ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

Context attention network for occluded pedestrian detection

Shiyang Zhao[1, *]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
**Corresponding author*

Abstract: *Pedestrian detection in occluded scenes has always been a thorny problem in computer vision. In this case, due to the large difference in scale of occluded pedestrians and low visibility, it usually brings great challenges to detection. To solve this problem, this paper proposes a model structure for pedestrian occlusion detection, which improves the pedestrian detection method based on anchor-free. Specifically, we introduce a structure for extracting multi-scale context information to learn a better feature representation, and a channel attention module on each decoder layer to provide global context as a guidance of low-level features to select category localization details. Experimental results show that this method achieves 41.93% of MR^{-2} on the occlusion subset of Caltech pedestrian dataset, which is better than other contrast detectors.*

Keywords: *Pedestrian detection, multi-scale context, channel attention, anchor-free*

1. Introduction

Pedestrian detection is a significant research topic in object detection, which benefits many applications, driverless cars, intelligent robotics and intelligent transportation. It is quite common to utilize the methods proposed in object detection to detect pedestrians directly. However, these methods can hardly obtain the optimal performance. The main reason is that pedestrians always gather together and are easily obscured by other objects in reality. Therefore, it is challenging and meaningful to deal with occlusion problems in pedestrian detection.

In this regard, in order to improve the accuracy of pedestrian occlusion detection, some models designed for occlusion have also been proposed. The most common strategy to handle occlusionis based on part models, in which a series of part detectors corresponding to specific occlusion patterns are built. Then, an ensemble model is learned to integrate the part scores of these detectors. DeepParts [1] constructed a series of part detectors corresponding to specific occlusion mode, but this method based on part model is usually time-consuming and difficult to train. Bi-Box [2] uses two different branches: the visible part prediction branch and the pedestrian overall prediction branch. The two branches complement each other and perform occlusion detection based on the visible part of the pedestrian. Repulsion Loss [3] introduced a new regression loss function to make the generated prediction box as close to the specified target box as possible and far away from other surrounding target boxes. FasterRCNN+ATT-vbb[4] finds that different channels in CNN are corresponding to different body parts. Based on this perception, an attention mechanism across channels is employed to represent occlusion patterns. Among the various pedestrian detection approaches, there cently proposed CSP (Center and Scale Prediction) in [5] is a promising anchor-free detector, which can detect both center and scale for pedestrian detection. Detection is performed by predicting the center and scale of pedestrians. Although the CSP detector solves the challenges of various scales in pedestrian detection, it does not explicitly tackle the problem of pedestrian occlusion. To address the crowd occlusion challenge, we use a multi-scale context module and attention module to improve the stste-of-the-art CSP anchor-free detector. The main contributions of this work are summarized as follows:

(1) We propose a multi-scale context extraction module in the high-level features to integrate multi-scale context from different regions by multi-branch convolution layers with multiple receptive fields, which is able to make detectors more robust to occlusion.

(2) Then, We develop channel attention mechanism, an effective decoder module for pedestrian detection. The channel attention module is introduced to extract high-level global context to adjust low-level information.

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

(3) The proposed method achieved state-of-the-art performance on Caltech benchmark while maintaining computation efficiency.

2. Related Works

2.1 Center and Scale Prediction (CSP)-Based Detector

CSP was proposed by Wei Liu and Shengcai Liao in 2019. They first introduced anchor-free method into pedestrian detection area. More specifically, CSP includes two parts: feature extraction and detection head. In feature extraction part, a backbone, such as ResNet-50, MobileNet, is used to extract different levels of features [6-7]. Shallower feature maps can provide more precise localization information while deeper feature maps can provide high-level semantic information. In detection head part, convolutions are used to predict center, scale, and offset respectively.The objective function of CSP-based detector, denoted as L is the weighted sum of three losses as:

$$L=\lambda_c L_c + \lambda_s L_s + \lambda_o L_o \tag{1}$$

Where λ_c, λ_s and λ_o respectively represents weights for center classification, scale regression, and offset regression loss. We set these values to 0.01, 0.1, and 1.0, respectively, with reference to the original CSP-based detectors.

Three loss functions in (1) are defined as follows. The centerpoint classification loss function, denoted as L_c, is defined using the focal loss as:

$$L_c = -\frac{1}{K}\sum_{i=1}^{W/r}\sum_{j=1}^{H/r}\alpha_{ij}\left(1-\hat{p}_{ij}\right)^{\gamma} log\left(\hat{p}_{ij}\right) \tag{2}$$

Where K is the number of objects in an image, and $\hat{p}_{ij}$ indicates the existence of a center point of pedestrian.

The scale regression loss, denoted as L_s, is formulated as:

$$L_s = \frac{1}{K}\sum_{k=1}^{K} SmoothL1\left(s_k, t_k\right) \tag{3}$$

Where $SmoothL1$ represents the L1 smoothing loss function, and s_k and t_k respectively represent the prediction result and ground truth. The offset loss, denoted as L_o, is formulated by smooth L1 loss in the similar manner as Eq. (3).

2.2 Context modeling

In the real world, the target cannot exist alone, it must have more or less relationship with other objects around it, which is usually called context information. Context information is generally understood as the ability to perceive and apply some or all of the information that affects the objects in the scene and image. Based on the simulation of the human visual system, the human brain has excellent recognition performance, and the human visual system can still quickly identify and classify a large number of targets in the case of complex targets and backgrounds. It has good adaptability to illumination, attitude, texture, deformation and occlusion of target imaging.

Context information is perceived to be a helpful guidance for handling occlusion. In recent years, various types of context information have been applied to various fields of computer vision: such as object detection, semantic segmentation, human keypoint detection, etc. Therefore, how to extract multi-scale context information is an important problem in solving occlusion. The Inception block[8] adopts multiple branches with different kernel sizes to capture multi-scale information. However, all the kernels are sampled at the same center, which requires much larger ones to reach the same sampling coverage and thus loses some crucial details. For ASPP[9], dilated convolution varies the sampling distance from the center, but the features have a uniform resolution from the previous convolution layers of the same kernel size, which treats the cluesat all the positions equally, probably leading to

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

confusion between object and context. Deformable CNN [10] learns distinctive resolutions of individual objects,unfortunately it holds the same downside as ASPP. RFB Net [11] improved on the basis of two methods, stacking convolution kernels with different sizes and atrous rates to increase the size of the receptive field and capture more contextual information. Inspired by the above methods, this paper combines context and multi-scale information to combine dilated convolution and shortcut connection to enhance the expression of features to solve the problem of occlusion.

2.3 Attention mechanism

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, object detection, scene segmentation, image captioning, medical image segmentation, remote sensing imagery analysis, etc.

SENet [12] realizes channel reconstruction on feature graph by simulating the correlation between channels. Inspired by Senet and Inception, SKNet [13] has improved by combining the channel attention modules of Senet and Inception with the multi-branch convolutional layer. In addition, CBAM [14] proposed a dual attention structure integrated channel attention mechanism and spatial attention mechanism. The attention mechanism will guide the network to pay attention to the useful information and suppress the useless information when extracting features, so that the network can realize what features need attention and where features need attention. Inspired by the above methods, We introduced a channel attention module to extract high-level global context to adjust low-level information.

3. Methodology

3.1 Overall structure

The overall framework of our model is shown in Figure 1. The blue and red lines represent the downsample and upsample operators respectively. The backbone network is ResNet50, which is similar to the CSP. The detection head part mainly includes three 1×1 convolutional layers, which are used to predict the center position,, scale information, and offset. The ResNet-50 is divided into 5 stages. We define the output feature maps of 2 to 5 stages as φ_2, φ_3, φ_4 and φ_5, respectively. the input image is downsampled by 4, 8, 16, and 16, respectively. Because of φ_5 has advanced semantic features, a multi-scale context extraction block (MCB) is added afterwards to extract the context information of advanced features. After that, the upsampling and the features of the previous layer are input to the channel attention module(CA), and the global context information of the high-level features is extracted as a guide to adjust the low-level features, Finally high-level features are added with the weighted low-level features and upsampled gradually. the output feature map enters the detection head for prediction. In what follows, we describe the modeling process of two modules in detail.

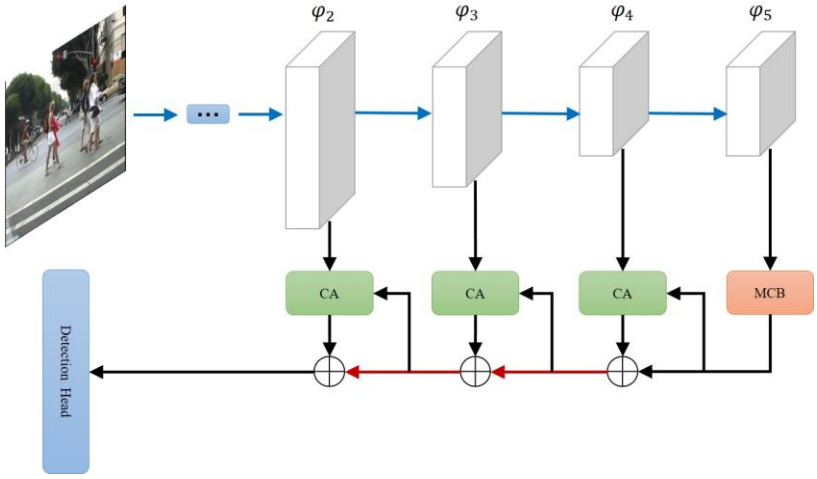


Figure 1: Overall architecture of model

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

3.2 Multi-scale context block

The proposed MCB is a multi-branch convolutional block. Its inner structure can be divided into two components: the multi-branch convolution layer with different kernels and the trailing dilated pooling or convolution layers.

To be specific, first, we employ the bottleneck structure in each branch, consisting of a 1×1 conv-layer, to decrease the number of channels in the feature map plus an n×n conv-layer. Second, since the use of 3×3 and 5×5 convolution may increase the amount of calculation, in order to reduce the number of parameters and increase the nonlinear capacity of the model, we replace the 5×5 conv-layer by two stacked 3×3 conv-layers to reduce parameters and deeper non-linear layers. For the same reason, we use a 1×n plus an n×1 conv-layer to take place of the original n×n conv-layer. Ultimately, In order to prevent network degradation, we apply the shortcut design from ResNet. At each branch, the convolution layer of a particular kernel size is followed by a pooling or convolution layer with a corresponding dilation. The dilated convolution layer uses the 3×3 kernel, and the atrous rate of 1, 3 and 5 is used to carry out the dilated convolution in the three branches respectively, and the receptive fields are 3, 9 and 15 respectively. The basic intention of this structure is to generate feature maps of a higher resolution, capturing information at a larger area with more context while keeping the same number of parameters. This design has also been adopted by some well-known object detectors, such as SSD[15] and R-FCN[16], to improve accuracy. The network structure of this module is shown in Figure 2.

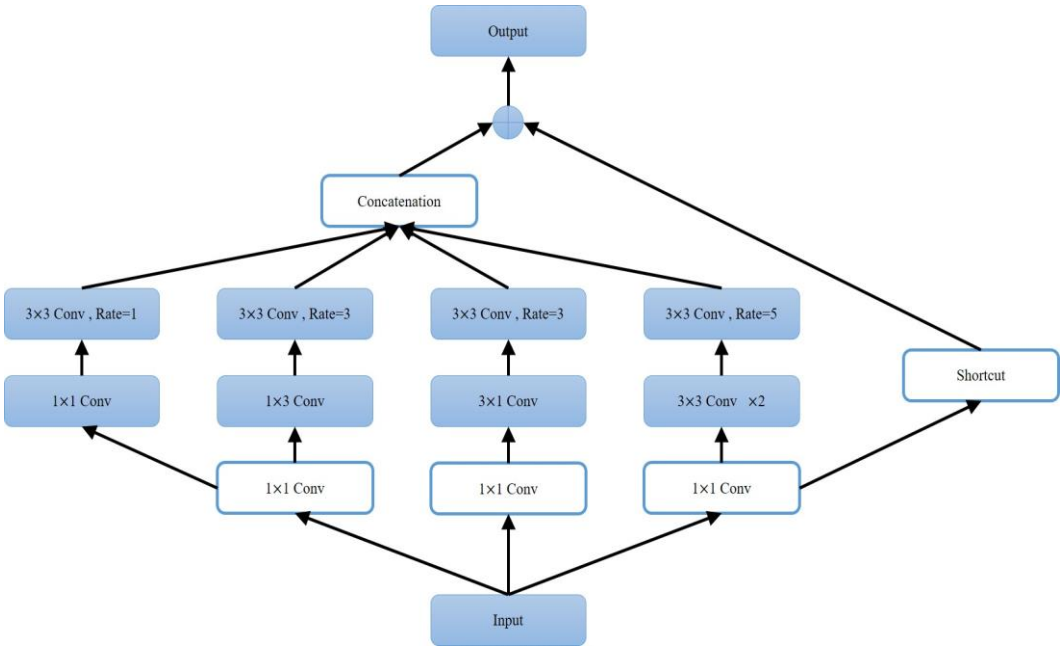


Figure 2: Multi-scale context block

3.3 Channel attention module

Combining CNNs with well-designed context module can obtain considerable performance and capability to obtain category information. Furthermore, high-level features with abundant category information can be used to weight low-level information to select precise resolution details.

Our channel attention module performs global average pooling to provide global context as a guidance of low-level features to select category localization details. In detail, we perform 3×3 convolution on the low-level features to reduce channels of feature maps from CNNs. The global context generated from high-level features is through a 1×1 convolution with batch normalization and ReLU non-linearity, then multiplied by the low-level features. Finally high-level features are added with the weighted low-level features and upsampled gradually. This module deploys different scale feature maps more effectively and uses high-level features provide guidance information to low-level feature maps in a simple way. The network structure of this module is shown in Figure 3.

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

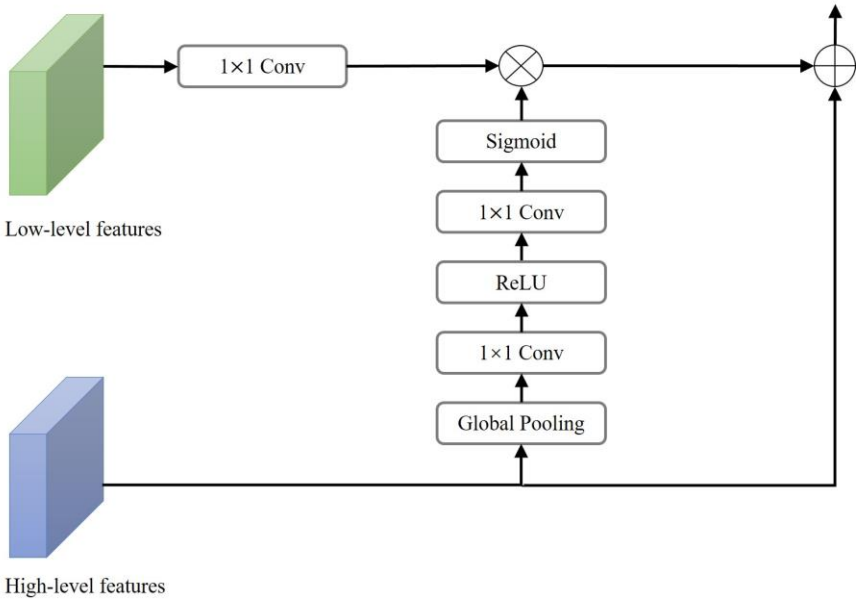


Figure 3: Channel attention module

Specifically, the channel attention module consists of two main operations: Squeeze and Excitation. Among them, the Squeeze part is mainly to obtain the global information of each channel feature graph and generate a feature vector. This step is achieved by using Global Average Pooling (GAP). Set the original feature as $F=\{f_1, f_2, ..., f_c\}$, where f_c represents the pixel value of channel c, then the GAP can be expressed as:

$$m_c = F_{sq}(f_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_c(i, j) \tag{4}$$

Where m_c is the average value of each channel feature graph.

To make use of the information aggregated in the squeeze operation, we follow it with a second operation which aims to fully capture channel-wise dependencies. To fulfil this objective, the function must meet two criteria: first, it must be flexible(in particular, it must be capable of learning a nonlinear interaction between channels) and second, it must learn a non-mutually-exclusive relationship since we would like to ensure that multiple channels are allowed to be emphasised (rather than enforcing a one-hot activation). To meet these criteria, we opt to employ a simple gating mechanism with a sigmoid activation:

$$e = F_{ex}(m, W) = \sigma(g(m, W)) = Sigmoid(W_2 ReLU(W_1 m)) \tag{5}$$

$W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ are respectively the weight matrices of two full connection layers (in order to maintain the image space structure, two 1×1 convolution are substituted). C represents the channel dimension, and r represents the number of hidden layer nodes in the middle of the full connection layer.

The final output of the block is obtained by rescaling F with the activations e:

$$\bar{X} = F_{scale}(f_c, e_c) = e_c \bullet f_c \tag{6}$$

Where $\bar{X} = \{x_1, x_2, ..., x_c\}$ and $F_{scale}(f_c, e_c)$ refers to channel-wise multiplication between the scalar e_c and the feature map $f_c \in R^{H \times W}$.

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

4. Results and discussion

4.1 Dataset

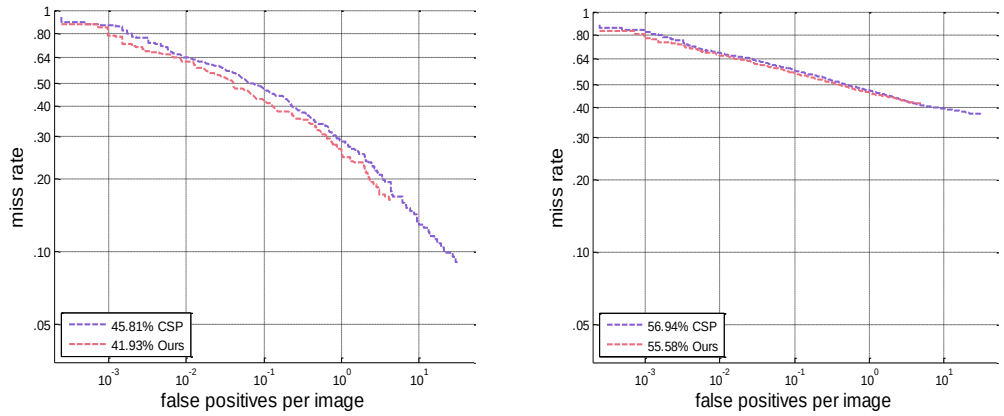
In order to prove the effectiveness of the proposed method, an evaluation was made on the Caltech pedestrian detection dataset, which is popular at present. The Caltech benchmark is composed of 10 hours video of urban driving with the image size of 640×480. In total itcontains about 350,000 bounding boxes around 2300 unique pedestrians. The evaluation metric is the log-average missrate sampled at a false positive per image (FPPI) range of $[10^{-2},10^{0}]$. Caltech has various evaluation settings and weconsider three frequently-used subsets for evaluation: Overall(height≥20 pixels), Heavy(35%≤occlusion≤80%). For training, we sample from the standard training set which contains 42,782 frames, while for inference we evaluate our framework on the 4,024 frames in the standard test set.

4.2 Implementation Details

We implement the proposed method in Keras. The backbone is ResNet50 pretrained on ImageNet unless otherwise stated. Adam is applied to optimize the network. We also apply the strategy of moving average weights proposed in [15] to achievemore stable training. To increase the diversity of the training data, standard data augmentation techniques are adopted. Firstly, random color distortion and horizontal flipare applied, followed by randomly scaled in the range of [0.4, 1.5]. Secondly, a patch is cropped or expanded by zero-padding such that the shorter side has a fixed number of pixels (336 for Caltech). Note that the aspect ratio of the image is kept during this process. For Caltech, a mini-batch contains 8 images with one GPU (GTX 1080Ti), the learning rate is set as 10^{-4} and training is stopped after 15K iterations.

4.3 Experimental results and analysis

In order to prove the effectiveness of the improved method proposed in this paper, a comparison was first made with the original CSP algorithm on the Caltech pedestrian data set. The calculated FPPI-MR curve is shown in Fig.4 , where 4(a), 4(b) respectively represent the detection results of heavy occluded subset and all subset. It can be observed from Fig.4 that This model has the greatest improvement on HO subset compared with CSP. This is because pedestrians with large occlusion ratio can extract fewer features and need more context information to help detect. MCB can extract multi-scale context information to learn a better feature representation, and CA on each decoder layer to provide global context as a guidance of low-level features to select category localization details. Our model achieves MR^{-2} of 41.93% on the HO setting, which outperforms CSP by 3.88%. For the ALL subset, Our model achieves MR^{-2} of 55.58%, MR^{-2} reduces by 1.36% compared to CSP, which effectively improves the detection performance of the model.



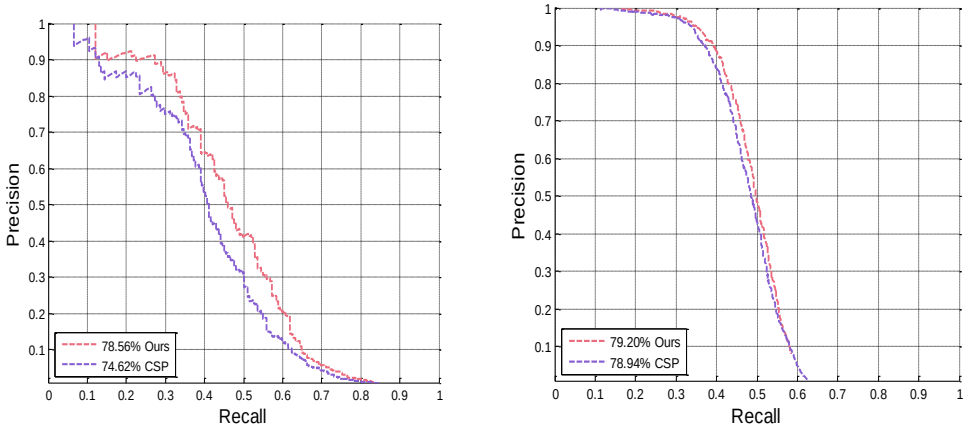
(a) HO *(b) ALL*

Figure 4: FPPI-MR curve

In order to better evaluate the model, the recall ratio and precision ratio of the algorithm were compared, and the PR curve obtained was shown in Fig. 5. The accuracy of the improved algorithm on

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

HO and ALL subsets is increased by 3.9% and 0.3% respectively, and the false detection rate is reduced, which proves the effectiveness of the proposed algorithm.



(a) HO *(b) ALL*

Figure 5: PR curve

In addition, we produce the results on the HO and ALL subset of the Caltech, and compare our model with other state-of-the-art models including RPN+BF[16], MS-CNN[17], SDS-RCNN[18], GDFL[19], ALFNet[20], DeepParts[1] , Bi-Box[2], RepLoss[3], ATT-part[4]. The experimental comparison results are shown in Table 1.

Table 1: Comparisons of state-of-the-art detections on Caltech

Method	Backbone	HO	ALL
RPN+BF	VGG-16	74.4	64.7
DeepParts	AlexNet	60.4	64.8
MS-CNN	VGG-16	59.9	60.9
SDS-RCNN	VGG-16	58.6	61.5
ALFNet	ResNet-50	51.0	59.1
RepLoss	ResNet-50	47.9	59.0
ATT-part	VGG-16	45.2	-
Bi-Box	VGG-16	44.4	-
GDFL	ResNet-50	43.2	-
Ours	ResNet-50	41.9	55.6

Table 1 reports the detailed experimental results on Caletch, suggesting that our model significantly outperforms the competitors in accuracy. When compared with anchor-based methods, the advantage of our model lies in two aspects. Firstly, this model does not require tedious configurations on anchors specifically for each dataset. Secondly, anchor-based methods detect objects by overall classifications of each anchor where background information and occlusions are also included and will confuse the detector's training. However, our model overcomes this drawback by scanning for pedestrian centers instead of boxes inan image, thus is more robust to occluded objects.

Figure 6 shows the detection result of CSP model and our proposed model in the face of occlusion. Where, the left side is the detection result of the CSP model, and the right side is the detection result of our proposed model. The red bounding box represents missed detection, and the green bounding box represents correct detection. It can be seen that the improved model in this paper is more robust in occlusion scenarios.

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

Figure 6: Comparisons of occlusion scene detection

5. Conclusion

In this paper, we proposes a model structure for pedestrian occlusion detection, design a structure for extracting multi-scale context information to learn a better feature representation, and a channel attention module on each decoder layer to extract high-level global context to adjust low-level feature. As a result, the proposed detector achieves the state-of-the-art performance on Caltech benchmark. In future works, we will work on the design of simple yet effective networks to further contribute to the realization of real-time detection.

References

[1] Y. Tian, P. Luo and X. Wang (2015). Deep learning strong parts for pedestrian detection. International Conference on Computer Vision, p.1904-1912.
[2] C.L. Zhou and J.S. Yuan (2018). Bi-box regression for pedestrian detection and occlusion estimation. European Conference on Computer Vision, p.138-154.
[3] X. Wang and T. Xiao (2018). Repulsion loss: Detecting pedestrians in a crowd. IEEE Conference on Computer Vision and Pattern Recogni-tion, p.7774-7783.
[4] S.S. Zhang, J. Yang and B. Schiele (2018). Occluded pedestrian detection through guided attention in cnns. IEEE Conference on Computer Vision and Pattern Recognition, p.6995-7003.
[5] W. Liu and S. Liao (2019). High-level semantic feature detection: A new perspective for pedestrian detection. IEEE Conference on Computer Vision and Pattern Recognition, p.5187-5196.
[6] K. He, X.Y. Zhang and S.Q. Ren (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, p.770-778.

ISSN 2616-5775 Vol. 4, Issue 1: 66-74, DOI: 10.25236/AJCIS.2021.040111

[7] A.G. Howard, M. Zhu and B. Chen (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.
[8] C. Szegedy and S. Loffe (2017). Inception-v4, inception-resnet and the impact of residual connections on learning.The AAAI Conference on Artificial Intelligence.
[9] L.C. Chen, G. Papandreou and F. Schroff (2017). Rethinking atrous convolution for semantic image segmentatio. arXiv:1706.05587
[10] J. Dai (2017). Deformable convolutional network. Proceedings of the IEEE Iternational Conference on Computer Vision.
[11] S. Liu and D. Huang (2018). Receptive field block net for accurate and fast object detection. IEEE Conference on Computer Vision and Pattern Recognition, p.385-400.
[12] J. Hu, L. Shen and G. Sun (2018). Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition, p.7132-7141.
[13] X. Li, W. Wang and X. Hu (2019). Selective Kernel Networks. IEEE Conference on Computer Vision and Pattern Recognition, p.510-519.
[14] S. Woo, J. Park and J.Y. Lee (2018). Cbam: Convolutional block attention module. European Conference on Computer Vision, p.3-19.
[15] A. Tarvainen and H. Valpola (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning re-sults. Advances in neural information processing systems, p.1195-1204.
[16] L. Zhang, L Lin and X. Liang (2016). Is faster r-cnn doing well for pedestrian detection?. European Conference on Computer Vision, p.443-457.
[17] Z.W. Cai and Q.F. Fan (2016). A unified multi-scale deep convolutional neural network for fast object detection. European Conference on Computer Vision.
[18] M. Cordts, M. Omran and S. Ramos (2016). The cityscapes dataset for semantic urban scene understanding. IEEE Conference on Computer Vision and Pattern Recognition, p.3213-3223
[19] C.Z. Lin, J.W. Lu and G. Wang (2018). Graininess-aware deep feature learning for pedestrian detection. European Conference on Computer Vision.
[20] W. Liu, S. Liao and W. Hu (2018). Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. European Conference on Computer Vision, p.618-634