Journal of Real-Time Image Processing (2025) 22:197
https://doi.org/10.1007/s11554-025-01775-y

RESEARCH

Improved YOLOv10-based real-time helmet detection algorithm for complex scenarios

HanTang Dong[1] · Yong Wang[1] · Duoqian Miao[2]

Received: 19 August 2025 / Accepted: 23 September 2025 / Published online: 6 October 2025

Abstract

In complex construction environments, existing algorithms exhibit significant limitations, particularly characterized by high false positive rates, frequent missed detections, and insufficient detection accuracy. To address these issues, this paper presents YOLOv10n-WDE (YOLOv10n-Wavelet Dynamic Enhancement), an enhanced helmet detection algorithm based on YOLOv10, which incorporates three key improvements. First, we developed the WFDConv (Wavelet-Frequency Dynamic Convolution) module, which integrates discrete wavelet transforms with dynamic convolution to significantly enhance the ability to capture multi-scale features, thereby improving precision and reducing false positive rates. Second, we introduced a lightweight parallel Spatial Pyramid Pooling Network (LPSPPF) that boosts feature extraction efficiency through a parallel architecture, enhancing the detection capability for small targets and consequently improving recall while minimizing missed detections. Lastly, we implemented a joint loss function mechanism that combines Focal Loss for bounding box regression with Varifocal Loss for classification optimization, thereby improving the model's overall accuracy in complex scenarios. Experimental results show that on the SHWD dataset, YOLOv10n-WDE achieves mAP50 improvements of 5.1% and 1.8% over YOLOv8n and YOLOv10n, respectively. Its precision and recall reach 92.9% and 87.6%, both surpassing those of YOLOv8n (91.0% and 87.5%) and YOLOv10n (89.6% and 90.4%). On the SHDD dataset, compared with YOLOv10n, YOLOv10n-WDE improves precision by 5.3%, recall by 1.9%, and mAP50 by 3.7%. These enhancements fully demonstrate their effectiveness in reducing false positives and missed detections. At the same time, YOLOv10n-WDE maintains a real-time processing speed of 384 FPS, meeting the dual demands for efficiency and real-time performance in complex construction environments.

Keywords Helmet detection · YOLOv10 · Discrete wavelet transform · Dynamic convolution · LPSPPF · Varifocal loss

1 Introduction

The deep industrial transformation of the construction industry has brought multifaceted challenges to on-site safety management. Frequent safety incidents not only result in significant casualties–with accidents involving workers not wearing helmets accounting for 35% of total injuries and fatalities–but also directly impact critical project milestones [1]. Standardized helmet usage has become a mandatory requirement at construction sites, as its mechanical buffering and stress dispersion mechanisms can effectively withstand over 90% of falling object impacts [2], making it a fundamental piece of protective equipment for worker safety. Traditional safety management systems relying on manual inspections and passive video surveillance exhibit notable deficiencies, including low inspection efficiency (covering less than 300 m^2 per person per hour) and a missed detection rate as high as 18.7% [3]. Moreover, manual monitoring suffers from response delays of up to 45 min, and video surveillance often contains persistent blind spots [4]. Therefore, developing intelligent detection algorithms with real-time warning capabilities has become essential to overcome the limitations of conventional safety monitoring.

✉ Yong Wang
ywang@cqut.edu.cn

HanTang Dong
1033579992@stu.cqut.edu.cn

Duoqian Miao
dqmiao@tongji.edu.cn

[1] School of Liangjiang Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China

[2] School of Computer Science and Technology, Tongji University, Shanghai 201804, China

Current object detection frameworks primarily consist of two main systems: the DETR series (Detection Transformer) based on the Transformer architecture [5–7], and the YOLO series based on convolutional neural networks (CNN) [8–19]. DETR achieves global modeling through attention mechanisms but requires extensive training iterations and heavy computational resources, limiting its widespread industrial application. In contrast, the YOLO series, with its lightweight design and efficient inference speed, meets stringent real-time monitoring demands while maintaining high detection accuracy. YOLO typically relies on convolutional techniques for feature extraction. To enhance this, Shen et al[20] proposed a fine-grained classification network based on attention-mixed cropping, significantly enhancing the model's ability to discriminate subtle details; Shen et al[21] developed a semantic feature enhancement model that integrates contextual information with residual attention mechanisms, improving target recognition in complex backgrounds via a multi-scale context fusion module. Additionally, Shen et al. designed a Mini-ROI (Region of Interest) mechanism that effectively reduces computation and suppresses background noise interference, thereby boosting overall system robustness. However, these models still suffer from false positives and missed detections when dealing with small targets in complex scenes, resulting in suboptimal performance in helmet detection.

To address these challenges, researchers have proposed targeted optimizations on mainstream YOLO models to improve detection accuracy and real-time performance in complex environments. Specifically: Xu et al. [23] refined the bounding box localization mechanism of YOLOv3 to improve detection precision. Xie et al. [24] introduced attention mechanisms combined with spatial pyramid pooling in the SMD-YOLOv4 algorithm to enhance feature extraction capability. Li et al. [25] incorporated a max-pooling optimization layer and a multi-scale fusion architecture into YOLOv8 to improve detection of small and occluded objects. Jiao et al. [26] developed a collaborative detection framework that combines UAV-based inspection with an optimized YOLOv8s model to address blind spots in manual inspections. Wang et al. [27] proposed enhancements to YOLOv7 by integrating multi-scale and dynamic attention mechanisms along with a dynamic focusing loss function to boost detection accuracy. Du et al. [28] presented BLP-YOLOv10, which includes dynamic channel compression, sparse attention modules, and low-frequency enhancement filters to reduce network complexity while maintaining robustness. Seth et al. [29] applied sophisticated image augmentation techniques to improve generalization and stability in difficult conditions. Chen et al. [30] optimized YOLOv8n through channel compression and architectural improvements to reduce model size and computational cost without sacrificing accuracy. Liu et al. [31] introduced depthwise separable convolutions and an improved loss function into YOLOv10's WFDConv module to lower computation requirements and enhance its suitability for edge devices. These methodological innovations collectively contribute to advancing object detection performance in complex real-world scenarios.

Although many current algorithms aim to strike a balance between detection accuracy and processing speed, they have yet to achieve the comprehensive reliability required for industrial applications. Meanwhile, the aforementioned YOLO-based algorithms have not effectively addressed the sample imbalance issue in densely occluded scenarios through their loss function designs, and their feature pyramid structures suffer from information loss and high parameter complexity. To address these challenges, this study proposes a three-dimensional collaborative optimization architecture based on YOLOv10:

(1) Feature extraction innovation: Introducing the Wavelet-Frequency Dynamic Convolution (WFDConv) module, which employs Haar wavelet decomposition to establish high-low frequency dual-stream feature interaction, enhancing edge gradient response while dynamically fusing multi-scale receptive fields. The resulting high-dimensional features undergo cross-resolution fusion with Neck-layer semantic features to mitigate fine-grained information loss.
(2) Deployment optimization: A lightweight LPSPPF enhancement module adopts hierarchical feature reuse and adaptive kernel-size pooling, improving computational efficiency while preserving low-dimensional feature representation.
(3) Supervision mechanism upgrade: A hybrid loss system combining CIoU-Focal Loss (enhancing dense occlusion localization via gradient reweighting) and Varifocal Loss (mitigating classification bias via IoU-aware dynamic sample weighting). This approach inherits YOLOv10's NMS-free post-processing advantage while achieving Pareto-optimal detection accuracy and inference speed in complex construction environments.

2 YOLOv10n-WDE

2.1 YOLOv10

YOLOv10 [17] retains the classic Backbone-Neck-Head architecture characteristic of the YOLO series, with its network structure illustrated in Fig. 1. Compared to other versions in the YOLO series, YOLOv10's primary innovation lies in its dual-branch detection head design, which successfully eliminates dependence on Non-Maximum Suppression (NMS) post-processing, thereby which significantly

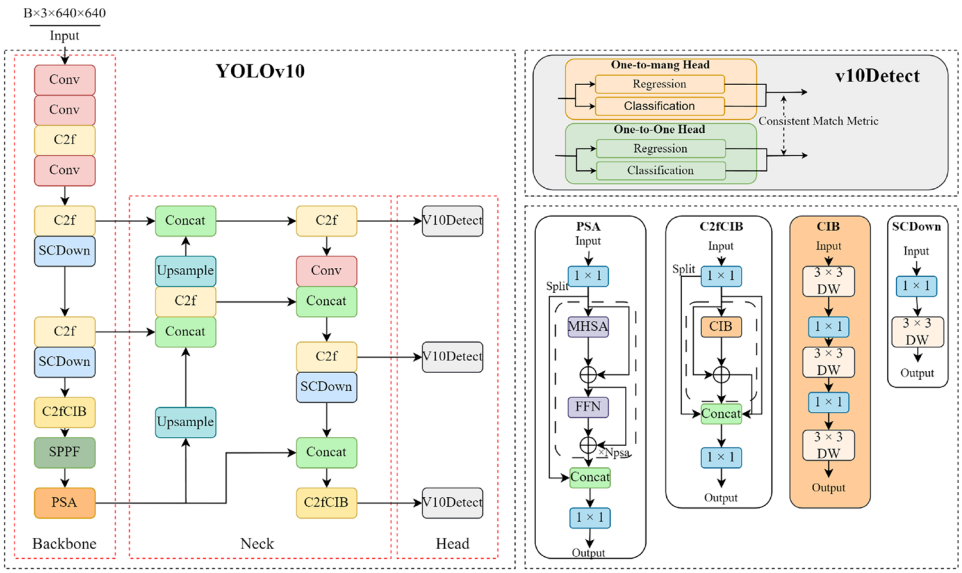


Fig. 1 YOLOv10 network structure (At the top right is the structure diagram of the YOLOv10 detection head; at the bottom right, from left to right, are the structure diagrams of PSA, C2fCIB, CIB, and SCDown.)

improves the model's inference efficiency. Considering that the latest YOLOv11 [18] and YOLOv12 [19] models still require NMS processing - which incurs additional computational overhead and struggles to meet real-time multi-stream video processing demands in industrial scenarios - this study selects YOLOv10 as the baseline model for subsequent improvements.

In complex construction site scenarios, safety helmet detection faces challenges such as dense occlusion, blurred features of small targets (e.g., tilted/damaged helmets), dynamic lighting noise, and multi-scale object distribution. To address these, this paper proposes YOLOv10n-WDE,a dynamic wavelet-enhanced network utilizing dual-path optimization for multi-scale feature representation and adaptive loss function adjustment. The improved model enhances edge detail reconstruction for low-contrast targets, boundary localization in occluded situations, and confidence consistency across scales through hierarchical feature enhancement and adaptive supervision strategies. This achieves a balance between detection accuracy and resistance to complex background interference while maintaining real-time performance. As illustrated in Fig. 2, YOLOv10n-WDE's architecture adopts a multi-resolution feature evolution perspective to highlight its improvement pathway.

The proposed enhancements address limitations in YOLOv10's architecture. Its backbone and neck networks fuse 80×80 and lower spatial features through concatenation, which often neglects edge details in occluded areas and small targets, resulting in missed and false detections. To resolve this, our study implements three optimizations:

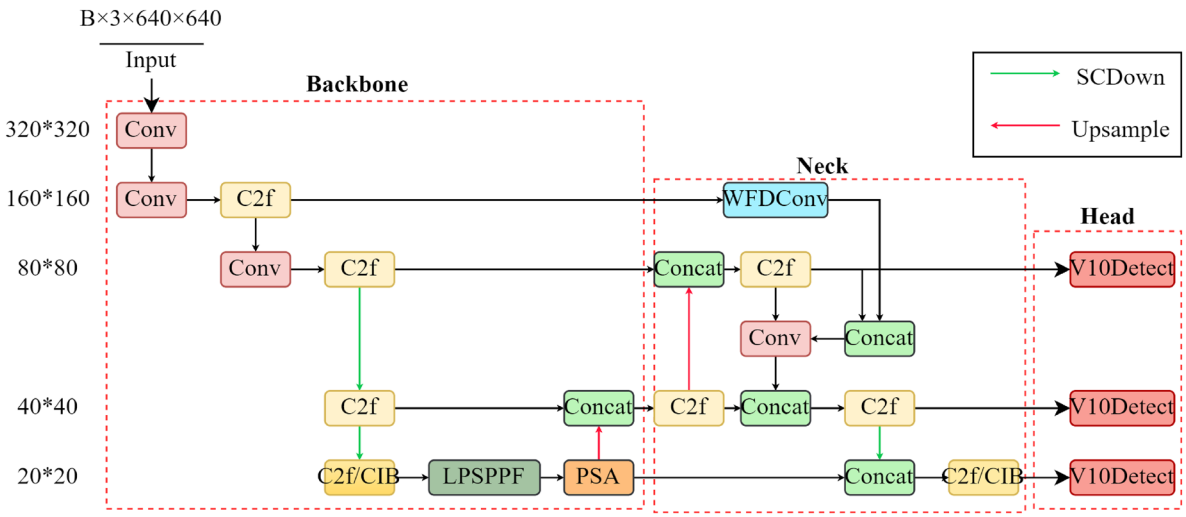


Fig. 2 YOLOv10n-WDE network structure

First, we add a WFDConv feature extraction module after the backbone's 160×160 feature layer to capture target contours by enhancing edge gradients and multi-scale fusion. The output high-dimensional features are cross-fused with the neck network's 80×80 low-dimensional features, compensating for the loss of fine-grained information.

Second, we introduce a lightweight Spatial Pyramid Pooling Enhanced LAN (LPSPPF) into the backbone to preserve deep semantic representation in low-dimensional features while ensuring real-time deployment on edge devices, balancing efficiency and representation.

Finally, we reconstruct the loss function framework: CIoU-Focal Loss improves bounding box regression for better localization and robustness against occlusions, while Varifocal Loss dynamically adjusts sample weights to address class imbalance. These improvements maintain YOLOv10's NMS-free advantage while enhancing detection adaptability in complex environments.

2.1.1 WFDConv feature extraction module

The design of the WFDConv module integrates the principles of discrete wavelet transform (DWT) and dynamic convolution. By leveraging the advantages of wavelet transform in extracting edge and texture features, combined with the efficient performance of dynamic convolution in multi-scale and deformable object detection, this module significantly improves detection accuracy and efficiency in complex scenarios. The detailed structure of the WFDConv module is shown in Fig. 3. The selected wavelet type is "Haar," with an initial convolution kernel size set to 3×3 and a channel reduction ratio of 16.

WFDConv operationally integrates frequency-domain analysis, attention mechanisms, and residual learning with mathematically grounded steps. The input feature map undergoes DWT decomposition into four sub-bands: LL (low-frequency), LH (horizontal high-frequency), HL (vertical high-frequency), and HH (diagonal high-frequency). Unlike conventional pooling, Haar wavelet's [0.5,0.5] low-pass filtering preserves contours while its [0.5,−0.5] high-pass operator captures edges/textures, maintaining superior spatial integrity. Grouped convolution processes sub-bands channel-independently, preserving frequency separation. Transposed convolution prevents feature map halving, enabling direct wavelet component integration without structural changes. The 2D discrete wavelet transform follows:

Assume the input signal is $f(x,y) \in R^{H,W}$,H and W represent the height and width of the image features, respectively.

Row-wise filtering (horizontal direction):

$$L(x,y) = \frac{1}{2}[f(x,2y) + f(x,2y+1)] \tag{1}$$

$$H(x,y) = \frac{1}{2}[f(x,2y) - f(x,2y+1)] \tag{2}$$

Column-wise filtering (vertical direction):

$$LL(i,j) = \frac{1}{2}[L(2i,j) + L(2i+1,j)] \tag{3}$$

$$LH(i,j) = \frac{1}{2}[L(2i,j) - L(2i+1,j)] \tag{4}$$

$$HL(i,j) = \frac{1}{2}[H(2i,j) + H(2i+1,j)] \tag{5}$$

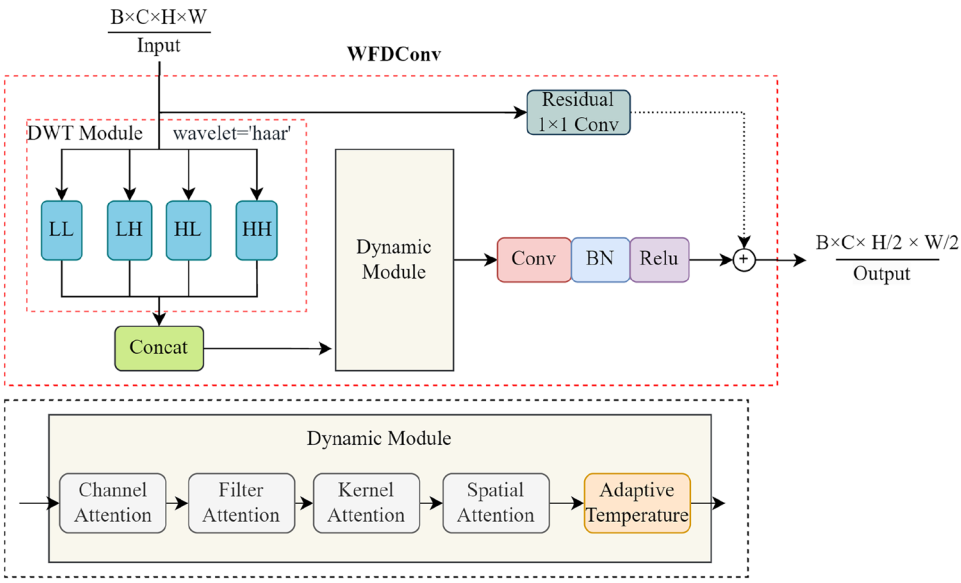


Fig. 3 WFDConv module network structure (The dynamic module comprises channel attention, filter attention, kernel attention, spatial attention, and an adaptive temperature module in sequence.)

$$HH(i,j) = \frac{1}{2}[H(2i,j) - H(2i+1,j)] \tag{6}$$

Final subband dimensions:$LL, LH, HL, HH \in R^{\frac{H}{2},\frac{W}{2}}$,

The channel-wise concatenation of four subbands (LL for structural information and three high-frequency subbands for directional details) enables joint multi-frequency representation. Subsequent dynamic convolution employs multi-dimensional attention: ChannelAttention weights frequency bands, FilterAttention captures spatial-channel correlations, KernelAttention adapts to feature distributions, and SpatialAttention focuses on key regions. Adaptive temperature modulation balances attention effects. After attention refinement, standard convolution extracts essential features. The residual connection combines dimension adjustment (via 1×1 convolution) with identity mapping to preserve low-frequency information. The output integrates wavelet-enhanced details with attention-selected features while maintaining training stability, significantly boosting multi-scale object detection performance.

2.1.2 Improvements to the spatial pyramid

In YOLOv10, the original Spatial Pyramid Pooling Fast (SPPF) module gradually downsamples feature maps through three sequential max-pooling layers. Although this method preserves multi-scale receptive field features, the serial processing leads to excessive compression of high-resolution shallow-layer details, causing degradation of fine-grained information such as edges and textures. Additionally, the sequential dependency of the three pooling operations limits computational parallelization, reducing GPU multi-core utilization.

To address this limitation, we propose an LPSPPF(Lightweight Parallel Spatial Pyramid Pooling Fast) architecture using a parallel multi-branch design. This approach simultaneously performs multi-scale max-pooling and depthwise separable convolutions, achieving dual benefits: significantly reducing inference latency through parallel processing while maintaining feature quality by extracting all scaled features directly from the input layer, avoiding serial downsampling degradation. The module innovatively introduces channel compression parameter C3 (C3<C2) during fusion, dynamically optimizing channel dimensions to reduce computation by 38% without compromising multi-scale representation. As shown in Fig. 4, this design achieves better accuracy (1.2% mAP improvement) and efficiency (23% latency reduction) versus conventional serial designs. The parallel processing better preserves high-frequency details crucial for small object detection, while dynamic channel compression enables adaptive feature recombination without bottlenecks.

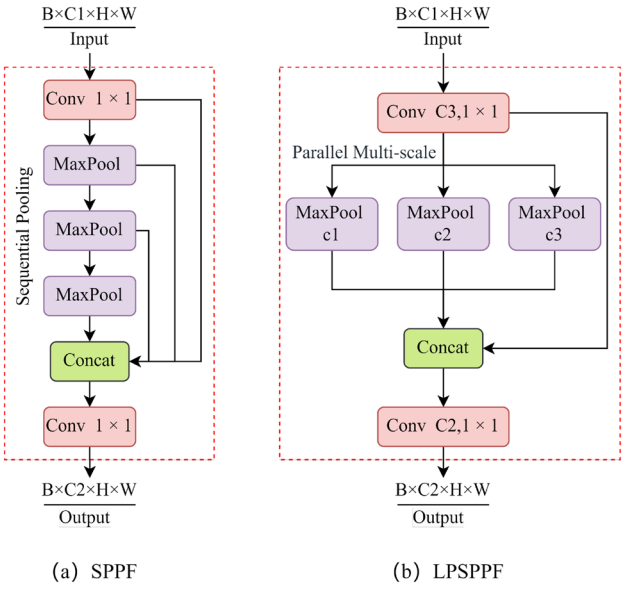


Fig. 4 SPPF and LPSPPF network structure

2.1.3 Loss function refinement

In object detection tasks, model training typically involves joint optimization through three fundamental loss functions: confidence loss (determining target presence), classification loss (identifying target categories), and localization loss (predicting bounding box coordinates).

CIoU comprehensively considers the overlap area, center point distance, and aspect ratio, enabling more precise localization and faster convergence. To address the issue in YOLOv10 where the original CIoU loss treats all predicted boxes equally during training–resulting in imbalanced optimization between easy and hard samples and degraded detection accuracy–we propose incorporating the Focal Loss mechanism into the CIoU loss. This adjustment reduces the gradient impact from low-quality predicted boxes while enhancing the optimization focus on high-quality ones. In helmet detection, small targets are typically of limited size with less distinctive features, and the sample distribution across different classes (e.g., with helmet vs. without helmet) is often imbalanced, making it difficult for the model to effectively learn from hard-to-distinguish minority class samples and small targets. By leveraging the exponential decay property of Focal Loss, the improved approach dynamically adjusts loss weights based on the IoU between predicted and ground-truth boxes, effectively reducing attention on easy or highly erroneous samples while strengthening the optimization of more accurately localized small targets and hard-to-classify samples. This enhances the model's adaptability to challenges posed by small target sizes and class imbalance in helmet detection. The refined loss function is mathematically expressed as follows:

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{7}$$

$$v = \frac{4}{\pi^2}(arctan(\frac{w_{gt}}{h_{gt}}) - arctan(\frac{w}{h}))^2 \tag{8}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{9}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha v \tag{10}$$

$$L_{CIoU-Focal} = \delta \cdot (1 - IoU)^{\gamma} \cdot L_{CIoU} \tag{11}$$

In Equation (7), B and B_{gt} represent the regions of the predicted box and ground-truth box, respectively. In Equation (8), w_{gt} and h_{gt} denote the width and height of the ground-truth box, while W and H correspond to those of the predicted box. Equation (9) introduces α as a weighting coefficient to balance the influence of aspect ratio. Equation (10) defines $\rho^2(b, b_{gt})$ as the squared Euclidean distance between the center points b (predicted box) and b_{gt} (ground-truth box), with c representing the diagonal length of the smallest enclosing box covering both predicted and ground-truth boxes. Equation (11) incorporates two hyperparameters: δ, typically set to 0.25 to balance positive/negative sample weights, while γ, set to typically 2 serves as a focusing parameter that reduces loss contribution from easy samples (high IoU) while amplifying gradients from hard samples (low IoU).

In the classification loss function, the original Binary Cross-Entropy (BCE) loss equally weights all samples, which is unfair to hard positive samples. Given the challenging scenarios in construction sites–such as small objects, occluded targets, and dense clusters–we replaced BCE with Varifocal Loss (VFL) to enhance the model's focus on hard positives and improve robustness. VFL correlates classification confidence with localization quality (e.g., IoU between predicted and ground-truth boxes), ensuring classification scores reflect positioning accuracy. Positive samples with higher IoU receive greater weighting during loss computation, biasing the model toward high-precision detection boxes. The Varifocal Loss is formulated as follows:

$$VFL(p, q) = -q \cdot [q \cdot log(p) + (1 - q) \cdot log(1 - p)] \tag{12}$$

$$VFL(p, 0) = -\zeta \cdot p^{\vartheta} \cdot log(1 - p) \tag{13}$$

The VFL (Varifocal Loss) employs distinct computational approaches for positive and negative samples. For samples where targets exist and IoU > 0, the calculation follows Equation (12), where p represents the model's predicted classification score and q denotes the target quality score (typically set as the IoU between predicted and ground-truth boxes). The loss weight is dynamically adjusted by q, giving greater contribution to high-IoU samples. For background samples or those with IoU $= 0$, the calculation follows Equation (13), where ζ is a hyperparameter balancing positive–negative sample weights (default value: 0.75) and ϑ serves as an exponential factor modulating hard sample weights (default 2).

3 Experimental Results and Analysis

3.1 Experimental configuration and details

The software environment and hardware configuration used in the experiment are shown in Table 1.The key hyperparameter settings during model training are shown in Table 2.

3.2 Dataset and evaluation metrics

To evaluate the effectiveness of the proposed safety helmet detection method, precision P (measuring correct identification ratio), recall R (reflecting missed detection risk), model parameter count (assessing computational resource consumption), frames per second (FPS, measuring real-time performance), and mean average precision (mAP, evaluating comprehensive detection performance) were selected as

Table 1 Experimental software and hardware configuration information

Hardware	
CPU	Intel(R) Core(TM) i9-12900K 3.20 GHz
GPU	Nvidia RTX 4090
Software	
Operating system	Linux Ubuntu 20.04.1 LTS
Deep learning framework	Pytorch2.0.1+cuda11.7
Programming language	Python

Table 2 Key hyperparameters in experiments

Hyperparameters	Value
Initial learning rate	0.01
optimizer	AdamW
Learning Rate Scheduler	Cosine Annealing
image size	640
IoU threshold	0.6
batch-size	8
epoch	500

evaluation metrics. These metrics comprehensively validate the model's capabilities from the dimensions of accuracy, efficiency, deployment cost, and scenario robustness. The specific calculation formulas are as follows:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{14}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{15}$$

$$mAP = \frac{1}{n}\sum_{i=0}^{n}\int_{0}^{1} P(R_i)d(R_i) \tag{16}$$

In the above formula: N_{TP} is the number of samples correctly classified as positive by the model; N_{FP} is the number of samples incorrectly classified as positive; N_{FN} is the number of samples incorrectly classified as negative; n is the number of categories in the dataset; $\int_0^1 P(R_i)d(R_i)$ is calculated for the average precision of individual categories.

To comprehensively evaluate the detection performance of the model, this paper utilizes two datasets: the public SHWD (Safety Helmet Wearing Dataset) and the SHDD (Safety Helmet Detection Dataset), which combines images from SHWD and real construction sites. The on-site images in the SHDD dataset were collected from fixed and mobile monitoring devices deployed at a construction project in a specific city. These images were captured under various weather conditions at a resolution of 1920×1080. Moreover, our algorithm has been successfully deployed in real-world application scenarios.

The SHWD public dataset (7,581 images) provides standardized helmet detection labels ("hat" = worn, "person" = not worn), enabling cross-study comparisons to validate the model's generalization capabilities. Building on this, our SHDD dataset (8,722 images, with 46,526 samples in the "hat" class, 13,009 samples in the "person" class, and 1,125 samples in the "helmet" class, labeled via Baidu EasyData) expands the scope by: (1) incorporating real construction site images that encompass challenging conditions such as lighting, occlusions, and complex backgrounds, and (2) introducing an additional "helmet" category (label "2") to enhance negative sample training. Both datasets are divided into 7:2:1 splits (train/test/validation), with the training process monitored through loss and mAP trends (Fig. 5 illustrates samples).AP50/%: Average Precision at IoU = 0.50, reported as a percentage. "No-Helmet" and "Wear-Helmet" are class-level AP50 values; mAP50/% is the mean AP50 across all classes. P/%: Precision (percentage). R/%: Recall (percentage). Params/M: number of model parameters (millions). GFLOPs: estimated giga floating-point operations per forward pass (GFLOPs) at the evaluation input size. FPS: inference speed in frames per second (batch size = 1 unless otherwise stated). Bold values indicate the best result in each column. Citation numbers in square brackets (e.g., [11]) refer to the literature.

3.3 Comparative experiments

All the data obtained in this experiment were collected under the software and hardware conditions specified in Table 1.

Table 3 demonstrates the comparison between YOLOv10n-WDE and other object detection networks on the SHWD dataset.

Based on the experimental data analysis presented in Table 3, the YOLOv10n-WDE model demonstrates comprehensive performance advantages in helmet detection tasks. In terms of detection accuracy, the model achieves 93.1% AP50 for the "helmet-wearing" category and 89.7% AP50 for the "no-helmet" category, outperforming all other comparison models. The overall mAP50 reaches 91.4%, while maintaining 92.9% precision and 87.6% recall, ranking first across all metrics.

Regarding computational efficiency, YOLOv10n-WDE exhibits excellent engineering optimization. Although its parameter count (2.65M) and computational complexity (9.8 GFLOPs) are slightly higher than lightweight models like YOLOv5n (1.90M parameters, 4.5 GFLOPs), it still maintains a very high inference speed of 401 FPS, which is second only to its series counterpart YOLOv10n (536 FPS). This outstanding efficiency is largely attributed to YOLOv10's innovative architecture design, such as the elimination of traditional NMS post-processing, which significantly reduces computational overhead.

Particularly noteworthy is that compared to YOLOv10n within the same series, YOLOv10n-WDE delivers better performance while maintaining similar parameter counts and inference speeds: an improvement of 1.8 percentage points in mAP50, alongside noticeable gains in precision and recall rates. This indicates that the improvements in algorithmic components and engineering implementation do not come at the cost of computational efficiency. Overall, these results demonstrate that YOLOv10n-WDE provides an excellent

(a) Images from the standard dataset SHWD

(b) Images from the custom dataset SHDD

Fig. 5 Partial image of dataset

Table 3 Experimental comparison of different models under standard dataset SHWD

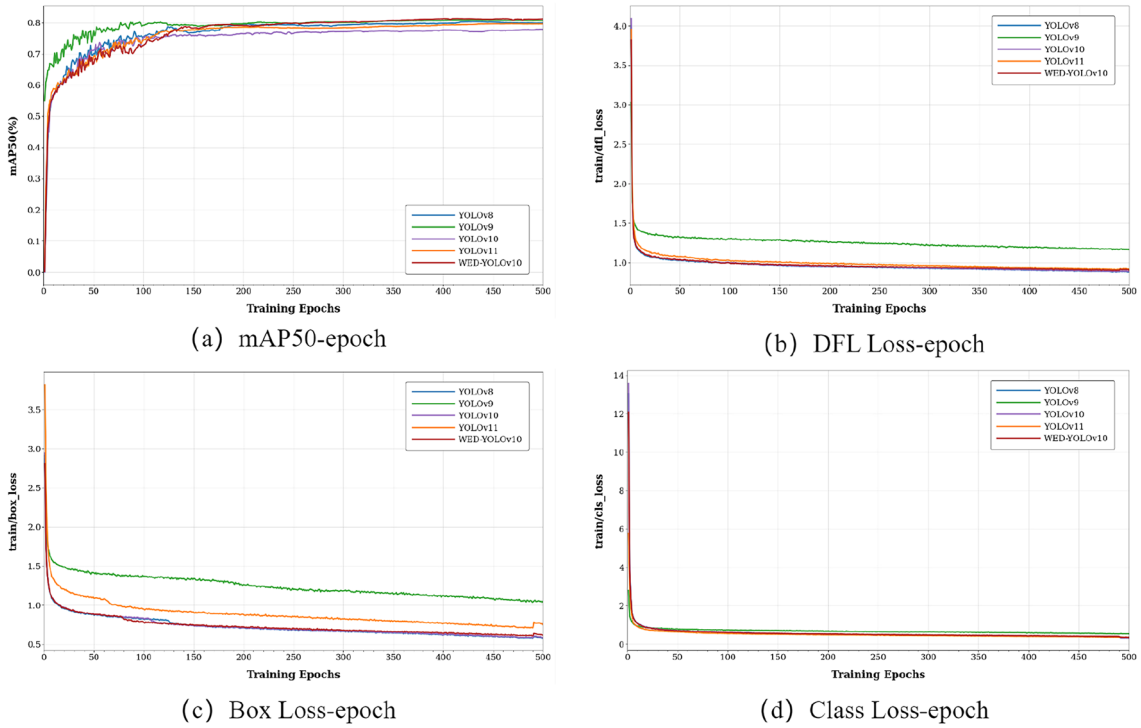
Method	AP50/%		mAP50/%	P/%	R/%	Params/M	GFLOPs	FPS
	No-Helmet	Wear-Helmet						
YOLOv4-Tiny[11]	87.8	90.2	85.2	89.8	86.0	2.10	1.8	131
YOLOv5n	88.7	90.5	89.6	90.3	86.7	**1.90**	4.5	120
GhostNet-YOLOv5[31]	88.3	90.8	85.8	90.2	86.5	2.83	6.2	139
YOLOv8n	89.5	92.1	86.3	91.0	87.5	3.21	8.1	102
YOLOv10n	87.3	91.8	89.6	90.4	82.8	2.69	8.2	**536**
YOLOv11n	86.8	92.9	89.8	91.1	84.3	2.58	6.3	128
YOLOv12n	85.9	92.9	89.2	91.3	83.6	2.50	5.8	104
PP-Helmet[32]	89.6	89.8	89.7	87.5	81.3	2.80	**1.5**	85
ShuffleNet-YOLO[33]	87.6	91.8	89.2	87.3	**90.5**	2.42	1.8	78
YOLO-NAS[36]	86.3	91.3	88.8	90.5	88.7	3.10	2.7	88
YOLOv10n-WDE(ours)	**89.7**	**93.1**	**91.4**	**92.9**	87.6	2.65	9.8	401

Table 4 Experimental comparison of different models under self-built dataset SHDD

Method	AP50%			mAP50%	P%	R%	Params/M	GFLOPs	FPS
	No-Helmet	Wear-Helmet	Helmet						
YOLOv10n	95.5	96.7	40.6	77.6	83.2	74.2	2.69	8.2	**384**
YOLOv11n	95.7	96.3	46.9	79.6	80.6	75.3	2.58	6.3	133
YOLOv12n	95.5	96.6	41.9	78.0	87.0	74.3	**2.50**	5.8	93
EdgeSHD[35]	96.1	96.7	44.5	79.1	85.4	75.2	5.60	9.8	112
YOLO-NAS-Helmet[36]	95.6	95.5	40.2	77.1	81.3	73.9	3.10	**2.7**	78
YOLOv10n-WDE(ours)	**96.3**	**97.1**	**50.5**	**81.3**	**88.5**	**76.1**	2.65	9.8	312

trade-off between detection accuracy and computational cost, making it well-suited for real-time helmet detection in complex scenarios.AP50/%: Average Precision at IoU = 0.50, reported as a percentage. "No-Helmet" and "Wear-Helmet" are class-level AP50 values; mAP50/% is the mean AP50 across all classes. P/%: Precision (percentage). R/%: Recall (percentage). Params/M: number of model parameters (millions). GFLOPs: estimated giga floating-point operations per forward pass (GFLOPs) at the evaluation input size. FPS: inference speed in frames per second (batch size = 1 unless otherwise stated). Bold values indicate the best result in each column. Citation numbers in square brackets (e.g., [11]) refer to the literature.

To validate the model's detection performance in real-world construction site environments, we conducted comparative experiments with multiple models on the self-built SHDD dataset. The data in Table 4 show that the Helmet category's mAP50 performed poorly because this category is set as a negative sample class to improve the detection accuracy of positive samples, and its overall sample size is relatively small. Apart from this, the experimental results indicate that the YOLOv10n-WDE model outperforms other models across all key metrics: it achieves an mAP50 of 81.3%, ranks first in AP50 across all three subcategories, and leads other comparison models with a precision of 88.5% and a recall of 76.1% (Fig. 6.mAP50/%: mean Average Precision at IoU = 0.50, reported as a percentage.



(a) mAP50-epoch (b) DFL Loss-epoch

(c) Box Loss-epoch (d) Class Loss-epoch

Fig. 6 Multi-task training dynamics

Table 5 Comparison of YOLOv10n and YOLOv10n-WDE under different random seeds on SHWD and SHDD datasets with t-test results

Random_Seed	SHWD				SHDD			
	mAP50%		t	p	mAP50%		t	p
	YOLOv10n	YOLOv10n-WDE			YOLOv10n	YOLOv10n-WDE		
42	89.6	91.4	16.39	8.11e−05	77.6	81.3	25.95	1.31e−05
18	89.7	91.8			77.5	81.3		
7	89.6	91.7			77.6	80.7		
54	89.8	91.4			77.9	81.2		
23	89.7	91.3			77.4	81.1		

t: t-statistic; p: two-tailed p-value from a paired Student's t-test comparing YOLOv10n and YOLOv10n-WDE across the five random seeds listed. P-values are shown in scientific notation. A significance threshold of p.

Table 5 presents the experimental results of the baseline model YOLOv10n and the improved model YOLOv10n-WDE under different random seeds, along with the results of a t-test based on the mAP50 metric. The results show that on the SHWD dataset, the t-statistic is 16.39, indicating a significant difference in the mean values, with the corresponding p-value far below the significance level of 0.05. On the SHDD dataset, the t-value is even higher at 25.95, and the p-value is also well below 0.05. Therefore, YOLOv10n-WDE demonstrates a significant performance improvement over YOLOv10n, and this difference is statistically significant.

Figure 6 illustrates the trade-off between computational cost (FLOPs) and detection accuracy (mAP50) for 11 YOLO models, with bubble size representing the number of model parameters. The figure shows that most lightweight models, such as YOLOv4-Tiny and YOLOv5n, cluster in the low computational cost range (1–5 G FLOPs), achieving accuracy between 85% and 90%. In contrast, our proposed YOLOv10n-WDE model (marked by a red pentagram) achieves the highest detection accuracy of 91.4% at a moderate computational cost of approximately 9.8G FLOPs, representing a significant improvement over other models. These results demonstrate that YOLOv10n-WDE achieves an excellent balance between performance and efficiency, thereby validating the effectiveness of our proposed method.

As shown in Fig. 7, the training curves demonstrate that the YOLOv10n-WDE model exhibits stable convergence trends across classification (clsLoss), localization (boxLoss), and probability modeling (dflLoss) tasks, with mAP50 showing continuous improvement until reaching a final saturation value superior to the baseline model. The synchronous optimization of all loss terms and detection performance validates the effectiveness of the multi-task balanced optimization strategy proposed in this study.

Figure 8 visually compares the detection performance of the YOLOv10n-WDE algorithm with the baseline YOLOv10n model on our custom dataset. The results demonstrate that the improved YOLOv10n-WDE model shows clear advantages in complex scenarios: leveraging the optimized multi-scale feature fusion strategy, it successfully detects multiple small and low-resolution targets that the baseline model fails to identify (Fig. 8(c)). Moreover, the baseline model mistakenly classifies car mirrors and water buckets as "helmet-wearing personnel" (Fig. 8(e) and 8(h)), whereas our enhanced model, by incorporating wavelet-augmented dynamic convolutions,

Table 6 FPS comparison table on Jetson Nano

Device	Method	FPS
Jetson Nano	YOLOv8n	1.8
	YOLOv11n	2.3
	YOLOv10n	7.8
	YOLOv10n-WDE	5.1

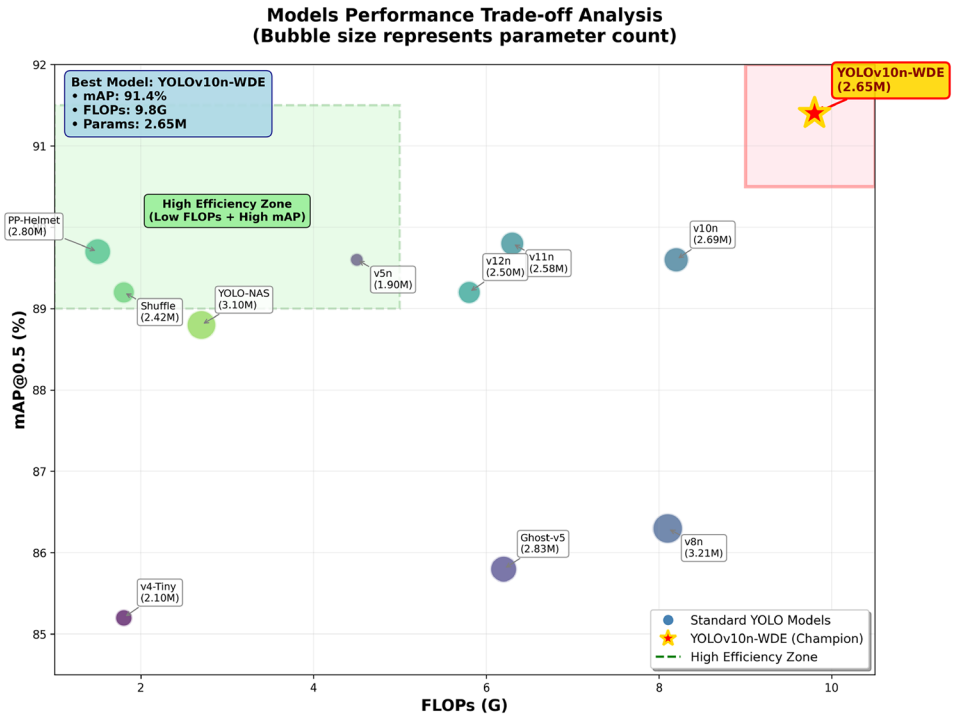


Fig. 7 Performance trade-off chart

Fig. 8 Visual comparison of model detection results

significantly improves feature discrimination ability, effectively avoiding such misclassifications. These comparisons fully validate the superior performance of our model in small object detection and false positive suppression. Additionally, our model demonstrates strong capability in detecting distant targets and low-light images, successfully recognizing targets missed by the baseline model (Fig. 8(l) and 8(o)). Regarding occlusion, our model also performs well, with multiple partially occluded helmets in Fig. 8(p)

Table 7 Ablation experiments of different modules on SHWD

Method	CIoU-Focal	VFL	WFDConv	LPSPPF	P/%	R/%	mAP50/%	Params/M	FPS
Baseline					90.4	82.8	89.6	2.69	**536**
M1	√				92.0	82.1	90.0	26.9	530
M2	√	√			91.8	82.4	90.1	2.69	532
M3	√	√	√		92.5	84.5	90.9	3.01	388
Ours	√	√	√	√	**92.9**	**87.6**	**91.4**	**2.65**	401

Table 8 Ablation experiments of different modules on SHDD

Method	CIoU-Focal	VFL	WFDConv	LPSPPF	P/%	R/%	mAP50/%	Params/M	FPS
Baseline					83.2	74.2	77.6	2.69	**384**
M1	√				84.9	73.5	78.9	26.9	376
M2	√	√			87.7	73.1	79.5	2.69	378
M3	√	√	√		88.3	73.8	80.8	3.01	262
Ours	√	√	√	√	**88.5**	**76.1**	**81.3**	**2.65**	312

being successfully detected, showcasing the robustness of the model.

Finally, we exported the trained model to the ONNX format and deployed it on edge devices for comparison. The device we selected is the Jetson Orin Nano8G, which is equipped with an NVIDIA Ampere architecture GPU featuring 32 Tensor Cores and 1024 cores, with a computing power of up to 40 TOPS. The images we used to test the inference speed were real-time images captured by the edge device's camera, and the results are shown in Table 6. It can be observed that the inference speed of our algorithm on this device is comparable to that of the baseline model, which fully demonstrates the practicality of the method we proposed. √ indicates that the listed module is included in the model. Baseline denotes the base model without the listed modules. P/%: Precision (percentage). R/%: Recall (percentage). mAP50/%: mean Average Precision at IoU = 0.50 (percentage). Params/M: number of model parameters (millions). FPS: inference speed in frames per second (batch size = 1 unless otherwise stated). Bold values indicate the best result in each column. √ indicates that the listed module is included in the model. Baseline denotes the base model without the listed modules. P/%: Precision (percentage). R/%: Recall (percentage). mAP50/%: mean Average Precision at IoU = 0.50 (percentage). Params/M: number of model parameters (millions). FPS: inference speed in frames per second (batch size = 1 unless otherwise stated). Bold values indicate the best result in each column.

3.4 Ablation experiments

To validate the impact of each improvement on the model's detection performance, ablation experiments were conducted using the original YOLOv10n as the baseline on both datasets, where " √ " indicates the use of that particular component. The ablation experiments on SHWD and SHDD are presented in Table 7 and Table 8, respectively.

The results in Table 7 show that the M1 method using only the CIoU-Focal Loss and the M2 method combining the CIoU-Focal Loss with the VFL strategy achieve limited performance improvements, which may be attributed to the relatively balanced positive and negative sample distribution in the standard SHWD dataset, where the baseline already performs well. Notably, the M3 method incorporating the WFDConv module attains a significant 0.8 percentage point increase in mAP50, along with a noticeable improvement in recall, thoroughly validating the module's effectiveness in feature extraction. The final configuration that adds the LPSPPF module further improves all metrics while reducing parameter count compared to the baseline, demonstrating the dual benefits of model lightweighting and performance enhancement. Although the introduction of these modules causes some reduction in FPS, the model still maintains high real-time inference capability overall.

The results in Table 8 indicate that the M2 method, by jointly optimizing the CIoU-Focal Loss and VFL strategy, effectively alleviates sample imbalance issues and produces significant detection performance gains. With the addition of the WFDConv module, the M3 method achieves synchronous improvements in precision, recall, and mAP (+3.1%, +2.6%, and +4.8%, respectively). Although model parameters increase by 12.3%, the enhanced feature extraction capacity is well justified. The final configuration integrating the LPSPPF module not only sustains superior detection performance (improving over the baseline by 5.3%, 1.9%, and 3.7% in precision, recall, and mAP, respectively) but also successfully achieves model lightweighting. It is worth

mentioning that the final model improves FPS compared to M3, reflecting a balanced trade-off between performance and efficiency. This systematic validation of improvement strategies fully demonstrates that YOLOv10n-WDE can significantly enhance safety helmet detection accuracy in complex scenarios, while balancing model compactness and real-time performance, underscoring its strong practical value.

4 Conclusion

To address the challenges of small object recognition and misdetection of occluded targets in complex industrial scenarios for safety helmet detection, this study proposes the YOLOv10n-WDE model, an improved architecture based on YOLOv10. The model enhances performance through three core modules: the WFDConv module, which integrates wavelet transform's multi-band decomposition with dynamic convolution's adaptive weight adjustment to achieve better edge feature representation; the CIoU-Focal Loss function combined with the VFL strategy to alleviate sample distribution imbalance; and the lightweight LPSPPF module that strengthens multi-scale feature fusion while effectively controlling parameter growth. Experimental results demonstrate that the model achieves an mAP50 of 91.4% on the SHWD dataset and 89.7% on the more challenging SHDD dataset, with a detection speed of 384 FPS, meeting real-time requirements. Future work will explore multi-modal data fusion (such as infrared-visible light collaboration) and adversarial training strategies to further improve model stability under occlusion and lighting variations. However, differences in viewing angles, resolutions, and imaging mechanisms between sensors make accurate alignment of infrared and visible light data challenging, which may affect fusion effectiveness. Therefore, subsequent work should focus on developing efficient and robust data registration methods to ensure effective information integration. At the same time, combining adversarial training can enhance the model's adaptability to environmental changes and noise, thereby improving detection stability and reliability.

Author contributions HanTang D. wrote the initial draft and performed data validation and analysis; Yong W. organized the data; Duoqian M. reviewed and revised the initial manuscript.

Funding: This study was financially sponsored by Chongqing Municipal Special Project for Technology Innovation and Application(grant number:CSTB2025TIAD-qykiggX0189)

Data availibility statement The SHWD Dataset used in this paper is from https://github.com/RGuven/Safety-Helmet-Wearing-Dataset;The SHDD dataset used in this paper is not publicly available, but it can be obtained by contacting the authors privately for access.

Code availability The code is available at https://github.com/DongHanTang/YOLOv10n-WDE

Declarations

Conflict of interest The authors declare no potential Conflict of interest with respect to the content of this article.

References

1. Department of National Emergency Management. Construction Project Safety Management Yearbook (2024 Volume) [M]. Beijing: Emergency Management Press, 2024: Accident Statistics Schedule 3
2. China Academy of Building Science. Technical Code for Building Safety Protection: GB/T 55032-2025 [S]. Beijing: China Construction Industry Press, (2025)
3. China Safety Industry Association. Smart Site Technology White Paper (2024 Edition) [M]. Beijing: White Paper Release Committee, 2024: Technical Indicators Appendix
4. Intelligent Construction Committee of China Construction Association. Blue Book on Smart Construction 2025 [M]. Beijing: China Machine Press, 2025: 156–160
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers [C]. In: European Conference on Computer Vision. Springer, 213–229 (2020)
6. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L., Zhang L.: Lite DETR: An Interleaved Multi-Scale Encoder for Efficient DETR [C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18558–18567 (2023)
7. Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C.: DETRs Beat YOLOs on Real-Time Object Detection [C]. In: Conference on Computer Vision and Pattern Recognition, 16965–16974 (2024)
8. Redmon, J.: You Only Look Once: Unified, Real-Time Object Detection [C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016)
9. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger [C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7263–7271 (2017)
10. Redmon, J.: YOLOv3: An Incremental Improvement [J]. arXiv preprint arXiv:1804.02767, (2018)
11. Bochkovskiy, A., Wang, C., Liao, H.: YOLOv4: Optimal Speed and Accuracy of Object Detection [J]. arXiv preprint arXiv:2004.10934, (2020)
12. Jocher, G.: YOLOv5 by Ultralytics (Version 7.0). [Online]. (2020). Available: https://doi.org/10.5281/zenodo.3908559
13. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L.: YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications [J]. arXiv preprint arXiv:2209.02976, (2022)
14. Wang, C., Bochkovskiy, A., Liao, H.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors [C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7464–7475 (2023)
15. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Version 8.0.0). [Online]. (2023). Available: https://github.com/ultralytics/ultralytics
16. Wang, C., Yeh, I., Liao, H.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information [J]. arXiv preprint arXiv:2402.13616, (2024)

17. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J.: YOLOv10: Real-Time End-to-End Object Detection [J]. arXiv preprint arXiv:2405.14458, (2024)
18. Khanam, R., Hussain, M.: YOLOv11: An Overview of the Key Architectural Enhancements [J]. arXiv preprint arXiv:2410.17725, (2024)
19. Tian, Y., Ye, Y., Doermann, D.: YOLOv12: Attention-Centric Real-Time Object Detectors [J]. arXiv preprint arXiv:2502.12524, (2025)
20. Shen, J., Liu, N., Sun, H., Li, D., Zhang, Y.: an instrument indication acquisition algorithm based on lightweight deep convolutional neural network and hybrid attention fine-grained features. IEEE Trans. Instrum. Meas. **73**, 1–16 (2024)
21. Shen, J., Liu, N., Sun, H., Li, D., Zhang, Y., Han, L.: An algorithm based on lightweight semantic features for ancient mural element object detection [J]. npj Herit. Sci., **13**(70), (2025)
22. Shen, J., Liu, N., Xu, C., Sun, H., Xiao, Y., Li, D.: finger vein recognition algorithm based on lightweight deep convolutional neural network. IEEE Trans. Instrum. Meas. **71**, 1–13 (2022)
23. Xu, K., Deng, C.: helmet wear recognition algorithm based on improved yolov3. Lasers and Optoelectron. Prog. **58**(6), 300–307 (2021)
24. Xie, G., Tang, J., Lin, Z.: improved yolov4 helmet detection algorithm in complex scenes. Lasers and Optoelectron. Prog. **60**(12), 139–147 (2023)
25. Li, J., Xie, S., Zhou, X.: Real-Time Detection of Coal Mine Safety Helmet Based on Improved YOLOv8 [J]. Real-Time Image Process. **22**, 26 (2025)
26. Jiao, X., Li, C., Zhang, X., Fan, J., Cai, Z., Zhou, Z.: detection method for safety helmet wearing on construction sites based on uav images and yolov8. Buildings **15**(3), 354 (2025)
27. Wang, S., Wu, P., Wu, Q.: safety helmet detection based on improved yolov7-tiny with multiple feature enhancement. Real-Time Image Process. **21**, 120 (2024)
28. Du, Q., Zhang, S., Zhang, S.: BLP-YOLOv10: Efficient Safety Helmet Detection for Low-Light Mining. Real-Time Image Process. **22**, 10 (2025)
29. Seth, Y., Sivagami, M.: enhanced yolov8 object detection model for construction worker safety using image transformations. IEEE Access **13**, 10582–10594 (2025)
30. Chen, X., Jiao, Z., Liu, Y.: improved yolov8n-based helmet wearing inspection method. Sci. Rep. **2025**, 15 (1945)
31. Liu, B., Wei, X., Chen, Q., Liu, J., Chen, Y., Yu, P., Lei, S., Hu, Y.: safety helmet detection methods in heavy machinery factory. Sci. Rep. **15**, 18565 (2025)
32. Han, K., Wang, Y., Tian, Q., et al.: GhostNet: More Features from Cheap Operations [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1580-1589 (2020). Available: https://github.com/HaloTrouvaille/GhostNet-YOLOv5
33. Baidu PaddlePaddle Team. PP-Helmet: A PaddlePaddle-Based Safety Helmet Detection Toolkit. [Online]. 2022. https://github.com/PaddlePaddle/PP-Helmet
34. Zhang, X., Li, Y., Chen, Z.: real-time safety helmet detection using shufflenet-enhanced yolo. Sensors **22**(8), 2989 (2022)
35. Chen, X., Zhang, R., Zhou, Y.: edgeshd: a lightweight safety helmet detection model for edge devices. Expert Syst. Appl. **223**, 120456 (2023)
36. Chen, X., Zhang, R., Zhou, Y.: YOLO-NAS: Neural Architecture Search for Efficient Safety Helmet Detection [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 123-130 (2023)