Small-scale Pedestrian Detection Method based on Attention Mechanism and Adaptive Feature Fusion

Yuhao You¹, Houjin Chen¹*, Yanfeng Li¹, Minjun Wang¹ and Jinlei Zhu²

¹ School of Electronic Information Engineering, Beijing Jiaotong University, Beijing, China
² Synthesis Electronic Technology Co., Ltd., Jinan, Shandong, China
*Corresponding author’s e-mail: hjchen@bjtu.edu.cn

Abstract. In pedestrian detection, high detection accuracy has been achieved for large-scale pedestrians, but for small-scale pedestrians, the detection effect needs to be further improved. In order to improve the detection accuracy of small-scale pedestrians, this paper proposes a YOLOv4 small-scale pedestrian detection algorithm based on the fusion of attention and weighted features. In order to enhance effective features and suppress ineffective features, an attention mechanism that adapts features is introduced in the backbone network CSPdarknet53 to reduce the interference to small-scale pedestrian detection. In order to better integrate features with inconsistent semantics and scales, adaptive channel-weighted feature fusion is used in the feature pyramid, so that the deep and shallow features focus on pedestrian targets of corresponding scales. The logarithmic mean missed detection rate (LAMR) on the small-scale pedestrian test sets "Far" and "Medium" of the Caltech public data set has decreased: 9.67% and 7.38%. Compared with other pedestrian detection algorithms, it has obvious advantages for small-scale pedestrian detection.

1. Introduction

Pedestrian detection is one of the most important research topics in the computer vision in these days. The combination of pedestrian detection and deep learning has further improved the detection accuracy. It has a wide range of applications in assisted driving systems, human behavior analysis, robot development, video surveillance and other fields. At this stage, the detection of large-scale pedestrians has achieved high accuracy, but the detection of small-scale pedestrians is not ideal. Improving the detection performance of small-scale pedestrians has become an important direction in the field of pedestrian detection research, which is useful for broadening the scope of application of pedestrian detection. And it is of great significance to improve the application performance of pedestrian detection.

Small-scale pedestrians have insufficient semantic information and small coverage area. The main reasons for the low accuracy of small-scale pedestrian detection are: The complex environment interferes with small-scale pedestrian detection: Under complex background noise, the information of small-scale pedestrians will be masked by the noise of other larger objects. The existing research on small-scale pedestrian detection algorithms are optimized based on general pedestrian detection algorithms. In 2017, Feature Pyramid Networks (FPN) [1] was proposed. The feature map with high resolution and high semantics significantly improves the accuracy of small target detection. Scale-Transfer Module (STM) [2] obtains semantic features of different scales without adding additional parameters. Then, Deconvolutional Single Shot Detector (DSSD) algorithm [3] was proposed, using
ResNet-101 as the basic network, and forming a "wide-narrow-wide" "hourglass" structure by adding a deconvolution layer, which significantly improved the detection accuracy of small targets. The RetinaNet network [4] combines FPN with a fully convolutional neural network, and proposes that the focus loss successfully solves the problem that the target detection loss is dominated by negative samples when the positive and negative samples are extremely imbalanced. Scale Normalization for Image Pyramids (SNIP) algorithm [5] is improved on the basis of multi-scale training, by limiting the gradient back propagation of targets of different scales in the training process, and weakening the interference of very large or very small targets in the training sample on the training effect. Design Backbone for Object Detection (DetNet) [6] pointed out the problems in the pre-training mechanism used by the existing target detection algorithms, and designed a new network for target detection tasks. Based on the idea of hollow convolution, a low-complexity expansion bottleneck structure was proposed to achieve higher resolution features. The semantic fusion network [7] combines target detection with semantic segmentation tasks, uses semantic segmentation to supervise the detection process, restores the output feature map to the pixel level, and subdivides each point on the feature map to improve the network’s response Detection performance of small targets. The above research mainly designed a multi-layer network structure and fuse deep and shallow features to adapt to the detection of small-scale pedestrians, but did not consider the fusion of deep and shallow features during feature fusion. In addition, feature extraction is the previous step of feature fusion, and the extraction of more effective small-scale pedestrian features will help improve detection performance.

In this paper, we proposed a small-scale pedestrian detection method based on attention mechanism and adaptive feature fusion for small-scale pedestrian detection. By introducing an attention mechanism that adapts to features in the YOLOv4 [8] backbone network CSPdarknet53, and in the feature fusion part of the feature pyramid, using self the feature fusion method that adapts to channel weighting improves the detection accuracy of small-scale pedestrians. Experiments on Caltech public dataset show that our method has better detection performance for small-scale pedestrians than the original YOLOv4.

Figure 1. YOLOv4 network based on the fusion of attention and weighted features
2. Methods

Attention module and Weighted Feature Fusion YOLOv4 structure is shown in Figure 1. According to the structural characteristics of the backbone network, by introducing an attention mechanism that adapts to features in CSPDarknet53, the model will pay more attention to the extraction of pedestrian features when extracting features, while suppressing invalid features and noise, making more effective input into the feature pyramid; the PA-Net feature pyramid module adopts an adaptive channel weighted feature fusion method to weight the features of different channels, so that the fusion features of each layer focus on the pedestrians of its corresponding scale, and for the shallow fusion feature focuses more on the detection of small-scale pedestrians.

2.1 Attention mechanism

Convolutional Block Attention Module (CBAM) [9] is a general hybrid attention model that can be integrated into various convolutional neural networks for end-to-end training. For the feature map generated by the convolutional neural network, CBAM calculates the feature map and the attention map from the two dimensions of channel and space, and then multiplies the attention map with the input feature map to realize the adaptive learning of features.

The backbone network of YOLOv4 is CSPDarknet53, which mainly includes 5 CSP modules, which are stacked by small residual networks. When the size of the input image is (608×608×3), the size of the output feature maps of the 5 CSP modules is expressed in the form of \((w\times h \times c)\) as \((304\times 304\times 64), (152\times 152\times 128), (76\times 76\times 256), (38\times 38\times 512), (19\times 19\times 1024)\), where \(c\) represents the channel number of the feature map, \(w, h\) represent the width and height of the feature map of each channel. As the features are transferred to the deep CSP module, the number of channels increases, the feature information of the channel dimension increases, the width and height of each channel feature map decrease, and the feature information of the spatial dimension decreases. Therefore, in different feature layers, according to each The channel characteristics and spatial characteristics of the CSP module feature map introduce the corresponding attention mechanism, as shown in Figure 2. The first two CSP modules introduce the spatial attention mechanism, the middle two CSP modules introduce the hybrid attention mechanism, and the last CSP The module introduces the channel attention mechanism. The attention mechanism is used to learn the weight distribution from the feature, and the learned weight distribution is applied to the original feature to enhance the effective feature and suppress the invalid feature and noise.

![Figure 2. Attention mechanism that adapts to features](image)

2.2 Adaptive feature fusion

In the PANet feature pyramid module, the deep feature map contains more information about large-scale targets, and the shallow feature map contains more information about small-scale targets. However, in the feature fusion, the deep and shallow features are fused in equal proportions, which is not conducive to the detection of targets related to its scale in each layer. The feature fusion method of adaptive channel weighting is used to add weights to the features of different layers to control the features transmitted from the deep layer to the shallow layer, so that the deep and shallow features are no longer proportionally fused, and each layer focuses more on detection and Objects related to this level of scale.

The image is extracted from the backbone network to obtain feature maps of different scales. After convolution, the features in the FPN in the feature pyramid are \((76\times 76\times 256)\) and \((38\times 38\times 512)\) from the shallow layer to the deep layer. \((19\times 19\times 1024)\). In the PANet feature pyramid module, the specific process of the feature fusion method of adaptive channel weighting is shown in Figure 3. Take the fusion of two features of different sizes \((19\times 19\times 1024)\) and \((38\times 38\times 512)\) as an example. First, the channel is compressed through convolution, and the output feature maps are respectively \((19\times 19\times 256)\) and
(38×38×256), and then the (19×19×256) features are up-sampled. At this time, two The feature map size is the same (38×38×256), and then the channel-based concat operation is performed to output the feature map size (38×38×512), and finally sent to the weighted feature fusion module SE[13], for each The feature of the channel is weighted, and the size of the feature map after output fusion is (38×38×512).

![Figure 3. Feature fusion with adaptive channel weighting](image)

3. Results and Discussion

3.1 Dataset

The Caltech dataset is the most commonly used dataset in the field of pedestrian detection. The picture resolution is 640×480, set00-set05 is the training dataset, and set06-set10 is the test set. In set00-set05, one frame is extracted from every 3 frames to form the training set of this network, which contains a total of 42782 pictures. In set06-set10, one frame is extracted from every 30 frames for network performance evaluation, which contains a total of 4024 pictures.

When evaluating network performance on the Caltech dataset, pedestrians of different scales are evaluated separately according to the pixels occupied by pedestrians in the picture. The algorithm in this paper is mainly aimed at improving the detection performance of small-scale pedestrians, so evaluate the detection performance of this algorithm on the subsets of "Far" (pedestrian width of 8-12 pixels) and "Medium" (pedestrian width of 12-32 pixels).

3.2 Qualitative analysis

![Figure 4. An example of some test results on the Caltech dataset](image)
The algorithm in this paper is tested on the Caltech dataset, and some representative results are selected for display, as shown in Figure 4. (A) The figure is the real box, (b) and (c) are the prediction results of the YOLOv4 network and the algorithm in this paper, respectively. In the first test scenario, the algorithm in this paper detected small-scale pedestrians on the sidewalk on the right rear; in the second test scenario, the algorithm in this paper successfully detected small-scale pedestrians on the middle and rear highway. In these two scenarios, the original YOLOv4 algorithm missed the detection of small-scale pedestrians in the above areas, indicating that the improved algorithm in this paper can effectively improve the detection performance of small-scale pedestrians.

3.3 Quantitative analysis

3.3.1 Evaluation index

In order to evaluate the detection performance of the algorithm more objectively, this paper adopts a common pedestrian detection evaluation index—Log-Average Miss Rate (LAMR). The calculation method is the average missed detection rate value under 9 FPPI values (evenly spaced in the logarithmic space within the value range [0.01, 1.0]). The specific calculation method is as shown in formula 1.

\[
LAMR = \exp\left(\frac{1}{9} \sum_{k=0}^{8} \log(MR_{FPPI=10^{2k+1}})\right)
\]

Where: missed detection rate MR=FN/NP, false detection rate FPPI=FP/NI. FN means false negative, FP means false positive, NP is the total number of rows in the test set, and NI is the total number of pictures in the test set.

3.3.2 Ablation experiment

The effectiveness of the improved algorithm is verified by ablation experiments. The test results are shown in Table 1. "Attentional" represents the attention mechanism of adaptive features, and "Fusion" represents adaptive channel weighted feature fusion. The original YOLOv4 network has a LAMR of 78.85% and 38.29% on the small-scale pedestrian test set "Far" and "Medium", respectively, 13.45% on the conventional-scale pedestrian test set "Reasonable", and on all test sets "All" is "51.34". By increasing the attention mechanism of adaptive features, the LAMR on "Far" and "Reasonable" is reduced by 2.24% and 4.31% respectively, and the improvement effect is not obvious for the "Medium" and "All" datasets; by adding adaptive feature fusion in terms of "Far", "Medium", "Reasonable" and "All", LAMR has been reduced by 8.54%, 3.15%, 3.46% and 4.09% respectively. Finally, two methods were added at the same time, and the LAMR on "Far", "Medium", "Reasonable" and "All" was reduced by 9.67%, 7.38%, 5.32% and 6.37% respectively. The results show that the improved algorithm has improved the detection performance on each subset of Caltech, especially for small-scale pedestrians.

| Table 1 Comparison of test results of network comprehensive improvement in Caltech dataset |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                 | Caltech         | LAMR(%)         |                 |                 |
|                                 | Far             | Medium          | Reasonable      | All             |
| YOLOv4                          | 75.85           | 38.29           | 13.45           | 51.34           |
| YOLOv4+Attentional              | 73.61           | 38.70           | 9.14            | 51.32           |
| YOLOv4+Fusion                   | 67.31           | 35.14           | 9.99            | 47.25           |
| YOLOv4+Attentional+Fusion       | 66.18           | 30.91           | 8.13            | 44.97           |

3.4 Comparison with different algorithm

The results of our algorithm and other pedestrian detection algorithms are shown in Table 2. There is a 0.78% gap between our algorithm and the SDS-RCNN algorithm in the "Reasonable" subset, but in other subsets, it has a lower log-average missed detection rate. It can be seen that our algorithm has better pedestrian detection performance on each subset than other algorithms, especially for small-scale pedestrian detection, it shows that the algorithm in this paper has obvious advantages for the detection of small-scale pedestrians.
Table 2 Comparison of test results of different algorithms on the Caltech dataset

| Algorithm   | Far    | Medium | Reasonable | All    |
|------------|--------|--------|------------|--------|
| SDN        | 100.0  | 74.58  | 37.87      | 78.42  |
| SD-RCNN    | 100.0  | 50.88  | 7.35       | 61.50  |
| SA-FastRCNN| 100.0  | 51.83  | 9.68       | 62.59  |
| F-DNN      | 77.47  | 33.27  | 8.65       | 50.55  |
| GDFL       | 70.97  | 32.49  | 7.85       | 48.14  |
| MS-CNN     | 97.22  | 49.12  | 9.95       | 60.95  |
| RPN+BF     | 100.0  | 53.93  | 9.58       | 64.66  |
| Faster R-CNN| 72.28 | 34.72  | 12.38      | 50.79  |
| YOLOv3     | 76.99  | 41.05  | 14.58      | 53.07  |
| YOLOv4     | 75.85  | 38.29  | 13.45      | 51.34  |
| **Propose**| **66.18**| **30.91**| **8.13**  | **44.97**|

4. Conclusions
This study proposed a small-scale pedestrian detection method based on attention mechanism and adaptive feature fusion and experiments on the public dataset Caltech. In order to improve the detection accuracy of small-scale pedestrians, the attention module of adaptive features is introduced into the backbone network, and the feature fusion method of adaptive channel weighting is adopted in the feature pyramid module of PANet to improve the detection performance of small-scale pedestrians. In addition, our algorithm has certain advantages in detecting small-scale pedestrians compared with other pedestrian detection algorithms. The next step is to use data enhancement methods to generate more small-scale pedestrian dataset to fundamentally solve the problem of small-scale pedestrian detection.

Acknowledgments
The work was supported in part by the Major Science and Technology Innovation Project of Shandong Province [grant number 2019TSLH0206].

References
[1] Lin T, Dollár P, Girshick R, He K, Hariharan B, and Belongie S. (2017) Feature pyramid networks for object detection. arXiv:1612.03144v2 [cs.CV] for this version
[2] Zhou P, Ni B, Geng C, et al. (2018) Scale-transferrable object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018:528-537
[3] Fu C Y, Liu W, Ranga A, et al (2017). DSSD: Deconvolutional Single Shot Detector[J]. arXiv:1701.06659.
[4] T. Lin, P. Goyal, R. Girshick, et al. (2020) Focal Loss for Dense Object Detection[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(2):318-327.
[5] SINGHB and DAVISLS. (2018) An analysis of scale invariance in object detection-SNIP[C]// 2018 IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2018. 3578-3587.
[6] Li Z M, Peng C, Yu G, et al. (2018) DetNet: Design Backbone for Object Detection[C]// European Conference on Computer Vision. Springer, Cham, 2018: 339-354.
[7] Brazil G, Yin X, Liu X. (2017) Illuminating Pedestrians via Simultaneous Detection and Segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 4960-4969.
[8] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934v1 [cs.CV] 23 Apr 2020
[9] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon. (2018) CBAM: Convolutional Block Attention Module. arXiv:1807.06521v2 [cs.CV] for this version.