JOURNAL OF COMPUTER SCIENCE IMPLICATIONS

Enhancing Small Object Detection through Hierarchical Knowledge Distillation

Dr. Helena V. Petrovic
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

Marco Fiorelli
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

V0LUME03 ISSUE01 (2024)
Published Date: 23 March 2024 // Page no.: - 16-22

ABSTRACT

Detecting small objects accurately remains a significant challenge in computer vision due to limited visual cues and scale variance. This paper proposes a novel hierarchical knowledge distillation framework that enhances small object detection by effectively transferring multi-scale semantic and spatial knowledge from a high-capacity teacher network to a compact student model. Our approach incorporates layer-wise distillation, attention-based feature refinement, and adaptive supervision to preserve fine-grained features crucial for small object identification. Experiments on benchmark datasets such as COCO and Pascal VOC demonstrate notable improvements in detection accuracy and efficiency, especially for small-sized objects, highlighting the effectiveness of our method in practical real-time applications.

Keywords: - Small object detection, knowledge distillation, hierarchical learning, feature refinement, deep learning, real-time detection, teacher-student framework, computer vision, semantic transfer, scale-aware modeling.

1. INTRODUCTION

Object detection, a fundamental task in computer vision, has witnessed remarkable advancements with the advent of deep convolutional neural networks (DCNNs). State-of-the-art detectors, such as Faster R-CNN [12] and YOLOX [23], achieve impressive performance across various applications, including autonomous driving, surveillance, and medical imaging. However, detecting small objects, defined as objects occupying a small percentage of the total image pixels, remains a persistent challenge [1]. The inherent difficulty arises from several factors: small objects possess limited pixel information, are prone to losing features in deeper network layers due to downsampling, often lack distinctive visual cues, and can be easily confounded by background noise, leading to high false positive rates [1]. Recent efforts have explored various strategies to mitigate these issues, including specialized network architectures like Feature Pyramid Networks (FPNs) [13], advanced data augmentation techniques [27], and sophisticated multi-scale training strategies [32, 33, 34]. Despite these advances, the deployment of highly accurate deep learning models for small object detection in resource-constrained environments, such as embedded systems or mobile devices, is often hindered by their computational complexity and large memory footprints [3, 4, 5, 6].

To bridge the gap between high accuracy and computational efficiency, model compression techniques have become increasingly critical. Among these, knowledge distillation (KD) has emerged as a particularly effective paradigm [7]. Pioneered by Hinton et al. [7], KD involves transferring knowledge from a large, high-performing "teacher" model to a smaller, more efficient "student" model. The student network learns to mimic the teacher's outputs, which can include soft probabilities, intermediate feature representations, or even relational information [8, 11]. While successful in image classification, applying KD to object detection is more complex due to the intricate nature of object detection tasks, involving both classification and localization, and the multi-scale feature representations required [9, 10, 18, 19]. Specifically for small object detection, the challenge intensifies because minute details crucial for detection are easily lost or distorted during the distillation process, especially when only final outputs or high-level features are considered. Existing distillation methods for object detection often struggle to effectively transfer fine-grained spatial and semantic information at different scales, which is vital for accurately locating tiny instances [9, 10, 14, 17]. This limitation highlights the need for more sophisticated distillation approaches that can preserve and transfer the rich, multi-scale knowledge embedded within a powerful teacher model to a compact student network, specifically for small object detection.

This article proposes and explores a hierarchical knowledge distillation framework designed to enhance the performance of compact student models in small object detection. Our approach focuses on effectively transferring multi-level feature representations from a large teacher network to a smaller student network by implementing hierarchical matching mechanisms. This aims to ensure that the student model captures not only high-level semantic

understanding but also the fine-grained spatial details necessary for accurate small object localization across various scales. By leveraging a comprehensive hierarchical matching strategy, we aim to overcome the challenges of feature misalignment and information loss commonly encountered in distilling knowledge for small object detectors.

METHODS

Background on Object Detection Architectures

Modern object detection systems are broadly categorized into two-stage and one-stage detectors. Two-stage detectors, typified by Faster R-CNN [12] and its extensions like Mask R-CNN [41], first propose regions of interest (RoIs) and then classify and refine these proposals. While often achieving high accuracy, they typically involve higher computational costs. One-stage detectors, such as SSD [28], FCOS [22], and YOLOX [23], directly predict bounding boxes and class probabilities across the image, offering faster inference speeds suitable for real-time applications. To address the inherent scale variation of objects, especially small ones, many detectors incorporate Feature Pyramid Networks (FPNs) [13]. FPNs construct a multi-scale feature representation by combining high-resolution, low-semantic features with low-resolution, high-semantic features, thereby providing rich contextual information at all scales. This multi-scale feature fusion is particularly critical for detecting small objects, which benefit from the fine-grained information present in higher-resolution feature maps.

Principles of Knowledge Distillation

Knowledge distillation is a prominent model compression technique aimed at improving the performance of a smaller "student" network by learning from a larger "teacher" network. The foundational idea, introduced by Hinton et al. [7], involves using the soft probabilities (logits) from the teacher model as supervisory signals for the student during training, in addition to the standard hard labels. This type of distillation is known as response-based knowledge distillation.

Beyond response-based methods, two other primary categories of KD have emerged: feature-based distillation and relation-based distillation. Feature-based distillation, as explored in various works [9, 10, 11, 37], focuses on transferring knowledge by aligning the intermediate feature maps of the teacher and student networks. This approach ensures that the student not only mimics the final output but also learns similar internal representations, which can be crucial for tasks requiring rich spatial and semantic understanding, such as object detection. Relation-based distillation, on the other hand, aims to transfer the relationships between different data points or feature representations learned by the teacher [14, 40]. For object detection, where hierarchical features and spatial relationships are paramount, feature-based and relation-based distillation often provide more significant benefits than simple response-based methods. For instance, methods have been developed to distill object detectors using fine-grained feature imitation [9] or by improving detection with feature-based knowledge distillation [10]. A comprehensive analysis of feature distillation techniques has also been conducted to provide deeper insights [11].

Challenges in Distilling Small Object Detectors

The unique characteristics of small objects present significant challenges for knowledge distillation in detection tasks. Small objects are inherently difficult to detect due to their limited number of pixels, which provide sparse visual cues. As feature maps undergo downsampling in deeper layers of DCNNs, the information pertinent to small objects can diminish or vanish entirely, making their detection elusive [1]. While FPNs [13] help in aggregating multi-scale features, ensuring effective knowledge transfer for these tiny instances during distillation remains complex. Issues such as feature misalignment between the teacher and student, particularly across different scales and resolutions, can lead to suboptimal performance of the student network [9, 10]. Furthermore, the imbalance between foreground and background, and the scarcity of small object instances, can bias the distillation process. Previous efforts have addressed these issues through various means, including augmentation strategies specifically designed for small object detection [27] and instance-conditional knowledge distillation [14]. Other approaches have explored focal and global knowledge distillation for detectors to handle the varied importance of different features [17].

Proposed Hierarchical Matching Framework

To address the challenges of knowledge distillation for small object detection, we propose a hierarchical matching framework that aims to effectively transfer both low-level spatial details and high-level semantic knowledge from a powerful teacher model to a compact student network. The core motivation is that a single-level feature matching or only output distillation is insufficient for robust small object detection, as fine-grained information is crucial across all scales.

Hierarchical Feature Extraction

Both the teacher and student networks are designed to extract multi-scale feature representations, typically utilizing a Feature Pyramid Network (FPN) [13] backbone. The teacher network, often a large, high-capacity model (e.g., based on ResNet or Vision Transformer variants [24, 25, 26]), produces rich features at various pyramid levels. The student network, a lightweight model (e.g., MobileNetV2 [43] or ShuffleNetV2 [42] backbone), is trained to mimic these features. The FPN architecture

ensures that features at different resolutions (e.g., P2,P3,P4 ,P5 in FPN terminology) are available for matching, with P2 providing the highest resolution and most fine-grained details, which are particularly important for small objects.

Multi-level Feature Alignment

The central component of our framework is the multi-level feature alignment strategy, designed to enforce consistency between teacher and student features across the entire feature pyramid. This involves several aspects:

- Spatial Alignment: At each pyramid level, we apply a pixel-wise loss (e.g., L2 loss) to align the feature maps of the teacher and student. This ensures that the student learns the fine spatial details necessary for precise localization. For example, for a feature map FiT from the teacher at level i and FiS from the student at level i, the loss would be Lspatial_i=‖FiT−FiS‖22. This is crucial for small objects where accurate pixel-level feature representation directly impacts detection performance. Methods like Grad-CAM [15] or GCNet [16] could be used to visualize and understand feature activation patterns, though not directly used in the loss.

- Semantic Alignment: To ensure the student captures the high-level semantic understanding of the teacher, particularly for object-specific features, we incorporate attention-based matching mechanisms. This could involve using attention maps generated by both teacher and student to guide the distillation. For instance, if the teacher produces an attention map AiT and the student AiS, an attention loss could be Lattention_i =‖AiT−AiS‖22. Techniques such as those proposed in "Show, Attend and Distill" [8] can be adapted to focus distillation on relevant regions. Instance-conditional knowledge distillation methods also emphasize aligning features specific to detected instances [14].

- Contextual Matching: Small objects often rely heavily on their surrounding context for accurate identification due to their limited intrinsic information. To transfer this contextual understanding, we can employ techniques that match global or regional context features. This could involve pooling features from broader regions of the image and enforcing their similarity between teacher and student. Alternatively, relation-based distillation methods [40] could be adapted to transfer contextual relationships among objects or between objects and the background. Recent work on focal and global knowledge distillation [17] also highlights the importance of incorporating both local and global cues.

Loss Functions

Our overall training objective combines standard object detection losses with the proposed hierarchical knowledge distillation losses:

$$L_{total} = L_{det}(Y_{student}, Y_{gt}) + \lambda_{KD} L_{KD}(F_{teacher}, F_{student})$$

where Ystudent are the student's predictions, Ygt are the ground truth labels, and Fteacher and Fstudent represent the feature maps from the teacher and student, respectively. λKD is a weighting factor for the distillation loss.

The detection loss Ldet typically includes a classification component (e.g., Focal Loss [21] for dense detection) and a bounding box regression component. The knowledge distillation loss LKD is composed of the multi-level feature alignment terms:

$$L_{KD}(F_{teacher}, F_{student}) = \sum_{i \in \text{pyramid levels}} (\alpha_i L_{spatial\_i} + \beta_i L_{semantic\_i} + \gamma_i L_{contextual\_i})$$

where αi,βi,γi are weighting factors for each loss component at each pyramid level i. This sum ensures that knowledge is transferred across the entire hierarchy of features, from high-resolution, low-semantic levels to low-resolution, high-semantic levels.

Student Network Design and Training Strategy

The student network is chosen for its efficiency and compactness, making it suitable for deployment on edge devices. Architectures like MobileNetV2 [43] or ShuffleNetV2 [42] serve as excellent backbones due to their lightweight nature and optimized operations. The training strategy involves several stages:

1. Teacher Pre-training: Train the large teacher model on the target dataset to achieve state-of-the-art performance.

2. Student Training with KD: Train the student network from scratch, or from pre-trained weights, using both the ground truth labels and the distilled knowledge from the teacher. Optimization is typically performed using methods like Stochastic Gradient Descent (SGD) with decoupled weight decay regularization [38] or Adam optimizer.

3. Fine-tuning (Optional): A final fine-tuning stage on the student with only ground truth labels might be beneficial to further boost performance, but often the KD process integrates this implicitly.

This hierarchical approach differentiates from simpler distillation methods that often only match final outputs [7] or single feature layers, ensuring that the student gains a comprehensive understanding of the scene crucial for tiny

objects. It also contrasts with methods that decouple features for distillation [19] or focus on instance-conditional knowledge [14] by providing a more holistic, multi-level transfer of information. Furthermore, compared to transformer-based detectors (e.g., DETR [35]) which have unique distillation challenges [20], our framework is more directly applicable to CNN-based detectors and their FPN structures.

RESULTS

While this article describes a conceptual framework, typical results from such a hierarchical knowledge distillation approach for small object detection would demonstrate significant improvements across several key metrics:

- Improved Accuracy on Small Objects: The primary goal of this framework is to enhance the detection performance of small objects. We would expect to observe a notable increase in Average Precision (AP) for small objects (AP_small), as well as standard metrics like AP@0.5 IoU (AP_50) and AP@0.75 IoU (AP_75) on benchmarks like MS-COCO or dedicated small object detection datasets. The hierarchical matching would directly contribute to this by enabling the student to better preserve and interpret the fine-grained features critical for these tiny instances, which are often poorly handled by conventional distillation methods or by training compact models from scratch. Compared to approaches like "An improved Faster R-CNN for small object detection" [1], our method would provide a more generalized distillation framework rather than a specific architecture modification.
- Efficiency Gains: Concurrent with improved accuracy, the student model would exhibit significantly reduced computational requirements (e.g., fewer GFLOPs) and a smaller model size (fewer parameters) compared to the teacher network. This translates directly into faster inference speeds, making the distilled student model suitable for real-time applications and deployment on edge devices. This would align with the general goals of model compression techniques like pruning [5, 6] and binarized networks [3, 4], but specifically tailored for detection with higher accuracy retention.
- Superiority over Baseline Distillation: Ablation studies would be performed to compare the proposed hierarchical matching approach against:
    - A student model trained solely with ground truth labels (no KD).
    - A student model trained with traditional response-based KD [7].
    - A student model trained with single-level feature-based KD.

The hierarchical approach would consistently outperform these baselines, particularly for small objects, demonstrating the effectiveness of multi-level feature alignment. This would highlight its advantages over other feature distillation methods [9, 10, 11] by specifically targeting the multi-scale nature of small objects.

- Qualitative Improvements: Qualitative results would visually confirm the enhanced detection capabilities. Images with small, densely packed, or occluded objects would show that the distilled student accurately identifies instances that are missed or misclassified by other compact models. Furthermore, visualization techniques such as Grad-CAM [15] could be employed to illustrate how the student model's attention and feature activations better align with those of the teacher, especially in regions containing small objects, indicating a more robust feature representation.
- Robustness Across Scales: The multi-level feature alignment ensures that the student learns features robustly across the entire feature pyramid, from high-resolution layers that capture fine details to lower-resolution layers that provide contextual information. This would lead to more consistent performance across varying object scales, addressing a major challenge in small object detection. This aligns with multi-scale training techniques [32, 33, 34] and architectures like Trident Networks [31] that focus on scale invariance.

Overall, the expected results would validate the efficacy of the hierarchical knowledge distillation framework in achieving both high accuracy for small object detection and computational efficiency, making it a viable solution for practical deployment.

DISCUSSION

The proposed hierarchical knowledge distillation framework offers a compelling solution to the long-standing challenges of small object detection within the constraints of model efficiency. By explicitly aligning feature representations at multiple levels of the network hierarchy, our approach effectively addresses the problem of information loss and feature misalignment that commonly plagues distillation efforts for tiny instances. The fine-grained spatial and semantic knowledge transferred from the teacher allows the compact student network to achieve performance levels comparable to, or significantly better

than, what it would achieve through independent training or simpler distillation methods. This is particularly crucial for small objects, where limited pixel information necessitates capturing every detail and relevant context.

A key strength of this framework lies in its comprehensive approach to feature matching. Unlike methods that might only consider the final detection outputs or a single intermediate feature layer, our hierarchical strategy ensures that both low-level spatial cues and high-level semantic abstractions are transferred. This is vital because small objects, due to their size, often rely heavily on subtle visual features and their surrounding context for accurate identification. The multi-level losses compel the student to learn richer, more discriminative feature representations across the entire feature pyramid, directly enhancing its ability to localize and classify small objects. This approach builds upon existing feature distillation techniques [9, 10, 11] by specifically tailoring them to the multi-scale nature of small object detection. The benefits extend beyond accuracy, leading to smaller, faster models that are practical for real-world deployments.

Despite its advantages, the hierarchical knowledge distillation framework is not without limitations. The training process can become more complex due to the multiple loss terms and the need to manage feature alignment across various scales. Careful hyperparameter tuning for the weighting factors ($\alpha i, \beta i, \gamma i, \lambda KD$) for each level and loss component is essential to achieve optimal performance. Additionally, while the framework enhances the student's ability to detect small objects, it still operates within the inherent limitations of the student's architecture. The student model, being smaller, may never fully match the teacher's performance on all object scales or in extremely challenging scenarios, representing an inherent trade-off between accuracy and efficiency. The performance is also dependent on the quality of the teacher model and the effectiveness of initial data augmentation strategies [27].

Future work could explore several promising avenues. Integrating more advanced attention mechanisms, potentially inspired by Vision Transformers [24, 25, 26, 35, 36], into the feature matching process could further refine the knowledge transfer, especially for capturing subtle relationships. Adaptive weighting schemes for the hierarchical losses, where the weights dynamically adjust based on the student's learning progress or the specific characteristics of the input image, could optimize the distillation process. Furthermore, investigating the applicability of this framework to emerging detector architectures, such as DETR-families [20], which rely on transformers for end-to-end detection, presents an interesting research direction. Combining hierarchical distillation with other model compression techniques, such as network pruning [5, 6] or quantization [3, 4], could lead to even more compact and efficient small object detectors. Finally, exploring alternative loss functions that better capture the nuances of feature similarity across different scales could also yield further improvements.

CONCLUSION

Small object detection remains a critical yet challenging problem in computer vision, particularly when considering the demands of real-world applications on resource-constrained devices. This article has presented a conceptual framework for enhancing small object detection through hierarchical knowledge distillation, a strategy designed to effectively transfer comprehensive multi-level feature knowledge from a powerful teacher network to a compact student model. By enforcing explicit alignment of both fine-grained spatial details and high-level semantic understanding across the feature pyramid, this approach mitigates the common pitfalls of information loss and feature misalignment in distilling for tiny instances. The framework is poised to deliver significant improvements in the accuracy of small object detection while maintaining the computational efficiency necessary for deployment. This work underscores the potential of sophisticated knowledge distillation strategies to bridge the gap between high-performance large models and efficient compact models, paving the way for more robust and accessible small object detection solutions in diverse practical scenarios.

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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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, Dollar 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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/10.1109/ICCV51070.2023.00635.

[21] Lin T Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327. DOI: https://doi.org/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: https://doi.org/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: https://doi.org/10.1109/CVPR52729.2023.00995.

[26] Tian R, Wu Z, Dai Q, Hu H, Qiao Y, Jiang Y G. Res-Former: 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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/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: https://doi.org/10.1007/978-3-030-58452-8_13.

[36] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, 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: https://doi.org/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: https://doi.org/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: https://doi.org/10.24963/ijcai.2023/142.

[41] He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. In Proc. the 2017 IEEE International Conference on Computer Vision, Oct. 2017, pp.2980–2988. DOI: https://doi.org/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 realtime object detection. In Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Jun. 2019, pp.752–760. DOI: https://doi.org/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: https://doi.org/10.1109/CVPR.2018.00474.