Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060609

ARTICLE

An Improved Knowledge Distillation Algorithm and Its Application to Object Detection

Min Yao[1,*], Guofeng Liu[2], Yaozu Zhang[3] and Guangjie Hu[1]

[1]School of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China
[2]Baidu, Beijing, 100000, China
[3]Shanghai Freesense Technology Co., Ltd., Shanghai, 200000, China

*Corresponding Author: Min Yao. Email: minyao@shmtu.edu.cn

Received: 06 November 2024; Accepted: 08 February 2025; Published: 16 April 2025

ABSTRACT: Knowledge distillation (KD) is an emerging model compression technique for learning compact object detector models. Previous KD often focused solely on distilling from the logits layer or the feature intermediate layers, which may limit the comprehensive learning of the student network. Additionally, the imbalance between the foreground and background also affects the performance of the model. To address these issues, this paper employs feature-based distillation to enhance the detection performance of the bounding box localization part, and logit-based distillation to improve the detection performance of the category prediction part. Specifically, for the intermediate layer feature distillation, we introduce feature resampling to reduce the risk of the student model merely imitating the teacher model. At the same time, we incorporate a Spatial Attention Mechanism (SAM) to highlight the foreground features learned by the student model. In terms of output layer feature distillation, we divide the traditional distillation targets into target-class objects and non-target-class objects, aiming to improve overall distillation performance. Furthermore, we introduce a one-to-many matching distillation strategy based on Feature Alignment Module (FAM), which further enhances the student model's feature representation ability, making its feature distribution closer to that of the teacher model, and thus demonstrating superior localization and classification capabilities in object detection tasks. Experimental results demonstrate that our proposed methodology outperforms conventional distillation techniques in terms of object detecting performance.

KEYWORDS: Deep learning; model compression; knowledge distillation; object detection

1 Introduction

In object detection [1–3], efficient model architectures typically consist of two core components: the backbone network responsible for extracting image features, and the detection head responsible for predicting object bounding boxes and classification. These algorithms are mainly divided into two major paradigms: anchor-based and anchor-free methods. Anchor-based algorithms are further divided into single-stage [4–7] and two-stage methods [8,9]; the former offers high computational efficiency but lower accuracy, while the latter generates proposals through Region Proposal Network (RPN) before classification, leading to higher accuracy but increased computational costs. Anchor-free algorithms [10–13] predict object locations through key points, balancing inference speed and accuracy while avoiding the computational overhead of anchor generation. Although these high-performance models excel in accuracy, they often face issues of slow inference speed and high computational costs, which limit their deployment in resource-constrained environments. To address this issue, Hinton knowledge distillation technology [14], which

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

enables small student models to learn from large teacher models without adding extra computational costs. Knowledge distillation is mainly divided into two categories: distillation based on logit outputs [14] and distillation based on intermediate layer features [15–18], with the latter being widely used in various tasks due to its rich semantic information.

FitNets [16] guides the training of the student network by utilizing the intermediate layer features of the teacher network, focusing on the scale of feature responses after ReLU activation and the activation state of each neuron. Chen et al. [19] were the first to apply knowledge distillation technology to the field of object detection, effectively addressing the imbalance between foreground and background by aligning the intermediate layer features of teacher and student models through prompt learning. Focal and Global Distillation (FGD) [20] proposed local and global distillation strategies. Feature-based Knowledge Distillation (FKD) [21] introduces attention-based foreground pixel extraction and non-local distillation. CrossKD [22] transmits the intermediate features from the student detection head to the teacher detection head and generates cross-head predictions for knowledge distillation, thereby resolving the target conflict issue. Cosine Similarity-Based Knowledge Distillation (CSKD) [23] employs Cosine Similarity (CS) as an additional distillation loss metric and distillation guidance map, facilitating comprehensive distillation across spatial and channel dimensions. Traditional feature-level distillation pursues the similarity between the student model's features and the teacher model's features, while Yang et al. [24] proposed the Masked Generative Distillation (MGD) method, which selectively generates the complete features of the teacher network from the student network, avoiding direct imitation and thus reducing the acquisition of redundant information.

Although MGD performs well on various models, its mask matrix is randomly generated and does not take into account the imbalance between foreground and background features, as well as the importance of highlighting foreground features in object detection. To address this issue, this study introduces a spatial attention mechanism [25] to optimize the mask matrix, increasing the sampling rate of foreground pixels in the student model, thereby improving distillation performance. Traditional logit distillation is usually performed by minimizing the Kullback-Leibler (KL) divergence between the teacher and student networks. This method, while universal and low in computational storage cost, may not achieve the best performance. To balance universality and performance, this study adopts a logit distillation method [26] designed for object detection networks, namely Decoupled Knowledge Distillation (DKD), which decomposes traditional logit distillation into two parts: distillation for target category objects and distillation for non-target category objects, and enhances overall performance by weighting these two types of distillation. In summary, this paper's main contributions are as follows:

(1) We propose a Combined Knowledge Distillation (CKD) method that integrates logit distillation and feature intermediate layer distillation, enabling the student network to effectively learn from the teacher and enhance detection performance.
(2) We enhance the intermediate layer feature distillation with a Spatial Attention Mechanism (SAM), modifying the mask matrix in Masked Generative Distillation (MGD) to improve foreground feature representation and address foreground-background imbalance.
(3) We propose a one-to-many matching distillation strategy based on Feature Alignment Module (FAM). This strategy allows the teacher network to guide multiple features in the student network, achieving more effective knowledge transfer by precisely aligning key features.

2 Research Methodology

Fig. 1 illustrates our proposed innovative Combined Knowledge Distillation (CKD) method, which integrates the latest feature intermediate layer and logit output layer distillation techniques aimed at enhancing model performance. During the feature intermediate layer distillation, we introduce a spatial

attention mechanism to improve the mask matrix in MGD, increasing the sampling rate of foreground pixels in the student network and making the distillation process more accurate and efficient. Furthermore, the teacher network guides the student network through a one-to-many matching distillation based on FAM, further optimizing the distillation effects.

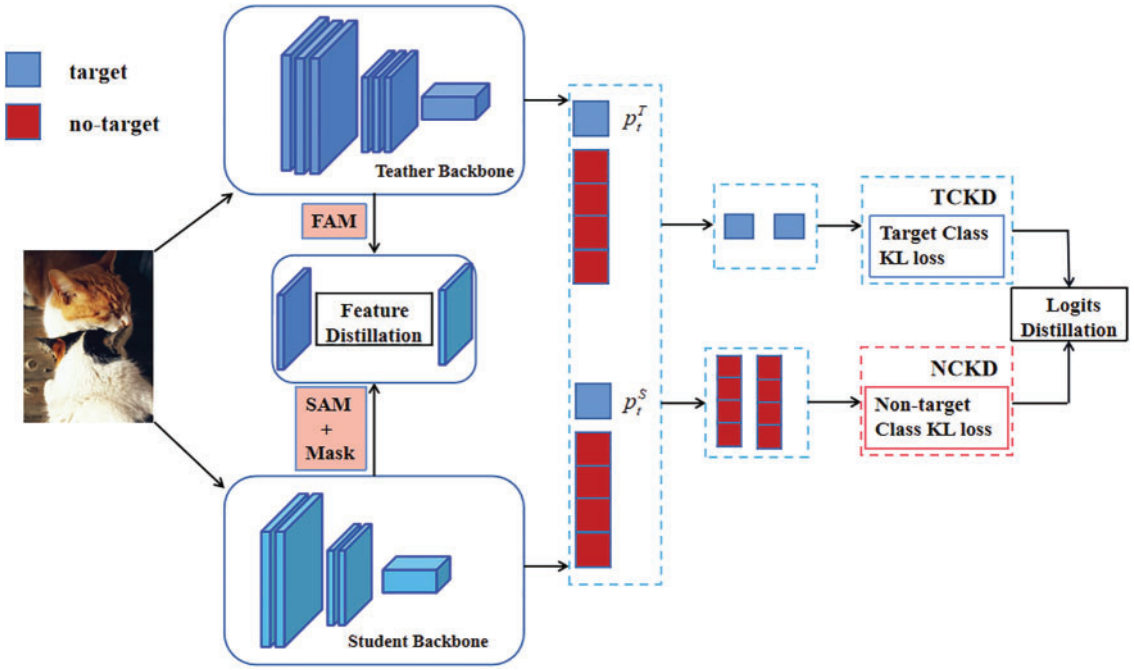


Figure 1: Combined Knowledge Distillation (CKD)

2.1 Decoupling Logit Output Layer Distillation

2.1.1 Logit Knowledge Distillation

The generalized softmax function proposed by Hinton can soften the output labels of the teacher network, making the logit distribution smoother, thereby highlighting the "dark knowledge" within the teacher network. The function is defined as in Eq. (1):

$$p_i = \frac{exp\left(\frac{z_i}{T}\right)}{\sum_{j=1}^{C} exp\left(\frac{z_i}{T}\right)}, \tag{1}$$

where p_i represents the output probability for class i, C represents the total number of classes, and z_i represents the logit output for class i. T represents the distillation temperature. When T is 1, the generalized softmax function degenerate into the regular softmax function. Increasing the distillation temperature T leads to a more uniform probability distribution in the softmax output of the model. In addition to the probability of the target class, the probabilities of the non-target classes represent the "dark knowledge" available for the student to learn from the teacher network. The following figure illustrates the difference between the probability distributions of the generalized softmax output and the regular softmax output.

The concept of "dark knowledge" refers to knowledge that is acquired through machine learning but not yet fully understood by humans. In the context of neural networks and knowledge distillation, "dark knowledge" specifically refers to information that is embedded in the logit outputs of neural networks but not explicitly provided to the model. For example, while the likelihood of misclassifying an image of a cat as

a dog is low, this possibility is still much higher than the likelihood of mistaking the cat for a car. This subtle probability of misclassification is part of the "dark knowledge".

As shown in Fig. 2, the first handwritten digit is quite similar to 3 and 9. Therefore, after being processed by the generalized softmax function, the probabilities of categories 3 and 9 in the soft labels increase. Similarly, for the second handwritten digit, which resembles both digits 1 and 9, the corresponding probabilities in the soft labels also increase. This representation of probabilities for non-target classes effectively captures the "dark knowledge" embedded in the logit outputs of the teacher network.

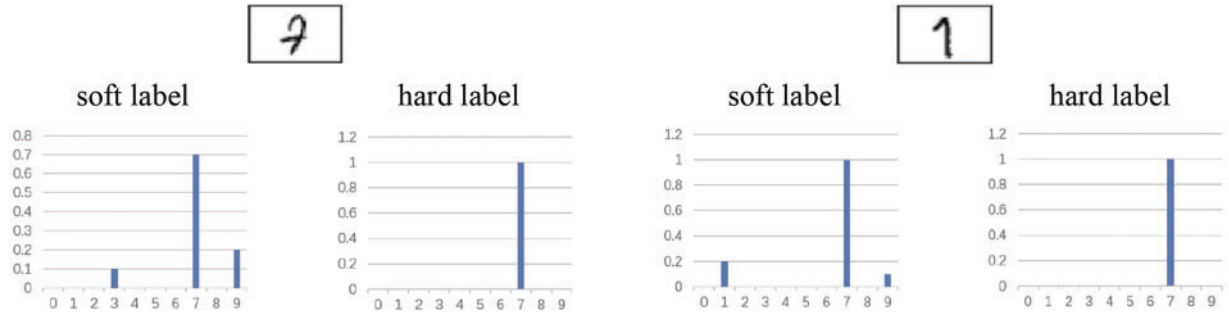


Figure 2: Soft labels and dark knowledge

During the process of knowledge distillation, the soft targets from the teacher network can convey this "dark knowledge" to the student network, helping the student network learn the implicit patterns and correlations within the teacher network. This knowledge is invaluable to the student network because it includes subtle and complex relationships between categories that are not visible in hard targets, such as one-hot encoded labels. In this way, the student network can not only learn the direct outputs of the teacher network but also absorb deep, subtle knowledge that may not be immediately apparent, potentially leading to better performance in certain situations.

2.1.2 Decoupled Class

In the original logit output layer distillation method, the probabilistic output for each category is usually obtained from a softmax function, if given a probability distribution such as Eq. (2):

$$P = [p_1, p_2, \cdots, p_t, \cdots, p_c] \in R^{1\times C}. \tag{2}$$

Typically, the probability output for each category is obtained by the softmax function, as shown in Eq. (3):

$$p_i = \frac{exp(z_i)}{\sum_{j=1}^{C} exp(z_j)}, \tag{3}$$

where p_i represents the output probability for class i, C represents the total number of classes, and z_i represents the logit output for class i. To separate the category probabilities into target category and non-target categories, define the binary categorical probability distribution $b = [p_t, p_{\setminus t}] \in R^{1\times 2}$. The formulas for the probabilities of the two kinds are as shown in Eqs. (4) and (5):

$$p_t = \frac{exp(z_t)}{\sum_{j=1}^{C} exp(z_j)}, \tag{4}$$

$$p_{\backslash t} = \frac{\sum_{k=1,k\neq t}^{C} exp(z_k)}{\sum_{j=1}^{C} exp(z_j)}, \tag{5}$$

where t denotes the target category and $\backslash t$ denotes all non-target categories. The probabilities of all non-target outputs of the teacher network are then independently modeled as probability distributions $\widehat{P}$ (as shown in Eq. (6)):

$$\widehat{P} = [\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_{t-1}, \widehat{p}_{t+1}, \cdots \widehat{p}_C,] \in R^{1\times(C-1)}. \tag{6}$$

The formula for calculating the probability of each non-target category can be obtained as shown in Eq. (7):

$$\widehat{p}_t = \frac{\exp(z_i)}{\sum_{j=1,j\neq t}^{C} \exp(z_j)}. \tag{7}$$

Assuming that the training sample is a non-target category, the sum of the probabilities of all non-target categories is computed as 1, and the probability distribution of that non-target category is completely independent of the target category probabilities, as shown in Eq. (8):

$$\widehat{p}_i = \frac{p_i}{p_{\backslash t}}. \tag{8}$$

Traditional logit distillation is defined by KL divergence, as shown in Eq. (9):

$$KD = KL(p^T||p^S) = p^T \log\left(\frac{p_t^T}{p_t^S}\right) + \sum_{i=1,i\neq t}^{C} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right), \tag{9}$$

where T represents the teacher model, S represents the student model. With the above definitions, Eq. (9) can be rewritten as Eq. (10) based on Eqs. (2), (6) and (7):

$$\begin{aligned} KD &= p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + p_{\backslash t}^T \sum_{i=1,i\neq t}^{C} \widehat{p}_i^T \left(\log\left(\frac{\widehat{p}_i^T}{\widehat{p}_i^S}\right)\right) + \log\frac{p_{\backslash t}^T}{p_{\backslash t}^S} \\ &= \underbrace{p_t^T \log\left(\frac{p_t^T}{p_t^S}\right) + p_{\backslash t}^T \log\left(\frac{p_{\backslash t}^T}{p_{\backslash t}^S}\right)}_{KL(b^T||b^S)} + \underbrace{p_{\backslash t}^T \sum_{i=1,i\neq t}^{C} \widehat{p}_i^T \left(\log\left(\frac{\widehat{p}_i^T}{\widehat{p}_i^S}\right)\right)}_{KL(\widehat{p}^T||\widehat{p}^S)} \\ &= KL(b^T||b^S) + (1-p_t^T)KL(\widehat{p}^T||\widehat{p}^S). \end{aligned} \tag{10}$$

Obviously, the first term of Eq. (10) is completely determined by the binary probability distribution b, while the second term is influenced by the non-target category probability ratio $\widehat{P}$. Therefore, the logit distillation is decoupled into target category knowledge distillation (TCKD) and non-target category knowledge distillation (NCKD). The distillation loss of the decoupled logit distillation method can be found as Eq. (11):

$$KD = TCKD + \left(1-p^T\right) NCKD = \alpha \cdot TCKD + \beta \cdot NCKD. \tag{11}$$

With the two hyperparameters α and β, it is possible to analyze and regulate the effect of both on the final distillation results. Subsequent experiments have shown that each of these components plays an important

role and they are to some extent interrelated. Decoupling them can improve the overall performance of logarithmic distillation.

Fig. 3 illustrates a comparison of the network structures between traditional logit distillation and decoupled logit distillation methods. As DKD optimizes the traditional logit distillation method and is suitable for deep learning models in multi-classification tasks, this paper adopts this distillation approach to distill the detection head of the target detection distillation model. However, DKD needs to take into account the imbalance between the foreground and background, as well as balance the knowledge transfer during the distillation process and the learning ability of the student network itself.

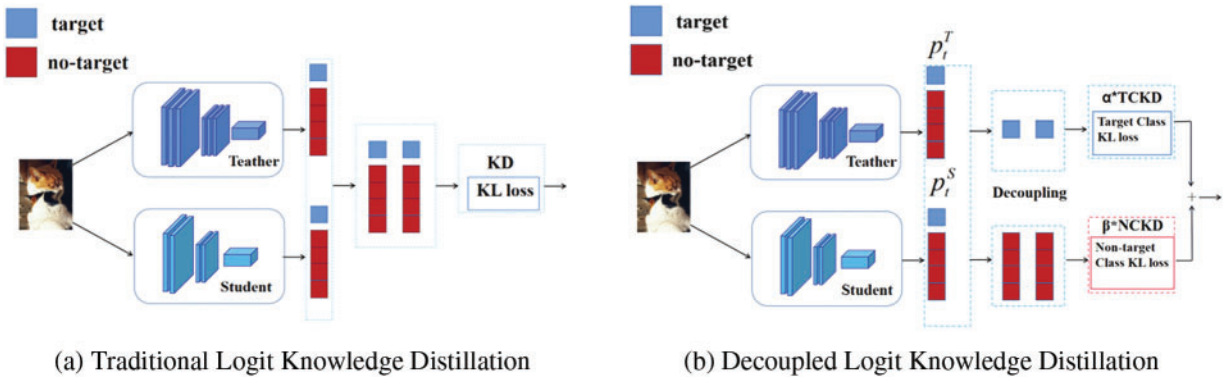


Figure 3: Decoupled Knowledge Distillation (DKD)

2.2 Distillation of Intermediate Layer Features Based on Feature Generation

2.2.1 Feature Resampling Distillation

In detection tasks, background features dominate, leading to an overemphasis on background loss during optimization, which can degrade performance. MGD shows that enhancing student network performance does not require an exact copy of the teacher network's feature maps. This paper resamples the student network's feature maps to eliminate unnecessary features, making the imitation process more efficient. SAM emphasize the importance of pixels at different spatial locations in the image. Thus, this paper integrates spatial attention during sampling to retain key foreground pixels crucial for object detection (Fig. 4a).

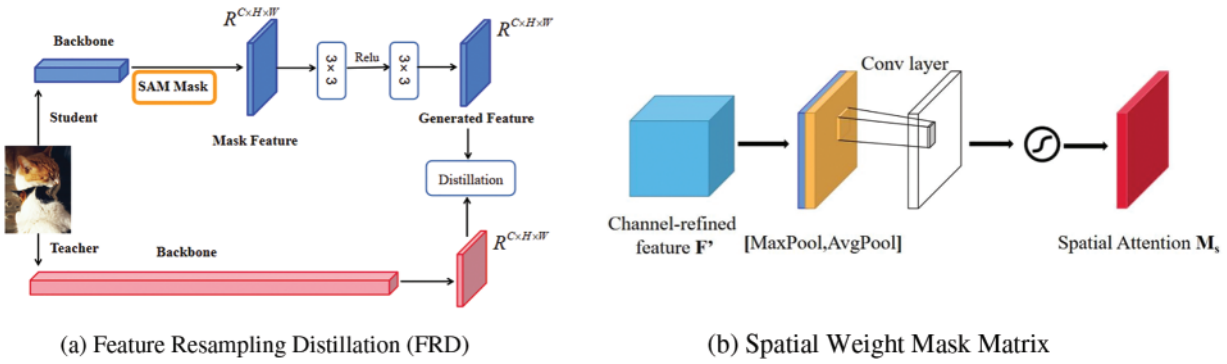


Figure 4: Structure of Feature Resampling Distillation (FRD)

The workflow is as follows: The student model's feature maps serve as inputs, and the spatial attention mechanism generates a spatial weight mask matrix. This matrix is applied to mask the student feature maps, achieving feature resampling. Next, alignment between student and teacher feature maps is performed using two 3 × 3 convolution layers and one adaptive layer. Finally, traditional feature-based distillation is applied to both sets of feature maps. Fig. 4a outlines the structure of feature resampling distillation, which is broadly applicable to CNNs extracting knowledge from intermediate layers. Experiments show that this method outperforms traditional feature-based distillation.

The process of generating spatial weight mask matrix is illustrated in Fig. 4b. First, the feature map of the student network is maximized and averaged over the channel dimensions to generate two new matrices, respectively. Then, concatenate the two into a convolutional layer with 2 channels. Finally, the spatial weight mask matrix M_s, with a final channel number of 1, is obtained through a 3 × 3 convolutional layer and a sigmoid function. The formula is as follows (Eq. (12)):

$$M_s = \sigma(Conv_{3\times3}([MaxPool(F'); AvgPool(F')])). \tag{12}$$

Based on the SAM module proposed by [25], this paper first extracts a relevant vector from the front-layer feature map of the feature maps to be sampled using convolution and pooling, in order to generate attention scores. Then, a lightweight fully connected layer or convolutional layer is used to obtain an attention score matrix that is consistent in width and height with the student model's feature map. The matrix is then normalized so that the size of each element in the matrix falls between 0 and 1. Finally, this normalized attention score matrix replaces the matrix used in the MGD method to sample the intermediate layer feature maps of the student model, completing the optimization of the MGD.

2.2.2 One-to-Many Matching Distillation

In CNN-based teacher models, deeper hidden layers carry more abstract semantic information, which becomes increasingly influential for the task. For object detection, shallow features often encode specific spatial geometric details, whereas deeper features capture high-level semantic information. Shallow knowledge benefits small object detection, while deep features, although rich in semantics, may suffer from lower resolution, making them better suited for large object detection. In distillation, if the student model's feature maps at each layer learn from multiple teacher feature maps through one-to-many matching, it can significantly enhance the student's learning capacity.

This paper proposes a one-to-many matching distillation algorithm, where each student feature map is guided by multiple teacher feature maps at each layer. The overall structure of the distillation model is shown in Fig. 5. First, teacher feature maps are recursively aligned by layer to reduce the parameter count when calculating distillation loss. The distillation loss uses L2 loss to capture the variance between teacher and student feature maps, enabling continuous adaptation of the student model to the teacher during training. Aligning high-level semantic features with low-level texture features avoids the limitation of shallow student features and enhances the detection of small targets.

The model introduces only a Feature Alignment Module (FAM) based on the teacher model. During training, only the parameters of this module are updated. The FAM is not included during inference, ensuring minimal impact on the teacher model's inference speed. Despite the increase in distillation targets (from single-layer to multi-layer feature maps), the additional parameter complexity remains manageable, achieving a balance between distillation performance and model complexity.

Fig. 6 illustrates the structural differences between this distillation method and traditional feature-based distillation methods. Fig. 6a illustrates the structure of logit-based distillation, where distillation occurs only

at the output layer of both networks. Fig. 6b presents the single-layer feature-based distillation structure, which distills only a single layer, usually the last layer feature map of both networks. Fig. 6c illustrates the multi-layer feature-based distillation structure, where each layer's feature map of the student model undergoes distillation based on the corresponding layers of the teacher model, building upon the single-layer distillation of Fig. 6b. Fig. 6d illustrates the proposed one-to-many matching distillation method, where each layer's feature map of the student model is distilled by simultaneously matching multiple layers of the teacher model's feature maps.

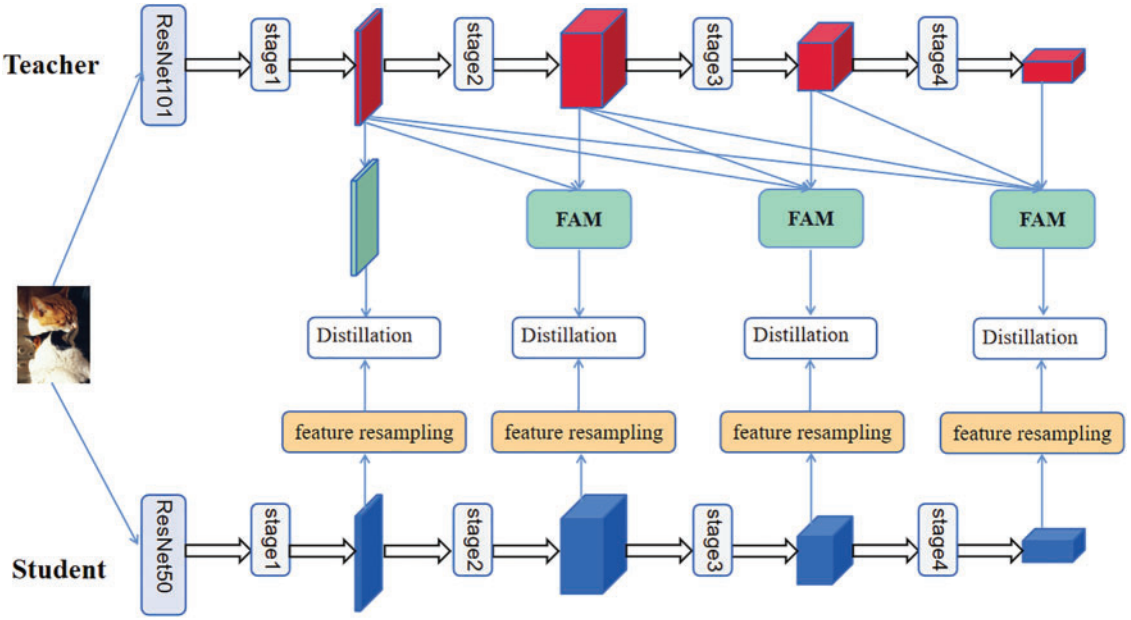


Figure 5: Feature-based distillation method based on feature resampling and one-to-many matching

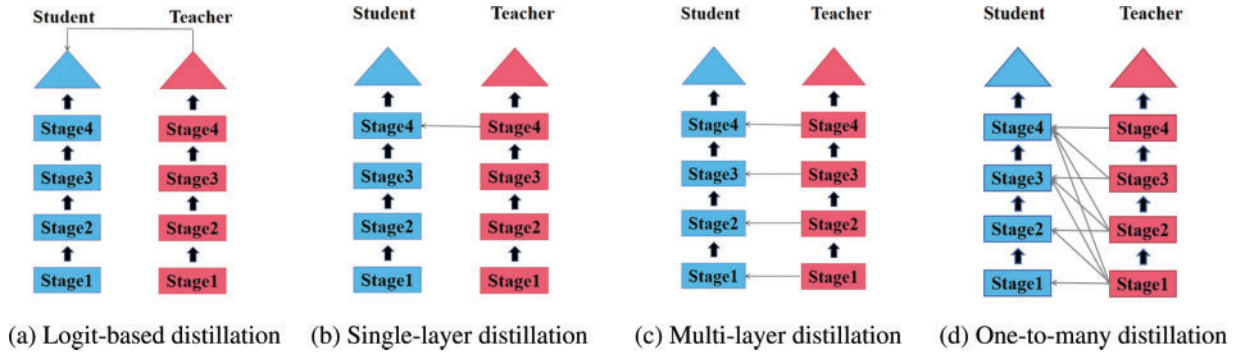


Figure 6: Structure of different feature-based distillation methods

As the number of distillations increases in the one-to-many matching method, the computational cost rises. Moreover, the features of different layers tend to exhibit substantial structural differences. To mitigate this, a FAM is used to align the teacher feature maps, enhancing distillation efficiency while maintaining manageable complexity. The alignment process is illustrated in Fig. 7.

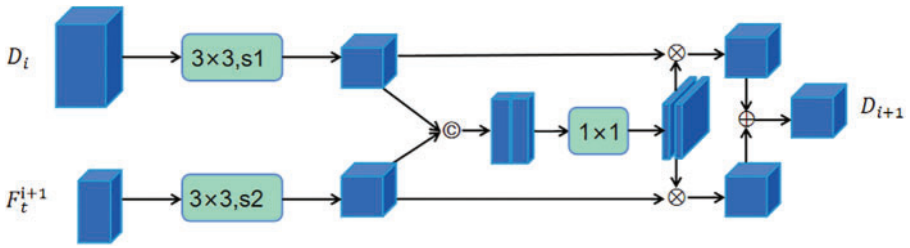


Figure 7: Feature Alignment Module (FAM)

As shown in Fig. 7, the alignment of feature maps $F_t^i, \cdots, F_t^n$ is treated as the alignment of F_t^{i+1} and D_i, where D_i denotes the alignment feature map for stage i. Since the feature maps from different stages have different scales and cannot be directly aligned, we first apply 3×3 convolution kernels with varying strides to match their sizes. Next, the resulting features are concatenated and passed through a 1×1 convolution to reduce the channel number, generating a weight matrix. Finally, the aligned feature map for stage $i + 1$ is obtained by masking and adding the feature map to the weight matrix.

3 Experiment

3.1 Dataset

To verify the effectiveness and efficiency of the proposed method, we conduct various experiments on the COCO (Common Objects in Context) dataset [27] which has emerged as the preeminent and most widely adopted benchmark in the field of object detection. All knowledge distillation methods were extensively evaluated on the COCO2017 dataset. The training set comprises 118K images, while the test set encompasses 4K images, collectively spanning the entirety of the 90 object categories.

3.2 Implement Details

All models were implemented using PyTorch on the NVIDIA GTX 3090 GPU in our lab device. The network training is facilitated by the Stochastic Gradient Descent (SGD) algorithm, with a learning rate of 0.001, a momentum of 0.9, and a batch size of 2. The training process spans a total of 24 epochs, and a linear warm-up strategy is employed to adjust the learning rate adaptively. Specific to the DKD algorithm, the parameters α and β in Eq. (11) are set to 1.0 and 0.25, respectively. For the MGD algorithm, the hyperparameter configurations differ based on the object detection paradigm. When applied to single-stage object detection algorithms, λ is set to 2e−7 and γ is set to 0.65. Conversely, for two-stage object detection algorithms, λ is set to 5e−7, and γ is set to 0.45. For the MGD algorithm that incorporates attention mechanisms, the parameter is the same as that in MGD [24], and there is no need to set the parameter.

Throughout the distillation process, pre-trained models are utilized as the teacher network, and their parameters are frozen, ensuring that the knowledge transfer occurs in a unidirectional manner from the teacher to the student network.

3.3 Evaluation Metrics

The metrics for evaluating target detection algorithms generally consist of two parts. One is performance metrics, including Average Precision (AP), mean Average Precision (mAP), etc. Specifically, mAP50 and mAP75 are the mAP calculations at IoU thresholds of 0.5 and 0.75, respectively. mAPs, mAPm, and mAP1 measure the detection performance for small, medium, and large objects, respectively. Performance metrics are used to evaluate the accuracy of the model detection results. The second is the model complexity metrics,

such as Floating-Point Operations (FLOPs), number of parameters and Frames Per Second (FPS), etc. which are used to measure the computation and storage consumption of the network model. Often the more complex the model, the higher its performance.

3.4 Experimental Result

3.4.1 Decoupled Logit Distillation Ablation Study

Since the logit-based distillation algorithm cannot transfer the knowledge of the target localization task to the student network, its performance enhancement for the object detection task is limited when used alone. However, it exhibits more noticeable enhancements in the classification task. Therefore, in this paper, we conducted ablation experiments on the CIFAR-100 dataset for image classification to analyze the impact of the two losses after decoupling on the distillation performance. The teacher network based on ResNet32 × 4, while the student network was based on ResNet8 × 4. The experimental results are shown in Table 1.

Table 1: Decoupled logit distillation ablation experiment

β	$1-p_t^T$	1.0	2.0	4.0	8.0	10.0
Top-1	73.41	74.52	75.42	75.72	**76.22**	76.01
α	0.0	0.2	0.5	1.0	2.0	4.0
Top-1	75.11	75.42	76.02	**76.21**	76.01	75.42

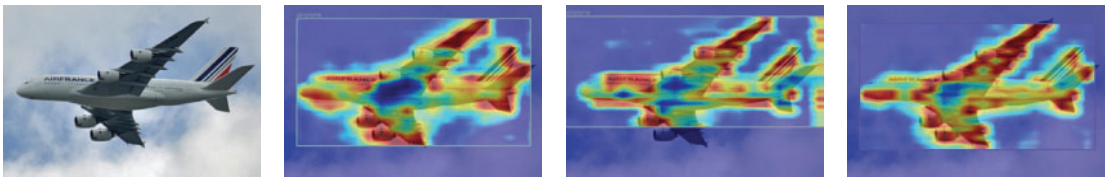
Through the above experiments, it can be seen that when the weight of NCKD is set to 8, the distillation algorithm exhibits the best performance index. This finding demonstrates that NCKD significantly contributes to the logit distillation performance, which aligns with the theory that negative samples of logit encapsulate "dark knowledge". Additionally, TCKD remains essential, proving effective as long as its weight is maintained at 1. Following this ablation experiment, the weights of NCKD and TCKD are established as 8 and 1, respectively, for future use in object detection.

3.4.2 Resampling Effect Visualization Ablation Study

To showcase the effectiveness of foreground pixel emphasis after resampling, this paper visualizes the feature maps of both the student and teacher networks within the Faster R-CNN model. Specifically, the last layer of the feature extraction network is chosen as the visualization target. The teacher model's backbone network is ResNet101, while the student model's backbone network is ResNet50. The results are presented in Fig. 8.

Through the teacher network (Fig. 8b), key components including the fuselage, wings, and engines are accurately segmented and clearly presented in heatmaps, with warmer colors such as red and yellow highlighting these areas. This reflects the rich object detection knowledge acquired by the teacher model through extensive training. In contrast, the baseline student network (Fig. 8c) manages to outline the aircraft's silhouette but falls short in detail handling, resulting in a lack of clarity in the separation between foreground and background objects, which is represented by cooler colors like blue and green in the heatmaps. However, the student network with feature resampling (Fig. 8d) delineates the aircraft's main structure more clearly and accurately captures intricate details such as the wings and tail, with an overall feature distribution that closely resembles that of the teacher network, as indicated by the warmer color regions in the heatmaps. This clearly demonstrates that the feature resampling distillation method is more effective than

traditional approaches in emulating the teacher network's features and significantly enhances the student network's fitting and knowledge acquisition capabilities, thereby improving its feature representation ability. Subsequent experiments further confirm that the feature resampling distillation method indeed enhances the overall distillation performance, which is visually evidenced by the heatmaps showing a distribution of warm color regions that is more akin to the teacher network.



(a) Original Image (b) Teachers Network (c) Benchmarking Student Networks (d) Resampling Student Networks

Figure 8: Comparison of Faster R-CNN model feature maps

3.4.3 Distillation's Student Network Ablation Study

In this section, we evaluate the three distillation strategies proposed in this paper, providing a detailed analysis and comparison. Through the ablation experiment, we investigate the contribution of each strategy to the final detection accuracy, finding that the feature-based distillation method exhibits more significant advantages in the object detection task. Simultaneously, the effects of the two enhancements, feature resampling and one-to-many matching, are quantitatively evaluated, providing valuable references for future related research. Moreover, the experiment results further emphasize the importance of addressing the foreground-background imbalance issue in improving object detection performance. By employing feature resampling, we achieve outstanding detection accuracy.

In this experiment, Faster R-CNN is still selected as the primary subject for the ablation experiment. The teacher model's backbone network adopts ResNet-101, while the student model's backbone network employs the lighter ResNet-50 architecture. The ablation experiments primarily focus on the feature-based distillation method introduced in Section 2.2 and the logit-based distillation method presented in Section 2.1. Two enhancements are proposed for the feature-based distillation method: feature resampling and one-to-many matching. Additionally, the logit-based distillation method introduces an improvement by decoupling the logit output. The results of the ablation experiments are shown in Table 2.

Table 2: Ablation results of student model after distillation

Feature resampling	One-to-many	Decoupled logit	mAP	mAP50	mAP75	mAPs	mAPm	mAP1
×	×	×	38.4	58.1	40.4	21.2	41.0	48.1
√	×	×	40.5	62.2	43.7	24.4	43.6	52.2
×	√	×	40.1	60.8	43.2	24.1	43.4	52.0
×	×	√	39.2	61.0	42.0	23.1	42.0	51.9
√	√	×	40.7	62.8	43.8	24.8	43.9	52.7
√	×	√	40.5	62.6	42.9	24.6	43.4	52.4
×	√	√	40.3	62.1	43.5	24.2	43.5	52.1
√	√	√	**41.5**	**63.7**	**44.2**	**25.5**	**44.2**	**53.2**

Firstly, it can be observed from the top four rows of the table that each distillation module contributes to enhancing the detection accuracy of the student model, thereby demonstrating the effectiveness of the three distillation schemes proposed in this paper. By comparing the degree of influence of these schemes on accuracy, it becomes evident that the feature-based distillation strategy exerts a more significant effect than the logit-based distillation. This observation is also analyzed earlier, primarily attributed to the characteristics of the object detection algorithm task. The backbone network contains more spatial and semantic knowledge that is beneficial for object detection bounding box regression and target classification. In contrast, the logit output layer of the detection head has a smaller effect on the object detection task because it lacks spatial positioning information, thereby only moderately affecting the accuracy of target classification. Furthermore, among the two enhanced feature-based distillation schemes, it is evident that the performance improvement of the feature resampling method surpasses that of the one-to-many matching method. This result also shows that the imbalance of foreground and background information is the key factor restricting the performance of object detection, and one-to-many matching is more focused on enhancing the student network's learning ability.

3.4.4 Comparison Results

In this part, we selected three detection frameworks, Faster R-CNN, RetinaNet, and RepPoints to observe the metrics of the student networks guided by different knowledge distillation modules, in order to validate the performance of the proposed knowledge distillation modules.

(1) Quantative results: Table 3 presents a comparative evaluation of our proposed distillation method against a variety of distillation algorithms from the three frameworks mentioned above. The research ideas underpinning these algorithms are largely aligned with the core principles of our proposed method, rendering them suitable candidates for comparison. It is noteworthy that for all object detection algorithms, the teacher and student networks employ ResNet101 and ResNet50 as their respective backbone architectures.

Table 3: Comparative experimental results after distillation of the student model

Detection model	Teacher-backbone	Student-backbone	Method	mAP	mAPs	mAPm	mAPl
Faster RCNN	ResNet101	ResNet50	FitNets	38.6	22.0	43.6	48.0
			FGD	**41.6**	23.8	46.4	55.5
			MGD	41.3	23.7	46.4	**56.1**
			FKD	41.5	23.5	45.0	55.3
			CSKD	41.0	23.8	45.0	53.0
			Ours	41.5	**24.5**	**47.2**	55.2
RetinaNet	ResNet101	ResNet50	FitNets	37.8	19.8	39.0	50.7
			FGD	37.4	20.6	40.7	49.7
			MGD	41.0	23.4	**45.3**	**55.7**
			FKD	39.6	22.7	43.3	52.5
			CSKD	39.9	23.0	44.5	53.2
			CrossKD	39.5	22.1	43.3	52.7
			Ours	**41.2**	**23.9**	45.2	55.5

(Continued)

Table 3 (continued)

Detection model	Teacher-backbone	Student-backbone	Method	mAP	mAPs	mAPm	mAPl
RepPoints	ResNet101	ResNet50	FitNets	39.6	22.5	42.3	51.9
			FGD	42.0	24.0	45.7	55.6
			MGD	**42.3**	**24.4**	46.2	**55.9**
			FKD	40.6	23.4	46.2	**55.9**
			Ours	**42.3**	24.3	**47.2**	54.5

As shown in the table, when employing different types of knowledge distillation models on the Faster R-CNN algorithm, our model achieved a mAP score of 41.5, outperforming FitNets, CSKD and MGD by 2.9, 0.5 and 0.2, respectively, and on par with FKD. Regarding mAPs and mAPl, our model surpassed other algorithms, with the highest mAPl score exceeding FitNets by 3.6. Subsequently, in the RetinaNet network, our proposed knowledge distillation model exhibited an overall superior performance. Notably, for the mAP metric, our scores consistently surpassed other knowledge distillation models. Specifically, our mAP score of 41.2% surpasses all other distillation models, including MGD (41.0%) and the lowest-performing FGD (37.4%) by 0.2% and 3.8%, respectively. Our model also attained commendable results on the mAPs, mAPm, and mAPl metrics. Finally, in the RepPoints network, our model demonstrated formidable competitiveness, as observed by its optimal performance based on the mAPm metric, surpassing FitNets by nearly 5 points. This substantive improvement underscores the effectiveness of our distillation algorithm in addressing medium-scale object detection challenges.

The combination of intermediate layer feature distillation, logit output layer distillation, and techniques tailored to address foreground-background imbalance and enhance knowledge transfer has proven effective in boosting object detection accuracy across various object scales and network architectures.

(2) Visual results: Fig. 9 demonstrates the visual comparison results with detection results marked by red boxes, false and missed detections denoted by blue boxes.

In scenarios involving overlapping or occluded objects, such as the elephants shown in the first column, our distillation method effectively reduces the false positive rate and prevents the generation of unnecessary bounding boxes. This is in stark contrast to the first two methods, which generated excess bounding boxes marked in blue during the detection process. When dealing with images of medium-sized objects, such as the horse in the second column, our distillation method effectively identifies and locates the objects without any instances of false positives or false negatives. When detecting smaller objects like horses and birds (as depicted in the last two columns of the document), despite some missed detections marked by blue squares, our model has shown a generally lower rate of false positives and more precise bounding box predictions overall. This confirms the effectiveness of our approach in transferring the teacher network's capability to identify and locate small objects to the student network.

Overall, the visualizations corroborate that the proposed distillation algorithm outperforms traditional distillation methods like FitNets and achieves performance comparable to or even surpassing recent state-of-the-art approaches like MGD. Notably, the distilled models exhibit superior performance on challenging samples, underscoring the efficacy of our distillation strategy in transferring valuable knowledge from the teacher to the student model, thereby enhancing its detection capabilities.

Figure 9: Visualization of distillation effect

To further validate the effectiveness of our proposed distillation algorithm, we visually analyze the curve of accuracy changing with epoch during training across three prominent object detection architectures (see Fig. 10): Faster R-CNN, RetinaNet, and RepPoints. The graphs depict the change in accuracy during the training process, contrasting the performance of the base student detector (base-detector) against the distilled student detector (distilled-detector) incorporated with our distillation methodology. Across all three detection frameworks, a consistent trend is observed: the detector that underwent knowledge distillation (orange line) performed better than the base detector (blue line) in the majority of training epochs. Both lines show an increasing trend in performance metrics as the number of training epochs increased. Particularly

after the 15th epoch, there is a notable surge in the orange line, indicating that the effects of knowledge distillation became particularly pronounced at this stage. This suggests that our method can effectively accelerate the model's learning process and enhance its ultimate performance.

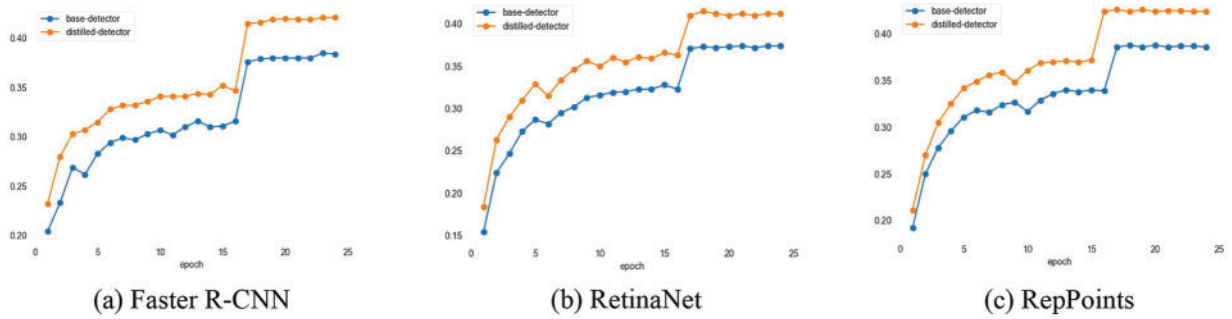


Figure 10: Distillation accuracy curve

(3) Efficiency results: Table 4 presents the experimental results and model complexity metrics for the used three object detection architectures: Faster R-CNN, RetinaNet, and RepPoints, after incorporating our proposed distillation algorithm. The reported metrics include mean Average Precision (mAP), FLOPs (floating-point operations), number of parameters, and FPS (frames per second).

Table 4: Comparison of object detection algorithms in performance and speed

Teacher	Student	mAP	FLOPs	Parameters	FPS
FasterR-CNN-ResNet101	FasterR-CNN-Res50 (baseline)	38.4	211.69G	41.53M	8.39
(mAP baseline:42.04)	**Ours**	**41.5**	**225.72G**	**43.25M**	**8.59**
RetinaNet-ResNet101	RetinaNet-Res50 (baseline)	37.4	244.89G	37.81M	8.78
(mAP baseline:41.0)	**Ours**	**41.2**	**251.22G**	**40.14M**	**8.98**
RepPoints-ResNet101	RepPoints-Res50 (baseline)	38.6	204.25G	33.94M	9.77
(mAP baseline:44.2)	**Ours**	**42.3**	**208.22G**	**35.23M**	**9.87**

In all three detection paradigms, using ResNet50 as the student model and ResNet101 as the teacher model consistently improves performance. Whether it's applied to two-stage (Faster R-CNN), one-stage (RetinaNet), or anchor-free (RepPoints) detection methods, the mAP accuracy of the distilled student model has significantly increased by 2–3 percentage points, approaching the performance of the teacher model. Comparisons of FLOPs and parameter counts show that the complexity introduced by the distillation module is extremely low, with fewer than 3 M additional parameters. This highlights the effectiveness of knowledge distillation as a model compression technique, enhancing performance without causing a significant computational overhead. Analysis of FPS indicates that the inference speed of the distilled model is almost identical to that of the baseline model, with a negligible difference of less than two frames per second. This suggests that our distillation method has a minimal impact on detection latency, making it suitable for low-resource devices and real-time applications.

In summary, the experimental results validate our distillation method's ability to enhance compact student model performance across various detection architectures, offering significant accuracy improvements

while keeping model complexity and inference speed changes minimal. This makes our approach a promising solution for efficient and accurate object detection on resource-constrained platforms.

4 Conclusion

This paper presents an innovative approach that combines the state-of-the-art logit distillation method DKD with the middle layer feature distillation technique MGD, specifically for the field of object detection. By integrating these two methods, we have significantly enhanced the performance of the student model in tasks related to object classification and bounding box regression. To further refine the distillation process, we introduced a SAM based on MGD, allowing the model to focus more on foreground pixels, thereby retaining more critical information. Additionally, we proposed a one-to-many matching distillation strategy based on FAM, which further improves the precision of knowledge transfer. While the combined distillation method and the spatial attention-enhanced MGD improve student model performance, there are areas for optimization. For example, combining both methods may increase memory usage and computational time, requiring a balance between resources and performance in future work. Furthermore, exploring data-free distillation techniques that rely only on the teacher network's parameters, without needing raw training data, presents an interesting direction for future research.

Acknowledgement: The authors would like to express their sincere gratitude to all those who contributed to this research. Their support and efforts were key factors in the success of this study.

Funding Statement: This research was funded by National Natural Science Foundation of China (61603245).

Author Contributions: Min Yao and Guofeng Liu drafted the manuscript and made substantial contributions to the conception, design, and analysis of the experiment. Yaozu Zhang and Guangjie Hu provided guidance for data collection and experiment, also revising the manuscript critically. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data provided in this study are available upon request from the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 2980–8. doi:10.1109/ICCV.2017.322.
2. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
3. Hu H, Gu J, Zhang Z, Dai J, Wei Y. Relation networks for object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 3588–97. doi:10.1109/CVPR.2018.00378.
4. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318–27. doi:10.1109/TPAMI.2018.2858826.
5. Liu W, Anguelov D, Erhan D, Szegedy C. SSD: single shot multibox detector. In: Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands: Springer International Publishing; 2016.
6. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767v1. 2018.
7. Wang H, Jia T, Ma B, Wang Q, Zuo W. Fully cascade consistency learning for one-stage object detection. IEEE Trans Circuits Syst Video Technol. 2023;33(10):5986–98. doi:10.1109/TCSVT.2023.3263557.

8. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 6154–62. doi:10.1109/CVPR.2018.00644.
9. Wang SY, Qu Z, Li CJ. A dense-aware cross-splitNet for object detection and recognition. IEEE Trans Circuits Syst Video Technol. 2023;33(5):2290–301. doi:10.1109/TCSVT.2022.3221658.
10. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea: IEEE; 2019. p. 9626–35. doi:10.1109/iccv.2019.00972.
11. Qiu H, Li H, Wu Q, Cui J, Song Z, Wang L, et al. CrossDet++: growing crossline representation for object detection. IEEE Trans Circuits Syst Video Technol. 2023;33(3):1093–108. doi:10.1109/TCSVT.2022.3211734.
12. Gao F, Cai Y, Deng F, Yu C, Chen J. Feature alignment in anchor-free object detection. IEEE Trans Circuits Syst Video Technol. 2023;33(8):3799–810. doi:10.1109/TCSVT.2023.3241993.
13. Wang Y, Du H, Cheng Z, Gao C, Wei L, Fang B, et al. KRRNet: keypoint relational regression network for bottom-up anchor-free object detection. IEEE Trans Circuits Syst Video Technol. 2024;34(4):2249–60. doi:10.1109/TCSVT.2023.3305289.
14. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531v1. 2015.
15. Zhou H, Song L, Chen J, Zhou Y, Wang G, Yuan J, et al. Rethinking soft labels for knowledge distillation: a bias-variance tradeoff perspective. arxiv:2102.00650v1. 2021.
16. Adriana R, Nicolas B, Ebrahimi KS, Chassang A, Gatta C, Bengio Y. FitNets: hints for thin deep nets. arXiv:1412.6550. 2014.
17. Heo B, Kim J, Yun S, Park H, Kwak N, Choi JY. A comprehensive overhaul of feature distillation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea: IEEE; 2019. p. 1921–30. doi:10.1109/iccv.2019.00201.
18. Chen P, Liu S, Zhao H, Jia J. Distilling knowledge via knowledge review. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE; 2021. p. 5006–15. doi:10.1109/cvpr46437.2021.00497.
19. Chen G, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation. In: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 3–9; Long Beach, CA, USA: MIT Press; 2017. p. 742–51.
20. 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); 2022 Jun 18–24; New Orleans, LA, USA: IEEE; 2022. p. 4633–42. doi:10.1109/CVPR52688.2022.00460.
21. Zhang L, Ma K. Improve object detection with feature-based knowledge distillation: towards accurate and efficient detectors. In: International Conference on Learning Representations, 2021 (ICLR 2021); 2021 May 4–8; Vienna, Austria; 2021.
22. Wang J, Chen Y, Zheng Z, Li X, Cheng MM, Hou Q. CrossKD: cross-head knowledge distillation for object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE; 2024. p. 16520–30. doi:10.1109/CVPR52733.2024.01563.
23. Park S, Kang D, Paik J. Cosine similarity-guided knowledge distillation for robust object detectors. Sci Rep. 2024;14(1):18888. doi:10.1038/s41598-024-69813-6.
24. Yang Z, Li Z, Shao M, Shi D, Yuan Z, Yuan C. Masked generative distillation. In: European Conference on Computer Vision (ECCV); 2022 Oct 23–27; Tel Aviv, Israel. Cham, Springer; 2022. p. 53–69. doi:10.48550/arXiv.2205.01529.
25. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany: Springer; 2018. p. 3–19.
26. Zhao B, Cui Q, Song R, Qiu Y, Liang J. Decoupled knowledge distillation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE; 2022. p. 11943–52. doi:10.1109/CVPR52688.2022.01165.
27. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. In: Computer Vision-ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland: Springer International Publishing.