COCO-Net: A Dual-Supervised Network With Unified ROI-Loss for Low-Resolution Ship Detection From Optical Satellite Image Sequences

Qizhi Xu, *Member, IEEE*, Yuan Li, Mingjin Zhang, *Member, IEEE*, and Wei Li, *Member, IEEE*

***Abstract*— Low-resolution ship detection from optical satellite image sequences is critical in high-orbit remote sensing satellite applications. However, it is still a difficult problem due to the following challenges: 1) the size of the ship is tiny in the low-resolution image; 2) the ship target is dim and the contrast with the background is low; and 3) the interference of cloud and fog covering is complex and changeable. For these reasons, the targets are easily lost during the detection. In fact, the Clearer the Objects against to the background, the more Confident the Observers can detect it. In light of these considerations, we propose a COCO-Net to detect the small dynamic objects on low-resolution images in this article. First, the multiframe images are associated by introducing motion information as an effective compensation for small object features. Second, an integrated dual-supervised network that processes single-level tasks hierarchically is presented to adaptively enhance the input data quality of object detection without being limited by diverse scene disturbances. Third, a unified region of interest (ROI)-loss scheme that modulates the loss function of the first component by introducing ROI-masks from the second component is utilized to make the first component also work for object detection. In addition, we construct a new dataset for the small dynamic object detection based on the GaoFen-4 satellite imagery. Comprehensive experiments on a self-assembled dataset from the GaoFen-4 satellite show the superior performance of the proposed method compared to state-of-the-art object detectors.**

***Index Terms*— Dual-supervised network, low-resolution imagery, optical remote sensing (RS) images, ship detection.**

I. Introduction

With the continuous development of modern remote sensing (RS) technology, many RS images are regularly produced, providing data for various research fields [1], [2]. High-resolution images from low-orbit RS satellites have attracted much attention because of their clear imaging properties [3]. However, the image's width is relatively small, and it takes a long time for the satellite to revisit the fixed area. High-orbit RS satellites can exactly compensate for this drawback due to their high temporal resolution. Accordingly, interpreting low-resolution images with larger widths is of great research value.

In the RS community, marine object detection is a large and active research area with many applications, including behavioral analysis, military surveillance, and border protection. Nevertheless, due to the ultralow spatial resolution images from RS satellites, the ship object contains only a few pixels of information, as shown in Fig. 1(a). Hence, no effective shape and texture features can be used as the discriminant basis. However, because of the high temporal resolution of the satellite, continuous image sequences usually can be acquired. It brings us inspiration to introduce temporal information to make up for the lack of spatial information. In addition, the complex cloud and fog may cover the targets and further weaken the object features. It is known that the greater the difference between the object and the background, the more prominent the object will be. And then, the target will be easier to detect. Therefore, how to mine more target information and enhance the object feature from limited data has come into the focus of research.

Previous researchers have made efforts in object enhancement through multiple frame image fusion. A common approach is to use background subtraction or frame difference methods to find objects in consecutive frames [4], [5], [6]. However, these methods suffer from their own costly drawbacks. Frame difference methods rely heavily on frame registration. For the images with unpredictable clouds, they may introduce much extra noise. Further, it usually requires the time difference between consecutive frames to be small. Thus, they are not suitable for RS images with changing cloudy backgrounds and longer frame difference time. Moreover, other works have attempted to extract multiframe information from video data by tracking methods, or optical flow [7], [8], [9]. These methods utilize the temporal context to supplement the lack of information in a single-frame image effectively, but they require more than five frames of image data. It is a luxury for wide-swath RS image processing with high timeliness requirements. Because the satellite has a fixed shooting time interval, the more frames required, the longer it takes. Therefore, the requirement of a simple and suitable method to integrate multiframe information is put forward.

Additionally, we discover through rigorous experiments that the object feature strength in the original images is closely related to the precision of the outcomes of the subsequent detection [10]. Thus, it is necessary to enhance input image

Manuscript received 28 June 2022; revised 11 August 2022; accepted 18 August 2022. Date of publication 25 August 2022; date of current version 12 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61972021 and Grant 61672076. *(Corresponding author: Yuan Li.)*

Qizhi Xu and Yuan Li are with the School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: qizhi@bit.edu.cn; liyuansme@bit.edu.cn).

Mingjin Zhang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: mjinzhang@xidian.edu.cn).

Wei Li is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: liwei089@ieee.org).

Digital Object Identifier 10.1109/TGRS.2022.3201530

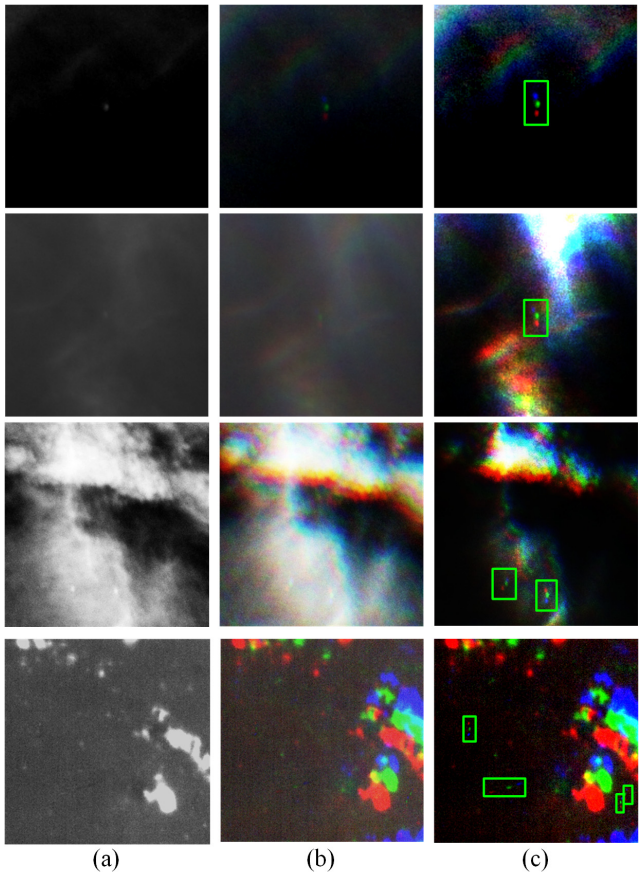

Fig. 1. Schematic of continuous frame fusion images and their enhancement results in different scenes. Columns (a) Illustrate original single-band image blocks. Columns (b) Show fused three-channel image blocks. Columns (c) Indicate the results after adaptive object enhancement processing.

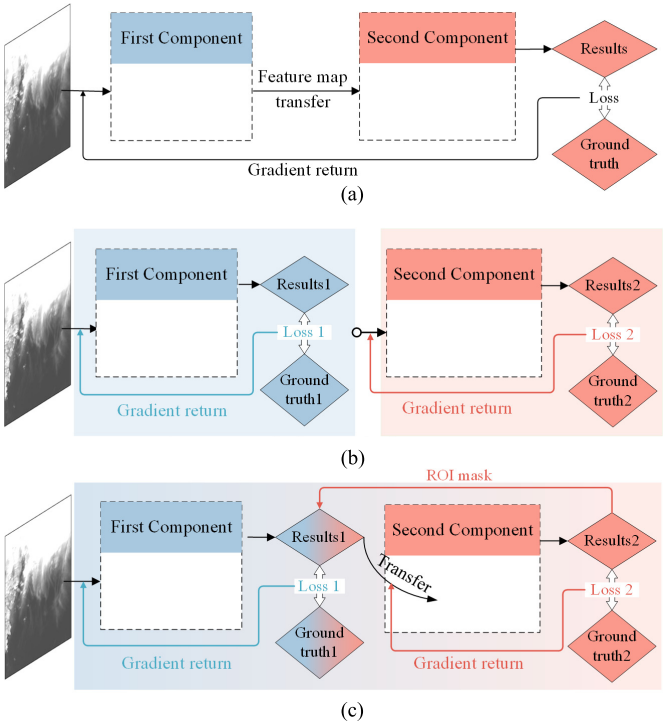

Fig. 2. Comparison between different approaches for multistage object detection. Pipeline (a) corresponds to classic cascade structure, in which the different components are connected by feature maps. Pipeline (b) represents the dual-supervised structures with separated components. In contrast, our proposed pipeline (c) uses dual-supervised joint mode for different component.

quality before performing detection tasks. In recent years, deep learning methods have shown impressive performance in this field [11], [12], [13]. Generally, they are combined with the object detection task as a preprocessing part, and the common structures are demonstrated in Fig. 2(a) and (b). Specifically, on the one hand, one of these structures connects the different components through feature maps and utilizes one-level supervision to regulate the whole network [14], [15]. They are highly integrated but cannot achieve the purpose of weakening the difficulty of single-level tasks by a divide-and-conquer strategy. On the other hand, the common multisupervised structure divides the tasks hierarchically [16], while they are separate from each other. They have their own loss function and ground truth to supervise the network separately, and there exists no information interaction during the training process. This leads to the fact that the previous components will only learn according to the ground truth defined on the basis of human vision rather than machine perception. Our ultimate goal is to get better object detection performance rather than obtaining human-defined sharper images. Consequently, an effective intermodulation mechanism with a novel loss is necessary to integrate the different components.

Inspired by the above analysis, we believe that the core idea of improving the detection rate of small dim objects in low-resolution images is: **C**learer **O**bjects, the more **C**onfident **O**bservers. Specifically, we devise a novel dual-supervised network with unified region of interest (ROI)-loss, called COCO-Net, for detecting ships in low-resolution optical satellite image sequences. First, we aggregate three consecutive frames to compensate for small object features by introducing the motion information, thus transforming the original single-band image into a three-channel image. The composed images are then fed into a novel dual-supervised network for feature extraction (FE), where the dual-supervised network consists of an object enhancement component (OEC) and an object detection component (ODC). The OEC consist of multiple attention modules stacked with a detail feature compensation (DFC) module added to integrate the features at different levels. The ODC is optimized for small target detection based on the you only look once (YOLO) architecture [17]. We apply an energy filter kernel in the multiscale feature cross fusion (MFCF) module of the ODC, which is obtained from high-level feature maps and used to filter out the noise of mid-level features. Thus, the multiscale features can be efficiently aggregated without introducing noise. Besides, we design a uniform ROI-loss scheme that constrains the OEC to focus more on the target region according to the ROI information obtained from the ODC. In addition, we construct a new dataset based on the GaoFen-4 satellite imagery, consisting of 3030 image patches in various scenes with accurate annotations. Experimental results on this dataset demonstrate that the proposed COCO-Net outperforms state-of-the-art (SOTA) methods in terms of different evaluation indicators.

The contributions of this study can be summarized as follows.

1) We propose an inspired idea to address the challenges in small object detection, i.e., incorporating suitable consecutive frames to compensate for small object features.
2) We develop a novel dual-supervised framework with two components named COCO-Net to hierarchically processes single-level tasks, which can adaptively enhance the input data quality of object detection without the limitation of diverse noise.
3) We design a unified ROI-Loss scheme to constrain the first component to simultaneously serve the final object detection task, which is the key to the composition of the integrated network.
4) Experiments on our newly constructed database illustrate the superior performance of the proposed method. Low-resolution ship detection database (LSD) dataset taken by the GaoFen-4 satellite consists of 3030 image patches in various scenes with accurate annotations.

The rest of this article is organized as follows. Section II investigates the related works, and Section III describes the details of the proposed COCO-Net for low-resolution ship detection. Experimental results and detailed comparisons are shown in Section IV to verify the superiority of our method. Finally, conclusions are drawn in Section V.

II. Related Works

A. Small Objects Detection Methods

Small object detection is an indispensable and challenging problem in image understanding and the computer vision field. In recent years, the compelling success of deep learning techniques has pushed small object detection forward to a research highlight. In general, there are two different definitions of small objects. One refers to objects with smaller physical sizes in the real world, and the other can be found in Microsoft COCO (MS-COCO) [18]. That is, objects occupying areas less than and equal to 32×32 pixels are regarded as "small objects". Since RS images always have lower resolution than natural images, ships with large physical sizes are also small objects occupying only a few pixels in RS images. Three difficulties are often encountered in constructing an accurate small target detector: the lack of appearance information separated from the background, the high requirements for localization accuracy, and the limited empirical knowledge [19]. Based on this situation, many researchers have made efforts in different aspects.

On the one hand, multiscale feature fusion is regarded as a crucial issue in improving the performance of small object detection. In [20], a scale-aware network is proposed to resize all objects on a similar scale and then train a single scale detector. Singh *et al.* [25] designed a new framework called scale normalization for image pyramids (SNIP), which trained multiple scale-dependent detectors. Each of them was in charge of a specific scale object. This is a roundabout strategy that avoids the difficulty of training one model that can accurately detect objects at all scales. Pang *et al.* [21] presented the aggregate interaction modules to integrate the features from adjacent levels. It can effectively cope with the great challenge of the variable scale of salient objects. In 2022, a new enhanced multiscale feature fusion method is developed [22]. The multiscaled atrous convolution operators are employed to make full use of context information. Liu *et al.* [23] also introduced a novel stereoscopically attentive multiscale (SAM) module to a lightweight network, which can adaptively fuse the features of various scales.

On the other hand, data augmentation and training strategy are also beneficial for small object detection. Kisantal *et al.* [24] found that one of the factors behind the poor detection performance for small objects is the lack of representation of small objects in a training set. First, they demonstrate that the detection rate can be effectively improved by oversampling images containing small objects. Second, they designed a data augmentation approach by copy-pasting small objects through the segmented mask. Besides, literature [25] proposed a novel model called Scale Normalization for Image Pyramids with Efficient Resampling (SNIPER). It only processed context regions around ground truth instances at the appropriate scale. Later, Kim *et al.* [26] designed a scale-aware network (SAN). It first maps the convolutional features obtained from the different scales onto a scale-invariant subspace. Then, SAN and detection network are trained simultaneously. In addition, Prakash and Karam [27] utilized a generative adversarial network (GAN) to generate features that provide robustness for object detection on reduced-quality images. Although these methods can improve the small object detection performance, they are still unsuitable for dim tiny marine object detection under unpredictable clouds background.

B. Image or Object Enhancement Methods

Vision-based methods including object detection, activity recognition, *etc.*, require visible images for superior performance. In early research, histogram equalization [28], gamma transform [29], *etc.* are the simple and straightforward methods. However, they sometimes prompt over-enhancement, and other limitations. In [30], Dynamic Histogram Equalization (HE)-based approaches are utilized to overcome the above shortcomings and enhance the contrast. Another prominent algorithm [31] enhance the image by introducing fuzzy contextual information about the images. In [32], the author presented an improved Retinex model to enhance the low-light images and reduce the intensive noise interference. In 2009, He *et al.* [33] designed a dark channel prior theory for image defogging, which has been widely applied.

Additionally, deep learning-based methods have also performed well in this field [34], [35], [36]. In 2020, a semi-supervised learning approach, deep recursive band network (DRBN) [37], for low-light image enhancement was developed. It was well designed to extract a series of band representations from coarse-to-fine and generated enhanced results with well-reconstructed details with the help of this two-stage design. In [38], an accurate and efficient single-shot object detector (FAENet) with feature aggregation and enhancement was proposed. They integrated a pair of novel feature aggregation modules and two feature enhancement blocks

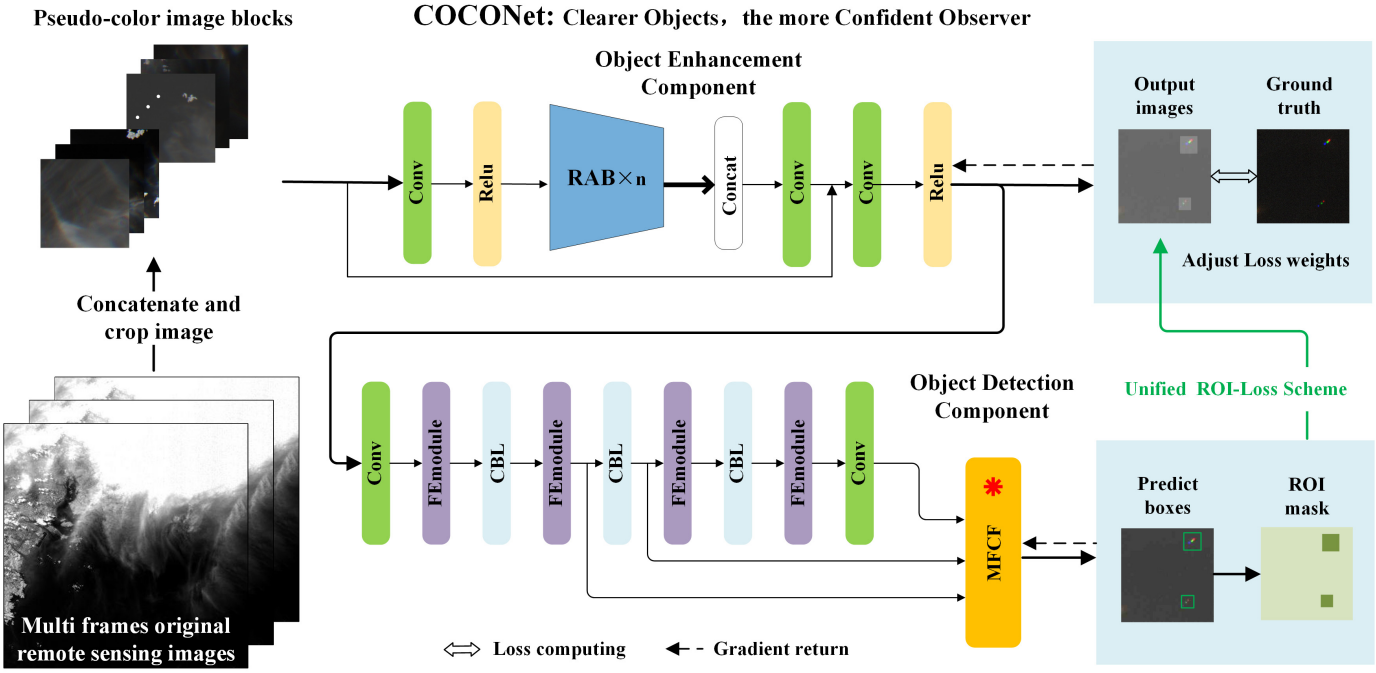

Fig. 3. Overview of the proposed ship detection method COCO-Net. The method consists of two components: OEC for object enhancement and ODC for object detection.

into the Single Shot MultiBox Detector (SSD) network to improve the detection performance. In order to improve the visual effects of weak vehicles in RS images, Gao *et al.* [39] proposed a detection-guided CycleGAN to enhance the weak targets for accurate detection. Besides, super-resolution methods are often chosen for reconstructing images [40], [41], [42]. However, most of these methods enhance objects at the extracted feature level, while it is also important to improve the input data quality for low-resolution object detection.

C. Approaches for Reducing Background Interference

In many image interpretation tasks, such as object detection, action recognition, *etc.*, background interference is one of the main reasons that affect the performance of the model. In [43], Shen *et al.* designed a residual learning structure incorporated with weakly supervised detection, which decomposes background noise and models clean data. In 2020, an improved RBox-based object detection model is proposed [44]. It can effectively reduce the interference of background pixels by locating the objects more finely. To overcome the challenge of detecting small infrared target under complex background, study [45] defined an enhanced local contrast measure method to enhance small targets and suppress complex background. In [46], Wang *et al.* presented a debackground detail convolutional network. Specifically, they enable the decomposer to produce a detail layer by subtracting background interference from the crowd images, which optimizes the learning process.

In addition to the above methods, multistage processing is also an important idea. Yang *et al.* [47] designed a preidentification mechanism and a cascaded detector for tiny faces. The prerecognition mechanism first preidentified face region candidates as regions of interest and then used them as inputs to subsequent networks, leading to reducing background and other extraneous information. In [48], a new pixel to global matching network (PG-Net) framework is proposed, consisting of a FE subnet and an object localization subnet. The PG-corr module integrated into the object localization subnetwork can effectively suppress background interference by narrowing the matching area. In [49], the author introduced a cascade region proposal network with soft-decision nonmaximal suppression, improving the performance under complex background. However, the above methods mainly address the interference of various ground objects rather than the cloud and fog occlusion. These two types of backgrounds have different influences on target detection; thus, they require further exploration.

III. Methodology

The proposed network model COCO-Net is composed of an OEC and an ODC. The overall framework is shown in Fig. 3. Since the object size in ultralow-resolution images is only a few pixels, it is necessary to perform effective feature compensation for dynamic objects by integrating sequence images. Following cropping, the image blocks are put into the OEC to automatically improve the saliency of objects in accordance with local scenes. Then the improved images with clearer objects are trained by the ODC, which is optimized based on YOLOv5 backbone. Finally, it is worth noting that the prediction boxes from the ODC are fed back to the OEC to adjust the loss weights. More detailed descriptions are given in the following subsections.

A. Multiframe Images Association Approach

The original RS images from the satellite are single-band 16-bit images with low spatial resolution, and the object only

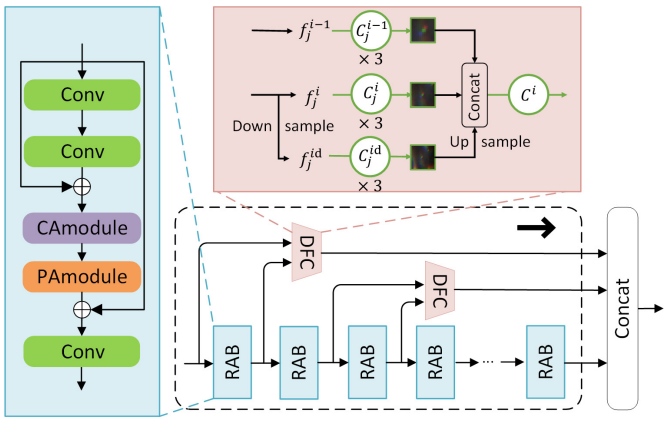

Fig. 4. RAB module and the concatenate method.

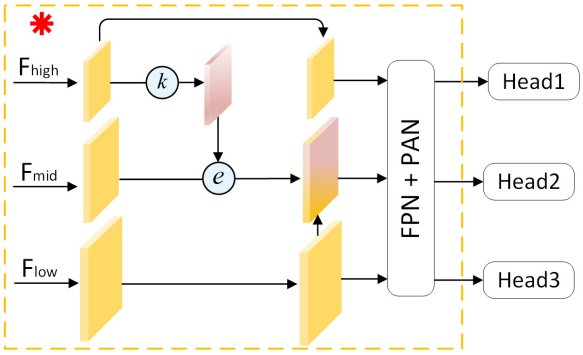

Fig. 5. Structure of MFCF module. Function k denotes the process of constructing the energy filter kernel, and function e denotes the process of filtering.

contains a few pixels. Accordingly, both the texture and color information of the target are extremely insufficient. In this case, it is necessary to compensate the input image with effective object features to obtain a better detection effect. A simple and practical approach is to correlate consecutive images, which can enhance the target features by introducing motion information. As shown in Fig. 1(a) and (b), we take the three consecutive frames of images as the three channels of the resulting image, respectively

$$I_r = I_1, \quad I_g = I_2, \quad I_b = I_3 \tag{1}$$

where I_r, I_g, and I_b are the R, G, and B band image in the final pseudo-color image, respectively. I_i is the ith frame of the image sequence. The ship target is visually a point in the original image, while the target becomes a sequence of colored dots after the association of the image. Since the speed of the ship is within a certain range, the distance between the points is also within a certain range. Therefore, this scheme can effectively enhances the ship target features, which can improve the detection accuracy rate.

How many frames to correlate is a critical issue that deserves careful consideration. On the one hand, integrating more frames of image data can increase target information, but it can also introduce more complex background information and more interference. Moving clouds in other frames may obscure the target in the current frame and weaken the target features. On the other hand, it usually takes a certain amount of time for a satellite to shoot an area. The more frames we use, the longer the data generation time. However, high timeliness is necessary for both military and civilian applications. Thus, it is not the case that the more frames of the fused image, the better the detection. Therefore, we use frame difference method to correlate more than three frames and verify the influence of the number of frames on the detection effect through experiments. For example, the sequence number of multiple frames is set to 1-k. The R and G band of fused three-channel images are "1" and "2" frame images, respectively. The B band is the resulting image of the 3-k frames calculated by the frame difference method, which can be described as

$$I_r = I_1 \tag{2}$$

$$I_g = I_2 \tag{3}$$

$$I_b = I_3 + |I_3 - I_4| + |I_4 - I_5| + \cdots + |I_{k-1} - I_k| \tag{4}$$

where $|\cdot|$ denotes the calculation of absolute value. In this study, we choose 1, 3, 5, and 7 frames image sequences to verify the detection effect. Experimental results show that using three consecutive frames has a better detection effect. So we final utilize three frame images to composite sample image. Take part in Section IV for specific experimental contents.

B. Framework of COCO-Net

The whole framework of COCO-Net consists of two components, namely OEC and ODC. As shown in Fig. 3, the OEC is utilized to sharpen the images and enhance the contrast between the object and background. It is mainly constructed by several residual attention blocks (RAB), which consist of channel attention (CA) modules, pixel attention (PA) modules, and multiple DFC modules for DFC. The ODC is used to capture object features and detect them. It is modified from YOLOv5 for small objects and consists of an improved MFCF module. The two components are tightly connected, and the output feature map of OEC is fed into the ODC directly, which can be expressed as

$$O = \Psi_{\mathrm{ODC}}(\Psi_{\mathrm{OEC}}(I_{\mathrm{mul}})) \tag{5}$$

where I_{mul} and O represent the multiframe image and the final detection result. Ψ_{ODC} and Ψ_{OEC} denote the calculation process of the ODC and OEC components, respectively. Each component has its own supervised ground truth, but is trained uniformly.

The OEC first goes through a 3 × 3 convolution layer. Then, it is constructed by stacking six RAB modules followed by another convolution layer to extract useful features. Next, it adds skip-connection to share the low-level details of the image to enhance the feature. After that, the obtained feature maps are fed into another convolution layer and a relu activation function again. The above process can be expressed as follows:

$$\Psi_{\mathrm{OEC}}(I_{\mathrm{mul}}) = C(C(F_{\mathrm{RAB}}(C(I_{\mathrm{mul}}))) + I_{\mathrm{mul}}) \tag{6}$$

where C denotes the convolution and F_{RAB} represents the operator constructed by the RAB module. At this point, the OEC calculation is completed. The output images are utilized to calculate loss along with ground truth and are treated as input of the ODC. As shown in Fig. 4, each RAB is composed of three convolution layers, a CA module and a PA module. The specific structures of CA and PA modules can be found in [50]. The output of each RAB is fed into the DFC module. The reason for using this module is mainly to minimize information loss of small targets in deep layers of the network. This loss is irreversible and will accumulate with the deepening of the layer, but it is very important for small target detection. The structure of DFC is shown in the red box in Fig. 4. The f_j^{i-1} and f_j^i denote $i-1$th and ith level input of the DFC module. And the f_j^{id} represent the downsampled feature of f_j^i. Then we utilized the convolution 3×3 (C^i) to extract features of the target in multiple levels. Finally, the different features are fused. We can express it visually with the following formula:

$$M_{DFC} = C^i\left(\prod_{j=1}^{3} C_j^{i-1}\left(f_j^{i-1}\right) \oplus \prod_{j=1}^{3} C_j^i\left(f_j^i\right) \oplus \prod_{j=1}^{3} C_j^{\mathrm{id}}\left(f_j^{\mathrm{id}}\right)\right) \quad (7)$$

where M_{DFC} denotes the output of each DFC module and the $\oplus$ represents the concatenate operator. $(\prod_{j=1}^{n} C_j)(f_j)$ is defined as

$$\left(\prod_{j=1}^{n} C_j\right)\left(f_j\right) = C_n\left(C_{n-1}, \ldots, C_2\left(C_1\left(f_j\right)\right)\right) \quad (8)$$

where $(\cdot)$ denotes the input data of each calculation module and f_j is represents the different level input of DFC module. Based on the DFC module, high-resolution features with detailed information can be used to compensate for the semantics for detection, and the detailed features of the small target can be preserved.

Furthermore, the ODC is constructed based on a one-stage detection framework, such as YOLO [17]. The backbone contains four FE modules, each of which includes five residual units. Meanwhile, each FE module is followed by a convolution layer, batch normlization layer and leaky relu function layer (CBL) block consisting of a convolutional layer, a batch normalization layer, and a relu activation function layer. After that, the output feature maps of the last three FE modules are fed into the MFCF module. This process can be formulated as

$$\Psi_{\mathrm{ODC}}\left(O'\right) = \underset{k=2}{\overset{4}{Fu}}\left(\prod_{i=1}^{k}\left(M_{\mathrm{CBL}}^i\left(M_{\mathrm{FE}}^i\left(O'\right)\right)\right)\right) \quad (9)$$

$$M_{\mathrm{CBL}}\left(I_f\right) = \mathrm{Relu}\left(\mathrm{BN}\left(C\left(I_f\right)\right)\right) \quad (10)$$

$$M_{\mathrm{FE}}\left(I_f\right) = \prod_{i=1}^{5} R_e^i\left(I_f\right) \quad (11)$$

where O' represents the output of OEC component and I_f denotes the input feature map for different module. M_{CBL} and M_{FE} denote the calculation process of FE module and CBL block. $(\prod_{i=1}^{n} R_e^i)(\cdot)$ is defined similar with $(\prod_{j=1}^{n} C_j)(\cdot)$, and R_e denotes the residual blocks. F_u represents the MFCF operator described later. Generally, target extraction is performed to effectively capture the features of the target region, while it may convolution out the object-like noise and lead to a high false alarm rate. To suppress noise and improve the signal-to-noise ratio of the target, we introduce an energy filter kernel function in the MFCF module.

Specifically, we first construct an energy kernel that relies on the energy distribution of the high-level feature. This step is mainly achieved by pooling and convolution operations on high-level features. Then, it is used to filter out background noise in mid-level features, which is achieved by convolution. Finally, the obtained pure target feature is added with the low-level features to enhance the contrast of the target area. After that, the signal-to-noise ratio of image can be effectively improved and the false alarms are also be suppressed. The specific process is formulated as

$$K_e = \phi\left(P_m\left(F_{\mathrm{high}}\right) * P_a\left(F_{\mathrm{high}}\right)\right) \quad (12)$$

$$F_e = K_e * F_{\mathrm{mid}} + F_{\mathrm{low}} \quad (13)$$

where P_m and P_a are max pooling layer and average pooling layer, respectively. $*$ denotes the convolution operator. $\phi(\cdot)$ denotes the sigmoid layer, and K_e is the energy filter kernel. $K_e * F_{\mathrm{mid}}$ is the pure feature map after energy filtering and F_e denotes the output feature map with high signal-to-noise ratio. Then the multiscale maps are fused through the feature pyramid network (FPN) and path aggregation network (PAN) module, which is popular in the YOLO detection framework. Finally, we utilize three multiscale detection heads to predict targets.

C. Unified ROI-Loss Scheme

A unified ROI-loss of pixel and bounding-box regression is used to optimize the COCO-Net jointly. The first supervised function for OEC is applied to mean absolute error (MAE), called L1loss, which calculates the difference of all corresponding pixels in the image. Although this function is simple, many studies have verified that L1loss-based image restoration tasks achieve better performance than L2loss in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics [51]. Therefore, L1loss is used by default in this work, which is expressed as

$$L_{\mathrm{OEC}} = \frac{1}{n}\sum_{i=1}^{n}\left|y_{\mathrm{gt}}^i - \mathrm{OEC}\left(\mathrm{y}_{\mathrm{pre}}^{\mathrm{i}}\right)\right| \quad (14)$$

where y_{gt}^i stands for the ith pixel in ground truth and y_{pre}^i denotes the ith pixel in predict image. n is the pixel number of the image.

It is worth noting that if a feature enhancement operator is performed on the entire image, it is likely to introduce unnecessary noise information. Hence, setting up a supervision mechanism for OEC so that the network pays more attention to the target area is very important for this dual-supervised network. Based on this consideration, we feed the detection boxes of ODC to OEC as ROI masks. Then assign different weights to the pixels according to whether they are in the target area when calculating the L1loss. When a pixel is within the ROI mask, its weight is larger; otherwise, it is smaller. The L_{OEC} loss is modified as

$$L_{\mathrm{uin}} = \frac{1}{n}\sum\left(\alpha\left|y_{\mathrm{gt}}^i - \mathrm{OEC}\left(\mathrm{y}_{\mathrm{pre}}^i\right)\right| + \beta\left|y_{\mathrm{gt}}^j - \mathrm{OEC}\left(\mathrm{y}_{\mathrm{pre}}^j\right)\right|\right) \quad (15)$$

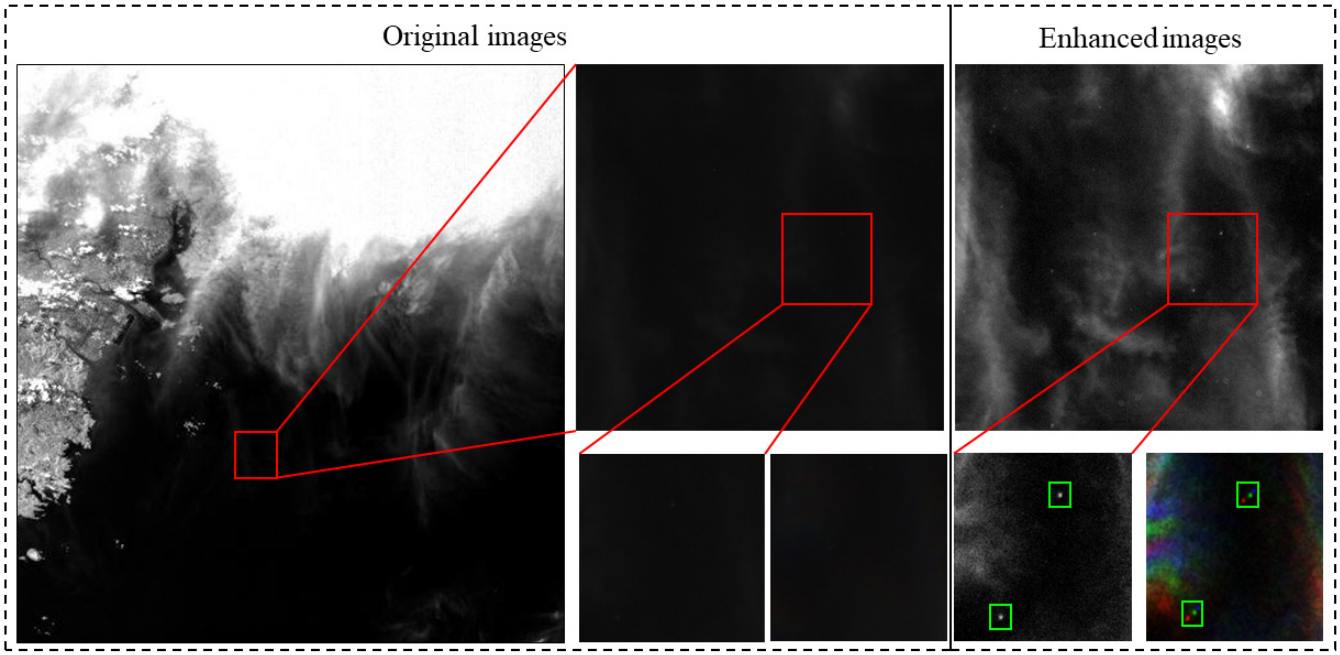

Fig. 6. Visualization of original RS images and the enhanced results for the LSD database. The pseudo-color image is obtained by superimposing multi frames of image block.

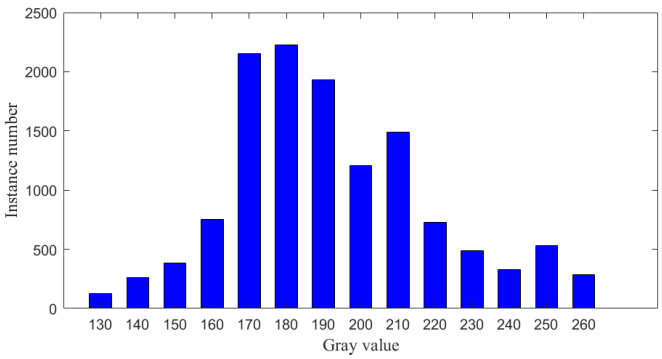

Fig. 7. Grayscale distribution of instance objects on the LSD database (16 bits).

TABLE I
TRAINING PARAMETERS

Paramers	OEC	ODC
Resized Image Size	416 pixels	416 pixels
Initial Learining rate	0.00001	0.001
Weight Decay	-	0.0005
Optimizer	Adam	Adam
Loss function	L1Loss	YOLOLoss
Max Iteration	500	500

where α is set 0.8 and β is set 0.2 in this work. y^i denotes that the pixel is inside the ROI-mask area, and y^j indicates that the pixel is outside the ROI-mask area. Afterward, the constructed ensemble COCO-Net is trained using this unified loss scheme. The experimental results verify the effectiveness of the scheme.

For the object detection task, the focal loss proposed recently performed well [52]. It adds a modulation factor to the cross-entropy loss to reduce the relative loss of easy samples and focus on hard samples. The focal loss can be expressed as

$$L_f = \begin{cases} -\alpha\left(1-y'\right)^{\gamma}\log\left(y'\right), & y=1 \\ -(1-\alpha)\left(y'\right)^{\gamma}\log\left(1-y'\right), & y=0 \end{cases} \tag{16}$$

where the factor $(1-y')^{\gamma}$ to the standard cross-entropy criterion. The balance factor α is added to balance the uneven ratio of positive and negative samples. Moreover, the total loss needs to be weighted with the classification loss (L_f), intersection over union loss (L_{IoU}) and confidence loss (L_{conf}). Among them, IoU loss is the Complete-IoU (CIoU) loss, including the aspect ratio factor and confidence loss is binary cross-entropy loss (BCEloss). More details can be seen in literature [53]. Consequently, the total loss L_{ODC} is defined as follows, and λ_1, λ_2 and λ_3 are the set parameters used to balance the weights between these losses

$$L_{\text{ODC}} = \lambda_1 \cdot L_{\text{conf}} + \lambda_2 \cdot L_{\text{IoU}} + \lambda_3 \cdot L_f. \tag{17}$$

According to the experiment experience of former researchers, the weight ratio of these three losses should be the same. It is reasonable because these losses are all equally important for accurate detection. Consequently, we set all three coefficients λ_1, λ_2, and λ_3 are 1.

IV. EXPERIMENTS

In this section, the efficacy of the proposed COCO-Net is verified, and we compare it with SOTA methods on our newly constructed LSD dataset from the GaoFen-4 satellite.

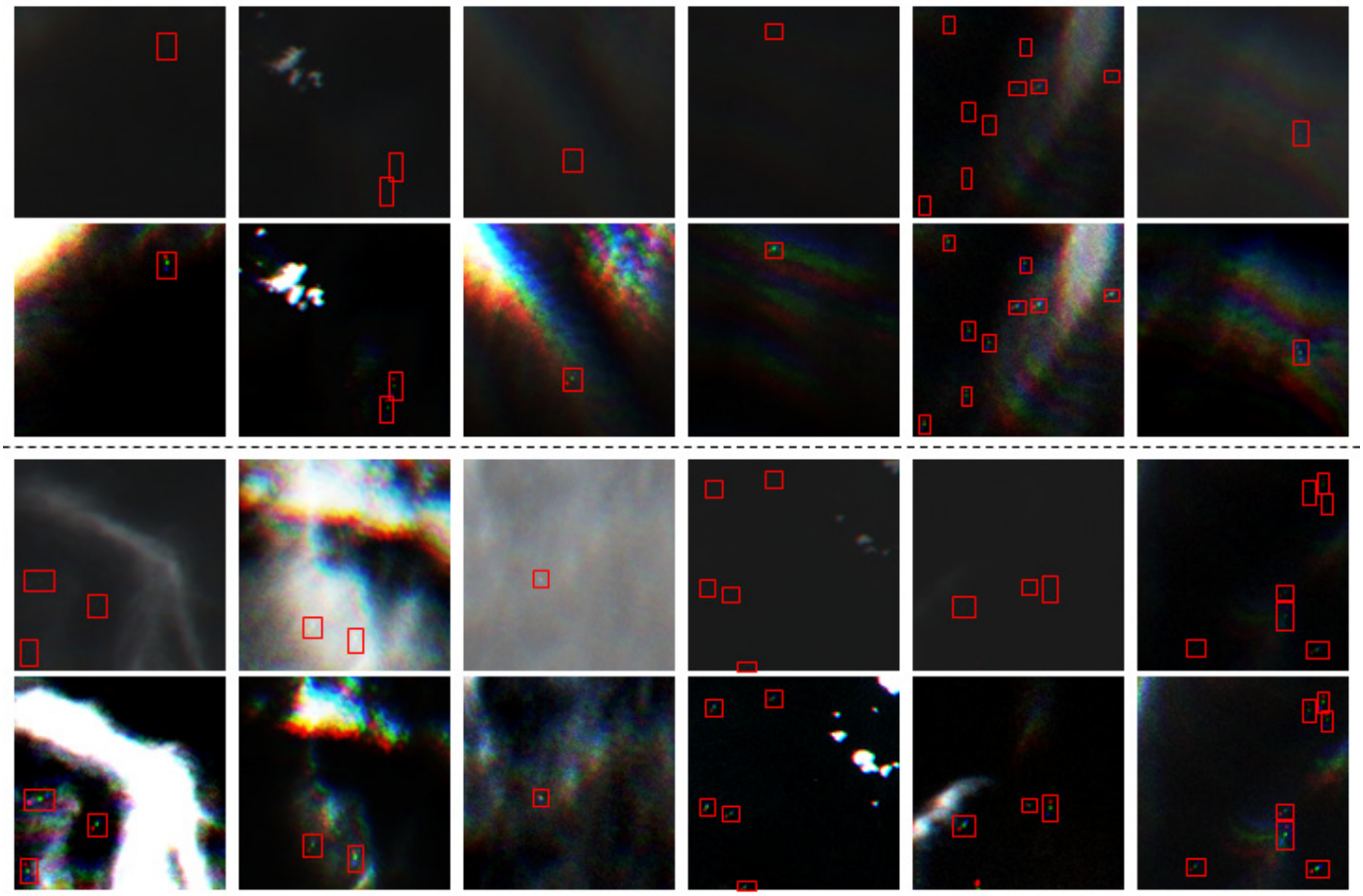

Fig. 8. Example images and their bounding box annotations on LSD dataset. The images in the first and third rows are the original RS image patches, and the images in the second and fourth rows are the corresponding target enhanced image patches.

This article mainly aims to detect moving ships in high-orbit satellites' ultralow resolution images. First, experimental conditions are described, including datasets, evaluation metrics, implementation details, and baseline. Second, detailed ablation experiments are performed, and the impact of each proposed scheme on the final detection performance is discussed. Finally, experiments comparing the proposed approach with SOTA methods are conducted on the LSD dataset, and the results are analyzed and visualized.

A. Experimental Conditions

1) LSD Database: To evaluate the performance of the proposed method, we conduct experiments on continuous frame image data of the GaoFen-4 satellite and develop a low-resolution ship detection database (namely, LSD Database). The original RS images can be seen in Fig. 6. The spatial resolution is 50 m, and the spectral range is 0.45–0.52 μm. The size of original RS images is 10240 $\times$ 10240 pixels. Furthermore, due to the low image resolution, our targets are mainly various types of aircraft carriers and large warships. Their size range is about 120–300m, which is about 2–6 pixels in a single frame image and falls into the category of dim tiny targets. At the same time, the real ship objects in RS images are annotated by professional ship target interpreters. They utilize more than one valid information to discriminate the ship object and the result is quite credible.

To correlate multiframe features, we construct a synthetic image with three sequence frames as R, G, and B channels, respectively. That is, the model input contains an image sequence of three adjacent frames. And the subsequent operations are all implemented based on these synthetic images. In addition, due to the large size of the original RS image, we cut the images to 256 $\times$ 256 pixels. At the same time, to avoid the breakage of the target sequences, we cut the images with a 15% overlap. Moreover, since the GaoFen-4 data obtained is limited, we utilize the simulation method for object augmentation [24] to simulate more samples. We first separate the sea background, various ships,, and clouds image blocks, then randomly select one of each category and fuse them through Poisson calculation. Thus, the final dataset contains a total of 3030 images containing a total of 3497 instances. Among them, there are 2490 real samples, including 460 positive sample images and 2030 negative sample images, and a total of 1877 ship instances. There are 540 simulation sample images, including 1620 simulated ship instances. The ratio of positive to negative sample images is 1:2. For the LSD database, 1939, 485, and 606 images are used for training, validation, and testing, respectively.

Since the unique properties of our dataset, detecting objects in our dataset have several distinctive challenges. First, low spatial resolution of objects. We can see from Fig. 6 that all instances have a small size ($<$ 15 pixels), and the multiframe-enhanced instances are less than 50 pixels in size. This poses a great challenge to detection methods. Second, the large grayscale difference of objects. In original 16-bit RS images, the gray value of instances varies from 130 to 260. Some dim objects in the original image cannot even be seen. Third, the number of instances in different grayscale intervals is not balanced, which can be seen as Fig. 7. The instances with grayscale values less than 160 are only about 5% of all instances. This brings great difficulty to accurate target detection.

We also constructed the ground truth of target enhancement image in addition to the ground truth of target box annotation, because the COCO-Net is a dual-supervised framework. Generally, some objects will be under-salient or over-exposed if the same algorithm is applied to process all image patches. Hence, we individually adjust the contrast and brightness of each image patch according to its background, making the target more salient. Specifically, We first divide the different image patches into several groups according to the scenes. Then, we set different parameters of the enhancement algorithm to adjust the brightness and contrast of the images in each group to make the target more salient. After that, we adjust the image patches with bad enhancement effect one by one, to ensure that all the samples get the best image enhancement effect. Consequently, the network can be trained using the ground truth, and perform adaptive target enhancement processing according to different scenes.

2) Implementation Details: The proposed method is implemented using the PyTorch deep learning framework and is trained on a workstation with an NVIDIA GeForce RTX 2080 graphics processing unit (GPU) with 8 GB memory. To compare the proposed COCO-Net with other methods more fairly, the training hyperparameters are set to be the same as the comparison methods. The specific experimental parameter settings are shown in Table I. It's worth noting that we followed YOLO's implementation trick, and fix the input image size to 416 $\times$ 416 pixels by BiCubic interpolation, which can achieve better detection results. The number of training epochs is 500. The IoU threshold is set to 0.2 to obtain better results because the object size was small. The confidence threshold is set to 0.3, and the nonmaximum suppression (NMS) threshold is 0.5. The initial learning rate of OEC and ODC are 1×10^{-5} and 0.001, and the final learning rate are 1×10^{-9} and 1×10^{-6}, respectively. The two learning rates are updated using a cosine update strategy. Assume the total number of batches is T, l is the initial learning rate, then at batch t, the learning rate l_t is computed as

$$l_t = \frac{l}{2}\left(1 + \cos\left(\frac{t\pi}{T}\right)\right). \tag{18}$$

3) Evaluation Metrics: To quantitatively evaluate the ship detection performance of these methods, we chose accuracy evaluation indexes from the RS community (P_d, P_m and P_f) and deep learning community [precision, recall, and average precision (AP)]. However, to compute these indicators, the true positives (TPs), false positives (FPs), false negatives (FNs), and true negatives (TNs) in the detection results need to be found first. Further, Intersection over Union (IoU) is required, which represents the overlap ratio between the prediction box S_p and ground truth box S_{gt}. It can be defined as

$$\text{IoU} = \left(S_p \cap S_{\text{gt}}\right)/\left(S_p \cup S_{\text{gt}}\right) \tag{19}$$

If IoU $>$ the setting threshold value, this predicted box is considered as TP, otherwise it is considered as FP. If no predicted box covers the target area, it is treated as a FN. Otherwise, the region is a TN.

Consequently, the detection probability (P_d), missed-detection probability P_m, and false alarm probability P_f are defined as

$$P_d = \text{TP/GT} \tag{20}$$

$$P_m = \text{FN/GT} \tag{21}$$

$$P_f = \text{FP}/(\text{TP} + \text{FP}). \tag{22}$$

The precision and recall can be calculated as follows:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{23}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{24}$$

where the GT is the number of true objects. It is not sufficient to evaluate the performance of the model only using above indexes.

Another comprehensive indicator, the precision-recall curve, shows the tradeoff between precision and recall for different thresholds. It is the average value of precision for each object category when the recall varies from 0 to 1. The closer the curve is to the upper right corner, the better the performance of the model. Furthermore, compared with other indexes, the AP score reflects the performance of the detection model more accurately and intuitively. It is shown as follows:

$$\text{AP} = \int_0^1 P(R)dR \tag{25}$$

where $P(R)$ is the precision-recall (P-R) curve. The model detection speed can be quantitatively evaluated using time (t) and frames per second (FPS)

$$\text{FPS} = 1/t. \tag{26}$$

B. Analysis of Different Scheme Settings

1) Analysis of Multiframe Correlation Scheme: Additionally, although GaoFen-4 is a geostationary satellite that remains stationary toward the Earth, clouds are also moved by atmospheric flow. Therefore, as the number of frames increases, the cloud will move and causing the cloud color separation, as shown in Fig. 8. This will complicate the background to some extent. Therefore, it is a problem to consider whether the more frames the better the detection effect. We conduct comparative experiments on different methods based on different frame numbers to verify the correctness of the proposed continuous frame correlation strategy. The specific results can be seen in Table III. For our approach,

TABLE II
QUANTITATIVE EVALUATION OF THE PROPOSED MODULES IN THIS ARTICLE

Methods	Cloudless	Thin cloud	Broken cloud	Thick cloud	Overall			Inference time
					Recall	Precision	AP	
SF + YOLOv5	76.56	73.27	70.58	71.93	72.95	94.21	72.64	23.22
SF + ODC	78.03	76.57	74.79	75.22	76.08	94.95	75.20	**21.03**
MF + ODC	82.23	80.71	76.81	79.73	80.82	94.79	81.78	21.08
MF + OEC + ODC	89.28	87.41	80.28	85.27	84.30	95.83	85.05	32.31
MF + COCO-Net	**93.30**	**90.33**	**85.42**	**87.25**	**90.65**	**96.69**	**89.61**	32.27

Note: The units of the above accuracy indexes are all percentages (%), and the time is given in milliseconds (ms / image).

TABLE III
COMPARATIVE EXPERIMENTAL RESULTS USING IMAGES WITH DIFFERENT FRAME NUMBERS (k)

Methods	k	Recall	Precision	AP
COCO-Net	1	79.89	96.31	80.23
	3	**90.65**	**96.69**	**89.61**
	5	86.72	94.22	85.23
	7	82.93	92.55	83.87
LSTS	3	83.29	95.33	84.23
	5	87.10	95.82	87.39
	7	86.44	94.79	86.76

TABLE IV
EXPERIMENTAL RESULTS OF HYPERPARAMETER α AND β ON LSD DATASET

Hyper-paramers		AP
$\alpha = 0.9$	$\beta = 0.1$	88.37
$\alpha = 0.8$	$\beta = 0.2$	**89.61**
$\alpha = 0.7$	$\beta = 0.3$	87.52
$\alpha = 0.6$	$\beta = 0.4$	86.33
$\alpha = 0.5$	$\beta = 0.5$	86.49

when the frame number (k) is larger than 3, the last channel image is obtained through the frame difference method. It can be observed that the recall rate of detection results based on single-frame data is about 11% lower than that of three-frame images. This means that many ships are not detected, and we find that most of them are very dim ships. Moreover, the results of five-frame images (the third row) denote that the addition of more frames increases the detection rate but introduces additional noise.

Compared to five-frame images, the seven-frame images causes coverage interference to some targets, thereby reducing the detection rate. We also compare COCO-Net with a good method [learnable spatio-temporal sampling (LSTS)] for treating sequence frame images as video streams [54]. Although the result of five-frame images is better than three-frame images, the detection rate of which is also 3% lower than that of COCO-Net. In the meantime, due to the characteristics of satellite shooting, the time consumption caused by each additional frame is also unacceptable. Thus, we choose three frame image to composite the final image and regarded them as R, G, and B bands of the synthesized images, respectively. The method of correlating multiframe images mentioned in this article is simple, but it is also very effective and necessary.

2) Analysis of Differentn Module: In this module comparison experiment, the constructed LSD dataset is utilized to verify the effectiveness of our proposed module, including the multiframe enhancement strategy, dual-supervised framework with OEC and ODC components, and ROI-Loss control strategy. We set YOLOv5 as the baseline network and added each module in order. The specific results are shown in Table II. "SF" and "MF" represent single-frame images and multiframe images, respectively. For the fairness of the experiment, except for the module for comparison, the parameters of other parts are consistent. We also give the AP value of targets in different scenes, including cloudless, thin cloud, broken cloud, and thick cloud. Different clouds have different interference to the target. It can be observed that the optimized ODC component is better for target detection in broken cloud scenes with more background interference. The dual-supervised framework improves the performance of target detection in cloud-free scenes more significantly.

Furthermore, we can also see the overall AP value of the improved ODC (second row in Table II) is 4% higher than the baseline network (first row in Table II). This shows that ODC component indeed has better detection performance for small ships in low-resolution imagery. Similarly, the third row indicates that the ODC component is used to process multiframe images. The AP and Recall rate is increased by about 5%. The fourth row represents that the OEC and ODC components process multiframe images sequentially, but they are separated during training. The AP and Recall rate is increased by about 8%, showing a substantial positive effect. The last row illustrated the results of COCO-Net for multiframe images. The AP and Recall rate is about 5% higher than the previous one, which denotes the ROI-Loss strategy for the dual-supervised framework is important. Although there is an increase in inference time, this is acceptable.

As shown in Fig. 9, we visualize the feature map from COCO-Net and other compared models. The images in column (a) are original RS images. We also give the corresponding enhanced images in column (b) for a more intuitive view of the ship's position. The images in (c), (d), and (e) are feature

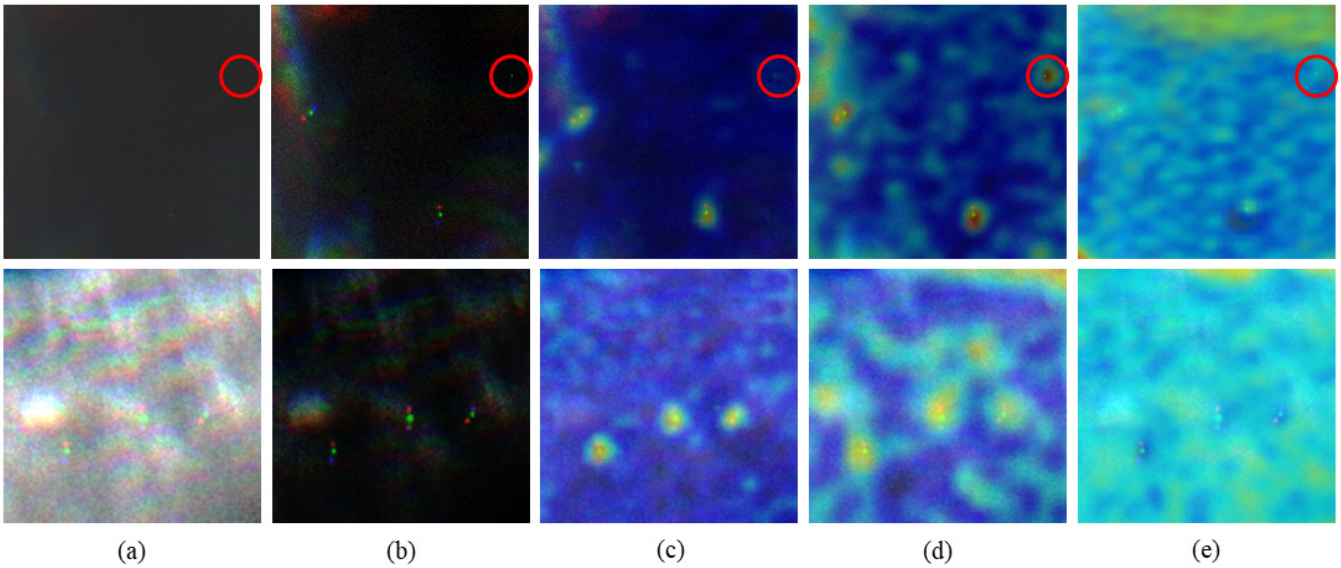

Fig. 9. Validity verification of the different modules. (a) and (b) Illustrate the original images and object-enhanced images, respectively. (c), (d), and (e) Show the feature maps of COCO-Net, COCO-Net without MFCF and detection model without OEC. Red boxes denote a noise point in images.

TABLE V
EVALUATION INDEXES OF DIFFERENT METHODS BASED ON LSD DATASETS

Method	Accuracy (Remote Sensing)			Accuracy (Deep learning)			Speed	
	P_d	P_m	P_f	Recall	Precision	AP	Time	FPS
FasterRCNN	76.46	23.54	9.99	76.46	90.01	75.64	38.11	26.24
SSD	78.32	21.68	7.22	78.32	92.78	77.57	**22.15**	**45.14**
FMSSD	79.38	20.62	6.84	79.38	93.16	78.85	34.49	28.99
R-DFPN	80.14	19.86	6.23	80.14	93.77	79.93	39.67	25.21
YOLOv5	80.26	19.74	5.62	80.26	94.38	80.33	23.32	42.89
R^3-Net	84.39	15.61	4.94	84.39	95.06	84.20	35.71	28.00
MSCNN	82.41	17.59	5.71	82.41	94.29	81.93	31.66	31.59
SME-Net	85.87	14.13	4.73	85.87	95.27	85.31	29.05	34.42
CC-Net	87.33	12.67	5.64	87.33	94.33	87.11	34.49	28.99
COCO-Net	**90.65**	**9.35**	**3.31**	**90.65**	**96.69**	**89.61**	32.27	30.98

Note: Except for AP, which has no units, the units of the above accuracy indexes are all percentages (%), and the time is given in milliseconds (ms / image).

maps of different models and are illustrated by superimposing them on the enhanced images. Among them, the images in (c) are the detected feature map from COCO-Net. It is obvious that the ship feature has been well extracted. Compared with the feature maps generated from COCO-Net, the feature maps processed by the dual-supervised model without the MFCF module show more background noises. In particular, there is an obvious noise in the red circle. The image in column (b) shows that it has comparable feature strength to the ships. But the COCO-Net feature map can not only better emphasize the targets' features but also effectively suppress the interference information. In addition, the images in column (e) illustrate the feature maps generated by a single detection model without OEC. That is, the input images are the original unenhanced data. Obviously, the captured target features in feature maps are weaker than those of the COCO-Net. In conclusion, the effectiveness of the proposed dual-supervised framework and the MFCF module can be verified.

3) Analysis of the ROI-Loss Scheme: This section focuses on finding the best hyperparameters α and β of the proposed ROI-Loss. We effectively correlate OEC and ODC via ROI-mask. It can be seen in Table IV that our model with different parameters consistently improves over the baseline. The best performance is achieved when the loss with $\alpha = 0.8$ and $\beta = 0.2$. This shows that the target enhancement component should pay more attention to the target region and increase the weight of these pixels. However, when $\alpha = 0.9$ and $\beta = 0.1$, the AP value has dropped. It indicates that the background cannot be completely discarded either, and it still has an impact on the final detection performance.

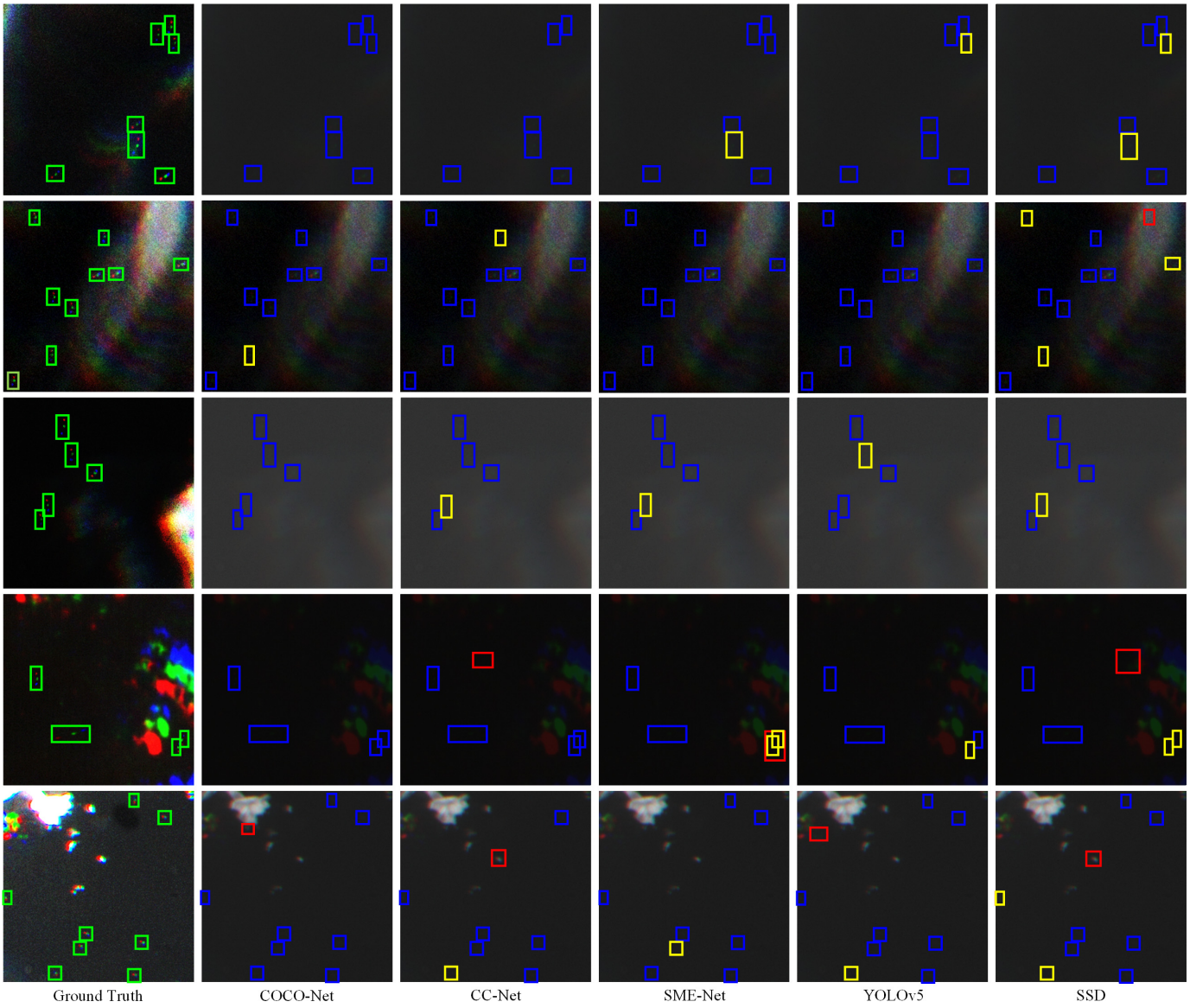

Fig. 10. Detection results of different methods on LSD database. Green boxes denote the ground truth boxes, which illustrated based on the enhanced images. Blue boxes represent the detected positive objects, while orange boxes show the undetected positive samples. Red boxes denote the false alarms.

C. Detection Results and Comparison

To illustrate the detection performance of COCO-Net, we conduct comparative experiments on the LSD database to compare the proposed method with SOTA object detection methods. These comparison methods include classic object detection methods such as faster-region convolution neural network (Faster-RCNN) [55], SSD [56], improved feature-merged single-shot detection (FMSSD) [57], and YOLOv5. There are also some small object detection methods, including R^3-Net [58], multiscale convolutional neural network (MSCNN) [59] and split-merge-enhancement network (SME-Net) [10]. In addition, the multistage method chained cascade network (CC-Net) [14] and the method specifically ship detection rotation dense feature pyramid networks (R-DFPN) [60] are also added for comparison. For effective validation of the different methods, all experiments are performed based on three-frame superimposed data. To be fair, the hyperparameters of all methods are set the same except for different modules. The comparative experiments are also extensive. For methods with open-source code, we leverage them directly for testing. And for popular detectors like Faster R-CNN, the MMDetection project is utilized.

We compare the COCO-Net with other SOTA methods based on the LSD database. The results are shown in Table V, where the optimal results are shown in bold. It can be seen that the AP of COCO-Net for small marine ship detection reaches 89.61% and is significantly higher than that of the other detectors investigated. Moreover, the recall and precision are also increased. In comparison with the models designed for detecting small objects, a multistage network such as CC-Net shows better performance. Our approach combines the multistage network architecture with small target detection technology, which achieves better detection results. For the RS evaluation community, COCO-Net also performs well. Compared with early classic detection algorithms, such as FasterRCNN and SSD, our approach produces

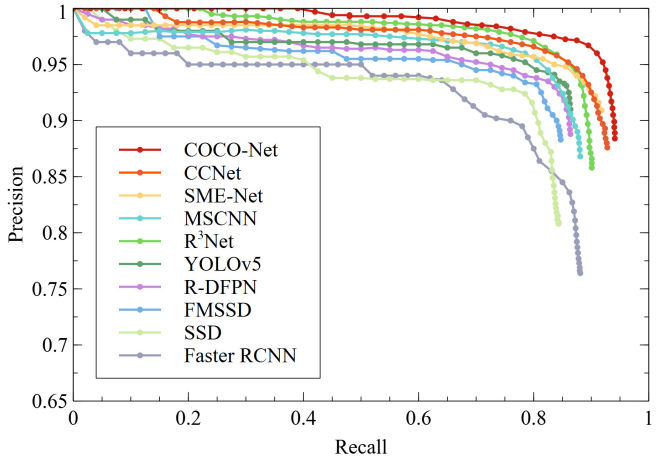

Fig. 11. PR curves achieved by different methods on LSD dataset.

an approximately 13% improvement in detection probability (P_d). At the same time, the P_d of COCO-Net is 5% higher than the best-performing small target detection method SME-Net. In addition, although its inference speed is not the fastest, COCO-Net still performed well in the studies we investigated. Besides, the PR curve for each model is illustrated in Fig. 11. PR curve is a curve with precision as the vertical axis and recall as the horizontal axis. Generally, precision and recall rate vary with different confidence thresholds, and the higher the accuracy, the lower the recall rate. It is difficult to define the true performance of a model from a single set of Precision and Recall values. Therefore, the precision and recall values under all the confidence threshold values are used to construct a curve, and the larger the area contained in the curve, the better the performance of the model. Then, it can be clearly seen that COCO-Net has a stronger performance advantage than other compared methods.

We also give the intuitive visual detection results, as shown in Fig. 10. The ground truth boxes are illustrated on the enhanced images to clearly exhibit the ship. For the results of compared methods, the detection boxes are illustrated on the original RS images. Notably, YOLOv5 or other methods in the YOLO family have superior detection performance in nature images containing multiscale objects. But it does not have a processing mechanism for dim and small targets in ultralow resolution images, so it generally performs in RS images. CC-Net achieves relatively good results through the chained cascade structure, which can handle easy and difficult samples by level. In SME-Net, a split-and-merge module is proposed to eliminate salient information about large targets and highlight the features of small objects. Hence, it can be obtained better detection results for small targets. However, their detection effects are still not good enough for some dim objects whose gray value is close to the background. Among all the methods evaluated, COCO-Net has better detection performance than other methods. It has fewer missed targets (orange boxes) and false alarm targets (red boxes).

D. Discussion

The proposed dual-supervised COCO-Net provides a universal learning framework capable of hierarchically processing target tasks. It is suitable for different domains, and more layers of supervision can also be added. However, it is important to note that although all levels have their own supervision, a regulatory mechanism is required to keep it still serving the final task. This is also the key to distinguishing it from other multisupervised networks. Additionally, a very difficult problem in the marine object detection from optical RS images is the complex and changeable cloud and fog interference. Earlier approaches attempted to process it differently through scene classification. However, class division is a discrete process, so it is still difficult to effectively process the intermediate scene. In our approach, the target enhancement component can adaptively improve the contrast between the targets and background in different scenes. We can observe from Table II that the approach proposed in this study shows a huge advantage. Further, we will conduct more in-depth exploratory research from the view of the lightweight model in the follow-up research. It is also important to put the method into practice.

V. Conclusion

In this study, a well-designed dual-supervised framework is proposed for low-resolution optical satellite imagery ship detection. First, the superimposed sequence images are generated through a preprocessing module, which can enhance small object features by introducing the moving optical flow information of the target. Second, we treat the object detection task hierarchically. The target region contrast is adaptively adjusted first to enhance its saliency through the target enhancement component. Then, the small ODC with a novel MFCF module avoids incorporating background noise and achieves accurate object detection. Finally, an ROI-Loss scheme is proposed to regularize the whole network so that the more ideal input data for object detection can be obtained. We train the whole network uniformly to make the object enhancement network serve the object detection task. The experiments on our newly constructed database: LSD, demonstrate that the proposed COCO-Net is effective and important.

References

[1] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS-J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.

[2] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021.

[3] Z.-Z. Wu, J. Xu, Y. Wang, F. Sun, M. Tan, and T. Weise, "Hierarchical fusion and divergent activation based weakly supervised learning for object detection from remote sensing images," *Inf. Fusion*, vol. 80, pp. 23–43, Apr. 2021.

[4] M. Teutsch and M. Grinberg, "Robust detection of moving vehicles in wide area motion imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 27–35.

[5] S. S. Sengar and S. Mukhopadhyay, "Moving object detection based on frame difference and W4," *Signal, Image Video Process.*, vol. 11, no. 7, pp. 1357–1364, Oct. 2017.

[6] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer, "A survey on moving object detection for wide area motion imagery," in *Proc. IEEE winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[7] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7210–7218.

[8] L. He *et al.*, "End-to-end video object detection with spatial-temporal transformers," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1507–1516.

[9] L. Jiao *et al.*, "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3195–3215, Aug. 2021.

[10] W. Ma *et al.*, "Feature split–merge–enhancement network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[11] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10561–10570.

[12] C. Guo *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1780–1789.

[13] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3025–3038, Dec. 2020.

[14] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Chained cascade network for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1938–1946.

[15] Q. Zhong, C. Li, Y. Zhang, D. Xie, S. Yang, and S. Pu, "Cascade region proposal and global context for deep object detection," *Neurocomputing*, vol. 395, pp. 170–177, Jun. 2020.

[16] Z. Zhang, L. Zhao, Y. Liu, S. Zhang, and J. Yang, "Unified density-aware image dehazing and object detection in real-world hazy scenes," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2020, pp. 1–17.

[17] 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.*, Jun. 2016, pp. 779–788.

[18] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[19] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image Vis. Comput.*, vol. 97, May 2020, Art. no. 103910.

[20] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang, "Recurrent scale approximation for object detection in CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 571–579.

[21] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[22] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[23] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021, doi: 10.1109/TIP.2021.3065239.

[24] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, *arXiv:1902.07296*.

[25] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[26] Y. Kim, B.-N. Kang, and D. Kim, "San: Learning relationship between convolutional features for multi-scale object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 316–331.

[27] C. D. Prakash and L. J. Karam, "It GAN do better: GAN-based detection of objects on images with varying quality," *IEEE Trans. Image Process.*, vol. 30, pp. 9220–9230, 2021, doi: 10.1109/TIP.2021.3124155.

[28] M. Kaur, J. Kaur, and J. Kaur, "Survey of contrast enhancement techniques based on histogram equalization," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 7, pp. 1–5, 2011.

[29] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, "An adaptive gamma correction for image enhancement," *EURASIP J. Image Video Process.*, vol. 2016, no. 1, pp. 1–13, Dec. 2016.

[30] A. S. Parihar and O. P. Verma, "Contrast enhancement using entropy-based dynamic sub-histogram equalisation," *IET Image Process.*, vol. 10, no. 11, pp. 799–808, Nov. 2016.

[31] A. S. Parihar, O. P. Verma, and C. Khanna, "Fuzzy-contextual contrast enhancement," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1810–1819, Apr. 2017.

[32] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust Retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[33] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2010.

[34] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[35] C. Li, J. Guo, F. Porikli, and Y. Pang, "LightenNet: A convolutional neural network for weakly illuminated image enhancement," *Pattern Recognit. Lett.*, vol. 104, pp. 15–22, Mar. 2018.

[36] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[37] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3063–3072.

[38] W. Li and G. Liu, "A single-shot object detector with feature aggregation and enhancement," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3910–3914.

[39] P. Gao, T. Tian, L. Li, J. Ma, and J. Tian, "DE-CycleGAN: An object enhancement network for weak vehicle detection in satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3403–3414, 2021.

[40] M. Zhang, J. Xin, J. Zhang, D. Tao, and X. Gao, "Microscope chip image super-resolution reconstruction via curvature consistent network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 28, 2022, doi: 10.1109/TNNLS.2022.3168540.

[41] E. Koester and C. S. Sahin, "A comparison of super-resolution and nearest neighbors interpolation applied to object detection on satellite data," 2019, *arXiv:1907.05283*.

[42] M. Zhang, Q. Wu, J. Zhang, X. Gao, J. Guo, and D. Tao, "Fluid micelle network for image super-resolution reconstruction," *IEEE Trans. Cybern.*, early access, Apr. 20, 2022, doi: 10.1109/TCYB.2022.3163294.

[43] Y. Shen *et al.*, "Noise-aware fully webly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11326–11335.

[44] Q. An, Z. Pan, L. Liu, and H. You, "DRBox-v2: An improved detector with rotatable boxes for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8333–8349, Nov. 2019.

[45] X. Guan, Z. Peng, S. Huang, and Y. Chen, "Gaussian scale-space enhanced local contrast measure for small infrared target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 327–331, Feb. 2020.

[46] L. Wang, B. Yin, X. Tang, and Y. Li, "Removing background interference for crowd counting via de-background detail convolutional network," *Neurocomputing*, vol. 332, pp. 360–371, Mar. 2019.

[47] Yang, Li, Min, and Wang, "Real-time pre-identification and cascaded detection for tiny faces," *Appl. Sci.*, vol. 9, no. 20, p. 4344, Oct. 2019.

[48] B. Liao *et al.*, "PG-Net: Pixel to global matching network for visual tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 429–444.

[49] S. Li, Y. Xu, M. Zhu, S. Ma, and H. Tang, "Remote sensing airport detection based on end-to-end deep transferable convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1640–1644, Oct. 2019.

[50] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11908–11915.

[51] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

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

[54] Z. Jiang *et al.*, "Learning where to focus for efficient video object detection," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 18–34.

[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 28, 2015, pp. 91–99.

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

[57] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, Dec. 2019.
[58] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R3-Net: A deep network for multioriented vehicle detection in aerial images and videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5028–5042, Jul. 2019.
[59] Q. Yao, X. Hu, and H. Lei, "Multiscale convolutional neural networks for geospatial object detection in VHR satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 23–27, Jan. 2020.
[60] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, 2018.

Qizhi Xu (Member, IEEE) received the B.S. degree from Jiangxi Normal University, Nanchang, China, in 2005, and the Ph.D. degree from Beihang University, Beijing, China, in 2012.

He was a Post-Doctoral Fellow with the University of New Brunswick, Fredericton, NB, Canada. He is currently an Associate Professor with the Beijing Institute of Technology, School of Mechatronical Engineering, Beijing, China. His research interests include image fusion, image understanding, and big data analysis of remote sensing.

Dr. Xu was a recipient of the Technological Invention Award First Prize from the Chinese Institute of Electronics for his image fusion research in 2017.

Yuan Li received the B.S. and M.S. degrees from the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, in 2018 and 2021, respectively. She is currently pursuing the Ph.D. degree with the School of Mechatronical Engineering, Beijing Institute of Technology.

Her research interests include remote sensing image process and pattern recognition.

Mingjin Zhang (Member, IEEE) received the B.Sc. degree in electronic and information engineering and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

She is currently an Associate Professor with the State Key Laboratory of Integrated Services Networks, Xidian University, Key Laboratory of Spectral Imaging Technology of Chinese Academy of Sciences, and Science and Technology on Reliability Physics and Application Technology of Electronic Component Laboratory. From October 2015 to October 2016, she has been a Visiting Ph.D. Student with the University of Technology, Sydney, NSW, Australia. Her research interests include computer vision, pattern recognition, and machine learning. She has published more than ten articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON COMPUTERS (TC), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), the Association for the Advancement of Artificial Intelligence (AAAI), and the International Joint Conference on Artificial Intelligence (IJCAI).

Wei Li (Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

Subsequently, he spent one year as a Post-Doctoral Researcher with the University of California, Davis, CA, USA. He was a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, from 2013 to 2019. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing. His research interests include hyperspectral image analysis, pattern recognition, and data compression.

Dr. Li is currently serving as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He has served as the Guest Editor for special issue of the *Journal of Real-Time Image Processing*, *Remote Sensing*, and the IEEE JSTARS. He received the 2015 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS) for his service for the IEEE JSTARS and the Outstanding Paper Award at the IEEE International Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers), 2019.