HOFA-Net: A High-Order Feature Association Network for Dense Object Detection in Remote Sensing

YunPeng Xu, *Member, IEEE*, Xin Wu, *Senior Member, IEEE*, Li Wang, *Senior Member, IEEE*, Lianming Xu, Zhengyu Shao, and Aiguo Fei

***Abstract*—In the remote sensing field, deep learning-based methods have become mainstream for remote sensing image object detection in recent years. However, traditional methods, such as convolutional neural networks (CNNs), mainly ignore the dependencies between features, failing to capture the spatial relationships and relative positions of objects, which affects the detection performance of dense objects, especially small-size objects. To this end, a high-order feature association network (HOFA-Net) for dense object detection in remote sensing has been proposed to better capture the interdependencies between features of channel and spatial dimensions, yielding more distinguishable features. First, we employ CNNs to learn high-level but low-resolution features of objects. To capture feature interdependencies while retaining crucial information, we design a feature association module based on size adaptation nonlocal. This module partitions the low-resolution and high-level features into local regions and utilizes nonlocal residual connections to capture the local contextual information of objects. In addition, we introduce a high-order feature association (HFA) module designed to learn nonlinear feature correlations and interdependencies within the features. In addition, a covariance normalization acceleration strategy is introduced to accelerate computation. Experimental results on two public remote sensing datasets, including the DOTA dataset and the Tiny Person dataset, demonstrate the superiority and effectiveness of the proposed method through comparative experiments.**

***Index Terms*—Convolutional neural networks (CNNs), covariance normalization, high-order feature association, object detection, remote sensing.**

Manuscript received 28 May 2023; revised 15 September 2023; accepted 7 November 2023. Date of publication 28 November 2023; date of current version 15 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62101045, Grant 62171054, and Grant 62201071, in part by the National Key Research and Development Program of China under Grant 2020YFC1511801, in part by the Natural Science Foundation of Beijing Municipality under Grant L222041, in part by the Fundamental Research Funds for the Central Universities under Grant 24820232023YQTD01, and Grant 2023RC96, and in part by the Double First-Class Interdisciplinary Team Project Funds under Grant 2023SYLTD06. *(Corresponding author: Xin Wu.)*

The authors are with the School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xuyunpeng@188.com; 040251522wuxin@163.com; liwang@bupt.edu.cn; xulianming@bupt.edu.cn; 2437886958@qq.com; 475208092@qq.com).

Digital Object Identifier 10.1109/JSTARS.2023.3335288

I. Introduction

REMOTE sensing image object detection, as a key issue in remote sensing image interpretation, has been widely investigated and applied in various civilian and military tasks, such as geological disaster detection [1], remote precision strikes [2], [3], and aerospace and maritime defense [4]. Its objective is to determine whether a given aerial or satellite image contains one or more objects belonging to the interested category and to locate the position of each predicted object in the image. In recent years, more and more researchers are continuously trying to utilize data-driven deep learning methods to improve the accuracy and robustness of remote sensing image object detection, and are constantly pushing the development of remote sensing image object detection by introducing the aid techniques, such as super-resolution, attention mechanism, and multiscale fusion [5], [6].

In general, remote sensing image object detection can be categorized based on the density of the arrangement of objects. It can be divided into different types, such as dense object detection and sparse object detection. Dense object detection typically involves detecting numerous object instances in the image, whereas sparse object detection focuses on detecting only a few object instances in the image. In addition, dense objects are often difficult to distinguish from the background due to their high density and multiscale size, especially small or tiny size. Moreover, objects often appear in close proximity to each other with high similarity, making them prone to mutual occlusion and interference, thus increasing the difficulty of detection. Current densely-packed object detection methods in remote sensing imagery can be roughly categorized into four types: two-stage detectors, single-stage detectors, multiscale feature fusion methods, and attention-based methods. Specifically, two-stage detectors, such as region proposals convolutional neural network (RCNN) [7], Fast-RCNN [8], and Faster-RCNN [9], were the earliest detection methods applied in remote sensing images. Although they have high detection accuracy, their detection speed is slow, which makes them hard to handle large-scale remote sensing images in real-time. Single-stage detectors, such as You Only Look Once (YOLO) [10], Single Shot MultiBox Detector (SSD) [11], and RetinaNet [12], are currently mainstream detection networks, but they often focus on the background information and ignore objects themselves, so their detection performance is limited. Despite continuous updates in recent

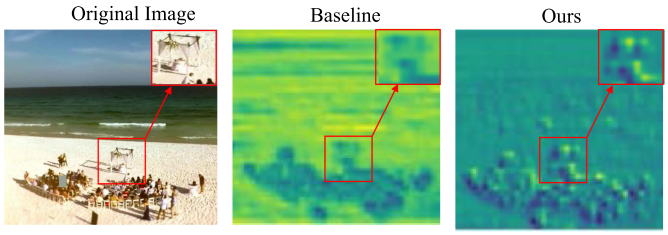

Fig. 1. Comparative of feature visualization between the baseline method and HOFA-Net using an example image with dense object.

years for single-stage models, such as YOLOv5, which has made improvements and optimizations in network architecture, model size, accuracy, and training methods. Fig. 1 shows the feature visualization of YOLOv5 for dense object detection, which indicates that there is still room for improvement in dense and even small-size object detection [13].

The imaging mechanism of remote sensing images often results in the presence of specific viewing angles, complex backgrounds, and diverse scales for objects within the images [14]. One effective approach to addressing these issues is the utilization of multiscale feature fusion methods. They can obtain multiscale information by introducing pyramid structure or multiscale feature fusion methods, such as FPN [15] and PAN [16]. Attention-based methods have shown significant advantages for object detection in remote sensing, as they can effectively capture contextual information and dependencies. They also allow the model to better focus on the location and features of objects, such as residual channel attention networks [17] and scale-aware networks [18] have been proven to perform well in small object detection in remote sensing images.

The attention mechanism can be traced back to neural science research in the 1980 s. Bahdanau et al. [19] first introduced the attention mechanism into machine translation tasks, proposing a neural machine translation model, now known as "attention-based neural machine translation." In 2015, the paper "Faster R-CNN: toward real-time object detection with region proposal networks (RPN)" proposed an RPN based on the attention mechanism [9], which was first applied to object detection and improved detection performance. Since then, the attention mechanisms are commonly used in object detection, and various forms have emerged, such as squeeze-and-excitation networks (SENet) [20], convolutional Block attention modules (CBAM) [21], and selective kernel networks (SKNets) [22], all of which have achieved good results and brought significant performance improvements to remote sensing object detection. For example, SENet [20] adaptively readjusts the channel weights of the feature maps by learning channel attention coefficients to enhance the response of useful features. The CBAM [21] based on spatial and channel attention mechanisms can capture spatial and channel correlations in the feature maps. The SKNet [22] attention mechanism proposed by the team at Tsinghua University could adaptively select different sizes of convolutional kernels to extract features. The nonlocal networks [23] proposed by the team at Nanjing University can learn the correlations between different positions and the cross-channel correlations. The global context network module [24] proposed at CVPR 2019 can learn global contextual information to improve object detection performance. Composite backbone network [25] calculates the weights of the channels as learnable parameters. However, most attention mechanism methods are used for high-resolution remote sensing images, which leads to complex models and large computational costs.

To this end, a high-order feature association network (HOFA-Det) for dense object detection in remote sensing has been proposed, shorted for HOFA-Det. The HOFA-Det network aims to better capture the interdependencies between features of channel and spatial dimensions, resulting in more distinctive and discriminative features. Specifically, HOFA-Det first learns the high-level features of the objects through convolutional neural networks (CNNs). To capture interdependencies between features without losing discriminative information from high-level but low-resolution features, we have designed a feature association module based on size adaptation nonlocal (SANL) and a high-order channel-wise attention (HCFA) module. It starts by partitioning the low-resolution features into local regions and employing nonlocal residual connections to capture the local contextual information of the objects. Behind it, the HFA module learns the interdependencies of features, while improving feature resolution. In addition, a covariance normalization acceleration strategy is introduced to accelerate computation. In detail, our contributions to this article can be summarized as follows.

1) A high-order feature association network (HOFA-Det) for dense object detection in remote sensing has been proposed. By improving feature resolution and generating more distinguishable features, our network effectively captures the interdependencies between features of channel and spatial dimensions to compensate the representation capability of single-level features, which helps to improve the model's detection performance for dense objects.
2) A SANL module has been developed to acquire size-adaptive nonlocal high-order features. This allows us to capture the interdependencies between features of channel and spatial dimensions, enhancing contextual information modeling while mitigating computational complexity. In addition, we have introduced a HFA module aimed at capturing nonlinear feature correlations. This module helps us establish the object distribution within the feature space, ultimately boosting the accuracy of dense object detection.
3) We evaluate the detection performance of the proposed HOFA-Det on two public remote sensing object detection datasets, i.e., DOTA and TinyPerson datasets, yielding significant advantages compared to several existing methods.

The rest of this article is organized as follows. In Section II, we provide a detailed introduction to the proposed HOFA-Det framework. This framework encompasses high-level feature learning, SANL learning, high-order feature association learning, as well as its associated loss function. Section III delves into the quantitative and visual analytics conducted on two public remote sensing object detection datasets, namely, the DOTA and TinyPerson datasets. We highlight the substantial advantages

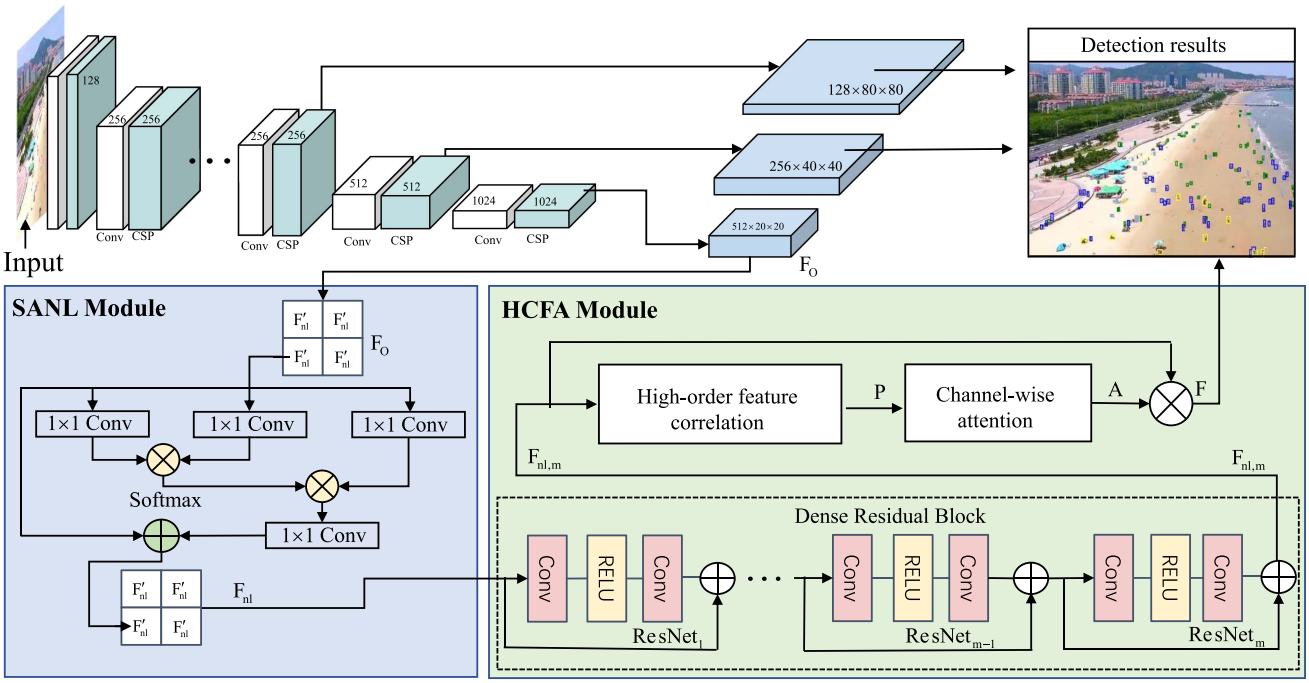

Fig. 2. Flowchart of the proposed HOFA-Det for dense object detection in RS images.

of HOFA-Net compared with several existing methods. Finally, Section IV concludes this article and discuss the future prospects of the HOFA-Det.

II. Methodology

Fig. 2 illustrates the network architecture of the proposed HOFA-Det. In detail, HOFA-Det is a detection network specifically developed for dense objects in remote sensing, especially suited for small-size objects. We first use CSPDarknet53 to learn high-level features with more details. Then, the generated feature is sent to the SANL module to establish long-range dependencies while reducing the computational burden. To further capture the nonlinear feature correlation and enhance model expressiveness, we introduced a HCFA module with its corresponding covariance normalization acceleration strategy.

A. Feature Learning

The cross-stage partial (CSP) structure first proposed is to reduce computational burden and enhance gradient performance. CSPDarknet53, as a representative of the CSP-Net series, consists of multiple layers of convolutions and residual blocks to capture image details, textures, and semantic information. CSPDarknet53 also enhances feature reuse and information flow within the network. This design offers several potential advantages. 1) CSPDarknet53 efficiently reuses features across different stages of the network, allowing for better exploitation of hierarchical features. This can be particularly beneficial when detecting dense objects with various scales and complexities, as it helps in capturing both low-level and high-level features effectively. 2) Enhanced information flow: The CSP connections facilitate the flow of information between different parts of the network, reducing information loss and improving the network's ability to handle dense and intricate object arrangements.

Given $\mathbf{X}$ as the input data and $\mathbf{F}$ as the learned features. At the beginning stage of feature learning, we involve a series of operations, such as convolution layers, pooling layers, and residual blocks, to learn high-level features $\mathbf{F}$ while preserving details and textures information. Following it, the CSP module splits the feature maps $\mathbf{F}$ into two paths: the main branch and the cross branch, that is $\mathbf{F}(x) = [\mathbf{F}'(x), \mathbf{F}''(x)]$

$$\begin{aligned} \mathbf{F_k} &= \omega_{\mathbf{k}} * [\mathbf{F_0}'' \mathbf{F_1} \cdots, \mathbf{F_{k-1}}] \\ \mathbf{F}_T &= \omega_T * [\mathbf{F}_0'' \mathbf{F}_1 \cdots, \mathbf{F}_k] \\ \mathbf{F}_O &= \omega_O * [\mathbf{F}_0' \mathbf{F}_T] \end{aligned} \tag{1}$$

where $\mathbf{F}_k$, $\mathbf{F}_T$, and $\mathbf{F}_O$ are the dense layer, transition layer, and output layer, respectively.

B. SANL Module

CNNs primarily focus on local region features when processing images, but neglecting the long-range dependencies among features. By incorporating nonlocal operations, we can effectively capture these relationships between features, improving the modeling of contextual information. However, the global-level nonlocal operations impose an impractical computational burden, particularly when dealing with large feature sizes. Dense objects are also influenced by their surrounding context. To address this, we partition the feature map into a grid of regions, with each $k \times k$ region-based nonlocal operation. The choice of k depends on the size of the feature map for that layer with the SANL module added. Assuming the size of the size adaptation

feature map is denoted as $\mathbf{F'}_O \in \mathbb{R}^{\frac{M}{k}\times\frac{N}{k}\times C}$, the size adaptation factor k is exemplified as follows:

$$k = \max(M/r, N/r, L) \tag{2}$$

where r is an empirical value and L is the smallest among the values of k. In this article, r and k are set to 20 and 2, respectively.

The SANL operation $\mathbf{F'}_{\text{nl}}(i)$ for specified position can be formulated as follows:

$$\mathbf{F'}_{\text{nl}}(i) = \frac{1}{C(x)}\sum_{\forall j} f(\mathbf{F'}_{\text{O}}(i), \mathbf{F'}_{\text{O}}(j))g(\mathbf{F'}_{\text{O}}(j)) \tag{3}$$

where the index i is the specified position and j represents the index of all possible positions. The function f is the pairing computation function, which calculates the correlation between the ith position and all other positions j. g is a unary input function designed for information transformation, we use a 1×1 convolution, representing linear embedding. $C(x)$ is a normalization function that ensures the overall information to remain unchanged before and after the transformation.

The final output $\mathbf{F}_{\text{nl}}$ of the SANL module is the matrix splicing of k^2 $\mathbf{F'}_{\text{nl}}$. The size adaption nonlocal module not only reduces the computational burden but also enables the capture of contextual information within and between local blocks. This enhancement improves the understanding and representation capacity of feature maps of varying sizes.

C. HCFA Module

To better capture nonlinear feature correlations, enhance model expressiveness, and further mitigate the risk of overfitting, this section introduces a HCFA module.

First, multiple simplified residual blocks are stacked together to achieve feature association because this stacking allows for the progressive refinement of features and the extraction of increasingly abstract and complex patterns from the input data. Assuming the stacking quantity is M, the generated dense residual blocks can be defined as

$$\mathbf{F}_{\text{nl},m} = H_m\left(\mathbf{F}_{\text{nl},m-1}\right) \tag{4}$$

where $H_m(\cdot)$ is the function of the mth residual units, and $\mathbf{F}_{\text{nl},m-1}$, $\mathbf{F}_{\text{nl},m}$ are the corresponding inputs and outputs.

In addition, we incorporate the high-order feature transformer [26] to capture nonlinear feature correlations effectively. This method indirectly allows us to obtain the object distribution in the feature space, leading to enhanced accuracy and precision in object detection [27], [28].

Taking into account that the size of $\mathbf{F}_{\text{nl},m}$ is $\mathbb{R}^{M\times N\times C}$, we reshape it to $Q \times C$, where $Q = M \times N$. The resulting normalized covariance matrix of $\mathbf{F}_{\text{nl},m}$ serves as the high-order nonlinear correlated feature. The covariance matrix is calculated as follows:

$$\mathbf{F}_{\text{nl},m} \mapsto \mathbf{P}, \mathbf{P} = \mathbf{F}_{\text{nl},m}\bar{\mathbf{I}}\mathbf{F}^T_{\text{nl},m} \tag{5}$$

where $\bar{\mathbf{I}} = \frac{1}{Q}(\mathbf{I} - \frac{1}{Q}\mathbf{1}\mathbf{1}^T)$, $\mathbf{I}$ is the $Q \times Q$ identity matrix, $\mathbf{1} = [1, \ldots, 1]^T$ is a Q-dimensional vector, and T stands for matrix transpose. The sample covariance matrix P is symmetric and positive semidefinite. It can be decomposed into eigenvalues as follows:

$$\mathbf{P} \mapsto (\mathbf{U}, \Lambda), \mathbf{P} = \mathbf{U}\Lambda\mathbf{U}^T \tag{6}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_C)$ is a diagonal matrix, $\lambda_i, i = 1, \ldots, C$ are eigenvalues arranged in nondecreasing order, and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_C]$ is an orthogonal matrix, where the $\mathbf{u}_i$ column corresponds to the eigenvector associated with λ_i. By using eigenvalue decomposition (EIG), the covariance normalization can be transformed into the power of eigenvalues

$$(\mathbf{U}, \Lambda) \mapsto \mathbf{Y}, \mathbf{Y} \triangleq P^\alpha = \mathbf{U}\mathbf{F}(\Lambda)\mathbf{U}^T \tag{7}$$

where α is a positive real number, $\mathbf{F}(\Lambda) = \text{diag}(\mathbf{f}(\lambda_1), \ldots, \mathbf{f}(\lambda_C))$, $\mathbf{f}(\lambda_i) = \lambda_i^\alpha$. Inspired by the technique of element-wise power normalization technique [27], [29], [30], we set $\alpha = \frac{1}{2}$ in this article.

Let $\mathbf{Y} = [y_1, \ldots, y_C]$, and the channel-wise statistics $\mathbf{z} \in \mathbb{R}^{C\times 1}$ can be obtained by squeezing the dimensions of $\mathbf{Y}$. Then, the cth dimension of $\mathbf{z}$ is computed as

$$z_c = \frac{1}{C}\sum_{i}^{C}\mathbf{y}_c(i) \tag{8}$$

To exploit the interdependencies between features of channel and spatial dimensions from squeezing information, we introduce the adaptive recalibration $\mathcal{A}$

$$\mathcal{A} = \sigma(\mathcal{B}(\mathbf{W}_2\delta(\mathcal{B}(\mathbf{W}_1\mathbf{z})))) \tag{9}$$

where $\sigma(\cdot), \delta(\cdot)$, and $\mathcal{B}(\cdot)$ are the sigmoid function, rectified linear unit (ReLU), and batch normalization, respectively. $\mathbf{W}_1 \in R^{1\times 1\times\frac{C}{r}}$ and $\mathbf{W}_2 \in R^{1\times 1\times C}$ stands for weight, respectively. The final high-order feature is $\mathbf{F} = \mathcal{A} \cdot \mathbf{F}_{\text{nl},m}$.

Considering that covariance normalization relies heavily on eigenvalue decomposition, and the GPU platform requires a large number of iterations, which leads to low training efficiency. To address this issue, we introduce a fast matrix normalization method based on Newton–Schulz iteration [31], [32], [33], [34] to achieve an approximate solution with fewer iterations, thus accelerating the normalization process.

D. Loss Function

We employ three loss functions for network optimization and updating. These consist of an object classification loss L_{cls}, a location loss L_{loc}, and an objectness loss L_{obj}.

L_{cls} is the binary cross-entropy loss, quantifying the disparity between the prediction box's class and the ground truth box's class. L_{obj} is also the binary cross-entropy loss, determining whether there are targets within the prediction box

$$L = -\sum_{i=0}^{N}\left[y_i \log p_i + (1 - y_i)\log\left(1 - p_i\right)\right] \tag{10}$$

where p_i is the predicted probability of the ith sample, between 0 and 1, y_i is the true label of the ith sample, either 0 or 1.

The location loss assesses the disparity between the position and shape of the prediction box and the ground truth box. It

utilizes a complete intersection over union (CIoU) metric [35]

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2\left(p, p^{gt}\right)}{c^2} + \alpha v$$
$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}$$
$$\alpha = \frac{v}{1 - \text{IoU} + v}$$
$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{11}$$

where L_{CIoU} is the loss value of CIoU. A is the forecast box. B is the actual box. ρ is the Euclidean distance. p is the center point of the prediction box. p^{gt} is the center point of the target box. c is the diagonal distance of the minimum bounding rectangle between the boxes. α is the weight function. Intersection over Union (IoU) is the intersection ratio of the prediction box to the actual box. v is the aspect ratio metric function. w^{gt} is the width of the target box. h^{gt} is the height of the target box. w is the width of the prediction box. h is the height of the prediction box.

The total loss is

$$L_{\text{total}} = \frac{1}{N}\sum_{i=0}^{N}\left(\lambda_1 \cdot L_i^{\text{cls}} + \lambda_2 \cdot L_i^{\text{loc}} + \lambda_3 \cdot L_i^{\text{obj}}\right) \tag{12}$$

where N is the sample number and $\lambda_{1\sim3}$ is an adjustable weight.

III. Experimental Results and Analysis

A. Data Description

In this section, we construct two multimodal object detection benchmark datasets in remote sensing that focus on the properties of dense and irregular parking, similar appearance, and similar color.

1) DOTA Dataset[1]*:* DOTA 1.0 dataset containing 2806 aerial images from 800 × 800 pixels to 4000 × 4000 pixels, in which more than 188 282 objects falling into 15 categories are annotated. To fit the network, the input images in the DOTA1.0 dataset were cropped into one size: 1024 × 1024 pixels, with a 200 pixel overlap to maximize object information. The categories of the objects in DOTA are plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). In the experiment, the training set to the testing set data ratio for the network is 3:1.

2) Tiny Person[2]*:* TinyPerson [36] is a dataset specially prepared for small object detection recently launched by the University of the Chinese Academy of Sciences. The images in TinyPerson are mainly from the Web. The researchers collect high-definition videos from various websites (images were collected every 50 frames). Duplicate images are removed. Finally, a total of 72 651 object enclosures are manually marked. The dataset contains two tags, sea_person, and dust_person, which are very small in absolute size. According to Tiny Benchmark, focusing on the person detection task, the earth_person and sea_person are both treated as person, and the size range of them are divided: tiny [2, 20], small [20, 32], and all [2, inf]. The tiny [2, 20] is further divided into 3 intervals: tiny1 [2, 8], tiny 2[8, 12], and tiny3 [12, 20].

To fit the network, since most of the images in the TinyPerson dataset are quite large, the input images are cropped into subimages (512 × 640 or 640 × 512 pixels) with overlapping during training and test. The training set to the testing set data ratio for the network is 1:1.

B. Experimental Setup

All the experiments were implemented on a computer with a 13th Gen Intel Core i7-13700 K, 64 GB of memory, and two NVIDIA GeForce RTX 3090 GPUs (2 × 24 GB). For the DOTA 1.0 experiment, we utilized a stochastic gradient descent (SGD) optimizer with an initial learning rate of 1×10^{-2}, which decays to 1×10^{-3}. The total training round is 150 epochs, and the batch size is 16. The threshold of confidence and the threshold of nonmaximum suppression are 0.25, and 0.45, respectively. For the Tinyperson experiment, all experiments were the same, except we used an SGD optimizer that decayed to 2×10^{-3}, and the total training round is 200 epochs.

C. Evaluation Metrics

The common object detection evaluation criteria AP is used for quantitative analysis. AP is a global indicator and can fairly be used to compare various detection methods. In the experiment, we utilized $AP25$, $AP50$, and $AP75$ to calculate the AP at the IoU threshold of 0.25, 0.5, and 0.75, respectively

$$AP = \sum_{k=1}^{n} P(k)\Delta R(k) \tag{13}$$

where k is the threshold., $P(k)$ is the precision at the kth threshold. $\Delta R(k) = R(k) - R(k-1)$ stands for the differences in *precision* and *recall*.

D. Comparison With State-of-the-Art MVD Models

In the experiment, several state-of-the-art methods related to object detection of RS images are selected for quantitative and qualitative comparisons, namely, learning RoI transformer (ROI Trans.) [37], rotation sensitivity detector (RSDet) [38], small, cluttered, and rotated detector (SCRDet) [39], refined single-stage rotation detector (R^3Det) [40], gliding vertex [41], oriented objects detection network (O^2-DNet) [42], box boundary aware vectors (BBAVectors) [43], dynamic refinement network (DRN) [44], circular smooth label (CSL) [45], critical

[1][Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html.

[2][Online]. Available: https://universe.roboflow.com/chris-d-dbyby/tiny person/dataset/1.

TABLE I
COMPARATIVE EXPERIMENTS ON THE DOTA DATASET

DOTA Dataset									
CSPDarknet53	**SANL**	**HFA**	Anchor	mAP					
✓			AF	74.65					
✓	✓		AF	73.91					
✓		✓	AF	75.39					
✓	✓	✓	AF	**75.53**					
Tiny Person Dataset									
CSPDarknet53	**SANL**	**HFA**	Anchor	AP_{50}^{tiny1}	AP_{50}^{tiny2}	AP_{50}^{tiny3}	AP_{50}^{tiny}	AP_{50}^{small}	
✓			AF	25.45	45.80	53.86	42.57	58.08	
✓	✓		AF	26.34	46.09	**54.52**	43.70	58.22	
✓		✓	AF	26.05	46.55	54.38	43.54	58.41	
✓	✓	✓	AF	**26.44**	**47.10**	54.16	**43.71**	**58.96**	

The best results are shown in bold.

TABLE II
COMPARATIVE EXPERIMENTS ON THE DOTA DATASET

Methods	Year	Backbone	Anchor	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP	Speed (fps)
ROI Tran. [37]	2019	ResNet101	AB	88.53	77.91	37.63	74.08	66.53	62.97	66.57	90.50	79.46	76.75	59.04	56.73	62.54	61.29	55.56	67.74	7.80
RSDet [38]	2019	ResNet101	AB	89.80	82.90	48.60	65.20	69.50	70.10	70.20	90.50	85.60	83.40	62.50	63.90	65.60	67.20	68.00	72.20	-
SCRDet [39]	2019	ResNet101	AB	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61	9.51
R^3Det [40]	2019	ResNet152	AB	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.67	62.68	67.53	**78.56**	72.62	76.47	10.53
Gliding Vertex [41]	2019	ResNet101	AB	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	**70.91**	72.94	70.86	57.32	75.02	13.10
O^2-DNet [42]	2020	Hourglass-104	AF	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04	-
BBAVectors [43]	2020	ResNet101	AF	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32	18.37
DRN [44]	2020	Hourglass-104	AF	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23	-
CSL [45]	2020	ResNet152	AB	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17	-
CFC-Net [46]	2021	ResNet50	AB	89.08	80.41	52.41	70.02	76.28	78.11	87.21	**90.89**	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50	17.81
RIL [47]	2021	ResNet101	AB	88.94	78.45	46.87	72.63	77.63	80.68	88.18	90.55	81.33	83.61	64.85	63.72	73.09	73.13	56.87	74.70	13.36
RetinaNet-GWD [48]	2021	ResNet152	AB	86.14	81.59	**55.33**	75.57	74.20	67.34	81.75	87.48	82.80	85.46	69.47	67.20	70.97	70.91	74.07	75.35	11.65
GGHL [49]	2021	DarkNet53	AF	89.74	**85.63**	44.50	**77.48**	76.72	80.45	86.16	90.83	**88.18**	86.25	67.07	69.40	73.38	68.45	70.14	**76.95**	42.30
ARSD [53]	2022	ResNet101	AB	86.90	69.64	46.38	56.85	**80.60**	66.96	78.96	90.76	72.15	78.96	39.67	61.27	72.23	72.39	50.64	68.28	43.00
HOFA-Net	-	CSPDarknet53	AF	**90.42**	76.64	47.57	59.01	73.41	**85.64**	**89.29**	90.76	73.30	**89.44**	**71.15**	69.39	**75.16**	67.05	**74.77**	75.53	**44.14**

The best results are shown in bold.

feature capturing network (CFC-Net) [46], representation invariance loss (RIL) [47], RetinaNet-Gaussian wasserstein distance (RetinaNet-GWD) [48], general gaussian heatmap label (GGHL) [49], fully convolutional one-stage object detector (FCOS) [50], RetinaNetS [12], dual shot face detector (DSFD) [51], adaptive freeAnchor [52], and adaptive reinforcement supervision distillation (ARSD) [53]. To avoid over-fitting, data preprocessing, and augmentation are performed on all images, and the CSP-Darknet53 serves as the backbone network for all datasets.

E. Ablation Study

To validate the increment of each module, in this section, we conducted ablation experiments to evaluate the contribution of different modules, including the SANL module and the HFA module, as given in Table I. It can be observed that when the SANL module is used independently, there is a slight decrease in performance on the DOTA dataset. However, the performance on tiny and small objects in the tiny person dataset improves. One possible reason for this is that using SANL alone is more favorable for small objects by partitioning the $k \times k$ region. Building upon this, the HFA module, with its capacity to capture nonlinear feature correlations and enhance model expressiveness, leads to a 1% improvement.

F. Results and Analysis on the DOTA Data

Table II gives the quantitative results of 14 widely used methods and the proposed method on the DOTA dataset. It can be observed that the proposed method shows significant improvements for densely packed objects, such as PL, LV, SH, ST, SBF, HA, and HC. However, its performance is relatively weaker for large-scale or elongated objects, such as BDs, ground field tracks, BRs, BCs, and SP. Specifically, methods, such as ROI Trans. RSDet, R^3Det, Gliding vertex, O^2-DNet, BBAVectors, CSL, RIL, and RetinaNet-GWD, focus specifically on detecting rotated objects and employ different strategies to improve detection accuracy. These strategies include deforming the region of interest, using specific loss functions, rotating anchor boxes, determining the number of anchor points, implementing multilevel and multiscale designs, and so on. However, these methods often lack an effective strategy to separate densely packed objects, particularly small or tiny ones.

SCRDet addresses this limitation by adapting the size of the receptive field based on the density information around the objects. In addition, it incorporates adaptive attention-based feature fusion methods to enhance the representation of small objects. Nevertheless, challenges remain in recovering and enhancing fine details. RIL and RetinaNet-GWD enhance the accuracy of detecting rotated objects through the use of representation invariance loss, Gaussian Wasserstein distance, etc. GGHL improves the detection performance of dense small objects by utilizing a Gaussian distribution-based heatmap label assignment method. However, improper parameter settings may result in inaccurate localization or missed detection. CFC-Net and ARSD introduce an adaptive receptive field mechanism and a multiscale core features imitation (MCFI) module to adapt to objects of different scales and sizes, enabling the network to perceive better and

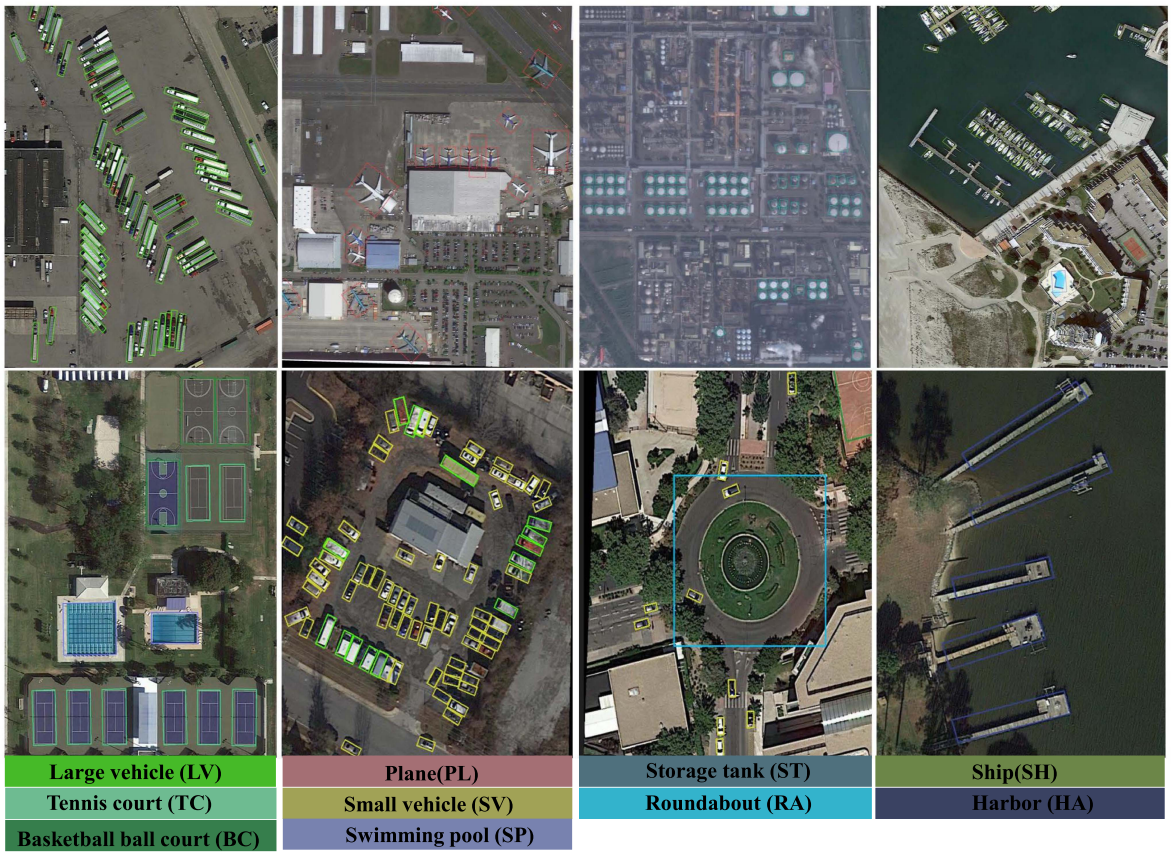

Fig. 3. Visualization results of the proposed HOFA on part example images in DOTA dataset.

capture the features of objects at various scales. However, it does not effectively address the challenges posed by low-density and crowded objects.

Overall, these existing methods face difficulties in detecting dense objects, especially tiny objects that are densely overlapping. The depth of the network often significantly reduces the resolution of small objects. Although some methods, such as DRN, gradually improve the accuracy and robustness of detecting densely packed objects through multilevel refinement modules, they still struggle to effectively detect densely packed small objects with low resolution and indistinct texture features. The proposed method addresses these challenges by designing SANL feature learning and high-order feature association, enhancing the correlation between channel and spatial features, indirectly improving the resolution of dense and even small objects, and achieving the best results.

Fig. 3 shows the visual results of the proposed method on some example images in DOTA data. We can observe that the quantitative (see Table II) and qualitative (see Fig. 3) results are generally consistent. It is worth noting that the proposed method demonstrates good separability in detecting densely packed objects. The performance of detecting SV is not very good, one possible reason for this could be that the discriminability between different vehicle sizes has not been effectively learned. However, for SH in the port, the complexity of detection is increased due to the various types and sizes of vessels. The surrounding environment of the port is typically complex, which contributes to a higher proportion of false detection and missed detection.

G. Results and Analysis on the Tinyperson Data

Table III gives the quantitative results of four comparative methods and the proposed HOFA-Net on the Tinyperson dataset. It shows that the proposed HOFA-Net demonstrates significant enhancements in detecting dense and even occluded small objects. As expected, the classic fully convolutional FCOS network exhibits the lowest detection performance. RetinaNetS and DSFD methods utilize a multiscale feature pyramid network and an adaptive downsampling strategy to extract features of varying scales from different levels of feature maps, effectively adapting to objects of different sizes. FreeAnchor enables multiple anchor boxes to be matched with each target and determines the best match through learning. However, these methods have limitations in detecting tiny objects in high-level semantic feature maps with low resolutions. HOFA-Net effectively enhances high-level feature map resolution through SANL and high-order feature association, thereby improving the performance of detecting tiny objects.

Fig. 4 shows the visual results of the proposed method on some example images in tiny data. It shows that the proposed method demonstrates better results in detecting objects in challenging

TABLE III
COMPARATIVE EXPERIMENTS ON THE TINYPERSON DATASET

Methods	Year	Backbone	Anchor	AP_{50}^{tiny1}	AP_{50}^{tiny2}	AP_{50}^{tiny3}	AP_{50}^{tiny}	AP_{50}^{small}	AP_{25}^{tiny}	AP_{75}^{tiny}
RetinaNetS [12]	2017	ResNet50	AB	11.80	37.14	44.12	31.65	44.19	58.10	2.95
FCOS [50]	2019	-	AB	3.74	13.03	30.18	17.53	36.64	41.38	1.52
DSFD [51]	2019	ResNet-50	AB	14.55	35.89	47.23	31.72	52.56	60.37	2.23
FreeAnchor [52]	2019	ResNet50	AB	25.34	**48.62**	52.13	42.15	54.24	**64.69**	4.37
HOFA-Net	-	DarkNet53	AF	**26.44**	47.10	**54.16**	**43.71**	**58.96**	61.76	**6.22**

The best results are shown in bold.

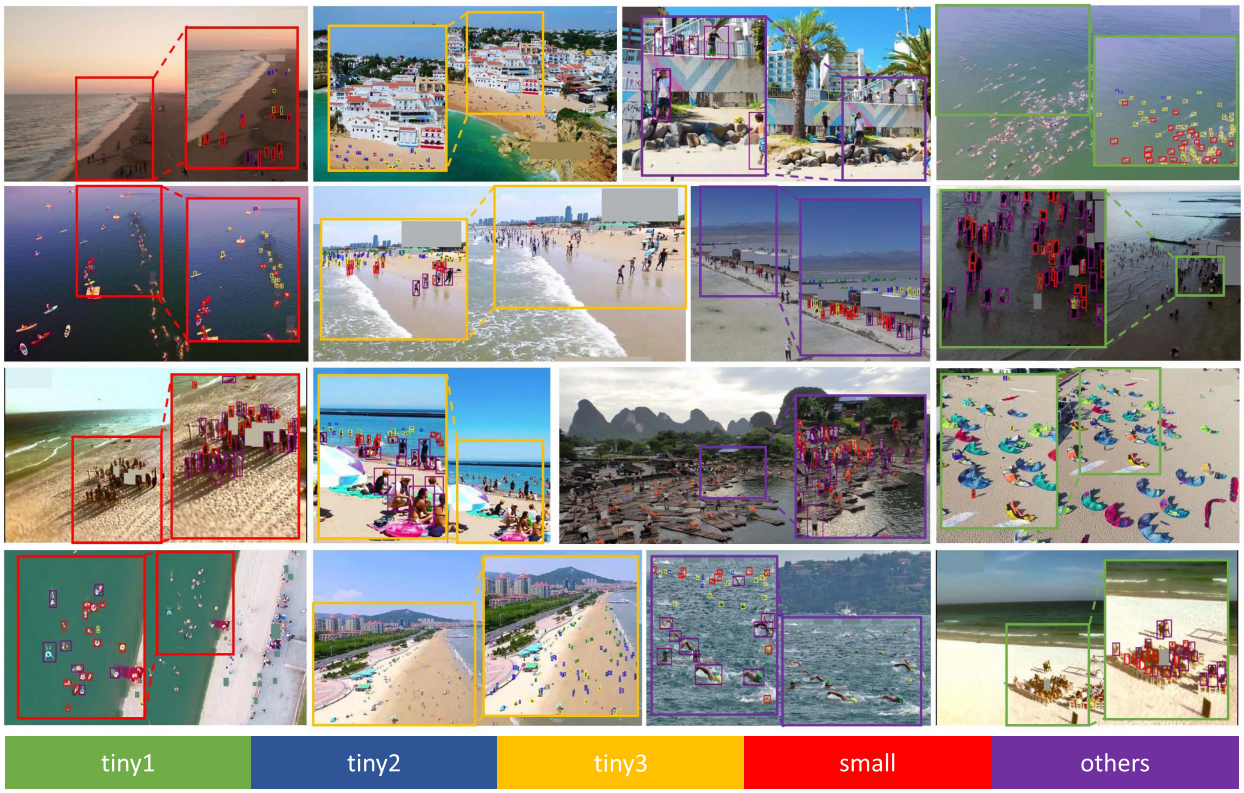

Fig. 4. Visualization results of the proposed HOFA on part example images in tiny person dataset.

scenarios, such as low-resolution images, small individuals with low signal-to-noise ratios, and subjects exhibiting noticeable scale variations. Nonetheless, there is still potential for enhancing the accuracy of detection, particularly when dealing with small individuals in low-light conditions where instances overlap and occlusion occurs. Furthermore, the dataset comprises a wide range of deformations and poses, including individuals standing, lying down, swimming, and sitting in diverse scenes like beaches and waterfronts. Consequently, these deformations introduce a notable number of missed detections. Improving the precise detection performance in these demanding scenarios is an area of focus for future research endeavors.

IV. CONCLUSION

In this article, we propose a high-order feature association (HOFA) network for dense object detection in remote sensing, yielding capturing of the interdependencies between features of channel and spatial dimensions. In HOFA-Net, CNNs are utilized to generate high-level but low-resolution features. To capture long-range dependencies of local features, a SANL learning module has been proposed. The high-order correlation feature association is then employed to capture interdependencies within channel and spatial features while incorporating a covariance normalization acceleration strategy. Although the proposed method significantly improves the detection performance of dense objects, there is still potential for further improvement. In the future, we will mainly focus on the wide range of deformations and pose object detection, including individuals in various positions such as standing, lying down, swimming, and sitting, within diverse scenes, such as beaches and waterfronts.

REFERENCES

[1] H. Wang et al., "Multi-source remote sensing intelligent characterization technique-based disaster regions detection in high-altitude mountain forest areas," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Aug. 2022, Art. no. 3512905.

[2] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[3] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5412012.

[4] R. Hang, G. Li, M. Xue, C. Dong, and J. Wei, "Identifying oceanic eddy with an edge-enhanced multiscale convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9198–9207, Sep. 2022.

[5] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[6] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 5527812.

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

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

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[10] 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.*, 2016, pp. 779–788.

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

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

[13] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5513412, doi: 10.1109/TGRS.2023.3279834.

[14] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, Dec. 2023.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[16] Z. Ma, M. Li, and Y. Wang, "PAN: Path integral based convolution for deep graph neural networks," 2019, *arXiv:1904.10996*.

[17] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[18] 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.*, 2018, pp. 316–331.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[21] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[22] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[23] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[24] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 1971–1980.

[25] Y. Liu et al., "CBNet: A novel composite backbone network architecture for object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, pp. 11653–11660, 2020.

[26] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jun. 2023, Art. no. 5514415.

[27] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2070–2078.

[28] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[29] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, pp. 222–245, 2013.

[30] D. Hong et al., "SpectralGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[31] N. J. Higham, *Functions of Matrices: Theory and Computation*. Philadelphia, PA, USA: SIAM, 2008.

[32] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 947–955.

[33] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11065–11074.

[34] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[35] Z. Zheng et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022.

[36] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny person detection," in *Proc. IEEE/CVF winter Conf. Appl. Comput. Vis.*, 2020, pp. 1257–1265.

[37] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

[38] W. Qian, X. Yang, S. Peng, Y. Guo, and C. Yan, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2458–2466, May 2021.

[39] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8232–8241.

[40] X. Yang, Q. Liu, J. Yan, and A. Li, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3163–3171, May 2021.

[41] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[42] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 268–279, 2020.

[43] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. N. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2150–2159.

[44] X. Pan et al., "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11207–11216.

[45] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 677–694.

[46] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2021.

[47] Q. Ming, Z. Zhou, L. Miao, X. Yang, and Y. Dong, "Optimization for oriented object detection via representation invariance loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art no. 8021505, doi: 10.1109/LGRS.2021.3115110.

[48] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian wasserstein distance loss,"in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 11830–11841.

[49] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general Gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 1895–1910, Sep. 2022.

[50] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[51] J. Li et al., "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5060–5069.

[52] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, "FreeAnchor: Learning to match anchors for visual object detection," 2019, *arXiv:1810.10220*.

[53] Y. Yang et al., "Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5623715.

YunPeng Xu (Member, IEEE) received the B.Sc. degree in network engineering from Hebei University, Baoding, China, in 2007, and the M.E. degree in 2013 from the Beijing University of Posts and Telecommunications, Beijing, China, where he is currently working toward the Ph.D. degree in software engineering.

His research interests include intelligent algorithms, machine learning, and computer vision.

Xin Wu (Senior Member, IEEE) received the M.Sc. degree in computer science and technology from the College of Information Engineering, Qingdao University, Qingdao, China, in 2014, and the Ph.D. degree in information and communication engineering from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2020.

She is currently an Assistant Professor with the School of Computer Science, Beijing University of Posts and Telecommunications (BUPT), Beijing. Her research interests include artificial intelligence, signal/image processing, and geospatial object detection and tracking.

Dr. Wu is a Topical Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS). She was a recipient of the Best Reviewer Award of the IEEE TGRS in 2023 and the IEEE JSTARS in 2022 as well as the Jose Bioucas Dias award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2021. She is also a Leading Guest Editor of the IEEE JSTARS and Remote Sensing.

Li Wang (Senior Member, IEEE) received the Ph.D. degree in circuits and systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2009.

She is currently a Full Professor with the School of Computer Science, National Pilot Software Engineering School, BUPT, where she is also an Associate Dean and the Head of the High Performance Computing and Networking Laboratory. She is also a Member of the Key Laboratory of the Universal Wireless Communications, Ministry of Education, China. She is also a rotating Director of the Key Laboratory of Application Innovation in Emergency Command Communication Technology, Ministry of Emergency Management, China. She also held Visiting Positions with the School of Electrical and Computer Engineering, Georgia Tech, Atlanta, GA, USA, from 2013 to 2015, and with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, from 2015, and 2018. She has authored or coauthored almost 70 journal papers and four books. Her research interests include wireless communications, distributed networking and storage, vehicular communications, social networks, and edge AI.

Dr. Wang is currently serves on the Editorial Boards for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE INTERNET OF THINGS JOURNAL, AND CHINA COMMUNICATIONS. She was an Associate Editor for IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, the Symposium Chair of IEEE ICC 2019 on Cognitive Radio and Networks Symposium, and a Tutorial Chair of IEEE VTC 2019. She is also the Chair of the Special Interest Group on Sensing, Communications, Caching, and Computing in Cognitive Networks for the IEEE Technical Committee on Cognitive Networks. She was the Vice Chair of the Meetings and Conference Committee for the IEEE Communication Society Asia Pacific Board for the term of 2020–2021. She was the recipient of the 2013 Beijing Young Elite Faculty for Higher Education Award, best paper awards from several IEEE conferences, IEEE ICCC 2017, IEEE GLOBECOM 2018, and IEEE WCSP 2019, and the Beijing Technology Rising Star Award in 2018. Shewas the TPC of multiple IEEE conferences, including IEEE Infocom, Globecom, International Conference on Communications, IEEE Wireless Communications and Networking Conference, and IEEE Vehicular Technology Conference in recent years.

Lianming Xu received the B.E. degree from the Hefei University of Technology, Hefei, China, in 2003, and the Ph.D. degree in electronic science and technology from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2009.

He is currently an Assistant Professor with the School of Electronic Engineering, BUPT. His research interest includes edge intelligence, the Internet of Things, caching, and collaborative computing.

Zhengyu Shao received the B.E. degree in electronic commerce and law in 2023 from the Beijing University of Posts and Telecommunications, Beijing, China, where he is currently working toward the Ph.D degree in computer science and technology.

His research interests include computer vision, and machine learning.

Aiguo Fei received the M.S. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1981, and the Ph.D. degree in control science and engineering from the University of Science and Technology Beijing, Beijing, in 2004.

He is currently a Professor with the School of Computer Science (National Pilot Software Engineering School), BUPT. He is also a Member of the State Key Laboratory of Networking and Switching Technology and the Academician of the Chinese Academy of Engineering, Beijing. His current research interests include the Internet of Things, intelligent emergency communication systems, intelligent information systems, Big Data, cloud computing, and intelligent software development and testing.