SVSDet: A Fine-Grained Recognition Method for Ship Target Using Satellite Video

Shanwei Liu, Xi Bu, Mingming Xu, *Member, IEEE*, Hui Sheng, Zhe Zeng, and Muhammad Yasir

***Abstract*—Target recognition from remote sensing images is commonly challenging because of large-scale variations and small objects, and these challenges are more prominent in satellite video images. The current object detection algorithms have some difficulties in fine-grained feature extraction and classification for multiscale and small objects. We propose a novel model called the SVSDet method based on YOLOv5 improvement to address the above-mentioned issues. In this method, we have introduced the space-to-depth module into the backbone of the network, which enhances the network's ability to extract fine-grained features. The neck structure is improved by using the bidirectional feature pyramid network to enhance the network's ability to extract features at multiple scales, thereby improving its overall multiscale feature extraction ability. Subsequently, we have replaced the C3 module in the original network's neck with the C2f module to obtain more abundant gradient flow information. This helps to improve the network's performance further. Finally, the coordinate attention module is introduced into the cross-scale feature connection path, which effectively enhances the network's target detection and recognition performance. We have conducted extensive comparative experiments and ablation experiments on the publicly available datasets ShipRSImageNet and SAT-MTB to confirm the effectiveness of our proposed SVSDet method. The performance of this approach is then evaluated using Jilin 1 satellite video data, and it outperforms the main YOLO series algorithms currently used.**

***Index Terms*—Attention mechanism, deep learning, multiscale feature fusion, satellite video, ship recognition.**

I. Introduction

THE sea is an important area for human activity and plays a crucial role in national defense, military operations, economic growth, and transportation connectivity. The most essential ways of transportation for people to engage in a variety of activities in the ocean are ships, and it is crucial for coastal surveillance and defense that ships be rapidly and precisely detected and recognized [1]. Furthermore, due to the vast and diverse types and sizes of ships distributed across the sea, it is necessary and technically challenging to quickly and accurately detect and identify them [2]. The observation of the Earth from space using high-resolution satellite remote sensing technology has developed into a significant observation approach and is frequently employed in ship detection and recognition. As a "staring" observation of a sensitive area, video satellites can continuously observe the changes in the field of view and obtain more dynamic information about the target area by means of "space video recording" than the traditional Earth observation satellites [3]. The continuous imaging mode of satellite video makes a strong correlation between the front and back frames and can provide rich contextual information. The fine-grained detection and recognition of ship targets in satellite video and the full use of satellite video context information can more accurately monitor and track ship dynamics in the target area.

Traditional methods for video object detection include the frame difference method, background modeling method, optical flow method, and others. Kopsiaftis and Karantzalos [4] used the background difference method to calculate the current frame and background template for the Skysat-1 satellite video to get the target to be detected. Zhang [5] used the video data of the "Jilin-1" satellite based on the classical algorithm random neighborhood and region matching method, combined with the motion vector assistance of the optical flow method and refined processing to obtain detection results. To address errors caused by global scene motion and local pseudomotion, Xu et al. [6] proposed a method of global motion compensation and local dynamic updating. Zhang et al. [7] integrated known satellite attitude motion information and unknown object motion information to detect target objects in satellite videos. Yang et al. [8] introduced a method that combines dynamic scene motion heat maps to enhance the detection of moving vehicles in video satellite images using saliency background models. Du et al. [9] proposed a multiframe optical flow tracker for object tracking. Although these methods have achieved good results in target detection in satellite video data, they cannot identify the target in fine-grained classification.

Convolutional neural networks (CNNs), which have an excellent ability to represent features, have experienced significant success in the field of computer vision since the deep-learning approach was proposed. Deep-learning-based object detection and recognition algorithms are continually being upgraded due to the rapid development of computer vision technology, and they are progressively taking over as the most common method. Several target detection algorithms have been proposed; you only look once (YOLO) [10] series and R-CNN series [11], [12], [13] algorithms have become the representative algorithms of one-stage detectors and two-stage detectors. Bousetouane and Morris [14] extracted ship target features based on CNNs

Manuscript received 24 August 2023; revised 15 October 2023, 10 November 2023, and 26 December 2023; accepted 22 January 2024. Date of publication 29 January 2024; date of current version 20 February 2024. This work was supported by the National Key Research and Development Program of China (No. 2017YFC1405600). *(Corresponding authors: Shanwei Liu; Mingming Xu.)*

The authors are with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China (e-mail: shanweiliu@163.com; buxiupc@163.com; xumingming@upc.edu.cn; sheng@upc.edu.cn; zengzhe@upc.edu.cn; ls1801004@s.upc.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3359252

and realized ship identification by comparing the target features with the template library. Guo et al. [15] proposed a rotation-balanced R-CNN method that uses three levels of rotational angle information to balance neural networks to predict ship positions. Zhang et al. [16] used the improved faster R-CNN for ship detection, adding the steps of image preprocessing by using a support vector machine and effectively improving the detection accuracy of small target ships. Sun et al. [17] carried out ship target detection based on the YOLO algorithm and marked the target position with a rectangular box corresponding to the ship target to realize multitarget detection, which significantly improved the detection performance. Among them, YOLO is favored in target detection and recognition because of its faster detection speed and accuracy no less than the two-stage networks. These algorithms have been introduced to the application of remote sensing images and have performed well as a result of the effective implementation of these algorithms on natural images among researchers in the field of remote sensing. Currently, CNN-based target detection methods can be classified into anchor-based and anchor-free (AF) methods according to whether the anchor-frame mechanism is used [18]. The anchor-based method reduces the training difficulty by incorporating prior knowledge in the form of preset anchor boxes for different feature maps. However, the size of the anchor box needs to be set manually and fixed, which limits its applicability for remote sensing images with large-scale changes. In addition, the use of a large number of preset anchor boxes can lead to an imbalance between positive and negative samples, further affecting the performance of the network. Compared with anchor-based methods, the AF approach does not require the production of anchors, which reduces the complexity of manual intervention and design and offers greater flexibility for multiscale target processing.

While CNN-based target detection algorithms have been successful in natural images and have also been developed for remote sensing images, current research in the field of remote sensing is primarily based on static high-resolution optical remote sensing images and synthetic aperture radar (SAR) images. In remote sensing images, ship targets have the characteristics of large aspect ratio and large-scale variation. At the same time, ship targets in remote sensing images are often small targets that occupy only a few pixels. Neural networks based on CNNS are challenging to deal with these problems [19]. Compared with the traditional static image, the video satellite image has the characteristics of real-time dynamic and high-time resolution, but the spatial resolution is less than that of a high-resolution optical image. For target detection in satellite video, multiscale and small target problems are more challenging. The authors in [20] and [21] employed CNN methodologies for the detection of vehicle targets in satellite videos. Meanwhile, Yan et al. [22] proposed a CNN object detection approach based on deep regression and incorporating transfer learning, facilitating the detection and categorization of moving objects. Zhang et al. [1] improved the SSD algorithm and applied it to the ship target detection of Jilin-1 satellite video. However, as far as we know, there are few applied researches on fine-grained ship recognition and classification of satellite video.

Based on the characteristics of video satellite data, a fine-grained ship target detection method SVSDet applied to video satellite data is proposed in this article. This method is improved by using AF YOLOv5 implemented under the YOLOv8 framework as the baseline network. The space-to-depth (SPD) module has been introduced into the backbone of the network to enhance its ability to extract fine-grained features. Bidirectional feature pyramid network (BiFPN) is used to improve the neck structure and enhance the multiscale feature extraction capability of the network. The coordinate attention (CA) module is introduced into the cross-scale connection, which enhances the feature expression ability of the network effectively. According to the test results and a substantial number of experiments, our proposed method has demonstrated excellent performance.

The main contributions of this work are as follows.

1) Adding the SPD module to the backbone structure of the network to enhance the ability of fine-grained feature extraction and the difficulty of small target detection.
2) The neck structure is enhanced by using the idea of BiFPN, and multiscale feature information is fused to enhance the multiscale feature extraction ability of the network. At the same time, introduced CA in the neck to make the model more accurately locate and identify the target area and enhance the detection and recognition performance of the network.
3) Replaced the C3 module in the original neck structure with the C2f module to obtain more abundant gradient flow information.

The rest of this article is organized as follows. Section II reviews a series of existing related works. Section III explains the proposed method in detail. Section IV designs the relevant experiments and analyzes the experimental results. Section V discusses the proposed method. Finally, Section VI concludes this article.

II. Related Work

A. You Only Look Once

Among various object detection algorithms, the YOLO framework stands out for its excellent balance of speed and accuracy, which enables fast and dependable identification of objects in images [10]. Since it was proposed, YOLO has been updated and developed many times, and has become one of the most advanced object detection frameworks. In 2015, Redmon et al. [23] first proposed a real-time end-to-end object detection method, YOLO, and published it in CVPR 2016. Compared to extracting the interested area and then running the classifier of the two-phase detection method, YOLO can pass a network to complete the detection task. Subsequently, Redmon and Farhadi made a series of improvements on the basis of YOLO, releasing YOLOv2 [24] and YOLOv3 [25]. The success of YOLO has led more researchers to participate in the research of YOLO series algorithms. Although different authors have published some versions of YOLO, they have largely maintained the same YOLO philosophy. So far, the mainstream YOLO series algorithm has been updated to YOLOv8 [26], [27], [28]. At the same time, the author of YOLOv8 implemented the AF version of YOLOv5 in

this framework, which has better performance than the original anchor base (AB). Of course, there are many other derivative versions of YOLO, such as Scald-YOLOv4 [29], YOLOX [30], YOLOR [31], DAMO-YOLO [32], and PP-YOLO [33] The continuous iteration of these algorithms has made the YOLO series more powerful time after time.

B. Remote Sensing Image Ship Detection

Feature descriptors and other techniques to extract ship candidate areas were also frequently used in the early stages of ship target detection and classification recognition, as were threshold segmentation based on statistical models, background separation based on significance, and feature classifiers, such as SVM and other feature classifiers [34]. Later, the development of CNNs provides a more efficient and effective way for ship target detection. Qu et al. [35] improved YOLOv3 and achieved satisfactory improvement in speed and accuracy. Hu et al. [36] improved YOLOv4 to reduce missed detection and false positives in complex scenarios and improve network performance. The method of deep learning relies on the establishment of datasets. Many scholars have carried out corresponding research and published their established datasets for the detection and fine-grained classification of ships in remote sensing images. Zhang et al. [2] published ShipRSImageNet, a ship dataset with multiple levels of mission types, and conducted a large number of experiments using popular target detection algorithms, such as faster R-CNN, FCOS [37], and SSD [38], to verify the feasibility of the dataset. Li et al. [39] proposed a dynamic soft label assignment strategy, which uses the dynamic anchor quality score threshold to replace the fixed IOU threshold, effectively improving the detection ability of ship targets in any direction. X. Zhang and Zhang [40] introduced the ASPP module and CBAM attention mechanism in YOLOv5 to strengthen spatial and semantic features of different scales, and used focal loss as a loss function to improve data imbalance. Aiming at ship target detection in SAR images, Kong et al. [41] proposed a multiscale ship feature extraction module and an overall detection strategy based on adaptive threshold, which suppressed background interference and improved detection accuracy. Yasir et al. [42] used C3 and FPN+PAN structures and attention mechanisms to enhance the backbone and neck of the YOLOv5 model to achieve a high recognition rate of the SAR ship.

C. Satellite Video Object Detection

In comparison to the image-based detection method, the video-based object detection method incorporates the integration of space–time features among frames and includes contextual information to enhance the algorithm's detection capability [43]. The application of deep learning has made great achievements in image target detection and has also been developed in video target detection tasks. T-CNN [44] combines time information in dynamic object tracking and spatial information in image object detection to improve the performance of video object detection. FGFA [45] fuses the features of adjacent video frames into the optical flow information extracted by FlowNet to improve the algorithm's ability to distinguish target features. DFF [46] only runs convolutional subnetworks on sparse key frames and propagates feature maps to other frames through streamlight information, which effectively improves the speed of the network. SELSA [47] uses semantic similarity to guide global feature fusion at the full sequence level, instead of relying on optical flow information, and improves the robustness of detection through feature fusion of completely random sampling frames.

Compared with general video, satellite videos have the characteristics of poor data continuity, global motion due to platform movement, large redundancy between video frames, etc. [48]. These greatly increase the difficulty of satellite video processing. The traditional methods for objects detection in satellite video mainly use background modeling method [49], frame difference method [50], and saliency-based method [51]. These methods capture the changing regions in satellite video sequence images, extract moving objects, and complete object detection tasks [48]. Lei et al. [49] proposed a background modeling method combining vital interframe temporal information to optimize detection results to detect vehicle targets in satellite video. Li et al. [50] propose a method for detecting and tracking moving ships of multiple sizes based on interframe differences. Li and Man [51] proposed a method based on visual attention saliency combined with optical flow information to effectively extract moving ship targets. Zhang et al. [52], [53], [54] established a series of methods for moving vehicle detection based on low-rank structured sparse decomposition. Traditional methods for detecting targets in satellite videos can only detect objects with motion states in the video. For stationary targets, such as ships docked at the shore, targets tend to be ignored. In addition, the traditional method is based on the detection of the difference between different frames, which cannot distinguish the target in categories. At present, the research on object detection in satellite video based on deep learning is still in its infancy [43]. At the same time, some existing research object for satellite video target detection datasets mainly focus on moving vehicles [55], [56], [57], [58]. Li et al. [55] proposed a motion-driven RGB image differential fusion network to detect moving objects in optical remote sensing satellite videos by combining time information from adjacent frames and spatial features from image pixels. Xiao et al. [56] proposed a network DSFNet, which combines static features obtained from single-frame images and motion features obtained from continuous frames, to detect moving targets in satellite videos. Feng et al. [57] combined the cross-frame key point detection network with the space motion information guidance and tracking network to detect and track moving vehicle targets in satellite video. Li et al. [43] released the dataset SAT-MTB, which not only contained ship targets but also classified ship targets in a fine-grained manner, making a great contribution to solving the difficulty of the lack of public data in fine-grained ship detection of satellite video.

D. Attention Mechanism

The attention mechanism can be considered a dynamic weight adjustment process for input image features that can effectively find significant areas in complex scenes by imitating human vision [59]. Currently, the attention mechanism is an essential

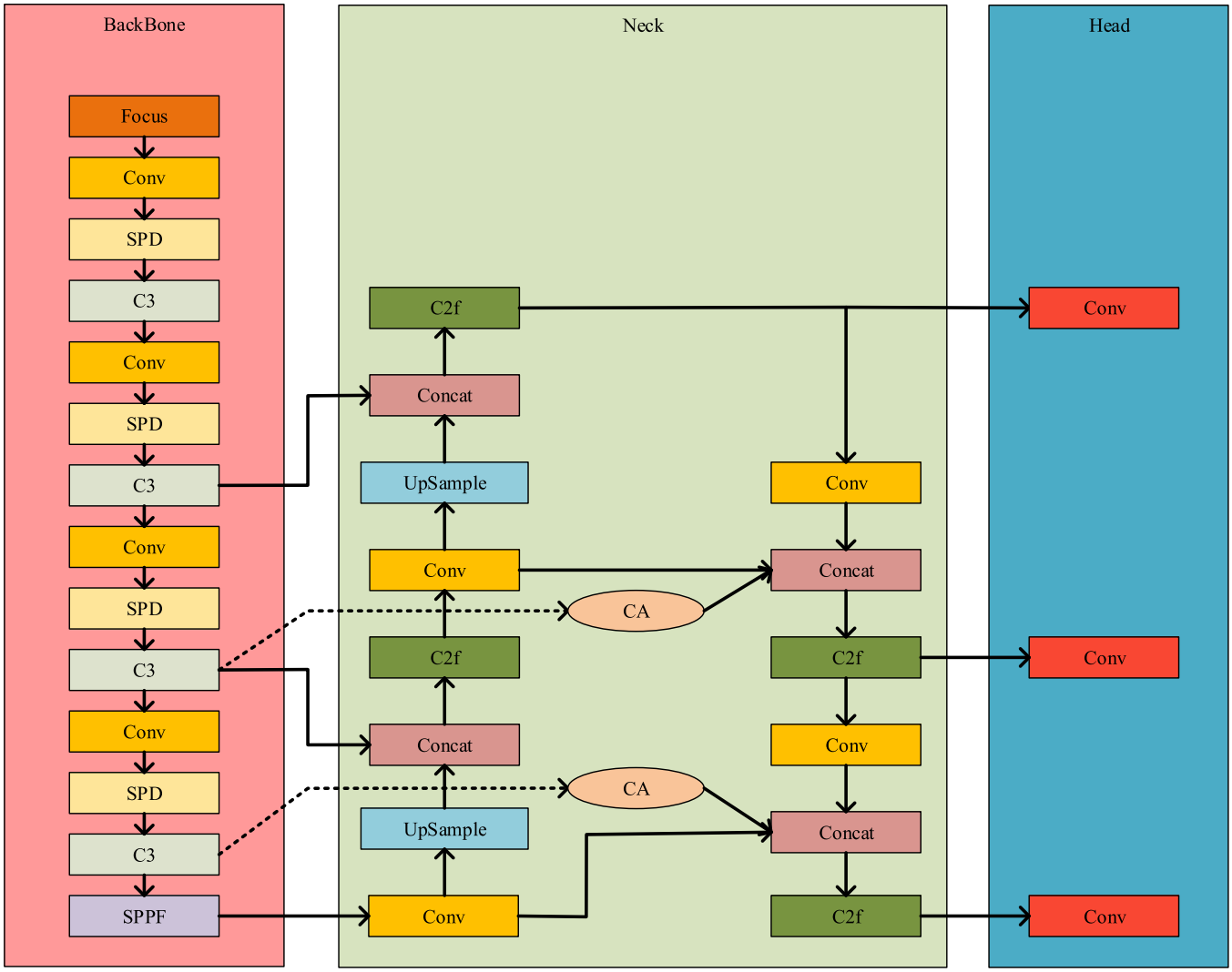


Fig. 1. Overall SVSDet network structure.

component of deep learning. In object detection tasks, common attention mechanisms include channel attention, spatial attention, and channel and spatial attention. Channel attention can capture different channels of information flow and enhance the channel, and the information transmission can adaptively adjust the importance of each channel, capturing useful features, such as SENet [60], FcaNet [61], and ECANet [62]. A spatial attention mechanism is an adaptive spatial region selection mechanism that relies on spatial relations to select regions of interest, such as RAM [63] and DCN [64]. CBAM [65], CA [66], and other integrated attention mechanisms that consider both channel and spatial relations of features have been proposed, which can give full play to the advantages of the two attention mechanisms and adaptively select the features and regions of interest.

III. Proposed Method

The complete network structure of SVSDet proposed in this article is shown in Fig. 1. In this network, an SPD convolution designed for extracting small target features is added to the backbone, a redesigned neck structure is used instead of the original path aggregation network (PANet) to aggregate the feature maps of different feature layers extracted by the backbone network, and a C2f module is used to obtain more abundant gradient flow information. Finally, the CA mechanism is used in the cross-scale connection of the neck to select important features and regions and then converge to improve network performance

A. SPD Model Into Backbone

How to effectively improve the detection ability of small targets has always been a concern in the field of target detection. Due to the unique imaging method and the resolution limitation, the difficulty caused by small targets has always been an important research topic in remote sensing image target detection. Lin et al. [67] propose that the poor performance of the existing CNN network in low-resolution or small objects is due to the loss of fine-grained features caused by the use of stride convolution

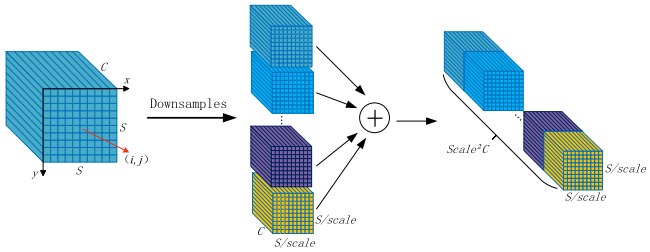


Fig. 2. SPD downsampling.

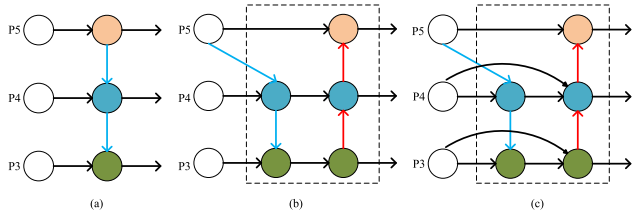


Fig. 3. Different feature fusion network structures.

and/or pooling layers. To solve this problem, the authors of this article propose a new CNN building block called SPD. Considering the difficulty of detecting small- and medium-sized targets in remote sensing images and the limitation of resolution, this article introduces the SPD module into the backbone of YOLOv5 to improve the network's ability to extract fine-grained information.

SPD module subsamples the feature mapping of CNN. For any feature mapping X with size $S\times S\times C$, the subfeature mapping series is sliced, as given in, (1) shown at the bottom of the this page.

In general, given any feature map X, a submap $f_{x,y}$ is formed by all the entries $X(i, j)$ that $i+x$ and $i+y$ are divisible by scale. Therefore, each subgraph downsamples X by a scale factor. Downsampling X according to the scaling factor yields feature subgraphs $f_{0,0}$, $f_{0,1}$, …, $f_{0,\text{scale}-1}$, $f_{\text{scale}-1,\text{scale}-1}$. The shape of each feature subgraph is $(\frac{S}{\text{scale}}, \frac{S}{\text{scale}}, C)$, then connecting these subfeature maps along the channel dimension to obtain a new feature map X′. Thus, the original feature map X ($S\times S\times C$) is converted to $X'(\frac{S}{\text{scale}}, \frac{S}{\text{scale}}, \text{scale}^2C)$, which reduces the spatial dimension by a factor of scale and increases the channel dimension by a factor of scale^2. The SPD downsampling process is shown in Fig. 2.

B. Improved Feature Fusion Neck Network

The large-scale difference of the detected object is another feature of remote sensing image target detection. Due to the large size difference of different ships, the scale change of ship targets is more prominent under the influence of different imaging angles and other factors in remote sensing images. In order to solve the multiscale problem of the detected target, feature pyramid networks (FPN) are integrated into the target detection network. By adopting up and down and connecting horizontally, FPN fuses feature maps of different levels to improve the detection capability of targets of different scales. PANet [68] adopts a top-down path to propagate features, fuse feature maps of upper and lower adjacent layers, and sum the fusion results. This method can retain more detailed information and be more efficient. Tan et al. [69] believe that different features are of different importance and propose an efficient BiFPN. We adopted the idea of bidirectional cross connection in BiFPN to improve PANet and added two cross-scale connection paths for feature map fusion to obtain more feature information of different scales. Different feature fusion network structures are shown in Fig. 3.

C. C2f Module

The C2f module, introduced in YOLOv8, serves as a novel feature extraction module, replacing the original C3 module in YOLOv5. The C2f module retains the core concept of the C3 module and is designed with reference to the layer aggregation architecture for effective gradient propagation, as proposed in ELAN [70]. C2f employs a split operation to divide feature maps, sending one-half of the feature maps into the BottleNeck module and merging the other half with the output feature maps and residuals of each BottleNeck module. This process enhances the network's ability to capture richer gradient flow information. After extracting features comprehensively from the backbone network, feature fusion takes place in the neck network. Replacing the C3 module with the C2f module in the neck structure allows for the preservation of finer textures and other features, ultimately improving the network's classification capabilities for recognizing targets. The bottleneck in C3 and C2f is a residual network structure block, which effectively reduces network parameters and promotes network optimization

$$
\begin{aligned}
f_{0,0} = X\left[0:S:\text{scale}, 0:S:\text{scale}\right],\ f_{1,0} =\ & X\left[1:S:\text{scale}, 0:S:\text{scale}\right],\\
\ldots, f_{\text{scale}-1,0} =\ & X\left[\text{scale}-1:S:\text{scale}, 0:S:\text{scale}\right];\\
f_{0,1} =\ X\left[0:S:\text{scale}, 1:S:\text{scale}\right],\ & f_{1,1}, \ldots, f_{\text{scale}-1,1} =\ X\left[\text{scale}-1:S:\text{scale}, 1:S:\text{scale}\right];\\
&\vdots\\
f_{0,\text{scale}-1} =\ & X\left[0:S:\text{scale}, \text{scale}-1:S:\text{scale}\right], f_{1,\text{scale}-1},\\
f_{\text{scale}-1,\text{scale}-1} =\ & X\left[\text{scale}-1:S:\text{scale}, \text{scale}-1:S:\text{scale}\right].
\end{aligned}
\tag{1}
$$

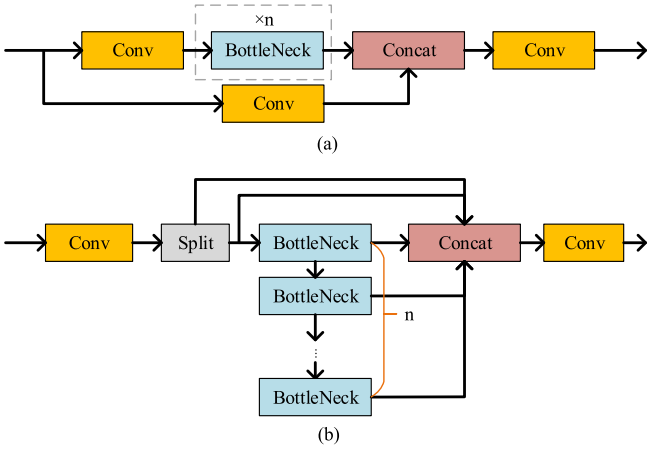


Fig. 4. Bottleneck module structure diagram of C3 and C2f modules, where *n* indicates the number of bottleneck modules. C3 show as (a) and C2f show as (b).

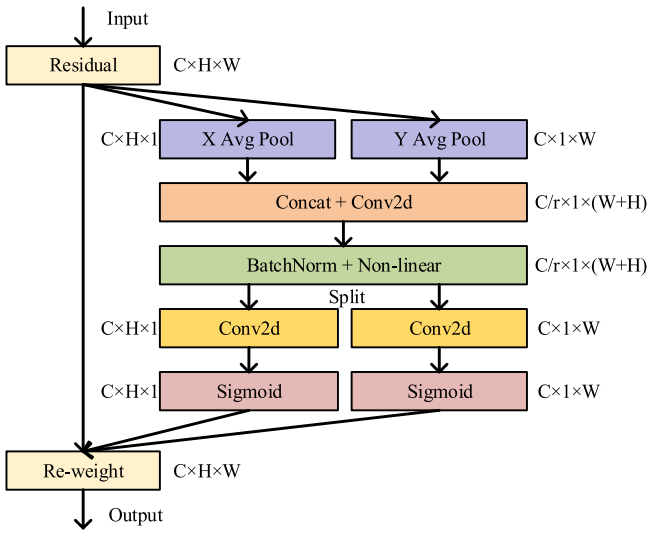


Fig. 5. CA module network structure.

in deep neural networks. The structures of C3 and C2f show in Fig. 4.

D. Coordinate Attention

An attention mechanism has been proposed, which greatly improves the feature extraction capability of deep-learning networks and is widely used in object detection, semantic segmentation, natural language processing, and other tasks. Channel attention focuses on important features, spatial attention focuses on the areas of interest, and mixed channel and space attention have both features. Hou et al. [66] proposed that CA is embedded with position information in channel attention to solve the problem that mixed attention, such as CBAM, is difficult to capture long-distance dependence. This mechanism can obtain information on a larger area at a lower computational cost so as to infer the location and attribute of the target more accurately. The CA structure is shown in Fig. 5.

CA is coded for each channel using a pooling check of size *(H,1)* or *(1,W)* in the horizontal and vertical directions, respectively, embedding coordinate information and connecting the output of the two pooling layers using a shared 1×1 convolution transform function for CA generation. It is described by formulae (2)–(8)

$$z^h = \mathrm{GAP}^h\ (X) \tag{2}$$

$$z^w = \mathrm{GAP}^w\ (X) \tag{3}$$

$$f = \delta\left(\mathrm{BN}\left(\mathrm{Conv}_1^{1\times1}\left(\left[z^h; z^w\right]\right)\right)\right) \tag{4}$$

$$f^h,\ f^w = \mathrm{Split}\,(f) \tag{5}$$

$$s^h = \sigma\left(\mathrm{Conv}_h^{1\times1}\left(f^h\right)\right) \tag{6}$$

$$s^w = \sigma\left(\mathrm{Conv}_w^{1\times1}\left(f^w\right)\right) \tag{7}$$

$$Y = X s^h s^w \tag{8}$$

where h and w represent the horizontal and vertical directions, respectively, GAP is the global average pooling function, and $s^h \in \mathbb{R}^{C\times1\times W}$ and $s^h \in \mathbb{R}^{C\times H\times1}$ are the weights of the corresponding direction coordinates.

IV. Experiment and Result

To verify the effectiveness of our proposed method, we conducted comprehensive experiments on the public fine-grained ship classification dataset ShipRSImageNet and tested the practical application effect on the video data collected by the Jilin-1 video satellite. In this section, we will explain our experimental environment, experimental data, and evaluation indicators, conduct detailed ablation experiments to verify the effectiveness of each improvement, and conduct comparative tests with other popular methods to verify the advanced nature of the proposed method. Then, the performance of the model was improved through pretraining on the large aerial remote sensing dataset DOTA [71]. Finally, the model obtained by the proposed method is tested in practice to verify its feasibility.

A. Datasets

1) ShipRSImageNet: ShipRSImageNet integrates multiple existing ship datasets and optical remote sensing images from different data acquisition platforms and accurately classifies and labels ships in the images [2]. The dataset contains more than 3400 images and 17 500 ship instances, enabling multilevel detection and fine-grained classification tasks. Except for a Dock category, Level 0 tasks are single-class ship detection tasks. Level 1 tasks are divided into three ship categories: Other ship, Warship, and Merchant; Level 2 tasks are divided into 24 ship categories; Level 3 tasks are divided into 49 ship categories; and Other ships are ships that cannot be identified as merchant ships. Because of the large differences in size, resolution, scene, and other aspects of the image in the dataset, it is challenging to use the dataset to detect the performance of the method. Different levels of classification in different scenarios are shown in Fig. 6.

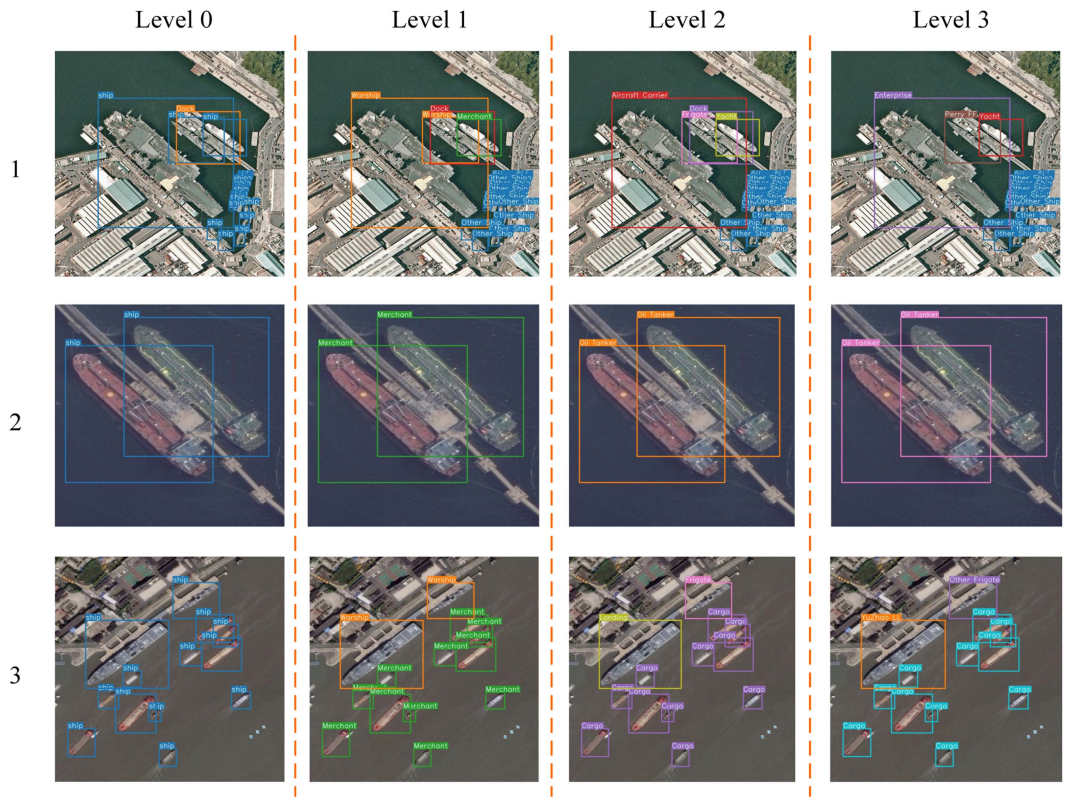


Fig. 6. Classification of ships at different levels of tasks in three scenarios.

Fig. 7. Instance of the Jilin-1 video dataset used in the current research.

2) Satellite Video: The satellite video data used in this article were obtained by Jilin-1 Video 03 satellite with a spatial resolution of 0.92 m. The video duration was 10 s and a total of 100 frames, each frame size being 6786 × 2528 instance, as shown in Fig. 7.

3) SAT-MTB: The dataset consists of 249 satellite videos from Jilin-1, annotating 4 coarse-grained categories of aircraft, ships, cars, and trains, and 14 corresponding fine-grained target categories. Among them, there are six fine-grained categories of ships, including speed boat, yacht, cruise, freighter, naval vessels, and other ships. The dataset can meet the needs of three tasks: detection, tracking, and segmentation. To accommodate the target tracking task, only moving ship targets offshore are annotated. Some instances are given in Fig. 8.

4) DOTA: The DOTA dataset comprises 2806 aerial images with dimensions of 4000 × 4000 pixels, each containing objects with varying proportions, orientations, and shapes. The images underwent expert tagging and annotation, capturing a total of 188 282 instances spanning 15 common object categories.

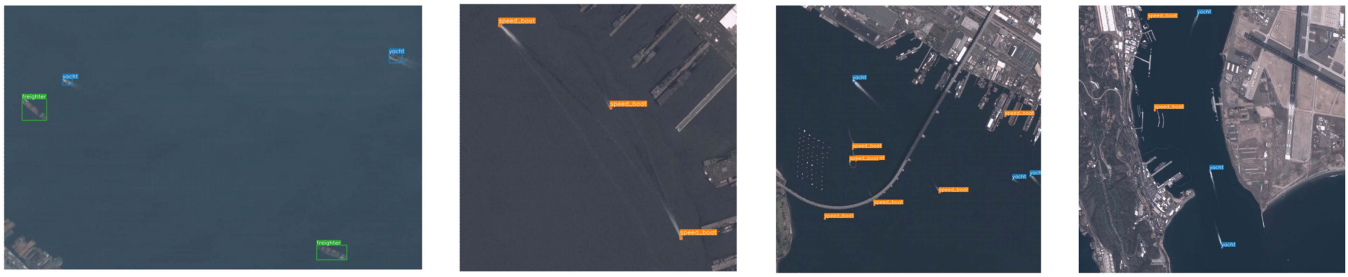


Fig. 8. Some instances of the SAT-MTB.

TABLE I
CONFIGURATION PARAMETERS RELATED TO THE EXPERIMENTAL ENVIRONMENT

Configuration	Parameters
Operating system	Windows10
CPU	Intel(R) Core(TM) i5-12490F
RAM	16G
GPU	Nvidia Gforce RTX3060 12G
CUDA	11.3
Python	3.7
Pytorch	1.10

B. Experimental Environment

All experiments in this article were carried out in the same environment, and the relevant configuration parameters of the specific environment are shown in Table I.

C. Evaluation Index

The two evaluation indices that are widely used in target detection tasks are precision–recall (*P*-*R*) curve and average accuracy (AP). The precision is expressed as (9), and the recall is expressed as (10). AP is the average accuracy of all unique recall levels, and only a single category is included in the calculation. There are usually K classes in multicategory detection tasks, so the AP average mAP of K classes is introduced as the evaluation index of multicategory target detection. The precision is expressed as (11)

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

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

$$\text{mAP} = \frac{\sum_{i=1}^{K}\text{AP}_i}{K} \tag{11}$$

where TP represents the number of true positives, FP represents the number of false positives, and FN represents the number of false negatives.

The intersection ratio (IOU) calculates the ratio of the intersection and union between the predicted bounding box (Box_p) and the real bounding box (Box_{gt}). Different IOU thresholds have a great impact on the mAP result. Normally, the default threshold is $\text{Threshold}_{\text{IOU}} = 0.5$, which is mAP@50, so as to improve the positioning accuracy of the detector, MS-COCO adopted [0.50:0.05:0.95] ten mAP values with different IOU thresholds to evaluate the performance of the detector, i.e., mAP@50:95. The IOU calculation formula is given as follows:

$$\text{IoU} = \frac{\text{area}\left(\text{Box}_p \cap \text{Box}_{gt}\right)}{\text{area}\left(\text{Box}_p \cup \text{Box}_{gt}\right)}. \tag{12}$$

In addition, $F1$ is a common index used to measure the accuracy of binary classification models, which can take into account both accuracy rate and recall rate. In multiclass problems, the generalization method Micro-$F1$ is often used as an evaluation metric for multiclassifiers. First, calculate the average precision and recall for each category, as shown in (13) and (14). Then, the Micro-$F1$ is calculated by (15)

$$\text{Recall}_m = \frac{\sum_{i=1}^{K}\text{TP}_i}{\sum_{i=1}^{K}\text{TP}_i + \sum_{i=1}^{K}\text{FN}_i} \tag{13}$$

$$\text{Precision}_m = \frac{\sum_{i=1}^{K}\text{TP}_i}{\sum_{i=1}^{K}\text{TP}_i + \sum_{i=1}^{K}\text{FP}_i} \tag{14}$$

$$\text{micro} - F1 = 2\frac{\text{Recall}_m \times \text{Precision}_m}{\text{Recall}_m + \text{Precision}_m}. \tag{15}$$

D. Comparison Experiments

1) Comparison Experiments on ShipRSImageNet: In this part, we compare the proposed SVSDet with several popular YOLO series target detection methods on all levels of the ShipRSImageNet dataset and verify the detection performance of SVSDet. Comparison methods include YOLOv5, YOLOv6, YOLOv7, YOLOv8, and YOLOX, in which YOLOv6, YOLOv8, and YOLOX are AF methods, and YOLOv7 is AB method. YOLOv5 includes the original AB method and the AF method implemented in the YOLOv8 framework. The YOLO series of networks are typically categorized into several versions, namely *n, s, m, l*, and *X*, which represent different network sizes. In this study, with the exception of YOLOv7, all other models were evaluated using the *s*-sized network version. YOLOv7 does not have a version with size *s* and uses Tiny version with a relatively close size to other models. Tables II –V list the comparative experimental results.

In the three tasks of 0,2,3, our proposed method achieves the best performance in each index. Specifically, for level 0 tasks, SVSDet's precision (P) is 81.2%, which is on par with YOLOv5-AF but gets the best score compared with other models. The recall (R), mean average precision (mAP) of SVSDet, and the

TABLE II
COMPARES THE RESULTS OF EACH MODEL ON THE LEVEL0 TASK

Method	P/%	R/%	mAP/%	mAP@.50:95/%
YOLOv5-AB	73.5	71.9	75.8	51.2
YOLOv5-AF	81.2	77.9	82.1	61.5
YOLOv6	79.3	65.1	71.5	52.1
YOLOv7	64.2	62.3	64.8	40.2
YOLOv8	80.4	76.4	81.9	63.0
YOLOX	77.4	76.0	79.3	53.0
SVSDet	**81.2**	**79.8**	**84.5**	**63.6**

The highest values of our model are bold.

TABLE III
COMPARES THE RESULTS OF EACH MODEL ON THE LEVEL1 TASK

Method	P/%	R/%	mAP/%	mAP@.50:95/%
YOLOv5-AB	65.3	66.2	67.1	44.6
YOLOv5-AF	67.5	68.7	69.5	51.9
YOLOv6	62.4	67.3	65.1	47.0
YOLOv7	62.8	58.7	60.9	38.6
YOLOv8	70.7	69.4	71.3	54.7
YOLOX	**71.5**	68.1	71.6	49.5
SVSDet	68.5	**73.4**	**72.9**	**55.1**

The highest values of our model are bold.

TABLE IV
COMPARES THE RESULTS OF EACH MODEL ON THE LEVEL2 TASK

Method	P/%	R/%	mAP/%	mAP@.50:95/%
YOLOv5-AB	51.4	57.0	54.6	39.7
YOLOv5-AF	54.8	56.8	58.1	46.0
YOLOv6	48.8	50.1	49.7	38.5
YOLOv7	41.6	45.9	42.9	31.5
YOLOv8	57.4	58.7	59.7	48.5
YOLOX	53.7	57.8	56.3	41.5
SVSDet	**59.7**	**59.1**	**61.2**	**48.6**

The highest values of our model are bold.

TABLE V
COMPARES THE RESULTS OF EACH MODEL ON THE LEVEL3 TASK

Method	P/%	R/%	mAP/%	mAP@.50:95/%
YOLOv5-AB	41.0	59.2	53.1	39.2
YOLOv5-AF	53.7	63.4	61.9	51.3
YOLOv6	54.2	55.7	56.8	46.6
YOLOv7	35.2	51.3	40.3	30.0
YOLOv8	59.3	63.3	65.1	54.7
YOLOX	57.5	57.4	60.0	45.6
SVSDet	**64.2**	**64.7**	**68.5**	**56.3**

The highest values of our model are bold.

mean average precision at different thresholds mAP@.50:95 are significantly improved. Compared with the baseline network YOLOv5-AF, *R* increased by 1.9%, mAP increased by 2.4%, and mAP@.50:95 increased by 2.1%. SVSDet can detect more targets without the loss of detection precision and has the best detection capability under different IOU thresholds. In the task of fine-grained ship target recognition at level 2 and level 3, each index of SVSDet outperforms other comparison models, indicating that our method can obtain more fine-grained features and accurately classify ship targets with more detailed features.

For level 1 tasks, compared with the baseline model YOLOv5-AF, SVSDet has improved all performance indices, *P* by 1%, *R* by 4.7%, mAP by 3.4%, and mAP@.50:95 by 3.2%. In the comparison of other comparison models, all of them are optimal except *P*. The precision of SVSDet is lower than that of YOLOv8 and YOLOX models, which may be due to the fact that the model obtains a large number of fine-grained features and merges multiscales while bringing some feature redundancy, resulting in certain interference noise. In addition, the increase in feature diversity will increase the complexity of the model, which increases the difficulty of the model to accurately learn and judge the target. However, in the case of a combination of multiple indicators, SVSDet is still a more balanced and excellent method in all aspects.

Figs. 9 and 10 depict the performance of each model in ShipRSImageNet. The results demonstrate that the proposed method excels in detecting and accurately identifying targets in complex backgrounds. Furthermore, it exhibits superior robustness compared with other methods in detecting and recognizing densely packed targets, and those with a large aspect ratio. Misclassified targets are indicated by red triangles, while missing detections are marked with green triangles. It can be seen from the results that SVSDet has fewer missed targets and misclassification cases when detecting the densely arranged targets in the scene of Fig. 9, and only one repeated detection box of misclassification appears in the eight targets. Among them, the YOLOv5-AB method with poor performance only detected three targets and made classification errors. When detecting the target in the scenario in Fig. 10, SVSDet misses a target and a correctly classified but repeated detection box. Other methods have more cases of missing detection and wrong classification, among which YOLOv7 and YOLOv8 two methods failed to detect the targets. At the same time, SVSDet can surpass other methods in the detection performance of various types of vessels, although the difference in the target scale of different types of vessels in these two scenarios is huge. This shows that our improvements are effective in solving multiscale problems.

The above experimental visualization results demonstrate that the proposed method performs well in the fine-grained ship classification task. In the multicategory fine-grained detection task, each index is the average score of a single category and is the comprehensive result of the evaluation index scores corresponding to different categories. When the detection performance of each category is improved, the comprehensive evaluation score of the model is also improved. Tables II–V list the comprehensive ability of each model to perform fine-grained classification at different classification levels, indicating the superiority of the classification recognition performance of our proposed model. It can also be seen in Figs. 9 and 10.

2) Comparison Experiments on SAT-MTB: In this section, we conduct detailed comparative experiments on the dataset SAT-MTB. Li et al. [39] conducted a large number of experiments on this dataset, including image-based object detection methods and video object detection methods. Therefore, we will use the optimal method based on image object detection YOLOv3 and the optimal method based on video object detection DFF provided by the author as the baseline results and

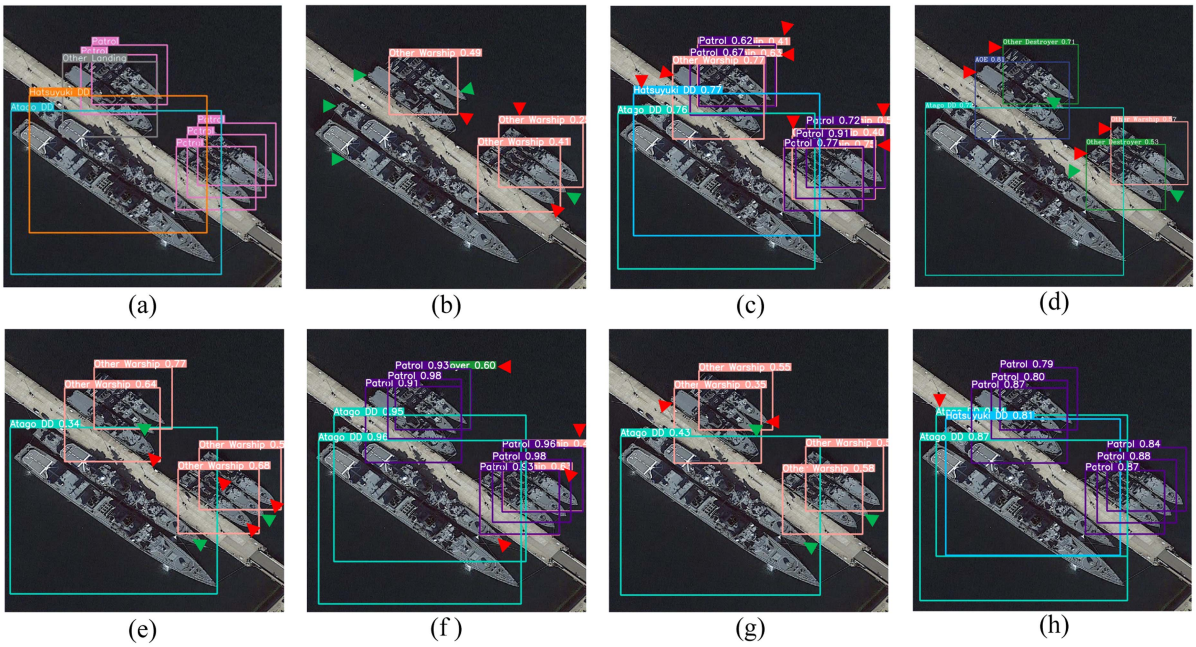


Fig. 9. (a) Ground truth. (b) Result of YOLOv5-AB. (c) Results of YOLOv5-AF. (d) Results of YOLOv6. (e) Results of YOLOv7. (f) Results of YOLOv8. (g) Results of YOLOX. (h) Results of SVSDet.

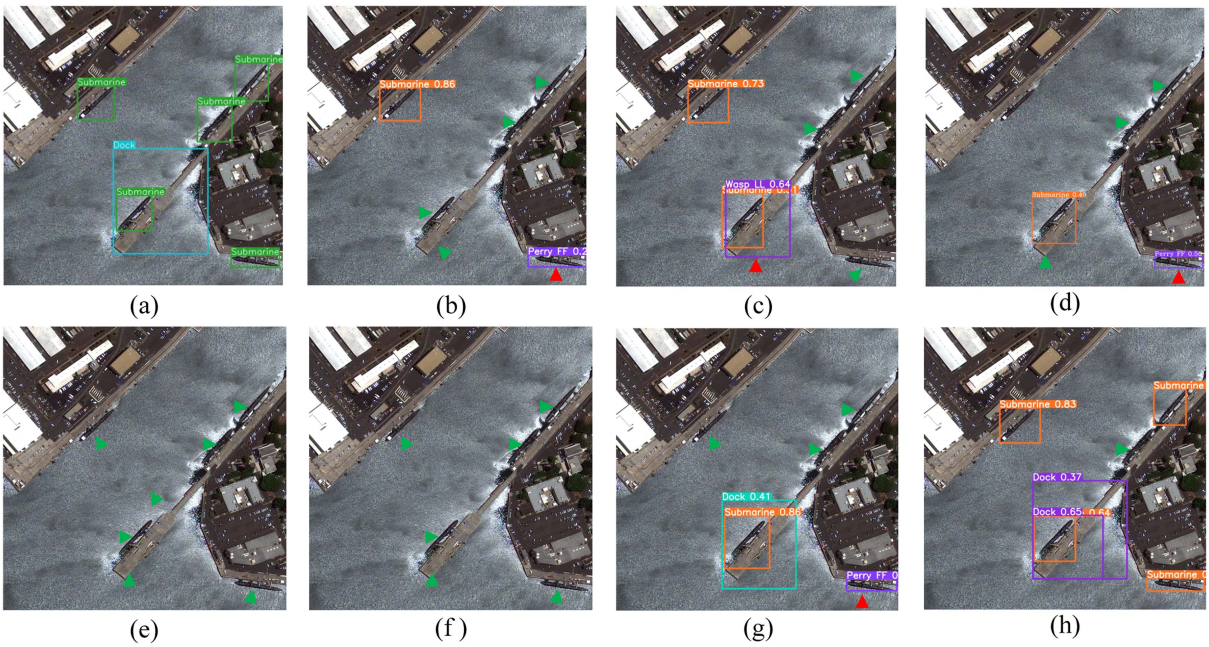


Fig. 10. (a) Ground truth. (b) Result of YOLOv5-AB. (c) Results of YOLOv5-AF. (d) Results of YOLOv6. (e) Results of YOLOv7. (f) Results of YOLOv8. (g) Results of YOLOX. (h) Results of SVSDet.

make a comprehensive comparison with our proposed method to verify its effectiveness. We compared mAP(%) with FPS, and the results of each method are shown in Table VI. As can be seen from the table, different models have different learning abilities for different target features. Among them, YOLOv3 demonstrates outstanding detection capabilities for the "othership" category but exhibits subpar performance in other categories. YOLOv7 demonstrates strong performance in detecting two small target classes, namely speed boat and yacht, indicating its proficiency in capturing features of diminutive objects. Although our SVSDet method is not optimal in a single class, it surpasses other methods in comprehensive performance. SVSDet balances the extraction of small objects and the multiscale problem between different classes and achieves good results in the detection of different scale objects. Therefore, the proposed method is more balanced and robust in

TABLE VI
COMPARES THE RESULTS OF EACH MODEL ON THE SAT-MTB

	DFF	v3	v5-AB	v5-AF	v6	v7	v8	x	SVSDet
Speed Boat	0	8.8	7.3	16.5	14.5	**27.0**	16.1	5.75	25.7
Yacht	6.7	29.4	46.1	42.9	41.4	**49.0**	48.4	48.2	45.1
Cruise	3.1	0.4	38.0	43.3	36.5	32.0	**44.2**	39.4	39.0
Freighter	21.2	5.5	**30.1**	22.3	17.8	20.5	20.1	24.0	27.8
Naval vessels	0	0	0.1	0	**0.9**	0	0	0.1	0.4
Other Ship	28.9	**65.7**	4.5	9.4	1.1	7.7	3.5	15.1	14.6
mAP	10.0	18.3	21.0	22.4	18.7	22.8	22.0	22.1	**25.4**
FPS	38.9	39.7	90.9	**113.6**	53.8	103.1	105.3	76.3	75.8

The highest values of our model are bold.

TABLE VII
COMPARISON OF THE RESULTS OF IMPROVED ABLATION EXPERIMENTS IN EACH PART

Method	P/%	R/%	mAP/%	mAP@.50:95/%
YOLOv5-AF	53.7	63.4	61.9	51.3
+SPD	53.4	64.2	63.6	53.4
Improve Neck	59.3	63.3	65.3	54.6
C2f	61.4	61.7	65.1	54.9
+CA	64.2	64.7	68.5	56.3

fine-grained detection tasks. At the same time, compared with the baseline method YOLOv5-AF, SVSDet's ability to detect different objects has been improved almost comprehensively. Compared with the method DFF for video target detection, it has obvious advantages in terms of detection ability and running speed.

E. Ablation Experiment

In this part, we conduct detailed ablation experiments on the dataset to verify the effectiveness of each part of the improvement. In this part of the experiment, we successively introduced SPD, improved the neck network, and added CA attention module to carry out the experiment, and compared it with the baseline network YOLOv5-AF. Specific experimental results are shown in Table VII.

For convenience, we provide a detailed description of the experimental results on the task at the highest level 3, which also yielded similar results on other level tasks.

Based on the experimental results, after implementing the SPD module, all metrics show improvement, with the most significant improvement observed in the mAP and mAP@.95 metrics, with an increase of 1.7% and 2.1%, respectively. This suggests that the implementation of the SPD module significantly enhances the network's feature extraction capability. After improving the neck and integrating multiscale features, there has been a decrease in the recall rate, which could be attributed to the interference caused by the inclusion of different scale features, resulting in the introduction of varying background information details. After replacing the C3 module with C2f in the neck network, the presence of abundant gradient flow information provides further improvement to P and mAP@.50:95. By incorporating the CA attention module into the cross-scale connection path, the feature information from different scales can be more effectively integrated, allowing for more accurate capture of the positional information of the target. This, in turn, enhances both the localization accuracy and overall target accuracy.

TABLE VIII
PERFORMANCE OF THE NETWORK AT ALL LEVELS OF TASKS AFTER USING PRETRAINING WEIGHTS

Task	P/%	R/%	mAP/%	mAP@.50:95/%
Level0	84.9	83.8	89.2	69.6
Level1	74.0	72.0	74.3	58.6
Level2	69.4	59.6	64.8	53.5
Level3	73.9	67.6	74.8	63.6

F. Pretraining

Pretraining is a commonly used method in the construction and use of deep-learning models. By pretraining on large-scale datasets with the same or similar features to capture common features, you can effectively improve network performance. In this part, we pretrain the proposed SVSDet network on the large aerial remote sensing dataset DOTA and use the pretraining results as the initial weights for training on the ShipRSImageNet dataset. Compared with other natural image datasets, the images in the DOTA dataset are sourced from various satellite remote sensing platforms. The imaging angles and other characteristics of the targets make it more suitable for remote sensing tasks, such as ship detection and recognition. Compared to training the network from scratch, the performance is significantly improved after using pretrained weights. Performance at all levels of tasks is shown in Table VIII.

G. Validation on Satellite Video

We apply the model trained on the level 1 task of the ShipRSImageNet dataset with the proposed method to the video data obtained by the JIL-1 video 03 satellite and compare the application results of each model. In the application, micro-$F1$ is used as an evaluation index to evaluate the model, and the micro-$F1$ results of each model in the video data are shown in Table IX. The micro-F1 of SVSDet proposed by us in practical application is 0.76, which is optimal compared with other models. Through pretraining of the model, the performance of the model is further improved. Due to the lack of accurate AIS data for the detailed classification of ships in the videos, we utilized

Fig. 11. Video data annotation.

TABLE IX
COMPARISON OF MICRO-$F1$ RESULTS OF EACH MODEL ON VIDEO DATA

Method	Micro-F1
YOLOv5-AB	0.72
YOLOv5-AF	0.71
YOLOv6	0.67
YOLOv7	0.72
YOLOv8	0.61
YOLOX	0.68
SVSDet	**0.74**
SVSDet-Pretained	**0.76**

The highest values of our model are bold.

TABLE X
EFFECT OF EACH PART IMPROVEMENT ON THE MODEL SPEED

Method	FPS
YOLOv5-AF	113.6
+SPD	79.4
Improve Neck	78.7
C2f	77.5
+CA	75.8

the classification criteria at level 1 from the ShipRSImageNet dataset to visually interpret and annotate the ships in the videos. The annotation results are depicted in Fig. 11, encompassing a total of 16 ship targets, including ten military vessels and six merchant ships.

The actual application results of each model are shown in Fig. 12. In the video, there are 16 ship targets, including 10 military ships and 6 merchant ships. It is not difficult to see from the experimental results that AF methods, such as YOLOv5-AF, YOLOv6, YOLOv8, and YOLOX, have a higher recall rate and can detect more targets because they are not limited by preset prior boxes, but there will be more false alarm targets. The AB methods of YOLOv5-AB and YOLOv7 are more likely to miss targets. Our proposed SVSDet method can effectively improve these problems. The addition of the SPD module effectively improves the model's ability to extract fine-grained features, which not only helps to detect small targets but also helps to improve the model's classification and recognition performance of targets with rich detailed features. The improved neck network fused feature maps from different levels of backbone to enhance the model's learning of multiscale information and improve the detection ability of multiscale targets.

Fig. 12(g) shows the results of SVSDet's lack of pretraining, when two merchant ships were not properly identified, two were classified as Othership, and one was misidentified as a warship. In port, one warship target was missed, and one warship was repeatedly detected and incorrectly identified as Othership. Fig. 12(h) shows the results of pretraining for SVSDet, where nine of the ten military ships in port were detected and correctly classified, and four of the six merchant ships offshore were detected and correctly classified, but two were misclassified as military ships. Although SVSDet still has a small number of missed and misclassified cases in ship fine-grained identification of video data, our proposed method obviously has stronger generalization performance than other models.

V. DISCUSSION

Obviously, the application prospect of fine observation of Earth's surface by high-resolution remote sensing image is broad. High-resolution video satellite provides a high dynamic and real-time observation method for Earth observation. By integrating with prevailing deep-learning methods, it becomes possible to acquire a richer set of deep features, thereby enhancing the feasibility of remote sensing technology. Most of the remote sensing satellites are optical and SAR remote sensing satellites, which acquire static images. So, the available satellite video data are limited. The current popular image-based deep-learning object detection method has proven its powerful object detection capability and has considerable efficiency with the support of GPU technology. Satellite video and static remote sensing images are essentially remote sensing data of optical imaging. It is meaningful to make full use of static remote

(a)

(b)

(c)

Fig. 12. Application results of each model on the frequency data of Jilin-1 satellite video. (a) Result of YOLOv5-AB. (b) Results of YOLOv5-AF. (c) Result of YOLOv6.

sensing images from a wider range of sources to build rich ship sample datasets and deep-learning models for satellite video.

In this article, we proposed a fine-grained ship target detection method based on YOLOv5 improvement by enhancing the network of fine-grained and multiscale feature fusion feature extraction ability to adapt to the small target and multiscale problems in satellite video. Through extensive experimental comparison, it is proved that our proposed method not only improves the detection capability of small objects but also has better multiscale target detection capability and is more robust in fine-grained target detection tasks.

However, we also regret to find that our improvements inevitably have some impact on the speed of the model. To examine

(d)

(e)

(f)

Fig. 12. (Continued..) Application results of each model on the frequency data of Jilin-1 satellite video.(d) YOLOv7 results. (e) Results of YOLOv8. (f) YOLOX result.

the specific factors, we conducted tests on the speed variations of the enhanced model across different components, and the corresponding outcomes are presented in Table X. The table clearly indicates that the incorporation of SPD modules in the backbone leads to the most significant decrease in speed. In order to better retain the detailed information in low-resolution images and small objects, the SPD module splits a large number of feature maps and increases the number of feature maps, which has a great impact on the model speed. The SPD module is utilized to enhance the preservation of detailed information in low-resolution images and small objects by partitioning and increasing the number of feature maps, resulting in a huge impact on the speed of the model. The neck improvement of the network, the fusion of multiscale features, and the use of C2f to obtain

(g)

(h)

Fig. 12. (Continued..) Application results of each model on the frequency data of Jilin-1 satellite video. (g) Results of SVSDet without pretraining. (h) Results of pretraining for SVSDet. The red circle is marked as the wrong classification, and the green circle is marked as missing detection.

more abundant gradient flow information also have a certain impact on the model speed. The addition of CA increases the computational complexity of the model to a certain extent, which also results in the reduction of network speed.

The balance between speed and precision is the goal of target detection. But we think accuracy is the more important factor. The problems caused by model speed can be improved by hardware upgrading or key-frame detection. Of course, it is still worthwhile to explore how to improve the efficiency of the model while enhancing its detection performance. Therefore, in future research, we will focus on this aspect to make the model more efficient and effective.

VI. Conclusion

In this article, an improved method SVSDet based on YOLOv5 is proposed to detect and identify ship targets in video satellite data. SVSDet incorporates SPD into the backbone network of YOLOv5 to enhance small object detection capabilities. It further improves the neck network architecture by introducing cross-layer connections for the fusion of multiscale information, addressing the challenge of detecting ships with varying sizes in remote sensing images. The C3 module is replaced by the C2f module to capture richer gradient flow information while preserving finer grained features. A CA attention mechanism is introduced in the cross-layer connections to improve the model's precision in locating targets during multiscale feature extraction and fusion. The experimental results demonstrate that our proposed method can achieve outstanding performance not only in single-class detection of ship targets but also in multiclass fine-grained recognition detection. On the satellite video dataset SAT-MTB, our approach is more robust across all categories of detection. The performance of all levels of tasks in the dataset ShipRSImageNet exceeds the current mainstream detection methods, and the application of video images also shows better generalization, with the highest micro-$F1$ index comparison of various models, reaching 0.76. The results of this study will promote the further development of the application of video satellites for ship detection and recognition. In future research work, we will study how to further improve the detection and recognition ability of the model and the light weight of the model so as to make it more lightweight and realize the detection and recognition of ship targets more accurately and efficiently while ensuring the detection ability of the model. In addition, how to effectively use the rich context information of satellite video to improve the stability of target recognition and real-time target tracking is also worth studying.

REFERENCES

[1] C. Zhang, Z. Wang, and H. Sheng, "Research on ship target detection based on satellite video," *Mar. Sci.*, vol. 45, no. 5, pp. 9–15, 2021.

[2] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "ShipRSImageNet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8458–8472, Aug. 2021.

[3] Z. Sun, L. Zhang, G. Jin, K. Xu, and M. Chen, "Simulation and experiment on attitude tracking control of small TV satellite," *Opt. Precis. Eng.*, vol. 19, no. 11, pp. 2715–2723, 2011.

[4] G. Kopsiaftis and K. Karantzalos, "Vehicle detection and traffic density monitoring from very high resolution satellite video data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1881–1884.

[5] G. Zhang, "Satellite video processing and applications," *J. Appl. Sci.*, vol. 34, no. 4, pp. 361–370, 2016.

[6] A. Xu et al., "Motion detection in satellite video," *J. Remote Sens. GIS*, vol. 6, no. 2, 2017.

[7] X. Zhang, J. Xiang, and Y. Zhang, "Space object detection in video satellite images using motion information," *Int. J. Aerosp. Eng.*, vol. 2017, 2017, Art. no. 1024529.

[8] T. Yang et al., "Small moving vehicle detection in a satellite video of an urban area," *Sensors*, vol. 16, no. 9, Sep. 2016, Art. no. 1528.

[9] B. Du, S. Cai, and C. Wu, "Object tracking in satellite videos based on a multiframe optical flow tracker," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3043–3055, Aug. 2019.

[10] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO: From YOLOv1 and beyond," Jun. 2023, *arXiv:2304.00501*

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[14] F. Bousetouane and B. Morris, "Off-the-shelf CNN features for fine-grained classification of vessels in a maritime environment," in *Advances in Visual Computing*. Berlin, Germany: Springer, 2015, pp. 379–388.

[15] H. Guo, X. Yang, N. Wang, B. Song, and X. Gao, "A rotational libra R-CNN method for ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5772–5781, Aug. 2020.

[16] S. Zhang, R. Wu, K. Xu, J. Wang, and W. Sun, "R-CNN-based ship detection from high resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 6, Jan. 2019, Art. no. 631.

[17] X. Sun, H. Jiang, T. Huo, and W. Yang, "A fast multi-target detection method based on improved YOLO," *MIPPR, Autom. Target Recognit. Navig.*, vol. 11429, pp. 190–197, Feb. 2020.

[18] D. Zhang, C. Wang, and Q. Fu, "OFCOS: An oriented anchor-free detector for ship detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Mar. 2023, Art. no. 6004005.

[19] J. Ma, Z. Zhou, B. Wang, H. Zong, and F. Wu, "Ship detection in optical satellite images via directional bounding boxes based on ship center and orientation prediction," *Remote Sens.*, vol. 11, no. 18, Jan. 2019, Art. no. 2173.

[20] Q. Tan, J. Ling, J. Hu, X. Qin, and J. Hu, "Vehicle detection in high resolution satellite remote sensing images based on deep learning," *IEEE Access*, vol. 8, pp. 153394–153402, 2020.

[21] R. Chen, X. Li, and S. Li, "A lightweight CNN model for refining moving vehicle detection from satellite videos," *IEEE Access*, vol. 8, pp. 221897–221917, 2020.

[22] Z. Yan, X. Song, H. Zhong, and F. Jiang, "Moving object detection for video satellite based on transfer learning deep convolutional neural networks," in *Proc. 10th Int. Conf. Pattern Recognit. Syst.*, 2019, pp. 106–111.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, May 2016, pp. 779–788.

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, 2017, pp. 6517–6525.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018, *arXiv:1804.02767*.

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," Apr. 2020, *arXiv:2004.10934*.

[27] C. Li et al., "YOLOv6: A single-stage object detection framework for industrial applications," Sep. 2022, *arXiv:2209.02976*.

[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, 2023, pp. 7464–7475.

[29] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, 2021, pp. 13024–13033..

[30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," Aug. 2021, *arXiv:2107.08430*.

[31] J. Redmon et al., "You only learn one representation: Unified network for multiple tasks," *J. Inf. Sci. Eng.*, vol. 39, no. 3, pp. 691–709, May 2023.

[32] X. Xu, Y. Jiang, W. Chen, Y. Huang, Y. Zhang, and X. Sun, "DAMO-YOLO: A report on real-time object detection design," Apr. 2023, *arXiv:2211.15444*.

[33] X. Long et al., "PP-YOLO: An effective and efficient implementation of object detector," Aug. 2020, *arXiv:2007.12099*.

[34] B. Li, X. Xie, X. Wei, and W. Tang, "Ship detection and classification from optical remote sensing images: A survey," *Chin. J. Aeronaut.*, vol. 34, no. 3, pp. 145–163, Mar. 2021.

[35] Z. Qu, F. Zhu, and C. Qi, "Remote sensing image target detection: Improvement of the YOLOv3 model with auxiliary networks," *Remote Sens.*, vol. 13, no. 19, Jan. 2021, Art. no. 3908.

[36] J. Hu, X. Zhi, T. Shi, W. Zhang, Y. Cui, and S. Zhao, "PAG-YOLO: A portable attention-guided YOLO network for small ship detection," *Remote Sens.*, vol. 13, no. 16, Jan. 2021, Art. no. 3059.

[37] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9626–9635.

[38] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 9905, pp. 21–37.

[39] Y. Li, C. Bian, and H. Chen, "Dynamic soft label assignment for arbitrary-oriented ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1160–1170, Dec. 2022.

[40] X. Zhang and Z. Zhang, "Ship detection based on improved YOLO algorithm," in *Proc. 3rd Int. Conf. Consum. Electron. Comput. Eng.*, 2023, pp. 98–101.

[41] W. Kong, S. Liu, M. Xu, M. Yasir, D. Wang, and W. Liu, "Lightweight algorithm for multi-scale ship detection based on high-resolution SAR images," *Int. J. Remote Sens.*, vol. 44, no. 4, pp. 1390–1415, Feb. 2023.

[42] M. Yasir et al., "Multi-scale ship target detection using SAR images based on improved Yolov5," *Front. Mar. Sci.*, vol. 9, 2023, Art. no. 1086140, Accessed on: Oct. 30, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmars.2022.1086140

[43] S. Li et al., "A multitask benchmark dataset for satellite video: Object detection, tracking, and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5611021.

[44] K. Kang et al., "T-CNN: Tubelets with convolutional neural networks for object detection from videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2896–2907, Oct. 2018.

[45] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 408–417.

[46] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4141–4150.

[47] H. Wu, Y. Chen, N. Wang, and Z.-X. Zhang, "Sequence level semantics aggregation for video object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9216–9224.

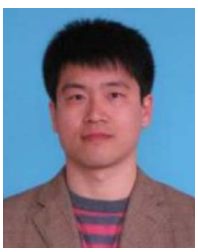
[48] S. Li et al., "Recent advances in intelligent processing of satellite video: Challenges, methods, and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6776–6798, Jul. 2023.

[49] J. Lei, Y. Dong, and H. Sui, "Tiny moving vehicle detection in satellite video with constraints of multiple prior information," *Int. J. Remote Sens.*, vol. 42, no. 11, pp. 4110–4125, Jun. 2021.

[50] H. Li, L. Chen, F. Li, and M. Huang, "Ship detection and tracking method for satellite video based on multiscale saliency and surrounding contrast analysis," *J. Appl. Remote Sens.*, vol. 13, no. 2, Jun. 2019, Art. no. 026511.

[51] H. Li and Y. Man, "Moving ship detection based on visual saliency for video satellite," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1248–1250.

[52] J. Zhang, X. Jia, and J. Hu, "Error bounded foreground and background modeling for moving object detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2659–2669, Apr. 2020.
[53] J. Zhang, X. Jia, J. Hu, and J. Chanussot, "Online structured sparsity-based moving-object detection from satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6420–6433, Sep. 2020.
[54] J. Zhang, X. Jia, J. Hu, and K. Tan, "Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5185–5198, Sep. 2022.
[55] Y. Li, L. Jiao, X. Tang, X. Zhang, W. Zhang, and L. Gao, "Weak moving object detection in optical remote sensing video with motion-drive fusion network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5476–5479.
[56] C. Xiao et al., "DSFNet: Dynamic and static fusion network for moving object detection in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2021, Art. no. 3510405.
[57] J. Feng et al., "Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 116–130, Jul. 2021.
[58] R. Pflugfelder, A. Weissenfeld, and J. Wagner, "Deep vehicle detection in satellite video," Apr. 2022. *arXiv:2204.06828*.
[59] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, Sep. 2022.
[60] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[61] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, New York, NY, USA, 2021, pp. 763–772.
[62] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," Apr. 2020, *arXiv:1910.03151*.
[63] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, La Jolla, San Diego, USA, 2014, vol. 27. Accessed: Feb. 6, 2024. [Online]. Available: https://www.webofscience.com/wos/alldb/full-record/WOS:000452647102103
[64] J. Dai et al., "Deformable convolutional networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, New York, NY, USA, 2017, pp. 764–773.
[65] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Comput. Vis. - ECCV* , Cham, Switzerland, 2018, vol. 11211, pp. 3–19.
[66] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, 2021, pp. 13708–13717.
[67] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, 2017, pp. 936–944.
[68] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.
[69] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," Jul. 2020, *arXiv:1911.09070*.
[70] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," *J. Inf. Sci. Eng.*, vol. 39, no. 3, pp. 975–995, May 2023.
[71] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, 2019, pp. 3974–3983.



Shanwei Liu received the Ph.D. degree in environmental science from the Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China. He has authored or coauthored more than ten high-level papers, and is the inventor or coinventor of five patents. His research interests include satellite altimetry, ocean remote sensing, hyperspectral image processing, and GIS application.

Xi Bu received the bachelor's degree in geomatics engineering in 2021 from the China University of Petroleum (East China), Qingdao, China, where he is currently working toward the master's degree in geomatics engineering with the College of Oceanography and Space Informatics.

His research focuses on ship detection.

Mingming Xu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from the China University of Petroleum, Qingdao, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum. Her research interests include hyperspectral image processing and wetland remote sensing.

Hui Sheng received the Ph.D. degree in geological resources and geological engineering from the China University of Petroleum, Qingdao, China, in 2010.

He is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China. He has authored or coauthored seven high-level papers and has undertaken a number of scientific research projects. His research interests include ocean remote sensing, hyperspectral image processing, and photogrammetry.

Zhe Zeng received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2010.

He is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China. His research interests include spatial data analysis and decision support technology, intelligent path planning methods, hyperspectral image processing, and analysis and application of remote sensing images.

Muhammad Yasir received the B.S. degree in geology from the University of Peshawar (UOP), Peshawar, Pakistan, in 2018, and the master's degree in geological engineering in 2021 from the China University of Petroleum, Qingdao, China, where he is currently working toward the Ph.D. degree in marine resources and information engineering with the College of Oceanography and Space Informatics.

He has several research publications in well-reputed international journals as a first and coauthor. He is a reviewer of the *Environment of Remote Sensing*, *International Journal of Applied Earth Observation and Geoinformation*, *ISPRS*, *MPDI* different journals, and IEEE Journal. His research interests include computer vision and remote sensing image object detection, remote sensing image processing techniques, and systematic literature review.