Mech. Sci., 16, 877–886, 2025
https://doi.org/10.5194/ms-16-877-2025

Mechanical Sciences
Open Access

A lightweight optimization framework for real-time pedestrian detection in dense and occluded scenes

Cui Chen[1], Jun Li[1,2], Zequn Shuai[2], Yiyun Wang[1], and Yaohong Wang[3]

[1]Chongqing Vocational and Technical University of Mechatronics, Chongqing 402760, China
[2]School of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China
[3]Chongqing Academy of Metrology and Quality Inspection, Chongqing 400020, China

Correspondence: Cui Chen (78935883@qq.com) and Jun Li (cqleejun@cqjtu.edu.cn)

Received: 22 July 2025 – Revised: 5 October 2025 – Accepted: 10 October 2025 – Published: 21 November 2025

Abstract. Pedestrian detection is one of the most widely applied tasks in industrial computer vision. It encapsulates three core challenges of object detection: detecting small objects, handling heavy occlusion, and balancing speed and accuracy for deployment on mobile devices. In targeting scenarios relating to the Internet of Things (IoT), we propose a dedicated lightweight pedestrian detector that is robust to occlusions. First, we redesign the decoupled prediction head with a hierarchical structure, separating classification confidence estimation from bounding box regression. We then decode the offsets from the regression branch, extract features from high-confidence predictions, and fuse these with classification feature maps to enhance the local reliability of semantic features. Furthermore, we introduce a label-dynamic matching strategy that increases the number of high-quality positive samples, particularly improving matching for small and occluded objects. Finally, an optimized knowledge distillation framework significantly boosts the prediction accuracy of the compact model, facilitating deployment on edge devices. Experimental results on the CrowdHuman test set show that our proposed approach achieves comparable accuracy to the baseline (53.8 %) with an inference latency of only 7.1 ms – 281.7 % faster than the baseline.

1 Introduction

Pedestrian detection is a fundamental yet challenging task in computer vision and object detection (Zaidi et al., 2022). Unlike generic object categories, pedestrians exhibit high intra-class similarity, dynamic poses, dense spatial layouts, and frequent occlusions, and often appear as small-scale objects. These challenges are especially prominent in real-world applications such as urban surveillance (Mo et al., 2023; Choi and Kim, 2023), intelligent transportation, and public safety monitoring (Hamzenejadi and Mohseno, 2023; Zhu and Ji, 2005; Dong et al., 2023), where detection failures can have serious consequences.

Despite significant progress in object detection, most modern detectors are general purpose and struggle with pedestrian-centric scenarios (Chowdhury et al., 2018). Two-stage methods like R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2016) achieve high accuracy through powerful backbones and deep features, but their complex pipelines and high latency hinder real-time applications. To address efficiency, one-stage detectors such as SSD (Liu et al., 2016) and YOLO (Redmon et al., 2016) unify detection into a single forward pass. DETR (Carion et al., 2020) further proposed a transformer-based end-to-end paradigm that removes handcrafted anchors and post-processing, although at the cost of slow convergence and high computational demands.

YOLO-based models have gained popularity for their speed and simplicity. Over successive versions, the YOLO family has integrated components like CSPNet (Wang et al., 2020), new activation functions (Li and Huang, 2024), advanced augmentation (Lu et al., 2018), and improved loss functions (He et al., 2015), culminating in YOLOv5 (Jocher, 2020), which has become a strong baseline widely deployed in both academia and industry. Although later versions introduce additional modules, particularly transformer- or attention-based operators, such designs are not well supported on many industrial edge-computing devices (Khanam

and Hussain, 2025). Therefore, we adopt YOLOv5 as our backbone to balance accuracy, efficiency, and hardware compatibility.

Nevertheless, lightweight variants such as YOLOv5s still face difficulties in detecting small and occluded pedestrians in crowded scenes. Direct model compression typically results in significant accuracy loss, further limiting deployment on resource-constrained devices. To address these limitations, we propose a pedestrian detection framework based on YOLOv5s that integrates three core strategies:

1. a hierarchical prediction head with cross-branch guidance, where localization cues enhance classification of small or occluded pedestrians;

2. a dynamic label assignment strategy combined with a response-aware feature module to improve representation learning in lightweight backbones; and

3. a knowledge distillation scheme that transfers rich semantics from a large teacher to a compact student model, effectively balancing accuracy and efficiency.

The rest of the paper is organized as follows: Sect. 2 reviews related work, Sect. 3 details the proposed methodology, Sect. 4 presents extensive experiments, and Sect. 5 concludes with discussions on future work.

2 Related work

2.1 General object detection

Pedestrian detection has long been a core task in computer vision, with early approaches relying on handcrafted features such as HOG (Dalal and Triggs, 2005) and decision forests (Dollar et al., 2011). With the advent of deep learning, convolutional neural networks (CNNs) brought significant performance gains. Two-stage detectors like Faster R-CNN (Ren et al., 2016) offer strong accuracy but suffer from high computational cost, limiting deployment on resource-constrained devices. One-stage detectors, such as SSD (Liu et al., 2016) and YOLO (Redmon et al., 2016), introduced faster architectures that allow real-time inference. Despite this, pedestrian detection remains difficult due to the small object size, frequent occlusion, and dense scenes – conditions under which generic object detectors typically underperform (Zhang et al., 2017).

2.2 Lightweight detectors and model compression

In latency-sensitive applications such as autonomous driving and edge surveillance, lightweight detectors are preferred (Li et al., 2025). Some YOLO variants (Bochkovskiy et al., 2020; Jocher, 2020; Wang et al., 2023) employ compact backbones like CSPDarknet (Wang et al., 2020) or MobileNets (Howard, 2017) to reduce model size and accelerate inference. However, these simplifications often lead to accuracy degradation, especially on small or overlapping pedestrians. Beyond lightweight design, another mainstream direction for efficient deployment is model compression, such as pruning, quantization, or designing lightweight operators (Tan et al., 2020; Chen et al., 2019). While these methods reduce memory and computation, they typically require complex tuning and may sacrifice detection robustness. In contrast, our framework maintains the YOLOv5 backbone and instead improves accuracy through label assignment, feature enhancement, and knowledge distillation, offering a complementary solution to compression-based strategies. Table 1 presents a brief comparison.

2.3 Occlusion-robust detection

Occlusion remains a primary challenge in pedestrian detection. Approaches such as part-based models (Tian et al., 2015) and attention mechanisms (Zhang et al., 2018) improve robustness by focusing on visible regions or contextual cues. Other methods incorporate pose estimation (Fang et al., 2017) or semantic segmentation (Liu et al., 2019) for auxiliary supervision, although at significant computational cost. More recently, mutual learning between classification and localization branches has shown promise in improving feature alignment under occlusion (He et al., 2021). Our design extends this line by explicitly leveraging cross-branch guidance in the prediction head.

2.4 Knowledge distillation

Knowledge distillation (KD) has emerged as an effective technique to transfer knowledge from a large teacher model to a compact student model. Initially explored in classification (Hinton et al., 2015), KD has been extended to detection through feature-level (Chen et al., 2017), response-level (Li et al., 2017), and relation-based distillation (Wang et al., 2019). These methods allow compact models to inherit semantic richness from larger networks without increasing the inference cost significantly. We build on this paradigm to improve lightweight pedestrian detectors for edge deployment.

3 Methodology

To transform YOLOv5 from a general-purpose detector into a pedestrian-specific model, we introduce three core enhancements:

1. a hierarchical decoupled prediction head to improve detection precision;

2. a dynamic label assignment strategy to address occlusion challenges (Zou et al., 2020); and

3. knowledge distillation to boost accuracy while maintaining a lightweight structure (Chu et al., 2020).

Table 1. Comparison of different strategies for pedestrian detection on edge devices.

Category	Representative methods	Advantages	Limitations
High-accuracy transformer-based models	PedFormer (Rasouli and Kotseruba, 2022), RT-DETR (Zhao et al., 2024), MSTF (Hou et al., 2025)	Strong performance on small/occluded pedestrians; robust in dense scenes	High computational cost; slow inference; not suitable for edge devices
Lightweight models	YOLOv4-tiny (Bochkovskiy et al., 2020), YOLOv5s (Jocher, 2020), MobileNets (Howard, 2017)	Fast inference; low memory footprint; suitable for real-time edge deployment	Limited accuracy on small/occluded pedestrians; struggles in crowded scenes
Our proposed model	YOLOv5s + hierarchical head + dynamic label assignment + knowledge distillation	Balances accuracy and efficiency; better handling of small/occluded pedestrians; edge-friendly	Depends on teacher model quality

These methods are elaborated below.

3.1 Hierarchical guided decoupling prediction head

Classic YOLO models (v1–v5) adopt a coupled prediction head architecture, in which both classification and bounding box regression tasks share a common feature representation and a single output layer. Following feature extraction through the backbone (e.g., CSP-Darknet) and neck modules (e.g., PANet), a 1×1 convolutional layer is applied to project high-dimensional feature maps into final detection predictions (Carion et al., 2020). This output tensor encodes spatial location, objectness scores, class probabilities, and box offsets in a unified structure (Wang et al., 2022). While this design offers a compact and efficient detection pipeline, it imposes significant constraints on task-specific learning. Classification and regression are inherently different in terms of learning objectives – classification seeks to separate semantic categories in high-level feature space, whereas regression requires precise spatial and geometric localization. Forcing these two tasks to share parameters and gradients within the same branch can lead to feature interference and sub-optimal convergence, particularly for small or occluded objects. Motivated by recent studies (Hiller et al., 2020) emphasizing the need for decoupled optimization in detection heads, we redesign the standard YOLO prediction structure and introduce a hierarchical guided decoupling prediction head (HGDPH). This architecture explicitly separates the detection head into two specialized branches – one dedicated to classification and the other to regression. The decoupled structure allows each task to learn independent, task-specific features, which improves generalization and robustness in complex visual scenes. Moreover, to bridge the semantic gap between the two branches, we introduce a hierarchical guidance mechanism: regression outputs are utilized to guide the classification branch via spatial alignment and foreground enhancement.

As illustrated in Fig. 1, the output feature maps from the neck are first fed into two independent branches. The regression branch predicts bounding box coordinates and object confidence scores. These outputs are then decoded and passed through a feature extraction module (FEM) to select top-k high-confidence boxes. These regions are projected back to the classification branch, where they act as semantic anchors, guiding attention toward informative foreground areas. This spatial correspondence not only suppresses background noise but also enriches the semantic representation through localized feature fusion. Ultimately, the proposed HGDPH significantly improves detection accuracy and confidence estimation, especially in scenarios involving dense layouts and occluded objects.

3.2 Revised label assignment strategy

In crowded pedestrian detection scenarios, the challenges posed by occlusion, small-scale objects, and dense object overlap remain a persistent obstacle for accurate detection. Traditional label assignment strategies, such as static IoU thresholding, are often insufficient under such conditions. These approaches typically define a fixed intersection-over-union (IoU) threshold (e.g., 0.5) and assign predicted boxes to ground truth objects based solely on spatial overlap. However, in highly congested scenes where occluded or small objects abound, such rigid strategies fail to generate sufficient high-quality positive samples, resulting in poor supervision signals and degraded performance during training (Krishnaveni, 2023; Sandler et al., 2018). To overcome these limitations, we propose a cost-aware, dynamic label assignment strategy that reformulates the positive sample selection problem from the perspective of optimal transport theory. Specifically, we draw an analogy to a transportation cost minimization problem (Ge et al., 2021), where predicted bounding boxes are interpreted as goods to be delivered, and ground truth boxes serve as destination warehouses. The objective, in

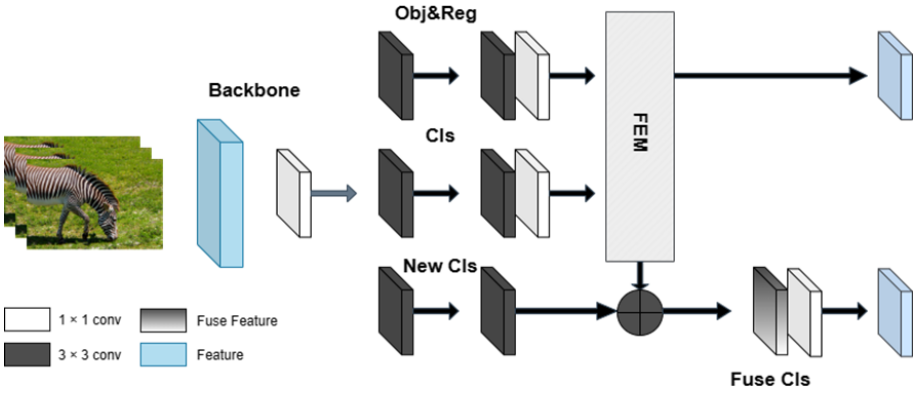


Figure 1. Illustration of the redesigned YOLOv5 prediction head.

this metaphor, is to minimize the overall delivery cost by optimizing the mapping between predicted boxes and their corresponding objects. This formulation leads to the construction of a cost matrix that encodes both spatial and semantic mismatches. As defined in Eq. (1), the total matching cost between a prediction and a ground truth box is composed of two terms: the IoU-based localization loss and the binary cross-entropy classification loss:

$$\mathcal{L}_{\text{iou\_loss}} = \text{IoU}(P_{\text{bbox}}, T_{\text{bbox}}),$$
$$\mathcal{L}_{\text{cls\_loss}} = \text{BCE}(P_{\text{cls}}, T_{\text{cls}}). \quad (1)$$

Here, P_{bbox} and T_{bbox} denote the predicted and ground truth bounding box coordinates, respectively, while P_{cls} and T_{cls} represent the predicted and target classification scores. A lower IoU loss indicates better spatial alignment between predicted and ground truth boxes, thereby increasing the likelihood of the prediction being considered a valid positive sample. To further improve adaptability, we introduce a dynamic positive sample selection mechanism. For each ground truth object, we first project its bounding box across the multi-scale feature maps generated by the backbone. Then, we evaluate all anchor points in its vicinity using the joint cost metric defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls\_loss}} + 3.0 \times \mathcal{L}_{\text{iou\_loss}}. \quad (2)$$

The factor of 3.0 emphasizes the importance of spatial alignment in positive sample matching. All candidate anchors are ranked in ascending order of this composite cost. The number of positive samples, denoted as K, assigned to each object is dynamically determined based on the size and quality of the candidate anchors. Larger or well-aligned predictions are granted more positive samples, reflecting their greater learning potential, while smaller or poorly aligned anchors receive fewer to reduce noise, illustrated in Fig. 2. This adaptive sampling strategy offers several benefits: (1) it increases robustness to spatial ambiguity in crowded scenes, (2) it avoids over-penalizing small objects, and (3) it enhances the diversity and relevance of the positive sample pool. Collectively, these enhancements enable the detector to more effectively learn fine-grained distinctions between foreground and background, ultimately boosting both precision and recall in dense pedestrian environments.

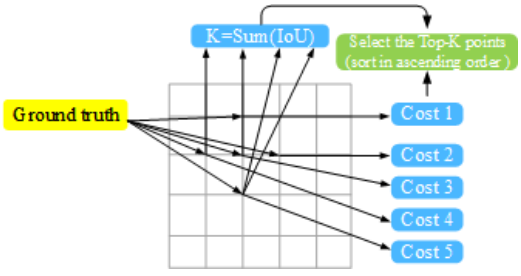


Figure 2. Illustration of the dynamic positive sample allocation strategy.

3.3 Knowledge distillation based on labels

Despite the significant accuracy gains from the hierarchical guided decoupling head (HGDPH) and dynamic label assignment strategies in Sect. 3.1 and 3.2, challenges remain. A key issue is the trade-off between model size and performance, especially for deployment on edge devices or embedded systems. To address this, we adopt knowledge distillation (KD) to compress the model without compromising accuracy and to accelerate convergence. KD, introduced by Hinton et al. (2015), transfers knowledge from a large teacher network to a smaller student model using soft output probabilities. These soft targets provide more informative supervision than one-hot labels by capturing inter-class relationships and confidence levels, helping the student to learn more generalizable features. In classification-based object detection, the teacher's logits z_i are converted to soft labels q_i using a temperature-scaled softmax:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}. \quad (3)$$

Table 2. A comparative analysis for the YOLOv5 series on the custom dataset.

Model	Input size	AP50 (%)	AP95 (%)
YOLOv5s	640	56.7	33.4
YOLOv5m	640	64.2	38.5
YOLOv5l	640	64.9	41.1

Here, T is a temperature parameter that controls the sharpness of the softmax distribution. When $T = 1$, the output reduces to a standard softmax distribution. As T increases, the output distribution becomes softer, exposing subtle differences among class probabilities. These softened outputs are particularly beneficial for improving generalization in low-capacity student networks. The temperature T thus serves as a crucial hyperparameter, enabling the tuning of the supervisory signal's granularity during distillation.

To realize label-based knowledge distillation, we adopt a response-based feature extraction strategy, where only the final logits of the teacher network are used to supervise the student. This form of distillation is computationally efficient and easy to implement, as it avoids the need to match intermediate feature maps or attention weights. The primary goal is to enhance the semantic sensitivity of the student network while maintaining a favorable speed–accuracy trade-off.

3.3.1 Dataset configuration and training setup

We conduct experiments on a custom pedestrian detection dataset of 10 000 images from public scenes (ports, campuses, transport hubs) with dense crowds, occlusions, and cluttered backgrounds – conditions challenging for lightweight detectors. A randomly sampled validation set of 1500 images covers diverse occlusions, lighting, and pedestrian scales. For knowledge distillation, we use three YOLOv5 variants – YOLOv5s (small), YOLOv5m (medium), and YOLOv5l (large) – as teacher and student candidates. Table 2 shows baseline performance: YOLOv5l yields the best accuracy due to its depth, YOLOv5s offers the fastest inference for real-time deployment, and YOLOv5m strikes a balance between accuracy and complexity. We adopt YOLOv5s as the student and test both YOLOv5m and YOLOv5l as teachers to assess optimal distillation under parameter constraints.

3.3.2 Effectiveness of knowledge transfer

As shown in Table 3, YOLOv5s without distillation achieves 79.4 % AP50 and 48.5 % AP95. With YOLOv5m as teacher, soft-label distillation improves AP50 by 4.5 % and AP95 by 2.9 %. Although a 3 % AP50 gap remains, this demonstrates effective transfer. Using YOLOv5l as the teacher yields less gain due to the semantic gap from its deeper architecture – hindering the student's ability to replicate high-level features. This supports prior findings that over-large teacher models can reduce transfer effectiveness. Thus, YOLOv5m proves to be the optimal teacher, balancing knowledge richness and transferability.

3.3.3 Training strategies for enhanced distillation

To further reduce the teacher–student gap, we test training strategies like early stopping and supervision decay (Table 4):

1. ES-1: disable distillation in the final five epochs;
2. ES-2: disable it in the final 40 epochs;
3. DS-1: linearly decay the teacher supervision to zero (YOLOv5m as teacher); and
4. DS-2: same as DS-1 but with YOLOv5l as the teacher.

ES-2 leads to early convergence with minimal gain (< 0.1 %), suggesting early stopping impairs learning. In contrast, DS-1 with YOLOv5m reduces the AP50/AP95 gap to 2.4 %/1.3 %. ES-1 achieves a 1.9 %/1.5 % gap, showing that late-stage independent learning can refine task-specific features. DS-2 underperforms DS-1, reaffirming the incompatibility between YOLOv5l and YOLOv5s. These results highlight the role of teacher–student compatibility in effective distillation.

3.3.4 Discussion and implications

A promising finding is the potential of a hierarchical approach: introducing an intermediate "teaching network" (e.g., YOLOv5l-m) as a semantic bridge to ease capacity transitions. This could benefit multi-stage compression pipelines. Although YOLOv5m has twice as many parameters than YOLOv5s, our ES-1-trained student approaches its accuracy with half the size, delivering strong trade-offs in compactness and speed. Overall, knowledge distillation emerges not just as a compression method but as a crucial strategy for designing accurate, efficient detectors in resource-limited settings.

4 Experiments

4.1 Datasets and experimental setup

In this study, we adopt YOLOv5s as the baseline detector due to its efficient trade-off between size, speed, and accuracy, making it suitable for resource-constrained applications like embedded or mobile systems. Leveraging COCO-pretrained weights, we apply transfer learning to accelerate convergence and improve generalization. Training is conducted using stochastic gradient descent (SGD) with an initial learning rate of 0.01. For bounding box regression, we use complete-IoU (CIoU) loss, which incorporates center

Table 3. Distillation experiment based on the YOLOv5.

Teacher model	Student model	AP50 (%)	Gap (%)	AP95 (%)	Gap (%)
YOLOv5m	YOLOv5s	61.2 (+4.5)	3.0	36.3 (+2.9)	2.2
YOLOv5l	YOLOv5s	57.8 (+1.1)	7.1	35.1 (+2.7)	6.0

Table 4. Experimental results of different distillation strategies.

Distillation strategy	Teacher model	Student model	AP50 (%)	AP95 (%)
Assistant network	YOLOv5l-m	YOLOv5s	61.1	36.1
Early stopping strategy – 1	YOLOv5m	YOLOv5s	62.3	37.0
Early stopping strategy – 2	YOLOv5m	YOLOv5s	61.3	36.3
Decay strategy – 1	YOLOv5m	YOLOv5s	61.8	36.8
Decay strategy – 2	YOLOv5l	YOLOv5s	61.4	35.6

distance, aspect ratio, and overlap, enabling more accurate localization than simpler IoU-based losses. To ensure robustness, our training dataset integrates three sources: CrowdHuman (Shao et al., 2018) for dense pedestrian detection and a custom pedestrian-vehicle dataset from real-world scenes for comprehensive evaluation. The implementation is based on Python and C++ using PyTorch 1.7.1, running on Ubuntu 20.04 with an NVIDIA RTX 3090 GPU (24 GB RAM). CUDA 10.1, NumPy, OpenCV, and Pandas support efficient preprocessing and data handling.

4.2 Ablation study and competitive experiment

To comprehensively evaluate our model in dense pedestrian detection, we conducted comparative experiments against several mainstream lightweight real-time detectors on the CrowdHuman dataset – a challenging benchmark featuring dense human instances and heavy occlusion. Our evaluation adopts two standard metrics: average precision at IoU = 0.5 (AP50) and at IoU = 0.5–0.95 (AP95), which together assess both coarse and fine-grained localization accuracy.

Besides detection accuracy, we assessed computational cost, including parameter count, network depth, and inference speed. Reported inference time reflects only the network's forward pass on a single GPU image, excluding preprocessing, decoding, and non-maximum suppression (NMS), ensuring a fair comparison of architectural efficiency. Notably, inference speed depends on more than parameter count alone – it also reflects the design of convolution types, fusion strategies, input resolution, activation functions, and memory patterns (Li et al., 2024). Models with similar parameter sizes can differ significantly in graph complexity, resulting in practical speed gaps. Hence, a multi-dimensional evaluation across accuracy, efficiency, and scalability offers a fairer and more informative performance benchmark (Li, 2024).

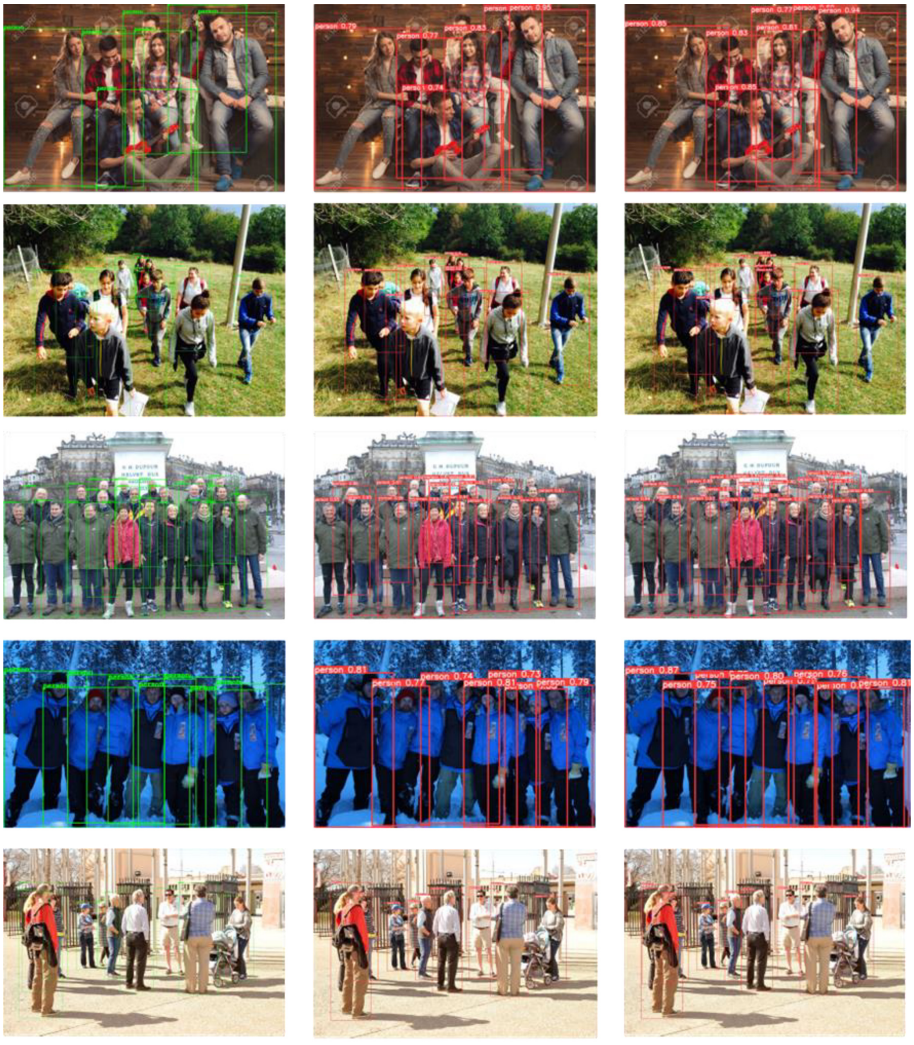
As shown in Table 5, YOLOv3-tiny, the most lightweight baseline, features a shallow architecture and achieves the fastest inference time (4.6 ms per image) due to its simplified backbone and detection heads. However, this speed comes at the cost of low accuracy (30.5 % AP50), with high false-negative and false-positive rates in crowded scenes, making it unsuitable for safety-critical applications. YOLOv5-Ghost (Han et al., 2020) improves upon YOLOv3-tiny by using Ghost convolution modules to reduce redundancy and enhance speed. It cuts over 50 % of parameters and improves AP50 by 9.7 points. However, its precision remains insufficient for detecting occluded pedestrians. Among all tested models, YOLOv5l achieves the highest accuracy (52.7 % AP95) due to its deep architecture with 607 layers and 76.8M parameters, enabling fine-grained localization. However, its heavy design limits real-time deployability on edge devices due to increased memory and computation costs. To address this, we propose a lightweight yet high-performance model with re-parameterization (Ding et al., 2021), *Ours-reparam*, which incorporates knowledge distillation and structural re-parameterization. The former enables a compact student model to mimic a larger teacher's predictions, while the latter merges multi-branch structures into an efficient single-path model at deployment. By removing 75 redundant layers via re-parameterization, our model reduces the parameter count to 9.8M, making it more suitable for edge platforms. Despite its lightweight nature, it achieves 83.9 % AP50 and 53.8 % AP95, outperforming all lightweight baselines. Moreover, its 7.1 ms inference time is 281.7 % faster than YOLOv5l, setting a new benchmark for real-time pedestrian detection in dense and occluded scenarios.

4.3 Visual comparison

To assess model performance visually, we compare YOLOv5l and our detector on representative images from CrowdHuman, VisDrone2019, and our custom dataset in

Table 5. The competitive experiments on the CrowdHuman dataset.

Model	Layers	Weight size (MB)	AP50 (%)	AP95 (%)	Latency (ms)
YOLOv3 (Redmon and Farhadi, 2018)	333	61.9	81.6	49.4	30.2
YOLOv3-tiny (Redmon and Farhadi, 2018)	59	8.9	60.5	30.5	4.6
YOLOv4-csp (Wang et al., 2020; Bochkovskiy et al., 2020)	513	52.9	81.2	49.5	27.8
YOLOv4-tiny (Bochkovskiy et al., 2020)	93	6.1	65.7	35.6	4.9
YOLOv5l (Jocher, 2020)	607	76.8	83.9	52.7	27.1
YOLOv5s-ghost (Han et al., 2020; Jocher, 2020)	453	3.9	73.9	40.2	5.8
YOLOv7-tiny (Wang et al., 2023)	255	6.2	75.1	41.6	5.0
Ours-m	396	28.9	85.7	53.1	16.0
Ours-s	297	10.2	83.9	53.8	8.4
Ours-s-reparam	222	9.8	83.9	53.8	7.1



(a) Ground truth (b) YOLOv5l (c) Ours-s-reparam

Figure 3. The visual prediction comparison of five test images. All images are from CrowdHuman (Shao et al., 2018).

Fig. 3. The baseline fails to detect heavily occluded pedestrians, often missing partially hidden targets. In contrast, our model, with a compact structure, effectively detects occluded pedestrians and maintains high confidence for large unoccluded objects.

5 Conclusions

In this work, we address the challenges associated with pedestrian detection in crowded scenarios, including the increased number of small-scale objects, severe occlusions, and the deployment limitations imposed by edge-computing devices. To tackle these issues, we propose a model optimization scheme that integrates several key strategies: a hierarchical structure that decouples classification confidence prediction and bounding box regression into separate branches, a dynamic matching strategy to enhance positive sample assignment for small and occluded objects, and a knowledge distillation framework to balance speed and accuracy in lightweight models. Our proposed detection model achieves an accuracy of 53.8 % AP95 and an inference speed of 7.1 ms per image on the CrowdHuman test set, delivering state-of-the-art performance among real-time detection models.

Code availability. The code for this project is available upon request from the corresponding authors, Cui Chen (78935883@qq.com) and Jun Li (cqleejun@cqjtu.edu.cn).

Data availability. The publicly available dataset CrowdHuman used in this study is available at https://www.crowdhuman.org/ (last access: 9 September 2025).

Author contributions. CC and JL designed the experiments and ZS, YaW, and YiW carried them out. CC and ZS developed the model code and performed the simulations. CC prepared the paper, with contributions from all co-authors.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support. This work is supported by the National Natural Science Foundation of China (52172381) and the Chongqing Major Technological Innovation Project (CSTB2022TIAD-STX0003).

Review statement. This paper was edited by Pengyuan Zhao and reviewed by two anonymous referees.

References

Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M.: YOLOv4: Optimal speed and accuracy of object detection, arXiv [preprint], https://doi.org/10.48550/arXiv.2004.10934, 23 April 2020.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S.: End-to-end object detection with transformers, European Conference on Computer Vision, Springer International Publishing, 213–229, https://doi.org/10.1007/978-3-030-58452-8_13, 2020.

Chen, G., Choi, W., Yu, X., Han, T., and Chandraker, M.: Learning efficient object detection models with knowledge distillation, Advances in Neural Information Processing Systems, 30, https://dl.acm.org/doi/10.5555/3294771.3294842, 2017.

Chen, Y., Yang, T., Zhang, X., Meng, G., Xiao, X., and Sun, J.: DetNAS: Backbone search for object detection, Advances in Neural Information Processing Systems, 32, https://dl.acm.org/doi/10.5555/3454287.3454883, 2019.

Choi, J. D. and Kim, M. Y.: A sensor fusion system with thermal infrared camera and LiDAR for autonomous vehicles and deep learning based object detection, ICT Express, 9, 222–227, https://doi.org/10.1016/j.icte.2021.12.016, 2023.

Chowdhury, S. A., Kowsar, M. M. S., and Deb, K.: Human detection utilizing adaptive background mixture models and improved histogram of oriented gradients, ICT Express, 4, 216–220, https://doi.org/10.1016/j.icte.2017.11.016, 2018.

Chu, X., Zheng, A., Zhang, X., and Sun, J.: Detection in crowded scenes: One proposal, multiple predictions, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12214–12223, https://doi.org/10.1109/CVPR42600.2020.01223, 2020.

Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, 886–893, https://doi.org/10.1109/CVPR.2005.177, 2005.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J.: RepVGG: Making VGG-style convnets great again, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13733–13742, https://doi.org/10.1109/CVPR46437.2021.01352, 2021.

Dollar, P., Wojek, C., Schiele, B., and Perona, P.: Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34, 743–761, https://doi.org/10.1109/TPAMI.2011.155, 2011.

Dong, J., Song, C., Sun, Y., and Zhang, T.: DAON: A decentralized autonomous oracle network to provide secure data for smart contracts, IEEE Transactions on Information Forensics and Security, 18, 5920–5935, https://doi.org/10.1109/TIFS.2023.3318961, 2023.

Fang, H. S., Xie, S., Tai, Y. W., and Lu, C.: RMPE: Regional multi-person pose estimation, Proceedings of the IEEE

International Conference on Computer Vision, 2334–2343, https://doi.org/10.1109/ICCV.2017.256, 2017.

Ge, Z., Liu, S., Li, Z., Yoshie, O., and Sun, J.: OTA: Optimal transport assignment for object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 303–312, https://doi.org/10.1109/CVPR46437.2021.00037, 2021.

Girshick, R.: Fast R-CNN, Proceedings of the IEEE International Conference on Computer Vision, 1440–1448, https://doi.org/10.1109/ICCV.2015.169, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587, https://doi.org/10.1109/CVPR.2014.81, 2014.

Hamzenejadi, M. H. and Mohseni, H.: Fine-tuned YOLOv5 for real-time vehicle detection in UAV imagery: Architectural improvements and performance boost, Expert Systems with Applications, 231, 120845, https://doi.org/10.1016/j.eswa.2023.120845, 2023.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C.: GhostNet: More features from cheap operations, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1580–1589, https://doi.org/10.1109/CVPR42600.2020.00165, 2020.

He, K., Zhang, X., Ren, S., and Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 1904–1916, https://doi.org/10.1109/TPAMI.2015.2389824, 2015.

He, Y., Zhu, C., and Yin, X. C.: Occluded pedestrian detection via distribution-based mutual-supervised feature learning, IEEE Transactions on Intelligent Transportation Systems, 23, 10514–10529, https://doi.org/10.1109/TITS.2021.3094800, 2021.

Hiller, M., Ma, R., Harandi, M., and Drummond, T.: Rethinking classification and localization for object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10186–10195, https://doi.org/10.1109/CVPR42600.2020.01020, 2020.

Hinton, G., Vinyals, O., and Dean, J.: Distilling the knowledge in a neural network, arXiv [preprint], https://doi.org/10.48550/arXiv.1503.02531, 9 March 2015.

Hou, S., Yang, M., Zheng, W. S., and Gao, S.: MultiSpectral Transformer Fusion via exploiting similarity and complementarity for robust pedestrian detection, Pattern Recognition, 162, 111383, https://doi.org/10.1016/j.patcog.2025.111383, 2025.

Howard, A. G.: MobileNets: Efficient convolutional neural networks for mobile vision applications, arXiv [preprint], https://doi.org/10.48550/arXiv.1704.04861, 17 April 2017.

Jocher, G.: YOLOv5, GitHub [code], https://github.com/ultralytics/yolov5 (last access: 22 November 2022), 2020.

Khanam, R. and Hussain, M.: A review of YOLOv12: Attention-based enhancements vs. previous versions, arXiv [preprint], https://doi.org/10.48550/arXiv.2504.11995, 16 April 2025.

Krishnaveni, K.: A novel framework using binary attention mechanism based deep convolution neural network for face emotion recognition, Measurement: Sensors, 30, 100881, https://doi.org/10.1016/j.measen.2023.100881, 2023.

Li, H.: Rethinking Features-Fused-Pyramid-Neck for Object Detection, European Conference on Computer Vision, Springer Nature Switzerland, 74–90, https://doi.org/10.1007/978-3-031-72855-6_5, 2024.

Li, H., Li, J., Wei, H., Liu, Z., Zhan, Z., and Ren, Q.: Slim-neck by GSConv: A lightweight design for real-time detector architectures, Journal of Real-Time Image Processing, 21, 62, https://doi.org/10.1007/s11554-024-01436-6, 2024.

Li, H., Ren, Q., Li, J., Wei, H., Liu, Z., and Fan, L.: A biologically inspired separable learning vision model for real-time traffic object perception in dark, Expert Systems with Applications, 129529, https://doi.org/10.1016/j.eswa.2025.129529, 2025.

Li, Q., Jin, S., and Yan, J.: Mimicking very efficient network for object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6356–6364, https://doi.org/10.1109/CVPR.2017.776, 2017.

Li, S. and Huang, C.: Using convolutional neural networks for image semantic segmentation and object detection, Systems and Soft Computing, 6, 200172, https://doi.org/10.1016/j.sasc.2024.200172, 2024.

Liu, S., Huang, D., and Wang, Y.: Adaptive NMS: Refining pedestrian detection in a crowd, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6459–6468, https://doi.org/10.1109/CVPR.2019.00662, 2019.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C.: SSD: Single shot multibox detector, Computer Vision–ECCV 2016, Springer International Publishing, 21–37, https://doi.org/10.1007/978-3-319-46448-0_2, 2016.

Lu, Y., Lu, J., Zhang, S., and Hall, P.: Traffic signal detection and classification in street views using an attention model, Computational Visual Media, 4, 253–266, https://doi.org/10.1007/s41095-018-0116-x, 2018.

Mo, W., Zhang, W., Wei, H., Cao, R., Ke, Y., and Luo, Y.: PVDet: Towards pedestrian and vehicle detection on gigapixel-level images, Engineering Applications of Artificial Intelligence, 118, 105705, https://doi.org/10.1016/j.engappai.2022.105705, 2023.

Rasouli, A. and Kotseruba, I.: PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning, arXiv [preprint], https://doi.org/10.48550/arXiv.2210.07886, 14 October 2022.

Redmon, J. and Farhadi, A.: YOLOv3: An incremental improvement, arXiv [preprint], https://doi.org/10.48550/arXiv.1804.02767, 2018.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: You only look once: Unified, real-time object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788, https://doi.org/10.1109/CVPR.2016.91, 8 April 2016.

Ren, S., He, K., Girshick, R., and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031, 2016.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C.: MobileNetV2: Inverted residuals and linear bottlenecks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510–4520, https://doi.org/10.1109/CVPR.2018.00474, 2018.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd, arXiv [preprint], https://doi.org/10.48550/arXiv.1805.00123, 30 April 2018.

Tan, M., Pang, R., and Le, Q. V.: EfficientDet: Scalable and efficient object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10781–10790, https://doi.org/10.1109/CVPR42600.2020.01079, 2020.

Tian, Y., Luo, P., Wang, X., and Tang, X.: Deep learning strong parts for pedestrian detection, Proceedings of the IEEE International Conference on Computer Vision, https://doi.org/10.1109/ICCV.2015.221, 2015.

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., and Yeh, I. H.: CSPNet: A new backbone that can enhance learning capability of CNN, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 390–391, https://doi.org/10.1109/CVPRW50498.2020.00203, 2020.

Wang, C. Y., Bochkovskiy, A., and Liao, H. Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721, 2023.

Wang, T., Yuan, L., Zhang, X., and Feng, J.: Distilling object detectors with fine-grained feature imitation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4933–4942, https://doi.org/10.1109/CVPR.2019.00507, 2019.

Wang, Y., Zhang, X., Yang, T., and Sun, J.: Anchor-DETR: Query design for transformer-based detector, Proceedings of the AAAI Conference on Artificial Intelligence, 36, 2567–2575, https://doi.org/10.1609/aaai.v36i3.20158, 2022.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., and Lee, B.: A survey of modern deep learning based object detection models, Digital Signal Processing, 126, 103514, https://doi.org/10.1016/j.dsp.2022.103514, 2022.

Zhang, S., Benenson, R., and Schiele, B.: CityPersons: A diverse dataset for pedestrian detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3213–3221, https://doi.org/10.1109/CVPR.2017.474, 2017.

Zhang, S., Wen, L., Bian, X., Lei, Z., and Li, S. Z.: Occlusion-aware R-CNN: Detecting pedestrians in a crowd, Proceedings of the European Conference on Computer Vision, 637–653, https://doi.org/10.1007/978-3-030-01219-9_39, 2018.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J.: DETRs beat YOLOs on real-time object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16965–16974, https://doi.org/10.1109/CVPR52733.2024.01605, 2024.

Zhu, Z. and Ji, Q.: Robust real-time eye detection and tracking under variable lighting conditions and various face orientations, Computer Vision and Image Understanding, 98, 124–154, https://doi.org/10.1016/j.cviu.2004.07.012, 2005.

Zou, T., Yang, S., Zhang, Y., and Ye, M.: Attention guided neural network models for occluded pedestrian detection, Pattern Recognition Letters, 131, 91–97, https://doi.org/10.1016/j.patrec.2019.12.010, 2020.