An Anchor-free Detector Based on Residual Feature Enhancement Pyramid Network for UAV Vehicle Detection

Jianghuan Xie, Dianwei Wang*, Pengfei Han, Jie Fang
School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Jiaxing Guo*
School of Electrical Engineering, Xidian University, Xi'an 710071, China

Zhijie Xu
School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK

ABSTRACT

Vehicle detection in Unmanned Aerial Vehicle (UAV) images is a challenging task because there are many small objects in UAV images, and the scale of objects varies greatly, which brings great difficulty to vehicle detection using existing algorithms. This paper proposes an anchor-free detector called Residual Feature Enhancement Pyramid Network (RFEPNet) for UAV vehicle detection. RFEPNet contains a Cross-Level Context Fusion Network (CLCFNet) and a Residual Feature Enhancement Module (RFEM) based on pyramid convolution. Specifically, CLCFNet utilizes the densely connected structure and Dual Attention Fusion Module (DAFM) to increase the sensitivity of high-resolution feature maps to small objects. Simultaneously, RFEM exploits pyramid convolution and residual connection structure to enhance the semantic information of the feature pyramid. In addition, the anchor-free head is used for classification and bounding box regression. The experimental results on the UAVDT dataset show that the proposed RFEPNet achieves state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**; • **Object detection**;

KEYWORDS

UAV vehicle detection, Cross-Level Context Fusion Network, Residual Feature Enhancement, Anchor-free

ACM Reference Format:
Jianghuan Xie, Dianwei Wang*, Pengfei Han, Jie Fang, Jiaxing Guo*, and Zhijie Xu. 2021. An Anchor-free Detector Based on Residual Feature Enhancement Pyramid Network for UAV Vehicle Detection. In *2021 4th International Conference on Artificial Intelligence and Pattern Recognition (AIPR 2021), September 24–26, 2021, Xiamen, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3488933.3488936

AIPR 2021, September 24–26, 2021, Xiamen, China

ACM ISBN 978-1-4503-8408-7/21/09…$15.00
https://doi.org/10.1145/3488933.3488936

1 INTRODUCTION

Unmanned Aerial Vehicles (UAVs), with the advantages of high mobility and rapid deployment, have been widely used in various fields, such as traffic surveillance [1] and disaster rescue [2], etc. As an important problem in computer vision, vehicle detection in UAV images has received significant attention. Compared with the common detection task, UAV vehicle detection tasks have many challenges. Firstly, UAV images contain many objects with large scale-variation, such as large trucks and small vehicles, which leads to weak feature representation. It is difficult for existing methods to learn those multi-scale features; Secondly, UAV images usually cover a larger area and have many small objects with low resolution. These small objects are often difficult to detect; Thirdly, to meet high detection accuracy and high-speed performance requirements of real-time operation, it is necessary to reduce the size and parameters of the detection model.

Traditional UAV object detection methods usually extract image features manually, and then the classifiers are used to identify the objects. However, hand-crafted features have weak representation power and poor generalization ability. In recent years, object detection methods based on the convolutional neural network (CNN) [3] have gradually emerged and been widely used in UAV object detection due to the strong feature representation power and better detection performance. Although CNN-based object detection methods [4, 5] show good performance in general object detection tasks, they don't perform very well in detecting large scale-variation targets in high-resolution UAV images. This is due to the difference between general images and UAV images.

To address this problem, we propose a novel anchor-free detector called Residual Feature Enhancement Pyramid Network (RFEPNet). Specifically, we propose a Cross-Level Context Fusion Network (CLCFNet) including the densely connected structure and the Dual Attention Fusion Module(DAFM), which makes the shallow feature maps more sensitive to small objects and fuses the semantic information in different levels according to the channel and spatial importance. At the same time, we propose a Residual Feature Enhancement Module (RFEM) based on pyramid convolution [6] and residual connection [7], which can effectively deal with the objects with large scale-variation. Compared to the currently popular methods, the proposed RFEPNet achieves state-of-the-art performance on the UAVDT dataset, achieving 84.5 AP.

The rest of this paper is organized as follows: the related work is presented in Section 2, the proposed method in Section 3, the

experimental details and results in Section 4, and the conclusions in Section 5

2 RELATED WORK

2.1 Anchor-based Methods

CNN-based object detection methods can be divided into two-stage methods and single-stage methods. The two-stage methods (such as Faster R-CNN [8], FPN [9], Cascade R-CNN [10], etc.) get cursory predicted boxes named proposal through the Region Proposal Network (RPN), thereby filtering out a large number of negative samples. Then the proposals are refined to obtain the final detection result. For single-stage methods (such as YOLO [11], SSD [12], RetinaNet [13], EfficientDet [14], etc.), the detection result of classification and bounding box regression are directly obtained via end-to-end neural networks. Compared with the two-stage detectors, the single-stage detectors eliminate the region proposal stage and focus on classification and regression, so the single-stage detectors have higher computational efficiency and faster detection speed.

2.2 Anchor-free Methods

For the anchor-based methods, it is difficult to match targets with large scale-variation. To eliminate the design of anchors and alleviate the imbalance between positive and negative samples, CornerNet [15] and ExtremeNet [16] directly detect the key points of objects to get the bounding boxes. FCOS [17] predicts the classification confidence and location information at each coordinate. With continuous improvement, anchor-free methods have surpassed anchor-based methods in detection precision.

2.3 Object Detection in UAV Images

CNN-based object detection methods in natural images have achieved unprecedented achievements, which brought great inspiration to the counterparts in UAV images. Recently, UAV Vehicle detection methods based on CNN are emerging, such as FPN-based methods [18, 19], YOLO-based methods [20, 21], etc. . Inspired by anchor-free methods, Yang et al. [18] adopt the idea of per-pixel prediction in FCOS [17] to solve the imbalance between positive and negative samples in UAV images. However, merely introducing CNN-based detectors into UAV vehicle detection will not significantly improve the detection performance because the characteristics of UAV images are not considered. In this paper, a novel anchor-free detector called Residual Feature Enhancement Pyramid Network (RFEPNet) is proposed to deal with the above problems.

3 PROPOSED METHOD

The overall architecture of the proposed network is shown in Figure 1, which mainly consists of the backbone(Figure 1(a)), the Cross-Level Context Fusion Network(Figure 1(b)), the Residual Feature Enhancement Module(Figure 1(c)), and anchor-free head(Figure 1(d)). In the beginning, we use ResNet50 [7] as the backbone for extracting the feature maps. Then, the extracted feature maps are fused by CLCFNet. Next, the output of CLCFNet undergoes RFEM to obtain the enhanced multi-level features. Finally, we adopt the anchor-free head to get bounding boxes of UAV vehicles. The details of the proposed model and loss functions will be explained in the rest of this section.

3.1 Cross-Level Context Fusion Network

To achieve better performance on small objects, FPN [9] builds a top-down pathway to improve the semantic information of shallow feature maps, as shown in Figure 2(a). However, it cannot enhance semantic information according to the importance of features. We propose a densely connected cross-level structure and a fusion method based on channel and spatial attention mechanism, constructing a Cross-Level Context Fusion Network (CLCFNet) to solve the above problems.

Figure 2(b) illustrates a detailed structure of the proposed CLCFNet. In Figure 2, $C_i (i = 2, 3, 4, 5)$represents the feature maps extracted from the backbone. Firstly, we adopt a lateral 1×1 convolution to reduce the channels of these feature maps to 256, as in Equation 1). After that, we assume a densely connected cross-level structure to build a top-down path, which connects deep features to shallow features across levels. It is worth noting that we did not use the element-wise sum to process the upsampled feature map and the lateral feature map. Instead, we use a fusion method based on the channel and spatial attention mechanism, called Dual Attention Fusion Module (DAFM).

As in Equation(2), the element-wise sum of the upsampled feature maps and the shallow feature maps are sent to DAFM together.

$$L_i = Conv_{1\times1}(C_i) \tag{1}$$

$$T_i = DAFM\left(\sum_{j=i+1}^{5} UpSampling\left(T_j\right), L_i\right) \tag{2}$$

$$P_i = Conv_{3\times3}(T_i) \tag{3}$$

DAFM includes two sequential processes: Channel Attention Fusion Block (CAFB) and Spatial Attention Fusion Block (SAFB).

As shown in Figure 3(a), the CAFB firstly cascades the two input feature maps, and a global average pooling layer is used to squeeze the feature map into a one-dimensional vector. Then, two fully connected layers are used to obtain the channel attention weights of two feature maps. Finally, the channel-wise multiplication of the two inputs and their respective weights is added to the two inputs to get the channel-sensitive feature maps.

Figure 3(b) shows the structure of SAFB. It takes the channel-sensitive feature maps as the inputs, replaces FC layers in CAFB with 1×1 convolution layers to get the space weights, and uses softmax to limit the weights. As a result, the two inputs are multiplied by their respective weights in the form of the spatial-wise product and add up to get the output.

To reduce the aliasing effect of upsampling, we append a 3×3 convolution on the merged feature map to obtain the final feature maps with 256 channels, as in Equation 3). The above process can be summarized as Equation 4).

$$P_i = Conv_{3\times3}\left[T_i = DAFM\left(\sum_{j-i+1}^{5} Upsample(T_j), Conv_{1\times1}(C_i)\right)\right] \tag{4}$$

Finally, a 3×3 convolution with the stride of 2 is applied to C_5 to get a high-resolution feature mapP_6 to adapt small objects, so CLCFNet outputs $P_i (i = 2, 3, 4, 5, 6)$.

The densely connected structure in CLCFNet propagates the rich semantic information from deep layers to shallow layers, providing

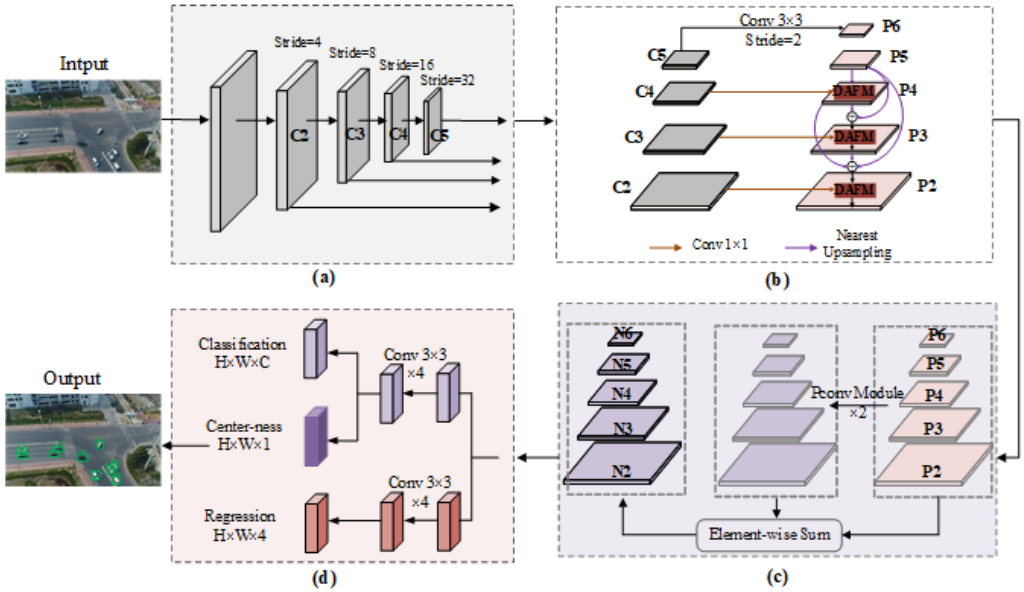

Figure 1: The overall architecture of the RFEPNet, including (a) Backbone, (b) Cross-Level Contextual Fusion Network, (c) Residual Feature Enhancement Module, and (d) Anchor-free head.

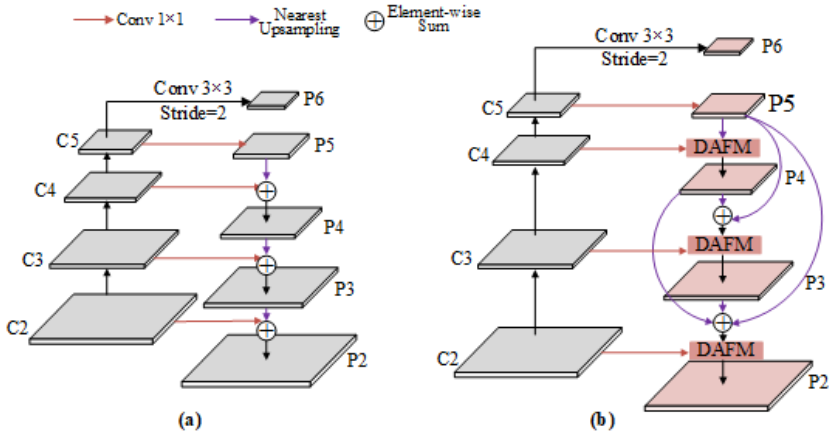

Figure 2: Comparison between FPN and CLCFNet: (a) The structure of FPN, (b) The structure of CLCFNet.

better classification and localization features for small objects. At the same time, the DAFM is used to fuse features at different levels according to the information importance in channel and space, which can effectively enhance the useful information and suppress invalid information.

3.2 Residual Feature Enhancement Module

Although the top-down pathway in FPN can effectively fuse semantic information of the deep features to the shallow layers, it does not take full advantage of more detailed information in shallow features beneficial for locating objects. In an attempt to solve the above difficulty, inspired by pyramid convolution (Pconv) [6], we propose a Residual Feature Enhancement Module (RFEM) to fuse the features of different levels. Specifically, RFEM contains two pyramid convolution modules and a residual connection structure, as in Figure 4.

The pyramid convolution module is indeed a 3-D convolution across both scale and spatial dimensions. The pyramid convolution can be represented as N different 2-D convolutional kernels. We set different strides for the N different kernels when convolving in different layers. In this paper, N is set to 3. Then the output of the pyramid convolution module is as Equation 5).

$$Pconv(x^i) = w_1 \otimes_{s0.5} x_{i+1} + w_0 \otimes x_i + w_{-1} \otimes_{s2} x_{i-1}\, i = 2, 3, 4, 5, 6 \quad (5)$$

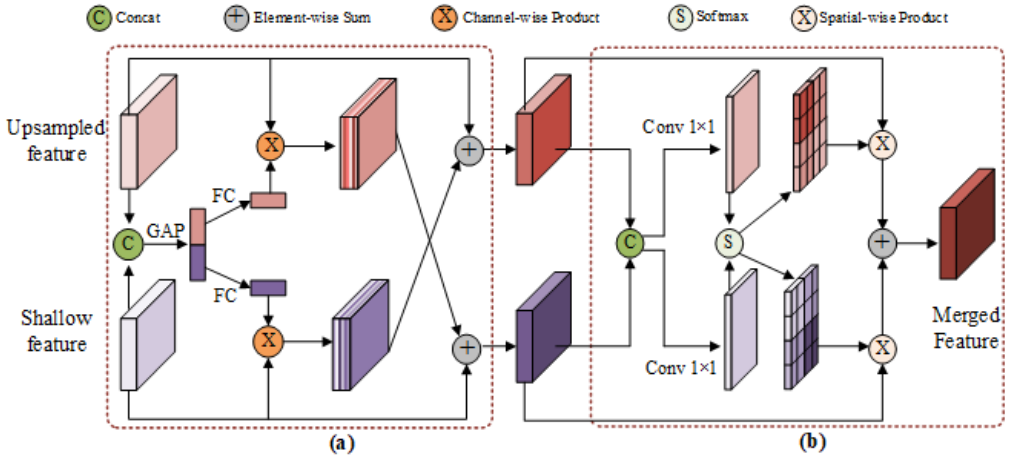

Figure 3: The structure of DAFM. (a) CAFB, (b) SAFB. The upsampled features and shallow features are sequentially fed into CAFB and SAFB to obtain the fused feature maps. FC is the fully connected layer, and GAP is the global average pooling layer.

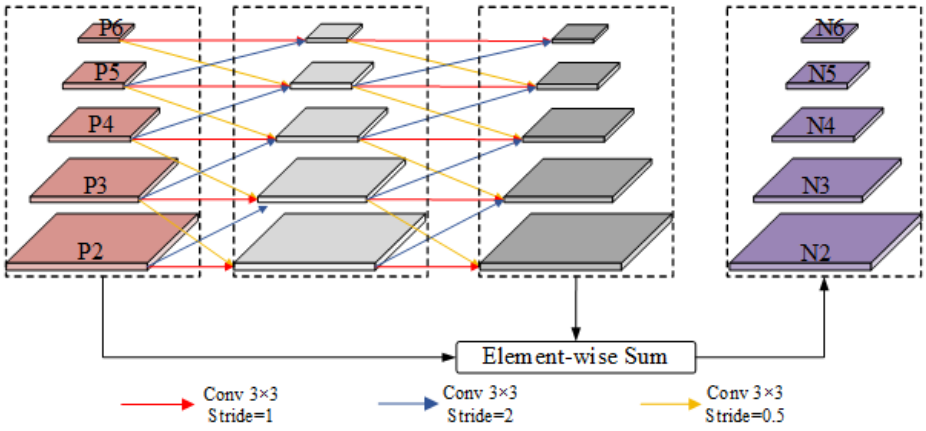

Figure 4: The structure of RFEM.

where $i(=2,3,4,5,6)$ denotes the current feature level, x_i is feature map of the i-th level, w_0, w_{-1} and w_1 are three independent 2-D convolutional kernels and $\otimes_{s2}$ means convolution with the stride of 2.

As shown in Figure 4, the red, orange, and blue arrows represent 3×3 convolutions with the stride of 1, 0.5, and 2, respectively. The convolution with the stride of 0.5 is achieved by bilinear upsampling and a normal convolution. As for the bottom pyramid level (i=2), the last term in Equation 5) is unnecessary, while the first term is ignored for the top-most level. In this paper, two repeated pyramid convolution modules are used for multi-scale information fusion. In addition, considering that the enhanced features may lose some of the original features that are beneficial to prediction, we introduce a residual connection structure to preserve the original beneficial features, as shown in Equation 6).

$$N_i = x_i + Pconv\left(Pconv\left(x_i\right)\right) \quad i = 2,3,4,5,6 \tag{6}$$

3.3 Anchor-free Head

In this paper, we introduce the anchor-free network FCOS into UAV vehicle detection. FCOS is a fully convolutional single-stage detection network, and it can output the predicted result in a per-pixel manner. The prediction layer will output classification confidence and a four-dimensional vector at each position. The four-dimensional vector represents the distance between the current position and the boundaries of the ground truth, as is defined in Equation 7).

$$t^*_{x,y} = \left\{l^* = x - x_t, t^* = y - y_t, r^* = x_b - x, b^* = y_b - y\right\} \tag{7}$$

where $(x_t\ , y_t)$ and $(x_b\ , y_b)$ represent the coordinates of the top-left and the down-right of the ground truth box, respectively.

Like FCOS, we introduced the Center-ness branch to suppress the predicted bounding boxes far away from the center. Center-ness

Table 1: Effects of the proposed modules

FCOS	CLCFNet	RFEM	AP	AP50	AP75	APs	APm	APl
√			82.6	98.9	95.5	74.0	87.4	92.0
√	√		83.6	98.9	95.9	76.3	87.9	92.2
√		√	84.2	98.9	96.3	76.6	88.5	92.8
√	√	√	**84.5**	**98.9**	**96.4**	**76.9**	**88.9**	**92.9**

Table 2: Performance comparisons of the state-of-the-art methods on the UAVDT Dataset

	Method	AP	AP50	AP75	APs	APm	APl	Time(ms)
Two-stage	Faster R-CNN C4 [8]	79.0	98.5	92.4	68.8	85.6	91.1	86
	Faster R-CNN FPN [9]	83.1	98.3	95.6	75.9	87.7	92.3	92
	PANet [24]	83.0	98.0	95.6	75.0	87.9	92.2	98
	Cascade R-CNN [10]	84.3	98.7	96.1	76.6	88.8	**94.2**	105
Single-stage	YOLOv3 [11]	77.8	98.5	93.2	66.7	84.0	84.8	**34**
	SSD [12]	78.8	98.3	92.4	69.3	84.7	88.6	50
	RetinaNet [13]	82.0	98.3	93.6	72.6	87.6	92.3	76
	Free-anchor [25]	81.7	95.7	94.4	73.3	87.3	93.1	77
Anchor-free	FSAF[26]	82.4	**99.0**	95,4	75.2	87.1	91.9	69
	Reppoints [27]	81.3	98.9	94.2	71.4	86.5	92.2	78
	FCOS [17]	82.6	98.9	95.5	74.0	87.4	92.0	68
	RFEPNet (ours(	**84.5**	98.9	**96.4**	**76.9**	**88.9**	92.9	76

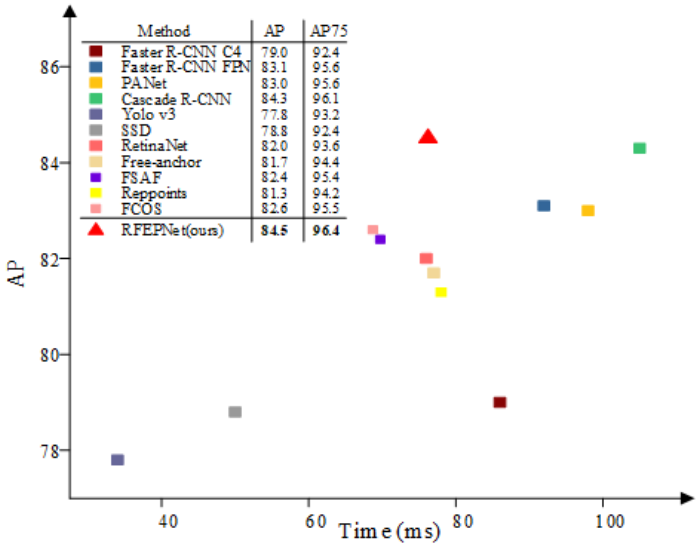

Figure 5: Accuracy (AP) vs. speed (ms) comparison.

is defined as Equation 8).

$$\text{Center} - ness = \sqrt{\frac{\min(l,r)}{\max(l,r)} \times \frac{\min(t,b)}{\max(t,b)}} \tag{8}$$

3.4 Loss Function

We applied the supervised signals to the three branches of the FCOS head during the training phase, including classification, regression, and Center-ness. The total loss is calculated in a per-pixel manner, as in Equation 9).

$$\begin{aligned} L\left(p_{x,y}, t_{x,y}, cn\right) = & \frac{1}{N_{pos}} \sum_{(x,y)} L_{cls}\left(p_{x,y}, c^*_{x,y}\right) \\ & + \frac{1}{N_{pos}} \sum_{(x,y)} c^*_{x,y} L_{cn}\left(cn, cn^*\right) \\ & + \frac{1}{N_{pos}} \sum_{(x,y)} c^*_{x,y} L_{reg}\left(t_{x,y}, t^*_{x,y}\right) \end{aligned} \tag{9}$$

where $c^*_{x,y}$ is the category of ground truth, $p_{x,y}$ is the predicted category, $t_{x,y}$ is the four-dimensional vector used to represent the

bounding box, and *cn*denotes center-ness. All characters marked with * indicate that the value is calculated by ground truth. In addition, $N_{pos}$is the number of positive samples, L_{cls} is classification loss, which uses focal loss, L_{reg} is regression loss, which uses IoU loss in this paper, and L_{cn} is Center-ness loss, which adopts cross-entropy loss. The total loss is the summation of the losses of all levels$N_i (i = 2, 3, 4, 5, 6)$.

4 EXPERIMENTS AND RESULTS

4.1 Experimental Datasets and Settings

4.1.1 Dataset and Metrics. This paper uses a large-scale UAV object detection and multi-target tracking public dataset UAVDT [22] to train and test the network, which includes 40409 images with a total object number of 798795. We randomly divided the dataset into the training set and test set, in which 24245 images are used for training and 16164 images for test.

The metrics of the MSCOCO dataset [23] are used to evaluate our detection results, including AP, AP50, AP75, APs, APm, and APl, which provide a rigorous evaluation of the detection results from different perspectives.

4.1.2 Training Details. In this paper, 4 × NVIDIA TITAN Xp (11GB) GPUs are used to train our detector with a batch size of 8 (2 images per GPU). We resize the short side of the input image to 800 and the long side to 1333. Our network is optimized by SGD, the learning rate is set to 0.01, the weight decay is set to 0.0001, and the momentum is set to 0.9. We train the network for 12 epochs, and after the 8th and 11th epochs, the learning rate is reduced by ten times.

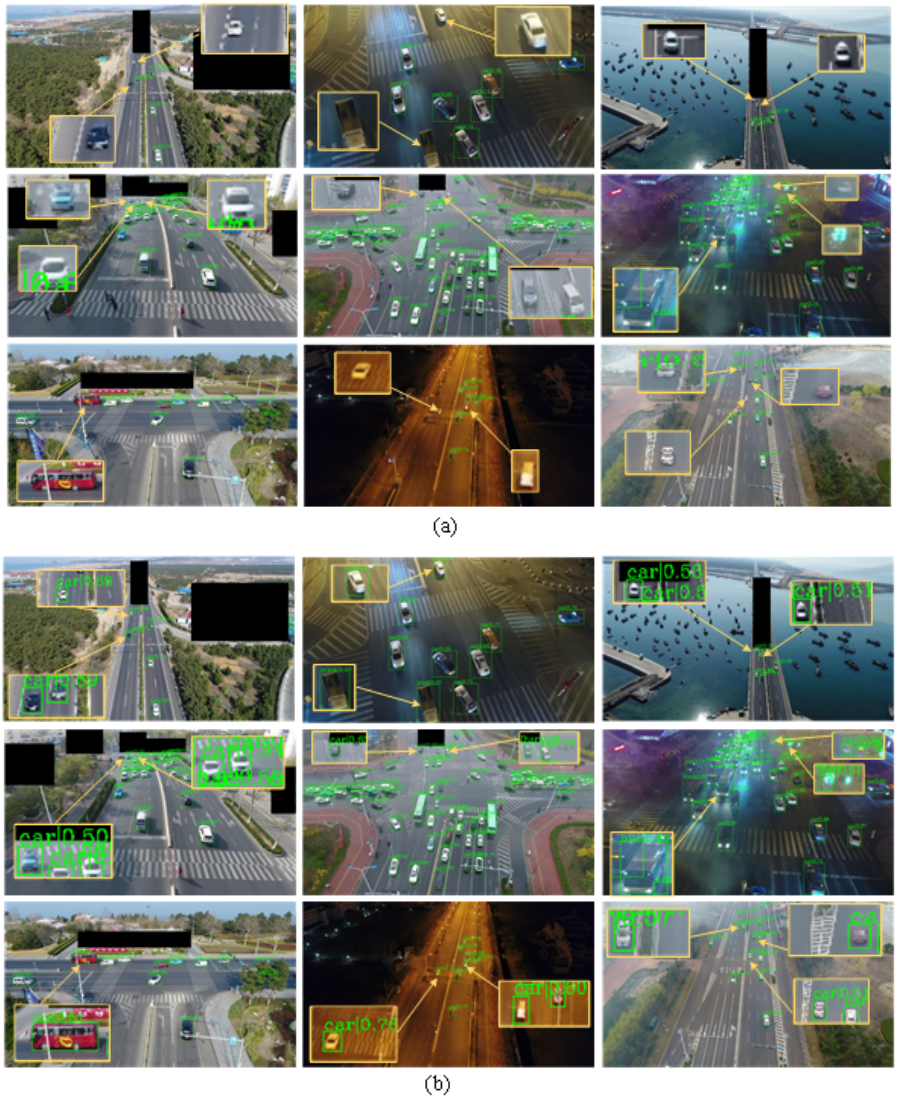

Figure 6: Comparison results on the UAVDT dataset. (a) Results of FCOS, (b) Results of our method. The black regions represent the ignored areas in the UAVDT dataset.

4.2 Ablation Experiments

We conducted some ablation experiments on the UAVDT dataset to study the effectiveness of the main modules of RFEPNet. All the experiments were based on FCOS, and ResNet50 [7] is used as the backbone network.

As shown in the 2nd row of Table 1, after replacing the feature fusion method of FPN with the CLCFNet module, the AP reaches 83.6, surpassing the FCOS baseline (82.6 AP). It is worth noting that for APs of small vehicles, CLCFNet obtains a significant gain of 1.7% compared to the baseline, which proves CLCFNet provides better spatial location information and classification features for small targets. The 3rd row of Table 1 shows when RFEM is added to the FCOS baseline, the AP reaches 84.2. Specifically, due to the introduction of RFEM, APs for the small objects is increased by 2.6%, APm for medium objects is increased by 1.1%, and APl for large objects is increased by 0.8%. Finally, combining CLCFNet and RFEM modules, 84.5 AP is achieved. Compared with the FCOS baseline, our method increases AP by 1.9%, increases APs by 2.9%, increases APm by 1.5%, and increases APl by 0.9%.

4.3 Comparisons with State-of-the-art Methods

We compare the proposed method with several state-of-the-art methods, including two-stage methods, single-stage methods, and anchor-free methods. All experiments take ResNet50 as the backbone.

Table 2 shows the results of the comparison experiment. The best performances are highlighted in bold.

Figure 5 shows the detection performance of our method on the UAVDT dataset. It can be seen that RFEPNet outperforms other methods by achieving 84.5 AP, achieving state-of-the-art performance on UAVDT datasets. Compared to Cascade R-CNN, an absolute gain of 0.2% is obtained at AP, and 1.38 times speedup is achieved, demonstrating the superiority of our RFEPNet.

4.4 Visual Analysis

This section shows some detection results of our method and FCOS baseline on the UAVDT dataset.

Figure 6 shows a qualitative comparison between the FCOS baseline and the proposed method. Compared with FCOS (Figure 6(a)), our method (Figure 6(b)) significantly reduces the false-negative rate, especially in the scene with dense vehicles

5 CONCLUSIONS

This paper proposes an anchor-free Residual Feature Enhancement Pyramid Network (RFEPNet) for UAV vehicle detection. Specifically, we design a Cross-Level Contextual Fusion Network, which uses densely connected structure and Dual Attention Fusion Module to improve the sensitivity of shallow features to small objects. Next, we propose a Residual Feature Enhancement Module based on pyramid convolution, which can effectively fuse multi-level features of the object and further maintain the useful features of the original feature pyramid. Finally, the experimental results on the UAVDT dataset show the superiority and practicability of the proposed method.

ACKNOWLEDGMENTS

This research was supported by the Special Project of Strengthening Police with Science and Technology of Ministry of public security under Grant Nos.2019GABJC42.

REFERENCES

[1] <number>[1] Majid Azimi S. ShuffleDet: Real-Time Vehicle Detection Network in On-board Embedded UAV Imagery[C]. Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018: 0-0.
[2] <number>[2] Alotaibi E T, Alqefari S S, and Koubaa A J I A. LSAR: Multi-UAV Collaboration for Search and Rescue Missions[J].2019. IEEE Access, 2019, 7: 55817-55832. https://doi.org/10.1109/ACCESS.2019.2912306
[3] <number>[3] Lecun Y, Bengio Y J T H O B T, and Networks N. Convolutional Networks for Images, Speech, and Time-Series[J].1995. The handbook of brain theory and neural networks, 1995, 3361(10): 1995
[4] <number>[4] Girshick R, Donahue J, Darrell T, and Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 580-587.
[5] <number>[5] Redmon J, Divvala S, Girshick R, and Farhadi A. You Only Look Once: Unified, Real-Time Object Detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.
[6] <number>[6] Wang X, Zhang S, Yu Z, Feng L, and Zhang W. Scale-Equalizing Pyramid Convolution for Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 13359-13368.
[7] <number>[7] He K, Zhang X, Ren S, and Sun J. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.
[8] <number>[8] Ren S, He K, Girshick R, and Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J].2017. IEEE Trans Pattern Anal Mach Intell, 2017, 39(6): 1137-1149. https://doi.org/10.1109/tpami.2016.2577031
[9] <number>[9] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, and Belongie S. Feature Pyramid Networks for Object Detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2117-2125.
[10] <number>[10] Cai Z, and Vasconcelos N. Cascade R-CNN: Delving Into High Quality Object Detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6154-6162.
[11] <number>[11] Redmon J, and Farhadi A J a P A. YOLOv3: An Incremental Improvement[J].2018. arXiv preprint, 2018, arXiv:1804.02767
[12] <number>[12] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, and Berg A C. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision, 2016: 21–37.
[13] <number>[13] Lin T-Y, Goyal P, Girshick R, He K, and Dollár P. Focal Loss for Dense Object Detection[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2980-2988.
[14] <number>[14] Tan M, Pang R, and Le Q V. EfficientDet: Scalable and Efficient Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10781-10790.
[15] <number>[15] Law H, and Deng J. CornerNet: Detecting Objects as Paired Keypoints[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 734-750.
[16] <number>[16] Zhou X, Zhuo J, and Krahenbuhl P. Bottom-Up Object Detection by Grouping Extreme and Center Points[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 850-859.
[17] <number>[17] Tian Z, Shen C, Chen H, and He T. FCOS: Fully Convolutional One-Stage Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9627-9636.
[18] <number>[18] Yang J, Xie X, Shi G, and Yang W. A Feature-Enhanced Anchor-Free Network for UAV Vehicle Detection[J].2020. Remote Sensing, 2020, 12(17): 2729. https://doi.org/10.3390/rs12172729
[19] <number>[19] Wang H, Wang Z, Jia M, Li A, Feng T, Zhang W, and Jiao L. Spatial Attention for Multi-Scale Feature Refinement for Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 0-0.
[20] <number>[20] Liu M, Wang X, Zhou A, Fu X, Ma Y, and Piao C. UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective[J].2020. Sensors, 2020, 20(8): 2238. https://doi.org/10.3390/s20082238
[21] <number>[21] Zhang P, Zhong Y, and Li X. SlimYOLOv3: Narrower, Faster and Better for Real-Time UAV Applications[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 0-0.
[22] <number>[22] Du D, Qi Y, Yu H, Yang Y, Duan K, Li G, Zhang W, Huang Q, and Tian Q. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 370-386.

[23] <number>[23] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, and Zitnick C L. Microsoft COCO: Common Objects in Context[C]. European Conference on Computer Vision, 2014: 740-755.
[24] <number>[24] Liu S, Qi L, Qin H, Shi J, and Jia J. Path Aggregation Network for Instance Segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 8759-8768.
[25] <number>[25] Zhang X, Wan F, Liu C, Ji X, and Ye Q. Learning to Match Anchors for Visual Object Detection[J].2021. IEEE Trans Pattern Anal Mach Intell, 2021, Pp. https://doi.org/10.1109/tpami.2021.3050494
[26] <number>[26] Zhu C, He Y, and Savvides M. Feature Selective Anchor-Free Module for Single-Shot Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 840-849.
[27] <number>[27] Yang Z, Liu S, Hu H, Wang L, and Lin S. RepPoints: Point Set Representation for Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9657-9666.