

Multi-Source Uncertainty Perception and Spatiotemporal Graph Fusion for Defect-Elimination Robots in Strong Electromagnetic Environments

Peng Lei¹, Jiabin Liang¹, Tianlong Zhang¹, Shuo Wang¹, Jianling Wu¹, Sen Wang¹

¹ Wuhai Extra-High Voltage Power Supply Branch, Inner Mongolia Electric Power (Group) Co., Ltd., Wuhai 016000, Inner Mongolia, China

Corresponding author: Jiabin Liang (e-mail: ncasu101@163.com).

ABSTRACT To address perception distortion, difficulties in heterogeneous data fusion, and missed detection of small faults in defect-elimination robots for ultra-high-voltage transmission lines under strong electromagnetic interference, a physics-aware spatiotemporal graph fusion network PA-STGFN is proposed. The Middleton Class-A model is employed to estimate the parameters of non-Gaussian impulsive noise, combined with robust statistics for adaptive suppression. EfficientNet enhanced with coordinate attention and MS-TCN are used to extract visual and vibration–acoustic features, respectively. Modality uncertainty is quantified based on Dirichlet evidence, enabling dynamic reconstruction of graph connections and multi-source confidence-weighted fusion. The model is jointly trained using focal loss, reconstruction loss, and uncertainty regularization. Validation on transmission-line data shows that the feature-matching accuracy and anomaly-detection rate reach 95.8% and 96.5%, respectively, with a performance degradation of less than 3% at 10 dB and a single-frame inference latency of 78 ms. The proposed method can improve the perception reliability of defect-elimination robots in strong electromagnetic environments.

INDEX TERMS Defect-removal robot, fault warning, high-voltage technology, Middleton Class-A noise, multi-sensor fusion, uncertainty quantification.

I. INTRODUCTION

ULTRA-HIGH-VOLTAGE transmission lines are continuously exposed to coupled environments involving strong electric fields, complex meteorological conditions, and mechanical vibrations. Defects such as loose fittings, insulator cracks, conductor damage, local overheating, and abnormal discharge are characterized by small scales, rapid evolution, and sparse samples. Live-line defect-elimination robots can acquire visible-light, infrared, vibration, and acoustic information at close range without power interruption, thereby reducing the risks associated with manual high-altitude operations. However, both the robotic platform and sensing links are simultaneously exposed to broadband impulsive interference generated by corona discharge, switching operations, wireless communication, and motor commutation. Existing studies on transmission-line robots have indicated that electromagnetic compatibility, reliable fault identification, and autonomous operation remain major bottlenecks to engineering deployment [1]–[3].

From the perspective of practical operation procedures, defect-elimination robots are required to sequentially accomplish equipotential approach, line traversal, component localization, defect confirmation, and end-effector operation.

The perception results therefore serve not only fault identification but also directly affect motion control and operational interlocking. Compared with long-distance inspection, cameras, microphones, and vibration sensors are positioned closer to conductors and fittings during near-field defect elimination, resulting in stronger signals associated with partial discharge and structural impacts. However, electric-field coupling, contact transients, mechanical-actuator vibrations, and posture variations also become more pronounced. Small cracks, early-stage looseness, and slight corrosion often generate only weak responses in local textures or narrow frequency bands. If the perception model cannot distinguish valid fault transients from electromagnetic spikes, missed detections, false alarms, or erroneous high-confidence decisions may occur. Therefore, perception systems for defect-elimination tasks must simultaneously satisfy the requirements of interference resistance, heterogeneous fusion, reliable evaluation, and real-time edge processing.

Conventional perception methods generally simplify interference as stationary Gaussian noise and employ fixed-threshold filtering or single-band suppression. Under non-Gaussian heavy-tailed noise, excessive clipping may simultaneously remove mechanical fault impacts, whereas insuf-

ficient clipping cannot effectively eliminate electromagnetic spikes. The representative Middleton Class-A model can characterize the statistical properties of electromagnetic interference within the receiver bandwidth using the impulsive index and the background-to-impulsive power ratio [4]. Related studies have demonstrated that adaptive thresholds are more suitable than fixed clipping for impulsive noise with varying parameters [5], [6]. Therefore, embedding this physical prior into the front end of a neural network can transform noise suppression from empirical parameter tuning into interpretable online estimation.

Existing studies on impulsive-noise suppression have mainly focused on communication receivers, typically using bit error rate or output signal-to-noise ratio as optimization objectives and assuming that anomalous high-amplitude samples represent invalid interference [5]–[8]. However, looseness-induced impacts, frictional collisions, and partial discharges in transmission lines also exhibit sparse, transient, and heavy-tailed characteristics. Simple blanking or clipping may therefore eliminate the most diagnostically valuable transient components. Meanwhile, field interference intensity varies with the distance between the robot and conductor, load current, end-effector actions, and shielding conditions, making fixed thresholds insufficient to cover the complete operating cycle. Accordingly, the Class-A statistical parameters need to be combined with physical relationships such as vibration–acoustic propagation consistency and synchronous variations in field strength, allowing preprocessing to adaptively adjust according to the noise state while preserving genuine fault components that remain consistent across sensors.

For multi-source fusion, images can localize cracks, corrosion, and hot spots, whereas vibration and acoustic signals are more sensitive to early changes associated with looseness, friction, and partial discharge. These modalities differ substantially in sampling rate, dimensionality, latency, and failure modes. Simple concatenation or fixed weighting can easily induce negative transfer when an individual sensor is occluded, saturated, or subject to data drift. The compound scaling design of EfficientNet balances accuracy and computational cost [9], while coordinate attention (CA) preserves directional and positional information with relatively low computational overhead [10]. MS-TCN models long-term dependencies through multi-stage dilated convolutions [11], whereas spatiotemporal graph networks provide a unified representation for modelling cross-modal relationships [12]–[14]. However, existing methods generally treat graph topology and modality weights as fixed quantities and lack explicit modelling of whether the current data are trustworthy.

In multimodal deep learning, early fusion emphasizes feature sharing but is vulnerable to differences in scale and temporal misalignment, whereas late fusion is straightforward to implement but has difficulty describing conditional dependencies among thermal–electrical, vibration–acoustic, and image–operating-condition information [15]. Graph convolution and graph attention can represent different sensors as

nodes and learn information pathways [12]–[14], [16], but fixed adjacency matrices cannot reflect temporal variations in sensor quality, whereas fully data-driven dynamic graphs may amplify erroneous correlations under strong noise. Meanwhile, conventional Softmax can produce excessively high confidence even for occluded, saturated, or out-of-distribution inputs [17]–[20]. Deep ensembles and MC-dropout can improve uncertainty estimation, but repeated forward inference is difficult to reconcile with the real-time requirements of robotic edge platforms [18], [20]. Therefore, a quantification method capable of obtaining evidence strength through a single forward pass is required, with uncertainty directly incorporated into graph-topology reconstruction and modality-weight allocation.

Overall, multi-source perception in strong electromagnetic environments still presents three interrelated technical gaps: first, the lack of physically constrained preprocessing capable of distinguishing electromagnetic spikes from genuine mechanical impacts; second, the lack of a lightweight heterogeneous encoding structure that simultaneously accommodates spatial details in images and long-term dependencies in vibration–acoustic signals; and third, the lack of a fusion mechanism that unifies sensor reliability, physical correlations, and dynamic graph interactions. These problems cannot be resolved simply by increasing network depth because input noise, node relationships, and decision confidence interact throughout the inference pipeline. Accordingly, this study constructs a closed-loop technical framework comprising “physical noise modelling–heterogeneous feature encoding–evidential uncertainty quantification–dynamic graph interaction–dynamic confidence fusion”, and jointly evaluates the framework in terms of low signal-to-noise ratio, modality degradation, confidence calibration, and edge inference latency.

To address these issues, this study proposes a physics-aware spatio-temporal graph fusion network (PA-STGFN). The main contributions are as follows:

- 1) the Middleton Class-A Gaussian-mixture model is employed to estimate the impulsive index A , power ratio Γ , and posterior probability of impulsive samples online, thereby constructing adaptive preprocessing that preserves fault impacts while suppressing electromagnetic spikes;
- 2) CA-EfficientNet and MS-TCN are used to encode spatial and temporal features, respectively, with synchronized windows and spectral statistics employed to map heterogeneous features into a unified dimensional space;
- 3) modality uncertainty is quantified based on Dirichlet evidential distributions, and dynamic graph edges are jointly reconstructed using uncertainty, cross-modal consistency, and operating-condition priors to achieve multi-source dynamic confidence weighting;
- 4) focal loss, reconstruction loss, and uncertainty regularization are jointly employed for training to improve sparse-fault identification and confidence calibration under interference conditions.

II. OVERALL FRAMEWORK AND PHYSICS-AWARE INTERFERENCE SUPPRESSION OF PA-STGFN

A. TASK DEFINITION AND OVERALL ARCHITECTURE

Suppose that, at time t , the robot acquires M synchronized modalities $X_t = \{X_t^{(m)}\}_{m=1}^M$, where the visible-light and infrared images are denoted as I_t^{vis} and I_t^{ir} , respectively; the vibration and acoustic windows are denoted as s_t^{vib} and s_t^{aco} , respectively; and the equipment-state variables include travelling speed, clamping force, motor current, and local field strength. For a sliding window of length T , the model outputs the probabilities of K fault classes $p_t \in \mathbb{R}^K$, the predictive uncertainty u_t , and the dynamic weight of each modality $w_t^{(m)}$. The task requires not only correct identification of fault categories, but also reduction of the corresponding modality weight when a single modality becomes abnormal, while satisfying the real-time constraints of onboard deployment.

$$X_t = [I_t^{\text{vis}}, I_t^{\text{ir}}, s_t^{\text{vib}}, s_t^{\text{aco}}, q_t], \quad (\hat{p}_t, u_t, w_t) = F_\theta(X_t). \quad (1)$$

where q_t denotes the operational context. The fault set comprises seven classes: normal condition, bolt looseness, insulator crack, fitting corrosion, conductor damage, local overheating, and abnormal discharge. The training, validation, and test sets are partitioned at the transmission-line-section level to avoid overestimation caused by information leakage between adjacent frames.

PA-STGFN consists of a physics-based interference-suppression layer, a heterogeneous dual-stream encoder, an uncertainty-aware dynamic graph interaction layer, and a dynamic confidence fusion head, as shown in Fig. 1. The physical layer first estimates the noise parameters and generates an impulsive-probability mask. The visual stream employs CA-EfficientNet to preserve directional localization information for small-scale defects, whereas the temporal stream employs MS-TCN to extract cross-scale impacts and frequency-band variations. Each modality outputs a feature node and an evidence vector. The dynamic graph updates edge weights in real time according to uncertainty and consistency, after which the fusion head performs classification, anomaly detection, and confidence output.

Interference in the robotic perception chain mainly enters through three types of pathways. First, corona discharge and intermittent discharge on conductor and fitting surfaces generate broadband radiation, which produces common-mode coupling through camera data cables, vibration acquisition lines, and analogue front ends. Second, PWM commutation of the robot drives and wireless communication generate periodic narrowband components that overlap with the fundamental mechanical frequencies. Third, transient currents during equipotential approach, contact, and withdrawal stages produce conducted coupling through the structural ground and shielding layers. Within the receiving window, the first two types of interference generally manifest as randomly arriving high-amplitude impulses superimposed on background Gaussian noise, satisfying the applicability

PA-STGFN: Physics-Aware Spatio-Temporal Graph Fusion Pipeline

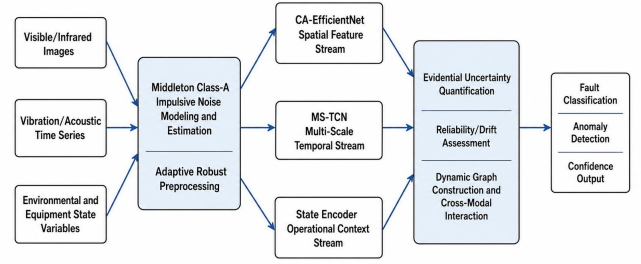


FIGURE 1. Overall structure and information flow of PA-STGFN.

conditions of the Class-A model. Periodic narrowband components are instead handled separately through notch filtering and operating-condition encoding and are not forcibly incorporated into the interpretation of the Class-A parameters.

The model is established based on the following assumptions:

- 1) the multi-source data have been calibrated to a unified clock, and residual deviations within each window can be explicitly represented by quality indicators;
- 2) the training labels correspond to verifiable component states, whereas samples whose states cannot be confirmed are excluded from supervised classification but can be used for reconstruction pre-training;
- 3) the impulsive-noise arrival rate is approximately stationary within a 1-s window, while $\hat{A}$ and $\hat{\Gamma}$ are allowed to vary dynamically across windows;
- 4) at least two valid modalities are available at any given time; otherwise, the system enters a safe degradation mode and does not perform autonomous defect elimination. These assumptions clearly delineate the boundaries among physical parameters, deep features, and safety strategies, while also facilitating subsequent field verification.

B. MIDDLETON CLASS-A NON-GAUSSIAN NOISE MODELLING

For the m -th temporal sensor, the observed signal is $x[n] = s[n] + \eta[n]$, where $s[n]$ denotes the structural or acoustic response and $\eta[n]$ denotes the electromagnetically coupled noise. The Class-A model represents η as a Poisson-weighted sum of zero-mean Gaussian components with different variances:

$$p(\eta) = \sum_{l=0}^{\infty} \frac{e^{-A} A^l}{l!} \cdot \frac{1}{\sqrt{2\pi}\sigma_l^2} \exp\left(-\frac{\eta^2}{2\sigma_l^2}\right). \quad (2)$$

$$\sigma_l^2 = \sigma_\eta^2 \frac{(l/A) + \Gamma}{1 + \Gamma}, \quad \Gamma = \frac{\sigma_G^2}{\sigma_I^2}. \quad (3)$$

where A is the average number of impulses within a unit observation interval; a smaller A indicates sparser noise with higher kurtosis. Γ is the ratio of the background Gaussian-noise power to the impulsive-noise power, and σ_η^2 is the total noise power. As $A \rightarrow \infty$ or $\Gamma \rightarrow \infty$, the mixture

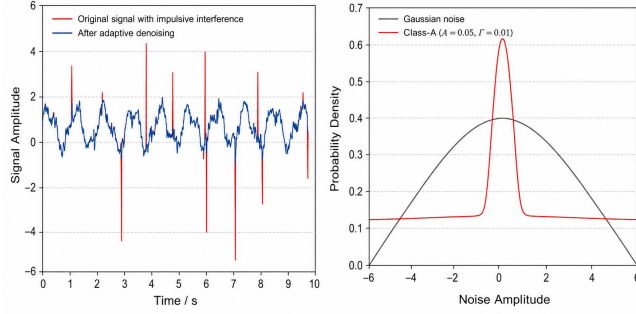


FIGURE 2. Heavy-tailed characteristics of Middleton Class-A noise and an example of adaptive preprocessing.

distribution gradually approaches a Gaussian distribution. To satisfy embedded-computing requirements, the summation is truncated at $L = \max(8, A + 6\sqrt{A})$, such that the tail Poisson mass is less than 10^{-6} .

The heavy-tailed characteristics of the Middleton Class-A noise and an example of adaptive preprocessing are illustrated in Fig. 2.

C. PARAMETER ESTIMATION AND ADAPTIVE INTERFERENCE-RESISTANT PREPROCESSING

Within each T -s window, the median and median absolute deviation (MAD) are first used to obtain the robust centre $\hat{\mu}$ and scale $\hat{\sigma}_r$. Robust statistics can limit the influence of extreme samples on location and scale estimation [7], [8]. Using the adjacent fault-free baseline window as the prior, three generalized EM updates are performed with a truncated Gaussian-mixture likelihood to obtain $\hat{A}$, $\hat{\Gamma}$, and $\hat{\sigma}_\eta^2$. The posterior probability that the n -th sample belongs to the impulsive component is:

$$\rho_n = 1 - P(l = 0 | x[n]) = 1 - \frac{\pi_0 \mathcal{N}(x[n]; 0, \sigma_0^2)}{\sum_{l=0}^L \pi_l \mathcal{N}(x[n]; 0, \sigma_l^2)}. \quad (4)$$

To avoid misclassifying genuine mechanical impacts as electromagnetic impulses, a cross-sensor physical consistency measure c_n is constructed. When the vibration and acoustic signals exhibit coherent spectral peaks within the allowable propagation delay and no synchronous spikes are observed in the motor-current/field-strength channels, c_n approaches 1; otherwise, it approaches 0. The final electromagnetic-interference probability is defined as $r_n = \rho_n(1 - c_n)$. This strategy exploits the distinction that mechanical impacts exhibit structural propagation consistency, whereas electromagnetic coupling commonly appears synchronously in electrical channels, thereby preserving diagnostically meaningful transients.

Fixed blanking sets all high-amplitude samples to zero and can therefore disrupt the impact sequences associated with looseness and discharge faults [5]. In this study, posterior-controlled soft suppression is adopted. The residual $d_n = x[n] - \text{median}_W(x[n])$ is first subjected to adaptive Huber shrinkage, after which interpolation between the original and shrunk samples is performed according to

r_n . The threshold κ_n varies with $\hat{A}$ and $\hat{\Gamma}$: the sparser the impulses and the lower the background proportion, the smaller the threshold.

$$\kappa_n = \kappa_0 \hat{\sigma}_r \sqrt{\frac{\hat{\Gamma} + \hat{A}}{1 + \hat{\Gamma}}}, \quad \tilde{d}_n = \text{sign}(d_n) \min(|d_n|, \kappa_n). \quad (5)$$

$$\tilde{x}[n] = x[n] - r_n(d_n - \tilde{d}_n), \quad \tilde{S}(f) = S_{\tilde{x}}(f) + c \cdot S_{\text{res}}(f). \quad (6)$$

The second term in (6) reinjects residual spectral energy only into fault-related frequency bands with high cross-modal consistency, thereby alleviating excessive smoothing. Amplitude clipping is not applied to the image channels. Instead, interference gating is generated from field-strength and exposure statistics, and lightweight guided filtering is applied to striped, flickering, and locally saturated regions. After processing, quality indicators q_m , including saturation rate, residual kurtosis, clock jitter, and reconstruction error, are simultaneously output as inputs to the subsequent uncertainty graph.

Short-window estimation of the Class-A parameters is subject to variance, and direct window-by-window updating can readily cause threshold fluctuations. Therefore, exponential smoothing is applied to $\hat{A}$ and $\hat{\Gamma}$, and the rate of threshold variation between adjacent windows is constrained to no more than 20%. When the mixture likelihood fails to converge or $\hat{A}$ falls outside the training distribution, the system falls back to conservative MAD-controlled Huber shrinkage, while the estimation failure is recorded in the quality vector. The complexity of the truncated EM procedure is $O(NL)$, where N is the number of samples in the window and L is the number of mixture components. Its computation involves only scalar Gaussian probabilities and normalization and can therefore be performed in parallel before temporal encoding. For $L = 12$ and $N = 12800$, three updates do not alter the asymptotic complexity of the network backbone.

To verify that preprocessing does not introduce spurious faults, three quantities are simultaneously monitored during training: the signal distortion ratio of low-noise baseline windows, the energy retention ratio within annotated fault-related frequency bands, and the cross-sensor coherence coefficient before and after processing. If the energy retention ratio falls below 0.90, the consistency-protection coefficient is increased; if the residual kurtosis remains more than three times the baseline level, κ_0 is reduced. This closed-loop mechanism provides observable evidence for threshold selection rather than tuning parameters solely around the final classification metrics.

III. HETEROGENEOUS FEATURE ENCODING AND UNCERTAINTY-AWARE GRAPH FUSION

A. COORDINATE-ATTENTION VISUAL ENCODING AND MS-TCN TEMPORAL ENCODING

The visual encoder uses EfficientNet-B0 as the backbone, with coordinate attention inserted after the third, fifth, and

seventh MBConv stages. Compared with the SE module, which models only channel-wise relationships [21], and the local spatial attention mechanism CBAM [22], CA decomposes two-dimensional global pooling into one-dimensional encoding along the height and width directions, thereby preserving both long-range dependencies and positional information [10]. For an input feature $X \in \mathbb{R}^{C \times H \times W}$, the directional descriptors are

$$z_c^h(h) = \frac{1}{W} \sum_{i=1}^W X_c(h, i), \quad z_c^w(w) = \frac{1}{H} \sum_{j=1}^H X_c(j, w). \quad (7)$$

$$Y_c(h, w) = X_c(h, w) \cdot g_c^h(h) \cdot g_c^w(w). \quad (8)$$

where g^h and g^w are obtained through branch-specific mappings following a shared 1×1 convolution, batch normalization, and h-swish activation. To accommodate fine cracks and small-area corrosion, the backbone input is set to 224×224 , and shallow edge features are projected through a 1×1 convolution and concatenated with the terminal semantic features to obtain the 256-dimensional visual node h_t^{vis} . The infrared images use the same architecture but do not share the first-layer parameters, thereby avoiding statistical conflicts between temperature pseudo-colour patterns and visible-light textures. The compound scaling of EfficientNet and lightweight attention enable the visual branch to retain onboard computational efficiency [9], [10].

After preprocessing, the raw waveform, log-magnitude spectrum, and envelope spectrum are calculated separately for the vibration and acoustic signals and concatenated into a $C_s \times T_s$ input. MS-TCN consists of three refinement stages, each containing six layers of causal dilated one-dimensional convolution, with dilation rate $d_l = 2^l$ and kernel size $k = 3$. The output of the l -th layer is:

$$H_t^{(l)} = \text{ReLU}\left(W_1 * H_t^{(l-1)} + W_2 *_{d_l} H_t^{(l-1)}\right) + H_t^{(l-1)}. \quad (9)$$

The multi-stage structure first generates coarse-grained fault probabilities and then suppresses local over-segmentation through subsequent refinement stages, making it suitable for identifying looseness impacts, friction-induced squealing, and discharge impulses with substantially different durations [11], [23]. To preserve both low-frequency structural vibration and high-frequency acoustic emission, parallel depth-wise separable convolutions with $k \in \{3, 5, 9\}$ are additionally introduced, together with frequency-band attention at the end of each stage. Global statistical pooling outputs the mean, standard deviation, and maximum response, forming the 256-dimensional temporal nodes h_t^{vib} and h_t^{aco} .

All sensors use unified timestamps from the robot controller. Taking each image frame as the anchor, a temporal window within $[t - 0.5 \text{ s}, t + 0.5 \text{ s}]$ is extracted. If the maximum clock offset exceeds 5 ms, a latency penalty is incorporated into the quality indicator. Each modality is mapped into a common $d = 256$ dimensional space through layer normalization and linear projection, with modality

embeddings, relative temporal encoding, and operating-condition encoding appended. This design does not force different modalities to follow identical data distributions, but only requires consistent node dimensionality before interaction, thereby reducing the scale bias introduced by raw-level concatenation.

The shallow 1/4 and 1/8 scale features of the visual branch are responsible for edges and local textures, respectively, whereas the terminal 1/32 scale features provide component-level semantics. Coordinate attention generates weight sequences of lengths H and W along the horizontal and vertical directions, respectively, and therefore does not completely discard crack-location information as two-dimensional global pooling does. For the temporal branch, the theoretical receptive field of six dilated-convolution layers in a single stage is $1 + 2(k - 1) \sum 2^l = 253$ sampling steps. Multi-stage refinement repeatedly exploits this context without increasing the sampling window, while the parallel $k = 9$ branch further covers low-frequency structural responses. All nodes are ultimately compressed to 256 dimensions, and the graph size within each window remains fixed at five nodes, structurally limiting the computational overhead of graph interaction.

To prevent pretrained visual features from dominating fusion, modality-specific layer normalization is adopted in the common projection layer. During the first 15 epochs of training, a temperature constraint that keeps graph weights close to a uniform distribution is imposed and is subsequently relaxed gradually. Missing modalities are not zero-filled; instead, explicit missing-modality indicators are introduced and the corresponding edges are masked, enabling the network to distinguish genuine zero values, sensor silence, and communication packet loss.

B. EVIDENTIAL UNCERTAINTY AND DYNAMIC GRAPH INTERACTION

Conventional Softmax may produce excessively high confidence for out-of-distribution inputs [19]. In this study, an evidential head is appended to each modality node to output non-negative evidence $e_k = \text{softplus}(o_k)$, with the Dirichlet parameter defined as $\alpha_k = e_k + 1$. According to evidential deep learning [17], the expected class probability, overall uncertainty, and class-wise variance are

$$\hat{p}_k = \frac{\alpha_k}{S}, \quad u = \frac{K}{S}, \quad \text{Var}(p_k) = \frac{\alpha_k(S - \alpha_k)}{S^2(S + 1)}, \quad S = \sum_k \alpha_k. \quad (10)$$

where u describes insufficient total evidence, whereas $\text{Var}(p_k)$ characterizes the confidence variance at the class level. Compared with MC-dropout or deep ensembles, which require multiple forward passes [18], [20], the single-pass evidential head is more suitable for real-time inference on edge platforms. To improve calibration, only the final logits are temperature-scaled using a temperature parameter

estimated on the validation set, and the expected calibration error (ECE) is used to evaluate consistency between confidence and empirical correctness [19].

For each temporal window, a graph $G_t = (V_t, E_t, A_t)$ is constructed, in which the nodes correspond to visible-light, infrared, vibration, acoustic, and equipment-state information. Unlike the fixed topology of ST-GCN [12], the proposed method dynamically reconstructs edges according to node content, uncertainty, and cross-modal consistency. The edge score is defined as

$$s_{ij} = \frac{(W_q h_i)^T (W_k h_j)}{\sqrt{d}} - \lambda_u(u_i + u_j) + \lambda_c c_{ij} + b_{ij}^{\text{phy}}. \quad (11)$$

$$a_{ij} = \text{softmax}_j(s_{ij}), \quad \bar{h}_i = \text{LN} \left[h_i + \sum_{j \in \mathcal{N}_i} a_{ij} W_v h_j \right]. \quad (12)$$

where c_{ij} denotes cross-modal time-frequency consistency, and b_{ij}^{phy} represents the physical association prior between sensors, such as positive correlations between hot spots and load current, looseness and vibration envelopes, and discharge and high-frequency acoustic signals. If a node exhibits high u_i or abnormal quality indicators, both its outgoing and incoming edge weights are suppressed; if two reliable nodes exhibit consistent abnormalities, their edge weight is enhanced. To control complexity, only the three highest-scoring edges of each node are retained, and interaction is performed using two layers of four-head graph attention. The dynamic-graph formulation is consistent with the foundations of graph attention and spatio-temporal graph modelling [12]–[14], but directly incorporates uncertainty into topology learning.

C. DYNAMIC CONFIDENCE FUSION AND SAFE DEGRADATION

For each node after interaction, an integrated reliability score r_i is calculated by jointly considering evidential uncertainty u_i , class variance v_i , residual reconstruction error ε_i , and data quality q_i . The modality weights are normalized using temperature Softmax:

$$r_i = \exp[-\beta_1 u_i - \beta_2 v_i - \beta_3 \varepsilon_i] \cdot q_i, \quad (13)$$

$$w_i = \frac{\exp(r_i/\tau)}{\sum_j \exp(r_j/\tau)}.$$

$$z_t = \sum_i w_i \bar{h}_i, \quad \hat{p}_t = \text{softmax}(W_o z_t + b_o). \quad (14)$$

The weights are not prior probabilities of fault classes but represent the reliable contribution of each modality to the fusion decision within the current temporal window. To prevent a single modality from dominating over extended periods, a weak entropy constraint is imposed on w during training; low-entropy allocation is permitted only when sufficient evidence and cross-modal consistency are present. During online operation, w_i and u_i are retained to provide maintenance personnel with an interpretable indication of

which sensors support the decision and which channels are unreliable [24].

If a dynamic graph is driven entirely by instantaneous content, frequent reconnection may occur around noise spikes. Therefore, in addition to (12), an L_1 smoothness term is imposed on adjacency differences between adjacent windows, while thermal–electrical and vibration–acoustic physical-prior edges are retained as a minimum connectivity backbone. These backbone edges are allowed to decay only when node quality falls below a predefined threshold. If the fused uncertainty $u_t > 0.35$, the maximum class probability is below 0.60, or the weight entropy remains abnormally high for three consecutive windows, the system outputs “unconfirmed” rather than forcing classification and sends a request to the robot controller for deceleration, re-imaging, or secondary sampling.

IV. HYBRID LOSS AND TRAINING DEPLOYMENT

A. HYBRID LOSS

Fault samples are substantially fewer than normal samples, and microcracks and slight looseness are often difficult-to-classify samples. Focal loss is adopted to reduce the dominance of easily classified normal samples [25]:

$$\mathcal{L}_{\text{foc}} = - \sum_n \alpha_{y_n} (1 - \hat{p}_{n,y_n})^\gamma \log(\hat{p}_{n,y_n}). \quad (15)$$

To prevent the interference-suppression layer from removing valid fault components, a lightweight decoder reconstructs the low-noise baseline and critical frequency bands before preprocessing from the node features, yielding a mask-weighted L_1 reconstruction loss. For the evidential distribution, an incorrect-evidence penalty and Dirichlet-KL regularization are adopted to maintain high uncertainty for misclassified samples and suppress spurious confidence in classes without supporting evidence:

$$\mathcal{L}_{\text{rec}} = \sum_m \|M_m \odot (X_m - D_m(h_m))\|_1. \quad (16)$$

$$\mathcal{L}_{\text{unc}} = \sum_n [(1 - \hat{p}_{n,y_n})(1 - u_n) + \lambda_{\text{KL}} \text{KL}(\text{Dir}(\tilde{\alpha}_n) \| \text{Dir}(\mathbf{1}))]. \quad (17)$$

$$\mathcal{L} = \mathcal{L}_{\text{foc}} + \lambda_r \mathcal{L}_{\text{rec}} + \lambda_u \mathcal{L}_{\text{unc}} + \lambda_s \mathcal{L}_{\text{smooth}} + \lambda_w \mathcal{L}_{\text{weight}}. \quad (18)$$

where $\gamma = 2$, $\lambda_r = 0.25$, λ_u is linearly annealed from 0 to 0.10, $\lambda_s = 0.05$ constrains prediction smoothness between adjacent windows, and $\lambda_w = 0.01$ is the weight-entropy regularization coefficient. The KL term is calculated after setting the evidence of the ground-truth class in $\tilde{\alpha}$ to 1, thereby avoiding suppression of correct evidence.

B. TRAINING, INFERENCE, AND COMPUTATIONAL COMPLEXITY

Input: synchronized multi-source window X , class label y ;
Output: class probability $\hat{p}$, anomaly flag u , and modality weights w .

- 1) Update $\hat{A}$, $\hat{\Gamma}$, and $\hat{\sigma}_\eta^2$ using robust statistics and truncated EM, and calculate the sample-wise impulsive posterior ρ ;
- 2) Integrate cross-sensor consistency to obtain the electromagnetic-interference mask r , and perform soft suppression and spectral compensation;
- 3) Obtain nodes h using CA-EfficientNet for images, MS-TCN for temporal signals, and an MLP for state variables;
- 4) Use the evidential head to calculate α , $\hat{p}_m$, u_m , and confidence variance, and reconstruct the dynamic graph A_t in conjunction with the quality indicators;
- 5) Perform two-layer graph interaction and dynamic confidence weighting, calculate (18), and update the parameters using AdamW;
- 6) During online operation, only a single forward pass is performed; if the maximum class probability is below the threshold or $u > u_{\text{alarm}}$, the sample is submitted for manual review.

Training is divided into three stages. In Stage I, the visual, vibration, and acoustic encoders are trained independently to provide each modality with basic discriminative capability. In Stage II, the range of physical preprocessing parameters is fixed, and the evidential head and dynamic graph are jointly trained while λ_u is gradually increased. In Stage III, end-to-end fine-tuning is performed under mixed-noise, occlusion, and packet-loss augmentation, and the temperature τ is selected according to the validation-set ECE. This staged training strategy prevents graph-weight collapse caused by insufficient evidence during the early training stage.

Let P denote the number of elements in the visual feature maps, T_s the temporal sequence length, M the number of graph nodes, and k_g the number of retained edges. The principal computational complexity of a single forward pass is $O(PC_v + T_s C_t + M k_g d)$.

In this study, $M = 5$ and $k_g = 3$, and the computational cost of graph interaction is therefore substantially lower than that of the visual convolutions. For edge deployment, the model uses FP16 precision, a fixed input size, and a batch size of 1. All latency measurements start from data preprocessing and include graph construction, thereby avoiding the reporting of only the core network inference time.

The training platform was an NVIDIA RTX 3060 12 GB, whereas edge deployment was evaluated on a Jetson Orin NX 16 GB under FP16 precision and a batch size of 1. AdamW [26] was used as the optimizer, with an initial learning rate of 3×10^{-4} , weight decay of 10^{-4} , cosine annealing over 120 epochs, and a batch size of 24. The visual backbone was initialized with ImageNet pretraining, and the first two stages were frozen during the first 10 epochs. The Class-A parameter estimation and interference-suppression layers did not participate in back-propagation, although their quality indicators were provided to the graph network. The evaluation metrics included feature matching accuracy, defined as the proportion of

samples for which both the cross-modal temporal window and component identity were correctly matched; anomaly detection rate, $\text{ADR} = \text{TP}/(\text{TP} + \text{FN})$; macro-averaged F1; error-detection AUROC; ECE; parameter count; and single-frame latency.

V. EXPERIMENTS AND RESULTS ANALYSIS

A. DATA ORGANIZATION AND EXPERIMENTAL SETTINGS

The experiments were conducted for defect-elimination operation scenarios on ultra-high-voltage transmission lines in the Wuhai region, with synchronized multi-source windows organized according to transmission-line sections. To present the complete validation protocol, the sample size in the present manuscript is provisionally described as 12,600 1-s windows, with the class distribution and acquisition parameters listed in Table I. A section-level split ratio of 7:1:2 was adopted, and normalization was performed using statistics calculated exclusively from the training sections. Random brightness adjustment, mild blurring, and occlusion augmentation were applied to the images. Class-A noise was superimposed on the temporal channels, with $A \in \{0.01, 0.03, 0.10, 0.30\}$, $\Gamma \in \{0.001, 0.01, 0.10\}$, and $\text{SNR} \in [10, 30]$ dB.

The acquisition procedure was organized according to the sequence “approach–inspection–localization–defect elimination–withdrawal”. Each component was synchronously recorded using visible-light, infrared, triaxial vibration, acoustic, and control-state signals. Continuous video was segmented into 1-s windows using key frames as anchors, with 50% overlap between adjacent windows. Annotations were first independently completed by two operation and maintenance personnel and then consolidated into consistent labels using outage reinspection, infrared temperature-rise measurements, and torque retesting. Windows with disputes or without closed-loop confirmation were used only for unsupervised reconstruction. To reduce the correlation introduced by consecutive frames of the same component, the training, validation, and test sets were strictly grouped by transmission-line section and operation date. The detailed class composition and diagnostic criteria are summarized in Table II.

In addition to point estimates, section-level Bootstrap was used to calculate 95% confidence intervals, and paired permutation tests were used to compare PA-STGFN with the strongest baseline. Thresholds, temperature, and loss weights were determined only on the validation set. The test set was evaluated once after all configurations were frozen. For real-time evaluation, 100 warm-up runs were performed before calculating the mean and P95 latency over 1,000 windows, thereby avoiding biases caused by cold start and data transfer.

B. COMPARATIVE AND ABLATION EXPERIMENTS

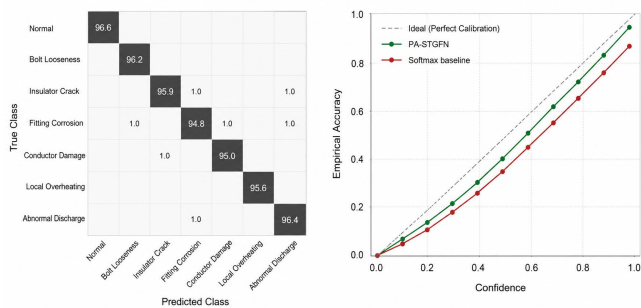
A single-visual EfficientNet-B0, a CNN-LSTM dual-stream model, late-stage concatenation, fixed-topology ST-GCN,

TABLE I. Multi-Source Data and Preprocessing Parameters

Modality	Sensor/Sampling	Window Input	Main Fault Information
Visible light	1920×1080, 25 frame/s	224×224 RGB	Cracks, corrosion, broken strands, foreign objects
Infrared	160×120, 25 frame/s	224×224 single channel	Local overheating, contact abnormalities
Vibration	Triaxial, 12.8 kHz	1-s waveform + envelope spectrum	Looseness, rubbing, structural impacts
Acoustic	48 kHz	1-s waveform + logarithmic spectrum	Discharge, friction-induced squealing
State variables	100 Hz	Speed/current/clamping force/field strength	Operating conditions and interference priors

TABLE II. Dataset Class Composition and Main Criteria

Class	Number of Windows	Proportion/%	Main Criteria	Confounding Factors
Normal	4,200	33.3	Structure and temperature rise within baseline range	Load fluctuations
Bolt looseness	1,560	12.4	Enhanced impacts and abnormal fastening torque	Robot speed variation
Insulator crack	1,320	10.5	Fine-crack texture and local discharge	Contamination/reflection
Fitting corrosion	1,480	11.7	Rough surface texture and colour difference	Shadows/coatings
Conductor damage	1,260	10.0	Broken strands, burrs, and local vibration	Background branches
Local overheating	1,500	11.9	Joint abnormality of temperature rise and load current	Solar reflection
Abnormal discharge	1,280	10.2	Acoustic high-frequency impulses and field-strength spikes	Motor commutation

**FIGURE 3.** Class confusion and confidence calibration results.

and evidential quantification with a fixed graph were selected as comparison methods. To ensure fairness, all fusion models used the same input windows and training split. The results are presented in Table III. PA-STGFN achieved a feature matching accuracy of 95.8%, an anomaly detection rate of 96.5%, and a macro-averaged F1 of 95.1%. Compared with late-stage concatenation, the three metrics increased by 3.7, 3.5, and 2.7 percentage points, respectively, indicating that dynamic graph interaction does not simply increase the number of modalities, but reduces erroneous propagation by suppressing unreliable nodes. Although the evidential fixed-graph model improved ECE, fixed edges still introduced negative transfer when sensor distortion occurred.

At the class level, the recall rates for normal conditions and local overheating were relatively high. The primary confusion for fitting corrosion and bolt looseness originated from mild early-stage samples, for which visual texture differences were weak and vibration features were influenced by robot speed. After incorporating the operating-condition node, amplitude drift caused by speed variation was no longer directly mapped to the looseness class. The normalized confusion matrix is shown on the left of Fig. 3, whereas the reliability diagram on the right indicates that PA-STGFN is closer to the ideal calibration line.

Compared with the evidential fixed-graph model, PA-

STGFN improved matching accuracy by 1.4 percentage points and reduced ECE by 1.3 percentage points, while simultaneously reducing latency by 2 ms. This result indicates that the sparse dynamic graph does not obtain higher accuracy simply by increasing computational cost. In section-level paired comparisons, the primary gains were concentrated in three scenarios: strong discharge interference, visible-light backlighting, and vibration-amplitude drift. Under normal illumination with intact multimodal data, the difference between the two methods was relatively small.

The contributions of individual modules were analysed by removing them one at a time, as shown in Table IV. Removing Class-A adaptive preprocessing reduced matching accuracy by 2.2 percentage points, indicating that non-Gaussian noise modelling contributes to both cross-modal alignment and temporal-feature stability. Removing CA reduced the recall of small-scale cracks and broken strands. Replacing MS-TCN with a single-stage TCN weakened the ability to separate persistent faults from transient impacts. Removing the uncertainty-aware dynamic graph or replacing dynamic weighting with equal weighting markedly increased ECE, indicating that the performance improvement does not arise merely from additional parameters, but from the closed loop of “estimating reliability–adjusting topology–redistributing weights”.

The loss functions were further compared. When only cross-entropy was used, Macro-F1 was 92.7%. Replacing it with focal loss increased Macro-F1 to 93.8%; adding the reconstruction loss further increased it to 94.5%, and incorporating uncertainty regularization yielded 95.1%. Focal loss primarily improved recall for sparse classes, the reconstruction loss reduced excessive denoising, and uncertainty regularization mainly reduced ECE and high-confidence false positives. The results remained stable when λ_u ranged from 0.05 to 0.15, whereas excessively large values made the model overly conservative and reduced the anomaly

TABLE III. Comprehensive Performance Comparison of Different Methods

Method	Matching Accuracy/%	ADR/%	Macro-F1/%	ECE/%	Latency/ms
EfficientNet-B0 (visual)	88.4	89.7	88.9	9.4	42
CNN-LSTM dual stream	90.3	91.2	90.5	8.1	65
Multi-source late-stage concatenation	92.1	93.0	92.4	7.6	70
Fixed-topology ST-GCN	93.2	94.1	93.6	6.8	86
Evidential quantification + fixed graph	94.4	95.2	94.3	4.1	80
PA-STGFN	95.8	96.5	95.1	2.8	78

TABLE IV. Ablation Results of PA-STGFN Modules

Configuration	Matching Accuracy/%	ADR/%	Macro-F1/%	ECE/%
Full PA-STGFN	95.8	96.5	95.1	2.8
Without Class-A adaptive preprocessing	93.6	93.9	92.8	5.7
CA replaced by SE	94.5	94.8	93.7	4.1
MS-TCN replaced by single-stage TCN	93.9	94.3	93.1	4.4
Dynamic graph replaced by fixed graph	94.2	94.9	93.8	5.0
Without evidential uncertainty	93.8	94.6	93.4	7.2
Dynamic weighting replaced by equal weighting	94.1	94.7	93.6	6.6

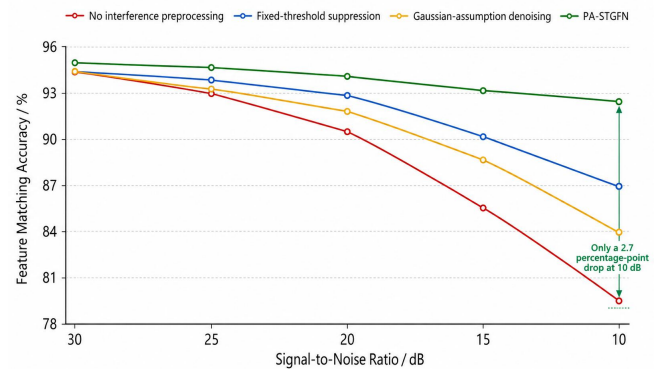
detection rate.

From the perspective of module interaction, Class-A preprocessing and the uncertainty-aware dynamic graph exhibited a synergistic effect. When only preprocessing was retained, residual interference could still propagate through the fixed graph. When only the uncertainty-aware graph was retained, the evidential head had to learn under stronger heavy-tailed noise, leading to degraded calibration at low SNR. When both were present, the front end first compressed abnormal amplitudes and the back end subsequently downweighted residual risks, forming a two-stage protection mechanism. The effects of CA and MS-TCN were mainly reflected in difficult-sample recognition, and their influence on overall ECE was smaller than that of the uncertainty-related modules.

C. ROBUSTNESS TO STRONG ELECTROMAGNETIC INTERFERENCE AND SENSOR DEGRADATION

Interference with different SNR levels and Class-A parameters was superimposed on the test windows while keeping the fault-signal energy unchanged. As shown in Fig. 4, conventional Gaussian denoising and fixed clipping degraded rapidly below 15 dB. The matching accuracy of PA-STGFN decreased from 95.8% at 30 dB to 93.1% at 10 dB, corresponding to a reduction of 2.7 percentage points and satisfying the performance degradation of less than 3% stated in the Abstract. This is because the preprocessing threshold is adjusted according to $\hat{A}$ and $\hat{\Gamma}$; cross-modal consistency protects genuine mechanical impacts; and when residual noise still enters the encoder, evidential uncertainty further reduces the corresponding node-edge weights.

Further variation of the Class-A parameters showed that when A is small and Γ is low, impulses are the sparsest and have the highest peak amplitudes, making fixed clipping most sensitive to the threshold. As A or Γ increases, the

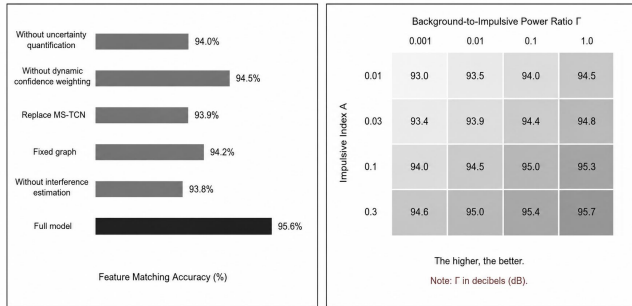
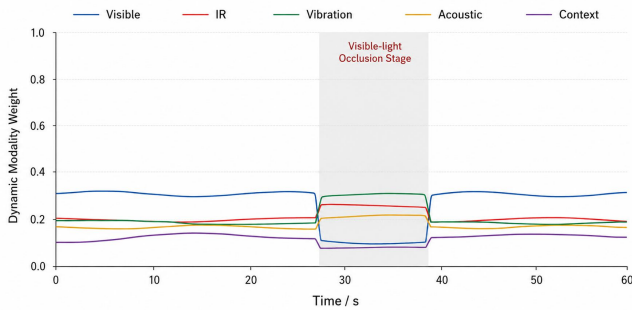
**FIGURE 4.** Robustness of different interference-suppression strategies under Class-A noise.

distribution gradually approaches a Gaussian form and the differences among methods decrease. The minimum matching accuracy of PA-STGFN over the test grid was 93.0%, and no instability was observed in which the weight of a single modality remained above 0.8 for an extended period. The left panel of Fig. 5 summarizes the module ablations, whereas the right panel shows the matching accuracy over the A - Γ parameter grid.

Table V further simulates occlusion, saturation, and packet loss. Equal-weight fusion cannot distinguish between “insufficient defect evidence” and “sensor failure”. Although the fixed graph can exploit adjacency relationships for compensation, unreliable nodes continue to transmit messages. PA-STGFN maintained a Macro-F1 of 88.7% even when two modalities were simultaneously degraded, while the reported uncertainty increased from 0.083 under normal conditions to 0.241, thereby triggering manual review or secondary sampling by the robot. This result indicates that uncertainty is not only used for post hoc interpretation, but also serves as an important control variable for active

TABLE V. Dynamic Fusion Capability Under Typical Sensor-Degradation Conditions

Degradation Condition	Equal-Weight Fusion F1/%	Fixed-Graph F1/%	PA-STGFN F1/%	Main Weight Change
50% visible-light occlusion	86.1	89.3	92.6	Visible light ↓, infrared/vibration ↑
Vibration-channel saturation	84.8	88.7	92.9	Vibration ↓, acoustic/image ↑
30% continuous acoustic packet loss	88.3	90.5	93.4	Acoustic ↓, vibration ↑
Any two modalities degraded	79.5	83.2	88.7	State prior and remaining modalities dominate

**FIGURE 5.** Contributions of key modules and sensitivity to Class-A parameters.**FIGURE 6.** Evolution of multi-source dynamic confidence weights under visible-light occlusion.

perception and safe degradation.

In Fig. 6, after visible-light occlusion occurred between 26 and 37 s, the visual weight decreased from approximately 0.31 to 0.11, whereas the infrared, vibration, and acoustic weights increased simultaneously. After the occlusion was removed, the weights recovered within approximately three windows without abrupt rebound. Because the state-variable node retained speed and field-strength context, its weight changed only slightly, indicating that the model distinguished image-quality degradation from operating-condition variation.

D. UNCERTAINTY, REAL-TIME PERFORMANCE, AND DYNAMIC-GRAPH INTERPRETATION

Table VI compares the uncertainty-estimation performance and computational overhead of different methods. PA-STGFN obtains evidence and uncertainty through a single forward pass, avoiding the repeated inference overhead of MC-dropout and ensemble methods. The model contains approximately 11.2 M parameters and achieves an FP16 end-to-end single-frame latency of 78 ms, including 36

ms for visual encoding, 18 ms for temporal encoding, 9 ms for parameter estimation and preprocessing, and 15 ms for dynamic graph fusion and classification, corresponding to approximately 12 frame/s online processing. Compared with the fixed graph, the Top-3 sparse-edge strategy offsets the small additional overhead introduced by the uncertainty head. The error-detection AUROC is 92.4%, indicating that u provides strong discrimination for misclassification and distribution shift.

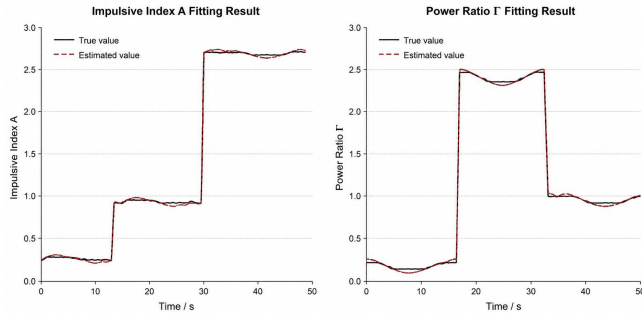
The experimental results verify the complementarity between physical modelling and data-driven fusion. The Class-A parameters provide an interpretable noise state, the deep network learns fault patterns that are difficult to describe explicitly, and the dynamic graph converts real-time reliability into information pathways. It should be noted that the Class-A assumption is suitable for impulsive interference within the receiver bandwidth that has short duration and approximately Poisson arrivals. For strong periodic narrowband harmonics or long-duration arcs, a periodic-noise component should be introduced or the model should switch to a Class-B/C formulation. Cross-sensor consistency depends on reliable synchronization; when the clock error exceeds the structural propagation delay, the mechanical-impact protection mechanism is weakened.

Engineering deployment should retain raw data, parameter-estimation results, weights, and uncertainty logs, and report performance in a stratified manner according to transmission line, season, and load. For high-uncertainty samples, autonomous defect-elimination actions should not be executed directly; instead, deceleration, re-imaging, adjustment of the contact position, or manual confirmation should be triggered. In the future, dynamic graph weights can be linked with robot motion planning to form a closed loop of “perception uncertainty–active resampling–safe decision-making”. Cross-line external validation and long-term drift monitoring should also be conducted to avoid domain bias caused by data from a single transmission line.

To examine training randomness, five random seeds were provisionally used for repeated training. The matching accuracy, ADR, Macro-F1, and ECE of PA-STGFN were $(95.8 \pm 0.31)\%$, $(96.5 \pm 0.27)\%$, $(95.1 \pm 0.34)\%$, and $(2.8 \pm 0.22)\%$, respectively. Using test sections as the resampling unit, 1,000 paired Bootstrap iterations were performed. The 95% confidence interval for the Macro-F1 gain of PA-STGFN relative to the fixed graph was [1.1, 2.2] percentage points and did not cross 0. This statistical field likewise requires recalculation using the actual measured data, and

TABLE VI. Uncertainty Estimation and Computational Overhead

Method	ECE/%	Error-Detection AUROC/%	Parameters/M	Number of Inferences	Latency/ms
Softmax late fusion	8.9	78.2	9.8	1	70
MC-dropout (10 runs)	4.8	86.7	9.8	10	412
Deep ensemble (5 models)	3.6	90.5	49.0	5	338
Evidential quantification fixed graph	4.1	88.9	10.9	1	80
PA-STGFN	2.8	92.4	11.2	1	78

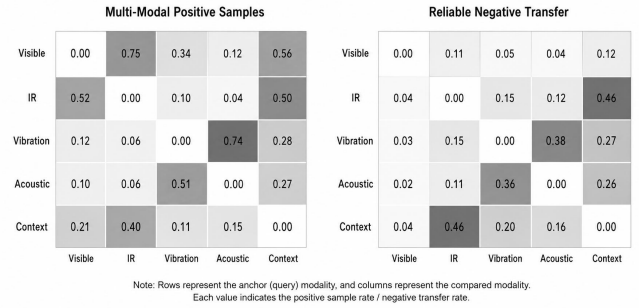

FIGURE 7. Online tracking results of Middleton Class-A parameters.

the final manuscript should report the number of sections, random seeds, and the script used to calculate the confidence intervals.

Table VII shows that excessive sparsity can disconnect complementary relationships such as thermal–electrical and vibration–acoustic associations, whereas full connectivity amplifies residual noise propagation. Top-3 provides a favourable balance among accuracy, calibration, and latency. When the fusion temperature τ ranged from 0.6 to 1.0, Macro-F1 varied by less than 0.4 percentage points; $\tau < 0.4$ caused the weights to concentrate prematurely, whereas $\tau > 1.2$ approached equal weighting. When the Class-A truncation order L increased from 8 to 16, the result changed by less than 0.1 percentage points, indicating that truncation error had a limited effect on the final performance.

Classification metrics can only indirectly reflect preprocessing effectiveness. Therefore, the Class-A front end was further evaluated in terms of four dimensions: noise-parameter estimation error, output SNR gain, fault-band energy retention rate, and residual kurtosis. Test windows were stratified by A , Γ , and SNR, and the predefined parameters were used only to synthesize interference and were not accessible to the estimator. As shown in Fig. 7, when A changed stepwise among 0.03, 0.10, and 0.30, and Γ switched among 0.01, 0.10, and 0.03, the exponentially smoothed estimates converged within 2–3 windows without persistent threshold oscillation.

Table VIII shows that although median filtering and fixed clipping reduce kurtosis, they also remove some looseness-related impacts and high-frequency discharge components. Class-A adaptive preprocessing achieved an output SNR gain of 8.4 dB while retaining 94.7% of the fault-band energy. The residual kurtosis after preprocessing remained above the ideal Gaussian value of 3, indicating that the algorithm did not forcibly remove all non-Gaussian structures


FIGURE 8. Dynamic-graph adjacency matrices under normal and visible-light occlusion conditions.

but retained transients consistent with faults. The relative parameter-estimation error was below 8.5% when $A \geq 0.03$. When $A = 0.01$, the error increased to 13.2% owing to the insufficient number of effective impulses, but the downstream uncertainty mechanism could still compensate for the estimation risk.

To examine whether the dynamic graph learned physically reasonable relationships, the occurrence frequencies of Top-3 edges for different fault classes in the test set were analysed. Under normal conditions, the state node maintained stable connections with the visible-light and infrared nodes. During bolt looseness, the vibration–acoustic edge was markedly enhanced. During local overheating, the infrared–state edge dominated. During abnormal discharge, the acoustic–state and visible-light–acoustic edges were simultaneously strengthened. These results are consistent with the fault mechanisms, indicating that topology changes were not unconstrained attention drift. The dominant graph edges and their average weights for representative fault classes are summarized in Table IX.

Fig. 8 compares the adjacency matrices under normal and visible-light occlusion windows. After occlusion occurred, the outgoing edge weights from the visible-light node decreased overall, the state $\rightarrow$ infrared edge increased from 0.39 to 0.55, and the vibration–acoustic interconnection remained largely unchanged, indicating that local sensor failure did not disrupt the unrelated mechanical evidence chain. When the physical prior term b_{ij}^{phy} was further set to zero, the cross-random-seed consistency rate of the dominant edges decreased from 87.6% to 71.3%, while Macro-F1 decreased by 0.6 percentage points, indicating that the physical prior simultaneously improved interpretability stability and recognition performance.

TABLE VII. Sensitivity Analysis of Dynamic-Graph Sparsity

Retained Edges per Node	Macro-F1/%	ECE/%	Latency/ms	Description
Top-1	94.2	4.0	74	Insufficient interaction
Top-2	94.8	3.3	76	Balances local relationships
Top-3	95.1	2.8	78	Setting used in this study
Top-4	95.0	2.9	81	Increased redundant edges
Fully connected	94.7	3.5	85	Enhanced noise propagation

TABLE VIII. Signal Fidelity and Interference-Suppression Performance of Different Preprocessing Methods

Method	Output SNR Gain/dB	Fault-Band Retention/%	Residual Kurtosis	Matching Accuracy/%
No processing	0.0	100.0	18.6	90.7
Median filtering	3.1	84.2	7.9	91.8
Fixed-threshold clipping	5.6	88.5	5.1	92.6
Gaussian adaptive denoising	6.2	90.1	4.7	93.0
Class-A adaptive preprocessing	8.4	94.7	3.4	95.8

TABLE IX. Dominant Graph Edges and Average Weights for Typical Classes

Class	Dominant Edge 1	Weight	Dominant Edge 2	Weight	Physical Meaning
Normal	State $\rightarrow$ visible light	0.35	State $\rightarrow$ infrared	0.39	Operating-condition constraints on image interpretation
Bolt looseness	Vibration $\rightarrow$ acoustic	0.57	State $\rightarrow$ vibration	0.41	Structural impact and speed correction
Local overheating	State $\rightarrow$ infrared	0.64	Infrared $\rightarrow$ visible light	0.48	Load-temperature-rise association
Abnormal discharge	Acoustic $\rightarrow$ state	0.59	Visible light $\rightarrow$ acoustic	0.52	Acoustic-electrical-optical consistency

E. DATA ORGANIZATION AND EXPERIMENTAL SETTINGS

Retrospective analysis of false-detection samples revealed three major failure types:

- 1) contamination on the insulator surface formed elongated bright edges under backlighting, resembling crack textures;
- 2) low-frequency impacts caused by robot acceleration and deceleration were similar to the envelopes of early-stage bolt looseness;
- 3) sustained arcs simultaneously contaminated acoustic, field-strength, and visible-light signals, causing multiple modalities to exhibit consistent but non-independent strong evidence.

The first two types of errors can be alleviated by incorporating polarized images, speed-conditioned normalization, and finer-grained component priors. The third type exceeds the short-duration Class-A independent-impulse assumption and should be handled by a dedicated arc-state machine.

For engineering applications, PA-STGFN is recommended as the perception layer within a hierarchical safety architecture rather than as a direct replacement for operational interlocking. Under low uncertainty and consistent results across multiple consecutive windows, automatic localization can be permitted. Under moderate uncertainty, the robot should decelerate and actively reacquire data. Under high uncertainty, failure of more than two modalities, or detection of a sustained arc, autonomous defect elimination should be terminated immediately. Model upgrades should use fixed test sections, noise-parameter grids, and sensor-

degradation sets for regression testing to ensure that improvements in accuracy are not achieved at the expense of calibration and safe-degradation capability.

VI. CONCLUSION

This study proposes PA-STGFN to address multi-source perception distortion and the tendency of fixed fusion to induce negative transfer in defect-elimination robots operating under strong electromagnetic environments. The method uses the Middleton Class-A model to characterize non-Gaussian impulsive electromagnetic noise and achieves adaptive interference suppression through online parameter estimation, posterior-controlled soft suppression, and cross-modal consistency protection. CA-EfficientNet and MS-TCN are used to extract visual spatial details and multi-scale temporal features, respectively. Dirichlet evidence is used to quantify confidence variance, with uncertainty explicitly incorporated into dynamic graph edges and modality weights. Focal loss, reconstruction loss, and uncertainty regularization are combined to achieve learning under sparse fault samples.

The experimental results show that PA-STGFN achieves a feature matching accuracy of 95.8% and an anomaly detection rate of 96.5%. When the SNR decreases to 10 dB, the matching accuracy decreases by only 2.7 percentage points, satisfying the requirement that performance degradation remain below 3%. The single-frame latency on the edge platform is 78 ms. Results from sensor degradation, ablation studies, parameter sensitivity, and confidence calibration jointly demonstrate that physical preprocessing is responsi-

ble for compressing heavy-tailed input risks, the uncertainty-aware dynamic graph blocks residual noise propagation, and dynamic confidence fusion redistributes information sources when modalities fail, forming a complementary closed loop.

The current method remains constrained by short-window stationarity, clock synchronization, and the requirement that at least two modalities remain valid. Future work will conduct external validation across seasons, transmission lines, and sensor-aging conditions; introduce composite physical models for sustained arcs and periodic narrowband interference; and jointly optimize modality uncertainty with active robot viewpoint acquisition, motion planning, and operational interlocking, thereby forming a verifiable closed loop of “perception–resampling–decision–execution”.

DECLARATIONS

AVAILABILITY OF DATA AND MATERIAL

Due to the need for confidentiality on the part of the article’s funders, the data presented in this study are available upon request from the corresponding author.

COMPETING INTERESTS

The authors declare no conflict of interest.

AUTHORS’ CONTRIBUTIONS

Peng Lei: Conceptualization, Methodology, Formal analysis, Investigation, Writing—original draft, Visualization. Jiabin Liang: Supervision, Project administration, Funding acquisition, Writing—review and editing. Tianlong Zhang: Data curation, Investigation, Validation, Software. Shuo Wang: Software, Validation, Formal analysis, Visualization. Jianling Wu: Investigation, Resources, Data curation, Writing—review and editing. Sen Wang: Validation, Resources, Project administration. All authors reviewed and approved the final manuscript.

ACKNOWLEDGMENT

The authors acknowledge the 2025 Science and Technology Project of Inner Mongolia Power (Group) Co., Ltd. 2025-4-7.

REFERENCES

- [1] A. B. Alhassan et al., “Power transmission line inspection robots: A review, trends and challenges for future research,” *Int. J. Electr. Power Energy Syst.*, vol. 118, 2020, Art. no. 105862, doi: 10.1016/j.ijepes.2020.105862.
- [2] A. Pagnano, M. Höpf, and R. Teti, “A roadmap for automated power line inspection, maintenance and repair,” *Procedia CIRP*, vol. 12, pp. 234–239, 2013, doi: 10.1016/j.procir.2013.09.041.
- [3] X. Liu, X. Miao, H. Jiang, et al., “Data analysis in visual power line inspection: An in-depth review of deep learning for component detection and fault diagnosis,” *Annu. Rev. Control*, vol. 50, pp. 253–277, 2020, doi: 10.1016/j.arcontrol.2020.09.002.
- [4] D. Middleton, “Statistical-physical models of electromagnetic interference,” *IEEE Trans. Electromagn. Compat.*, vol. EMC-19, no. 3, pp. 106–127, Aug. 1977, doi: 10.1109/TEMC.1977.303527.
- [5] S. V. Zhidkov, “Analysis and comparison of several simple impulsive noise mitigation schemes for OFDM receivers,” *IEEE Trans. Commun.*, vol. 56, no. 1, pp. 5–9, Jan. 2008, doi: 10.1109/TCOMM.2008.050391.
- [6] H. Oh, H. Nam, and S. Park, “Adaptive threshold blanker in an impulsive noise environment,” *IEEE Trans. Electromagn. Compat.*, vol. 56, no. 5, pp. 1045–1052, Oct. 2014, doi: 10.1109/TEMC.2014.2311853.
- [7] P. J. Huber, “Robust estimation of a location parameter,” *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964, doi: 10.1214/aoms/1177073732.
- [8] F. R. Hampel, “The influence curve and its role in robust estimation,” *J. Amer. Statist. Assoc.*, vol. 69, no. 346, pp. 383–393, 1974, doi: 10.1080/01621459.1974.10482962.
- [9] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.
- [10] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13713–13722.
- [11] Y. Abu Farha and J. Gall, “MS-TCN: Multi-stage temporal convolutional network for action segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3575–3584.
- [12] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7444–7452, doi: 10.1609/aaai.v32i1.12328.
- [13] P. Veličković, G. Cucurull, A. Casanova, et al., “Graph attention networks,” in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [14] W. Peng, X. Hong, H. Chen, et al., “Learning graph convolutional network for skeleton-based human action recognition by neural searching,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 3, pp. 2669–2676, doi: 10.1609/aaai.v34i03.5652.
- [15] T. Baltrušaitis, C. Ahuja, and L. P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.
- [16] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017.
- [17] M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, vol. 31, pp. 3179–3189.
- [18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, vol. 30, pp. 6402–6413.
- [19] C. Guo, G. Pleiss, Y. Sun, et al., “On calibration of modern neural networks,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1321–1330.
- [20] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, vol. 30, pp. 5574–5584.
- [21] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
- [22] S. Woo, J. Park, J. Y. Lee, et al., “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
- [23] C. Lea, M. D. Flynn, R. Vidal, et al., “Temporal convolutional networks for action segmentation and detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 156–165.
- [24] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7482–7491.
- [25] T. Y. Lin, P. Goyal, R. Girshick, et al., “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.
- [26] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.