International Core Journal of Engineering Volume 7 Issue 7, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

Multiscale Feature Fusion Pyramid Networks for Object Detection in UAV-captured Images

Haoran Li[1], Boyong He[2]

[1]School of Information Engineering, China University of Geosciences, Beijing 100083, China;

[2]School of Aerospace Engineering, Xiamen University, Xiamen 361102, China.

Abstract

Due to the change of flight altitude and attitude of UAV, the object scale in UAV images exists difference which leads to a great challenge for object detection and has drawn wide attention. In this paper, an improved object detection network named MR Cascade-RCNN is proposed to deal with the multi-scale problem in object detection task for UAV images. Furthermore, we propose a novel method called Smooth Multiscale Feature Fusion Pyramid Networks (SEMLPCN), which is aimed at obtaining rich features as much as possible, improving the information propagation and reuse. Specifically, the dense connection is designed to fully utilize the representation from the different convolutional layers. Furthermore, cascade architecture is applied in the second stage to enhance the localization capability. Experiments on the drone-based datasets named VisDrone suggest a competitive performance of our method.

Keywords

Object Detection; Feature Fusion; UAV Images.

1. Introduction

With the rapid development of aerial technology, especially for UAVs which have found a wide range of applications in the commercial field, including agricultural, aerial photography, fast delivery, environmental monitoring, etc. [1]. Consequently, more and more attention has been paid to the research of general computer vision algorithms, such as object detection in aerial images. In previous years, many works are mainly focused on the sliding window search [2] and the handcrafted features [3], which normally require a lot of prior knowledge and formula derivation. In recent years, object detection based on deep learning algorithms has become the dominant technique, these methods, such as R-CNN series [4-6], YOLO series [7-10], SSD series [11,12], etc. have achieved great success in natural image detection (e.g., images in Pascal VOC [13], MS COCO [14]). However, these approaches have been to result in undesirable performance when detecting objects in images or videos captured from UAVs. [15]

UAVs (or Drones) have been usually deployed in the large scene, that means there are lots of objects in a single image and most of these objects are very small size, which is a big characteristic for aerial images and remains an open challenge. Generally, some detectors, such as Faster R-CNN, SSD, and YOLO, only utilize the feature map from a single layer of CNN networks, which has limited the representation capability of the feature information. Recent works focus on feature fusion for object detection, Feature Pyramid Network (FPN) [16] is one of a classical method that combines low-level and high-level features information by adopting a top-down architecture and lateral connections. This typical feature fusion method greatly improves the detecting performance for objects with small size, since the low-level features have sufficient location information, which is quite important to small objects for both classification and localization. However, considering that each pyramid layer of FPN

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

only focuses on the lateral connections from the corresponding feature map as shown in the lower part of figure 1, feature information of these other layers is still available to utilize and it is necessary to extract more contextual semantic information for small objects.

To obtain more sufficient representations of feature maps, we propose a novel feature fusion method named Smooth Multiscale Feature Fusion Pyramid Networks (SEMLPCN) to more efficiently fuse the low-level and high-level features information for feature maps. Besides, cascade architecture is used to refine the bounding box prediction in the second stage. A detailed outline of our framework is presented in Section 3. In general, the main contributions of this work are as follows:

1. We design a simple but effective feature fusion method called SEMLPCN, which can fully exploit feature propagation, feature reuse, and enhance the performance for the prediction of objects, especially for the small objects in aerial images.

2. We adopt Cascade architecture in the second stage to refine the bounding box regression and overcome the difficulty of locating small objects in complex backgrounds.

3. Extensive experiments on the aerial image datasets named VisDrone [17] demonstrates the validity and stability of the proposed framework.

The rest of this paper is organized as follows. Section 2 gives the related work of natural image detection and small object detection. Section 3 shows the analysis and description of the proposed framework. Section 4 presents experiment of the drone-based dataset. The last section concludes this paper.

2. Related Work

Nowadays, Object detection has made hot interests among researchers in the field of computer vision. Generally, Traditional detection methods usually need much prior knowledge and complex formula derivations and thus have defects of generalization ability. But CNN-based object detection methods can learn feature information automatically and are more likely to apply in real life because of the strong ability of generalization. In general, the detection methods based on deep learning can be generally divided into one-stage methods and two-stage methods. One-stage methods are good at high speed, including SSD, YOLO, and two-stage methods such as Fast-RCNN, Faster-RCNN, and R-FCN, have a good performance on accuracy.

RCNN [4] is the first successful work that leads the methods of deep learning into object detection, it adopts the selective search algorithm and uses SVM as the classifier. Fast R-CNN improves by using ROI pooling method. Based on that, Faster RCNN adopts RPN (region proposal networks) to generate proposals. However, Faster R-CNN makes the prediction only on the final feature map, which is unfavorable to detect the small objects.

Considering the drawback in Faster R-CNN, SSD utilizes the multi-level features to predict on multiscale objects. Specifically, it uses low-level feature maps to predict the small objects and use high-level feature maps to predict the large objects, yet it does its prediction on the intermediate feature maps without using the shallow feature maps, which is important to detect small objects. General speaking, it is not beneficial to predict the small object on low resolution, because the large stride of feature extractor has made the semantic information of small objects vanished, so it is quite hard to detect its features in high-level feature maps.

In order to solve this problem, FPN addressed the problem in SSD by fusing the low-level and high-level features, including the bottom-up pathway, top-down pathway, and lateral connection. As everyone knows that the high-resolution maps have strong location information and low-resolution maps have strong semantic information, both of which are of vital importance for object detection especially for detecting small objects. Similar to the feature fusion methods of FPN, relevant works like RetinaNet [19] and Mask RCNN [20] also adopt the same structures as the baseline networks for better detection results. Recent works on feature fusion method HRNet [21] maintains high-resolution feature maps, gradually fuse high-to-low resolution sub-networks in parallel architecture to obtain

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

rich representations. PANet [22] makes additional bottom-up path augmentation to shortens the transmission path of the feature information.

Compared with natural images, aerial images have many unique characteristics, such as small and densely distributed objects, which takes a large proportion in a single image. Because small objects have few pixels, which makes an unfavorable condition to the representation of feature information. Therefore, there is a hard difficulty for both classification and localization when detecting small objects and the detection results are still far from satisfactory. Our method combines different feature maps by adopting a dense connection to fully exploit the features of each layer, which can bring more contextual information and keep the small extra cost of computation.

3. Proposed Method

In this section, we will detail each part of the proposed work, which mainly include SEMLPCN for the first stage and Cascade architecture for the second stage. Specifically, SEMLPCN efficiently utilizes the feature information of each layer to generate the feature map that has a more powerful semantic representation. Next, getting proposals from the region proposal networks (RPN) for the second stage. Then, high-quality regression and classification of proposals are processed by Cascade R-CNN [23]. Finally, we get the detection result.

3.1 Smooth Multiscale Feature Fusion Pyramid Networks

As we all know, the low-level location information and high-level semantic information attach equal importance to object detection, especially to the aerial images detection, which has more small objects. FPN is an effective method that fusing multi-level information from low and high-level feature maps via the bottom-up, top-down, and lateral connection.

Since some objects like bicycle, tricycle and motor, both of them are generally smaller size compared with others. Besides, there are many similar classes such as car and van in aerial images, which cannot be distinguished effectively. Moreover, the complexity of background increases the

difficulty of the recognition process, for example, there are lots of vehicles-like disturbances due to the large scene of aerial images such as the shape of the roof, which is likely to confuse.

To solve the problems mentioned above, we propose a novel feature fusion method that makes a dense connection between Top-down pathway and bottom-up pathway. Figure 1 shows the architecture of SEMLPCN based on ResNets [24], the details are as follows.

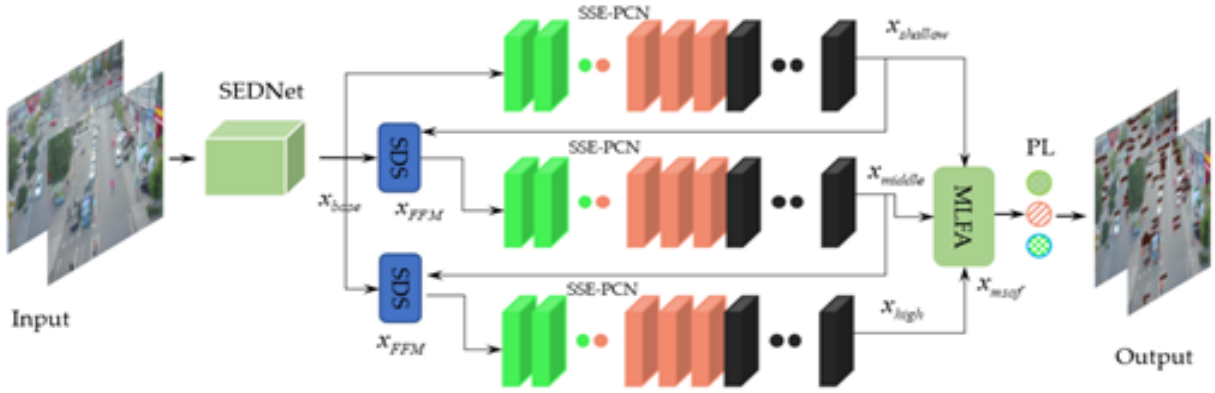

Figure 1. SEMLPCN includes dense connection parts

Feature Pyramid Networks (FPN). We use the feature maps of the bottom-up pathway as $\{C2, C3, C4, C5\}$, which is from the last feature maps of each residual block with the strides {4,8,16,32} pixels. In top-down architecture, the pyramid layers are described as $\{p2, p3, p4, p5\}$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

Dense Multiscale Feature Fusion. The upper part described in figure 2 is the dense connection from the bottom-up pathway. We integrate all the valid and non-redundant connections to$\{P2, P3\}$before finally predicting, note that we don't make connections to$\{P4, P5\}$, because there are none of the available connections from $\{C2, C3, C4, C5\}$, The specific definition is as follows.

$$\begin{array}{ccccccccc} Conv_{31}^{1} \searrow & & \nearrow Conv_{31}^{2} \searrow & & \nearrow Conv_{31}^{3} \searrow & & \nearrow Conv_{31}^{4} \searrow & \\ Conv_{32}^{1} \rightarrow & \varepsilon_3^1 \longrightarrow & Conv_{32}^{2} \rightarrow & \varepsilon_3^2 \longrightarrow & Conv_{32}^{3} \rightarrow & \varepsilon_3^3 \longrightarrow & Conv_{32}^{4} \rightarrow & \varepsilon_3^4 \\ Conv_{33}^{1} \nearrow & & \searrow Conv_{33}^{2} \nearrow & & \searrow Conv_{33}^{3} \nearrow & & \searrow Conv_{33}^{4} \nearrow & \end{array} \quad (1)$$

$$P_5^* = Conv_{3*3}[Conv1 * 1(C_5)] \quad (2)$$

$$P_4^* = Conv_{3*3}[Conv1 * 1(C4) + Upsample(P_5)] \quad (3)$$

$$P_i^* = Conv_{3*3}\{\sum_{j=i+1}^{5} Upsample[Conv1 * 1(C_j)] \oplus P_i\} \quad (4)$$

where $Convk * k$ represents the operation of convolution, which adopts kernel size 1×1 or 3×3. Upsample represents the up-sampling operation, in this paper, we apply bilinear up-sampling in all experiments. $\oplus$is the operation of concatenation. $P_i^*$is the final prediction of all fused feature maps from C_j and P_i. So we can finally get $\{P_2^*, P_3^*, P_4^*, P_5^*\}$.

3.2 Loss Function

The multi-task loss is used and defined as follows.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{cls}(t_i, t_i^*) \quad (5)$$

Here,L_{cls} and L_{reg} respectively represents the loss function of classification and bounding box regression at each stage t. Note that, softmax cross-entropy is defined as the classification loss, bounding box regression loss is defined as smooth L1 loss which follows the setting in Faster R-CNN.T is the total number of the cascaded stage. λ controls the balance between the different task.

3.3 SEDNet

In order to further reduce the operating parameters of the backbone network and improve the multi-scale feature extraction capabilities of the backbone network, the Sequence Squeeze-and-Excitation (Sq-SE) is embedded in each densely connected block of DenseNet. Build a fast and powerful feature extraction backbone network, which can be called SEDNet. The structure of this backbone network is shown as in Fig. 2.

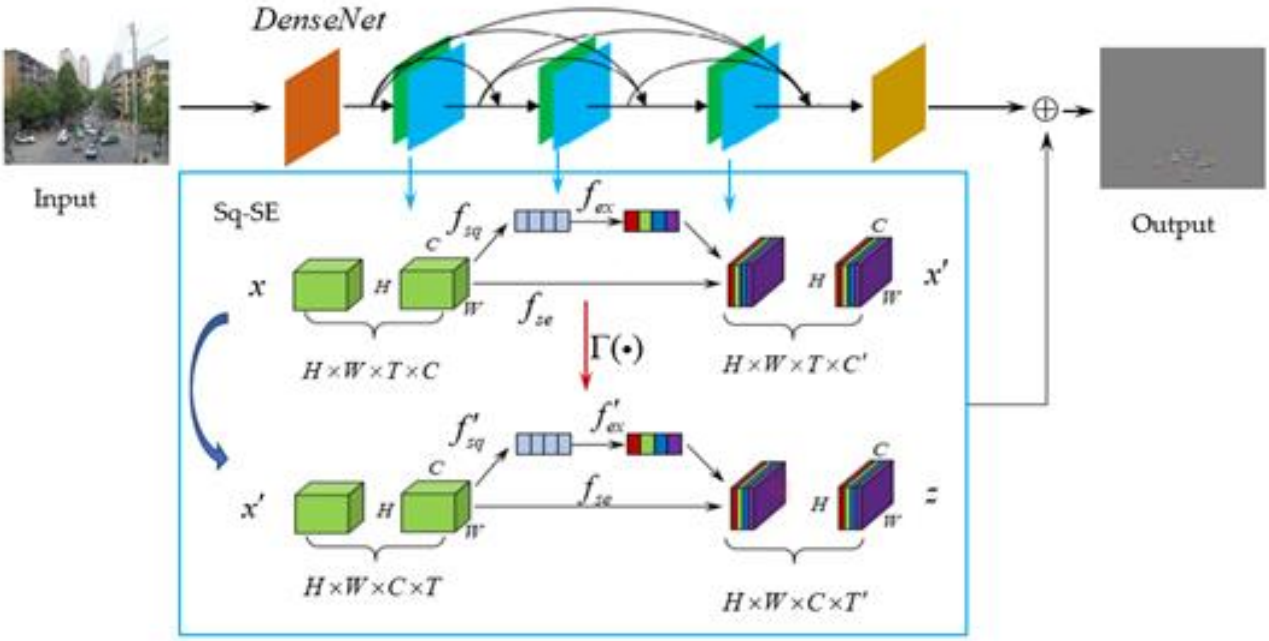

Figure 2. The structure of SEDNet

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

The Sequence Squeeze-and-Excitation (Sq-SE) module enhances channel interconnection through low computational cost to improve feature representation, strengthens the interdependence between channels, and thus improves the sensitivity of the network to information. Capture the global information of the image, before data conversion, use squeeze and excitation to reset the filter to achieve more information acquisition.

Different from the previous SE module, we not only consider the channel (channel), but also include the capture of sequence information, that is, the channel and sequence information are embedded in each layer of the network, so that the subsequent network layer can obtain more effective detailed information. For multi-scale structure information, the squeeze excitation operation will be performed again on each sequence SE block of the channel, that is, the squeeze excitation aggregation operation. The formula is shown below.

$$\begin{cases} \tilde{x}_c = f_{se}(v_c, t_c) = t_c v_c \\ t_c = f_{ex}(T, w) \end{cases} \tag{6}$$

Among them, $f_{se}()$represents the basic multi-scale SE feature.

In short, in order to further improve the utilization of the information obtained from the DenseNet structure, this chapter embeds the designed sequence squeeze excitation module (Sq-SE) into the densely connected block, and re-fuses the multi-scale structure information. Compared with the original densely connected convolutional network, SEDNet can simplify the network structure and increase the speed of the feature extraction module. At the same time, it captures more diverse structural information and strengthens the nonlinearity and generalization capabilities of the backbone network.

3.4 Smooth scale equalization pyramid convolution

In the process of capturing the multi-scale pyramid convolution features of the smooth pyramid convolution network, the size of the convolution kernel always remains the same as the scale increases, but the size of the feature map is gradually decreasing, which is structurally similar to the Gaussian pyramid. Certainly similar, that is, deploying the smooth pyramid convolution kernel designed in this paper on the Gaussian pyramid can also capture the features of the Gaussian pyramid, that is, the smooth pyramid convolution can capture the invariance of pedestrians and vehicles from the Gaussian pyramid structure. Multi-scale structure information.

In short, when the smooth pyramid convolution network processes high-order spatio-temporal feature maps, it will predict the deformation offset based on the current spatio-temporal feature layer, so that the deformable convolution operation can be used for each The multi-scale spatiotemporal features obtained by the smooth pyramid convolution are equalized. In the subsequent transfer of feature information, the shared smooth pyramid convolution will continue to be operated. It is worth noting that in the process of capturing the bottom multi-scale spatiotemporal features of pedestrian vehicles, the smooth scale-balanced pyramid convolution kernel is fixed at a size of 3*3. When N=1, the smooth scale equalization pyramid convolution operation process is shown in formula 7.

$$\begin{cases} x' = \sum_q G(p,q) * x(q) \\ y^{(l)} = BN(Upsample(w_1 x'^{(l+1)})) + BN(w_0 x'^{(l)}) + BN(w_{-1}(* s\_2)x'^{(l-1)}) \end{cases} \tag{7}$$

In Equation7, G () represents the bilinear interpolation operation; * represents the dot multiplication operation; q represents the possible offset of p.

3.5 Aggregation of multi-level features

In order to better capture the multi-scale and spatial feature information of pedestrians and vehicles in the image, and to describe the pedestrian and vehicle targets in the image in detail from different angles and levels, the multi-scale spatial structure features extracted by the smooth-scale balanced pyramid convolution of different levels The information is aggregated to form a multi-level pyramid convolution feature with a balanced scale. The aggregation process uses bidirectional attention as the guidance of different levels of features, that is, using dual attention networks (Dual attention, DA) to

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

refine the features of different levels to form the final refined multi-level features. It is worth noting that the use of the bidirectional attention network as the guiding aggregation layer to gradually refine the multi-level feature information captured by the smooth scale equalization pyramid convolution module can further weaken the irrelevant redundant information, and at the same time, it is also conducive to The multi-scale and spatial features of pedestrians and vehicles in the image are locally and globally coded to obtain more effective spatial context semantics. Suppose that the smooth scale equalization pyramid convolution module outputs three levels of multi-scale and spatial structure information, and they are defined as $x_{shallow}, x_{middle}, x_{high}$,Through the Dual attention-oriented layer, different levels of aggregate information can be obtained x_{msaf}, The aggregation operation can be expressed as Equation 8.

$$x_{msaf} = MLFA(FC(x_{shallow}), FC(x_{middle}), FC(x_{high})) \quad (8)$$

In formula 8, MLFA () represents the use of Dual attention-oriented layer; FC() represents a fully connected operation. The purpose of this operation is to force the feature information of different levels to the same dimension.

4. Experiments and Results

4.1 Datasets and Evaluation Criteria

4.1.1 Drone-based Datasets

VisDrone, Drone-based Datasets, is collected by the AISKYEYE team, Tianjin University, China. The benchmark dataset focuses on four core problems, i.e., object detection in images, object detection in videos, single object tracking, and multi-objects tracking. In this paper, we are mainly aimed at object detection in images, which include 10,209 static images(6,471 images used for training, 548 images for validation and 3,190 images for testing), 54.2k labels and 10 common objects (car, van, bus, pedestrian, tricycle, etc.) are involved.

4.1.2 Evaluation Criteria

Due to VisDrone has its own evaluation method, in this paper, we compare our method with state-of-art algorithms by using the corresponding criteria. Following the criteria of MS COCO, VisDrone uses AP0.5:0.95, AP0.5, AP0.75, AR1, AR10, AR100 and AR500 metrics to evaluate the detection results. Specifically, AP0.5:0.95 is computed by making an average value of 10 IoU thresholds from 0.5 to 0.95 with the step size 0.05. AP0.5 and AP0.75 are computed in a single IoU threshold 0.5 and 0.75. Note that, The max detections per images are 500, which is different from the metrics of MS COCO. Moreover, AR1, AR10, AR100 and AR500 are the maximum recalls of 1, 10, 100, 500 objects per images.

4.2 Implementation Details

All of the results of our experiments are performed on the validation set of VisDrone2019. Firstly, Tensorflow 1.12 is used as deep learning framework, and ResNets-101 is the pre-training model to initialize the network. Besides, short side of input images is resized to 800 pixels, Momentum Optimizer is selected as the optimizer, weight decay is 0.0001 and Momentum is set to 0.9. Meanwhile, we sample 256 batch size of anchors with positive-to-negative samples 1:1 at RPN stage and set the batch size of RoIs to 512 at Fast R-CNN stage where the ratio of positive and negative samples is 1:3. We train the model on a single GPU (NVIDIA GTX2070 8G) with a learning rate of 0.00125 for the first 70k iterations, 0.000125 for the next 15k iterations. flipping image randomly is used as the data augmentation. Finally, to match the objects in aerial images, we set the base size of the prior anchor to {16, 32, 64,128, 256} and the anchor ratio is {1: 2,1:1, 2 :1}.

4.3 Quantitative analysis results

Compare the proposed algorithm with the other 9 advanced detectors in the above data set. The detection results are shown in Table 1. Mark the best and second-best scores in each evaluation result with red and green respectively.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

As shown in the above table, MR Cascade-RCNN performs well in the data set taken by drones. AP, AP50, AP75 are 34.6, 58.1 and 35.8, respectively, which are better than other algorithms. Analyzing its network architecture, the main reason lies in the multi-layer cascading structure detection part of the framework. It consists of a series of detectors trained with increasing IoU thresholds, making it more selective for approaching false positives. Generally speaking, the detector can be of high quality only when high-quality suggestions are made. The cascaded detection of thresholds can ensure that the thresholds are adaptive in the detection process, so that the network can obtain higher-quality regional suggestions. The overall accuracy is improved by 11% compared to the original Faster RCNN detection accuracy. Compared with the single-stage RetinaNet [59], YOLO, etc., the detection accuracy has been improved by more than 20%.

Table 1. Comparison of accuracy between this method and other algorithms on the VisDrone dataset

	Backbone	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**
One-stage methods							
SSD512	ResNet-101-SSD	15.8	30.8	14.6	7.5	26.2	34.4
YOLOv3	DarkNet-53	13.0	26.5	11.5	5.9	22.2	29.9
RetinaNet	ResNet-101-FPN	16.0	29.0	15.9	6.7	28.5	35.7
Anchor-free method							
FCOS	ResNet-50-FPN	20.2	36.6	19.7	11.0	32.3	38.1
FSAF	ResNet-50-FPN	21.2	42.0	19.1	14.6	29.3	28.1
Foveabox	ResNet-50-FPN	21.3	40.2	20.3	14.6	30.3	28.2
Two-stage methods							
Faster RCNN	ResNet-50-FPN	23.6	46.7	21.1	18.0	32.0	25.8
DH RCNN	ResNet-50-FPN	24.6	47.6	22.5	19.0	33.1	26.8
Grid RCNN	ResNet-50-FPN	24.7	46.0	23.5	18.6	33.3	29.0
MR Cascade-RCNN(ours)	HrNetv2p	**34.6**	**58.1**	**35.8**	**26.3**	**48.0**	**61.2**
SE-MLPC(Our)	**SEDNet**	**39.5**	**60.3**	**41.2**	**28.3**	**50.6**	**63.6**

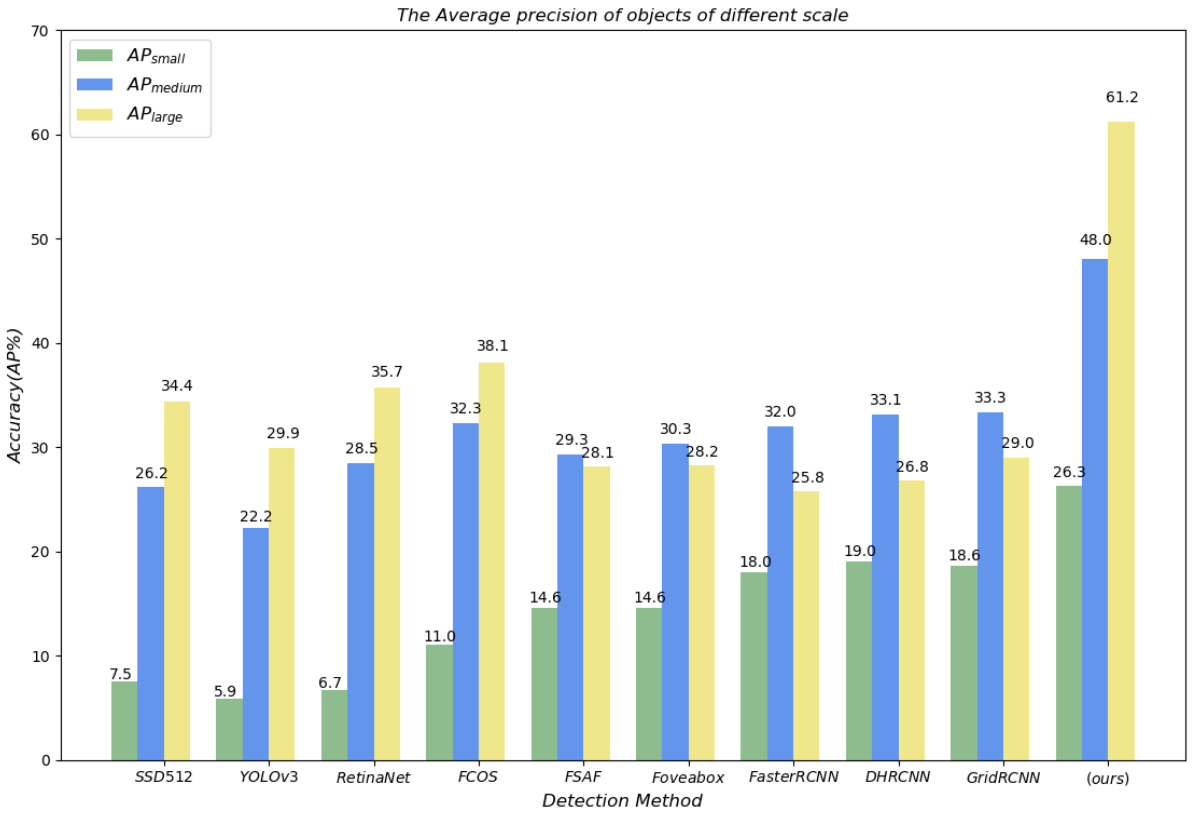

Figure 3. Comparison of target detection accuracy at different scales

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

As shown in the figure 3, it is obvious that compared with other detection methods, our proposed algorithm has a significant improvement in the detection of large, medium and small targets. The main reason for the analysis is HRNetV2p, which has been carried out in the feature extraction stage. The fusion of scale features makes the network robust to targets with various scale changes. Small-scale targets increased by 7.3%, medium-sized targets increased by 14.7%, and large-scale targets increased by 23.1%, reaching a higher detection rate of 61.2%. In addition, the recall rate of this algorithm in the data set detection is compared with the same 9 detection algorithms, which are all higher than other algorithms. It proves that our algorithm is robust to the detection accuracy of positive examples in images.

4.4 Result in each category

We investigate the effect of the dense connection in each category. Cascaded R-CNN with ResNets-101 as the baseline network is constructed. The detection result is listed in Table 2. We can see that the detection accuracy of MR CascadeRCNN is 34.60% to AP0.5:0.95, which is improved by 2.1% points compared with the initial FPN. Furthermore, the result shows that the AP in the most categories is much higher than the initial method, especially in these with smaller size objects such as bicycle, tricycle, and so on. It demonstrates a consistent improvement in the method we proposed.

Table 2. Results in each category. All categories are evaluated in AP0.5:0.95.

category	Car	Motor	Bus	truck	pedestrian
AP	0.648	0.360	0.593	0.405	0.359
category	Bicycle	Tricycle	Van	Awning-tricycle	people
AP	0.242	0.308	0.458	0.197	0.254

The algorithm shows excellent detection accuracy in cars and buses, reaching 0.648 and 0.593 respectively, and the detection rate is greater than 0.4 in vans and trucks. It can be seen that the algorithm is suitable for regular vehicles, especially large vehicles. The detection has reached a more ideal state. There are still shortcomings. For example, the detection accuracy of the sheltered tricycle is 0.197. The reasons for the analysis are as follows. First, in the data set, the sheltered tricycle has fewer target labels. During the road shooting process, it is rarely possible to capture the relevant As a result, there is not enough data to extract its features. Secondly, the addition of sheds brings a lot of negative effects on detection, and problems such as insufficient features have caused the final detection accuracy to be unsatisfactory.

4.5 Qualitative analysis results

In order to more intuitively show the effectiveness and reliability of the SE-MLPC pedestrian vehicle detection framework proposed in this paper, the detection results of different algorithms are given. The test results are shown in Figure 4, 5.

In Figure 4, (a) represents the original input image; (b) represents the detection effect of yolov3; (c) represents the detection effect of RetinaNet; (d) represents the detection effect of SSD.

It can be seen from the figure that the convergence effect of yolov3 is better, there is no large false alarm frame, but the detection effect of small-scale targets and large-scale pedestrian vehicles is poor, and its detection accuracy is also poor.

In Figure 5, (e) represents the detection effect of faster R-CNN; (f) represents the detection effect of Mask R-CNN; (g) represents the detection effect of Cascade R-CNN; (h) represents this The detection effect of SE-MLPC mentioned in the chapter.

It can be seen from the figure that compared to the single-stage pedestrian vehicle detection algorithm, the two-stage detection algorithm has achieved better results, and the regression effect also shows better detection results for small-scale and large-scale pedestrian and vehicle targets. But the accuracy still needs to be improved. At the same time, it can also be clearly seen that the algorithm proposed in this chapter is optimal for both small-scale and large-scale targets.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

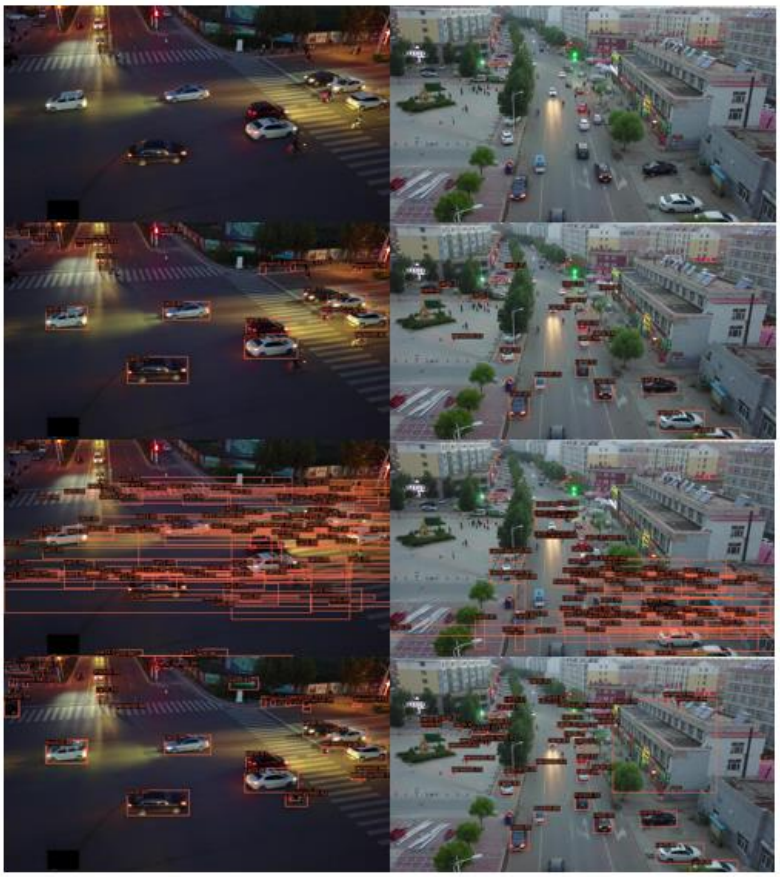

Figure 4. Single-stage pedestrian vehicle detection

Figure 5. Two-stage pedestrian vehicle detection

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

Figure 6 gives the visual result of the proposed framework compared with the ground-truth. It can be seen that our method can effectively process the most objects with different scale and viewpoint.

Figure 6. Visualization result on VisDrone validation datasets. (a) detected results of the proposed framework (b) Ground-truth.

5. Conclusion

In this paper, we have proposed an end-to-end object detection framework, which fully utilizes the multi-scale feature information as much as possible, which is beneficial to detect small objects in aerial images. Meanwhile, the Cascade architecture in the second stage refines the bounding box regression to enhance the localization capability for objects. Experiment on VisDrone datasets demonstrates the effectiveness of the framework proposed and reaches a state-of-the-art performance in object detection for UAV-captured images. Despite achieving good performance in most classes, there are still some issues that categories like awning-tricycle, bicycle, reached a much lower result than others. One possible reason is that the number of its samples is not enough to be trained well and heavy occlusion happened in these categories. We need to explore the methods of resolution in the future.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

References

[1] A. C. Watts, V. G. Ambrosia, and E. A. Hinkley, “Unmanned aircraft 534 systems in remote sensing and scientific research: Classification and con- 535 siderations of use,” Remote Sens., vol. 4, no. 6, pp. 1671–1692, Jun. 2012, 536 doi: 10.3390/rs4x061671.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[3] P. Felzenszwalb, D. McAllester and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587597.

[4] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014.

[5] R. Girshick, “Fast R-CNN,” international conference on computer vision, 2015..

[6] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, pp. 91-99, 2015.

[7] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You only look once: Unified real-time object detection,”? Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, December 2016..

[8] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,”?2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6517-6525, doi: 10.1109/ CVPR. 2017. 690

[9] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv 2018, arXiv:1804.02767.

[10] A. Bochkovskiy, Y.W. Chien, M. L. Hong, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv 2020, arXiv:2004.10934.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SSD: Single shot multiBox detector,” in Proc. ECCV, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[12] Y. F. Cheng, L. Wei, R. Ananth, T. Ambrish, C. Alexander, DSSD: Deconvolutional Single Shot Detector. arXiv 2017, arXiv:1701.06659.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The pascal visual object classes (voc) challenge,” IJCV, 2010.

[14] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., “Microsoft COCO: Common Objects in Context,” ECCV, 2014.

[15] D. Du et al., “VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results,” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 213-226, doi: 10.1109/ICCVW.2019.00030.

[16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, “Feature pyramid networks for object detection,” arXiv 2016, arXiv:1612.03144.

[17] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, H. Ling, “Vision Meets Drones: A Challenge,” arXiv 2018, arXiv: 1804. 07437.

[18] J. Dai, Y. Li, K. He, J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” arXiv 2016, arXiv:1605.06409.

[19] T Y Lin, P Goyal, R Girshick et al., "Focal loss for dense object detection", IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. PP, no. 99, pp. 2999-3007, 2017.

[20] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN", Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 2980-2988, Oct. 2017.

[21] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, et al., "High-resolution representations for labeling pixels and regions" in arXiv:1904.04514, 2019.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0071

[22] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 8759-8768, doi: 10.1109/CVPR.2018.00913.

[23] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2019.2956516.

[24] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[25] Z. Li, C. Peng, G. Yu, "Detnet: A backbone network for object detection," arXiv 2018, arXiv:1804.06215.

[26] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li, "Single-Shot Refinement Neural Network for Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 4203-4212, doi: 10.1109/CVPR.2018.00442.