Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083923

ARTICLE

BMGKD: A High Precision Object Detection Knowledge Distillation Method for Bridging Multi-Dimensional Gaps

Tianqi Wang, Yang Li and Zhisong Pan*

College of Command and Control Engineering, Army Engineering University of PLA, Nanjing, China

*Corresponding Author: Zhisong Pan. Email: panzhisong@aeu.edu.cn

Received: 13 April 2026; Accepted: 09 June 2026; Published: 23 July 2026

ABSTRACT: Existing knowledge distillation methods for object detection struggle to bridge the teacher-student capacity gap and overlook the inherent differences between classification and regression subtasks. To address these issues, we propose a Bridging Multi-dimensional Gaps Knowledge Distillation (BMGKD) method, which comprises two core modules: a feature difference distillation module and a response difference distillation module. The feature difference distillation module achieves global feature structural alignment via improved centered kernel alignment and performs local key feature alignment using joint spatial and channel-wise cosine similarity masks. The response difference distillation module constructs a dynamic classification mask and a high-quality prediction box selection mechanism, along with a classification and regression co-optimization loss function. On the MS COCO dataset, BMGKD achieves the highest mAP across all three teacher-student configurations on two-stage Faster R-CNN, single-stage anchor-free GFL, and single-stage anchor-based RetinaNet detectors. In the ResNet101→ResNet50 setting, it surpasses baselines by 2.7, 4.2, and 2.8 percentage points, respectively, and maintains consistent optimality under varying capacity gaps and cross-architecture scenarios. On Pascal VOC 2007, BMGKD attains mAP of 56.2%, 58.3%, and 58.0% on the same three detectors, outperforming all compared methods. Ablation studies confirm each component's contribution. These results demonstrate that BMGKD effectively resolves the distillation accuracy degradation caused by capacity gaps and task discrepancies between teacher and student models. It provides an effective solution for knowledge distillation across diverse detection architectures.

KEYWORDS: Object detection; knowledge distillation; model compression; feature fusion

1 Introduction

In recent years, deep learning has achieved significant progress and remarkable results in various applications, particularly in visual tasks such as image classification [1] and object detection [2]. Benefiting from the superior performance of deep learning models, existing relevant methods have been widely integrated and deployed in practical scenarios, including autonomous driving, smart retail, and drone operations. Although deep learning models have demonstrated excellent performance, their deployment on mobile and edge devices remains a major challenge, primarily due to the severe constraints on computing power and storage resources of such devices. To address this issue, knowledge distillation (KD) [3] has been proposed as a solution. It effectively reduces the parameter count and computational overhead of deep convolutional neural networks (CNNs) while maximizing the preservation of model performance.

Knowledge distillation was initially applied to the lightweight compression of classification models, and numerous researchers have since extended it to the field of object detection. Li et al. [4] directly selected

features from the teacher model as distillation supervision targets. Building upon the soft labels introduced in [3], Chen et al. [5] directly applied them to the classification branch of object detection models. Wu et al. [6] designed refined feature masks to enable feature imitation learning between teacher and student models. Lan and Tian [7] utilized gradient information to filter and distill high-contribution features. Despite the significant progress made by the aforementioned works in object detection knowledge distillation, none of these studies have thoroughly explored the model capacity differences between teacher and student networks, nor have they conducted targeted research on the distinctions between the classification and regression tasks in object detection.

Regarding the intermediate feature layers of object detection models, existing feature mask designs rely excessively on the teacher model. These masks are typically generated by extracting feature maps from the intermediate layers of the teacher network and performing simple attention calculations [8]. This approach has two major limitations. First, mask generation relies solely on the teacher model's perspective, completely ignoring the student model's learning state. Second, attention is extracted only based on the basic information of feature maps, failing to uncover their deep-level correlations. Consequently, during training, the student model cannot fully perceive and understand the semantic meaning of key features filtered by the teacher model through masking. Furthermore, regarding the final output layer of object detection models, traditional knowledge distillation algorithms [9] use kullback-leibler (KL) divergence to uniformly align the outputs of the classification and regression heads, failing to account for the distinct characteristics of these two subtasks. This leads to the underutilization of a substantial amount of valuable output information, ultimately resulting in limited distillation performance.

To address these issues, we propose a knowledge distillation method for object detection that bridges multidimensional gaps between teacher and student models. BMGKD constructs a feature difference module that synchronizes the correlation between the global and local feature maps of the teacher and student models in key regions. This enhances the student model's ability to perceive and understand features, thereby bridging the performance gap between teacher and student networks. Additionally, we design a response difference module to achieve the joint optimization of classification and regression in object detection tasks. Experimental results on the MS COCO dataset demonstrate that BMGKD achieves a detection accuracy of 44.4% on a single-stage anchor-free detector, representing a 4.2% improvement over the baseline model.

Our main contributions are summarized as follows:

(1) We propose the BMGKD framework, which bridges feature discrepancies through global alignment via the Gram matrix and local alignment via space-channel joint masking, and bridges response discrepancies through dynamic classification masking and region selection.

(2) Within the response difference distillation module, we establish a collaborative optimization mechanism for classification and regression, implementing joint teacher-student decision-making and a task-consistency loss design.

(3) Across three detectors, multiple capacity gaps, and two datasets, BMGKD outperforms over a dozen mainstream distillation methods from recent years, with ablation studies and visualizations confirming the effectiveness of each component.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work. The details of BMGKD are described in Section 3. In Section 4, two public datasets are used to evaluate the performance of BMGKD. Section 5 provides a summary of the paper.

2 Related Work

2.1 Object Detection

Object detection is one of the most fundamental and actively researched topics in the field of computer vision, with significant practical value in industrial applications. Based on algorithmic architecture and detection processes, existing object detectors can generally be categorized into three types: two-stage detectors, single-stage anchor-based detectors, and single-stage anchor-free detectors. Two-stage detectors use neural networks to extract candidate regions and then perform precise classification and bounding box regression on each candidate region. Representative algorithms of this class of detectors include R-CNN and its improved version, Fast R-CNN [10]. Among these, Faster R-CNN achieved end-to-end training by introducing a region proposal network(RPN). Building on this foundation, various RPN-based improved algorithms have emerged, including Cascaded R-CNN and Mask R-CNN [11].

Two-stage detectors achieve higher detection accuracy but have slower inference speeds. Compared with two-stage detectors, single-stage detectors can simultaneously perform object classification and regression tasks, significantly improving inference speed but at the cost of reduced detection accuracy. Single-stage anchor-based detectors, such as RetinaNet [12] and YOLO [13] complete detection tasks using predefined anchor boxes, but this introduces additional hyperparameters and requires redundant anchor box pruning mechanisms during the processing. In contrast, single-stage anchor-free detectors, represented by FCOS [14] and GFL [15], abandon the anchor box design and use keypoint regression to directly predict bounding boxes.

2.2 Knowledge Distillation

Knowledge distillation is an efficient model compression technique that extracts and transfers implicit supervisory knowledge from a high performance teacher model to train a lightweight student model. It enables the student model to achieve performance close to that of the teacher model while maintaining its lightweight nature. Existing methods are primarily divided into two categories: feature-based distillation and response-based distillation.

Feature-based distillation focuses on the multi-scale features output by feature pyramid networks, extracting the implicit knowledge embedded in these intermediate-layer features. As an early knowledge distillation model targeting intermediate features, FitNet [16] demonstrated that the intermediate-layer features of the teacher network contain rich local structural and pattern information. This information can serve as an effective supervisory signal to guide the training of the student model. Lv et al. [17] introduced feature map transformation and alignment mechanisms to optimize the distribution differences between teacher and student features. Yang et al. [8] proposed a feature masking knowledge distillation method that can adaptively select key regions from the teacher's feature maps. Li et al. [18] proposed CSAKD, which utilized the teacher's attention information to generate spatial masks and enhanced student features by adding teacher global context. However, CSAKD relied on the teacher's unilateral attention to determine the masking regions, ignoring the real time learning state and capability gap of the student model.

Response-based distillation focuses on the final network outputs generated by classification and regression tasks. The LD method [19] adopted the bounding box probability distribution proposed by GFL, used the KL divergence to align the outputs of the teacher and student models. It also designed a region-weighting strategy based on the intersection over union (IoU) ratio to optimize knowledge distillation performance. BCKD [20] reconstructed the classification branch output logic into multiple sets of binary classification results, effectively alleviating the mismatch between detection tasks and classification loss functions. Zhang et al. [21] proposed SAMKD, which introduced a spatial pyramid mechanism and used the difference between teacher and student features to weight logit distillation. Although SAMKD achieved hierarchical logit alignment at multiple scales, it uniformly applied KL divergence for logit distillation across

all levels, thereby overlooking the fundamental task disparity between classification and regression heads in object detection.

However, feature-based distillation methods fail to fully exploit the capacity differences between teacher and student models, resulting in insufficient alignment between feature supervision and the student model. Response-based distillation methods overlook the task differences between the classification head and the regression head in detection tasks. In contrast, the method proposed in this paper further refines the features of the neck network and the final response outputs of the object detector by introducing feature difference and response difference modules.

3 Methodology

3.1 Overall Structure

The proposed BMGKD framework consists of a feature difference distillation module and a response difference distillation module. During the model training phase, both the teacher and student networks perform forward passes simultaneously. The parameters of the teacher network are fully frozen, while the student network learns by transferring knowledge from the teacher network and utilizing supervision from the detection task, iteratively updating its own parameters through backpropagation. Subsequently, feature maps are extracted from each layer of the teacher-student feature pyramid network and processed through the feature difference module to achieve dual alignment of global and local features. Meanwhile, the response difference module selectively distills response values from the teacher and student network branches, focusing on optimizing regions where the outputs of the teacher and student networks differ significantly in classification and regression. Additionally, we designed a dedicated loss function to bridge the task gap between the classification head and the regression head. The overall network architecture of BMGKD is shown in Fig. 1.

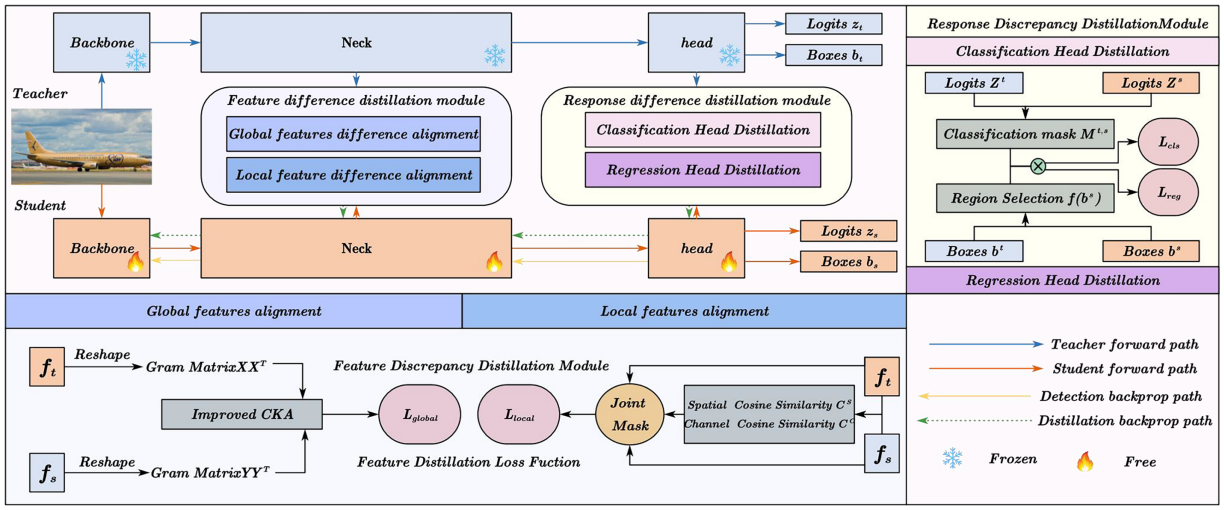


Figure 1: Overall network architecture of BMGKD.

3.2 Feature Difference Distillation Module

To ensure that the student model can fully learn from the teacher model, we propose a feature difference distillation module. This module uses a Gram matrix-based CKA loss function to distill the structural relationships between features rather than their absolute values. Furthermore, by combining spatial and channel-wise cosine similarity masks, it dynamically assigns weights to each distillation region based on the

differences between the teacher and student models. This achieves comprehensive alignment between the global and local features of the teacher and student models, thereby improving the student model's learning efficiency and its ability to understand features.

3.2.1 Global Feature Difference Alignment

In the global feature processing stage, the feature difference distillation module does not perform direct numerical alignment on the global feature map. Instead, it introduces Gram matrices to model the relationships between features.

Let the teacher model be denoted by t and the lightweight student model by s. The feature map output by the i-th layer of the teacher model's neck network is denoted by $f_t^i \in \mathbb{R}^{b\times c\times h\times w}$, where c, h and w represent the number of channels, height, and width of the teacher model's feature map, respectively. The feature map output by the i-th layer of the student model's neck network is denoted by $f_s^i \in \mathbb{R}^{b\times c\times h\times w}$, where c, h and w represent the number of channels, height, and width of the student model's feature map, respectively. b is the training batch size. Use $X = T(f_t^i)$ and $Y = T(f_s^i)$ to transform $\mathbb{R}^{b\times c\times h\times w}$ into $\mathbb{R}^{b\times chw}$, where T is transformation function used to unify the dimensions of the teacher and student feature maps. The CKA method is effective for knowledge distillation tasks [22], but the computational complexity of the traditional CKA is high. Therefore, we have improved and optimized it, as shown in Eq. (1):

$$\begin{aligned} CKA &= \frac{\left\|Y^TX\right\|_F^2}{\left\|X^TX\right\|_F^2\left\|Y^TY\right\|_F^2} = \frac{tr\left((Y^TX)^TY^TX\right)}{\left\|X^TX\right\|_F^2\left\|Y^TY\right\|_F^2} = \frac{tr\left(X^TYY^TX\right)}{\left\|X^TX\right\|_F^2\left\|Y^TY\right\|_F^2} \\ &= \frac{tr\left(XX^TYY^T\right)}{\left\|X^TX\right\|_F^2\left\|Y^TY\right\|_F^2} = \frac{vec(XX^{\mathrm{T}})^{\mathrm{T}}vec(YY^{\mathrm{T}})}{\left\|vec(XX^{\mathrm{T}})\right\|_2\left\|vec(YY^{\mathrm{T}})\right\|_2} \end{aligned} \tag{1}$$

where XX^{T} and YY^{T} are Gram-Matrix. $vec(\cdot)$ is a vectorization operation. For the feature matrix $X \in \mathbb{R}^{b\times d}$, the similarity between the i-th and j-th samples in the feature space of its Gram matrix $G_X = XX^{\mathrm{T}} \in \mathbb{R}^{b\times b}$ is denoted by $G_{ij} = x_i^{\mathrm{T}}x_j$. As can be seen from the formula, the improved CKA metric is equivalent to the cosine similarity of teacher-student features after vectorizing the Gram matrix.

The Gram matrix elevates traditional eigenvalue matching to the level of feature relationship structure matching, enabling the student network to simultaneously learn both the teacher's feature representations and the logic of feature organization. The distillation loss function used for global feature alignment is shown in Eq. (2):

$$L_{\text{Global}} = 1 - \frac{vec(XX^{\mathrm{T}})^{\mathrm{T}}vec(YY^{\mathrm{T}})}{\left\|vec(XX^{\mathrm{T}})\right\|_2\left\|vec(YY^{\mathrm{T}})\right\|_2} \tag{2}$$

3.2.2 Local Feature Difference Alignment

In the phase of aligning local feature differences, to ensure that the student model focuses on the most discriminative feature regions in the teacher model, we use the extracted teacher and student feature maps to calculate spatial cosine similarity and channel cosine similarity, respectively. We then construct a local key information mask based on these two metrics.

First, the feature mapping layer $f_s' = \phi(f_s)$ is used to transform the dimensions of the student network's feature map so that they match the number of channels in the teacher network. Subsequently, the spatial cosine similarity $C^S \in \mathbb{R}^{h\times w}$ and channel cosine similarity $C^C \in \mathbb{R}^{h\times w}$ are calculated for the teacher feature maps and student feature maps at the corresponding levels, respectively. For each spatial location (i, j) on the feature map, the teacher and student channel vectors at that location are defined as follows:

$$v_{ij}^{t} = \left(f_t(1,i,j), f_t(2,i,j), \ldots, f_t(C,i,j)\right)^{\top}. \tag{3}$$

$$v_{ij}^{s} = \left(f_s'(1,i,j), f_s'(2,i,j), \ldots, f_s'(C,i,j)\right)^{\top}. \tag{4}$$

The spatial cosine similarity C_{ij}^{s} is the cosine similarity between the vectors of these two channels:

$$C_{ij}^{S} = \frac{\mathbf{v}_{ij}^{t} \cdot \mathbf{v}_{ij}^{s}}{\left\|\mathbf{v}_{ij}^{t}\right\| \left\|\mathbf{v}_{ij}^{s}\right\|} = \frac{\sum_{k=1}^{C} f_t(k,i,j) \cdot f_s'(k,i,j)}{\sqrt{\sum_{k=1}^{C} \left(f_t(k,i,j)\right)^2} \cdot \sqrt{\sum_{k=1}^{C} \left(f_s'(k,i,j)\right)^2}}. \tag{5}$$

where $f_t(k,i,j)$ represents the value of the teacher feature map at channel k-th and position (i,j). $f_s'(k,i,j)$ represents the value of the student feature map at the corresponding position. For each channel k, flatten the spatial feature map on that channel into an $h \times w$ dimensional vector:

$$u_k^{t} = \left(f_t(k,1,1), f_t(k,1,2), \ldots, f_t(k,h,w)\right)^{\top} \tag{6}$$

$$u_k^{s} = \left(f_s'(k,1,1), f_s'(k,1,2), \ldots, f_s'(k,h,w)\right)^{\top} \tag{7}$$

The channel cosine similarity C_k^{C} is the cosine similarity between the vectors of these two spatial:

$$C_k^{C} = \frac{\mathbf{u}_k^{t} \cdot \mathbf{u}_k^{s}}{\left\|\mathbf{u}_k^{t}\right\| \left\|\mathbf{u}_k^{s}\right\|} = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} f_t(k,i,j) \cdot f_s'(k,i,j)}{\sqrt{\sum_{i=1}^{h} \sum_{j=1}^{w} \left(f_t(k,i,j)\right)^2} \cdot \sqrt{\sum_{i=1}^{h} \sum_{j=1}^{w} \left(f_s'(k,i,j)\right)^2}}. \tag{8}$$

By combining the spatial cosine similarity Eq. (5) with the channel cosine similarity Eq. (8) to construct a joint feature mask, we achieve precise screening of key regions in the teacher-student feature maps. The calculation formula is shown in Eq. (9):

$$M_J^{ijk} = \left|\left(1 - C_{ij}^{S}\right) \times \left(1 - C_k^{C}\right)\right| \tag{9}$$

The joint mask, constructed based on cosine similarity, integrates prior knowledge of teacher features with the real-time learning state of the student model. It addresses the shortcoming of traditional methods, where region selection is solely teacher-driven, while enhancing the student model's understanding of the relationships among teacher features, thereby minimizing the differences between teacher and student feature maps. The final loss function for local feature differences is shown in Eq. (10):

$$L_{local} = \frac{1}{chw} \sum_{i} \sum_{j} \sum_{k} M_J^{ijk} \left\| f_t^{i,j,k} - f_s'^{i,j,k} \right\|_2 \tag{10}$$

3.3 Response Difference Distillation Module

Traditional knowledge distillation methods for object detection optimize the response outputs of the classification and regression heads independently. This results in a significant amount of useful information being overlooked in both branches, further degrading the distillation performance. To overcome this limitation, the response difference distillation module enables collaborative optimization and joint distillation of the classification and regression tasks. For the predicted scores output by the classification head, we define the base classification mask $M^{t} = \max_{0<c\leq C} |p_c^{t}|$, where p_c^{t} is the prediction score of the teacher model for the c-th target class. This mask is a static mask and cannot be adapted during the model training process. Therefore, based on the design of the feature-differential distillation module, we construct a dynamic classification mask $M^{t,s} = \max_{0<c\leq C} |Z_c^{t} - Z_c^{s}|$ that employs joint teacher-student decision-making.

Based on the coordinate values output by the regression head, we use a region selection function to identify high quality prediction boxes and filter out low quality ones during the regression output refinement stage. $B^s = \{b_r^s\}_{r=1}^{R}$ is the set of prediction boxes for a single image in the student model, containing a total of R prediction boxes. $B^{gt} = \left\{b_{r'}^{gt}\right\}_{r'=1}^{R'}$ is the set of ground-truth bounding boxes corresponding to the image, containing a total of R' ground-truth boxes. The region selection function for the r-th prediction box b_r^s generated by the student model is defined as:

$$f(b_r^s) = I\left(\max_{r' \in \{1,\ldots,R\}} \mathrm{IoU}(b_r^s, b_{r'}^{gt}) > t\right) \tag{11}$$

where $\mathrm{IoU}(b_r^s, b_{r'}^{gt})$ is the formula for calculating the standard merger ratio. I is the indicator function. When the maximum intersection ratio exceeds the set threshold t, the prediction box is considered a positive sample. Otherwise, it is considered a negative sample.

To ensure consistency between distillation and classification tasks, binary cross-entropy loss is used as the basis for designing the classification head distillation loss function. This minimizes potential inconsistencies arising from differences in loss functions, thereby achieving more efficient knowledge distillation. A joint masked distillation operation is performed on the classification head distillation loss function to resolve these discrepancies, as shown in Eq. (12):

$$L_{\mathrm{cls}} = \frac{1}{N_c}\sum_{r=1}^{R} M^{t,s} f(b_r^s) \sum_{k=1}^{K} L_{\mathrm{BCE}}\left(p_{r,k}^t, p_{r,k}^s\right) \tag{12}$$

where L_{cls} represents the classification distillation loss, and L_{BCE} represents the binary cross-entropy loss. N_c is the normalization factor. $p_{r,k}^t$ and $p_{r,k}^s$ represent the prediction scores for class k at position i in the teacher detector and student detector, respectively.

In addition, for the regression loss of the object detection head, we adopt the $GIoU$ loss function, which takes into account not only the position and size of the bounding box but also its shape. The formula for the $GIoU$ loss is shown in Eq. (13):

$$L_{GIoU}\left(b_r^t, b_r^s\right) = 1 - \frac{I\left(b_r^t, b_r^s\right)}{U\left(b_r^t, b_r^s\right)} + \frac{A\left(b_r^t, b_r^s\right) - U\left(b_r^t, b_r^s\right)}{A\left(b_r^t, b_r^s\right)} \tag{13}$$

where b_r^t and b_r^s represent the r-th bounding box from the teacher detector and the student detector, respectively. $I(\cdot)$ and $U(\cdot)$ denote the intersection area and union area of the two bounding boxes, respectively. $A(\cdot)$ represents the area of the smallest bounding box that contains the two rectangles. The regression head distillation loss function is given by Eq. (14):

$$L_{\mathrm{reg}} = \frac{1}{N_r}\sum_{r=1}^{R} M^{t,s} f(b_r^s) L_{\mathrm{GIoU}}\left(b_r^t, b_r^s\right) \tag{14}$$

3.4 Overall Loss Function

The overall loss function of the method proposed in this paper consists of five components: L_{det} represents the original detection task loss of the student model. L_{global} represents the global feature alignment loss of the feature difference distillation module. L_{local} represents the local feature alignment loss of the feature difference distillation module. L_{cls} represents the classification head distillation loss of the response difference distillation module. L_{reg} represents the regression head distillation loss of the response difference distillation module. The overall loss function is defined as shown in Eq.(15):

$$L_{\text{loss}} = L_{\text{det}} + \alpha L_{\text{global}} + \beta L_{\text{local}} + \gamma L_{\text{cls}} + \eta L_{\text{reg}} \tag{15}$$

where α, β, γ, and η are hyperparameters used to balance the individual loss components. The overall distillation loss is calculated solely based on the feature maps and response outputs, without involving the internal structure of the detector. Consequently, the proposed method exhibits excellent adaptability and can be applied to various mainstream object detection architectures.

4 Results and Discussion

4.1 Dataset and Experimental Environment Configuration

To verify the effectiveness and efficiency of the proposed method, We conducted experimental validation using the MS COCO object detection dataset [23] and Pascal VOC 2007 dataset [24]. The training phase utilized the trainval135k subset, comprising 115,000 labeled images. The validation phase employed the minival validation subset, containing 5000 images, to assess model performance during training. Finally, the MS COCO test-dev 2019 dataset, which contains 20,000 images, was used for evaluation. Pascal VOC 2007 dataset contains 4952 images covering 20 object categories, including people, animals, vehicles, and indoor objects, among others.

The model was trained using the stochastic gradient descent (SGD) optimizer throughout the entire process, with a momentum coefficient of 0.9 and a weight decay coefficient of 0.0001. BMGKD is compatible with various mainstream object detection architectures, including the two-stage detector Faster-RCNN, the anchor-based single-stage detectors RetinaNet and GFL, and the anchor-free single-stage detector FCOS. All methods were trained on four NVIDIA A100 GPUs for 12 training epochs. The evaluation adopted the standard COCO metric of average precision (AP), and reporting the mean average precision (mAP) values at intersection-over-union (IoU) thresholds of 0.5 and 0.75, as well as the AP values for small, medium, and large-scale targets. All results are based on three independent replicates and report the average performance. The experiments demonstrate that the improvements made to BMGKD are stable and reliable, with a standard deviation of mAP between runs of less than 0.2%.

4.2 Comparison Experiments

As shown in Table 1, we conducted a comprehensive performance comparison between our proposed method and other state-of-the-art methods (TBD [25], SamKD [21], CSAKD [18], FGD [8], BCKD [20], DrKD [26], CrossKD [27], HMKD [28], MLD [29], and PCD [30]) on the COCO dataset. On the two-stage Faster R-CNN, single-stage anchor-free GFL, and single-stage anchor-based RetinaNet detector architectures, we evaluate the performance of BMGKD under three teacher-student configurations: the standard capacity gap ResNet101 to ResNet50, the larger capacity gap ResNet101 to ResNet18, and the cross-architecture ResNet50 to MobileNetV3. In all experiments, BMGKD achieved the highest mAP and outperformed other models across all evaluation metrics, including AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. Specifically, in the ResNet101→ResNet50 configuration, BMGKD achieved mAPs of 41.1%, 44.4%, and 40.4% on the three detectors, respectively, representing improvements of 2.7, 4.2, and 2.8 percentage points over the baseline, significantly outperforming existing methods. When the capacity gap between teachers and students widened further from ResNet101 to ResNet18, BMGKD still achieved mAPs of 37.4%, 40.0%, and 34.6%, respectively, maintaining a clear lead. In the more challenging cross-architecture scenario from ResNet50 to MobileNetV3, BMGKD ranked first with mAP scores of 36.9%, 39.4%, and 34.8%. These results fully demonstrate that BMGKD can consistently handle different detection architectures, varying degrees of capacity differences, and cross-backbone family transfer requirements, showcasing exceptional consistency and generalization capabilities.

Table 1: Comparison of distillation results across different detectors and teacher-student configurations.

Detector	Teacher	Student	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	ResNet101	ResNet50	Baseline	38.4	59.0	42.0	21.5	42.1	50.3
			TBD (2023)	39.8 (+1.4)	57.2	40.1	21.9	42.3	50.6
			SamKD (2025)	40.6 (+2.2)	61.0	44.3	23.5	44.7	51.0
			CSAKD (2025)	40.7 (+2.3)	61.1	44.4	23.6	44.7	51.4
			BMGKD	41.1 (+2.7)	61.5	44.7	23.8	45.0	52.5
	ResNet101	ResNet18	Baseline	34.5	54.6	37.2	19.2	36.8	45.2
			TBD (2023)	35.4 (+0.9)	55.7	38.1	20.4	37.9	46.7
			SamKD (2025)	36.0 (+1.5)	56.2	38.6	20.7	38.3	46.8
			CSAKD (2025)	36.6 (+2.1)	56.7	39.5	21.6	39.4	47.4
			BMGKD	37.4 (+2.9)	57.4	40.0	21.9	39.8	48.1
	ResNet50	MobileNetV3	Baseline	34.3	54.3	37.0	18.9	36.5	45.0
			FGD (2022)	35.4 (+1.1)	54.8	37.4	19.4	38.6	46.3
			BCKD (2023)	36.2 (+1.9)	56.4	39.1	20.9	39.0	47.0
			DrKD (2025)	36.6 (+2.3)	56.5	39.4	21.1	39.3	47.4
			BMGKD	36.9 (+2.6)	57.0	39.9	21.5	39.6	47.6
GFL	ResNet101	ResNet50	Baseline	40.2	58.4	43.3	23.3	44.0	52.2
			BCKD (2023)	42.1 (+1.9)	60.2	45.5	25.3	46.5	55.6
			CrossKD (2024)	42.4 (+2.2)	59.4	45.0	24.3	45.6	53.4
			HMKD (2026)	42.9 (+2.7)	61.6	46.9	25.7	47.3	55.9
			DrKD (2025)	43.2 (+3.0)	61.3	46.8	26.3	47.8	56.3
			BMGKD	44.4 (+4.2)	62.7	48.3	27.4	49.0	56.8
	ResNet101	ResNet18	Baseline	35.7	53.5	38.0	19.7	38.7	46.6
			BCKD (2023)	37.8 (+2.1)	55.7	40.3	21.8	40.9	48.9
			CrossKD (2024)	38.4 (+2.4)	56.1	40.7	22.0	41.2	49.4
			HMKD (2026)	38.7 (+3.0)	56.8	41.2	22.4	41.9	50.0
			DrKD (2025)	38.6 (+2.9)	56.8	41.3	22.4	42.0	49.8
			BMGKD	40.0 (+4.3)	57.9	42.3	23.5	43.2	51.4
	ResNet50	MobileNetV3	Baseline	35.4	53.3	37.7	19.4	38.5	46.4
			SamKD (2025)	38.0 (+2.6)	56.1	40.4	22.1	41.0	49.2
			DrKD (2025)	38.0 (+2.6)	56.2	40.3	22.0	41.2	49.4
			HMKD (2026)	38.7 (+3.3)	56.6	41.2	23.1	42.0	50.3
			BMGKD	39.4 (+4.0)	57.3	41.9	23.7	42.8	51.0
RetinaNet	ResNet101	ResNet50	Baseline	37.4	55.4	39.1	20.4	40.3	48.1
			FGD (2022)	39.1 (+1.7)	59.0	42.3	22.8	43.1	52.3
			CrossKD (2024)	39.7 (+2.3)	58.9	42.5	22.4	43.6	52.8
			HMKD (2026)	39.9 (+2.5)	59.1	42.8	22.3	43.9	53.6
			BMGKD	40.4 (+2.8)	59.4	43.0	23.1	44.1	54.0
	ResNet101	ResNet18	Baseline	31.6	49.6	33.3	16.1	34.2	43.0
			FGD (2022)	33.9 (+1.9)	51.6	35.2	18.3	36.3	44.9
			CrossKD (2024)	34.1 (+2.5)	52.2	36.1	18.6	36.9	47.4
			HMKD (2026)	34.2 (+2.6)	52.3	35.9	18.7	37.0	47.5
			BMGKD	34.6 (+3.0)	52.5	36.1	19.0	37.3	47.9
	ResNet50	MobileNetV3	Baseline	31.4	49.3	33.1	15.8	34.1	42.7
			MLD (2023)	34.1 (+2.7)	52.0	35.9	18.4	37.0	46.9
			PCD (2025)	34.2 (+2.8)	52.1	36.0	18.6	37.1	46.9
			CSAKD (2025)	34.5 (+3.1)	52.4	36.3	18.8	37.2	47.1
			BMGKD	34.8 (+3.4)	52.8	36.5	19.4	37.6	47.6

As shown in Table 2, on the Pascal VOC 2007 dataset, when using ResNet101 as the teacher network to train the student network ResNet50, BMGKD achieved the best performance across all three detectors. Faster R-CNN, RetinaNet, and GFL achieved mAPs of 56.2%, 58.0%, and 58.3%, respectively. These results not only significantly outperformed the student baselines but also comprehensively surpassed the corresponding teacher models, maintaining a clear lead over existing methods such as AMD, BCKD, and SamKD. These results fully validate BMGKD's effectiveness across different datasets, its generalization capability, and its ability to elevate student models to a level surpassing that of the teacher.

Table 2: Performance comparison on Pascal VOC 2007 dataset.

Method	Faster R-CNN			RetinaNet			GFL		
	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Teacher	56.3	82.7	62.6	58.2	82.0	63.0	58.4	81.6	64.3
Student	54.2	82.1	59.9	56.1	80.9	60.7	56.1	80.0	61.6
AMD [31] (2023)	55.3	82.3	61.4	57.2	81.2	62.1	57.5	80.6	62.9
BCKD [20] (2023)	55.6	82.4	61.9	57.5	81.4	62.5	57.9	80.9	63.3
SamKD [21] (2025)	55.9	82.5	62.3	57.8	81.6	62.9	58.1	81.1	63.7
BMGKD	56.2	82.6	62.4	58.0	81.8	63.1	58.3	81.3	64.0

Table 3 compares the training overhead and accuracy gains of various distillation methods under the GFL ResNet101→ResNet18 setting. None of the methods alter the student architecture, and there are no additional parameters or computational costs during the inference stage. Compared to the response-based method CrossKD, the feature-based method FGD, and the hybrid method SamKD, BMGKD requires only 5.9 GFLOPs of additional computational effort and 1.14 times the training duration, achieving the highest mAP improvement of 4.6% and striking the optimal balance between training efficiency and accuracy.

Table 3: Training cost and performance improvement of different KD methods.

Model	Type	Additional GFLOPs (Training)	Training Time (Relative to Baseline)	Δ of mAP (%)
CrossKD [27]	Response-based	+1.9	1.03×	+2.7%
FGD [8]	Feature-based	+4.8	1.12×	+2.5%
SamKD [21]	Response + Feature	+9.5	1.18×	+3.9%
BMGKD	Response + Feature	+5.9	1.14×	+4.6%

4.3 Hyperparameter Analysis

To determine the optimal loss weights, this section conducted hyperparameter sensitivity experiments using the control variable method based on the GFL detection framework, with ResNet-101 serving as the teacher model and ResNet-50 as the student model. As shown in Table 4, the optimal hyperparameter settings for BMGKD were determined to be $\alpha = 0.04$, $\beta = 0.02$, and $\gamma = \eta = 1.0$. Under this configuration, the student model achieved a mAP of 44.4% on the GFL detection framework. Searches in different directions within the neighborhood of the optimal values did not reveal any performance gains. All deviations led to a consistent decline in detection accuracy.

Table 4: Parameter sensitivity analysis.

α	β	γ	η	mAP (%)	Δ from Best
0.04	0.02	1.0	1.0	44.4	–
0.04	0.02	0.75	0.75	44.0	−0.4
0.04	0.02	1.25	1.25	44.1	−0.3
0.04	0.02	1.5	1.5	44.2	−0.2
0.04	0.02	1.25	0.75	44.2	−0.2
0.02	0.03	1.0	1.0	44.0	−0.4
0.03	0.04	1.0	1.0	43.9	−0.5
0.06	0.03	1.0	1.0	43.8	−0.6
0.06	0.04	1.0	1.0	43.9	−0.5

4.4 Ablation Experiments

To validate the effectiveness of each module in the BMGKD method, we used ResNet-101 and ResNet-50 as teacher-student backbone networks and conducted systematic ablation experiments based on the single-stage anchor-free detector GFL. The experimental results are shown in Table 5. When using only the global feature difference alignment module, the model's detection accuracy improved by 2.8%. When using only the local feature difference alignment module, the model's detection accuracy improved by 2.6%. After combining the global and local feature difference modules, the accuracy improvement increased to 3.8%. Furthermore, when the feature difference module was integrated with the response output difference module, the model achieved optimal performance, with an mAP improvement of 4.2% compared to the baseline model. The experimental results fully demonstrate that the complementary collaboration among the modules in the proposed method comprehensively improves the model's final detection accuracy.

Table 5: Ablation experiments.

Global	Local	Response	mAP	AP_S	AP_M	AP_L
×	×	×	40.2	23.3	44.0	52.2
✓	×	×	43.0 (+2.8)	26.0	46.5	55.1
×	✓	×	42.8 (+2.6)	26.2	46.5	54.9
✓	✓	×	44.0 (+3.8)	26.9	47.3	55.7
✓	✓	✓	44.4 (+4.2)	27.4	49.0	56.8

4.5 Data Visualization

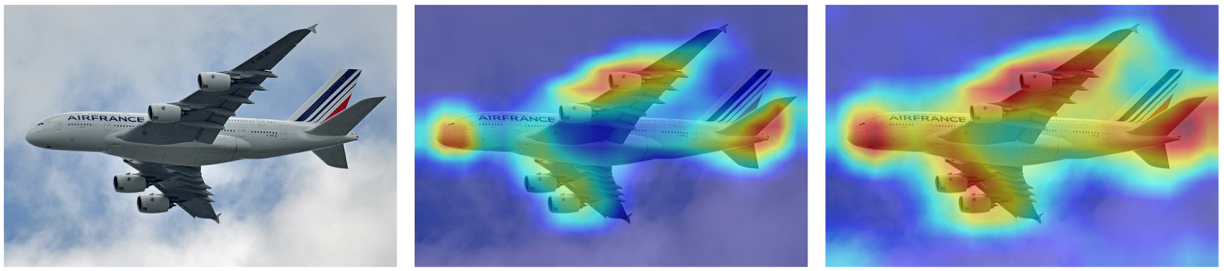
Fig. 2 shows the object detection visualization results of the teacher model, student model, and the proposed BMGKD method across different scenarios. The pure student model exhibits significant performance shortcomings. It not only makes redundant errors such as misclassifying a bench as an object but also suffers from issues such as inflated confidence scores and insufficient bounding box accuracy, resulting in a significant gap in detection performance compared to the teacher model. In contrast, the BMGKD method, through its bidirectional mask-guided knowledge distillation mechanism, successfully transfers the teacher model's target discrimination capability to the student model. This not only eliminates false background detections in the student model but also calibrates the detection confidence to a level highly consistent

with that of the teacher model, demonstrating greater robustness and detection reliability in scenarios with multiple targets and complex background interference.

(a) Original image **(b)** Teacher Model **(c)** Student Model **(d)** BMGKD

Figure 2: Comparison of predicted results. (**a**) Original image. (**b**) Teacher model. (**c**) Student model. (**d**) BMGKD.

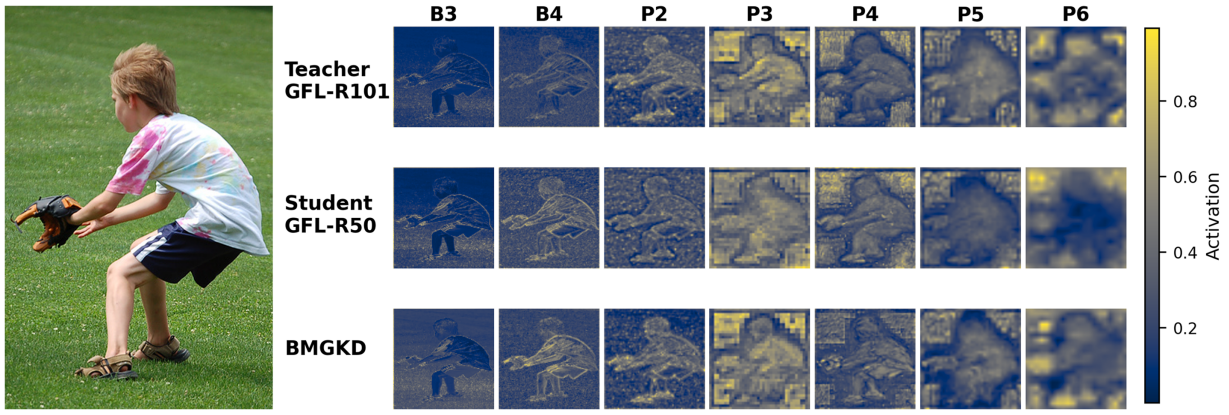
To visually demonstrate the proposed method's ability to focus on key regions of interest, we selected a representative test image from the MS COCO dataset and visualized the attention heatmap extracted from the GFL-ResNet50 student model after 12 epochs of training. The results are shown in Fig. 3. The intensity of the colors in the heatmap indicates the level of attention weighting, with red areas representing the highest level of attention and blue areas representing the lowest. As shown in the attention map for aircraft class detection, BMGKD can accurately focus on three key feature regions, the aircraft nose, engines, and tail, while filtering out redundant background interference. This fully validates that the collaborative masking design in the feature difference module can effectively enhance the student network's ability to understand and distinguish target features.



(a) Original image (b) DMKD (c) BMGKD

Figure 3: Attention comparison diagram. (**a**) Original image. (**b**) DMKD. (**c**) BMGKD.

As shown in Fig. 4, we extracted feature maps from seven layers B3 and B4 in the backbone, and P2 through P6 in the neck of the GFL detection network. We calculated the channel averages layer by layer to obtain two-dimensional activation matrices, which were then visualized using a Cividis color map, where higher brightness indicates stronger activation. After BMGKD distillation, the feature activation distributions of the student model at each layer closely align with those of the teacher model. In the deeper layers P3 to P5, the student model successfully reproduces the semantic regions and object contours of interest to the teacher, whereas the activations of the original student model were relatively diffuse and lacked sufficient semantic distinctiveness. These results demonstrate that the global and local alignment mechanisms within the feature difference distillation module effectively transfer the teacher's tacit knowledge to the student model.



(a) Original image **(b)** Heatmap of feature channel activation values

Figure 4: Feature visualization. (**a**) Original image. (**b**) Heatmap of feature channel activation values.

5 Conclusion

To effectively improve the performance of knowledge distillation, we propose an object detection knowledge distillation algorithm designed to bridge the multidimensional differences between student and teacher networks. By simultaneously processing global and local features, this algorithm specifically addresses the differences in feature representations between the teacher network and the student network. In addition, at the model's response output layer, we designed a differentiated knowledge distillation strategy specific to the classification and regression tasks of object detection. This further bridges task-specific differences and enhances the student network's ability to understand the implicit knowledge of the teacher network.

Extensive comparative experiments, ablation experiments, and visualization results demonstrate that the proposed BMGKD algorithm delivers outstanding performance, with each module contributing positive gains, and exhibits strong competitiveness across various mainstream object detection frameworks. In future research, we will continue to advance lightweight feature distillation and loss function simplification to reduce resource consumption during training. At the same time, we will focus on exploring feature alignment and masking strategies that are more robust to extreme scales, occlusions, and long-tail classes, designing class-aware distillation mechanisms, and introducing learnable intermediate adaptation modules to further bridge the semantic gap across architectures.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by Army Engineering University of PLA. The authors would like to acknowledge Army Engineering University of PLA for funding this work.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Tianqi Wang; data collection: Yang Li; analysis and interpretation of results: Tianqi Wang and Zhisong Pan; draft manuscript preparation: Zhisong Pan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in MS COCO at https://cocodataset.org.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu H, Cao P, Jiao L, Li X, Hou B, Yi X, et al. A progressive semi-distillation model for dual-source remote sensing image classification. IEEE Trans Cybern. 2026;56(1):67–80. doi:10.1109/TCYB.2025.3616504.
2. Tang Z, Wu Z, Li M, Wen J, Zhang B, Xu Y, et al. Adaptive fine-grained fusion network for multimodal UAV object detection. IEEE Trans Image Process. 2026;35:1870–82. doi:10.1109/tip.2026.3661868.
3. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531. 2015.
4. Li Q, Jin S, Yan J. Mimicking very efficient network for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2017. p. 7341–9.
5. Chen G, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 742–51.
6. Wu K, Wang L, Duan S. Head-adaptive knowledge distillation for dense object detection. In: 2025 8th International Symposium on Big Data and Applied Statistics (ISBDAS). Piscataway, NJ, USA: IEEE; 2025. p. 857–60.
7. Lan Q, Tian Q. Gradient-guided knowledge distillation for object detectors. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE; 2024. p. 423–32.
8. Yang Z, Li Z, Jiang X, Gong Y, Yuan Z, Zhao D, et al. Focal and global knowledge distillation for detectors. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2022. p. 4633–42.
9. Wang R, Xu Y, Wu Z, Wei Z. Distilling object detectors with scale-conscious knowledge in remote sensing images. IEEE Geosci Remote Sens Lett. 2026;23:1–5. doi:10.1109/lgrs.2025.3634080.
10. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2015. p. 1440–8.
11. Ricki R, Dicky D, Apriyansyah B. Fast region-based convolutional neural network in object detection: a review. Intl J Adv Comput Inform. 2025;2(1):34–40. doi:10.71129/ijaci.v2i1.pp34-40.
12. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2017. p. 2980–8.
13. Jia N, Yang J, Liu X, Sun Y. F-YOLO: delving into fuzzy YOLO for improved traffic object detection. IEEE Trans Fuzzy Syst. 2026;34(2):440–52.
14. Tian Z, Shen C, Chen H, He T. Fcos: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2019. p. 9627–36.
15. Li X, Wang W, Hu X, Li J, Tang J, Yang J. Generalized focal loss V2: learning reliable localization quality estimation for dense object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2021. p. 11627–36.
16. Romero A. Fitnets: hints for thin deep nets. arXiv:1412.6550. 2014.

17. Lv J, Yang H, Li P. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. Adv Neural Inf Process Syst. 2024;37:65445–75.
18. Li M, Liang X, Hu Q, Ye L, Xia C. Multi-scale feature fusion with knowledge distillation for object detection in aerial imagery. Eng Appl Artif Intell. 2025;158(2):111518. doi:10.1016/j.engappai.2025.111518.
19. Zheng Z, Ye R, Hou Q, Ren D, Wang P, Zuo W, et al. Localization distillation for object detection. IEEE Trans Pattern Anal Mach Intell. 2023;45(8):10070–83. doi:10.1109/tpami.2023.3248583.
20. Yang L, Zhou X, Li X, Qiao L, Li Z, Yang Z, et al. Bridging cross-task protocol inconsistency for distillation in dense object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2023. p. 17129–38.
21. Zhang Z, Li J, Li J, Xu J. SAMKD: spatial-aware adaptive masking knowledge distillation for object detection. arXiv:2501.07101. 2025.
22. Saha A, Bialkowski A, Khalifa S. Distilling representational similarity using centered kernel alignment (CKA). In: Proceedings of the the 33rd British Machine Vision Conference (BMVC 2022). London, UK: British Machine Vision Association; 2022. p. 1–12.
23. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2014. p. 740–55.
24. Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes challenge: a retrospective. Intl J Comput Vis. 2015;111(1):98–136. doi:10.1007/s11263-014-0733-5.
25. Tang R, Liu Z, Li Y, Song Y, Liu H, Wang Q, et al. Task-balanced distillation for object detection. Pattern Recognit. 2023;137(2):109320. doi:10.1016/j.patcog.2023.109320.
26. Lv Y, Cai Y, He Y, Li M. DrKD: decoupling response-based distillation for object detection. Pattern Recognit. 2025;161:111275. doi:10.1016/j.patcog.2024.111275.
27. Wang J, Chen Y, Zheng Z, Li X, Cheng MM, Hou Q. CrossKD: cross-head knowledge distillation for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 16520–30.
28. Chen Y, Gao Z, Yan J, Liu Y. Remote sensing object detection through hierarchical feature mining and multivariate head collaboration with knowledge distillation. Neural Netw. 2026;195(10):108205. doi:10.1016/j.neunet.2025.108205.
29. Jin Y, Wang J, Lin D. Multi-level logit distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2023. p. 24276–85.
30. Li J, Li J, Zhang Z, Xu J. Progressive class-level distillation. arXiv:2505.24310. 2025.
31. Yang G, Tang Y, Li J, Xu J, Wan X. Amd: adaptive masked distillation for object detection. In: 2023 International Joint Conference on Neural Networks (IJCNN). Piscataway, NJ, USA: IEEE; 2023. p. 1–8.