Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2025.060114
ISSN 2789-7168 Vol. 6, Issue 1:144-151

Object Detection and Image Segmentation Algorithm Optimization in High-Resolution Remote Sensing Images

Chuying Lu

University of Michigan, University of Michigan, Michigan 48109, USA

Keywords: High-resolution remote sensing; Object detection; Image segmentation

Abstract: High-resolution remote sensing images play an important role in urban decision-making, natural resource supervision and environmental assessment, etc. To meet the demands of such fields, deep detection methods such as FasterRCNN and YOLOv5 were deeply analyzed. Multi-scale feature fusion was achieved using FPN, attention expression was enhanced through Transformer and CBAM, and detection efficiency was improved through MobileNet and pruning techniques. In terms of segmentation, the semantic description of spatial information is highlighted by using dilated convolution and jump connection. Multimodal fusion can improve the recognition accuracy of complex ground objects. The semantic model ability is enhanced through the Transformer module. Finally, a new edge detection branch is proposed to improve the contour extraction effect.

1. Introduction

In recent years, remote sensing image technology has been widely applied, and both its resolution and data acquisition frequency have been significantly improved. Especially after high-resolution remote sensing images have been widely accepted, the recognition accuracy of ground objects has been enhanced. However, the improvement in image quality also means that the difficulty of data processing has increased. Because high-resolution images are rich in detailed information, with complex textures and a large amount of ground object information, for traditional image processing methods, it will cause difficulties in improving efficiency and accuracy. Especially for the tasks of object detection and image segmentation, what is required is not only powerful feature extraction techniques, but also to deal with problems such as blurred boundaries between different scales and categories. Therefore, seeking technical laws that are more effective, more stable and more applicable to high-resolution images is currently the focus of intelligent remote sensing processing research.

2. The characteristics of high-resolution remote sensing images

2.1 Data Spatial resolution

High-resolution remote sensing images have an extremely strong ability to express details. They have smaller pixel sizes and can more truly reflect the contours, shapes and distribution correlations

of ground elements. High-resolution images in regional scenes such as complex terrains, urban streets, and rivers can clearly reflect the detailed differences of each object in the scene, which is helpful for the boundary distinction and classification recognition of the target. In medium-resolution or low-resolution images, the terrain structure is relatively discrete, the lines are blurred, and the texture is rough. In contrast, high-resolution images are more suitable for detailed inspection, separation and other processing. This type of image can also reduce the probability of small targets being missed in the image and enhance the model's recognition ability for small-scale terrains. In the processing of deep learning models, high-resolution images input more comprehensive spatial information features for the network, providing a more solid premise for the proposal of multi-scale modeling and feature fusion.

2.2 Semantic detail enhancement

In addition to the improvement of spatial resolution, the semantic capabilities of high-resolution remote sensing images have also been further enhanced. It is manifested as clearer edges of ground object differences, more observable features in the distinction of small categories, and the overall image is constructed based on more information. For example, for cities, it is possible to further distinguish between residences and factories, and understand their functions by observing the style of the roof. Farmland monitoring can more accurately distinguish semantic information such as the arrangement of different crop planting, the sequence of leaves, and the differences caused by different color reflections. The improvement of the above-mentioned semantic capabilities can make satellite images more "readable", but at the same time, it also increases the difficulty of semantic interpretation. Because the model not only needs to extract the geometric information of the original pixels from it, but also needs to mine deeper semantic features in order to better understand the complex environment. Therefore, in applications such as image segmentation and object classification, how to achieve the accurate extraction and effective preservation of semantic detail features of high-resolution images has also become the top priority of deep network structure and feature fusion.

3. Object detection in high-resolution remote sensing images

3.1 Analysis of the Mainstream Detection Algorithms of Faster R-CNN and YOLOv5

At present, the remote sensing target recognition applications that are widely used and have superior performance include FasterRCNN and YOLOv5. FasterRCNN completes the task in a two-stage manner, that is, first obtaining the candidate target positions using the region proposal network, and then performing feature extraction, classification and bounding box adjustment on the candidate target regions. Therefore, it can be adapted to the requirements of high-precision remote sensing target analysis applications. High accuracy is the effect achieved by adopting candidate area selection and stage processing, but the computational efficiency is slightly low. YOLOv5 regards target localization as a regression problem and directly obtains the bounding box and category of the target on the image. It is not only efficient but also convenient, and is particularly suitable for remote sensing target recognition tasks that are limited by resources and real-time applications. The objective function of YOLOv5 is composed of multi-part losses. Among them, the bounding box regression part often adopts the CIoU loss function, which is expressed as:

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

Among them, CIoU is the intersection and union ratio, ρ represents the Euclidean distance between the center of the prediction box and the center of the real box,c is the diagonal length of the minimum circumscribed rectangle of the two boxes, and αv represents the consistency penalty term of the aspect ratio.This loss enhances the fitting ability of the target position and shape, making YOLOv5 perform more stably in high-resolution remote sensing images.

3.2 Utilize the FPN structure to enhance the performance of multi-scale object detection

Ground objects in high-resolution remote sensing images often have uneven scale characteristics, ranging from large urban main roads and large-scale farms to small vehicles, street lamps or temporary buildings, etc. The coexistence of ground objects of different scales leads to challenges for ground object detection models. Traditional convolutional neural networks have difficulty in handling both the appearance of large-scale targets and the details of small-scale targets in a unified architecture. However, the feature pyramid network adopts context features to enhance the path and horizontal connection structure, achieving the feature fusion of small-scale spatial details and high-level semantic information, and has good perception ability in large-scale and wide viewing angles. The multiple feature outputs established based on FPN can ensure that the feature maps at each stage are focused on different object detection tasks, especially in object detection, the detection effect on small objects is obvious. After introducing the FPN network structure into popular detection frameworks such as YOLOv5 and FasterRCNN, the model has more stable performance and higher accuracy for complex remote sensing tasks such as building boundary extraction, road intersection recognition, and small-scale farm segmentation. To further accelerate the information flow and enhance the representation ability of features, the improvement of structures such as PANet and BiFPN can also effectively improve the effects in semantic consistency construction and context modeling.

3.3 Attention Enhancement Methods of Transformer and CBAM

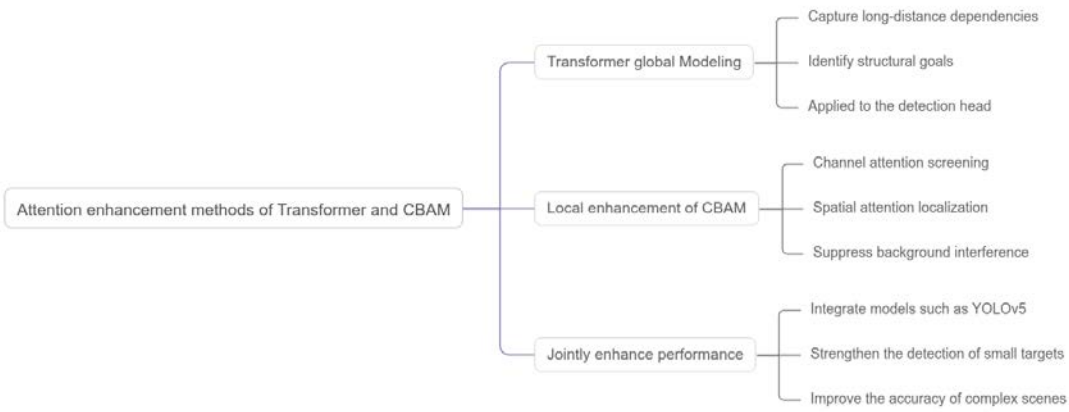

Figure 1. Shows the attention enhancement methods of Transformer and CBAM

In remote sensing image target detection, introducing the attention mechanism to assist in improving network performance has gradually become the mainstream idea. Because the global modeling structure adopted by the Transformer has an excellent ability to represent and process remote relationships, it is particularly suitable for the recognition of structural features in high-resolution images, such as large areas of houses and roads. If it is introduced into the detection

head or the feature aggregation part, it can enhance the network's understanding of the entire image. In contrast, the convolutional attention module is a shallow local enhancement mechanism. It relies on a dual-path scheme of channel attention and spatial attention to obtain useful features while removing background noise interference. The combination of the two can achieve better results in remote sensing environment scenarios, especially in aspects such as small targets, occluded targets, and low-contrast targets, where significant detection improvement effects have been achieved. Nowadays, Transformer and CBAM are respectively introduced into the YOLOv5 and RetinaNet detection networks to achieve rapid detection under complex remote sensing human figures. This approach optimizes the perception performance of the model at the global level and also optimizes the local areas, playing an important role in the intelligent recognition of high-resolution remote sensing images (see Figure 1).

3.4 Lightweight Detection Model of MobileNet and Pruning Technology

Table 1. Application Characteristics of Mobilenet and Pruning Techniques in Remote Sensing Target Detection

Technical name	Core principle	Advantage	Application scenarios
MobileNet	Use depth-separable convolution	It greatly reduces the number of parameters and computational load of the model, and increases the detection efficiency	Mobile devices or drones
Pruning of the passage	Remove the channels with smaller weights in the convolutional layer	Reduce the redundancy of the model and improve the reasoning efficiency.	The structure was streamlined before the model was deployed
Network pruning	Perform global deletion or reconstruction of the model by layer or structure	According to the requirements of the hardware platform, reduce or enlarge the data model.	Integrated remote sensing data and situation monitoring system.
Lightweight detection framework	Take MobileNet as the backbone network and combine pruning to optimize the structure	By reducing the model size, the detection speed is also improved.	Unmanned aerial vehicle (UAV) remote sensing, emergency disaster relief, edge computing environment.

Due to the fact that high-resolution remote sensing images contain a large amount of information and the models are relatively complex, the previous detection networks were unable to achieve efficient detection with limited hardware. To alleviate this problem, MobileNet and its model pruning methods can be used as favorable tools to reduce the load of heavyweight detection models. MobileNet uses depth-separable convolution to divide normal convolution operations into single-channel convolution operations and single-point convolution operations, significantly reducing the number of parameters and computational load of the model, and increasing detection efficiency. It is suitable for mobile devices or drones. Model pruning simplifies the model scale from different perspectives. For example, channel pruning is to remove channels with smaller contributions to achieve partial simplification, and network pruning is to modify the entire network

structure to achieve the goal of comprehensive simplification. Using MobileNet as the backbone network of the detection network like YOLOv5, combined with the corresponding pruning technology, can improve efficiency while ensuring the detection accuracy rate. This lightweight architecture is designed for disaster monitoring, low-power device deployment and on-site remote sensing emergencies, providing support for the practical application of intelligent remote sensing analysis technology (see Table 1).

To sum up, while high-resolution remote sensing images improve the accuracy of ground object recognition, they also pose higher requirements for target detection and image segmentation algorithms.

4. Optimization of image segmentation algorithms in high-resolution remote sensing images

4.1 Hollow convolution and jump connection enhance the expression of spatial details

In high-resolution remote sensing images, the object boundaries are clear, the shapes are diverse, and the scale differences are large, which puts forward higher requirements for the spatial details and semantic description of the image segmentation model. In the multi-layer sampling process of traditional convolution operations, it is easy to cause boundary blurring and information loss, especially when it comes to small targets or terrain boundaries, the effect is limited. In this regard, dilated convolution is introduced into the segmentation network as a way to expand the window. That is, without increasing the computing time, the simulation of data in the remote environment is improved through sparse convolution to enhance semantic consistency and the ability of global structure recognition. Meanwhile, jump connection, as a core component of structures such as U-Net, breaks through the semantic gap between shallow and deep features. Enable the detailed information such as edge and corner lines obtained by the encoder to be precisely restored during the decoding process. The combination and application of the two in networks such as DeepLabv3+ can effectively improve the cutting accuracy of complex targets such as the corner lines of buildings, road intersections and field gullies in remote sensing images.

4.2 Multimodal fusion enhances segmentation performance in complex scenarios

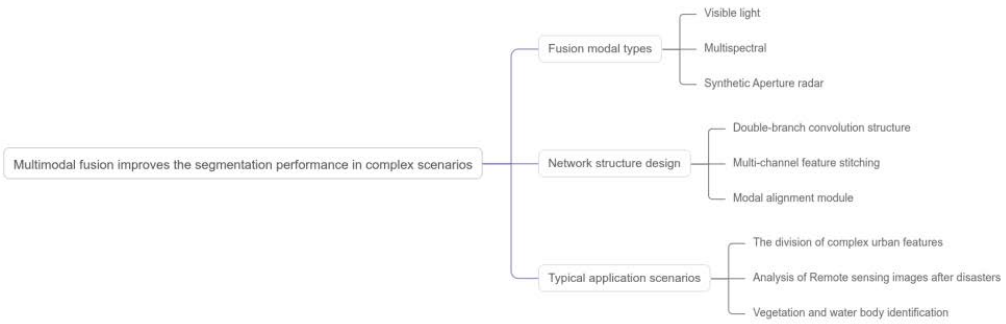

Figure 2. shows that multimodal fusion enhances the segmentation performance in complex scenarios

The problem of high-resolution remote sensing image segmentation has received extensive attention because it can effectively enhance the feature detection ability of complex ground objects. Generally, various types of images such as visible light, multispectral and synthetic aperture radar images are adopted to display the complementarity of surface structure, plant conditions or

penetration capabilities at different levels. To handle such information sources, a dual-branch convolutional neural network is usually adopted to learn the characteristics of different patterns. Multi-channel feature stitching is carried out in the middle layer or output stage, and a modal alignment module is added to ensure semantic consistency. This structure maintains the independence of image expression in different patterns and effectively integrates the features in different patterns. In practical applications, for complex geographical scenarios such as the identification of urban complex ground features and road intersection areas; In the process of remote sensing image analysis after disasters, SAR images can break through the occlusion of smoke and clouds. Furthermore, in the segmentation process of lush vegetation and water areas, etc., the optical resolution of multispectral further improves the classification accuracy (See Figure 2).

4.3 The Transformer module enhances the global semantic modeling capability

In the task of high-resolution remote sensing image cutting, the Transformer module effectively improves the model's ability to construct global semantics through the self-attention mechanism, making up for the deficiencies of traditional convolutional networks in field of view and background understanding. Specifically, structures such as SwinTransformer are used as the backbone structure of the encoder. Through the window-level attention mechanism, cross-view-domain semantic connection is achieved without reducing the operation rate, thereby enhancing the model's global understanding ability of the entire geographic object. Embedding the Transformer module into the existing U-Net or DeepLab models can integrate the middle and high-level semantic information with the low-level spatial details, further enhancing the boundary localization and internal consistency capabilities of the model. Guided and supervised by using the attention map, the attention of the model is focused on the important areas to improve the classification accuracy and cutting accuracy of the model. Relative coordinate coding is adopted to enhance the spatial structure recognition ability of the model to adapt to the problem of obvious layout of geographic objects in remote sensing images. In practical work, this method can obtain higher mIoU values on various remote sensing data, and can demonstrate better global recognition ability and detail reconstruction ability in the segmentation of ground objects with similar semantics but uneven distribution, such as forest areas, buildings, and water areas.

4.4 Edge detection branches enhance the accuracy of target contour recognition

Table 2. Functional Role of Edge Detection Branches in Spatial Structure Restoration

Objective and task	Description of auxiliary functions
Spatial boundary restoration	The boundary surface joint refers to the geometric edge line to assist in analyzing and solving the boundary misalignment.
Multi-scale structure alignment	Guidance for maintaining spatial consistency at different scales of the backbone network.
Contour semantic enhancement	Fusing the main segmentation results can enhance the accuracy of interpreting the boundary categories of ground features.
Typical applicable objects	Regular structural features refer to roads, water systems, cultivated land and building surfaces, etc.

For the task of high-resolution remote sensing image segmentation, the model needs to be able to distinguish multiple ground object categories and accurately restore their shapes and boundaries. Traditional deep learning methods often lead to edge blurring and structural deformation during feature processing, which affects the accuracy of spatial positioning and the continuity of semantics.

In order to enhance the sensitivity of the model to edge lines, methods in recent years have begun to introduce edge detection modules externally to adjust the structural positioning and thereby improve the effect of image segmentation. This module can be achieved by simultaneously capturing edge characteristics and cooperating with the main module to take a further step in the decoding process. Another example is the adoption of edge guidance strategies, that is, the use of edge cross-entropy or structural dissimilarity loss, etc., which can make the model more flexible in response to the changes of edge lines. The adoption of attention mechanisms, such as gradient guidance in egnet, can enhance the effect of edge perception and further increase the performance of structural edges (see Table 2).

This table summarizes the core role of the edge detection branch in the segmentation of high-resolution remote sensing images. Besides the function of extracting boundary lines, it also includes necessary functional AIDS such as restoring spatial structure, multi-scale simultaneous operation, and enhancing boundary semantic representation. When integrated into large network branches, the edge detection branch can effectively enhance the model's structural restoration ability and segmentation accuracy on targets with clear geometric contours such as buildings, and is suitable for remote sensing scenarios that require high levels of precise boundary detection.

5. Conclusion

To further improve the main technical ideas, advantages and disadvantages of object detection and image segmentation in high-resolution remote sensing images, specific optimization ideas in aspects such as structural improvement, feature fusion, attention mechanism and edge perception are proposed. For remote sensing target detection, F-PN, Transformer and lightweight structure are applied to improve the accuracy and execution speed of multi-scale target recognition; For remote sensing image segmentation, caved convolution, modal fusion and edge branching design are adopted to enhance the coping ability in complex environments and fine structures. Further exploration and research are conducted in combination with large-scale remote sensing models, self-supervised learning, cross-modal collaboration and other aspects to improve the accuracy level, universality and application degree of intelligent remote sensing interpretation.

Reference

[1] *Gong H, Sun Q, Fang C, et al. TreeDetector: Using Deep Learning for the Localization and Reconstruction of Urban Trees from High-Resolution Remote Sensing Images. Remote Sensing, 2024, 16(3):22.*

[2] *Wang M, Shen L. High-Resolution Remote Sensing Imagery for the Recognition of Traditional Villages. Journal of Architectural Research and Development, 2024, 8(1):75-83.*

[3] *Sun Y, Chen J, Huang X Z H. Multi-Level Perceptual Network for Urban Building Extraction from High-Resolution Remote Sensing Images. Photogrammetric Engineering & Remote Sensing: Journal of the American Society of Photogrammetry, 2023, 89(7):427-434.*

[4] *Fan L, Zeng C, Li Y, et al. GRC-Net: Fusing GAT-Based 4D Radar and Camera for 3D Object Detection. SAE International Journal of Advances and Current Practices in Mobility, 2024(5):6.*

[5] *Cuevas E, Héctor Becerra, Luque A, et al. Fast multi-feature image segmentation. Applied Mathematical Modelling, 2021, 90(5):742-757.*

[6] *Zou, Y. (2025). Design and Implementation of a Cloud Computing Security Assessment Model Based on Hierarchical Analysis and Fuzzy Comprehensive Evaluation. arXiv preprint arXiv:2511. 05049.*

[7] Su H, Luo W, Mehdad Y, et al. Llm-friendly knowledge representation for customer support[C]//Proceedings of the 31st International Conference on Computational Linguistics: Industry Track. 2025: 496-504.

[8] Liu, B. (2025). Design and Implementation of Data Acquisition and Analysis System for Programming Debugging Process Based on VS Code Plug-In. arXiv preprint arXiv: 2511. 05825.

[9] Zhu, P. (2025). The Role and Mechanism of Deep Statistical Machine Learning In Biological Target Screening and Immune Microenvironment Regulation of Asthma. arXiv preprint arXiv:2511. 05904.

[10] Sun, Jiahe. "Research on Sentiment Analysis Based on Multi-source Data Fusion and Pre-trained Model Optimization in Quantitative Finance." (2025).

[11] Chang, Chen-Wei. "Compiling Declarative Privacy Policies into Runtime Enforcement for Cloud and Web Infrastructure." (2025).

[12] F. Liu, "Transformer XL Long Range Dependency Modeling and Dynamic Growth Prediction Algorithm for E-Commerce User Behavior Sequence," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-6.

[13] F. Liu, "Architecture and Algorithm Optimization of Realtime User Behavior Analysis System for Ecommerce Based on Distributed Stream Computing, " 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-8.

[14] Q. Hu, "Research on Dynamic Identification and Prediction Model of Tax Fraud Based on Deep Learning," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-6.

[15] D. Shen, "Complex Pattern Recognition and Clinical Application of Artificial Intelligence in Medical Imaging Diagnosis," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-8.

[16] Wu Y. Optimization of Generative AI Intelligent Interaction System Based on Adversarial Attack Defense and Content Controllable Generation. 2025.

[17] Sun J. Quantile Regression Study on the Impact of Investor Sentiment on Financial Credit from the Perspective of Behavioral Finance. 2025.

[18] Wang Y. Application of Data Completion and Full Lifecycle Cost Optimization Integrating Artificial Intelligence in Supply Chain. 2025.

[19] Chen M. Research on Automated Risk Detection Methods in Machine Learning Integrating Privacy Computing. 2025.

[20] Wei, X. (2025). Deployment of Natural Language Processing Technology as a Service and Front-End Visualization. International Journal of Engineering Advances, 2(3), 117-123.