Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 34(9): 6846–6876.
DOI: 10.3934/era.2026298
Received: 18 June 2026
Revised: 08 July 2026
Accepted: 28 July 2026
Published: 12 August 2026

Research article

Real-time detection method for photovoltaic electroluminescence images targeting weak and multi-scale defects

Xiang Chen[1,2,3,*], Liming Sun[4], Haiyang Sun[4] and Yelin Deng[2]

[1] School of Big Data, LiaoNing Finance and Trade College, 167 Xinghai North Street, Xingcheng 125105, China

[2] School of Rail Transportation, Soochow University, 8 Jixue Road, Suzhou 215137, China

[3] School of Mechanical Engineering, Nantong University, 9 Seyuan Road, Nantong 226019, China

[4] School of Computer Science and Engineering, Northeastern University, 195 Chuangxin Road, Shenyang 110169, China

* **Correspondence:** Email: xiang33.Chen@ntu.edu.cn.

Abstract: Defect detection in photovoltaic (PV) modules is crucial for long-term reliable operation of solar plants, reducing levelized cost of energy, and preventing safety incidents. Electroluminescence (EL) images suffer from low contrast, complex background noise, and frequent overlapping defects, hindering the balance between detection accuracy and real-time processing speed for current algorithms. To address this, a novel real-time object detector is proposed. First, a convolutional additive self-attention block (CAS-Block) reconstructs the backbone by replacing dot-product with additive similarity, enhancing the response sensitivity to weak defect features in low-contrast images while maintaining linear computational complexity. Second, an efficient dynami-scale intra-feature interaction (EDII) module based on frequency-domain learning employs adaptive spectral filtering to suppress environmental noise interference, such as illumination variations. Finally, a gated cross-scale feature fusion (GCFF) module built on gated linear unit (GLU) uses a dynamic gating mechanism to precisely disentangle feature confusion arising from overlapping defects. Experimental results show that the proposed model achieves a mean average precision (mAP)@0.5 of 93.2% and a detection speed of 104.2 frames per second (FPS), outperforming the state-of-the-art (SOTA) YOLOv12-L and EER-DETR. The method achieves an optimal accuracy–speed trade-off, offering an efficient, cost-effective solution for intelligent inspection of large-scale PV plants.

Keywords: photovoltaic cells; defect detection; electroluminescence images; deep learning; real-time object detection

1. Introduction

The global PV industry has stepped into the terawatt-scale era. High-altitude PV power plants feature exorbitant labor costs and harsh working environments, rendering traditional manual handheld inspection gradually obsolete. Although intelligent unmanned aerial vehicle inspection constitutes an effective alternative solution, utility-scale PV plants at megawatt and even gigawatt levels are equipped with millions of PV modules, thereby producing massive volumes of EL imaging data of PV cells. It should be clarified that field EL imaging requires the inspected PV strings to be forward-biased; in practice, such inspection is typically conducted at night or under low ambient illumination, where strings are electrically excited by external power supplies or grid-connected inverters while near-infrared cameras mounted on handheld devices, tripods, or unmanned aerial platforms capture the emitted luminescence. Against this backdrop, edge-deployed defect detection algorithms that enable real-time fault diagnosis and identification are poised to become the dominant technical paradigm for PV inspection.

EL imaging serves as a mainstream non-destructive testing technique for identifying defects in PV cells. By applying a forward bias, radiative carrier recombination is stimulated within PV cells, and the emitted near-infrared luminescence is captured to generate EL images, thereby enabling defect detection through subsequent image analysis [1]. EL images can intuitively reflect internal carrier dynamic behaviors and defect distributions [2], supporting the effective identification of cracks, finger interruptions, and other defects that impair the electrical performance of PV cells [3]. In recent years, deep learning methods have significantly advanced EL image-based defect detection by autonomously extracting discriminative features from large-scale EL image datasets to achieve accurate defect identification, classification, and localization [4, 5]. Among these, CNN-based methodologies still dominate the field, primarily leveraging advanced object detection frameworks [6–8]. Notably, existing studies have validated the performance of representative CNN models, including AlexNet [9], VGG16 [10], ResNet [11], and MobileNetV2 [12], for defect detection in PV cells [13]. Furthermore, lightweight CNN architectures such as the MobileNet series [14] and improved U-Net models [15] have also been explored and applied for efficient defect detection in PV cell EL images.

Despite their strong feature extraction capability, conventional CNN-based models exhibit inherent limitations in characterizing sparsely distributed and ambiguous defect features within PV cell EL images, which frequently result in missed detections [16]. These drawbacks accordingly highlight the applicability and research potential of Transformer-based architecture [17]. The detection Transformer (DETR) was first proposed in 2020 and has rapidly emerged as a prominent research hotspot for PV cell defect detection tasks [18]. In contrast to classical CNNs, which depend on local receptive fields and are inherently incapable of capturing long-range spatial dependencies, DETR adopts a global self-attention mechanism. This mechanism enables effective modeling of spatially dispersed PV defects, including inter-cell cracks and vertical dislocations, thereby substantially alleviating the missed detection issues prevalent in CNN-based approaches.

Current research efforts are centered on the advancement and optimization of DETR-series models, the integration of Transformer architectures with mainstream object detection algorithms (e.g., YOLO), and task-specific model customization, thereby establishing a diversified technological roadmap. A variety of tailored improvement strategies have been reported in the literature. Liu Z. et al. introduced the Swin Transformer, which reduces the computational overhead of conventional

vision Transformers through shifted window operations, thus enhancing real-time inference capability [19]. In 2023, DINO-DETR, equipped with contrastive denoising training and mixed query selection, achieved superior detection accuracy and faster convergence, outperforming CNN-based detectors on the common objects in context benchmark [20]. Baidu's RT-DETR, the first real-time Transformer detector, provides an effective trade-off between speed and accuracy, making it highly suitable for defect detection in PV EL images [21]. More recently, Dun J. et al. proposed EER-DETR, which achieved better detection performance than RT-DETR for PV defect inspection, demonstrating strong practical engineering application potential [22].

Although SOTA Transformer-based visual models have exhibited distinct advantages over conventional CNN-based approaches, existing research still encounters multiple critical challenges in EL image-based PV cells defect detection, owing to the inherent properties of EL imagery [23]. First, partial defect types exhibit tiny pixel scales and low visual contrast, which stem from the intrinsic constraints of EL imaging principles. Such defects are readily masked by background textures including electrode gridlines [24], resulting in insufficient discriminability between defect regions and the background. This issue not only impedes models from eliminating redundant information and extracting representative discriminative semantic features of defects efficiently, but also elevates the probability of missed detection, particularly for small-scale defect targets. Second, EL images acquired by near-infrared imaging systems are vulnerable to random disturbances originating from the imaging environment and acquisition devices, accompanied by illumination fluctuations, image noise, and PV cell surface artifacts [16, 25]. Unlike intrinsic structural characteristics of PV cells, such unstructured stochastic noise deteriorates image quality and directly degrades the accuracy of defect recognition. Third, multiple defects frequently coexist on a single PV cell. As several defects possess correlated generation mechanisms, their morphological features tend to overlap or exhibit derivative relationships [26]. Mutual interference among heterogeneous defect features greatly complicates model discrimination between primary defects and their derived defects, thereby further deteriorating overall detection accuracy.

To tackle the challenges, this paper proposes a dedicated object detection algorithm termed PVEL-DETR for PV cell EL images defect inspection. The proposed method focuses on three core research directions: strengthening the model's feature extraction capability, enhancing its discriminative ability toward multi-scale defect features, and optimizing intra-scale feature interaction. The objective is to effectively improve the model's capability in capturing fine-grained defect features and its robustness against complex background interference. Meanwhile, by satisfying the lightweight constraints for real-time edge deployment, PVEL-DETR further elevates detection accuracy for practical PV cell EL image defect detection scenarios. The main contributions of this work are summarized as follows:

(1) A novel CAS-Block is proposed. By employing additive fusion operations, this module reconstructs the feature extraction pathway upon the original backbone network. By introducing an additive similarity function and a dynamic gating mechanism, it overcomes the single and insufficient feature representation of conventional residual blocks in low-contrast scenarios, enlarges the effective receptive field, and achieves efficient collaborative modulation of spatial and channel features while maintaining linear computational complexity, thereby suppressing background noise, amplifying weak defect responses, and alleviating the missed detection problem induced by blurred and redundant defect features.

(2) An EDII module is introduced for intra-scale interaction of high-dimensional features.

Specifically, an efficient dynamic feed-forward network (EDFFN), which exploits the global feature-decoupling capability of frequency-domain transforms to replace the computation-intensive spatial fully-connected operations of the conventional feed-forward network, is integrated into the original AIFI module. This improvement substantially enhances the model's robustness against interferences including illumination variations, while strengthening the feature stability for small and weak defects at a reduced computational cost.

(3) A GCFF structure is designed. This architecture enhances the modeling accuracy of local defect features and reduces the probability of confusion caused by overlapping multiple defects. Through a feature-splitting strategy, it also minimizes redundant computations, ensuring that the cross-scale feature fusion process does not compromise the model's real-time inference speed.

2. Problem definition

As a non-destructive inspection technique, EL imaging relies on the physical principle that forward bias injection of minority carriers in PV cells induces subsequent radiative carrier recombination. Mathematically, the intensity of an EL image $I(x, y)$ can be described as a mapping function of the local carrier recombination efficiency $\eta(x, y)$ and the voltage distribution $V(x, y)$:

$$I(x, y) \propto \eta(x, y) \cdot \exp\left(\frac{qV(x, y)}{kT}\right). \tag{2.1}$$

Here, q denotes the electron charge, k is the Boltzmann constant, and T is the absolute temperature. Defects appear in EL images as an abrupt reduction in local luminescence intensity $I(x, y)$ accompanied by distinct textural disturbances. As depicted in Figure 1, the observed image I_{obs} constitutes a severely ill-posed inverse problem, which is affected by multiple factors including grain boundary noise in polycrystalline silicon, non-uniform doping of the PV cell, and thermal noise generated by the imaging sensor:

$$I_{\text{obs}} = H(I_{\text{true}}) + N(\mu, \sigma^2), \tag{2.2}$$

where H denotes the degradation operator and N represents additive noise. Conventional computer vision methods encounter severe mathematical difficulties when processing such images, which feature a low signal-to-noise ratio, weak contrast, and multi-scale coupling of defect features. These challenges mainly involve inadequate feature representation extraction and the difficulty of suppressing background interference, exactly the core issues targeted by the model design methodology proposed in this paper.

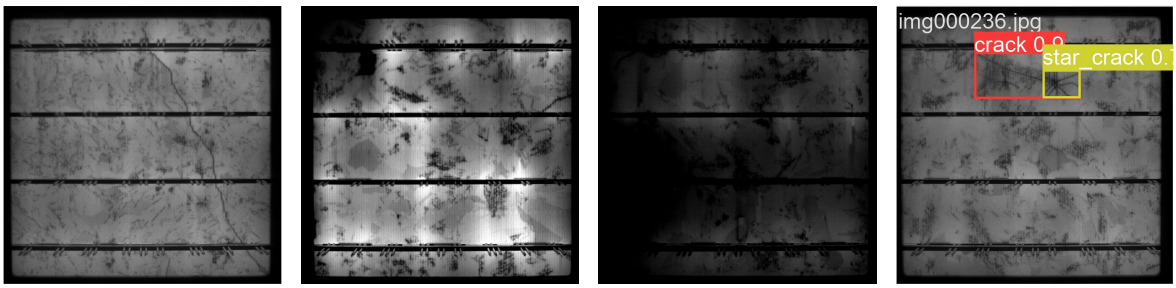

Figure 1. EL images of PV cells captured by a near-infrared imaging device.

As illustrated in Figure 2, a total of 3600 PV cell EL images were randomly selected to conduct a comprehensive statistical analysis covering all defect samples in the dataset. The statistics quantitatively characterize the instance quantity, spatial distribution, and dimensional characteristics of each defect category. Statistical results indicate that the spatial positions of defects are widely distributed over the entire cell surface, whereas the instance quantities of different defect categories are markedly imbalanced. This imbalance faithfully reflects the natural occurrence frequencies of different defect types in real industrial production lines, and its influence on the experimental design is discussed in detail in Section 4.2. Moreover, the results reveal that most defects belong to small-scale targets, which highlights the critical necessity of robust small-defect detection. Accordingly, improving the model's fine-grained feature learning capability is particularly vital for PV cell EL image defect detection tasks.

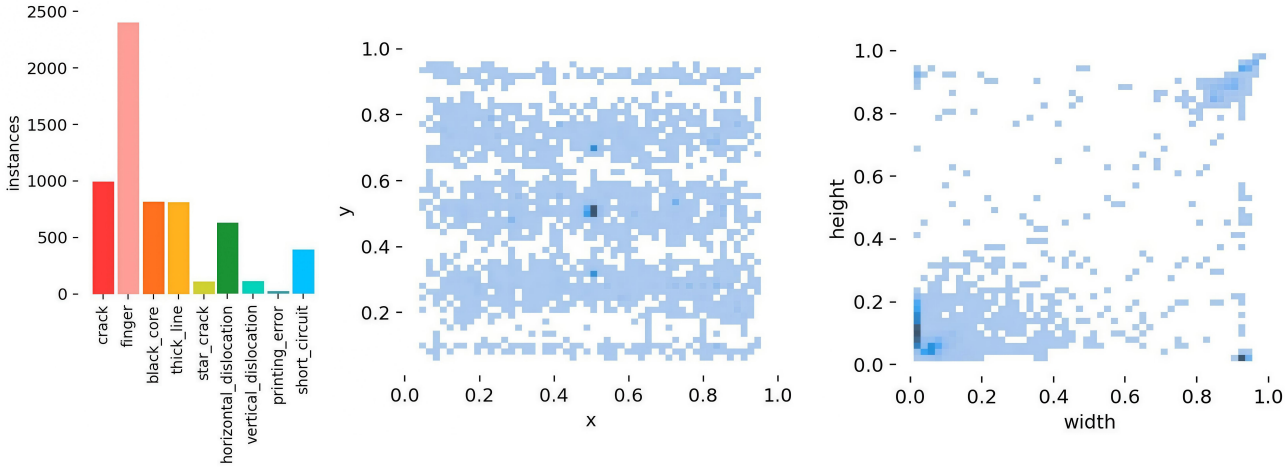

Figure 2. Statistical analysis results of defect instances in EL images of PV cells.

3. The architecture of the network

To tackle the practical challenges, the proposed PVEL-DETR algorithm incorporates targeted optimization designs for its core functional modules. These designs are specifically tailored to the intrinsic properties of PV cell EL images, including low contrast, overlapping multi-defect regions, and high susceptibility to noise interference. Accordingly, a real-time detection framework dedicated to PV cell defect inspection is constructed, and its overall network architecture is illustrated in Figure 3. The PVEL-DETR algorithm processes images through four key stages. The first stage corresponds to feature extraction. After pixel normalization, a PV cell EL image with a resolution of 640×640 is fed into the backbone network to extract multi-level features. A ResNet-based network is adopted as the fundamental backbone, preserving its original four-stage hierarchical feature output structure. By virtue of the synergistic mechanism implemented in the convolutional additive token mixer, which integrates local convolution and additive attention-style gating, this stage effectively overcomes the limitations of conventional convolutions in capturing low-contrast defect features and aggregating contextual semantics beyond the constrained local receptive field. This stage

ultimately generates three critical-scale feature maps: P_3 ($160 \times 160 \times C_3$), P_4 ($80 \times 80 \times C_4$), and P_5 ($40 \times 40 \times C_5$), corresponding to different down sampling rates. Specifically, P_3 preserves rich shallow detailed characteristics, whereas P_5 encodes high-level deep semantic information. Collectively, these multi-scale features provide a comprehensive and discriminative foundation for subsequent cross-scale feature fusion.

The second stage corresponds to an efficient hybrid encoder, which acts as the core bridge connecting the feature extraction backbone and the decoder within the PVEL-DETR framework. This stage integrates two elaborately optimized sub-modules, which are respectively responsible for intra-scale dynamic feature interaction and multi-scale feature fusion. The first is the EDII module, which operates exclusively on high-level semantic features extracted from P_5. By employing a global self-attention mechanism, this module establishes long-range feature dependencies. Furthermore, a frequency-domain filtering mechanism is introduced to suppress interference caused by noise and illumination variations. The module outputs an enhanced image feature sequence, which is subsequently reshaped into a 2D spatial feature map denoted as F_5. The second is the GCFF module, which takes multi-scale features S_3 and S_4 from the backbone network as well as the refined F_5 feature from the EDII module as inputs. It utilizes gated convolution units to dynamically filter discriminative defect features at each scale, thereby alleviating feature confusion induced by overlapping defects. Meanwhile, a PANet-based cross-stage fusion strategy is adopted to guarantee effective information propagation and integration across different scales.

The third and fourth stages correspond to the query initialization scheme and the decoder & prediction head, respectively. The query initialization module adopts an uncertainty-minimization query selection strategy. It selects a fixed set of 300 representative feature vectors from the output sequence of the hybrid encoder to serve as learnable object queries. Each object query comprises two components: a content query derived from the selected feature vector and a position query corresponding to the predicted bounding box of the feature. By prioritizing features with low prediction uncertainty, this strategy ensures that the initialized queries possess high classification confidence and accurate localization ability, thus providing a reliable initialization for iterative optimization in the decoder. The decoder and prediction head is constructed using a 6-layer Transformer decoder, where each decoder layer is integrated with an auxiliary prediction head. The core operation of the decoder involves cross-attention between the object queries and the enhanced image features generated by the encoder, followed by feed-forward networks to strengthen nonlinear feature representation. The auxiliary heads generate intermediate defect predictions at each layer, which are jointly supervised by a multi-task loss function to guide the progressive optimization of object queries. Through such layer-wise iterative refinement, the network gradually regresses precise defect categories and bounding boxes.

Finally, the 300 refined object queries output by the decoder are directly matched with ground-truth annotations to generate the final defect classes and bounding box coordinates. This design eliminates the requirement for conventional non-maximum suppression, thus enabling a fully end-to-end detection pipeline. When applied to PV cell EL image defect detection, the proposed PVEL-DETR algorithm exhibits significant advantages in terms of feature extraction accuracy, multi-defect discrimination capability, and anti-interference robustness.

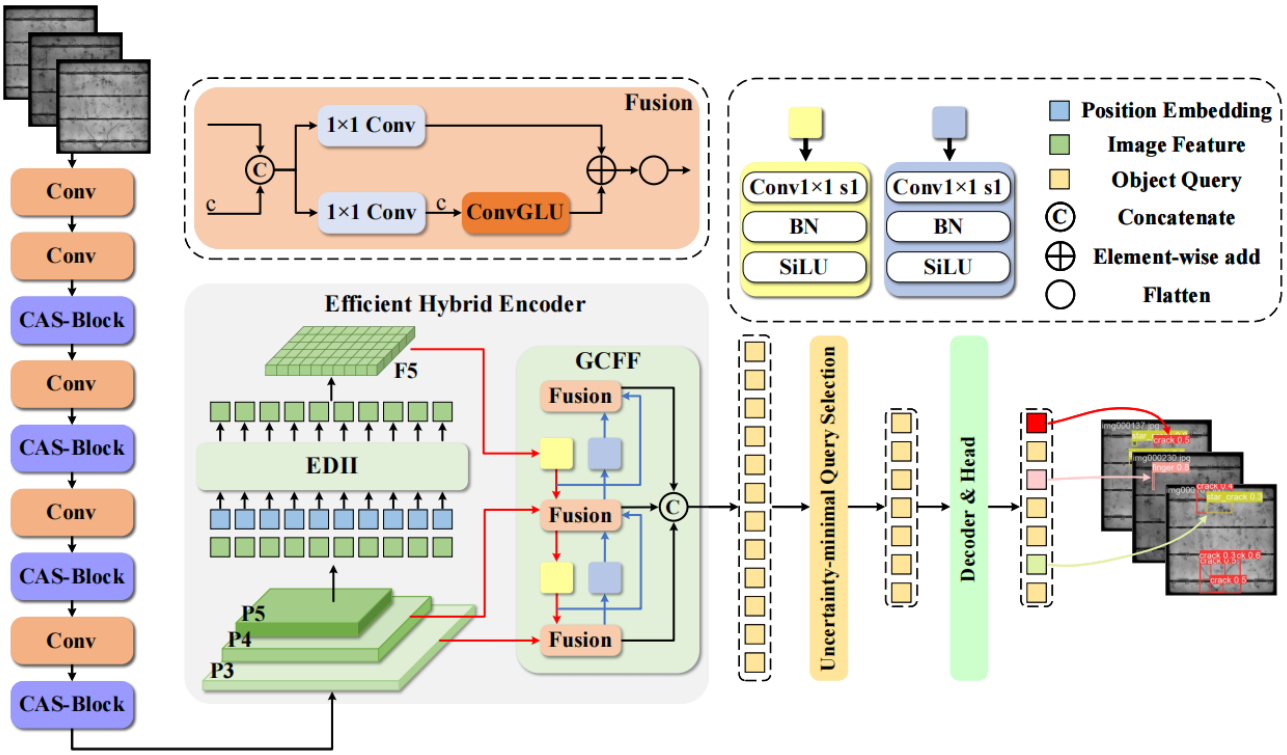

Figure 3. Structure diagram of the PVEL-DETR defect detection algorithm.

It is worth clarifying the relationship between the overall framework in Figure 3 and that of EER-DETR [22]. Since both PVEL-DETR and EER-DETR are constructed upon the RT-DETR paradigm, their top-level pipelines inevitably share the same four-stage organization, namely the backbone, the hybrid encoder, the uncertainty-minimal query selection, and the decoder with auxiliary prediction heads, which explains the apparent similarity of the structure diagrams. Nevertheless, the two methods differ fundamentally in the improved components and design motivations. EER-DETR enhances RT-DETR mainly from the perspective of lightweight re-parameterization: it introduces a wavelet-based diverse branch block (WDBB) for feature reconstruction of small targets and an enhanced hierarchical feature pyramid network (EHFPN) with dynamic weight allocation to reduce model complexity. In contrast, PVEL-DETR is designed around the imaging characteristics of PV cell EL images: (1) in the backbone, the proposed CAS-Block replaces the standard residual blocks with an additive-similarity token mixer tailored to low-contrast weak defects, whereas EER-DETR retains conventional convolutional feature extraction with re-parameterized branches; (2) in intra-scale interaction, PVEL-DETR introduces the frequency-domain EDII module to suppress illumination and noise interference, a component that has no counterpart in EER-DETR; and (3) in cross-scale fusion, PVEL-DETR employs the gated GCFF structure to disentangle overlapping defect features, while EER-DETR relies on weighted feature-pyramid aggregation. Consequently, although the two frameworks appear similar at the diagram level, their improved modules, design motivations, and targeted challenges are entirely different, which is also reflected by the performance comparison in Section 4.4.

3.1. Feature extraction module

The backbone of RT-DETR adopts standard residual blocks, which rely on 3×3 convolutions and residual connections to conduct feature extraction. Nevertheless, when processing PV cell EL images, the constrained local receptive field inherent in convolutional operations is inadequate to capture global semantic dependencies of faint defects within low-contrast scenarios, frequently giving rise to localization errors. Moreover, the exclusive dependence on spatial-domain information yields a single and insufficient feature representation, which is incapable of simultaneously modeling discriminative channel-wise characteristics. This further introduces severe feature redundancy, whereby background noise and irrelevant texture patterns occupy substantial feature dimensions and consequently prevent subsequent modules from concentrating on authentic defect features.

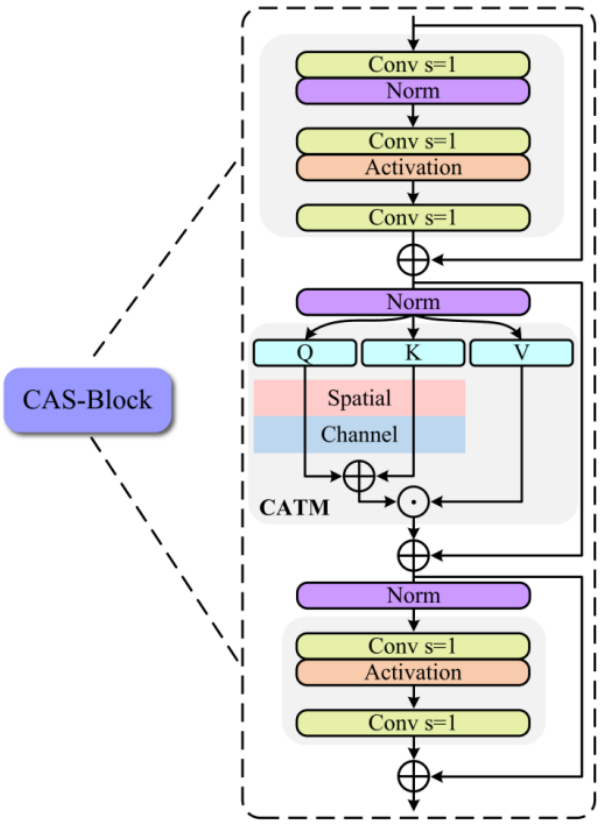

Figure 4. Convolutional additive self-attention block structure diagram.

To overcome the limitations, including the inadequate capture of low-contrast defect features and the constrained effective receptive field of conventional convolutions for contextual semantic aggregation, this paper designs a novel feature extraction unit termed the CAS-Block. The proposed block enables adaptive information interaction across both spatial and channel dimensions, thereby effectively improving the precision and efficiency of defect feature extraction. The design of the CAS-Block draws on the convolutional additive self-attention paradigm recently proposed in CAS-ViT [27], and its superiority over existing feature extraction units can be highlighted from three

aspects. First, compared with the standard residual block adopted in the RT-DETR backbone, which relies solely on fixed 3×3 convolutions and thus yields a single spatial feature representation, the CAS-Block jointly models spatial and channel saliency and adaptively gates informative regions. Second, compared with the standard multi-head self-attention in vision Transformers [17], whose computational complexity grows quadratically with the number of tokens, the additive similarity function maintains linear complexity and is therefore compatible with the real-time constraint of the detector. Third, compared with window-based attention such as the Swin Transformer [19], which confines interactions within fixed local windows, the additive token mixer aggregates contextual information through cascaded depthwise convolutions and sigmoid-gated modulation without hard window boundaries, making it better suited to weak, low-contrast EL defect features. The quantitative benefit of this design is verified by the ablation study in Section 4.5, where introducing the CAS-Block alone improves mAP@0.5 from 0.891 to 0.926. The detailed network architecture of the CAS-Block module is illustrated in Figure 4.

3.1.1. Integration subnet

The CAS-Block achieves efficient multi-dimensional information interaction while circumventing the excessive computational overhead imposed by conventional self-attention, which is realized via a three-stage pipeline: integration subnet, additive self-Attention, and non-linear enhancement. The integration subnet acts as a feature preprocessing unit that delivers refined feature inputs for subsequent attention-based interaction. This subnet is composed of a 1×1 pointwise convolutional layer for inter-channel interaction, followed by three stacked depthwise convolutional layers, each configured with a kernel size of 3×3 and a stride of 1. Given an input feature tensor $X_{\text{in}} \in \mathbb{R}^{H \times W \times C}$, where H and W are the spatial dimensions and C is the number of channels, the integration subnet utilizes a depthwise convolutional operator F_{dw} to perform channel-wise spatial filtering. The purpose is to map the original feature space into one where defect features are more readily separable. The overall computational flow of the integration subnet S_{int} is as follows:

$$X_{\text{int}} = S_{\text{int}}(X_{\text{in}}) = \sigma_2\Big(\text{BN}(F_{\text{dw}}^{3\times3}(F_{\text{dw}}^{3\times3}(F_{\text{dw}}^{3\times3}(\sigma_1(\text{BN}(F_{\text{conv}}^{1\times1}(X_{\text{in}}))))))\Big). \tag{3.1}$$

Here, $\sigma(\cdot)$ denotes the nonlinear activation function, BN represents the batch normalization operation, and $F_{\text{conv}}^{1\times1}$ stands for the 1×1 convolution. This pipeline first enables inter-channel information interaction via the 1×1 convolution and then performs local spatial feature extraction through three cascaded 3×3 depthwise convolutions, which progressively enlarge the local receptive field at a negligible parameter cost. After processing by the integration subnet, a preliminarily enhanced feature X_{int} with unchanged dimensions is output. By exploiting the local receptive field inherent in depthwise convolution, this step effectively suppresses background noise while enhancing discriminative local features, thereby establishing a cleaner and more representative feature basis for subsequent attention-based interaction.

3.1.2. Convolutional additive token mixer

The convolutional additive token mixer (CATM) functions as the core component of the CAS-Block. It replaces the dot-product similarity metric in conventional self-attention with an additive similarity function, which supports the joint modeling of spatial and channel attention to enlarge the

effective receptive field and realize discriminative feature selection. Within the CATM, the input feature X_{int} is projected into three distinct vectors (Query (Q), Key (K), and Value (V)) via three separate 1×1 convolutional layers:

$$Q = X_{\text{int}} W_Q, \quad K = X_{\text{int}} W_K, \quad V = X_{\text{int}} W_V. \tag{3.2}$$

Here, $W_Q, W_K, W_V \in \mathbb{R}^{C \times C}$ are learnable weight matrices. To capture local contextual information and enable feature filtering, we define a context mapping function $\Phi(Z)$, where $Z \in \{Q, K\}$. This function captures spatial context through a depthwise convolutional operator T_s and generates a spatial gate via a Sigmoid activation function. In the PVEL-DETR model, spatial attention is adopted as the primary mechanism, and its computation proceeds as follows:

$$\Phi(Z) = \text{Sigmoid}(W_{1\times1} * (W_{\text{dw}} * Z)) \odot Z. \tag{3.3}$$

Here, $\odot$ denotes the Hadamard product (i.e., element-wise multiplication). During inference, for a feature vector at position ij, if its local neighborhood structure matches the defect pattern learned by the convolutional kernel W_{dw}, the Sigmoid activation value at this position will approach 1, thereby allowing the feature to be preserved. Conversely, regions corresponding to background noise will be suppressed.

3.1.3. Gradient analysis of the additive similarity function

The traditional Transformer architecture uses dot product similarity $\text{Sim}(Q, K) = QK^\top$ with a computational complexity of $O(N^2C)$. CAS-Block proposes an additive similarity model:

$$\text{Sim}_{\text{add}}(Q, K) = \Phi(Q) \oplus \Phi(K), \tag{3.4}$$

where $\oplus$ represents element-wise addition, the final attention output O_{attn} is calculated as follows:

$$O_{\text{attn}} = \Gamma(\Phi(Q) + \Phi(K)) \odot V. \tag{3.5}$$

Here, Γ represents a linear transformation of a 1×1 convolution, which integrates the information of the query and the key.

Regarding the complexity analysis, since $\Phi(\cdot)$ only involves convolution and element-wise multiplication, its complexity is $O(N \cdot C)$. The addition operation and the final Hadamard product with V also have a complexity of $O(N \cdot C)$. Therefore, the total complexity is linear $O(N)$, which is much lower than the standard attention's $O(N^2)$.

Regarding the modeling behavior for weak features, the traditional dot-product attention couples the query and the key multiplicatively and normalizes the similarity scores with a global Softmax over all spatial tokens. In low-contrast EL images dominated by approximately isotropic background noise, the K values of the massive background tokens are numerically close to those of the sparse weak-defect tokens; the global Softmax then tends to produce over-smoothed attention weights, which dilutes the contribution of weak defect features to the aggregated output and correspondingly weakens the effective gradient signal $\partial L/\partial Q$ at weak-feature positions. In contrast, the additive similarity model proposed in this paper decouples the query and the key in the geometric space: the similarity is constructed by element-wise addition followed by a locally gated mapping, rather than by a globally normalized inner product. In the additive model, the gradient is calculated as follows:

$$\frac{\partial L}{\partial Q} = \frac{\partial L}{\partial O_{\text{attn}}} \cdot \left(\frac{\partial \Gamma}{\partial \Phi} \cdot \frac{\partial \Phi}{\partial Q} \right) \odot V. \tag{3.6}$$

Since Φ includes the sigmoid derivative and is modulated by the local neighborhood context rather than solely by the magnitude of K, a comparatively stable gradient can still be backpropagated through $\Phi'(Q)$ even in the low-contrast region of the PV cell where $K \approx 0$. It should be emphasized that this analysis does not claim the complete elimination of gradient attenuation: If the value tensor V is extremely weak or the sigmoid gate saturates, the gradients can still become small. Rather, Eq (3.6) indicates that the additive formulation avoids the dilution effect introduced by the global Softmax normalization and maintains relatively more stable gradient propagation in low-contrast (dark) regions, thereby improving the response sensitivity of the model to weak-feature defects. This theoretical tendency is empirically supported by the ablation results in Section 4.5.

3.2. Efficient dynamic attention-based intra-feature interaction

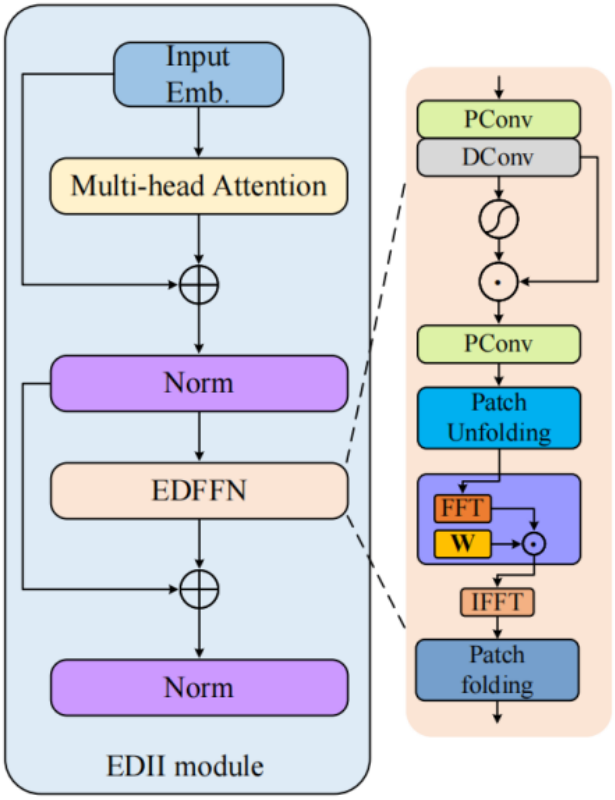

Figure 5. Network structure diagram of the EDII module.

The original AIFI module adopts a self-attention mechanism to implement intra-scale interaction for high-level semantic features (P_5). Its main function is to establish long-range global dependencies among high-dimensional semantic features, thereby supplying high-quality feature representations for subsequent query selection. Nevertheless, it exhibits obvious limitations when applied to PV cell EL images. Specifically, fully-connected layers are highly sensitive to noise; interference caused by illumination variations and other disturbances can induce shifts in feature distribution within the embedding space, thus degrading the accuracy of global dependency modeling. Moreover, during

feature interaction, the construction of global dependencies depends entirely on spatial-domain information while ignoring informative frequency-domain characteristics. This leads to inadequate capture of global structural relationships for faint and weak defects.

To overcome these drawbacks, an EDFFN is introduced to substitute the original feed-forward network (FFN), yielding the proposed efficient dynamic intra-scale feature interaction (EDII) module. Unlike the original AIFI module, whose FFN performs purely spatial-domain fully-connected transformations, the EDFFN conducts feature transformation in the frequency domain, and its superiority over existing frequency-domain designs lies in its dynamic and localized nature. Compared with GFNet [28], which learns static global spectral filters shared by all input images and therefore cannot adapt to sample-specific interference, the EDFFN generates an input-conditioned spectral gate from the global context vector, enabling adaptive suppression of scene-specific disturbances such as illumination variations. Compared with the frequency-domain feed-forward design adopted in FFTformer [29] for image deblurring, the EDFFN inherits the patch-wise FFT formulation but couples it with a global-context-driven dynamic gating branch, and applies it for the first time to intra-scale feature interaction in a real-time detection Transformer for PV cell EL images. The detailed structure of the proposed module is illustrated in Figure 5.

3.2.1. Frequency-domain transform and spectral distribution of defect features

PV cell EL images exhibit distinctive frequency-domain characteristics. Normal gridlines manifest as periodic high-frequency components, while the background corresponds to low-frequency components. Defects, in contrast, appear as non-periodic high-frequency mutations. The EDFFN module utilizes a two-dimensional discrete fourier transform (2D DFT) to convert spatial features into the frequency domain.

We employ the standard 2D DFT algorithm to transform the spatial feature $X \in \mathbb{R}^{H\times W\times C}$ into its frequency-domain representation $F(u, v, c)$, where u, v denotes the frequency coordinates and c is the channel index. This transformation effectively separates low-frequency illumination variations from high-frequency defect features that are otherwise entangled in the spatial domain.

3.2.2. Patch unfolding and local spectrum analysis

To reduce computational cost while capturing local frequency characteristics, the EDII module first partitions the input features into patches. Let the patch size be $p \times p$; the feature map is reshaped into a tensor $X_{\text{patch}} \in \mathbb{R}^{M\times p\times p\times C}$, where $M = \frac{HW}{p^2}$. A local FFT operation is then applied to each patch:

$$P_{\text{fft}}^{(m)} = \text{FFT2D}(X_{\text{patch}}^{(m)}) \in \mathbb{C}^{p\times p\times C}. \tag{3.7}$$

This operation thereby preserves both the spatial localization information indexed by m and the frequency information within the $p \times p$ spectrum of each patch. In our implementation, the patch size is set to $p = 8$ and the patches are non-overlapping, with the default FFT normalization of PyTorch adopted. Since non-overlapping patch unfolding and folding constitute an exact partition and reconstruction of the feature map, no windowing operation is required to suppress boundary artifacts, and the residual connection in Eq (3.10) further smooths potential discontinuities across patch borders. All remaining hyperparameters are kept identical to those of the original RT-DETR.

3.2.3. Dynamic frequency-domain filtering

The key to dynamic frequency-domain filtering lies in the construction of a dynamic gating matrix W_{gate}. This matrix must be dynamically generated based on the global statistical characteristics of the input features, enabling the network to adapt to defect detection tasks under varying interference conditions, such as changes in illumination.

Let the global context vector be $g = \text{GlobalAvgPool}(X_{\text{patch}})$. The dynamic weight W_{gate} is generated by a lightweight MLP:

$$W_{\text{gate}} = \text{Reshape}\Big(\sigma(W_2 \cdot \text{ReLU}(W_1 \cdot g))\Big) \in \mathbb{R}^{p\times p\times C}. \tag{3.8}$$

The filtering operation is performed as the Hadamard product between the frequency-domain features and the weighting matrix. PVEL-DETR employs a real-valued spectral gating scheme applied to the complex-valued spectrum, which is formulated as follows:

$$\widetilde{P}^{(m)}_{\text{fft}} = P^{(m)}_{\text{fft}} \odot W_{\text{gate}}. \tag{3.9}$$

The physical significance of this step is to construct an adaptive band-pass filter. Specifically, W_{gate} learns to suppress isotropic high-frequency components corresponding to background noise and low-frequency components associated with illumination non-uniformity, while simultaneously preserving the specific directional high-frequency components that represent defect features. It is worth clarifying the mathematical properties of this filtering scheme. The dynamic gating matrix generated by Eq (3.8) is a real-valued spectral mask rather than a complex-valued weight: when it multiplies the complex FFT coefficients in Eq (3.9), the real part and the imaginary part of each coefficient are scaled by the same real factor. Consequently, the phase of every frequency component is preserved, and the Hermitian symmetry of the spectrum of a real-valued feature map is not violated. In implementation, the forward transform is computed by the real FFT (`torch.fft.rfft2`), which exploits the Hermitian symmetry of real-valued inputs and stores only the non-redundant half of the spectrum, and the inverse transform is computed by `torch.fft.irfft2`, which guarantees a strictly real-valued output tensor.

3.2.4. Inverse transform and feature reconstruction

The filtered spectrum is transformed back to the spatial domain via the inverse FFT, and the original dimensions are restored through a patch folding operation, yielding $X_{\text{recon}} \in \mathbb{R}^{H\times W\times C}$. Finally, the output feature is obtained by adding the result to the original input via a residual connection:

$$X_{\text{out}} = X_{\text{in}} + \text{Conv}_{1\times 1}(X_{\text{recon}}). \tag{3.10}$$

The resulting feature X_{out} contains high-quality semantic information processed in the frequency domain. It will serve as the output of the encoder, providing high-quality initial queries for the decoder.

3.3. GCFF module

The basic module enhances feature representation during the training phase via structural reparameterization, thereby improving inference efficiency. Nevertheless, it still presents distinct limitations in PV cell defect detection scenarios. Specifically, the straightforward stacking of convolutions and reparameterization may readily give rise to severe feature confusion among

overlapping defects. To tackle this problem, a gated convolution unit (ConvGLU) is incorporated into the proposed GCFF module. The superiority of the GCFF structure over existing fusion designs can be clarified by explicit comparison. Compared with the RepBlock-based CCFF fusion unit in the original RT-DETR, which aggregates adjacent-scale features through static convolutions and structural re-parameterization, the GCFF introduces an input-dependent gating path derived from the gated linear unit (GLU) family [30], so that the contribution of each scale is dynamically modulated by the feature content instead of being fixed after training. Compared with classical cross-scale structures such as PANet-style aggregation, which sum or concatenate features indiscriminately and thus tend to blend overlapping defect responses, the gating mask in GCFF selectively passes discriminative defect features while suppressing redundant background components, which is the key to disentangling overlapping defects. Moreover, by adopting the cross-stage-partial splitting strategy [31], only part of the channels are processed by the gated unit, so the dynamic-selection capability is obtained at a reduced computational cost rather than at the expense of inference speed. The detailed structure of the GCFF module is illustrated in Figure 6.

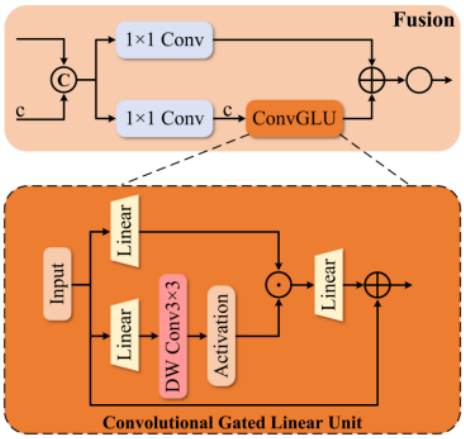

Figure 6. The improved network structure diagram of the fusion module.

3.3.1. Convolutional gated linear unit

The GCFF module optimizes the fusion unit by leveraging the gating mechanism of ConvGLU to achieve dynamic filtering and precise integration of multi-defect features. The specific workflow is as follows.

The inputs to GCFF are adjacent-scale features from different stages of the backbone and the output X_{out} from the EDII module after processing F5. The standard GLU is $\text{GLU}(x) = (xW + b) \odot \sigma(xV + c)$. In GCFF, this formulation is extended to a convolutional version to effectively handle spatially structured features.

Given the input feature F_{in}, it is split into two parallel convolutional pathways: the content branch and the gate branch. The content branch is responsible for extracting feature information, while the gate branch generates a feature-selection mask. The corresponding computational expressions are as follows:

$$V = F_{\text{val}}(F_{\text{in}}) = W_v * F_{\text{in}} + b_v, \tag{3.11}$$

$$G = F_{\text{gate}}(F_{\text{in}}) = W_g * F_{\text{in}} + b_g. \tag{3.12}$$

Typically, W_g uses depthwise convolution to reduce the number of parameters and expand the receptive field.

The final output expression of ConvGLU is:

$$F_{\text{glu}} = \text{ConvGLU}(F_{\text{in}}) = V \odot \text{Sigmoid}(G). \tag{3.13}$$

The component form expansion expression is:

$$F_{\text{glu}}(i, j, c) = V(i, j, c) \cdot \frac{1}{1 + \exp(-G(i, j, c))}. \tag{3.14}$$

3.3.2. The cross stage partial architecture and gradient propagation in GCFF

The GCFF module incorporates the concept of cross stage partial networks [31]. It splits the input feature F along the channel dimension into two parts, F_1 and F_2. Here, F_1 directly proceeds to subsequent operations, while F_2 is processed by the ConvGLU unit before being combined with F_1. This design, widely adopted in modern deep learning architectures, not only reduces computational cost, but also enhances gradient flow.

To analyze the gradient propagation, consider the gradient of the loss $\mathcal{L}$ with respect to the input G of the gating branch:

$$\frac{\partial \mathcal{L}}{\partial G} = \frac{\partial \mathcal{L}}{\partial F_{\text{glu}}} \odot V \odot \sigma(G)(1 - \sigma(G)). \tag{3.15}$$

The presence of V in the computational rule implies that the learning of the gating network is directly modulated by the amplitude of the content branch's features V. If V contains large-magnitude defect features, the gradient increases, prompting the gate G to open. Conversely, if V corresponds to noise, the gradient drives G to close. This adaptive feature-selection mechanism enables the network to dynamically suppress or enhance features at specific scales based on their semantic importance, a mathematically more refined behavior than that of a simple ReLU activation.

3.3.3. Multi-scale feature fusion

The final output of GCFF is the fusion of features from two branches. Let the high-level input feature be denoted as F_{high} and the low-level feature as F_{low}. Feature alignment is first performed to ensure spatial consistency for the subsequent fusion. To align features across different scales, an upsampling operator $\mathcal{U}$ and a downsampling operator $\mathcal{D}$ are introduced. The aligned and fused feature F_{fused} is then given by:

$$F_{\text{align}} = \text{Concat}(F_{\text{low}}, \mathcal{U}(F_{\text{high}})). \tag{3.16}$$

After processing by the gated convolutional unit ConvGLU,

$$F_{\text{out}} = \text{Conv}_{1\times1}\Big(\text{Concat}(F_1, \text{ConvGLU}(F_2))\Big), \tag{3.17}$$

where F_1 and F_2 are obtained by splitting F_{align}. This process ensures that high-level semantic information about defect categories and low-level detail information about defect boundaries can be combined in an alias-free manner, regulated by the gating mechanism. The output feature F_{out} will then serve as the input for the next stage of cross-scale fusion.

3.4. Loss function

To ensure that the PVEL-DETR algorithm achieves precise category classification and bounding box regression for PV cell defect detection, the total loss function is formulated as a combination of the bounding box loss $\mathcal{L}_{\text{box}}$ and the classification loss $\mathcal{L}_{\text{cls}}$. A feature uncertainty term is introduced to dynamically weight the classification loss, ensuring that high-quality features play a dominant role during training. The detailed calculation is given by:

$$\mathcal{L}(X, Y, \hat{Y}) = \mathcal{L}_{\text{box}}(b, \hat{b}) + \mathcal{L}_{\text{cls}}(U(X), c, \hat{c}), \tag{3.18}$$

where $Y = \{c, b\}$ represents the model's prediction result; $c \in [0, 1]^{K\times C}$ represents the probability distribution of defect category prediction, with the number of queries K set to a fixed value of 300; C represents the number of defect categories set to 9; $b \in \mathbb{R}^{K\times 4}$ represents the bounding box coordinate prediction; $\hat{Y} = \{\hat{c}, \hat{b}\}$ represents the ground truth label, $\hat{c} \in \{0, 1\}^{K\times C}$ represents the one-hot encoding of the defect category, and $\hat{b} \in \mathbb{R}^{K\times 4}$ represents the ground truth coordinates of the bounding box; $X \in \mathbb{R}^{H\times W\times C}$ represents the feature map output by the encoder, and $U(X)$ represents the feature uncertainty, used to measure feature quality; the smaller the value, the better the consistency between feature classification and localization.

Three additional clarifications are provided for the loss formulation. First, following the original RT-DETR [21], the feature uncertainty is explicitly defined as the discrepancy between the predicted probability distributions of localization and classification, i.e., $U(\hat{X}) = \|P(\hat{X}) - C(\hat{X})\|$, so that encoder features with consistent classification and localization quality are preferentially selected as initial object queries. Second, the label assignment between the K predicted queries and the ground-truth objects follows the standard set-prediction paradigm of the DETR family. Given a variable number m of ground-truth objects in an image (typically $m \ll K$), the ground-truth set is padded to the fixed size K with the "no-object" (background) class, and the Hungarian algorithm is applied to obtain the optimal one-to-one assignment:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{K} L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \tag{3.19}$$

where the matching cost combines the classification term with the L1 and generalized IoU bounding-box terms. Queries that are not matched to any real object are assigned to the no-object class and contribute only to the classification loss, so redundant predictions are automatically suppressed without non-maximum suppression. The number of queries $K = 300$ is kept identical to the original RT-DETR and does not influence the detection results, provided that it exceeds the maximum number of objects per image; accordingly, the $K \times C$ and $K \times 4$ ground-truth dimensions refer to the padded ground-truth set after bipartite matching rather than to 300 physical defect instances. Third, the classification loss adopts the IoU-aware varifocal loss, and the bounding-box loss is composed of the L1 loss and the generalized IoU loss, with weighting coefficients of 1, 5, and 2 for the classification, L1, and generalized IoU terms, respectively, identical to the original RT-DETR configuration.

The bounding-box loss employs the generalized intersection over union (IoU) loss. Unlike the standard IoU, generalized IoU considers not only the overlap area between the predicted and ground-truth boxes but also the area of their minimum enclosing box. This design effectively mitigates the gradient-vanishing problem inherent in traditional IoU loss, thereby improving both the convergence speed and the localization accuracy of bounding-box regression.

3.5. Characteristics of the proposed method

To conclude the methodological part, the characteristics of the proposed PVEL-DETR are summarized from four aspects. (1) Task-driven module design: each of the three improvements corresponds to one intrinsic challenge of PV cell EL imagery. The CAS-Block addresses weak and low-contrast defect features through additive-similarity gating in the backbone; the EDII module addresses environmental noise and illumination interference through dynamic frequency-domain filtering in intra-scale interaction; and the GCFF structure addresses overlapping-defect confusion through gated cross-scale fusion. (2) Replacement-based lightweight architecture: rather than stacking additional blocks onto the baseline, all three modules replace the corresponding components of RT-DETR, namely the standard residual blocks in the backbone, the FFN inside the AIFI module, and the RepBlock-based fusion unit in the CCFF module, respectively. Accuracy improvements are therefore obtained while the overall parameter count (18.3 M) remains lower than that of the 19.9 M baseline. (3) Linear-complexity contextual modeling: both the additive token mixer and the patch-wise frequency-domain filtering operate with linear computational complexity, so richer contextual information is exploited without sacrificing the real-time inference capability (104.2 FPS). (4) End-to-end NMS-free detection: the uncertainty-minimal query selection and the Hungarian-matching-based set prediction retain the fully end-to-end nature of the DETR family, which avoids hand-crafted post-processing and the associated extra latency.

4. Experimental results and discussion

4.1. Experimental setup

The implementation details and experimental settings of the proposed PVEL-DETR model are summarized as follows. All experiments are conducted on a hardware platform equipped with an NVIDIA RTX 4090 GPU (24 GB) and an Intel(R) Xeon(R) W7-3455 CPU with a base frequency of 2.50 GHz. The software environment is built upon Python 3.10, PyTorch 2.4.0, and CUDA 11.8. To optimize the experimental results, the initial learning rate is set to 0.0002, with a batch size of 8, 200 training epochs, and an input image size of 640×640. The AdamW optimizer is employed with a weight decay of 0.0001 and dynamic parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to balance training stability. A cosine annealing scheduler is used for learning rate scheduling. All other parameter settings and pre-training strategies remain consistent with the design in the original RT-DETR model. For fair comparison, all competing detectors in Section 4.4 are trained and evaluated on the same hardware and software platform with the identical data split and 640×640 input resolution, and their remaining hyperparameters follow the officially recommended configurations of their original implementations.

4.2. Dataset description

The PV cell EL anomaly detection dataset is a large-scale and specialized benchmark for defect detection in PV cell EL images. It was jointly constructed and released by the School of Artificial Intelligence and Data Science at Hebei University of Technology and the School of Automation Science and Electrical Engineering at Beihang University [1]. This dataset fills the gap of large-scale annotated data in the field of PV cell defect inspection and provides a standardized evaluation platform for model training and industrial application of deep learning algorithms.

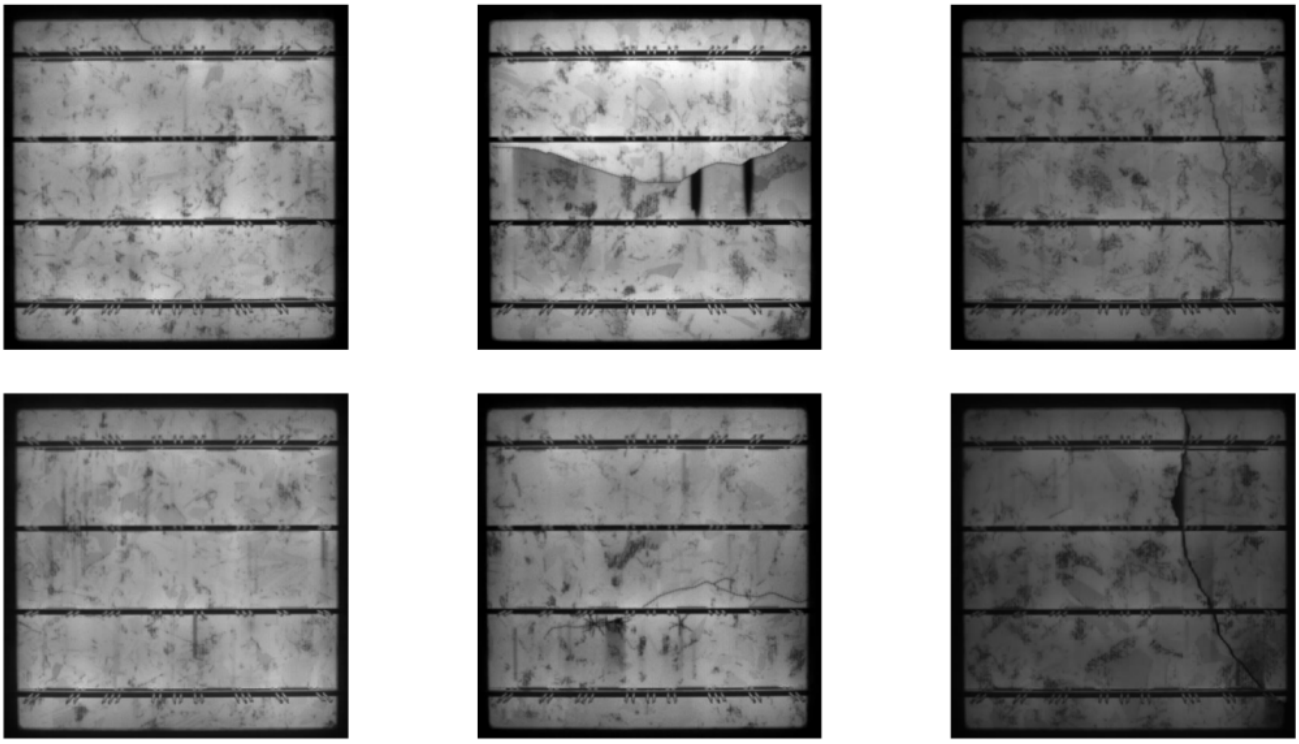

Figure 7. Example of the PV cell EL anomaly detection dataset.

The dataset covers nine common industrial defect categories, namely: crack, finger interruption, black core, thick line, star crack, horizontal dislocation, vertical dislocation, printing error, and short circuit. All experiments in this work are conducted on a single dataset: from the complete PVEL-AD benchmark, a total of 3600 EL images covering all nine defect categories are selected, and this subset is randomly divided into training, validation, and test sets with a ratio of 8:1:1, i.e., 2880 images for training, 360 for validation, and 360 for testing. The random split is performed once at the image level and kept fixed throughout all experiments, so every compared model is trained and tested on exactly the same data; each image corresponds to an individual PV cell, and every defect instance is annotated with a rectangular bounding box and its category label. The test set contains 1570 defect instances in total, whose per-category statistics are listed in Table 2. It should be noted that the instance quantities across categories are imbalanced (e.g., 560 finger instances versus 16 printing-error instances), which faithfully reflects the natural occurrence frequencies of defects in industrial production. The rare categories, such as printing error and vertical dislocation, exhibit large scales and highly distinctive structures and are comparatively easy to identify, so only a limited number of such samples is required; selecting a representative subset of the full PVEL-AD dataset for experiments is also a common practice in recent studies [32]. Typical sample images from the PV cell EL anomaly detection dataset are illustrated in Figure 7.

4.3. Experiment metrics

In this paper, mAP, giga floating-point operations per second (GFLOPs), and FPS are adopted as the key evaluation metrics for quantitative assessment of defect detection performance. Among these, mAP serves as the most critical metric for detection accuracy. It is calculated by first obtaining the average precision (AP) for each defect category, which corresponds to the area under the precision-recall curve. Precision is defined as the ratio of true positive predictions to all predicted positive samples, whereas recall denotes the ratio of true positive predictions to all ground-truth positive samples. The mAP is then computed as the arithmetic mean of AP values across all categories.

Specifically, mAP@0.5 represents the mAP value evaluated at an IoU threshold of 0.5, where IoU measures the overlapping ratio between the predicted bounding box and the ground-truth box. Meanwhile, mAP@0.5:0.95 refers to the average mAP over multiple IoU thresholds from 0.5 to 0.95 with a step of 0.05. This metric provides a more comprehensive evaluation of model robustness under high-precision requirements. On the other hand, model parameters and GFLOPs are utilized to evaluate the lightweight level and hardware resource consumption of the model. FPS is adopted to quantify the inference efficiency, defined as the number of image frames that the model can process per second.

4.4. Model performance comparison results

Under identical hardware and software configurations, comprehensive quantitative and qualitative comparisons with several SOTA real-time detectors and end-to-end detectors are conducted to validate the effectiveness of the proposed method. The quantitative experimental results on the dataset are summarized in Table 1. The proposed PVEL-DETR model achieves a superior trade-off among detection accuracy, inference speed, and model lightweightness for PV cell EL image defect detection, yielding more favorable overall performance than existing advanced object detection frameworks.

In terms of detection accuracy, PVEL-DETR achieves an mAP@0.5 of 0.932, which is 2.4 percentage points higher than the second-best model EER-DETR, 4.1 percentage points higher than the baseline RT-DETR, and 4.9 percentage points higher than the best-performing YOLO-series variant (YOLOv12-L). This substantiates its stronger recognition capability for nine defect categories, including cracks, black cores, and finger interruptions. On the stricter multi-IoU metric mAP@0.5:0.95, PVEL-DETR surpasses YOLOv12-L and EER-DETR by 3.4 and 3.8 percentage points, respectively, while also outperforming the baseline RT-DETR. This indicates its stable and reliable detection performance across diverse defect scales and various bounding box matching criteria. For million-level PV power stations, where PV modules are usually connected in series such that one defective module may disable the entire string, such accuracy improvement effectively reduces power generation losses and operation and maintenance costs. The relatively high mAP@0.5:0.95 values observed across all models will be further discussed in subsequent ablation experiments, with detailed analysis provided in Table 2.

Since EER-DETR is the strongest competitor, the concrete advantages of PVEL-DETR over it deserve particular emphasis. Quantitatively, with an equal parameter budget of 18M, PVEL-DETR improves mAP@0.5 from 0.908 to 0.932 (+2.4 percentage points), improves the stricter mAP@0.5:0.95 from 0.522 to 0.560 (+3.8 percentage points), and increases the inference speed

from 98.1 to 104.2 FPS (+6.1 FPS). The larger margin on mAP@0.5:0.95 than on mAP@0.5 indicates that the improvement stems mainly from higher-quality localization of difficult targets rather than from easier samples. Qualitatively, the advantages concentrate exactly on the challenges that EER-DETR does not explicitly address: as shown in Figure 8, EER-DETR produces ambiguous bounding boxes on faint defect edges and occasional duplicate detections, whereas PVEL-DETR yields tighter boxes for weak, small-scale defects and cleaner separation of densely overlapping defects, benefiting from the additive-similarity feature extraction, the frequency-domain denoising interaction, and the gated cross-scale fusion, respectively.

In terms of inference efficiency, tested under the same experimental platform, PVEL-DETR achieves a high frame rate of 104.2 FPS, ranking first among all compared detectors. This represents an improvement of 6.1 FPS over EER-DETR and 11.6 FPS over RT-DETR, exceeding both conventional DETR-based models and high-performance YOLO variants. Therefore, the proposed model fully satisfies the real-time detection requirements in practical industrial inspection scenarios.

In terms of model lightweight design, PVEL-DETR has only 18M parameters, which is comparable to RT-DETR and lower than most mainstream real-time and end-to-end detectors. Its computational complexity is 56 GFLOPs, which is slightly higher than that of EER-DETR, but accompanied by a significant accuracy improvement. Compared with the baseline RT-DETR, the computational cost is reduced by 8.2%, which effectively lowers the hardware resource consumption and facilitates practical deployment on edge computing devices.

Table 1. Quantitative comparison of experimental results on the dataset.

Model	Parameters (M)	GFLOPs	mAP@0.5	mAP@0.5:0.95	$FPS_{bs=1}$
YOLOv5-L [33]	46	109	0.712	0.402	54.4
YOLOv6-L [34]	59	150	0.821	0.443	92.4
YOLOv7-L [35]	37	104	0.728	0.411	55.9
YOLOv8-L [36]	43	165	0.822	0.451	65.6
YOLOv8-X [36]	68	257	0.885	0.493	37.4
YOLOv9-E [37]	57	189	0.890	0.511	42.3
YOLOv10-L [38]	25	120	0.866	0.514	72.8
YOLOv11-L [39]	25	87	0.879	0.522	87.7
YOLOv12-L [40]	27	89	0.883	0.526	85.4
DINO-DETR [20]	47	279	0.844	0.468	9.6
RT-DETR [21]	20	61	0.891	0.496	92.6
EER-DETR [22]	18	49	0.908	0.522	98.1
PVEL-DETR	**18**	**56**	**0.932**	**0.560**	**104.2**

Compared with defect detection tasks in other industrial fields, the overall mAP@0.5:0.95 values achieved by all models in the comparative experiments are relatively high. Since the overall mAP@0.5:0.95 is computed as the arithmetic mean of the per-category AP values, it is strongly influenced by the categories that remain accurate under strict IoU thresholds. As reported in Table 2, black core, horizontal dislocation, and vertical dislocation achieve AP@0.5:0.95 values of 0.705, 0.660, and 0.619, respectively. These three defect types possess large spatial scales, regular geometric

structures, and salient contrast against the background, so their predicted bounding boxes remain tightly aligned with the ground truth even under the strictest IoU criteria, which raises the averaged metric for all detectors. Detailed detection performance for each defect subcategory is listed in Table 2.

Table 2. Detailed performance metrics for each defect category.

Defect Category	Instances	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Overall	1570	0.902	0.870	0.932	0.560
Crack	268	0.848	0.585	0.787	0.436
Finger	560	0.947	0.779	0.933	0.523
Black_core	216	0.989	0.949	0.993	0.705
Thick line	174	0.942	0.862	0.936	0.528
Star_crack	27	0.695	0.778	0.806	0.466
Horizontal_dislocation	171	0.994	1.000	0.995	0.660
Vertical_dislocation	32	0.872	0.875	0.948	0.619
Printing_error	16	0.833	1.000	0.991	0.542
Short_circuit	106	0.998	1.000	0.995	0.565

As illustrated in Figure 8, to comprehensively validate the detection performance, visual comparisons of defect detection results are further conducted between the proposed PVEL-DETR and all competing models. The visualization results reveal clear and distinct performance gaps among different detectors.

Early real-time detectors such as YOLOv5-L and YOLOv7-L exhibit serious missed detections for tiny and faint defects. YOLOv8 and YOLOv10 achieve better recall performance but still generate duplicate and overlapping bounding boxes, mainly due to their limited capability in distinguishing real defects from cluttered EL image backgrounds. YOLOv12-L significantly improves defect recall; nevertheless, it still suffers from obvious label confusion under severe overlapping defect scenarios, indicating insufficient capability in handling complex defect distributions.

Within the DETR-based family, DINO-DETR yields high-precision localization for defects with salient features but is easily disturbed by noise, frequently misclassifying noisy regions as real defects. RT-DETR maintains favorable accuracy while ensuring high inference speed, yet it still lacks sensitivity to small-scale and faint defects. EER-DETR shows relatively stable recognition for medium-size defects with few redundant annotations. However, its bounding box annotations for faint defect edges are often ambiguous, and duplicate bounding boxes still occasionally occur.

In sharp contrast, the proposed PVEL-DETR displays more remarkable visual advantages. It can accurately capture complete contours of faint defects, and its predicted bounding boxes exhibit higher tightness with ground-truth defects without obvious missed detections. In addition, PVEL-DETR shows stronger anti-noise robustness and superior capability in recognizing densely overlapping defects.

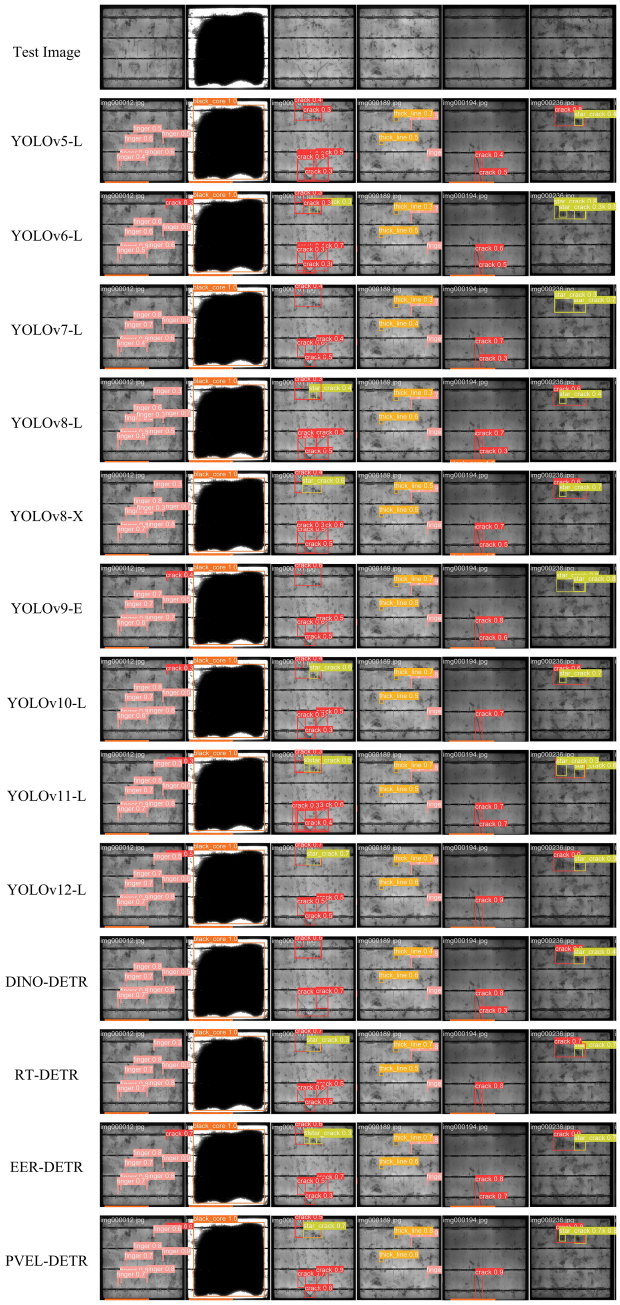

Figure 8. Visual comparison of defect detection results between the proposed PVEL-DETR and the competing models.

In summary, compared with existing SOTA real-time detectors and end-to-end detectors, the proposed model achieves higher detection accuracy without introducing excessive parameters or computational overhead. It achieves significantly improved overall performance over the baseline RT-DETR. Although its computational complexity is slightly higher than that of the lightweight EER-DETR, PVEL-DETR achieves higher detection accuracy and faster inference speed with a comparable model size. The reliability and advancement of the proposed method are further strongly verified by the above visual results.

Therefore, the proposed PVEL-DETR achieves a more favorable and comprehensive trade-off among detection accuracy, real-time inference speed, model lightweightness, and computational complexity, making it more suitable and practical for industrial real-time defect detection on PV cell EL images.

4.5. Ablation study results

To verify the individual effectiveness and complementary advantages of each proposed module, a comprehensive set of ablation experiments is carried out based on the baseline RT-DETR framework. It should be clarified first that the three proposed modules are replacements of the corresponding baseline components rather than additionally stacked blocks: the CAS-Block replaces the standard residual blocks in the backbone, the EDII module replaces the FFN inside the original AIFI module, and the GCFF structure replaces the RepBlock-based fusion unit in the CCFF module. In each ablation variant, the corresponding component of the baseline is substituted by the proposed module for quantitative comparison and evaluation. The ablation results are summarized in Table 3, where A represents the introduction of the CAS-Block, B denotes the EDII module, and C indicates the GCFF structure.

The experimental results demonstrate that integrating the CAS-Block alone increases mAP@0.5 to 0.926, which significantly enhances the feature extraction ability for faint and small defects and effectively reduces missed detections, at the cost of a slight increment in model parameters and computational complexity. Replacing the FFN with the EDII module alone raises mAP@0.5 to 0.912 while slightly increasing the speed to 93.1 FPS. The GCFF structure is primarily an efficiency-oriented replacement: Used alone, it trades a slight accuracy decrease (0.882 versus the 0.891 baseline in mAP@0.5) for a substantial efficiency gain (109.3 versus 92.6 FPS, and 18.2M versus 19.9 M parameters); its accuracy benefit emerges only when the preceding modules supply more discriminative multi-scale features. The parameter and computation changes in Table 3 also directly reflect this replacement design: Because both the EDII module and the GCFF structure are lighter than the original components they replace, the variants Base+B (18.6 M), Base+C (18.2 M), and Base+B+C (16.9 M) contain fewer parameters than the 19.9 M baseline, whereas the CAS-Block introduces a moderate increase (Base+A, 21.3 M); the complete PVEL-DETR (18.3 M) therefore combines the accuracy gain of the CAS-Block with the parameter savings of the other two replacements and remains lighter than the baseline.

Furthermore, the combination experiments should be interpreted from the replacement perspective as well. When individual replacements are combined, mild metric fluctuations are observed; for example, Base+A+B yields a lower mAP@0.5:0.95 than Base+A, and Base+A+C yields a lower mAP@0.5 than Base+A, which is expected, because each replacement alters the feature statistics received by the subsequent stages and re-balances the accuracy–efficiency trade-off. Nevertheless, the complete model achieves the best overall performance (0.932 mAP@0.5 and 0.560 mAP@0.5:0.95 at 104.2 FPS) with fewer parameters than the baseline, surpassing every single- and dual-module variant on both accuracy metrics. This indicates that the three replacements are complementary rather than conflicting: the CAS-Block strengthens weak-feature extraction, the EDII module stabilizes high-level semantics against noise, and the GCFF structure recovers the efficiency cost while preserving the accuracy gains, jointly achieving the best trade-off among detection accuracy, inference speed, and model complexity.

Table 3. Experimental results of ablation study.

Model	Parameters (M)	GFLOPs	mAP@0.5	mAP@0.5:0.95	$FPS_{bs=1}$
Base	19.9	61	0.891	0.496	92.6
Base+A	21.3	65	0.926	0.543	84.4
Base+B	18.6	59	0.912	0.517	93.1
Base+C	18.2	54	0.882	0.476	109.3
Base+A+B	20.0	62	0.928	0.522	86.7
Base+A+C	19.6	58	0.917	0.533	99.5
Base+B+C	16.9	53	0.908	0.526	107.2
PVEL-DETR	18.3	56	0.932	0.560	104.2

To further illustrate the intermediate iterative process of training, Figure 9 presents the evolution of the total training loss and the validation accuracy of PVEL-DETR and the baseline RT-DETR over the 200 training epochs. The total loss decreases rapidly within the early epochs and then converges smoothly under the cosine annealing schedule without oscillation or divergence, while the validation mAP@0.5 rises correspondingly and stabilizes in the later training stage. Moreover, PVEL-DETR converges to a lower loss level and a higher accuracy plateau than the baseline, indicating that the proposed modules not only improve the final detection performance, but also preserve a stable and efficient optimization behavior.

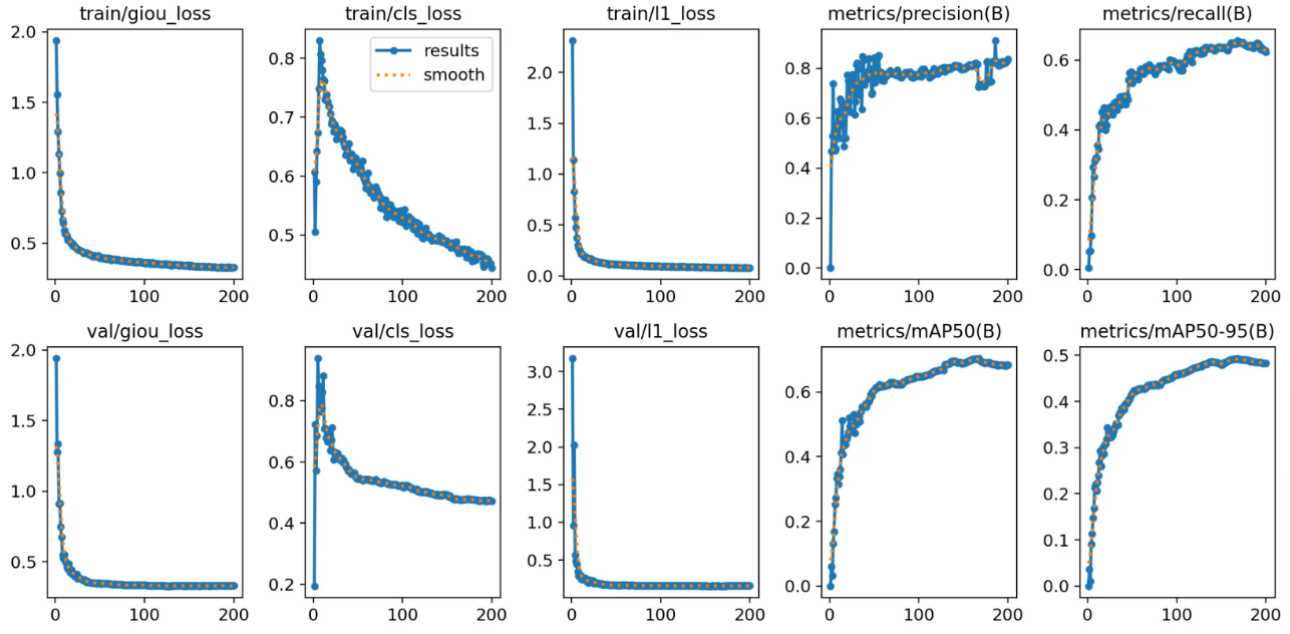

Figure 9. Training loss and validation accuracy curves of PVEL-DETR and the baseline RT-DETR.

4.6. Discussion

4.6.1. Detection mechanism and performance analysis of complex defect features

This study aims to tackle the critical detection challenges in PV cell EL images, including faint defect features, serious noise interference, and severe multi-defect overlapping. Extensive experimental results consistently confirm that the proposed PVEL-DETR yields substantial and stable improvements across all key evaluation metrics. The underlying detection mechanisms and working principles are further analyzed and summarized as follows.

First, to address the difficulties caused by low contrast and poor imaging quality in EL images, the ablation study demonstrates that the introduction of the CAS-Block improves the model's mAP@0.5 by 3.5%. This significant improvement strongly validates the theoretical design proposed in this work: conventional ResNet blocks rely on fixed spatial convolutions and tend to lose high-frequency edge information in low signal-to-noise ratio regions due to weak feature amplitudes. In contrast, the additive attention mechanism embedded in the CAS-Block decouples the query and the key geometrically and relies on sigmoid-gated local context instead of globally Softmax-normalized dot products; as analyzed in Section 3.1.3, this avoids the dilution of weak defect responses by massive background tokens and maintains a comparatively stable feature sensitivity in dark regions. Visualization results clearly show that PVEL-DETR outlines fine crack boundaries more sharply and completely than the baseline RT-DETR, which verifies the superiority of the additive similarity function in solving ill-posed inverse imaging problems under low-contrast conditions.

Second, the integration of the EDII module significantly enhances the noise-robustness of the detection framework. In actual industrial field inspection, non-uniform illumination, grain boundaries, and texture noise are frequently misidentified as real defects. By projecting features into the frequency domain for decoupling processing, the EDII module effectively separates low-frequency components (mainly corresponding to illumination variations) from high-frequency components (mainly corresponding to real defect structures). Benefiting from the dynamic global context filtering realized by the frequency-domain filtering mechanism, PVEL-DETR exhibits stronger anti-interference capability than YOLOv12-L, which suffers from a high false detection rate in complex backgrounds. Such frequency-domain feature interaction essentially acts as an adaptive band-pass filter, which suppresses environmental interference to a large extent while retaining effective defect features.

Finally, regarding the challenge of severe overlapping defects, the gating mechanism in the GCFF module plays a decisive role. Quantitative experimental results indicate that for the easily confused defect categories, the proposed model achieves detection precisions of 84.8% and 94.7%, respectively, accompanied by excellent recall performance. The ConvGLU unit adaptively learns channel-wise feature weights and successfully suppresses the interference caused by redundant background information during multi-scale feature fusion, thereby realizing accurate target decoupling for densely overlapping defects.

4.6.2. Engineering feasibility analysis of real-time and edge computing

In practical industrial inspection scenarios, detection speed is equally critical as detection accuracy. With the rapid construction and large-scale deployment of terawatt-level ultra-large PV power plants, unmanned aerial vehicle (UAV)-based inspection systems generate an extremely large

amount of real-time image data. If the intelligent detection algorithm cannot support synchronous on-board processing, the transmission and storage of such massive data will introduce huge communication bandwidth costs and unacceptable detection latency.

The proposed PVEL-DETR achieves an ultra-high inference speed of 104.2 FPS, which implies that it only takes less than 10 milliseconds to process a single EL image. This speed far exceeds the commonly used video frame rates (30–60 FPS) in industrial inspection. Compared with the advanced YOLO-series model YOLOv12-L, PVEL-DETR is not only faster in inference but also lighter in parameters, and such a compact design indicates favorable potential for future deployment on resource-constrained platforms. It must be emphasized, however, that all speed measurements in this work were obtained on a desktop-grade RTX 4090 GPU; the latency, memory footprint, power consumption, and quantized-model performance on embedded GPUs, edge accelerators, or UAV onboard computers have not yet been evaluated. Therefore, the results reported here should be interpreted as evidence of algorithm-level real-time capability rather than as a demonstration of embedded deployment, and the corresponding edge-side validation is explicitly identified as future work in Section 4.6.3. It is also worth clarifying the application scenario: field EL inspection requires the PV strings to be forward-biased and is typically performed at night, with near-infrared cameras carried by handheld devices, ground vehicles, or UAV platforms; the massive EL images collected in this way are processed either on the inspection terminal or on nearby edge servers, which is the deployment context targeted by the lightweight design of this work.

Furthermore, different from the anchor-based mechanism widely used in the YOLO series, the anchor-free architecture adopted in PVEL-DETR effectively eliminates the computational overhead caused by post-processing steps, such as NMS, thereby further reducing the end-to-end detection latency.

4.6.3. Limitations and future work

Although the proposed PVEL-DETR achieves superior and reliable detection performance on the constructed PV cell EL dataset, several limitations still exist and deserve further in-depth investigation.

First, the current experimental dataset mainly focuses on poly-silicon PV cells. The generalization ability of the trained model to defect characteristics in emerging high-efficiency cell technologies, such as heterojunction (HJT), TOPCon, and perovskite solar cells, still requires sufficient experimental verification. Variations in surface texture, luminescence characteristics, and defect morphology across different cell types may affect the stability and effectiveness of feature extraction.

Second, although the frequency-domain filtering mechanism endows the model with certain robustness to illumination variations, its detection performance under extreme real-world environmental conditions (e.g., heavy rain, severe dust coverage, and snow shielding) has not been fully tested. Collecting representative EL images under harsh conditions and further optimizing the model's anti-interference ability will be crucial challenges for practical engineering deployment.

Accordingly, future research will concentrate on the following three directions:

(1) Validating and optimizing edge deployment: deploying the model on representative embedded platforms (e.g., NVIDIA Jetson-class devices) with quantization and TensorRT acceleration, and systematically evaluating the end-to-end latency, memory footprint, and power consumption under realistic inspection workloads, so as to provide direct experimental evidence for the edge-side applicability that is currently claimed only at the algorithm level.

(2) Improving generalization under limited supervision: Exploring effective strategies to utilize massive unlabeled EL images or multi-modal data for semi-supervised or self-supervised training, so as to enhance model generalization and reduce dependence on high-cost, high-quality manual annotations.

(3) Developing temporal defect prediction models: Integrating long-term historical inspection data to establish a dynamic degradation model for predicting defect propagation trends and formulating intelligent maintenance strategies. This will advance the research from single-image defect detection to full-life-cycle health state prediction of PV cells, providing more proactive and intelligent decision support for smart operation and maintenance of large-scale PV power plants.

5. Conclusions

This paper focuses on solving the critical challenges in PV cell EL images defect detection, including low imaging quality, strong background noise interference, and difficulty in detecting multi-scale and faint defects. To this end, an efficient real-time detection algorithm named PVEL-DETR is proposed based on the improved RT-DETR architecture. By redesigning the feature extraction backbone, constructing a dynamic frequency-domain feature interaction mechanism, and optimizing the cross-scale feature fusion structure, the proposed model achieves remarkable improvements in fine-grained feature capture, anti-noise robustness, and multi-target discrimination ability. Extensive experimental results on a large-scale benchmark dataset demonstrate that PVEL-DETR achieves 93.2% mAP@0.5 and a real-time inference speed of 104.2 FPS. Compared with SOTA detectors such as YOLOv12-L and EER-DETR, the proposed method not only significantly reduces missed detections of tiny and weak defects but also satisfies the strict requirements of industrial real-time inspection with lower model parameters and computational complexity.

Use of AI Tools Declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgment

The research was funded by the National Natural Science Foundation of China (Grant No. 62477218). The authors gratefully acknowledge the great help of the fund.

Author Contributions

The authors confirm contribution to the paper as follows: Xiang Chen: Conceptualization, methodology, validation, writing—original draft preparation. Haiyang Sun: formal analysis, writing—review and editing. Liming Sun: formal analysis, writing—review and editing. Yelin Deng: Funding acquisition, supervision, writing—review and editing. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials

Data will be made available on request.

Conflicts of Interest

The authors declare no conflicts of interest to report regarding the present study.

References

1. B. Su, Z. Zhou, H. Chen, PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection, *IEEE Trans. Ind. Inform.*, **19** (2023), 404–413. https://doi.org/10.1109/TII.2022.3162846
2. V. Sharma, S. Chandel, Performance and degradation analysis for long term reliability of solar photovoltaic systems: A review, *Renew. Sust. Energ. Rev.*, **27** (2013), 753–767. https://doi.org/10.1016/j.rser.2013.07.046
3. S. Gallardo-Saavedra, L. Hernández-Callejo, M. Alonso-García, J. Santos, J. Morales-Aragonés, V. Alonso-Gómez, et al., Nondestructive characterization of solar PV cells defects by means of electroluminescence, infrared thermography, I–V curves and visual tests: Experimental study and comparison, *Energy*, **205** (2020), 117930. https://doi.org/10.1016/j.energy.2020.117930
4. S. A. Memon, Q. Javed, W. G. Kim, Z. Mahmood, U. Khan, M. Shahzad, A machine-learning-based robust classification method for PV panel faults, *Sensors*, **22** (2022), 8515. https://doi.org/10.3390/s22218515
5. J. Fioresi, D. J. Colvin, R. Frota, R. Gupta, M. Li, H. P. Seigneur, et al., Automated defect detection and localization in photovoltaic cells using semantic segmentation of electroluminescence images, *IEEE J. Photovolt.*, **12** (2022), 53–61. https://doi.org/10.1109/JPHOTOV.2021.3131059
6. J. Tian, C. Chen, W. Shen, F. Sun, R. Xiong, Deep learning framework for lithium-ion battery state of charge estimation: Recent advances and future perspectives, *Energy Storage Mater.*, **61** (2023), 102883. https://doi.org/10.1016/j.ensm.2023.102883
7. H. Tella, A. Hussein, S. Rehman, B. Liu, A. Balghonaim, M. Mohandes, Solar photovoltaic panel cells defects classification using deep learning ensemble methods, *Case Stud. Therm. Eng.*, **66** (2025), 105749. https://doi.org/10.1016/j.csite.2025.105749
8. S. Shah, D. Suraj, S. M. Reza, M. A. R. B. A. Salam, A. Ashraf, S. F. Ferdous, Comparative analysis of deep learning models for defect detection in additive manufacturing using thermal imaging, *Results Eng.*, **28** (2025), 108359. https://doi.org/10.1016/j.rineng.2025.108359
9. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., (2012), 1097–1105.
10. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
12. M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2018), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
13. W. Tang, Q. Yang, X. Hu, W. Yan, Deep learning-based linear defects detection system for large-scale photovoltaic plants based on an edge-cloud computing infrastructure, *Sol. Energy*, **231** (2022), 527–535. https://doi.org/10.1016/j.solener.2021.11.016
14. A. Howard, M. Sandler, B. Chen, W. Wang, L. C. Chen, M. Tan, et al., Searching for MobileNetV3, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2019), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
15. L. Pratt, D. Govender, R. Klein, Defect detection and quantification in electroluminescence images of solar PV modules using U-net semantic segmentation, *Renew. Energy*, **178** (2021), 1211–1222. https://doi.org/10.1016/j.renene.2021.06.086
16. U. Hijjawi, S. Lakshminarayana, T. Xu, G. P. M. Fierro, M. Rahman, A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations, *Sol. Energy*, **266** (2023), 112186. https://doi.org/10.1016/j.solener.2023.112186
17. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Curran Associates, Inc., (2017), 5998–6008.
18. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
19. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
20. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, et al., DINO: DETR with improved denoising anchor boxes for end-to-end object detection, preprint, arXiv:2203.03605.
21. Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, et al., DETRs beat YOLOs on real-time object detection, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2024), 16965–16974. https://doi.org/10.1109/CVPR52733.2024.01605
22. J. Dun, H. Yang, S. Yuan, Y. Tang, EER-DETR: An improved method for detecting defects on the surface of solar panels based on RT-DETR, *Appl. Sci.*, **15** (2025), 6217. https://doi.org/10.3390/app15116217
23. M. W. Akram, G. Li, Y. Jin, CNN based automatic detection of photovoltaic cell defects in electroluminescence images, *Energy*, **189** (2019), 116319. https://doi.org/10.1016/j.energy.2019.116319

24. M. Y. Demirci, N. Beşli, A. Gümüşçü, Efficient deep feature extraction and classification for identifying defective photovoltaic module cells in electroluminescence images, *Expert Syst. Appl.*, **175** (2021), 114810. https://doi.org/10.1016/j.eswa.2021.114810
25. J. Zhang, Y. Wang, B. Jiang, H. He, S. Huang, C. Wang, et al., Realistic fault detection of Li-ion battery via dynamical deep learning, *Nat. Commun.*, **14** (2023), 5940. https://doi.org/10.1038/s41467-023-41226-5
26. Q. Liu, M. Liu, C. Wang, Q. M. Wu, An efficient CNN-based detector for photovoltaic module cells defect detection in electroluminescence images, *Sol. Energy*, **267** (2024), 112245. https://doi.org/10.1016/j.solener.2023.112245
27. T. Zhang, L. Li, Y. Zhou, W. Liu, C. Qian, J. N. Hwang, et al., CAS-ViT: Convolutional additive self-attention vision transformers for efficient mobile applications, *IEEE Trans. Image Process.*, **35** (2026), 1899–1909. https://doi.org/10.1109/TIP.2026.3655121
28. Y. Rao, W. Zhao, Z. Zhu, J. Lu, J. Zhou, Global filter networks for image classification, in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, Curran Associates, Inc., (2021), 980–993.
29. L. Kong, J. Dong, J. Ge, M. Li, J. Pan, Efficient frequency domain-based transformers for high-quality image deblurring, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2023), 5886–5895. https://doi.org/10.1109/CVPR52729.2023.00570
30. Y. N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, (2017), 933–941.
31. C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, I. H. Yeh, CSPNet: A new backbone that can enhance learning capability of CNN, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2020), 390–391. https://doi.org/10.1109/CVPRW50498.2020.00203
32. A. Mohamed, Y. Nacera, B. Ahcene, A. Teta, E. O. Belabbaci, A. Rabehi, et al., Optimized YOLO based model for photovoltaic defect detection in electroluminescence images, *Sci. Rep.*, **15** (2025), 32955. https://doi.org/10.1038/s41598-025-13956-7
33. G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, et al., Ultralytics YOLOv5 (v7.0)–YOLOv5 sota realtime instance segmentation, 2022. https://doi.org/10.5281/zenodo.7347926
34. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, et al., YOLOv6: A single-stage object detection framework for industrial applications, preprint, arXiv:2209.02976.
35. C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2023), 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
36. G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8 (v8.0), 2023. Available from: https://github.com/ultralytics/ultralytics.

37. C. Y. Wang, I. H. Yeh, H. Y. M. Liao, YOLOv9: Learning what you want to learn using programmable gradient information, in *Lecture Notes in Computer Science*, Springer, (2024), 1–21. https://doi.org/10.1007/978-3-031-72751-1_1

38. A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al., YOLOv10: Real-time end-to-end object detection, in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, Curran Associates, Inc., (2024), 107984–108011.

39. R. Khanam, M. Hussain, YOLOv11: An overview of the key architectural enhancements, preprint, arXiv:2410.17725.

40. Y. Tian, Q. Ye, D. Doermann, YOLOv12: Attention-centric real-time object detectors, in *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, Curran Associates, Inc., (2025), 78433–78457.