Guided Hybrid Quantization for Object Detection in Remote Sensing Imagery via One-to-One Self-Teaching

Jiaqing Zhang, Jie Lei, *Member, IEEE*, Weiying Xie, *Member, IEEE*, Yunsong Li, *Member, IEEE*, Geng Yang, and Xiuping Jia, *Fellow, IEEE*

***Abstract*— Deep convolutional neural networks (CNNs) have improved remote sensing image analysis, but their high computational demands may limit their deployment on low-end devices with limited resources, such as intelligent satellites and unmanned aerial vehicles. Considering the computation complexity, we propose a guided hybrid quantization with one-to-one self-teaching (GHOST) framework. More concretely, we first design a structure called guided quantization self-distillation (GQSD), an innovative idea for realizing a lightweight model through the synergy of quantization and distillation. The training process of the quantization model is guided by its full-precision model, which is time-saving and cost-saving without preparing a huge pretrained model in advance. Second, we put forward a hybrid quantization (HQ) module that automatically acquires the optimal bit-width by imposing a threshold constraint on the distribution distance between the center point and samples in the weight search space, aiming to retain more shallow detail information that is advantageous for small object detection. Third, to improve information transformation, we propose a one-to-one self-teaching (OST) module to give the student network the ability to self-judgment. A switch control machine (SCM) builds a bridge between the student and teacher networks in the same location to help the teacher reduce wrong guidance and impart vital knowledge about objects without vast background information to the student. This distillation method allows a model to learn from itself and gain substantial improvement without any additional supervision. Extensive experiments on a multimodal dataset (VEDAI) and single-modality datasets (DOTA, NWPU, and DIOR) show that object detection based on GHOST outperforms the existing detectors. The tiny parameters (<9.7 MB) and bit-operations (BOPs) (<2158 G) compared with any remote sensing-based, lightweight, or distillation-based algorithms demonstrate the superiority in the lightweight design domain. Our code and model will be released at https://github.com/icey-zhang/GHOST.**

***Index Terms*— Distillation, object detection, quantization, remote sensing image.**

Manuscript received 6 April 2023; revised 20 May 2023 and 21 June 2023; accepted 27 June 2023. Date of publication 7 July 2023; date of current version 19 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071360. *(Corresponding author: Jie Lei.)*

Jiaqing Zhang, Jie Lei, Weiying Xie, Yunsong Li, and Geng Yang are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: jqzhang2@stu.xidian.edu.cn; jielei@mail.xidian.edu.cn; wyxie@xidian.edu.cn; ysli@mail.xidian.edu.cn; gengyang@stu.xidian.edu.cn).

Xiuping Jia is with the School of Engineering and Information Technology, The University of New South Wales, Canberra, ACT 2600, Australia (e-mail: xp.jia@ieee.org).

Digital Object Identifier 10.1109/TGRS.2023.3293147

I. Introduction

In THE field of remote sensing, object detection plays a critical role in many applications by identifying objects of interest [1]. Although universal detectors based on deep learning for natural images have been introduced in remote sensing, specialized detectors for remote sensing scenes have been designed and improved to meet specific object detection tasks. Despite the significant strides made in deep learning, deploying models on resource-constrained devices remains daunting. Within the satellite domain, high-performance computing hardware, such as graphics processing units (GPUs) and field-programmable gate arrays (FPGAs), are commonly employed for real-time image analysis and processing, along with the acceleration of machine learning algorithms [2], [3]. However, given the peculiar working conditions, satellite platforms impose stringent constraints on onboard hardware's size and power consumption, often necessitating bespoke designs that cater to application-specific requirements.

One of the foremost challenges posed by existing deep neural network-based object detection architectures is their computational and storage requirements, which can be prohibitively taxing for remote sensing devices, such as satellites, drones, and airplanes, with limited computing resources. For instance, the VGG-16 [4] model consists of 138.34 million parameters, occupies over 500 MB of storage, and requires 15.5 billion floating-point operations to analyze a single image. Such high-complexity models can easily exceed the computational limitations of most remote sensing devices, hindering practical deployment in resource-limited scenarios and increasing the burden of processing large volumes of images. To address this issue, several compression schemes have been proposed, including pruning [5], quantization [6], [7], [8], and distillation [9], [10].

Quantization algorithms [11], [12] directly compress the cumbersome network, effectively reducing the computation cost and model size with a great compression potential. However, trivially applying quantization to convolutional neural networks (CNNs) usually leads to inferior performance if the compression bit decreases to a low level. Some knowledge distillation (KD) methods [13], [14] are proven to be valid to elevate the performance of the lightweight model but have to pre-train a huge teacher model as a guide of the student model, which is time-consuming and resource-consuming [15].

Self-distillation methods [16], [17] overcome this problem via the transfer of information inside the model itself without introducing extra huge storage and time consumption from the teacher model.

The above view naturally leads to two questions: 1) what research results will we get if we combine quantization and distillation by using a small full-precision network to guide the learning process of a quantization network for this full-precision network in the remote sensing field? and 2) can the obtained quantization network overcome the problems of small target detection difficulties and complex background interference in remote sensing object detection? In this way, the formidable compression capability of quantized networks can complement and cooperate with the performance of full-precision networks, making it feasible to utilize quantized networks in remote sensing object detection.

In this article, we design an adaptive one-to-one educational policy pertaining to the full-precision network and the quantization network. We propose a simple yet novel approach that allows the quantization network to reinforce presentation learning of itself relative full-precision network without the need for additional labels and external supervision. Our approach is named *guided hybrid quantization with one-to-one self-teaching* (GHOST) based on the guided quantization self-distillation (GQSD) framework. As the name implies, GHOST allows a network to exploit useful and vital knowledge derived from its own full-precision layers as the distillation targets for its quantization layers. GHOST opens up new possibilities of training accurate tiny object detection networks.

Instead of training a large teacher model comes first, followed by distilling the knowledge from it to the student model, we propose a two-step mixed-bit self-distillation framework, in which the training process of the second quantization step is based on the pretrained small full-precision model. In other words, our proposed method can be applied in the lightweight model to obtain a smaller network to achieve detection acceleration. And the one-to-one self-training (OST) module creates distillation relationships automatically by switch control machine (SCM), thereby aiding the student model in detecting small objects in remote sensing images and reducing the impact of complex background information during the feature extraction phase via this distillation technology. Additionally, our hybrid quantization (HQ) method adaptively sets a larger bit-width in the shallow layers of the network to retain more detailed information in these layers, thereby improving the detection of small objects in remote sensing images. It also performs better compared with the standard fixed bit quantization. As shown in Fig. 1, the proposed framework not only requires less computation cost (from 17 024 to 692 G BOPs on the VEDAI dataset, a 24× reduction in training complexity) but also can accomplish much higher accuracy (from 75.92% in traditional quantization to 80.31% on SuperYOLO).

The main contributions of our work are as follows.

1) We propose a unified guided quantization method GQSD, which can tackle the lightweight object detectors' quantization optimization problem in remote sensing. We are the first to formulate an adaptive one-to-one education policy between the full-precision network and the quantization network at the same structure in object detection.
2) Based on the finding of weight value distribution features of remote sensing images, we design an HQ module whose adaptive selection of the core information of the weight values for quantization with a constrained preset condition can keep the balance between accuracy and efficiency.
3) Aiming to offset the loss of the quantization information, the SCM is adopted to enable the student to distinguish and close the teacher's wrong guidance and mine the correct and vital knowledge from self-distillation.

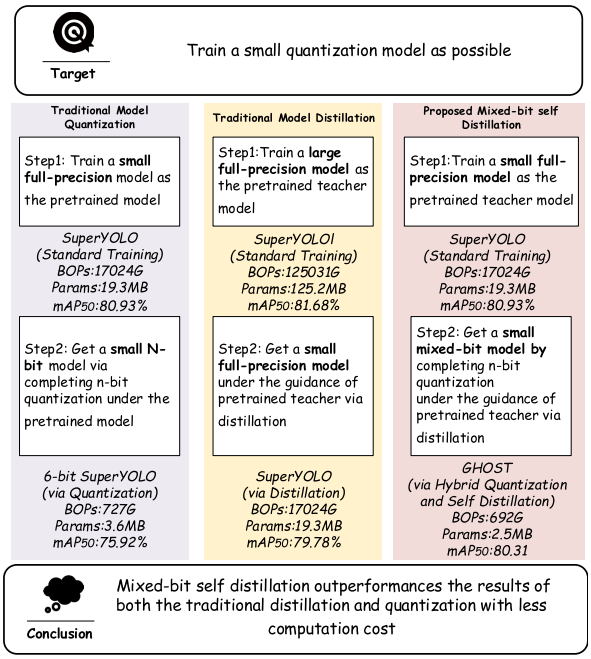

Fig. 1. Comparison of training complexity, and accuracy between traditional distillation, traditional quantization and proposed mixed-bit self-distillation (reported on VEDAI fold one validation).

The rest of this article is organized as follows. Section II gives a rough overview of the specific related work to this article. Section III presents our proposed method in detail. Section IV introduces experimental results and analysis. Section V concludes this article and discusses future work.

II. Related Work

In this section, we reviewed related work from object detection and network compression and acceleration in detail.

A. Object Detection in Remote Sensing

Various CNN-based object detection architectures, both two-stage [18], [19], [20] and one-stage [21], [22], showed promising performance, bringing the natural image object detection field to a new level. To solve the dilemma of the category imbalance, RetinaNet [23] reduced the weight of massive amounts of simple negative samples in training by designing a focal term for cross-entropy loss. As an anchor-free method, FCOS [24] has eliminated the need for adjusting hyperparameters and calculations related to anchor boxes.

ATSS [25] selected positive samples adaptively to enhance the detection performance.

Remote sensing images owned complex backgrounds and multiscale objects compared to natural images. Therefore, researchers proposed many object detection algorithms based on deep learning for remote sensing images and have devoted themselves to improving the feature extraction performance of deep learning networks and the regression method of the detector's bounding boxes. Cheng et al. [26] developed a feature enhancement network to improve object detection in remote sensing images. Wu et al. [27] proposed a global context aggregation module to combine high-level and low-level features through feature weaving and a feature refinement module to enhance feature distinction at different scales for detecting rotating objects with a large aspect ratio and dense arrangement. Hou et al. [28] introduced the asymmetric feature pyramid network with a dynamic feature alignment module and area-IoU regression loss to detect multiclass objects with arbitrary orientations in remote sensing images. Xu et al. [29] addressed the issue of spatial misalignment between ground truth and anchors with a new pseudo-anchor proposal module. Cheng et al. [30] proposed a novel anchor-free oriented proposal generator that abandons horizontal box-related operations from the network architecture. Huang et al. [31] proposed an anchor-free object-adaptation label assignment strategy to define positive candidates based on 2-D oriented Gaussian heatmaps, reflecting the shape and direction features of objects in the detection.

However, these detectors usually demanded significant computational resources to reach satisfactory detection results, which hindered their deployment on intelligent terminals with limited computing power. To accelerate object detection algorithms, we introduce a guided HQ approach with a one-to-one self-teaching (OST) framework. This technique allows us to obtain a lightweight quantized network from high-performance neural networks, facilitating faster inference.

B. Deep Network Compression and Acceleration

Although the speed of one-stage detection networks is superior, their large model size and high computational complexity still requires exploration. Some researches focused on the design of a lightweight backbone. MobileNetV2 [32] utilized the depthwise separable convolutions to build a lightweight model. ShuffleNet [33] and SqueezeNet [34] also effectively reduced the memory footprint during inference and speed up the detection. In the literature, a potential direction of model compression was KD which concentrates on transferring knowledge from a heavy model (teacher) to a light one (student) to improve the light model's performance without introducing extra costs [35]. Whereas the KD enabled utilizing the larger network in a condensed manner, the pretraining of the large network requires extra substantial computation resources to prepare the teacher network [16]. The preparation of the pretrained teacher network is time-consuming and cost-consuming. The self-KD [15], [16], [17] can overcome this problem by distilling its own knowledge without prior preparation of the teacher network. Quantization was another way to compact the model directly and compress the ponderous network by using low-bit representation. Mixed-precision quantization method used different numbers of bits for a given data type to represent values in weight tensor. Many works [11], [36], [37] showed that the mixed-precision method is efficient for quantizing network layers that have different importance and sensitiveness for the bit-width. However, trivially applying quantization to CNNs usually led to inferior performance if the compression bit decreases to a low level.

In the remote sensing field, the distillation [38], [39] was gradually utilized in many tasks. Li et al. [40] proposed a relationship construction approach to enhance the learning of intraclass compactness and interclass dispersion. Yang et al. [14] proposed an adaptive reinforcement supervision distillation framework to promote the detection capability of the lightweight model. Yang et al. [41] introduced a category correlation and adaptive distillation method for accurate and compact cloud detection.

Although these distillation algorithms have brought many improvements in accuracy to remote sensing models, there is still great potential for model compression. In our article, we first combine the synergy of the quantization and distillation in remote sensing to compress detection models with significant compression capabilities without the need for a large pretrained model, thereby saving time and cost.

III. Network Architecture

In this section, we first revisit conventional KD and describe the proposed GHOST framework in Section III-A. Then, we present the details of the inspired HQ algorithm (Section III-B), and this quantization training process is guided by an OST method illustrated in Section III-C.

A. Overview

KD is a widely applied method that can be expressed as a knowledge transformer from teacher to student. Given a teacher model T and a student model S, the $\mathbf{X}$ denotes the data examples of models. Here, they can be the same for the teacher and the student model. In general, the KD machine can be uniformly expressed as

$$\min \mathcal{L}_{\mathrm{KD}} = \min \sum_{\mathbf{x}_i \in X} \mathcal{L}(\mathrm{T}(\mathbf{x}_i), \mathrm{S}(\mathbf{x}_i)) \tag{1}$$

where $\mathcal{L}$ is the loss function that penalizes the differences between the teacher and the student. $\mathbf{x}_i$ denotes the one sample from $\mathbf{X}$.

The student model is commonly designed in a small size to achieve the purpose of model compression in which the performance of the student can chase the teacher but consumes a computing-friendly resource. Distillation improves model performance, robustness, and generalization for tasks requiring high precision and high performance [13], while quantization can reduce the model size and computational complexity and is suitable for small devices [6]. This indicates that there is significant room for improvement in KD-based quantization algorithms [36] that combine quantization and distillation.

We aim to develop a novel and generic baseline network with a focus on the learnable knowledge characteristics, making it well-applicable to the highly accurate and fine object

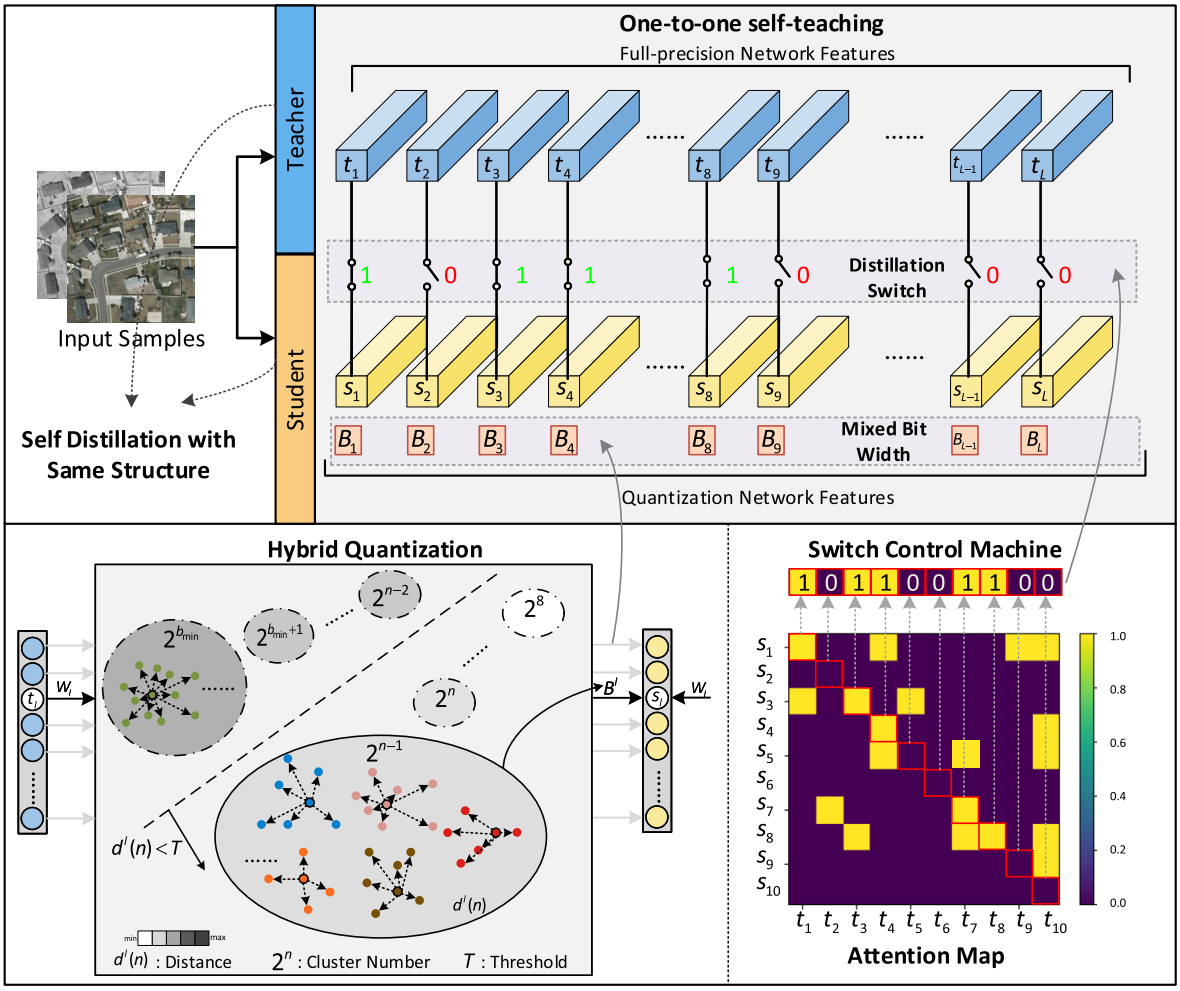

Fig. 2. Overview of our proposed framework. An attention-based model determines similarities between the teacher and student features. Knowledge from each teacher feature is transferred to the student with similarities identified by the SCM by self-distillation with the same structure. The mixed bit-widths of the student network for quantization are based on the search results of the full-precision weights research space of the teacher network in the same layer.

detection of RS images with less computational costs. The key to model quantization with knowledge learning is to reduce the discrepancy, which can be punished by distance or angle loss function between full-precision model P (teacher) and low-precision model Q (student) through optimizing Q, which can be expressed as

$$\mathrm{Q}^* = \min_{\mathrm{Q}} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathcal{L}(\mathrm{P}(\mathbf{x}_i), \mathrm{Q}(\mathbf{x}_i)). \tag{2}$$

The weights of the teacher are frozen without gradient propagation when the teacher network guides the training of the student network. Based on the above presentation, we design an effective teacher–student distillation framework called GQSD which can be represented as

$$\begin{aligned} &\min \mathcal{L}_{\mathrm{KD}} = \min \sum_{\mathbf{x}_i \in x} \mathcal{L}(\mathrm{P}(\mathbf{x}_i), \mathrm{Q}(\mathbf{x}_i)) \\ &\text{s.t. } \mathbf{W}_{\mathrm{Q}} = \mathbf{W}, \quad \mathbf{B}_{\mathrm{Q}} = \mathbf{B}. \end{aligned} \tag{3}$$

Specifically, a fully accurate network plays a pivotal role as the foundation for the teacher model, with trained weights $\mathbf{W}$ and bit-width settings $\mathbf{B}$ calculated within the weight constraint space of the network. The initial weights of the quantized model $\mathbf{W}_{\mathrm{Q}}$ and bit-width $\mathbf{B}_{\mathrm{Q}}$ are derived from the full-precision network, which guides the quantization process and facilitates the discovery of essential knowledge related to specific features by the teacher.

The loss function is designed on the features of the intermediate layer or soft object predictions of region proposals in the classification head to imbue the student with dark knowledge inside the teacher [42], [43]. Intermediate features are employed to enhance the performance of KD to extend the proposed method to various object prediction representations, such as oriented bounding box (OBB) and quadrilateral bounding box (QBB).

As shown in Fig. 2, we propose a GHOST framework that concludes an HQ module and an OST module. The mixed bit-widths of the student network for quantization are based on the search results of the full-precision weights research space of the teacher network in the same layer. We design an SCM as R to generate an attention map that gains intermediate feature similarities between intermediate features of the teacher and student to improve the performance in KD. The SCM controls the distillation switch and determines which knowledge should be delivered dynamically. Knowledge from each teacher feature is transferred to the student with similarities identified by SCM by self-distillation with the same structure. With a

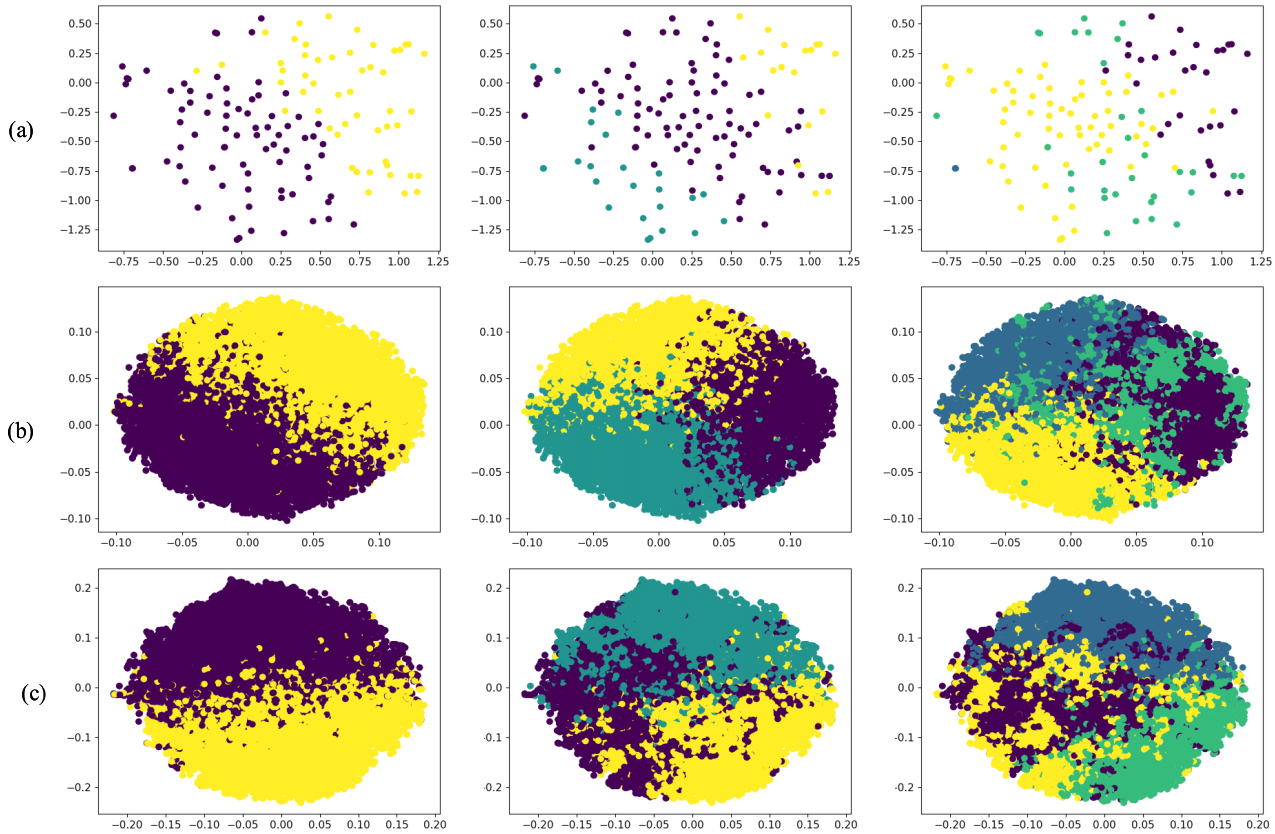

Fig. 3. (a) Distance between different categories is relatively far which indicates that the weight distribution is dispersed and complicated in the initial layer. This is due to the fact that the color and texture features, which are detailed and multifarious, are captured in the shallow layer. As the layer propagates forward (b) and (c), the convolution weight becomes converging gradually.

pretrained full-precision model as initial weight, the quantization and distillation processes are conducted simultaneously to ultimately obtain a small, lightweight model with little loss of accuracy. The details of the modules will be described as follows.

B. Hybrid Quantization

Powerful deep networks normally benefit from large model capacities but induce high computational and storage costs. Model quantization is a promising approach to compress deep neural networks, making it possible to be deployed on edge devices. The quantization operator divides the weight into different fixed values by a quantization function which can be regarded as a cluster of convolution kernels in substance. The different scale weights are clustered to a certain value.

To illustrate this intuition explicitly, a SuperYOLO [44] network model, which consists of 60 convolutional layers, is trained based on the VEDAI dataset. After training, test samples are fed into the model. The convolution weight is first clustered into different categories by k-means and then transformed into 2-D by t-SNE [45] to realize the visualization. As shown in Fig. 3, the convolution kernel weights in (a) 0th, (b) 26th, and (c) 52nd convolutional layer are clustered in different numbers. In Fig. 3(a), the distance between different categories is relatively far, which indicates that the weight distribution is dispersed and complicated in the initial layer. This is due to the fact that the color and texture features, which are detailed and multifarious, are captured in the shallow layer, and as the layer propagates forward [Fig. 3(b) and (c)], the convolution weight becomes converging gradually. In other words, the semantic features in the deep layer are more robust and condensed so that the clustering categories of weights can be relatively reduced with the deepening of the network layers. When a larger bit-width is set in the shallow layers of the network, the quantized network can retain more detailed information in these layers, which is beneficial to the detection of small objects in remote sensing images.

Based on this finding, the HQ approach is introduced to determine the optimal bit-width definition in the weight value space. To this end, a hyperparameter T is first defined as a threshold to restrict the search space and control the compression rate of the quantization model. The bit-width search strategy of each convolution layer is then described as follows:

$$B^l = \text{argmax} \ (d(n^l)) \quad \text{s.t. } d(n^l) < T. \tag{4}$$

The notation l is used to represent the lth convolution layer, while the function $d(\cdot)$ measures the clustering extent of the weight cluster in each layer. The clustering category is set to 2^n corresponding to the 2^n number at the n bit-width quantization. The symbol B^l represents the selected superior and adaptive bit-width of the lth convolution layer. The aim of this definition is to find the limited maximum clustering extent $d(n^l)$, which corresponds to the minimum clustering categories of 2^n (indicating the minimum bit-width n) for each layer.

The initial bit-width is set to 8, and the (4) can be deduced accordingly

$$B^l = \min(n | d^l(n) < T), \quad n = b_{\min}, b_{\min}+1, \ldots, 8 \tag{5}$$

where the $b_{\min}$ is a limit on the minimum bit-width allowed in the quantization process, and its value is set in advance. Initially, the clustering extent $d^l(n)$ is computed for all bit-widths $n = b_{\min}, b_{\min}+1, \ldots, 8$. Subsequently, the smallest bit-width n that satisfies the condition $d^l(n) < T$ is identified and set as the bit-width B^l for the corresponding convolution layer. Essentially, this entails searching for the minimum bit-width that achieves a clustering extent lower than the threshold T. This process is then repeated for each convolution layer, resulting in an optimized bit-width for each layer. The activation following this convolution layer keeps the same bit-width.

To determine the final bit-width for quantizing the lth convolution layer weight, we use a distribution distance defined as follows:

$$d^l(n) = \frac{1}{M} \sum_{j=0}^{M} \sum_{i=0}^{2^n} \left(w_{ij}^l - c_i^l\right)^2 \tag{6}$$

where M is the total number of kernel weights, given by $M = C_{\text{in}} \times C_{\text{out}} \times K \times K$, and C_{in}, C_{out}, and K are the input channels, output channels, and kernel size of the convolution layer, respectively. The k-means++ algorithm is applied to all weight values of each convolution layer for different cluster categories. c_i^l and w_{ij}^l are the cluster centers and samples, respectively. The HQ of the whole network definitely can be collected as follows:

$$\mathbf{B} = [B^1, B^2, \ldots, B^L] \tag{7}$$

where the L is the total number of convolution layers, and the bit-width decreases progressively as the network propagates forward.

Taking the distance threshold $T = 50$ as an example, Fig. 4 demonstrates the judgment results of bit-width for each convolution layer. It can be indicated that the values of bit-width progressively decrease with the deepening of the network layer. In addition, the bit-width of the convolution layer before the detection process may be relevantly large to maintain more location discrimination information.

We use a simple-yet-effective quantization method which refers to [8] for both weights and activations. The uniform quantization function $q(\cdot)$ is defined as

$$q(v, k) = \frac{1}{2^k - 1} \text{round}((2^k - 1)v) \tag{8}$$

where v is a real number indicating the full-precision (float32) value, $v \in [0, 1]$. The output $q(v, k)$ of quantization function is a k bits real number, $q(v, k) \in [0, 1]$. The quantization calculations of lth convolution layer weight and activation are defined as follows:

$$\bar{w}_{ij}^l = 2q\left(\frac{\tanh\left(w_{ij}^l\right)}{2\max\left(\left|\tanh\left(w_{ij}^l\right)\right|\right)} + \frac{1}{2}, B^l\right) - 1 \tag{9}$$

$$\bar{a}_i^l = q\left(a_i^l, B^l\right). \tag{10}$$

Algorithm 1 HQ Method

Input: The weights of l^{th} certain convolution layer $\mathbf{W} \in \mathbb{R}^{H \times W \times K \times K}$, The manual distance threshold T and the minimum bit-width $b_{\min}$.

Output: The bit-width of the current convolution layer and activation B^l.

1: Initialize the B^l as 8.
2: **for** n in range ($b_{\min}$,8) **do**
3: Cluster weights into 2^n clusters via the kmeans++ algorithm and then get the centers c_i and samples w_{ij} of the i^{th} cluster.
4: Calculate the distribution distance according to (6).
5: Update the bit-width B^l by (5).
6: Complete the quantization for the convolution layer weight and activation by (9) and (10), respectively.

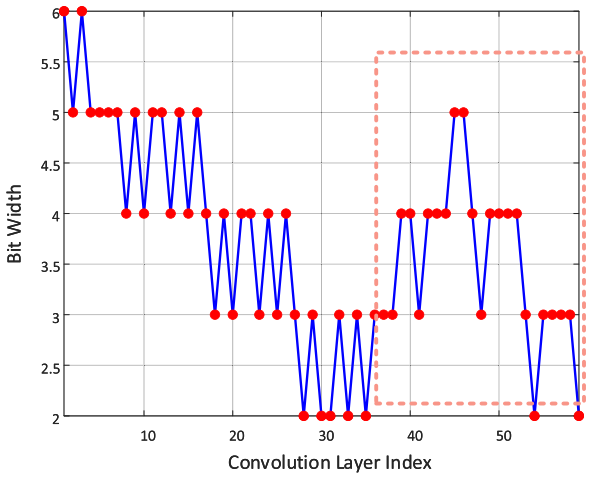

Fig. 4. Bit-width results of each convolution layer at the threshold $T = 50$. The values of bit-width progressively decrease with the deepening of the network layer. In addition, the bit-width of the convolution layer before the detection module may be relevant large to maintaining more location discrimination information.

The activation a_i^l is the range in $[0, 1]$ determined by a bounded activation function, while the weight w_{ij}^l is not restricted in a limit boundary. Here, the quantization result of weight $\bar{w}_{ij}^l$ is the range in $[-1, 1]$, and the quantization result of activation $\bar{a}_i^l$ is the range in $[0, 1]$. The Algorithm 1 clarifies the process of the HQ method. As described in [8], the first and last layers in the network are sensitive to performance during the process of quantization. Based on this intuition, the last detection layer keeps intact to avoid potential degradation of detection performance.

C. One-to-One Self-Teaching

Previous mixed quantization approaches pay more attention to the bit-width selection [7], which costs a lot of resources to obtain the optimal decision. Our HQ method can make a quick decision with less computation cost. The performance loss is mitigated through distillation. In general, previous distillation algorithms are a full precision network, so the network weights are in the same order of magnitude, and the feature maps generated by the teacher network and the student network

are similar. However, in the case of a quantized network, the student network's feature map will exhibit weight information loss due to an increase in zero content. As a result, differences may arise between the feature map of the teacher network and that of the student network, making it difficult for the teacher network to directly guide the student network's feature layer. Therefore, we proposed an OST to conquer this question.

Our distilled decision algorithm is designed based on the SCM mechanism, allowing the teacher and student networks to adopt a network-adaptive selection strategy for distilled decisions across different depth layers. In this mechanism, the distillation switch (DS) is inclined to retain detailed information at the shallow layers for object detection and operates according to adaptive selection when transferring knowledge from favorable teacher network layers. This selection strategy is conducive to detecting small objects in remote sensing images and reduces feature extraction and interference from complex background information through this distillation method.

Let $\mathbf{s} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_l]$ represent a set of multiscale feature maps for the student network and $\mathbf{t} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_l]$ for the teacher. To calculate the attention map similar to [46] between the student feature and teacher feature, we define that each teacher feature generates a query $\mathbf{q}_i$, and each student feature produces a key $\mathbf{k}_j$

$$\mathbf{q}_i = \mathbf{W}_i \cdot \mathrm{GAP}(\mathbf{t}_i) \tag{11}$$

$$\mathbf{k}_j = \mathbf{W}_j \cdot \mathrm{GAP}(\mathbf{s}_j) \tag{12}$$

where GAP$(\cdot)$ denotes global average pooling. The parameters $\mathbf{W}_i$ and $\mathbf{W}_j$ are the linear transition matrices for the ith query and the jth key. The teacher and student sequences are concatenated as

$$\mathbf{q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_i, \ldots, \mathbf{q}_L] \tag{13}$$

$$\mathbf{k} = [\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_j, \ldots, \mathbf{k}_L]. \tag{14}$$

The attention map that reveals the internal relationships between teacher and student features is defined as

$$\mathbf{a} = (\mathbf{q} \cdot \mathbf{k}^{\mathbf{T}})/\sqrt{d} \tag{15}$$

where d is the dimensionality of $\mathbf{q}$ and $\mathbf{k}$.

Here, we introduce the Gumbel-Sigmoid trick [47] to convert the values greater than the threshold to 1 and the rest to 0. The transformation is defined by the equation

$$\mathbf{A} = \frac{1}{1 + e^{-\frac{\log(\mathbf{a}) + \mathbf{G}}{\tau}}} \tag{16}$$

where τ is a temperature parameter that governs the degree of smoothing, and the Gumbel noise $\mathbf{G}$ is represented as $-\log(-\log(\mathbf{u}))$, where $\mathbf{u}$ is a vector of samples drawn from the uniform distribution on the interval [0, 1]. The temperature parameter τ regulates the smoothness of $\mathbf{A}$.

With a better attention map to mining the internal correlation of features, we generate a DS mask that can automatically determine whether to transfer the information from the teacher to the student at the same site in the network

$$\alpha = \mathrm{Diag}(\mathbf{A}) \tag{17}$$

where SCM digs out the diagonal elements from the attention map matrix $\mathbf{A}$. We devise the self-feature distillation loss as follows:

$$\mathcal{L}_{\mathrm{s}} = \sum_{i=0}^{m} \alpha_i \| \mathrm{CAP}(\mathbf{t}_i) - \mathrm{CAP}(\mathbf{s}_i) \|_2. \tag{18}$$

Here, CAP represents channel-wise average pooling, and m is the total number of features utilized for distillation.

Finally, the distillation loss terms are combined with detection loss and minimized in an end-to-end manner as

$$\mathcal{L}_{\mathrm{total}} = \beta \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{dec}} \tag{19}$$

where $\mathcal{L}_{\mathrm{dec}}$ includes the objectness, location, and classification. The hyperparameter β indicates the impact balance between the detection and distillation.

IV. Experimental Results

In this section, we evaluate the proposed method on the four datasets for remote sensing object detection. We first demonstrate the experiment setup, including the introductions of datasets and networks, implementation details, and evaluation metrics. Then, we report the performance of our method on a dataset in detail, and the mean average precision and compression ratio are calculated to measure the comprehensive performance in the accuracy and computation cost.

A. Dataset Description

The publicly available dataset VEDAI [48], (DOTA [49], NWPU [50], and DIOR [51]) are utilized in experiments to verify the generation of our proposed algorithm.

1) VEDAI: The VEDAI is a multimodal (RGB and IR) dataset consisting of 1246 smaller images cropped from the much larger Utah Automated Geographic Reference Center (AGRC) dataset. Each image collected from the same altitude in AGRC has approximately 16 000 × 16 000 pixels, with a resolution of about 12.5 × 12.5 cm per pixel. The main scenes of VEDAI include grass, highway, mountains, and urban areas. The VEDAI dataset contains 11 classes of vehicles. In this work, we operate on the 512 × 512 images of eight vehicle classes. We do not consider classes with fewer than 50 instances in the dataset, such as planes, motorcycles, and buses. The final selected classes include car, pickup, camping, truck, other, boat, and van. The VEDAI dataset is devised to tenfold cross-validation. In each split, 1089 images are used for training, and another 121 images are used for testing. Our ablation experiments are conducted on the first fold of the dataset, while the comparisons with previous methods are performed on the ten folds by averaging their results.

2) DOTA: The DOTA dataset was proposed by Xia et al. [49] for object detection of remote sensing. It contains 2806 large images and 188 282 instances, which are divided into 15 categories. The size of each original image is 4000 × 4000, and the images are cropped into 1024 × 1024 pixels with an overlap of 200 pixels in the experiment. We select half of the original images as the training set, 1/6 as the validation set, and 1/3 as the testing set. The size of the image is fixed to 512 × 512.

TABLE I
TRAINING STRATEGY

Dataset	Image Size	Batch Size	Learning Rate	Epoch
VEDAI	512	2	0.01	300
DOTA	512	16	0.01	100
NWPU	512	8	0.01	150
DIOR	512	16	0.01	150

3) NWPU VHR-10: The dataset of NWPU VHR-10 was proposed by Cheng et al. [50]. It contains 800 images, of which 650 pictures contain objects, so we use 520 images as the training set and 130 images as the testing set. The dataset contains ten categories, and the size of the image is fixed to 512 × 512.

4) DIOR: The DIOR dataset was proposed by Li et al. [51] for the task of object detection, which involves 23 463 images and 192 472 instances. The size of each image is 800 × 800. We choose 11 725 images as the training set and 11 738 images as the testing set. The size of the image is fixed to 512 × 512.

B. Implementation Details

1) Networks: To demonstrate superior performance, SuperYOLO [44] is tested as a teacher model with our method. For the multimodal VEDAI dataset, the number of convolution layers is 47, including one detection layer on a small scale. For a single modal dataset (DOTA, NWPU, and DIOR), the number of convolution layers is 61, including three detection layers on the small, medium, and large scales. To verify the superiority of the GHOST proposed in this article, we selected some generic methods for comparison:

One-stage algorithms (YOLOv3 [52], YOLOv4 [21], YOLOv5 [22], YOLO-Fine [53], YOLOFusion [54], CFT [55], SuperYOLO, FCOS [24], ATSS [25], RetainNet [23], GFL [56]); Two-stage algorithms (Faster R-CNN [20]); Lightweight algorithms (MobileNetV2 [32] and ShuffleNet [33]); Distillation-based algorithm (ARSD [14]); Remote sensing designed algorithms (FMSSD [57], O2DNet [58]), and S2A-Net [59]).

2) Training Strategy: Our proposed framework is implemented in PyTorch and runs on a workstation with an NVIDIA A100-SXM4-80GB GPU. We use different training strategies for different datasets, and the detail is illustrated in Table I. In addition, data is augmented with hue saturation value (HSV), multiscale, translation, left-right flip, and mosaic. The augmentation strategy is canceled in the test stage. The standard stochastic gradient descent (SGD) is used to train the network with a momentum of 0.937 and a weight decay of 0.0005 for the Nesterov accelerated gradients utilized. The learning rate is set to 0.01 initially. All the baseline training process is completed from scratch without any pretrained model while the GHOST is carried on the baseline model. In the test stage, the IoU threshold of non-maximum suppression is 0.6 on NWPU VHR-10 and VEDAI, and it is 0.4 on DOTA and DIOR.

3) Evaluation Metric: For the detection result, the IoU is defined as the ratio of the intersection and union of two boxes. During the evaluation, according to the IoU of predicted boxes and ground truths, each sample will be assigned attributes: true positive (TP) for correctly matching, false positive (FP) for wrongly predicting the background as an object, and false negative (FN) for the undetected object. During the evaluation, all the detection boxes are sorted in order of confidence score from high to low and then traversed. In the traversed process, the calculations of the precision and recall metrics can be defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{21}$$

The precision and recall are correlated with the commission and omission errors, respectively. The AP values use an integral method to calculate the area enclosed by the precision–recall curve and coordinate axis of all categories. Hence, the AP can be calculated by

$$\text{AP} = \int_0^1 p(r)dr \tag{22}$$

where p denotes precision, and r denotes recall. The mAP is a comprehensive indicator obtained by averaging APs for all classes. Moreover, we choose bit-operations (BOPs) count [60] and parameters to measure the compression performance. The BOPs of convolution are calculated as follows:

$$\text{BOPs}_l = c_{l-1} \times c_l \times w_l \times h_l \times k_w \times k_h \times b_{w,l} \times b_{a,l-1}. \tag{23}$$

The h_l, w_l, and c_l are the height, width, and several channels of the lth layer output feature map, respectively. $b_{w,l}$ and $b_{a,l}$ denote lth layer weight and activation bit-widths. The k_h and k_w are the convolution kernel sizes. The parameters (params) are defined as

$$\text{params} = \frac{c_{l-1} \times c_l \times k_h \times k_w \times b_{w,l}}{8 \text{ bits}}(B). \tag{24}$$

C. Ablation Study

In this section, we conduct the ablation experiments of our GHOST framework. We explore how each module (HQ and OST) compresses the model and promotes the performance of the small student model. Besides, the experiments of different distillation algorithms and distillation hyperparameters' optimization are carried out. We conduct ablation experiments on the dataset of VEDAI for object detection.

1) Validation of HQ: Distribution distance HQ can integrate device n-bit settings for the network, so we experiment with such variations of integrated hybrid n-bit quantization. To analyze performance differences, we compare fixed DoReFa-Net [8] and various hybrid DoReFa-Net quantization methods on the SuperYOLO detection network. As illustrated in Table II, the HQ is the proposed hybrid quantization method, and the ·W·A presents the traditional unified bit-width quantization algorithm except 32W32A represents the full precision network. Max and Min denote the maximum and

TABLE II
COMPARISON RESULT OF THE TRADITIONAL QUANTIZATION METHOD AND OUR MIXED-BIT QUANTIZATION AND WE USE THE SAME ABBREVIATION IN THE FOLLOWING SECTIONS ON THE VEDAI DATASET

Bit Width	T	Min	Max	mAP50	mAP	Params(MB)	BOPs(G)
32W32A	-	32	32	**80.93**	**50.80**	**19.30**	**17023.76**
8W8A	-	8	8	75.87	47.08	4.83	1201.10
HQ	0.1	7	8	**79.57**	**48.55**	**4.34**	**1123.12**
6W6A	-	6	6	75.92	45.63	3.63	727.04
HQ	4	3	8	**78.42**	**46.32**	**2.49**	**691.88**
4W4A	-	4	4	74.14	44.85	2.43	382.91
HQ	70	3	8	**75.76**	**46.00**	**1.87**	**371.23**
2W2A	-	2	2	45.08	25.21	1.22	168.69

TABLE III
VALIDATION RESULT OF THE SELF-DISTILLATION METHOD IN THE DIFFERENT MODEL SIZE ON THE VEDAI DATASET

HQ	T	Min	Max	OST	mAP50	mAP	Params(MB)	BOPs(G)
✓	0.1	2	8		78.00	46.59	4.12	1107.95
✓	0.1	2	8	✓	**78.59**	**47.14**	4.12	1107.95
✓	4	2	8		78.42	46.32	2.49	691.88
✓	4	2	8	✓	**79.46**	**48.57**	2.49	691.88
✓	52	2	8		74.64	44.29	1.46	377.54
✓	52	2	8	✓	**76.99**	**46.91**	1.46	377.54

TABLE IV
COMPARISON WITH SOTA DISTILLATION METHOD FOR DETECTORS ON THE VEDAI DATASET

Distillation	HQ	mAP50	mAP
-	✓	78.00	46.59
ZAQ [63]	✓	76.12	46.48
AFD [46]	✓	75.86	45.44
ReviewKD [64]	✓	77.75	47.35
OST	✓	**78.59**	**47.14**

TABLE V
MAP COMPARISONS OF DIFFERENT β VALUE ON THE VEDAI DATASET

	β	0	100	200	300	400	500
T	0.1	46.59	47.17	48.26	46.72	**49.17**	46.29
	4	46.32	48.57	47.05	47.96	47.29	**49.05**
	52	44.29	46.91	44.20	45.14	46.48	**47.03**

minimum bit-width in the quantization model. HQ enables the quantization model to preserve the significant information to achieve the minimal accuracy loss possible. According to Table II, the HQ module achieved higher detection accuracy, with values of 79.57%, 78.42%, and 75.76%, respectively, compared to fixed quantization at different computation orders. The improvement in mAP50 is significant, with increases of 3.7%, 2.5%, and 1.62%. Additionally, the HQ requires less memory than fixed quantization at different quantize levels, with 4.34, 2.49, and 1.87 MB parameters, which are 0.49, 1.14, and 0.56 MB less than fixed quantization at 8W8A, 6W6A, and 4W4A quantize levels, respectively. Furthermore, the computation cost of the HQ is also lower, with 1123.12, 691.88, and 371.23 G BOPs, which are 77.98, 35.16, and 11.68 G BOPs less than fixed quantization at 8W8A, 6W6A, and 4W4A quantize levels, respectively. The HQ achieves better accuracy in the VEDAI dataset than the accuracy of fixed quantization, costing fewer computation resources (parameters and BOPs). The accuracy of the detection model decreases as the number of quantization bits decreases. This can be attributed to the fact that the reduction in the number of model parameters results in the loss of some information at some important layer, leading to a decline in accuracy. The impact is particularly significant when using only two bits, which results in an accuracy of only 45.08% mAP50. HQ can effectively address this issue by allocating larger bit-widths to the most important layers, thereby reducing the amount of information loss during quantization. Hence, it is significant to design an HQ framework that can reduce information loss with a small number of acceptable parameters to improve performance.

2) Effect of OST Module: After the HQ module has been added to the network, we also adopt OST within the three-quantization scale. Table III is based on SuperYOLO, which is used as the teacher network and student network simultaneously. The experiments are carried out on the VEDAI dataset. The distillation algorithm does not introduce extra parameters and computation. The results demonstrate that the OST module is effective in restoring the detection performance of the typical network that has been quantized. Specifically, the mAP50 and mAP of the quantized network with OST are improved by 0.59%, 1.04%, 2.35%, and 0.55%, 2.25%, and 2.62%, respectively, at different computation orders.

3) Comparison With the SOTA Distillation Method: In addition, Table IV shows the comparisons between the proposed OST module and existing distillation frameworks. The HQ model ($T = 0.1$, Min $= 2$, and Max $= 8$) is used in all the methods to search the distillation method for an adaptive quantization network. OST achieves superior performance mAP50 and mAP in the field of remote sensing with values of 78.59% and 47.14%, respectively, under the premise of the same computation cost, while the ZAQ, AFD, and ReviewKD lead to an accuracy degradation of the quantization network. And also, it can be proved that OST distillation is a benefit for the guidance between the full-precision model and the quantization model.

4) Hyperparameters' Optimization: As presented in Section III-C, β is the distillation weights of the OST module, so we compare the performance of the distillation in the different weights, which are shown in Table V. We conduct the hyperparameter experiment on the VEDAI dataset on GHOST to find the best β. As shown in Table V,

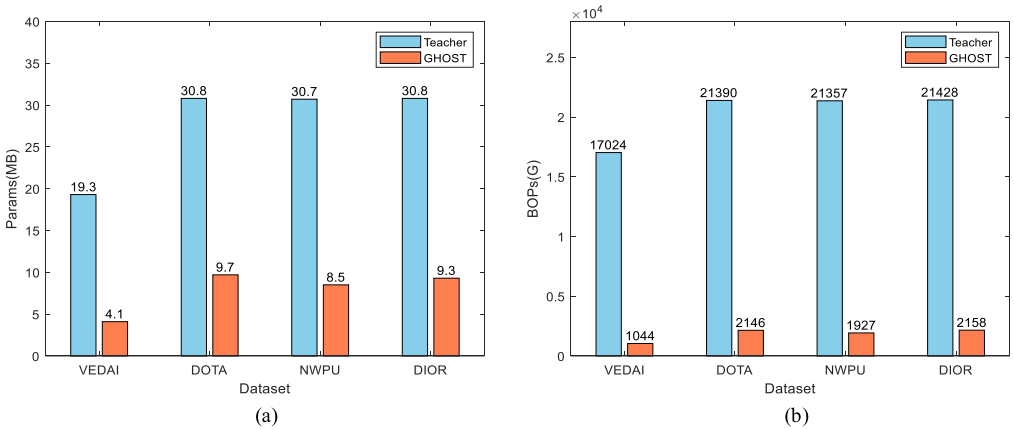

Fig. 5. Comparisons of the teacher model and lightweight model by parameters and BOPs on the four datasets (VEDAI, DOTA, NWPU, and DIOR). The BOPs and parameters of the lightweight model are smaller, and the inference speed is faster. (a) Params (MB). (b) BOPs.

the model reaches the best when $\beta = 400$ at the $T = 0.1$ and when $\beta = 500$ at the $T = 4$ or $T = 52$.

5) Lightweight Analysis: Owning to the GQSD design idea, our student model is very lightweight. We compare the GHOST and the teacher model regarding parameters and BOPs on the four datasets. As Fig. 5 shows, GHOST has smaller model parameters and fewer BOPs than the teacher model regardless of which dataset. Hence, our compression strategy for the lightweight model is more practical to be deployed on intelligent terminals.

6) Visualization, Analysis and Discussion About Distillation: To investigate the learned relation of information between the teacher and student model, we depict the attention maps using EigenCAM [61]. There are three important observations that align with our aforementioned analyses in Fig. 6.

1) The network acquires finer details surrounding the object at the shallow layers, such as the third and fifth modules depicted in the figure. These finer details are beneficial in detecting smaller objects. To ensure that knowledge is transferred effectively between the teacher and student networks at these layers, the α coefficient for distillation is assigned to 1.
2) The middle layers in the network, specifically the 8th–13th modules illustrated in the figure, learn about the context of the object with a broader surrounding area, and the holistic information that includes the object nearby. The network dynamically makes distillation decisions to preserve comprehensive information about the object while filtering out irrelevant background information. The network adapts its decision-making based on the distillation relationship between the teacher and student networks and retains the distillation relationship of the 10th and 12th modules, which contain the holistic information of the object while eliminating the distillation relationship of the 8th module, which focuses on the background information.
3) Moving to the deeper network layers, such as 16th–23rd modules in the figure, the network obtains high-level semantic information that is no longer object-specific. However, object-specific details are essential for the critical task of object localization in the detection network. To prioritize learning complex and abstract features for the student network, the network suppresses the influence of the teacher network in these layers by setting the α coefficient to 0.

D. Results and Discussions

In this section, we compare our lightweight model with other classic heavy object detection methods. As shown in Tables VI and VII, experiments on the four datasets prove the efficiency and efficacy of the GHOST framework. Not only does our lightweight have higher accuracy, but also it has strong information retention capability under extreme model compression.

1) VEDAI: Our GHOST achieves 71.31% mAP50 compared with other detectors, surpassing the one-stage series mentioned in Table VI. Our model achieves the lowest model parameters (4.1 MB) and BOPs (1044 G). It is evident from the results that SuperYOLO outperforms all other frameworks except YOLOFusion. Although YOLOFusion has slightly better results, it uses pretrained weights from MS COCO [65] and has approximately three times more parameters than SuperYOLO. On the other hand, YOLO-Fine performs well on a single modality but lacks the development of multimodality fusion techniques. The usage of GHOST can assist SuperYOLO in reducing the model size by 78.8% and computation cost by 93.9% at the expense of a 3.78% mAP50 decrease in accuracy. Nevertheless, even with this trade-off, GHOST still outperforms YOLOv3, YOLOv4, YOLOv5, and YOLO-Fine. The detection accuracy of GHOST is 0.81% mAP50 lower than CFT, but its model size and computational cost are only 0.5% of CFT.

2) DOTA-v1.0: As presented in Table VII, our GHOST achieves the optimal detection result (69.02% mAP50), and the model parameters (9.7 MB) and BOPs (2146 G) are much smaller than other detectors regardless of the two-stage, one-stage, anchor-free, or distillation-based method. We also compare two detectors designed for remote sensing imagery: FMSSD, O2DNet, and S2A-Net. Although these models have superior performance compared with our lightweight model, the huger parameters and BOPs seem to be a massive cost in

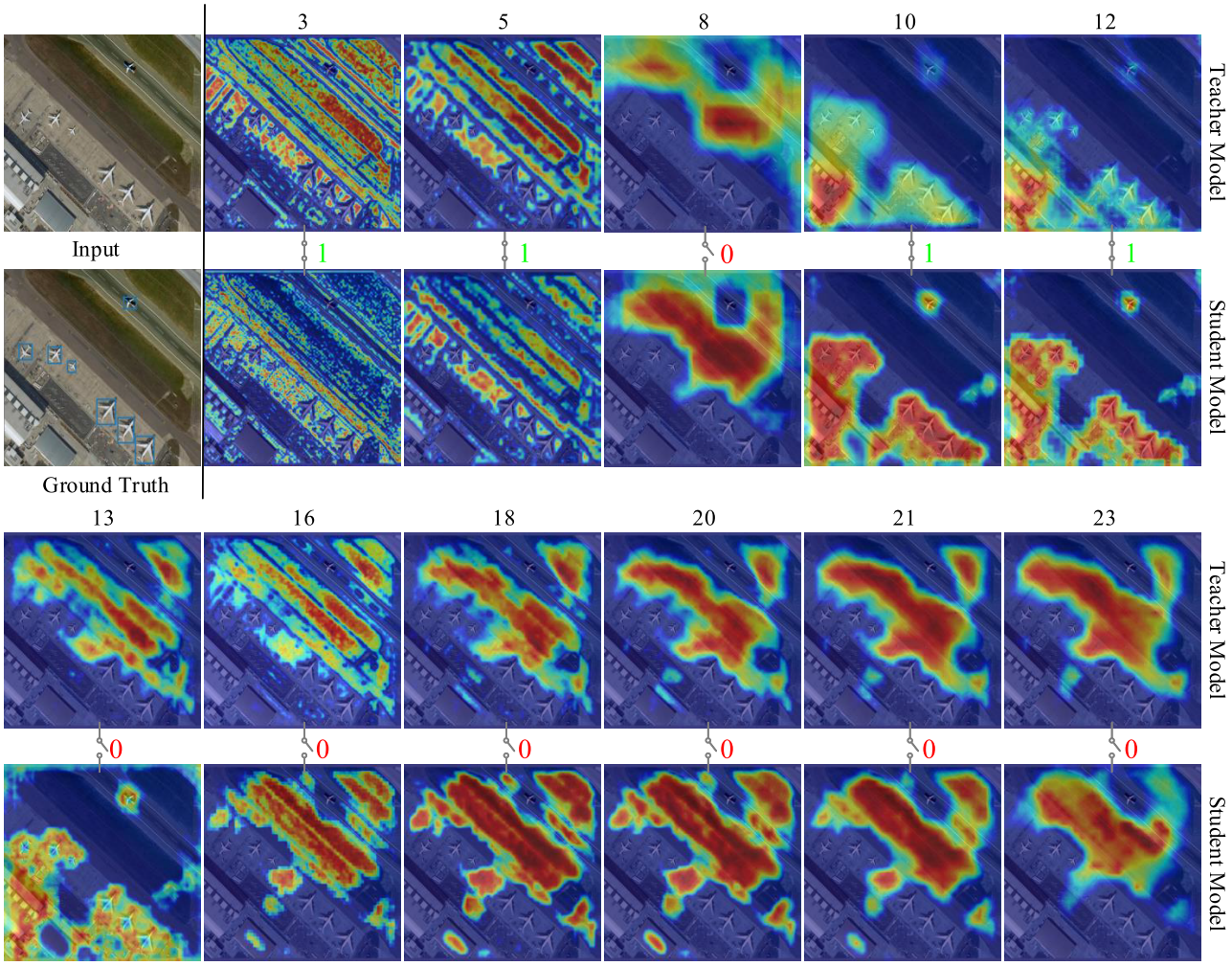

Fig. 6. Visualizations of learned attention map using EigenCAM [61], [62] of the teacher full-precision model and student quantization model. In each group of them, the numbers in black represent the nth module in the network. The numbers in green and red represent DS mask α which means whether to transfer the information from the teacher to the student at the same site in the network.

TABLE VI

PERFORMANCE OF DIFFERENT ALGORITHMS ON VEDAI ON THE TEN FOLDS TESTING SET. † AND ‡ REPRESENT THAT THE METHODS USE THE WEIGHT OF YOLOV5S AND YOLOV5L, RESPECTIVELY, ON THE MS COCO [65] DATASET AS THE PRETRAINING WEIGHT OF NETWORK. * REPRESENTS THE IMAGE SIZE USED IN NETWORK IS 640 × 640

Method	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP50	Params(MB)	BOPs(G)
YOLOv3 [52]	84.57	72.68	67.13	61.96	43.04	65.24	37.10	58.29	61.26	246	50,749
YOLOv4 [21]	85.46	72.84	72.38	62.82	48.94	68.99	34.28	54.66	62.55	210	39,076
YOLOv5s [22]	80.81	68.48	69.06	54.71	46.76	64.29	24.25	45.96	56.79	28	5,427
YOLOv5m [22]	82.53	72.32	68.41	59.25	46.20	66.23	33.51	57.11	60.69	84	16,599
YOLOv5l [22]	82.83	72.32	69.92	63.94	48.48	63.07	40.12	56.46	62.16	186	37,530
YOLOv5x [22]	84.33	72.95	70.09	61.15	49.94	67.35	38.71	56.65	62.65	349	71,301
YOLO-Fine [53]	79.68	74.49	77.09	80.97	37.33	70.65	60.84	63.56	68.83	-	-
YOLOFusion ∗† [54]	91.7	85.9	78.9	78.1	54.7	71.9	71.7	75.2	75.9	50	-
CFT∗‡ [55]	87.88	79.93	74.20	66.60	56.11	74.31	66.99	70.93	72.12	824	229,785
SuperYOLO [44]	91.13	85.66	79.30	70.18	57.33	80.41	60.24	76.50	75.09	19.3	17,024
SuperYOLO+GHOST	89.15	83.57	76.19	59.55	53.05	78.70	59.58	70.71	**71.31** -3.78	**4.1**	**1,044**

computation resources. Hence, our model has a better balance considering detection efficiency and efficacy. Compared to the distillation-based method ARSD (−3.37%), the GHOST obtains an even smaller accuracy gap (−0.97%) between

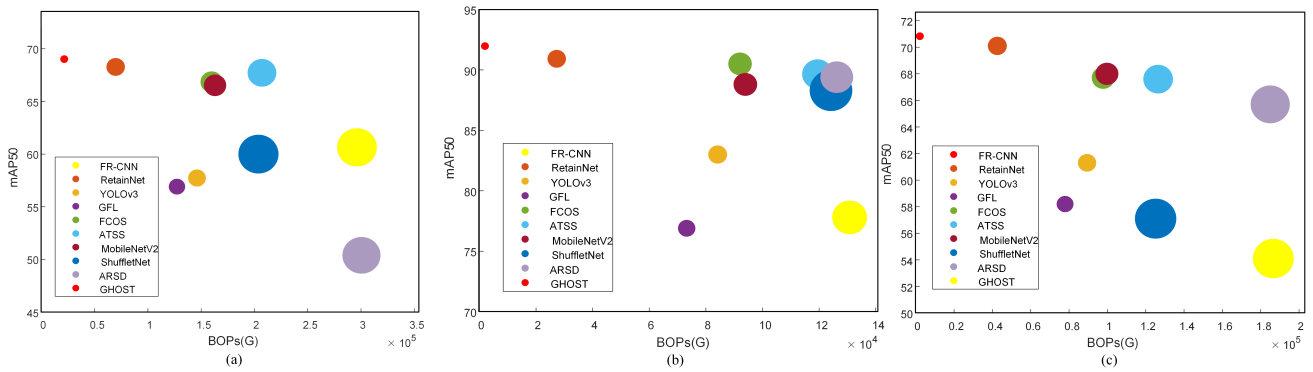

Fig. 7. Comparison of the efficiency between the current methods and our method on the three datasets. The bigger size of cycles represents costing more parameters. (a) DOTA. (b) NWPU. (c) DIOR.

TABLE VII
PERFORMANCE OF DIFFERENT ALGORITHMS ON DOTA, NWPU AND DIOR TESTING SET

	DOTA-v1.0			NWPU			DIOR		
Method	**mAP50**	**Params(MB)**	**BOPs(G)**	**mAP50**	**Params(MB)**	**BOPs(G)**	**mAP50**	**Params(MB)**	**BOPs(G)**
Faster R-CNN [20]	60.64	240	296,192	77.80	164	130,764	54.10	240	186,572
RetainNet [23]	50.39	221	300,400	89.40	145	126,228	65.70	221	184,954
YOLOv3 [52]	60.00	246	203,694	88.30	246	124,180	57.10	247	125,153
GFL [56]	66.53	76	163,000	88.80	76	93,931	68.00	76	99,768
FCOS [24]	67.72	126	207,001	89.65	127	119,429	67.60	127	126,474
ATSS [25]	66.84	75	159,754	90.50	75	92,057	67.70	75	97,792
MobileNetV2 [32]	56.91	41	127,221	76.90	41	73,205	58.20	41	77,926
ShuffleNet [33]	57.73	48	146,022	83.00	48	84,142	61.30	48	89,405
O2-DNet [58]	71.10	836	-	-	-	-	68.3	836	-
FMSSD [57]	72.43	544	-	-	-	-	69.5	544	-
S2A-Net [59]	**74.15**	142	100,346	-	-	-	-	-	-
Teacher for ARSD [14]	71.65	203	291,491	93.21	127	122,234	71.7	203	178,176
ARSD [14]	68.28 -3.37	52	69,662	90.92 -2.29	46	27,289	70.10 -1.60	52	42,598
SuperYOLO [44]	69.99	30.8	21,390	93.30	30.7	21,357	71.95	30.8	21,428
SuperYOLO+GHOST	69.02 -0.97	**9.7**	**2,146**	**91.97 -1.33**	**8.5**	**1,927**	**71.53 -0.42**	**9.3**	**2158**

the student and teacher networks. It demonstrates that our GHOST method can transfer sufficient knowledge to guide the learning of the student model, and the misunderstanding between both models can be reduced by the SCM training strategy.

3) NWPU: We compare the results of our method with other approaches on the NWPU dataset. As shown in Table VII, our GHOST obtains the best result (91.97% mAP50) with the smallest amount of model parameters (8.5 MB) and the fewer BOPs (1927 G).

4) DIOR: As illustrated in Table VII, our GHOST achieves the optimal detection result (71.53% mAP50), and the model parameters (9.3 MB) and BOPs (2158 G) are much smaller than other detectors regardless of the two-stage, one-stage, anchor-free lightweight, distillation-based methods. It reveals the solid ability for compress models and the power capacity of object detection in remote sensing imagery. The accuracy of the student GHOST is only a bit less 0.4% than the teacher network, compared with ARSD (1.6%).

To show the performance of our algorithm more intuitively, we compare the mAP50, parameters, and BOPs of various algorithms in Fig. 7. It can be evident that GHOST has a better trade-off between performance and lightweight. The visualization results on the three datasets are illustrated in Fig. 8, in which we can see that the GHOST can easily detect objects at different scales.

Furthermore, we discuss the generation in the different object representations. As described in (3), the proposed distillation loss function is designed on the features of the intermediate layer without being limited by the target representation. To validate the generation in the different representation methods, such as OBB, we applied our proposed method to

Fig. 8. Three sets of visualization results. (a) DOTA. (b) NWPU. (c) DIOR.

TABLE VIII
COMPARISON IN THE DIFFERENT OBJECT REPRESENTATION (OBB) ON THE DOTA DATASET

Method	GHOST	mAP50	mAP	Params(MB)	BOPs(G)
S2A-Net [59]		74.15	40.68	142.8	100,346
S2A-Net [59]	✓	73.61	40.76	67.9	58,253

TABLE IX
COMPARISON OF TIME COST OF SUPERYOLO METHOD WITHOUT AND WITH GHOST PER 512 × 512 IMAGE

Method	SuperYOLO	SuperYOLO+GHOST
time (ms)	38.127	**30.403**

the S2A-Net [59] as shown in Table VIII. GHOST slightly reduces the mAP50 score by 0.54 but reduces half the number of parameters from 142.8 to 67.9 MB and half the number of BOPs from 100 346 to 58 253 G, making the network more efficient. The results demonstrate that incorporating the GHOST module can significantly reduce the model size and computation cost without sacrificing the detection performance. Overall, the proposed method achieves competitive performance with fewer parameters and BOPs, demonstrating its effectiveness for object detection with OBB representations. We also test the real speed of our implementation when quantizing SuperYOLO with GHOST on GPU. As shown in Table IX, the inference speed using GHOST is much faster than that of the SuperYOLO full-precision method.

V. CONCLUSION

In this article, we propose a GHOST framework for a lightweight object detection method in remote sensing imagery. We first design a GQSD structure which is not only a training technique to preserve model performance but also a method to compress and accelerate models. Although most of the previous research focuses on knowledge transfer among different models, we believe that inside distillation is also very promising. Second, we propose an HQ that captures the optimal bit-width selection based on an adaptive way in the weight value research space to break the limit of the fixed quantization model accuracy. Third, the proposed OST module gives the student network of self-judgment through an SCM that accurately handles the knowledge transformation. It can dynamically discriminate the wrong guidance and mine the effective knowledge from the teacher. The experiments based on the VEDAI, DOTA, NWPU, and DIOR datasets certify that our GHOST achieves superior performance compared with other detectors. It can well balance the tradeoff between accuracy and specific resource constraints.

REFERENCES

[1] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[2] J. Lei, G. Yang, W. Xie, Y. Li, and X. Jia, "A low-complexity hyperspectral anomaly detection algorithm and its FPGA implementation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 907–921, 2021.

[3] G. Yang et al., "Algorithm/hardware codesign for real-time on-satellite CNN-based ship detection in SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5226018.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[5] J. Luo and J. Wu, "Neural network pruning with residual-connections and limited-data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1455–1464.

[6] R. Li, Y. Wang, F. Liang, H. Qin, J. Yan, and R. Fan, "Fully quantized network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2810–2819.

[7] Z. Wang, H. Xiao, J. Lu, and J. Zhou, "Generalizable mixed-precision quantization via attribution rank preservation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5291–5300.

[8] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 2016, *arXiv:1606.06160*.

[9] G. Wang, Y. Ge, and J. Wu, "Distilling knowledge by mimicking features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8183–8195, Nov. 2022.

[10] X. Dai et al., "General instance distillation for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7842–7851.

[11] H. Yang, L. Duan, Y. Chen, and H. Li, "BSQ: Exploring bit-level sparsity for mixed-precision neural network quantization," 2021, *arXiv:2102.10462*.

[12] Z. Liu, Z. Shen, M. Savvides, and K.-T. Cheng, "ReActNet: Towards precise binary neural network with generalized activation functions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 143–159.

[13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

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

[15] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3713–3722.

[16] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10664–10673.

[17] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self attention distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1013–1021.

[18] 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.*, Jun. 2014, pp. 580–587.

[19] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

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

[22] G. Jocher et al. (2021). *Ultralytics/YOLOv5: V5.0*. [Online]. Available: https://github.com/ultralytics/yolov5

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

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

[25] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[26] G. Cheng, C. Lang, M. Wu, X. Xie, X. Yao, and J. Han, "Feature enhancement network for object detection in optical remote sensing images," *J. Remote Sens.*, vol. 2021, Jan. 2021, Art. no. 9805389.

[27] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619912.

[28] L. Hou, K. Lu, and J. Xue, "Refined one-stage oriented object detection method for remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 1545–1558, 2022.

[29] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607117.

[30] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.

[31] 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, 2022.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[33] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[34] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.

[35] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11953–11962.

[36] Y. Cai, Z. Yao, Z. Dong, A. Gholami, M. W. Mahoney, and K. Keutzer, "ZeroQ: A novel zero shot quantization framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13169–13178.

[37] Z. Dong, Z. Yao, D. Arfeen, A. Gholami, M. W. Mahoney, and K. Keutzer, "HAWQ-V2: Hessian aware trace-weighted quantization of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18518–18529.

[38] Y. Zhang, Z. Yan, X. Sun, W. Diao, K. Fu, and L. Wang, "Learning efficient and accurate detectors with dynamic knowledge distillation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5613819.

[39] S. Pande, A. Banerjee, S. Kumar, B. Banerjee, and S. Chaudhuri, "An adversarial approach to discriminative modality distillation for remote sensing image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[40] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602011.

[41] Z. Yang, Z. Yan, X. Sun, W. Diao, Y. Yang, and X. Li, "Category correlation and adaptive knowledge distillation for compact cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623318.

[42] J. Guo et al., "Distilling object detectors via decoupled features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2154–2164.

[43] Z. Yang et al., "Focal and global knowledge distillation for detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4643–4652.

[44] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605415.

[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[46] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 9, pp. 7945–7952.

[47] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," in *Proc. SIAM Int. Conf. Data Mining (SDM)*. Philadelphia, PA, USA: SIAM, 2022, pp. 424–432.
[48] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016.
[49] G. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[50] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[51] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[52] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[53] M.-T. Pham, L. Courtrai, C. Friguet, S. Lefèvre, and A. Baussard, "YOLO-fine: One-stage detector of small objects under various backgrounds in remote sensing images," *Remote Sens.*, vol. 12, no. 15, p. 2501, Aug. 2020.
[54] F. Qingyun and W. Zhaokui, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108786.
[55] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," 2021, *arXiv:2111.00273*.
[56] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21002–21012.
[57] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.
[58] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.
[59] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.
[60] Y. Wang, Y. Lu, and T. Blankevoort, "Differentiable joint pruning and quantization for hardware efficiency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 259–277.
[61] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.
[62] J. Gildenblat. (2021). *PyTorch Library for CAM Methods*. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam
[63] Y. Liu, W. Zhang, and J. Wang, "Zero-shot adversarial quantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1512–1521.
[64] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5008–5017.
[65] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

Jiaqing Zhang received the B.E. degree in telecommunications engineering from Ningbo University, Ningbo, Zhejiang, China, in 2019. She is currently pursuing the Ph.D. degree with the Image Coding and Processing Center, State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China.

Her research interests include multimodal image processing, remote sensing object detection, and network compression.

Jie Lei (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 2006 and 2010, respectively.

From 2014 to 2015, he was a Visiting Scholar with the Department of Computer Science, University of California, Los Angeles, CA, USA. Currently, he is a Professor with the School of Telecommunications Engineering, Xidian University, a member at the Image Coding and Processing Center, State Key Laboratory of Integrated Services Networks, and also with the Science and Technology on Electrooptic Control Laboratory, Luoyang, China. His research interests focus on image and video processing, computer vision, and customized computing for big-data applications.

Weiying Xie (Member, IEEE) received the B.S. degree in electronic information science and technology from the University of Jinan, Jinan, China, in 2011, the M.S. degree in communication and information systems from Lanzhou University, Lanzhou, China, in 2014, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2017.

Currently, she is an Associate Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. Her research interests include neural networks, machine learning, hyperspectral image processing, and high-performance computing.

Yunsong Li (Member, IEEE) received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1999 and 2002, respectively.

In 1999, he joined the School of Telecommunications Engineering, Xidian University, where he is currently a Professor. He is also the Director at the Image Coding and Processing Center, State Key Laboratory of Integrated Service Networks. His research interests focus on image and video processing and high-performance computing.

Geng Yang received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2019. She is currently pursuing the Ph.D. degree with the Image Coding and Processing Center, State Key Laboratory of Integrated Services Networks, Xidian University.

Her research interests include remote sensing image processing, computer vision, and efficient deep learning.

Xiuping Jia (Fellow, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in January 1982, and the Ph.D. degree in electrical engineering and the Graduate Certificate in Higher Education from the University of New South Wales, Canberra, ACT, Australia, in 1996 and 2005, respectively, via part-time study.

She has had a lifelong academic career in higher education, for which she has continued passion. She is currently an Associate Professor at the School of Engineering and Information Technology, The University of New South Wales. She has published widely addressing various topics, including data correction, feature reduction, and image classification using machine learning techniques. She has coauthored the remote sensing textbook *Remote Sensing Digital Image Analysis* (Springer-Verlag, third edition, 1999, and fourth edition, 2006). She is the author of book titled *Field Guide to Hyperspectral/Multispectral Image Processing* (SPIE, 2022). Her research interests include remote sensing, hyperspectral image processing, and spatial data analysis.