Innovative Research on Small Object Detection and Recognition in Remote Sensing Images Using YOLOv5

Shan Jiang [1], He Huang [1], Junxing Yang [1] *, Xin Zhang [1], Siqi Wang [1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China -
2108570023131@stu.bucea.edu.cn,huanghe@bucea.edu.cn,
zx2806404866@163.com,yangjunxing@bucea.edu.cn,w2248183443@163.com

KEY WORDS: Remote Sensing Satellite Imagery, Target Detection, Deep Learning, YOLOv5

ABSTRACT:
With the increase of remote sensing image acquisition methods and the number of remote sensing image data, the traditional manual annotation and recognition methods can no longer meet the needs of the present production life. This study explores the use of deep learning techniques to improve the efficiency and accuracy of target detection in remote sensing satellite images, especially for small targets. Traditional target detection methods often face challenges in recognition accuracy and processing speed due to the specificity and complexity of satellite images, such as large size, variable lighting conditions and complex background. Therefore, this paper adopts the YOLOv5 model and introduces the CBAM (Convolutional Block Attention Module) attention mechanism, which significantly improves the detection of small and dense targets. Experimental validation, using the improved YOLOv5 model on the VisDrone2021 dataset, demonstrates that the model improves the mean average percision (mAP) by 1.9% while maintaining real-time performance. This paper provides new ideas for remote sensing image processing, especially for applications in the fields of urban planning and automatic driving, etc. Despite the progress made in this study, the detection of small targets in remote sensing images, the limited classification accuracy and the detection of dynamic targets still need further research.

1.INTRODUCTION

Remote sensing satellite images play a very important role in modern life and production activities, through which images rich in geographic information can be obtained, enabling researchers to analyze their areas of interest in depth and extract valuable data (Lin et al., 2014). In view of the fact that remote sensing satellite images can quickly and extensively monitor large areas, the identification of objects in them is particularly critical to the accuracy of data processing and analysis. Accurate detection of objects is a complex and critical task in remote sensing satellite image analysis, which requires not only the accurate identification of targets of interest, but also their precise positioning (Everingham et al., 2010). With the continuous development and improvement of remote sensing technology, this technology is of great help and has a wide range of application scenarios for smart city construction, land resources investigation and management, disaster management, and military reconnaissance (Cordts et al., 2016).

In the early days, traditional target detection algorithms mainly relied on manually labelled features for data extraction, which consisted of three steps: region of interest selection, feature extraction, and classifier classification. The traditional manual annotation method is relatively simple, but it is generally only suitable for specific scenes or a small range of scenes, once encountered in a large scene, the classification of more categories, the labour cost will increase and the quality of accuracy will be reduced (Girshick et al., 2014). Moreover, because of the complexity and specificity of satellite images, it is also a challenge to accurately identify the target objects in remote sensing satellite images, such as the large image size, changing lighting conditions and the complexity of the background information, so that the traditional methods of [1]accuracy and efficiency are affected (Zhang et al., 2011). In recent years, with the rapid development of machine learning and advanced deep learning techniques, these challenges have been effectively addressed, and the efficiency of target detection and recognition in satellite images has been significantly improved. Although deep learning methods show great potential in image recognition and detection though, this method is still limited by the complexity of image features and training datasets (Russakovsky et al., 2015). Due to the complexity of the actual scene of remote sensing images, small object targets appear densely and frequently in the data, which requires the recognition and detection task to reach a finer and deeper level, and at the same time helps to address the research on how to overcome the challenges of detecting and recognising objects in the images due to objective factors (Bochkovskiy et al., 2020).

The current data processing process for remote sensing satellite images usually divides deep learning-based target detection algorithms into two categories: region-based target detection algorithms and end-to-end target detection algorithms. The former generates a series of candidate boxes as samples through the algorithm, and then classifies and locates the samples through convolutional neural networks, and determines whether each candidate box contains the target of interest and the exact location of the target; this type of algorithm avoids the repeated calculation of a large number of sliding windows, and it has a great advantage in the detection accuracy and localisation accuracy, but the disadvantage is that it needs to generate a large number of region proposals, and it needs to extract and classify features for each proposal separately; its processing time is long and relatively slow, and its representative algorithms are R-CNN, Fast R-CNN, etc. (Ren et al., 2015). The latter converts the target bounding box localisation and classification problem into a regression problem, omits the intermediate process of region detection, and directly performs in-depth processing of the bounding box, and clusters the feature extraction, candidate classification and regression problems into the same deep convolutional network to achieve

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-77-2024 | © Author(s) 2024. CC BY 4.0 License.

the project requirements (Liu et al., 2016). This type of algorithm is superior in detection speed, the representative algorithms are YOLO series and SSD algorithm, etc., due to the research focus on speed, omitting the process of region detection, so that the accuracy is not as good as the former detection algorithm (Redmon and Farhadi, 2018).

Meanwhile, with the continuous development of remote sensing technology, the size, type, and background complexity of the target are constantly changing, and the detection speed and versatility of target recognition have become the primary consideration. Moreover, the current research on target detection in remote sensing images is not very perfect, and the following difficulty points are now encountered: firstly, small target detection has always been the difficult point of target detection in remote sensing images, because the small target itself contains little information and is affected by the background information, which leads to the poor processing of the small target by the whole network, and it brings a great challenge to the detection task; secondly, the problem of high-quality detection, which is solved in the target detection task by comparing the predicted Second, the problem of high-quality detection, in the target detection task, by comparing the intersection and integration ratio (IOU) between the predicted frame and the real frame, the IOU value is usually thresholded according to the task, and usually the threshold is set to 0.5, once the threshold is raised, the algorithm will be inaccurate due to regression localisation, and the detection accuracy is not satisfactory (Bochkovskiy et al., 2020). In order to solve these challenges, comparing the commonly used algorithms for target detection, according to the requirements and scene characteristics, this study adopts the YOLOv5 detection algorithm based on the detection of small and medium-sized objects in remote sensing satellite images, conducts experimental tests on the VisDrone2021 public dataset, analyses the problems of the existing algorithms in detail, and proposes the improvement ideas for the detection of small objects, which enhances the overall detection accuracy of the algorithm. The overall detection accuracy of the algorithm is improved.

2.Review of Related Works

2.1 Characteristics of remote sensing satellite images

First, the low utilization rate of remote sensing images: according to available research data, the utilization rate of aerial remote sensing data is less than 10 per cent, while the actual utilization rate of aerospace remote sensing data is even lower than 5 per cent, which makes it an urgent need at present to develop smarter methods of understanding high-resolution remote sensing images to improve the real utilization rate of remote sensing data. In addition, adverse weather conditions such as cloudy, foggy, rainy, etc. will affect the quality of remote sensing images to a certain extent, further affecting the utilisation rate of remote sensing images.

Secondly, the scale diversity and perspective specificity of remote sensing images: aerial remote sensing images can usually be taken from hundreds of meters to nearly 10,000 meters at different heights, and there will be different sizes and patterns of ground targets, with the same kind of targets, such as the same kind of buildings, having large buildings as big as thousands of square meters, and also small buildings of several square meters; and the size differences between different kinds of targets, such as bridges and automobiles. Remote sensing images are classified into tilt photography and orthophotography due to the different angles of image acquisition, both of which are usually high altitude overhead views, which are different from the conventional ground-based horizontal view pattern. A well-trained detector on conventional data may also have different degrees of aberrations and distortions due to the influence of internal state changes of the sensor, external state and ground conditions during imaging, resulting in poor remote sensing satellite images.

Third, the small target problem: In remote sensing satellite images, many target objects are small targets with limited information. Although CNN-based target detection methods perform well on conventional datasets, detection of small targets is still a challenge. pooling layers in CNN may reduce the amount of target information, e.g., a 24x24 pixel target may have only one pixel left after multi-layer pooling, which makes classification difficult.

Fourth, uncertainty of target orientation: since remote sensing satellite images are taken with a top-down perspective, they are usually classified into two types, horizontal target detection and directed target detection, which are used to determine how to extract rotationally invariant features and how to accurately identify the target angle, respectively, as a way to reduce the impact of the multi-directional problem on target recognition.

Fifth, high background complexity: remote sensing satellite images have a wide field of view, the degree of coverage can span several square kilometres, and the amount of data they collect is large, making the image information very complex. In addition, when objects are densely arranged, the bounding box prediction between similar targets with very close positions may be inaccurate and easily overlapped, thus affecting the efficiency and accuracy of detection.

2.2 Target detection based on YOLOv5

Target detection technology is not only a core research topic in the field of computer vision, but also a key technology in practical applications, which are widely used in a variety of applications such as security surveillance, self-driving vehicles, industrial automation, smart retailing, and drones. The development of these technologies has driven innovation and progress in many fields. In the past, target detection mainly relied on traditional manual feature methods, such as SIFT (Scale Invariant Feature Transform) and HOG (Histogram of Orientation Gradients), which achieved some success at that time. However, as the dataset continues to expand and the computational power increases, the traditional methods gradually show their limitations, especially the saturation of performance when dealing with complex or changing image environments. In recent years, the introduction of deep learning techniques has marked a major revolution in target detection technology. Deep learning methods, especially Convolutional Neural Networks (CNN), have become mainstream techniques in the field of target detection. These methods significantly improve the accuracy and efficiency of target detection by learning a large amount of labelled data to automatically extract effective features.

A typical algorithmic process for target detection can be divided into three main steps: candidate region generation, image feature extraction, and classification with candidate regions. In the candidate region generation stage, the network analyses the image by the sliding window method to identify regions that may contain targets. This method may result in a large number of duplicate candidate boxes, generating redundant data, which increases the model computation and affects the computational speed. In the image feature extraction stage, according to detecting the features of the target object, it is necessary to extract the features according to the manual requirements, because it is affected by the subjective factors of manual operation and objective factors such as the weather environment, which leads to poor robustness of the target detection model, and the detection effect is not good. Finally, in the candidate region classification step, each candidate region is further analysed and classified to determine its specific category. In addition, the development of target detection techniques has led to some new research directions, such as real-time target detection, multi-scale detection and unsupervised or semi-supervised learning methods for target detection. These technological advances provide more powerful and flexible tools for solving real-world visual recognition problems.

YOLOv5 was developed by Glenn Jocher et al. It is an open source project of Ultralytics, Inc. It has been updated with 7 major versions from its release in June 2020 to November 2022, with semantic segmentation added in v7. YOLOv5 is currently one of the most popular and widely used target detection algorithms, and has gone through many years of iterative development, YOLOv5 is the iteration of YOLOv1 over a number of years and is divided into, according to the number of parameters YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x four models, the number of parameters in order to rise, of course, its effect is also better and better. The network structure is more complete, and the detection performance and accuracy are also significantly improved over the previous version. Its main improvement ideas have four aspects: first, in the input side stage, it mainly proposes the improvement ideas of Mosaic data enhancement, adaptive anchor frame calculation, and adaptive picture scaling; second, it integrates some new ideas from other detection algorithms, mainly including the Focus structure and the CSP structure; third, the Neck network of YOLOv5 adopts the FPN+PAN structure; fourth, the The anchor frame mechanism of YOLOv5 output layer is the same as that of YOLOv4, and its main improvements are the loss function GIOU_LOSS for training, and DIOU_nms for prediction frame screening.

YOLOv5 adds Mosaic data enhancement on the basis of other data enhancement methods used for target detection. The algorithm is improved on the basis of CutMix data enhancement method, CutMix only uses two images for splicing, while Mosaic data enhancement method uses four images and splices them into a single image by random scaling, random cropping and random arrangement, which not only enriches the dataset, but also greatly improves the training speed of the network while reducing the memory requirements of the model. This not only enriches the dataset, but also greatly improves the training speed of the network while reducing the memory requirements of the model, and the method is also very advantageous for the detection of small target objects. Because many images have different aspect ratios, the black edges on both sides of the processed image are of different sizes, and if more information is filled in, it will lead to redundancy of information, which affects the inference speed, therefore, by modifying the letter box function, the least amount of black edges are added to the original image adaptively to reduce the amount of computation, i.e., the speed of the target detection is also improved.

As can be seen from Figure 1, YOLOv5 proposes the Focus structure in the first layer of the benchmark network, which crops the input image by slice operation, the original image size from 608*608*3 after the operation outputs a feature mapping of 304*304*12, and after the processing of the Conv layer, it outputs a feature mapping of 304*304*32 size, that is, it adopts the The slicing operation disassembles the high-resolution image into multiple low-resolution images.The role of the Focus structure is to transform the high-resolution image information from the spatial dimension to the channel dimension, so as to maximise the retention of the input information and reduce the size of the input image, which is conducive to improving the efficiency of network training and detection.

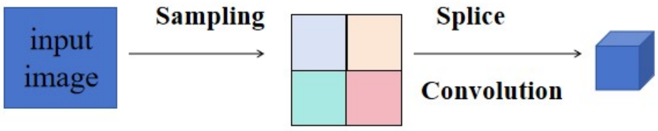


Figure1.Schematic diagram of the Focus structure of YOLOv5

The difference between YOLOv5 and YOLOv4 is that two CSP structures are designed, one is CSP1_X structure applied in Backbone backbone network, and the other is CSP2_X structure applied in Neck part.The purpose of the CSP structure is to split the feature map into two parts, one part of the feature map continues to undergo the convolutional operation to obtain more profound feature information, and the other part is the current feature map is processed by splicing with the previous part of the feature map that has undergone convolutional operation.The advantages of this design include: reducing the amount of computation, improving the inference speed, reducing the cost of memory and guaranteeing the accuracy rate.

In the Neck part, YOLOv5 still adopts the combined structure of feature pyramid network and path aggregation network to fuse feature maps with low resolution and high semantic information with high resolution and high geometrical information, but at the same time draws on the structure of CSP2 designed by CSPnet, which strengthens the ability of the network feature fusion, and fuses the different layers of feature maps more effectively with high quality, which is conducive to the feature information extraction.

YOLOv5 uses CIOU_Loss as the loss function of Bounding box for model training, which contains classification loss and edge regression loss. The loss function, also known as the cost function, plays an important role in both machine learning and deep learning, it is used to quantify the difference or error between the model's predicted value and the actual value, and it is a measure of the degree of the model's prediction error. The loss function helps the network in deep learning to adjust the weights by back propagation algorithm to reduce the prediction error.YOLOv5 chooses the CIoU loss function for model training in border regression.IoU, which is fully known as the intersection and concatenation ratio, is one of the most commonly used performance measures in target detection, and is equal to the ratio of the intersection and concatenation between predicted borders and the real border, which is computed as shown in Equation 1.

$$IOU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

Iou schematic is shown below:

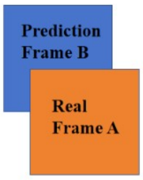


Figure2. Schematic diagram of Iou

In practical applications, the screening of positive and negative samples is performed by whether the IoU is greater than the threshold, and in the process of non-extremely large value suppression, the IoU greater than the threshold with the current candidate box is discarded, and in the calculation of the evaluation index AP, only when the IoU between the detected box and the real box exceeds the threshold and is classified correctly, it is defined as a correct detection. When the real frame and the prediction frame do not overlap, the IoU is constant equal to 0, and when they completely overlap, the IoU is constant equal to 1. When the value of IoU is larger, the larger the overlap area between the real frame and the prediction frame, the prediction of the prediction frame predicts the more accurate effect.CIoU is based on DIoU, the aspect ratio of the bounding box will be taken into account, and the loss calculation function of the CIoU is shown in Eq. 2 and Eq. 3.

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \qquad (2)$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \qquad (3)$$

Where v is used to represent the similarity in aspect ratio and α is the weighting factor.

3. Methods

3. 1 Dataset and experimental environment

In recent years, with the continuous progress of deep learning and remote sensing data processing, various types of datasets for different tasks have been increasing. Currently, the public datasets commonly used for many deep modelling studies are:

(1) PASCL VOC dataset: a commonly used dataset by many researchers at present, the more important ones are the datasets of the years PASCAL VOC 2007 and PASCAL VOC 2012, which have 20 categories and hierarchical structures, and become an important benchmark for evaluating the performance of target detection or segmentation algorithms.

(2) DOTA aerial image dataset: a large dataset for target detection in aerial images, now available in v1.0, v1.5.v2.0. Compared with other datasets, the DOTA dataset is unique in that its objects are labelled with oriented bounding boxes (OBB for short), which can better encapsulate targets and distinguish crowded objects.

(3) VisDrone2021 dataset: a UAV remote sensing image dataset produced by the AISKYEYE team from Tianjin University, which contains a lot of small target objects covering a wide range of locations, and provides some important attributes including scene visibility, object class and occlusion to improve data utilisation.

(4) KITTI dataset: It is a benchmark dataset widely used in the field of autonomous driving, created jointly by Karlsruhe Institute of Technology and Toyota Technological Institute in Germany. This dataset is widely used for its complex real-world scenarios and exhaustive annotation, advancing the prospects of autonomous driving.

The rapid development of UAV provides a large number of high resolution remote sensing images for small target object detection (e.g. vehicle detection, pedestrian detection), UAV remote sensing technology is less affected by weather and has the advantages of long working time, simple operation and flexibility. So in this paper, VisDrone2021 is used to validate the algorithm of this paper effectively, the dataset consists of 288 video clips consisting of 261,908 frames and 10,209 still images, which cover many aspects, including location, environment, objects and so on. In this paper, 3685 images from VisDrone2021 dataset are selected as training set, 268 images as test set and 268 photos as validation set. The images contain three categories of pedestrians, bicycles and vehicles.VisDrone2021 dataset contains rich vehicle targets, dense target distribution, a large number of small targets and can meet the detection task of this paper, and can well verify the effectiveness of the algorithm of this paper. Before target detection on remote sensing image data, each image is preprocessed to remove some pictures affected by external factors such as sensor noise.

In order to ensure the reliability and fairness of each experiment, all experiments in this paper are completed in the following uniform configuration. In the training and testing phases, this paper uses an operating system of Windows 10, a graphics processor of NVIDIA GTX1080Ti, a Python version of 3.8.18, a Torch 1.12.1 framework, and a cuda 11.3 accelerated environment for training and testing. The reliability of the experimental results is ensured by the above choices of operating system, graphics processor, Python version and acceleration environment.

3. 2 Methods

Because of the advantages of YOLOv5 high efficiency and high accuracy and easy to use, target detection processing of remote sensing images based on YOLOv5, but the limitation of YOLOv5 is that it needs to rely on a large amount of data of various kinds, and the effect of detecting smaller objects is generally effective, so in order to make YOLOv5 have a better performance performance in the task of target detection and to improve the accuracy of detection of small targets, the attention mechanism is introduced. Attention mechanism in target detection is a method to enhance the accuracy of object detection, which is similar to the attention mechanism in human vision, putting more attention into the region of interest and the object to be detected itself, according to the different domains of attention to the attention mechanism is classified into five categories: channel domain, spatial domain, layer domain, hybrid domain, and temporal domain.Convolutional Block Attention Module (CBAM) is a simple but effective attention module for feed-forward convolutional neural networks.The advantages of the CBAM attention mechanism are as follows: firstly, it has a powerful feature representation capability, which can adaptively learn the importance of the input feature values, thus improving the feature representation capability. Second, it can be embedded and used with different convolutional neural network structures, such as ResNet, DenseNet, etc. Third, the design of CBAM attention mechanism is relatively simple, and the addition to the existing convolutional neural network China has a smaller increase in the number of computations and parameters, and less pressure on computer equipment. In this paper, according to the characteristics of CBAM, it is added to the Backbone part of the YOLOv5 detection model, and the network structure of the model is improved and optimised appropriately, as a way to achieve better results in detection.The CBAM attention module contains two independent sub-modules, the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), and the basic principle is that the features to be processed are The basic principle is that the feature map to be processed is strengthened in the channel domain, and then the processed feature map is multiplied with the original feature map, and then some features are strengthened and weakened through the spatial domain.

The advantage of this operation lies in the effective use of network resources and the automatic introduction of feature strengthening and weakening regions. Its structure is shown in Figure.3.

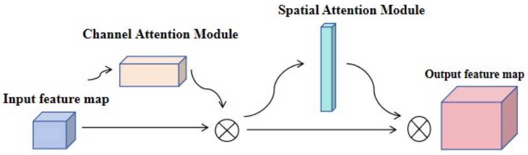


Figure 3. CBAM module structure

Adding CBAM to the algorithmic model backbone network brings many benefits. Firstly, the backbone network plays a role in YOLOv5 in extracting features from the input image, and by adding the CBAM attention mechanism, it makes it possible to improve the perception of key features and semantic information in the image. Secondly, the introduction of CBAM into the backbone part can reduce the influence of background information on recognising the target by using the attention mechanism to link between the target and the background at an early stage of feature extraction. In summary, it is shown that the introduction of CBAM attention mechanism can reduce the amount of computation while achieving better results, and the feature enhancement operation is performed in a feature map with more complete information, and the completeness and expressiveness of the feature information is improved compared with the previous algorithm.

4.Experiment

4.1 Evaluation indicators

Commonly used performance evaluation metrics in the field of target detection are: accuracy, precision, recall, average precision (AP), mAP, P-R curve, etc. In the experimental detection accuracy, mean Average Precision (mAP) is used as an evaluation metric in detection accuracy, so that the IoU threshold is greater than 0.5, i.e., mAP@0.5. mAP is the mean of Average Precision (AP) of all target categories. mAP value reflects the model's performance on the mAP value reflects the detection accuracy of the model on the dataset. The calculation of mAP is shown in Equation 4 below:

$$mAP = \frac{1}{c}\sum_{(i=1)}^{c} AP_i \tag{4}$$

where c denotes the number of detection target categories and i is the category index. mAP is closely related to the average detection precision, recall, precision and average accuracy of all target categories. Recall represents the proportion of the number of samples correctly predicted as positive cases by the classification model to the number of samples of all true positive cases, which measures the correctness of the model in predicting positive cases, and is calculated as in Equation 5:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Where TP is true cases, i.e., the number of samples correctly predicted as positive cases by the model, and FN is false negative cases, i.e., the number of samples incorrectly predicted as negative cases by the model. The precision rate, also known as the checking rate, indicates the proportion of the number of samples correctly predicted as positive cases by the classification model to the number of samples predicted as positive cases, which measures the accuracy of the model in predicting positive cases, and is calculated as in Equation 6：

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where FP is the number of false positive cases, i.e., the number of samples that the model incorrectly predicts as positive.

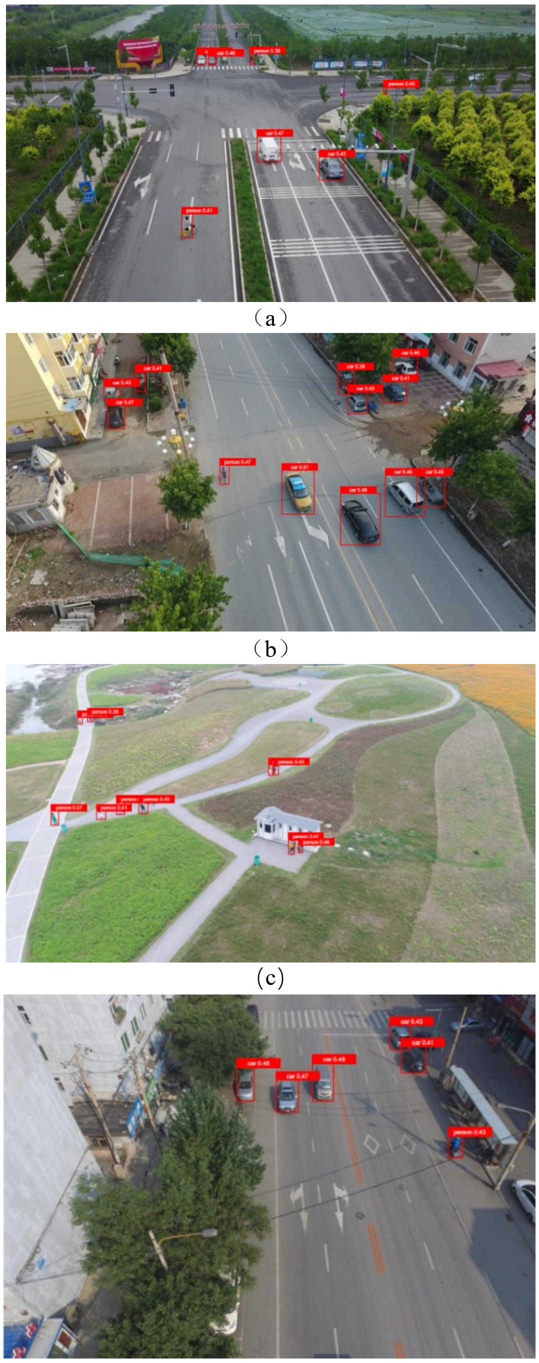
4.2 Ablation Experiment

In this paper, two sets of ablation experiments are designed on the image dataset to verify the validity of the model, and the results of the ablation experiments are shown in Table 1.

Group	Methods	Params (M)	Precision (%)	Recall (%)	mAP@ 0.5 (%)
1	YOLOv5	30.12	45.7	41.2	49.3
2	YOLOv5 +CBAM	30.53	46.1	43.6	51.2

Table 1. Results of ablation experiments

From Table 1, it can be concluded that in the experiment without adding the CBAM attention mechanism, the number of parameters is 30.12M, the precision rate is 0.457, the recall rate is 0.412, and the average precision is 0.493, which reflects that YOLOv5 is a target detection algorithm with high accuracy, easy to use, and performs well in real-time scenarios. After adding the CBAM attention mechanism, the precision and recall increased by 0.4% and 2.4% respectively, and the mAP also increased by 1.9%. From the experimental results, it can be seen that the algorithm with the addition of CBAM attentional mechanism can correctly detect the target object in target detection without false detection as well as missed detection in different situations. And it can be seen in the experimental results that the enhancement of this algorithmic improvement can be well reflected in the detection of small targets. The following are the results of two groups of ablation experiments, as shown in Fig. By comparing Fig. 4 and Fig. 5, it is concluded that Fig. 4.a identifies puddles on the road as pedestrians, while the addition of the CBAM attention mechanism detects all the target objects that are misdetected as well as missed by YOLOv5 algorithm, and Fig. 5.d identifies and detects vehicles parked under the shade of the trees, which are not detected by Fig. 4.d.



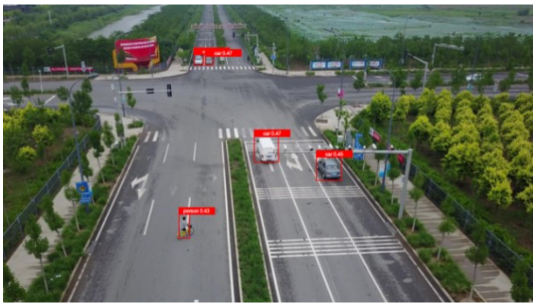
(a)

(b)

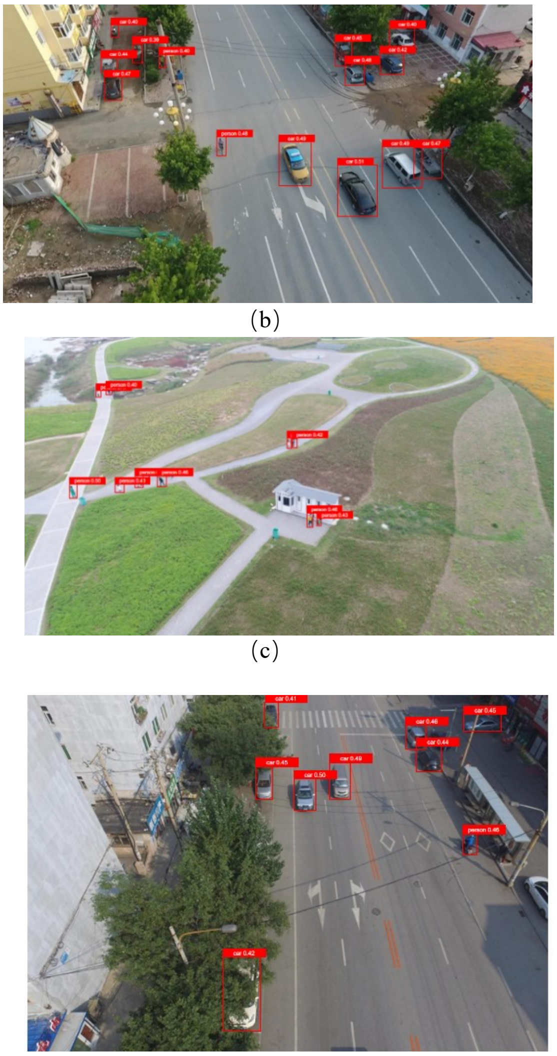
(c)

(d)

Figure 4. Graph of results for YOLOv5



(a)



(b)

(c)

(d)

Figure 5.YOLOv5+CBAM result map

5.Conclusions

With the rapid development of remote sensing technology, the acquisition and use of remote sensing data have become more and more common, especially in the fields of urban planning, environmental monitoring, agricultural management and automated driving, which show their unique value. Therefore, the development of efficient and accurate automated processing methods is crucial to enhance the processing capability of remote sensing data. In this paper, based on the YOLOv5 model, CBAM (Convolutional Block Attention Module) attention mechanism is integrated into the backbone part of the model, aiming to enhance the feature extraction ability of the model for small target objects in remote sensing images, which effectively reduces the difficulty of detecting small targets and improves the accuracy of detection.

By integrating CBAM into the backbone of YOLOv5, we not only strengthen the model's attention to small targets, but also improve the information flow efficiency of the overall network.CBAM optimises the important parts of the feature map through the spatial and channel attention mechanisms, which enables the network to focus more on those critical features, thus improving the detection performance. The experimental results show that the YOLOv5 model with integrated CBAM exhibits excellent performance on the remote

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-77-2024 | © Author(s) 2024. CC BY 4.0 License.

sensing image target detection task, both in terms of real-time and accuracy. Specifically, the model's mean accuracy (mAP) was improved by 1.9% over the use of the traditional YOLOv5 model, an improvement that is an enhancement for remote sensing data processing.

However, some problems and challenges were also exposed during the study. Firstly, there are numerous small targets in remote sensing images, and these targets are easily neglected during network training, resulting in the loss of important information during feature extraction. This problem needs to be solved by further optimising the network structure or introducing more effective data enhancement methods. Secondly, remote sensing images are very rich in information, and current manual classification methods are limited by the number of categories, which affects the comprehensive performance of target detection. Future research could explore more automatic classification techniques to handle large and complex data types more accurately. Finally, for dynamic target detection in remote sensing images, such as moving vehicles or other changing objects, the existing models still need to be optimised and improved in terms of handling time series data.

In conclusion, although this study has achieved some results in the field of small target detection in remote sensing images, in-depth research on the proposed method is still needed to overcome the existing limitations and further improve the performance. Looking ahead, with the continuous advancement of machine learning techniques, all these problems and challenges should be important issues that require further in-depth research and improvement in the future.

References

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision (IJCV), 88(2), 303-338.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European conference on computer vision.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In European conference on computer vision.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3), 211-252.

Zhang, L., Li, W., & Nevatia, R. (2011). Global data association for multi-object tracking using network flows. In Proceedings of the IEEE conference on computer vision and pattern recognition.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-77-2024