Knowledge Distillation via Cross Supervising with Attention for Remote Sensing Object Detection

Kefan Zhan
kefanzhan@smail.xtu.edu.cn
An Luo
anluo@smail.xtu.edu.cn
Yunpeng Zeng
yunpz@smail.xtu.edu.cn
Jiaxin Li
JXLi@smail.xtu.edu.cn
Yuan Zhang*
yuanz@xtu.edu.cn
Kai Hu*
kaihu@xtu.edu.cn

School of Computer Science,
Xiangtan University,
Hunan, P. R. China

Abstract

Knowledge distillation, which serves as a model compression approach that centers on transferring knowledge from a teacher model to a more compact student model, has been extensively employed to derive lightweight models. However, when applied to remote sensing object detection (RSOD), the unique characteristics of remote sensing images, such as small objects and complex background, substantially impede the student model in effective assimilation of knowledge from its superior teacher model. In this paper, we present a novel self-distillation framework designated as Knowledge Distillation via Cross Supervising with Attention (*CSAKD*), in which the teacher will offer module-level mentoring and adaptive guidance to eliminate the performance discrepancy between the teacher and the student. Specifically, the teacher will have access to the intermediate features of the student, generate a new set of features, and provide additional Cross Supervising for the student. Meanwhile, we propose an Adaptive Attention Weight Modulation (AAWM) to dynamically adjust the intensity of Cross Supervising, which ensures that the student can receive targeted guidance where it is most needed. In addition, we introduce an Attention-Guided Knowledge Alignment (AGKA), enabling the student to concentrate on the area that the teacher deems crucial instead of learning from the teacher's features blindly. To verify the effectiveness of the proposed method, we perform extensive experiments on two publicly available datasets, *i.e.*, DIOR and NWPU VHR-10. The experimental results show that our *CSAKD* outperforms existing state-of-the-art knowledge distillation approaches in the realm of RSOD.

* Corresponding authors

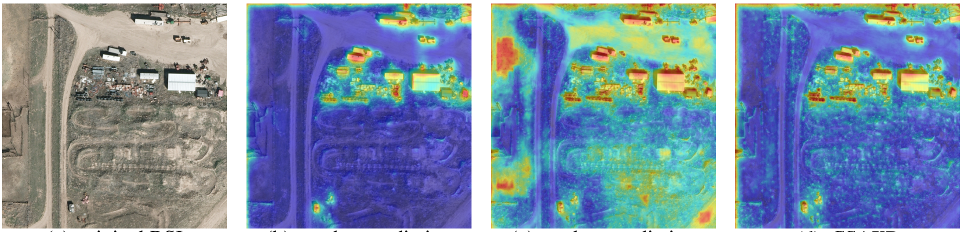
1 Introduction

Remote sensing object detection (RSOD), which is capable of identifying the category to which the target objects belong within a remote sensing image (RSI) and locating their positions, has played a fundamental role in the field of remote sensing [16]. In recent years, the rapid advancement of deep learning, particularly the maturation of convolutional neural networks (CNNs), has led to the proposal of numerous powerful object detectors, which have significantly improved RSOD techniques [27]. As computationally resource-constrained devices (*e.g*. smartphones and edge devices) become prevalent in daily life, the demand to deploy real-time applications on them in practical scenarios increases accordingly. However, the limited computing power and storage capacity of these devices have become significant bottlenecks for the deployment of remote sensing object detection models [22].

To address the above issue, numerous approaches have been proposed, including quantization [4, 8, 13, 19], pruning [5, 6, 14, 15], and knowledge distillation (KD) [7, 10, 17, 18], with the aim of designing models that are more lightweight and easier to deploy. Compared to quantization and pruning that directly modify the model (*e.g*., reducing parameter precision or eliminating redundant weights) to compress model size and lower computational complexity, KD adopts a distinct approach focusing on transferring knowledge. Specifically, KD focuses on transferring knowledge from a more complex and powerful teacher model to a simpler student model, with the goal of improving the accuracy, generalizability, and overall performance of the student model. Based on the knowledge transferred from the teacher model, existing KD methods can be broadly categorized into two groups: prediction mimicking and feature imitation. Prediction mimicking aims to align the student model's prediction distribution with that of the teacher model. In contrast, feature imitation focuses more on achieving consistency in feature representations between teacher and student models. In the field of object detection, since intermediate features contain hybrid information that includes classification and localization, feature imitation is generally observed to outperform prediction mimicking.

Unfortunately, when feature imitation is introduced into RSOD, it encounters a series of non-trivial obstacles. RSI (*e.g*. Figure 1(a)), in contrast to conventional photographic images, possesses distinct characteristics such as small objects and complex background, making it extremely difficult for the student model, which has a simpler architecture and lower model capacity, to understand and assimilate the intermediate features of the teacher model. As a result, the effectiveness of feature imitation in RSOD is significantly limited. As shown in Figure 1(b) and Figure 1(c), the student trained with the conventional feature imitation method (*e.g*. OST [24]) exhibits scattered attention patterns in complex backgrounds compared to the teacher's precise localization. This suggests that the inherent challenges of RSOD critically degrade the efficacy of feature imitation. More visualization details are shown in the Appendix Section 2.

Aware that it is challenging for the lightweight student model to directly mimic the intermediate knowledge of the teacher model, we devise a straightforward yet innovative self-distillation framework, namely Knowledge Distillation via Cross Supervising with Attention (*CSAKD*), aiming to boost the student model's performance by providing module-level mentoring and adaptive guidance. By adopting this approach, we conceptually simplify the model's architecture by treating each structural component (*e.g*. convolution layers) as discrete 'modules', a terminology that will be consistently employed throughout the following discussion. Specifically, this module-level mentoring, denoted as Cross Supervising, comes in the form of a new pair of cross features, which is generated when the teacher processes



(a) original RSI (b) teacher prediction (c) student prediction (d) ***CSAKD***

Figure 1: Visualizations of gradient heatmaps from the YOLOv8s. (a) is an original RSI. (b) and (c) are outputs predicted by the teacher and the student, which is trained with OST [24]. (d) is predicted by the student model trained with the proposed *CSAKD*.

the intermediate features of the student. Since these cross features originate from the student itself and are combined with the teacher's wisdom, they are much easier for the student to comprehend. In our adaptive guidance, an Adaptive Attention Weight Modulation (AAWM) is deployed to modulate the intensity of Cross Supervising. Moreover, we introduce an Attention-Guided Knowledge Alignment (AGKA) to enable the student to concentrate on the area that the teacher considers essential. Extensive experiments demonstrate that our proposed method surpasses the existing state-of-the-art KD methods in the realm of RSOD. In summary, the main contributions of our work are as follows.

- We propose a novel self-distillation framework, termed Cross Supervising with Attention Knowledge Distillation (*CSAKD*). Specifically designed for RSOD, this framework offers module-level mentoring, denoted as Cross Supervising, allowing the student model to acquire knowledge more effectively from the teacher model. In addition, a new Adaptive Attention Weight Modulation (AAWM) is proposed to dynamically adjust the intensity of Cross Supervising.

- We introduce an Attention-Guided Knowledge Alignment (AGKA), which enables the student to concentrate on the area that the teacher deems crucial and have better comprehension of the feature knowledge instead of blindly learning from the intermediate features of the teacher.

- We incorporate the proposed method with the YOLOv8 framework [21], allowing the YOLOv8 model to be compact and much easier to deploy while maintaining its strong performance. Through extensive experiments on publicly available DIOR [12] and NWPU VHR-10 [3] datasets, we have verified the effectiveness of *CSAKD* over existing state-of-the-art KD methods in the realm of RSOD.

2 Related Work

Knowledge Distillation was first proposed in [7] for the classification task. In general, the student is supposed to achieve an approximate performance as the teacher by learning from the dark knowledge provided by the teacher. The dark knowledge usually is the prediction distribution of the teacher. In contrast, FitNet [20] introduced the concept of "hint learning", where hint was the information of the intermediate features of the teacher. Since intermediate features contain richer hybrid information, including classification and localization, numerous feature-imitation-based KD methods have emerged in the realm of object detection. In [9], Attention-based Feature Matching (AFD) was introduced to adaptively transfer knowledge from the teacher's features to the student. In [2], Chen *et al.* introduced a method to distill knowledge through what they called Knowledge Review (ReviewKD), focusing on the factor of connection path across different levels between the teacher and the student. Motivated by the observation that feature imitation often surpasses prediction mimicking in object detection tasks, Wang *et al.* [23] presented a novel prediction mimicking distillation scheme, termed CrossKD, which delivered the intermediate features of the student's detection head to the teacher's detection head and produced cross-head knowledge.

Unlike traditional KD methods, self-distillation methods [10, 24, 25, 26] distill knowledge within the network itself. These techniques abandon the extra teacher model required in previous KD methods and provide an adaptive depth architecture for time-accuracy trade-offs at runtime. Rather than a method for compressing or accelerating models, self-distillation is more like a training technique to boost model performance. Numerous studies have demonstrated the efficacy of integrating self-distillation with other lightweight techniques. In [1], sharing the same parameters with the student, the teacher model was formed by assigning stochastic precision to the activation of each layer and producing soft labels for the student to learn. In [24], Zhang *et al.* proposed a two-step mixed-bit self-distillation framework, in which an One-to-one Self Teaching (OST) was introduced to improve the performance of the quantized student model.

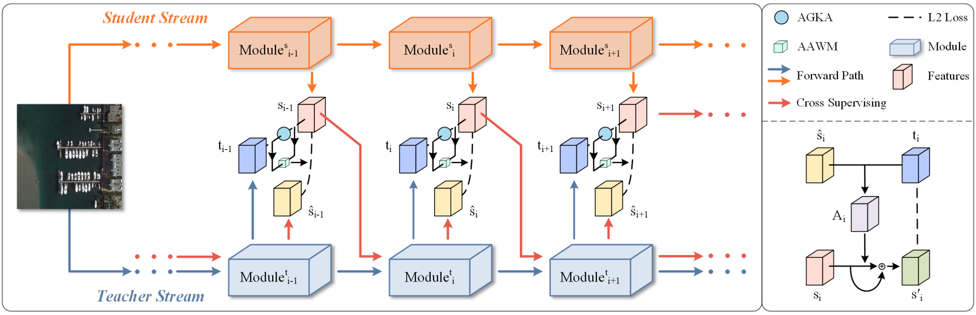
3 Proposed Method

In this section, we will first introduce the general framework of the proposed *CSAKD* and Cross Supervising. Then, an introduction about the AAWM and AGKA will be given in detail.

3.1 Overall Pipeline

In this paper, we propose *CSAKD*, a self-distillation framework to improve the performance of the student model through module-level mentoring. It incorporates a cross-feature imitation approach, inspired by CrossKD and adapted to the specific characteristics of RSI, as shown in Figure 2. In the proposed *CSAKD*, the teacher model assumes multifaceted roles. On the one hand, the teacher imparts knowledge by calculating between its intermediate features and the student's intermediate features and dynamically adjusts it based on attention mechanism termed AGKA, enabling the student to concentrate on the area the teacher deems crucial. On the other hand, the teacher is responsible for providing module-level mentoring through Cross Supervising, which enables the student to acquire more fine-grained module-level knowledge. This module-level knowledge is expected to compensate for the efficiency gap between the teacher and the student in each module and thereby improve the student's overall performance. In addition, we propose AAWM to dynamically control the intensity of Cross Supervising, ensuring that each module in the student model receives targeted guidance. The overall loss of *CSAKD* framework can be represented as:

$$\mathcal{L}_{total} = \mathcal{L}_{Det} + \alpha \cdot \sum_{i=0}^{N} \mathcal{L}_{AGKA_i} + \beta \cdot \sum_{i=0}^{N} \mathcal{L}_{Cross_i} \quad (1)$$



(a) Overview of *CSAKD* (b) AGKA

Figure 2: (a) Overview of the proposed Knowledge Distillation via Cross Supervising with Attention (*CSAKD*). Through Cross Supervising, the student will acquire module-level mentoring, which eliminates the performance discrepancy between the teacher and the student. The Adaptive Attention Weight Modulation (AAWM) is introduced to regulate the intensity of Cross Supervising. (b) Attention-Guided Knowledge Alignment (AGKA), which involves the cross features obtained in Cross Supervising, is capable of enabling the student to focus on the area that the teacher deems crucial.

where $\mathcal{L}_{Det}$ includes object loss, classification loss, and localization loss. The hyperparameters α and β indicate the impact balance between object detection and distillation.

Among the proposed framework, the core is Cross Supervising. To be concrete, let t_{i-1} and s_{i-1} denote the input of the teacher stream and the student stream, which are intermediate features, respectively. To achieve Cross Supervising, we transfer s_{i-1} to the teacher's i^{th} module to obtain cross features. The cross features $\hat{s}_i$ can be represented as:

$$\hat{s}_i = \begin{cases} s_0 & i = 0 \\ Module_i^t(s_{i-1}) & i \neq 0 \end{cases} \tag{2}$$

and of course we have the teacher's and the student's original output of the i^{th} module, t_i and s_i, which can be represented as:

$$t_i = Module_i^t(t_{i-1}) \tag{3}$$

$$s_i = Module_i^s(s_{i-1}) \tag{4}$$

Unlike CrossKD that obtains cross-head knowledge by calculating the difference between the teacher's original prediction and cross-head prediction, the Cross Supervising loss in our framework is obtained by calculating the *L2* loss between s_i and $\hat{s}_i$ as:

$$\mathcal{L}_{Cross_i} = \|\hat{s}_i - s_i\|_2 \tag{5}$$

where $\|\cdot\|_2$ denotes L2 norm loss. That is because in Cross Supervising, the teacher's and the student's i^{th} module share identical input, so the loss of the L2 norm between s_i and $\hat{s}_i$ will directly indicate the gap between the teacher's and the student's i^{th} module, which can be regarded as module-level knowledge. If the loss of Cross Supervising is calculated between t_i and $\hat{s}_i$, it will function on the student's $(i-1)^{th}$ module instead of i^{th} module according to back propagation, resulting in inaccurate module-level mentoring.

In addition, since s_i and $\hat{s}_i$ derive from the same input of the student, the knowledge imparted through Cross Supervising is considerably more accessible for the student to comprehend.

3.2 Adaptive Attention Weight Modulation

In traditional feature imitation methods, the intensity of knowledge distillation between each pair of teacher-student modules is typically uniform. However, given the distinctive characteristics of RSI, this coarse-grained approach results in a deficiency in fine-grained perception of the disparities between the intermediate features of the teacher model and those of the student model. In [], Kang *et al.* introduced a Difficulty Adaptive Training (DAT) strategy that forced the model to pay more attention to hard cases. Motivated by DAT, we introduce AAWM, which is capable of dynamically regulating the intensity of Cross Supervising for all distilled modules. The calculation of AAWM is based on the similarity between the corresponding intermediate features of the teacher and the student. Initially, we operate Global Average Pooling (GAP) on t_i and s_i respectively, and accordingly we gain T_i and S_i. Then a similarity number a_i is calculated basing on T_i and S_i:

$$T_i = GAP(t_i) \tag{6}$$

$$S_i = GAP(s_i) \tag{7}$$

$$a_i = \frac{T_i \cdot S_i^T}{\sqrt{d}} \tag{8}$$

where d is the dimensionality of T_i and S_i. When the value of a_i is higher, it indicates a greater similarity between the intermediate feature maps of the teacher and the student, suggesting that less Cross Supervising is required in the student's i^{th} module. To this end, after calculating a_i, we apply the function $Sigmoid(\cdot)$ to constrain its output within the interval $[0,1]$ and reverse mapping to a_i, through which the magnitude of a_i can reflect the degree of Cross Supervising needed. The final attention weight can be represented as:

$$weight_i = 1 - Sigmoid(a_i) \tag{9}$$

By utilizing the acquired attention weight, we are able to adaptively adjust the intensity of Cross Supervising between every pair of teacher-student modules. This allows for a more refined and targeted transfer of module-level knowledge, enabling student modules to better learn from the corresponding teacher modules. Then the Cross Supervising loss can be represented as:

$$\mathcal{L}_{Cross_i} = weight_i \cdot \|\hat{s}_i - s_i\|_2 \tag{10}$$

3.3 Attention-Guided Knowledge Alignment

When KD is applied in RSOD, the student model faces a multitude of challenges due to the distinctive characteristics of RSI, such as small objects and complex background. Although module-level mentoring has improved the learning ability of the student model, it falls short of fastidiously guiding the student to capture crucial information from the teacher. Especially when dealing with small objects, the less-capable student finds it difficult to extract important information from the teacher's knowledge due to the complex background. To address the above issue, we propose a new solution based on the attention mechanism, termed Attention-Guided Knowledge Alignment (AGKA). By introducing the attention of the teacher, the goal of this approach is to let the student focus on the area that the teacher deems crucial instead of blindly imitating the whole intermediate features of the teacher. Not only does this approach facilitate the efficiency of KD, it also achieves a more precise transfer of the feature information on the pixel level. To be concrete, we first operate $ReLU(\cdot)$ on t_i and $\hat{s}_i$ to obtain q_i and k_i, respectively.

$$q_i = ReLU(t_i) \tag{11}$$

$$k_i = ReLU(\hat{s}_i) \tag{12}$$

where $ReLU(\cdot)$ is used to avoid incorrect attention information that might be caused by the multiplication of negative numbers. Assuming that the teacher is well trained in advance, t_i and $\hat{s}_i$ are expected to contain meaningful feature representations, where higher positive values in their activations correspond to more semantically important information. The teacher's attention map A_i can be represented as:

$$A_i = \frac{Softmax(q_i \cdot k_i^T)}{\sqrt{d_{k_i}}} \tag{13}$$

where d_{k_i} denotes the dimensionality of q_i and k_i.

Instead of directly computing self-attention with q_i, the participation of k_i allows dynamic adaptation of the attention map. Since the teacher's parameters are frozen during *CSAKD*, q_i remain constant. In contrast, k_i varies according to s_{i-1}. As the student model progressively absorbs knowledge of the teacher during distillation, s_{i-1} will converge to t_{i-1}. Consequently, the attention matrix A_i will offer more accurate guidance. Then a new s_i' is calculated as follows:

$$s_i' = s_i + A_i \cdot s_i \tag{14}$$

where residual operation is employed to maintain the student's original features and prevent the student from becoming overly reliant on the guidance provided by the teacher. Since s_i' is combined with the teacher's attention, it enables the student to better understand the feature knowledge from the teacher and place more emphasis on the area that the teacher deems to be more closely related to the object. Finally, the loss of AGKA can be represented as:

$$\mathcal{L}_{AGKA_i} = \left\| t_i - s_i' \right\|_2 \tag{15}$$

4 Experimental results and Discussion

4.1 Datasets

4.1.1 DIOR

The DIOR dataset was proposed by Li et al. [12] for the task of object detection, which involves 23,463 images and 192,472 instances. We choose 11,725 images as the training set and 11,738 images as the testing set. The dataset contains twenty categories, and the size of the image is fixed to 512×512.

4.1.2 NWPU VHR-10

The NWPU VHR-10 dataset [3] contains 800 Very-High-Resolution images, of which 650 images 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.

Table 1: Comparison with the state-of-the-art knowledge distillation methods

Method	DIOR		NWPU VHR-10	
	mAP50	**mAP50-95**	**mAP50**	**mAP50-95**
YOLOV8s (Teacher)	76.57	52.45	90.58	52.68
YOLOv8s+HQ (Student)	73.84	51.44	88.05	51.53
AFD [9]	75.72	52.30	89.41	52.97
ReviewKD [2]	75.27	51.59	91.00	52.92
OST [24]	75.92	**52.65**	90.25	52.74
CrossKD [23]	75.79	52.42	90.17	53.30
***CSAKD* (Ours)**	**76.27**	52.56	**91.15**	**53.69**

4.2 Implementation Details

Our framework is implemented in PyTorch and runs on a workstation with an NVIDIA RTX4090-24GB GPU. We use different training strategies for different datasets. The standard stochastic gradient descent (SGD) is used to train the network with a momentum of 0.937 and the weight decay is set to 0.0005. The learning rate is initially set to 0.01. The batch size is 16 on the DIOR dataset and 8 on the NWPU VHR-10 dataset. The epoch is 150 for both the DIOR and NWPU VHR-10 datasets. Considering the power and size of the YOLOv8 model, we train YOLOv8s [21] as the teacher model and obtain the student model, which is the baseline by applying Hybrid Quantization (HQ) [24] on the teacher model. All training processes are completed from scratch without any pre-trained model, and we obtain knowledge from the teacher model's backbone and neck. In the test stage, the Intersection over Union (IoU) threshold of non-maximum suppression is 0.4 on the DIOR dataset, and 0.6 on the NWPU VHR-10 dataset.

4.3 Results and Discussion

To validate the effectiveness of the proposed method, we carry out extensive experiments in the following aspects: (1) We compare the proposed framework with other state-of-the-art knowledge distillation methods. (2) We conduct some ablation experiments to verify the effectiveness of key components of the proposed framework. (3) Experiments are carried out to find the best combination of distillation hyperparameters. (4) We visualize some detection results on the two datasets.

4.3.1 Comparison with the state-of-the-art Knowledge distillation methods

We apply the proposed *CSAKD* framework and other existing state-of-the-art KD methods on the student model, serving as training strategies and transferring the knowledge from the teacher model to elevate the student model's performance. AFD [9], ReviewKD [2], OST [24] and CrossKD [23] are selected to compare with our method. And we have reproduced these methods based on the YOLOv8 framework at [1]. Among these KD methods, AFD, ReviewKD and OST are feature imitation methods, meanwhile CrossKD is prediction mimicking method. The numerical results are shown in Table1. Except for the teacher, all HQ-quantized students are trained with different KD methods. In particular, the baseline is

[1] https://github.com/KefanZhan/YOLOv8-KD

Table 2: Ablation study of key components in our *CSAKD* framework

Key Components		DIOR		NWPU VHR-10	
Cross Supervising	**AGKA**	**mAP50**	**mAP50-95**	**mAP50**	**mAP50-95**
×	×	74.21	50.52	87.91	51.02
✓	×	74.28	50.49	88.50	51.08
×	✓	76.14	52.54	88.74	52.26
✓	✓	**76.27**	**52.56**	**91.15**	**53.69**

Table 3: mAP50 comparisons of different α and β value on the NWPU VHR-10 dataset.

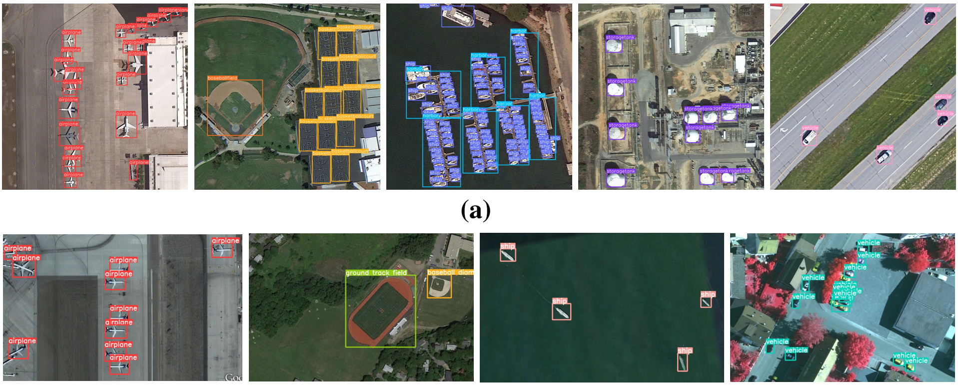
β \ α	0	100	200	300	400
0	88.05	90.33	90.56	89.52	88.74
10	90.42	90.47	90.40	91.05	**91.15**
20	89.40	89.66	90.11	90.71	89.82
30	90.44	89.88	89.99	90.83	90.73
40	90.00	90.43	90.23	90.41	91.01
50	90.57	90.29	90.78	90.10	90.67

trained only with the ground truth. It is evident that our *CSAKD* outperforms other KD methods without introducing additional parameters or GFLOPs. On the DIOR dataset, compared to other KD methods, our *CSAKD* reaches the highest mAP50 76.27%, with 2.43% promotion compared to the student. On the NWPU VHR-10 dataset, due to its limited scale and the high efficiency of knowledge transfer of our framework, the student model trained with our approach outperforms the teacher model by 0.57% in mAP50. This superiority indicates that, via the *CSAKD* framework, the teacher model can transfer more valuable and understandable knowledge to the student model in the realm of RSOD. We further compare the student model trained with the proposed *CSAKD* with some state-of-the-art object detectors in the Appendix Section 3.

4.3.2 Ablation experiments

In this section, we conduct ablation experiments with the proposed *CSAKD* framework. Initially, we explore how each component (Cross Supervising and AGKA) promotes the performance of the compact student model. Then optimization of the distillation hyperparameters is performed.

Effectiveness of key components of our CSAKD framework. We compare the key components of the proposed framework with directly learning from the intermediate feature knowledge of the teacher model, and the results are presented in Table 2. In particular, by replacing direct feature learning with the proposed AGKA mechanism, the student model achieves significant performance improvements across all datasets. On the DIOR dataset, AGKA increases mAP50 by 1.93% and while on the NWPU VHR-10 dataset, AGKA boosts mAP50 by 0.83%. These results demonstrate that AGKA effectively enhances the student's ability to comprehend and utilize the knowledge distilled from the teacher, compared to directly learning from the teacher's intermediate features. Relatively, although the improvement brought by Cross Supervising seems to be less than AGKA, Cross Supervising provides more extra



(a)

(b)

Figure 3: Visualization of detection results on two datasets. (a) DIOR. (b) NWPU VHR-10.

guidance for the student when combining with AGKA. Especially, on the NWPU VHR-10 dataset, mAP50 increases by 2.41% (from 88.74% to 91.15%), while mAP50-95 also improves by 1.43% (from 52.26% to 53.69%). This convincingly demonstrates the efficacy of the entire *CSAKD* framework. We further apply Cross Supervising on other feature imitation methods in the Appendix Section 4 to verify its effectiveness.

Hyperparameters' optimization. As presented in Section 3.3, α is used to control the loss of distillation obtained in AGKA and β is used to control the intensity of Cross Supervising, so we compare the performance of the distillation in the different weights. To determine the optimal combination of α and β, we perform ablation experiments on the NWPU VHR-10 dataset. The corresponding mAP50 results are summarized in Table 3. From the results, we can find that we identify α=400 and β=10 as the most favorable combination in the proposed *CSAKD* framework.

4.3.3 Visualization

To further show the detection performance of the student model trained with the proposed *CSAKD*, we also visualize the detection results on the two datasets, as shown in Fig. 3, which shows the student model trained with *CSAKD* can easily detect objects at different scales.

5 Conclusion

In this paper, we propose a novel and efficient self-distillation method, called Knowledge Distillation via Cross Supervising with Attention, denoted as *CSAKD*, for remote sensing object detection. In our *CSAKD* framework, the teacher has access to the student's intermediate features and offers module-level mentoring through Cross Supervising. Meanwhile, we introduce an Adaptive Attention Weight Modulation to dynamically adjust the intensity of Cross Supervising. In addition, an Attention-Guided Knowledge Alignment is proposed to enable the student to concentrate on the area that the teacher deems crucial and make the direct knowledge distilled from the teacher easier to comprehend. The results of publicly available datasets DIOR and NWPU VHR-10 show that our *CSAKD* outperforms other state-of-the-art knowledge distillation methods in the field of remote sensing object detection.

6 Acknowledgments

This work was supported by the Research Foundation of Education Department of Hunan Province of China under Grant 23A0146, the Research Project on Teaching Reform of Colleges in Hunan Province under Grant 202401000574, and the Innovation and Entrepreneurship Training Program for China University Students under Grants S202410530007 and S202410530071.

References

[1] Yoonho Boo, Sungho Shin, Jungwook Choi, and Wonyong Sung. Stochastic precision ensemble: Self-knowledge distillation for quantized deep neural networks. *ArXiv*, abs/2009.14502, 2020.

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5008–5017, June 2021.

[3] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.

[4] Yifu Ding, Weilun Feng, Chuyan Chen, Jinyang Guo, and Xianglong Liu. Reg-ptq: Regression-specialized post-training quantization for fully quantized object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16174–16184, June 2024.

[5] Song Guo, Lei Zhang, Xiawu Zheng, Yan Wang, Yuchao Li, Fei Chao, Chenglin Wu, Shengchuan Zhang, and Rongrong Ji. Automatic network pruning via hilbert-schmidt independence criterion lasso under information bottleneck principle. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17412–17423, 2023.

[6] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In *Neural Information Processing Systems*, 2015.

[7] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[8] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. *ArXiv*, abs/2102.02973, 2021.

[10] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10664–10673, June 2021.

[11] Weitai Kang, Luowei Zhou, Junyi Wu, Changchang Sun, and Yan Yan. Visual grounding with attention-driven constraint balancing. *ArXiv*, abs/2407.03243, 2024.

[12] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ArXiv*, abs/1909.00133, 2019.

[13] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Haokun Lin, Haoli Bai, Zhili Liu, Lu Hou, Muyi Sun, Linqi Song, Ying Wei, and Zhenan Sun. Mope-clip: Structured pruning for efficient vision-language models with module-wise pruning error metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27370–27380, June 2024.

[15] Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2658–2668, January 2024.

[16] An Luo, Kai Hu, and Kai Jiang. Rethinking dual-stream super-resolution for enhancing remote sensing object detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[17] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4233–4241, 2024.

[18] Roy V Miles, Adrian Lopez, and Krystian Mikolajczyk. Information theoretic representation distillation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[19] Jaehyeon Moon, Dohyung Kim, Junyong Cheon, and Bumsub Ham. Instance-aware group quantization for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16132–16141, June 2024.

[20] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2014.

[21] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.

[22] Ching-Hao Wang, Kang-Yang Huang, Yi Yao, Jun-Cheng Chen, Hong-Han Shuai, and Wen-Huang Cheng. Lightweight deep learning: An overview. *IEEE Consumer Electronics Magazine*, 13(4):51–64, 2024.

[23] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16520–16530, June 2024.

[24] Jiaqing Zhang, Jie Lei, Weiying Xie, Yunsong Li, Geng Yang, and Xiuping Jia. Guided hybrid quantization for object detection in remote sensing imagery via one-to-one self-teaching. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[25] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721, 2019.

[26] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2022.

[27] Xiangrong Zhang, Tianyang Zhang, Guanchun Wang, Peng Zhu, Xu Tang, Xiuping Jia, and Licheng Jiao. Remote sensing object detection meets deep learning: A metareview of challenges and advances. *IEEE Geoscience and Remote Sensing Magazine*, 11(4): 8–44, 2023.