SE-CBAM-YOLOv7: An Improved Lightweight Attention Mechanism-Based YOLOv7 for Real-Time Detection of Small Aircraft Targets in Microsatellite Remote Sensing Imaging

Zhenping Kang , Yurong Liao , Shuhan Du , Haonan Li , Zhaoming Li *

Posted Date: 18 June 2024

doi: 10.20944/preprints202406.1287.v1

Keywords: aircraft detection; YOLOv7; CBAM; SENet; microsatellite

Article

SE-CBAM-YOLOv7: An Improved Lightweight Attention Mechanism-Based YOLOv7 for Real-Time Detection of Small Aircraft Targets in Microsatellite Remote Sensing Imaging

Zhenping Kang, Yurong Liao, Shuhan Du, Haonan Li and Zhaoming Li *

Department of Electronic and Optical Engineering, Space Engineering University, Beijing 101416, China; zhenpingkang@hgd.edu.cn (Z.K.); liaoyr@hgd.edu.cn (Y.L.); lhn_links@hgd.edu.cn (S.D.)
* Correspondence: lizm@hgd.edu.cn

Abstract: Addressing real-time aircraft target detection in microsatellite-based visible light remote sensing video imaging requires considering the limitations of imaging payload resolution, complex ground backgrounds, and the relative positional changes between the platform and aircraft. These factors lead to multi-scale variations in aircraft targets, making high-precision real-time detection of small targets in complex backgrounds a significant challenge for the detection algorithms. Hence, this paper introduces a real-time aircraft target detection algorithm for remote sensing imaging using an improved lightweight attention mechanism that relies on the You only look once version 7 (YOLOv7) framework (SE-CBAM-YOLOv7). The proposed algorithm replaces the standard convolution (Conv) with a lightweight convolutional Squeeze-and-Excitation convolution (SEConv) to reduce the computational parameters and accelerate the detection process of small aircraft targets, thus enhancing real-time onboard processing capabilities. Besides, the SEConv-based Spatial Pyramid Pooling and Connected Spatial Pyramid Convolution (SPPCSPC) module extracts image features. It improves detection accuracy while the feature fusion section integrates the Convolutional Block Attention Module (CBAM) hybrid attention network, forming the Convolutional Block Attention Module Concat (CBAMCAT) module. Furthermore, it optimizes small aircraft target features in channel and spatial dimensions, improving the model's feature fusion capabilities. Experiments on public remote sensing datasets reveal the proposed SE-CBAM-YOLOv7 improves detection accuracy by 3% and mAP value by 4.2% compared to YOLOv7, significantly enhancing the detection capability for small-sized aircraft targets in satellite remote sensing imaging.

Keywords: aircraft detection; YOLOv7; CBAM; SENet; microsatellite

1. Introduction

Detecting small targets within complex backgrounds is a research hotspot, typically applied in real-time detection of small aircraft targets in satellite-based visible light remote sensing video imaging. Remote sensing images contain complex background texture information [1,2], where aircraft targets appear as multi-scale variation targets and are influenced by the satellite platform's imaging resolution and relative positional changes. This is incredibly challenging for high-precision detection, especially when small targets are superimposed with complex background interference. Therefore, it is necessary to investigate the real-time detection of small aircraft targets in satellite remote-sensing imaging.

Object detection is a crucial task in computer vision, aiming at automatically identifying and locating specific objects in images or videos using computer algorithms and techniques [3,4]. In recent years, the continuous development of aerospace, computer, sensor, and data processing technology has pushed the boundaries of object detection, demonstrating significant capabilities in military, civilian, and intelligent applications [5]. Traditional object detection algorithms include region selection, feature extraction, and classification. Deep learning-based object detection algorithms

 doi:10.20944/preprints202406.1287.v1

outperform traditional object detection algorithms in complex scenes with convolutional neural network (CNN)-based object detection algorithms divided into two-stage and single-stage detection algorithms. The two-stage detection algorithms first form proposal boxes to predict proposed regions. They afford a high detection accuracy but have a complex structure and long training times. Classic two-stage detection algorithms include R-CNN [6], Fast R-CNN [7], Faster R-CNN [8], and Mask R-CNN [9]. The single-stage detection algorithms directly generate class probabilities and position coordinates on objects. They have a simple structure and shorter training times but slightly lower accuracy. Classic single-stage detection algorithms include SSD [10] and YOLO [11].

Despite the high level of current technology, there is still room for improvement in detecting small objects. Particularly, detecting small aircraft targets in aerial images is quite challenging as their small size often results in them being filtered out by the pooling layers in the CNNs. Hence, this addresses this challenge by proposing a new deep-learning algorithm, SE-CBAM-YOLOv7, which improves traditional convolution techniques and incorporates state-of-the-art attention mechanisms. The main contributions of this paper are as follows:

1. Replacing the standard convolution (Conv) process with a new lightweight convolution (SEConv) to reduce the network's computational parameters and speed up the detection process for small aircraft targets.
2. Designing the SESPPCSPC module that integrates the channel attention mechanism network SENet. This achieves multi-scale spatial pyramid pooling on the input feature maps, enhances the model's receptive field and feature expression capabilities, and improves the network's feature extraction capability.
3. Introducing CBAMCAT, a new feature fusion layer that sequentially infers attention maps along two independent dimensions (channel and spatial). The attention maps are multiplied with the input feature maps for adaptive optimization, improving the model's feature fusion capability.

This paper is organized as follows. Section 2 discusses the existing work on the YOLO algorithm for target detection in remote sensing images. Section 3 overviews the proposed SE-CBAM-YOLOv7 network. Section 4 presents the experiments conducted, tests the algorithm's performance on a small aircraft target dataset, and analyzes the results. Finally, Section 5 concludes this paper.

2. Related Work

In 2016, inspired by the GoogLeNet architecture [12], Redmon et al. introduced the YOLO (You Only Look Once) structure in the CVPR paper You Only Look Once: Unified, Real-Time Object Detection [11]. They replaced the initial module of GoogLeNet with 1×1 convolutions followed by 3×3 convolution filters. The main features of YOLO are integrating object localization and classification predictions into a single neural network model, thereby achieving fast object detection and recognition with high accuracy. Since introducing YOLOv1 in 2016, the YOLO algorithm has undergone continuous updates and optimizations. Each subsequent version has demonstrated advancements in innovative architectures, leading to increased speed and accuracy in object detection.

In recent years, YOLO has been widely applied to target detection in remote sensing images. However, real-time detection of small objects in remote sensing images captured by drones is challenging, as the various drone shooting angles present the target objects under varying scales, densities, and shapes. Hence, Zhang et al. focused on real-time small vehicle detection in drone-captured remote sensing images and proposed a depth-wise attention mechanism network (DAGN) based on YOLOv3. This method combines feature concatenation and attention modules, allowing the model to distinguish between important and unimportant features, thereby improving vehicle detection and promoting real-time detection of small objects in drone imagery [13]. Nevertheless, current research on aircraft target detection and classification in remote sensing images suffers from data sample imbalance, significant variations in target scales and backgrounds, and target occlusion, leading to low average precision and slow detection speeds. Spurred by these concerns, Liu et al. proposed the YOLO-Extract model to detect small, dense, and occluded targets. Specifically, they optimized the Mish activation function and the Conv module using representative batch

 doi:10.20944/preprints202406.1287.v1

normalization. Furthermore, they improved the classification loss function using a VariFocal loss to overcome the precision issues caused by data sample imbalance. Finally, they designed the RepVGG module in the Backbone, further enhancing the model's detection accuracy [14]. Sun et al. introduced a ship target detection algorithm based on YOLOv5, achieving ship target detection in remote sensing images with complex backgrounds. Their method relied on an improved K-means clustering method, with the experimental results demonstrating that the enhanced network significantly improved performance on images with densely distributed small targets over the original YOLOv5 network at the expense of a slightly reduced detection speed [15].

3. Method

The YOLO algorithm is currently one of the most used object detection and recognition algorithms, adopting a single-stage deep learning approach [16]. Compared to traditional two-stage object recognition algorithms, YOLO does not require selecting candidate regions for classification but relies on a unified architecture that simultaneously predicts bounding boxes and class probabilities, enabling end-to-end object detection at a very high speed. Its main features are real-time performance and accuracy. YOLO transforms the object detection task into a regression problem, where a single neural network simultaneously predicts object classes and bounding boxes. The core idea of YOLO is to divide the input image into a fixed-size grid and predict multiple bounding boxes within each grid cell using convolutional layers. Each bounding box contains an object and predicts the object's class and position. Using CNN for feature extraction and prediction allows YOLO to perform object detection and classification in a single forward pass.

This study proposes SE-CBAM-YOLOv7, a remote-sensing aircraft target detection algorithm based on YOLOv7. The network model comprises the head, neck, and backbone components, as illustrated in Figure 1. Specifically, after resizing and normalizing the remote-sensing aircraft images, these are input into the proposed SE-CBAM-YOLOv7 network model. The backbone network first extracts feature information, which is input into the neck network for feature fusion, producing three feature maps of different sizes (large, medium, and small). Finally, the fused feature maps are processed by the head network, which has three detection heads that output the predicted bounding boxes and class information. The backbone network of SE-CBAM-YOLOv7 comprises convolution modules, Cross-Stage Partial Network Block(CBS), an Efficient Layer Aggregation Network (ELAN) module, an Mixed-Precision Convolution (MPConv) module, and an SPPCSPC module. This study introduces the SENet attention mechanism at the feature layer located at the backbone's output to enhance feature extraction capability and increase the critical information of small aircraft targets on the feature map. The neck network of SE-CBAM-YOLOv7 adopts a Path Aggregation Feature Pyramid Network (PAFPN) structure, performing feature fusion for the upsampling and downsampling parts. Furthermore, this study incorporates a lightweight CBAM network in the fusion module to comprehensively capture critical information on small aircraft targets in channel and spatial dimensions. The head network of SE-CBAM-YOLOv7 incorporates three sizes of Identity Detection (IDetect) detection heads, which detect and recognize small aircraft targets based on the critical feature information output by the neck network. The following sections detail the principles of network optimization.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2024 doi:10.20944/preprints202406.1287.v1

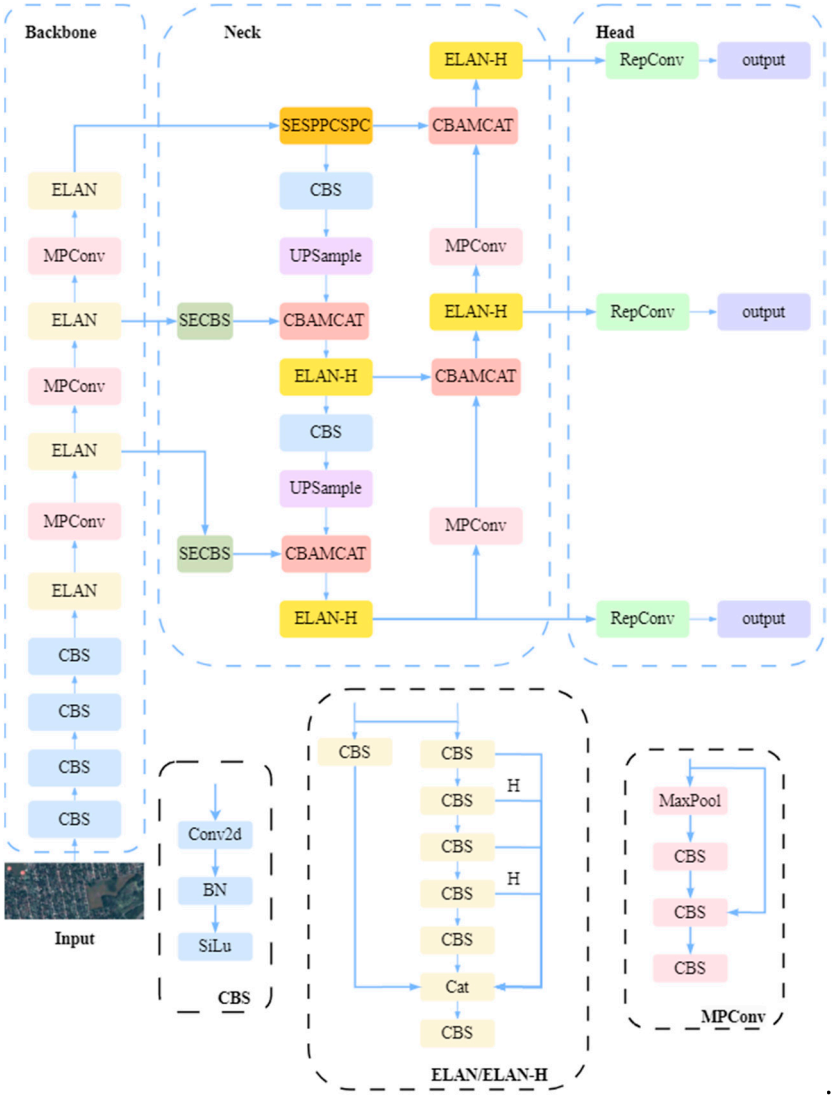


Figure 1. SE-CBAM-YOLOv7 model structure.

3.1 SEConv

In neural network models, as the number of network layers increases, the parameters the model must learn increase, gradually accumulating information that must be stored during model training. This result leads to information overload, with typical solutions utilizing attention networks. These networks help the model focus on the most crucial information relevant to the training task, reducing the attention given to less important information and filtering out irrelevant data. This strategy effectively solves the information overload problem and improves the model's classification accuracy in the later stages. Regarding SENet (Squeeze-and-Excitation Network), it first performs global average pooling on the input features across the spatial dimensions, generating C weights between 0 and 1 in a fully connected network structure. Then, it captures the inter-channel dependencies through fully connected layers. Different channel features are essential and are obtained using the Sigmoid function, which serves as weight coefficients. These weights are multiplied with the input feature signals, automatically assigning weights to each channel [17]. Figure 2 depicts SENet.

SENet is divided into four steps: firstly, Transformation, after transforming F_{tr}, maps the input X to the feature map U. The calculation formula is shown in Equation (1):

$$U_c = V_c * X \quad (1)$$

where $U \in R^{H' \times W' \times C'}$ denotes the input feature map, $X \in R^{H \times W \times C}$ denotes the transformed output feature map, V_c denotes the parameter of the cth filter, and * denotes the convolution operation.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2024 doi:10.20944/preprints202406.1287.v1

The Squeeze operation then compresses the $H \times W \times C$ feature map containing global information into a $1 \times 1 \times C$ feature vector defined as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \quad (2)$$

where z_c is the cth element of z.

This is followed by the Excitation operation, where the features are adaptively corrected to obtain the weight coefficients s through two fully connected layers.

$$s = F_{ex}(z, W) \quad (3)$$

Finally, Scale, which multiplies each feature map in the feature map U by the corresponding weight to obtain the final output of the SE model $\tilde{X}$.

$$\tilde{X}_c = F_{Scale}(u_c, s_c) \quad (4)$$

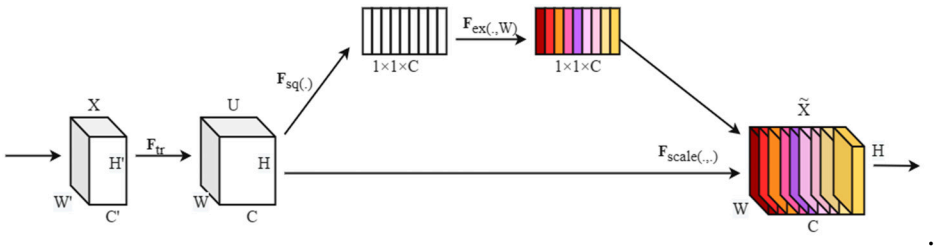


Figure 2. SENet model

Figure 3 illustrates the proposed lightweight SEConv. The convolutional layer is connected to Batch Normalisation, which solves the problem of vanishing gradient to some extent. After that, SiLU activation function is connected, SiLU is the combination of Relu and sigmoid. It can be regarded as a smooth Relu, which solves the disadvantage that Relu has negative input and output of 0, and the problem of gradient dispersion does not occur. Finally, the SENet network is connected. The expression for SiLU is as follows:

$$SiLU(x) = x \cdot \sigma(x) \quad (5)$$

where $\sigma(x)$ denotes the sigmoid function,$\sigma(x) = \frac{1}{1+e^{-x}}$.

The use of the SEConv convolution instead of the standard convolution Conv reduces the computational parameters of the network. Additionally, SEConv optimizes the output of small aircraft target features by the ELAN module, thus accelerating the model's detection speed.

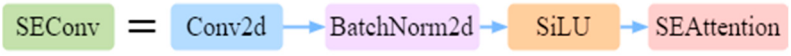


Figure 3. SEConv structure

3.2 SESPPCSPC

YOLOv7 incorporates the SPPCSPC (Spatial Pyramid Pooling and Connected Spatial Pyramid Convolution) module. SPPCSPC was initially presented in YOLOv5 [18], extracts image features, and enhances target detection performance. This module combines the Spatial Pyramid Pooling (SPP) and Channel Pyramid Convolution (CSPC) techniques. The former performs pooling operations on feature maps at different scales to capture contextual information of targets of various sizes, thus improving target detection accuracy. On the other hand, SPP does not change the size of the feature map but uses pooling kernels of different sizes to pool the feature map. Then, it concatenates these pooled results into a fixed-length feature vector. This process retains multi-scale information, enabling the network to better adapt to objects of different sizes. CSPC performs pyramid convolution operations along the channel dimension by introducing more nonlinear transformations to extract richer feature representations. Additionally, CSPC splits the input feature map into two parts, applies pyramid convolution to one part, and then concatenates it with the other part. This process increases the network's width and enhances its expressive capability. The SPPCSPC module combines SPP and CSPC to improve target detection performance. First, the input feature map is sent

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2024 doi:10.20944/preprints202406.1287.v1

to the SPP layer for spatial pyramid pooling, resulting in a fixed-length feature vector. This vector then undergoes several convolution operations, including channel pyramid convolution, to extract more discriminative feature representations. Finally, the resulting feature map is used for the target detection task.

This paper introduces the SENet attention mechanism network into this module, as depicted in Figure 4. Besides, adding parallel Maximum Pooling(MaxPool) layers to the continuous convolution layers partially addresses image distortion problems due to image preprocessing. MaxPool is the process of selecting the maximum value in each region of an image or signal feature map, thereby reducing the dimensionality of the data and retaining the most important features while preventing overfitting. Instead of calculating detail at the pixel level, this process uses a fixed-size window to scan the input and take the maximum value, and is commonly used in feature extraction and image processing. Hence, SESPPCSPC performs multi-scale feature fusion on the small aircraft target images and fine-tunes these features at each scale, effectively capturing information of different scales. This strategy significantly enhances the model's ability to detect small objects.

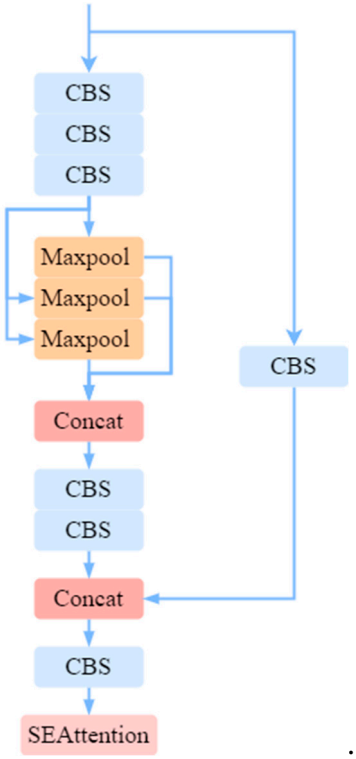


Figure 4. SESPPCSPC structure

3.3 CBAMCAT

The CBAM network [19] is a lightweight attention mechanism network comprising the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). These modules extract attention maps from the input feature map in the channel and spatial dimensions, respectively, which are then multiplied with the input feature map for adaptive feature refinement. The specific structure is presented in Figure 5. The feature map $X \in R^{H\times W\times C}$ is used as input, and CBAM sequentially infers a one-dimensional channel attention map $M_c \in R^{1\times 1\times C}$ and a two-dimensional spatial attention map $M_s \in R^{H\times W\times 1}$, as shown in Figure 1. The whole attention process can be summarised as:

$$X' = M_c(X) \otimes X,$$
$$X'' = M_s(X') \otimes X' \quad (6)$$

where ⊗ denotes element-wise multiplication, X'' stands for final output.

 doi:10.20944/preprints202406.1287.v1

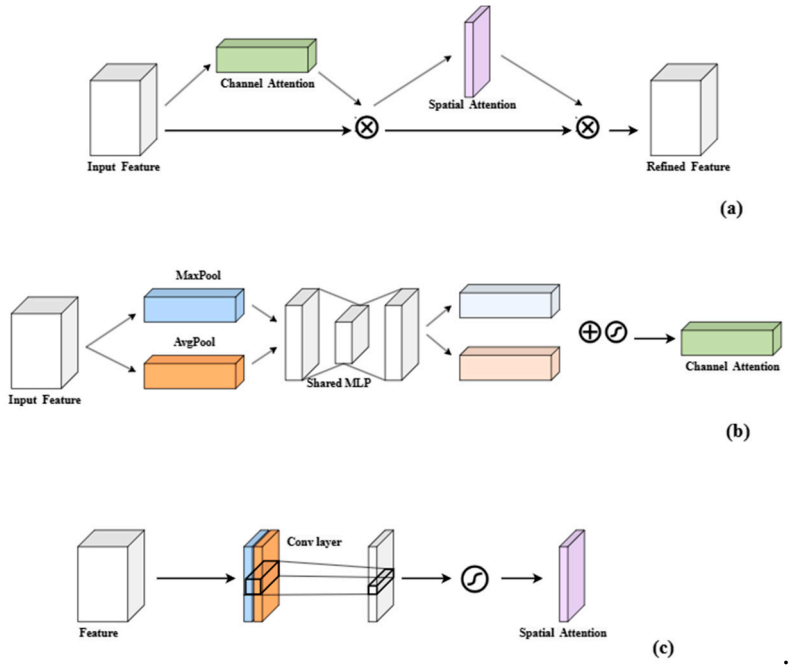


Figure 5. (a)CBAM model (b)CAM (c)SAM

The innovation of CAM is the use of mean-pool and max-pool operations to aggregate spatial information of feature maps. The CAM sub-module compresses the input feature map along the spatial dimension, then performs max and average pooling operations. The outputs are input into shared network to form the channel attention map, which is then multiplied with the feature map to highlight the essential target features. The shared network consists of a multilayer perceptron (MLP). This process is mathematically formulated as follows:

The calculation formula is shown in Equation (7).

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (7)$$

where F represents the input feature map, AvgPool denotes global average pooling, MaxPool denotes max pooling, MLP stands for multi-layer perceptron, and σ is the Sigmoid activation function.

SAM is used to aggregate the channel information of a feature map by using two pooling operations to generate two 2D mappings. The SAM sub-module uses the channel attention map as the input feature map, compressing it along the channel dimension, followed by max pooling and average pooling operations. The results are concatenated along the channel axis and passed through a 7×7 convolution to form the spatial attention map, which is then multiplied with the feature map to emphasize important positional information about the target. The calculation formula is shown in Equation (8).

$$M_s(F) = \sigma(f^{7\times7}[AvgPool(F); MaxPool(F)]) \quad (8)$$

where $f^{7\times7}$ denotes a 7×7 convolution.

This study integrates the CBAM attention mechanism into the fusion layer of the neck network in the proposed SE-CBAM-YOLOv7, replacing the ordinary concatenation (CAT) layer with a CBAMCAT layer. This integration suppresses complex background noise and optimizes and fuses critical features of small aircraft targets in both channel and spatial dimensions.

4. Experiments

4.1. Experimental Data

The data used in this study is sourced from publicly available satellite remote sensing datasets, comprising 697 images with a resolution of 1283×521 pixels. During the experiment, the data was split into a training set and a testing set in a ratio of 7:3.

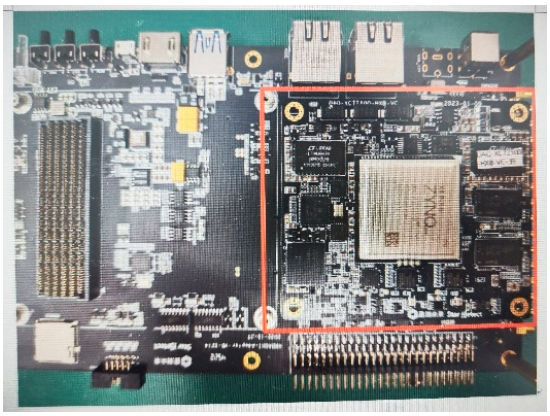
 doi:10.20944/preprints202406.1287.v1

4.2. Space-Based Intelligent Processing Platform

Combined with the industry development trend and actual architecture evolution, the star-carried intelligent computing simulation platform in this paper adopts a super heterogeneous body system of ECU unit (with FPGA module inside) and SCC unit (with GPU module inside). The platform is connected to a satellite data simulator and executes algorithms based on mission commands and data transmitted by the simulator and transmits the results of the processing back. The platform adopts the ZYNQ chip, as shown in Figure 6(a), and the prototype of the starboard intelligent computing platform is shown in Figure 6(b). Table 1 lists the configuration parameters of satellite-borne intelligent computing platform.

The ECU unit is mainly composed of an FPGA SoC with an internal FPGA unit and an ARM core. The FPGA unit is mainly responsible for performing interface adaptation and receiving external instructions and data, while the ARM side is mainly responsible for carrying out the functional management of the load, and is able to carry out flexible scheduling of functions based on the task instructions.

The SCC unit mainly consists of a GPU SoC, which contains an ARM core and a GPU processing unit inside. the ARM side mainly carries out the task management, and can schedule different algorithmic models according to different task requirements to meet different applications. the GPU unit, as the main processing unit, provides a large amount of general-purpose computation arithmetic and AI computation arithmetic for performing efficient image general-purpose processing and AI processing. With the highly parallel processing architecture of the GPU unit and mature GPU acceleration tool libraries, fast algorithm processing can be achieved with low energy consumption, and we develop and deploy the SE-CBAM-YOLOv7 algorithm in this unit.



(a)

(b)

Figure 6. On-board Intelligent Computing Simulation Platform (a) ZYNQ Chip (b)satellite-borne intelligent computing platform prototype

 doi:10.20944/preprints202406.1287.v1

Table 1. Configuration parameters of satellite-borne intelligent computing platform.

Satellite-borne Intelligent Computing Platform		
Basic parameter	volumetric	208*125*55mm ±5mm
	weights	1.5kg±0.2
	electricity supply	28±3 V
ECU Modules Central Control Unit	microchip	ZYNQ 7100
	main frequency	766MHz （dual core）
	random access memory (RAM)	512MB×2，DDR3，1066MHz
	stockpile	32GB eMMC×2
SCC Module Central Computing Unit	microchip	Jetson AGXi Xavier
	main frequency	CPU: 2.0GHz（8 core） GPU: 1.2GHz
	random access memory (RAM)	32GB，LPDDR4x，136.5GB/s
	stockpile	1TB SSD
	arithmetic power	30 TOPS

4.3. Ground Link Experiment Syetem

The flowchart of the ground link experiment in this study is shown in Figure 7, and the experimental platform is shown in Figure 8. The experimental platform, which includes four main components: a satellite ground simulator, a satellite data simulator, a satellite-borne intelligent computing platform prototype, and an upper computer. The satellite data simulator simulates the camera payload of a real satellite platform, generating data that mimics the output of onboard cameras and transmitting this data to the computing payload. The satellite-borne intelligent computing platform prototype comprises an ECU (Electronic Control Unit) and an SCC (Satellite Control Center). The intelligent computing platform connects to the satellite data simulator, executes algorithms based on the simulator's task instructions and data transmitted, and then returns the processed results. The upper computer connects to the satellite data simulator, simulating a ground station command control and scheduling functions. Table 2 reports the hardware setup used in the satellite data simulator. The input image configuration is set to 640×640 pixels, and the model's training process involves 200 epochs, with a batch size of 16 and an initial learning rate of 0.01.

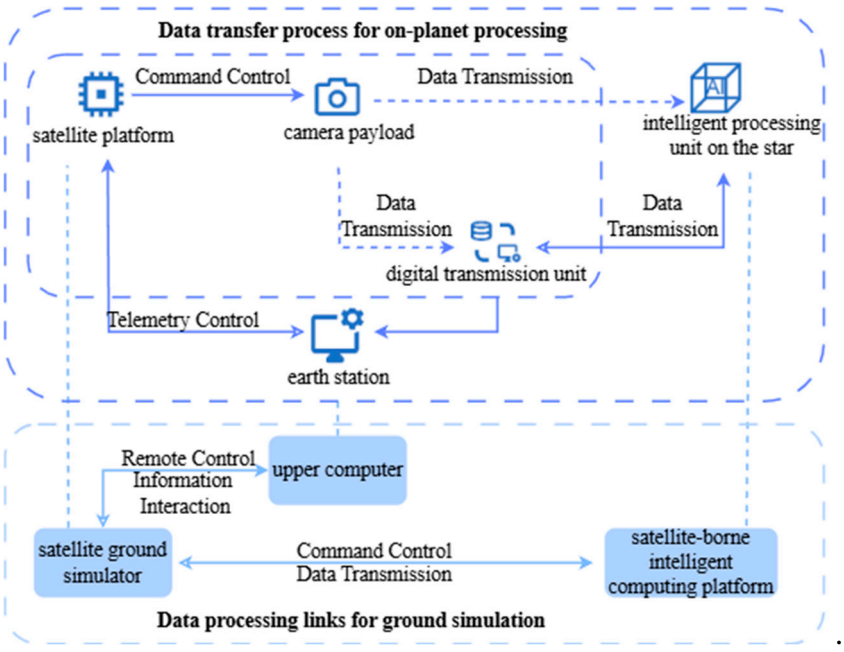
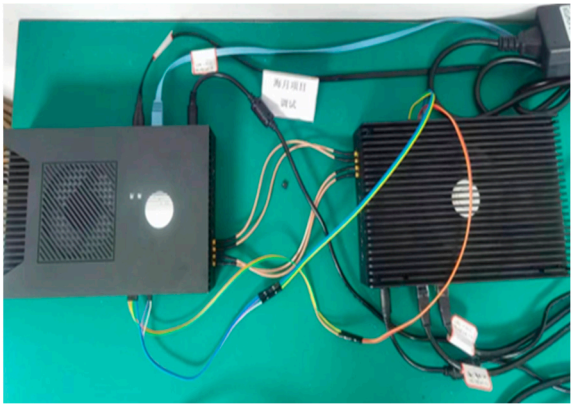


Figure 7. Ground link experiment flow chart.

 doi:10.20944/preprints202406.1287.v1

(a)



(b)

Figure 8. Experimental platform (a)satellite ground simulator (b)satellite-borne intelligent computing platform prototype.

Table 2. Configuration parameters of satellite data simulator.

Satellite Data Simulator		
Basic parameter	volumetric	208*125*55mm ±5mm
	weights	1.5kg±0.2
	electricity supply	28±3 V
OBC On-Board Computing Unit	microchip	ZYNQ 7100
Storage Module	stockpile	1TB SSD

4.4. Evaluation Metrics

Several evaluation metrics are used to analyze and assess the model's prediction results, thus determining the performance of various models from multiple perspectives. The performance of SE-CBAM-YOLOv7 is evaluated based on Precision [20], Recall [21], F1-Score [22], and mean Average Precision (mAP) [23,24]. All models are evaluated on the same training and testing datasets.

Precision is the ratio of correctly predicted positive samples to the total predicted positive samples, formulated in Eq. (9). Here, TP (True Positive) denotes the case where the actual and predicted values are positive. FP (False Positive) denotes the case where the actual value is negative, but the expected value is positive.

 doi:10.20944/preprints202406.1287.v1

Recall is the ratio of the correctly predicted positive samples to the actual positive samples and is calculated as in Eq. (10). Here, FN (False Negative) denotes the case where the exact value is positive, but the predicted value is negative.

The F1 Score is a comprehensive evaluation metric that considers precision and recall and is mathematically formulated as in Eq. (11)

The P-R Curve (Precision-Recall Curve) visualizes the relationship between Precision and Recall, with Precision on the y-axis and Recall on the x-axis. The area under the P-R Curve is called the AP value (see Eq. (12)). A higher AP value indicates better model performance.

The mAP (mean Average Precision) represents the mean of AP values across different categories and is calculated using Eq. (13).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

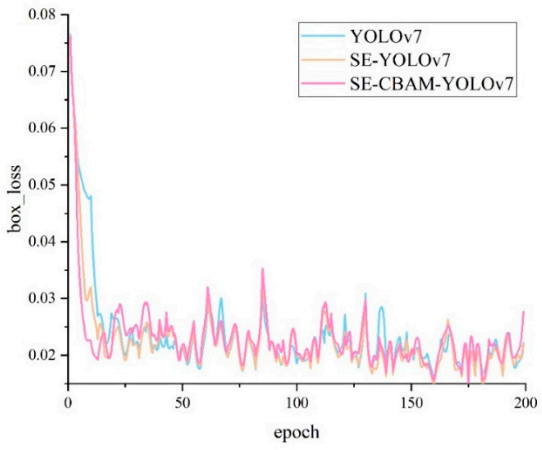
$$\text{F1} = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{11}$$

$$\text{AP} = \int_0^1 P(R)dR \tag{12}$$

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{13}$$

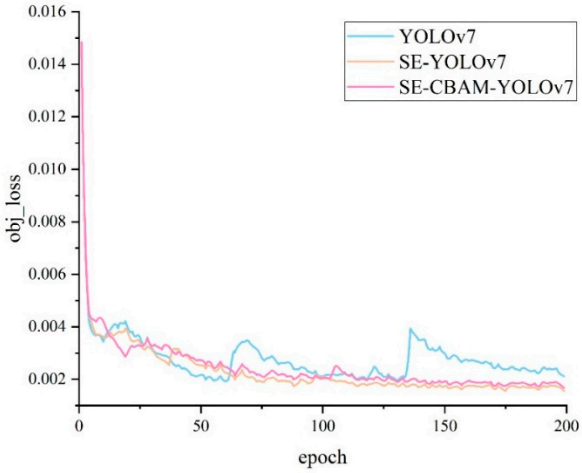
4.5 Experimental Results and Analysis

The network must have high accuracy and fast convergence, which are crucial for the model's robustness and stability. The YOLOv7, SE-YOLOv7, and SE-CBAM-YOLOv7 models were trained and tested for 200 epochs. Figure 9 illustrates the convergence curves of the three models regarding the bounding box loss and object detection loss. The vertical axis represents the loss value during network training, and the horizontal axis represents the iteration rounds of the network. The experimental results indicate that the loss values of the three models are relatively high in the early stages of training. Within the first 25 epochs, the loss values of all three models exhibit a rapid decrease trend. Still, as the number of iterations increases, their loss values decrease, indicating that the network fits the training data. The SE-CBAM-YOLOv7 model maintains the loss value at a lower level while ensuring convergence speed, exhibiting better convergence performance, robustness, and stability.



(a)

 doi:10.20944/preprints202406.1287.v1



(b)

Figure9. Convergence curves (a) box loss (b) object detection loss

Figure 10 presents the test set output results of YOLOv7 and SE-CBAM-YOLOv7, and Table 2 reports the corresponding data results. Figure 10 highlights that YOLOv7 encounters difficulty in recognizing small airborne targets in complex backgrounds, resulting in many miss-detections due to insufficiently extracting features from small airborne targets in the early stages of the network. On the contrary, SE-CBAM-YOLOv7 can accurately detect the targets and exhibits higher sensitivity to smaller-sized aircraft objects. This indicates that the SENet and CBAM attention networks enhance the model's ability to recognize low-density targets. From Table 3, it is evident that the SE-CBAM-YOLOv7 model performs better in recognizing small airborne targets, as the recognition accuracy of the SE-CBAM-YOLOv7 model is 3% higher than that of YOLOv7 and 1.8% higher than that of SE-YOLOv7. Moreover, the mAP value of SE-CBAM-YOLOv7 is 4.2% higher than that of YOLOv7. This is because after the SENet and CBAM attention mechanisms autonomously learn the weighting coefficients, they reinforce important features and simultaneously suppress unimportant feature information using dynamic weighting. This makes the deep learning networks focus better on the critical features and improves their sensitivity to small target recognition, thereby enhancing the model's ability to recognize small airborne targets in complex backgrounds.

(a) (b) (c)

Figure 10. Test set output results (a) YOLOv7 (b)SE-YOLOv7 (c)SE-CBAM-YOLOv7.

 doi:10.20944/preprints202406.1287.v1

Table 3. Experimental results.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	F1 (%)
YOLOv7	79.4	72.3	57.7	68
SE- YOLOv7	80.6	80	70.1	77
SE-CBAM-YOLOv7	82.4	72.4	61.9	68

5. Conclusions

This paper proposes the SE-CBAM-YOLOv7 optimization algorithm to facilitate the real-time detection of small airborne targets in complex background remote sensing video. Specifically, we introduce SENet, a lightweight attention mechanism, and design the SESPPCSPC module to improve the model's efficiency in feature extraction. Additionally, a hybrid attention mechanism CBAM is introduced, and the CBAMCAT module is designed to effectively suppress complex background noise and enhance the model's ability to integrate important feature information of small airborne targets. The SE-CBAM-YOLOv7 model is tested on remote sensing datasets, achieving a detection accuracy of 82.4%, 3% higher than YOLOv7. This lays the algorithmic foundation for subsequent deployment applications in satellite missions, and the potential application of the proposed algorithm includes the support to ground radar for Air Traffic control in congested airports and runway incursion prediction etc.

Author Contributions: Formal analysis, Y.L.; investigation, Z.K. and Z.L.; software, S.D. and H.L.; validation, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Du C, Cui J, Wang D, et al. Prediction of aquatic vegetation growth under ecological recharge based on machine learning and remote sensing. Journal of Cleaner Production, 2024, 452.
2. Yang F ,Men X ,Liu Y , et al. Estimation of Landslide and Mudslide Susceptibility with Multi-Modal Remote Sensing Data and Semantics: The Case of Yunnan Mountain Area. Land, 2023, 12 (10):
3. Braun A ,Warth G ,Bachofer F , et al. Mapping Urban Structure Types Based on Remote Sensing Data—A Universal and Adaptable Framework for Spatial Analyses of Cities. Land, 2023, 12 (10).
4. A. J R ,M. H C ,Miguel R V . Analysis of Spacecraft Materials Discrimination Using Color Indices for Remote Sensing for Space Situational Awareness. The Journal of the Astronautical Sciences, 2023, 70 (5):
5. Bai, Liang et al. Remote Sensing Target Detection Algorithm based on CBAM-YOLOv5. Frontiers in Computing and Intelligent Systems, 2023.
6. Girshick R , Donahue J , Darrell T ,et al.rich feature hierarchies for accurate object detection and semantic segmentation tech report (v5). 2017.
7. Girshick R .Fast R-CNN.Computer Science, 2015.
8. Ren S , He K , Girshick R ,et al.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
9. Johnson J W .Adapting Mask-RCNN for Automatic Nucleus Segmentation. 2018.
10. Liu W, Anguelov D,Erhan D,et al.SSD: Single Shot MultiBox Detector//Proceedings of 14th European Conference on Computer Vision.Amsterdam: Springer, 2016: 21-37.
11. Redmon J , Divvala S , Girshick R ,et al.You Only Look Once: Unified, Real-Time Object Detection//Computer Vision & Pattern Recognition.IEEE, 2016
12. Szegedy C , Liu W , Jia Y ,et al.Going Deeper with Convolutions.IEEE Computer Society, 2014.
13. Zhang Z Y, Liu Y P, Liu T C, et al. DAGN: a real-time UAV remote sensing image vehicle detection framework. IEEE Geoscience and Remote Sensing Letters, 2020, 17(11): 1884- 1888.

 doi:10.20944/preprints202406.1287.v1

14. Liu Z , Gao Y , Du Q .YOLO-Class: Detection and Classification of Aircraft Targets in Satellite Remote Sensing Images Based on YOLO-Extract.IEEE Access, 11，2024.
15. M X S ,J Y Z ,H W , et al. Research on ship detection of optical remote sensing image based on Yolo V5. Journal of Physics: Conference Series, 2022, 2215 (1).
16. Wang C Y , Bochkovskiy A , Liao H Y M .YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.arXiv e-prints, 2022.
17. Jie H, Li S, Gang S, et al. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).
18. Huang Y , Fan J Y , Hu Y ,et al.TBi-YOLOv5: A surface defect detection model for crane wire with Bottleneck Transformer and small target detection layer.Proceedings of the Institution of Mechanical Engineers, Part C. Journal of mechanical engineering science, 2024(6):238.
19. WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
20. Wang C , Bi F , Zhang W ,et al. An Intensity-Space Domain CFAR Method for Ship Detection in HR SAR Images.IEEE Geoscience & Remote Sensing Letters, 2017, 14(4):529-533.
21. Ai J , Luo Q , Yang X ,et al. Outliers-Robust CFAR Detector of Gaussian Clutter Based on the Truncated-Maximum-Likelihood- Estimator in SAR Imagery.IEEE Transactions on Intelligent Transportation Systems, 2020, 21(5):2039-2049.
22. Karvonen J ,Gegiuc A , Niskanen T ,et al. Iceberg Detection in Dual-Polarized C-Band SAR Imagery by Segmentation and Nonparametric CFAR (SnP-CFAR).IEEE Transactions on Geoscience and Remote Sensing, 2021, PP(99):1-12.
23. Hou X , Ao W , Song Q ,et al. FUSAR-Ship: building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition. Science China Information Sciences, 2020, 63(4):1-19.
24. Ao W , Xu F , Li Y ,et al. Detection and Discrimination of Ship Targets in Complex Background From Spaceborne ALOS-2 SAR Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(2):536-550.