MKD: Multi-Knowledge Distillation for Real-Time Object Detection on Edge Devices

Yanjun Xu jesse@tongji.edu.cn
Chunqi Tian tianchunqi@tongji.edu.cn
Xuhui Xia xiaxuhui@tongji.edu.cn
Yaoru Sun yaoru@tongji.edu.cn
Tongji University, No. 4800 Caoan Road, Shanghai

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Real-time object detection on resource-constrained edge devices presents a significant challenge in balancing performance and efficiency. This paper introduces a novel knowledge distillation framework designed to enhance the capabilities of lightweight student models for object detection tasks. Our approach, Multi-Knowledge Distillation (MKD), integrates three key components: multi-scale distillation, frequency domain mask distillation, and feature alignment distillation. Multi-scale distillation enables the student to learn feature representations at various levels of granularity. Frequency domain mask distillation improves the student's ability to focus on relevant regions. Feature alignment distillation facilitates the transfer of channel-wise knowledge from teacher to student. We combine these techniques with a traditional detection loss to form a comprehensive loss function, balanced by a hyperparameter α. Experimental results across various scenarios demonstrate that MKD significantly improves detection accuracy while reducing computational and storage costs. Our approach clearly presents significant performance gains and faster inference speeds.

Keywords: object detection,knowledge distillation,feature alignment.

1. Introduction

Object detection is a fundamental task developed in the computer vision community over the past few years. Especially small object detection has gained significant attention due to its wide range of applications Wei et al. (2022); Qin et al. (2022); Wei et al. (2023) in fields such as surveillance, autonomous driving, and medical imaging. However, detecting small objects accurately remains challenging, mainly due to their limited spatial information and low signal-to-noise ratio. Current works strive to refine feature fusion modules Ghiasi et al. (2019), devise novel training schemes Singh and Davis (2018) to explicitly train on small objects, design new neural architectures Yang et al. (2022a) to extract small objects' features better, and leverage increased input resolution to enhance representation quality Bai et al. (2018). However, these approaches struggle to balance detection performance on small objects with computational costs at the inference stage.

Knowledge distillation Hinton et al. (2015) is a technique used to transfer knowledge from a large, complex model (known as the teacher model) to a smaller, more efficient model (known as the student model). The goal is to enable the student model to achieve performance similar to the teacher model while reducing computational requirementsDong

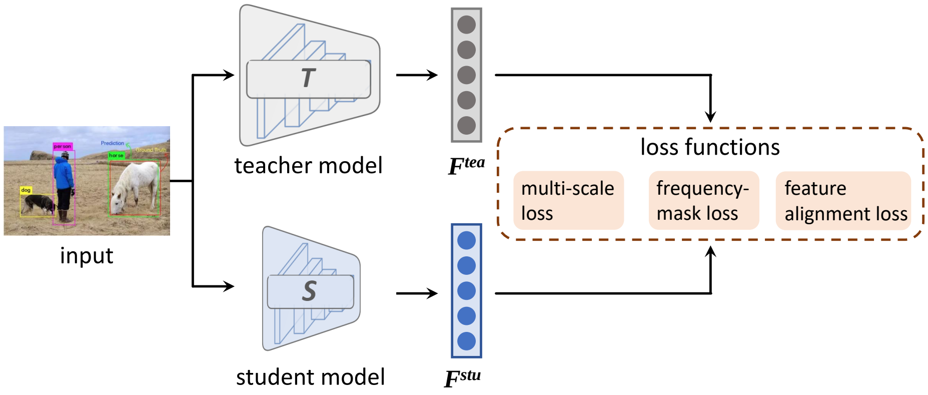

Figure 1: A schematic overview of our method. Our method incorporates multiscale, frequency domain mask, and feature alignment distillation. Multiscale distillation enables the student model to capture information at different scales. Frequency domain mask distillation improves the accuracy of generated masks. Feature alignment distillation aligns channel-wise features between the teacher and student models, allowing the student model to focus on informative regions and capture object details effectively.

et al. (2023b). Knowledge distillation has been applied in object detection to improve the performance of object detection models. The idea is to leverage the knowledge learned by a teacher model, which is typically a more accurate and computationally expensive model, and transfer that knowledge to a smaller and faster student modelLiu et al. (2022). This allows the student model to benefit from the teacher model's expertise and achieve competitive performance in object detection tasksFeng et al. (2021). However, there are some issues with knowledge distillation in object detection, such as the misalignment between the spatial dimensions of the teacher and student features and the challenge of effectively transferring channel-wise knowledgeDong et al. (2024); Chen et al. (2022). These issues can limit the knowledge transfer process and hinder the performance improvement of the student model.

In this paper, we propose a knowledge distillation method to improve the performance of real-time object detection via multiple distillation techniques, which comprise the multiscale distillation loss, the frequency domain mask distillation loss, and the feature alignment distillation loss. Our multiscale distillation involves processing the input data at different scales or resolutions. This allows the student model to capture information at different levels of detail and granularity. By aligning the feature maps of the teacher and student models at multiple scales, the student model can learn to effectively utilize multiscale features, which is particularly important in tasks such as object detection, where objects can vary in size and scale. Another technique employed is frequency domain mask distillation, which incorporates mask information into the distillation process. The teacher model may generate more accurate or precise masks than the student model, and by distilling this mask information, the student model can learn to generate masks that closely resemble those of the teacher model. This helps improve the student model's ability to identify and attend to relevant regions or objects in the input data, leading to better localization and segmentation performance. Additionally, feature alignment distillation is utilized to enhance the matching of channel-wise features between the teacher and student models. Alignment mechanisms are introduced to guide knowledge transfer, focusing on informative regions. By optimizing the disparity of the channel correlation matrix, the student model can learn to focus on the most discriminative channel-wise features, improving its ability to capture object details and channel-specific knowledge. The total loss function is used during training to optimize the student model. This loss function combines the detection loss, which measures the discrepancy between the predicted and ground truth bounding boxes and class probabilities, with the distillation loss. The distillation loss comprises the multiscale distillation loss, the frequency domain mask distillation loss, and the feature alignment distillation loss. The hyperparameter alpha is used to balance the relative importance of the two components during optimization. This comprehensive approach enhances the student model's ability to predict image recognition and object detection tasks accurately.

In the experiment, we validate the effectiveness of our proposed method through extensive experiments. Our proposed method consistently outperforms various state-of-the-art object detection models across different model sizes, such as YOLOv5, YOLOX, YOLOv6, and PPYOLOE. It achieves higher AP and AP50 scores, indicating improved detection accuracy. Comparing with other detection distillation methods like FGD and MGD, our method consistently achieves higher AP and AP50 improvements, showcasing its effectiveness in enhancing detection performance. These results demonstrate the efficacy of our approach in addressing the challenges of object detection, potentially leading to more reliable and accurate detection systems in various applications.

To sum up, our contributions are the following:

* We introduce a novel frequency domain mask distillation to align the spatial dimensions of the teacher and student features using masked image reconstruction. By randomly generating a mask matrix and reconstructing the student feature, we promote performance by strengthening pixel interdependencies.

* We propose a feature alignment distillation mechanism. This component focuses on matching channel-wise features by optimizing the disparity of the channel correlation matrix. This alignment-based distillation helps the student model focus on discriminative channel-wise features, leading to improved object details and channel-specific knowledge capture.

* We verify the effectiveness of our method via various experiments. The student model achieves significant improvements in object detection.

2. Related work

2.1. Object Detection

The current mainstream object detection algorithms are divided into two-stage and one-stage detectors. Two-stage methods Lin et al. (2017a) represented by Faster R-CNN **?** maintain the highest accuracy in the detection field. These methods utilize region proposal networks (RPNs) and refine classification and location procedures to obtain better performance. However, high demands for lower latency bring one-stage detectors Liu et al. (2016) under the spotlight, which directly achieve classification and location of targets through the feature map. Recently, another criterion divides detection algorithms into anchor-based and anchor-free methods. Anchor-based detectors such as Redmon and Farhadi (2017) solve object detection tasks with the help of anchor boxes, which can be viewed as pre-defined sliding windows or proposals. Nevertheless, all anchor-based methods need to be meticulously designed, and many anchor boxes need to be calculated, which takes much computation. To avoid tunning hyper-parameters and calculations related to anchor boxes, anchor-free methods Tian et al. (2019a) predict several key points of target, such as center and distance to boundaries, reach a better performance with less cost. Object detection is one of the most challenging tasks in computer vision and involves locating and classifying objects on the images. For object detection tasks, the imbalance of foreground and background is a crucial problem. R-CNN-like detectors use a two-stage cascade and sampling heuristics to narrow the number of background samples, or online hard example mining(OHEM) Shrivastava et al. (2016) to ensure the balance between foreground and background. One-stage approaches, such as RetinaNet Lin et al. (2017b) proposed focal loss to solve this problem. Anchor-free detectors Tian et al. (2019b); Kong et al. (2020) attempt to discard the design of anchor boxes to avoid time-consuming box operations and cumbersome hyper-parameter tuning. Dynamic label assignment methods Nguyen et al. (2022) are proposed better to define the positive and negative samples for model learning. Recently, attributing to the strong ability of the transformer block to encode expressive features, DETR family Zhu et al. (2021); Li et al. (2022a) has become a new trend in the object detection community.

2.2. Knowledge Distillation for Object Detection

Knowledge distillation is one of the model compression methods Hu et al. (2021); Lu et al. (2024); Zhu et al. (2024). Recently, some works have applied knowledge distillation to object detection tasks. Previous works Li et al. (2024); Li (2022); Shao et al. (2023); Dong et al. (2023a) mainly focus on designing proper distillation areas to cope with the extreme unbalance between positive and negative instances in object detection. Chen et al. (2017) first deals with this problem by underweighting the background distillation loss in the classification head. Li et al. (2017) applies the L2 distillation loss to the negative and positive ROI features sampled by RPN in a certain proportion. Wang et al. (2019) proposes a fine-grained feature imitation, distilling the near objects regions. Guo et al. (2021) proposes to decouple the foreground and background features for distillation, adjusting each loss weight and temperature separately. Dai et al. (2021) further proposes to distill the discriminative patches between students and teachers. However, all of those previous works take one fixed teacher for experiments, and they need to explore the relationship between teacher performance and student performance for object detection tasks. FGD Yang et al. (2022b) aligns the attention place for the teacher-student pair, while PKD Cao et al. (2022) maximizes the Pearson Correlation Coefficient between the feature representations. As the earliest distillation strategy proposed in Hinton et al. (2015), prediction mimicking plays a vital role in classification distillation. Recently, some improved prediction-mimicking methods have been proposed to adapt to object detection. For example, Rank Mimicking Li et al. (2022b) regards the score rank of the teacher as a kind of knowledge and aims to force the student to rank instances as the teacher. LD Zheng et al. (2022) proposes to distill the localization distribution of bounding box to transfer localization knowledge. Our method differs from other methods in several key aspects. We employ multiple distillation techniques, including multiscale, frequency domain mask, and feature alignment distillation. This combination allows for more comprehensive knowledge transfer and enhances the student model's performance in object detection tasks.

3. Methodology

In this section, we present our proposed approach, which includes three key components: multiscale distillation, frequency domain mask distillation, and feature alignment distillation. The overall framework is schematically shown in Figure 1.

3.1. Multiscale Distillation

Multiscale knowledge distillation (KD) involves processing the input data at different scales or resolutions and capturing information at different levels of granularity. The motivation behind multiscale KD stems from the fact that the teacher and student models may have differences in their ability to extract and utilize multiscale features because of architectural discrepancies or variations in training data. These discrepancies can arise from factors such as the depth of the network, the receptive field sizes, or the training data distributions. Specifically, the feature maps from the teacher and student models are processed using pooling functions of various scales, denoted as $pool_k$. The multiscale distillation loss, $L_{Multiscale}$, is then calculated as the mean L2 distance between the pooled teacher and student feature maps across all scales:

$$L_{Multiscale} = \frac{1}{HW} \cdot \sum_{k=1}^{m} l_2(pool_k(F^{tea}), pool_k(F^{stu})), \tag{1}$$

where l_2 is the L2 distance, F^{tea} and F^{stu} represent the feature maps of the teacher and student models, respectively, m is the number of scales considered, and H and W are the height and width of the feature maps. By minimizing the multiscale distillation loss, $L_{Multiscale}$, during training, the student model is encouraged to learn feature representations that align with the teacher's multiscale features, enabling it to capture and utilize the information at different levels of granularity effectively.

3.2. Frequency Domain Mask Distillation

In knowledge distillation, masks can provide spatial information or highlight specific regions or objects of interest in the input data. The teacher model might generate more accurate or precise masks than the student model due to architectural disparities or variations in training data. By incorporating mask information in the distillation process, the student model can learn to generate masks that closely resemble the teacher model. This enables the student model to improve its ability to identify and attend to relevant regions or objects in the input data, resulting in improved localization and segmentation performance. Specifically, our frequency domain mask distillation loss, denoted as L_{FMask}, can be formulated as:

$$\hat{F}^s = \mathcal{F}^{-1}(M \odot \mathcal{F}(F^s)), \tag{2}$$

where F^s represents the feature maps of the student model, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and inverse Fourier transform operations, respectively, and $\odot$ represents element-wise multiplication. The mask $M \in \mathbb{R}^{H \times W \times 1}$ is a binary matrix generated randomly with a mask ratio $\zeta \in [0, 1)$, defined as:

$$M_{i,j} = \begin{cases} 0, & R_{i,j} < \zeta \\ 1, & R_{i,j} \geq \zeta \end{cases} \tag{3}$$

Here, $R_{i,j}$ is a random value sampled from a uniform distribution $\mathcal{U}(0, 1)$ for each spatial location (i, j). The mask M is applied to the Fourier domain representation of the student's feature maps, $\mathcal{F}(F^s)$, by element-wise multiplication, effectively masking out certain frequency components. The masked features, $\hat{F}^s$, are then obtained by applying the inverse Fourier transform, $\mathcal{F}^{-1}$. The frequency domain mask distillation loss, L_{FMask}, is computed as the mean squared error between the masked student features, $\hat{F}^s$, and the teacher features, F^{tea}, across all spatial locations and channels:

$$L_{FDMD} = \frac{1}{C^t \times H \times W} \sum_{k=1}^{C^t} \sum_{i=1}^{H} \sum_{j=1}^{W} (\hat{F}^{stu}_{i,j,k} - F^{tea}_{i,j,k})^2, \tag{4}$$

where C^t is the number of channels in the teacher's feature maps, and H and W are the height and width of the feature maps, respectively. By minimizing the frequency domain mask distillation loss, L_{FMask}, during the training process, the student model is encouraged to learn feature representations that are consistent with the teacher model's features in the frequency domain while also accounting for spatial information provided by the randomly generated masks. This approach can potentially enhance the student model's ability to capture and attend to relevant regions or objects in the input data, leading to improved performance in tasks such as localization and segmentation.

3.3. Feature Alignment Distillation

Recent distillation works also pay attention to the knowledge contained in each channel. To further enhance the distillation process, alignment mechanisms focus on informative regions and guide knowledge transfer. Alignment mechanisms improve the matching of channel-wise features by optimizing the disparity of the channel correlation matrix. In this method, the KL divergence between the probability map is minimized and calculated by normalizing each channel's feature map. Feature Alignment Distillation uses the sum of teacher and student attention to focus the student on changeable areas. Feature Alignment Distillation proposes focal distillation, which forces the student to learn the teacher's crucial parts, and global distillation compensates for missing global information. This channel-wise knowledge can be derived from the following formula:

$$\mathcal{L}_{FAD} = \frac{\mathcal{T}^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{W \cdot H} \phi\left(G_{c,i}^{T}\right) \cdot \log\left[\frac{\phi\left(G_{c,i}^{T}\right)}{\phi\left(G_{c,i}^{S}\right)}\right] \tag{5}$$

where ϕ represents the softmax function and $\mathcal{T}$ denotes the temperature coefficient. It is important to note that channel distillation and L_2 distance are combined when the normalization is achieved in the channel dimension.

The alignment mechanisms guide knowledge transfer from the teacher to the student during the distillation process. The alignment weights are applied element-wise to the corresponding feature maps, emphasizing informative regions while suppressing less relevant information. This alignment-based distillation helps the student network focus on the most discriminative channel-wise features and improves its ability to capture object details and channel-specific knowledge.

3.4. Total Optimization and Inference

Our approach encourages the student network to match the behavior of the teacher network, either in terms of output probabilities or intermediate representations. This is achieved through the total loss function, which is formulated as:

$$L_{all} = L_{det} + \alpha \cdot (L_{fea} + +L_{Multiscale} + +L_{FDMD} + L_{FAD}) \tag{6}$$

In this equation, L_{det} represents the detection loss, which is a combination of the localization loss and the classification loss in object detection tasks. The localization loss, typically implemented using the smooth L1 loss, quantifies the difference between the predicted bounding box for an object and the ground truth bounding box for that object. On the other hand, the classification loss, often calculated using the cross-entropy loss, measures the discrepancy between the predicted class probabilities for an object and the ground truth class probabilities for that object. The second term in the total loss function is L_{fea}, which is the distillation loss. This loss component captures the difference between the feature maps extracted from the teacher and student models, encouraging the student model to learn representations similar to those of the teacher model. The hyperparameter α is used to balance the relative importance of the detection loss and the distillation loss during optimization.

During the inference stage, the student model generates object classification and location information using the features extracted from the input image. These features learned through the distillation process enable the student model to make accurate predictions for object detection tasks, leveraging the knowledge transferred from the teacher model.

4. Experiments

Datasets We conduct our object detection experiments on the COCO2017 dataset Lin et al. (2014), which comprises approximately 1.5 million object instances across 80 different categories, with 41.43% of these instances being small objects. Our evaluation utilizes 120K training images and 5K validation images. The performance of the models is assessed using Average Precision (AP) across various settings.

Table 1: Comparison of various object detection methods on the number of parameters, FLOPS, latency, and accuracy on COCO `val2017` set.

Model	Input shape	Params(M) ↓	FLOPs(G) ↓	Latency(ms) ↓	AP(%) ↑	AP50(%) ↑
YOLOv5-n glenn jocher et al. (2021)	640(LB)	1.9	2.3	1.51	28.0	45.7
RTMDet-tiny	640×640	4.8	8.1	0.98	**41.1**	**57.9**
RTMDet-tiny+FGD	640×640	4.8	8.1	0.98	**41.2**	**58.0**
RTMDet-tiny+MGD	640×640	4.8	8.1	0.98	**41.4**	**58.1**
RTMDet-tiny+Ours	640×640	4.8	8.1	0.98	**41.8**	**58.4**
YOLOv5-s	640(LB)	7.2	8.3	1.63	37.4	56.8
YOLOv6-s	640(LB)	17.2	22.1	0.92	43.5	60.4
RTMDet-s	640×640	8.99	14.8	1.22	**44.6**	**61.9**
RTMDet-s+FGD	640×640	8.99	14.8	1.22	**44.9**	**62.2**
RTMDet-s+MGD	640×640	8.99	14.8	1.22	**45.2**	**62.5**
RTMDet-s+Ours	640×640	8.99	14.8	1.22	**45.7**	**63.2**
YOLOv5-m	640(LB)	21.2	24.5	1.89	45.4	64.1
YOLOX-m	640×640	25.3	36.9	1.68	46.9	65.6
RTMDet-m	640×640	24.7	39.3	1.62	**49.4**	**66.8**
RTMDet-m+FGD	640×640	24.7	39.3	1.62	**49.6**	**67.0**
RTMDet-m+MGD	640×640	24.7	39.3	1.62	**49.8**	**67.3**
RTMDet-m+Ours	640×640	24.7	39.3	1.62	**50.2**	**68.3**
YOLOv5-l	640(LB)	46.5	54.6	2.46	49.0	67.3
YOLOX-l	640×640	54.2	77.8	2.19	49.7	68.0
YOLOv6-l	640(LB)	58.5	72.0	1.91	51.0	-
YOLOv7 Wang et al. (2022)	640(LB)	36.9	52.4	2.63	51.2	-
PPYOLOE-l	640×640	52.2	55.0	2.57	51.4	68.6
RTMDet-l	640×640	52.3	80.2	2.40	**51.5**	**68.8**
RTMDet-l+FGD	640×640	52.3	80.2	2.40	**51.6**	**68.9**
RTMDet-l+MGD	640×640	52.3	80.2	2.40	**51.8**	**69.0**
RTMDet-l+Ours	640×640	52.3	80.2	2.40	**52.3**	**69.9**

Implementation Details We conduct experiments on COCO dataset, which contains about 118K images in the `train2017` set and 5K images in the `val2017` set for training and validation, respectively. For ablation studies, we trained our models on the `train2017` set for 300 epochs and validated them on the `val2017` set. For hyper-parameters, object detection, and instance segmentation, the optimizer used is AdamW Kingma and Ba (2015), with a base learning rate of 0.004. A weight decay 0.05 is applied, excluding bias and norm parameters He et al. (2019). The optimizer momentum is set to 0.9, and the batch size is 256. The learning rate schedule follows a Flat-Cosine pattern with 300 training epochs. Additionally, there are 1000 warmup iterations. The input size for these tasks is 640x640. As for augmentation, cached Mosaic and MixUp techniques are employed for the first 280 epochs, followed by LSJ Ghiasi et al. (2022); Wu et al. (2019) for the last 20 epochs. The exponential moving average (EMA) decay is set to 0.9998. All our object detection and instance segmentation models are trained on 8 NVIDIA A100 GPUs. We evaluate the model performance on object detection and instance segmentation by box AP and mask AP.

Compare to real-time detectors. Table 1 comprehensively compares our method's performance against various state-of-the-art object detection models, including YOLOv5, YOLOX, YOLOv6, and PPYOLOE. The comparison encompasses the detection accuracy metrics and the model complexity regarding parameters, FLOPs, and inference latency. For the tiny model configuration, RTMDet-tiny+Ours achieves an impressive AP of 41.8 and

AP50 of 58.4. Compared to other tiny models like YOLOv6-tiny (AP 40.3, AP50 56.6) and YOLOX-tiny (AP 32.8, AP50 50.3), our method demonstrates a significant improvement in both AP and AP50 scores, outperforming these baselines. Moving to the small model configuration, RTMDet-s+Ours achieves an AP of 45.7 and an AP50 of 63.2. These results surpass the performance of other small models such as PPYOLOE-s (AP 43.0, AP50 59.6), YOLOv6-s (AP 43.5, AP50 60.4), and YOLOX-s (AP 40.5, AP50 59.3), indicating the effectiveness of our proposed method in enhancing object detection accuracy. RTMDet-m+Ours achieves an impressive AP of 50.2 and an AP50 of 68.3 for the medium model configuration. These results outperform other medium models like PPYOLOE-m (AP 49.0, AP50 65.9), YOLOv6-m (AP 48.5), and YOLOX-m (AP 46.9, AP50 65.6), further highlighting the efficacy of our approach in improving object detection performance. In the large model configuration, RTMDet-l+Ours achieves the highest AP of 52.3 and an AP50 of 69.9. This result surpasses the performance of other large models, including YOLOv7 (AP 51.2), YOLOv6-l (AP 51.0), PPYOLOE-l (AP 51.4, AP50 68.6), and YOLOX-l (AP 49.7, AP50 68.0), demonstrating the superiority of our method in enhancing object detection accuracy for large models. Overall, the results demonstrate that our proposed method consistently improves the AP and AP50 scores across various model configurations, from tiny to large, compared to other state-of-the-art object detection models. This improvement in object detection accuracy is a notable achievement, as it showcases the effectiveness of our approach in enhancing the performance of object detectors, potentially leading to more reliable and accurate detection systems in various applications.

Compare to detection distillation methods. The results in Table 1 demonstrate the effectiveness of our method (referred to as ”Ours”) in improving the performance of the RTMDet models. The table presents the comparison of different models with and without the application of various detection distillation methods, including FGD and MGD. We observe that across all the RTMDet variants, the inclusion of our method consistently leads to an increase in AP. For example, in the case of RTMDet-tiny, the AP improves from 41.1 to 41.8 when our method is applied. This trend is also observed in RTMDet-s, RTMDet-m, and RTMDet-l models, with AP values increasing from 44.6 to 45.7, 49.4 to 50.2, and 51.5 to 52.3, respectively. The improvement is further evident when considering AP50, which measures the Average Precision at an IoU threshold of 0.5. Similar to AP, our method consistently leads to an increase in AP50. For instance, in the case of RTMDet-tiny, AP50 increases from 57.9 to 58.4 when our method is applied. This trend is also observed in RTMDet-s, RTMDet-m, and RTMDet-l models, with AP50 values increasing from 61.9 to 63.2, 66.8 to 68.3, and 68.8 to 69.9, respectively. Comparing the performance of our method with the other detection distillation methods (FGD and MGD), it is evident that our method consistently achieves the highest AP and AP50 improvements across all models. This indicates that our method outperforms the other methods in terms of enhancing the detection performance. Overall, the results demonstrate that the inclusion of our method significantly improves the detection performance of the RTMDet models, as evidenced by the increase in both AP and AP50. Our method consistently outperforms the other detection distillation methods, highlighting its effectiveness in improving the model's performance.

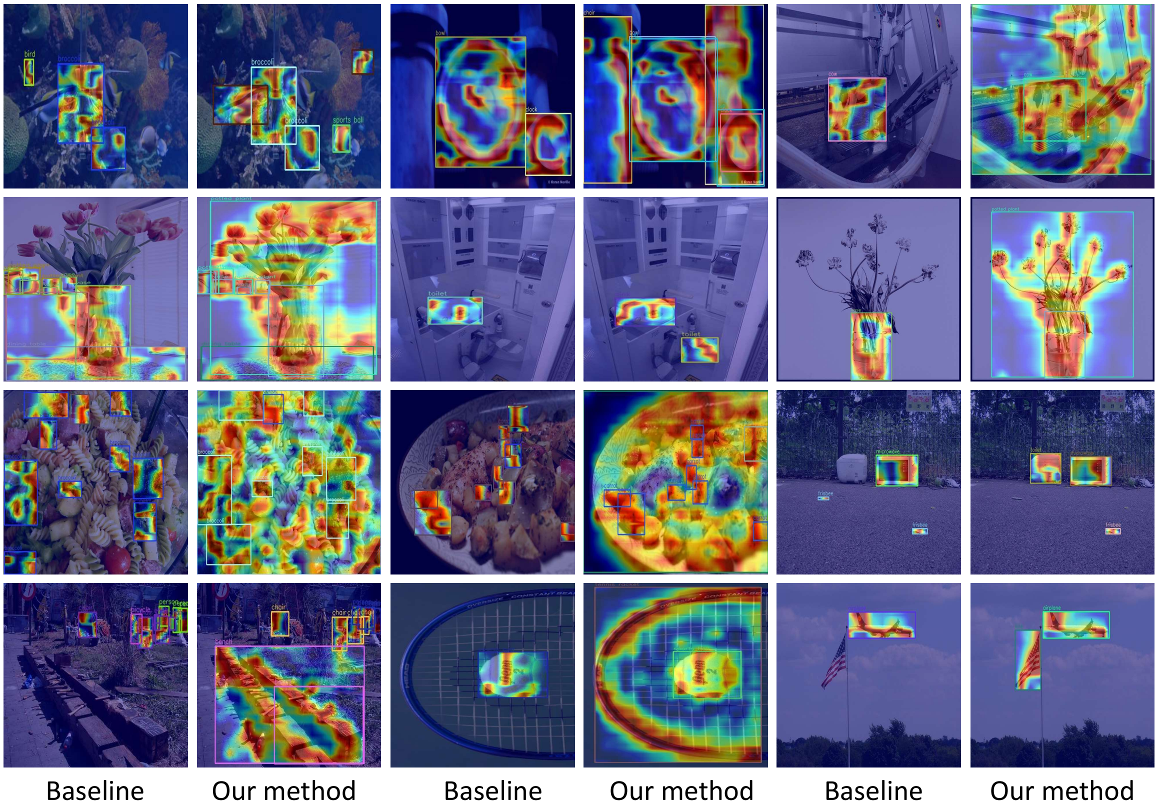

Figure 2: Qualitative analysis of Baseline model and our approach on COCO benchmarks.

4.1. Instance Segmentation

We also apply our method to the more challenging instance segmentation task. We use Mask R-CNN as our baseline models and distill between backbone architectures. The models are trained on the COCO2017 training set and are evaluated on the validation set. Table 2. provides a detailed analysis of the experimental results, such as segmentation on the MS COCO 2017 dataset. The table presents the performance of various knowledge distillation methods, including the proposed Shared Knowledge Distillation (our method) approach, in comparison with the baseline student model (Mask RCNN-R50) and other state-of-the-art techniques such as FKD, FGD, and MGD. Our method achieves an impressive score of 41.3, significantly outperforming the baseline student model by a substantial margin of 5.9 (41.3 vs. 35.4). This remarkable improvement highlights the effectiveness of our method in boosting the performance of the student model, for instance, segmentation tasks.

Comparison results. Our knowledge distillation method achieves superior performance compared to other techniques like FKD, FGD, and MGD for instance segmentation on the MS COCO 2017 dataset. It significantly outperforms these methods in overall Average Precision (AP), surpassing FKD by 3.9%, FGD by 3.5%, and MGD by 3.2%. Further analysis reveals our method's strengths in handling objects of varying scales. For small objects, it

Table 2: Experiments of instance segmentation on MS COCO2017. AP means average precision.

Distillation Method	AP	AP_S	AP_M	AP_L
Rtmdet-Tiny (S)	35.4	19.1	38.6	48.4
FKD	37.4	19.7	40.5	52.1
FGD	37.8	17.1	40.7	56.0
MGD	38.1	17.1	41.1	56.3
Ours	41.3	23.1	45.0	55.2

Table 3: Results of different Teacher for RTMDet-tiny+Ours.

Teacher	Rtmdet-L	Rtmdet-S
AP	40.8	41.8
AP_{50}	57.6	58.4
AP_{75}	43.8	45.4
AP_S	20.8	21.7
AP_M	44.7	46.2
AP_L	58.5	59.7

Table 4: Results of different components.

RTMDet-tiny	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Ours	41.8	58.4	45.4	21.7	46.2	59.7
no $L_{Multiscale}$	41.1	57.0	43.5	21.0	44.0	57.5
no L_{FDMD}	41.5	60.9	43.9	23.0	44.5	54.0
no L_{FAD}	41.2	58.4	43.3	23.3	44.0	52.2
RTMDet-s+Ours	45.7	63.2	49.7	25.2	50.1	63.5
no $L_{Multiscale}$	45.1	60.0	46.5	23.0	47.0	62.5
no L_{FDMD}	44.5	60.9	46.9	23.0	47.5	62.0
no L_{FAD}	44.2	60.4	46.3	23.3	46.0	62.2

Table 5: Results of different KD loss weight.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RTMDet-tiny+Ours	41.8	58.4	45.4	21.7	46.2	59.7
KD Loss Weight=0.1	40.4	56.3	43.2	19.0	43.1	57.3
KD Loss Weight=1.0	41.8	58.4	45.4	21.7	46.2	59.7
KD Loss Weight=5.0	41.2	58.4	43.3	23.3	44.0	57.2
RTMDet-s+Ours	45.7	63.2	49.7	25.2	50.1	63.5
KD Loss Weight=0.1	44.3	62.4	47.1	24.6	47.2	61.5
KD Loss Weight=1.0	45.7	63.2	49.7	25.2	50.1	63.5
KD Loss Weight=5.0	43.5	62.5	47.0	23.5	48.0	61.3

obtains an AP of 23.1%, outperforming the baseline student by 4.0% and other methods. For medium objects, it achieves an impressive AP of 45%, substantially higher than the baseline and other techniques. While slightly behind FGD and MGD for large objects, our method demonstrates robust performance across different object scales, particularly excelling at small and medium instances which are crucial in real-world scenarios. Results clearly highlight our knowledge distillation method's effectiveness in boosting the student model's instance segmentation performance.

4.2. Ablation Study

Analysis for different components in our method. In our method, experiments were conducted to assess the impact of different design components, as detailed in Table 4. The results across various model configurations highlight that our complete method, integrating all components, yields the highest AP and AP50 scores. Removal of any component, including the multiscale component ($L_{Multiscale}$), feature distillation and mask distillation component (L_{FDMD}), or feature alignment distillation component (L_{FAD}), results in performance decline in terms of AP and AP50 metrics. Notably, the multiscale component enhances the model's object detection abilities at different scales, as evidenced by the significant decrease in AP and AP50 scores upon its exclusion. The feature distillation and mask distillation component also significantly contribute to performance, particularly in the tiny model configuration. The feature alignment distillation component is crucial for maintaining high AP and AP50 scores.

Analysis for different Teacher. In our ablation study, we investigated the impact of different Teacher models on the performance of our method, as presented in Table 3. We compared the results of our full model with two different Teacher models: Rtmdet-L and Rtmdet-S. The results indicate that using the Rtmdet-S Teacher model led to a slightly higher mAP of 41.40, compared to 40.7 achieved by the Rtmdet-L Teacher model. This improvement is consistent across most evaluation metrics. These findings suggest that the choice of Teacher model has a noticeable effect on the overall performance of our method, with the Rtmdet-S model demonstrating better results than the Rtmdet-L model.

Analysis for loss weight of Knowledge Distillation We also experimentally analyze the impact of weight loss for knowledge distillation on detection results. Across both model configurations in Table 5, the results indicate that our proposed KD loss weight of 1.0 strikes the optimal balance, leading to the highest AP and AP50 scores. Deviating from this value, either by decreasing (0.1) or increasing (5.0) the KD loss weight, negatively impacts the detection performance. Furthermore, the table provides insights into the performance across different object scales (small, medium, and large). Our proposed method with a KD loss weight of 1.0 consistently achieves the highest AP scores across all object scales, demonstrating its effectiveness in detecting objects of varying sizes. In summary, the results suggest that our proposed loss weight of 1.0 is optimal for achieving the best detection performance, as measured by AP and AP50 metrics, across different model configurations and object scales. This finding highlights the importance of carefully tuning the KD loss weight to maximize the effectiveness of our knowledge distillation approach for object detection tasks.

Qualitative analysis. Figure. 2 shows significant improvements in detecting small objects and reducing false positives in object detection. Small objects are challenging to detect due to their limited visual information and reduced spatial context. Traditional methods struggle to detect and locate these objects accurately, leading to lower detection rates. Our method addresses this issue by implementing advanced techniques specifically designed for detecting small objects. We enhance the representation and discrimination of small objects by incorporating features such as context aggregation and multi-scale analysis. This allows our model to capture important details and subtle visual cues, resulting in higher detection rates and improved localization accuracy for small objects. Additionally, our method effectively reduces false positives, where the model incorrectly identifies background regions or irrelevant objects as targets. We achieve this by integrating advanced filtering mechanisms and context-aware reasoning. Our model can distinguish between true positive detections and false positives by considering contextual information, semantic consistency, and spatial relationships. This leads to a significant reduction in false positives.

5. Conclusion

In this paper, we propose a method that combines multi-layer feature distillation and feature alignment distillation to enhance the knowledge transfer process. The multi-layer feature distillation aligns the spatial dimensions of the teacher and student features, enabling the student model to capture spatial relationships and improve performance in tasks like image recognition and object detection. Additionally, the feature alignment distillation focuses on matching channel-wise features, guiding knowledge transfer, and helping the student model capture object details and channel-specific knowledge. By leveraging these two components, our method aims to improve the performance of the student model by effectively utilizing the knowledge and expertise of the teacher model. Our work contributes to advancing small object detection techniques and opens up opportunities for further research in this area.

References

Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–221, 2018.

Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. In *Advances in Neural Information Processing Systems*, 2022.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 742–751. Curran Associates, Inc., 2017.

Kunlong Chen, Liu Yang, Yitian Chen, Kunjin Chen, Yidan Xu, and Lujun Li. Gp-nas-ensemble: a model for the nas performance prediction. In *CVPRW*, 2022.

Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7842–7851, June 2021.

Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *CVPR*, 2023a.

Peijie Dong, Xin Niu, Lujun Li, Zhiliang Tian, Xiaodong Wang, Zimian Wei, Hengyue Pan, and Dongsheng Li. Rd-nas: Enhancing one-shot supernet ranking ability via ranking distillation from zero-cost proxies. In *ICASSP)*, 2023b.

Peijie Dong, Lujun Li, Xinglin Pan, Zimian Wei, Xiang Liu, Qiang Wang, and Xiaowen Chu. Parzc: Parametric zero-cost proxies for efficient nas. *arXiv preprint arXiv:2402.02105*, 2024.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3510–3519, October 2021.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin Cubuk, Quoc Le, Barret Zoph, Google Research, Brain Team, and U Berkeley. Simple copy-paste is a strong data augmentation method for instance segmentation. 2022.

glenn jocher et al. yolov5. https://github.com/ultralytics/yolov5, 2021.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2154–2164, June 2021.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yiming Hu, Xingang Wang, Lujun Li, and Qingyi Gu. Improving one-shot nas with shrinking-and-expanding supernet. *Pattern Recognition*, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29: 7389–7398, 2020.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13619–13627, June 2022a.

Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1306–1313, 2022b.

Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *ECCV 2022*, 2022.

Lujun Li, Peijie Dong, Anggeng Li, Zimian Wei, and Ya Yang. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *NeuIPS*, 2024.

Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017b.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oMI9PjOb9Jl.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

Liming Lu, Zhenghan CHen, Lu Lu, Xiaoyu, Yihang Rao, Lujun Li, and Shuchao Pang. Uniads: Universal architecture-distiller search for distillation gap. In *AAAI*, 2024.

Chuong H. Nguyen, Thuy C. Nguyen, Tuan N. Tang, and Nam L.H. Phan. Improving object detection by label assignment distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1005–1014, January 2022.

Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *AAAI*, 2022.

Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Shitong Shao, Xu Dai, Shouyi Yin, Lujun Li, Huanran Chen, and Yang Hu. Catch-up distillation: You only need to train once for accelerating sampling. *arXiv preprint arXiv:2305.10769*, 2023.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019a.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. *CoRR*, abs/1904.01355, 2019b.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *CoRR*, abs/2207.02696, 2022.

Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *CVPR*, 2019.

Zimian Wei, Hengyue Pan, Lujun Li Li, Menglong Lu, Xin Niu, Peijie Dong, and Dongsheng Li. Convformer: Closing the gap between cnn and vision transformers. *arXiv preprint arXiv:2209.07738*, 2022.

Zimian Wei, Hengyue Pan, Lujun Li, Peijie Dong, Zhiliang Tian, Xin Niu, and Dongsheng Li. Tvt: Training-free vision transformer search on tiny datasets. *arXiv preprint arXiv:2311.14337*, 2023.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 13668–13677, 2022a.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4643–4652, June 2022b.

Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9407–9416, June 2022.

Chendi Zhu, Lujun Li, Yuli Wu, and Zhengxing Sun. Saswot: Real-time semantic segmentation architecture search without training. In *AAAI*, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.