Journal of Computer Science & Technology

Indexed in: SCIE, EI, INSPEC, CBST, DBLP, etc.

Sponsored by: ICT, CAS & CCF

JCST Papers

Only for academic and non-commercial use

Thanks for reading!

JCST WeChat Subscription Account

JCST URL: https://jcst.ict.ac.cn

SPRINGER URL: https://www.springer.com/journal/11390

E-mail: jcst@ict.ac.cn

Online Submission: https://mc03.manuscriptcentral.com/jcst

Twitter: JCST_Journal

LinkedIn: Journal of Computer Science and Technology

Ma YC, Ma X, Hao TR *et al.* Knowledge distillation via hierarchical matching for small object detection. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(4): 798-810 July 2024. DOI: 10.1007/s11390-024-4158-5

Knowledge Distillation via Hierarchical Matching for Small Object Detection

Yong-Chi Ma (马永驰), Xiao Ma (马　啸), Tian-Ran Hao (郝天然), Li-Sha Cui (崔丽莎), *Member, CCF*
Shao-Hui Jin (靳少辉), and Pei Lyu* (吕　培), *Member, CCF*

School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450000, China

E-mail: mycsir@gs.zzu.edu.cn; iemaxiao@gs.zzu.edu.cn; 202011171030162@gs.zzu.edu.cn; ielscui@zzu.edu.cn
ieshjin@zzu.edu.cn; ielvpei@zzu.edu.cn

Received January 29, 2024; accepted June 20, 2024.

Abstract Knowledge distillation is often used for model compression and has achieved a great breakthrough in image classification, but there still remains scope for improvement in object detection, especially for knowledge extraction of small objects. The main problem is the features of small objects are often polluted by background noise and not prominent due to down-sampling of convolutional neural network (CNN), resulting in the insufficient refinement of small object features during distillation. In this paper, we propose Hierarchical Matching Knowledge Distillation Network (HMKD) that operates on the pyramid level P2 to pyramid level P4 of the feature pyramid network (FPN), aiming to intervene on small object features before affecting. We employ an encoder-decoder network to encapsulate low-resolution, highly semantic information, akin to eliciting insights from profound strata within a teacher network, and then match the encapsulated information with high-resolution feature values of small objects from shallow layers as the key. During this period, we use an attention mechanism to measure the relevance of the inquiry to the feature values. Also in the process of decoding, knowledge is distilled to the student. In addition, we introduce a supplementary distillation module to mitigate the effects of background noise. Experiments show that our method achieves excellent improvements for both one-stage and two-stage object detectors. Specifically, applying the proposed method on Faster R-CNN achieves 41.7% mAP on COCO2017 (ResNet50 as the backbone), which is 3.8% higher than that of the baseline.

Keywords knowledge distillation, object detection, small object detection, machine learning

1 Introduction

In recent times, there has been a significant surge in computer vision research, with a particular upswing in the prominence accorded to the domain of small object detection. Previous studies[1, 2] have often employed complex and large networks to enhance the detection accuracy of small objects. However, these performance gains typically come at the cost of increased computational demand, making it challenging to deploy such models on mobile and edge computing devices, which generally lack the necessary processing power. To address the computational limitations, a lot of work has focused on developing lightweight neural networks. Techniques such as quantization[3, 4], network pruning[5, 6], and knowledge distillation[7, 8] have proven effective in reducing the complexity of these networks. Knowledge distillation, in particular, involves compressing a trained large-scale model into a smaller, yet well-performing, lightweight model. In this process, the large model, referred to as the teacher, imparts its knowledge to

Regular Paper

Special Section of CVM 2024

This work was supported in part by the Joint Fund of the Ministry of Education for Equipment Pre-Research of China under Grant No. 8091B032257, the National Natural Science Foundation of China under Grant Nos. 62106232 and 62372415, the China Postdoctoral Science Foundation under Grant No. 2021TQ0301, and the Outstanding Youth Science Fund of Henan Province of China under Grant No. 242300421050.

*Corresponding Author

the smaller model, known as the student. The student, without altering its network architecture, not only learns the ground truth from the training data, but also benefits from the teacher's refined knowledge, thereby improving its generalization capabilities.

Most knowledge distillation methods are based on object features[9–11]. These researchers distill the suggested regions predicted by the region proposal network (RPN)[12] or the features extracted from the feature pyramid network (FPN)[13]. Researchers are continually striving to determine how to effectively locate and distill useful knowledge. For instance, Kang *et al.*[14] employed a conditional distillation network to identify the necessary knowledge and facilitate the transfer of instance-specific knowledge.

Although these methods have achieved significant breakthroughs, they often overlook the inherent differences in object sizes, which implies that objects of different scales contribute varying degrees of knowledge within the feature space. This variation leads to differing levels of difficulty in knowledge extraction. Consequently, objects of different scales should not be treated equally. As illustrated in Fig.1, different methods for small object detection are compared. Specifically, as shown in Fig.1(b), conventional knowledge distillation methods are evidently insufficient for accurately recognizing small objects. This insufficiency is apparent in the missed detections of small objects such as cars, people, and kites. The fundamental challenge in learning small object features lies in their potential dilution during the down-sampling process of convolutional neural networks (CNNs). Moreover, small objects are more susceptible to interference from background noise, making their knowledge distillation particularly challenging. Just as teachers should pay extra attention to knowledge that is ambiguous or difficult for students to grasp, similar care must be taken in distilling knowledge related to small objects to avoid omissions or misunderstandings. These issues pose a significant challenge in the effective refinement of small object knowledge.

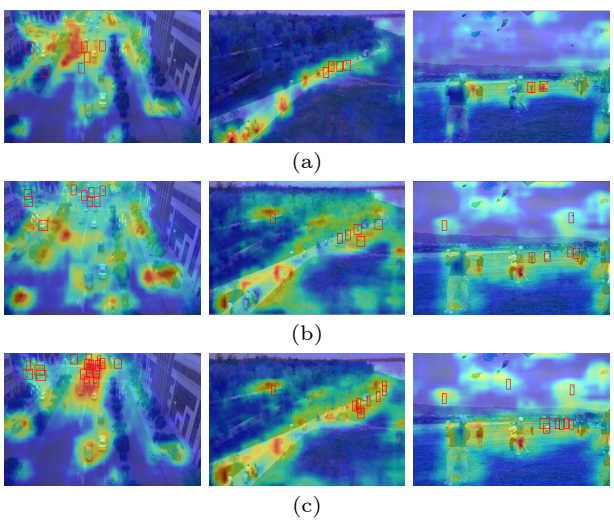


Fig.1. Detection results comparison of small objects using Grad-CAM heatmaps[15]. Small objects are highlighted with red boxes. It is evident that our method demonstrates a significantly higher sensitivity to small objects compared to the others. (a) Conventional Detection. (b) General KD. (c) Our HMKD.

To enhance students' understanding of small object knowledge, we propose Hierarchical Matching Knowledge Distillation Network (HMKD) as shown in Fig.2. This network is designed to improve the student model's learning of small object features in the shallow, high-resolution layers of the FPN, specifically focusing on the P2 to P4 layers. Through hierarchical matching, a decoding network is introduced to identify and extract challenging knowledge. Specifically, we first separate the foreground and background of the image to prevent small objects from being polluted by the background during down-sampling. Then we encode the strong semantic information of small objects at low resolution as inquiries and use the fine-grained feature values at high resolution as key values. As shown in Fig.1(b), most of the background area is highlighted. We observe that distilling knowledge for the foreground only is not the optimal approach; the relationship between foreground and background must also be considered. Therefore, we design an additional supplementary distillation module to impart the relationships between the foreground and background to the student as part of the distillation process. The contributions of this paper are as follows:

- We develop the Hierarchical Matching Knowledge Distillation Network (HMKD) that distills knowledge separately from the foreground and background, significantly improving the performance of small object detection.
- We design a distillation strategy that encodes high-semantic information at low-resolution of FPN as inquiries and represent fine-grained graph feature values at high-resolution as key-values to improve the refinement of small object features. Additionally, we introduce a supplementary distillation module to reduce background interference.
- We conduct extensive and comprehensive experiments on both one-stage and two-stage object detectors. HMKD shows consistent and significant improve-

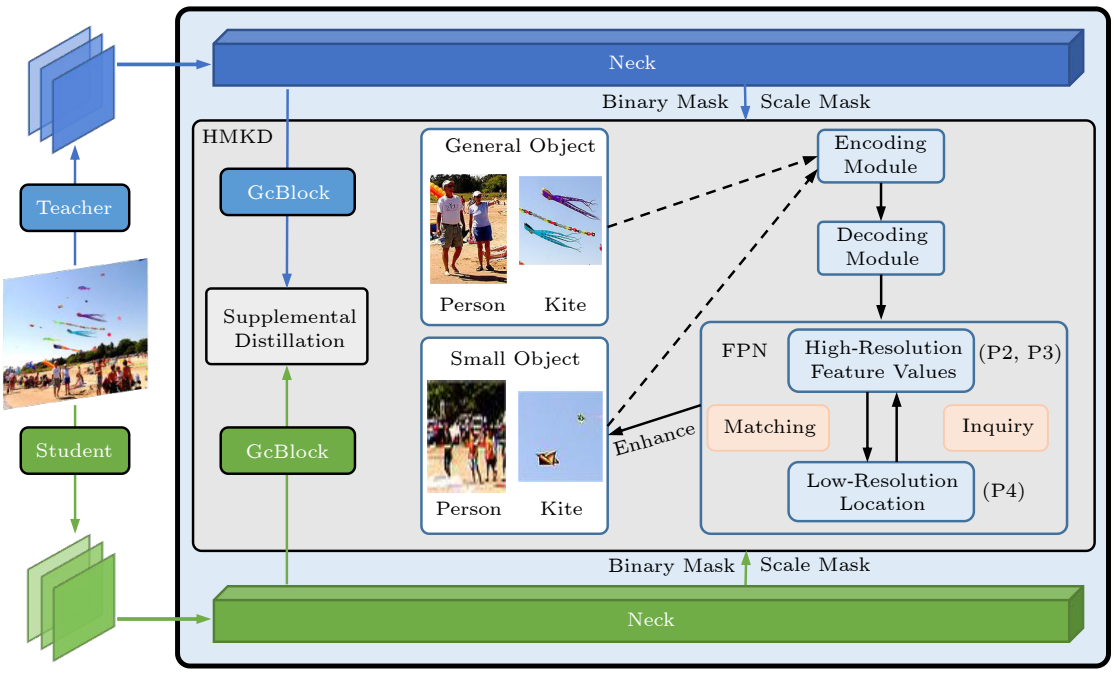


Fig.2. Pipeline of our method. Firstly, the foreground and background are separated to prevent the background information from excessively interfering with the extraction of small objects. Then HMKD is used to enhance the small object features, thus improving the detection ability of the student model. Meanwhile, GcBlock[16] is used to distill the background information.

ments on the COCO2017① and VisDrone② datasets.

2 Related Work

2.1 Knowledge Distillation

Recently, there have been many researches using knowledge distillation for object detection with notable results. Kang *et al.*[14] proposed a novel approach for extracting knowledge by using instance search to transfer image features from the instructor to the student. Yang *et al.*[17] addressed the issue of uneven differences between feature maps, which can negatively affect feature extraction. They optimized the separation of background and foreground to encourage the student model to focus on the teacher's channels and pixels. Chen *et al.*[18] proposed prediction-based methods to extract features from RPN regions and distill foreground knowledge. Wang *et al.*[9] focused on extracting regions with maximum IoU between anchor points and ground truth. Guo *et al.*[19] introduced a new loss function to address the imbalance between the background and objects. However, these methods are heuristic-based and require predefined rules, making them inflexible. In contrast, Zhang and Ma[10] incorporated the attention mechanism into knowledge rectification and established spatial channel attention for distillation. Chang *et al.*[20] proposed the DETRDistill method, a general knowledge distillation framework tailored for DETR. They utilized target-aware feature distillation to help the student model learn from the teacher model's object-centric features and introduced query prior distribution distillation to expedite the student model's learning from well-trained queries and stable distributions of the teacher model.

None of these methods pay specific attention to the challenge of detecting small objects. In the distillation process, objects of different sizes cannot be treated equally, as smaller objects are inherently more difficult to detect and are more prone to information loss. To address this issue, focused distillation is necessary to prevent the loss of critical information from small objects and to enhance detection accuracy. To overcome these limitations, we propose a distillation method that specifically enhances knowledge transfer for small objects. Our method uses hierarchical matching to improve the student model's understanding of difficult concepts, such as small objects, thereby ensuring better retention and detection performance.

①https://cocodataset.org/#download, Jul. 2024.
②https://github.com/VisDrone/VisDrone-Dataset, Jul. 2024.

2.2 Object Detection

CNN-based object detection frameworks can be categorized into one-stage[21], two-stage[12], and anchor-free[22] methods. A notable example of one-stage detector is RetinaNet[21], which applies focal loss to address the imbalance between the background and foreground. When combined with FPN, its performance is comparable to that of two-stage detectors. The YOLO[23] series, currently under active development, directly regresses box coordinates and class probabilities from image pixels, offering a significant speed advantage.

In contrast, a classical two-stage detector is Faster R-CNN[12], which employs a region proposal network (RPN) to efficiently generate proposal regions. FPN is used to capture multi-scale feature maps through lateral connections, offering superior detection accuracy, though it typically lags behind one-stage detectors in speed. Anchor-based detection methods utilize anchor boxes of various aspect ratios to label objects and apply heads to classify and regress each anchor.

Recently, an anchor-free detection framework FCOS[22] has been introduced, predicting label coordinates and candidate boxes using a fully convolutional network. Since all these frameworks require input features, our knowledge distillation approach can be effectively applied to enhance these detectors.

Huang *et al.*[24] proposed a simple yet powerful super token attention (STA) mechanism. STA decomposes global attention into sparse correlation graphs and low-dimensional attention multiplication, thereby creating a hierarchical visual transformer. Zhu *et al.*[25] introduced a novel dynamic sparse attention mechanism through dual-layer routing, enabling more flexible computation allocation based on content awareness. This new general-purpose visual transformer is named BiFormer. Tian *et al.*[26] addressed the challenge of fragile scalability with respect to target resolution and proposed the ResFormer framework. This framework is built upon the innovative concept of multi-resolution training, aimed at enhancing performance across various testing resolutions.

2.3 Small Object Detection

Small object detection is particularly challenging due to factors such as limited semantic information and vulnerability to interference from complex scenes. Most current research has improved small object detection performance through various approaches, including data augmentation[27, 28], enhancing input feature resolution[29, 30], multi-scale information fusion[31, 32], and leveraging contextual semantic information[1].

In recent years, scale regularization strategies such as SNIP[32] and SNIPER[33] have been developed to address the issue of varying object sizes across images of different resolutions. Chen *et al.*[34] introduced a feedback-driven data provider, aiming to enhance small object detection by balancing detection loss. Although this technique provides a promising solution to the challenges of small object detection, its application within the knowledge distillation paradigm has been largely overlooked. Similarly, TridentNet[31] is a parallel multi-branch approach, utilizing different perceptual fields to generate more accurate and discriminative features for small objects.

3 Method

3.1 Overview

In contemporary object detection methodologies that hinge on detector architecture, the extraction of features pertaining to diminutive entities frequently entails recourse to high-resolution feature maps. However, a student may have weaker capabilities in extracting such features compared with its teacher. To address this issue, we propose an algorithm that enhances the student's ability to learn features of small objects. Specifically, we strengthen the extraction of high-resolution features in the shallow stages of the FPN. We adopt a hierarchical matching approach for knowledge transfer, taking high-level semantics of small objects at low-resolution as inquiries and fine-grained graph features proposed by teachers at high-resolution layers as key-values. Finally, the student's knowledge is updated by an attention-weighted feature extraction loss.

3.2 Hierarchical Matching Knowledge Distillation

3.2.1 Foreground and Background Separation

In this subsection, our approach HMKD is elaborated, which is an enhanced small object knowledge transfer algorithm based on hierarchical matching. Carion *et al.*[35] applied the idea of encoder and decoder to object detection in DETR. Kang *et al.*[14] al-

so applied it to knowledge distillation. Inspired by this, we notice the difficulty in transferring information about small objects in knowledge distillation. HMKD utilizes the idea of encoder and decoder. It uses the high semantic information of small objects at low resolution as inquiries, and the fine-grained graph feature values presented by the teacher in the high-resolution layer are used as keys. Finally, the student is updated through feature distillation losses based on attention weighting, whose goal is to make these difficult knowledge available to the student.

Since background noise may have an effect on small object features, we decide to remove it as much as possible at first. This allows the model to focus on the pure distillation of the foreground. We separate the background and foreground using the binary mask M.

$$M_{x,\,y} = \begin{cases} 1, & \text{if } (x,y) \in g, \\ 0, & \text{otherwise}, \end{cases}$$

where g indicates the ground-truth area of objects, x is the horizontal coordinate of the feature map, and y is the vertical coordinate. If (x, y) matches the ground-truth, then $M_{x,\,y} = 1$, otherwise it is 0. The proportions of the background and objects are different in different images. Large-scale objects cause more losses because they take up more pixels. Thus we treat each different goal equally in order not to affect the extraction of small goals. The proportional mask K is to balance the loss in separation:

$$K_{x,\,y} = \begin{cases} \dfrac{1}{H_g W_g}, & \text{if (x, y)} \in \text{g}, \\ \dfrac{1}{N_{bg}}, & \text{otherwise}, \end{cases}$$

$$N_{bg} = \sum_{x=1}^{H} \sum_{y=1}^{W} (1 - M_{x,\,y}),$$

where H_g and W_g denote the height and width of the boxes respectively. If a pixel is part of more than one object, K takes the minimum value.

3.2.2 Distillation Strategy

The one-stage RetinaNet classifies objects using a single feature pyramid network (FPN) module, while the two-stage faster R-CNN employs two detection heads for localization. Given an input image of size $H \times W$, the feature maps produced by the FPN can be denoted as a set $\boldsymbol{P} = \{\boldsymbol{P}_l \mid \boldsymbol{P}_l \in \mathbb{R}^{(H/2^l)\times(W/2^l)\times C}\}$, where l indicates the level of the pyramid, C represents the number of channels in each feature map, corresponding to different types of learned feature representations such as edges, textures, or color information at each level of the pyramid.

We introduce a new head module, called Inquiry Head, which works in parallel with the original detection head to predict the approximate location of small objects. The Inquiry Head receives feature maps as input with a stride of 2^l, and outputs a probability map $\boldsymbol{MAP}_l \in \mathbb{R}^{(H/2^l)\times(W/2^l)}$), where $\boldsymbol{MAP}_l^{x',\,y'}$ denotes the probability that the network contains a small object at location $(x',\ y')$.

If the area occupied by an object is less than 32×32 pixels, it is considered a small object. The distance between the central point of the object and other positions on the feature map is encoded as part of the Inquiry Head's object mapping. Distances less than a certain threshold s_l are encoded as 1, while the others are set to 0. We train the Inquiry Head using focal loss, and select positions with prediction results higher than a critical value t as potential locations for small objects.

Next, the student needs to learn the more difficult parts of the knowledge. We suggest focusing on delivering small object information from the teacher to the student, as illustrated in Fig.3. $\boldsymbol{I}_i^{\mathrm{S}}$ and $\boldsymbol{I}_i^{\mathrm{T}}$ correspond to the i-th piece of information for the student and teacher, respectively.

$$L_{\text{small-distill}} = \sum_{i=1}^{N} \sum_{x=1}^{H} \sum_{y=1}^{W} M_{x,\,y} K_{x,\,y} L_s(\boldsymbol{I}_{i(l-1)}^{\mathrm{S}}, \boldsymbol{I}_{i(l-1)}^{\mathrm{T}}).$$

The knowledge of the teacher is denoted by $\boldsymbol{T}$, and the knowledge of the condition $\boldsymbol{z}_i$ can be expressed as $\boldsymbol{I}_i^{\mathrm{T}} = D(\boldsymbol{T}, \boldsymbol{z}_i)$, with $\boldsymbol{I}_i^{\mathrm{S}}$ being defined similarly, where D represents the decoding module. Since currently commonly used detector contains a feature pyramid network (FPN), we denote the multi-scale features as:

$$\boldsymbol{T} = \left\{ \boldsymbol{Z}_\gamma \in \mathbb{R}^{C\times H_\gamma \times W_\gamma} \;\middle|\; \gamma \in E \right\},$$

where $\boldsymbol{T}$ represents the collection of multi-scale features from the teacher, E denotes the set of different scales (resolutions), and C indicates the channel dimension. We obtain $\boldsymbol{X}^{\mathrm{T}} \in \mathbb{R}^{U\times C}$ by concatenating features from different scales, where $U = \sum_{\gamma\in E} H_\gamma W_\gamma$ represents the total number of pixels across multiple scales. After extracting the information related to small objects, we need to annotate it, denoted by $A = \{a_i\}_{i=1}^{Q}$, where Q is the number of objects and a_i

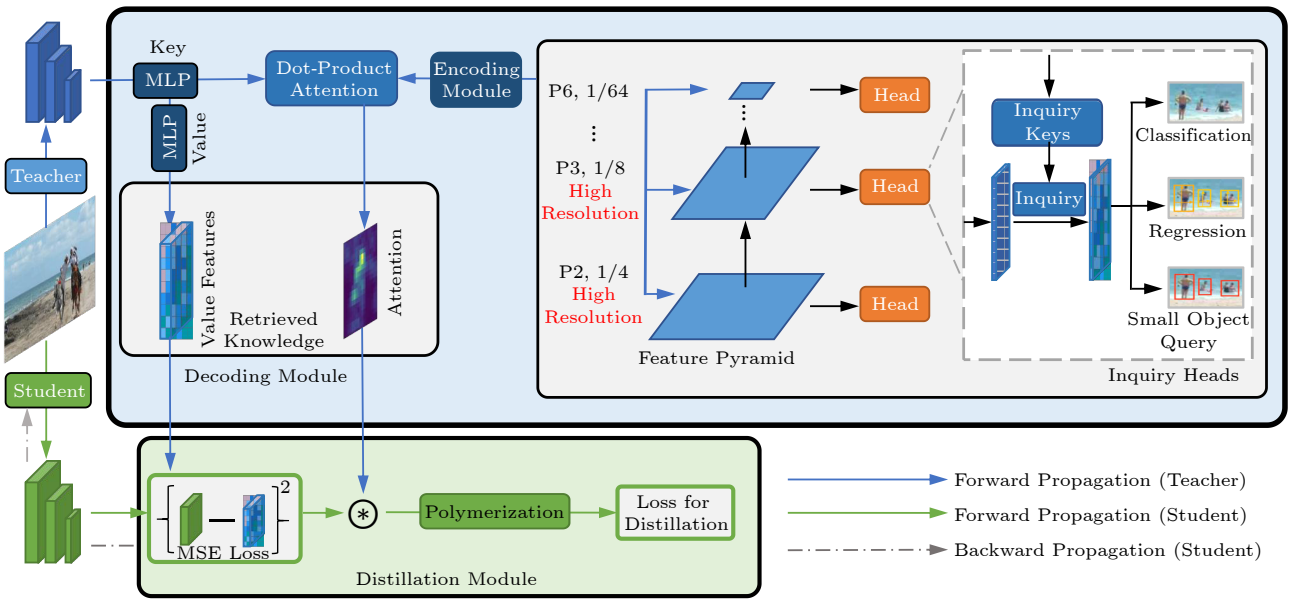


Fig.3. Illustration of HMKD in detail. The high semantic information of small objects at low resolution is encoded as a inquiry, and the fine-grained feature values at high resolution are used as key values. In this way, the detection ability of the student for small objects gradually converges to that of the teacher.

represents the annotation of each object, including category and size information.

To generate a learnable embedding of localization knowledge for each small object, we need to annotate the inquiry feature vector. This inquiry feature vector $\boldsymbol{b}_i$ specifies the conditions for collecting the required knowledge:

$$\boldsymbol{b}_i = F_b(\delta(a_i)), \quad \boldsymbol{b}_i \in \mathbb{R}^C,$$

where $\delta(*)$ is the encoding function and F_b is the multi-layer perceptron network (MLP). This process involves the dot product attention mechanism[36] with N_m heads, which manages the inquiry key attention. Each head j consists of three linear layers $(\boldsymbol{F}_j^{\mathrm{k}}, \boldsymbol{F}_j^{\mathrm{q}}, \boldsymbol{F}_j^{\mathrm{v}})$ responsible for generating keys, inquiries, and values, respectively. The eigenvalue $\boldsymbol{I}_i^{\mathrm{T}}$ is computed by integrating $\boldsymbol{X}^{\mathrm{T}}$ and the location embedding $\boldsymbol{V}$, where $\boldsymbol{V} \in \mathbb{R}^{U\times C}$ represents the spatial information that is projected through the teacher model's network to enhance contextual relevance.

$$\boldsymbol{I}_{i(l-1)}^{\mathrm{T}} = \boldsymbol{F}_j^{\mathrm{k}}\left(\boldsymbol{X}^{\mathrm{T}} + F_{pe}(\boldsymbol{V})\right), \quad \boldsymbol{F}_j^{\mathrm{k}} \in \mathbb{R}^{U\times c},$$

$$\boldsymbol{O}_{j(l-1)}^{\mathrm{T}} = \boldsymbol{F}_j^{\mathrm{v}}\left(\boldsymbol{X}^{\mathrm{T}}\right), \quad \boldsymbol{O}_j^{\mathrm{T}} \in \mathbb{R}^{U\times c},$$

$$\boldsymbol{b}_{ij(l-1)} = \boldsymbol{F}_j^{\mathrm{q}}\left(\boldsymbol{b}_i\right), \quad \boldsymbol{b}_{ij} \in \mathbb{R}^c,$$

$$m_{ij} = \mathrm{softmax}\left(\frac{\boldsymbol{I}_i^{\mathrm{T}}\boldsymbol{b}_{ij}}{\sqrt{c}}\right), \quad m_{ij} \in \mathbb{R}^U,$$

where F_{pe} denotes the linear projection on the location embedding. The value features $\boldsymbol{O}_{j(l-1)}^{\mathrm{T}}$ are projected onto a subspace with channel dimension $c = C/N_m$ using the linear mapping $\boldsymbol{F}_j^{\mathrm{v}}$, and the inquiry features $\boldsymbol{b}_{ij(l-1)}$ are similarly projected using $\boldsymbol{F}_j^{\mathrm{q}}$. Both projections are performed on the feature space $\boldsymbol{P}_{l-1}$. The perceptual attention mask m_{ij} for the j-th head of the i-th information is obtained by the normalized dot product of $\boldsymbol{I}_{i(l-1)}^{\mathrm{T}}$ and $\boldsymbol{b}_{ij(l-1)}$. In summary, the inquiries along the key and value features describe the correlation between results and the small object information. We gather $\boldsymbol{I}_{i(l-1)}^{\mathrm{T}} = \{(m_{ij}, \boldsymbol{O}_{j(l-1)}^{\mathrm{T}})\}_{j=1}^{N_m}$ as the localization information extracted for small objects from $\boldsymbol{T}$, which encodes the knowledge corresponding to the i-th information.

It should be noted that there are fundamental differences between our method and Querydet[2]. The inquiry mechanisms of the two are different. Our approach focuses on enhancing the transfer of knowledge for small objects from the teacher to the student during the knowledge distillation process.

3.2.3 *Supplemental Distillation*

Next is the supplemental distillation module, where we use Gcblock[16] to extract background knowledge. In Subsection 3.2.1 and Subsection 3.2.2, we first separated the background and distilled the foreground knowledge, but this approach neglects the interaction between the foreground and background. Therefore, we utilize this module to supplement miss-

ing knowledge regarding the overall relationship between objects and backgrounds, and transfer it from the teacher to the student. The loss function related to the background is as follows:

$$L_{\text{background}} = \alpha \times \sum \left(f'\left(|\boldsymbol{F}^{\text{T}} - \boldsymbol{F}^{\text{S}}|\right)\right)^2,$$
$$f'(\boldsymbol{F}) = \boldsymbol{F} + \boldsymbol{C}_{v2} ReLU\left(LN\left(\boldsymbol{C}_{v1}\sum_{j=1}^{N_p}\frac{e^{\boldsymbol{C}_k \boldsymbol{F}_j}}{\sum_{m=1}^{N_p} e^{\boldsymbol{C}_k \boldsymbol{F}_m}}\boldsymbol{F}_j\right)\right),$$

where $\boldsymbol{C}_k$, $\boldsymbol{C}_{v1}$, and $\boldsymbol{C}_{v2}$ are the corresponding convolution layers, LN denotes layer normalization, N_p represents the total number of pixels, and α is the balance factor.

3.3 Distillation Loss

Finally, we present the final knowledge distillation formula. The attention mask m_{ij} quantifies the correlation between the features and each specific piece of information during this transfer.

$$L_{\text{small-distill}} = L_{\text{others}} + \frac{1}{N_m N_s}\sum_{j=1}^{N_m}\sum_{i=1}^{N_s}\langle m_{ij}, L_{\text{MSE}}(\boldsymbol{I}^{\text{S}}_{i(l-1)}, \boldsymbol{I}^{\text{T}}_{i(l-1)})\rangle,$$

where N_s represents the total number of distinct information pieces (or features) that are considered in the distillation process, and $L_{\text{MSE}}(\boldsymbol{I}^{\text{S}}_{i(l-1)}, \boldsymbol{I}^{\text{T}}_{i(l-1)})$ is the mean squared error of pixels in the hidden dimension, which stabilizes the normalized feature. $\langle -, - \rangle$ is the Dirac notation for the inner product, and L_{others} is the feature distillation for the other dimensional object information. Combined with the supervised learning loss $L_{\text{detection}}$, the formula is summarized as follows:

$$L_{\text{total}} = L_{\text{detection}} + L_{\text{background}} + \beta L_{\text{small-distill}},$$

where β is a hyper-parameter. As described above, we divide the small object knowledge into key knowledge taught to the student by the teacher while distilling the knowledge of the object based on hierarchical matching. We use the low-resolution high-semantic information of the small object in the neck stage as an inquiry, and the high-resolution fine-grained feature map values as keys to enhance the student's learning of small object knowledge.

4 Experiments

4.1 Datasets

MS COCO 2017 Dataset. Our primary experiments were conducted on the COCO dataset[③], which consists of 80 object classes. The training set includes 120 000 images, while the validation set comprises 5 000 images. All subsequent results were evaluated on this validation set, using average precision as the metric to assess the performance of different detectors.

VisDrone Dataset. We also conducted experiments on the VisDrone dataset[④], which was collected by the AISKYEYE team at Tianjin University, and contains 11 categories of drone-captured images. The training set includes 6 471 images across 10 object classes, predominantly featuring small objects.

4.2 Implementation Details

Our experiments were all set up in the widely used Detectron2 library[⑤] and AdelaiDet library[⑥]. All programs were executed on a single NVIDIA RTX 3 060 GPU, and due to memory constraints, we set the batch size to 2. We followed the criterion in Detectron2 where 1x scheduler denotes 9 000 training sessions. For optimizing the transformer decoder during knowledge distillation, we used the AdamW optimizer[38] as the decoder. The MLP used the regular settings[35, 36]. Other hyper-parameters were set with reference to DETR[35], where the learning rate and weight decay were set to 0.001. We set the hidden dimension of the decoder and all MLPs to 256, and the decoder had eight heads in parallel. In the supplemental distillation module, the background distillation hyper-parameters $\alpha = 1.0 \times 10^{-3}$ and small object distillation hyper-parameters $\beta = 1.0 \times 10^{-3}$.

4.3 Main Results

Results on Dataset MS COCO. Our method demonstrates strong generality and can be readily applied to various detection frameworks. We began by conducting experiments with popular detectors, as detailed in Table 1. In the table, the bold values indicate the best performance among all methods in each

③https://cocodataset.org/#download, Jul. 2024.
④https://github.com/VisDrone/VisDrone-Dataset, Jul. 2024.
⑤https://github.com/facebookresearch/detectron2, Jul. 2024.
⑥https://git.io/adelaidet, Jul. 2024.

Table 1. Results on Dataset MS COCO 2017

Method	Faster R-CNN				RetinaNet			
	mAP	AP_S	AP_M	AP_L	mAP	AP_S	AP_M	AP_L
ResNet101(teacher) 3x	42.0	25.2	45.6	54.6	40.4	24.0	44.3	52.2
ResNet50(student) 1x	37.9	22.4	41.1	49.1	37.4	23.1	41.6	48.3
+FitNet[37]	39.3	22.7	42.3	51.7	38.2	23.1	41.6	48.8
+FGFI[9]	39.3	22.5	42.3	52.2	38.6	21.4	42.5	51.5
+ICD[14]	40.9	24.5	44.2	53.5	40.7	24.2	45.0	52.7
+FGD[17]	40.5	22.6	44.7	53.2	39.7	22.0	43.7	53.6
+TinyKD[39]	33.1	15.8	36.2	45.1	–	–	–	–
+DRKD[40]	41.6	24.2	45.3	55.3	40.3	23.4	44.2	53.4
+Ours	**41.7 (+3.8)**	**24.8(+2.4)**	44.9	54.2	**40.7 (+3.3)**	**24.6 (+1.5)**	44.3	52.1

metric, and the values in parentheses (e.g., +3.8) indicate the improvement over the student model. AP_S (average precision for small objects) measures how well the model detects small-sized objects. AP_M (average precision for medium objects) evaluates the detection performance for medium-sized objects. AP_L (average precision for large objects) assesses the accuracy for detecting large-sized objects. These metrics help understand how the model performs across different object scales. This is also the case for all the following results tables. The pre-trained models used in these experiments were sourced from the official release of Detectron2[⑦]. Specifically, the teacher model employed a ResNet101 backbone trained for 3x, while the student model utilized a ResNet50 backbone trained for 1x. Comparison with several advanced methods revealed that our approach showed a notable advantage. For Faster R-CNN, our method improved the mAP value by 3.8 over the baseline, with the AP_S (for objects smaller than 32×32) increasing by 2.4. This underscores our method's effectiveness in enhancing small object detection. The relatively poor performance of TinyKD[39] may be attributed to its design, which is tailored specifically for tiny person detection and may not be suitable for the broader COCO dataset scenarios. This limitation highlights the broader applicability of our method. Additionally, as shown in Table 2, we applied our method to instance segmentation (Mask R-CNN)[41] and FCOS[22], and the experimental results further demonstrated its effectiveness in these tasks. According to the results in Table 2, it is evident that anchor-free detectors, such as FCOS, are significantly influenced by knowledge distillation. HMKD compels the student to mimic the teacher's ability to extract and comprehend small object features. Meanwhile, FCOS continuously optimizes and progressively aligns with the ground truth based on anchors generated by the detection network. We believe this explains the substantial improvement observed in the FCOS detector, where the performance of the student model surpasses that of the teacher model.

Results of Heterogeneous Distillation Using Varied Backbone Networks. To further assess the versatility of our method, we substituted the teacher and student networks in Faster R-CNN with VoVNetV2. VoVNetV2, introduced by Lee *et al.*[42] in 2019, is an efficient backbone designed for real-time object detec-

Table 2. Results of Different Detectors

Detector	Setting	Type	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN	ResNet101(teacher) 3x	BBox	42.9	63.3	46.8	26.4	46.6	56.1
	ResNet50(student) 1x		38.6	59.5	42.1	22.5	42.0	49.9
	Ours		**41.0 (+2.4)**	61.5	45.0	**25.2 (+2.7)**	44.2	53.1
	ResNet101(teacher) 3x	Mask	38.6	60.4	41.3	19.5	41.3	55.3
	ResNet50(student) 1x		35.2	56.3	37.5	17.2	37.2	50.3
	Ours		**37.2 (+2.0)**	58.6	40.1	**19.3 (+2.1)**	40.0	53.2
FCOS	ResNet101(teacher) 3x	BBox	43.2	62.4	46.8	26.1	46.2	52.8
	ResNet50(student) 1x		38.6	57.4	41.4	22.3	42.5	49.8
	Ours		**43.6 (+5.0)**	62.3	47.3	**27.4 (+5.1)**	47.5	55.6

Note: Results are reported for bounding boxes (BBox) and instance masks (Mask), respectively.

[⑦]https://github.com/facebookresearch/detectron2, Jul. 2024.

tion, optimizing GPU computational efficiency. Thus, we used VoVNetV2 to replace ResNet in our experiments. The pre-trained models used are from the official release: VoVNetV2-57 (3x) for the teacher model and VoVNetV2-19 (1x) for the student model, with all the other settings remaining unchanged. As shown in Table 3, the experimental results confirm that our method effectively performs heterogeneous network distillation. However, it is notable that the detection performance of the student model does not surpass the original results even with a more powerful teacher model. This may be attributed to differences in network architectures.

Results with Different Mobile Networks. The primary goal of using a lightweight network is to deploy the model on devices with limited computing power, and mobile networks are commonly used for this purpose. Therefore, we employed MobileNetV2[43], a well-established lightweight network, in our experiment. We used ResNet101 (3x) as the teacher model and MobileNetV2 (1x) as the student model. The teacher model was sourced from the official release in Detectron2, while the student model was trained using our setup. Our experimental results, presented in Table 3, compare the performance of students across different frameworks.

Results on Dataset VisDrone. The experimental results, obtained from the validation set, are presented in Table 4. These results demonstrate that our proposed method is effective in enhancing the features of small objects and outperforms the baseline as well as other methods in detection capabilities. Notably, the detection accuracy of our student model even exceeds that of the teacher model. Hyper-parameters were set identically to those used for the COCO experiments. Additionally, Fig.4 provides a comparison of visualization results on the VisDrone dataset, illustrating that our method significantly improves the detection of small targets, particularly those at the far end of the image.

4.4 Ablation Studies

Effect of Supplemental Distillation. The aim of separating the foreground from the background is to mitigate the impact of background noise on the detection of small objects. However, background information also contains valuable context that can enhance model learning. Therefore, we introduce a supplemental distillation module to impart background knowledge to the student model, enriching its knowledge framework. To assess the effectiveness of this module, we conducted a series of experiments, as detailed in Table 5. The results reveal a significant decline in performance when the supplemental distillation module is omitted, underscoring its essential role within the HMKD framework.

Number of Heads in the Decoder. The number of heads in the decoder is a crucial factor influencing detection performance. Heads balance the dimensions and spaces within the subspace. Our experiments indicate that the optimal number of heads remains around 8, as shown in Table 6, which aligns with the original number of probes.

Hierarchical Matching Starting Layer Analysis. In the hierarchical matching process, selecting the appropriate starting layer for the query is crucial. We conducted a series of ablation experiments to determine the optimal starting layer, whose results are detailed in Table 7. The experiments show that starting the query at either P2 or P3 yields the best detection performance for small targets, achieving an accuracy of

Table 3. Results with Different Backbone Networks

Detector	Setting	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	ResNet101(teacher) 3x	42.0	62.5	45.9	25.2	45.6	54.6
	VoVNetV2-19(student) 1x	32.0	51.4	34.0	18.4	34.4	40.8
	Ours	**36.4 (+4.4)**	56.8	39.1	**21.6 (+3.2)**	39.1	46.3
Faster R-CNN	VoVNetV2-57(teacher) 3x	43.3	64.3	47.0	27.5	46.7	55.3
	ResNet50(student) 1x	37.9	58.8	41.1	22.4	41.1	49.1
	Ours	**40.8 (+2.9)**	61.4	44.7	**24.2 (+1.8)**	44.2	53.3
Faster R-CNN	ResNet101(teacher) 3x	42.0	62.5	45.9	25.2	45.4	54.6
	MobileNetV2(student) 1x	26.4	45.0	27.2	14.9	28.6	33.2
	Ours	**29.8 (+3.4)**	47.7	31.9	**16.7 (+1.8)**	32.0	38.8
RetinaNet	ResNet101(teacher) 3x	40.4	60.3	43.2	24.0	44.3	52.2
	MobileNetV2(student) 1x	20.4	33.3	21.6	10.7	22.1	26.1
	Ours	**23.4 (+3.0)**	37.2	24.7	**13.4 (+2.7)**	25.0	29.2

Table 4. Results on Dataset VisDrone Using RetinaNet

Method	mAP	AP_S	AP_M	AP_L
Resnet101(teacher) 3x	23.8	14.0	36.3	58.0
Resnet50(student) 1x	20.6	11.5	31.9	55.0
+ICD[14]	23.6	13.9	36.1	56.5
+FGD[17]	22.9	11.8	37.3	57.5
+Ours	**24.0 (+3.4)**	**14.2 (+2.7)**	36.4	56.6

Note: Both the teacher and student are trained by ourselves.

24.8. Given the actual training time, using P3 as the starting layer proves more efficient than P2. Thus, we conclude that setting the starting layer to P3 is optimal. Starting the query from lower-resolution layers (e.g., P4 or P5) results in degraded performance for small target detection due to the challenge of distinguishing small targets on very low-resolution feature maps. The experiments were conducted using the COCO dataset and the Faster R-CNN detector.

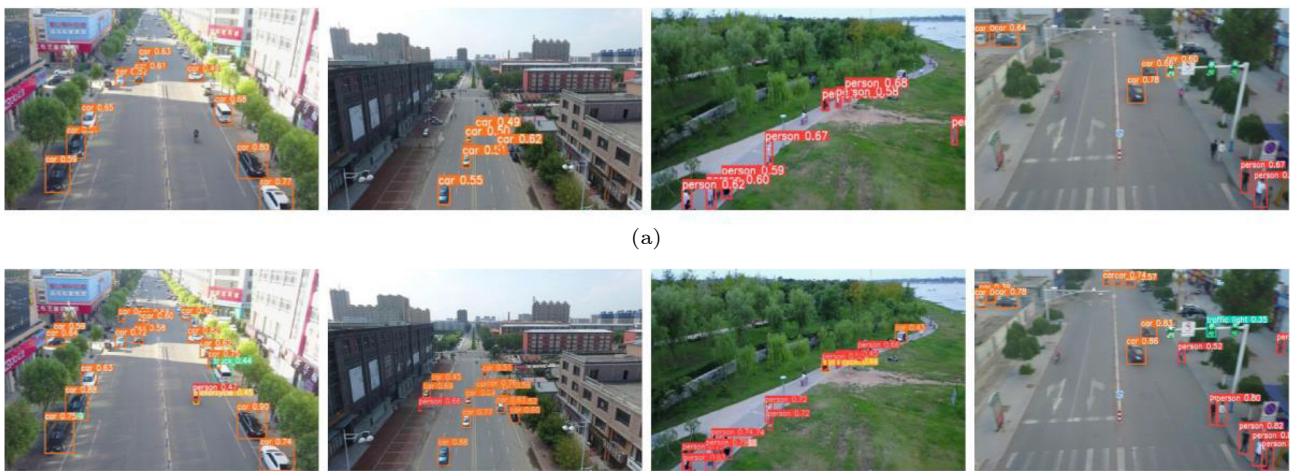
Speed Test. In addition to detection accuracy, model speed is also a crucial factor in evaluating model quality. We specifically measured the FPS (frames per second) for ResNet50 and MobileNetV2 using Faster R-CNN and RetinaNet, respectively. The results are detailed in Table 8. Faster R-CNN with MobileNetV2 demonstrated superior detection speed, meeting the requirements for real-time detection.

Model Analysis. The computational power of a model is typically measured in giga-floating point operations per second (GFLOPS), which reflects the amount of computational work the model can perform per second. As shown in Fig.5, the MobileNet model in the Faster R-CNN network does not achieve significant improvements in lightweight performance. Although MobileNet has the same number of model parameters as ResNet50, its computational power is lower. This indicates that MobileNet may not be the most effective option for lightweighting within the Faster R-CNN framework. In contrast, the RetinaNet network demonstrates a more significant reduction in computational power with a lighter model, suggesting that one-stage models might be better suited for deployment on edge devices.

5 Conclusions

We noted that it is not easy to transfer small object knowledge to students during knowledge distillation. Therefore, we designed Hierarchical Matching Knowledge Distillation Network (HMKD) to enhance students' knowledge learning of small objects. We encoded high-semantic information at low-resolution of the FPN (feature pyramid network) as inquiries, and represented fine-grained graph feature values at high-resolution as key-values. Extensive experiments demonstrated that our method effectively enhances the student's understanding of small objects detection capability and is suitable for mainstream object detectors or instance segmentation models. The training time increases due to the additional augmentation design for small objects, which we will optimize in the future work. At the same time, we will investigate differences in knowledge extraction or learning effectiveness between teachers and students.

Conflict of Interest The authors declare that they have no conflict of interest.



(a)

(b)

Fig.4. Visualization of detection results on the dataset VisDrone. (a) Without knowledge distillation. (b) Using the HMKD Method. The HMKD method shows improved detection of distant small objects.

Table 5. Ablation Study of the Supplemental Distillation Module

Separate+Supplement	mAP	AP_S	AP_M	AP_L
×	40.6	24.3	43.9	53.1
√	41.7	24.8	44.9	54.2

Table 6. Effect of Different Numbers of Heads

Number of Heads	mAP	AP_S	AP_M	AP_L
1	38.4	23.0	41.6	49.9
4	40.0	23.8	44.4	51.3
8	40.7	24.6	44.3	52.1
16	39.8	23.5	43.9	50.9

Table 7. Ablation Study Results of the Starting Layer

Starting Layer	mAP	AP_S	AP_M	AP_L
None	37.9	22.4	41.1	49.1
P5	40.9	23.4	44.6	54.3
P4	41.3	24.1	44.7	54.3
P3	41.7	24.8	44.9	54.2
P2	41.6	24.8	44.8	53.9

Table 8. Model Detection Speed

Method	FPS
Faster R-CNN + ResNet50	14.77
RetinaNet + ResNet50	13.00
Mask R-CNN + ResNet50	11.49
Faster R-CNN + MobileNetV2	21.98
RetinaNet + MobileNetV2	17.36

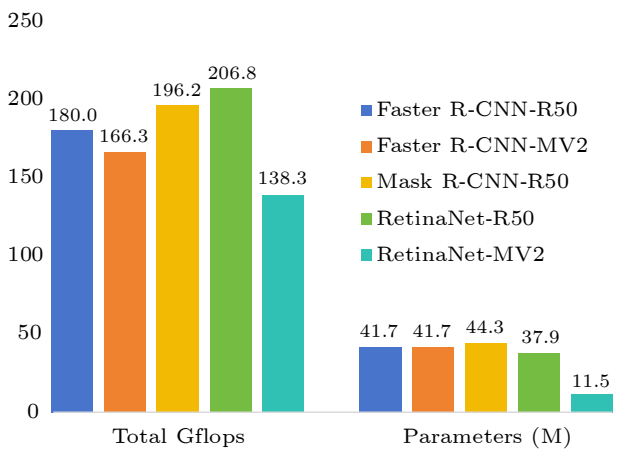


Fig.5. Gflops and parameters of different methods.

References

[1] Cao C, Wang B, Zhang W, Zeng X, Yan X, Feng Z, Liu Y, Wu Z. An improved faster R-CNN for small object detection. *IEEE Access*, 2019, 7: 106838–106846. DOI: 10.1109/ACCESS.2019.2932731.

[2] Yang C, Huang Z, Wang N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.13658–13667. DOI: 10.1109/CVPR52688.2022.01330.

[3] Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. Binarized neural networks. In *Proc. the 30th International Conference on Neural Information Processing Systems*, Dec. 2016, pp.4114–4122.

[4] Rastegari M, Ordonez V, Redmon J, Farhadi A. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.525–542. DOI: 10.1007/978-3-319-46493-0_32.

[5] Han S, Pool J, Tran J, Dally W J. Learning both weights and connections for efficient neural network. In *Proc. the 28th International Conference on Neural Information Processing Systems*, Dec. 2015, pp.1135–1143.

[6] He Y, Zhang X, Sun J. Channel pruning for accelerating very deep neural networks. In *Proc. the 2017 IEEE International Conference on Computer Vision*, Oct. 2017, pp.1398–1406. DOI: 10.1109/ICCV.2017.155.

[7] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv: 1503.02531, 2015. https://arxiv.org/abs/1503.02531, Jul. 2024.

[8] Ji M, Heo B, Park S. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proc. the 35th AAAI Conference on Artificial Intelligence*, Feb. 2021, pp.7945–7952. DOI: 10.1609/aaai.v35i9.16969.

[9] Wang T, Yuan L, Zhang X, Feng J. Distilling object detectors with fine-grained feature imitation. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.4928–4937. DOI: 10.1109/CVPR.2019.00507.

[10] Zhang L, Ma K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *Proc. the 9th International Conference on Learning Representations*, May 2021.

[11] Heo B, Kim J, Yun S, Park H, Kwak N, Choi J Y. A comprehensive overhaul of feature distillation. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27-Nov. 2, 2019, pp.1921–1930. DOI: 10.1109/ICCV.2019.00201.

[12] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. the 28th International Conference on Neural Information Processing Systems*, Dec. 2015, pp.91–99.

[13] Lin T Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In *Proc. the 2017 IEEE conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.936–944. DOI: 10.1109/CVPR.2017.106.

[14] Kang Z, Zhang P, Zhang X, Sun J, Zheng N. Instance-conditional knowledge distillation for object detection. In *Proc. the 35th International Conference on Neural Information Processing Systems*, Dec. 2021, Article No. 1259.

[15] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. the 2017 IEEE International Conference on Computer Vision*, Oct. 2017, pp.618–626. DOI: 10.1109/ICCV.2017.74.

[16] Cao Y, Xu J, Lin S, Wei F, Hu H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In

Proc. the 2019 IEEE/CVF International Conference on Computer Vision workshop, Oct. 2019, pp.1971–1980. DOI: 10.1109/ICCVW.2019.00246.

[17] Yang Z, Li Z, Jiang X, Gong Y, Yuan Z, Zhao D, Yuan C. Focal and global knowledge distillation for detectors. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.4633–4642. DOI: 10.1109/CVPR52688.2022.00460.

[18] Chen G, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation. In *Proc. the 31st International Conference on Neural Information Processing Systems*, Dec. 2017, pp.742–751.

[19] Guo J, Han K, Wang Y, Wu H, Chen X, Xu C, Xu C. Distilling object detectors via decoupled features. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.2154–2164. DOI: 10.1109/CVPR46437.2021.00219.

[20] Chang J, Wang S, Xu H M, Chen Z, Yang C, Zhao F. DETRDistill: A universal knowledge distillation framework for DETR-families. In *Proc. the 2023 IEEE/CVF International Conference on Computer Vision*, Oct. 2023, pp.6875–6885. DOI: 10.1109/ICCV51070.2023.00635.

[21] Lin T Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2020, 42(2): 318–327. DOI: 10.1109/TPAMI.2018.2858826.

[22] Tian Z, Shen C, Chen H, He T. FCOS: Fully convolutional one-stage object detection. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27-Nov. 2, 2019, pp.9627–9636. DOI: 10.1109/ICCV.2019.00972.

[23] Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO series in 2021. arXiv: 2107.08430, 2021. https://arxiv.org/abs/2107.08430, Jul. 2024.

[24] Huang H, Zhou X, Cao J, He R, Tan T. Vision transformer with super token sampling. arXiv: 2211.11167, 2024. https://arxiv.org/abs/2211.11167, Jul. 2024.

[25] Zhu L, Wang X, Ke Z, Zhang W, Lau R. BiFormer: Vision transformer with bi-level routing attention. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp.10323–10333. DOI: 10.1109/CVPR52729.2023.00995.

[26] Tian R, Wu Z, Dai Q, Hu H, Qiao Y, Jiang Y G. ResFormer: Scaling ViTs with multi-resolution training. In *Proc. the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2023, pp.22721–22731. DOI: 10.1109/CVPR52729.2023.02176.

[27] Kisantal M, Wojna Z, Murawski J, Naruniec J, Cho K. Augmentation for small object detection. arXiv: 1902.07296, 2019. https://arxiv.org/abs/1902.07296, Jul. 2024.

[28] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: Single shot MultiBox detector. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.21–37. DOI: 10.1007/978-3-319-46448-0_2.

[29] Cai Z, Fan Q, Feris R S, Vasconcelos N. A unified multi-scale deep convolutional neural network for fast object detection. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.354–370. DOI: 10.1007/978-3-319-46493-0_22.

[30] Kong T, Yao A, Chen Y, Sun F. HyperNet: Towards accurate region proposal generation and joint object detection. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.845–853. DOI: 10.1109/CVPR.2016.98.

[31] Li Y, Chen Y, Wang N, Zhang Z X. Scale-aware trident networks for object detection. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27-Nov. 2, 2019, pp.6054–6063. DOI: 10.1109/ICCV.2019.00615.

[32] Singh B, Davis L S. An analysis of scale invariance in object detection—SNIP. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.3578–3587. DOI: 10.1109/CVPR.2018.00377.

[33] Singh B, Najibi M, Davis L S. SNIPER: Efficient multi-scale training. In *Proc. the 32nd International Conference on Neural Information Processing Systems*, Dec. 2018, pp.9333–9343.

[34] Chen Y, Zhang P, Li Z, Li Y, Zhang X, Qi L, Sun J, Jia J. Dynamic scale training for object detection. arXiv: 2004.12432, 2021. https://arxiv.org/abs/2004.12432, Jul. 2024.

[35] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.213–229. DOI: 10.1007/978-3-030-58452-8_13.

[36] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In *Proc. the 31st International Conference on Neural Information Processing Systems*, Dec. 2017, pp.6000–6010.

[37] Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. In *Proc. the 3rd International Conference on Learning Representations*, May 2015. DOI: 10.48550/arXiv.1412.6550.

[38] Loshchilov I, Hutter F. Decoupled weight decay regularization. In *Proc. the 7th International Conference on Learning Representations*, May 2017.

[39] Liu H, Liu Q, Liu Y, Liang Y, Zhao G. Exploring effective knowledge distillation for tiny object detection. In *Proc. the 2023 IEEE International Conference on Image Processing*, Oct. 2023, pp.770–774. DOI: 10.1109/ICIP49359.2023.10222589.

[40] Ni Z L, Yang F, Wen S, Zhang G. Dual relation knowledge distillation for object detection. In *Proc. the 32nd International Joint Conference on Artificial Intelligence*, Aug. 2023, pp.1276–1284. DOI: 10.24963/ijcai.2023/142.

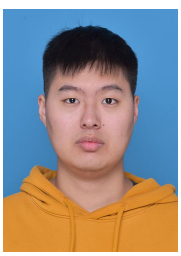
[41] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In *Proc. the 2017 IEEE International Conference on Computer Vision*, Oct. 2017, pp.2980–2988. DOI: 10.1109/ICCV.2017.322.

[42] Lee Y, Hwang J W, Lee S, Bae Y, Park J. An energy and GPU-computation efficient backbone network for real-time object detection. In *Proc. the 2019 IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition Workshops, Jun. 2019, pp.752–760. DOI: 10.1109/CVPRW.2019.00103.

[43] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proc. the 2018 IEEE/CVF conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.4510–4520. DOI: 10.1109/CVPR.2018.00474.

Yong-Chi Ma received his B.S. degree and M.S. degree from the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, in 2020 and 2024, respectively, both in computer science and technology. His research interests are computer vision and knowledge distillation.



Xiao Ma received his B.S. degree in computer science and technology from Henan University, Kaifeng, in 2022. Currently, he is now a Master student in the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou. His current research interests mainly focus on object detection and knowledge distillation.

Tian-Ran Hao received his B.S. degree from the Information Engineering Institute, Zhengzhou University, Zhengzhou, in 2017. He is currently pursuing his Ph.D. degree in the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou. His current research interests include object detection and computer vision.

Li-Sha Cui received her Ph.D. degree in software engineering from Zhengzhou University, Zhengzhou, in 2020. She is currently an associate professor with the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou. Her current research interests include object detection, deep learning, and computer vision.

Shao-Hui Jin received her Ph.D. degree in control science and engineering from Xidian University, Xi'an, in 2016. She now works at the School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou. Her current research interests include image processing, computer vision, and non-line of sight imaging.

Pei Lyu received his Ph.D. degree from State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, in 2013. He is currently a full professor with the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou. His research interests include computer vision and computer graphics.