Journal of Artificial Intelligence Practice (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2024.070105
ISSN 2371-8412 Vol. 7 Num. 1

Vehicle Target Detection Algorithm Based on Improved Faster R-CNN for Remote Sensing Images

Yiran Yang

Beijing University of Civil Engineering and Architecture, Beijing, 102616, China

Keywords: Remote sensing imagery; target detection; vehicles; feature extraction; attention mechanism

Abstract: Aiming at the problems that remote sensing image vehicle targets are susceptible to complex background interference, multi-scale differences, and difficulties in detecting small targets, this paper proposes a remote sensing image vehicle target detection algorithm based on improved Faster R-CNN. In this paper, based on the framework of Faster R-CNN, firstly, a multi-scale feature extraction network (EM-FPN) is designed by using the FPN structure and ResNet50 network, so that the network extracts rich target features; secondly, the ECA attention mechanism is introduced, so that the feature extraction network focuses on the target features, suppresses the interference of irrelevant background information, and constructs the multirate dilated convolution module (MDCM) to enhance the network's ability to perceive the contextual information of the target; finally, ROI Align is used instead of ROI Pooling to reduce the feature quantization error. The experimental results prove that the accuracy of the proposed algorithm reaches 88.6%, which can effectively detect vehicle targets in remote sensing images.

1. Introduction

With the continuous development of satellite remote sensing technology, the application of remote sensing images in the fields of traffic monitoring, urban planning, and environmental monitoring is becoming more and more widespread [1-2]. Vehicle detection, as one of the important tasks of remote sensing image processing, automatically identifies and locates vehicle targets through remote sensing images, which is of great significance for realizing traffic management, intelligent transportation system, and urban planning and management. However, remote sensing image vehicle targets are susceptible to complex background interference, multi-scale differences, small target detection difficulties, to achieve accurate detection is very difficult.

In recent years, with the rapid development of computer technology, target detection methods based on deep learning have become the mainstream methods for remote sensing image vehicle detection, which can be mainly summarized into two types: single-stage detection algorithm (One Stage) and two-stage detection algorithm (Two-Stage). The One Stage detection algorithm predicts the bounding box and category probabilities of all targets in one step, which is characterized by a faster speed than the Two-Stage detection algorithm, but with a loss of accuracy, representative algorithms such as SSD [3], YOLO series [4-5], and so on. Two-stage detection algorithms generate candidate regions through feature extraction by convolutional neural networks, and then localize

and classify the candidate regions by classifiers, representative algorithms such as R-CNN [6] series, RetinaNet [7] and so on. However, the existing mainstream detection algorithms are often ineffective in detecting vehicle targets in remote sensing images, and many scholars modify and optimize the existing mainstream algorithms based on them to improve the detection accuracy. qu et al [8] further strengthened the target features by designing a bi-directional multi-scale fusion feature extraction network fusing feature information at different levels and constructing a global high-efficiency attention module with an average detection accuracy of 92%; Zhao et al [9] used YOLOv5 as the basic architecture, extracted different layers of feature information by designing a multiple pyramid network, and used data enhancement and K-means to significantly improve the accuracy of small-target vehicle detection; Zhang et al [10] proposed a coarse-to-fine target detection framework, which progressively enhances the sample representations with the training samples, and ultimately achieves a better detection effect; Jia et al. et al [11] used the K-means++ algorithm to optimize the initial clustering points to obtain a more suitable anchor frame for the target size, used the CA attention mechanism and BiFPN structure to enhance the target feature extraction ability of the network, and increased the detection head of the small target, which effectively improved the detection effect.

In this paper, we analyze the related research and literature on the subject, and propose a remote sensing image vehicle target detection algorithm based on improved Faster R-CNN for remote sensing image vehicle target detection difficulties, as shown in Fig. 1. Aiming at the problem of multi-scale differences in remote sensing image vehicle targets, a multi-scale feature extraction network (EM-FPN) is designed based on the FPN structure, which integrates different levels of deep and shallow features to improve the network's ability to extract target features; aiming at the problem of remote sensing image vehicle targets being susceptible to background interference, the ECA attention mechanism is introduced to highlight the target features and reduce the interference of irrelevant background information; aiming at the problem of difficulty in detecting small targets of remote sensing image vehicles, an improved Faster R-CNN algorithm is constructed based on Faster R-CNN. For the problem of vehicle small target detection difficulty in remote sensing images, a multi-rate dilated convolution module is constructed to enhance the context information perception capability of small targets, and ROI Align is used instead of ROI Pooling to reduce the impact of feature quantization error on small target detection and improve the target detection accuracy.

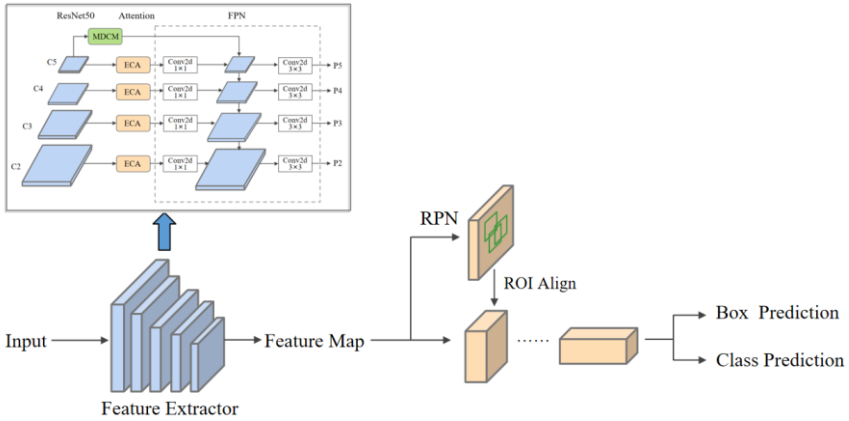

Figure 1: Improvement of Faster R-CNN algorithm

2. Methods

2.1 A Design of multi-scale feature extraction network structure (EM-FPN)

Faster R-CNN uses VGG16 for feature extraction, which improves the detection performance of the model by increasing the network depth. However, remote sensing image vehicle target size is small, contains less feature information, image after many convolution and pooling operation deep features are easy to produce aggregation points, difficult to detect.ResNet [12] network through the construction of the residual block, in order to avoid the loss of the deep feature information on the premise, but also to a certain extent to solve the network degradation problem, and effectively improve the ability of the network feature extraction. Through the experimental comparison of ResNet18, ResNet50, and ResNet101, ResNet50 with better experimental results is selected as the feature extraction network in this paper, and the structure is shown in Figure 2.

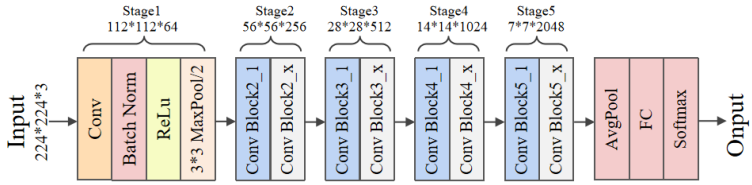

Figure 2: ResNet50 structure

Based on the FPN structure, this paper proposes a multiscale feature extraction network, notated as EM-FPN, as shown in Fig. 3. Firstly, this paper uses ResNet50 as the feature extraction network to extract the multilevel feature maps C2, C3, C4 and C5; secondly, the multilevel feature maps C2-C5 are inputted into the FPN structure after the ECA attention mechanism with 1*1 convolution, and the top-level feature map C5 is processed by the Multi-Rate Dilated Convolution Module (MDCM) and inputted into the FPN structure with the other feature maps; the feature map The deep and shallow features of the feature maps are fully fused, and the input feature maps P2, P3, P4 and P5 are finally obtained.The multiscale feature extraction network proposed in this paper is able to fuse the deep semantic features of different layers with the shallow detailed features, so that the network obtains richer target features.

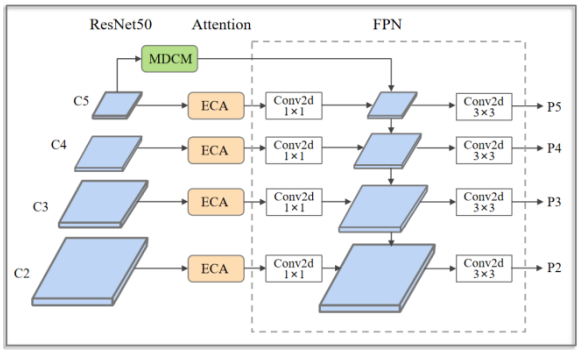

Figure 3: EM-FPN structure

2.2 B Introduction of the ECA attention mechanism

Complex backgrounds often exist in remote sensing images, which can easily cause vehicle targets to be difficult to distinguish from the background, resulting in detection difficulties. The attention mechanism can make the network focus on important target features and reduce the interference of irrelevant background. In this paper, we introduce the ECA attention mechanism [13], which adopts a local cross-channel interaction strategy without dimensionality reduction. Under the

premise of avoiding the impact of dimensionality reduction on the learning effect of channel attention, the network can focus on the important channel information more effectively without introducing significant computational burden, through adaptively weighting the features within each channel, the structure of which is shown in Figure 4.

The ECA module first compresses the input feature map through global average pooling (GAP) operation to obtain a feature map of size 1*1*C; secondly, it calculates the 1D convolution kernel size k according to the number of channels, as shown in equation (1); then it uses 1D convolution to calculate the channel weights of the pooled feature map and maps the weights to the (0,1) interval using a sigmoid activation function; finally, it maps the input feature map is multiplied with the computed weights to obtain the weighted new feature map.

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

where k denotes the convolutional kernel size, $\psi(C)$ is the mapping from C to k, C denotes the number of channels, and γ and b are hyperparameters used to change the scaling relation and bias of the C to k mapping.

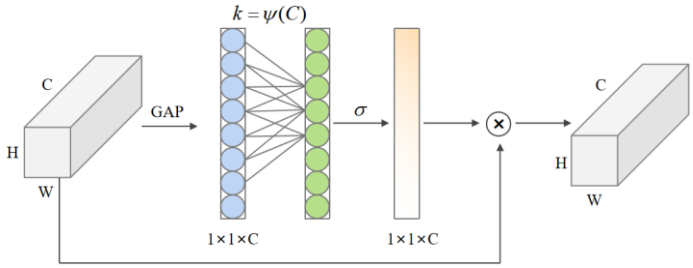

Figure 4: ECA Attention Mechanism

2.3 Construction of multi-rate dilated convolution module (MDCM)

Vehicle small targets in remote sensing images have small size and limited feature expression, relying on the local information of the target is not enough to accurately recognize and detect them. It has been experimentally demonstrated that dilated convolution [14] can enable the network to better capture the contextual relationship of the target by expanding the receptive field of the convolution kernel. Dilated convolution can effectively aggregate the global feature information of the image without losing the resolution of the feature map, which effectively improves the target detection effect.

In this paper, the multi-rate dilated convolution module (MDCM) is proposed, as shown in Fig. 5. In this paper, the dilated rate r is set to 1, 2, and 3, respectively, which enables the network to obtain the contextual information of the target in a larger range while avoiding the checkerboard effect caused by the convolution of a single-size dilated. In particular, for small targets with smaller sizes, it is able to acquire more contextual information, thus compensating for its own problem of fewer features. The multi-rate dilated convolution module performs dilated convolution of input feature maps with different dilated rates to obtain feature maps with different sensory fields, and after splicing using Concat, the final 1*1 convolutional layer is used for information fusion and reducing the number of channels.

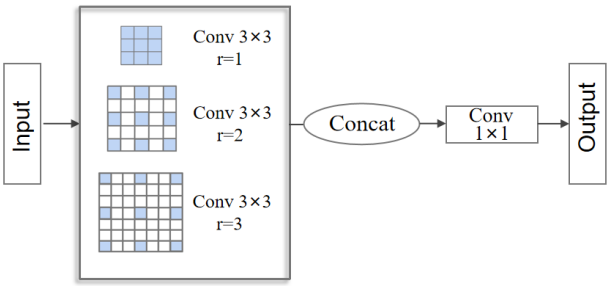

Figure 5: Multi-rate Dilated Convolution Module Architecture

2.4 Using ROI Align

Faster R-CNN uses ROI Pooling to transform the input feature map into a fixed-size feature representation, however, ROI Pooling is prone to pixel loss and positional bias in the process of quantization of candidate box bounding. For small targets, the positional deviation only produces a small error, both of which can lead to a large impact on the accuracy of the final target detection. Therefore, in this paper, we use the method of ROI Align, which adopts the bilinear interpolation method to obtain the image values on the pixels with floating-point coordinates to avoid the quantization process. The specific process is in the pooling process will be divided into four parts of each bin evenly, take its center position for the sampling point, using the method of bilinear difference to calculate the value of the sampling point and thus carry out the maximum pooling operation, as shown in Figure 6.

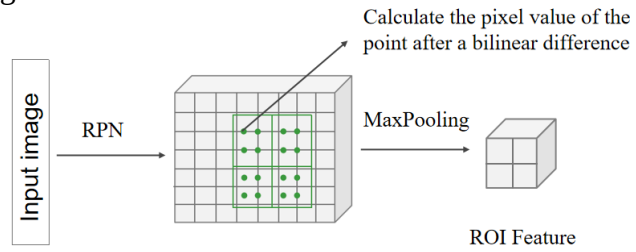

Figure 6: ROI Align Principle

3. Experiment

3.1 Data set and environment configuration

In this study, vehicle images containing vehicle images were selected from the publicly available remote sensing datasets NWPU VHR, DIOR and DOTA, and 3895 images were obtained. The dataset images were labeled with the location and category information of the targets using LabelImg, and the training set, test set and validation set were assigned in the ratio of 7:2:1, so as to construct a target dataset of remote sensing images vehicles. The dataset includes nine categories: car, truck, tractor, camping_car, van, vehicle, plane, boat, and pick-up. All experimental environments in this study are based on Ubuntu 16.04 operating system with Intel(R) Xeon(R) CPU E5-2620, NVIDIA GeForce RTX 3090 graphics card, and Python 3.6, Pytorch 1.8.0, and CUDA 11.1 environments for model construction, training, and testing. The initial learning rate of this paper is 0.01, momentum decay is 0.0005, momentum parameter is 0.9, batch is 16, and iteration number is 400.

3.2 Experimental result

We conducted experiments on the test set and Figure 7 shows some of the results.

Figure 7: Partial test results

In order to verify the detection effect of the algorithm proposed in this paper, we use the more mainstream four feature extraction network models VGG16, MobileNetV2, ResNet50, ResNet101 . The multiscale feature extraction network proposed in this paper as the backbone network of the Faster R-CNN to conduct comparative experiments, respectively.We use the Mean Average Precision (mAP) as the evaluation index to test the effectiveness and accuracy of the experimental results, and the results are shown in Table 1.

Table 1: Comparison of different backbone networks

Backbone	VGG16	MobileNetV2	ResNet50	ResNet101	Ours
mAP(%)	73.5	73.6	81.7	83.2	85.6

Through Table 1, it can be found that the detection mAP of the Faster RCNN network built by the multiscale feature extraction network structure proposed in this paper is significantly improved compared with the rest of the Faster RCNN networks built by VGG16, MobileNetV2, ResNet50, and ResNet101. This result shows the superiority of the multiscale feature extraction network proposed in this paper compared to other feature extraction networks.

In order to verify the contribution of the proposed improvement module in this paper, we analyzed the contribution values of different improvement methods by performing ablation tests on the test dataset, as shown in Table 2.

Table 2: Comparison of results of different optimization methods

Experiment	ResNet+FPN	ECA	MDCM	ROI Align	maP(%)
1	√				85.64
2	√	√			87.02
3	√	√	√		88.15
4	√	√	√	√	88.61

Through Table 2, we can find that the improvement methods proposed in this paper are able to improve the detection results of this paper to different degrees. Through the superposition of the above improvement methods, we find that the detection of mAP is significantly improved, and the final mAP reaches 88.6%, respectively, which verifies the effectiveness of the proposed improvement algorithm in this paper.

4. Conclusion.

In this paper, a remote sensing image vehicle target detection algorithm based on improved Faster R-CNN is proposed using remote sensing image vehicle target as the detection object. Aiming at the problem that remote sensing image vehicle targets are susceptible to complex background interference, multi-scale differences, and difficulties in detecting small targets, by combining the ResNet and FPN structures, introducing the ECA attention mechanism, constructing the Multi-Rate dilated Convolutional Module (MDCM), and using the ROI Align algorithm, the network effectively improves the feature extraction capability of the remote sensing image vehicle targets, thus improving the network's detection Effect. The experimental results show that the algorithm in this paper achieves better detection effect, can meet the needs of remote sensing image vehicle target detection task, and provides technical ideas for traffic management, urban planning and other practical work.

References

[1] Li K., Wan G., Cheng G., Meng L., Han J. (2020) Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark 2019. ISPRS J. Photogramm. Remote Sens. 159, 296–307.

[2] Lv Y., Zhu H., Meng J., Cui C., Song Q. (2022) A review and adaptability study of deep learning models for vehicle detection based on high-resolution remote sensing images. Remote Sensing for Natural Resources, 34(04):22-32.

[3] Liu W., Anguelov D., Erhan D., et al. (2016) Ssd: Single shot multibox detector. European conference on computer vision, 21-37.

[4] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection [J]. arXiv preprint arXiv: 2004.10934, 2020.

[5] Zhu, X., Lyu, S., Wang, X., Zhao, Q. (2021) TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. Proceedings of the IEEE/CVF international conference on computer vision, 2778-2788.

[6] Girshick, R., Donahue, J., Darrell, T., Malik, J., Berkeley, UC. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 580-587.

[7] Ren, S., He, K., Girshick, R., Sun, J. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[8] Qu H., Wang M., Chai R. (2023) Efficient Vehicle Detection in Remote Sensing Images with Bi-directional Multi-scale Feature Fusion. Computer Engineering and Applications, 1-13.

[9] Zhao Q., Yang Y. (2023) Small object detection algorithm for lightweight remote snesing vehicles with multiple pyramids. Electronic Measurement Technology, 46(13):88-94.

[10] Zhang C., Lam K M., Wang Q. (2023) Cof-net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-17.

[11] Tian Z., Huang J., Yang Y., Nie W. (2023) KCFS-YOLOv5: A High-Precision Detection Method for Object Detection in Aerial Remote Sensing Images. Applied Sciences, 13(1): 649.

[12] Targ S., Almeida D., Lyman K. (2016) Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

[13] Wang Q., Wu B., Zhu P., et al. (2020) ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 11534-11542.

[14] Yu F., Koltun V. (2015) Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.