LGC-YOLO: Local-Global Feature Extraction and Coordination Network With Contextual Interaction for Remote Sensing Object Detection

Qinggang Wu, Yang Li, Junru Yin, and Xiaotian You

***Abstract*—Object detection in high-resolution remote sensing image (HRRSI) faces great challenges of large-scale variations in object size, densely distributed small objects, and complex background interferences. To address these challenges, we propose an innovative single-stage local-global feature extraction and coordination network (LGC-YOLO) to improve the detection accuracy of objects in HRRSIs. LGC-YOLO mainly comprises three modules of local-global spatial feature extraction (LGSFE), gradient optimized spatial information interaction (GOSII), and edge-semantic feature coordination fusion (ESFCF), which synergistically improves the feature extraction and object detection capabilities of LGC-YOLO. First, LGSFE captures local and global features of dense objects through receptive-field attention convolution and global pooling in a multibranch structure, which effectively alleviates the misalignment between the extracted features of objects and their intrinsic characteristics, thereby providing more accurate and abundant features for subsequent object detection. Second, GOSII is designed to dynamically adjust the weights of each feature channel through combining SRU blocks and the SimAM attention mechanism, which are further optimized and embedded into C2f to enhance the representation ability of contextual features. GOSII captures crucial features from complex backgrounds and improves information transmission. Finally, ESFCF integrates the edge and semantic information within shallow feature maps to address the issue of inaccurate localization for small objects, and further improves object detection accuracy by compensating for the loss of edge details in feature extraction. Extensive experiments on three commonly used remote sensing datasets of NWPU VHR-10, VisDrone 2019, and DOTA demonstrate the superiority of our method in object classification and localization compared to other state-of-the-art methods.**

***Index Terms*—Attention mechanism, edge features, remote sensing object detection (RSOD), small object, YOLO.**

I. Introduction

NOWADAYS, remote sensing technology provides a powerful way for Earth observation tasks, which gathers vast coverage of Earth's surface. Recently, remote sensing image processing [1], [2], [3], [4], [5] has been significantly revolutionized with rapid advances in deep learning methods [6], [7], [8], [9]. Remote sensing object detection (RSOD) has been widely utilized in both military and civilian fields, such as military reconnaissance [10], urban planning [11], environmental monitoring [12], [13], and so on.

RSOD aims to automatically identify and localize objects from high-resolution remote sensing image (HRRSI). Motivated by natural object detection methods [14], current RSOD methods based on deep learning can be roughly categorized into supervised [15], weakly supervised [16], and unsupervised [17] methods. Although the latter two types of methods have been studied for RSOD in recent years, they often suffer from the problems of high model complexity and instability. Therefore, the supervised methods become the most widely used ones. Generally, supervised object detection methods are mainly divided into single-stage and two-stage approaches. In two-stage methods [18], [19], [20], the candidate boxes or regions are initially generated, and then object classification and location regression are performed for each candidate box or region. These methods obtain higher object detection accuracy but struggle to improve detection speeds. On the other hand, single-stage methods [21], [22], [23] perform object detection and localization in an end-to-end manner instead of explicitly generating candidate boxes, which have the advantages of real-time detection speed. Although these methods have achieved good results for RSOD, they still face significant challenges in detecting the densely arranged small objects from complex HRRSIs captured by spaceborne or airborne cameras.

Recently, most studies have been conducted to obtain rich and diverse spatial features to detect densely arranged objects from HRRSIs. Zhu et al. [24] proposed an object detection method of GMDRA-net based on global multilevel perception (GMP) and dynamic region aggregation (DRA) to address the problems of lacking global information while detecting densely distributed small objects. Li et al. [25] introduced a GALDET method for remote sensing micro-object detection based on convolutional (Conv) block as well as global and local attention mechanism (GAL), to address the problems of insufficient contextual features and poor localization capabilities for densely distributed small targets. Yu et al. [26] designed a cascaded significant attention network (CSAN) by combining context and pixel attention maps to enhance dense salient objects and suppress background

Received 7 February 2025; revised 26 April 2025; accepted 22 May 2025. Date of publication 30 May 2025; date of current version 1 July 2025. This work was supported in part by the Young Backbone Teacher Training Program of Henan Province under Grant 2023GGJS090, in part by the Scientific and Technological Research Project of Henan Provincial Department of Science and Technology under Grant 242102210013, in part by the National Natural Science Foundation of China under Grant 61502435 and Grant 62476255, in part by the Key Scientific Research Projects of Colleges in Henan Province under Grant 23A520001, and in part by the Natural Science Foundation of Henan Province under Grant 252300420389. *(Corresponding author: Qinggang Wu.)*

The authors are with the College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China (e-mail: wuqinggang323@126.com).

Digital Object Identifier 10.1109/JSTARS.2025.3575239

interferences. Zhang et al. [27] proposed a unified object contour detector of UniconDet to integrate multigrained object detection requirements of horizontal bounding box (HBB), oriented bounding box (OBB), and instance segmentation (InSeg) into a challenging task of arbitrary-shaped object contour detection. As a generic vision backbone, Vision Mamba (Vim) incorporates bidirectional state space modeling (SSM) for data-dependent global visual context modeling, and becomes the first pure-SSM-based model to handle dense prediction tasks. However, it struggles to capture fine-grained local details of small objects. Zhang et al. [28] utilize rotated enclosing boxes and pixel-refinedwise module (PRW) to enhance the representation of densely aligned objects by encoding a wider range of contextual information into localized features. Although the above methods improve the detection performance of densely arranged objects in HRRSI, they often suffer from the feature misalignment across different scales.

In addition, many researchers have paid much attention in extracting crucial features from HRRSIs with complex backgrounds. Li et al. [29] employed a semantic transfer block (STB) to reduce noise interference and recover semantic information, which not only adapts to objects across different scales and obtains accurate bounding boxes, but also effectively reduces the influence of complex backgrounds. Zhang et al. [30] designed a framework for object detection in remote sensing images, i.e., coarse-to-fine feature adaptation (COF-FA) and sample allocation (COF-SA), which aims to progressively enhance feature representations and select stronger training samples to extract useful information from complex backgrounds. Xie et al. [31] proposed a Swin-DETR method based on DETR to improve the small object detection performance from complex environments. Zhang et al. [32] introduced an RSOD framework named MSHEMN to optimize the feature pyramid structure of FPN by integrating a multiscale region proposal network (MSRPN) with lateral connection blocks (LCBs) and adaptive feature merging (AFM), which fuses high-resolution and semantically rich features of objects from complex backgrounds. Liu et al. [33] proposed a YOLO-SSP model based on YOLOv8 by introducing a small object detection layer and a pyramid spatial attention mechanism, which enhances the detection performance of small objects in complex backgrounds. However, these methods are still susceptible to background noise and redundant features.

Furthermore, it brings significant difficulties in detecting small-sized objects since they occupy only dozens of pixels. Qi et al. [34] designed a single-stage small object detection network (SODNet) by integrating an adaptively spatial parallel convolution (ASPConv) module and a fast multiscale fusion (FMF) module to improve small object detection accuracy. Ma et al. [35] utilized a multitask joint training method to provide richer semantic structural features for bounding box localization of small objects. Zhang et al. [36] proposed a controllable generative knowledge-driven few-shot object detection (CGK-FSOD) method based on visual-textual prompts to address the challenges of detecting unseen objects in optical remote sensing images. Cao et al. [37] integrated the SIoU loss function into YOLOv5 to achieve precise localization for densely aligned small objects. Wu et al. [38] eliminated the unnecessary residual modules from traditional cross-layer spatial (CSP) layer in YOLOv5s, and also present an improved residual coordinate attention (RCA) module. By incorporating residual structures and mix pooling (i.e., local max pooling or average pooling) to replace existing average pooling, it enhances the feature representation for densely distributed small objects in remote sensing images. However, the small objects in HRRSI are often missed by existing methods due to insufficient detailed feature extraction and utilization.

The detection of densely arranged small objects from complex HRRSIs has emerged as a prominent research topic [39]. Although many methods achieve promising object detection results, there are still several challenges. First, it brings great difficulties in effectively extracting local feature and model global contextual features. Conventional backbone networks predominantly capture local features of objects, and subsequent attempts to integrate global representations through multiscale fusion often induce spatial misalignment in feature maps, which will result in false positive predictions. Second, enhanced spatial resolution of remote sensing images introduces discriminative details. Current researches prioritizing extracting more abundant spatial features inadvertently bring redundant noise and background interferences, thereby degrading the separability between objects and intricate backgrounds. Finally, with the continuous advancement of deep learning techniques, researchers increasingly concentrate on extracting semantically rich features from deep layers, which inevitably leads to the loss of geometrically critical shallow-layer features. This will result in insufficient feature extraction for small targets that occupy relatively small number of pixels in remote sensing images, causing them indistinguishable from environmental textures and ultimately leading to miss detections.

These challenges highlight the need for innovative methods that extract comprehensive features, suppress redundant information, and preserve details and edge features. Our research focuses on developing an effective method that not only extracts local-global features and crucial ones, but also retains small object features. To address these challenges, this article proposes a novel single-stage RSOD method, i.e., LGC-YOLO, to improve the detection performance of densely arranged small objects from complex HRRSIs. The proposed LGC-YOLO primarily includes three key modules, i.e., the local-global spatial feature extraction (LGSFE) module, the gradient optimized spatial information interaction (GOSII) module, and the edge-semantic feature coordination fusion (ESFCF) module. First, LGSFE adopts a multibranch structure to capture local and global spatial features through receptive field attention mechanism, which effectively enhances the model's ability to capture object features in densely arranged scenes, thereby mitigating the potential inconsistencies between extracted features and inherent object characteristics. Second, GOSII is designed to enhance the understanding of contextual information by optimizing the spatial features of objects against complex backgrounds. It effectively increases information flow and reduces spatial redundancy by recombining feature maps, as well as extracting the important information from complex backgrounds by combining attention

mechanism. Finally, ESFCF is presented to ensure that the extracted features are not only rich in semantic information but also integrated with abundant edge and texture details during feature extraction, effectively avoiding the loss of edge features to help with accurate object detection. The contributions of this article can be summarized as follows:

1) We propose a novel method, i.e., LGC-YOLO, to detect densely arranged small objects from complex HRRSIs by effectively extracting local-global spatial features through contextual information interaction. The proposed LGC-YOLO significantly reduces parameters while improving object detection accuracy via contextual feature coordination, demonstrating strong potential remote sensing applications.
2) We design a LGSFE module to maintain the consistency between extracted features and target ones through multibranch feature interaction. In this module, one branch highlights the extraction of local spatial features, while remaining branches emphasize the mining of global spatial features, which enables our method to accurately capture and align object features and enhance object detection performance.
3) We propose a GOSII module by incorporating spatial reconstruction unit (SRU) block to effectively capture the crucial spatial features from complex backgrounds, which are further enhanced by information flow through well-designed gradient propagation to obtain the abundant contextual features. This design not only maintains the lightweightness of our method but also improves information transmission.
4) To preserve the accurate localization of small objects, we design the ESFCF module by integrating attention mechanism to enhance shallow features and supplement abstract features. Apart from preserving rich semantics features, ESFCF captures important edge information, which effectively compensates for the loss of edge details in feature extraction.

The rest of this article is organized as follows. Section II introduces related techniques and works. In Section III, the proposed LGC-YOLO method is elaborated in detail. Extensive comparative experiments are conducted in Section IV. Section V discusses the limitations and future research directions. Finally, Section VI concludes this article.

II. Related Works

A. Receptive-Field Attention Convolution

The dense arrangement of objects in remote sensing images brings great challenges to accurate object detection due to partial occlusions or overlaps. Most existing studies leverage multifeature fusion or InSeg to mine distinct features for these densely distributed objects, which may result in inconsistency between the extracted features and ground truth ones. To address this problem, this article proposes an innovative solution by designing a multibranch local and global feature extraction and fusion module to enhance the feature alignment and object detection performance. In this module, an improved receptive-field attention convolution (RFAConv) [40] block is adopted to perform convolution operations, which greatly enriches the varieties and details of extracted features by shifting attention mechanism from traditional spatial features to those in receptive fields.

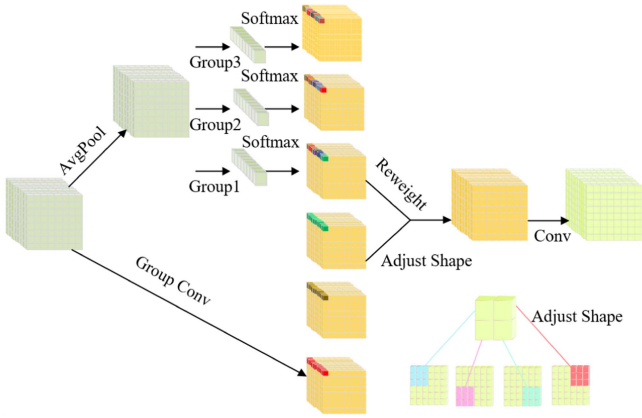

Fig. 1. Detailed structure of RFAConv, which dynamically assesses the importance of each feature within the receptive field.

As shown in Fig. 1, RFAConv first employs AvgPool to aggregate the global information for each receptive field, which minimizes the computational overhead and parameter numbers. Then, a $1\times$ 1 group convolution is adopted to strengthen information interaction. Finally, Softmax is utilized to emphasize the importance of each feature in receptive field. The formula of RFAConv is shown in the following:

$$F = Softmax\left(g^{i\times i}\left(AvgPool(X)\right)\right) \times ReLU\left(Norm\left(g^{k\times k}(X)\right)\right) = A_{rf} \times F_{rf} \quad (1)$$

where $g^{i\times i}$ represents the group convolution of size i $\times$ i, k is the size of convolution kernel, $Norm$ denotes the normalization operation. X is the input feature map. Feature map F is obtained by multiplying attention map A_{rf} with space feature F_{rf} in transformed receptive field.

The spatial features in receptive field obtained by RFAConv do not overlap with each other after the adjust shape operation. Therefore, the learned attention map aggregates the features in each receptive field slider. RFAConv enhances the feature details and diversities by incorporating group convolution and spatial sharing receptive field features, which may lead to some additional storage and computational overheads. However, due to the superior local spatial feature extraction capability, RFAConv favors in extracting the features of densely arranged objects from complex remote sensing images.

B. SRU Block

As we know, it often includes a vast coverage of Earth surfaces as the spatial resolution of HRRSIs increases, which often leads to complex background interferences and even much redundant information. This unnecessary information often causes more sever false detections, where some backgrounds are erroneously

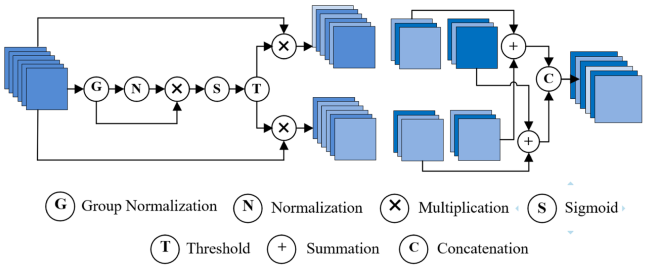

Fig. 2. Architecture of the SRU block. It mainly includes the separation and reconstruction operations to suppress redundant information.

identified as an object and further decreases the detection performance. To address these challenges, this article proposes an enhanced module by incorporating SRU [41] to suppress spatial feature redundancies. In the SRU block, a weighting mechanism is performed for separating redundant features and reconstructing useful information, which suppress the redundant components in spatial features while enhancing the representation of effective features. This strategy helps to improve the accuracy and robustness of object detection in HRRSIs.

As shown in Fig. 2, SRU utilizes separation and reconstruction operations. First, the information contained in different feature maps is evaluated using scale factor in group normalization (GN) layer. Then, the feature maps with abundant information are separated from less informative feature maps in terms of spatial content by utilizing separation operation. Subsequently, the cross-reconstruction operation is employed to combine two weighted features and enhance the information flow inside them, which effectively suppress spatial redundancy. The input feature X is normalized by first subtracting mean μ and then divided by standard deviation θ as follows:

$$X_{\text{out}} = GN(X) = \gamma \frac{X - \mu}{\sqrt{\theta^2 + \varepsilon}} + \beta \tag{2}$$

where ε is a small positive constant in denominator to maintain the stability of division calculation while γ and β are trainable affine transformations

$$W = Gate\left(Sigmoid\left(W_\gamma\left(GN(X)\right)\right)\right). \tag{3}$$

The weights of the feature map W_γ are mapped between 0 and 1 by $Sigmoid$ function and selected via a threshold. The weights above threshold are set to 1 to obtain informative weights $W1$, while 0 below threshold to obtain noninformative weights $W2$. The whole process of achieving W can be expressed as (3).

Finally, the input feature X is multiplied by $W1$ and $W2$ to obtain the informative features X_{W1} and less informative features X_{W2}, respectively. Thus, the input is divided into two parts of X_{W1} with informative spatial content and X_{W2} with less valuable or redundant information.

C. Coordinate Attention (CA) Block

Generally, it tends to achieve inaccurate object detection results depending only on high-level semantic information for small objects, since the important features of them are often

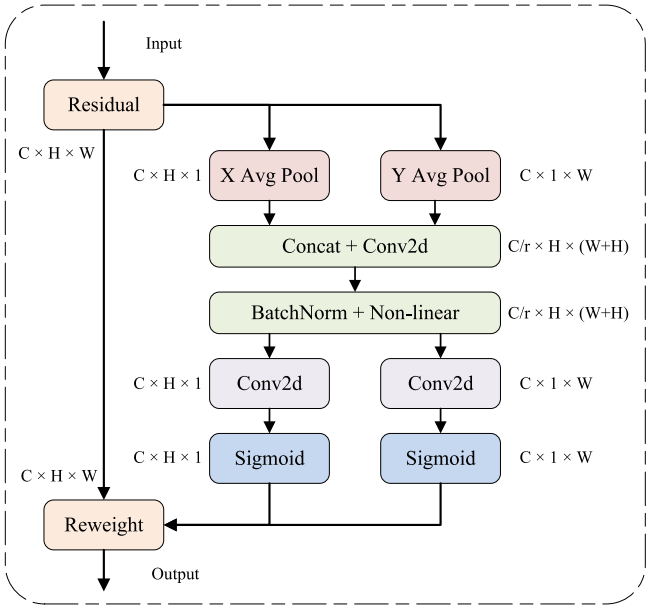

Fig. 3. Structure of CA mechanism. CA embeds coordinate information to capture local features along horizontal and vertical dimensions, respectively.

embedded in edge and texture information of complex backgrounds. To address this problem, most researchers employ attention mechanism to focus on important features, and accordingly we design a novel attention module by extending CA [42] to accurately capture and enhance edge and texture information. Specifically, it improves the detection accuracy of small objects by combining spatial coordinate information with channel attention mechanism.

The structure of CA mechanism is shown in Fig. 3. To obtain the attention of an image in width and height and encode the location information, global average pooling is performed on input feature map (size of C × H × W) in *X*-direction (width) and *Y*-direction (height), which generates the feature maps of size $C\times$ H $\times$ 1 and C $\times$ 1 $\times$ W, respectively. Subsequently, the feature maps Z^h, Z^w in the width and height of receptive field are spliced together, and then fed into a convolution module with shared kernel of $1\times$ 1 to reduce their dimensions to original C/r. Afterward, the batch of normalized feature maps F_1 are fed into Sigmoid function δ to obtain the feature map f with shape of $1\times$ (W+H)$\times$ C/r, as shown in the following:

$$f = \delta\left(F_1\left(\left[Z^h, Z^w\right]\right)\right). \tag{4}$$

The feature map f is split along spatial dimension. Then, $1\times$ 1 convolution is performed according to original height and width to obtain the feature map F_h, F_w, which have the same number of channels as the original feature map. After Sigmoid function, the attention weights g^h, g^w are obtained in height and width of feature map, respectively. The formulas are shown in the following:

$$g^h = \alpha\left(F_h\left(f^h\right)\right) \tag{5}$$

$$g^w = \alpha\left(F_w\left(f^w\right)\right). \tag{6}$$

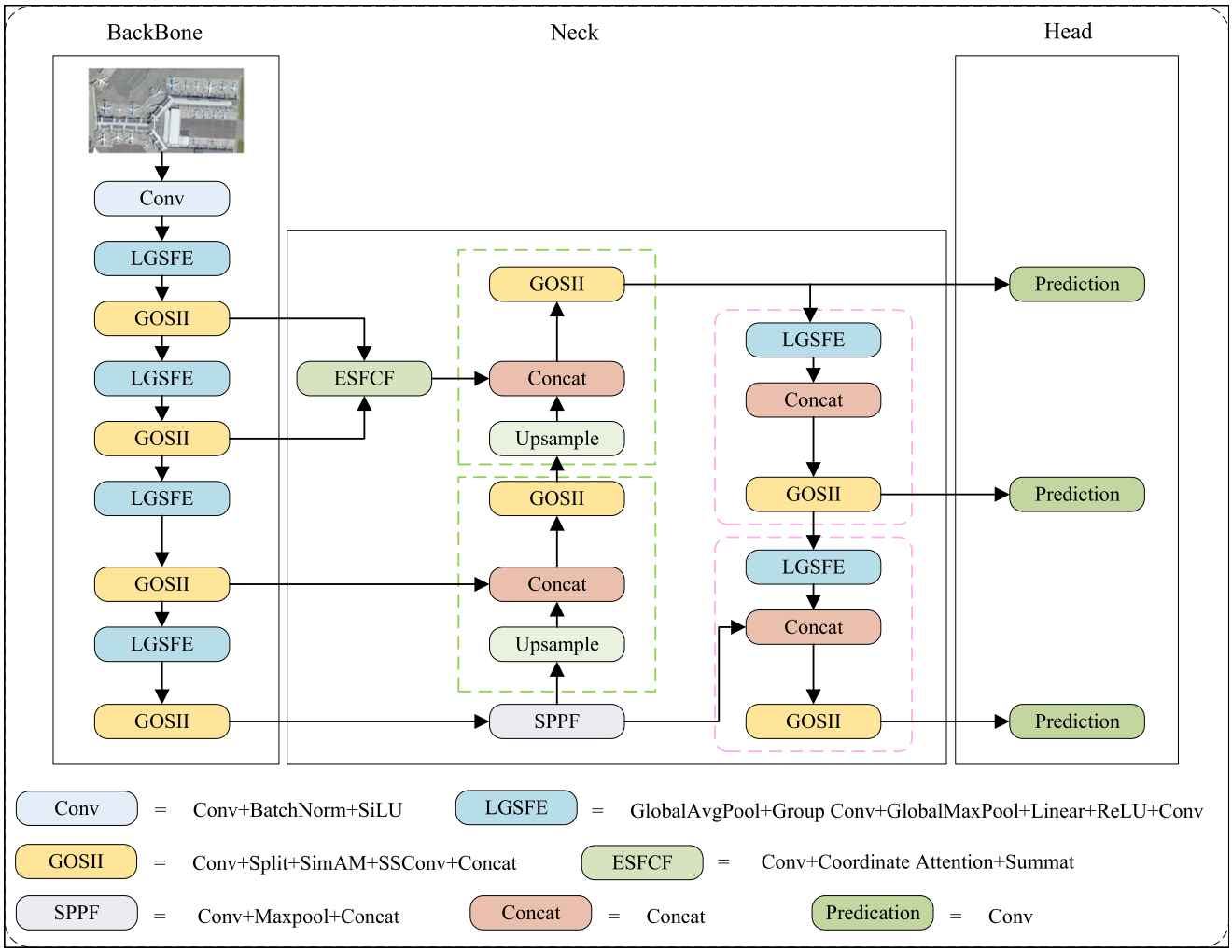

Fig. 4. Overall architecture of the proposed LGC-YOLO method, which mainly consists of three components of BackBone, Neck, and Head. This article put forward the blue LGSFE and yellow GOSII modules for the backbone as well as the green ESFCF module for the neck.

Finally, the feature map will be obtained with attention weights in width and height directions by multiplied with original feature map.

III. Proposed Method

In this article, we propose a new one-stage method, i.e., LGC-YOLO, to improve the object detection accuracy from HRRSIs. The overall architecture of our method is illustrated in Fig. 4. LGC-YOLO primarily includes the three components of Backbone, Neck, and Head. This section will elaborate on the newly proposed three modules of LGSFE, GOSII, and ESFCF that are mainly integrated in Backbone and Neck. The prediction head will adopt the commonly used decoupled structure [43].

A. LGSFE Module

The accurate detection of densely arranged small objects from HRRSIs faces great challenges due to insufficient extraction of detailed object features for distinguishing edge information. To address this problem, researchers employ the strategies of multiscale or adaptive-scale feature extraction and fusion via atrous convolutions with different dilation rates. Although these methods [44], [45], [46] achieve better object detection accuracy in specific scenarios, there are still difficulties in accurately aligning the extracted features with inherent object characteristics, especially for the densely arranged small objects in HRRSIs.

To address these challenges in accurately detecting densely arranged small objects from HRRSIs, a multibranch local and global spatial feature extraction module, i.e., LGSFE, is proposed by incorporating RFAConv block. The structure of LGSFE is illustrated in Fig. 5, and it combines three branches to comprehensively improve the accuracy and robustness of feature extraction. Supposing $X \in R^{C\times H\times W}$ is an input feature map, where C, H, and W denote the channel numbers, height, and width, respectively. The first branch primarily captures contextual information with the size of C$\times$ 1 via global average pooling operation. The second branch dynamically focuses on important local information in different regions by adjusting the features from RFAConv block to the dimension of C$\times$ KH$\times$

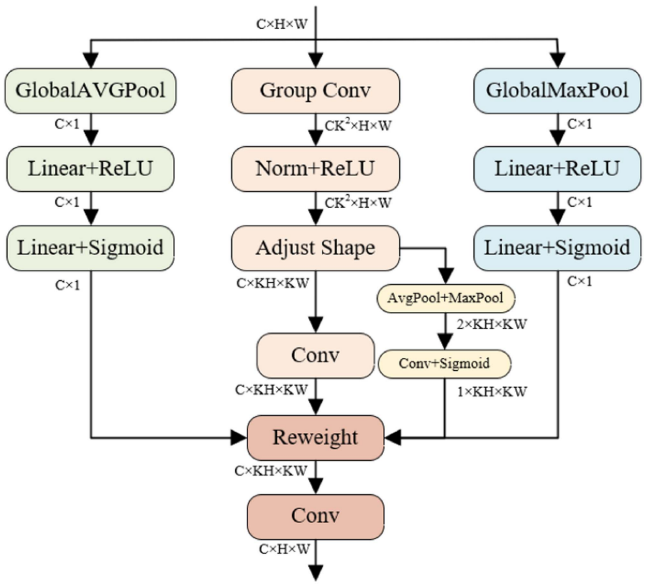

Fig. 5. Proposed LGSFE module simultaneously captures local and global features through a multibranch structure, which effectively maintains the feature alignment.

KW according to the weights from spatial attention mechanism with the dimension of 1× KH× KW. RFAConv improves the alignment of local spatial feature extraction through the final convolution operation. Similar to the first branch, the third branch emphasizes on the extraction of global contextual features with the size of C× 1 via global max pooling operation. The first and third branches establish the robust long-range connections to improve the ability to capture global spatial structures. Finally, a reweighting mechanism is introduced to further enhance the feature representation ability from these three branches by multiplication operation, which refines the enhancement of spatial features through attention maps from different branches and ensures LGSFE to extract richer spatial object features. The operations in LGSFEcan be formulated as follows:

$$F_{LGSFE} = R\left(AvgPool(X) \times ReLU\left(Norm\left(g^{3\times3}(X)\right)\right) \times MaxPool(X)\right) \tag{7}$$

where $AvgPool$ and $MaxPool$ are the global average and max pooling operations, respectively. g indicates the groupwise convolution with kernel size of 3×3. $Norm$ is the normalization operation, $ReLU$ signifies the activation function, $R(\cdot)$ denotes the weighting operation on the combined features in above three branches. Finally, F_{LGSFE} indicates the output of both local and global spatial features.

The proposed LGSFE module improves the local and global feature representation, which extracts spatial features by combining receptive-field attention mechanism. It notes that the features extracted from different branches of LGSFE are effectively aligned for densely arranged small objects in HRRSIs with complex backgrounds, since it captures various scales of features due to the dynamic adjustment of convolution kernel size in the receptive field of RFAConv block. In all, LGSFE highlights spatial features within receptive field, which allows the model to more effectively extract the local features from RFAConv instead of restricting to traditional spatial features.

B. GOSII Module

Abundant local and global spatial features have been extracted in the previous module, which will inevitably bring some background interferences and redundant noise, which is not conducive to the detection of small objects from HRRSIs with complex backgrounds. To address this problem, we propose the GOSII module to optimize features by integrating SRU and SimAM to alleviate the interference of backgrounds and redundant noise and distinguish interested objects from complex backgrounds. These features are further integrated with the C2f block to promote the interaction between them to extract contextual information and address the problem of small object detection from HRRSIs with complex backgrounds.

The structure of the proposed GOSII module is shown in Fig. 6(a), which includes a series of operations of Convolution, Splitting, SRAConv, and Concatenation. Among them, SRAConv plays a vital role in primarily extracting important features, which is composed of SRA and Conv blocks, as shown in Fig. 6(b). Here, the proposed SRA block is employed to replace the Convolution operation in original BottleNeck of C2f to achieve the purpose of extracting crucial features. The structure of SRA is illustrated in Fig. 6(c). First, the features are fed into SRU to suppress the complex background interferences and redundant noise in them. During this operation, the weight processing is performed on compressed spatial features through spatial attention to suppress the interference of less important backgrounds and enhance the focus on crucial regions. Second, these features are further fed into SimAM to highlight crucial features, thereby separating the important features from background. The calculation for SimAM is as follows:

$$E_t\left(w_t, b_t, y, x_i\right) = \frac{1}{M-1}\sum_{M-1}^{i=1}\left(-1-\left(w_i x_i + b_i\right)\right)^2 + \left(1-\left(w_i x_i + b_i\right)\right)^2 + \lambda w_t^2 \tag{8}$$

where t and x_i are the target neuron and other neurons in the input feature x, respectively. i is the index along spatial dimension, and M denotes the total number of neurons in a channel. w_t and b_t represent the weights and biases in the transformation for a single neuron. λ refers to a hyperparameter with a value of 0.0001.

On the other hand, as an efficient and lightweight attention network (ELAN) [47], C2f demonstrates excellent ability in capturing crucial features. To further enhance the capture of global contextual information and transmission efficiency of gradient flow, we introduce the C2f block in our proposed GOSII module to improve information flow while reducing model complexity and increase the ability of the model to establish long-range feature relationships.

The GOSII module suppresses redundant information and enhances the interaction of contextual features, which provides

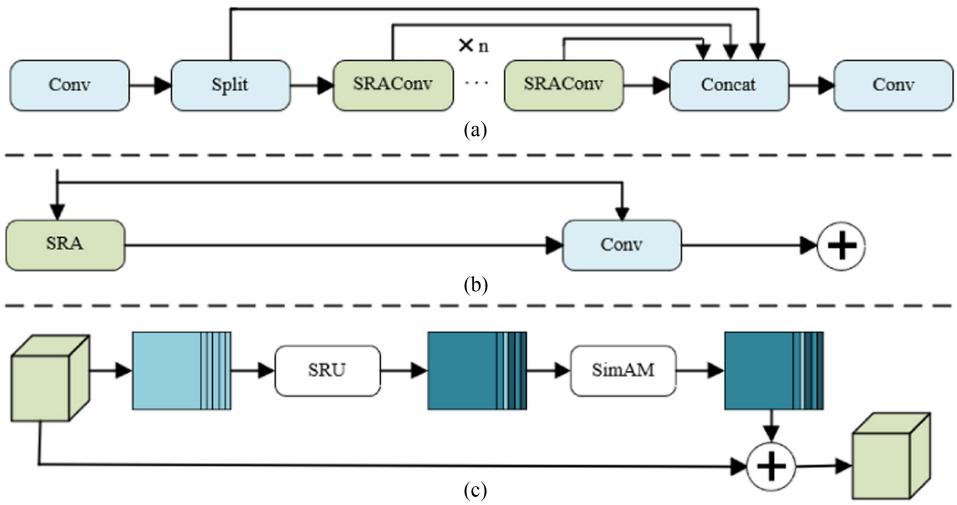

Fig. 6. (a) Detailed structure of the proposed GOSII module. (b) Structure of the SRAConv block within GOSII module. (c) Structure of the SRA block within SRAConv by integrating SRU and SimAM blocks.

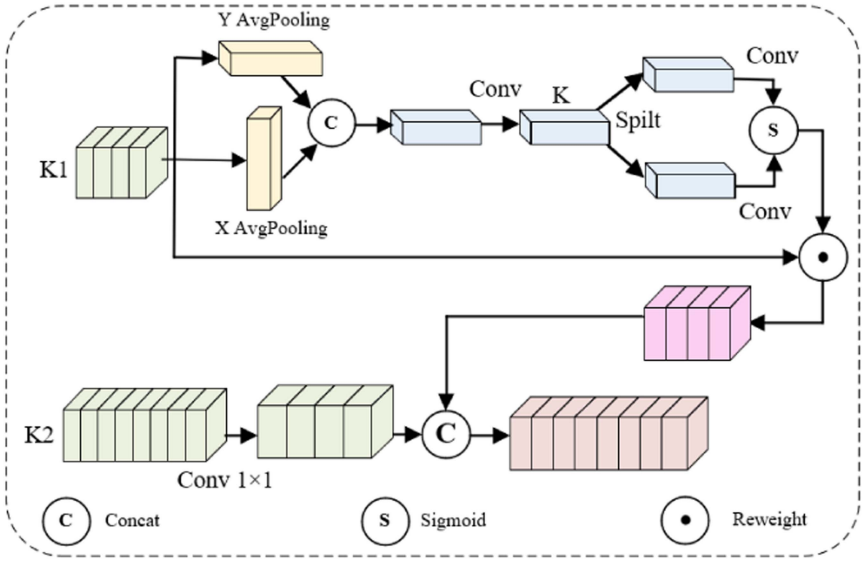

Fig. 7. Detailed structure of the proposed ESFCF module, which consists of a dual-branch structure including K1 and K2 to capture edge and texture features as well as semantic information, respectively.

more effective information and improves the ability of the model to classify and locate small objects in complex remote sensing images. Note that GOSII is also effective for detecting objects from HRRSIs with simple backgrounds or less interference, since the redundancy removal strategy seldom causes the loss of valuable information. In fact, complex backgrounds benefit from strong filtering, while simple backgrounds from weak filtering.

C. ESFCF Module

As the number of layers in deep learning network increases, the shallow edge information tends to be lost, which yet plays a vital role for small object detection. To mitigate this problem, we design the ESFCF module to integrate shallow information and deep semantic information, which is integrated between backbone and neck as shown in Fig. 4. CA improves the localization accuracy and enhances the ability to suppress background interferences to small objects, thereby increasing the detection accuracy of small objects. By focusing on the features across different channels, CA significantly boosts the feature representation ability to extract edge and texture information from HRRSIs.

The proposed ESFCF module consists of two input branches, i.e., K_1 and K_2, as shown in Fig. 7. Since the number of input channels in these two branches does not correspond with each other, we first adjust the number of channels in K_2 to coincide with that of K_1 by using the $1\times$ 1 convolution of $T(\cdot)$, which can be expressed as follows:

$$K_2^* = T\left(K_2\right) \tag{9}$$

where K_2^* is the feature map obtained via a $1\times$ 1 convolution, which preserves the semantic information of input feature map.

The K_1 branch with rich edge and texture information is divided into two subbranches for global average pooling in height and width directions to obtain two feature maps corresponding to each direction. Subsequently, these feature maps are concatenated via Concat operation to obtain a feature map containing detailed information and semantic information, which is convolved using a shared $1\times$ 1 convolution to obtain the feature map K. Then the split operation is applied to K, and a $1\times$ 1 convolution is performed separately according to original height and width to obtain the feature maps F_h and F_w with the same number of channels as the original feature map. After applying Sigmoid function, the attention weights g_X and g_Y for the feature map in height and width directions are, respectively, obtained as follows:

$$g_X = \alpha\left(F_h\left(K_1\right)\right), g_Y = \alpha\left(F_w\left(K_1\right)\right). \tag{10}$$

The adjusted feature maps in both branches of K_1 and K_2 are spliced together by Concat operation

$$F_{\text{ESFCF}} = C\left(K_2^*, R\left(g_X, g_Y\right)\right) \tag{11}$$

where $C(\cdot)$ represents Concat operation and $R(\cdot)$ denotes the Reweight operation. The final feature map F_{ESFCF} contains rich shallow edge information and deep semantic information.

The ESFCF module integrates the high-level semantic information with low-level texture information. Although CA primarily relies on the average aggregation of global features, it also captures the large local variations. In other words, CA not only allocates the weights for important channels through attention mechanism but also incorporates spatial positional information. Thus, the combination of spatial and channel information allows CA to more effectively extract crucial features within local regions. Meanwhile, ESFCF provides an in-depth understanding concerning image content through the high-level semantic information, which helps the model to establish connections between local and global features. In all, ESFCF not only allows the model to extract global structures, but also reducing the loss of local information.

IV. Experiments and Analysis

To validate the effectiveness of the proposed LGC-YOLO method, this section conducts extensive quantitative and qualitative experiments on NWPU VHR-10 [48], Visdrone 2019 [49], and DOTA [50] datasets by comparing with other SOTA methods. In addition, various ablation studies are performed to verify the efficacy of each component, i.e., LGSFE, GOSII, and ESFCF, in the proposed method for object detection in remote sensing images.

A. Datasets and Evaluation Metrics

1) Datasets: Three remote sensing datasets of NWPU VHR-10, VisDrone 2019 and DOTA are employed in the experiments to validate the proposed method. The first NWPU VHR-10 dataset is collected by Northwestern Polytechnical University of China, which consists of 10 object categories with 800 HRRSIs in total. Among them, 650 images contain objects, while the remaining 150 are background images without any objects. The ten classes of objects include airplane (AE), ship (SP), storage tank (SK), baseball diamond (BD), tennis court (TT), basketball court (BT), ground track field (GD), harbor (HR), bridge (BE), and vehicle (VE). The image size ranges from 500 $\times$ 500 to 1100 $\times$ 1100 pixels.

The second VisDrone 2019 dataset consists of 10 209 high-resolution aerial images captured by drones across 14 different Chinese cities. These images are collected by Tianjin University of China and have dimensions of 2000 $\times$ 1500 pixels. Compared to NWPU VHR10, the VisDrone 2019 dataset poses great challenges for object detection methods, including viewpoint variations, severe occlusions, and dense clustering of small objects. This dataset encompasses ten different object categories: pedestrian (PN), people (PE), bicycle (BY), car (CR), van (VN), truck (TK), tricycle (TE), awning-tricycle (AT), bus (BS), and motor (MR).

The third DOTA dataset is a comprehensive collection of aerial and satellite images captured by different sensors and platforms, including Google Earth imagery, GF-2 satellite images, and aerial remote sensing images. It consists of 2806 images with resolutions from 800×800 to 4000×4000 pixels and includes 188,282 instances of objects belonging to 15 different categories: baseball diamond (BD), basketball court (BT), bridge (BE), harbor (HR), helicopter (HP), ground track field (GD), large-vehicle (LE), plane(PE), ship (SP), small-vehicle (SE), soccer-ball-field (SD), storage tank (SK), swimming-pool (SL), tennis court (TT), and roundabout (RT).

2) Evaluation Metrics: To quantitatively evaluate the performance of the proposed LGC-YOLO method, we employ the mean average precision (mAP) to evaluate object detection accuracy of different methods, average precision (AP) calculates the average area under the precision-recall curve for each object category, which takes into account the varied accuracies under different recall rates.

Precision measures the detection accuracy by the percentage of true positives versus all predicted targets, and the formula is shown in the following:

$$P = \frac{TP}{TP + FP}. \tag{12}$$

Recall measures the percentage of correctly detected objects, as shown in the following:

$$R = \frac{TP}{TP + FN}. \tag{13}$$

In above two equations, TP, TN, FP, and FN refer to True Positive, True Negative, False Positive, and False Negative, respectively.

To account for the overall object detection performance across different categories, the APs of all object categories are averaged to obtain mAP, as shown in the following:

$$\text{AP} = \int_0^1 PR(r)dr \tag{14}$$

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{15}$$

TABLE I
ABLATION STUDIES TO EVALUATE THE EFFECTIVENESS OF EACH PROPOSED COMPONENT OF LGSFE, GOSII, AND ESFCF IN OUR METHOD ON THE NWPU VHR-10 DATASET IN TERMS OF MAP, PARAMETERS(M), GFLOPS, AND FPS

Methods	AE	SP	SK	BD	TT	BT	GD	HR	BE	VE	mAP	Parameters	GFlops	FPS
Net-1	99.3	82.7	56.5	98.8	89.4	99.5	99.5	94.9	65.5	99.0	88.5	14.11	37.2	47.2
Net-2	99.3	81.6	79.8	99.0	89.6	99.5	99.5	95.5	99.5	99.5	94.3	14.22	37.7	46.9
Net-3	99.1	81.6	64.8	98.7	91.3	99.5	99.5	93.9	73.1	93.6	89.5	11.13	37.2	52.2
Net-4	97.8	86.9	57.4	99.0	91.6	99.5	99.5	96.1	85.0	99.3	91.2	14.12	37.3	46.7
Net-5	98.8	82.0	92.3	99.0	95.5	99.5	99.5	98.8	99.5	89.8	95.5	11.24	29.0	51.1
Net-6	98.3	89.3	93.5	99.3	98.9	99.6	99.6	96.6	99.5	98.8	97.3	11.25	29.0	51.1

where C denotes the number of categories, P and R are the precision and recall for the ith object category, respectively.

Additionally, the metrics of parameters (M), Giga floating-point operations per second (FLOPs), and frames per second (FPS) are adopted to measure model complexity and computational overheads, with smaller values being preferable.

B. *Experimental Settings*

To guarantee the fairness of experimental comparisons, the proposed LGC-YOLO method and baseline methods are all implemented using the PyTorch framework on a workstation equipped with an Intel Core i5-12400 processor (16 GB) and an NVIDIA 1660 GPU (6 GB). Our method is trained from scratch, while the baseline methods are fine-tuned on the NWPU VHR-10, VisDrone 2019, and DOTA datasets. These three datasets are divided into training, testing, and validation sets according to the ratios of 8:1:1, 13:1:3, and 8:1:1, respectively. The training process consists of 200 epochs with a consistent image size of 640×640. The batch size is set to 8, and the learning rate is initialized as 0.01 and gradually decreases at 0.0001.

C. *Ablation* Studies

To evaluate the impact of the newly designed LGSFE, GOSII, and ESFCF modules on object detection performance, ablation experiments are conducted on NWPU VHR-10 dataset. These experiments investigate various network combinations that are gradually augmented by incorporating the above three modules. The details are as follows:

1) NET-1: The CSPDarknet53 backbone, classical FPN, and Decoupled-Head detection head are adopted as the baseline network.
2) NET-2: NET-1+LGSFE. LGSFE is introduced to substitute all convolutions in NET -1.
3) NET-3: NET-1+GOSII. GOSII is utilized to replace all BottleNeck modules in the CSPDarknet53 backbone of NET-1.
4) NET-4: NET-1+ ESFCF. ESFCF is added into NET-1.
5) NET-5: NET-2+GOSII. GOSII is employed to replace all BottleNeck modules in the CSPDarknet53 backbone of NET-2.
6) NET-6: NET-5+ESFCF (i.e., LGC-YOLO). ESFCF is added into NET-5 to produce the proposed LGC-YOLO method.

The quantitative ablation experimental results of aforementioned six network combinations are reported in Table I. As observed, NET-1, i.e., CSPDarknet53+FPN+Decouple-Head, leads to very poor object detection performance with only 88.5% mAP and relatively higher parameter amounts and GFLOPs for the densely arranged small objects in HRRSI. By contrast, the proposed LGSFE, GOSII, and ESFCF modules progressively enhance object detection performance.

1) Effectiveness of LGSFE by Comparing NET-2 With NET-1*:* It is observed that LGSFE improves object detection accuracy by 5.8% mAP at the cost of slight increase of parameter numbers and computational overheads as well as decreasing the inference speed of FPS. The reason mainly lies in that LGSFE enhances the ability of network to capture long-range information by extracting local and global features. The advantage of LGSFE is that it improves feature representation capabilities by combining and aligning the local and global spatial features through receptive field attention mechanism.

2) Effectiveness of GOSII by Comparing NET-3 with NET-1*:* NET-1 utilizes the traditional BottleNeck module in CSPDarknet53 backbone, while NET-3 employs the proposed GOSII module to enhance the understanding of contextual information by optimizing spatial features. GOSII performs well in optimizing feature extraction based on SRAConv by incorporating SRU and SimAM into C2f framework to reduce complex or simple background interferences and spatial feature redundancy as well as improving the efficiency of information transmission and reducing model complexity.

3) Effectiveness of ESFCF by Comparing NET-4 With NET-1*:* Although ESFCF marginally increases the parameter numbers and computational overheads by 0.01 and 0.1 as well as decreasing the inference speed of FPS by 0.51, the mAP greatly increases by 2.7% due to the incorporation of shallow local detailed features via CA block in ESFCF module. It effectively supplements the detailed local information especially with large variations to abstract features for subsequent feature aggregation operation, especially for small objects that are often overlooked by deeper layers.

4) Effectiveness of LGSFE+GOSII by Comparing NET-5 With NET-2 or NET-3*:* By simultaneously introducing LGSFE and GOSII into NET-1, it increases mAP by 1.2% and 6.0% compared to NET-2 and NET-3, respectively. Meanwhile, it also surpasses them in terms of parameters, GFLOPs, and FPS except for a negligible increasing of parameters and decreasing of FPS compared to NET-3. These results indicate that both modules collaboratively enhance the small object detection performance, especially for SK and BE that are often affected by background interference and geometric deformation.

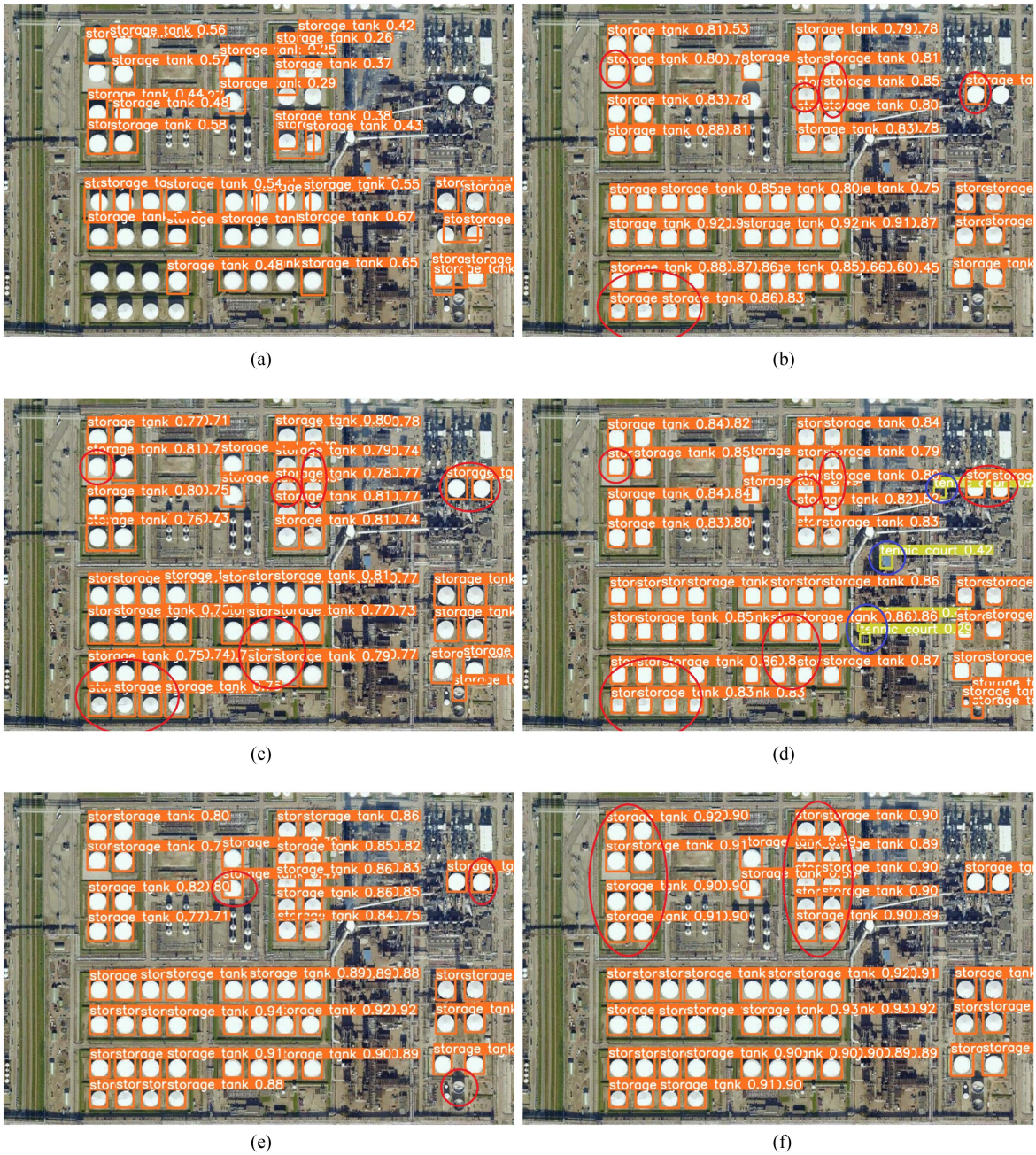

Fig. 8. Ablation experiments qualitatively validate the effectiveness of each component in our method by visual detection results on a remote sensing image with densely arranged small SK objects in complex backgrounds from NWPU VHR-10 dataset. (a) NET-1 (CSPDarknet53 + FPN + Decoupled-Head). (b) NET-2 (NET-1 + LGSFE). (c) NET-3 (NET-1 + GOSII). (d) NET-4 (NET-1 + ESFCF). (e) NET-5 (NET-1 + LGSFE + GOSII). (f) NET-6 (NET-1 + LGSFE + GOSII + ESFCF, i.e., our proposed LGC-YOLO method). The advantages of each module are marked by red circles., (a) NET-1. (b) NET-2. (c) NET-3. (d) NET-4. (e) NET-5. (f) NET-2.

5) Effectiveness of LGSFE+GOSII+ESFCF by Comparing NET-6 With NET-5*:* After integrating all the proposed three modules, LGC-YOLO further increases mAP by 1.8% with similar parameters, GFLOPs, and FPS when compared to NET-5. Thus, the proposed LGC-YOLO strikes a better balance between detection accuracy and model complexity as well as inference speed in detecting densely arranged small objects from HRRSIs with complex backgrounds.

On the whole, NET-6, i.e., the proposed LGC-YOLO method, outperforms all other network combinations in terms of mAP with small number of parameters and GFlops as well as high inference speed of FPS.

Fig. 8 qualitatively validates the effect of each component in our method by visually comparing the object detection results on a complex HRRSI from NWPU VHR-10 dataset. As observed from Fig. 8(a), NET-1 only detects some of the densely arranged

TABLE II
QUANTITATIVE EXPERIMENTAL COMPARISONS OF OBJECT DETECTION RESULTS ON THE NWPU VHR-10 DATASET BETWEEN OUR METHOD AND ELEVEN SOTA METHODS OF FASTER R-CNN (FR-N)[51], TINY-YOLOV4(T-V4)[52], SWIN-TRANSFORMER-TINY(S-TY)[53], YOLOV5S(Y-5 S), YOLOX-S(Y-S)[54], YOLOV7(Y-V7), YOLOV8S(Y-8 S), RT-DETR[55](RT-R), YOLOV9S-C[56](Y-9 S), YOLO-WORLD(Y-WD)[57] AND SWIN-DETR (S-R) [31] IN TERMS OF MAP, PARAMETERS(M), GFLOPS AND FPS; HERE, EACH METHOD IS ABBREVIATED IN A BRACKET DUE TO SPACE LIMITATIONS

Methods	FR-N	T-v4	S-Ty	Y-5s	Y-S	Y-v7	Y-8s	RT-R	Y-9s	Y-Wd	S-R	Ours
AE	99.3	98.6	99.5	97.7	96.6	98.8	99.5	92.5	99.3	62.6	99.2	98.3
SP	87.5	84.8	89.2	86.4	92.1	93.8	91.1	74.0	87.5	64.6	81.2	89.3
SK	59.1	86.7	97.9	96.0	97.2	96.3	80.2	90.3	85.2	70.9	76.3	93.5
BD	98.6	97.8	98.2	98.9	94.8	98.9	99.3	99.3	99.3	65.2	99.3	99.3
TT	87.3	78.2	97.4	87.6	98.5	99.3	99.3	54.3	89.4	66.2	87.5	98.9
BT	90.1	99.5	85.6	99.5	98.5	99.1	99.5	33.8	99.5	59.4	99.5	99.5
GD	95.4	99.5	99.5	99.5	98.5	90.5	99.5	99.2	99.5	60.3	99.5	99.5
HR	99.3	95.8	92.4	97.6	93.4	99.0	96.8	90.2	89.9	67.2	95.6	99.5
BE	87.9	99.5	80.7	81.0	89.0	93.2	97.2	99.5	99.5	63.5	83.7	96.8
VE	82.4	89.7	86.1	98.3	96.6	84.0	99.0	82.0	98.3	58.6	99.3	99.5
mAP	88.7	93.0	92.6	94.3	95.5	95.3	96.1	82.5	94.7	63.9	92.1	97.3
Parameters	137.10	62.18	34.24	7.05	8.96	37.21	11.13	32.83	13.14	4.05	62.8	11.3
GFlops	370.2	15.4	44.5	16.0	26.9	105.2	28.5	110.0	60.7	15.9	50.1	29.0
FPS	5.2	82.4	22.1	110.0	93.3	132.3	80.2	75.2	150.2	67.4	17.6	51.1

SK objects. By introducing LGSFE, NET-2 effectively extracts and fuses the aligned local and global information, which favors in detecting more SK objects compared to NET-1, as shown in the four red circles of Fig. 8(b). Similarly, NET-3 enhances the detection performance for the densely arranged SK objects by integrating GOSII to alleviate the spatial feature redundancy and capture contextual information, as observed by the red circle in the upper-right corner of Fig. 8(c). In fact, GOSII performs well in this remote sensing scene, which is relatively simple or medium complex HRRSI compared to other images in VisDrone 2019 or DOTA datasets. Meanwhile, NET-4 successfully detects all SK objects by incorporating ESFCF that combines shallow local features and texture information, especially for those with large local variations. However, some buildings and swimming pools are mistakenly identified as tennis court due to excessive feature extraction from this HRRSI, as shown in the three blue circles of Fig. 8(d). Furthermore, by simultaneously introducing LGSFE and GOSII modules, the detection performance is further enhanced for the densely arranged SK objects in this HRRSI with similar backgrounds and target occlusions when compared to NET-2 or NET-3. However, NET-5 may fail to detect the highly cluttered small objects in other extremely complex HRRSI with multiple target occlusions and large-scale background interferences from Visdrone 2019 or DOTA datasets. Finally, by further introducing ESFCF module that combines shallow edge and texture information, all the SK objects are successfully detected with higher detection accuracy, as shown in Fig. 8(e). These visual object detection results validate the effects of each component in identifying the densely arranged small objects in HRRSIs with complex backgrounds.

D. Experimental Results and Comparative Analyses

To quantitatively and qualitatively evaluate the effectiveness of the proposed LGC-YOLO method in RSOD task, we compare our method with eleven SOTA methods, including the two-stage method of Faster R-CNN [51], and the single-stage methods of Tiny-YOLOv4 [52], Swin-Transformer-Tiny [53], YOLOv5s, YOLOX-S [54], YOLOv7, YOLOv8s, RT-DETR [55], YOLOv9s-c [56], YOLO-World [57], and Swin-DETR [31] on NWPU VHR-10, VisDrone 2019, and DOTA datasets.

1) Comparisons on the NWPU VHR-10 Dataset: We initially perform a quantitative comparison between our proposed LGC-YOLO method and eleven advanced methods, as indicated in Table II, to assess the object detection results on NWPU VHR-10 in terms of AP, mAP, parameters, GFlops, and FPS. As observed, the proposed LGC-YOLO achieves an mAP of 97.3, which consistently outperforms the mainstream two-stage Faster R-CNN method and one-stage YOLO series methods (e.g., Tiny-YOLOv4, YOLOv5s, YOLOX-S, YOLOv7, YOLOv8s, YOLOv9s-c, and YOLO-World) as well as transformer-based methods (e.g., Swin-Transformer-Tiny, RT-DETR, and Swin-DETR). Among them, YOLO-World achieves the lowest mAP of 63.9% since it aims to improve the real-world open-vocabulary detection performance without considering the detailed features of small-sized objects in HRRSI that tends to be easily lost by many methods. Although Swin-DETR improves the mAP up to 92.1% by combining CNN and Swin-Transformer in backbone network, which significantly increases the parameters and decreases inference speed of FPS. In terms of model size, our method is larger or comparable to those of YOLOv5s, YOLOX-S, YOLOv8, and YOLO-World, but the detection accuracy of our method is superior to theirs by 3.0, 1.8, 1.2, and 33.4 percentage point, respectively. This can be primarily attributed to the introduction of LGSFE and ESFCF modules in our method, which effectively extract local-global information and small-object features, thereby improving the detection accuracy of densely arranged small objects in HRRSI. As for the computational overhead, the GFOPs of our method is 29.0, which is superior to those of Faster R-CNN, Swin-Transformer-Tiny, YOLOv7, RT-DETR, YOLOv9s-c, and Swin-DETR. The reason mainly lies in that the SRU block utilized in the proposed

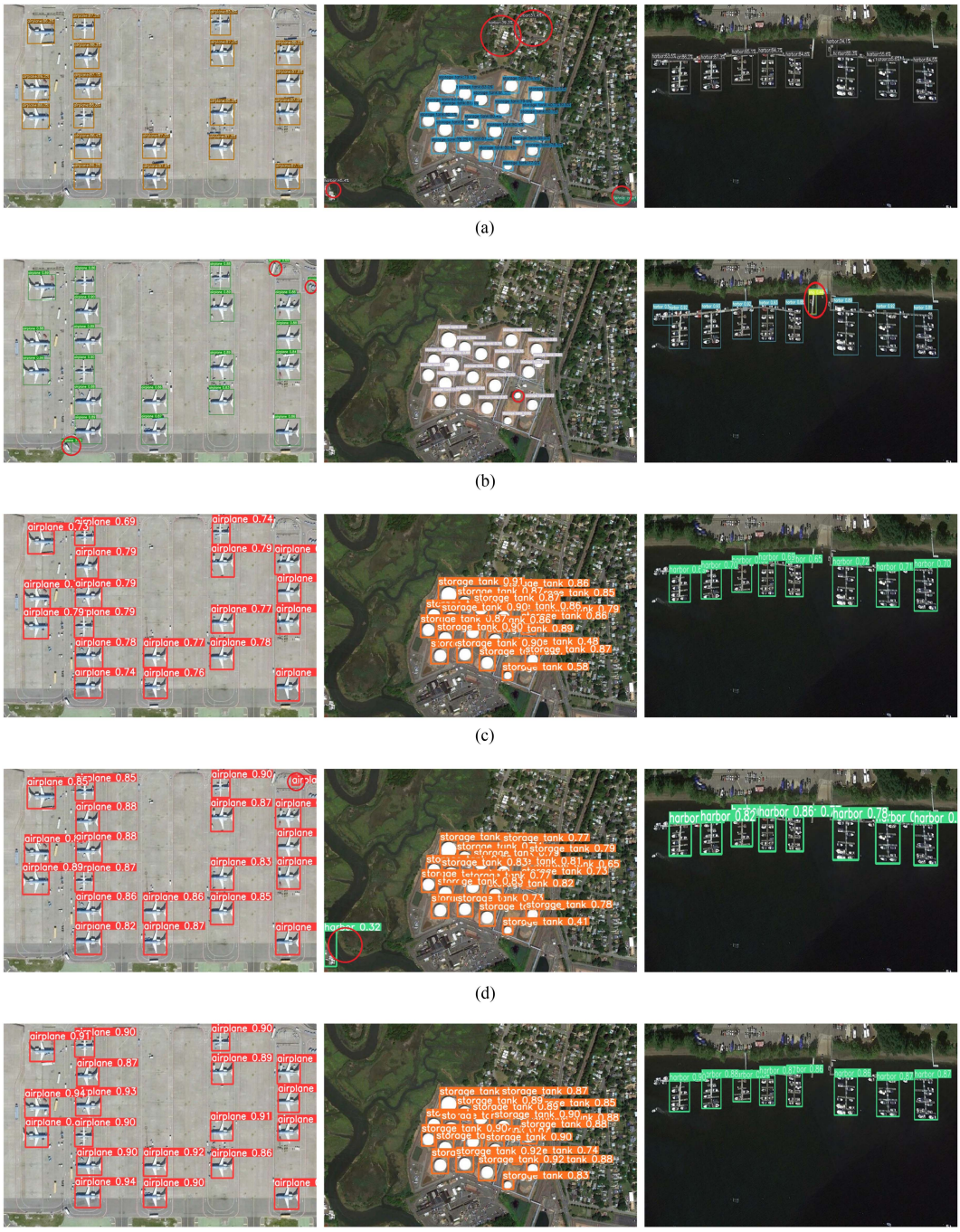

Fig. 9. Object detection results by our method and four state-of-the-art methods on three sample remote sensing images from NWPU VHR-10 dataset, which consists of densely arranged small objects in complex backgrounds. The first four rows are the object detection results by (a) YOLOX-S, (b) YOLOv7, (c) YOLOv8s, and (d) YOLOv9s-c, respectively. The last row presents the object detection results by (e) our proposed LGC-YOLO method. Note that the missed or false detections in different methods are marked by red circles. (a) YOLOX-S. (b) YOLOv7. (c) YOLOv8s. (d) YOLOv9s-c. (e) Our proposed LGC-YOLO method.

GOSII module suppresses spatial redundant features and reduces model complexity as well as computational burden. Our method achieves a relatively higher inference speed of FPS at 51.1, which indicates a great potential for actual deployment in real applications. On the whole, our method strikes a better balance in terms of mAP, parameters, GFLOPs, and inference speed of FPS.

To qualitatively evaluate the proposed LGC-YOLO method, Fig. 9 presents the visual object detection results on three remote sensing images from NWPU VHR-10 dataset by comparing our method with the top four methods of YOLOX-S, YOLOv7, YOLOv8s, and YOLOv9s-c in Table II. These images are characterized by dense small objects and complex backgrounds. As observed in the red circles, the counterpart YOLOv7

TABLE III
QUANTITATIVE EXPERIMENTAL COMPARISONS OF OBJECT DETECTION RESULTS ON THE VISDRONE 2019 DATASET BETWEEN OUR METHOD AND ELEVEN SOTA METHODS OF FASTER R-CNN (FR-N)[51], TINY-YOLOV4(T-V4)[52], SWIN-TRANSFORMER-TINY(S-TY)[53], YOLOV5S(Y-5 S), YOLOX-S(Y-S)[54], YOLOV7(Y-V7), YOLOV8S(Y-8 S), RT-DETR[55](RT-R), YOLOV9S-C[56](Y-9 S), YOLO-WORLD(Y-WD)[57] AND SWIN-DETR (S-R)[31] IN TERMS OF MAP, PARAMETERS(M), GFLOPS AND FPS; HERE, EACH METHOD IS ABBREVIATED IN A BRACKET DUE TO SPACE LIMITATIONS

Methods	FR-N	T-v4	S-Ty	Y-5s	Y-S	Y-v7	Y-8s	RT-R	Y-9s	Y-Wd	S-R	Ours
PN	6.0	39.7	45.1	43.6	44.0	45.2	40.6	41.5	41.7	26.3	33.5	42.6
PE	1.0	33.7	35.8	34.6	35.1	35.4	31.4	31.3	32.0	12.7	24.0	33.0
BY	1.3	11.6	14.4	13.7	13.7	15.1	11.3	8.4	11.6	3.8	10.8	11.0
CR	38.1	73.9	76.1	75.6	75.9	76.4	78.9	73.5	79.4	63.4	73.2	79.6
VN	30.6	37.7	38.5	37.4	37.9	39.6	44.7	31.0	45.0	35.2	41.2	36.2
TK	32.2	30.8	35.7	34.7	35.1	38.2	33.6	28.5	35.3	32.5	41.2	36.2
TE	7.8	21.9	22.1	20.6	22.1	22.5	26.2	12.8	27.0	12.9	21.6	27.9
AT	6.5	12.1	12.0	11.6	11.3	13.0	16.4	8.1	16.6	12.2	19.5	15.6
BS	49.5	43.5	45.6	42.5	44.2	45.5	54.7	32.9	57.9	43.9	59.3	56.9
MR	4.2	39.2	42.4	41.7	42.2	42.4	42.4	38.1	43.3	22.0	35.7	44.4
mAP	17.7	34.4	36.8	35.6	36.2	37.3	38.0	30.6	39.0	26.5	36.4	39.2
Parameters	137.10	62.18	34.24	7.06	8.96	37.21	11.13	32.83	13.14	4.05	62.8	11.25
GFlops	370.2	15.4	44.5	16.0	26.9	105.2	28.5	110.0	60.7	15.9	50.1	29.0
FPS	4.9	81.9	21.7	108.2	92.5	130.6	79.2	130.3	79.5	148.3	66.4	50.2

misclassifies some buildings in the upper-right corner of the image in the first column as AE, while the buildings in the second column by YOLOX-S as HR, demonstrating poor detection performance for AE and SK objects. By contrast, the proposed LGC-YOLO successfully avoids these misclassifications and accurately detects all AE and SK objects in these images. This can be primarily attributed to the introduction of LGSFE and ESFCF modules. Specifically, LGSFE captures the spatial local and global information through multibranch feature extraction and models the capability of long-term dependence between dense objects, which favors in distinguishing the densely distributed SK objects. ESFCF effectively extracts rich edge and texture information from shallow layers, especially for the small objects that are often overlooked by deeper layers. In addition, the cluttered marine scenes pose significant challenges for object detection. YOLOX-S and YOLOv7 mistakenly identify some buildings as SP and HR in these marine scenes, as observed in the red circles of the second and third columns, respectively. On the other hand, although YOLOv8s and YOLOv9s-c successfully detect all objects, the classification and localization results are not satisfactory due to spatial feature redundancy, which is effectively alleviated by the proposed GOSII module for its ability to suppress redundant features. These qualitative comparisons further demonstrate the superiority of our method in detecting the densely arranged small objects from remote sensing images with complex backgrounds.

2) Comparisons on the VisDrone 2019 Dataset: To further validate the effectiveness of our method, we conduct more comparative experiments on the challenging VisDrone 2019 dataset, which presents intricate urban remote sensing scenes. Table III reports the quantitative detection results by our LGC-YOLO and other SOTA methods. As observed, many methods struggle to achieve high detection accuracy for individual object categories. For these challenging scenarios, our method consistently achieves the highest mAP value of 39.2% . Compared with the latest YOLOv9s-c and Swin-DETR approaches, our method increases mAP by 0.2% and 2.8% with a remarkable reduction of parameters and GFlops, respectively. Although our method is not optimal in terms of model complexity, the parameters of our method are similar to the fourth best of 11.13 M, while the GFlops similar to the fourth best of 26.9. In addition, our method achieves a high inference speed of FPS at 50.2. By comprehensively considering these evaluation metrics, the proposed LGC-YOLO method strikes a superior balance on this challenging dataset.

Fig. 10 presents the visual object detection results on three remote sensing images from VisDrone 2019 by qualitatively comparing with the top four methods of SwinTransformer-Tiny, YOLOv7, YOLOv8s, and YOLOv9s-c in Table III. These images depict the densely populated or traffic-congested areas in complex urban environments, and are characterized by highly cluttered complex backgrounds and dense arrangement of small objects. In particular, the obstructions such as nighttime lighting severely affect the object detection accuracy, which results in significant feature misalignment and missed detections. For example, in the first column, all the counterpart methods fail to detect the blue TE, as shown in the red circle on the left part of the image, whereas our method successfully detects the missed TE object. In addition, as shown in the red circles, Swin-Transformer-Tiny misclassifies the traffic sign as PN in the third column, YOLOv7 misclassifies PN and PE as BY in the second column, YOLOv8s falsely detects the pole as PN in the third column, YOLOv9s-c omits MR in the lower-right corner of the first column. However, our method effectively avoids these missed or false detections and accurately recognizes these objects. In general, for the densely arranged small objects, our method exhibits excellent detection performance compared to other methods.

3) Comparisons on the DOTA Dataset: To evaluate the robustness of our method on different types of remote sensing images, we quantitatively conduct more comparative experiments on the large-scale DOTA dataset that includes Google

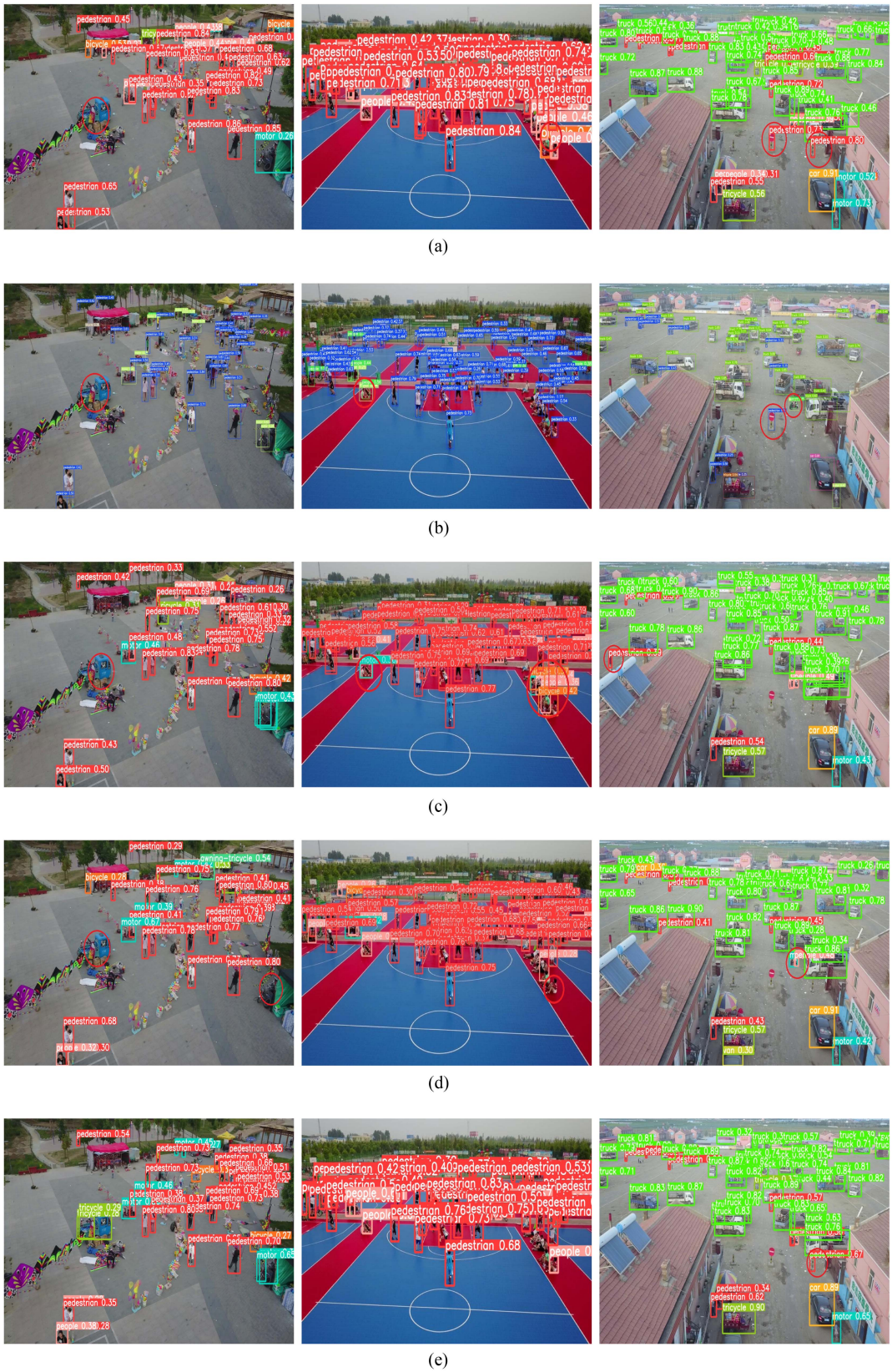

Fig. 10. Object detection results by our method and four state-of-the-art methods on three sample remote sensing images from VisDrone 2019 dataset, which are more challenging compared to NWPU VHR-10 dataset due to the complex background interferences and cluttered objects. The first four rows are the object detection results by (a) Swin-Transformer-Tiny, (b) YOLOv7, (c) YOLOv8s, and (d) YOLOv9s-c, respectively. The last row presents the object detection results by (e) our proposed LGC-YOLO method. Note that the missed or false detections in different methods are marked by red circles. (a) Swin-Transformer-Tiny. (b) YOLOv7. (c) YOLOv8s. (d) YOLOv9s-c. (e) Our proposed LGC-YOLO method.

TABLE IV
QUANTITATIVE EXPERIMENTAL COMPARISONS OF OBJECT DETECTION RESULTS ON THE DOTA DATASET BETWEEN OUR METHOD AND ELEVEN SOTA METHODS OF FASTER R-CNN (FR-N)[51], TINY-YOLOV4(T-V4)[52], SWIN-TRANSFORMER-TINY(S-TY)[53], YOLOV5S(Y-5 S), YOLOX-S(Y-S)[54], YOLOV7(Y-V7), YOLOV8S(Y-8 S), RT-DETR[55](RT-R), YOLOV9S-C[56](Y-9 S), YOLO-WORLD(Y-WD)[57] AND SWIN-DETR (S-R)[31] IN TERMS OF MAP, PARAMETERS(M), GFLOPS AND FPS; HERE, EACH METHOD IS ABBREVIATED IN A BRACKET DUE TO SPACE LIMITATIONS

Methods	FR-N	T-v4	S-Ty	Y-5s	Y-S	Y-v7	Y-8s	RT-R	Y-9s	Y-Wd	S-R	Ours
BD	71.8	82.4	80.2	74.6	76.6	82.9	80.0	83.1	69.7	76.6	84.5	74.7
BT	71.9	79.2	83.9	84.5	90.8	85.6	64.0	84.7	85.9	79.5	85.3	85.6
BE	47.9	49.4	43.5	43.9	45.1	48.6	63.6	50.9	94.7	45.1	59.7	95.2
HR	67.3	62.0	67.6	66.7	62.3	65.6	76.7	66.9	63.1	69.1	72.7	63.5
HP	64.2	62.2	52.5	52.2	60.1	68.0	73.6	53.9	47.4	60.1	71.1	53.3
GD	54.3	73.5	63.4	70.2	66.8	65.2	55.6	67.3	57.1	66.8	73.9	61.8
LE	78.0	63.5	70.7	67.3	76.8	70.1	79.1	80.4	72.4	76.8	76.4	77.1
PE	77.4	87.8	88.9	88.9	88.7	89.8	86.2	88.8	77.7	88.7	87.4	81.2
SP	82.1	76.6	84.9	79.8	79.7	70.2	83.9	86.7	86.1	79.7	86.5	87.6
SE	62.8	71.1	73.5	67.3	67.0	69.5	65.9	76.2	75.1	67.0	77.6	81.8
SD	65.7	48.4	50.1	54.6	78.5	62.5	60.1	62.0	40.1	57.7	52.5	46.0
SK	61.4	73.3	84.1	78.5	79.5	83.4	79.6	83.2	92.9	78.5	84.3	93.7
SL	65.7	67.0	68.6	68.0	72.0	67.2	79.5	70.6	82.3	73.1	69.6	84.6
TT	88.5	90.9	90.1	90.9	79.7	90.5	85.7	90.8	88.0	90.8	90.8	89.5
RT	54.2	60.9	58.4	62.6	57.7	63.9	66.6	61.4	71.5	62.3	65.1	71.4
mAP	67.5	69.9	70.7	70.0	72.1	72.2	73.3	73.8	73.6	71.5	75.8	76.5
Parameters	137.1	62.18	34.24	7.06	8.96	37.21	11.13	32.83	13.14	4.05	62.8	11.25
GFlops	370.2	15.4	44.5	16.0	26.9	105.2	28.5	110.0	60.7	15.9	50.1	29.0
FPS	4.9	81.3	21.5	107.9	92.1	129.9	79.1	129.7	79.1	147.6	66.1	49.9

Earth imagery, GF-2 satellite images, and aerial remote sensing images. The quantitative object detection results are presented in Table IV. As observed, our method achieves the highest mAP of 76.2% . Compared to the Swin-DETR method that ranks second, our method increases mAP by 0.7% . Meanwhile, the parameters of our method are only one-sixth of the Swin-DETR method. For the densely arranged small objects of SE and SK, the proposed method not only achieves the highest AP values but also surpasses Swin-DETR by 4.2% that ranks second for SE, and YOLOv9s-c by 0.8% that ranks second for SK. Although the mAP for the methods of Tiny-YOLOv4, YOLOv5s, Swin-Transformer-Tiny, YOLO-World, YOLOX-S, YOLOv7, and YOLOv8 steadily increases, our proposed method consistently outperforms them. In addition, our method also achieves better performance in terms of GFLOPs and FPS compared to other methods. On the whole, these experimental comparisons demonstrate that our method has strong robustness to different kinds of remote sensing images.

To visually demonstrate the detection performance of the proposed LGC-YOLO method, Fig. 11 presents some object detection results on randomly selected fifteen images for the object categories from DOTA dataset. As observed, LGC-YOLO exhibits excellent performance in detecting dense objects, such as LE, SP, SE, SK, and so on. This can be primarily attributed to the multibranch structure of the proposed LGSFE module, which captures local and global spatial features through a receptive field attention mechanism. LGSFE effectively enhances the model's ability to model the capability of long-term dependence for capturing the features of dense objects and mitigates the potential inconsistencies between extracted features and inherent object ones. Meanwhile, it also produces good detection results for the objects with complex backgrounds, such as BD, BT, BE, HR, GD, SD, SL, TT, RT, and so on. GOSII suppresses the complex background interference and redundant noise through SRU and SimAM blocks, and further optimizes the spatial features of objects to enhance the understanding of contextual information. For the small objects like PE and HP in DOTA dataset, ESFCF ensures that the high-level features are sufficiently supplemented with rich edge and texture details, which effectively avoids the loss of edge features and further facilitates accurate object localization. Thus, through simultaneously integrating these three modules, the proposed LGC-YOLO method achieves satisfactory detection performance for the densely arranged small objects from HRRSIs with complex backgrounds.

V. DISCUSSION

The aforementioned experiments validate the effectiveness of the proposed LGC-YOLO method in RSOD, particularly in achieving a better balance between detection accuracy and model complexity for detecting the densely arranged small objects from complex environments. This is primarily attributed to the three proposed LGSFE, GOSII, and ESFCF modules. First, LGSFE employs a multibranch structure combined with RFA-Conv to extract local-global information, thereby effectively capturing richer spatial features. Second, GOSII introduces SRU to suppress redundant spatial features and retain effective ones, and utilizes the C2f framework to enhance feature interaction, thus obtaining abundant contextual information. Lastly, ESFCF leverages an attention mechanism to enhance shallow feature information and supplement abstract features, thereby acquiring crucial edge details and rich semantic features. Experimental

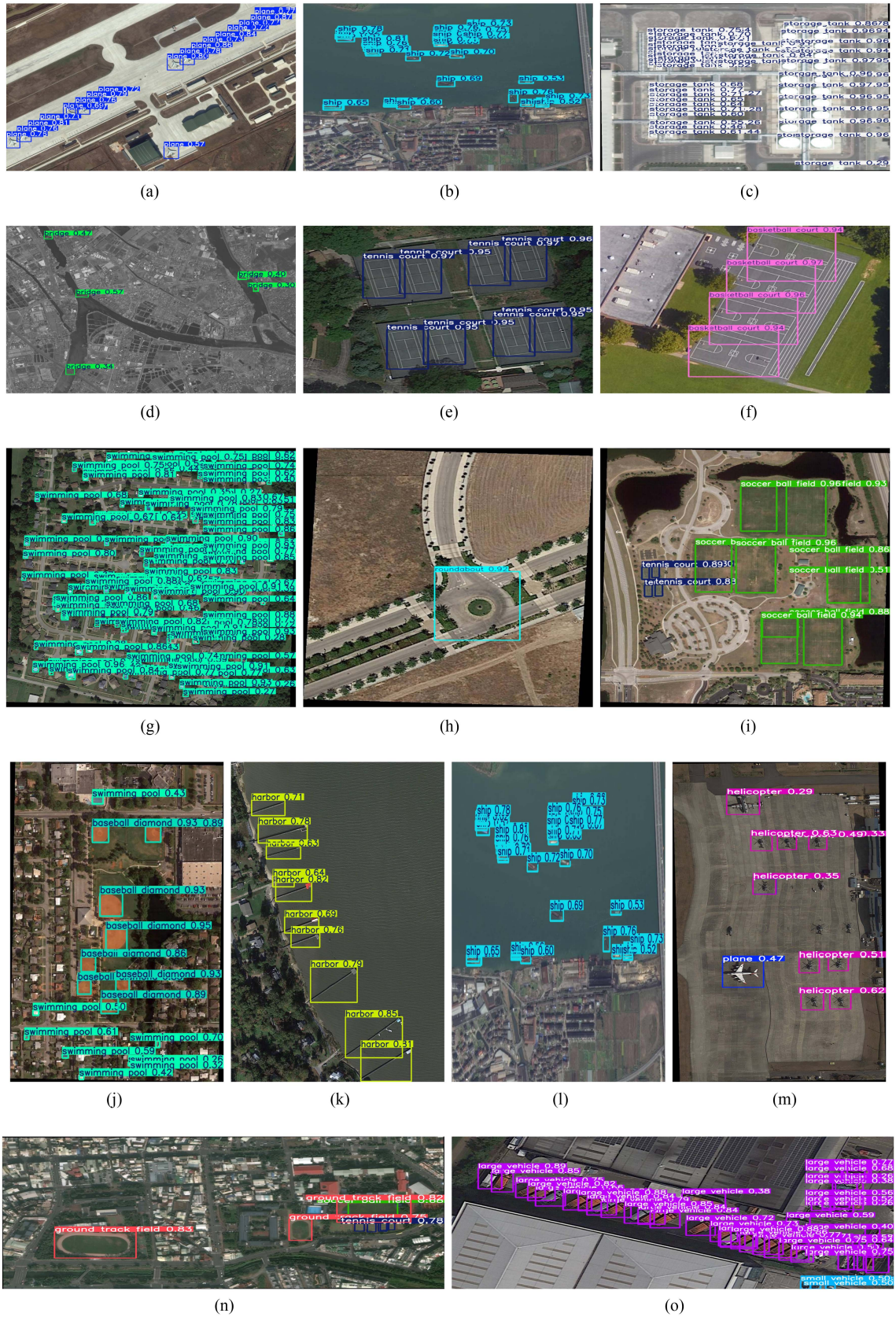

Fig. 11. Object detection results by our method on remote sensing images belonging to 15 categories of DOTA dataset, which are more challenging due to the complex background interferences, cluttered objects, and large-scale remote sensing images. (a) PE. (b) SP. (c) SK. (d) BE. (e) TT. (f) BT. (g) SL. (h) RT. (i) SD. (j) BD. (k) HR. (l) SE. (m) HP. (n) GD. (o) LE.

results demonstrate that the proposed LGC-YOLO achieves an average increase of 0.7% in terms of mAP across NWPU VHR-10, VisDrone 2019, and DOTA datasets compared with other state-of-the-art RSOD methods.

However, although the proposed LGC-YOLO method significantly improves the accuracy of feature extraction and object detection, it still has some limitations. For the case of extremely complex HRRSI with multiple target occlusions and large-scale background interferences as well as geometric distortions, it is difficult to effectively extract the features of highly cluttered small objects. For example, "SP" in the NWPU VHR-10 dataset, "BY" and "AT" in the VisDrone 2019 dataset, as well as "SD" and "GD" in the DOTA dataset, bring great challenges for our method to effectively distinguish important features from extremely complex backgrounds. In addition, "PE" and "PN" in the VisDrone 2019 dataset, as well as "HP" and "HR" in the DOTA dataset, often represent significant geometric distortions, leading to difficulties in recognizing and integrating effective features. The detection accuracies for these kinds of objects that are often with large-scale geometric distortions or surrounded by cluttered backgrounds are significantly lower than others. In the future, we will explore the object priors within YOLO framework to further improve the detection accuracy of densely arranged small objects in extremely complex remote sensing images.

VI. Conclusion

In this article, a novel single-stage object detection method, i.e., LGC-YOLO, has been proposed to identify the densely arranged small objects in remote sensing images with complex backgrounds. The core of this method lies in the accurate feature extraction to address the challenges of inconsistency between extracted object features and ground truth ones. LGC-YOLO primarily consists of three key modules of LGSFE, GOSII, and ESFCF. First, LGSFE enhances the feature perception capability through a multibranch feature extraction and fusion strategy, where the RFAConv mechanism is embedded to capture and fuse important information. LGSFE comprehensively extracts object features through different branches, and effectively alleviates the problem of inconsistency between extracted features and actual object ones. Second, GOSII utilizes SRU and C2f for feature cross-reconstruction to reduce redundant spatial features and increase information flow efficiently, as well as SimAM for better focus on crucial features to further improve the model ability to identify and locate objects. Finally, ESFCF leverages CA to learn detailed features and semantic information from shallow and deep layers, aiming to address the challenges of insufficient feature extraction from small objects.

References

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G. -S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[2] Y. Zhao, J. Liang, S. Huang, and P. Huang, "Hierarchical deep features progressive aggregation for remote sensing images scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*,vol. 17, pp. 9442–9450, 2024.

[3] C. X. Zhang, M. Zhang, and Jiang, "Research on parallel detection technology of remote sensing object based on deep learning," *IEEE Access*, vol. 11, pp. 114146–114154, 2023.

[4] L. Jiang et al., "MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 5015214.

[5] X. Wang A. Wang J. Yi Y. Song, and A. Chehri, "Small object detection based on deep learning for remote sensing: A comprehensive review," *Remote Sens.*, vol. 15, 2023, Art. no. 3265.

[6] S. Zhu, Y. Song, Y. Zhang, and Y. Zhang, "ECFNet: A siamese network with fewer FPs and fewer FNs for change detection of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6001005.

[7] J. Wang et al., "SSCFNet: A spatial-spectral cross fusion network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4000–4012, 2023.

[8] J. Zheng et al., "MDESNet: Multitask difference-enhanced siamese network for building change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 14, 2022, Art. no. 3775.

[9] Q. Wu, Y. Li, W. Huang, Q. Chen, and Y. Wu, "C3TB-YOLOv5: Integrated YOLOv5 with transformer for object detection in high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 45, no. 8, pp. 2622–2650, 2024.

[10] W. Xue, J. Qi, G. Shao, Z. Xiao, Y. Zhang, and P. Zhong, "Low-rank approximation and multiple sparse constraint modeling for infrared low-flying fixed-wing UAV detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4150–4166, 2021.

[11] M. Knura, F. Kluger, M. Zahtila, J. Schiewe, B. Rosenhahn, and D. Burghardt, "Using object detection on social media images for urban bicycle infrastructure planning: A case study of dresden," *ISPRS Int. J. Geo- Inf.*, vol. 10, 2021, Art. no. 733.

[12] J.-I. Watanabe, Y. Shao, and N. Miura, "Underwater and airborne monitoring of marine ecosystems and debris," *J. Appl. Remote Sens.*, vol. 13, no. 4, Oct. 2019, Art. no. 044509.

[13] S. Xu, H. Tang, J. Li, L. Wang, X. Zhang, and H. Gao, "A YOLOW algorithm of water-crossing object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16901–16911.

[14] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 62429–62442.

[15] Z. Li et al., "Deep learning-based object detection techniques for remote sensing images: A survey," *Remote Sens.*, vol. 14, 2022, Art. no. 2385.

[16] C. Fasana, S. Pasini, F. Milani, and P. Fraternali, "Weakly supervised object detection for remote sensing images: A survey," *Remote Sens.*, vol. 14, 2022, Art. no. 5362.

[17] Z. He, Z. Zhang, M. Guo, L. Wu, and Y. Huang, "Adaptive unsupervised-shadow-detection approach for remote-sensing image based on multichannel features," *Remote Sens.*, vol. 14, 2022, Art. no. 2756.

[18] J. Kang, J. Kwon, and H. Kim, "Accelerating a two-stage object detector for high quality in-orbit remote sensing," in *Proc. SPIE*, vol. 12267, 2022, Art. no. 122670B.

[19] P. Gao, T. Tian, T. Zhao, L. Li, N. Zhang, and J. Tian, "Double FCOS: A two-stage model utilizing FCOS for vehicle detection in various remote sensing scenes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4730–4743, 2022.

[20] Z. Zhao, P. Tang, L. Zhao, and Z. Zhang, "Few-shot object detection of remote sensing images via two-stage fine-tuning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8021805.

[21] Y. Li, X. Pei, Q. Huang, L. Jiao, R. Shang, and N. Marturi, "Anchor-free single stage detector in remote sensing images based on multiscale dense path aggregation feature pyramid network," *IEEE Access*, vol. 8, pp. 63121–63133, 2020.

[22] Y. Yuan, Z. Li, and D. Ma, "Feature-aligned single-stage rotation object detection with continuous boundary," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5538011.

[23] Y. Li, C. Kong, L. Dai, and X. Chen, "Single-stage detector with dual feature alignment for remote sensing object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6503605.

[24] Z. Zhu, R. Zheng, G. Qi, S. Li, Y. Li, and X. Gao, "Small object detection method based on global multi-level perception and dynamic region aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 10011–10022, Oct. 2024.

[25] Y. Li, Z. Zhou, G. Qi, G. Hu, Z. Zhu, and X. Huang, "Remote sensing micro-object detection under global and local attention mechanism," *Remote Sens.*, vol. 16, 2024, Art. no. 644.

[26] D. Yu, R. Zhang, and S. Qin, "Cascade saliency attention network for object detection in remote sensing images," in *Proc. 25th Int. Conf. Pattern Recognit.*, Milan, Italy, 2021, pp. 217–223.
[27] T. Zhang, Y. Zhuang, G. Wang, H. Chen, L. Li, and J. Li, "A unified remote sensing object detector based on Fourier contour parametric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5611225.
[28] Z. Zhang et al., "Rotated YOLOv4 with attention-wise object detectors in aerial images," in *Proc. 4th Int. Conf. Robot Syst. Appl.*, 2021, pp. 21–23.
[29] Y. Li, H. Mao, R. Liu, X. Pei, L. Jiao, and R. Shang, "A lightweight keypoint-based oriented object detection of remote sensing images," *Remote Sens.*, vol. 13, 2021, Art. no. 2459.
[30] C. Zhang, K. -M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.
[31] Y. Xie, X. Ma, and Q. Zhao, "Research on target detection network based on improved Swin-DETR," in *Proc. 4th Int. Conf. Big Data, Artif. Intell. Internet of Things Eng.*, 2023, pp. 324–328.
[32] L. Zhang, Y. Wang, and Y. Huo, "Object detection in high-resolution remote sensing images based on a hard-example-mining network," *IEEE Trans. Geosci. Remote Sens.*, no. 10, vol. 59, pp. 8768–8780, Oct. 2021.
[33] Y. Liu et al., "YOLO-SSP: An object detection model based on pyramid spatial attention and improved downsampling strategy for remote sensing images," *Vis. Comput.*, vol. 41, no. 3, pp. 1467–1484, 2025.
[34] G. Qi et al., "Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion," *Remote Sens.*, vol. 14, 2022, Art. no. 420.
[35] Y. Ma et al., "Aircraft-LBDet: Multi-task aircraft detection with landmark and bounding box detection," *Remote Sens.*, vol. 15, 2023, Art. no. 2485.
[36] T. Zhang et al., "Controllable generative knowledge-driven few-shot object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5612319.
[37] F. Cao et al., "An efficient object detection algorithm based on improved YOLOv5 for high-spatial-resolution remote sensing images," *Remote Sens.*, vol. 15, 2023, Art. no. 3755.
[38] Q. Wu, Y. Wu, Y. Li, and W. Huang, "Improved YOLOv5s with coordinate attention for small and dense object detection from optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2543–2556, 2024.
[39] G. X. Cheng et al., "Towards large-scale small object detection: Survey and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 11, vol. 45, pp. 13467–13488, Nov. 2023.
[40] Z. Xin et al., "RFAConv: Innovating spatial attention and standard convolutional operation," 2023, *arXiv:2304.03198*.
[41] J. Li, Y. Wen, and L. He, "SCConv: Spatial and channel reconstruction convolution for feature redundancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 6153–6162.
[42] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 13708–13717.
[43] J. Terven, D.-M Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extraction*, vol. 5, pp. 1680–1716, 2023.
[44] L. Ouyang, L. Fang, and X. Ji, "Multigranularity self-attention network for fine-grained ship detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9722–9732, 2022.
[45] C. Liu, S. Zhang, M. Hu, and Q. Song, "Object detection in remote sensing images based on adaptive multi-scale feature fusion method," *Remote Sens.*, vol. 16, 2024, Art. no. 907.
[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis.*. Munich, Germany, 2018, vol. 11211, pp. 3–19.
[47] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 7464–7475.
[48] K. Li et al., "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
[49] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 213–226.
[50] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
[51] 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, pp. 1137–1149, Jun. 2017.
[52] E. Yildirim and T. Kavzoglu, "Ship detection in optical remote sensing images using YOLOv4 and tiny YOLOv4," in *Proc. 6th Int. Conf. Smart City Appl.*, Safranbolu, Turkey, 2021, vol. 393, pp. 913–924.
[53] Z. Y. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 9992–10002.
[54] B. Liu, J. Huang, S. Lin, Y. Yang, and Y. Qi, "Improved YOLOX-S abnormal condition detection for power transmission line corridors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Harbin, China, 2021, pp. 13–16.
[55] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16965–16974.
[56] C. Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 1–21.
[57] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-time open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 12345–12356.

Qinggang Wu received the Ph.D. degree in computer science from Dalian Maritime University, Dalian, China, in 2012.

He is currently an Associate Professor with the College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, China. His current research interests include remote sensing image processing, hyperspectral image classification, artificial intelligence, and deep learning.

Yang Li received the B.S. degree in Internet of Things engineering from the Henan University of Animal Husbandry and Economy, Zhengzhou, China, in 2022. He is currently working toward the M.E.s degree in computer technology with the Zhengzhou University of Light Industry, Zhengzhou.

His research interests include remote sensing object detection, machine learning, and deep learning.

Junru Yin received the Ph.D. degree in forest management from the Chinese Academy of Forestry, Beijing, China, in 2015. She is currently an Associate Professor with Zhengzhou University of Light Industry, Zhengzhou, China.

Her current research interests include hyperspectral image classification, pansharpening, and forest management.

Xiaotian You received the B.S. degree in computer science and technology in 2023 from the Zhengzhou University of Light Industry, Zhengzhou, China, where she is currently working toward the M.E.s degree in network and information security.

Her research interests include remote sensing object detection and deep learning.