scientific reports

OPEN

A PV cell defect detector combined with transformer and attention mechanism

Du Lang[1]✉ & Zhenzhen Lv[2]

Automated defect detection in electroluminescence (EL) images of photovoltaic (PV) modules on production lines remains a significant challenge, crucial for replacing labor-intensive and costly manual inspections and enhancing production capacity. This paper presents a novel PV defect detection algorithm that leverages the YOLO architecture, integrating an attention mechanism and the Transformer module. We introduce a polarized self-attention mechanism in the feature extraction stage, enabling separate extraction of spatial and semantic features of PV modules, combined with the original input features, to enhance the network's feature representation capabilities. Subsequently, we integrate the proposed CNN Combined Transformer (CCT) module into the model. The CCT module employs the transformer to extract contextual semantic information more effectively, improving detection accuracy. The experimental results demonstrate that the proposed method achieves a 77.9% mAP50 on the PVEL-AD dataset while preserving real-time detection capabilities. This method enhances the mAP50 by 17.2% compared to the baseline, and the mAP50:95 metric exhibits an 8.4% increase over the baseline.

Keywords Anomaly detection, Photovoltaic modules, Transformer, Attention mechanism, Deep learning

Photovoltaic (PV) solar cells are primary devices that convert solar energy into electrical energy. However, unavoidable defects can significantly reduce the modules' photoelectric conversion efficiency and lifespan, leading to substantial economic losses. Inspecting solar cells during the intelligent manufacturing process can substantially reduce the impact of defects in photovoltaic (PV) solar cells on the final products[1,2]. Manual electroluminescence (EL) image inspection is exceedingly cumbersome and necessitates specialized expertise. Consequently, a vision-based automated inspection process[3–6] is highly desirable.

Electroluminescence (EL) imaging technology[7] is vital for image acquisition, as certain anomalies are often embedded within the interior of PV cells, rendering them undetectable by optical or infrared cameras. However, visual inspection using EL imaging technology enables the easy identification of anomalies in solar cells, whether caused by external environmental influences such as impacts during the manufacturing process or by pre-existing material defects.

Computer vision-based methods can effectively meet the urgent demand for quality monitoring across various fields. Conventional computer vision methods for defect recognition primarily encompass two approaches: filter-based methods[4] and handcrafted feature-based methods[8]. Filter-based methods employ filters to extract defect features that differ from the background information, thereby achieving defect detection. Handcrafted feature-based methods manually extract features from images, such as texture, color, and morphological characteristics, then use trained classifiers to recognize defects. Traditional defect recognition methods require significant effort in feature extraction, and the extracted features often have limitations. In contrast, deep learning methods can automatically learn valuable features from data, enabling superior performance in defect recognition.

Deep learning-based methods are gaining increasing popularity in various fields, such as defect recognition[9–11], fault diagnosis[12,13], and soft sensing of industrial processes[14,15], owing to their high accuracy, broad generalization capabilities and malleability compared to traditional methods. This paper focuses exclusively on deep learning methods for defect recognition. CNN modules, such as Defect-Net[16] and PreAugNet[17], have demonstrated prominent performance in classifying defects in industrial manufacturing. The success of deep learning has gradually led to replacing traditional pattern recognition techniques with image inspection tasks. Deitscha et al.[18] proposed an end-to-end deep CNN for classifying defects in EL images of solar cells. Chen

[1]School of Information and Design, Zhejiang Industry Polytechnic College, Shaoxing 312000, China. [2]Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110167, China. ✉email: 20230037@zjipc.edu.cn

et al.[19] developed a novel solar CNN architecture to classify defects in visible light images of solar cells. Han et al.[20] proposed a deep learning-based defect segmentation method for polycrystalline silicon solar cells. This method employs an RPN to generate underlying defect regions, which are then resized and fed into an enhanced segmentation network. Otamendi et al.[21] used Faster R-CNN to detect defective regions in PV images, cropped the defect areas, and classified them using EfficientNet. Although this method exhibits excellent performance, it is too complex to meet the real-time requirements of the production line. Su et al.[22] proposed the complementary attention network (CAN) to enhance the recognition accuracy of the Faster RCNN algorithm for PV anomaly detection. Shin et al.[23] developed a solar distribution panel anomaly detection system using thermal imaging based on Faster RCNN. El Yanboiy et al.[7] implemented real-time solar cell defect detection using the YOLOv5 algorithm, improving the stability and efficiency of solar systems. Jha et al.[24] conducted a comprehensive review of defect detection research, emphasizing the application of high-performance detection algorithms like Faster RCNN and Cascade RCNN, as well as real-time detection algorithms such as the YOLO series in this domain.

A vital characteristic of the human visual system is its inability to process entire scene images simultaneously. Instead, humans utilize a series of partial glimpses to better capture visual structures, selectively focusing on salient parts[25]. Neural networks possess a mechanism akin to the human visual system known as the attention mechanism. In the industrial field, neural networks have widely applied attention mechanisms to improve performance. Zhao et al.[26] introduced the Squeeze and Excitation (SE) module into SSD. They proposed the SE-SSD algorithm to enhance the model's performance detecting surface defects on fabrics. Chen et al.[27] designed an efficient LCD defect detection algorithm based on the Efficient Channel Attention (ECA) module and Mask R-CNN, improving small object detection accuracy. Chen et al.[28] combined the Convolutional Block Attention Module (CBAM) with MobileNetv3 to design a lightweight, high-performance crack detection algorithm to achieve real-time crack localization. Liu et al.[29] proposed Polarized Self Attention (PSA) to model remote dependencies of high-resolution input/output features, subsequently estimating highly nonlinear pixel semantics. In various computer vision tasks, models designed using PSA modules have achieved better performance[29–31]. The attention mechanism enhances the performance of deep learning models by prioritizing significant information with high weights and disregarding irrelevant information with low weights.

Recently, researchers have increasingly utilized transformer-based vision models in diverse computer vision tasks, including image classification[32], object detection[33], and semantic segmentation[34]. Leveraging the powerful contextual information extraction capabilities of Transformers, Transformer-based models have outperformed pure CNN models across various vision tasks[35]. Transformers exhibit impressive performance, yet their computational efficiency is slow, posing challenges for applications requiring high real-time performance. Introducing the transformer module into CNN models can enhance performance[36].

Leveraging the attention mechanism and transformer, we introduce a novel model based on YOLOv8[37] designed to enhance performance in PV defect detection tasks significantly. The primary contributions of this study include:

(1) We introduce the CNN Combined Transformer (CCT) module, which integrates the Transformer with CNN to enhance the model's capability in extracting global contextual information. CCT improves the mAP50:95 metric of YOLOv8 on the validation set by 4.7%.
(2) We employ the Polarized Self Attention (PSA) mechanism to address feature fusion conflicts across various levels within the deep learning model, thereby enhancing detection accuracy across different detection heads. PSA improves the mAP50:95 metric of YOLOv8 on the validation set by 2.2%.
(3) We analyzed the performance metrics, frames per second (FPS), and model size of various PV defect detection algorithms, demonstrating that our proposed method achieves high-quality real-time monitoring in the PV cell production process.

The structure of this study is as follows: "Related work" section reviews relevant studies related to the proposed methodology. "Methodologies" section provides a detailed description of the proposed methods. "Experiments" section presents an in-depth analysis of the experiments conducted and their interpretations. Finally, "Conclusion" section offers concluding remarks and a discussion.

Related work

YOLOv8 baseline framework

Real-time object detection has long been a central focus of research in computer vision, aiming to accurately predict object categories and positions in images with minimal latency. YOLO models have garnered increasing popularity due to their adept balance of performance and efficiency. YOLOv3 identifies the typical detection architecture comprising three main components: backbone, neck, and head[38]. YOLOv4[39] introduces the CSPNet[40], coupled with data augmentation strategies, enhanced PAN, and an expanded range of model scales, among other improvements. YOLOv6[41] presents BiC and SimCSPSPPF for the neck and backbone, respectively, incorporating anchor-aided training and a self-distillation strategy.

YOLOv8[37], the latest iteration in the YOLO series, continues to advance object detection capabilities. Building on the success of its predecessors, YOLOv8 employs a multi-category approach and precise positional information to facilitate end-to-end model training, resulting in accelerated detection speeds and high average precision. By incorporating innovative features like Focus[42], CSP-Darknet[40], and SPP[43], YOLOv8 enhances detection accuracy and improves overall speed. YOLOv8 comes in n/s/m/l/x versions, each varying in parameters and performance; the baseline model used in this study is YOLOv8s.

Attention module

Methods aimed at directing attention to crucial image regions while disregarding irrelevant parts are called attention mechanisms. In a vision system, an attention mechanism is a dynamic selection process that adaptively weights features based on input importance. Various attention mechanism modules, including Squeeze-and-Excitation (SE)[26], Convolutional Block Attention Module (CBAM)[28], Efficient Channel Attention (ECA)[27], and Spatial and Channel Squeeze & Excitation (SCSE)[44], have been proposed to enhance neural network model performance. These methods strengthen expressiveness by employing attention mechanisms that prioritize significant features and suppress irrelevant features in channels and spaces. The emphasis or suppression of information also contributes to information flow within the network.

The Polarized Self Attention (PSA)[29] mechanism introduces polarization into self-attention mechanisms, allowing the model to focus on informative elements while suppressing distractions selectively. PSA enhances the discriminative power of self-attention mechanisms by assigning polarities to attention weights, thereby improving feature representation and task performance. PSA incorporates softmax-sigmoid composition in both channel-only and spatial-only attention branches to align with output distributions typical of fine-grained regression, thereby benefiting pixel-wise regression tasks.

Transformer

The transformer architecture has revolutionized deep learning, especially in tasks involving natural language processing (NLP) and sequence modeling. Introduced by Vaswani et al.[45], the transformer model departs from traditional recurrent and convolutional neural networks by exclusively utilizing self-attention mechanisms to capture dependencies across input sequences. The self-attention mechanism allows the model to evaluate the significance of different elements in the input sequence when computing the representation of each component. In contrast to traditional recurrent models that process sequences sequentially, self-attention allows for parallel computation across all elements in the sequence, thereby enhancing efficiency and scalability.

The primary advantage of the transformer lies in its effective capture of long-range dependencies facilitated by the self-attention mechanism. This capability renders it particularly suitable for tasks involving modeling complex relationships across input sequences, such as machine translation, text generation, and language understanding[46]. In computer vision tasks, the transformer can directly acquire contextual semantic information, a capability absent in CNNs. Owing to these advantages, transformer-based models outperform CNN in most computer vision tasks[47].

Methodologies

Proposed network

Despite the well-designed architecture of YOLOv8 for general object detection, detecting minute objects like PV anomaly defects remains highly challenging. In this study, we design an effective PV defect detection model leveraging the PSA attention mechanism and transformer architecture. Figure 1 illustrates the overall framework of the proposed model. Figure 2 presents the detailed architecture of each module within the YOLOv8 model. The Conv module, the most fundamental component, comprises a convolutional layer, a batchnorm layer, and the silu activation function. The Bottleneck module consists of multiple Conv units interconnected by residual connections. The SPPF module employs maximum pooling with kernels of varying sizes, concatenating the outputs to expand the receptive field. The C2f module comprises Bottleneck units linked through dense residual concatenation, enabling its output to encapsulate feature information from multiple layers.

Although the introduction of the feature pyramid network (FPN) and path aggregation network (PANet)[48] enhances feature representation in the neck of YOLOv8, it also generates conflicting and redundant information due to the direct fusion of multi-level features, thereby impeding detection performance enhancement. To address the issue of feature conflicts across different levels, we integrated the PSA attention mechanism, which autonomously selects optimal features during parameter optimization, thereby mitigating feature conflict problems during multi-level feature fusion.

As the number of layers in a convolutional neural network (CNN) increases, the receptive field associated with each feature in the deep feature map expands. However, the intrinsic properties of CNNs cause the peripheral regions of the receptive field to be less emphasized, thereby diminishing the global semantic information in the deep feature map. The transformer adaptively captures long-term dependency information through its multi-head attention mechanism. However, the computational load of the transformer module is substantial, making it challenging for pure transformer models to meet the real-time requirements of various modeling applications. We introduce a CNN Combined Transformer (CCT) module following the backbone of the YOLOv8 model to extract global semantic information, effectively addressing the issue of diminished global semantic information in pure CNN models. Positioning the CCT in the backbone's final layer minimally increases the model's overall computational burden, ensuring that the real-time performance remains largely unaffected.

CNN combined transformer module

The self-attention mechanism fundamentally empowers the transformer architecture to process input sequences in parallel rather than sequentially. This parallelization markedly reduces training times and enhances the model's ability to manage long-range dependencies within the data. The matrices Q (Query), K (Key), and V (Value) are the feature matrices, $1/\sqrt{d_k}$ serving as the scaling factor. The calculation process of the attention mechanism is as Eq. (1):

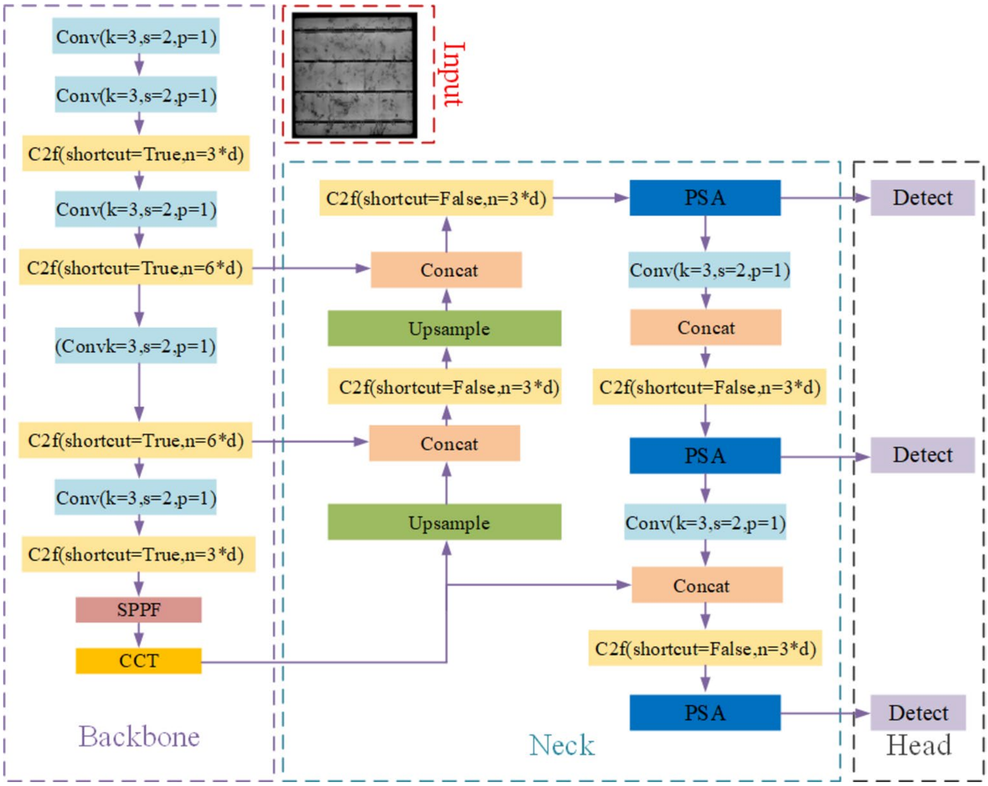

Fig. 1. The proposed network.

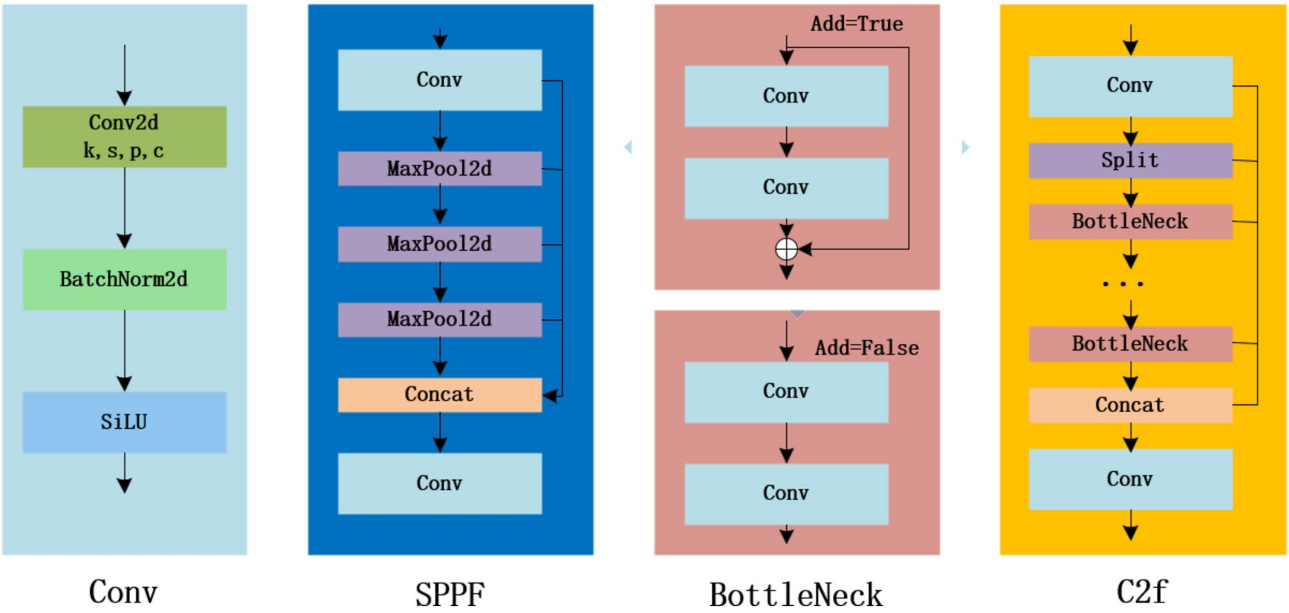

Fig. 2. Modules in YOLOv8.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

The computational complexity of the transformer model is contingent upon the dimensions of the input patches. In computer vision tasks, the number of patches is typically determined by the product of the width and height of the feature map, potentially resulting in substantial computational resource demands for the neural network model. To harness the global contextual modeling capabilities of the transformer while mitigating the computational complexity, we propose the CNN Combined Transformer (CCT) module. This module takes the feature map with the lowest resolution from the backbone as input, significantly reducing the computational resource demands introduced by the transformer on the neural network model. Concurrently, it exploits the formidable global contextual modeling capabilities of the transformer.

As depicted in Fig. 3, the CCT module initially expands the number of channels and subsequently partitions the feature map into two segments. One segment serves as the input to the transformer module, whose output feature map encompasses global contextual semantic information. The transformer-generated feature map directly concatenates with the other segment. The output feature map of the CCT module comprises both the contextual features extracted by traditional CNNs and the global features derived from Transformers. Subsequent neural network layers can thus learn the most pertinent features from this fused feature map.

Polarized self attention

The neck module of the YOLOv8 architecture integrates the polarized self attention (PSA) mechanism. This mechanism is an autonomous module designed to decouple and learn spatial and semantic associations in feature graphs. Through the polarization of the feature map, the PSA effectively segregates spatial localization information from semantic features. These decoupled features are subsequently recombined with the original features to amplify the model's representational power and perceptual accuracy. This recombination augments the representational power of the feature map, enabling the network to more precisely capture target features in complex scenes and enhance target localization and regression. Furthermore, incorporating the PSA allows the model to concentrate more effectively on crucial regions surrounding the target, thereby enhancing object detection accuracy.

The PSA integrates two pivotal designs, polarized filtering and high dynamic range (HDR), to enhance PV defect feature extraction. In this study, polarized filtering sustains high internal resolution in channel and spatial attention computations of PV defect features. It converts input features to corresponding dimensions, enabling simultaneous spatial and channel attention computations, thus reducing computational complexity. Softmax normalization of the smallest feature tensor within the attention module, facilitated by high dynamic range in PV defect detection, broadens the attention range and enhances information capture, efficiently isolating critical defective features. This operation directs attention to essential regions, accentuates faulty parts, and suppresses background noise, improving detection accuracy and robustness. Subsequently, projection mapping using a sigmoid function refines model output, bringing it closer to real-world observations. PSA enables more precise identification of minor and complex defects in PV images, improving detection accuracy and effectiveness. Applying high dynamic range processing within the polarized self-attention mechanism significantly enhances the model's capacity to manage PV defects, rendering detection results more detailed and precise.

In Fig. 4, the input feature X transforms into W_q and W_v via 1×1 convolution. W_q undergoes full channel compression while W_v maintaining a relatively high dimension (C/2). Given the compressed nature of W_q, HDR augmentation is crucial, achieved through Softmax. Following this, W_q and W_v undergo matrix multiplication,

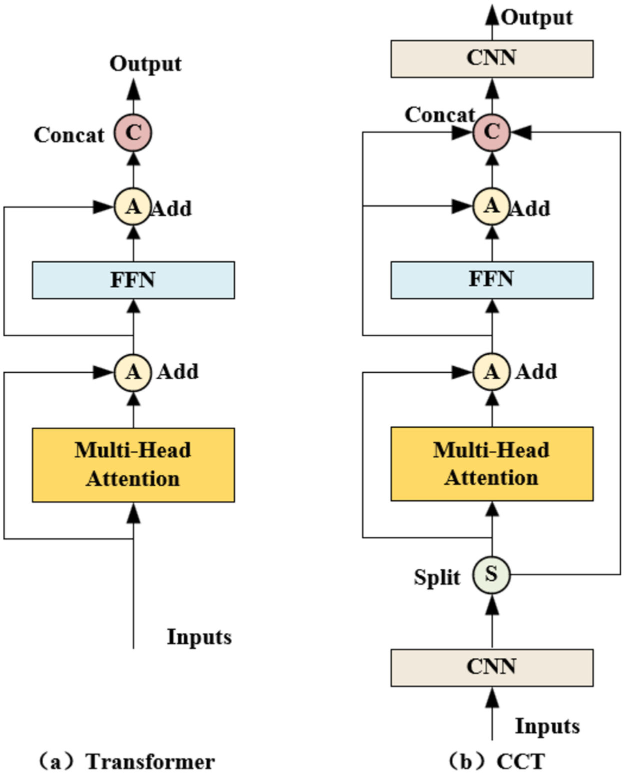

Fig. 3. Transformer and CNN combined transformer.

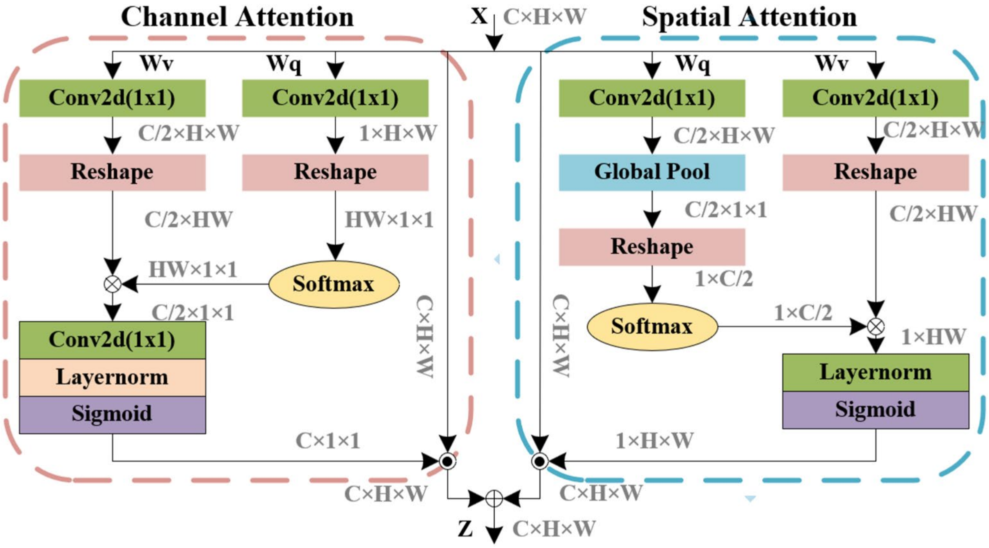

Fig. 4. Parallel distributed structure of PSA.

succeeded by 1 × 1 convolution and layer normalization, expanding the channel dimension from C/2 to C. Lastly, a sigmoid function confines all parameters within the range of 0–1. The PSA module comprises outputs from two parallel branches, as depicted in Eq. (2):

$$PSA_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \odot^{ch} X + A^{sp}(X) \odot^{sp} X. \tag{2}$$

Here, $A^{ch}(X) \in R^{C\times 1\times 1}$ represents the expression of the channel attention mechanism, detailed in Eq. (3):

$$\mathrm{A}^{ch}(X) = F_{SG}[W_{Z|\theta_1}((\sigma_1(W_v(X)))) \times F_{SM}(\sigma_2(W_q(X)))]. \tag{3}$$

W_q, W_k, and W_v refer to 1 × 1 convolutional layers, σ_1 and σ_2 act as reshape operators, F_{SM} signifies the softmax operator, and × denotes the matrix dot product operation. The number of internal channels between W_v and W_q, denoted as W_z, is C/2 and $\odot^{ch}$ serves as the channel multiplication operator. $A^{sp}(X) \in R^{1\times H\times W}$ represents the expression for the spatial attention mechanism, outlined in Eq. (4).

$$\mathrm{A}^{sp}(X) = F_{SG}[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X)))]. \tag{4}$$

σ_1, σ_2 and σ_3 denote the three reshape operators, F_{SM} signify the softmax operator, and $F_{GP}(\cdot)$ stand for the global pooling operator. $\odot^{sp}$ symbolizes the spatial multiplication operator. F_{SM} denotes the sigmoid operator.

Experiments

Dataset

This study derives the dataset from the PVEL-AD[49] dataset, which consists of 4500 annotated images at a resolution of 1024 × 1024 pixels. It encompasses 12 distinct defect categories, including *crack, finger, black_core, thick_line, star_crack, corner, fragment, scratch, horizontal_dislocation, vertical_dislocation, printing_error*, and *short_circuit*. Figure 5 depicts visual examples of the dataset. The dataset is partitioned into training and validation sets in an 8:2 ratio, comprising 3600 images for training and 900 images for validating the model's performance.

The dataset encompasses 12 defect categories, and Table 1 summarizes the distribution of defects across these categories. Of the 4500 images, there are 7842 annotated defective instances, with fewer instances noted in *corner*,

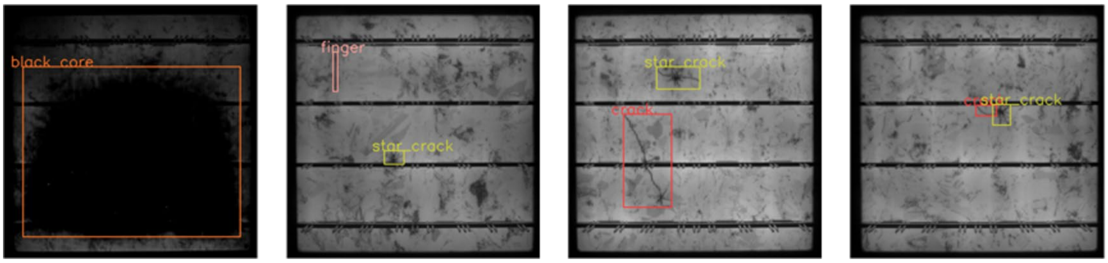

Fig. 5. Dataset examples.

Category	Anomaly object number			Image number
	Train	Val	Total	
crack	992	268	1260	4500
finger	2398	560	2958	
black_core	812	216	1028	
thick_line	807	174	981	
star_crack	108	27	135	
corner	8	1	9	
fragment	6	1	7	
scratch	4	1	5	
horizontal_dislocation	627	171	798	
vertical_dislocation	105	32	137	
printing_error	16	16	32	
short_circuit	386	106	492	
All	6269	1573	7842	4500

Table 1. Statistical overview of the PVEL-AD dataset.

fragment, scratch, and *printing_error* categories. Figure 6 illustrates further details regarding the distribution of defect categories in the training set.

Evaluation metric

To demonstrate the effectiveness of our method, we evaluate our model on the split validation dataset. We present the classical confusion matrix comprising true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). In anomaly detection tasks, TP indicates the number of true defective objects correctly predicted as defective, FP indicates the number of true normal objects correctly predicted as normal, and FN indicates the number of true defective objects incorrectly predicted as normal. Once determined, precision and recall can be calculated using Eqs. (5) and (6). Precision indicates how many of the objects predicted as defective by the model are truly defective, while recall indicates how many of the truly defective objects are predicted as

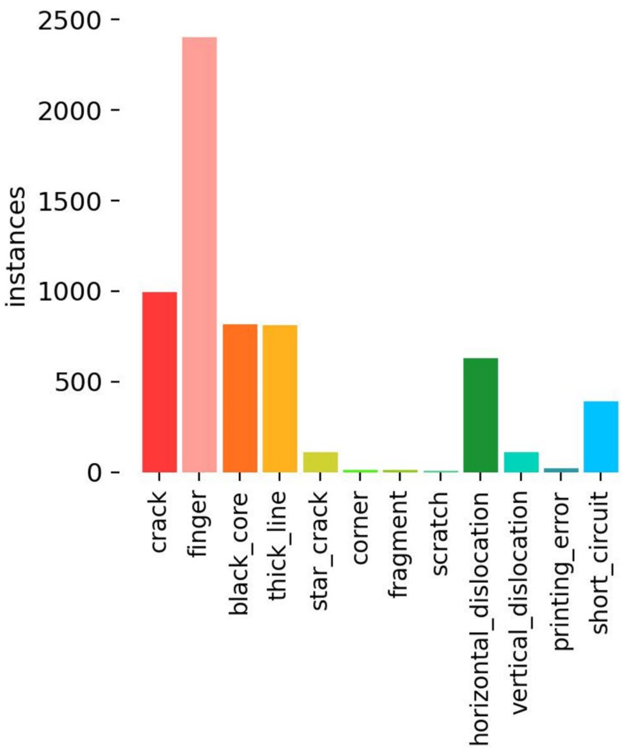

Fig. 6. The distribution of each defect category objects.

defective by the model. Finally, the F1 score (Eq. (7)) serves as the ultimate indicator, representing the harmonic average of precision and recall.

$$precision = \frac{TP}{TP + FP}, \tag{5}$$

$$recall = \frac{TP}{TP + FN}, \tag{6}$$

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall}. \tag{7}$$

We utilize Average Precision (AP) and mean Average Precision (mAP) to evaluate defect detection results. AP evaluates individual class objects, while mAP represents the mean AP across all classes. The Precision/Recall (P/R) curve provides an intuitive visualization to assess anomaly detection's effectiveness directly. Precision, recall, and Intersection over Union (IoU) calculate the AP value. We derive the AP value by computing the area under the Precision/Recall (P/R) curve. We assess AP across various IoUs, calculating it for 10 IoU thresholds ranging from 50 to 95% in 5% increments, typically reported as AP50:95. We use parameters and frames per second (FPS) as metrics to evaluate temporal efficiency.

Experimental environment

We conduct experiments on the curated PVEL-AD dataset, training model parameters on the training set, and validation on the validation set. To ensure experimental fairness, identical hardware and software configurations are employed across all experiments. The experimental hardware includes an Intel® Core™ i9-12900 K Processor, 32 GB RAM, a single RTX 3060 GPU, and the Windows 11 operating system.

We trained all experimental models using identical hyperparameters. Hyperparameters include a maximum of 100 epochs, a base learning rate of 0.000625, a batch size of 16 for a single RTX 3060 GPU, an AdamW optimizer, and an input resolution of 640 × 640 for model training and evaluation.

Ablation study

To demonstrate the effectiveness of the PSA module, we employ Gradient Class Activation Mapping (Grad-CAM)[50] to visualize the learned features of the model in the dataset. Figure 7 depicts the results of feature map visualization using Grad-CAM for the YOLOv8 and YOLOv8 model incorporating the PSA module, highlighting the enhanced focus on defective regions within the PV images.

To illustrate the effectiveness of the proposed model, we assess the performance metrics of the YOLOv8 variants, encompassing mAP50, F1 score, and mAP50:95. Table 2 presents the experimental results. As shown in Table 2, including the CCT module in YOLOv8 enhances the mAP50 metric by 5.4% and the mAP50:95 metric by 4.7% compared to the baseline YOLOv8. Incorporating the PSA module into YOLOv8 enhances the mAP50

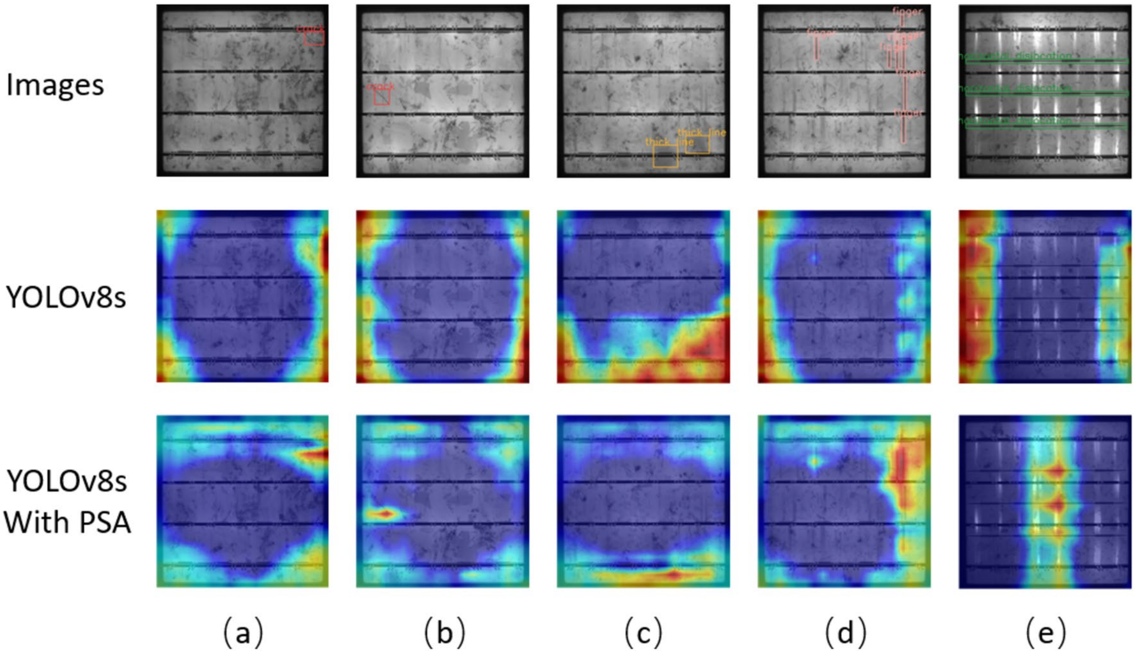

Fig. 7. Gradient Class Activation Mapping (Grad-CAM) visualization results (Grad-CAM is available at: https://github.com/jacobgil/pytorch-grad-cam). We compare the visualization results of our proposed method PSA-integrated network (YOLOv8 with PSA) with the baseline network (YOLOv8) on the validation dataset.

Method	CCT	PSA	F1 score	mAP50 (%)	mAP50:95 (%)
YOLOv8			0.61	0.64	0.41
	✓		0.667	0.694	0.457
		✓	0.679	0.648	0.432
	✓	✓	0.697	0.779	0.496

Table 2. Ablation study.

metric by 0.8% and the mAP50:95 metric by 2.2% compared to the baseline YOLOv8. The YOLOv8 with CCT and PSA modules improves the mAP50 metric by 13.9% and the mAP50:95 metric by 8.6% over the original YOLOv8. The ablation study demonstrates that our CCT and PSA modules enhance the detection accuracy of YOLOv8 in photovoltaic cell anomaly detection tasks.

Results comparison and analysis

Our primary objective is to develop an algorithm suitable for defect detection on industrial production lines. Despite the outstanding performance of Faster RCNN and Cascade RCNN, their detection speeds are insufficient for production lines with constrained computing resources. Therefore, we center our analysis on YOLO algorithms capable of real-time defect detection.

We compared the proposed model with Faster RCNN, Faster RCNN with CAN[22], Cascade RCNN, YOLOv3, YOLOv5u, YOLOv6u, and YOLOv8 in Table 3. Compared to YOLOv3-tiny/YOLOv5u-s/YOLOv6u/YOLOv8, the proposed method significantly improves F1 score by 8.4%/2.4%/6.5%/8.7%, increases mAP50 by 15.5%/17.2%/20%/16.9% and mAP50:95 by 22.1%/8.4%/9.7%/8.6%. Based on the experimental results, our proposed method achieves a mAP50 of 77.9%, surpassing YOLOv8, which achieves 61.0% mAP50, demonstrating superior performance in F1 score, mAP50, and mAP50:95 compared to other YOLO detectors. Although the FPS of the proposed method slightly decreases compared with the baseline, it still meets the real-time demands of production lines and achieves a significant performance improvement.

Figure 8 depicts the progression of mAP50 and mAP50:95 metrics across YOLOv3-tiny, YOLOv5u-s, YOLOv6u-s, YOLOv8s, and our proposed model during training on the validation set. It is apparent that as the metrics stabilize, our proposed model consistently surpasses the others in both mAP50 and mAP50:95 metrics.

Table 4 presents the mAP50:95 metrics of various models across each category in the validation set. We attribute the suboptimal detection performance for *corner, fragment, scratch*, and *vertical_dislocation* across all models to the limitations of the training samples and the model's difficulty in learning the corresponding features. The proposed model consistently outperforms others in the remaining categories.

Figure 9 illustrates the confusion matrix depicting the detection results of the proposed model on the validation dataset, offering detailed insights into its performance across different categories. The horizontal axis represents the number of defects in each category within the validation set, while the vertical axis shows the number of defects predicted by the model. It demonstrates the model's ability to detect defects such as *crack, finger, black_core, thick_line, star_crack, horizontal_dislocation, printing_error,* and *short_circuit* across most targets. However, it also indicates instances where the model may misidentify or fail to detect targets in specific other categories. For further comparative analysis across models, Fig. 10 showcases the prediction results of different models on three distinct PV defect images. The figure demonstrates that our proposed model achieves detection results closer to the ground truth (GT) across different images. Compared to other models, ours exhibits fewer erroneous and missed detections.

Although deep learning-based defect detection algorithms generally achieve good results, difficult cases still occur under certain conditions. Figure 11 illustrates the problematic cases detected by various models on PV defect images. The most frequent types of difficult cases are missed detections and misclassifications. As shown

Method	FPS	Params (M)	F1 score	mAP50 (%)	mAP50:95 (%)
Faster RCNN	15.06	41.4	0.73	68.6	47.0
Faster RCNN_CAN	12.54	42.1	0.82	77.9	51.9
Cascade RCNN	9.832	69.2	0.77	79.7	44.9
YOLOv3-tiny	136.41	8.7	0.61	62.4	27.5
YOLOv5u-s	151.52	9.1	0.67	60.7	41.2
YOLOv6u-s	140.13	16.3	0.63	57.9	39.9
YOLOv8s	149.22	11.7	0.61	61.0	41.0
YOLOv8s_CCT	142.81	12.1	0.66	69.4	45.7
YOLOv8s_PSA	135.15	11.9	0.67	64.8	43.2
Ours	133.33	12.9	0.69	77.9	49.6

Table 3. The metrics of models.

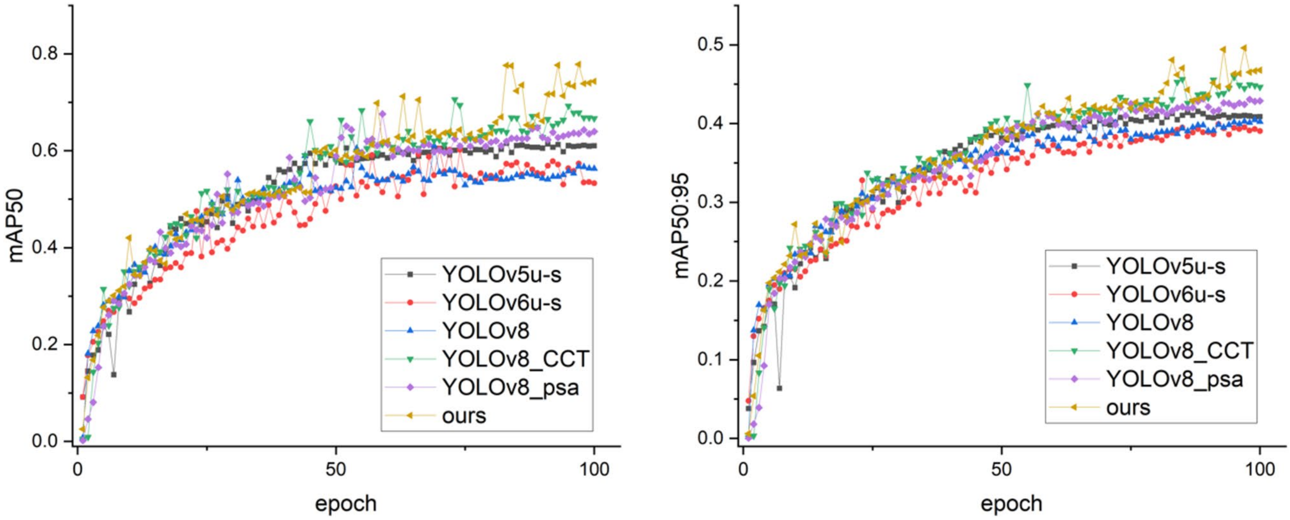

Fig. 8. The change process of mAP50 and mAP50:95 metrics of different models on the validation dataset across epochs.

Defeat classes	Yolov5u-s	Yolov6u-s	Yolov8s	Ours
crack	0.479	0.498	0.47	0.526
finger	0.557	0.551	0.542	0.559
black_core	0.956	0.962	0.96	0.965
thick_line	0.49	0.486	0.5	0.493
star_crack	0.483	0.504	0.455	0.491
corner	0.0225	0.0995	0.398	0.126
fragment	0.0221	0.0663	0.0193	0.796
scratch	0	0	0	0
horizontal_dislocation	0.328	0	0	0.376
vertical_dislocation	0	0	0	0
printing_error	0629	0.639	0.62	0.641
short_circuit	0.983	0.982	0.955	0.984
All	0.412	0.399	0.41	0.496

Table 4. mAP50:95 for different PV detection models on the validation dataset.

in Fig. 11a, missed detection occurs when the model fails to detect defects due to significant differences in visual features, such as uneven brightness distribution, between the image and the rest of the dataset. Figure 11b,c depict misclassifications, where the model fails to accurately identify defects due to the scarcity of training samples for specific defect types, such as *corner* and *fragment* defects. Current models exhibit poor performance when there is a significant difference between the distribution of prediction data and training data or when training samples are minimal. These two scenarios pose a substantial challenge for the industrial application of deep learning technology. Nevertheless, developing zero-shot and few-shot learning techniques may provide new solutions to these issues.

Conclusion

This paper proposes a novel PV defect detection method using attention mechanisms and transformers within the YOLOv8 object detection framework. The method enhances spatial and semantic information by employing the polarization self-attention mechanism during feature extraction, which integrates with original features to augment the network's representative capacity. Furthermore, the CCT module enhances the model's capability to extract contextual information, thereby improving the localization accuracy of PV defects and enhancing detection performance. Experimental results on the PVEL-AD dataset demonstrate that our method enhances the mAP50 metric by 13.9%, mAP50:95 by 8.6%, and F1 by 8.7% compared to the baseline. Existing methods still face significant limitations when there is an inconsistent distribution between predicted and training samples and when training samples are limited, highlighting a valuable problem to address.

In conclusion, our method exhibits the potential for direct deployment in actual production lines. The code is available at https://github.com/justld/PV-Detector.

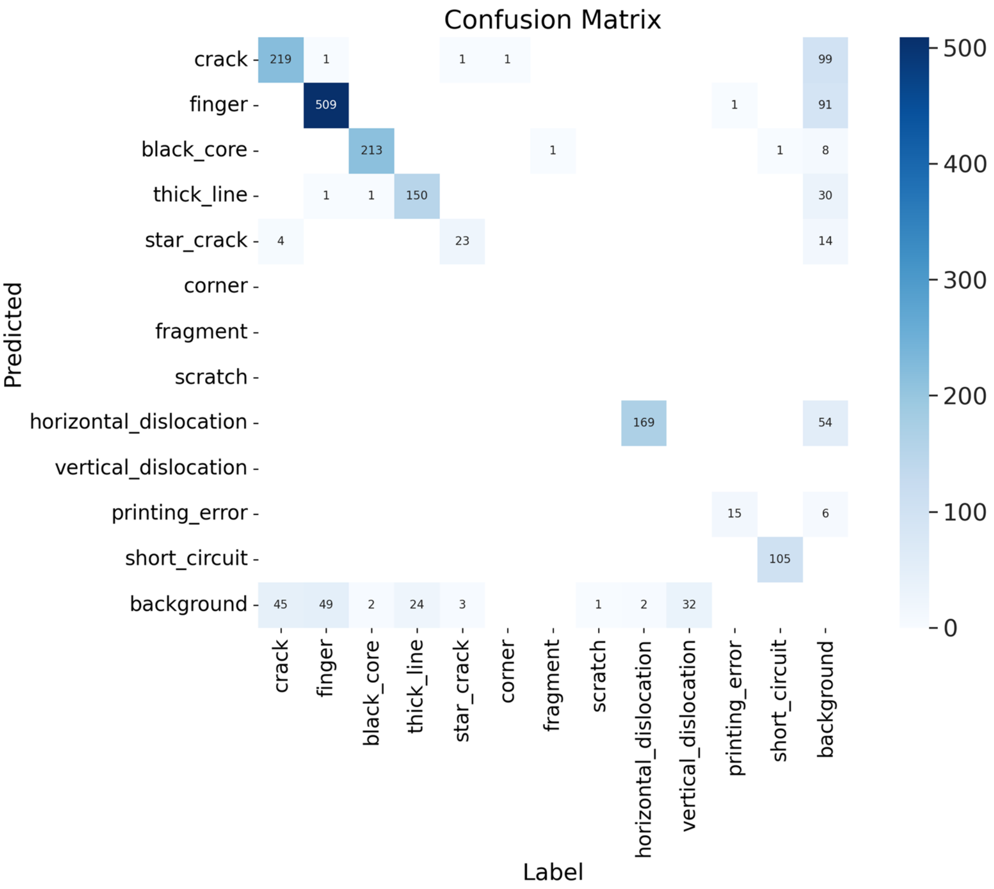

Fig. 9. The confusion matrix of the proposed model.

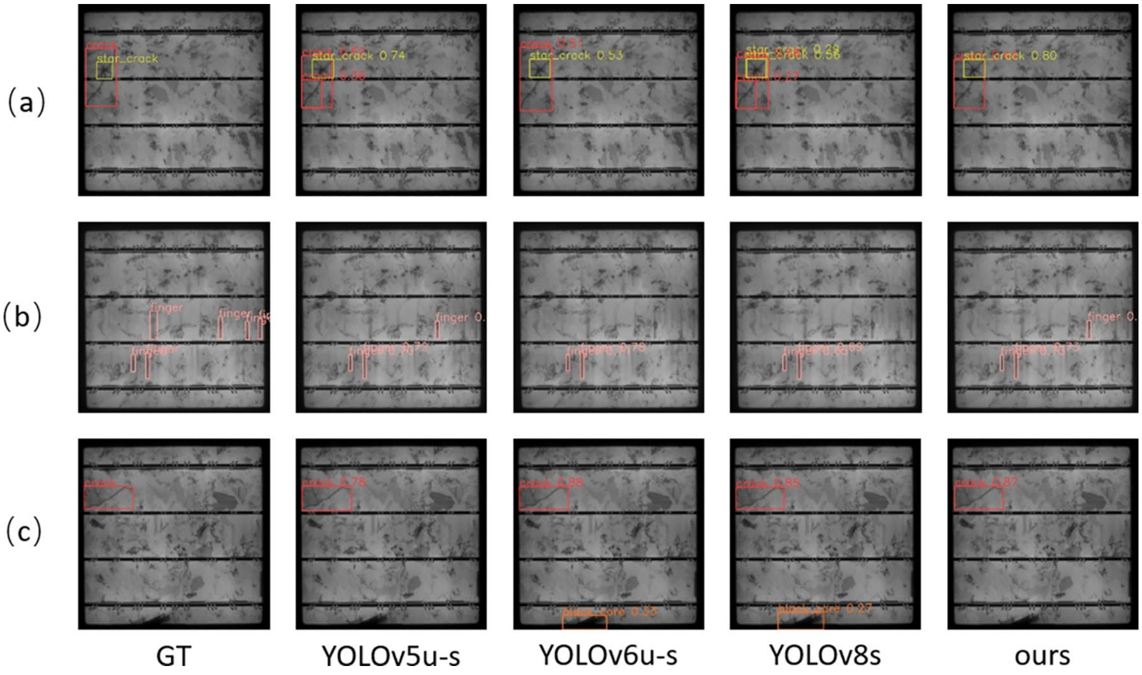

Fig. 10. The predicted results of different models on different defect images.

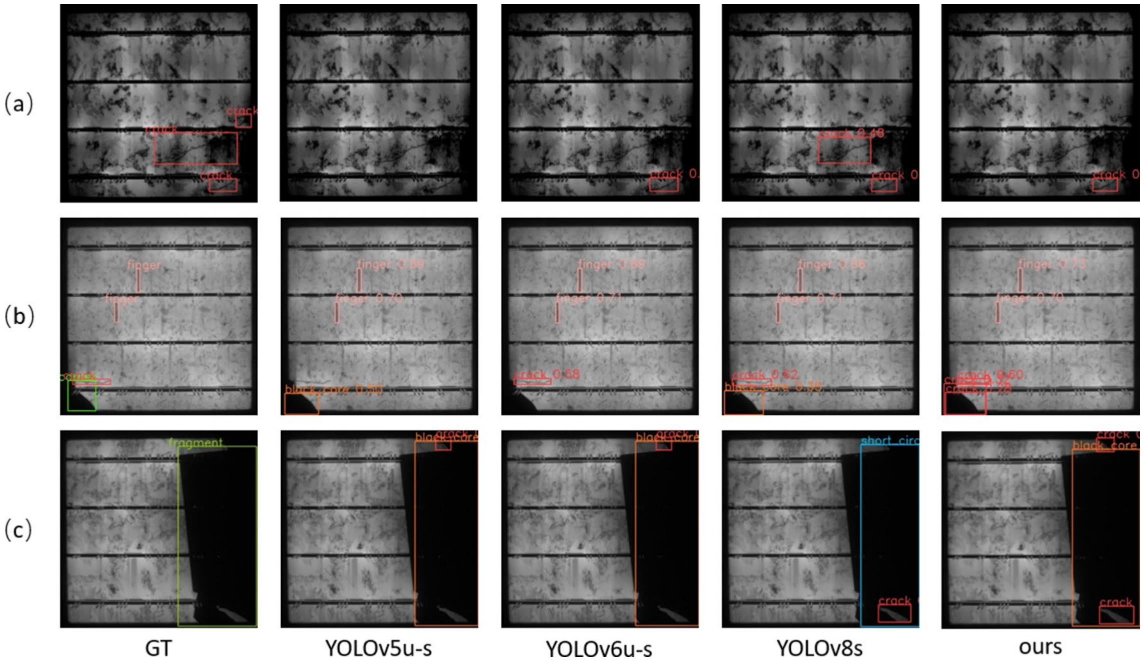

Fig. 11. Difficult cases predicted by different models on different defect images.

Data availability

The datasets analysed during the current study are available in the [PVEL-AD] repository [https://github.com/binyisu/PVEL-AD].

Received: 8 July 2024; Accepted: 3 September 2024
Published online: 05 September 2024

References

1. Fuyuki, T. & Kitiyanan, A. Photographic diagnosis of crystalline silicon solar cells utilizing electroluminescence. *Appl. Phys. A* **96**, 189–196 (2009).
2. Dallan, B. S., Schumann, J. & Lesage, F. Performance evaluation of a photoelectric–thermoelectric cogeneration hybrid system. *Solar Energy* **118**, 276–285 (2015).
3. Anwar, S. A. & Abdullah, M. Z. Micro-crack detection of multicrystalline solar cells featuring an improved anisotropic diffusion filter and image segmentation technique. *EURASIP J. Image Video Process.* **2014**, 1–17 (2014).
4. Mukherjee, S. & Acton, S. T. Oriented filters for vessel contrast enhancement with local directional evidence. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* 503–506 (IEEE, 2015).
5. Chen, H., Song, M., Zhang, Z. & Liu, K. Detection of surface defects in solar cells by bidirectional-path feature pyramid group-wise attention detector. *IEEE Trans. Instrum. Meas.* **71**, 1–9 (2022).
6. Gan, D., Quan, L., Zhu, F., Xie, K. & Bai, J. Orthogonal modulation based light beam induced current method for anti-noise defect detection in photovoltaic cells. *Solar Energy* **245**, 158–164 (2022).
7. El Yanboiy, N. *et al.* Enhancing the reliability and efficiency of solar systems through fault detection in solar cells using electroluminescence (EL) images and YOLO version 5.0 algorithm. In *Sustainable and Green Technologies for Water and Environmental Management* 35–43 (Springer, 2024).
8. Mery, D. & Arteta, C. Automatic defect recognition in X-ray testing using computer vision. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1026–1035 (IEEE, 2017).
9. Chen, J., Wen, Y., Nanehkaran, Y. A., Zhang, D. & Zeb, A. Multiscale attention networks for pavement defect detection. *IEEE Trans. Instrum. Meas.* **72**, 1–12 (2023).
10. Huang, H., Tang, X., Wen, F. & Jin, X. Small object detection method with shallow feature fusion network for chip surface defect detection. *Sci. Rep.* **12**, 3914 (2022).
11. Wang, X., Gao, H., Jia, Z. & Li, Z. BL-YOLOv8: An improved road defect detection model based on YOLOv8. *Sensors* **23**, 8361 (2023).
12. Gawde, S. *et al.* Multi-fault diagnosis of industrial rotating machines using Data-driven approach: A review of two decades of research. *Eng. Appl. Artif. Intell.* **123**, 106139 (2023).
13. Jiang, J. *et al.* A digital twin auxiliary approach based on adaptive sparse attention network for diesel engine fault diagnosis. *Sci. Rep.* **12**, 675 (2022).
14. Jia, M., Xu, D., Yang, T., Liu, Y. & Yao, Y. Graph convolutional network soft sensor for process quality prediction. *J. Process Control* **123**, 12–25 (2023).
15. Liu, S. & Sun, W. Attention mechanism-aided data-and knowledge-driven soft sensors for predicting blast furnace gas generation. *Energy* **262**, 125498 (2023).
16. Abbes, W., Elleuch, J. F. & Sellami, D. Defect-Net: A new CNN model for steel surface defect classification. In *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)* 1–5 (IEEE, 2024).
17. Farady, I., Lin, C.-Y. & Chang, M.-C. PreAugNet: Improve data augmentation for industrial defect classification with small-scale training data. *J. Intell. Manuf.* **35**, 1233–1246 (2024).
18. Deitsch, S. *et al.* Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy* **185**, 455–468 (2019).

19. Chen, H., Pang, Y., Hu, Q. & Liu, K. Solar cell surface defect inspection based on multispectral convolutional neural network. *J. Intell. Manuf.* **31**, 453–468 (2020).
20. Han, H. *et al.* Polycrystalline silicon wafer defect segmentation based on deep convolutional neural networks. *Pattern Recogn. Lett.* **130**, 234–241 (2020).
21. Otamendi, U. *et al.* Segmentation of cell-level anomalies in electroluminescence images of photovoltaic modules. *Solar Energy* **220**, 914–926. https://doi.org/10.1016/j.solener.2021.03.058 (2021).
22. Su, B. *et al.* Deep learning-based solar-cell manufacturing defect detection with complementary attention network. *IEEE Trans. Ind. Inform.* **17**, 4084–4095. https://doi.org/10.1109/tii.2020.3008021 (2021).
23. Shin, K.-S., Kim, J.-C. & Shin, S.-Y. Anomaly detection system for solar power distribution panels utilizing thermal images. *J. Inf. Commun. Converg. Eng.* **22**, 1 (2024).
24. Jha, S. B. & Babiceanu, R. F. Deep CNN-based visual defect detection: Survey of current literature. *Comput. Ind.* **148**, 103911 (2023).
25. Larochelle, H. & Hinton, G. E. Learning to combine foveal glimpses with a third-order Boltzmann machine. *Adv. Neural Inf. Process. Syst.* **23**, 1 (2010).
26. Zhao, H. & Zhang, T. Fabric surface defect detection using SE-SSDNet. *Symmetry* **14**, 2373 (2022).
27. Chen, M., Chen, S., Wang, S., Cui, Y. & Chen, P. Accurate segmentation of small targets for LCD defects using deep convolutional neural networks. *J. Soc. Inf. Display* **31**, 13–25 (2023).
28. Chen, L., Yao, H., Fu, J. & Ng, C. T. The classification and localization of crack using lightweight convolutional neural network with CBAM. *Eng. Struct.* **275**, 115291 (2023).
29. Li, T., Xu, H. & Bai, J. A lightweight safety helmet detection network based on bidirectional connection module and polarized self-attention. In *International Conference on Neural Information Processing* 253–264 (Springer, 2023).
30. Wang, H.-K., Zhu, P.-J., Du, J. & Zhang, X. EPF-Net: An anomaly detection model with an enhanced polarized self-attention. In *13th International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE 2023)* 500–504 (IET, 2023).
31. Wu, P., Hua, Z. & Li, J. PATN: Polarized attention based transformer network for multi-focus image fusion. *KSII Trans. Internet Inf. Syst.* **17**, 4 (2023).
32. Dosovitskiy, A. *et al.* An image is worth 16 × 16 words: Transformers for image recognition at scale (2020).
33. Carion, N. *et al.* End-to-end object detection with transformers. In *European Conference on Computer Vision* 213–229 (Springer, 2020).
34. Xie, E. *et al.* SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).
35. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision* 10012–10022 (2021).
36. Mehta, S. & Rastegari, M. J. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer (2021).
37. Sohan, M., Sai Ram, T., Reddy, R. & Venkata, C. A review on YOLOv8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics* 529–545 (Springer, 2024).
38. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement (2018).
39. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection (2020).
40. Wang, C.-Y. *et al.* CSPNet: A new backbone that can enhance learning capability of CNN. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 390–391 (2020).
41. Li, C. *et al.* Yolov6 v3.0: A full-scale reloading (2023).
42. Tian, Z., Chu, X., Wang, X., Wei, X. & Shen, C. Fully convolutional one-stage 3d object detection on lidar range images. *Adv. Neural Inf. Process. Syst.* **35**, 34899–34911 (2022).
43. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1904–1916 (2015).
44. Yan, P. *et al.* Detection of coal and gangue based on improved YOLOv5.1 which embedded scSE module. *Measurement* **188**, 110530 (2022).
45. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
46. Choi, S. R. & Lee, M. Transformer architecture and attention mechanisms in genome data analysis: A comprehensive review. *Biology* **12**, 1033 (2023).
47. Wang, X. & Zhu, Z. Context understanding in computer vision: A survey. *Comput. Vis. Image Understand.* **229**, 103646 (2023).
48. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 8759–8768 (2018).
49. Su, B., Zhou, Z. & Chen, H. PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection. *IEEE Trans. Ind. Inform.* **19**, 404–413. https://doi.org/10.1109/tii.2022.3162846 (2023).
50. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision* 618–626 (2017).

Author contributions

Du Lang wrote the main manuscript text and completed all the experiments. Zhenzhen Lv drew the figures in the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.