International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 1-16

https://doi.org/10.70917/ijcisim-2025-0320
Article

Application and validation of RT-DETR target detection model for site safety management in real-time helmet wearing monitoring system

Zeyu Hu[1,*] and Yue Zhang[2]

[1] College Of Architecture And Engineering, Guangdong Baiyun University, Guangzhou, Guangdong, 510450, China
[2] Guangdong Nanhua Vocational College of Industry and Commerce, Guangzhou, Guangdong, 510000, China
* Correspondence author: hugohzy@126.com

Abstract: Safety helmet wearing detection using video real-time monitoring system is important for site safety. The existing safety helmet wearing detection algorithms have more application scenario condition limitations, and it is difficult to meet different scenario requirements at the same time. In order to solve the above problems, the article proposes an improved safety helmet wearing target detection model based on RT-DETR model. The model reconstructs the original backbone network BackBone with ConvNeXt, introduces the F-CBAM attention mechanism to improve the feature extraction effect of the model on small targets and low-resolution images, and improves the original GIoU loss function to CIoU loss function, so as to improve the convergence efficiency and accuracy of the model. The results show that the size of the improved RT-DETR model is only increased by 1.41M compared with the YOLOv7 model, the mAP reaches 93.04%, and the inference time of the model is only 34 frames/ms.Relying on the improved RT-DETR model, the detection effect of helmet wearing can be significantly improved, and the overall generated model is smaller, which can satisfy the practical deployment in different scenarios.

Keywords: RT-DETR model; ConvNeXt; F-CBAM attention mechanism; CIoU; helmet detection

1. Introduction

With the rapid development of China's economy, high-risk industries such as construction, electric power, and mining are developing rapidly, but safety accidents still occur from time to time due to the shortage of protective awareness and irregularities in the wearing of safety protective equipment by construction workers [1-3]. Since the twentieth century, the protection measures for workers in these high-risk industries have been strengthened, and the rate of safety accidents in the industry has declined, but it is still three times higher than the average level of other industries, and has a high degree of danger [4-6].

Human head is an important nervous system center of the body, in the construction site, about the head around the traumatic injury rate is higher, so the helmet as a common and effective protection of the head of the safety protection tools, its correct wearing is very important [7-8]. Relevant enterprises stipulate that relevant personnel must wear correctly when entering the construction site, and relevant staff have also received relevant training before starting work, and these measures have strongly reduced the probability of construction, but due to insufficient supervision and lack of personal awareness of prevention, safety accidents caused by failure to wear helmets correctly still occur from

time to time [9-12]. According to relevant statistics, nearly 85% of the head-injured people in construction sites are injured due to the behavior of not wearing helmets correctly [13].

At present, factories have basically installed monitoring equipment, which can monitor the construction behavior of workers in real time, but factories still need to employ more monitors and inspectors to monitor the construction site regularly, which is easy to produce fatigue resulting in omission and misdiagnosis, and it is difficult to comprehensively and timely to prevent the violation of workers' behaviors, and this traditional way is labor-consuming and inefficient, which does not solve the problem fundamentally [14-17]. Therefore, real-time monitoring and alarming through intelligent technology is a necessary and efficient initiative.

In recent years, the rapid development in the field of computer vision has prompted the application of intelligent processing technology in image acquisition facilities, making it possible to automatically detect helmet wearing. Among them, the target detection technology carries out dynamic monitoring of the site scene, which reduces the incidence of accidents by prompting unsafe behaviors in time [18-19]. For the monitoring of the helmet wearing scene, the intelligent recognition system can analyze the behavioral characteristics of the staff in the monitoring screen, record and supervise the violations in real time through the positioning, tracking and other technologies, and if necessary, it can also automatically alarm, so as to realize the intelligent automatic monitoring [20-23]. Therefore, combined with the actual complex industrial environment, the design of a helmet wearing detection and tracking algorithm for small targets can effectively reduce the incidence of accidents in factories, which is of great significance for safe production.

At present, based on the target detection technology of helmet wearing real-time monitoring system, due to the angle of the monitoring equipment installation, the target is generally far away from the image acquisition equipment, the proportion of the screen personnel is small, the scale changes, etc., the problem of accuracy reduction caused by small target detection is particularly obvious [24-26]. At the same time, changes in gaze illumination, occlusion, and the appearance of foreign objects in the environment in the actual scene are all prone to omission and misdetection. In response to the above situation, researchers have made a variety of improvement measures on various benchmark target detection techniques, and are committed to realizing all-round, real-time, and accurate monitoring of helmet wearing to help site safety management. Under the continuous maturation of deep learning technology, the target detection model for helmet wearing monitoring has experienced three main stages, two-stage detector, single-stage detector, and Transformer series.

In the two-stage detector, helmet wearing monitoring is realized with algorithms such as Faster R-CNN (Regional Convolutional Neural Network), Mask R-CNN, and other algorithms by processing the region after candidate region generation. Literature [27] proposes a two-step helmet wearing detection algorithm that uses CNN to first recognize the human head and then classify the head into helmet, head and hat as a way of determining whether the helmet is being worn correctly or not. Literature [28] first preprocesses the grayscale video stream image using weighted average method, introduces the improved Faster R-CNN algorithm to optimize the sample weight allocation and feature extraction accuracy, and the safety helmet recognition accuracy is 97%, and the recognition speed is 30 frames/second. Literature [29] combined Faster R-CNN and Long Short-Term Memory Network (LSTM) to extract human motion gesture features at spatial and temporal levels, respectively, and recognized whether the helmet was worn or not after classification, and the detection accuracy of Faster R-CNN-LSTM was improved by 15% compared to the CNN-LSTM network framework, whereas the accuracy rate was as high as 99.99%. Literature [30] added Multi-NMS (Non-Maximum Suppression) algorithm and Soft-NMS to the Mask R-CNN algorithm by removing the redundant masks and labels when detecting the target in the Mask R-CNN and binding the helmet to the head, which resulted in 98% detection in the case of having a masked and hand-held helmet Accuracy. Although the two-stage detector can obtain high detection accuracy, the detection speed is slow, which is difficult to effectively meet the real-time helmet wearing monitoring.

Among the single-stage detectors, the YOLO algorithm (You Only Look Once) series and the main helmet wearing monitoring predicts the bounding box and categories from the image, and the SSD (Single Shot MultiBox Detector) model also belongs to this type of method. Literature [31] identifies staff from surveillance camera by SSD neural network model, introduces HSV (Hue, Saturation, Value) color space and morphology, and detects helmet wearing from the perspective of human geometric features. Literature [32] combined SSD algorithm, CNN and LSTM, SSD algorithm was used to extract helmet wearers, and both networks performed real-time detection based on the extraction results, and the accuracy of detection was more than 90%. Literature [33] improved the YOLOv3 network based on spatial attention module and distance intersection and union ratio (IoU) loss function, which improved the small object detection accuracy (96.5%) and convergence speed, while the detection speed was 27 frames/second. Literature [34] improved the YOLOv5 detection model using three loss functions,

generic IoU, distance IoU, and complete IoU, and improved mosaic-9 data, and the accuracy of helmet wearing detection reached 93.16%. Literature [35] constructed an intelligent helmet recognition system based on multi-target tracking algorithm and DeepSort and YOLOv5, which improved the recognition accuracy to 94.5% while guaranteeing the real-time detection requirements in complex environments and maintained the recognition speed of 40 frames/second. Literature [36] added the attention mechanism, full IoU loss function, and Mish activation function to the YOLOv5 detection model, which strengthened the inter-feature information transfer, bounding box regression, model detection accuracy, and adaptability, and improved the accuracy of helmet wearing detection to 96.7%. Literature [37] developed a lightweight helmet detection algorithm based on YOLOv4, Path Aggregation Network, Expanded Convolutional Cross-Stage Part with X res Unit Module, which improves the detection accuracy and reduces the model parameters for better robustness and deployability. Literature [38] proposes a real-time automated helmet detection system based on YOLOv8, which can adapt to complex environments with high accuracy, maintains a low false alarm rate, and provides feedback and corrective measures for recognized violations.The YOLO family of algorithms has improved the detection speed and the model performance, but it is difficult to significantly improve the accuracy in complex environments and small target detection situations.

In the Transformer series, the DETR (Detection Transformer) model based on the Transformer architecture is applied in helmet wearing monitoring with an end-to-end object detection architecture. Literature [39] used the DETR architecture to capture the helmet situation of construction site workers, to realize object detection, group prediction and combination, and combined with IoU to obtain the results of the detection, which can realize fast detection in the case of IoU=0.5. Literature [40] evaluated the performance of the DETR model for detecting helmets in a construction environment, and the model can achieve 99% accuracy and higher than 95% confidence under different image qualities (grayscale, blur, etc.) and scenes. Literature [41] improved the DETR model by eliminating the model's converter encoder module (to improve detection efficiency) and introducing a transformer network based on a deformable attention module (to sense the context at multiple scales and to improve computational efficiency), accomplishing a detection speed of 20 frames/second and a computational efficiency of 13.335 billion times/second. The lightweight DETR model proposed in literature [42] adds a large kernel selection network, a frequency-space fusion transformer, and a boundary-aware selection aggregation module, which is capable of capturing contextual features and long-distance dependencies in more detail, and enhances the edge-sensitive feature representations, which improves the model's accuracy and computational efficiency for the detection of small helmets under complex conditions and reduces the pixel occupancy rate. Literature [43] proposed a helmet and seatbelt monitoring system based on DETR model with recognition, tracking, and attribute recognition functions, fusing image features and semantic attribute features, which can effectively deal with occlusion, image background clutter, and human body multi-posture changes. Literature [44] utilizes Res2Net to replace the backbone network of the DN-DETR (DeNoising DETR) model and introduces a loss function and a positive and negative sample comparison denoising training algorithm to accelerate the convergence of the helmet detection model for higher accuracy and speed.The end-to-end detection of the DETR model does not require the additional addition of redundant feature detection and processing frameworks for higher high detection accuracy, but the model is computationally intensive and real-time performance is difficult to ensure. Therefore, RT-DETR (Real-Time DETR) model is proposed.

The RT-DETR model significantly improves the detection speed and accuracy by removing innovations such as non-maximum suppression, end-to-end training, hybrid encoder, and optimized decoder, and performs especially well in complex scenes and small object detection. In a related study, the literature [45] combined the attitude estimation technique with RT-DETR to significantly improve the detection rate of PPE compliance under different lighting conditions in underground mines, automatically removing erroneous detection results and reducing the false alarm rate. Literature [46] designed a foggy helmet detection framework named “DST-DETR”, which adds a convolutional module to the integrated defogging network model to repair multi-fog images, and introduces a loss function to optimize the robustness and generalization ability of the model, and optimizes the recognition of small objects by RT-DETR with the ST-DETR segmentation module. Small Objects. Literature [47] improves the RT-DETR model with the help of inverse residual block technique, efficient attention mechanism, and dynamic up-sampling method to improve the model's performance (precision, accuracy) in detecting helmet wearing for small targets and also reduces the computational cost. Literature [48] introduced perceptual feature extraction module, deformable attention mechanism, and dynamic multi-scale feature fusion network to improve the detection accuracy and adaptive performance of RT-DETR model at different scales in a complex power operation dress environment. Literature [49] implements a lightweight detection algorithm for PPE in small-size industrial

environments by improving the RT-DETR model, which improves the average progress of the model by 2.4%, reduces the operational parameters and computational complexity, and has a strong generalization ability. In summary, few studies have been conducted on helmet detection in extreme environments, such as bright light and blizzard, etc., and the detection stability and low false alarm rate in this type of weather are the difficulties of the current target detection model. In addition, the compatibility of the model with the existing security monitoring system of the enterprise and employee privacy issues need to be considered.

In order to solve the problems of imprecise detection results and low detection efficiency in the current real-time helmet wearing monitoring system, this paper uses the improved RT-DETR model for helmet wearing target detection. The article innovatively replaces the original BackBone network using ConvNeXt network, and improves the feature extraction capability and convergence accuracy of the model through F-CBAM attention mechanism and CIoU loss function. This paper provides decision support for promoting the efficiency of construction site safety management, and also lays a solid foundation for ensuring the correct helmet wearing and personal safety of construction workers.

2. Improved RT-DETR helmet wearing detection model

With the rapid rise of infrastructure construction and real estate industry, construction safety gradually attracts the attention of the society. Safety is the most important and basic need of workers, and helmet, as a kind of practical labor protection appliance, can effectively protect workers' safety. Therefore, in the process of engineering construction and safety management, detecting the wearing of construction workers' helmets is an important measure to correct workers' unsafe behaviors, reduce head injury safety accidents, and ensure construction safety. Based on the RT-DETR model, this chapter proposes an improvement method of the RT-DETR model under the consideration of the needs of the real-time monitoring system of helmet wearing, aiming to further enhance the effect of helmet wearing monitoring and ensure the safety of construction sites.

2.1. DETR target detection model

DETR uses Transformer to replace the complex traditional sets of target detection, such as mainstream deep learning networks that use multi-scale feature fusion, special convolutions such as deformable convolution to extract features.DETR uses CNN extracted features to encode and decode to get the output results. In target detection, predicting an object requires two necessary conditions, an ensemble prediction loss that uniquely matches the predicted ensemble to the real frame, and the network structure.

2.1.1. Pooled prediction of losses

DETR produces far more N fixed predictions than the number of targets in the image after the decoder, and in order to evaluate these N predictions, the predictions are best bisected matched to the true values by means of a loss function, which is then optimized for bounding box loss [50].

Assuming that the true target set is $y=\{y_i\}_{i=1}^{N}$ and the prediction set is $\hat{y}=\{\hat{y}_i\}_{i=1}^{N}, N$ is much larger than the number of targets in the image, using $\varnothing$ to populate the true set y and find the lowest cost binary match with $\hat{y}$ by the following equation. I.e:

$$\hat{\sigma}=\arg_{\sigma}\min\sum_{i}^{N}L_{match}\left(y_i,\hat{y}_{\sigma(i)}\right) \tag{1}$$

where $L_{match}\left(y_i,\hat{y}_{\sigma(i)}\right)$ is the matching cost between the true target y_i and the exponential $\sigma(i)$ prediction, which takes into account both the category prediction and the bounding box prediction.

Calculate the Hungarian loss for all matched pairs in the previous step with the following formula:

$$\begin{aligned}L_{Hungarian}\left(y_i,\hat{y}_i\right)&=\sum_{i=1}^{N}\left[-\log p_{\hat{\sigma}(i)}\left(c_i\right)\right]\\&*\sum_{i=1}^{N}\left[\gamma_{\{c_i\neq\varnothing\}}L_{box}\left(b_i,\hat{b}_{\hat{\sigma}(i)}\right)\right]\end{aligned} \tag{2}$$

where $\hat{\sigma}$ is the lowest cost bisection match computed in the first step, and each element i of the

real target is denoted by $y_i = (c_i, b_i)$, c_i is the target category label, and b_i is a vector that contains the center coordinates of the real bounding box as well as the height and width relative to the height and width of the image size. For the prediction of c_i, the probability of the category is $p_{\hat{\sigma}(i)}(c_i)$, and the probability of the predicted box is $\hat{b}_{\hat{\sigma}}(i)$.

Finally, a linear combination of l_1 loss and generalized IoU loss is used to solve the problem of relative error of bounding boxes at different scales and to directly predict the location of the bounding box, i.e:

$$L_{box}\left(b_i, \hat{b}_{\hat{\sigma}(i)}\right) = \lambda_{iou} L_{box}\left(b_i, \hat{b}_{\hat{\sigma}(i)}\right) + \lambda_{L1}\left\|b_i - \hat{b}_{\hat{\sigma}(i)}\right\|_1 \tag{3}$$

where $\lambda_{iou}, \lambda_{L1}$ are hyperparameters that normalize the number of objects in the batch by two loss calculations.

2.1.2. Network structure

DETR consists of three components, a CNN backbone for feature extraction, an encoder-decoder Transformer and a feed-forward network FFN.The original image is passed through the CNN to output $f \in R^{(C\times H\times W)}$ features. Then a 1×1 convolution reduces the channel dimension from 2048 to a smaller dimension, generating a new feature map $z_0 \in R^{d\times H\times W}$. The z_0 dimension is then compressed to one dimension, combined with the location information and input to the encoder, and each decoder layer decodes N different objects in parallel, adding them to the input of each attention layer. The decoder converts the N object queries into outputs containing location information, which are then fed into the feedforward network. The feedforward network consists of a layer containing the ReLU activation function, hidden dimensions, and a linear mapping layer. The linear layer uses the SoftMax function to predict the category labels, and the N objects are independently decoded into bounding box coordinates and category labels through the feedforward network to generate the N predicted objects. Using the pairwise relationship between the objects, global inference is performed on all objects, and the undetected objects among the N predicted objects are categorized as background, resulting in the final prediction result [51].

2.2. Improved RT-DETR detection models

2.2.1. Helmet wear detection framework

The Transformer-based DETR algorithm transforms the target detection problem into an ensemble prediction problem, establishes a one-to-one matching relationship between the prediction frames and the labeled frames by means of a dichotomous matching algorithm and defines the prediction loss of the algorithm on the basis of the matched prediction frames. Thus realizing the removal of artificially designed references in the pure convolutional algorithm, the predicted output does not need to be post-processed by the non-maximal value algorithm.The RT-DETR algorithm has a detection accuracy that exceeds similar algorithms, and at the same time it can maintain a high degree of real-time performance, and it has a greater potential for real-time detection of helmet wearing applications in engineering scenarios.

The RT-DETR algorithm consists of four main parts, which are the backbone network model (BackBone), a hybrid encoder, a decoder and a prediction header. The backbone network of the standard RT-DETR algorithm is ResNet-50, and its hybrid encoder in turn consists of a single-scale feature fusion module (AIFI) and a cross-scale feature fusion module (CCFM), with the AIFI module being a single self-attention layer computation and the CCFM module borrowing the traditional feature pyramid computation. The algorithm extracts the last 3 scales of feature maps through the backbone network, and passes the extracted top-level features through the single-layer self-attention layer of the AIFI module for single-scale feature fusion, and then passes the 3 scales of feature maps through the CCFM module for cross-scale feature fusion. A fixed number of image features are filtered from the hybrid encoder output via IoU-aware queries as initialization for the decoder target queries. Finally, the decoder and prediction head's optimize the output target frame and category confidence by iteration [52].

In order to meet the requirements of real-time high-precision detection of helmet wearing in site construction scenarios, the improved RT-DETR algorithm introduces the ConvNeXt model with better

feature extraction capability as BackBone on the basis of the original algorithmic architecture, and designs a feature channel compression module (ESCM) based on the channel attention after the feature maps S3, S4, and S5, in order to increase the number of fewer parametrics improve the effectiveness of the fused channel features, accelerate the training convergence of the algorithm, and improve the detection accuracy of the algorithm. Figure 1 shows the helmet wearing detection framework based on the improved RT-DETR model.

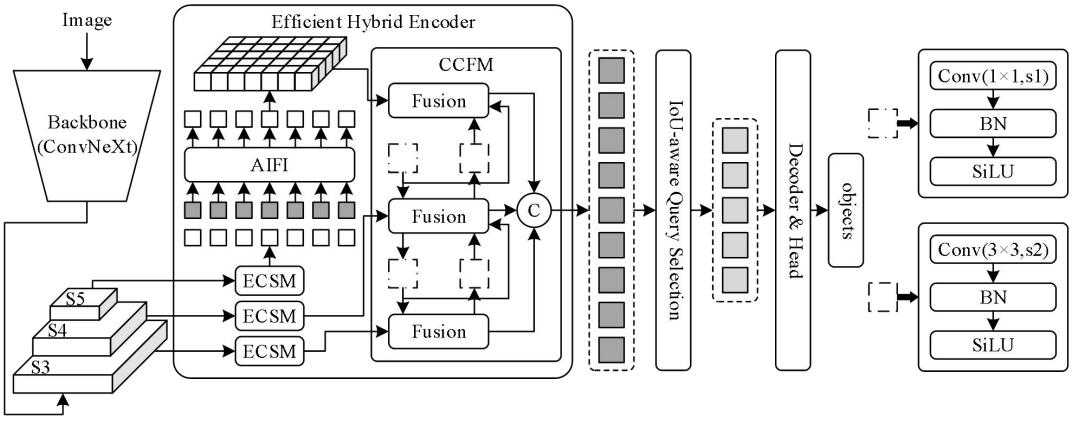

Figure 1. Improve the RT-DETR model framework

2.2.2. Reconfiguration of the backbone network ConvNeXt

The backbone network of the RT-DETR model uses the BackBone structure, which splits the input into two branches, one for residual operations and the other for convolutional operations, and then combines the two branches to make the input and output the same size. The model has more residual structures for more efficient feature extraction, but it also comes with more number of parameters, which makes the model slower to detect, limited in applications, and difficult to deploy in some real-world scenarios such as mobile devices.

The proposed model applies the core concept of residual network and uses ConvNeXt to reconstruct the BackBone network. In contrast to traditional convolutional methods, ConvNeXt adopts Deep Separable Convolution (DSC), which can be viewed as a high-dimensional convolutional kernel, each of which is responsible for only one channel in the input feature matrix. by combining multiple such convolutional kernels, ConvNeXt maintains the ability to globally perceive the input feature matrix, reduces the number of model parameters, and improves the model training speed.

In ConvNeXt, the input is first passed through one DSC, then convolved by two 1×1 convolution kernels and activated by GELU activation function, and then the output is normalized by layer scale layer, which limits the weight of the output to a certain range, which can avoid the problems of gradient explosion and gradient vanishing. And make the model training has a high stability, and finally connect a drop path to reduce the number of model parameters.ConvNeXt network not only improves the accuracy of the detection of helmet wearing target, but also makes the network's floating point operation number reduced.

2.2.3. F-CBAM attention mechanism

CBAM is an efficient attention module that consists of two parts, the channel attention module (CAM) and the spatial attention module (SAM). In this paper, the ReLU activation function in CBAM is replaced by the FReLU activation function, which is specialized for the task of target detection, to improve the accuracy of target detection by capturing the complex visual layout in two dimensions, and the new attention mechanism designed is called F-CBAM attention mechanism.

CBAM contains an important module - multilayer perceptual machine (MLP), the features in the perceptual machine model use ReLU as the activation function for nonlinear activation, and solve the gradient vanishing problem that occurs in the back-propagation algorithm through the ReLU function. In order to achieve pixel-level spatial information modeling capability at the activation function stage and improve the accuracy, this paper replaces the original ReLU activation function with the FReLU activation function, which is specifically designed for visual tasks. The ReLU is extended to a two-dimensional activation function by adding negligible spatial condition overhead.ReLU is denoted as:

$$y = \max(x, 0) \tag{4}$$

And FReLU is of the form:

$$y = \max(x, T(x)) \tag{5}$$

where $T(\cdot)$ is a two-dimensional spatial representation that captures the complex visual layout in two dimensions and improves the accuracy of target detection ReLU's condition is a manually set zero value. FReLU, on the other hand, is a spatial context-dependent two-dimensional condition that helps to extract the spatial layout of target features. The 2D condition depends specifically on the spatial context of each pixel. Finally the maximum value between x and the condition is obtained using $\max(\bullet)$.

The FReLU activation function is defined as follows:

$$f\left(x_c, i, j\right) = \max\left(x_c, i, j, T\left(x_c, i, j\right)\right) \tag{6}$$

where $T\left(x_c, i, j\right)$ is the two-dimensional condition that creates spatial dependence using a parameterized pool window, and spatial dependence is implemented using a highly optimized depth-separable convolution operator and a BN layer. The $T\left(x_c, i, j\right)$ unfolds as follows:

$$T\left(x_c, i, j\right) = x_{c,i,j}^{w} \bullet p_c^{w} \tag{7}$$

where (i, j) denotes the pixel position in 2D space, c denotes the c th channel, $x_{c,i,j}^{w}$ denotes the parameterized pooling window centered on the input pixel of the nonlinear activation function on the c th channel at position (i, j) in 2D space, and p_c^{w} denotes the window this coefficients shared in the same channel.

The CBAM structure that incorporates the FReLU activation function is called F-CBAM, and FReLU has better context capture and therefore better understanding of objects with fuzzy and small targets. In complex situations, the F-CBAM structure can capture irregular and detailed object layouts better than the CBAM structure.

2.2.4. Loss function optimization design

The original RT-DETR model loss function consists of 3 parts: localization loss, confidence loss and category loss. It can be expressed as:

$$\begin{aligned} Loss_{object} &= Loss_{loc} + Loss_{conf} + Loss_{class} \\ Loss_{loc} &= 1 - GIoU \end{aligned} \tag{8}$$

where confidence loss and category loss are calculated using a binary cross-entropy loss function, i.e.:

$$\begin{aligned} Loss_{conf} &= -\sum_{i=0}^{K\times K} I_{ij}^{obj}\left[\hat{C}_i^j \log C_i^j + \left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right] \\ &\quad -\lambda_{noobj}\sum_{i=0}^{K\times K}\sum_{j=0}^{M} I_{ij}^{noobj}\left[\hat{C}_i^j \log C_i^j + \left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right] \\ Loss_{class} &= -\sum_{i=0}^{K\times K} I_{ij}^{obj}\sum_{c\in classes}\left[\hat{P}_i^j \log P_i^j + \left(1-\hat{P}_i^j\right)\log\left(1-P_i^j\right)\right] \end{aligned} \tag{9}$$

where K denotes that the final output feature map of the network is divided into $K\times K$ lattices, M denotes the number of anchor frames corresponding to each lattice, I_{ij}^{obj} denotes anchor frames with targets, I_{ij}^{noobj} denotes anchor frames without targets, and λ_{noobj} denotes the confidence loss weight coefficient of the anchor frame without target.

The GIoU is used in the original RT-DETR model to compute the localization loss, i.e:

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|\mathrm{C} - \mathrm{GTUP}|}{|\mathrm{C}|} = \frac{|\mathrm{PIGT}|}{|\mathrm{PUGT}|} - \frac{|\mathrm{C} - \mathrm{GTUP}|}{|\mathrm{C}|} \tag{10}$$

Different from the original IoU, GIoU not only focuses on the overlap area between the real frame and the prediction frame, but also focuses on other non-overlapping areas, so GIoU can better respond to the degree of overlap between the two compared to the original IoU. However, GIoU always only considers the overlap rate between the real frame and the prediction frame as a factor, which can not describe the regression problem of the target frame well. When the prediction box is inside the real box and the size of the prediction box is the same, then GIoU will degenerate into IoU, which cannot distinguish the positional relationship between each prediction box.

In this paper, CIoU is chosen to replace GIoU as the loss function of target frame regression, which is calculated as:

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{\rho^2\left(b, b^{gt}\right)}{c^2} - \alpha v \tag{11}$$

$$\alpha = \frac{v}{(1 - \mathrm{IoU}) + v} \tag{12}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{13}$$

where α is a balancing parameter, which is not involved in the gradient calculation, and v is a parameter used to measure the consistency of the aspect ratio.CIoU integrally takes into account the overlap rate between the real frame and the predicted frame, the distance of the centroid, and the aspect ratio, so as to make the regression process of the target frame more stable, and the accuracy of the convergence is higher.

3. Experimental results and analysis of helmet wear detection modeling

In the process of site construction, helmets play a very important role in safeguarding the lives of workers, but some workers have weak safety awareness and do not wear helmets according to the requirements in the construction process, which poses a great threat to their lives. This study proposes a real-time state recognition algorithm for helmet wearing based on RT-DETR model. The helmet wearing detection is converted into a binary classification problem of whether to wear a helmet or not, which aims to better realize the detection accuracy of low-resolution, small targets, etc., to fully ensure the safety of construction workers and improve the efficiency of site safety management.

3.1. Experimental environment and evaluation indicators

3.1.1. Experimental environment

The experiments in this paper were done in Windows environment to build the algorithms.In the hardware environment, the processor is AMD Ryzen5 3600 3.6 GHz, the GPU graphics card is RTX4090SUPER, the memory of the graphics card is 16 GB, and the computer memory is 32 GB.In the software, Visual Studio, Python 3.7.2 was installed, and CUDA 10.3 and cudnn 9.0.3 were also installed to support the use of NVIDIA GPUs, and the deep learning framework was Caffe.

The network is trained using a stochastic gradient descent algorithm with momentum, an initial learning rate of 0.001, a momentum factor of 0.8, a learning rate decay strategy of Steps, a batch size of 15, an input image size of 512 pixels × 512 pixels, and 300 training rounds.

3.1.2. Data set production

(1) Data Collection

The current open source helmet dataset is only the SHWD dataset, which contains a total of 7,000 image data, labeled in Pascal VOC format, and is divided into two categories, Person (not wearing a helmet) and Hat (wearing a helmet). Through the study and analysis of this dataset, it was found that the dataset included 8500 targets wearing helmets and 100000 targets not wearing helmets, with a very

unbalanced sample of positive and negative examples. And there is a single type of data scene, the shooting angle is mostly straight shooting, the lack of easy helmet confusion hat data (such as sun hat, beret, etc.) and helmet is not correctly worn data (such as helmet placed in the hands or on the desktop, etc.). Therefore, in this paper, a total of 2,000 images of the above lack of data and some complex and difficult cases are collected and labeled, and a total of 9,000 images with the SHWD dataset are used for model training and validation.

(2) Data Enhancement

For helmet detection, most of the data are long-distance small target data, and dense small target data for target detection algorithms has always been a difficult problem, this paper uses the method of data mosaic for data enhancement. Data mosaicing technology refers to the specified image size, randomly selected 4 pictures converted to the same size, then in the image randomly selected point p as a cut point, the image is cut into 4 parts and respectively, from the above randomly selected 4 pictures were cropped to get section1, 2, 3, 4, and ultimately the four parts of the mosaic together to get the results. After that, it is further judged whether each section contains a detection target, and if it contains a target, the targets contained in each section are combined to form the labeled data of the mosaiced image. In addition, this paper also uses data enhancement means such as random zoom, random horizontal flip, random channel dithering. During the training process, image deflation of 0.5 to 1.5, horizontal flipping with 60% probability, and channel dithering with a magnitude of 0.65 are used.

3.1.3. Evaluation indicators

In the paper, the evaluation metrics mAP, which is commonly used in target detection, and FPS, which is the number of images that can be detected per second, are chosen to evaluate the model. The final detection results are categorized into four types, i.e., True Positive Example (TP), False Positive Example (FP), True Negative Example (TN) and False Negative Example (FN). The precision rate P as well as the recall rate R are important indicators for calculating the mAP, where the precision rate indicates the ratio of the total number of true predicted targets of the model to the total number of targets of the predicted results, and the recall rate indicates the ratio of the total number of true predicted targets to the total number of actual targets in the data set. The calculation formula is as follows:

$$P = \frac{a_{\mathrm{TP}}}{a_{\mathrm{TP}} + a_{\mathrm{FP}}} = \frac{a_{\mathrm{TP}}}{a_{\text{All prediction frames}}} \tag{14}$$

$$R = \frac{a_{\mathrm{TP}}}{a_{\mathrm{TP}} + a_{\mathrm{FN}}} = \frac{a_{\mathrm{TP}}}{a_{\text{All Marker Boxes}}} \tag{15}$$

The precision recall curve can be obtained by taking the precision rate as the horizontal axis and the recall rate as the vertical axis, which is called the $P-R$ curve. The area enclosed by the $P-R$ curve and the coordinate axis for each class of recognized objects is the accuracy value (AP) of the class, and mAP is the average of all APs. The formula for AP and mAP is as follows:

$$\varepsilon_{AP} = \int_0^1 P(R)\mathrm{d}R \tag{16}$$

$$\varepsilon_{\mathrm{maP}} = \frac{\sum_{i=1}^{N} \delta_{AP,i}}{N} \tag{17}$$

where N is 2, denoting 2 target classes.

3.2. Experimental results and analysis

3.2.1. Comparison of model improvement effects

Based on the SHWD dataset given in the previous section, the dataset is divided into a training set and a test set in the ratio of 8:2. In the training phase, data enhancement techniques are used to improve the performance of the network model, and the samples are cropped, panned, brightness changed, and noise added to achieve the data augmentation effect. The training is performed using a multi-scale

strategy, where the input image resolution size is resized every 10 training cycles to enhance the adaptability to different resolution images. The loss curve of the training process of the improved RT-DETR in this paper is shown in Fig. 2. From the figure, we can see that the model is trained nearly $3.5*10^5$ times, the loss value is stabilized at about 0.1, and the better robustness of the model can be inferred by the degree of fluctuation of the curve.

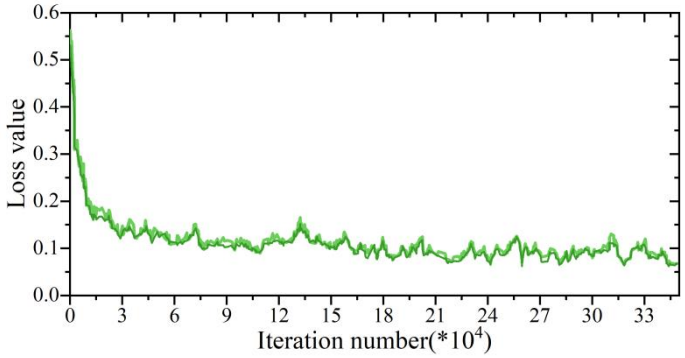

Figure 2. Training process loss curve

In order to compare the effect of the algorithm improvement, the original RT-DETR model is compared with the improved RT-DETR model on the test set, and the statistical results are shown in Table 1. Analyzing the data in the table, it can be found that the improved helmet wearing target detection model based on the RT-DETR model for the positive sample Hat and the negative sample Person has improved the test in precision and recall by 4.23% and 6.99% on average respectively, while the error rate has been reduced by 6.99% on average. It can be seen that the improved RT-DETR model is effective compared with the original RT-DETR model and the model performance is better than the original model, which can realize the accurate detection of helmet wearing.

Table 1. Comparison of model improvement effects

Model	Sample	TP	FP	FN	P/%	R/%	Error/%
Original	Hat	290	38	50	88.42%	85.29%	14.71%
RT-DETR	Person	212	29	42	87.97%	83.47%	16.53%
Improve	Hat	310	30	33	92.26%	90.91%	9.09%
RT-DETR	Person	225	18	24	92.59%	91.84%	8.16%

3.2.2. Testing in small target scenarios

In this paper, ConvNeXt is used to replace the original BackBone backbone network in the improved RT-DETR model and combined with a cross-scale feature fusion module to enhance the helmet wearing detection in small target scenes. In order to verify its detection effect on small target detection performance, this paper sorts the test data set according to the size of the target, respectively, 0-15%, 15%-30%, 30%-45%, 45%-60%, 60%-75%, 75%-90%, and 90%-100% of its target size into XXS, XS, S, M, L, XL, XXL, and XXL, a total of 7 subcategories, which represent the sizes of different targets. Figure 3 shows the effect of the model's detection performance on targets of different sizes. As can be seen from the figure, the improved RT-DETR model has higher detection accuracy for different target sizes than the original RT-DETR model, and its maximum detection accuracy reaches 98.11%. Therefore, using the strategy of multi-feature fusion, combining high-level features with low-level features in the detection of helmet wearing makes the whole algorithm achieve better results in the detection of small targets.

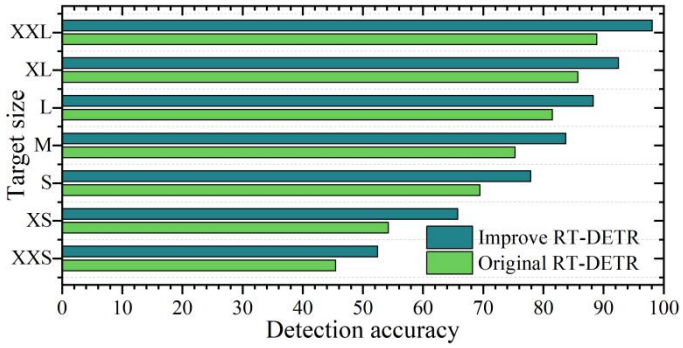

Figure 3. Target sensitivity analysis of the two model

In order to further illustrate the detection effect of this paper's model for helmet wearing at different resolutions, in this paper, during the testing process, the test set of images are all divided into four different resolution sizes, i.e., {128, 256, 512, and 1024}, which represent the four types of images, namely, ultra-low, low, medium, and high-resolution, respectively. Figure 4 shows the detection accuracy performance results of the two models for different resolution images. As can be seen from the figure, the detection accuracy of the improved RT-DETR model for different resolution images is higher than that of the original RT-DETR model, and even for the ultra-low resolution 128*128 helmet image, its detection accuracy can reach 86.74%. It can be seen that the multi-scale training detection strategy adopted in this paper can enhance the adaptability of the improved RT-DETR model to different resolution images.

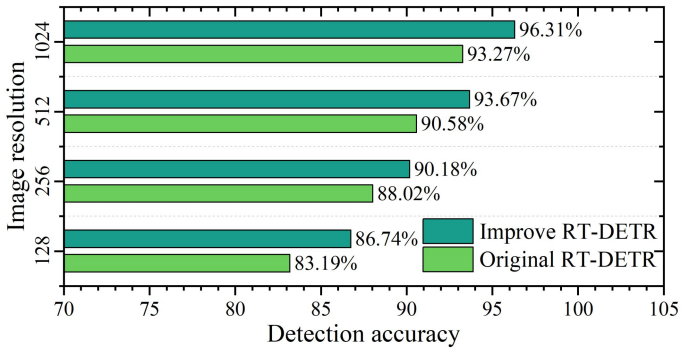

Figure 4. Sensitivity analysis of target image resolution

3.2.3. Mainstream model performance comparison

In order to demonstrate the advantages of the improved RT-DETR model in helmet wearing detection, we compared it with some of the current mainstream target detection methods, such as YOLOv5, SSD300, Faster-RCNN and YOLOv7 algorithms. Table 2 shows the results of the detection performance comparison between the improved RT-DETR model and other mainstream models.

As can be seen from the table, compared with other mainstream target detection methods, the improved RT-DETR models proposed in this paper all have a certain improvement in the mAP value, which reaches 93.04%.The Faster-RCNN model has a relatively high accuracy in helmet wearing detection, which reaches 90.37%, but its computational volume is too large (102.43MB), and the generated model is much larger than that of the YOLOv7 target detection algorithm model.The YOLOv5 algorithm model has a lower average accuracy and generates a larger model, which is not suitable for practical deployment. The algorithmic model proposed in this paper increases only 1.41M over the YOLOv7 model, while the mAP reaches 93.04%, and the average accuracies of both helmet-wearing detection and helmet-unwearing detection are improved.

In order to compare the detection efficiency of the models, the article tested the detection speed of different algorithms individually and on the same GPU. The results show that the detection speed of the improved RT-DETR model is slower than that of the YOLOv7 algorithm, but it is improved by 14 FPS compared with the YOLOv5 model and 25 FPS compared with SSD300.It also proves that the improved RT-DETR model is effective, which ensures that the model quickly detects helmet wearing and improves the site safety management Effectiveness.

Table 2. Performance comparison of mainstream models

Model	mAP /%	Model size/M	FPS	AP (%)		P (%)		R (%)	
				Hat	Person	Hat	Person	Hat	Person
Faster-RCNN	90.37	102.43	78	87.81	94.02	89.61	93.72	82.19	90.21
SSD300	82.49	95.76	92	80.53	86.37	82.35	86.28	78.05	83.15
YOLOv5	86.73	205.38	103	84.36	89.32	85.57	88.35	79.63	86.64
YOLOv7	90.25	14.61	121	87.42	94.16	89.83	93.06	81.94	89.39
RT-DETR	93.04	16.02	117	89.46	95.58	91.18	95.41	84.26	91.08

In order to further verify the inference time of the model in the paper under the real-time monitoring system of helmet wearing, a video containing 1000 frames of workers' operation in the construction site is selected for actual testing. Figure 5 shows the comparison results of the model inference time. The two-stage algorithm Faster RCNN inference of video frames consumes 93ms, while the inference time of the improved RT-DETR model in this paper is 34 frames/ms, which is 59ms faster than the former, but 8ms slower than the original YOLOv3 model. Although the introduction of the feature pyramid (SPP) and the attention module to the YOLOv3-SPP model adds some time overhead, the pooling operation is less computationally intensive and the overall inference time is moderate. Overall, the improved RT-DETR model constructed in this paper has strong inference efficiency and can realize the real-time monitoring of helmet wearing monitoring system.

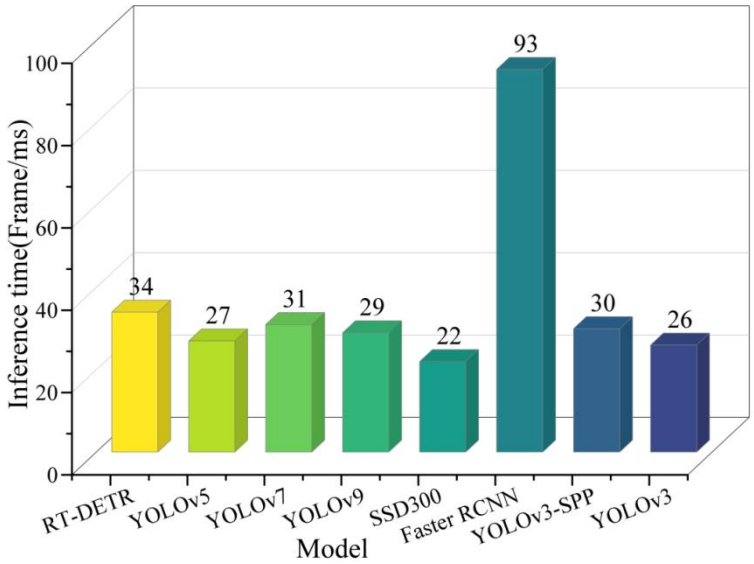

Figure 5. Model inference time comparison results

3.2.4. Analysis of model ablation experiments

To evaluate the extent to which different combinations improve the performance of the improved RT-DETR model, ablation experiments are designed in this paper. All hyperparameters in the ablation experiment are consistent throughout the training process, where the number of training rounds is 500, the batch size is 15, the input image size is 1024*1024, the model optimizer is SDGM, and the initial learning rate is set to 0.005.In this paper, three improvements are made to the original RT-DETR model, i.e., the ConvNeXt backbone network, F-CBAM Attention Mechanism and Loss Function Optimization, i.e., YH1~YH3. Different combinations of the three improvements are applied to obtain the ablation experiment results as shown in Table 3.

The experimental results show that the introduction of each strategy optimizes the model structure to varying degrees. Combination 2 shows that when the network is improved with the ConvNeXt backbone network, mAP@0.5 and mAP@0.5:0.95 only decreases by 1.02% and 1.05%. This indicates that the ConvNeXt backbone network improves the complexity and slightly reduces the accuracy. Combination 3 shows that the introduction of the F-CBAM attention mechanism enhances the feature extraction capability of the backbone network, which in turn improves the accuracy of model detection. In combination 4, the CIoU loss function increased mAP@0.5 and mAP@0.5:0.95 by 1.37% and 0.91% respectively without increasing the computational complexity. The above results show that the CIoU loss function can improve the model's localization accuracy, enhance the bounding box regression accuracy, improve the model's robustness, and perform better in the process of small target identification and detection, which enables the model to more accurately locate and identify the target,

and thus improves the overall detection performance of the model.

The combination of different strategies has a positive optimization effect on the network structure. By comparing the data of combination 5, combination 2 and combination 4, it can be observed that the network using ConvNeXt backbone network and CIoU loss improved by 4.13% and 2.62% respectively at mAP@0.5 and mAP@0.5:0.95 compared with the network using only ConvNeXt. It has increased by 1.74% and 0.66% respectively compared with the network that only uses CIoU loss. The comparison of Combination 6 and Combination 7 with Combination 2 shows that both the F-CBAM attention mechanism and the CIoU loss apply to the network after the introduction of the ConvNeXt backbone. After adding the F-CBAM attention mechanism and the CIoU loss mechanism, both mAP@0.5 and mAP@0.5:0.95 of the network have been improved to a certain extent. This indicates that simultaneously adopting these three strategies will to some extent slow down the optimization effect of a single strategy, but it can still maintain good model accuracy and network structure depth.

Table 3. Results of the ablation experiment

No.	YH1	YH2	YH3	mAP@0.5	mAP@0.5:0.95
1	×	×	×	75.21%	47.62%
2	√	×	×	74.19%	46.57%
3	×	√	×	75.76%	47.01%
4	×	×	√	76.58%	48.53%
5	×	√	√	78.32%	49.19%
6	√	√	×	76.51%	46.38%
7	√	×	√	76.51%	47.64%
8	√	√	√	77.94%	48.25%

4. Conclusion

In order to improve the accuracy of real-time monitoring of helmet wearing in the process of site safety management, this paper proposes a helmet wearing target detection model based on the improved RT-DETR model. In this study, ConvNeXt is introduced to reconfigure the backbone network and combined with the F-CBAM attention mechanism to improve the feature extraction capability for small targets and low resolution, and the CIoU loss function is utilized to accelerate the model convergence. The simulation results show that the improved RT-DETR model has high helmet wearing detection accuracy, and the overall inference time and detection efficiency are high, which can meet the application requirements of real-time helmet wearing monitoring system. Therefore, relying on deep learning technology can promote the real-time helmet wearing detection at construction sites and help improve the efficiency of construction site safety management.

About the Authors

Zeyu Hu (1985-), male, Han ethnicity, born in Jieyang City, Guangdong Province, master's degree, Institute of Civil Engineering, Guangdong Baiyun University, assistant, research direction: engineering project management, intelligent construction management technology.

Yue Zhang (1987-), Female, Han ethnicity, native to Lianzhou, Guangdong, master's degree, Guangdong Nanhua Vocational College of Industry and Commerce, research direction: art design.

References

1. Sehsah, R., El-Gilany, A. H., & Ibrahim, A. M. (2020). Personal protective equipment (PPE) use and its relation to accidents among construction workers. La Medicina del lavoro, 111(4), 285.
2. Gidiagba, J. O., Leonard, J., Olurin, J. O., Ehiaguina, V. E., Ndiwe, T. C., Ayodeji, S. A., & Banso, A. A. (2024). Protecting energy workers: A review of human factors in maintenance accidents and implications for safety improvement. Advances in Industrial Engineering, 15(2), 123-145.
3. Tian, S., Wang, Y., Ma, T., Mao, J., & Ma, L. (2024). Analysis of the causes and safety countermeasures of coal mine accidents: A case study of coal mine accidents in China from 2018 to 2022. Process Safety and Environmental Protection, 187, 864-875.
4. Zhang, J., Fu, J., Hao, H., Fu, G., Nie, F., & Zhang, W. (2020). Root causes of coal mine accidents:

Characteristics of safety culture deficiencies based on accident statistics. Process Safety and Environmental Protection, 136, 78-91.

5. Ranganathan, P. (2022). Occupational Accidents and need for worker safety in manufacturing and high Risk Industries–An Explorative Study with solutions. International Journal of Professional Business Review: Int. J. Prof. Bus. Rev., 7(6), 21.
6. Nygren, M., Jakobsson, M., Andersson, E., & Johansson, B. (2017). Safety and multi-employer worksites in high-risk industries: An overview. Relations industrielles, 72(2), 223-245.
7. Adade-Boateng, A. O., Fugar, F., & Adinyira, E. (2021). Framework to improve the attitudes of construction workers towards safety helmets. Journal of Construction in Developing Countries, 26(2), 65-86.
8. Li, X., Li, H., Skitmore, M., & Wang, F. (2022). Understanding the influence of safety climate and productivity pressure on non-helmet use behavior at construction sites: A case study. Engineering, Construction and Architectural Management, 29(1), 72-90.
9. Ebekozien, A. (2022). Construction companies' compliance to personal protective equipment on junior staff in Nigeria: issues and solutions. International Journal of Building Pathology and Adaptation, 40(4), 481-498.
10. Gupta, R., Yadav, D., Singh, D., Taneja, K., Sharma, A., & Dadich, B. (2024, September). Safety Helmet for Mining Workers. In 2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit) (pp. 789-794). IEEE.
11. Casey, T. W., Mason, H. M., Huang, J., & Franklin, R. C. (2021). Shaping frontline practices: a scoping review of human factors implicated in electrical safety incidents. Safety, 7(4), 76.
12. Xuecai, X., Gui, F., Shifei, S., Xueming, S., Jing, L., Lida, H., & Na, W. (2024). Accident case data-accident causation model driven safety training method: Targeted safety training empowered by historical accident data in coal industry. Process safety and environmental protection, 182, 1208-1226.
13. Abukhashabah, E., Summan, A., & Balkhyour, M. (2020). Occupational accidents and injuries in construction industry in Jeddah city. Saudi Journal of Biological Sciences, 27(8), 1993-1998.
14. Ammad, S., Alaloul, W. S., Saad, S., Qureshi, A. H., Sheikh, N., Ali, M., ... & Iskandar, S. (2020). Personal protective equipment in construction, accidents involved in construction infrastructure projects. Solid State Technology, 63(6), 4147-4159.
15. Baoju, L., Xiangqian, W., Qingshan, C., Jiaqi, L., Ye, C., Peng, Y., ... & Yongfeng, H. (2025). Safety helmet detection methods in heavy machinery factory. Scientific Reports, 15(1), 18565.
16. Mneymneh, B. E., Abbas, M., & Khoury, H. (2019). Vision-based framework for intelligent monitoring of hardhat wearing on construction sites. Journal of Computing in Civil Engineering, 33(2), 04018066.
17. Kondrateva, O. E., Loktionov, O. A., & Miroshnichenko, D. A. (2024, February). Analysis of regulatory requirements for providing personal protective equipment to electric power industry employees in Russia, the USA and Canada. In 2024 6th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE) (pp. 1-5). IEEE.
18. Kim, S., Hong, S. H., Kim, H., Lee, M., & Hwang, S. (2023). Small object detection (SOD) system for comprehensive construction site safety monitoring. Automation in Construction, 156, 105103.
19. Shen, J., Jiao, L., Zhang, C., & Peng, K. (2024). Monocular 3D object detection for construction scene analysis. Computer-Aided Civil and Infrastructure Engineering, 39(9), 1370-1389.
20. Merchán-Cruz, E. A., Moveh, S., Pasha, O., Tocelovskis, R., Grakovski, A., Krainyukov, A., ... & Petrovs, V. (2025). Smart Safety Helmets with Integrated Vision Systems for Industrial Infrastructure Inspection: A Comprehensive Review of VSLAM-Enabled Technologies. Sensors, 25(15), 4834.
21. Wong, G., Anizam, N., & Hadiana, N. (2025). An IoT-Enabled Smart Safety Helmet for Enhancing Worker Protection on Construction Sites. Politek. Kolej Komuniti J. Eng. Technol., 10, 105-116.
22. Hayat, A., & Morgado-Dias, F. (2022). Deep learning-based automatic safety helmet detection system for

construction safety. Applied Sciences, 12(16), 8268.

23. Vo, Q. B., Nguyen, T. H. T., Hoang, T. H. T., Tran, D. T., & Ly, H. B. (2025). Enhancing construction safety management efficiency with AI-Powered real-time helmet detection. Journal of Science and Transport Technology, 77-91.
24. Sun, X., Xu, K., Wang, S., Wu, C., Zhang, W., & Wu, H. (2021). Detection and tracking of safety helmet in factory environment. Measurement Science and Technology, 32(10), 105406.
25. Liang, H., & Seo, S. (2022). UAV low-altitude remote sensing inspection system using a small target detection network for helmet wear detection. Remote Sensing, 15(1), 196.
26. Zhou, J., Wu, Z., Huang, W., Liu, J., & Que, Z. (2024, November). Enhancing live working safety: a helmet detection method for operators. In Fourth International Conference on Advanced Algorithms and Neural Networks (AANN 2024) (Vol. 13416, pp. 528-535). SPIE.
27. Lee, J. Y., Choi, W. S., & Choi, S. H. (2023). Verification and performance comparison of CNN-based algorithms for two-step helmet-wearing detection. Expert Systems with Applications, 225, 120096.
28. Yang, X., Chen, L., Wang, Z., Lin, S., Luo, D., Cai, Z., & Lu, Y. (2025). Video stream safety helmet recognition method based on improved faster R-CNN. International Journal of Intelligent Systems Technologies and Applications, 23(1-2), 202-216.
29. Li, X., Hao, T., Li, F., Zhao, L., & Wang, Z. (2023). Faster R-CNN-LSTM construction site unsafe behavior recognition model. Applied Sciences, 13(19), 10700.
30. Chen, Z., & Su, M. (2021, November). Improved mask R-CNN method for intelligent monitoring of helmet in power plant. In 2021 Photonics & Electromagnetics Research Symposium (PIERS) (pp. 844-848). IEEE.
31. Wang, W., Gao, S., Song, R., & Wang, Z. (2020). A safety helmet detection method based on the combination of ssd and hsv color space. In IT Convergence and Security: Proceedings of ICITCS 2020 (pp. 123-129). Singapore: Springer Singapore.
32. Xu, B. (2023). Real-time helmet wearing status detection method for construction safety. In Frontiers of Civil Engineering and Disaster Prevention and Control Volume 2 (pp. 326-331). CRC Press.
33. Zhou, Q., Qin, J., Xiang, X., & Tan, Y. (2021). Algorithm of Helmet Wearing Detection Based on AT-YOLO Deep Mode. Computers, Materials & Continua, 69(1).
34. Liang, H., Yang, L., Chen, J., Liu, X., & Hang, G. (2024). Detection and tracking of safety helmet wearing based on deep learning. Open Computer Science, 14(1), 20240017.
35. Song, H., Zhang, X., Song, J., & Zhao, J. (2023). Detection and tracking of safety helmet based on DeepSort and YOLOv5. Multimedia Tools and Applications, 82(7), 10781-10794.
36. Li, Y., Zhang, J., Hu, Y., Zhao, Y., & Cao, Y. (2022). Real-time Safety Helmet-wearing Detection Based on Improved YOLOv5. Computer Systems Science & Engineering, 43(3).
37. Li, H., Wu, D., Zhang, W., & Xiao, C. (2024). YOLO-PL: Helmet wearing detection algorithm based on improved YOLOv4. Digital Signal Processing, 144, 104283.
38. Santi, R., Suwarningsih, W., & Sastrosubroto, A. S. (2025). AUTOMATED DETECTION OF HELMET WEARING WITH YOLOV8 AND REAL-TIME MONITORING FOR FACTORY SAFETY. Interdisciplinary Journal of Information, Knowledge & Management, 20.
39. Subhi, M. R., Rachmawati, E., & Kosala, G. (2023). Safety helmet detection on field project worker using detection transformer. Journal of Information System Research (JOSH), 4(4), 1316-1323.
40. Shanti, M. Z., An, B., Yeun, C. Y., Cho, C. S., Damiani, E., & Kim, T. Y. (2025). Enhancing Worker Safety at Heights: A Deep Learning Model for Detecting Helmets and Harnesses Using DETR Architecture. IEEE Access.
41. Chen, S., Sun, H., Wu, Y., Shang, L., & Ruan, X. (2025). A helmet detection algorithm based on transformers with deformable attention module. Chinese Journal of Electronics, 34(1), 229-241.

42. Xu, Z., & Liu, W. (2025, May). Small Target Detection Method for Safety Helmet Based on Lightweight Transformer. In 2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 1206-1210). IEEE.
43. Wu, X., Li, Y., Long, J., Zhang, S., Wan, S., & Mei, S. (2023). A remote-vision-based safety helmet and harness monitoring system based on attribute knowledge modeling. Remote Sensing, 15(2), 347.
44. Yan, Y., & Niu, K. (2023, November). Improved DN-DETR for Safety helmet wearing Detection. In 2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC) (pp. 874-877). IEEE.
45. Imam, M., Baïna, K., Tabii, Y., Ressami, E. M., Adlaoui, Y., Benzakour, I., ... & Abdelwahed, E. H. (2024). Ensuring Miners' Safety in Underground Mines through Edge Computing: Real-Time Pose Estimation and PPE Compliance Analysis. IEEE Access.
46. Liu, Z., Sun, C., & Wang, X. (2024). DST-DETR: image Dehazing RT-DETR for safety helmet detection in foggy weather. Sensors, 24(14), 4628.
47. Zhu, W., & Zhu, C. (2024, July). Improved Safety Helmet Detection Model Based on RTDETR: RTDETR-IHD. In 2024 7th International Conference on Computer Information Science and Application Technology (CISAT) (pp. 895-899). IEEE.
48. Li, T., Li, N., & Nie, Y. (2025, July). SDMD-RTDETR: An Improved Real-Time Transformer Detection Model for Detecting Violations in Power Operators' Dress. In International Conference on Intelligent Computing (pp. 84-95). Singapore: Springer Nature Singapore.
49. Wang, H., Ma, J., Chen, W., Han, Q., Lin, J., Li, J., & Yao, Z. (2025). Personal protective equipment detection for industrial environments: a lightweight model based on RTDETR for small targets. Engineering Research Express, 7(2), 0252a1.
50. Kanghui Zhao,Qinghong Yang & Xingang Miao. (2025). FR-DETR: a streamlined DETR for remote sensing object detection. Measurement Science and Technology,36(10),106008-106008.
51. Lifang Chen,Mingxu Chen,Xufeng Zhang & Zhenping Xie. (2025). LCPD-DETR: a lightweight object detection model based on RT-DETR for military camouflaged personnel. Journal of Real-Time Image Processing,22(6),198-198.
52. Manav Madan & Christoph Reich. (2025). Strengthening Small Object Detection in Adapted RT-DETR Through Robust Enhancements. Electronics,14(19),3830-3830.