EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Research on Intelligent Detection Method for Operation and Maintenance Violations of Power Distribution Equipment Based on YOLOv12

Yuexing Hu[1]

Skill Training Center of State Grid Shanxi Electric Power Company Limited, Taiyuan, 030025, China

Abstract

INTRODUCTION: With the emergence of new equipment and technologies, the difficulty of operation and maintenance (O&M) of power distribution equipment (PDE) has been continuously increasing. Traditional manual supervision and monitoring methods have been unable to meet the requirements of real-time performance and accuracy.
OBJECTIVES: In order to effectively reduce operational safety risks, we propose an intelligent O&M violation detection method.
METHODS: This paper optimizes the architecture of YOLOv12 and constructs three models: a security tool violation carrying recognition model, a general violation operation behavior recognition model, and a specific task violation operation behavior recognition model, this paper also uses the 3D electronic fence and real-time acquisition of each operator's 3D joint coordinates, and predicts the 3D joint coordinates of operation and maintenance personnel based on the Kalman filter.
RESULTS: The method achievies accurate detection of O&M violations. In addition, this paper successfully establishes a 3D electronic fence for the O&M environment of PDE, and also achieves the recognition and early warning of violations related to spatial locations.
CONCLUSION: The intelligent analysis and evaluation system for power distribution equipment operation and maintenance safety based on multimodal data fusion developed based on this method achieves intelligent recognition of violations in power distribution equipment operation and maintenance and significantly improves the level of intelligence in on-site safety control.

Keywords: Intelligent detection method; O&M violation; PDE; YOLOv12; Deep Learning; Object detection

Received on 05 November 2025, accepted on 10 April 2026, published on 22 April 2026

doi: 10.4108/eetsis.10801

1. Introduction

In the last few years, with the upgrading of the power system, new types of equipment, cutting-edge technologies and advanced operation modes have been continuously introduced, and the difficulty of power distribution equipment (PDE) operation and maintenance (O&M) has increased significantly[1]. When the operators operate distribution equipment, their decisions and behaviors are often influenced by multiple factors such as ability, experience, environment, etc., which may lead to an increase in safety risks[2-3]. Therefore, the introduction of efficient, automatic, and accurate supervision and control means has become particularly important[4]. The continuous progress of technologies such as high-definition video surveillance, Deep Learning, machine vision, robotics and drone inspection [5-6] provides conditions for automatic detection and identification of operator behavior [7-8].

The video monitoring content of the distribution O&M site mainly includes personnel status, equipment status, and environmental status [9]. Among them, the automatic detection of personnel status and equipment status mainly adopts the object detection method in machine vision. For

the special real-time and accuracy requirements of PDE O&M, YOLO algorithm is more applicable.

YOLO is a groundbreaking real-time object detection method that has sparked a technological revolution in the field of computer vision. This approach significantly enhances the speed of object detection, achieving real-time processing while maintaining high accuracy[10]. Since its inception, YOLOv1, has been updated to YOLOv12[11]. After YOLOv5 outputs detection boxes, DeepSORT achieves cross-frame tracking through Kalman filtering and appearance feature matching, achieving a MOTA of 68.5% on the MOT17 dataset[12]. YOLOv6-N compresses the model to 4.3×10^6 parameters through reparameterization technology, enabling real-time tracking of drone videos with an FPS of 30[13]. YOLOv9 introduces programmable gradient information to reduce feature degradation during long-term tracking, reducing the ID switching rate by 18%[14]. YOLO is not only used for object detection but can also be extended to multi-task learning. YOLOv8 integrates a Mask R-CNN branch to achieve instance segmentation[15]. YOLOv7 combines a keypoint detection head to support human pose estimation, applicable to security and medical monitoring[16]. YOLOv10 adopts spatial channel decoupling downsampling to optimize semantic segmentation[17].

YOLOv12, released by Ultralytics in 2025, is the latest iteration of the YOLO series[18]. It is the first to fully integrate an attention mechanism into a real-time object detection framework, breaking the dominance of traditional convolutional neural networks in the YOLO series and significantly improving detection accuracy while maintaining excellent inference speed. Based on the architecture foundation of YOLOv11, YOLOv12 achieves an excellent balance between accuracy and efficiency by introducing a regional attention module, a residual efficient layer aggregation network, and multiple architectural optimizations. It is suitable for various visual computer tasks such as object detection, instance segmentation, pose estimation, directional object detection, and image classification[19]. YOLOv12 further enhances performance through a series of architectural optimizations: removing positional encoding to simplify the model structure, adjusting the proportion of multi-layer perceptrons to balance the computational load of attention and fully connected layers, replacing linear layers with convolutional operations to improve computational efficiency, and integrating FlashAttention to optimize memory access, significantly reducing memory overhead during inference[20]. YOLOv12 significantly enhances the stability and applicability of the model in complex environments through refined training processes and diverse data augmentation techniques. Unlike the C2PSA module of YOLOv11, YOLOv12 introduces an adaptive regional attention mechanism that specifically optimizes the detection capability for small targets and high-density scenes. Through dynamic feature optimization and multi-level information integration, it successfully addresses the challenge of recognizing occluded objects[21].

In the power industry, Huang W et al. utilized AlphaPose and ResNet to achieve clothing detection for power workers[22]. However, this method relies heavily on the target detection algorithm. Any false or missed detection of the target person will directly affect the subsequent pose estimation results. Additionally, when multiple people cause occlusion, repeated convolution is required within the candidate boxes, resulting in low computational efficiency for such methods. Some research uses OpenPose to extract key skeletal points of people and detects smoking, falling, and climbing violations by calculating the relative positional relationships between specific points[23]. Another study also utilizes OpenPose to extract key points and completes behavior analysis for grounding detection tasks through support vector machines[24]. However, these methods only extract features from the image once, and even if there are more people in the image, they do not cause repeated convolution. Therefore, these models have high computational efficiency and are relatively small in size.

However, the aforementioned studies primarily focus on detection methods for a single type of violation operation, or merely investigate the methods of detection without considering their practical application. They lack practicality and fail to adapt risk detection standards according to different types of electrical operations. Furthermore, they do not consider conducting safety supervision for multiple different types of electrical operations simultaneously. Given that there are dozens of types of electrical operations, the safety production requirements for different types of electrical operations vary. For instance, welding operations require the use of protective glasses, while high-voltage operations necessitate the use of insulating gloves[25]. Consequently, the criteria for risk detection also differ.

This article presents a method based on the YOLOv12 detection algorithm, which optimizes the structure of YOLOv12 and constructs three different recognition models to realize the intelligent detection of O&M operation related and position independent violation behaviors. In addition, the method establishes a 3D model of O&M scene based on the YOLOv12-pose and 3D electronic fence, realizes the optimized reconstruction of 3D human body joint coordinates, and intelligently identifies the spatial position related violation. Based on this method, the intelligent detection and warning system for PDE O&M violations is constructed, which can be used to identify and predict whether the operator's behaviors violate the regulations according to the corresponding work tickets of different O&M tasks, thus realizing the intelligent on-site safety management and control function to flexibly cope with different tasks.

The contributions of this article are: (1) proposing an intelligent detection method for multiple types of violations in the O&M process of PDE; (2) optimizing the structure of YOLOv12, thereby enhancing the accuracy and real-time performance of recognition; (3) establishing a violation sample library and training and optimizing the model to improve the accuracy of the algorithm; (4)

Identify and predict the coordinates of human joint points to achieve early warning of violations.

In the rest of this article, Section 2 describes the improvement method for the object detection algorithm YOLOv12, as well as the intelligent recognition method for three types of operation-related violations based on YOLOv12; Section 3 describes the method of recognition and prediction of human joint points; Section 4 carries out the experiment and evaluates the experimental results; and Section 5 summarizes the essay.

2. Method for intelligently identifying operation related violations based on YOLOv12

In order to realize automatic and accurate identification of different types of distribution equipment O&M violations, this article has improved the object detection algorithm YOLOv12 and establishes three operation-related violation identification models based on the object detection algorithm YOLOv12. The models include safety tools violation carrying identification model, generic violation operation behavior identification model, task-specific violation operation behavior identification model.

2.1. Improvement methods for YOLOv12 algorithm

We propose an improved method for the YOLOv12 algorithm that combines a multi-layer perceptron. The improved YOLOv12 structure design consists of three parts: the backbone, neck, and head network. The backbone is responsible for feature extraction, the neck for multi-scale feature aggregation, and the head for generating the final detection results. The structural diagram is shown in Figure 1.

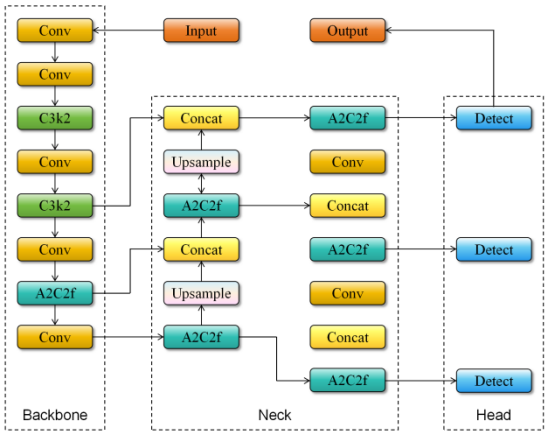

Figure 1. Improved YOLOv12 structure

The backbone part extracts basic textures and edge features of the image through the standard convolutional block of YOLOv12, introduces an improved C3k2 module, and improves the reuse rate of features through multi-branch residual connections while maintaining parameter lightness. In addition, 7×7 depthwise separable convolution is adopted to enhance the ability to capture subtle structures in complex scenes and improve spatial perception of small objects.

The neck network integrates an attention mechanism, double convolution, and a feature fusion module (A2C2f), combining regional attention with the Residual Effective Layer Aggregation Network (R-ELAN). Regional attention calculates local attention weights by dividing the feature map into multiple sub-regions, reducing global computational overhead while enhancing the model's focus on target regions. In scenarios where interference and multiple fragment types overlap, the A2C2f module significantly improves the sensitivity of target boundaries and the distinguishability between categories.

The head network utilizes a dynamic multi-layer perceptron (MLP) proportional adjustment mechanism and supports multi-scale prediction, allowing for simultaneous object classification and bounding box regression at different resolution levels. By employing strategies such as channel weighting and scale balancing, it enhances performance in small object detection tasks.

The convolutional module of YOLOv12 comprises two-dimensional convolution, two-dimensional batch normalization, and a SiLU activation function. The feature map is divided into local regions for computing self-attention; positional encoding is introduced through 7×7 depthwise convolution; a lightweight MLP implemented using 1×1 convolution performs channel expansion and compression to enhance nonlinear representation while maintaining compactness; residual connections merge the original input with processed features, facilitating stable gradient propagation and improving convergence, overcoming the limitations of traditional convolution in modeling long-range dependencies and extracting fine textures.

2.2. Safety tools violation carrying identification model

2.2.1 Identify requirement analysis and sample library construction

Table 1. Labels and collection quantity of image sample of safety tools

Labeling Number	Labeling Name	Number of Images
1	Insulated Boot	426
2	Insulated Clamp	433

3	Warning Sign-No Switching on	426
4	Warning Sign-Enter and Exit Here	434
5	Warning Sign-High Voltage	428
6	Safety Helmet	432
7	Insulated Pole	427
8	Grounding Wire	430
9	Electroscope	435
10	Insulated Gloves	425
11	Operating Lever	431
12	Safety Rope	424
13	Sliding T-Bar	436
14	Discharger	425
Total		5976

According to the PDE O&M specifications, it is a typical violation to fail to carry or incorrectly carry safety tools during operations. We identified 14 types of work tools and protective tools that need to be identified as shown in Table 1. The table displays the count of samples labeled in each category.

2.2.2 Optimized training parameter settings for the model

The hyperparameters listed in Table 2 were chosen based on the task's characteristics to enhance both the convergence speed and the model's performance.

Table 2. Training parameter settings

Parameter Type	Parameter Symbol	Parameter Size
Pre-training Model	weights	YOLOv12.pt
Training Rounds	epoch	100
Number of Batches	batch-size	4
Input Image Size	img-size	640
Learning Rate	lr0	0.01
Weight Decay	weight_decay	0.0005
Optimizer	adam	ture

2.3. Generic violation operation behavior identification model

2.3.1 Identify requirement analysis and sample library construction

Generic violation operation behavior refers to the violation operation not strongly related to the O&M task of distribution equipment. Through the analysis of the distribution equipment O&M safety regulations, O&M task operation process and relevant specifications, and typical violation cases, this article identifies seven common violations that need intelligent identification. They include smoking, not wearing safety helmet, no safety helmet chin strap, not wearing insulated gloves, not wearing insulated boots, no fence, and not hanging warning sign of "enter and exit here ".

Considering the complexity of generic violation operation behavior characteristics, this method establishes 10 kinds of label datasets. The number of image samples collected by all labels is more than 500.

2.3.2 Optimized training parameter settings for the model

Table 3. Model training parameter settings

Parameter Type	Parameter Symbol	Parameter Size
Pre-training Model	weights	YOLOv12.pt
Training Rounds	epoch	300
Number of Batches	batch-size	32
Input Image Size	img-size	640
Learning Rate	lr0	0.01
Weight Decay	weight_decay	0.0005
Optimizer	adam	ture

The increase in the number of targets to be measured in the generic violation behavior identification model increases the training difficulty. To allow the model to learn features more fully, the number of epochs was raised to 300. To enhance accuracy, accelerate model convergence, stabilize gradient estimation, and improve generalization performance, the batch size was increased to 32 and mixed precision training was employed. The model's training hyperparameter settings are detailed in Table 3.

2.4. Task-specific violation operation behavior identification model

Task-specific operation violation behavior refers to the operation violation behavior that is strongly related to the O&M tasks of PDE. Detection of task-specific violations behavior requires comprehensive identification of multiple states and sequences to be accomplished. Th method in this article identifies intelligent identification of 4 types of task-specific operation violation behaviors, including:

(i) Hanging grounding wire violation: when hanging the grounding wire, one end of the grounding wire has been hung on the high-voltage wire, but the other end has not been grounded.

(ii) Hanging warning signs violation: After switching off the machine, it is detected that the image of the warning sign hung by the operator does not coincide with the rectangular frame of the hanging position.
(iii) Electric test sequence error violation: During an electrical test operation, the 10kV HV normally energised point was not tested prior to conducting a phase-by-phase test on the transformer HV conductive bars.
(iv) Drop fuse power off and on sequence error violation: Based on the operation guideline of drop fuse power off and on, 8 different states of drop fuse are recognized by images to judge whether the operation is in violation of drop fuse power off and on.

In order to correctly detect the above violations involving multiple states and time sequences, a dataset of 20 labels is established. The number of image samples collected for all the labels is more than 500.

Taking Electric test sequence error violation as an example, we will illustrate the content of the dataset for violation behavior sequences.

In the electrical verification process, it is stipulated that a test must be conducted on the 10kV high-voltage constant current point first, as shown in Figure 2(a). Afterwards, align the high-voltage conductive rod of the transformer and perform phase by phase electrical testing, as shown in Figure 2(b). If the verification sequence is reversed, it will be identified as an abnormal verification error violation of the "incorrect verification sequence" type.

Figure 2. Testing the 10kV high-voltage constant point and aligning the high-voltage conductive rod of the transformer for phase by phase verification

To determine the Electric test sequence error violation, it is necessary to separately detect the actions of a: "testing the 10kV high-voltage constant point" and b: "aligning the high-voltage conductive rod of the transformer for phase by phase verification" from the state, and then detect the timing of both. If a occurs before b, it indicates that no violation has occurred; otherwise, it indicates that a violation has occurred Electric test sequence error violation。

Therefore, we establishes a dataset for Electric test sequence error behavior that includes both the actions of "testing the 10kV high voltage constant point" and "aligning the transformer high voltage conductive rod for phase by phase verification", and sorts the two to achieve strong temporal correlation.

3. Method of intelligent identification of spatial position related violations based on YOLOv12-Pose

The intelligent identification method for PDE O&M is not only to achieve accurate identification, but also to achieve early warning of illegal operations. Addressing the severe occlusion issue in multi-target human motion pose estimation within PDE O&M scenarios, we propose a multi-view human joint capture solution utilizing multi-camera stereo vision. In the first step, based on the YOLOv12-Pose 2D skeletal joint extraction model and the proposed preferred method of 3D reconstruction camera, the reliable reconstruction of 3D joints of multi-target operators complex scenes is realized. In the second step, the 3D coordinates of O&M joints are predicted based on kalman fliter. In the third step, based on the established 3D electronic fence and the real-time acquired and predicted 3D joint coordinates of each operator, the violation operation behaviors related to spatial position are automatically identified.

3.1. Optimal reconstruction of 3D human joint coordinates based on YOLOv12-Pose

3.1.1 YOLOv12-Pose based 2D joint point coordinate extraction for multiple persons

YOLOv12-Pose based 2D joint point coordinates extraction method from different cameras for human joint point coordinates and confidence level:

$$\{x(i)_n, y(i)_n, z(i)_n\}. \quad (1)$$

From (1), $\boldsymbol{i = 1 : 17}$, which represents the joint number; $\boldsymbol{n = 1 : N}$, which represents the number of the person in the O&M scene; $\boldsymbol{m = 1 : 4}$, which represents the camera number.

3.1.2 Matching of different person in different cameras

Before the 3D reconstruction of joints, matching individuals across different cameras is essential, and we used an approach based on the binocular camera polar constraint relationship to accomplish this task.

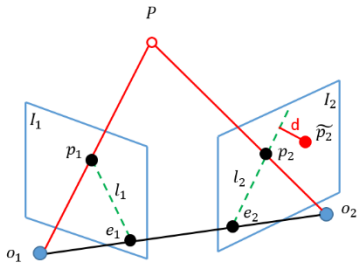

Figure 3. Schematic diagram of the principle of matching polar constraint relations

As illustrated in Figure 3, the center points of the cameras are denoted as O_1 and O_2, with point P in 3D space corresponding to pixels p_1 and p_2 in images I_1 and I_2 of the respective cameras. The lines O_1P and O_2P cross at the point P in 3D space. The points O_1, O_2, and P could be located with a flat known as polar flat. The connecting line of O_1 and O_2 is called the baseline. and the intersection points of the connecting lines of O_1 and O_2 with the image planes I_1 and I_2 are e_1 and e_2 respectively. e_1 and e_2 are called polar points. The lines l_1 and l_2, where the polar flat intersects with the image flats I_1 and I_2 , are referred to as polar lines.

If it is known that the 3D spatial joint point P corresponds to pixel p_1 in the first camera image I_1, it is not possible to determine the exact position of the spatial joint point P. It is also not known that the joint point P corresponds to the position on the second camera image I_2. However, since the point P is on the line O_1P, p_2 should be on the line e_2p_2. Thus, the correspondence between joint $\widetilde{p_2}$ in the second camera image I_2 and joint p_1 in the camera image I_1 could be quantified with measuring the distance d from $\widetilde{p_2}$ to the line e_2p_2.

Find the average of the distances calculated by matching all 17 points of the 1st personnel in image I_2 with all 17 points of the 1st personel in image I_1: $\overline{d_1} = \frac{1}{17}\sum_{i=1}^{17} d(i)_1$; and then, find the distance acquired by matching all 17 points of all the personnel in image I_2 (n = 1: N) with the 17 points of the 1st personnel in image I_1. The average of the distances obtained.

$$\overline{\boldsymbol{d}_n} = \frac{1}{17}\sum_{i=1}^{17} \boldsymbol{d}(i)_n. \quad (2)$$

The person corresponding to the minimum value of $\overline{d_n}$ in image I_2 matches the 1st person in image I_1. According to this principle, the matching relationship between people with different cameras can be obtained.

3.1.3 Camera preference in 3D reconstruction

The camera preference objective function based on polar constraint relations and 2D joint coordinate confidence values is:

$$\min_{j,k}\{a * \frac{1}{17}\sum_{i=1}^{17} \boldsymbol{d}(i)_n^{j,k} - b * (\frac{1}{17}\sum_{i=1}^{17} \boldsymbol{conf}(i)_n^{j} + \frac{1}{17}\sum_{i=1}^{17} \boldsymbol{conf}(i)_n^{k})\} \quad (3)$$

From (3), $\frac{1}{17}\sum_{i=1}^{17} d(i)_n^{j,k}$ denotes the average value of the distance to the pole line calculated from (2) for the n-th person, the 17 joint coordinates in the j-th camera and the kth camera. $\frac{1}{17}\sum_{i=1}^{17} conf(i)_n^{j}$ indicates the average confidence value of 17 joint coordinate of the n-th person in the k-th camera. When minimizing the value of (3) , the corresponding indices j and k, which represent the sequence numbers of the two selected cameras could be used for 3D reconstruction of the joint coordinates of the n-th person, are identified.

3.1.4 Multi-target 3D reconstruction of human joints

Using the camera calibration results, the camera matching results of (2), and the camera preference results of (3), the pixel coordinates p_1 and p_2 of each joint point in the two camera images are brought into the 3D reconstruction formula to complete the 3D reconstruction of the coordinates of each 2D joint point of multiple persons.

3.2. Kalman fliter-based prediction of 3D coordinates of the joints of operations personnel

In order to prevent O&M personnel from accidentally entering illegal areas, it is necessary to predict the movement trajectories of each joint of the body of O&M personnel in real time. Since human body movement is characterised by high dimensionality and strong sequentiality, this paper adopts the kalman fliter to forecast the 3D coordinates of the joint points of O&M personnel.

The state and observation equations of the kalman fliter are shown in equation (4) and equation (5):

$$\boldsymbol{x_k} = \boldsymbol{A}\boldsymbol{x_{k-1}} \quad (4)$$

$$\boldsymbol{y_k} = \boldsymbol{H}\boldsymbol{x_k} \quad (5)$$

where $x = [p_x, p_y, v_x, v_y, a_x, a_y]$ is the 2D coordinates , velocity and acceleration of the human joint in the 2D image coordinate system of the camera, $y = [p_x, p_y]$ is the 2D coordinates of the joint point in 2D image coordinate system, and w_k and v_k are the system noise and the measurement noise with gaussian distribution respectively.

$$\begin{cases} \boldsymbol{w_k} \sim \mathbf{N}(\mathbf{0}, \mathbf{Q}) \\ \boldsymbol{v_k} \sim \mathbf{N}(\mathbf{0}, \mathbf{R}) \end{cases} \quad (6)$$

The equation of state **A** and the observation equation ***H*** are

$$\mathbf{A} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{\Delta t} & \mathbf{0} & \mathbf{0.5\Delta t^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{\Delta t} & \mathbf{0} & \mathbf{0.5\Delta t^2} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{\Delta t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{\Delta t} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix} \quad (7)$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (8)$$

Δt is the sampling time interval from k–1 to k. Because of the high sampling rate of the motion capture camera, between adjacent frames, the distance of the joint point movement is small and the continuity is strong, so it is approximated that the joint point is in uniformly accelerated motion.

The kalman filter tracking algorithm is executed by predicting the state of the system at moment k using the state of the system at moment (k−1), while the parameters in the system are updated by using the observed values. The prediction formula is:

$$\breve{\boldsymbol{x}}_{\boldsymbol{k}} = \boldsymbol{A}_{\boldsymbol{k-1}}\hat{\boldsymbol{x}}_{\boldsymbol{k-1}} \quad (9)$$

$$\breve{\boldsymbol{P}}_{\boldsymbol{k}} = \boldsymbol{A}_{\boldsymbol{k-1}}\hat{\boldsymbol{P}}_{\boldsymbol{k-1}}\boldsymbol{A}_{\boldsymbol{k-1}}{}^{\boldsymbol{T}} + \boldsymbol{Q}_{\boldsymbol{k}} \quad (10)$$

The predicted state of the system is obtained through the above equation, and the kalman filter system is updated using the predicted state and the measurements. the optimal state estimate $\hat{x}_k$ at the moment k is derived from the measurements y_k:

$$\breve{\boldsymbol{x}}_{\boldsymbol{k}} = \boldsymbol{A}_{\boldsymbol{k-1}}\hat{\boldsymbol{x}}_{\boldsymbol{k-1}} \quad (11)$$

Kalman gain K_k is:

$$\boldsymbol{K}_{\boldsymbol{k}} = \breve{\boldsymbol{P}}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}{}^{\boldsymbol{T}}/(\boldsymbol{H}_{\boldsymbol{k}}\breve{\boldsymbol{P}}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}}{}^{\boldsymbol{T}} + \boldsymbol{R}_{\boldsymbol{k}}) \quad (12)$$

Find the covariance $\hat{\boldsymbol{P}}_{\boldsymbol{k}}$ corresponding to the optimal state estimate at moment $\boldsymbol{K}$:

$$\hat{\boldsymbol{P}}_{\boldsymbol{k}} = (\mathbf{1} - \boldsymbol{K}_{\boldsymbol{k}}\boldsymbol{H}_{\boldsymbol{k}})\breve{\boldsymbol{P}}_{\boldsymbol{k}} \quad (13)$$

According to the principle of 3D reconstruction of binocular stereo vision, the 2D coordinates of each joint point of each O&M personnel matched in the two cameras used for reconstruction are transformed into 3D coordinates $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. By tracking through the time sequences $\ldots, \boldsymbol{k-1}, \boldsymbol{k}, \boldsymbol{k+1}, \ldots$, the motion information of the joint points can be calculated, and the trend of their motion can be inferred. The displacement, velocity, and acceleration equations are shown below:

$$\begin{cases} \boldsymbol{d}_{\boldsymbol{x}} = \boldsymbol{X}_{\boldsymbol{k}} - \boldsymbol{X}_{\boldsymbol{k-1}} \\ \boldsymbol{d}_{\boldsymbol{y}} = \boldsymbol{Y}_{\boldsymbol{k}} - \boldsymbol{Y}_{\boldsymbol{k-1}} \\ \boldsymbol{d}_{\boldsymbol{z}} = \boldsymbol{Z}_{\boldsymbol{k}} - \boldsymbol{Z}_{\boldsymbol{k-1}} \end{cases} \quad (14)$$

$$\begin{cases} \boldsymbol{v}_{\boldsymbol{x}} = \boldsymbol{d}_{\boldsymbol{x}}/\Delta\mathbf{t} \\ \boldsymbol{v}_{\boldsymbol{y}} = \boldsymbol{d}_{\boldsymbol{y}}/\Delta\mathbf{t} \\ \boldsymbol{v}_{\boldsymbol{z}} = \boldsymbol{d}_{\boldsymbol{z}}/\Delta\mathbf{t} \end{cases} \quad (15)$$

$$\begin{cases} \boldsymbol{a}_{\boldsymbol{x}} = \Delta\boldsymbol{v}_{\boldsymbol{x}}/\Delta\mathbf{t} \\ \boldsymbol{a}_{\boldsymbol{y}} = \Delta\boldsymbol{v}_{\boldsymbol{y}}/\Delta\mathbf{t} \\ \boldsymbol{a}_{\boldsymbol{z}} = \Delta\boldsymbol{v}_{\boldsymbol{z}}/\Delta\mathbf{t} \end{cases} \quad (16)$$

Then the 3D space coordinate prediction equation of the joint is:

$$\begin{cases} \boldsymbol{X}_{\boldsymbol{k+1}} = \boldsymbol{X}_{\boldsymbol{k}} + \boldsymbol{v}_{\boldsymbol{x},\boldsymbol{k}} * \Delta\mathbf{t} + \mathbf{0.5}\boldsymbol{a}_{\boldsymbol{x},\boldsymbol{k}} * \Delta\mathbf{t}^{\mathbf{2}} \\ \boldsymbol{Y}_{\boldsymbol{k+1}} = \boldsymbol{Y}_{\boldsymbol{k}} + \boldsymbol{v}_{\boldsymbol{y},\boldsymbol{k}} * \Delta\mathbf{t} + \mathbf{0.5}\boldsymbol{a}_{\boldsymbol{y},\boldsymbol{k}} * \Delta\mathbf{t}^{\mathbf{2}} \\ \boldsymbol{Z}_{\boldsymbol{k+1}} = \boldsymbol{Z}_{\boldsymbol{k}} + \boldsymbol{v}_{\boldsymbol{z},\boldsymbol{k}} * \Delta\mathbf{t} + \mathbf{0.5}\boldsymbol{a}_{\boldsymbol{z},\boldsymbol{k}} * \Delta\mathbf{t}^{\mathbf{2}} \end{cases} \quad (17)$$

3.3. Identification method of spatial position related violations

Using the 3D joint coordinates of each operator acquired in real time in 3.1 and 3.2, this article proposes method to detect spatial position related violation and give early warning tips.

(i) Single person operation violation behavior identification: During the O&M of PDE, single person operation is strictly prohibited. Therefore, by determining whether the 3D joint coordinates of each operator are in the specified work area, and calculating the number of operators in the work area, we can detect whether single person operation violations have occurred.

(ii) Identification of violation behavior by mistakenly entering violation areas: During the O&M of PDE, the behavior of breaking into violation areas by mistake includes: distance from high-voltage equipment is less than the safety distance, mistaken out of the insulation pad, going to the wrong interval and so on. Through real-time judgment of each operator's 3D joint coordinates whether to comply with the violation conditions set by the 3D electronic fence, can be detected whether the violation occurred.

- Distance from high-voltage equipment is less than the safety distance: Firstly, set up a 3D electronic fence at a safe distance from the high-voltage equipment. When any skeletal point of the human body enters the 3D electronic fence partition area, a safety distance alarm is triggered.
- Mistaken out of the insulation pad: Firstly, inspect the insulation pad area and establish a 3D electronic fence around the insulation pad. When all the skeletal points of the human foot enter the 3D electronic fence partition area, the alarm function is activated. When any foot bone point leaves the 3D electronic fence partition area, a false departure from the insulation pad alarm is triggered. After completing the assignment, turn off the alarm function.
- Going to the wrong interval: Firstly, at the beginning of the operation and maintenance task, establish a 3D electronic fence for the correct interval of the work area. When the operation and maintenance personnel start or are working, if any bone point is outside the 3D electronic fence, an interval error alarm will be triggered. When the operation and maintenance task is completed, turn off the alarm function.

4. Application effects

The training and evaluation of the experimental model were conducted in the same experimental environment, and the detailed information of the hardware environment used is shown in Table 4. The same configuration environment can be adopted for on-site application.

Table 4. Experimental environment configuration

hardware environment	details
CPU	i7-7700 3.6G Hz
CPU core	4
GPU	NVIDIA RTX 4060
display memory	8G

hard drive	1T
image resolution	1280*960
fps of real-time image	25

4.1. Safety tools violation carrying identification

(i) Training results.

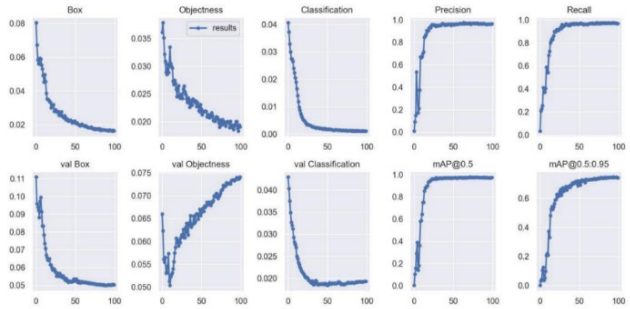

Figure 4. Training results graph

Figure 4 demonstrates the change process of the model's important metrics during the training process from 0 to epoch (100) iterations. After 100 iterations, it can be seen that the precision exceeds 90% and the recall exceeds 80%, indicating that the model has a good success rate in detection. model mAP@0.5 and mAP@0.5 0.95 both exceeded 60%, proving that the accuracy performance of the model is equally good. The change curves of the above metrics with the iterative process indicate that the safety tools violation carrying identification model developed by this method has small error, high accuracy and good generalization ability.

(ii) Model evaluation.

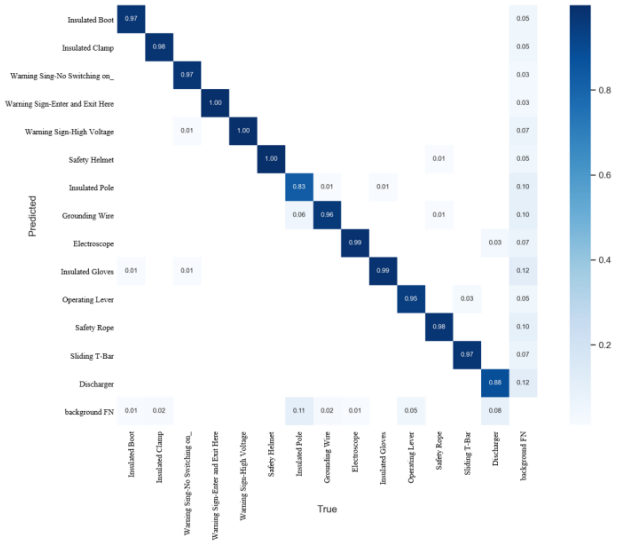

Figure 5. Confusion matrix of safety tools violation carrying identification model

After training the safety tools violation identification model, its property was assessed by using a test dataset. The evaluation results are displayed in a confusion matrix as shown in figure 5. There are 15 items in figure 5, the first 14 items are model detection types and the 15th item "background FN" is the background. The results show that the percentage of errors occurring in the test dataset of detection results is low, and the model can effectively identify whether the operator have correctly carried the safety tools required for the O&M tasks of PDE.

Table 5. Ablation experiment results

ID	C3k2	A2C2F	P(%)	R(%)	mAP@0.5
1	-	-	94.9	83.8	91.7%
2	√	-	95.1	84.9	92.9%
3	-	√	96.2	84.7	94.5%
4	√	√	98.0	85.8	94.8%

To explore the impact of various modules in the improved YOLOv12 model on object detection performance, ablation experiments were conducted in a unified environment. Table 5 presents the results of each ablation experiment, where "√" indicates the use of the module and "-" indicates its non-use.

We first use the experimental results of the original YOLOv12 model as the benchmark for subsequent experiments, so as to compare the performance improvement effects of each improved module.

After improving C3k2 in the backbone network, the ability to capture subtle structures in complex scenes has been enhanced, and the spatial perception of small objects has been improved. As a result, the mAP@0.5 of the model has increased by 1.2%.

The addition of the A2C2f module to the neck network has enhanced the sensitivity of target boundaries and the distinguishability between categories. As a result, the mAP@0.5 has increased by 2.8%.

After incorporating all the improvements, the mAP@0.5 increased by 3.1% compared to the baseline model.

In summary, the ablation experiment verifies the effectiveness of each improvement point and their combinations, proving the practicality of the improved method in object detection.

Table 6. Comparison results of different algorithms

Algorithm	mAP@0.5	FPS
Faster-RCNN	71.5%	8.9
SSD	61.3%	65.6

YOLOv7	88.9%	40.9
YOLOv10	92.7%	41.2
YOLOv12	94.8%	37.7

We select the popular target detection algorithms like Faster-RCNN, SSD, YOLOv7, YOLOv10 for comparative experiments on the test set, to prove the capability of YOLOv12 in the wearing of safety tools for typical distribution equipment operation and maintenance tasks,. mAP@0.5 is used as the accuracy standard of the model, and FPS is used as the detection speed standard of the model. Table 6 shows the comparison results. The YOLOv7, YOLOv10 and SSD algorithms have faster speeds compared to YOLOv12, because the YOLOv12 network structure is more complex and the computational complexity is larger. In summary, the YOLOv12 algorithm is suitable for the detection of the wearing of safety tools for typical distribution equipment operation and maintenance tasks with higher accuracy and detection speed.

4.2. Generic violation operation behavior identification

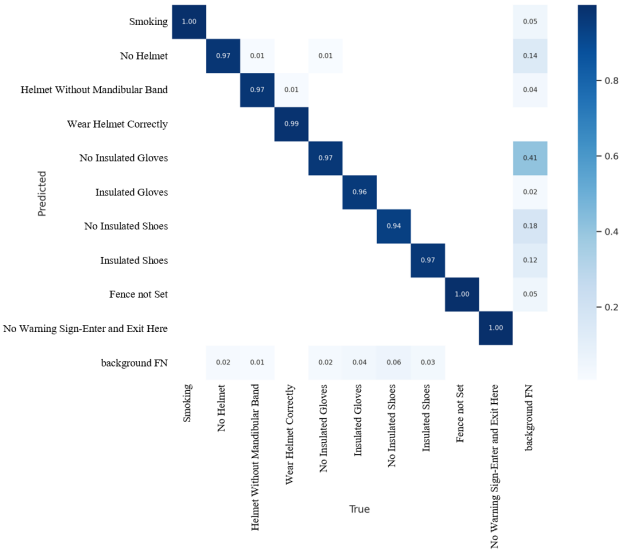

Figure 6. Confusion matrix of generic violation operation behavior identification model

Figure 6 displays the confusion matrix for the generic violation operation behavior identification model. The identification accuracy of the model for all the labels exceeds 94%, and the identification accuracy of the violations such as “smoking” and “No Warning Sign-Enter and Exit Here” reaches 100%.

4.3. Task-specific violation operation behavior identification

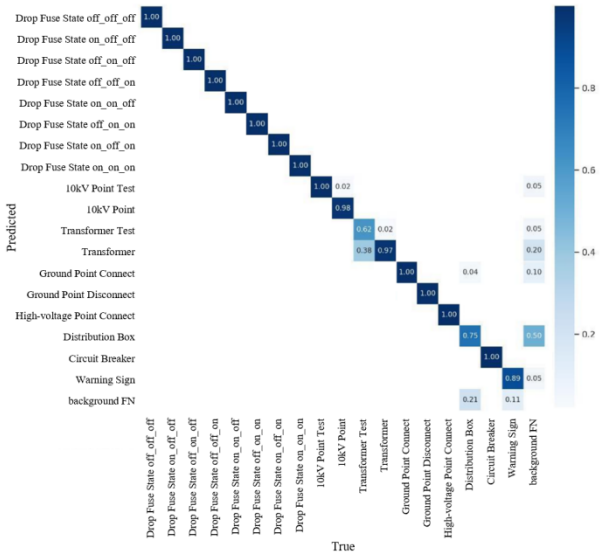

Figure 7. Confusion matrix of task-specific violation identification model

Figure 7 depicts the confusion matrix for the task-specific violation behavior identification model. The model has high identification accuracy for most of the labels. The label “Transformer test” has a slightly lower identification accuracy of 62%, due to the fact that there is little difference between this item and its control group “Transformer”. The label “Distribution Box” has a 21% chance of being recognized as “background FN”, due to the fact that “Distribution Box” is not well characterized and the area of interest is not distinguished from the “background FN”, which leads to a decrease in identification accuracy.

In order to test the effectiveness of detecting violations in occluded images in PDE O&M scenarios based on the method proposed in this paper, we set up scenarios with and without obstacle occlusion, and arranged four cameras to form a camera array in each scenario. Then, we used YOLOv12 to detect the imaging results of both single cameras and camera arrays, and the results are shown in Table 7.

Table 7. Comparison results of different algorithms

obstacle in scene	single camera			camera array		
	P(%)	R(%)	mAP@0.5	P(%)	R(%)	mAP@0.5
no obstacle	91.1	84.6	83.1%	91.8	83.4	83.4%

obstacle included	77.4	76.5	71.8%	86.1	79.9	83.9%

The results indicate that in the absence of obstructions, the use of a camera array for image acquisition has little impact on detection accuracy. However, when obstacles exist within the scene, employing a camera array can significantly mitigate the accuracy loss caused by regional obstruction.

5. Conclusions

In this article, according to the research and analysis of the O&M of PDE and the current status of violation detection technology, the architecture of YOLOv12 is optimized, and three operation-related violation recognition models are established. The aforementioned methods realize the identification of safety tools violation carrying behaviors, generic violation behaviors that is not specifically related to operation tasks, and violation behavior related to specific operation tasks. This article further uses YOLOv12-Pose 2D skeletal joint extraction algorithm to realize the reliable reconstruction of 3D joints of multi-target operators in complex scenes, and solves the serious problem of occlusion phenomenon existing in the estimation of multi-target human action postures. And by using the 3D electronic fence and the 3D joint coordinates of each operator acquired in real time, followed by the prediction of the joint 3D coordinates of the O&M personnel based on the kalman fliter, then the identification and early warning of violation behaviors related to spatial position is realized. In the future, we will conduct research on lightweight deployment methods for models, enhancing the adaptability of detection models to edge devices, aligning with the development trend of edge-oriented and intelligent grid equipment.

Acknowledgements.

The primary funding for this work is provided by the Science and Technology Project of State Grid Shanxi Electric Power Company Limited titled "Research and application of key technologies for intelligent analysis and evaluation of power distribution equipment operation and maintenance safety" (No. 520570220002).

References

[1] Wu T. Analysis of distribution operation and maintenance integration construction under the background of smart grid [J]. Electrical Equipment and Economy,2023, (06):237-239.

[2] Zhuang G. Exploration of distribution operation and maintenance integration construction under the background of smart grid [J]. Telecom Power Technologies,2019, 36(05):267-268.

[3] Zhao Z. Zhang W. Zhai Y. Concepts, Research Status and Outlook of Power Vision Technology [J]. Electric Power Science and Engineering, 2020, 36(01) : 1-8

[4] Yang Z. Research and application of lightweight target detection algorithms in relay protection inspection work [D]: Southwest Jiaotong University,2022.

[5] Zhang Z. Research on key technology of detecting and recognizing small targets in substation [D]: Southwest Jiaotong University,2022.

[6] Gu D. Research on Intelligent Identification of Substation Reconstruction and Expansion Violations Based on Artificial Intelligence Technology [D]: Nanjing University of Posts and Telecommunications,2022.

[7] Collins R T, Lipton A J, Kanade T, et al. A system for video surveillance and monitoring[J]. VSAM final report, 2000, 2000(1-68): 1.

[8] Bogaert M, Chelq N, Cornez P, et al. The PASSWORDS Project [intelligent video image analysis system] [C]. Proceedings of 3rd IEEE International Conference on Image Processing. IEEE, 1996, 3: 675-678.

[9] Y.H. Huang, B.L. Lyu, T.L. Gao, X.D. Wu, Y.G. Duan, CornMFN: a multimodal fusion network for corn phenology stage identification, Smart Agric. Technol. 12 (2025) 101202

[10] G. Kaur, J.S.S. Rajni, Development of deep and machine learning convolutional networks of variable spatial resolution for automatic detection of leaf blast disease of rice, Comput. Electron. Agric. 224 (2024) 109210

[11] B.S. Kusumo, A. Heryana, O. Mahendra, H.F. Pardede, Machine learning-based for automatic detection of corn-plant diseases using image processing, in: 2018 International conference on computer, control, informatics and its applications (IC3INA), IEEE, 2018, pp. 93–97

[12] L. Zhou, C. Zhang, M. Wu, D-LinkNet: linkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 182–186

[13] E.H. Wu, Y. Chen, R.J. Ma, X.D. Zhao, A review of weed image identification based on deep few-shot learning, Comput. Electron. Agric. 237 (2025) 110675

[14] H. Bahrami, K. Chokmani, S. Homayouni, V.I. Adamchuk, M. Saifuzzaman, M. Leduc, Alfalfa detection and stem count from proximal images using a combination of deep neural networks and machine learning, Comput. Electron. Agric. 232 (2025) 110115

[15] B.R. Pushpa, A. Ashok, S.H. AV, Plant disease detection and classification using deep learning model, in: third international conference on inventive research in computing applications (ICIRCA), IEEE, 2021, pp. 1285–1291

[16] X. Zhang, D. Zhu, R. Wen, SwinT-YOLO: detection of densely distributed maize tassels in remote sensing images, Comput. Electron. Agric. 210 (2023) 107905

[17] J. Wu, C. Wen, H. Chen, Z. Ma, T. Zhang, H. Su, C. Yang, DS-DETR: a model for tomato leaf disease segmentation and damage evaluation, Agronomy 12 (9) (2023)

[18] A. Zulfiqar, E. Izquierdo, K. Chandramouli, Harnessing deep learning for plant species classification: a comprehensive review, Comput. Electron. Agric. 237 (2025) 110663,

[19] R. Khanam, M. Hussain, Yolov11: an overview of the key architectural enhancements.2024, arXiv preprint arXiv:2410.17725.

[20] A. Thakuria, C. Erkinbaev, Real-Time canola damage detection: an End-to-End framework with semi-automatic crusher and lightweight shuffleNetV2_YOLOv5s, Smart Agric. Technol. 7 (2024) 100399

[21] X. Zhang, Y. Song, T. Song, D. Yang, Y. Ye, J. Zhou, L. Zhang, AKConv: Convolutional kernel with Arbitrary

Sampled Shapes and Arbitrary Number of Parameters, 2024.
[22] Huang W, Xu W, Zhang C, Dong C, Wan L. Power Construction Worker Dress Detection Model Combining Alphapose and ResNet [J]. Electric Power Information and Communication Technology, 2022, 20(03): 40-47
[23] Li B, Li Y, Sun Y, Gu S. Method for Surveillance Video Analysis Based on the Overlay of Object Detection and Human Pose Estimation Algorithms [J]. Electronic Technology and Software Engineering, 2020(07):143-147
[24] Yang X. Research on Behavior Recognition of Power Operation Personnel Based on Machine Learning [D]. University of Electronic Science and Technology of China, 2020
[25] Qiu H, Zhang W, Peng B, et al. Violation operation recognition algorithm based on YOLOv3 in specific power operation scenarios[J]. Journal of Electric Power Science and Technology, 2021, 36(3): 195-202