DOI: 10.54254/2755-2721/2026.34634

Confidence-Driven Region Refinement for Small Object Detection in UAV Image

Leyan Dong

Yangcun No.1 High School, Tianjin, China
dora27.successfall@gmail.com

Abstract. UAV target detection has become increasingly indispensable in the fields of public safety, traffic monitoring, and disaster relief. However, there are still some problems in practical application, such as the difficulty of small target detection and the computational power limitation of edge devices. To tackle the aforementioned challenges, a confidence-driven region adaptive YOLOv5 detection method is developed in this paper. This method detects the global low-resolution of the input image quickly and obtains the rough position and category information of the target. Through the dynamic region selection mechanism driven by confidence, the differentiated reasoning strategy is implemented for the detection regions with different confidence, and only local high-resolution fine detection is carried out for the regions with medium confidence to restore the target details. The non-maximum suppression method is used to fuse the global and local detection results to obtain the final detection output. The experimental results on the VisDrone dataset show that, compared with the baseline YOLOv5 algorithm, the average accuracy of this method is improved by 3.6%, and the detection frame rate is increased from 30 to 35. On the basis of ensuring the real-time detection performance, this method effectively improves the small target detection effect in the UAV scene, and provides a feasible technical scheme for the UAV target detection task on edge devices.

Keywords: confidence-driven, region adaptive reasoning, YOLOv5, small target detection

1. Introduction

In recent years, UAV technology has been widely used in many fields. In urban monitoring, traffic management, disaster relief, and other scenes, UAVs have gradually become an important carrier of information collection with the advantages of flexibility and a broad perspective. As the core task of the UAV vision system, target detection accuracy and reasoning speed directly affect the overall operation effect of the UAV system [1]. The special angle of view of UAV aerial photography leads to the characteristics of small-scale and sparse distribution of targets in the image [2]. Due to the low proportion of target pixels and fuzzy edge details, the traditional target detection algorithm is prone to the loss of effective information in the process of feature extraction, which leads to the decline of detection accuracy [3]. At the same time, most UAV platforms are equipped with edge computing devices, which have limited computing power and storage resources, and the deployment and application of complex depth detection models are greatly constrained [4]. How to achieve high-

DOI: 10.54254/2755-2721/2026.34634

precision real-time detection under the condition of limited computing power has become a problem to be solved in the field of UAV target detection.

Current studies have analyzed the UAV small target detection technology from three aspects: multi-scale feature fusion, image slice detection, and dynamic reasoning optimization. First, by constructing multi-scale detection branches, the capacity of detecting objects across various scales is enhanced [5]. YOLOv5 algorithm further optimizes the feature fusion structure and achieves a more effective trade-off between inference efficiency and detection performance [6]. The development of deep learning target detection provides a theoretical reference for the improvement of multiscale methods [7]. Second, in the direction of image slice detection, existing studies have effectively alleviated the issue of feature information degradation for tiny objects through dividing high-resolution aerial images into several smaller regions for independent detection [8]. This method is easy to introduce duplicate detection and significantly increases the computational overhead [9]. Third, the direction of dynamic reasoning optimization. The faster R-CNN algorithm realizes the dynamic screening of candidate boxes through the regional recommendation network [10].

The existing literature has made some progress in improving the detection accuracy, but there is still a research gap in the real-time optimization under the constraint of computational force. Although the existing multi-scale methods can enhance the performance of detecting small-scale objects, the model computational overhead remains considerable, making deployment on resource-constrained edge platforms challenging. In this paper, the small target detection in a UAV scene is taken as the research object, and a regional adaptive reasoning method based on confidence-driven is proposed. By constructing a joint framework of global low-resolution fast detection and local high-resolution fine detection, the balance between detection accuracy and computational efficiency is achieved. In theory, it enriches the application research of dynamic reasoning strategy in the field of small target detection. In practice, it provides a feasible technical scheme for UAV real-time target detection on edge equipment and has good engineering application value. The main contributions of this paper are as follows: A regional adaptive reasoning framework based on confidence-driven is proposed. By combining low-resolution global detection with local high-resolution fine detection, the balance between detection accuracy and reasoning efficiency is achieved. A dynamic region filtering mechanism is designed to adaptively select the region to be detected according to the confidence of the detection frame, which reduces the computational overhead caused by invalid high-resolution reasoning. The experimental results on the VisDrone dataset show that the proposed method is superior to the baseline YOLOv5 method in small target detection accuracy and real-time performance.

2. Methods

As seen in Figure 1, a confidence-driven region adaptive reasoning method is proposed in this paper. The main idea is to balance the detection accuracy and computational efficiency through a hierarchical detection strategy, which is composed of four modules: global low-resolution detection, confidence-driven region selection, local high-resolution fine detection, and result fusion. First, global low-resolution detection. The main function is to quickly screen out the suspected target area in the image and provide guidance for subsequent detection. Firstly, the original high-resolution image is scaled to the specified size according to the proportion of the long edge, and the insufficient part of the short edge is filled with gray value 0 to obtain the low-resolution input image. The image is sent to the pre-trained YOLOv5 model for reasoning, and the position coordinates, confidence scores, and category information of all detection frames are obtained. Because the total number of pixels in the low-resolution image is greatly reduced, the amount of calculation and memory

DOI: 10.54254/2755-2721/2026.34634

required for model reasoning is significantly reduced, which can meet the speed requirements of real-time detection. This stage can accurately identify most large-scale targets and initially locate the approximate area of small targets, providing basic data for subsequent area selection.

```
Input Image
    ↓
Low-Resolution Detection
    ↓
Confidence Classification
 ┌──────────────────────┐
 │ High Confidence │ → Keep
 │ Medium Confidence │ → ROI Refinement
 │ Low Confidence │ → Remove
 └──────────────────────┘
    ↓
High-Resolution Local Detection
    ↓
NMS Fusion
    ↓
Final Results
```

Figure 1. Regional adaptive reasoning framework (picture credit: original)

Second, confidence drives region selection. The subsequent processing flow of different regions is determined by grading the confidence of the global detection results. Based on the statistical analysis of target detection results in the UAV scene, the confidence of the detection frame is divided into three intervals: high, medium, and low. For the high confidence test box, it is considered that its target characteristics are obvious enough and the test results are reliable, and it is directly retained as part of the final results. For the medium confidence detection box, it is considered that it may be a small target with no obvious characteristics or a target with occlusion, which needs further fine detection to confirm. For the low confidence detection box, it is considered that its probability is background noise or false detection, and it is directly eliminated. Let the detection results be defined as model (1):

$$R = \{b_i \,|T_l < c_i < T_h\} \tag{1}$$

Among them, b_i represents the i-th detection box, c_i represents the confidence score of the corresponding detection box, T_l and Th represent the low confidence threshold and high confidence threshold, respectively, and R represents the set of candidate regions that need to enter the local high-resolution fine detection stage.

Third, local high-resolution fine detection. Recover the target details lost in low-resolution detection, and improve the detection accuracy of small targets. For the medium confidence detection box marked by the region selection module, a certain proportion of the pixel range is extended around based on its central coordinate, and the corresponding region of interest is extracted from the original high-resolution image. The extracted region of interest is separately sent to the YOLOv5 model for detection. Since this region only contains a small part of the original image, even if high-resolution input is used, the overall amount of calculation is lower than that of full image high-resolution detection. Figure 2 shows the process of local high-resolution detection. Compared with the full image multi-scale detection method, this method only performs high-resolution calculation for key areas, thus reducing the overall computational overhead.

DOI: 10.54254/2755-2721/2026.34634

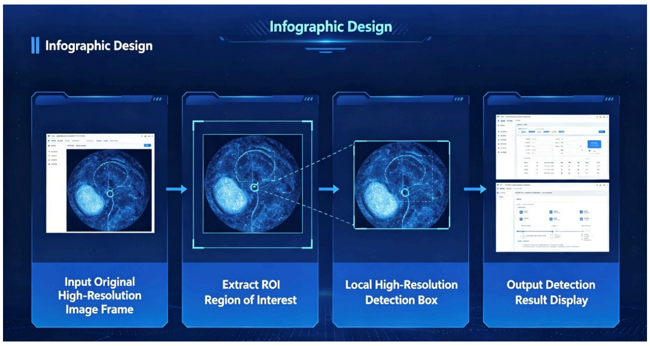

Figure 2. Local high-resolution detection process (picture credit: original)

Fourth, result fusion. The main task of the fusion phase is to integrate the results of global low-resolution detection and local high-resolution fine detection, eliminate duplicate detection, and get the final detection output. Since the global detection and local detection may produce multiple overlapping detection frames for the same target, the non-maximum suppression method is used for processing. The intersection over Union is defined as model (2):

$$IoU = \frac{|Bi \cap Bj|}{|Bi \cup Bj|} \tag{2}$$

Where Bi and Bj represent two candidate detection frames, respectively, and IoU represents the intersection and union ratio of the two. When the IoU between the detection boxes exceeds the set threshold, the detection box with high reliability is retained, and the duplicate box is deleted, so as to reduce the duplicate detection results. First, all the retained test boxes are sorted according to the confidence score from high to low, and the test boxes with the highest confidence are selected in turn. Calculate the intersection and union ratio with the remaining detection frames, eliminate the detection frames whose intersection and union ratio is greater than the set threshold, and repeat the process until all detection frames are processed. This method can effectively remove the duplicate detection results, while preserving the prediction frame with the highest reliability and the most accurate positioning.

3. Parameter settings and evaluation metrics

The experiment was verified by using the VisDrone dataset, which covers a variety of typical UAV shooting scenes such as urban streets, rural roads, mountain woodlands, etc., including 10 common targets such as pedestrians, vehicles, tricycles, bicycles, etc. The data set has a large target scale span, target occlusion, illumination change, and complex background, which has good representativeness and high detection challenge. The experiment uses the VisDrone2019-det data set, in which the training set contains 6471 images for model parameter learning and optimization. The verification set contains 548 images, which are used for model tuning and performance verification. The test set contains 1610 images, which are used to objectively evaluate the final detection performance of the model and guarantee that the model can fully extract target features from various complex scenes and scales during the training process. In this study, the average precision and frame rate are used as key evaluation indicators. The detection precision and real-time capability of the model are measured, respectively. The average precision includes two sub-indexes:

DOI: 10.54254/2755-2721/2026.34634

mAP_0.5 and mAP_0.5:0.95. YOLOv5 is selected as the baseline method and compared with the confidence-driven region adaptive reasoning method proposed in this paper.

The experiment was conducted on an NVIDIA RTX 3080 GPU using Python and the PyTorch framework. In the training process, data enhancement techniques consistent with the baseline method are used, including random cropping, random rotation, color space transformation, and mosaic enhancement, in which the random rotation angle ranges from -15° to 15°. Color transformation includes random adjustment of brightness, contrast, and saturation, and the adjustment coefficient ranges from 0.8 to 1.2. Diversified data enhancement is adopted to improve the generalization ability of the model. To maintain experimental fairness during comparative validation, all models adopt the same superparameter settings. The initial learning rate is set to 0.001, and the cosine annealing learning rate scheduling strategy is used for dynamic adjustment. The batch size is set to 16, combined with a 4-step accumulation gradient to simulate the training effect of a larger batch. The total number of training rounds is set to 50. The model converges basically in about 40 rounds, and the next 10 rounds are used to fine-tune the model parameters. In the global detection phase, the image scaling size is 416 × 416 to restore the details of small targets. In order to further verify the effectiveness of each module in this method, an ablation experiment and a parameter sensitivity experiment were designed. The ablation experiment mainly analyzes the contribution of low-resolution global detection, confidence-driven region screening, and the local high-resolution fine detection module to the overall performance. The parameter sensitivity experiment is used to analyze the influence of different confidence threshold settings on the detection accuracy and reasoning speed.

4. Results

The overall detection performance comparison results of this method and the baseline YOLOv5 method are shown in Table 1. The experimental results show that the proposed method is superior to the baseline method in detection accuracy and inference speed by introducing the confidence-driven adaptive reasoning mechanism. In terms of detection accuracy, the mAP_0.5 of this method is 3.6% higher than that of the baseline method, reaching 72.1%. mAP_0.5:0.95 is 2.8 percentage points higher than the baseline method, reaching 41.3%, indicating that this method not only improves the target recognition ability but also improves the target positioning accuracy. In terms of reasoning speed, the detection frame rate of this method is increased from 30fps to 35fps, with an increase of about 16.7%. This is because the global low-resolution detection phase significantly reduces the amount of calculation. Although the local fine detection phase increases some computational overhead, the overall amount of calculation is still lower than the full image high-resolution detection, realizing the synchronous improvement of detection accuracy and calculation efficiency.

Table 1. Performance comparison between YOLOv5

Method	mAP_0.5 (%)	FPS
YOLOv5	68.5	30
Method in this paper	72.1	35

Further statistics show that in most test scenarios, the number of candidate regions that need to enter the local high-resolution fine detection stage only accounts for 20% to 30% of the total detection regions. Therefore, although the local fine detection increases the additional computational overhead in some regions, due to the low-resolution input in the global detection phase, the overall reasoning calculation is still lower than that of the full image high-resolution detection method, thus

DOI: 10.54254/2755-2721/2026.34634

ensuring the real-time reasoning ability of the model. To investigate the impact of each constituent module on the final detection effect, further ablation experiments are performed in this paper. The experimental results are shown in Table 2. The experimental results show that when only using low-resolution global detection, the model reasoning speed can be significantly improved, but the accuracy of small target detection has a certain decline. After introducing the confidence-driven region filtering mechanism, it can effectively reduce the repeated calculation of invalid regions and further improve the detection performance while maintaining a high reasoning speed. Finally, after adding the local high-resolution fine detection module, the small target details are restored, and the model achieves the optimal overall detection performance.

Table 2. Ablation results of different modules

Method	mAP_0.5 (%)	FPS
YOLOv5	68.5	30
+Low-resolution global detection	69.4	38
+Confidence zone screening	70.3	37
+Local high-resolution fine detection	72.1	35

The experimental results show that the region adaptive reasoning mechanism proposed in this paper can effectively improve the performance of small target detection in a UAV scene on the basis of ensuring real-time performance. Among them, the local high-resolution fine detection can significantly improve the detection accuracy of small targets, while the confidence-driven region screening mechanism plays an important role in reducing invalid calculations. In order to visually show the detection effect of this method, Figure 3 shows the comparison of the detection results of the baseline method and this method in a single scene. It can be seen that for pedestrian targets with relatively small pixel proportions in the image, the baseline YOLOv5 method has obvious missed detection, and some pedestrian targets at a long distance cannot be detected. However, this method can accurately detect the traveler target, and the confidence of the detection frame is generally higher than that of the baseline method, indicating that local fine detection effectively enhances the feature expression ability of small targets.

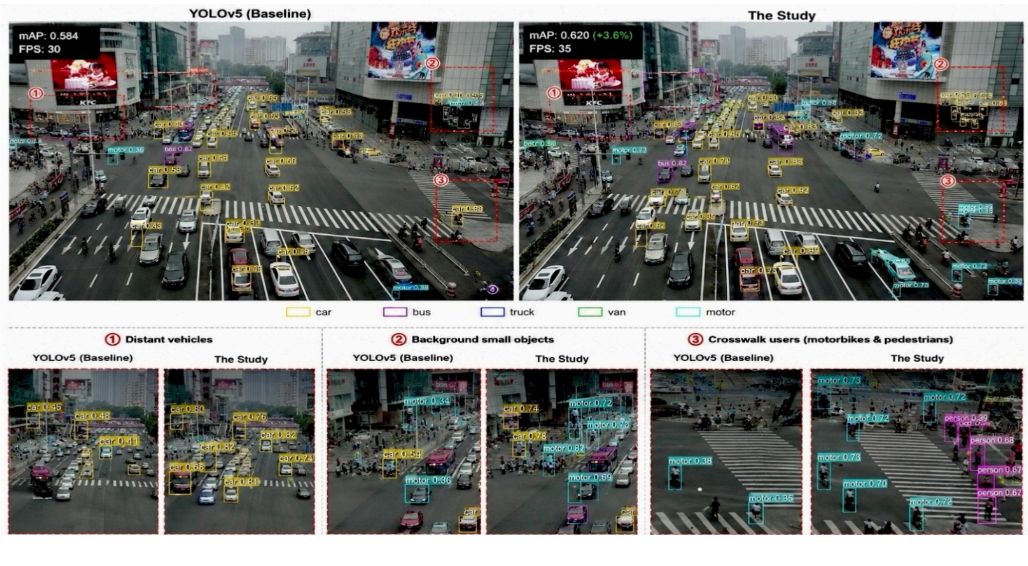

Figure 3. Example of small target detection in a UAV scene (source: VisDrone dataset)

DOI: 10.54254/2755-2721/2026.34634

Figure 4 illustrates the detection results of the proposed method under various environments, including typical UAV application scenarios such as urban streets, field areas, and cloudy light. The experimental results show that the proposed method has good generalization ability and can adapt to the complex and changeable practical application environment.

Figure 4. Example of small target detection in different environments (source: VisDrone dataset)

The division of the confidence threshold directly determines the number of regions that need local fine detection, and then affects the detection accuracy and reasoning speed. In this paper, the high confidence threshold is set to 0.7, and the low confidence threshold is set to 0.3. That is, the detection frame with confidence greater than 0.7 is directly retained, the detection frame with confidence less than 0.3 is directly eliminated, and the detection frame with confidence between 0.3 and 0.7 triggers local fine detection. By adjusting the threshold, it is found that when the high confidence threshold increases, more detection frames will enter the local fine detection stage, the detection accuracy will be improved, but the reasoning speed will be reduced accordingly. When the high confidence threshold is reduced, the reasoning speed will be accelerated, but some small targets with unclear characteristics will be directly retained, resulting in an increase in the false detection rate. Considering the requirements of detection accuracy and real-time, this paper selects 0.3 and 0.7 as the confidence threshold, which can achieve a good balance between the two. The region adaptive reasoning mechanism proposed in this paper has good scalability and can be applied to other target detection frameworks in the future, such as Faster R-CNN or SSD, to improve their detection performance. In order to analyze the impact of different confidence threshold settings on detection performance, this paper further carried out parameter sensitivity experiments, and the experimental results are shown in Table 3.

Table 3. Detection performance under different confidence threshold settings

Low/high threshold	mAP_0.5 (%)	FPS
0.2/0.6	71.8	32
0.3/0.7	72.1	35
0.4/0.8	72.4	28

The experimental results show that when the high confidence threshold gradually increases, the number of regions entering the local high-resolution fine detection phase increases, which can further boost the detection precision for small targets, but the overall reasoning speed of the model

DOI: 10.54254/2755-2721/2026.34634

will decrease accordingly. When the threshold value is set low, some uncertain target areas are directly retained, resulting in an increase in the false detection rate. Considering the two factors of detection accuracy and real-time performance, this paper finally selects 0.3 and 0.7 as the confidence threshold.

5. Conclusion

Aiming at the problems of low accuracy of small target detection and limited computing power of edge devices in the UAV scene, this paper proposes a regional adaptive YOLOv5 target detection method based on confidence-driven. By constructing a hierarchical reasoning framework combining global low-resolution fast detection and local high-resolution fine detection, the contradiction between small target feature loss and high computational overhead in traditional full image detection methods is solved. This method does not need to make complex modifications to the basic detection network, and can achieve performance improvement only by optimizing the reasoning process. It has a low deployment threshold and can adapt to the edge computing equipment carried by a UAV. It provides a feasible technical scheme for the real-time target detection task of UAVs in practical engineering.

The future research work can be further carried out from three aspects. First, explore the collaborative acceleration method of algorithm and hardware. According to the computational characteristics of edge-specific hardware such as FPGA and TPU, the reasoning process is optimized and parallelized to further improve the real-time performance of the algorithm on low-power devices. Second, the detection method of multimodal information fusion is studied, which combines RGB images with infrared, depth, and other sensor information, and uses the complementary characteristics of different modes to tackle the difficulties in small target detection under complex lighting, bad weather, and target occlusion conditions. Third, expand the application scenarios of the method, deeply integrate the regional adaptive reasoning mechanism with UAV autonomous navigation, dynamic obstacle avoidance, and other tasks, build an end-to-end UAV intelligent sensing system, and promote the landing application of UAV technology in more fields.

References

[1] Saeed, Z., Yousaf, M.H., Ahmed, R. et al. (2023) On-board small-scale object detection for unmanned aerial vehicles (UAVs). Drones, 7(5), 310.

[2] Lu, S., Guo, Y., Long, J. et al. (2025) Dense small target detection algorithm for UAV aerial imagery. Image and Vision Computing, 156, 105485.

[3] Kassab, M., Zitar, R.A., Barbaresco, F. et al. (2024) Drone detection with improved precision in traditional machine learning and less complexity in single-shot detectors. IEEE Transactions on Aerospace and Electronic Systems, 60(4), 3847-3859.

[4] McEnroe, P., Wang, S. and Liyanage, M. (2022) A survey on the convergence of edge computing and AI for UAVs: opportunities and challenges. IEEE Internet of Things Journal, 9(17), 15435-15459.

[5] Fang, H., Ding, L., Wang, L. et al. (2022) Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion. IEEE Transactions on Instrumentation and Measurement, 71, 1-20.

[6] Gao, S., Gao, M. and Wei, Z. (2024) MCF-YOLOv5: A small target detection algorithm based on multi-scale feature fusion improved YOLOv5. Information, 15(5), 285.

[7] Tang, G., Ni, J., Zhao, Y. et al. (2023) A survey of object detection for UAVs based on deep learning. Remote Sensing, 16(1), 149.

[8] Zhang, Z. (2023) Drone-YOLO: an efficient neural network method for target detection in drone images. Drones, 7(8), 526.

DOI: 10.54254/2755-2721/2026.34634

[9] Chen, Z., Ji, H., Zhang, Y. et al. (2023) High-resolution feature pyramid network for small object detection on drone view. IEEE Transactions on Circuits and Systems for Video Technology, 34(1), 475-489.
[10] Yang, Y. (2024) Vehicle target detection algorithm based on improved faster R-CNN for remote sensing images. J. Artif. Intell. Pract, 7(1), 27-33.