows the number of channel of the feature maps output by the last stage. #Para is the number of parameters of the model of CSANet. Test time is the runtime of one image with size of 640×480 pixels. Bold numbers indicate the best results.

As the number of fused feature maps increases, the detection accuracy of the model also increases. Under IoU = 0.5, the model with the fusion of φ_3, φ_4 and φ_5 has a significant improvement with MR^{-2} of 47% with the worst accuracy. Under IoU = 0.75, the fusion of φ_3, φ_4 and φ_5 has the best result. In general, we find that although deeper features are helpful for feature detection, they consume more running memory.

3) WHICH IS MORE EFFICIENT, ADDITION OR MULTIPLICATION?

Considering the real-time requirement of pedestrian detection, we use a different way of broadcasting attention maps from [38], [39]. The study of [41] modeled long-distance dependencies and used the addition operation to broadcast attention maps. In our proposed method, the pixel-by-pixel addition operation is used to broadcast attention maps, and the original feature maps are re-calibrated in turn by dual attention modules. In this set of experiments, we compare the effects of addition and multiplication under IoU =0.5.

It can be observed from Table 3 that the models with addition broadcasting is better than the models with multiplication broadcasting. The miss rate gap between "p3p4p5+add" and "p3p4p5+mul" is about MR^{-2} of 37%. In the second and third sets of experiments, the two gaps are about MR^{-2} of 17%. In addition, we find that the broadcast method of attention map hardly affects the test time. In fact, the runtime is mainly affected by the parameters of the model.

TABLE 3. Comparisons of different broadcasting operations.

Description	#Para (MB)	C	Test Time (ms/img)	MR^{-2} (%)	ΔMR^{-2} (%)
				IoU=0.5	
p3p4p5+add	41.5	2048	57.2	**3.88**	↓ 2.27
p3p4p5+mul	41.5	2048	58.0	6.15	
p3p4+add	16.58	1024	41.5	5.71	↓ 1.13
p3p4+mul	16.58	1024	42.2	6.84	
p2p3+add	4.76	512	**33.9**	7.36	↓ 1.56
p2p3+mul	4.76	512	34.8	8.92	

Comparisons of two broadcast operations: addition and multiplication. The description column summarizes the model. For example, "p3p4p5+add" means the $\phi_{conc} = \{\varphi_3, \varphi_4, \varphi_5\}$ and the addition broadcasting. The ΔMR^{-2} means the performance gap between them. The column of C shows the number of channels of the feature maps output by the last stage. #Para is the number of parameters of the model of CSANet. Test time is the runtime of one image with size of 640×480 pixels. Bold numbers indicate the best results.

The computational complexity of multiplication is much higher than that of addition [46]. While multiplication operation enhances the useful information expression of the feature maps, it also excessively enlarges the impact of noise. In addition, multiplication weighting operation excessively suppresses some contextual details, which is disadvantageous to locating pedestrians. In addition to accuracy, we also need consider the real-time indicators of the model. It should be

noted that the multiplication operation increases the runtime of networks.

4) HOW TO CONNECT CAM WITH SAM?

In this set of experiments, we compare three combinations of dual attention modules under IoU =0.5. There are three manners to connect the two attention modules, the CAM-first, the SAM-first and the parallel. These manners can effectively improve the accuracy of pedestrian detection, but there is a certain gap in effect.

As shown in Table 4, CAM+SAM represents that the channel attention module and the spatial attention module are connected in a sequential manner. We find that the sequential arrangement achieves a better result than a parallel arrangement. The manner of CAM-first has the best result with MR^{-2} of 3.88%. CAM//SAM represents that the two modules are arranged in parallel, which underperforms the CAM+SAM with MR^{-2} of 0.27%. In the third manner, SAM+CAM represents SAM-first in dual attention network, which has the worst performance with MR^{-2} of 4.57%. The results of Table 4 show that CAM-first has a better performance than others.

TABLE 4. Comparisons of different arrangement manners.

Description	#Para (MB)	C	Test Time (ms/img)	MR^{-2} (%)	ΔMR^{-2} (%)
				IoU=0.5	
CAM+SAM	41.5	2048	57.2	**3.88**	-
CAM//SAM	41.5	2048	56.4	4.15	↓ 0.27
SAM+CAM	41.5	2048	**52.9**	4.57	↓ 0.69

Comparisons of different arrangements of CAM and SAM. The column of channel shows the number of channels of the feature maps output by the last stage. #Para is the number of parameters of the model of CSANet. Test time is the runtime of one image with size of 640×480 pixels. The ΔMR^{-2} means the performance gap between them. Bold numbers indicate the best results.

C. NETWORK VISUALIZATION WITH GRAD-CAM

In this subsection, we apply Grad-CAM [47] to explain our model qualitatively. The interpretability of CNNs has improved to a certain level. The algorithm can derive class activation mapping, which can be used to locate the region of the class in the image. Grad-CAM mainly uses the gradient of the last convolution layer of networks to generate a heat map, and it can highlight important pixels in the input image.

We attempt to find how CSANet makes good use of features and enhances the expression ability of feature maps at the pixel level. In Fig. 3, we can clearly see that the masks of the model with dual attention network better cover the pedestrian areas than the model without dual attention network. In other words, the dual attention modules can better focus on the pixel information of the target areas. The visualization results qualitatively show that in the refined feature map, the pixel expression ability of the target areas is enhanced to a certain extent.

FIGURE 3. The visualization results on the test set of the Caltech dataset. We compare the visualization results with and without the use of dual attention network. The input images are from the test set of Caltech dataset. Pedestrians in the input image are marked with red circles.

D. COMPARISONS WITH STATE-OF-THE-ART METHODS ON TWO BENCHMARKS

In this subsection, we compare the model CSANet with state-of-the-art models on two benchmarks, Caltech and CityPersons. In this experiment, CSANet indicates that the initialization weights are derived from the ImageNet dataset model initialization weights. CSANet+City indicates that the initialization weight comes from CityPersons.

1) RESULTS ON CALTECH BENCHMARK

In this set of experiments, we produce the results on the Reasonable subset of the Caltech, and compare our two models with other state-of-the-art models including DeepParts [48], MS-CNN [49], FasterRCNN+ATT [2], SA-FasterRCNN [24], RPN+BF [8], SDS-RCNN [50], Adapt-FasterRCNN [9] and CSP+City [18]. The experiments are divided into two parts, IoU = 0.5 as shown in Fig. 3 and IoU = 0.75 as shown in Fig. 4. The FPPI curves are drawn from the method that is provided by [21].

As shown in Fig. 4, the model initialized from the CityPersons dataset has the best performance. Compared with the current advanced method CSP+City, CSANet achieves the best performance with MR^{-2} of 3.55%. The CSANet model initialized with the ImageNet dataset exceeds MR^{-2} of 3.88%, which indicates a significant improvement over the baseline models. As shown in Fig. 5, our models also achieve

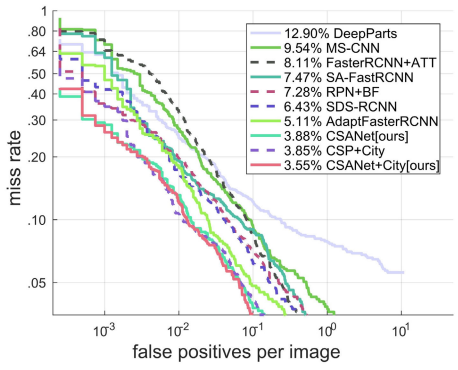


FIGURE 4. Comparisons of state-of-the-art detectors on the Reasonable subset of Caltech under IoU =0.5.

TABLE 5. Comparisons of state-of-the-art detectors on caltech.

Method	Hardware	Scale	Test Time (ms/img)	MR^{-2} (%)	
				IoU=0.5	IoU=0.75
DeepParts[48]	-		-	12.90	68.28
MS-CNN[49]	Titan X GPU		400	9.54	55.51
FasterRCNN+ATT[2]	-		-	8.11	64.08
SA-FasterRCNN[24]	Titan X GPU	×1.7	590	7.47	55.49
RPN+BF[8]	Tesla K40 GPU	×1.5	500	7.28	57.79
SDS-RCNN[50]	Titan X GPU		210	6.43	62.98
ALFNet[51]	GTX 1080 Ti GPU	×1	**50.0**	6.10	22.5
AdaptFastrerRCNN[9]	GPU	12 GB memory	-	5.11	24.97
CSP[18]	GTX 1080 Ti GPU	×1	59.6	4.54	28.80
ALFNet + City [51]	GTX 1080 Ti GPU	×1	**50.0**	4.50	**18.60**
CSANet[ours]	GTX 1080 Ti GPU	×1	57.2	3.88	20.05
CSANet+City[ours]	GTX 1080 Ti GPU	×1	54.9	**3.55**	18.86

Comparisons of state-of-the-art detectors on the Reasonable subset of Caltech. All data are reported from corresponding references. Hardware denotes the GPU device used for network training. Bold numbers indicates the best results.

TABLE 6. Comparisons of state-of-the-art detectors on citypersons.

Method	Hardware	Scale	Test Time (ms/img)	Reasonable	Heavy	Partial	Bare
AdaptFaster-RCNN[9]	GPU	12 GB Memory	-	15.4	-	-	-
RepLoss[52]	GPU	×4	-	14.6	60.6	18.6	7.9
TLL[26]	-	-	-	15.5	53.6	17.2	10.0
TLL(MRF)[26]	-	-	-	14.4	52.0	15.9	9.2
ALFNet[51]	GTX 1080Ti GPU	×2	**270**	12.0	51.9	11.4	8.4
CSP[18]	GTX 1080Ti GPU	×4	330	**11.0**	**49.3**	**10.4**	7.3
CSANet[ours]	GTX 1080Ti GPU	×1	320	12.0	51.3	11.9	**7.25**

Comparisons of state-of-the-art detectors on CityPersons under IoU=0.5. Hardware denotes the GPU device used for network training and the Scale shows the number of GPU. Bold numbers indicates the best results.

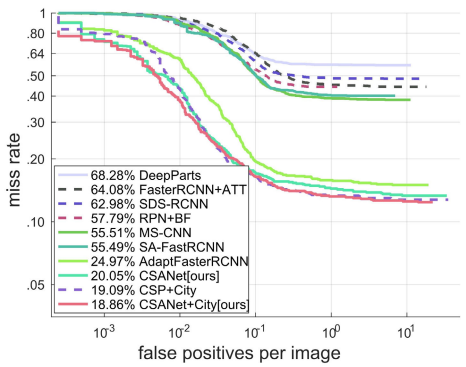


FIGURE 5. Comparisons of state-of-the-art detectors on the Reasonable subset of Caltech under IoU =0.75.

a smaller miss rate under stricter threshold settings, which means that the dual attention network also helps improve the quality of the bounding boxes.

Table 5 reports the detailed experimental results on Caletch, suggesting that CSANet significantly outperforms the competitors in accuracy while maintaining a high computational efficiency. The speed of the proposed method is about 18 FPS with the original 640 × 480 pixels, and the accuracy denotes MR^{-2} with 3.55% under IoU =0.5 and 18.86% under IoU =0.75. As can be seen from Table 5, the two-stage detectors have the tardy speed and our method achieves a better speed-accuracy trade-off.

2) RESULTS ON CITYPERSONS BENCHMARK

In this set of experiments, we produce the results on the CityPersons dataset under IoU =0.5, and we only use single NVIDIA GTX 1080Ti GPU and mini-batch =2. Table 6 shows that the CSANet detector achieves the state-of-the-art performance denoted MR^{-2} with of 7.25% on the Bare subset of CityPersons. On the Reasonable subset, the performance

of CSANet is second only to the CSP [18] model trained with mini-batch =8. Our detector is comparable to ALFNet [51] and produces about 2.6% improvement compared to the competitor RepLoss [52]. On two subsets of Heavy and Partial with different occlusion levels, the performance of our method still falls short of that of advanced detectors. In fact, within a reasonable range, a larger batch size makes the gradient descent direction more accurate.

V. CONCLUSION

In this paper, we propose a dual attention network to obtain larger receptive fields and contextual information, which can help CNNs better capture the image from a global perspective. We construct an anchor-free pedestrian detector based on dual attention network. As a result, the proposed detector CSANet achieves the state-of-the-art performance on Caltech benchmark and obtains mainstream performance on CityPersons benchmark.

It is worthy studying the introduction of attention mechanism into complex computer vision tasks. Our work also suggests that the attention mechanism can improve the performance of pedestrian detection to a certain extent. However, current detectors including CSANet only achieve about 20 FPS in detection speed, which is far from the standard of real-time detection. In future works, we will work on the design of simple yet effective attention networks to further contribute to the realization of real-time detection.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

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

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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

[9] 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.

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

[11] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[12] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 966–974.

[13] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[14] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[15] X. Zhou, D. Wang, and P. Krähenbähl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[16] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "FoveaBox: Beyond anchor-based object detector," 2019, *arXiv:1904.03797*. [Online]. Available: http://arxiv.org/abs/1904.03797

[17] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[18] 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.

[19] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "LFFD: A light and fast face detector for edge devices," 2019, *arXiv:1904.10633*. [Online]. Available: http://arxiv.org/abs/1904.10633

[20] X. Zhang, L. Cheng, B. Li, and H.-M. Hu, "Too far to see? Not really!—Pedestrian detection with scale-aware localization policy," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3703–3715, Aug. 2018.

[21] 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.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-CNN," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2017, pp. 2961–2969.

[23] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multi box detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[24] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2017.

[25] 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)*, 2018, pp. 637–653.

[26] 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. (ECCV)*, 2018, pp. 536–551.

[27] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[29] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 421–429.

[30] L. Mou, Y. Zhao, L. Chen, and J. Cheng, "CS-Net: Channel and spatial attention network for curvilinear structure segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2019, pp. 721–730.

[31] K. Xu, J. Ba, R. Kiros, and K. Cho, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2015, pp. 2048–2057.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[33] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2020.

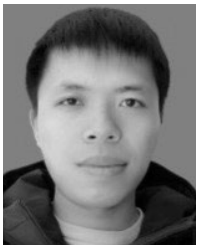
[34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31th AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 4278–4284.

[36] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[37] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.

[38] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.
[39] S. Woo, J. Park, and J. Y. Lee, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[40] T. Guo, D. Wang, Z. Jiang, A. Men, and Y. Zhou, "Deep network with spatial and channel attention for person re-identification," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.
[41] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 5659–5667.
[42] 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.
[43] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: A backbone network for object detection," 2018, *arXiv:1804.06215*. [Online]. Available: http://arxiv.org/abs/1804.06215
[44] 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.
[45] 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.
[46] H. Chen, Y. Wang, C. Xu, B. Shi, C. Xu, Q. Tian, and C. Xu, "AdderNet: Do we really need multiplications in deep learning?" 2019, *arXiv:1912.13200*. [Online]. Available: http://arxiv.org/abs/1912.13200
[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.
[48] 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.
[49] 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. (ECCV)*, 2016, pp. 354–370.
[50] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4960–4969.
[51] 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)*, 2018, pp. 618–634.
[52] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.



YUNBO ZHANG was born in Shangqiu, China, in 1993. He is currently pursuing the master's degree with the School of Software Engineering, Dalian University, and the Key Laboratory of Advance Design and Intelligent Computer (Ministry of Education). His research interests include artificial intelligence and pedestrian detection.

PENGFEI YI (Member, IEEE) was born in 1983. He received the Ph.D. degree from the Dalian University of Technology. He is currently a Lecturer with Dalian University, Dalian, China. His research interests include commuter graphics, artificial intelligence, and human–robot interaction.

DONGSHENG ZHOU (Member, IEEE) was born in 1978. He received the Ph.D. degree from the Dalian University of Technology. He is currently a Distinguish Professor of Liaoning Province. His research interests include CG, intelligence computing, and human–robot interaction. He is a member of ACM, CGS, and CCF.

XIN YANG (Member, IEEE) received the B.S. degree in computer science from Jilin University, in 2007, and the Ph.D. degree, in July 2012. From 2007 to June 2012, he was a joint Ph.D. student with Zhejiang University and UC Davis for Graphics. He is currently an Associate Professor with the Department of Computer Science, Dalian University of Technology, China. His research interests include computer graphics and robotic vision.

DEYUN YANG was born in Yuncheng, China. He received the Ph.D. degree in mathematics and information engineering from Nankai University. Since 2000, he has been a Professor with Taishan University. His main research interests are wavelet analysis, signal processing, and information technology. His research work has been supported by the National Natural Science Foundation of China and China Postdoctoral Science Foundation funded project.

QIANG ZHANG (Member, IEEE) was born in Xian, China, in 1971. He received the M.Eng. degree in economic engineering and the Ph.D. degree in circuits and systems from Xidian University, Xian, in 1999 and 2002, respectively. He was a Lecturer with the Center of Advanced Design Technology, Dalian University, Dalian, China, in 2003, and a Professor, in 2005. His research interests are bio-inspired computing and its applications. He has authored more than 70 articles in the above fields. So far, he has served in the editorial board of seven international journals and has edited special issues in journals, such as *Neurocomputing* and *International Journal of Computer Applications in Technology*.

XIAOPENG WEI (Member, IEEE) was born in Dalian, China, in 1959. He received the Ph.D. degree from the Dalian University of Technology, in 1993. He is currently a Professor with the Dalian University of Technology. His research areas include computer animation, computer vision, robot, and intelligent CAD. So far, he has coauthored about 200 articles published.