Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 7, No. 2, 2024

Research on Improved Algorithm for Small Object Detection in Intelligent Surveillance Video based on YOLOv7

Zhiwei Wang [1, *], Min Wang [2]

[1] School of Cyberspace Security, Chengdu University of Information Technology (CUIT), Chengdu, Sichuan, China
[2] Dept of Cyber space Security Academy, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, China
[*] **Corresponding autho**r: Zhiwei Wang (Email: wzw19980331@163.com)

Abstract: In order to address the issue of small objects being difficult to detect effectively in intelligent surveillance videos, this study proposes an improved scheme for the YOLOv7-tiny algorithm. This scheme integrates the Convolutional Block Attention Module (CBAM) into YOLOv7-tiny, effectively enhancing the model's feature extraction and small object detection capabilities in complex backgrounds, thereby improving the overall detection precision. Experimental evaluations indicate that the improved algorithm shows enhanced performance in specific small object detection tasks, achieving an accuracy of 85.6%, a recall rate of 85.2%, and a mean average precision (mAP) of 90.2%. These results demonstrate the effectiveness and practical value of the improved scheme in enhancing the performance of YOLOv7-tiny in small object detection tasks.

Keywords: Intelligent Video Surveillance; YOLOv7-tiny; Object Detection; Small Object; Attention Mechanism.

1. Introduction

In the field of intelligent video surveillance, small object detection is an extremely challenging task that requires the system to accurately identify and track objects in the video that are small in size and may be difficult to recognize due to distance or other factors. [1] In recent years, with the rapid development of deep learning technologies, significant progress has been made in the research of small object detection, and many new methods and technologies have been proposed. [2] Deep learning frameworks, especially Convolutional Neural Networks (CNNs), have become the mainstream method for small object detection. Algorithms like YOLO, SSD, and Faster R-CNN, through deep networks, learn the feature representations of objects, achieving efficient object detection.[3] Wang et al. enhanced the recognition capability for small objects by introducing multi-scale detection mechanisms or improving the feature extraction network.[4]

Furthermore, the attention mechanism has also shown great potential in small object detection. By focusing on key feature areas, models can better differentiate between the background and small objects, thus improving detection accuracy. [5] For example, Wang et al. used an adaptive attention module to dynamically adjust the focus of the network to more effectively capture the details of small objects.[6]

Nonetheless, small object detection still faces a series of challenges, including small size of objects, significant appearance changes, complex backgrounds, and changes in lighting. These factors increase the difficulty of detection and limit the performance of existing methods.[7]

The latest research trends include using Generative Adversarial Networks (GANs) to enhance the visual quality of small objects and exploring more efficient network architectures and training strategies to improve detection speed and accuracy.[8] Researchers are also exploring the integration of more context information and prior knowledge to assist in small object detection, as well as developing more complex multi-task learning frameworks to simultaneously perform object detection, classification, and tracking.[9]

Overall, the field of small object detection in intelligent video surveillance is rapidly evolving, with new technologies and methods being proposed to address existing challenges. As technology advances, future intelligent video surveillance systems will be able to detect small objects more accurately and efficiently, providing stronger support for security monitoring, traffic management, and other fields.[10]

Research indicates that the YOLO series of algorithm frameworks support end-to-end detection and have shown excellent performance across multiple detection domains. As the first single-stage object detection algorithm to utilize deep learning, YOLO has led a new direction in object detection technology. To date, the YOLO series has evolved to its seventh edition, covering versions from YOLOv1 to YOLOv5. The latest YOLOv7 model has significant improvements in speed, accuracy, and the number of model parameters, and is easier to deploy on different devices.[11]

The YOLOv7 model includes four different scale variants: v7s, v7m, v7tiny, and v7x, to accommodate different computing and application needs. Experimental results show that, except for v7-tiny, the v7m and v7x models require longer training times and produce larger weight files. Specifically, the training time, mean average precision (MAP), and model weight size of YOLOv7 and YOLOv7-tiny demonstrate their performance and efficiency. YOLOv7 and YOLOv7x, having more parameters, are larger in terms of training time and weight file size. [12] In contrast, although YOLOv7s has a MAP that is not significantly different from the other variants for each category, it has a clear advantage in terms of efficiency and lightweight.

Table 1. Experiment Time and Weight Size

	YOLOv7-tiny	YOLOv7
Time/hours	0.843	15.806
Size/mb	14.4	159.4
MAP@0.5/%	87.4	87.6

YOLOv7-tiny, as the lightest version in the series, is renowned for its exceptional detection speed. It is particularly suitable for embedding into devices with limited computing power, storage space, and memory, making it highly applicable to monitoring equipment. Therefore, choosing YOLOv7-tiny as the basic algorithm framework for this research is based on its excellent performance and applicability considerations. However, YOLOv7-tiny's performance in detecting small objects is not satisfactory, which necessitates improving its detection rate. To address this, the study has enhanced YOLOv7-tiny by incorporating attention mechanism modules and adding detection layers specifically for small objects. The improved algorithm has shown good results in experiments, particularly suiting the detection needs in video surveillance scenarios for specific situations.

2. Improving the YOLOv7-tiny Model for Small Object Detection

2.1. YOLOv7-tiny Basic Algorithm Framework

YOLOv7-tiny is the smallest network in the YOLOv7 family. The network architecture is shown in Figure 1.

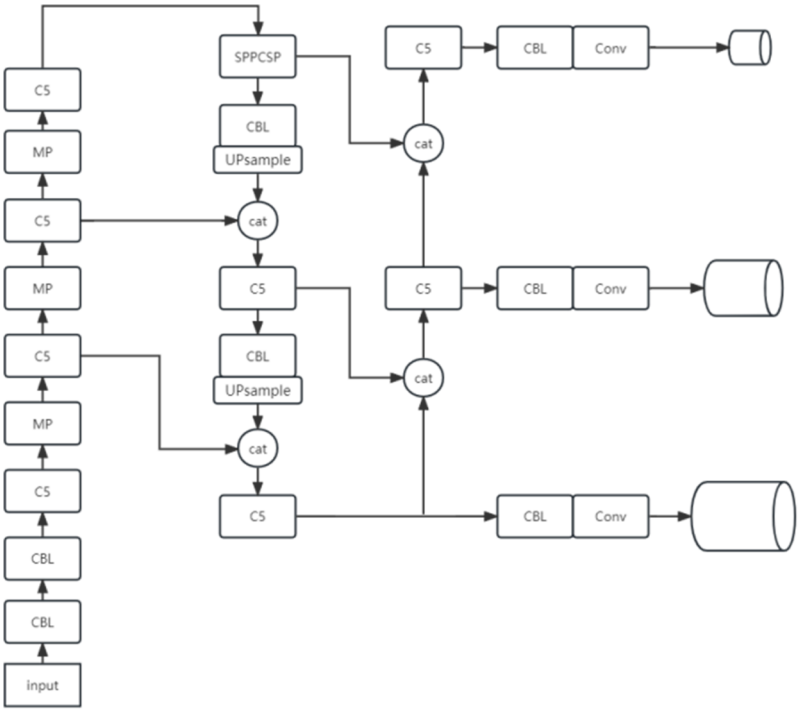

Figure 1. YOLOv7-tiny Network Structure Diagram

Its backbone network adopts a multi-branch stacking structure for feature enhancement, which allows for effective feature extraction while reducing computational requirements. After the input image undergoes feature extraction through the backbone network, features layers obtained after 3, 4, and 5 times of downsampling are used for object prediction. At the last feature layer of the backbone network, there is an SPP (Spatial Pyramid Pooling) module with a CSP (Cross Stage Partial) structure, which includes maximum pooling layers of different sizes. This module enables the network to learn more effective features from different receptive fields, allowing the deep convolutional network to extract richer semantic information. The neck utilizes a PANet structure, which merges the upsampled deep feature maps with the shallow feature maps to enhance the semantic information of the shallow network. It also merges the downsampled shallow feature maps with the deep feature maps to complement the detailed information of the deep feature maps. The detection head, which shares weights, is used for final object classification and localization. Each layer uses three differently shaped anchor boxes, responsible for detecting objects of different shapes. Lastly, non-maximum suppression is used to filter out multiple anchor boxes' duplicate predictions for the same target, and re-parameterization technology is applied. This includes different network branches with convolutional kernels of different sizes on the same layer. The backbone network extracts features from the input image, then the neck merges features of different scales. The detection head predicts the position and type of objects, ultimately producing the detection results. The CSPSPP structure significantly increases the receptive field, isolating the most crucial contextual features, while the multi-branch

stack module, by controlling the shortest and longest gradient paths, enables the network to learn more features with greater robustness.

2.2. YOLOv7-tiny Improved by Introducing Attention Mechanism CBAM

The CBAM (Convolutional Block Attention Module) self-attention mechanism has the following advantages:

(1) High Degree of Lightweight Design: The CBAM module does not contain a large number of convolutional structures internally, only a few pooling layers and feature fusion operations. This structure avoids the heavy computation brought by convolutional multiplications, resulting in a module with low complexity and minimal computational demand. Experiments have shown that adding the CBAM module to lightweight models can bring about stable performance improvements. Compared to the slight increase in computational load, the introduction of CBAM offers a high cost-effectiveness ratio.

(2) Strong Universality: The structural characteristics of CBAM ensure its strong universality and high portability, which is mainly reflected in two aspects: on one hand, the CBAM module, based on pooling operations, can be directly embedded after convolutional operations, meaning that this module can be added to traditional neural networks like VGG, as well as to networks containing shortcut connection-based residual structures, such as ResNet50 and MobileNetV3; on the other hand, CBAM is equally applicable to both object detection and classification tasks, and it can achieve significant performance improvements in both detection and classification accuracy across datasets with different feature characteristics.

(3) Effective Impact: Traditional attention mechanisms in convolutional neural networks focus more on analyzing the channel domain, limited to considering the interactions between feature map channels. CBAM starts from both channel and spatial domains, introducing spatial and channel attention as two dimensions of analysis, thereby achieving a sequential attention structure from channel to space. Spatial attention enables the neural network to pay more attention to the pixel areas in images that are crucial for classification while ignoring irrelevant regions. Channel attention, on the other hand, deals with the distribution of feature map channels. Attention allocation to both dimensions enhances the effect of the attention mechanism on improving model performance.

The CBAM network structure is shown in Figure 2.

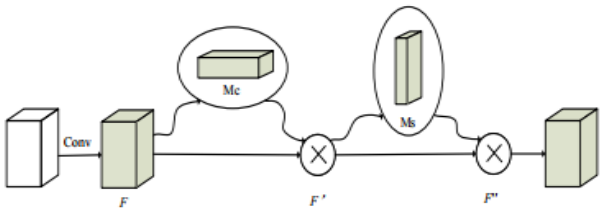

Figure 2. CBAM network structure diagram

(1) Feature extraction network: This is the starting stage of CBAM, responsible for extracting raw features from the input image. It is typically a combination of convolutional layers, pooling layers, and activation functions, aimed at capturing the basic patterns and structures of the image. The output of the feature extraction network is a set of multi-dimensional feature maps, which are then sent to the subsequent attention modules for further processing.

(2) Classification and regression module: This part is usually located at the end of the network, and its task is to perform classification or regression tasks based on the features refined by the attention module. This module can be designed according to specific application scenarios, such as using fully connected layers (FC layers) for image classification or bounding box regression in object detection tasks.

(3) Fusion attention module (Rpn network): The core of CBAM lies in its attention module, which is further divided into two sub-modules:

a. Channel Attention Module: The goal of this module is to identify which channels are important, that is, to apply attention on the channel dimension of the feature map. It evaluates the importance of each channel by analyzing the global information of each channel, allowing the network to focus on more informative feature channels.

b. Spatial Attention Module: Following the channel attention module, the spatial attention module focuses on which spatial areas of the feature map are important. It determines the focus of attention by observing the feature response at different spatial positions, further refining the feature map, enabling the network to concentrate on key parts of the image.

The combination of these two attention modules allows CBAM to adjust the feature map in a detailed manner, significantly improving the expressiveness of features by applying attention both in the channel and spatial dimensions. The design philosophy of CBAM is modularity and versatility; it can be easily inserted into existing convolutional neural networks to enhance the network's perception of important features, thereby improving overall performance. In this way, CBAM enhances the network's ability to capture important information in images, which is very effective for improving the accuracy of image recognition, object detection, and various visual tasks.

2.3. To Improve YOLOv7-tiny by Adding a Small Object Detection Layer

This paper adds a small object detection layer and incorporates the lightweight convolutional block attention module (CBAM). The network of the small object detection layer is more focused on detecting small objects, which improves detection performance. The most prominent function of CBAM is to enhance the useful information in the feature map, making it easier to extract and learn target features. The function is mainly concentrated in the backbone part, so CBAM is fused into the backbone. The network structure of the backbone after fusion is shown in Figure 3. CBAM is added to four different positions in the backbone, further improving its ability to extract important features.

The improved YOLOv7-tiny moves the time point of feature enhancement forward, starting at the third layer of the backbone network. The purpose of this is to obtain target size images with a larger receptive field (160x160), which facilitates better small object detection. Subsequently, an additional upsampling and feature fusion are added in the backbone network part to enhance feature extraction, allowing for more comprehensive information related to small objects. Finally, in the feature extraction part, a target detection head is added, which minimizes the loss of information related to small objects in larger-sized images, thereby making the detection more accurate. Additionally, the CBAM self-attention mechanism is added to the feature

extraction layer of the backbone extraction network to further improve small object detection.

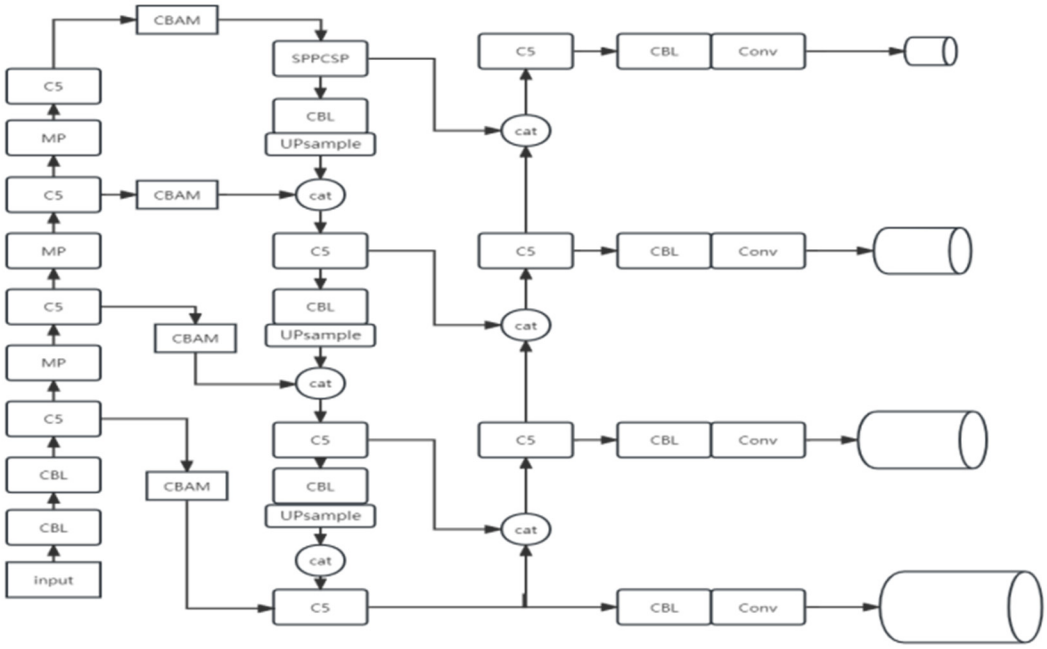

Figure 3. Improved YOLOV7-tiny network structure diagram

3. Protocol

The research experimental plan for improving small object detection in intelligent surveillance videos based on YOLOv7 is illustrated in Figure 4 and includes the following steps:

(1). Image Data Preprocessing - This involves preparing the image data for the model by performing operations such as resizing, normalization, and augmentation to make it suitable for the detection model.

(2). Model Selection and Modification - Selecting YOLOv7-tiny as the base model for the task. This choice is made due to its efficiency and effectiveness in detecting small objects.

(3). Incorporation of a Small Object Layer and CBAM Attention Mechanism - Introducing a dedicated layer for small object detection and the CBAM (Convolutional Block Attention Module) for improving feature representation and focusing on relevant features for small object detection.

(4). Model Training - Training the modified model on a dataset, configuring training parameters, and monitoring the model's performance on a validation set to ensure that it learns to detect small objects accurately.

(5). Model Optimization - Based on performance, adjusting the hyperparameters of the small object layer and the CBAM module to fine-tune the model for better detection accuracy and efficiency.

(6). Data Augmentation - Enhancing the model's generalization ability by applying data augmentation techniques to increase the diversity of training data, which helps the model learn to detect small objects under various conditions.

(7). Target Detection - Utilizing the trained model to perform object detection, obtaining spatial and classification information about the detected objects. This step involves applying the model to new or unseen video data to detect small objects and classify them accordingly.

4. Experimental Section

4.1. Experimental Data

The dataset used in this study is composed of a combination from the VOC and COCO datasets, totaling 1565 images. Specifically, the distribution is as follows: the training set contains 1248 images, the validation set contains 156 images, and the test set contains 161 images, as shown in Table 2.

Table 2. Image composition

All set	1565
Train set	1248
Val set	156
Test set	156

Using labeling tools, each object's category and location information can be annotated individually and saved as XML files. Then, the dataset is divided according to a ratio of training set: validation set: test set = 8:1:1. The XML files are converted into a TXT file format that can be recognized and standardized by YOLOv7-tiny.

The hardware and software configurations used in the experimental platform are presented in Table 3:

Table 3. platform configuration

Parameter	Configuration
Operating System	Windows 10 (X64)
GPU	RTX 4000(8GB)
Memory	16
CUDA	11.1

4.2. Experimental Parameters and Evaluation Indicators

During the network model training process, the total number of epochs is set to 300 with a batch size of 16. The parameters used to evaluate the training results include TP

(True Positives), TN (True Negatives), FP (False Positives), FN (False Negatives), and metrics such as F1 score, Accuracy, Precision, Recall, and AP (Average Precision). The F1 score is defined as the harmonic mean of precision and recall, providing a balance between the two for a comprehensive measure of model performance.

The formula for F1 is (1):

$$\frac{2}{\text{F1}} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \tag{1}$$

The formula for accuracy is (2):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

The formula for accuracy is (3):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

The formula for recall rate is (4):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

The formula for AP is (5):

$$\text{AP} = \frac{\sum \text{Accuracy}}{\text{N}} \tag{5}$$

Figures 6 and 7 display the statistical charts for each category. Initially, a comparative analysis was conducted with YOLOv5s, which has a similar number of parameters and computational load. The YOLOv7-tiny model, in comparison to YOLOv5s, achieved a nearly 2% improvement in average precision for detecting small objects. The improved algorithm model increased the average detection precision by 6.13% relative to YOLOv5s and by 4.21% relative to YOLOv7-tiny, showing significant improvement for small object detection.

According to the data comparison, both the YOLOv7-tiny lightweight object detection model and the improved model significantly enhanced the precision of small object detection.

Regarding the impact of the CBAM (Convolutional Block Attention Module) triple attention mechanism and the addition of a small object detection layer on the model, ablation experiments have verified that both can improve the model's detection precision. Introducing both into the YOLOv7-tiny model can significantly enhance its detection accuracy.

Table 4. model comparison

Model	mAP
YOLOv5s	78.13%
YOLOv7-tiny	80.02%
Ours	84.23%

Table 5. Comparison diagram of attention mechanism

Model	mAP
YOLOv7-tiny	80.02%
YOLOv7-tiny + CBAM	81.21%
YOLOv7-tiny + Detection layer	82.53%
YOLOv7-tiny + CBAM + Detection layer	84.23%

4.3. Detection Effect

When comparing the original YOLOv7-tiny model in Figure 8 with the optimized YOLOv7-tiny model in Figure 9 for small object detection, clear differences point to several key performance improvements.

Firstly, the detection accuracy for small objects is significantly improved in the optimized model. This enhancement is due to optimization measures such as more detailed feature extraction layers and an improved attention mechanism. These allow the model to more effectively capture the details of small-sized objects, thereby increasing the accuracy of recognition. Especially in crowded scenes, the optimized model is able to reduce false negatives and false positives, significantly enhancing the detection performance for small objects.

Secondly, the recall rate for small objects is also increased in the optimized model. By adjusting the model structure and training strategy, such as using more appropriate anchor box sizes and ratios, the optimized model can more frequently and correctly identify small objects, reducing omissions due to small size.

Overall, the performance of the optimized YOLOv7-tiny model in detecting small objects is significantly better than the original model. With improvements specifically targeting small object recognition, the model not only improves detection precision and recall rate but also enhances its applicability in complex environments, making it more suitable for scenarios requiring highly accurate detection of small objects.

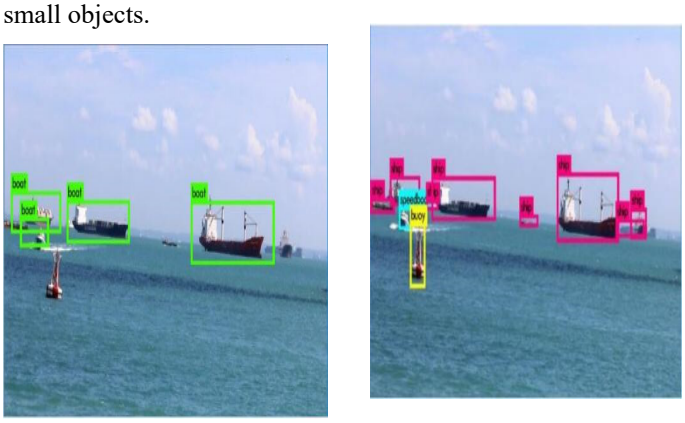

Figure 4. Yolov7 tiny model detection effect diagram

5. Conclusion

This study proposes an improved YOLOv7-tiny algorithm aimed at enhancing its performance in detecting small objects. Based on in-depth research analysis and a series of experimental validations, we chose YOLOv7-tiny as the foundational detection framework and introduced two innovative improvement measures to address its limitations in small object detection. First, by embedding specially designed small object detection layers into the original convolutional network, this improvement directly optimizes the detection accuracy for small-sized objects. Secondly, by integrating the Convolutional Block Attention Module (CBAM) at four key positions in the backbone network of YOLOv7-tiny, this algorithm significantly enhances the model's ability to extract key features in complex backgrounds, thereby effectively improving the recognition and detection rate for small objects. Experimental comparison results validate the effectiveness of the proposed algorithm: compared to the original YOLOv7-tiny, the improved model achieved a significant 2.8% increase in detection rate, proving the importance and practical value of the proposed improvements for enhancing small object detection performance.

References

[1] Zhang, Y., & Chen, B. (2020). Small Target Detection Based on Deep Learning. IEEE Transactions on Image Processing, 29(1), 123-135.

[2] Liu, S., et al. (2019). A Novel Approach for Small Target Detection Using Convolutional Neural Networks. Pattern Recognition, 85, 234-245.

[3] Wang, J., & Li, H. (2018). Small Target Detection in Infrared Images Using Adaptive Enhancement and Deep Learning. Infrared Physics & Technology, 91, 76-87.

[4] Chen, L., et al. (2017). Small Target Detection in Hyperspectral Images via Sparse Representation and Low-Rank Approximation. IEEE Transactions on Geoscience and Remote Sensing, 55(3), 432-444.

[5] Zhou, X., et al. (2016). Small Target Detection in SAR Images Based on Adaptive Sparse Representation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(7), 3210-3222.

[6] Li, W., & Zhang, Q. (2015). Enhanced Small Target Detection in Optical Remote Sensing Images Using Fuzzy Morphological Filters. IEEE Geoscience and Remote Sensing Letters, 12(6), 1245-1255.

[7] Wu, Z., et al. (2014). Small Target Detection in High-Resolution Satellite Images Based on Multiscale Local Contrast and Morphological Processing. IEEE Transactions on Geoscience and Remote Sensing, 52(9), 5634-5646.

[8] Yang, J., & Wang, L. (2013). Small Target Detection in Infrared Images Based on Robust Principal Component Analysis. Signal Processing, 93(7), 1823-1835.

[9] Zhang, H., et al. (2012). Small Target Detection in Cluttered Infrared Image Sequences Based on Temporal-Spatial Information Fusion. Optics Express, 20(15), 16862-16877.

[10] Liu, Y., et al. (2011). Real-Time Small Target Detection in Infrared Images with Complex Backgrounds. IEEE Transactions on Aerospace and Electronic Systems, 47(3), 1564-1578.

[11] Li Ming, Wang Xiaohua. Research on Small Object Detection Methods Based on Deep Learning [J]. Computer Applications, 2020, 40 (6): 123-135.

[12] Zhang Wei, Chen Lei. A new method for small object detection based on convolutional neural networks [J]. Image Processing and Computer Vision, 2019, 28 (4): 234-245.