Complex & Intelligent Systems (2025) 11:480
https://doi.org/10.1007/s40747-025-02111-4

ORIGINAL ARTICLE

LW-DETR: a lightweight transformer-based object detection algorithm for efficient railway crossing surveillance

Baoye Song[1] · Shihao Zhao[1] · Zidong Wang[2] · Jianyu Chen[1] · Weibo Liu[2] · Xiaohui Liu[2]

Received: 25 March 2025 / Accepted: 13 September 2025

Abstract
Object detection at coal transportation railway crossings is crucial for accident prevention and traffic efficiency improvement. However, the application of existing methods on resource-constrained devices has seldom been considered. To address these challenges, in this paper, we propose a lightweight railway crossing object detection algorithm based on the Transformer framework, referred to as Light-Weight DEtection TRansformer (LW-DETR). In this algorithm, the Paddle Paddle-Lightweight CPU Convolutional Network (PP-LCNet) is employed as the backbone network, where standard convolution is combined with depthwise separable convolution for multi-scale feature extraction. Furthermore, the cross-scale feature fusion module is optimized to reduce redundant calculations and enhance feature fusion efficiency. Moreover, the Scylla-Intersection over Union loss function is introduced to comprehensively evaluate bounding box similarity, thereby improving object detection accuracy. Ablation experiments conducted on a modified Pascal Visual Object Classes (Pascal VOC) dataset demonstrate that LW-DETR, while maintaining acceptable detection accuracy, achieves a 135.3% increase in frames per second, a 71.7% reduction in parameters, and a 73.7% decrease in computational load, leading to effective lightweight performance. Comparative experiments with other popular object detection algorithms further confirm that LW-DETR significantly enhances detection speed while maintaining high accuracy, considerably reducing model size and validating the effectiveness of these improvements.

Keywords Railway crossing · Object detection · Transformer framework · Lightweight algorithm · Deep learning · Real-time monitoring

Introduction

Railway transportation plays a crucial role in coal transport within local railway networks due to its high capacity, low risk, and cost-effectiveness. As critical junctions where railways and roads intersect, railway crossings are essential for ensuring the smooth and safe operation of both railway and road traffic near coal transport lines [17]. Due to their unique geographical locations and functions, railway crossings are susceptible to traffic accidents in the absence of effective monitoring methods. Traditional monitoring approaches have typically relied on human observation, which is labor-intensive, time-consuming, and may increase the risk of accidents [55]. Therefore, the effective monitoring of railway crossings is essential for accident prevention, traffic efficiency improvement, and overall traffic safety assurance.

In recent years, object detection technology has been recognized as a critical component in video surveillance systems [24, 30]. Consequently, its integration into unattended railway crossings has been of significant practical importance, particularly for coal mine railway crossings. Compared to conventional methods, the application of object detection technology in railway crossings requires a higher level of intelligence and real-time performance to promptly iden-

✉ Zidong Wang
Zidong.Wang@brunel.ac.uk

Baoye Song
songbaoye@sdust.edu.cn

Shihao Zhao
1464187988@qq.com

Jianyu Chen
1157615244@qq.com

[1] College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China

[2] Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, UK

Published online: 24 October 2025

tify vehicles and pedestrians, thereby reducing the risk of accidents and ensuring traffic safety [55]. These advanced methods contribute significantly to traffic management optimization, transportation efficiency enhancement, and traffic congestion alleviation, thereby playing a pivotal role in the intelligent development of transportation systems [30].

Despite its essential role in safety monitoring, the study of object detection algorithms continues to be confronted with significant challenges. Complex environments, influenced by factors such as variations in lighting and weather, impose stringent demands on the robustness and stability of these algorithms to ensure accurate performance under diverse and challenging conditions [55]. Concurrently, substantial challenges remain regarding real-time performance and efficiency, requiring improvements in processing speed while maintaining high accuracy to meet real-time monitoring requirements. However, existing object detection algorithms often consist of a substantial number of parameters, making their deployment on embedded devices impractical. Furthermore, although many algorithms exhibit commendable performance in controlled testing environments, issues such as low frames per second (FPS) frequently arise in practical applications, thereby significantly limiting their effectiveness in real-time monitoring scenarios [17].

In response to the aforementioned challenges, in this paper, we propose a lightweight railway crossing object detection algorithm based on the Transformer framework, referred to as Light-Weight DEtection TRansformer (LW-DETR). The algorithm incorporates a series of modifications to the feature extraction network, the feature fusion network, and the loss function of LW-DETR. Many existing Transformer-based methods, like DETR [2], are characterized by high model complexity and a large number of parameters. To reduce the required number of parameters and model complexity, LW-DETR adopts streamlined and efficient feature extraction and fusion strategies within its architecture. Evaluations are conducted on the augmented Pascal Visual Object Classes (Pascal VOC) dataset, Pascal VOC07+12, by integrating the Pascal VOC 2007 and 2012 datasets.

The main contributions of this paper can be highlighted as follows.

1. The Paddle Paddle-Lightweight CPU Convolutional Network (PP-LCNet) is employed to replace the High-Performance GPU Network (HGNet), enabling multi-scale feature extraction through standard convolution (Conv), depthwise separable convolution (DSConv), and squeeze-and-excitation (SE) modules. As a result, the model's complexity and computational costs are successfully reduced.
2. The cross-scale feature fusion module in Real-Time DEtection TRansformer (RT-DETR) is optimized to address issues related to insufficient fine-grained information capture, high computational costs, and inadequate feature utilization. Consequently, the expressive capabilities and inference speed of the model are enhanced.
3. The Scylla-Intersection over Union (SIoU) metric, incorporating angular loss, distance loss, shape loss, and Intersection over Union (IoU) loss, is introduced to optimize the representation of bounding box positional relationships and regression outcomes.
4. The performance of LW-DETR is extensively evaluated on the reconfigured Pascal VOC07+12 benchmark, demonstrating state-of-the-art results across multiple metrics in comparison with several other algorithms.

The remainder of this paper is organized as follows. Section Related work provides a comprehensive review of existing research on lightweight network architectures and their corresponding optimization strategies, with key advancements and remaining challenges highlighted. Section Development of LW-DETR presents the technical details of the proposed LW-DETR framework, including its novel architectural components and optimization mechanisms. In Section Experimental results, an in-depth evaluation of the proposed method is conducted through extensive experiments, followed by a detailed analysis and discussion of the results. Finally, Section Conclusion summarizes the key findings of this study and outlines potential directions for future research.

Related work

Lightweight networks

In recent years, deep neural networks have attracted attention due to their remarkable performance [4, 22, 42, 43]. The development of ResNet has effectively addressed the vanishing gradient problem in deep networks, allowing for further deepening of network architectures [14]. However, as network depth increases, model complexity and hardware demands escalate, prompting accelerated research into lightweight networks.

In [15], MobileNetV1 has been developed by Google, utilizing depthwise separable convolutions (DSConv) to reduce model parameters and computational load. In [29], MobileNetV2 has been introduced by incorporating inverted residuals and linear bottlenecks, thereby improving network accuracy despite a slight increase in parameters. In [16], MobileNetV3 has been further enhanced through the integration of attention mechanisms and neural architecture search (NAS), optimizing the network structure and boosting performance. In [25], ShuffleNetV2 has been proposed by Megvii

Technology, optimizing memory access costs to improve detection speed.

In [35], EfficientNet has been presented by Google, emphasizing the balance among network depth, width, and resolution to enhance accuracy. In [36], EfficientNetV2 has been released to address the long training time and slow inference speed of EfficientNet. In [12] and [13], C-GhostNet and G-GhostNet have been proposed by Huawei's Noah's Ark Lab, optimizing performance for CPU and GPU devices, reducing feature redundancy, and enhancing lightweight model efficiency. In [23], MicroNet has been introduced by Microsoft through employing micro-decomposition convolutions and dynamic activation functions to maintain high accuracy in lightweight models. In [37], MobileOne has been launched by Apple, decoupling training and inference frameworks through linear branches and new parameters, thereby reducing inference time. In [3], FasterNet has been proposed by the Hong Kong University of Science and Technology, analyzing the relationships among frames per second (FPS), floating point operations (FLOPs), and floating point operations per second (FLOPS), and introducing partial convolution to reduce computational redundancy.

Lightweight object detection algorithms

On resource-constrained mobile devices and embedded systems, conventional object detection algorithms frequently fail to meet real-time detection requirements due to excessive computational and memory demands. To address this limitation, significant research efforts have been dedicated to the development of lightweight object detection algorithms capable of delivering both accuracy and efficiency under hardware constraints. For example, a YOLOv8-based detector has been proposed in [21] for concealed object identification using active millimeter wave images. In the YOLOv8-based detector, a new backbone, the attention mechanism and the depth-wise convolutions are employed, and redundant branches are removed to detect concealed objects. In [34], an arbitrary-oriented detector has been introduced for remote sensing images by integrating DCNDarknet with deformable convolution and rotation detection, which significantly improves the ship detection performance in complex backgrounds. In [45], the YOLO model for efficient pedestrian detection (YOLO-EPD) has been developed by featuring a selective content-aware downsampling module, a crowded multi-head attention module, and knowledge distillation, which enhances pedestrian detection in dense scenes while maintaining real-time efficiency. In [40], the lightweight small object detection algorithm (LSOD-YOLO) has been presented, which employs a lightweight cross-layer output reconstruction module, spatial pyramid pooling layer, C2f-N module, and a lightweight Dysample upsampler to tackle the detection problem in complex scenarios.

It should be pointed out that most of the aforementioned optimizations incorporate depthwise separable convolution (DSConv), which, despite reducing model parameters and computational load on GPU devices, cannot efficiently enhance detection speed due to frequent memory accesses. Consequently, achieving an optimal balance between detection accuracy and speed on embedded devices remains a significant challenge.

Development of LW-DETR

Modified backbone of LW-DETR

RT-DETR is a real-time object detection method based on the Transformer architecture [52]. Although its backbone network, HGNet, has demonstrated outstanding performance across various standard datasets, significant challenges are encountered when detecting small and dense targets. Moreover, the high model complexity and computational costs of HGNet severely restrict the applicability of RT-DETR in resource-constrained environments.

To address the identified challenges, PP-LCNet is employed in this study as a replacement for HGNet [5]. PP-LCNet integrates standard convolutions (Conv) with depthwise separable convolutions (DSConv), enabling effective multi-scale feature extraction and demonstrating superior performance in detecting small and dense targets. The structures of PP-LCNet and DSConv are illustrated in Table 1 and Fig. 1, respectively. Specifically, PP-LCNet progressively reduces the feature map size through downsampling, facilitating the extraction of target information at different levels of granularity across multiple scales. The use of DSConv not only minimizes the number of parameters but also expands the model's receptive field, thereby enhancing the ability to capture complicated details.

Next, the SE attention module [18] is integrated into the last two DSConv layers of PP-LCNet. As illustrated in Fig. 2, the SE module adaptively recalibrates channel-wise feature responses, thereby enhancing the model's ability to emphasize salient features. This integration strengthens the model's capacity to distinguish subtle differences between dense objects and the background. By incorporating these structural designs and modules, PP-LCNet effectively overcomes the limitations of HGNet on the revised Pascal VOC07+12 dataset, significantly advancing object detection technology.

Remark 1 PP-LCNet has significantly reduced model complexity and computational demands through an optimized network architecture and the strategic integration of key modules. When small and dense targets are detected, feature information is effectively utilized by PP-LCNet to enhance object detection efficiency. The lightweight nature

Table 1 Structure of the PP-LCNet

Input	Operator	Output channels	SE	Stride
640 × 640 × 3	Conv2d 3 × 3	8	–	2
320 × 320 × 8	DSConv 3 × 3	16	–	1
320 × 320 × 16	DSConv 3 × 3	32	–	2
160 × 160 × 32	DSConv 3 × 3	32	–	1
160 × 160 × 32	DSConv 3 × 3	64	–	2
80 × 80 × 64	DSConv 3 × 3	64	–	1
80 × 80 × 64	DSConv 3 × 3	128	–	2
40 × 40 × 128	DSConv 5 × 5	128	–	1
40 × 40 × 128	DSConv 5 × 5	128	–	1
40 × 40 × 128	DSConv 5 × 5	128	–	1
40 × 40 × 128	DSConv 5 × 5	128	–	1
40 × 40 × 128	DSConv 5 × 5	128	–	1
40 × 40 × 128	DSConv 5 × 5	256	✓	2
20 × 20 × 256	DSConv 5 × 5	256	✓	1

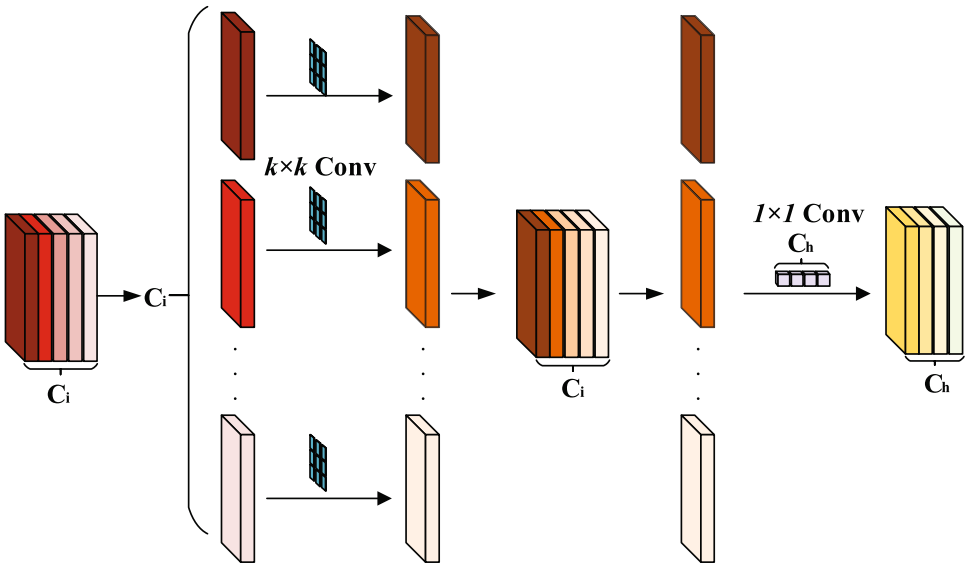

Fig. 1 Structure of the DSConv

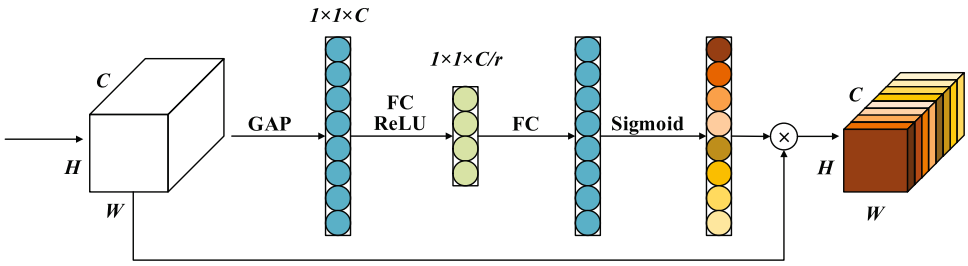

Fig. 2 Structure of the SE attention module

of its network structure substantially minimizes computational resource requirements, allowing PP-LCNet to operate efficiently in resource-constrained environments, including embedded systems and mobile devices.

Precise fusion module of LW-DETR

In PP-LCNet, DSConv significantly reduces the number of model parameters and computational load; however, it does not proportionally increase inference speed. This limitation arises because inference speed is influenced not only by the model's FLOPs but also by its FLOPS. Although DSConv has been highly effective in reducing FLOPs, frequent memory accesses result in lower FLOPS, thereby limiting improvements in inference speed. Furthermore, the cross-scale feature fusion module of RT-DETR exhibits several deficiencies. Specifically, the Reparameterized Block (RepBlock) [39] within this module fails to effectively capture fine-grained details, leading to suboptimal performance when detecting small and dense targets. This limitation adversely impacts feature extraction capabilities and ultimately compromises detection accuracy. Moreover, the repetitive structure of the RepBlock increases computational costs, making it less suitable for resource-constrained environments. Moreover, the simple element-wise addition method used by the RepBlock cannot fully leverage feature information, resulting in inadequate feature fusion and restricting the model's expressive power. These issues undoubtedly hinder the performance of RT-DETR in object detection tasks.

To address the limitations with the cross-scale feature fusion module, a precise fusion module (PFM) is proposed to enhance both precision and inference speed by overcoming the shortcomings of the existing fusion process. The PFM integrates a diverse set of operations, including standard convolution, partial convolution, pointwise convolution, residual connections, and element-wise addition, to effectively refine and extract intricate information from the input features. The architecture of the proposed precise fusion module is illustrated in Fig. 3.

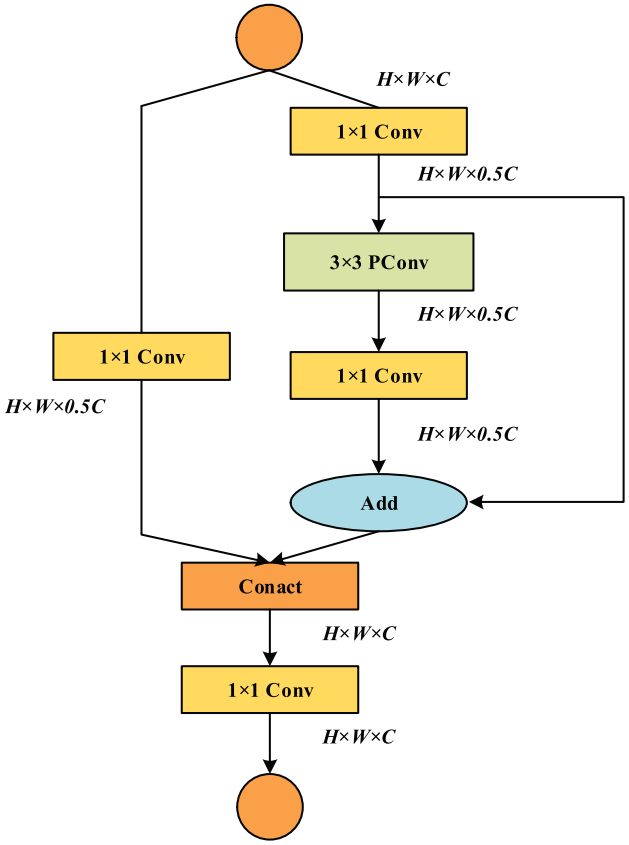

Fig. 3 Structure of the precise fusion module

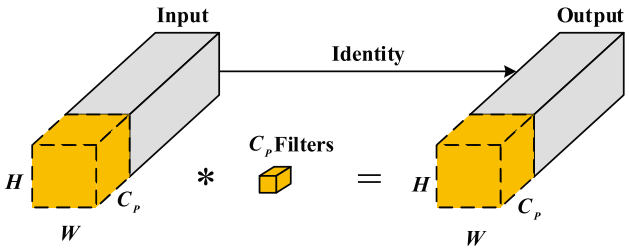

Fig. 4 Structure of PConv in the precise fusion module

The introduction of Partial Convolution (PConv), as illustrated in Fig. 4, represents a critical component of the PFM workflow. The PConv acts upon an input feature tensor with the dimension of $H \times W \times C$, where H, W, and C represent the height, width, and channel number, respectively. The core mechanism of PConv is to divide the C input channels into two subsets C_p and C_{id}. Here, a standard spatial convolution is applied to the C_p-channel subset, typically yielding an output feature map also containing C_p channels. The remaining C_{id} channels, defined as $C_{id} = C - C_p$, are processed via an identity mapping. Finally, the resulting features from the C_p channels (with the dimension of $H \times W \times C_p$) are concatenated along the channel with the features from the C_{id} channels (with the dimension of $H \times W \times C_{id}$). In this case, the number of channel in the final output feature map is C. Leveraging the operational principle of its overall process, PConv within the PFM workflow selectively processes significant regions of the feature map to preserve essential information while suppressing non-essential parts and noise. The PConv is subsequently employed to refine the input features, thereby improving feature extraction capabilities. The PConv mechanism not only optimizes convolution operations and feature extraction but also, in our PFM framework, effectively avoids vanishing and exploding gradients.

The final workflow of the proposed PFM can be summarized as follows. The input features are initially processed through two separate branches. In one branch, partial convolution is applied to target critical areas of the feature map,

retaining key information while filtering out irrelevant parts and noise. The features processed by partial convolution are then concatenated along the channel dimension with those from the other branch, which undergoes a 1×1 convolution. This concatenation significantly enhances the network's recognition performance and feature representation capabilities.

Remark 2 The proposed PFM significantly enhances feature extraction by integrating partial convolution and pointwise convolution. A substantial reduction in computational cost and model complexity is achieved through hierarchical adjustments and the elimination of redundant calculations, enabling more effective handling of inference demands for small and dense targets, particularly in resource-constrained environments. Furthermore, efficient element-wise addition and concatenation operations are utilized to improve the fusion of features extracted from different pathways. As a result, superior feature integration is achieved, further enhancing object detection accuracy.

Loss function of LW-DETR

The loss function of the traditional RT-DETR model comprises three distinct components: classification loss, regression loss, and bounding box loss. The classification loss is implemented using the VariFocal Loss (VFL) [48], which effectively addresses class imbalance. The regression loss is computed using the L_1 loss [51], selected for its robustness to outliers. The bounding box loss is based on the Generalized Intersection over Union (GIoU) loss [28], which enhances the accuracy of bounding box predictions.

Building upon this foundation, our work introduces modifications only to the bounding box loss. The original model employed the GIoU loss for this part. Specifically, the GIoU loss is defined by

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{\mathrm{C} - (\mathrm{A}\cup\mathrm{B})}{\mathrm{C}} \tag{1}$$

where A represents the predicted box, B denotes the ground truth box, and C is the minimum bounding box enclosing both the predicted and ground truth boxes, as illustrated in Fig. 5. However, our study has identified certain limitations of the GIoU loss function, particularly in cases where two predicted boxes are entirely overlapped, as depicted in Fig. 6.

To address the limitations of GIoU, SIoU [10] is employed as a replacement. SIoU integrates four loss components: angle loss, distance loss, shape loss, and intersection-over-union loss. By comprehensively incorporating these components, SIoU provides a more thorough evaluation of the similarity between bounding boxes, thereby improving the representation of spatial relationships and enhancing

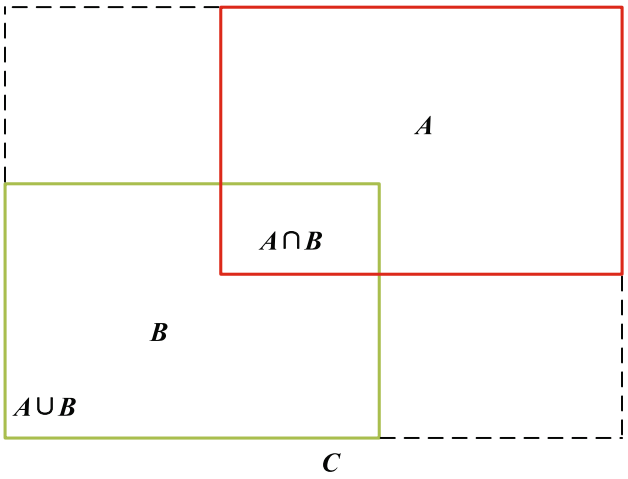

Fig. 5 Illustration of GIoU

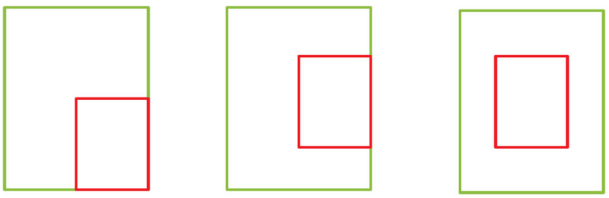

Fig. 6 Limitation of GIoU

regression results, and this ultimately leads to improved performance of the object detection model.

The SIoU loss function is defined as

$$\mathrm{SIoU} = 1 - \mathrm{IoU} + \frac{\Delta + \Omega}{2}, \tag{2}$$

where Δ represents the distance loss between the ground truth and predicted box; and Ω denotes the shape loss. The IoU loss is expressed as

$$\mathrm{IoU} = \frac{\mathrm{B}^{\mathrm{gt}} \bigcap \mathrm{B}}{\mathrm{B}^{\mathrm{gt}} \bigcup \mathrm{B}}, \tag{3}$$

where B^{gt} and B correspond to the areas of the ground truth and predicted boxes, as illustrated in Fig. 7. The distance loss Δ is given by

$$\Delta = \sum_{k=x,y} (1 - e^{-\gamma\rho_k}), \tag{4}$$

where $\rho_x = (\frac{B_{cx}^{gt} - B_{cx}}{C_w})$, $\rho_y = (\frac{B_{cy}^{gt} - B_{cy}}{C_h})$; $(B_{cx}^{gt}, B_{cy}^{gt})$ and (B_{cx}, B_{cy}) are the center coordinates of the ground truth and predicted boxes, respectively; C_w and C_h (which are shown in Fig. 7) are the width and height of the minimum bounding rectangle for the ground truth and predicted boxes, respectively; and γ is an intensity factor.

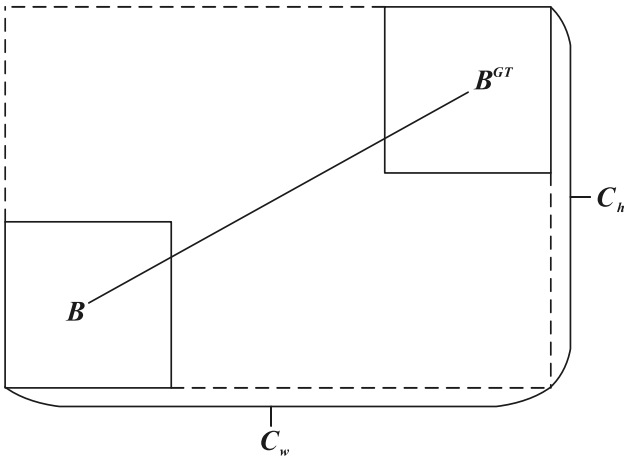

Fig. 7 Distance loss of SIoU

The shape loss Ω is defined as

$$\Omega = \sum_{k=w,h} (1 - e^{-\omega_k})^{\theta}, \tag{5}$$

where $\omega_w = \frac{|w^{gt} - w|}{\max(w^{gt}, w)}$, $\omega_h = \frac{|h^{gt} - h|}{\max(h^{gt}, h)}$; w^{gt} and h^{gt} are the width and height of the ground truth box, respectively; w and h are the width and height of the predicted box, respectively; and θ is a hyperparameter controlling the intensity of the shape loss.

Remark 3 It is worth noting that SIoU enhances the similarity measurement between bounding boxes by integrating shape features, allowing for a more precise description of their spatial relationships. As a result, a more comprehensive similarity assessment is facilitated, providing more accurate metrics, particularly for shape features and relative positions. This improvement leads to enhanced localization accuracy of target objects and further strengthens the performance of object detection algorithms.

Experimental results

Dataset description

The dataset employed in this study is derived from the Pascal VOC 07+12 benchmark, which encompasses diverse images across 20 object categories. To tailor the dataset for railway crossing safety analysis, focusing on the most relevant objects, a specific modification procedure is implemented. Initially, images and labels corresponding to five categories (i.e., person, car, bus, motorbike, and bicycle) are extracted. Subsequently, motorbike and bicycle classes are merged into one class, i.e., the motorcycle. Such consolidation aims to enhance robustness of the detector against potential misclassification between motorbike and bicycle (which are visually similar classes).

The resultant dataset comprises four object classes pertinent to railway crossing scenarios: person, car, bus, and motorcycle. Despite originating from a general visual dataset, the tailored subset (motorcycle) provides substantial visual diversity for the modified dataset, including variations in appearance, illumination, viewpoint, and occlusion. Such diversity is crucial for training detectors capable of robust performance within complex real-world railway crossing environments. For experimental purposes, the modified dataset is partitioned into the training and validation sets with an 8:2 ratio, yielding 21,671 training and 5417 validation images.

Experimental environment

The experimental environment is configured as detailed in Table 2. The operating system is Windows 10 Professional, with an AMD Ryzen 7 5700X 8-Core Processor running at 3.4 GHz and an NVIDIA GeForce RTX 3060 12 G GPU. The integrated development environment is set up using Anaconda and configured with Python 3.8.18, utilizing the PaddlePaddle 2.6 deep learning framework. Additionally, the CUDA 11.8 computing platform and the cuDNN 8.9.4 neural network library are employed for acceleration. The model uses the AdamW optimizer with a weight decay coefficient of 0.0001, processes input images at a resolution of 640×640, and operates with a batch size of 8. The initial learning rate is set to 0.001, and after each epoch, it decreases by a factor of 0.94 over a total of 100 epochs.

Ablation experiments

To evaluate the impact of the proposed enhancements on the network model's performance, a series of ablation studies have been performed on the modified Pascal VOC07+12 dataset. To ensure consistent and fair comparisons, all networks are trained and validated under identical hardware conditions. The results of these ablation studies are summarized in Table 3. In the table, 'Baseline' corresponds to RT-DETR; 'LC' denotes the incorporation of the lightweight backbone network PP-LCNet; 'PFM' represents the replacement of the fusion module with the precise fusion module; and 'SIoU' indicates the adoption of the SIoU-optimized loss function.

It can be seen in Table 3 that using PP-LCNet as the backbone network leads to a decrease in mAP@0.5:0.95 from 61.2% to 58% and mAP@0.5 from 79.3% to 76.6%, while the FPS of our model increases from 37.4 to 88.0. As a lightweight network, our model achieves a significant decrease in both the parameter count by 68.8% and the computational load (FLOPs) by 69.2%. After replacing

Table 2 Experimental environment

Name	Configuration
Operating System	Windows 10 Professional
CPU	AMD Ryzen 7 5700X 8-Core Processor 3.40 GHz
GPU	NVIDIA GeForce RTX 3060 12 G
Deep Learning Frameworks	Paddle 2.6.0
Integrated Development Environment	Python 3.8.18, CUDA 11.8, CUDNN 8.9.4

Table 3 Results of the ablation study

Method	mAP@0.5:0.95 (%)	mAP@0.5 (%)	FPS	Params (M)	FLOPs (G)
Baseline	61.2	79.3	37.4	19.3	60.4
Baseline + LC	58.0	76.6	79.6	6.02	18.6
Baseline + PFM	58.9	77.9	47.1	19.13	53.2
Baseline + SIoU	61.4	79.5	37.4	19.3	60.4
Baseline + LC + PFM	57.8	76.4	88.0	5.47	15.9
Baseline + LC + PFM + SIoU	58.2	77.0	88.0	5.47	15.9

the fusion module with C3-Faster, a slight degradation in detection accuracy is observed, but the model becomes more efficient, achieving faster detection speeds and better utilization of computational resources. As a result, mAP@0.5:0.95 and mAP@0.5 of our model decrease by only 0.2%, while the FPS of our model increases by 9.6%, with reductions in both model parameters and computational load. Finally, by replacing GIoU with the SIoU loss function, our model demonstrates accelerated convergence during training and improved robustness, achieving mAP@0.5:0.95 of 58.2% and mAP@0.5 of 77%. Compared to the RT-DETR (Baseline) algorithm, mAP@0.5:0.95 and mAP@0.5 of our model decrease by 6.3% and 6.6%, respectively. Although the proposed LW-DETR results in a slight degradation in detection accuracy, our model achieves a 135.3% increase in FPS, a 71.7% decrease in model parameters, and a 73.7% reduction in computational load, which enables real-time detection on mobile and embedded devices. We can conclude that the developed lightweight algorithm achieves higher detection accuracy, faster detection speeds, and smaller model sizes comparing with the baseline, which demonstrates the superiority of our model for efficient pedestrian and vehicle detection on mobile and embedded devices.

Comparison experiments

Comparison of loss functions

The performance of various IoU loss functions in the LW-DETR model is evaluated, including the original GIoU [28], Distance Intersection over Union (DIoU) [53], Complete Intersection over Union (CIoU) [54], Efficient Intersection over Union (EIoU) [50], Multi-Point Distance Intersection over Union (MPDIoU) [31], and SIoU [10]. The experiments are performed on the modified Pascal VOC07+12 dataset, with evaluation metrics including mAP@.5:.95 and mAP@.5. The testing results for each loss function on the modified Pascal VOC07+12 dataset are presented in Table 4.

A detailed analysis of Table 4 reveals that the performance differences among the loss functions are minimal for the mAP@0.5 metric but significant for mAP@0.5:0.95. Notably, SIoU achieves the highest mAP@0.5:0.95, demonstrating its superiority at finer precision scales. Although DIoU leads in mAP@0.5, it falls short compared to SIoU in the more comprehensive mAP@0.5:0.95 metric for object detection tasks. CIoU and EIoU exhibit identical performance in mAP@0.5:0.95, which is significantly lower than the other four loss functions. As a recently proposed loss function in 2023, MPDIoU performs exceptionally well; however, it trails SIoU by 0.1% in both mAP@0.5:0.95 and mAP@0.5. In conclusion, comparative experiments confirm SIoU's outstanding performance on the modified Pascal VOC07+12 dataset.

Table 4 Comparison of loss functions

Method	mAP@0.5:0.95 (%)	mAP@0.5 (%)
GIoU	57.8	76.4
DIoU	57.4	77.3
CIoU	56.9	77.1
EIoU	56.9	76.9
MPDIoU	58.1	76.9
SIoU	58.2	77.0

Comparison with different object detection algorithms

To validate the superiority of the LW-DETR algorithm, the same dataset is used for comparison with several other algorithms under identical hardware conditions. The compared algorithms include RT-DETR (Baseline) [52], YOLOv5s [19], YOLOv6n [20], YOLOv7-tiny [38], YOLOX-tiny [11], and LW-DETR (Ours). The comparative results of detection performance are summarized in Table 5.

On the modified Pascal VOC07+12 dataset, the performance of various algorithms is evaluated using the mAP@0.5:0.95 and mAP@0.5 metrics. Compared to the Baseline, the LW-DETR algorithm achieves a remarkable 135.3% improvement in FPS, along with a 71.7% reduction in model parameters and a 73.7% decrease in computational load, despite a 6.3% and 6.6% decline in mAP, thereby achieving lightweight performance. YOLOv5s demonstrates slightly higher detection accuracy than YOLOv6n and YOLOv7-tiny, but its adopted strategies adversely affect its detection speed, making it less suitable for mobile and embedded devices. YOLOv6n exhibits lower detection accuracy than most algorithms, although it has the smallest model size and offers better detection speed. YOLOv7-tiny, a compact version of YOLOv7, shows a slightly lower mAP than the Baseline and slower detection speeds. YOLOX-tiny achieves a relatively high mAP in this experiment, but its performance is suboptimal due to the dataset's incompatibility with its anchor-free detector and label assignment strategies. A comprehensive analysis of the comparative experiments, as presented in Table 5, indicates that the proposed LW-DETR algorithm delivers superior overall performance.

The algorithm's detection capabilities are demonstrated on images from three distinct scenarios: long-range visibility, foggy conditions, and occlusion. All selected images have a resolution of 192×108 pixels. The specific detection results are illustrated in Figs. 8, 9 and 10, where YOLOv7-tiny and YOLOX-tiny are abbreviated as YOLOV7t and YOLOXt, respectively, for simplicity. The final comparative experiments confirm the significant advantages of the proposed LW-DETR algorithm in terms of detection accuracy, speed, and model efficiency.

LW-DETR for object detection at railway crossings

To validate the effectiveness of the proposed LW-DETR in detecting targets at railway crossings, images of railway crossings under both sufficient and insufficient lighting conditions are selected for evaluation. The detection results, as illustrated in Figs. 11 and 12, demonstrate that the proposed LW-DETR outperforms other algorithms in detecting persons and vehicles at railway crossings. Compared to other algorithms, LW-DETR shows exceptional performance in real-time capability and detection accuracy, effectively addressing challenges related to lighting variations and target occlusions. Therefore, LW-DETR can be considered a more reliable and effective solution for detecting persons and vehicles at railway crossings.

From the discussion above, it is concluded that LW-DETR, a Transformer-based lightweight object detection algorithm, achieves an effective balance between high frame rates and accuracy in railway crossing object detection through a series of lightweight schemes. In future work, a key challenge is reducing accuracy degradation while maintaining high frame rates. Specific challenges include:

- To enhance accuracy while preserving FPS, further optimization of the feature extraction network could be pursued. Although PP-LCNet has demonstrated effectiveness as a lightweight backbone network, future research could explore more advanced feature extraction techniques. For example, integrating more efficient attention mechanisms could improve feature representation while maintaining a low computational load, thereby boosting detection accuracy.
- Feature fusion plays a critical role in enhancing object detection accuracy. While the current precise fusion module has improved the computational efficiency of feature fusion, future research could focus on more advanced feature fusion strategies to further enhance detection capabilities, particularly for small objects in complex environments. Additionally, the design of the loss function is crucial for detection accuracy. Although the SIoU loss function has shown promise in improving accuracy,

Table 5 Comparision results

Method	mAP@0.5:0.95 (%)	mAP@0.5 (%)	FPS	Params (M)	FLOPs (G)
Baseline	61.2	79.3	37.4	19.3	60.4
YOLOv5s	54.9	78.2	88.5	7.07	16.5
YOLOv6n	52.8	75.9	116.2	4.24	11.9
YOLOv7-tiny	53.7	79.3	83.8	6.02	13.2
YOLOX-tiny	51.3	78.5	77.3	5.05	15.1
LW-DETR	58.2	77.0	88.0	5.47	15.9

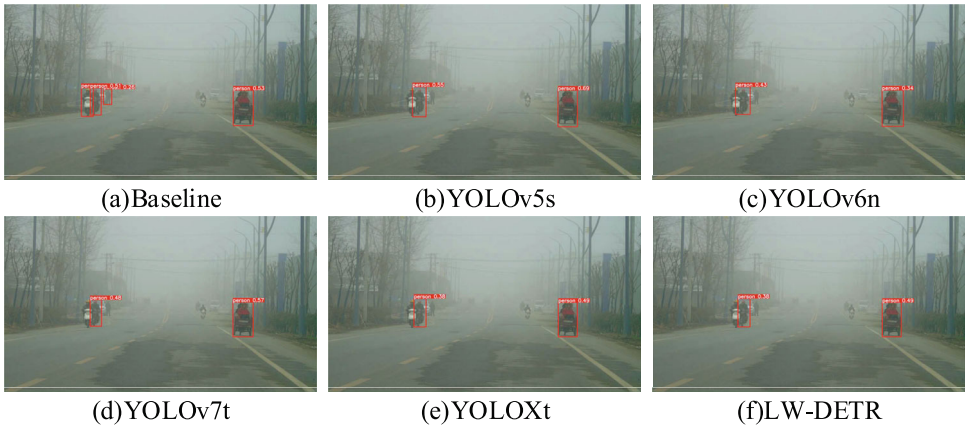

Fig. 8 Detection results in foggy scenes

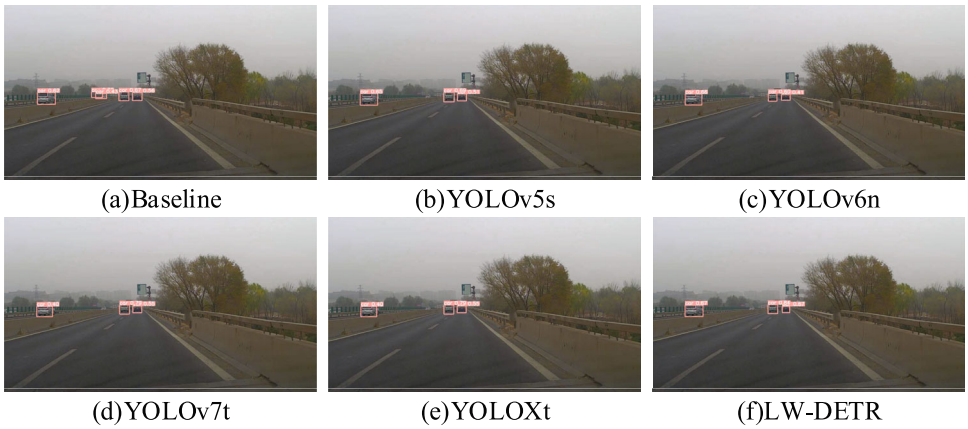

Fig. 9 Detection results in long-range scenes

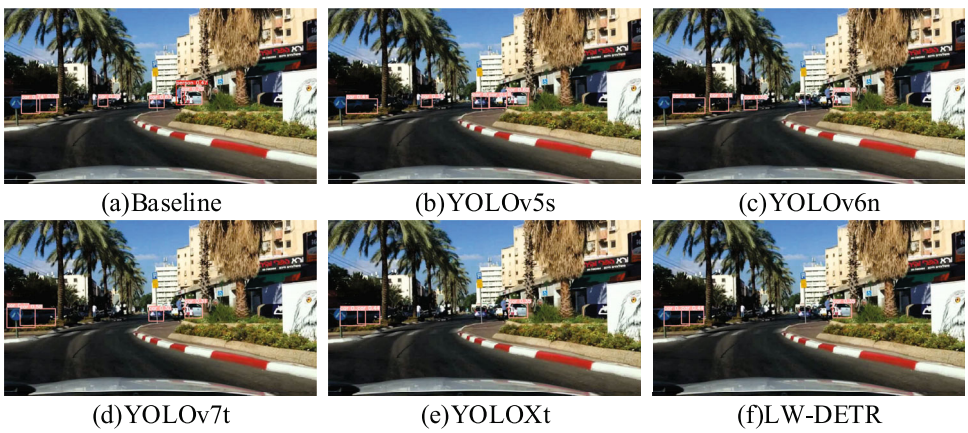

Fig. 10 Detection results in occlusion scenes

combining it with other loss functions could better guide the model in learning subtle object features, especially for objects against diverse backgrounds.

- Attention should also be directed toward deploying models on resource-constrained devices. By leveraging model pruning and distillation techniques, LW-DETR could further reduce computational demands on low-performance devices, thereby enhancing the user experience in practical applications, such as real-time monitoring.

Conclusion

In this paper, a transformer-based lightweight object detection algorithm, LW-DETR, has been proposed to address accuracy and efficiency challenges in railway crossing object detection, thereby mitigating the inefficiencies and missed detections associated with traditional manual monitoring methods. First, LW-DETR integrates the lightweight backbone network PP-LCNet, along with standard convolution, depthwise separable convolution, and an SE attention mod-

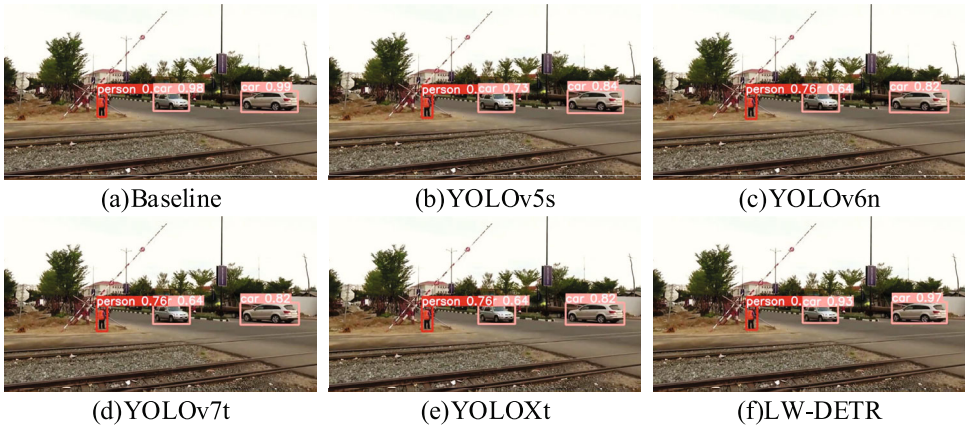

Fig. 11 Detection results in light scenes

Fig. 12 Detection results in low-light scenes

ule, to enhance feature extraction capabilities. This enables efficient detection of small and dense objects. Second, feature fusion is optimized through a precise fusion module, reducing computational costs while significantly improving the model's expressive capability and inference speed. Finally, to overcome the limitations of the traditional GIoU loss function, LW-DETR introduces the SIoU loss function to comprehensively evaluate bounding box similarity, further enhancing overall performance.

To further validate the effectiveness of the proposed LW-DETR, a series of ablation experiments have been conducted on the modified Pascal VOC07+12 dataset. The experiments systematically assess the independent contributions of each technique to the model's performance. The effectiveness of these improvements is demonstrated through the stepwise introduction of different enhancement modules and a quantitative analysis of multiple key performance metrics. Additionally, comparative experiments with other mainstream object detection algorithms have been performed, evaluating detection results in scenarios such as long-range visibility, fog, and occlusion. These comparisons highlight the significant advantages of LW-DETR in terms of detection accuracy, speed, and model compactness.

It is concluded that LW-DETR provides an efficient and accurate object detection solution through its novel network architecture design and optimization strategies, thereby advancing the development of intelligent monitoring technology for railway crossings. This algorithm is particularly well-suited for real-time detection applications on mobile and embedded devices, offering broad application prospects and practical value.

In the future research, we aim to focus on four core research topics to advance railway crossing object detection technologies: (1) constructing a dedicated dataset (that encompasses various weather conditions and complex backgrounds) with multimodal data (such as visible light and infrared objects) to enhance the model's environmental adaptability [27, 44, 47]; (2) designing physics-informed models that integrate sensor noise modeling with robustness-enhancement strategies to ensure stable detection performance under adverse weather conditions [8, 24, 26, 33, 43]; (3) developing a multi-source information fusion framework for signal processing, combined with state estimation techniques to enable real-time monitoring of system operation and anomaly detection [1, 7, 9, 32, 41, 49]; and (4) applying grid search or other optimization algorithms to effectively

automate the parameter tuning process and reduce manual intervention [6, 46].

Acknowledgements This work was supported in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2023MF067, the European Union's Horizon 2020 Research and Innovation Programme under Grant 820776 (INTEGRADDE), the Engineering and Physical Sciences Research Council (EPSRC) of the UK, the Royal Society of the UK, and the Alexander von Humboldt Foundation of Germany.

Author Contributions Baoye Song: Conceptualization, Methodology, Software, Writing- Original draft preparation. Shihao Zhao: Methodology, Software, Writing- Original draft preparation. Zidong Wang: Supervision, Conceptualization, Methodology, Writing- Reviewing and Editing. Jianyu Chen: Methodology, Software, Writing- Original draft preparation. Weibo Liu: Methodology, Software, Writing- Original draft preparation. Xiaohui Liu: Supervision, Conceptualization, Methodology.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Caballero-Águila R, Hu J, Linares-Pérez J (2024) Filtering and smoothing estimation algorithms from uncertain nonlinear observations with time-correlated additive noise and random deception attacks. Int J Syst Sci 55(10):2023–2035
2. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers, In: Proceedings of the European conference on computer vision, Glasgow, UK, pp 213–229
3. Chen J, Kao S,He H, Zhuo W, Wen S, LEE CH, Chan SHG (Jun. 2023) Run, don't walk: chasing higher FLOPS for faster neural networks, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12021–12031, Vancouver, Canada
4. Chen Z, Zhang L, Tang J, Mao J, Sheng W (2024) Conditional generative adversarial net based feature extraction along with scalable weakly supervised clustering for facial expression classification. Int J Netw Dyn Intell 3(4):100024
5. Cui C, Gao T, Wei S, Du Y, Guo R, Dong S, Lu B, Zhou Y, Lv X, Liu Q, Hu X, Yu D, Ma Y (2021) PP-LCNet: a lightweight CPU convolutional neural network, arXiv: 2109.15099
6. El-Shorbagy MA, Bouaouda A, Nabwey HA, Abualigah L, Hashim FA (2024) Bald eagle search algorithm: a comprehensive review with its variants and applications. Syst Sci Control Eng 12(1):2385310
7. Fu L, An L, Zhang L (2024) Attitude-position obstacle avoidance of trajectory tracking control for a quadrotor uav using barrier functions. Int J Syst Sci 55(16):3337–3354
8. Gao P, Jia C, Zhou A (2024) Encryption-decryption-based state estimation for nonlinear complex networks subject to coupled perturbation. Syst Sci Control Eng 12(1):2357796
9. Gao X, Deng F, Shang W, Zhao X, Li S (2024) Attack-resilient asynchronous state estimation of interval type-2 fuzzy systems under stochastic protocols. Int J Syst Sci 55(13):2688–2700
10. Gevorgyan Z (2022) SIoU loss: more powerful learning for bounding box regression, arXiv: 2205.12740
11. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) Yolox: exceeding yolo series in 2021, arXiv: 2107.08430
12. Han K, Wang Y, Tian Q, Guo J, Xu CJ, Xu C (2020) Ghostnet: more features from cheap operations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1580–1589, Virtually
13. Han K, Wang Y, Xu C, Guo J, Xu C, Wu E, Tian Q (2022) Ghostnets on heterogeneous devices via cheap operations. Int J Comput Vision 130(4):1050–1069
14. He K, Zhang X, Ren S, Sun J (Jun. 2016) Deep residual learning for image recognition, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778, Nevada, USA
15. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv:1704.04861
16. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H (2019) Searching for mobilenetv3, In: Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea, , pp 1314–1324
17. Huang Z, Yang S, Zhou M, Gong Z, Abusorrah A, Lin C, Huang Z (2022) Making accurate object detection at the edge: review and new approach. Artif Intell Rev 55(3):2245–2274
18. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141, Utah, USA
19. Jocher G (2022) YOLOv5 release v7.0, Available online: https://github.com/ultralytics/yolov5/tree/v7.0. Accessed on 31 Oct 2024
20. Li C, Li L, Jiang H,Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W, LiY, Zhang B, Liang Y, Zhou L, Xu X, Chu X, Wei XM, Wei XL (2022) YOLOv6: a single-stage object detection framework for industrial applications, arXiv: 2209.02976
21. Li C, Lyu H, Duan K (2025) A lightweight and efficient detector for concealed object inactive millimeter wave images. Knowl-Based Syst 310:112995
22. Li L-F, Hua Y, Liu Y-H, Huang F-H (2024) Study on fast fractal image compression algorithm based on centroid radius. Syst Sci Control Eng 12(1):2269183
23. Li Y, Chen Y, Dai X, Chen D, Liu M, Yuan L, Liu Z, Zhang L, Vasconcelos N (2021) Micronet: improving image recognition with extremely low flops, In: Proceedings of the IEEE/CVF international conference on computer vision, pp 468–477, Virtually
24. Liang Y, Tian L, Zhang X, Zhang X, Bai L (2024) Multi-dimensional adaptive learning rate gradient descent optimization algorithm for network training in magneto-optical defect detection. Int J Netw Dyn Intell 3(3):100016
25. Ma N, Zhang X, Zheng HT, Sun J (2018) Shufflenet v2: practical guidelines for efficient cnn architecture design, In: Proceedings of the European conference on computer vision, Munich, Germany, pp 116–131
26. Pan D (2024) String stable bidirectional platooning control for heterogeneous connected automated vehicles. Int J Netw Dyn Intell 3(4):100026

27. Qu B, Peng D, Shen Y, Zou L, Shen B (2024) A survey on recent advances on dynamic state estimation for power systems. Int J Syst Sci 55(16):3305–3321
28. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, California, USA, pp 658–666
29. Sandler M,Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Utah, USA, pp 4510–4520
30. Sikora P, Malina L, Kiac M, Martinasek Z, Riha K, Prinosil J, Jirik L, Srivastava G (2021) Artificial intelligence-based surveillance system for railway crossing traffic. IEEE Sens J 21(14):15515–15526
31. Siliang M, Yong X (2023) MPDIoU: a loss for efficient and accurate bounding box regression, arXiv: 2307.07662
32. Song W, Wang Z, Li Z, Han Q-L (2024) Particle-filter-based state estimation for delayed artificial neural networks: when probabilistic saturation constraints meet redundant channels. IEEE Trans Neural Netw Learn Syst 35(3):4354–4362
33. Song W, Wang Z, Li Z, Han Q-L, Yue D (2024) Maximum correntropy filtering for complex networks with uncertain dynamical bias: enabling componentwise event-triggered transmission. IEEE Trans Neural Netw Learn Syst 35(12):17330–17343
34. Su N, Huang Z,Yan Y, Zhao C, Zhou S (2022) Detect larger at once: large-area remote-sensiong image arbitraty-oriented ship detection. IEEE Geosci Remote Sens Lett, 19
35. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: Proceedings of the international conference on machine learning, California, USA, pp 6105–6114
36. Tan M, Le Q (2021) Efficientnetv2: smaller models and faster training, In: Proceedings of the international conference on machine learning, pp 10096–10106
37. Vasu PKA, Gabriel J, Zhu J,Tuzel O, Ranjan A (2023) Mobileone: an improved one millisecond mobile backbone. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Vancouver, Canada, pp 7907–7917
38. Wang CY, Bochkovskiy A, Liao HYM (2023) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Vancouver, Canada, pp 7464–7475
39. Wang CY, Mark Liao HY, Wu YH, Chen PY, Hsieh JW, Yeh IH (2020) CSPNet: a new backbone that can enhance learning capability of CNN, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, Washington, USA, pp 1571–1580
40. Wang H, Liu J, Zhao J, Zhang J, Zhao D (2025) Precision and speed: LSOD-YOLO for lightweight small object detection. Expert Syst Appl 269:126440
41. Wang W, Ma L, Rui Q, Gao C (2024) A survey on privacy-preserving control and filtering of networked control systems. Int J Syst Sci 55(11):2269–2288
42. Wang X, Kan X, Zhang Z, Sun W (2024) An automatic coke optical texture recognition method based on semantic segmentation model. Int J Netw Dyn Intell 3(4):100022
43. Wang Y, Wen C, Wu X (2024) Fault detection and isolation of floating wind turbine pitch system based on Kalman filter and multi-attention 1DCNN. Syst Sci Control Eng 12(1):2362169
44. Wei Y, Wang Y, Zhu B, Lin C, Wu D, Xue X, Wang R (2024) Underwater detection: a brief survey and a new multitask dataset. Int J Netw Dyn Intell 3(4):100025
45. Xu Y, Wen M, He W, Wang H, Xue Y (2024) An improved multi-scale and knowledge distillation method for efficient pedestrian detection in dense scense. J Real-Time Image Process, 21(4)
46. Xue Y, Li M, Arabnejad H, Suleimenova D, Jahani A, Geiger BC, Boesjes F, Anagnostou A, Taylor SJE, Liu X, Groen D (2024) Many-objective simulation optimization for camp location problems in humanitarian logistics. Int J Netw Dyn Intell 3(3):100017
47. Zhan Y, Yang R, You J, Huang M, Liu W, Liu X (2025) A systematic literature review on incomplete multimodal learning: techniques and challenges. Syst Sci Control Eng 13(1):2467083
48. Zhang H, Wang Y, Dayoub F, Sunderhauf N (2021) Varifocalnet: an iou-aware dense object detector, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8514–8523
49. Zhang Y, Liu G, Song X (2025) Unscented recursive three-step filter based unbiased minimum-variance estimation for a class of nonlinear systems. Int J Syst Sci 56(2):227–236
50. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T (2022) Focal and efficient iou loss for accurate bounding box regression. Neurocomputing 506:146–157
51. Zhao H, Gallo O, Frosio I, Kautz J (2017) Loss functions for image restoration with neural networks. IEEE Trans Comput Imaging 3(1):47–57
52. Zhao Y,Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J (2024) Detrs beat yolos on real-time object detection, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, USA, pp 16965–16974
53. Zheng Z,Wang P,Liu W, Li J, Ye R, Ren D (2020) Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence, New York, USA, vol. 34, no. 7, pp 12993–13000
54. Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, Zuo W (2022) Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans Cybern 52(8):8574–8586
55. Zhu S, Meng X (2021) Design of intelligent security early warning system for unguarded railway crossing in mining area. J Phys: Conf Ser 2029(1):012076