International Journal of Engineering and Technology Innovation, vol. 16, no. 3, 2026, pp. 352-369

Multi-Object Tracking and Detection of Power Grid Construction Workers Based on Pre-trained YOLOv11

Lingwen Meng, Guobang Ban, Jintong Ma, Jiangang Liu, Mingyong Xin*

Electric Power Research Institute of Guizhou Power Grid Co., Ltd., Guiyang, China

Received 20 March 2026; received in revised form 24 April 2026; accepted 07 May 2026

DOI: https://doi.org/10.46604/ijeti.2026.16292

Abstract

This study aims to develop a real-time safety equipment detection and multi-object tracking framework for power grid construction scenarios. To this end, a dedicated dataset, E-hat, is constructed by combining a public safety helmet dataset with newly collected images from real-world power grid construction environments. Based on a pretrained lightweight you only look once version (YOLOv11n) detector, this paper introduces scenario-oriented input enhancement strategies, including hue-saturation-value (HSV) perturbation, Mosaic augmentation, and Mixup augmentation, to improve robustness under complex lighting, dense background interference, and partial occlusion. In addition, the detector is integrated with the BoT-SORT algorithm to form a detection-tracking-warning pipeline for continuous worker monitoring, hazardous area intrusion recognition, and unsafe behavior warning. Experimental results show that the proposed system achieves 92.7% precision, 93.4% mAP@0.5, and 183 frames per second (FPS) on the E-hat dataset. Real-world video tests further demonstrate multiple object tracking accuracy (MOTA) values up to 97.71%, demonstrating its practical potential in power safety management.

Keywords: safety helmet detection, multi-object tracking, YOLOv11, power grid construction

1. Introduction

In power production and construction, worker safety is a fundamental requirement for safety management. Power grid companies generally require all personnel to wear personal protective equipment (PPE) properly when entering substations, power grid construction scenarios, and other operation sites. This requirement is closely related to the characteristics of power operation environments, where workers may be exposed to falling objects, electrical discharges, complex personnel mobility, and other hazards. As the most basic PPE, safety helmets can mitigate the impact of falling objects and reduce the risk of head injuries in accidents [1-2].

In complex power operational scenarios, dense equipment layouts, high-altitude operations, and severe environmental interference further increase safety risks. According to public statistics, 54 power safety accidents and 54 casualties were reported nationwide in 2019; falls from heights and electrical shocks were among the major causes, with many cases involving failure to wear or improper use of safety equipment [2-3]. In practical operations, protective gear frequently undergoes accidental dislodgement, temporary unmonitored removal, or remains undetected in multi-person collaborative scenarios. Therefore, real-time detection and warning of unsafe wearing behaviors are of great significance for reducing accidents and improving on-site safety management.

Traditional manual inspection is commonly adopted for safety equipment detection, exhibiting obvious drawbacks, including insufficient inspection coverage, response delays, and high manual misdetection rates [4]. With the rapid

* Corresponding author. E-mail address: jinshx3@163.com

development of computer vision, object detection has become an important technical means for intelligent safety monitoring. Compared with manual inspection, vision-based methods are more efficient and adaptable to complex environments, especially under changing illumination, partial occlusion, and other challenging conditions in a power scenario [4-5].

Among object detection methods, two-stage detectors typically provide high accuracy by first generating candidate regions and then refining them, but they incur high computational costs. Such methods have been applied to safety equipment detection, including region-based convolutional neural networks (R-CNN) and Fast R-CNN [6-7]. Chen et al. [8] improved Faster R-CNN for power construction sites by introducing Retinex image enhancement and K-means++-based anchor generation, which enhances robustness to lighting variations and improves small-object detection performance. Related work also addresses the error accumulation problem caused by detecting pedestrians first and then the safety equipment [9]. These studies demonstrate the feasibility of deep learning-based safety monitoring, but their computational burden limits real-time deployment in practical power environments.

To better balance accuracy and efficiency, researchers have increasingly adopted one-stage detectors represented by the you only look once (YOLO) series [8, 10-11]. These methods directly predict object categories and locations from input images, making them more suitable for real-time applications. Existing studies have applied YOLO-based models to safety helmet detection in different scenarios. For example, Zhou et al. [12] modified YOLOv5 for a two-class safety equipment detection task and improved both speed and accuracy. Tan et al. [13] enhanced YOLOv5 by increasing the detection scale and replacing conventional non-maximum suppression with distance-iou non-maximum suppression (DIoU-NMS), thereby improving detection accuracy and real-time performance. Jiang et al. [14] proposed an improved YOLOv8n-based self-inspection algorithm for field operation scenarios. Their study mainly emphasized small-target real-time detection and the lightweight deployment potential of the model on edge computing devices, thereby providing technical support for practical on-site operation monitoring.

Additionally, Lin et al. [15] improved YOLOv8 by introducing a slim-neck structure, coordinate attention, a small-object detection layer, and mosaic augmentation to strengthen robustness in complex backgrounds. Furthermore, YOLO11 provides improved architecture and training strategies while maintaining a favorable balance between accuracy and efficiency for real-time vision tasks [16]. Although these studies demonstrate the feasibility of deep learning-based helmet detection, most of them focus on frame-level detection only. These works offer limited consideration of continuous worker tracking, hazardous area intrusion warning, and scenario adaptation to power grid construction environments. Moreover, many existing studies focus on generic construction or industrial settings rather than highly specific power operation scenarios characterized by dense equipment layout, strong reflection, complex backgrounds, and strict real-time monitoring requirements.

Besides detection, multi-object tracking is also essential for safety management in real video scenes. This is because on-site safety supervision involves not only identifying workers and their equipment in single frames but also tracking their trajectories continuously over time [17]. Traditional tracking methods, including color-based, template-matching, and feature-point-based approaches, each has specific advantages but is often sensitive to illumination changes, deformation, or occlusion. Deep learning-based tracking methods have improved robustness and generalization, but they also require efficient detector support in practical deployment. In power grid construction sites, workers may enter hazardous areas or perform unsafe actions, and these behaviors cannot be fully captured by frame-level detection alone. Therefore, integrating a real-time detector with a reliable multi-object tracking algorithm is necessary for continuous worker monitoring and timely warning in practical safety scenarios.

To further clarify the position of the present study relative to representative prior work, a concise comparison is provided in Table 1, which is specifically designed to serve as a literature-level comparison to clarify the scope of this research against these benchmarks.

Table 1 Key differences between representative prior studies and the present study

Study	Technical emphasis	Scene type	Tracking	Warning	Key difference from the present study
Mahadi et al. [1]	YOLOv8+DeepSORT personal protective equipment (PPE) tracking and detection	General PPE/construction monitoring	Yes	Limited	Generic PPE tracking, not power-grid-specific monitoring
Chen et al. [8]	Improved Faster R-CNN for helmet detection	Power construction site	No	No	Detector-level improvement for frame-level detection
Zhou et al. [12]	Lightweight YOLOv5 helmet detection	General/limited scene	No	No	Efficient frame-level helmet detection
Tan et al. [13]	Improved YOLOv5 with DIoU-NMS	General construction/industrial scenes	No	No	Focus on detector accuracy and speed
Lin et al. [15]	Improved YOLOv8 in a complex background	Generic complex-background scenes	No	No	Robust frame-level helmet detection
Present study	Pretrained YOLOv11n with scenario-oriented input enhancement and BoT-SORT integration	Power grid construction scenarios	Yes	Yes	Scenario-oriented detection-tracking-warning framework with dedicated power-grid-relevant data and engineering validation

As shown in Table 1, representative prior studies mainly focus on detector-level optimization for frame-level helmet detection under general construction, industrial, or limited scenario settings. In contrast, the present study targets a power-grid-specific safety-monitoring scenario characterized by dense equipment layout, strong reflection, complex backgrounds, and stricter real-time warning requirements. The contribution of this work should therefore be understood at the system level rather than as a fundamentally new detector or tracker architecture. Specifically, the present study emphasizes scenario-oriented dataset construction, lightweight detector adaptation based on pretrained YOLOv11n. Furthermore, we integrate detection, tracking, and warning into a unified trajectory-aware framework for hazardous-area intrusion recognition and persistent no-helmet warning. The main contributions of this study are summarized as follows:

(1) A dedicated safety helmet detection dataset, E-hat, is constructed for power grid construction scenarios by integrating the public safety helmet wearing dataset (SHWD) with newly collected field images from real power operation sites.

(2) A scenario-adapted lightweight detection configuration, termed YOLOv11n-IE (YOLOv11n via image augmentation-enhanced strategy), is established based on pretrained YOLOv11n. This configuration incorporates power-scene-oriented input enhancement strategies, including hue-saturation-value (HSV) perturbation, mosaic augmentation, and Mixup augmentation, to improve robustness in complex power environments.

(3) By integrating the detector with the bag of tricks for the simple online and real-time tracking (BoT-SORT) algorithm, a real-time detection-tracking-warning pipeline is established for continuous worker tracking, hazardous area intrusion recognition, and unsafe behavior warning.

(4) Extensive experiments, including comparative studies, ablation analysis, and real-world video validation, demonstrate that the proposed system achieves a favorable balance between detection accuracy, tracking effectiveness, and real-time performance.

The remainder of this paper is organized as follows. Section 2 presents the overall framework of the proposed system, including the YOLOv11n-IE detection model, the detector-tracker integration based on BoT-SORT, and the training configuration. Section 3 describes the experimental setup and provides comprehensive evaluations, including detection performance, ablation studies, tracking results, and real-world application analysis. Section 4 concludes the paper and discusses limitations and future research directions.

2. Methods

This section describes the proposed safety monitoring method for power grid construction scenarios. The framework follows a detection-tracking-warning pipeline based on YOLOv11n and BoT-SORT The system integrates safety equipment detection, target tracking, hazardous-area intrusion recognition, and unsafe-behavior warning to enable continuous safety monitoring in complex power environments.

2.1. Overall framework

This study proposes a practical detection-tracking-warning framework for safety monitoring in power grid construction scenarios. Fig. 1 illustrates the overall workflow of the proposed framework. The processing pipeline consists of several key stages. In the initialization stage, the system is initialized with a pretrained YOLOv11n model and the BoT-SORT tracking module. For each incoming video stream, individual frames are sequentially extracted as input. Subsequently, the YOLOv11n-IE detector performs safety equipment detection, producing bounding boxes, category labels, and confidence scores for each worker and their protective gear. These detection results are then fed into the BoT-SORT tracker, which associates objects across consecutive frames and assigns consistent identity labels, thereby forming continuous motion trajectories. Based on the obtained trajectories and spatial information, the system further performs hazardous-area intrusion detection and unsafe behavior analysis. When predefined safety rules are violated, the system triggers an alarm output in real time.

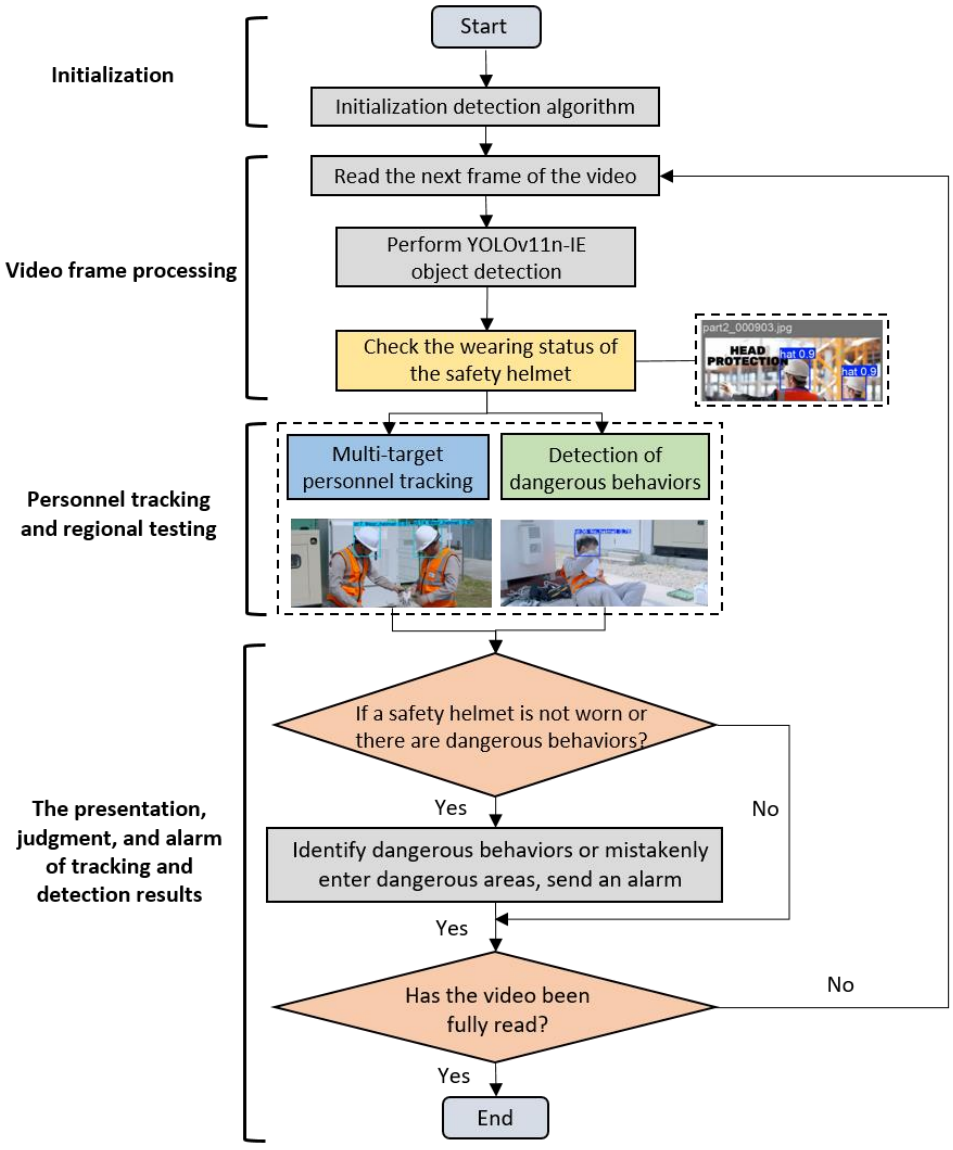


Fig. 1 Multi-target tracking and detection for power grid construction personnel

This integrated pipeline enables continuous monitoring, tracking, and safety assessment in complex power grid construction environments. To formally describe the detection-tracking process, let I_t denote the input video frame at time step t. The detector produces a set of detections for each frame, which is formulated as:

$$D_t = \left\{ \left(b_i, c_i, s_i \right) \mid i = 1, 2, ..., N_t \right\} \tag{1}$$

where b_i represents the bounding box coordinates of the i-th object, c_i denotes the predicted category label, s_i is the corresponding confidence score, and N_t is the number of detected objects in frame t.

These detections are then passed to the tracking module, which performs data association across consecutive frames to generate a set of trajectories:

$$T = \left\{ \tau_j \mid j = 1, 2, ..., m \right\} \tag{2}$$

where each trajectory τ_j consists of a sequence of observations corresponding to the same object over time, and m denotes the total number of tracked targets in the current video sequence. Based on the obtained trajectories and spatial-temporal information, the system performs hazard analysis, including hazardous area intrusion detection and unsafe behavior recognition. When predefined safety constraints are violated, an alarm signal is triggered.

2.2. Definition of YOLOv11n-IE

YOLOv11 is a recent member of the YOLO series designed for real-time object detection. In this study, the framework adopts it as the detector backbone because of its favorable balance between detection accuracy and inference efficiency. Herein, YOLOv11n-IE does not denote a newly redesigned YOLOv11 architecture. Instead, it refers to a scenario-adapted detection configuration built upon the pretrained YOLOv11n model. The backbone, neck, and head remain identical to those of the original YOLOv11n detector. However, the adaptation is mainly achieved through input-side enhancement strategies, including HSV perturbation, mosaic augmentation, and Mixup augmentation, together with the corresponding training configuration for power grid construction scenarios.

Specifically, the mosaic data augmentation technique generates synthetic images containing multiple targets and complex backgrounds by randomly cropping and stitching four images. This approach allows the model to learn diverse scene layouts during the training phase. This technique not only significantly increases the richness of small-target samples but also strengthens the model's feature extraction capability for complex compositions. It achieves this by simulating occlusion and overlapping states of targets in real scenes. Consequently, this strategy is particularly suitable for safety equipment detection tasks under multi-device interference in complex power scenarios. Additionally, the framework introduces the Mixup augmentation technique, which enhances the model's adaptability to complex scenes. This technique mixes and linearly interpolates two images and their annotation labels, further optimizing detection accuracy and robustness.

Furthermore, three-channel perturbations of brightness, contrast, and saturation are introduced on the basis of the RGB color space. and complex lighting conditions, such as strong light, backlighting, and cloudy days, are simulated by randomly adjusting relevant parameters. This technique enables model input samples to cover more diversified lighting-variant images. Therefore, it significantly improves detection stability in complex scenarios, such as strong light reflection in substations and high illumination during high-altitude operations.

These enhancement strategies are selected according to the visual characteristics of power grid construction scenarios. HSV perturbation is mainly used to simulate complex environmental illumination conditions, such as strong reflection, backlighting, and overcast weather. Mosaic augmentation enhances model robustness against dense backgrounds, small objects,

and multi-person layouts. Mixup augmentation improves the model's tolerance to ambiguous boundaries and hard-to-detect samples. Therefore, the augmentation design is not arbitrary. Instead, it directly addresses the practical challenges of safety equipment detection in power scenarios.

The design of the proposed framework is based on three considerations. First, power grid construction scenes often exhibit intense illumination variation, dense equipment layout, and partial occlusion. Therefore, the framework introduces training-time input enhancement to improve detector robustness under domain-specific visual perturbations. Second, frame-level detector outputs are susceptible to transient false positives and false negatives. To mitigate this volatility, the system incorporates trajectory-level association through BoT-SORT, which exploits temporal consistency to stabilize target identity and reduce frame-level fluctuation. Third, the system does not make warning decisions from isolated single-frame observations, but from temporally accumulated evidence along the same trajectory. This trajectory-level decision mechanism improves the reliability of hazardous area intrusion warning and persistent no-helmet warning in practical monitoring scenarios.

Fig. 2 shows the detector-side structure used in the present study. It should be noted that YOLOv11n-IE does not represent a newly redesigned detector architecture; rather, it retains the original YOLOv11n backbone, neck, and head, while the practical adaptation relies primarily on scenario-oriented training-time input enhancement. Therefore, Fig. 2 illustrates the detector backbone deployed within the framework, whereas the actual methodological adaptation of this study lies in the combination of this detector with scenario-oriented enhancement and the downstream tracking-warning pipeline.

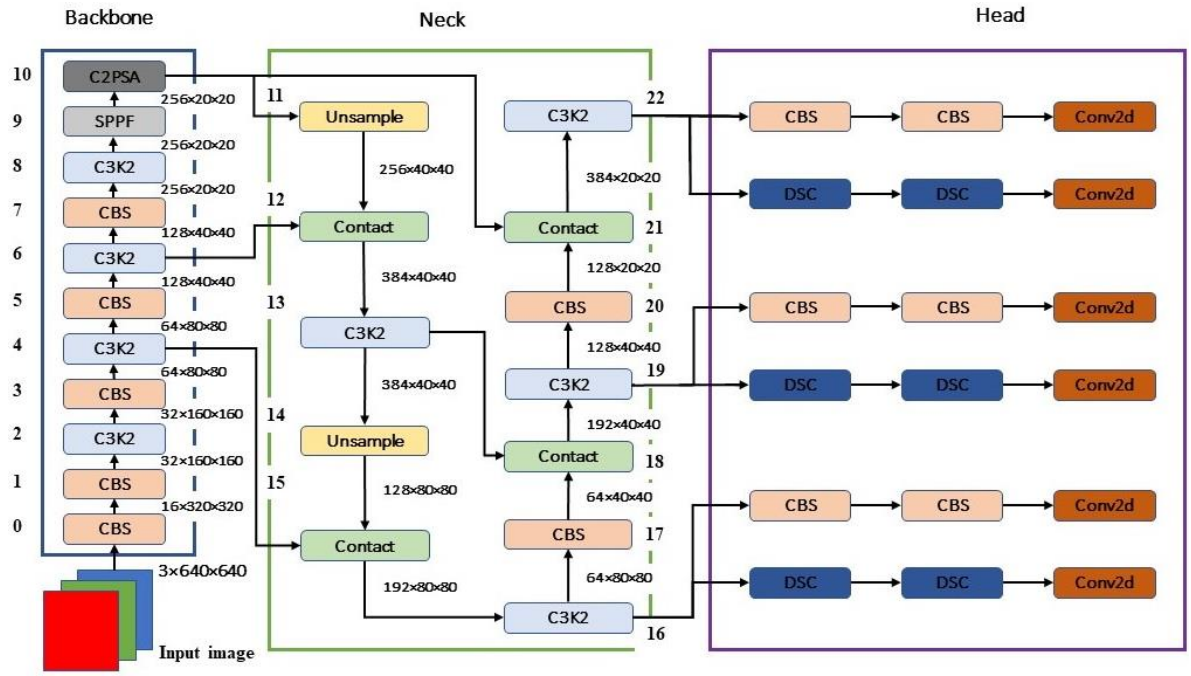


Fig. 2 Framework of the YOLOv11 model

(1) Backbone: the backbone is responsible for extracting multi-scale feature maps from raw image data. It progressively captures high-order features of targets by stacking CBS, C3k2, and C2PSA modules. The C3K2 block, an evolved version of the cross stage partial (CSP) bottleneck structure, splits feature maps and applies a series of smaller 3×3 convolutional kernels. This design reduces parameter count while enhancing feature representation, thereby maintaining the model's ability to capture key image features efficiently.

(2) Neck: the neck component acts as an intermediate processing stage, using dedicated layers to aggregate and enhance feature representations at different scales. It constructs a convolutional pyramid with Upsample, C3k2, and CBS modules, blending shallow features extracted by the backbone to enable the network to learn prominent target features. The neck design in YOLOv11 further optimizes feature fusion and information flow, making the model better suited for multi-scale target detection tasks.

(3) Head: the head component generates final target localization and classification outputs based on the feature maps processed by the neck. It consists of two branches: a “2many” branch (mapping model features to multiple output results) and a “2one” branch (aggregating multiple features or prediction results into a unified output), followed by non-maximum suppression-free training and inference. The head’s structure varies according to the specific visual task, such as object detection, pose detection, or instance segmentation.

YOLOv11’s head design prioritizes prediction accuracy and efficiency, ensuring rapid and precise target detection in real-time applications. Through architectural innovations and engineering optimizations, YOLOv11 achieves multi-dimensional performance breakthroughs, which include the following key aspects:

(1) Enhanced feature extraction: an improved backbone and neck structure boost feature extraction accuracy and robustness, enabling more efficient target detection in complex scenes.
(2) Optimized efficiency: a refined architecture and optimized training workflow maintain high accuracy while delivering faster inference speeds, successfully balancing detection efficiency and precision.
(3) Reduced parameter overhead: YOLOv11m achieves a higher mAP on the COCO dataset with 22% fewer parameters than YOLOv8m, significantly enhancing computational efficiency without sacrificing accuracy.
(4) Versatile deployment and task support: the model supports multiple environments (including edge devices, cloud platforms, and NVIDIA GPU-accelerated systems), adapting to diverse hardware requirements. Furthermore, it can handle object detection, instance segmentation, image classification, pose estimation, and oriented object detection, providing a unified solution for complex scene requirements.
(5) Real-time danger warning: high-precision target detection enables timely anomaly detection and alerting, which is critical for real-time application scenarios such as security monitoring and hazard warning.

2.3. Detector-tracker integration

YOLO-based detectors focus on frame-level object localization and classification, providing bounding boxes and category labels for each input frame. In contrast, tracking algorithms such as BoT-SORT perform cross-frame association by linking detection results over time and assigning consistent identity labels to targets.

In the proposed framework, YOLOv11n-IE is employed to detect safety equipment on workers by frame. Output detection results are delivered to the BoT-SORT module, which associates cross-frame bounding boxes and stabilizes target identities to form an integrated detection-tracking pipeline.

It should be noted that this study retains the original internal architecture of BoT-SORT and directly integrates it into the proposed system as a reliable tracking module for power grid worker monitoring. Adopting the unmodified off-the-shelf BoT-SORT serves as a stable, reproducible tracking baseline. This choice enables our research to concentrate on verifying the practical performance of trajectory-based early warning for power grid monitoring instead of developing a novel tracking algorithm. The obtained trajectories are further utilized to realize hazardous area intrusion identification and unsafe behavior alerts.

2.4. Training configuration

All experiments in this study are conducted under a unified training configuration to ensure fair comparison among different models. The YOLOv11n backbone is initialized with pretrained weights from the COCO dataset. The input image resolution is set to 640×640. The framework trains the model using the Adam optimizer with an initial learning rate of 0.001. A cosine annealing learning rate schedule is adopted to improve training stability. The batch size is set to 64, and the total number of training epochs is 200. The dataset is randomly split into training and validation sets with a ratio of 8:2.

During training, the proposed input enhancement strategies, including HSV perturbation, mosaic augmentation, and Mixup augmentation, are applied to improve the robustness of the detector. Conversely, during inference, all augmentation operations are disabled to ensure consistent and objective evaluation. Moreover, the default BoT-SORT implementation and parameter settings are used throughout this study. Table 2 presents the specific configuration of the model parameters. Crucially, all compared YOLO variants are trained under the exact same data split, input resolution, optimizer type, learning rate schedule, and number of epochs to guarantee absolute consistency across evaluations.

Table 2 Specific parameter settings for YOLOv11n-IE model

Input size	640×640	Augmentation	HSV+Mosaic+Mixup
Label	helmet/no-helmet	Pre-training dataset	COCO
Training cycles	200	Batch size	64
Optimizer	Adam	Initial learning rate	0.001
Scheduler	Cosine Annealing	Train/val split	8:2

Furthermore, the YOLOv11n detector is optimized using the default detection objective of the adopted implementation, which can be written in the general form as:

$$L_{det} = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} \tag{3}$$

where L_{det} denotes the detector training loss; L_{box} and λ_{box} represent the bounding-box regression term and corresponding weight coefficient; L_{cls} and λ_{cls} denot e the classification term and corresponding weight; as well as L_{reg} and λ_{reg} represent the localization refinement term and corresponding weight. The weighting coefficients follow the default implementation setting of the adopted YOLOv11n training framework.

3. Results

This section presents the experimental results and performance analysis of the proposed framework. Specifically, this section is organized as follows: (i) introducing the construction of the E-hat dataset, including its data sources, annotation process, and relevance to power grid construction scenarios; (ii) describing the experimental environment and evaluation metrics to clarify the basis for quantitative assessment; (iii) analyzing the training process and detection performance of YOLOv11n-IE are analyzed through loss curves, metric evolution, and visual prediction results; (iv) conducting comparative experiments with several lightweight YOLO variants to evaluate the accuracy, efficiency, and real-time performance of the proposed detector under the same settings; (v) performing ablation experiments to examine the contribution of different input enhancement strategies; and (vi) presenting real-world video tests to verify the effectiveness of the integrated detection-tracking-warning pipeline in worker tracking, hazardous-area intrusion recognition, and unsafe-behavior warning.

3.1. Construction of safety equipment detection dataset for power scenarios

In power industry safety production management, safety equipment wearing detection is a core measure to protect workers' lives. However, the lack of publicly available datasets for safety equipment detection in complex power scenarios (e.g., substations, patrol environments) poses challenges for model training [18]. To address this issue, a dedicated dataset for complex power scenarios, designated as E-hat, by integrating the public SHWD with real-world power scenario images.

The E-hat dataset is composed of two parts:

(1) The public dataset SHWD: this component contains 7,581 images of substation and industrial scenes, covering 9,044 safety helmet-wearing targets (positive samples) and 111,514 non-wearing targets (negative samples).

(2) The newly added special power scenario data (New_added): this component consists of 295 on-site images with significant power industry characteristics, obtained by field-collecting typical scenarios such as high-low voltage switchgear areas, transformer operation areas, and high-altitude hoisting operations.

Through data fusion, a total of 7,876 images form the complete dataset, which are divided into a training set (6,301 images) and a validation set (1,575 images) at an 8:2 ratio. These reliable data provide a rich foundation for downstream model training. Table 3 details the composition of each part of the dataset.

Table 3 Statistics of various indicators in the dataset

Dataset	Number of images	Positive samples	Negative samples	All samples
SHWD	7581	9044	111514	120558
New_added	295	724	36	760
E-hat	7876	9768	111550	121318

In the data preprocessing stage, the framework uses the LabelImg tool to carefully annotate the original images. The annotation format follows the visual object classes (VOC) standard, clearly classified into two categories: wearing a safety helmet ("helmet") and not wearing a safety helmet ("no-helmet"). Subsequently, the annotated data is converted into the training format of the YOLOv11 dataset to meet the requirements of model training.

It should be noted that the newly collected power-specific subset remains relatively limited in size. Therefore, the purpose of constructing E-hat is not to claim complete coverage of all possible power grid construction conditions, but to enhance scenario relevance under realistic data availability constraints. By combining a public safety helmet dataset with real field images, the dataset aims to improve the detector's adaptation to practical power-scene characteristics.

Furthermore, the class distribution of E-hat is imbalanced, especially in the public subset, where non-helmet/person instances are much more frequent than helmet instances. In this study, class imbalance is not addressed through a specially redesigned loss function. Instead, it is mitigated mainly through three practical strategies: pretrained initialization to improve representation stability, scenario-oriented augmentation to enhance data diversity, and multi-metric evaluation so that performance is not assessed by a single metric alone. Together, these strategies help improve the robustness of the detector under imbalanced data conditions. Nevertheless, the current handling of class imbalance remains limited, and future work should further explore larger and more balanced data collection or dedicated imbalance-aware optimization strategies.

3.2. Experimental environment and evaluation metrics

This study is conducted under the Ubuntu operating system environment, with an experimental platform constructed based on Python 3.10.14 and the PyTorch 2.6.0 + cuda12.4 framework. The experimental equipment utilizes an NVIDIA RTX 3090 GPU (with 24G memory), and the safety equipment detection model is trained and tested on the newly constructed dataset E-hat. The detailed detector training configuration is provided in Table 2 of Section 2.4.

To comprehensively evaluate model performance, a multi-dimensional evaluation system is constructed:

First, model performance evaluation: precision (P), recall (R), mean average precision (mAP@0.5), and the more strictly mean average precision (mAP@0.5:0.95) are used as evaluation metrics. Precision reflects the proportion of predicted positive samples that are truly positive, expressed as:

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Recall represents the proportion of actual positive samples correctly identified by the model, formulated as:

$$recall = \frac{TP}{TP + FN} \tag{5}$$

Herein, mAP@0.5 is the mean of the average precision at the intersection-over-union (IoU) threshold of 0.5, which is used to measure the overall performance of the model under the standard IoU threshold. Among them, Average Precision (AP)

is the result obtained by integrating the area under the precision-recall (P-R) curve, which is used to measure the detection performance of the model in a category. The IoU is used to measure the degree of overlap between the predicted box and the true box, and the formula is as follows:

$$IoU = \frac{\text{The intersection area of the prediction box and the true box}}{\text{The union area of the prediction box and the true box}} \tag{6}$$

mAP@0.5:0.95 is the average mAP over IoU thresholds ranging from 0.5 to 0.95 (step size 0.05), which is used to more rigorously evaluate the robustness of the model under different degrees of overlap.

Second, model complexity assessment: The complexity of the model is evaluated through the number of parameters (M) and the inference speed frames per second (FPS) to understand the computing resources required by the model and the actual detection speed of the model. The FPS formula is as follows:

$$FPS = \frac{\text{Total number of frames}}{\text{The consumed time}} \tag{7}$$

The parameter count reported in this study refers to the total number of learnable parameters of the detector. The FPS values are measured on an NVIDIA RTX 3090 GPU under an input resolution of 640×640 during inference. The reported speed corresponds to the detector running under the same experimental environment for all compared models. It should be noted that the reported FPS values are measured under a single fixed software setting and are intended mainly to provide a practical runtime reference for model comparison under the same environment.

3.3. Analysis of YOLOv11 model training results

Using the E-hat dataset, the YOLOv11n model with the parameters described is trained and tested. Fig. 3 presents both the optimization process and the metric evolution during training. The loss curves show a generally consistent downward trend for the training and validation sets, suggesting stable optimization under the adopted training setting. At the same time, the precision, recall, mAP@0.5, and mAP@0.5:0.95 curves gradually converge and become relatively stable in the later training stage, indicating that the detector performance does not rely on isolated fluctuations at a single epoch. Therefore, Fig. 3 supports the claim that the adopted training configuration enables stable detector learning on E-hat.

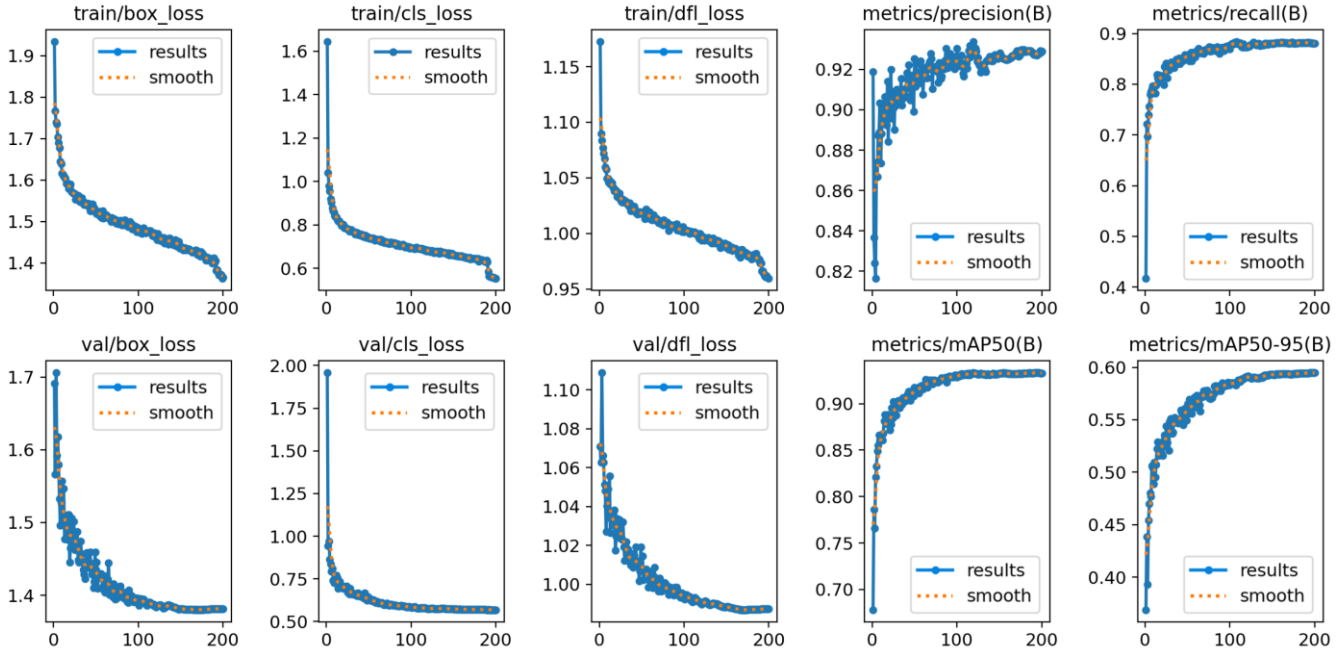


Fig. 3 Curves of training and test for the YOLOv11 model

Among them, the validation set curve of mAP@0.5:0.95 exhibits minimal fluctuations and is negligible after 150 epochs, further verifying the stability of the model learning. This also jointly verifies the adaptability of the model architecture to the training strategy, providing a reliable model basis for the engineering deployment of safety equipment detection tasks in power scenarios.

In addition, through the systematic evaluation of the annotation quality of the validation set and the visual analysis of the prediction results, this study finds that the real labels show a high degree of consistency with the prediction results. Fig. 4 presents qualitative examples of prediction results on the validation set. The figure is intended to visually compare ground-truth annotations and model predictions in representative power-scene images. It shows that the detector can generally localize target regions and distinguish helmet-wearing from non-helmet cases under typical scene conditions. Thus, Fig. 4 serves as qualitative support for the quantitative detection results reported in the comparison tables.

As a result, almost all the labeled samples are successfully predicted, and the overall precision rate reaches 0.93. This indicates that the model can capture the features of safety equipment during the learning proc0.93ess and accurately identify them from complex power scenarios. The highly accurate prediction results further prove the reliability and effectiveness of the model in this task, suggesting practical potential for safety equipment monitoring in actual power scenarios.



(a) The true labels on the validation set (b) The predicted labels on the validation set

Fig. 4 The true and predicted label results on the validation set

3.4. Comparative experiments

To better demonstrate the effectiveness of the proposed approach, this study considers both literature-level comparison and fair same-setting quantitative comparison. The former is used to position the present study relative to representative recent methods, whereas the latter is conducted among lightweight YOLO variants retrained on the same E-hat dataset and unified experimental settings. This strategy is adopted because direct numerical comparison across published studies may be influenced by differences in datasets, scene complexity, label definitions, and evaluation protocols.

Specifically, to evaluate the effectiveness of the proposed framework, comparative experiments are conducted using four lightweight YOLO variants, namely YOLOv5n, YOLOv8n, YOLOv10n, and YOLOv11n, under the same training settings and evaluation criteria on the E-hat dataset. All models are trained with identical input resolution, optimization strategy, and number of epochs to ensure a fair comparison. The evaluation metrics include P, R, mAP@0.5, mAP@0.5:0.95, model parameter count, and FPS.

In this context, YOLOv11n denotes the baseline pretrained detector using the default training configuration, while YOLOv11n-IE denotes the same pretrained YOLOv11n backbone further trained with the proposed scenario-oriented input enhancement strategies for power grid construction scenarios. Table 4 shows the specific difference between the YOLOv11n and YOLOv11n-IE models.

Table 4 The specific difference between the YOLOv11n and YOLOv11n-IE models

Item	YOLOv11n	YOLOv11n-IE
Backbone/Neck/Head	Original YOLOv11n	Original YOLOv11n
Pretrained weights	COCO	COCO
Loss setting	Default YOLOv11n setting	Default YOLOv11n setting
Optimizer	Adam	Adam
Initial learning rate	0.001	0.001
Learning-rate schedule	Cosine annealing	Cosine annealing
Input size	640×640	640×640
Batch size	64	64
Input enhancement	No additional enhancement	HSV+Mosaic+Mixup
Scenario adaptation	General baseline	Power grid construction scenario

The specific results are shown in Table 5. As a result, YOLOv11n achieves a more competitive performance than other versions of YOLO. In particular, YOLOv11n-IE outperforms the benchmark model on four model performance metrics of P, R, mAP@0.5, and mAP@0.5:0.95, while 2.58 M of parameters is 14.1% less than YOLOv8n. In addition, the detection speed of the YOLOv11n-IE model is 5.47 ms, and the inference speed is 183, which is higher than that of the previous YOLO series version comparison models, reflecting its real-time detection ability. Although the inference speed is slightly lower than that of the pre-trained YOLOv11n model, the processing speed of the model can still meet the actual real-time monitoring requirements. Compared with YOLOv11n, YOLOv11n-IE improves recall from 0.871 to 0.877 and mAP@0.5 from 0.926 to 0.934 with only a marginal speed drop from 186 to 183 FPS. The above indicator values reflect the powerful predictive performance of YOLOv11n-IE.

Table 5 Results of comparative experiments

Model	P	R	mAP@0.5	mAP@0.5:0.95	Parameter (M)	Inference speed (FPS)
YOLOv5n	0.914	0.860	0.919	0.571	2.503334	139
YOLOv8n	0.914	0.866	0.923	0.591	3.006038	156
YOLOv10n	0.910	0.848	0.916	0.578	2.695196	137
YOLOv11n	0.926	0.871	0.926	0.588	2.582542	186
YOLOv11n-IE	0.927	0.877	0.934	0.594	2.582542	183

The superior performance of YOLOv11n-IE may be attributed to two aspects. First, compared with earlier lightweight YOLO variants, YOLOv11n provides a more favorable balance between feature extraction efficiency and inference speed. This balance is important for real-time deployment in power grid construction scenarios. Second, the scenario-oriented input enhancement strategies further improve robustness to complex illumination, dense equipment backgrounds, small objects, and partial occlusion. This is supported by the comparison between YOLOv11n and YOLOv11n-IE in Table 5, where the latter achieves higher recall and mAP while maintaining nearly the same model size and inference speed. Therefore, the advantage of YOLOv11n-IE lies not only in the baseline detector itself, but also in its adaptation to the visual characteristics of power grid construction environments.

From a deployment perspective, YOLOv11n-IE achieves a favorable trade-off between complexity and performance. Since the proposed detector retains the lightweight YOLOv11n backbone unchanged, the parameter count remains low. Meanwhile, the additional scenario-oriented enhancement strategies improve detection robustness without introducing extra detector parameters. Although the integration of tracking and warning modules incurs system-level computation, the overall framework still satisfies real-time requirements in practical power grid monitoring scenarios.

Despite the performance advantages, several limitations should be noted. First, severe occlusion may still lead to missed detections or inaccurate helmet detection, especially when the head region is partially or fully blocked by equipment or other workers. Second, the dataset used in this study, although enhanced with real power grid images, is still limited in scale and

diversity, which may affect the generalization capability of the model in more complex real-world scenarios. These limitations indicate that further improvements in occlusion handling and dataset expansion are necessary for practical deployment. Furthermore, computational complexity should be understood from both the detector side and the system side.

From the detector perspective, YOLOv11n-IE keeps the same backbone, neck, and head as the baseline YOLOv11n; therefore, it does not introduce additional detector parameters or architectural complexity at inference time. The proposed HSV perturbation, mosaic augmentation, and Mixup augmentation are used only during training but are disabled during inference, thereby avoiding any increase in detector-side runtime complexity in deployment. From the system perspective, the additional computational overhead mainly comes from the BoT-SORT tracking module and the subsequent warning analysis based on trajectories. Nevertheless, the overall framework still maintains real-time performance, as reflected by the reported FPS values under the same evaluation environment. Therefore, the contribution of YOLOv11n-IE is to improve robustness under power-grid-specific visual conditions without increasing detector-side model complexity.

3.5. Ablation experiments

To verify the impact of different image augmentation modules on the performance of safety equipment detection in power grid construction scenarios, this study systematically conducted ablation experiments. These experiments are based on the E-hat dataset within the YOLOv11 model framework. This is done to comprehensively evaluate the independent contributions and synergistic effects of three major categories of strategies: HSV color space augmentation (including the hue (H), saturation (S), and value (V) channels), mosaic multi-image mosaic, and Mixup blending augmentation.

Keeping other training parameters consistent (pre-trained YOLOv11n architecture, 640×640 input size, Adam optimizer, and 200 training epochs), the H/S/V components of HSV augmentation and the mosaic and Mixup modules are removed. Subsequently, seven ablation experiments are implemented to analyze the independent effects of these augmentation modules. Specifically, the seven comparison models used in the experiments include the model YOLOv11n (baseline, with no augmentation), a model with the complete augmentation combination (YOLOv11n-IE), and five ablation groups (YOLOv11n-h, YOLOv11n-s, YOLOv11n-v, YOLOv11n-m, and YOLOv11n-mx). In addition, this study quantitatively assesses the models based on four metrics: precision (P), recall (R), mAP@0.5, and mAP@0.5:0.95. The specific results of the ablation experiments are shown in Table 6.

Table 6 Results of ablation experiments

Model	H	S	V	Mosaic	Mixup	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv11n	×	×	×	×	×	0.926	0.871	0.926	0.588
YOLOv11n-h	×	√	√	√	√	0.924	0.878	0.934	0.591
YOLOv11n-s	√	×	√	√	√	0.928	0.875	0.931	0.593
YOLOv11n-v	√	√	×	√	√	0.920	0.880	0.931	0.581
YOLOv11n-m	√	√	√	×	√	0.929	0.871	0.932	0.589
YOLOv11n-mx	√	√	√	√	×	0.922	0.875	0.930	0.590
YOLOv11n-IE	√	√	√	√	√	0.927	0.877	0.934	0.594

The experimental results show that the fully augmented model achieves the highest mAP@0.5:0.95 (59.4%) and tied-best mAP@0.5 (93.4%), indicating that the combined augmentation strategies contribute positively to detection performance. Although the baseline model achieves competitive precision (P, 0.926) and recall (R, 0.871) in a single scenario, its mAP@0.5 (92.6%) and mAP@0.5:0.95 (58.8%) remain lower than those of the fully augmented model. This difference further demonstrates the improvement of data augmentation in detection capability under complex IoU thresholds.

In terms of precision, the model with Mosaic augmentation removed (YOLOv11n-m) exhibits the highest performance. This suggests that with Mosaic augmentation increases the diversity of the training data by concatenating multiple images together. However, it may also introduce some irrelevant background information, thereby compromising detection precision.

In comparison, the complete model achieves the best mAP@0.5:0.95 and maintains balanced precision and recall. This indicates that the combination of augmentation techniques helps the model better learn equipment features in power grid construction scenarios and improve detection comprehensiveness.

Regarding recall, the experimental group with the V channel removed (YOLOv11n-v) achieves the highest R value of 0.880. This implies that removing V-channel perturbation helps preserve brightness-related features and increases the detection rate in some bright scenes; however, this comes at the expense of a 0.003 decrease in mAP@0.5, from 0.934 to 0.931. Disabling mixup results in a 0.004 decrease in mAP@0.5, from 0.934 to 0.930, and a 0.004 decrease in mAP@0.5:0.95, from 0.594 to 0.590. In addition, precision decreases by 0.005, from 0.927 to 0.922. This reflects that mixup helps enhance the model's ability to distinguish between safety equipment-wearing and non-wearing states, thereby improving detection stability.

Compared with the baseline model, models augmented with the HSV color space, Mosaic multi-image mosaic, and Mixup blending all exhibit higher performance. Overall, HSV augmentation, by adjusting hue, saturation, and brightness, helps the model adapt to equipment detection under different colors and lighting conditions. Mosaic augmentation, by increasing the diversity of training data, enhances the model's generalization capability. Mixup augmentation, by mixing image features, further strengthens the model's robustness. These results provide valuable references for optimizing YOLOv11 in practical safety equipment detection and recognition applications in complex power grid construction scenarios.

Overall, the ablation results indicate that the performance gain does not come from a single enhancement module alone, but from the complementary effects of multiple strategies. HSV augmentation mainly improves robustness under illumination variation, mosaic contributes to the understanding of cluttered layouts and small objects, and mixup helps reduce overfitting and improves tolerance to hard samples. The best overall performance of YOLOv11n-IE suggests that the full combination of these strategies is more suitable for power grid construction scenarios than any partial configuration.

3.6. Analysis of multi-object tracking and detection results in power grid construction sites

In the field operations of power grid construction, multi-object tracking and detection technology is of great significance for ensuring the safety of construction workers. By monitoring the locations and behaviors of workers in real time, it can effectively prevent the occurrence of safety accidents and optimize the workflow. In this study, based on the features of safety equipment and integrated with the pre-trained YOLOv11 model, this research conducts multi-object tracking and detection of workers in power grid construction scenarios. Subsequently, a detailed analysis of the experimental results is performed. The research data comes from two types of actual operation video sequences collected from the web in power grid scenarios.

Figs. 5-6 illustrate the two representative real-world video scenarios used for tracking evaluation. Specifically, scenario 1 (Fig. 5) records the process of two power grid construction workers entering the work area successively; scenario 2 (Fig. 6) focuses on the continuous behavior of workers taking off their helmets, wiping sweat, and putting them back on. Therefore, these two figures represent two different but complementary tracking challenges in practical power-grid monitoring. Each of the two scenarios contains 200 video sequences (frames), providing a rich sample for verifying the performance of the model. Moreover, the web-collected video sequences used for tracking and warning evaluation are not included in the training or validation subsets of E-hat.



Fig. 5 Power grid construction scenario 1



Fig. 6 Power grid construction scenario 2

In this study, the pre-trained YOLOv11n-IE model is employed for inference, and the number of detected people/total number of people, accuracy, multiple object tracking accuracy (MOTA), and FPS are selected as the basic evaluation criteria to quantitatively assess the model's actual performance in power grid scenarios. The specific calculation formula for MOTA is as follows:

$$MOTA = 1 - \frac{\sum_{t=1}^{N} FN_t + FP_t + IDSW_t}{\sum_{t=1}^{N} GT_t} \tag{8}$$

where N is the total number of frames in the video sequence (200 frames in this paper). False Negative at frame t (FNt) represents the number of missed detections in each frame. False Positive at frame t (FPt) represents the number of false detections in each frame. ID Switch (IDSWt) indicates the number of times the ID switches between frames (the same target is assigned different IDs in different frames). Ground Truth (GTt) represents the total number of true targets. t represents the frame index, and MOTA calculates the cumulative error ratio across all frames. Table 7 presents the multi-object tracking results under different scenarios.

Table 7 Multi-target tracking results in different power scenarios

	The number of detected/total number of people	Accuracy	MOTA	FPS
Scenarios 1	345/349	98.85%	97.71%	183
Scenarios 2	182/200	91%	86%	185

As can be seen from the results in Table 7, the accuracy of the multi-object tracking and detection by the model in this video segment reaches over 90%, indicating that YOLOv11n-IE can be effectively applied to personnel tracking and detection in real-world scenarios. Meanwhile, the MOTA values of the model in the two scenarios are 97.71% and 86%, respectively. This performance shows that the model has good overall tracking performance in different working scenarios, and issues such as detection errors, target loss, and ID switching have a relatively small impact on its performance. Overall, these results suggest that YOLOv11n-IE has practical potential for worker tracking and detection in power grid scenarios. However, because the current evaluation is based on only two short web-collected video sequences, further validation on larger and more diverse real-site videos is still necessary.

3.7. Analysis of detection results for workers' accidental entry into hazardous areas and hazardous behaviors in power grid construction sites

After detecting worker safety equipment with YOLOv11n-IE, BoT-SORT generates continuous cross-frame worker trajectories. Based on these trajectories, the system triggers two types of safety warnings: (i) an intrusion alert is activated when a worker's bounding-box center stays within manually predefined hazardous zones over consecutive frames. (ii) helmet-violation warnings adopt persistent multi-frame trajectory-linked detection instead of single-frame judgment. Benefiting from this detection-tracking-analysis pipeline, the system supports real-time localization and early warning of irregular on-site behaviors to support grid construction safety management. Typical warning cases are shown in Fig.7 to demonstrate the framework's trajectory-aware warning advantage over conventional single-frame detection.



(a) Workers' accidental entry into hazardous areas



(b) Hazardous behaviors of workers

Fig. 7 Case examples of power grid construction workers entering restricted hazard areas or performing unsafe operations

In the experimental part, video sequence data of workers in power grid construction sites accidentally entering hazardous areas or engaging in hazardous behaviors (such as not wearing safety equipment and abnormal behaviors on the construction site) are collected. The data include 100-frame video sequences for cases of accidental entry into hazardous areas and 50-frame sequences for cases of hazardous behaviors. Based on this, the trained and optimized YOLOv11n-IE model is used to identify and detect hazardous behaviors or accidental entry into hazardous areas in power grid construction sites and to issue warning information.

Table 8 Detection results of workers entering dangerous areas or performing unsafe operations

Case	The number of detected/total number of people	Accuracy	MOTA	FPS
Accidental Entry into Hazardous Areas	81/89	91.01%	87.64%	184
Hazardous Behaviors	39/45	86.67%	77.78%	183

The experimental results presents Table 8 show that in the detection of accidental entry into hazardous areas, 81 out of 89 actual cases are detected, with an accuracy rate of 91.01% and a MOTA value of 87.64%, and the frame rate reaches 184 FPS. In the detection of hazardous behaviors, 39 out of 45 cases are detected, with an accuracy rate of 86.67% and a MOTA value of 77.78%, and the frame rate is 183 FPS. It can be seen that the YOLOv11n-IE model performs well in identifying accidental entry into hazardous areas in power grid construction sites, with both high accuracy and real-time performance.

For hazardous behavior detection, the model can achieve effective identification, but there is still room for improvement in accuracy and MOTA value. The main reasons are the large-area occlusions in the video sequences used in this study and the relatively small number of frames selected in the video sequences. Subsequent work could further optimize the model parameters or adjust the data collection strategy to enhance the performance of hazardous behavior detection and better ensure the safety of power grid construction sites. Nevertheless, these warning results should be interpreted with caution, as the current evaluations of hazardous areas and unsafe behaviors are based on a limited number of short video sequences. More extensive real-world validation is required before drawing stronger conclusions about deployment-level robustness.

3.8. Failure cases and limitations

Although the proposed system achieves promising results in both worker tracking and hazardous-event warning, several limitations remain. First, severe occlusion may cause partial or complete disappearance of helmet regions, leading to missed detections or inconsistent helmet-wearing judgments. Second, in scenes with strong specular reflection, backlighting, or rapidly changing illumination, the visual characteristics of the head region may be weakened, which affects detection reliability. Third, multi-person overlap and motion crossing may still lead to identity switches in tracking. Fourth, the present study does not include a dedicated comparison with alternative tracking baselines; therefore, the relative advantage of the BoT-SORT-based tracking module over other tracker choices is not yet systematically verified under the current power-grid monitoring scenario. Fifth, although real-world video experiments are conducted, the validation scale is still limited, because the current evaluation is based on only a small number of short web-sourced video sequences and hazardous-behavior samples.

Therefore, the current real-world results should be interpreted as preliminary evidence of practical feasibility rather than comprehensive deployment-level validation. These limitations indicate that further studies should include broader on-site video collection, larger-scale warning evaluation, and comparative experiments with different tracking baselines under the same detector setting.

4. Conclusions

This study established a real-time detection-tracking-warning framework for safety monitoring of power grid construction workers. The key findings are summarized as follows:

(1) A dedicated E-hat dataset was constructed by integrating public SHWD images with newly collected field images from power operation scenarios, providing data support for helmet/no-helmet detection under complex backgrounds.

(2) Based on pretrained YOLOv11n, a scenario-adapted configuration, YOLOv11n-IE, was established using HSV perturbation, mosaic augmentation, and Mixup augmentation, while retaining the original YOLOv11n backbone, neck, and head.

(3) Experimental results showed that YOLOv11n-IE achieved 92.7% precision, 93.4% mAP@0.5, and 183 FPS, indicating a favorable balance between accuracy and real-time efficiency.

(4) Compared with other lightweight YOLO variants, the proposed configuration improved detection robustness without increasing detector-side model parameters. Ablation experiments further confirmed the complementary effects of the three augmentation strategies in handling illumination variation, dense backgrounds, small objects, and hard samples.

(5) By integrating BoT-SORT, the system extended frame-level detection to trajectory-level monitoring, enabling hazardous-area intrusion recognition and persistent unsafe-behavior warning in real-world video tests.

Overall, these results suggest the practical potential of the proposed framework for intelligent safety management in power grid construction sites.

However, the current study still has limitations, including limited power-specific data, small-scale real-world video validation, possible degradation under severe occlusion or strong backlighting, and the lack of systematic comparison with alternative trackers. Future work will focus on expanding real-site datasets, improving robustness under complex scenarios, and evaluating different tracking modules under unified settings.

Funding

This research was funded by Guizhou Power Grid Co., Ltd., grant number 060000KC23100011.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] M. K. Mahadi, R. Rahad, M. S. Mobassir, A. Rahman, A. Shafiullah, and M. M. Nishat, "Implementation of Personal Safety Equipment Tracking & Detection by Deepsort & Yolov8," Proceedings of 2024 International Conference on Inventive Computation Technologies (ICICT), IEEE, pp. 1969-1974, 2024.

[2] S. Rasouli, Y. Alipouri, and S. Chamanzad. "Smart Personal Protective Equipment (PPE) for Construction Safety: A Literature Review," Safety Science, vol. 170, article no. 106368, 2024.

[3] R. Yamaguchi, Y. Makino, S. Torimitsu, G. Inokuchi, F. Chiba, M. Yoshida, et al., "Occupational Accidental Injury Deaths in Tokyo and Chiba Prefectures, Japan: A 10-Year Study (2011–2020) of Forensic Institute Evaluations," Journal of Forensic Sciences vol. 68, no. 1, pp. 185-197, 2023.

[4] M. Çiftçi, M. U. Türkdamar, and C. Öztürk. "Leveraging Yolo Models for Safety Equipment Detection on Construction Sites," Journal of Computing Theories and Applications, vol. 1, no. 4, pp. 492-506, 2024.
[5] H. Wang, G. Wang, D. Lou, Y. Liu, L. Zhang and J. J. Fu, "Research on Power Scene Abnormal Recognition Technology Based on AI Edge Depth Algorithm Video Analysis Device," Power Systems and Big Data, vol. 24, no. 11, pp. 1-8, 2021.
[6] B. Liu, W. Zhao, and Q. Sun, "Study of Object Detection Based on Faster R-CNN," Proceedings of 2017 Chinese automation congress (CAC), IEEE, pp. 6233-6236, 2017.
[7] Sumit, B. Sumit, S. Joshi, and U. Rana. "Comprehensive Review of R-CNN and Its Variant Architectures," International Research Journal on Advanced Engineering Hub, vol. 2, no. 04, pp. 959-966, 2024.
[8] S. Chen, W. Tang, T. Ji, H. Zhu, Y. Ouyang, and W. Wang, "Detection of Safety Helmet Wearing Based on Improved Faster R-CNN," Proceedings of 2020 International joint conference on neural networks (IJCNN), IEEE, pp. 1-7, 2020.
[9] T. YaJie and P. Lian, "Improved Lightweight Helmet Wearing Detection Method for Yoloxs," Proceedings of 2022 China Automation Congress (CAC), IEEE, pp. 737-741, 2022.
[10] J. Deng, X. Xuan, W. Wang, Z. Li, H. Yao, and Z. Wang, "A Review of Research on Object Detection Based on Deep Learning," Journal of physics: Conference series, vol. 1684, no. 1, article no. 012028, 2020.
[11] W. Xu, Y. Zhao, X. Du, H. Ji, and L. Xing. "A Study on Bus Passenger Boarding and Alighting Detection and Recognition Based on Video Images and Yolo Algorithm," Sensors, vol. 26, no. 5, article no. 1418, 2026.
[12] F. Zhou, H. Zhao, and Z. Nie, "Safety Helmet Detection Based on Yolov5," Proceedings of 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), IEEE, pp. 6-11, 2021.
[13] S. Tan, G. Lu, Z. Jiang, and L. Huang, "Improved Yolov5 Network Model and Application in Safety Helmet Detection," Proceedings of 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR), IEEE, pp. 330-333, 2021.
[14] Y. Jiang, S. Guo, L. Li, Z. Sheng, X. Song, and Y. Wei, "Research on Self Inspection Algorithm of Field Operation Based on Improved Yolov8n," Proceedings of 2025 4th International Symposium on Robotics, Artificial Intelligence and Information Engineering (RAIIE), pp. 60-66, 2025.
[15] B. Lin. "Safety Helmet Detection Based on Improved Yolov8," IEEE Access, vol. 12, pp. 28260-28272, 2024.
[16] L.-h. He, Y.-z. Zhou, L. Liu, W. Cao, and J.-h. Ma. "Research on Object Detection and Recognition in Remote Sensing Images Based on Yolov11," Scientific Reports, vol. 15, no. 1, article no. 14032, 2025.
[17] X. He, X. Chen, X. Du, X. Wang, S. Xu, and J. Guan. "Maritime Target Radar Detection and Tracking via DTNet Transfer Learning Using Multi-Frame Images," Remote Sensing, vol. 17, no. 5, article no. 836, 2025.
[18] L. Huang, Q. Fu, M. He, D. Jiang, and Z. Hao. "Detection Algorithm of Safety Helmet Wearing Based on Deep Learning," Concurrency and Computation: Practice and Experience, vol. 33, no. 13, article no. e6234, 2021.



Copyright© by the authors. Licensee TAETI, Taiwan. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/).