Research on Engineering Structures & Materials
www.jresm.org
Research Group

Research Article

An identification method for device status combining video monitoring in substation main control rooms

Zheng Zhou [*,a]

School of Electrical Engineering, Zhengzhou Railway Vocational and Technical University, Zhengzhou, China

Article Info

Article History:

Received 04 Aug 2026

Accepted 08 Oct 2026

Keywords:

Substation main control room;
Video monitoring;
Device status identification;
Improved YOLO target detection

Abstract

The status of devices in substation main control rooms is jointly characterized by fine-grained components such as indicator lights, pressure plates, rotary knobs, display screens and alarm windows. To overcome the limitations of traditional single-frame detection in small object recognition, state continuity and component logical relations, this study proposes a device status identification model integrating improved YOLO target detection, bidirectional temporal modeling and graph relation reasoning. Based on fixed monitoring videos, the model performs state component annotation, local feature enhancement, multi-scale feature fusion and device status correlation graph construction to achieve device status judgment from component detection. Experimental results show that the proposed model achieves a device status accuracy of 96.28%±0.29%, an anomaly recall of 95.36%±0.41%, and a false alarm rate of 2.74%±0.18%. In real scenarios, the recall of state components reaches 96.34%±0.35% at a monitoring distance of 4 m, and the anomaly recall reaches 94.21%±0.46% under 20% occlusion. The proposed method provides an effective intelligent approach for substation video inspection and anomaly detection.

© 2026 MIM Research Group. All rights reserved.

1. Background

The advancement of smart grids, digital substations, and unattended operation and maintenance modes gradually shifts the perception of substation equipment status from manual periodic inspection to intelligent monitoring driven by video, images, and multi-source data [1]. In substation operation scenarios, the main control room undertakes functions such as protection control, dispatch monitoring, alarm display, and operation control. The device status directly affects the judgment and handling of power grid anomalies by operators [2]. Different from outdoor equipment, the device status in the main control room is mainly characterized by fine-grained visual components such as indicator lights, pressure plates, rotary knobs, display screens, and alarm windows [3]. Under continuous video monitoring conditions, factors such as illumination changes, screen refresh, camera angle deviation, and personnel occlusion easily cause missed detection, false detection, or status jitter in single-frame recognition [4-5]. Therefore, how to accurately extract key state components from the main control room video stream and infer the overall status of the device from the component states has become an important issue in the intelligent operation and maintenance of substations.

Focusing on problems such as missed detection of small targets, complex background interference, and insufficient model deployment efficiency in visual detection, defect recognition, and status perception of power equipment, existing studies proposed various deep learning improved detection methods. Pang F et al. proposed a defect detection method combining positive-unlabeled learning and class prior estimation to address incomplete annotation and unbalanced sample

*Corresponding author: zhouzheng20518@163.com
[a]orcid.org/0009-0004-0269-3960
DOI: http://dx.doi.org/10.17515/resm2026-2107vk0804rs

classes in electrical insulator defect detection, aiming to alleviate the impact of insufficient defect samples on model training [6]. Ji C et al. proposed a visual detection algorithm based on YOLO to address target scale variations and background interference in identifying construction machinery intrusion risks near transmission lines, in order to improve risk target recognition capabilities [7]. Wu T et al. proposed a single-stage YOLO real-time infrared detection model to address obvious noise, loss of small target features, and insufficient global information extraction in substation infrared images, which was used to improve the infrared target detection effect of substation equipment [8]. Yang Y et al. proposed a defect detection method combining a synthetic weather algorithm and improved YOLOv7 to address the decline in insulator defect detection accuracy under extreme weather, so as to enhance detection adaptability in complex meteorological backgrounds [9]. Bin F et al. proposed a lightweight detection model using the YOLO architecture to solve the problem of balancing detection accuracy and resource usage in insulator defects. The work enhanced detection capability by applying Bidirectional Feature Pyramid Network (BiFPN) and multi-scale attention mechanism [10].

These studies promoted the development of power equipment visual detection models in complex backgrounds, small target recognition, and insufficient sample processing, but limitations still exist. Most existing studies focused on outdoor equipment, transmission line insulators, construction machinery intrusion risks, or substation infrared defect detection, paying insufficient attention to secondary device status components in main control rooms. The status of devices in the main control room is not determined by a single defect target. Actually, it is jointly characterized by the states of multiple components, including running lights, alarm lights, pressure plates, rotary knobs, and display screens. Pure target detection cannot describe the subordination and state logical relations among components. At the same time, status changes in video monitoring have temporal continuity. Relying only on single-frame detection makes it hard to utilize the status evolution information in continuous frames. In addition, Mao Z et al. introduced a Graph Convolutional Network (GCN) into intelligent fault diagnosis of equipment, aggregating node correlation information through knowledge graphs and graph structures to enhance fault relationship reasoning capability [11]. Xu Z et al. constructed a bearing fault diagnosis model integrating convolutional neural networks, a Bidirectional Gated Recurrent Unit (Bi-GRU), and an attention mechanism. This study used Bi-GRU to capture the past and future dependencies in sequence features [12]. These methods show that graph structure modeling and bidirectional temporal modeling can provide technical support for complex equipment status identification, but their collaborative application in video device status identification within substation main control rooms still needs further development. Although YOLOv9, YOLOv11, and Vision Transformer-based detectors have shown strong feature representation capabilities, YOLOv10 was selected as the baseline because its end-to-end detection design reduces redundant post-processing and provides a favorable balance between detection accuracy, computational complexity, and inference latency. Compared with transformer-dominated architectures, its convolution-oriented structure also requires lower intermediate feature storage, which is more suitable for continuous video monitoring and subsequent edge deployment in substation main control rooms. Therefore, YOLOv10 provides an appropriate basis for further enhancing small-component perception and multi-scale feature fusion.

Accordingly, this study suggested a model to identify the device status in substation main control rooms. This technique begins with developing a dataset of state components using the fixed monitoring video of the main control room. Indicator lights, pressure plates, rotary knobs, screens, and alarm windows serve as recognition entities. It identifies the correlation between the state component annotation and device status labels. Subsequently, an Efficient Multi-Scale Attention (EMA) module and BiFPN are introduced into YOLOv10 to form Improved YOLOv10 (IYOLOv10), enhancing small-scale state component detection and multi-scale feature fusion capabilities. Finally, the component positions, categories, visual states, and confidence scores output by IYOLOv10 are converted into graph node features. According to spatial adjacency, device subordination, and state logical relationships, a device status correlation graph is constructed. The model uses Bi-GRU and GCN to complete temporal modeling and device-level correlation reasoning in continuous video frames. The purpose of this study is to realize the intelligent transformation

between images from the main control room video surveillance to device status grades, and offer an approach to perceive the operation status of devices in the substations' main control rooms. The innovations and contributions of this study mainly include:

Establishing a state component annotation and device-level status mapping method for main control room video monitoring, making visual objects such as indicator lights, pressure plates, rotary knobs, display screens, and alarm windows directly serve device status identification. The IYOLOv10 state component detection method combines EMA and BiFPN, improving feature response capability of the local features of small targets and cross-scale feature expression ability.

Constructing a Bi-GRU and GCN device status correlation identification framework, which jointly models continuous frame temporal information, component spatial adjacency relationships, device subordination relationships, and state logical relationships to achieve the expansion from single state component detection to overall device status identification.

Formally, given a main-control-room video sequence $\chi = \{I_t\}_{t=1}^{T}$, IYOLOv10 first maps each frame I_t to a set of detected state components $V_t = \{v_i^t\}$, where each node contains the component category, visual state, confidence score, spatial position, and affiliated device ID. Based on spatial adjacency, device subordination, and operational logic, the corresponding device graph is constructed as $G_t = (V_t, E_t)$. For a temporal window $X_{t-L+1:t}$, Bi-GRU models the temporal evolution of node states and GCN aggregates the relational information over G_t. The overall task is therefore formulated as learning a mapping $F_\theta: X_{t-L+1:t} \to y_m$, where $y_m \in$ {Normal, Warning,Abnormal} denotes the condition of device m.

2. Substation Main Control Room Device Status Identification Based on Yolov10

2.1 Video Monitoring Data and State Component Annotation

During the video monitoring process of the main control room, the device status is not determined solely by the overall appearance of the cabinet. Instead, it is jointly characterized by multiple key state components. To ensure that the detection object is consistent with the operational meaning of the device, this study uses indicator lights, pressure plates, rotary knobs, display screens, and alarm windows as basic recognition units, and establishes the mapping relationship between component-level annotations and device-level status labels. The workflow of video data construction and state component annotation is shown in Figure 1.

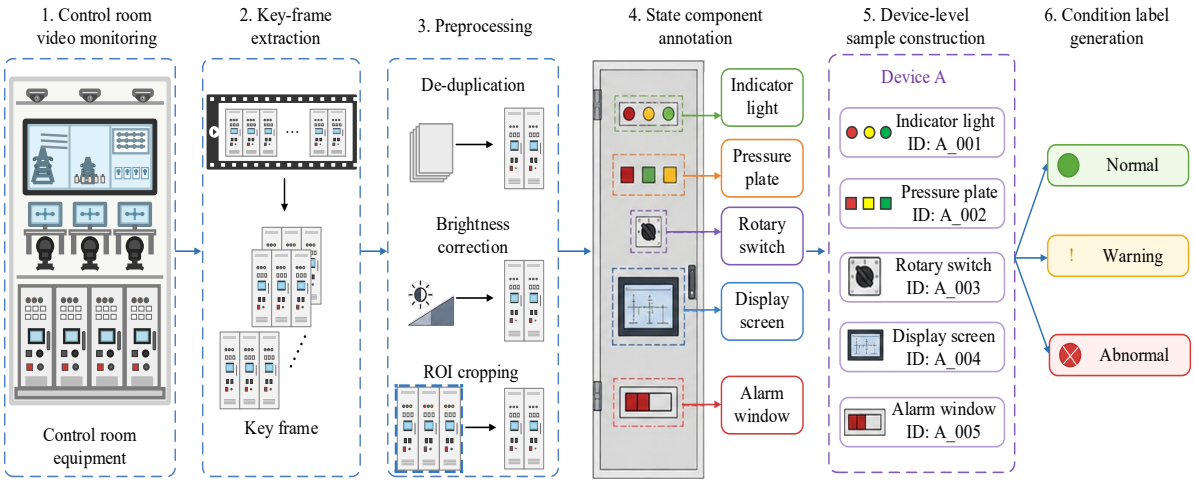


Fig. 1. Workflow of video data construction and state component annotation

In Figure 1, this study first performs keyframe extraction on the fixed monitoring video of the main control room, and executes de-duplication, brightness correction, and region cropping on the images. Following this, the state component annotations are carried out by cabinet number and the location of the device, taking note of the component type, spatial location, visual status, and the corresponding device number. Two experts independently annotated and reviewed the state components, while a third senior expert adjudicated inconsistent cases. Ambiguous samples

involving screen glare, dusty protective covers, dim indicator lights, or unclear platen positions were re-examined with reference to adjacent video frames and operation records. The Cohen's Kappa coefficient between the two primary annotators was 0.91, indicating high inter-rater agreement. Finally, multiple state components belonging to the same device are merged together to become device samples. Let the video sequence of the main control room be $V = \{I_1, I_2, \dots, I_T\}$, where I_t is the t -th frame image, and T is the total number of video frames. The keyframe extraction process is shown in Equation (1).

$$\chi = \{I_t | t \bmod \Delta t = 0 \ or \ D(I_t, I_{t-1}) > \tau\} \tag{1}$$

In Equation (1), χ represents the set of keyframes of the devices in the substation main control room, Δt is the frame extraction interval, $D(I_t, I_{t-1})$ is the difference between adjacent frames, and τ is the screen change threshold. This method can reduce redundant frames while preserving key screens where the device status changes. The frame extraction interval Δt and screen variation threshold τ were determined through sensitivity experiments on the validation videos. Specifically, Δt was tested at 5, 10, 15, and 20 frames, while τ was tested at 0.06, 0.09, 0.12, and 0.15 after normalizing the inter-frame variation to [0,1]. The combination of Δt =10 frames and τ =0.12 provided the best balance between redundant-frame removal and state-change retention. A smaller Δt retained excessive temporally adjacent frames, whereas an excessively large interval increased the probability of skipping short-duration state transitions. Similarly, a low τ preserved excessive visually similar frames, while a high τ could discard frames containing weak changes in indicator lights or display information. The corresponding sensitivity results are further analyzed in Section 3.1. For the i -th state component, its annotation vector is defined as shown in Equation (2).

$$a_i = (x_i, y_i, w_i, h_i, r_i, e_i, m_i) \tag{2}$$

In Equation (2), a_i is the annotation vector of the i -th state component, (x_i, y_i) is the coordinates of the center of the detection box, r_i is the component category, e_i is the visual state, and m_i is the affiliated device number. The component categories include indicator lights, pressure plates, rotary knobs, display screens, and alarm windows. The status categories of devices in the substation main control room include on/off, switched in/out, close/open positions, normal display, alarm prompt, and black screen [13]. For display-screen annotation, transient dark frames caused by camera–display refresh mismatch were distinguished from genuine black-screen states using temporal persistence. A dark display appearing only intermittently within adjacent frames was treated as flicker interference, whereas a black-screen label was assigned only when the low-luminance state persisted across at least three consecutive observations and was consistent with adjacent-frame information or operation records. Samples severely affected by specular reflection were additionally checked using neighboring frames to avoid introducing ambiguous black-screen labels. Based on component annotation, this study further generates device-level status labels. Let the set of state components contained in the m -th device be B_m. The device status label is expressed as shown in Equation (3).

$$y_m = f(\{r_i, e_i, m_i | b_i \in B_m\}), y_m \in \{Normal, Warning, Abnormal\} \tag{3}$$

In Equation (3), $f(\cdot)$ is the state mapping function, and y_m represents the device status label. If the states of the running lights, pressure plates, rotary knobs, and display screen are all consistent with the operational requirements, the label is marked as Normal. If conditions such as alarm lights being on or screen prompts of anomalies occur, the label is marked as Warning. If conditions such as incorrect pressure plate positions, inconsistent rotary knob states, or black display screens occur, the label is marked as Abnormal.

2.2 Improved YOLOv10 for State Component Detection

After completing the annotation of video state components, it is necessary to construct a detection model that can adapt to the complex visual environment of the main control room. The original YOLOv10 has high real-time performance. However, there are still problems of insufficient local feature response for small targets and inadequate cross-scale fusion in the detection of state

components in the main control room [14-15]. Therefore, this study introduces the Efficient Multi-Scale Attention (EMA) module and Bidirectional Feature Pyramid Network (BiFPN) feature fusion structure into YOLOv10 to form an improved YOLOv10 model (IYOLOv10) for detecting device state components in the main control room. First, to enhance the local response of small-scale components such as indicator lights, pressure plates, and rotary knobs, this study embeds the EMA attention module into the backbone feature layer. This module strengthens local texture, color boundaries, and spatial direction information through feature grouping and multi-branch attention calculation [16]. Its structure is shown in Figure 2.

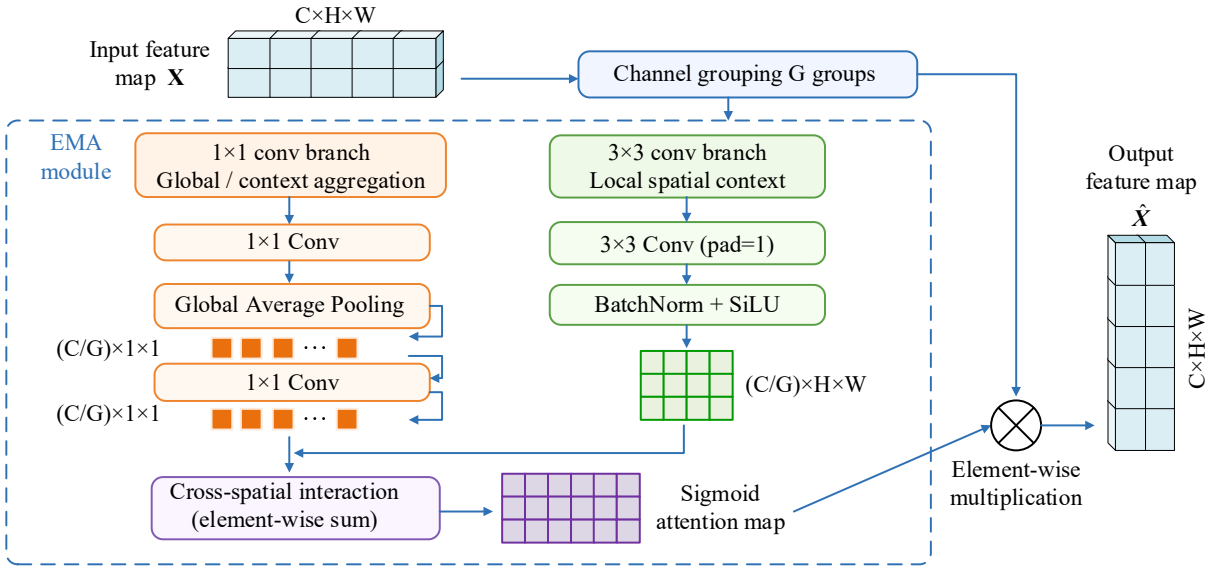


Fig. 2. Structure of the EMA attention module

As shown in Figure 2, the input feature map is first divided into multiple sub-feature groups according to the channel dimension. Each sub-group passes through a 1×1 branch and a 3×3 branch respectively to extract channel context and spatial context information. Subsequently, the input features are recalibrated by attention weights, making the model focus more on the on/off regions of indicator lights, boundary regions of pressure plates, and pointing regions of rotary knobs. Here, let the intermediate feature map output by the backbone network of YOLOv10 be $X \in \mathbb{R}^{C\times H\times W}$. Its grouping process is shown in Equation (4).

$$X = [X_1, X_2, \dots, X_G], X_g \in \mathbb{R}^{\frac{C}{G}\times H\times W} \tag{4}$$

In Equation (4), G is the number of feature groups, and X_g is the g -th sub-feature group. Through grouping processing, redundant computation can be reduced, and the capability of local detail modeling can be improved. Then, the output feature enhanced by EMA is shown in Equation (5).

$$\hat{X}_g = X_g \otimes \sigma\left(P_{1\times1}(X_g) + P_{3\times3}(X_g)\right) \tag{5}$$

In Equation (5), $\hat{X}_g$ represents the output feature of the g -th sub-feature group enhanced by EMA, $P_{1\times1}$ is the channel context feature extracted by the 1×1 branch, and $P_{3\times3}$ is the local spatial context feature extracted by the 3×3 branch [17]. The C2f-EMA module retains the residual feature propagation path of the original C2f structure and introduces attention recalibration within the feature branch. Therefore, the backbone can preserve the original gradient transmission path while assigning larger responses to informative local channels and spatial regions. This design reduces the risk that attention recalibration excessively suppresses useful low-level features during back-propagation. After local state feature enhancement, it is still necessary to solve the problem that the scales of different state components vary greatly. Display screens and alarm windows usually belong to larger targets, whereas indicator lights, pressure plates, and rotary knobs belong to small targets. A single-scale feature is difficult to balance semantic information and detailed information

simultaneously. Therefore, this study introduces the BiFPN structure into the Neck part to complete the bidirectional weighted fusion of multi-scale features. Its structure is shown in Figure 3.

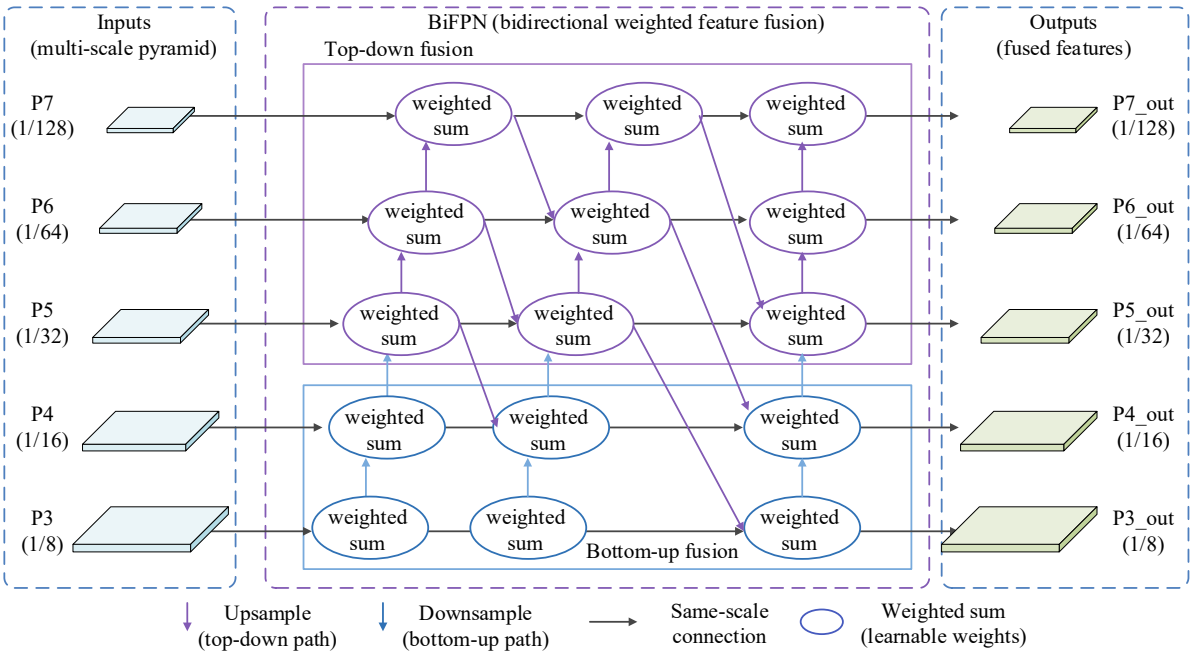


Fig. 3. Structure of the BiFPN feature fusion

As shown in Figure 3, BiFPN fuses features of different scales through top-down and bottom-up bidirectional paths [18]. High-level features are used to provide device categories and semantic information, and low-level features are used to preserve details of edges, colors, and textures. Learnable weights are introduced during the fusion of different scale features, enabling the model to adaptively adjust feature contributions based on the size of the state components. Before bidirectional fusion, the P3, P4, and P5 feature maps are projected to a unified channel dimension C_f using 1×1 convolutions. Spatial resolution is then aligned through up-sampling or down-sampling, after which the normalized weighted fusion in Equation (6) is performed. This channel projection prevents dimensional inconsistency among multi-scale features while introducing only limited additional computation. Let the multi-scale features participating in the fusion be F_j. The weighted fusion process of BiFPN is shown in Equation (6).

$$F^{out} = \frac{\sum_{j=1}^{n} \beta_j F_j}{\varepsilon + \sum_{j=1}^{n} \beta_j} \tag{6}$$

In Equation (6), F^{out} is the fused output feature, F_j is the j -th input scale feature, n is the number of scales participating in the fusion, β_j is the learnable fusion weight corresponding to the j -th scale feature, and ε is an extremely small constant to prevent the denominator from being zero. Finally, based on EMA and BiFPN, this study further constructs the IYOLOv10 state component detection network. Its structure is shown in Figure 4.

As shown in Figure 4, after the video frames of the main control room are input into the model, they first pass through the backbone network to extract basic features, and the responses of key state components are enhanced through EMA. Then, the cross-scale fusion is completed by BiFPN. The detection head outputs the positions, categories, states, and confidence scores of components. The output of this module is not the final device status. It is used as node input for subsequent correlation identification. To make the model meet both the positioning accuracy of state components and the requirements of visual state recognition, this study constructs a joint detection objective consisting of bounding box regression, effective target discrimination loss, component category classification, state classification, and distribution focal loss. Its expression is shown in Equation (7).

$$L_{det} = \frac{1}{N_+}\sum_{l=1}^{L}\sum_{i=1}^{N_l}\begin{bmatrix} I_i^{obj}\left(\lambda_{box}\left(1 - CIoU\left(\hat{b}_i^l, b_i^l\right)\right) + \lambda_{cls}CE\left(\hat{c}_i^l, c_i^l\right)\right) \\ +\lambda_{sta}CE\left(\hat{s}_i^l, s_i^l\right) + \lambda_{dfl}DFL\left(d_i^l, \hat{d}_i^l\right) + \lambda_{obj}BCE\left(\hat{p}_i^l, p_i^l\right)\end{bmatrix} \quad (7)$$

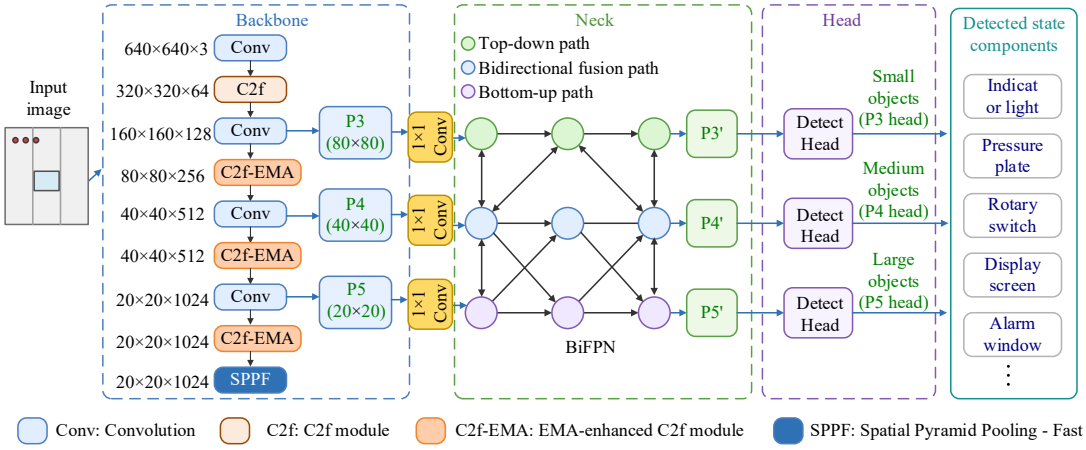


Fig. 4. Structure of the IYOLOv10 state component detection network

In Equation (7), L_{det} represents the joint detection loss of IYOLOv10, L represents the number of detection scales, N_l represents the number of candidate targets under the l -th scale, N_+ represents the number of positive samples, I_i^{obj} represents the target existence indicator function, $DFL(\cdot)$ represents the distribution focal loss, $CE(\cdot)$ represents the cross-entropy loss used for multi-classification constraints of component categories and visual state categories, and $BCE(\cdot)$ represents the binary cross-entropy loss, which determines whether an effective state component of the main control room exists in the candidate region. Considering the lower proportion of Abnormal samples in the field data, class imbalance was further considered during model training. A class-sensitive weighting strategy was introduced for device-condition supervision, assigning a larger penalty to errors involving Abnormal samples than to Normal samples. This treatment was used to reduce the tendency of the model toward the majority Normal class without changing the overall loss structure. The five loss weights were not adjusted using automatic weighting methods such as GradNorm or uncertainty weighting. Instead, λ_{box}, λ_{cls}, λ_{sta}, λ_{dfl}, and λ_{obj} were selected according to validation-set performance and were finally set to 7.5, 0.5, 1.0, 1.5, and 1.0, respectively. Particular attention was given to λ_{sta} because visual-state prediction directly affects subsequent device-level reasoning. Component-category prediction and visual-state prediction share the fused backbone features but use separate output branches in the detection head. Their losses are calculated independently before weighted aggregation, which reduces direct competition between the two prediction objectives. During training, no unstable convergence caused by obvious gradient conflict was observed.

2.3 Device State Association and Bi-GRU-GCN Identification Model

IYOLOv10 can complete the detection of state components, but the device status is usually determined by multiple components together. For example, obvious subordination and state logical relations exist among running lights, alarm lights, pressure plates, rotary knobs, and display screens. If only individual components are judged independently, the impact of component combinations on the overall device status can be easily ignored [19]. Therefore, this study utilizes Bi-GRU and GCN to further construct a device status correlation algorithm (Bi-GRU-GCN) to complete device status identification. The structure of Bi-GRU-GCN is shown in Figure 5.

As illustrated in Figure 5, the model takes the state component information generated by IYOLOv10 as input, and creates node sequences based on the device numbers. Bi-GRU is adopted to capture the temporal features of the changes in component state through continuous video frames. The temporal input of Bi-GRU was organized using a sliding window of T consecutive frames. The influence of sequence length was evaluated using T=5, 10, and 30. Considering the trade-off

between temporal context, recognition accuracy, and online inference latency, T=10 was finally adopted for subsequent experiments. In order to make sure that the temporal modeling maintains consistency, the current study conducts cross-frame matching for the same state component through continuous video frames according to the affiliated device number, category of the component, and its detection box position. Specifically, cross-frame association was implemented using IoU-constrained Hungarian matching.

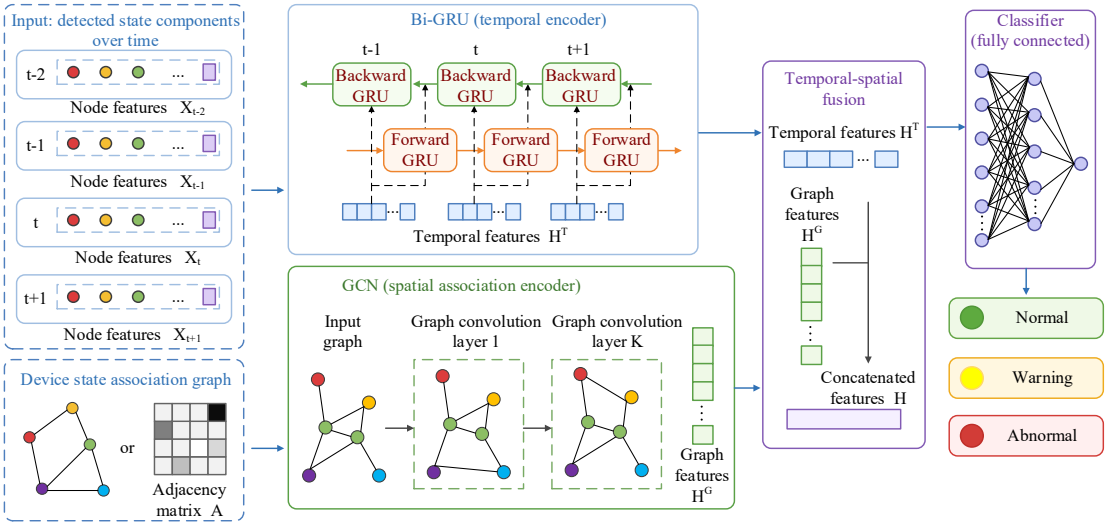


Fig. 5. Structure of the Bi-GRU-GCN algorithm

Candidate detections were first restricted to the same device ID and component category, and the matching cost was defined as 1-IoU between bounding boxes in adjacent frames. A match was accepted when IoU was not lower than 0.5. Low-confidence unmatched detections were removed as potential false positives, whereas a temporarily unmatched node was retained for up to two consecutive frames using its previous state feature. If no valid match was recovered thereafter, the node was removed from the current graph. This strategy maintains short-term graph continuity while limiting the propagation of false detections. GCN is used to aggregate the spatial subordination and state logical relations among different components within the same device. Finally, the fully connected layer outputs three status categories of the device: Normal, Warning, and Abnormal. Let the node feature of the i -th state component at the t -th moment be z_i^t. Its definition is shown in Equation (8).

$$z_i^t = \left[P_{r,i}^t, P_{e,i}^t, \alpha_i^t, \bar{x}_i^t, \bar{y}_i^t, \bar{w}_i^t, \bar{h}_i^t, m_i\right] \tag{8}$$

In Equation (8), $P_{r,i}^t$ is the category probability vector of component r_i, $P_{e,i}^t$ is the probability vector of visual state e_i, and α_i^t is the target confidence. $\bar{x}_i^t$, $\bar{y}_i^t$, $\bar{w}_i^t$, and $\bar{h}_i^t$ are the normalized center coordinates, width, and height of the detection box, respectively. $\bar{h}_i^t$ is the affiliated device number. Next, to describe the correlation relationships among state components, the device status adjacency matrix A is constructed. Its definition is shown in Equation (9).

$$A_{\mu\nu} = \begin{cases} 1, m_\mu = m_\nu \; and \; \left(d_{\mu\nu} < \tau_s \; or \; \eta_{\mu\nu} = 1\right) \\ 0, otherwise \end{cases} \tag{9}$$

In Equation (9), $A_{\mu\nu}$ represents whether a connection exists between node μ and node v, $m_\mu = m_\nu$ means the two state components belong to the same device, τ_s is the spatial adjacency threshold, and $\eta_{\mu\nu}$ is the state logical relation marker. When two components belong to the same device, and their spatial positions are adjacent or an operational logic relation exists, a graph connection is established. A binary static adjacency matrix was adopted because the spatial arrangement, device affiliation, and operational logic among secondary components in a fixed main control panel remain relatively stable during normal operation. Compared with GAT or a fully learnable adjacency matrix, this prior-constrained graph reduces additional parameter learning and computational

overhead, while preserving the interpretability of physical and logical connections. Dynamic state variations are mainly represented by the time-varying node features and further modeled by Bi-GRU. Nevertheless, the fixed adjacency matrix may be less flexible when operational logic changes under switching or reconfiguration conditions, and adaptive graph learning will be considered in future work. For example, correlation edges can be established between running lights and alarm lights, alarm lights and alarm windows, pressure plates and rotary knobs, and display screens and alarm windows. After the graph structure is established, GCN aggregates and updates the node state features. Its propagation process is shown in Equation (10) [20].

$$H^{k+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{k}W^{k}\right) \tag{10}$$

In Equation (10), H^{k} and H^{k+1} are the node feature matrices of the k -th layer and $k+1$-th layer respectively, $\tilde{A}$ is the adjacency matrix with self-connections added, $\tilde{D}$ is the degree matrix corresponding to $\tilde{A}$, and W^{k} is the graph convolution weight matrix of the k -th layer. Through this propagation process, each state component node can fuse information from adjacent components, components belonging to the same device, and logically related components, thereby expanding from a single component state to device-level correlation features. Based on the above content, this study combines IYOLOv10 and Bi-GRU-GCN to construct a device status identification model for substation main control rooms (Bi-GRU-GCN-IYOLOv10, BGG-IYOLOv10), as shown in Figure 6.

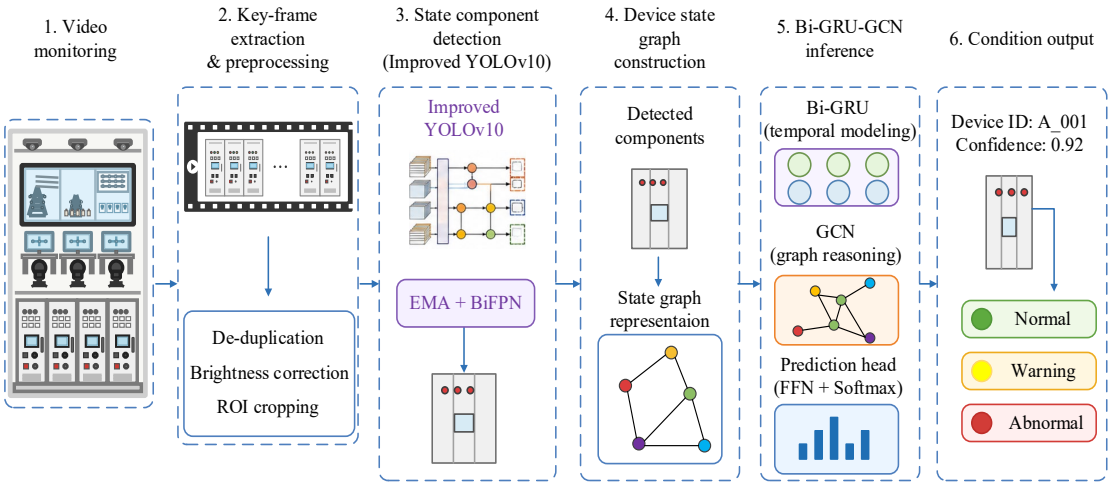


Fig. 6. Structure of the overall BGG-IYOLOv10 model

Note: FFN denotes Feed-Forward Network.

As shown in Figure 6, the final model takes the video stream of the main control room as input. After keyframe extraction, image preprocessing, and region cropping, the video is input into IYOLOv10 for detection. The detection results are further converted into graph node features, and a device status correlation graph is constructed based on spatial adjacency, device subordination, and state logical relationships. Subsequently, Bi-GRU extracts state change features in the continuous video, and GCN aggregates correlation information among components, finally outputting the device status level and the positions of anomalous components.

3. Performance Verification and Application Analysis of BGG-IYOLOv10 Model

3.1 Performance of BGG-IYOLOv10 Model

To verify the effectiveness of the proposed Bi-GRU-GCN-IYOLOv10 model (Bi-GRU-GCN-IYOLOv10, BGG-IYOLOv10), this study constructed a video state component dataset of a main control room. The data came from the fixed monitoring video of the substation main control room, with a video resolution of 1920×1080 and a frame rate of 25 fps. Upon undergoing processes such as keyframe extraction, duplication removal, brightening adjustment, and region cropping, 36800 keyframe images were extracted. The images comprised 128400 state components and 18240 device-level status groups. Examples of the state components include indicator lights, pressure plates, rotary

knobs, display screens, and alarm windows while the device status labels comprise Normal, Warning, and Abnormal. The 36,800 keyframes were obtained from existing fixed-camera video streams rather than being independently captured frame by frame. State-component annotations were generated according to cabinet IDs, device locations, and visual states, while device-level labels used in the field evaluation were further checked against manual review and corresponding operation records. The author was responsible for the research design, data processing and annotation protocol, model development, experimental analysis, and manuscript preparation. To avoid temporal data leakage caused by randomly splitting highly correlated adjacent frames, the dataset was partitioned according to independent video time intervals rather than at the individual-frame level. Continuous video segments from the same acquisition period were assigned entirely to only one subset. The 36,800 keyframes were divided into training, validation, and test sets at a ratio of 70:15:15, corresponding to 25,760, 5,520, and 5,520 keyframes, respectively. Adjacent frames originating from the same continuous video segment were not distributed across different subsets. In addition, the 6,240 samples collected from three substations during the 30-day field test in Section 3.2 were used only for independent real-scenario evaluation and did not participate in model training or parameter optimization.

The experiments were implemented based on Python 3.10 and PyTorch 2.1. The running platform consisted of an Intel Core i9-13900K CPU, 64 GB RAM, and an NVIDIA RTX 4090 GPU. The input size of the model was set to 640×640, the batch size was 32, and the initial learning rate was 0.001. The optimizer adopted AdamW, the training epochs were 200, and the weight decay coefficient was 0.0005. To reduce the impact of random initialization, each group of experiments was repeated 10 times, and results were expressed as mean±standard error. Differences between groups were analyzed using one-way ANOVA with Tukey's post-hoc test, and the significance level was set at $p<0.05$.

Before the ablation experiments, the sensitivity of the keyframe extraction parameters was evaluated to determine suitable values of Δt and τ. The redundant-frame removal rate, state-transition retention rate, and device condition accuracy were used to quantify the balance between data reduction and state-information preservation. When evaluating one parameter, the other parameter was fixed at its finally selected value. The results are shown in Table 1.

Table 1. Sensitivity analysis of keyframe extraction parameters

Δt (frames)	τ	Redundant-frame removal rate (%)	State-transition retention rate (%)	Device condition accuracy (%)
5	0.12	41.8±0.7	99.1±0.3	95.92±0.34
10	0.12	58.6±0.6	98.4±0.3	96.28±0.29
15	0.12	67.9±0.8	96.2±0.4	95.61±0.35
20	0.12	74.6±0.9	93.8±0.5	94.72±0.42
10	0.06	42.1±0.7	99.3±0.2	95.81±0.36
10	0.09	51.7±0.6	98.9±0.3	96.07±0.31
10	0.15	64.8±0.8	96.9±0.4	95.44±0.38

As shown in Table 1, increasing Δt continuously improved the redundant-frame removal rate, but excessive frame intervals caused a decline in state-transition retention. When Δt increased from 10 to 20 frames, the state-transition retention rate decreased from 98.4%±0.3% to 93.8%±0.5%, and the device condition accuracy decreased to 94.72%±0.42%. For the screen variation threshold, a low τ retained more redundant frames, whereas an excessively high threshold increased the risk of missing weak visual state changes. AtΔt =10 frames and τ =0.12, the redundant-frame removal rate reached 58.6%±0.6%, while the state-transition retention rate remained at 98.4%±0.3% and the device condition accuracy reached 96.28%±0.29%. Therefore, Δt =10 and τ =0.12 were adopted in the subsequent experiments.

The sequence length of Bi-GRU was further evaluated to determine an appropriate temporal window. Device condition accuracy and inference latency were compared at T=5, 10, and 30 frames, as shown in Table 2.

Table 2. Sensitivity analysis of Bi-GRU sequence length

Sequence length T (frames)	Condition accuracy (%)	Abnormal recall (%)	Inference latency (ms/frame)
5	94.73±0.41	93.18±0.48	17.6±0.5
10	96.28±0.29	95.36±0.41	18.7±0.6
30	96.35±0.31	95.48±0.39	24.9±0.7

As shown in Table 2, increasing T from 5 to 10 frames improved the device condition accuracy from 94.73%±0.41% to 96.28%±0.29% and the abnormal recall from 93.18%±0.48% to 95.36%±0.41%. When T was further increased to 30 frames, only a limited accuracy gain was obtained, whereas the inference latency increased to 24.9±0.7 ms/frame. Therefore, T=10 was selected to balance temporal information utilization and real-time performance. To verify the contribution of each module to model performance, this study designed ablation experiments. The results are shown in Table 3.

Table 3. Ablation and baseline comparison results

Model configuration	mAP@0.5 (%)	State F1 (%)	Condition accuracy (%)	Abnormal recall (%)	False alarm rate (%)	Inference latency (ms/frame)
YOLOv9	92.47±0.49	89.86±0.58	89.02±0.54	86.44±0.67	5.68±0.31	14.6±0.5
YOLOv10	91.26±0.52	88.74±0.61	87.85±0.58	84.62±0.73	6.31±0.34	12.8±0.4
YOLOv11	93.72±0.43	91.35±0.51	90.42±0.47	88.05±0.60	4.91±0.27	12.1±0.4
Swin Transformer-based detector	93.21±0.46	90.67±0.55	89.88±0.50	87.64±0.63	5.06±0.29	20.4±0.7
YOLOv10+EMA	93.08±0.44	90.53±0.55	89.47±0.51	86.78±0.66	5.42±0.29	13.6±0.4
YOLOv10+BiFPN	93.41±0.47	90.88±0.52	89.96±0.49	87.14±0.62	5.25±0.27	14.1±0.5
IYOLOv10	95.18±0.36	93.27±0.43	91.82±0.45	89.36±0.58	4.38±0.24	15.2±0.5
IYOLOv10+Bi-GRU	95.21±0.34	94.16±0.41	93.48±0.39	92.07±0.52	3.62±0.21	17.1±0.6
IYOLOv10+GCN	95.24±0.35	94.02±0.39	94.05±0.36	92.43±0.49	3.41±0.20	16.4±0.5
BGG-IYOLOv10	96.42±0.31	95.87±0.34	96.28±0.29	95.36±0.41	2.74±0.18	18.7±0.6

As shown in Table 3, the additional standalone baselines provided a direct comparison between the proposed detector and representative detection architectures. YOLOv9, YOLOv10, YOLOv11, and the Swin Transformer-based detector achieved mAP@0.5 values of 92.47%±0.49%, 91.26%±0.52%, 93.72%±0.43%, and 93.21%±0.46%, respectively, whereas IYOLOv10 reached 95.18%±0.36%. This result indicates that the integration of EMA and BiFPN provides additional gains in fine-grained state component detection. After Bi-GRU and GCN were further introduced, BGG-IYOLOv10 achieved the highest condition accuracy of 96.28%±0.29% and abnormal recall of 95.36%±0.41%, while reducing the false alarm rate to 2.74%±0.18%. The inference latency was 18.7±0.6 ms/frame at an input resolution of 640×640. Although the complete model introduced additional computational overhead relative to the original YOLOv10, the increase remained limited compared with the improvement in device-level condition identification.

Table 4. Edge deployment performance of BGG-IYOLOv10

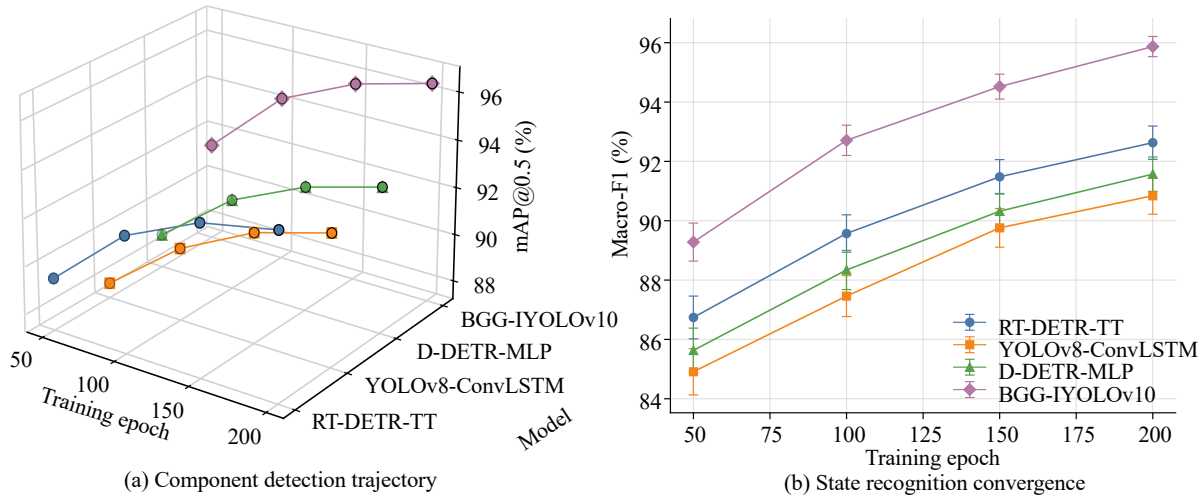
Hardware platform	Input size (pixel)	Latency (ms/frame)	FPS	Peak VRAM (GB)	Power consumption (W)
RTX 4090	896×896	29.6±0.9	33.8±1.0	3.12±0.08	186.4±5.7
Jetson AGX Orin	896×896	58.9±1.6	17.0±0.5	3.46±0.09	42.8±1.9
Jetson Orin Nano	896×896	94.1±2.8	10.6±0.3	3.31±0.10	14.6±0.8

To further evaluate the practical deployment capability of BGG-IYOLOv10, additional inference tests were conducted on Jetson AGX Orin and Jetson Orin Nano under the same input resolution of

896×896 pixels. In addition to inference latency, FPS, peak GPU memory usage, and average power consumption during continuous inference were recorded. The results are presented in Table 4.

As shown in Table 4, BGG-IYOLOv10 achieved an inference latency of 29.6±0.9 ms/frame and 33.8±1.0 FPS on the RTX 4090. When deployed on Jetson AGX Orin, the latency increased to 58.9±1.6 ms/frame, corresponding to 17.0±0.5 FPS, while the average power consumption decreased to 42.8±1.9 W. On the lower-power Jetson Orin Nano, the model maintained 10.6±0.3 FPS with an average power consumption of 14.6±0.8 W. Although edge hardware resulted in increased inference latency, the model retained continuous video processing capability with substantially reduced power consumption, supporting its potential deployment in substation main control room monitoring systems.

To further verify the performance of the BGG-IYOLOv10 model, this study compared it with a Real-Time Detection Transformer with Temporal Transformer (RT-DETR-TT) model, a YOLOv8 with Convolutional Long Short-Term Memory (YOLOv8-ConvLSTM) model, and a Deformable Detection Transformer with Multilayer Perceptron (D-DETR-MLP) model. All four models used the same training set, validation set, and test set, and had unified input sizes, training epochs, and optimization parameters. First, the state component detection mAP@0.5 and state recognition Macro-F1 of the four models were compared, as shown in Figure 7.



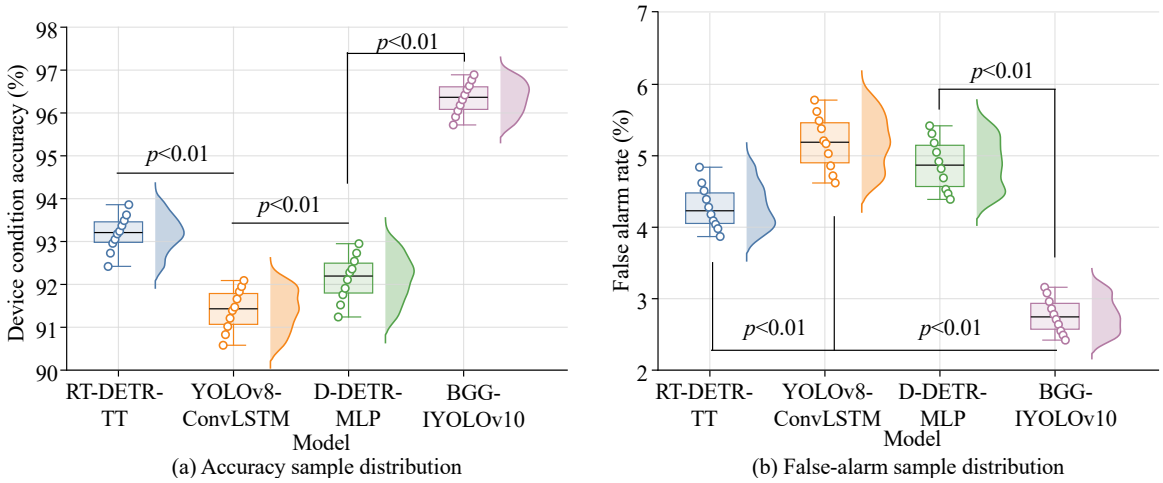
(a) Component detection trajectory (b) State recognition convergence

Fig. 7. Comparison of state component detection and state recognition

As shown in Figure 7(a), when the number of iterations was 200, the mAP@0.5 of BGG-IYOLOv10 was 96.42%±0.31%. This value was significantly higher than 94.18%±0.46% of RT-DETR-TT, 92.76%±0.51% of YOLOv8-ConvLSTM, and 93.35%±0.43% of D-DETR-MLP ($p<0.05$). This result indicated that EMA and BiFPN could enhance the local response of small-scale state components and cross-scale feature fusion capability respectively, giving the model a more stable detection effect. As shown in Figure 7(b), under the same conditions, the state recognition Macro-F1 of BGG-IYOLOv10 was 95.87%±0.34%, which was higher than that of all comparison models ($p<0.05$). This performance showed the model can recognize different visual states very well. Subsequently, this study compared the status accuracy and false alarm rate of the four models in the device status identification task, as shown in Figure 8.

As shown in Figure 8(a), in 10 repeated experiments, the device status accuracy of BGG-IYOLOv10 was 96.28%±0.29%, which was higher than that of other comparison models ($p<0.05$). This result indicated that relying only on detection results or temporal features was difficult to fully express the status logic inside the main control room devices. Conversely, Bi-GRU-GCN could jointly utilize continuous frame status changes and component correlation structures to improve device-level identification accuracy. As shown in Figure 8(b), the false alarm rate of BGG-IYOLOv10 was 2.74%±0.18%, which was lower than 4.26%±0.25% of RT-DETR-TT, 5.13%±0.31% of YOLOv8-ConvLSTM, and 4.82%±0.29% of D-DETR-MLP ($p<0.05$). This outcome showed that the proposed model could suppress instantaneous misjudgments caused by screen refresh, short-term occlusion,

and local reflection, thereby improving the stability of device status output under video monitoring conditions.



(a) Accuracy sample distribution (b) False-alarm sample distribution

Fig. 8. Comparison of device status accuracy and false alarm rate

3.2 Effect Verification under Video Monitoring Scenarios in Substation Main Control Rooms

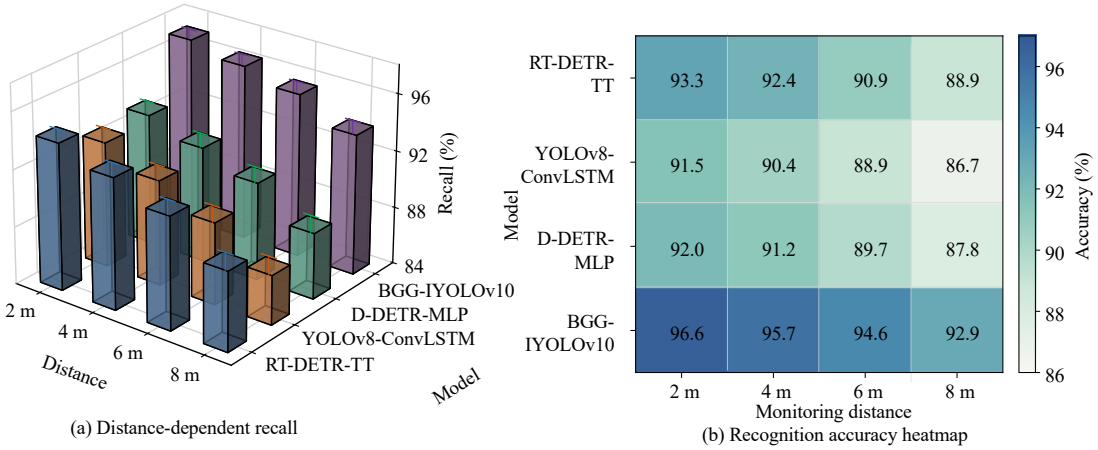
To further verify the application effect of BGG-IYOLOv10 in real substation main control room video monitoring scenarios, this study selected online monitoring videos of 3 substation main control rooms for 30 consecutive days for testing. The testing scenarios covered areas such as protection and control panels, automated monitoring panels, communication device panels, DC panels, and alarm display panels. There were 6240 testing samples in real scenarios, including 3910 Normal samples, 1520 Warning samples, and 810 Abnormal samples. All models adopted the optimal weights from the training phase, and inference testing was performed under the same hardware platform. Because the field dataset was imbalanced, overall accuracy alone was insufficient to characterize the recognition performance of safety-critical Abnormal samples. Therefore, a confusion matrix was additionally constructed for the three device-condition classes, and Precision, Recall, and F1-score were separately calculated for the Abnormal class. The confusion matrix was generated from the 6,240 independent field samples described above. The results are presented in Table 5.

Table 5. Confusion matrix of BGG-IYOLOv10 on the field dataset

Actual class	Predicted Normal	Predicted Warning	Predicted Abnormal	Total
Normal	3760	118	32	3910
Warning	23	1476	21	1520
Abnormal	8	30	772	810
Total	3791	1624	825	6240

As shown in Table 5, among the 6240 independent field samples, 3760 Normal, 1476 Warning, and 772 Abnormal samples were correctly classified, resulting in an overall condition accuracy of 96.28% (6008/6240). For the safety-critical Abnormal class, the Precision was 93.58% (772/825), indicating that 772 of the 825 samples predicted as Abnormal were correctly classified. The Recall was 95.31% (772/810), indicating that 772 of the 810 actual Abnormal samples were successfully detected. Based on the corresponding Precision and Recall, the F1-score reached 94.43% (1544/1635). These results indicate that the high overall accuracy was not dominated solely by the majority Normal samples and that the model retained stable recognition capability for the minority Abnormal class. For safety-critical evaluation, the study further examined the direct misclassification of actual Abnormal samples as Normal. Among the 810 Abnormal samples, only 8 were incorrectly classified as Normal, corresponding to a Critical Anomaly Miss Rate (CAMR) of 0.99%. In addition, 30 Abnormal samples were classified as Warning rather than Normal. Although

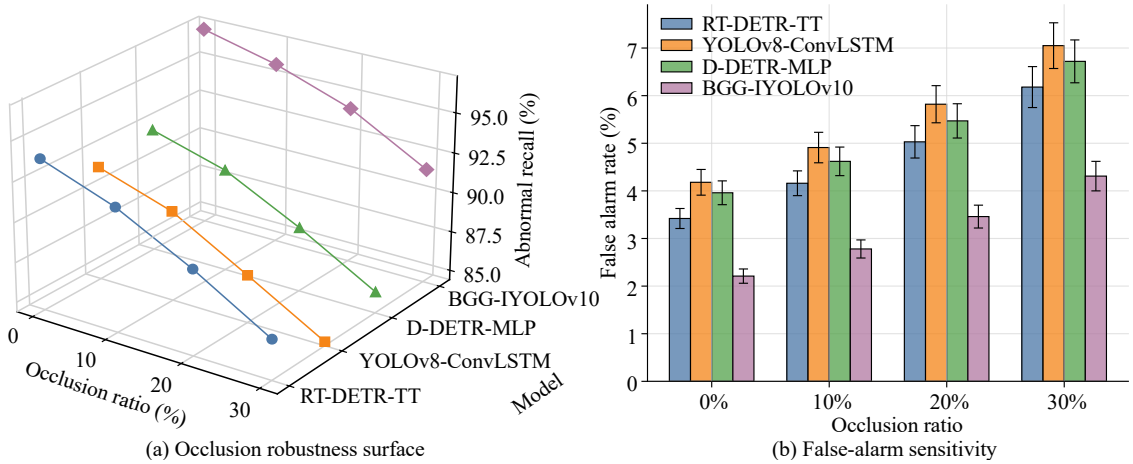
these cases were not correctly identified as Abnormal, they could still trigger operator attention and therefore posed a lower safety risk than direct Abnormal-to-Normal misclassification. The low CAMR indicates that the temporal continuity modeled by Bi-GRU and the component logical relations aggregated by GCN can effectively reduce safety-critical false-negative decisions caused by transient occlusion, screen refresh, or local visual ambiguity. Then, the state component recall and state recognition accuracy of the four models in real main control room scenarios were compared, as shown in Figure 9.



(a) Distance-dependent recall (b) Recognition accuracy heatmap

Fig. 9. Comparison of state component recall and state recognition accuracy

As shown in Figure 9(a), when the monitoring distance was 4 m, the state component recall of BGG-IYOLOv10 in real scenarios was 96.34%±0.35%. This value was significantly higher than 92.21%±0.47% of RT-DETR-TT, 90.63%±0.56% of YOLOv8-ConvLSTM, and 91.48%±0.52% of D-DETR-MLP (p<0.05). This performance was because BGG-IYOLOv10 enhanced local response through EMA and fused shallow details with deep semantics using BiFPN, which improved the detection capability of small-scale state components. As shown in Figure 9(b), when the monitoring distance was 8 m, the state recognition accuracy of BGG-IYOLOv10 was 92.94%±0.52%, which was higher than that of all comparison models (p<0.05). This data clearly proved the proposed model has strong robustness in real applications. Subsequently, this study analyzed the anomaly recall and false alarm rate of the models in real main control room scenarios, as shown in Figure 10.



(a) Occlusion robustness surface (b) False-alarm sensitivity

Fig. 10. Comparison of anomaly recall and false alarm rate

As shown in Figure 10(a), when the occlusion ratio was 20%, the anomaly recalls of BGG-IYOLOv10 reached 94.21%±0.46%, which was significantly higher than 90.77%±0.59% of RT-DETR-TT, 88.63%±0.68% of YOLOv8-ConvLSTM, and 89.51%±0.64% of D-DETR-MLP (p<0.05). The main reason is that Bi-GRU uses temporal information to keep state tracking continuous. As shown in Figure 10(b), when the occlusion ratios were 20% and 30%, the false alarm rates of BGG-IYOLOv10 were 3.46%±0.24% and 4.31%±0.31%, respectively. Both rates were lower than those of the comparison models (p<0.05). This result indicated that the modeling of continuous frame status

changes by Bi-GRU and the aggregation of component logical relationships by GCN could reduce instantaneous misjudgments under occlusion interference. Finally, to evaluate the deployment adaptability of the models in main control room online video monitoring, this study compared the state jitter rate and single-frame inference latency of the four models, as shown in Figure 11.

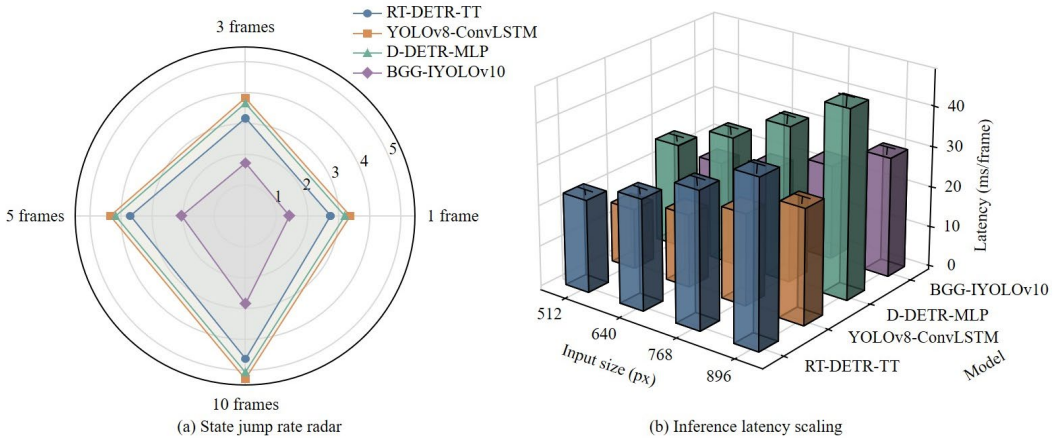


Fig. 11. Comparison of state jitter rate and single-frame inference latency

As shown in Figure 11(a), when the frame extraction interval was 10 frames, the state jitter rate of BGG-IYOLOv10 was 2.84%±0.24%, which was significantly lower than that of the comparison models (p<0.05). The decrease in state jitter rate indicated that the proposed model could suppress short-term status fluctuations caused by illumination disturbances, screen refreshes, and personnel occlusions in continuous video frames. As shown in Figure 11(b), when the input size was 896×896 pixels, the single-frame inference latency of BGG-IYOLOv10 was 29.6±0.9 ms, which was obviously lower than those of RT-DETR-TT and D-DETR-MLP. To evaluate robustness against small camera-angle drift caused by vibration or maintenance intervention, coordinate perturbations were further applied to the detected component bounding boxes before graph construction. The normalized center coordinates were simultaneously shifted in the horizontal and vertical directions by 0%, 2%, 5%, and 10% of the image size, while the remaining model parameters were unchanged. Device condition accuracy and abnormal recall were used to evaluate the influence of coordinate displacement. The results are shown in Table 6.

Table 6. Sensitivity of BGG-IYOLOv10 to bounding-box coordinate shifts

Coordinate shift	Condition accuracy (%)	Abnormal recall (%)
0%	96.28±0.29	95.36±0.41
2%	95.94±0.32	95.02±0.43
5%	95.21±0.36	94.28±0.47
10%	93.88±0.44	92.61±0.55

As shown in Table 6, small coordinate shifts had a limited influence on device condition identification. When the bounding-box displacement increased from 0% to 5%, the condition accuracy decreased from 96.28%±0.29% to 95.21%±0.36%, while the abnormal recall remained at 94.28%±0.47%. A more evident performance decline occurred at a 10% shift, with the condition accuracy and abnormal recall decreasing to 93.88%±0.44% and 92.61%±0.55%, respectively. This result indicates that the IoU-constrained cross-frame association and graph-based relation modeling can tolerate moderate camera displacement, whereas larger viewpoint changes may alter component spatial relationships and require additional image registration or viewpoint adaptation.

Combining the results of detection accuracy, anomaly recall, and false alarm rate, BGG-IYOLOv10 improved the accuracy and stability of substation main control room device status identification while maintaining real-time inference capability, which could meet the online recognition requirements in main control room video monitoring scenarios.

4. Conclusion and Future Work

To address the problems of small sizes of state components, similar backgrounds, complex status correlations, and insufficient single-frame recognition stability in video monitoring of substation main control rooms, this study proposed a device status identification model based on BGG-IYOLOv10. This model used indicator lights, pressure plates, rotary knobs, display screens, and alarm windows as basic recognition units. The EMA and BiFPN were introduced into YOLOv10 to enhance local state features and cross-scale fusion capabilities, and Bi-GRU-GCN completed status modeling of continuous frames and device-level correlation reasoning. The results showed that the mAP@0.5 of BGG-IYOLOv10 was 96.42%±0.31%, the state recognition F1 score was 95.87%±0.34%, the device status accuracy was 96.28%±0.29%, and the false alarm rate was 2.74%±0.18%. In real main control room scenarios, the state component recall was 96.34%±0.35% when the monitoring distance was 4 m. When the input size was 896×896 pixels, the single-frame inference latency was 29.6±0.9 ms.

Overall, the proposed model achieved the intelligent conversion from main control room video images to device status levels, which provided support for device status perception and anomaly auxiliary judgment. However, the study data mostly came from fixed views, so it did not test well for heavy occlusions. The adaptation verification under extreme occlusion, strong reflection, and equipment differences across different stations was insufficient. Future work will focus on three aspects. First, TensorRT-based FP16 and INT8 quantization will be investigated to further reduce inference latency and memory consumption on embedded edge devices. Second, the model will be extended from fixed-view cameras to PTZ cameras by incorporating viewpoint adaptation and dynamic region alignment. Third, additional data under strong reflection, screen flicker, illumination variation, and light-refraction interference will be introduced to improve the robustness of state component recognition under complex field conditions. These improvements are expected to further enhance the long-term deployment capability of the model in practical substations.

References

[1] Wang S, Tan W, Yang T, Zeng L, Hou W, Zhou Q. High-Voltage Transmission Line Foreign Object and Power Component Defect Detection Based on Improved YOLOv5. Journal of Electrical Engineering & Technology. 2024;19(1):851-866. https://doi.org/10.1007/s42835-023-01625-6

[2] Liu Y, Guo Y, Fan Y, Zhou J, Li Z, Xiao S, Wu G. Optical Imaging Technology Application in Transmission Line Insulator Monitoring: A Review. IEEE Transactions on Dielectrics and Electrical Insulation. 2024;31(6):3120-3132. https://doi.org/10.1109/TDEI.2024.3351093

[3] Cui B, Han C, Yang M, Ding L, Shuang F. DINS: A Diverse Insulator Dataset for Object Detection and Instance Segmentation. IEEE Transactions on Industrial Informatics. 2024;20(10):12252-12261. https://doi.org/10.1109/TII.2024.3417290

[4] Akella R, Gunturi SK, Sarkar D. Enhancing Power Line Insulator Health Monitoring with a Hybrid Generative Adversarial Network and YOLO3 Solution. Tsinghua Science and Technology. 2024;29(6):1796-1809. https://doi.org/10.26599/TST.2023.9010137

[5] He M, Qin L, Deng X, Liu K. MFI-YOLO: Multi-Fault Insulator Detection Based on an Improved YOLOv8. IEEE Transactions on Power Delivery. 2024;39(1):168-179. https://doi.org/10.1109/TPWRD.2023.3328178

[6] Pang F, Lei C, Zeng J. Electrical Insulator Defect Detection with Incomplete Annotations and Imbalanced Samples. IET Generation, Transmission & Distribution. 2024;18(4):694-715. https://doi.org/10.1049/gtd2.13107

[7] Ji C, Zhang F, Huang X, Song Z, Hou W, Wang B, Chen G. STAE-YOLO: Intelligent Detection Algorithm for Risk Management of Construction Machinery Intrusion on Transmission Lines Based on Visual Perception. IET Generation, Transmission & Distribution. 2024;18(3):542-567. https://doi.org/10.1049/gtd2.13093

[8] Wu T, Zhou Z, Liu J, Zhang D, Fu Q, Ou Y, Jiao R. ISE-YOLO: A Real-Time Infrared Detection Model for Substation Equipment. IEEE Transactions on Power Delivery. 2024;39(4):2378-2387. https://doi.org/10.1109/TPWRD.2024.3404621

[9] Yang Y, Yang S, Li C, Wang Y, Pi X, Lu Y, Wu R. Insulator Defect Detection under Extreme Weather Based on Synthetic Weather Algorithm and Improved YOLOv7. High Voltage. 2025;10(1):69-77. https://doi.org/10.1049/hve2.12513

[10] Bin F, Hou F, Chen D, Qiu K, Lu X, Sun Q. BCM-YOLO: An Improved YOLOv8-Based Lightweight Porcelain Insulator Defect Detection Model. High Voltage. 2025;10(4):876-889. https://doi.org/10.1049/hve2.70080
[11] Mao Z, Wang H, Jiang B, Xu J, Guo H. Graph Convolutional Neural Network for Intelligent Fault Diagnosis of Machines via Knowledge Graph. IEEE Transactions on Industrial Informatics. 2024;20(5):7862-7870. https://doi.org/10.1109/TII.2024.3367010
[12] Xu Z, Li YF, Huang HZ, Deng Z, Huang Z. A Novel Method Based on CNN-BiGRU and AM Model for Bearing Fault Diagnosis. Journal of Mechanical Science and Technology. 2024;38(7):3361-3369. https://doi.org/10.1007/s12206-024-0610-2
[13] Fang G, Zhao J, Chen H, Wu S, Ding T, Shi T. Verification Method of Relay Protection Platen State Based on the Improved YOLOX. Journal of Electrical Engineering (China). 2024;19(1):299-307. https://doi.org/10.11985/2024.01.032
[14] Li T, Lei L, Wang Z, Shi P, Wu Z. An Efficient Improved YOLOv10 Algorithm for Detecting Electric Bikes in Elevators. Electronic Research Archive. 2025;33(6):3673-3698. https://doi.org/10.3934/era.2025163
[15] Ji R, Xu Y, Wang X, Zhuang L, Zhang X, Tang X, Shi J. LBSD-YOLO: A Lightweight YOLOv10-Based Network with Multi-Attention Enhancement for Bridge Surface Defect Detection. ICCK Transactions on Sensing, Communication, and Control. 2026;3(1):39-53. https://doi.org/10.62762/TSCC.2025.718989
[16] Huo S, Yang C, Zhao X, Yu D, Zheng Y. An Improved Multi-Scale YOLOv8 for Apple Leaf Dense Lesion Detection. IET Image Processing. 2024;18(14):4913-4927. https://doi.org/10.1049/ipr2.13223
[17] Jooshin HK, Park J, Lee S, Kim H. Inception-YOLO: Computational Cost and Accuracy Efficient Object Detection Network. IET Image Processing. 2024;18(9):2914-2930. https://doi.org/10.1049/ipr2.13077
[18] Hua G, Li Y, Zhang Y, Wang J, Li X. An Enhanced YOLOv8-Based Bolt Detection Algorithm for Transmission Line. IET Generation, Transmission & Distribution. 2024;18(24):4065-4077. https://doi.org/10.1049/gtd2.13330
[19] Dong J, Chen C, Zhang C, Ma J, Peng K. Knowledge Graph Embedding With Graph Convolutional Network and Bidirectional Gated Recurrent Unit for Fault Diagnosis of Industrial Processes. IEEE Sensors Journal. 2025;25(5):8611-8620. https://doi.org/10.1109/JSEN.2025.3528223
[20] Yu X, Zhang Z, Tang B, Zhao M. Meta-Adaptive Graph Convolutional Networks With Few Samples for the Fault Diagnosis of Rotating Machinery. IEEE Sensors Journal. 2024;24(12):19237-19252. https://doi.org/10.1109/JSEN.2024.3392372