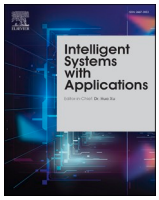
Intelligent Systems with Applications 24 (2024) 200443

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



Optimizing feature extraction and fusion for high-resolution defect detection in solar cells

Hoanh Nguyen *, Tuan Anh Nguyen, Nguyen Duc Toan

Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City 700000, Vietnam

ARTICLE INFO

Keywords:
Feature extraction
Feature fusion
Swin transformer
Defect detection
Electroluminescent images

ABSTRACT

In this paper, we propose a novel architecture for defect detection in electroluminescent images of polycrystalline silicon solar cells, addressing the challenges posed by subtle and dispersed defects. Our model, based on a modified Swin Transformer, incorporates key innovations that enhance feature extraction and fusion. We replace the conventional self-attention mechanism with a novel group self-attention mechanism, increasing the mAP50:5:95 score from 50.12 % to 52.98 % while reducing inference time from 74 ms to 62 ms. We also introduce a spatial displacement with shift convolution module, replacing the traditional Multi-Layer Perceptron, which further enhances the model's receptive field and improves precision and recall. Additionally, our fast multi-scale feature fusion mechanism effectively combines high-resolution details with high-level semantic features from different network layers, optimizing defect detection accuracy. Experimental results on the PVEL-AD dataset demonstrate that our model achieves the highest mAP50 score of 83.11 % and an F1-Score of 84.33 %, surpassing state-of-the-art models while maintaining a competitive inference time of 66.3 ms. These findings highlight the effectiveness of our innovations in improving defect detection accuracy and computational efficiency, making our model a robust solution for quality assurance in solar cell manufacturing.

1. Introduction

In recent years, the global energy landscape has witnessed a significant shift toward renewable energy sources, driven by the increasing awareness of environmental issues and the depletion of non-renewable resources such as fossil fuels (Zhang et al., 2023; Wei et al., 2022). Solar photovoltaic (PV) technology has emerged as a leading solution to address these challenges due to its inherent advantages, including low production costs, minimal environmental impact, and the ability to harness an abundant and renewable energy source: the sun. As a result, the adoption of solar energy has expanded rapidly, leading to a corresponding increase in the production and deployment of solar cells worldwide.

Among the different types of solar cells, polycrystalline silicon cells are widely used due to their cost-effectiveness and adequate efficiency levels. However, the manufacturing process of these cells is prone to various defects, such as cracks, grid breaks, stains, and other surface irregularities. These defects can arise from multiple factors, including material impurities, mechanical stresses during production, or thermal cycling. Detecting these defects is crucial, as they can significantly degrade the performance, reliability, and lifespan of PV systems. For instance, cracks can lead to power losses by interrupting the current flow, while stains and discolorations can reduce the cell's light absorption efficiency. Therefore, ensuring the quality of solar cells through effective defect detection is essential for maintaining the stability and efficiency of PV power generation systems.

Traditional defect detection methods, which often rely on manual inspection or simple image processing techniques, struggle to meet the demands of modern solar cell manufacturing. These methods are typically labor-intensive, time-consuming, and prone to human error, particularly when dealing with the complex and varied nature of defects in polycrystalline silicon cells. Moreover, the electroluminescent images (EL) used for defect detection in solar cells present additional challenges due to their high resolution and the subtle, dispersed nature of many defects. These challenges have spurred the development of automated, deep learning-based approaches that can provide more accurate, efficient, and scalable solutions for defect detection in solar cells.

Deep learning, with its ability to automatically learn and extract relevant features from large datasets, has shown great promise in addressing the limitations of traditional methods. In particular,

* Corresponding author.
E-mail address: nguyenhoanh@iuh.edu.vn (H. Nguyen).

https://doi.org/10.1016/j.iswa.2024.200443
Received 16 August 2024; Received in revised form 1 September 2024; Accepted 17 September 2024
Available online 26 September 2024

convolutional neural networks (CNNs) have been widely adopted for image-based defect detection tasks due to their powerful feature extraction capabilities (Chirgaiya & Rajavat, 2023; Qaddour & Siddiqa, 2023; Elhaija & Al-Haija, 2023; Xiao et al., 2023; Yu et al., 2023; Zhang et al., 2024). However, standard CNN architectures often face difficulties in detecting small and multi-scale defects, which are common in EL images of polycrystalline silicon solar cells. These defects may occupy only a small portion of the image or exhibit varying scales, making them challenging to detect with conventional CNN-based models. Furthermore, the diverse and complex textures of polycrystalline silicon surfaces, characterized by randomly shaped and sized crystalline particles, add another layer of difficulty to the defect detection task.

Recent studies have explored advanced deep learning techniques to improve feature extraction and defect detection across various domains. For instance, region-based convolutional neural networks (R-CNNs) have been employed effectively in the segmentation and detection of text in complex images with noisy backgrounds, as demonstrated in the work of (Preethi & Mamatha, 2023). Similarly, the effectiveness of CNNs in handling both structured and unstructured data, as shown in the recognition of Devanagari digits (Bhosle & Musande, 2023), further underscores the versatility of CNN architectures. Additionally, techniques such as semi-supervised learning and adaptive feature extraction have shown promise in handling small sample datasets, as seen in the classification of hyperspectral images in IoT applications (Chen et al., 2024). Moreover, addressing imbalanced data distributions and enhancing model robustness, as explored through variational autoencoders and CNNs with focal loss (Li et al., 2024), has influenced recent approaches to ensuring that models remain effective even when defect occurrences are rare or unevenly distributed. Finally, the integration of deep learning with computer-aided engineering (CAE) for simulating real-world conditions, as discussed in Akande, Alabi and Ajagbe (2022), emphasizes the potential for deep learning models to accurately predict and optimize physical phenomena.

To address the challenges of detecting subtle and dispersed defects in electroluminescent images of polycrystalline silicon solar cells, we propose a novel deep learning architecture. This architecture incorporates key innovations to enhance both the accuracy and efficiency of defect detection. A central feature of our model is the group self-attention mechanism, which improves upon the traditional self-attention mechanism by dividing the attention computation into smaller groups. This reduces computational complexity while improving the model's sensitivity to fine-grained details, making it particularly effective for identifying small defects in high-resolution images. In addition to group self-attention mechanism, our architecture includes a spatial displacement with shift convolution module, which expands the receptive field by introducing spatial shifts before convolution operations. This approach allows the model to capture broader contextual information, crucial for detecting dispersed defects that might span different areas of the solar cell image. Finally, we incorporate a fast multi-scale feature fusion mechanism, which combines high-resolution details from shallow layers with deep semantic information, optimizing the detection of defects across various scales. These components create a robust and efficient defect detection model capable of meeting the real-time demands of industrial applications. The primary contributions of this paper are summarized as follows:

- We propose a new deep learning architecture specifically designed for defect detection in electroluminescent images of polycrystalline silicon solar cells. This architecture integrates three key innovations: a group self-attention mechanism, a spatial displacement with shift convolution module, and a fast multi-scale feature fusion mechanism, leading to significant improvements in detection accuracy and computational efficiency.
- We introduce a novel group self-attention mechanism that replaces the traditional self-attention mechanism in the Swin Transformer block to enhance computational efficiency and capture finer details. This mechanism subdivides each window into smaller groups, which significantly reduces the computational complexity while maintaining the model's effectiveness. By focusing attention on these smaller groups, the mechanism improves the model's ability to detect fine-grained details, which is crucial for identifying subtle defects in high-resolution images.
- We propose the spatial displacement with shift convolution (SDSC) module, which expands the receptive field by incorporating spatial shifts. This allows the model to capture broader contextual information, improving its ability to accurately detect dispersed defects across the entire solar cell image.
- We develop a fast multi-scale feature fusion (FMFF) mechanism that effectively combines high-resolution details with high-level semantic information, enabling robust defect detection across various scales and ensuring that both small and large defects are accurately identified. The FMFF mechanism integrates two key components: a weighted fusion of features mechanism and an enhanced feature extraction layer. These components are meticulously designed to leverage features from different layers of the neural network, addressing the trade-off between capturing fine details and understanding broader contextual information.
- Extensive experiments on the PVEL-AD dataset demonstrate that our proposed model not only surpasses state-of-the-art methods in terms of accuracy (achieving a mAP50 score of 83.11 %) but also maintains speeds comparable to existing methods, making it suitable for real-time defect detection in industrial applications.

2. Related work

Defect detection in solar cells is a critical task that has attracted significant attention due to the increasing demand for high-quality solar photovoltaic systems. Traditional methods for detecting defects in solar cells often involve manual inspection or basic image processing techniques, which are labor-intensive, time-consuming, and prone to inaccuracies. With the advent of deep learning, more sophisticated and automated approaches have been developed, offering improvements in accuracy, speed, and scalability. This section reviews the existing work on solar cell defect detection, focusing on both conventional methods and recent advancements in deep learning-based approaches.

Early attempts at defect detection in solar cells largely relied on conventional image processing techniques, such as edge detection, thresholding, and morphological operations. In Tsai, Wu and Chiu (2012), the authors proposed a method for defect inspection in solar modules using electroluminescence images, which leverages independent component analysis to detect defects. Their approach involves training on defect-free subimages to create basis images and then using these to reconstruct test images, with classification based on the reconstruction error and linear combination coefficients. Su et al. (Su et al., 2019) introduced a novel feature descriptor, CPICS-LBP, which enhances defect detection in near-infrared electroluminescence images of multicrystalline solar cells by fusing gradient information and local binary patterns to address challenges from random crystal grain shapes and intensity variations. They further developed a robust feature extraction method, the bag of CPICS-LBP, which improves classification accuracy and time efficiency, achieving state-of-the-art results in defect recognition. Chen et al. (Chen, Zhao, Han & Liu, 2019) presented a robust crack detection scheme for multicrystalline solar cells in EL images, addressing challenges like heterogeneous backgrounds and low contrast. Their method employs a novel steerable evidence filter to enhance crack saliency, followed by segmentation and skeletonization techniques. Traditional methods, while straightforward, often struggled to handle the variability and complexity of defects in polycrystalline silicon solar cells, where defects can vary greatly in size, shape, and location. The rise of machine learning introduced more advanced techniques, such as support vector machines and random forests, which offered improvements in defect classification but still required extensive

feature engineering and were limited in their ability to detect small or subtle defects.

In recent years, deep learning has revolutionized the field of defect detection, particularly through the use of CNNs. CNNs have shown remarkable success in image-based tasks due to their ability to automatically learn hierarchical features from raw data. Among the various CNN architectures, two-stage networks like Faster R-CNN (Girshick, 2015) and one-stage networks like YOLO (Redmon, Divvala, Girshick & Farhadi, 2016; Redmon & Farhadi, 2017; Redmon & Farhadi, 2018; Bochkovskiy, 2020; Li et al., 2022; Wang, Bochkovskiy & Liao, 2023) and SSD (Liu et al., 2016) have been widely adopted for defect detection in solar cells. Two-stage networks outperform in accuracy by generating region proposals that are then refined, making them suitable for detecting defects in high-resolution images. However, their real-time deployment is limited due to high computational costs, especially in resource-constrained environments. One-stage networks have gained popularity for their balance between accuracy and speed. YOLO models are particularly favored in industrial applications where real-time processing is essential. The YOLOv5 model, for instance, has been extensively used in solar cell defect detection due to its efficient deployment on edge devices and its ability to maintain high detection accuracy. Despite these advancements, challenges remain in detecting small and multi-scale defects, which are prevalent in polycrystalline silicon solar cells. These defects often occupy a small portion of the image or are dispersed across different regions, making them difficult to detect with standard CNN architectures. Several researchers have proposed modifications to existing CNN architectures to address these challenges. For example, Su et al. (Su et al., 2020) proposed a novel complementary attention network (CAN) for automatic defect detection in solar cell EL images, combining channel-wise and spatial attention subnetworks to effectively suppress background noise and enhance defect features. By integrating CAN into a region proposal network within a faster R-CNN framework, they developed an end-to-end faster RPAN—CNN system, which demonstrated superior performance in defect classification and detection on a large-scale EL dataset. Similarly, Su et al. (Su, Chen & Zhou, 2021) developed a bidirectional attention feature pyramid network (BAFPN) to address the challenge of multiscale defect detection in photovoltaic cell EL images, enabling effective multiscale feature fusion through an attention-based top-down and bottom-up architecture. By embedding BAFPN into a novel object detector (BAF-Detector) within a Faster RCNN+FPN framework, they achieved high performance in multiscale defect classification and detection. In Deitsch et al. (2019), the authors introduced two approaches for automatic defect detection in photovoltaic module EL images: a hardware-efficient method using hand-crafted features classified by a SVM and a more hardware-demanding approach employing a deep CNN running on a GPU. While the CNN achieved higher accuracy at 88.42 %, the SVM, with an accuracy of 82.44 %, offers broader hardware compatibility, making both methods suitable for continuous, accurate monitoring of PV cells. Akram et al. (Akram et al., 2019) introduced a novel light convolutional neural network architecture for automatic defect detection in photovoltaic module EL images, achieving a state-of-the-art accuracy while requiring minimal computational power, making it suitable for real-time use on ordinary CPU computers. Extensive experimentation on various architectures and data augmentation techniques was conducted to address data scarcity and overfitting, ensuring the model's generalization and applicability in both lab and industrial settings. Qian et al. (Qian et al., 2020) proposed a novel micro-crack detection method for solar cells that combines short-term deep features, learned from the input image using a stacked denoising autoencoder, with long-term deep features, learned from natural scene images using convolutional neural networks. Ge et al. (Ge et al., 2020) developed a novel architecture called fuzzy convolution, which integrates fuzzy logic with convolution operations to handle uncertainties in PV cell data during defect detection, addressing the noise and subjectivity in human annotation. Chen et al. (Chen, Pang, Hu & Liu, 2020) designed a visual defect detection method using a multi-spectral deep CNN to address the challenges of detecting similar and indeterminate defects on solar cell surfaces with heterogeneous textures and complex backgrounds. Recently, Su et al. (Su, Zhou, Chen & Cao, 2021) proposed a novel approach for solar cell electroluminescence defect segmentation using a generative adversarial network that generates defect-free images while maintaining background consistency, allowing defects to be segmented through image subtraction and thresholding. In Rahman and Chen (2020), the authors proposed a precise defect inspection method for photovoltaic electroluminescence images using a multi-attention U-net architecture that incorporates channel and spatial attention to efficiently extract key features while suppressing background noise.

While deep learning-based methods have significantly advanced the field of solar cell defect detection, several challenges persist, particularly in detecting small and multi-scale defects in complex backgrounds. Recently, multi-scale feature extraction has been extensively explored in computer vision to improve detection and recognition tasks, as seen in the works of (Jiao et al., 2024) and (Dong, Zhang, Ji & Ding, 2020). These approaches have demonstrated the effectiveness of capturing features at different scales to enhance the detection of objects in varied contexts. For instance, Jiao et al. (Jiao et al., 2024) provided a comprehensive survey on multiscale deep learning, focusing on how multiscale representation is utilized in object detection and recognition tasks. They discuss the evolution of multiscale methods in deep learning, covering techniques such as pyramid representation, scale-space representation, and multiscale geometric representation, and compare their effectiveness across various tasks using CNNs and Vision Transformers (ViTs). Similarly, Dong et al. (Dong, Zhang, Ji & Ding, 2020) proposed a scale-aware CNN framework, named MMNet, specifically designed for crowd counting, which addresses challenges like occlusions and scale variations of people's heads. Their approach not only captures multi-scale features through varying filter sizes but also integrates these features across different network stages using multi-level density-based spatial information, significantly improving accuracy in crowd density estimation across diverse datasets. However, while these multi-scale methods have shown success in general computer vision tasks, they do not fully address the unique challenges presented by defect detection in high-resolution electroluminescent images of polycrystalline silicon solar cells. Our method builds on the foundation of multi-scale feature extraction by integrating a fast multi-scale feature fusion mechanism specifically tailored to the intricate and subtle defects found in solar cells. Unlike traditional approaches, our architecture incorporates a group self-attention mechanism that enhances the model's ability to detect fine-grained details, even within high-resolution images, and an SDSC module that expands the receptive field to better capture dispersed defects. These innovations enable our model to maintain high detection accuracy while also being computationally efficient, addressing the limitations of existing multi-scale techniques in this specialized domain.

3. Methodology

3.1. Overall architecture

Fig. 1 provides an overview of our model. The proposed model for defect detection in electroluminescent images of polycrystalline silicon solar cells is based on a modified Swin Transformer architecture. This model is designed to enhance both feature extraction and fusion, which are critical for accurately detecting defects across varying scales and complexities.

The Swin Transformer utilizes a hierarchical structure with shifted windows for local self-attention, which improves computational efficiency. However, the traditional self-attention mechanism can be computationally expensive, especially for high-resolution tasks. To address this, we introduce a novel group self-attention mechanism. This mechanism subdivides each window into smaller groups, thereby reducing the computational complexity significantly. By focusing

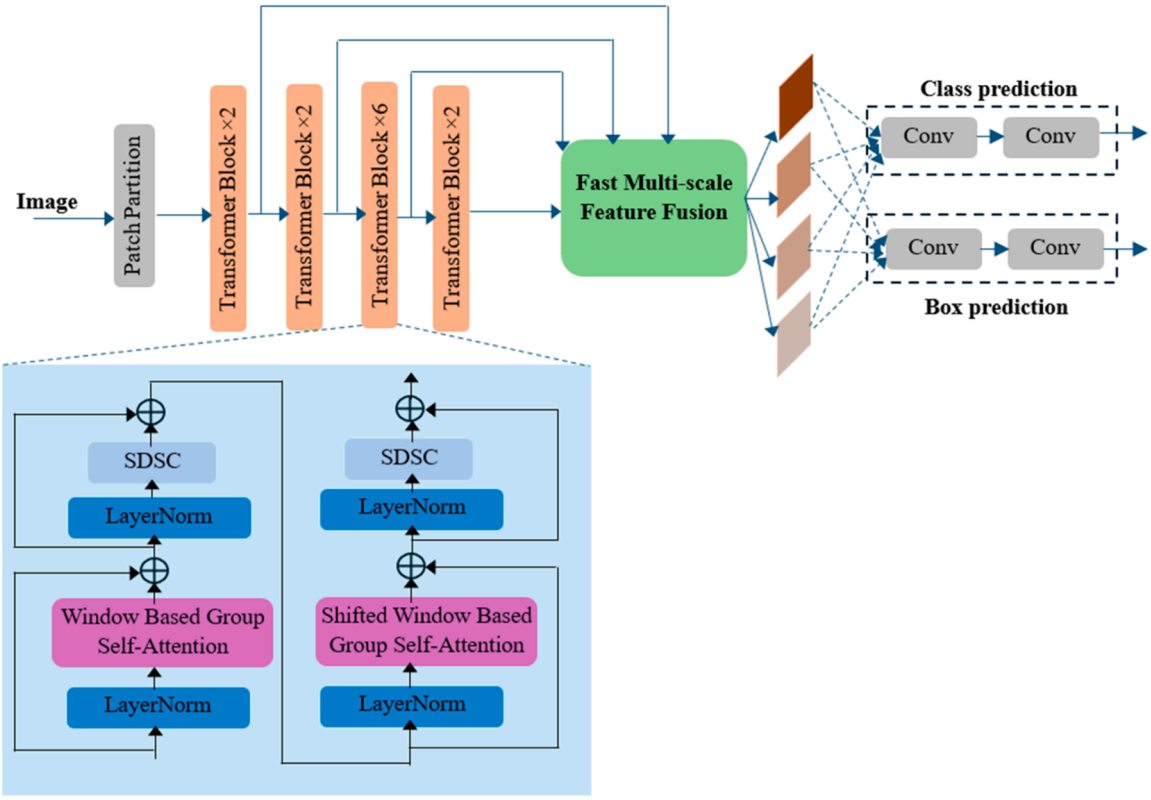


Fig. 1. The overall architecture of the proposed model.

attention on smaller regions within each window, the model captures finer details, which is crucial for detecting subtle defects in high-resolution images. The improved granularity in feature extraction ensures more accurate and reliable defect detection, while the enhanced computational efficiency allows the model to process larger images or batches more effectively. In traditional Swin Transformer blocks, the Multi-Layer Perceptron (MLP) plays a significant role in transforming features. However, it has limitations in capturing spatial relationships beyond the local scope. To overcome this, we propose the SDSC module, which replaces the MLP in the Swin Transformer block. The SDSC module enhances the receptive field by introducing spatial displacements before applying convolution operations. This module shifts the input feature maps in multiple directions, capturing broader contextual information from adjacent pixels. This process enables the network to detect subtle and dispersed defects more effectively, leading to improved performance in high-resolution defect detection tasks.

Defects in electroluminescent images can vary significantly in size, requiring a detection approach that is effective across multiple scales. To address this, we develop the FMFF mechanism, which integrates features from different layers of the network, combining high-resolution details with high-level semantic information. The FMFF mechanism includes a weighted fusion of features mechanism and an enhanced feature extraction layer (EFEL). These components are carefully designed to optimize the blending of shallow, high-resolution features with deeper, semantically rich features. The FMFF mechanism ensures robust detection of both small, intricate defects and larger, more apparent ones. This multi-scale approach significantly enhances detection precision while maintaining computational efficiency, making the model suitable for real-time industrial applications. Finally, the detection head leverages the fused multi-scale features to perform both class and box predictions, enabling accurate localization and classification of defects in the images.

Overall, the integration of these components within the modified Swin Transformer framework enhances the model's capability to detect a wide range of defects. The combination of improved feature extraction, expanded receptive fields, and efficient multi-scale feature fusion makes this model highly effective for defect detection in high-resolution electroluminescent images, providing a robust solution for quality assurance in solar cell manufacturing.

3.2. Improving feature extraction with group self-attention mechanism

The Swin Transformer (Liu et al., 2021) is a novel architecture designed to enhance the efficiency and scalability of Transformers for computer vision tasks. Unlike traditional Transformers that apply self-attention globally, the Swin Transformer introduces a hierarchical structure with shifted windows, allowing for local self-attention within non-overlapping windows and cross-window connections. This design significantly reduces the computational complexity while maintaining a high level of performance on various vision benchmarks. To understand the computational efficiency of the Swin Transformer, it is essential to delve into the complexity analysis of its window-based self-attention mechanism.

Given a feature map of dimensions $C \times H \times W$, the computational complexity of the window-based self-attention mechanism using $M \times M$ non-overlapping windows in the Swin Transformer can be derived as follows. First, the feature map is divided into non-overlapping windows, each of size $M \times M$, resulting in a total of $\frac{H \times W}{M^2}$ windows. Within each window, which contains M^2 tokens, the complexity of the self-attention operation is $O\left(\left(M^2\right)^2.C\right) = O\left(M^4.C\right)$, where M^2 represents the number of tokens in the window, and C is the number of feature channels. Summing the complexity across all windows, the total complexity for the self-attention mechanism is $\frac{H \times W}{M^2}.O\left(M^4.C\right) = O\left(HW.M^2.C\right)$. In a multi-head setup with h heads, the operations within each head are performed independently, thus the overall complexity remains $O\left(HW.M^2.C\right)$. This window-based approach significantly reduces computational complexity compared to the standard global self-

attention, which has a complexity of $O((H.W)^2.C)$, making the Swin Transformer more efficient for processing high-resolution feature maps.

In the context of Swin Transformers, M represents the size of the non-overlapping windows within which self-attention is computed locally. By partitioning a feature map of dimensions $C \times H \times W$ into smaller $M \times M$ windows, the model reduces the computational complexity of self-attention from a global perspective to a more manageable, localized one. Specifically, the total computational complexity of the window-based self-attention is $O(HW.M^2.C)$. If M is enlarged, each window will contain more tokens (M^2), increasing the computational burden within each window. Consequently, the self-attention complexity within each window rises quadratically ($O(M^4.C)$). Although the number of windows $\frac{H \times W}{M^2}$ decreases as M grows, the overall complexity still increases because the quadratic growth in M^2 for each window's computation outweighs the linear reduction in the number of windows. Therefore, while larger windows can capture more extensive contextual information within each window, they also significantly increase the computational demands.

In this paper, we design a novel group self-attention mechanism to replace the self-attention mechanism in the Swin Transformer block to enhance computational efficiency and capture finer details. This new mechanism subdivides each window into smaller groups, reducing the computational load and improving the model's ability to detect subtle features, which is particularly beneficial for high-resolution tasks such as defect detection in electroluminescent images of polycrystalline silicon solar cells.

The structure of the group self-attention mechanism is shown in Fig. 2. This mechanism operates by first dividing a feature map of dimensions $C \times H \times W$ into non-overlapping windows, each of size $M \times M$. Within each window, the mechanism further divides the tokens into G groups. Each group contains $\frac{M^2}{G}$ tokens, ensuring that the attention computation is performed on a smaller subset of tokens, thus reducing the computational complexity. Given a window of size $M \times M$, it is divided into G groups. Each group contains $\frac{M^2}{G}$ tokens. Self-attention is computed separately within each group, and the complexity of this operation is $O\left(\left(\frac{M^2}{G}\right)^2.C\right) = O\left(\frac{M^4}{G^2}.C\right)$. The results from all groups within a window are aggregated to form the final attention output for that window. This ensures that the local attention computations are integrated to provide a comprehensive representation of the window.

To calculate the computational complexity of the group self-attention mechanism, consider a feature map of dimensions $C \times H \times W$ divided into windows of size $M \times M$. The number of windows is $\frac{H \times W}{M^2}$. The self-attention within each group has a complexity of $O\left(\frac{M^4}{G^2}.C\right)$. The total complexity for all groups in a window is $G.O\left(\frac{M^4}{G^2}.C\right) = O\left(\frac{M^4}{G}.C\right)$. Summing across all windows, the total complexity for all windows is $\frac{H \times W}{M^2}.O\left(\frac{M^4}{G}.C\right) = O\left(\frac{HW.M^2}{G}.C\right)$.

In comparison, the original window-based self-attention in the Swin Transformer has a complexity of $O(HW.M^2.C)$. By introducing groups, the group self-attention mechanism can reduce the computational load by a factor of G, making it $O\left(\frac{HW.M^2}{G}.C\right)$. This demonstrates a significant improvement in computational efficiency, especially when G is large.

The group self-attention mechanism plays a crucial role in enhancing feature extraction for defect detection in electroluminescent images of polycrystalline silicon solar cells. These images often require high-resolution analysis to detect small and subtle defects. By subdividing the windows into groups, this mechanism allows the model to focus on smaller regions within each window, capturing finer details that might be missed by larger, non-grouped windows. This increased granularity in feature extraction leads to more accurate and reliable detection of defects, ultimately improving the overall performance of the defect detection system. The enhanced computational efficiency also allows for the processing of higher resolution images or larger batch sizes, further benefiting the defect detection pipeline.

3.3. Spatial displacement with shift convolution

In the Swin Transformer, the Multi-Layer Perceptron (MLP) is a crucial component used after the self-attention mechanism within each Swin Transformer block. The MLP is responsible for further transforming the features and enhancing the representation learning capabilities of the network. Typically, the MLP consists of two linear layers with a GELU activation function in between, and optionally a dropout layer for regularization. In practice, this MLP is implemented using 1×1 convolution layers to transform feature dimensions efficiently while maintaining the spatial structure of the input feature maps.

Despite its effectiveness, the MLP in Swin Transformer has certain limitations. The primary constraint is its inability to sufficiently capture spatial relationships and contextual information beyond the local scope of the feature map. While 1×1 convolutions are effective at channel-wise transformations, they do not inherently increase the receptive field, limiting the model's ability to integrate spatial context from neighboring regions. This limitation can be critical in tasks such as defect detection in high-resolution images, where capturing fine-grained spatial details is essential.

To address these limitations, we propose the spatial displacement with shift convolution (SDSC) module to replace the MLP in the Swin Transformer block. The SDSC module enhances the receptive field by introducing spatial displacement before applying convolution operations, thereby capturing broader contextual information. The SDSC module involves shifting the input feature maps in various directions

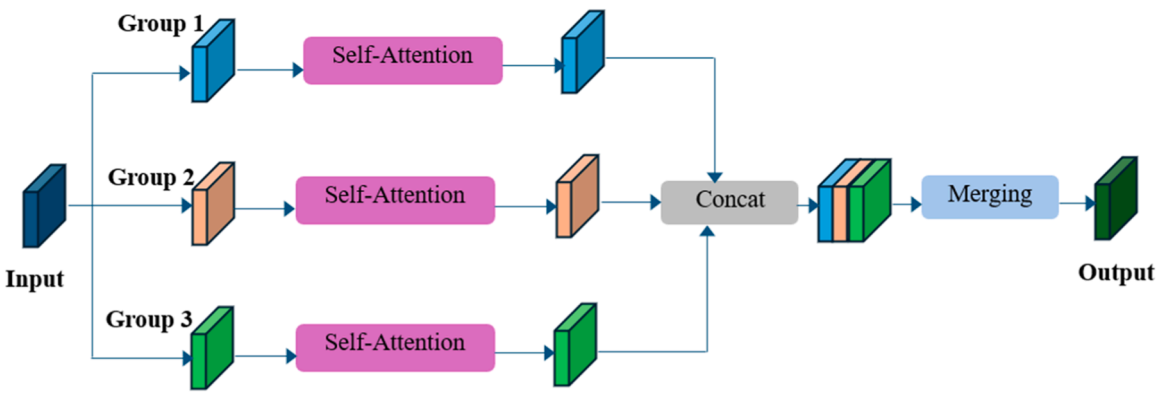


Fig. 2. Group self-attention mechanism.

and concatenating these shifted maps along the channel dimension. This approach allows the network to gather information from adjacent pixels, effectively increasing its ability to detect subtle and dispersed defects.

The detailed structure of the SDSC is shown in Fig. 3. The SDSC incorporates two shift convolutions (Wu et al., 2018), each followed by a Leaky ReLU activation. Each shift convolution operates by creating multiple shifted versions of the input feature map X of size $C \times H \times W$. These shifts are performed in four primary directions: left, right, up, and down. For instance, shifting X left by one pixel, denoted as X_l, involves moving each element of the feature map one position to the left. This process is mathematically represented as:

$$X_l(i,j) = X(i,j+1) \tag{1}$$

Similarly, we obtain X_r, X_u, and X_d by shifting X to the right, up, and down, respectively. These shifted maps are then concatenated along the channel dimension to form an augmented feature map X_{aug}:

$$X_{aug} = concat(X, X_l, X_r, X_u, X_d) \tag{2}$$

This augmented feature map, now of size $5C \times H \times W$, is then passed through a standard convolution layer. This convolution aggregates information from the spatially displaced feature maps, effectively increasing the receptive field and enabling the network to capture broader contextual information.

The SDSC module plays a pivotal role in feature extraction for defect detection in electroluminescent images of polycrystalline silicon solar cells. By leveraging spatial displacement, the SDSC module enhances the network's ability to aggregate contextual information from neighboring pixels. This improvement is crucial for detecting subtle defects that might be missed by conventional methods. The enhanced receptive field allows the model to capture fine-grained spatial relationships, leading to more accurate and reliable defect detection. Consequently, the SDSC module significantly improves the performance of the Swin Transformer in high-resolution imaging tasks, making it more effective for defect detection in polycrystalline silicon solar cells.

3.4. Fast multi-scale feature fusion

Electroluminescent images often contain defects of various scales and types, necessitating a detection approach that can accurately identify both small, subtle defects and larger, more apparent ones. Traditional neural networks struggle with this due to the trade-off between high-resolution features, which are good for small object detection, and high-level semantic features, which are better suited for larger object classification. In this paper, a fast multi-scale feature fusion (FMFF) is developed to overcome this trade-off by effectively combining features from different layers of the network, thereby enhancing detection precision across multiple scales. The structure of the FMFF integrates two key components: a weighted fusion of features mechanism and an enhanced feature extraction layer (EFEL). These components are meticulously designed to leverage features from different layers of the neural network, addressing the trade-off between high-resolution and high-level semantic information.

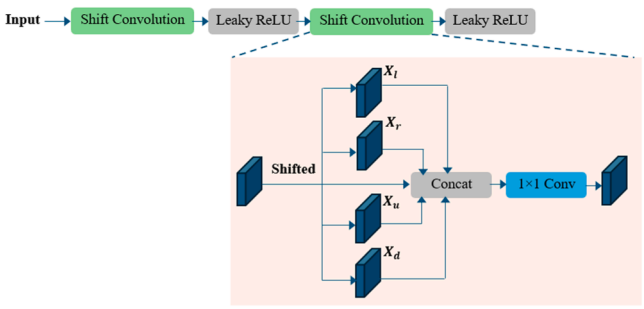


Fig. 3. The structure of the SDSC.

Fig. 4 provides the detailed structure of the FMFF. It captures and combines features from various hierarchical levels. Shallow feature maps (P_{in1} *and* P_{in2}) are high-resolution and contain detailed textural and positional information, making them essential for detecting small, intricate defects. Conversely, deeper feature maps (P_{in3} *and* P_{in4}) have a reduced resolution but a larger receptive field, carrying richer high-level semantic information that aids in identifying larger defects. The FMFF employs a dual feature fusion approach. Initially, it uses a direct connection between input and output layers, enriching the feature fusion without incurring additional computational costs. The features are processed through a series of EFEL layers and resizing operations:

$$P^{m}_{out_i} = \text{EFEL}\left(ReLU\left(\frac{w_a.P_{in_i} + w_b.Resize\left(P_{in_{i+1}}\right)}{w_c + w_d + \delta}\right)\right) \tag{3}$$

$$P^{end}_{out_i} = \text{EFEL}\left(ReLU\left(\frac{w_e.P_{in_i} + w_f.P^{m}_{out_i} + w_g.Resize\left(P^{end}_{out_{i-1}}\right)}{w_h + w_i + w_j + \delta}\right)\right) \tag{4}$$

where P_{in_i} is the input features of layer i within FMFF, $P^{m}_{out_i}$ and $P^{end}_{out_i}$ represent intermediate and output features respectively, EFEL represents the enhanced feature extraction layer, and *Resize* denotes up-sampling or down-sampling operations. The weighted fusion, inspired by the Bidirectional Feature Pyramid Network (BiFPN) (Tan, Pang & Le, 2020), ensures optimal blending of features from different levels. The fusion weights (w_i) are learnable, balancing the contributions of various features based on their significance, formulated as:

$$X_{output} = \sum_j \frac{w_j.X^{j}_{input}}{\sum_k w_k + \delta} \tag{5}$$

The enhanced feature extraction layer (EFEL), as illustrated in Fig. 5, optimizes the extraction of high-dimensional features while maintaining computational efficiency. EFEL addresses the need for rapid detection and accurate feature extraction in the dynamic environment of high-resolution images. The process begins with a 1×1 convolution to adjust the channel number of input features. The input features are split into two low-dimensional sets along the channel dimension, each further divided into blocks. These blocks are processed in parallel using 3×3 convolutions. The process can be represented by the following equations:

$$z_1, z_2 = split(conv_A(u)) \tag{6}$$

$$z'_i = split(conv_A(z_1)) \tag{7}$$

$$z = BN\left(conv_B\left(concat\left(z_1, \sum_{i=1}^{D} \varphi_m\left(z'_i\right), z_2\right)\right)\right) \tag{8}$$

where u represents the input feature, z signifies the output feature, $conv_A$ and $conv_B$ denote the 1×1 convolution, and z_1, z_2, and z'_i are intermediate features obtained from the split operation. The function φ_m represents the segmentation and parallel computation of lower-dimensional features. This architecture efficiently reduces complexity and boosts computational efficiency by processing features in parallel and using multi-branch convolutions.

By integrating weighted fusion of features mechanism and EFEL, FMFF effectively combines high-resolution details with high-level semantics, enhancing the network's ability to detect defects of various scales and types in electroluminescent images of polycrystalline silicon solar cells. The weighted fusion ensures optimal blending, while the parallel processing in EFEL maintains computational efficiency, making FMFF a robust and efficient solution for defect detection in this context.

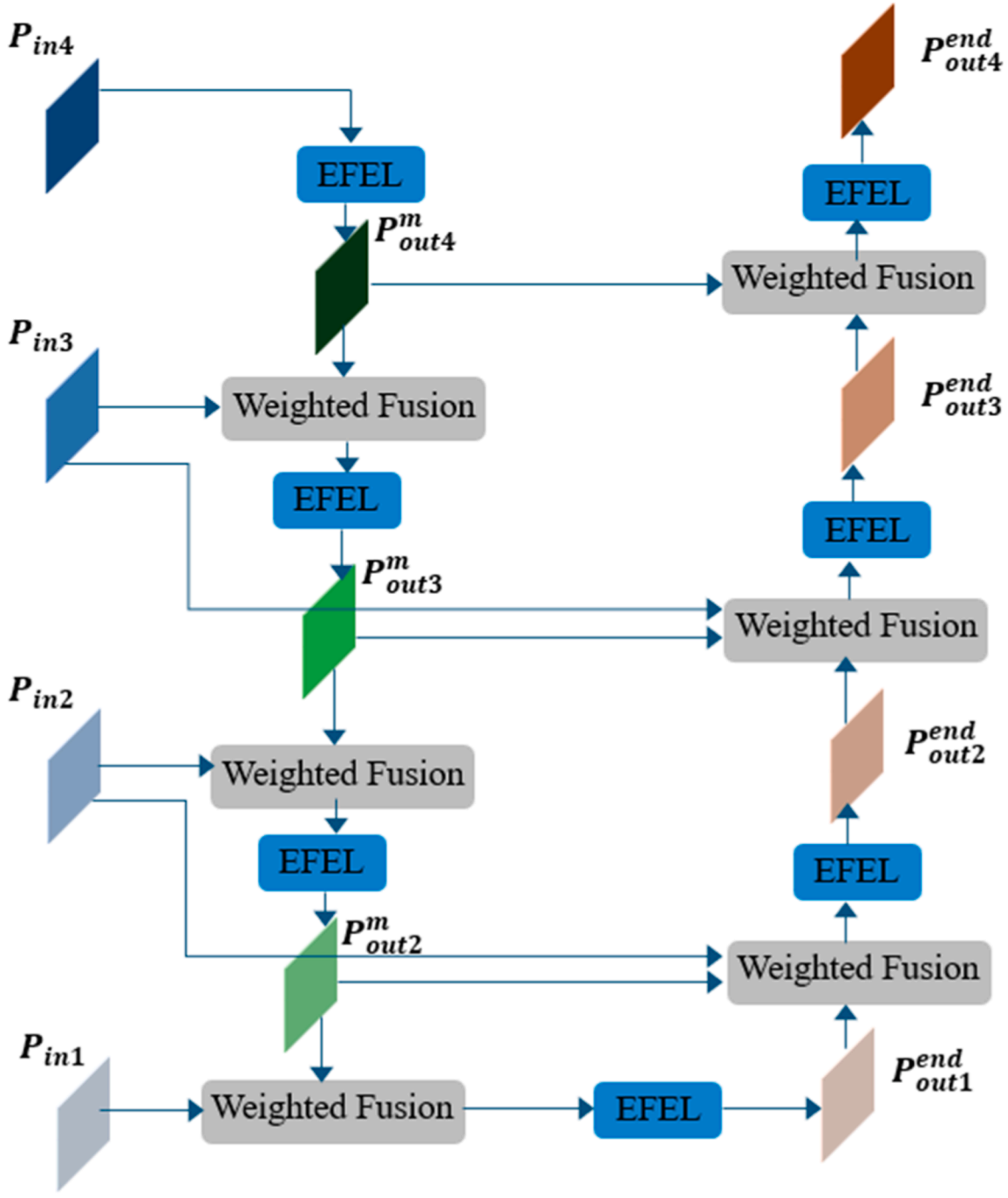


Fig. 4. The structure of the fast multi-scale feature fusion.

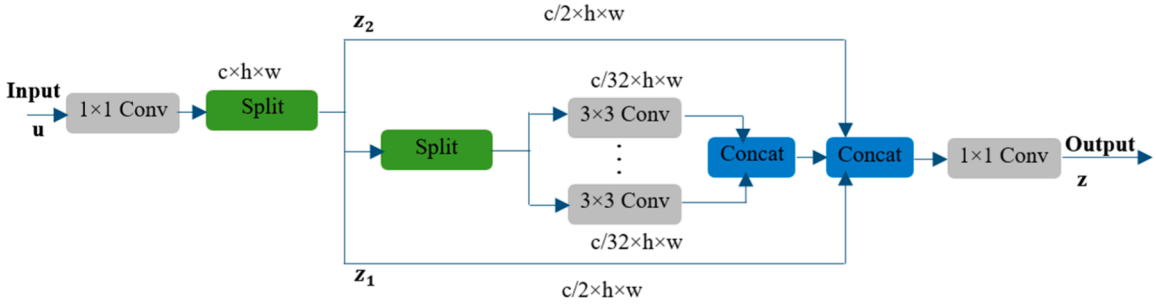


Fig. 5. Details architecture of the enhanced feature extraction layer.

4. Experimental results

4.1. Dataset and evaluation metrics

We evaluate the proposed model on the PVEL-AD (Photovoltaic EL Anomaly Detection) dataset (Su, Zhou & Chen, 2022). The PVEL-AD dataset is a significant contribution to the field of photovoltaic cell anomaly detection. This dataset comprises 36,543 near-infrared electroluminescence images of polycrystalline silicon solar cells, capturing a variety of internal defects and heterogeneous backgrounds, ensuring high-resolution imaging (1024×1024 pixels) suitable for detecting fine details of defects. The dataset includes both anomaly-free images and anomalous images classified into ten distinct categories. It provides detailed annotations with 37,380 ground truth bounding boxes covering eight specific types of defects: linear crack, finger interruption, black core, thick line, star crack, corner anomalies, fragment, and scratch. Fig. 6 provides sample defect images from this dataset.

PVEL-AD is the first publicly available dataset providing box-wise ground truth annotations for PV cell defect detection, enabling the training and evaluation of deep learning models. The comprehensive nature of the dataset, combined with its practical relevance to real-world manufacturing processes, makes it a valuable resource for researchers and developers aiming to improve the reliability and efficiency of solar cell quality control. In our study, we adopt the same data augmentation techniques as outlined in the PVEL-AD dataset paper to enhance the training of our defect detection model. Specifically, the data augmentation process includes horizontal flipping of images before the training phase, effectively doubling the size of the training dataset by creating mirror images of the original data. Additionally, random transformations are applied during the training process, such as resizing, cropping, and distortion, to simulate various conditions and further diversify the training data. These augmentations ensure that the model is exposed to a wide range of possible scenarios, enhancing its ability to generalize and perform accurately on real-world data, as well as preventing overfitting and improving the robustness of our deep learning model.

For evaluation, we use several key metrics: mAP50, mAP75, mAP50:5:95, precision, recall, F1-score, the number of parameters, and FPS. These metrics provide a comprehensive assessment of a model's accuracy, efficiency, and computational complexity.

Mean Average Precision (mAP) is a crucial metric in object detection tasks, representing the precision-recall tradeoff. Precision is the ratio of correctly predicted positive observations to the total predicted positives, while recall is the ratio of correctly predicted positives to all actual positives. The F1-score is the harmonic mean of precision and recall, providing a single measure that balances the two. Average Precision (AP) is the area under the Precision-Recall (P-R) curve, offering a single measure of a model's performance across different confidence thresholds. mAP50 calculates the AP with an Intersection over Union (IoU) threshold of 50 %, where IoU measures the overlap between the predicted bounding box and the ground truth box. Similar to mAP50, mAP75 uses a stricter IoU threshold of 75 %, emphasizing more precise bounding box predictions by requiring a higher degree of overlap between the predicted and ground truth boxes for a detection to be considered correct. mAP50:5:95 provides a more comprehensive evaluation by averaging AP scores across multiple IoU thresholds, ranging from 50 % to 95 % in increments of 5 %. This metric offers a better overall picture of the model's performance across a range of precision levels, from coarse to very fine localization of defects.

The number of parameters in a model refers to the total count of trainable variables, such as weights and biases, within the neural network. This metric directly indicates the model's complexity and memory requirements. A model with a high number of parameters can capture more intricate patterns in data but may require more computational resources and be prone to overfitting. Frames per second (FPS) measures the speed of the model, indicating how many images or frames it can process per second. This is a crucial metric for real-time applications, where the detection system needs to operate quickly to meet industrial throughput requirements. FPS is inversely related to the computational cost of the model; higher FPS indicates a faster, more efficient model.

4.2. Experimental settings

The experiments were conducted using the PyTorch framework, leveraging its flexibility and efficiency for implementing the proposed model. The training and testing were performed on a system equipped with an NVIDIA RTX 4080 GPU. The model was trained over 300 epochs with a batch size of 16 and an initial learning rate of 0.001, using the Adam optimizer for optimization, which provides adaptive learning rate adjustments for faster convergence. The values of key hyperparameters, including the batch size, learning rate, and number of epochs, were determined through a combination of grid search and cross-validation techniques. Specifically, a grid search was performed over a predefined range of values: the learning rate was tested between 0.0001 and 0.01, and the batch size was varied between 8 and 32. Cross-validation was used on a subset of the PVEL-AD dataset to evaluate the performance of different hyperparameter settings. The final values were chosen based on the best performance observed in terms of mAP50, F1-score, and training stability. The number of epochs was set to 300 after monitoring the learning curves and applying early stopping criteria to prevent overfitting.

For the model's architecture, the window size in the group self-attention mechanism was set to 7, and the shift size in the spatial displacement with shift convolution (SDSC) module was set to 1. These values were chosen based on a series of preliminary experiments that evaluated the trade-off between detection accuracy and computational efficiency. Smaller window allowed the model to focus on finer details, which is crucial for detecting small and dispersed defects in high-resolution electroluminescent images, but also increased the computational load. The selected values provided an optimal balance, ensuring high accuracy without significantly compromising processing speed.

We adopted the same data augmentation techniques as outlined in the PVEL-AD dataset paper to enhance the training of our defect

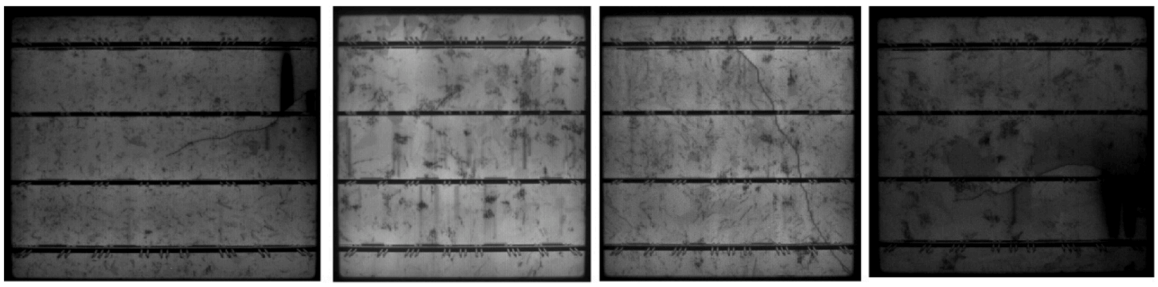


Fig. 6. Sample defect images from the PVEL-AD dataset.

detection model. Specifically, the data augmentation process includes horizontal flipping, random resizing, cropping, and distortion to simulate various conditions and prevent overfitting, thereby improving the robustness and generalization ability of the model.

4.3. Performance comparison

Table 1 provides a comprehensive comparison of various state-of-the-art models on the PVEL-AD dataset. The results highlight our model as achieving the highest scores across all key metrics. Notably, the mAP50 score of the proposed model stands at 83.11 %, surpassing all other models, including the previously high-performing YOLOv8 (82.52 %) and YOLOv5-l (81.93 %). This suggests that the proposed model's architecture, which integrates a group self-attention mechanism, spatial displacement with shift convolution, and fast multi-scale feature fusion, is highly effective at capturing and processing the intricate details necessary for defect detection in electroluminescent images of polycrystalline silicon solar cells.

In addition to mAP50, the proposed model excels in other performance metrics, achieving a precision of 85.57 %, recall of 83.12 %, and an F1-score of 84.33 %. These metrics indicate that the model is not only precise in identifying defects but also robust in recall, ensuring that most defects are detected. The F1-score further confirms the model's balanced performance, effectively managing the trade-off between precision and recall. The improvements in mAP75 and mAP50:5:95, where the proposed model scores 47.62 % and 52.98 % respectively, also demonstrate that it performs exceptionally well at detecting objects at higher IoU thresholds. This ability to maintain effectiveness as the IoU threshold increases is a direct result of the model's advanced feature extraction and fusion techniques, which traditional models struggle to match, particularly when dealing with subtle and dispersed defects.

Efficiency metrics also highlight the balanced performance of the proposed model. With 50.6 million parameters, it is more complex than YOLOv5 (22.0 M) but remains less resource-intensive than models like Faster RPAN—CNN (260.5 M). The model achieves an FPS of 66.3, indicating a strong balance between speed and accuracy, which is crucial for real-time defect detection scenarios. Compared to YOLOv8, which achieves 55 FPS, the proposed model is significantly faster, owing to the computational efficiency introduced by the group self-attention mechanism and the FMFF. These components reduce computational overhead while effectively combining multi-scale features, allowing the model to maintain high speed without sacrificing accuracy.

Table 2 presents a detailed breakdown of the detection performance for eight defect categories on the PVEL-AD dataset across several models, highlighting the strengths and weaknesses of each in detecting specific types of defects. The proposed model stands out with the highest overall performance, particularly in challenging categories such as line_crack and finger, where it achieves AP50 scores of 64.1 % and 93.8 % respectively. These results suggest that the proposed model's architecture is highly effective at detecting both small, subtle defects like line_crack and more apparent ones like finger. In comparison, YOLOv8 shows strong results, particularly in detecting black_core, where it scores an impressive 98.61 %. However, it falls slightly behind the proposed model in detecting other defect types such as line_crack and finger, indicating that while YOLOv8 is highly effective at certain defect types, it may not be as balanced across all categories. Similarly, the BAF-Detector, which achieves the highest score for line_crack detection at 62.05 %, struggles with detecting corner and fragment defects, where its performance is notably lower. This variability highlights the challenge of designing a model that performs consistently across all defect categories. DETR, known for its robust transformer-based architecture, demonstrates decent performance across most categories but generally lags behind the proposed model and YOLOv8, especially in categories requiring high precision like scratch and corner. Its AP50 scores in these categories suggest that while DETR is capable of handling more complex image features, it might be less efficient in capturing the finer details needed for detecting subtle defects.

4.4. Ablation study

We conduct several ablation studies to evaluate the effect of each proposed model. The first ablation study evaluates the role of the group self-attention mechanism in the proposed model. This mechanism is critical for balancing computational efficiency with the model's ability to capture intricate details necessary for high-resolution defect detection. In this study, the group self-attention mechanism is replaced with the standard self-attention mechanism used in the Swin Transformer. The results, as visualized in Fig. 7, show a decrease in the mAP50:5:95 score from 52.98 % to 50.12 % and an increase in inference time to 74 ms. This drop in accuracy and the increase in computational time underscore the effectiveness of the group self-attention mechanism in enhancing both the speed and precision of the model. By efficiently managing the attention mechanism, the group self-attention mechanism ensures that the model can focus on subtle defects without incurring a significant computational cost, making it a vital component of the architecture.

The second ablation study investigates the contribution of the SDSC module. The SDSC module is designed to expand the receptive field, enabling the model to capture broader contextual information, which is particularly important for identifying dispersed and subtle defects in electroluminescent images. When the SDSC is replaced with the traditional MLP used in the Swin Transformer, the mAP50:5:95 score decreases to 50.85 %, and the speed slows down to 69 ms. This modest decline in performance highlights the SDSC module's critical role in enhancing the model's ability to accurately detect defects while maintaining processing efficiency. The SDSC module's ability to expand the receptive field ensures that the model can integrate more contextual information, leading to more precise defect detection.

The third ablation study focuses on the FMFF mechanism, which is integral for effectively combining high-resolution details with high-level semantic features. This mechanism ensures the model can accurately detect defects across various scales, from small, subtle imperfections to larger, more apparent defects. In this study, the FMFF is replaced with a standard feature pyramid network (Lin et al., 2017), leading to a decrease in the mAP50:5:95 score to 49.73 % and a slower inference time of 80 ms. The reduction in both accuracy and speed emphasizes the

Table 1
Performance comparison on the PVEL-AD dataset.

Model	mAP50 (%)	mAP75 (%)	mAP50:5:95 (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	FPS
EfficientDet-D2 (Tan, Pang & Le, 2020)	80.17	44.47	48.99	81.21	78.58	79.87	8.1	45.1
YOLOv5 (Su, Zhou & Chen, 2022)	81.57	46.45	50.17	83.53	80.94	82.22	22.0	91.5
YOLOv5-l (Su, Zhou & Chen, 2022)	81.93	46.86	51.13	84.03	81.20	82.59	50.3	65.0
YOLOv8 (Reis, Kupec, Hong & Daoudi, 2023)	82.52	47.57	52.10	85.16	82.33	83.72	68.0	55.0
BAF-Detector (Su, Chen & Zhou, 2021)	80.77	41.80	47.69	79.15	76.59	77.85	120.87	8.4
Faster RPAN—CNN (Su et al., 2020)	73.24	32.30	40.66	75.41	70.80	73.04	260.50	7.9
DETR (Carion et al., 2020)	78.60	44.22	49.84	80.54	77.26	78.87	41.0	30.4
Our model	83.11	47.62	52.98	85.57	83.12	84.33	50.6	66.3

Table 2
Detection results of 8 categories in the PVEL-AD dataset.

Model	AP50(%)							
	line_crack	finger	black_core	thick_line	star_crack	corner	fragment	scratch
EfficientDet-D2 (Tan, Pang & Le, 2020)	54.46	53.37	76.26	76.06	73.14	0.06	0.02	8.33
YOLOv5 (Su, Zhou & Chen, 2022)	57.95	90.51	97.52	64.72	63.18	17.75	68.01	33.33
YOLOv5-l (Su, Zhou & Chen, 2022)	59.27	91.31	97.76	66.04	62.84	16.81	54.71	33.33
YOLOv8 (Reis, Kupec, Hong & Daoudi, 2023)	60.13	91.55	98.61	67.24	64.51	19.56	68.61	33.33
BAF-Detector (Su, Chen & Zhou, 2021)	62.05	88.47	90.85	68.71	73.11	27.69	52.53	0
Faster RPAN—CNN (Su et al., 2020)	59.69	88.37	90.77	68.67	68.04	28.49	52.07	0
DETR (Carion et al., 2020)	55.53	86.44	95.11	63.41	60.15	15.42	65.54	8.33
Our model	64.1	93.8	97.90	68.91	75.67	28.18	69.02	33.33

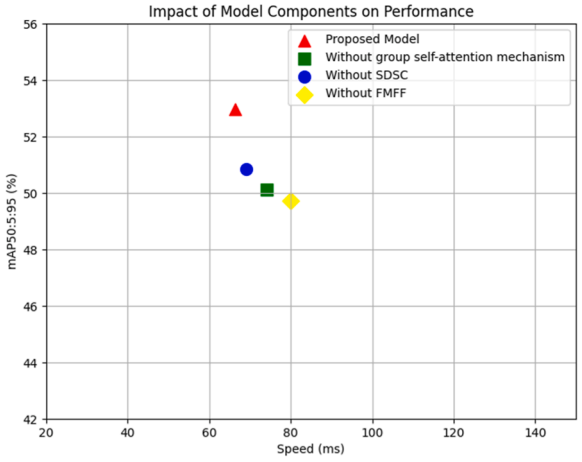


Fig. 7. The effect of each proposed model.

importance of FMFF in the proposed model. By balancing the detailed information from shallow layers with the broader contextual information from deeper layers, FMFF enables the model to maintain high detection accuracy while processing defects at multiple scales efficiently.

To further explore the impact of image resolution on model performance, we conducted experiments using input images at resolutions of 512×512, 1024×1024, and 2048×2048 pixels to evaluate the trade-offs between detection accuracy and computational efficiency. The results in Table 3 demonstrate how varying the resolution of input images impacts the performance of the proposed model. At the lower resolution of 512×512, the model experiences a slight decrease in mAP, precision, recall, and F1-score compared to the original 1024×1024 resolution. This reduction is due to the loss of fine details, which are critical for accurately detecting subtle defects in high-resolution electroluminescent images. However, the FPS increases significantly, indicating that the model processes lower-resolution images more quickly, making it more suitable for scenarios where speed is prioritized over absolute accuracy. Conversely, when the resolution is increased to 2048×2048, the model shows a modest improvement in precision and recall, with slight increases in mAP and F1-score compared to the 1024×1024 resolution. This suggests that the higher resolution allows the model to capture finer details more effectively, enhancing its ability to detect small and dispersed defects. However, this improvement comes at the cost of a significantly lower FPS, as the increased computational load slows down the processing speed. These results highlight the trade-off between accuracy and efficiency, suggesting that the choice of resolution should be based on the specific requirements of the application, whether it be real-time processing or achieving the highest possible detection accuracy.

4.5. Visualization results

Fig. 8 provides several detection results of the proposed model on the PVEL-AD dataset, demonstrating its capability to identify various types of defects within polycrystalline silicon solar cells. Each image contains bounding boxes around detected defects, with confidence scores displayed above the boxes. The model successfully detects a wide range of defects, such as line_crack, finger, star_crack, and black_core, among others, with varying confidence levels. In several instances, the model identifies multiple defects within a single image, such as the detection of both star_crack and line_crack in the top-left image, which indicates the model's ability to handle complex scenes with overlapping or multiple defect types. Additionally, the detection of subtle defects like fragment in the lower images highlights the model's sensitivity and precision in capturing fine-grained details, essential for thorough quality control in solar cell manufacturing. The varying sizes of the bounding boxes, particularly in detecting larger defects like black_core and smaller ones like finger, illustrate the model's proficiency in multi-scale detection. The model's ability to accurately pinpoint defects, regardless of their size or position, demonstrates the effectiveness of the FMFF mechanism, which balances high-resolution detail with semantic context from different layers of the network.

5. Conclusion

In this paper, we introduced a novel architecture for defect detection in electroluminescent images of polycrystalline silicon solar cells, addressing the challenges of detecting subtle and dispersed defects in high-resolution images. By modifying the Swin Transformer with a group self-attention mechanism and replacing the traditional MLP with the spatial displacement with shift convolution module, we significantly enhanced the model's feature extraction capabilities. Additionally, the fast multi-scale feature fusion mechanism effectively combines high-resolution details with high-level semantic information, enabling accurate detection of defects across various scales. Experimental results demonstrate that our approach outperforms traditional methods, providing improved detection accuracy and robustness. The model's

Table 3
Performance of the proposed model at different input image resolutions.

Resolution	mAP50 (%)	mAP75 (%)	mAP50:5:95 (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	FPS
512×512	81.45	45.20	50.76	83.52	81.11	82.23	50.6	88.2
1024×1024	83.11	47.62	52.98	85.57	83.12	84.33	50.6	66.3
2048×2048	83.78	47.95	53.60	85.85	83.70	84.76	50.6	38.7

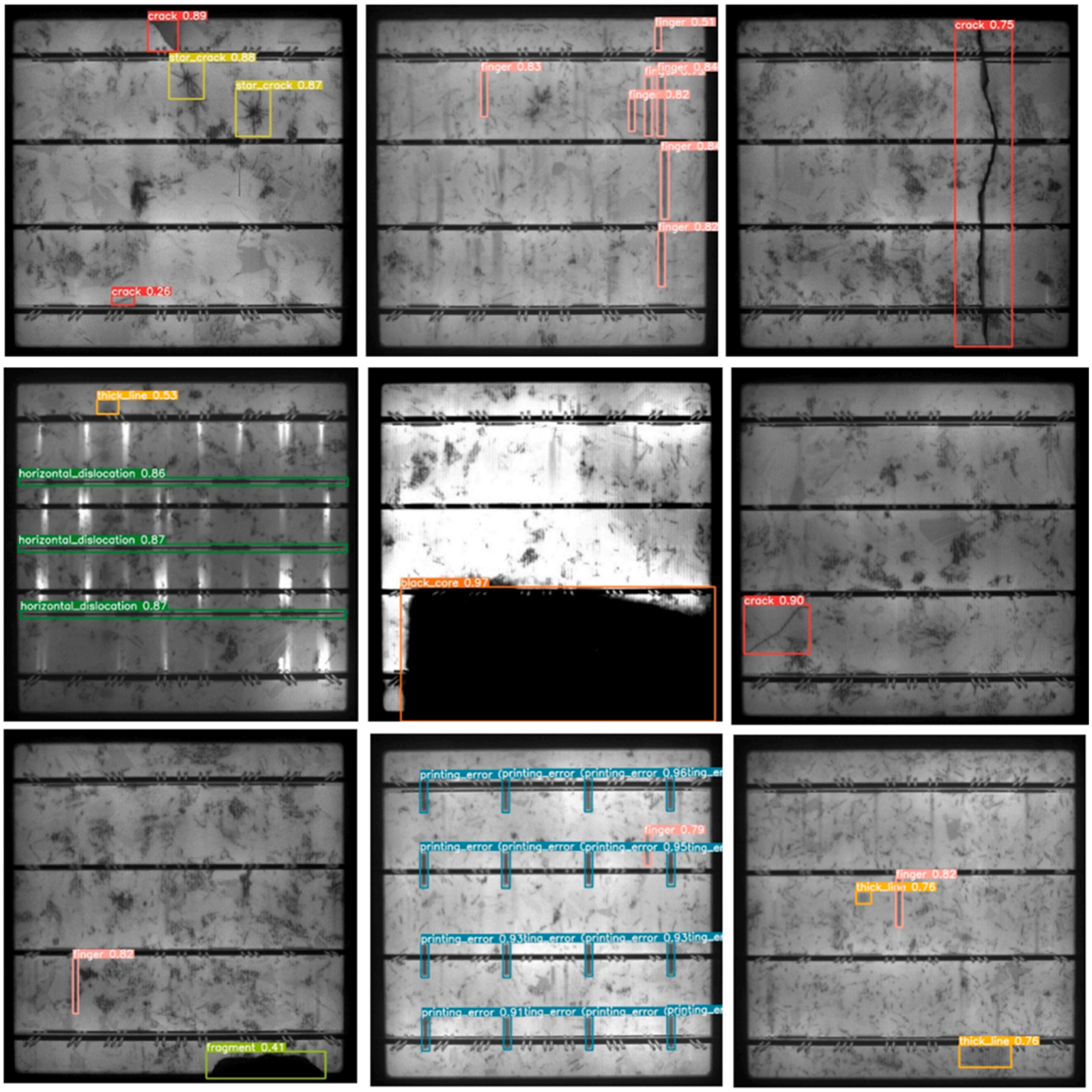


Fig. 8. Detection results of the proposed model.

ability to generalize well across different defect types and scales makes it a highly effective tool for quality assurance in solar cell manufacturing. However, the proposed method does have some limitations. One key limitation is the model's computational complexity, particularly during training, due to the use of advanced mechanisms such as group self-attention and multi-scale feature fusion. This could limit the scalability of the method, especially in environments with limited computational resources. To address this, future work could focus on optimizing the model's architecture to reduce computational overhead, possibly through model pruning, quantization, or other techniques that maintain accuracy while improving efficiency. Another limitation is the potential for the model to be sensitive to variations in the training data, particularly in terms of defect types that are underrepresented. While the model demonstrates good generalization across various defect types, further work could explore techniques such as data augmentation or the use of generative models to synthesize more diverse training samples, ensuring the model is robust even in cases of rare or novel defects. Finally, while the proposed method shows strong performance in detecting defects in polycrystalline silicon solar cells, its applicability to other types of solar cells or high-resolution imaging tasks may require additional adaptation. Future research could extend this architecture to other materials and imaging domains, exploring how the model can be adapted or fine-tuned to maintain its effectiveness across different contexts.

CRediT authorship contribution statement

Hoanh Nguyen: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Tuan Anh Nguyen**: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Writing – review & editing. **Nguyen Duc Toan**: Data curation, Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Akande, Timileyin Opeyemi, Alabi, Oluwaseyi O., & Ajagbe, Sunday A. (2022). A deep learning-based CAE approach for simulating 3D vehicle wheels under real-world conditions. *Artificial Intelligence and Applications*.

Akram, M. Waqar, Li, Guiqiang, Jin, Yi, Chen, Xiao, Zhu, Changan, Zhao, Xudong, et al. (2019). CNN based automatic detection of photovoltaic cell defects in electroluminescence images. *Energy, 189*, Article 116319.

Bhosle, Kavita, & Musande, Vijaya (2023). Evaluation of deep learning CNN model for recognition of devanagari digit. *Artificial intelligence and applications, 1*(2), 114–118.

Bochkovskiy, A. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *arXiv preprint arXiv:2004.10934* (2020).

Carion, Nicolas, Massa, Francisco, Synnaeve, Gabriel, Usunier, Nicolas, Kirillov, Alexander, & Zagoruyko, Sergey (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Cham: Springer International Publishing.

Chen, Huayue, Ru, Jie, Long, Haoyu, He, Jialin, Chen, Tao, & Deng, Wu (2024). Semi-Supervised Adaptive Pseudo-Label Feature Learning for Hyperspectral Image Classification in Internet of Things. *IEEE Internet of Things Journal*.

Chen, Haiyong, Pang, Yue, Hu, Qidi, & Liu, Kun (2020). Solar cell surface defect inspection based on multispectral convolutional neural network. *Journal of Intelligent Manufacturing, 31*(2), 453–468.

Chen, Haiyong, Zhao, Huifang, Han, Da, & Liu, Kun (2019). Accurate and robust crack detection using steerable evidence filtering in electroluminescence images of solar cells. *Optics and Lasers in Engineering, 118*, 22–33.

Chirgaiya, Sachin, & Rajavat, Anand (2023). Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). *Intelligent Systems with Applications, 18*, Article 200217.

Deitsch, Sergiu, Christlein, Vincent, Berger, Stephan, Buerhop-Lutz, Claudia, Maier, Andreas, Gallwitz, Florian, et al. (2019). Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy, 185*, 455–468.

Dong, Li, Zhang, Haijun, Ji, Yuzhu, & Ding, Yuxin (2020). Crowd counting by using multi-level density-based spatial information: A Multi-scale CNN framework. *Information Sciences, 528*, 79–91.

Elhaija, Wejdan Abu, & Al-Haija, Qasem Abu (2023). A novel dataset and lightweight detection system for broken bars induction motors using optimizable neural networks. *Intelligent Systems with Applications, 17*, Article 200167.

Ge, Chunpeng, Liu, Zhe, Fang, Liming, Ling, Huading, Zhang, Aiping, & Yin, Changchun (2020). A hybrid fuzzy convolutional neural network based mechanism for photovoltaic cell defect detection with electroluminescence images. *IEEE Transactions on Parallel and Distributed Systems, 32*(7), 1653–1664.

Girshick, Ross. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

Jiao, Licheng, Wang, Mengjiao, Liu, Xu, Li, Lingling, Liu, Fang, Feng, Zhixi, et al. (2024). Multiscale Deep Learning for Detection and Recognition: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Lin, Tsung-Yi, Dollár, Piotr, Girshick, Ross, He, Kaiming, Hariharan, Bharath, & Belongie, Serge (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Li, Chuyi, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li et al. et al. "YOLOv6: A single-stage object detection framework for industrial applications." *arXiv preprint arXiv:2209.02976* (2022).

Li, Weihan, Liu, Dunke, Li, Yang, Hou, Ming, Liu, Jie, Zhao, Zhen, et al. (2024). Fault diagnosis using variational autoencoder GAN and focal loss CNN under unbalanced data. *Structural Health Monitoring*, Article 14759217241254121.

Liu, Wei, Anguelov, Dragomir, Erhan, Dumitru, Szegedy, Christian, Reed, Scott, Fu, Cheng-Yang, et al. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21–37). Springer International Publishing.

Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Preethi, Padmaprabha, & Mamatha, Hosahalli Ramappa (2023). Region-based convolutional neural network for segmenting text in epigraphical images. *Artificial Intelligence and Applications, 1*(2), 119–127.

Qaddour, Jihad, & Siddiqa, Syeda Ayesha (2023). Automatic damaged vehicle estimator using enhanced deep learning algorithm. *Intelligent Systems with Applications, 18*, Article 200192.

Qian, Xiaoliang, Li, Jing, Cao, Jinde, Wu, Yuanyuan, & Wang, Wei (2020). Micro-cracks detection of solar cells surface via combining short-term and long-term deep features. *Neural Networks, 127*, 132–140.

Rahman, Muhammad Rameez Ur, & Chen, Haiyong (2020). Defects inspection in polycrystalline solar cells electroluminescence images using deep learning. *IEEE Access : Practical Innovations, Open Solutions, 8*, 40547–40558.

Redmon, Joseph, & Farhadi, Ali (2017). YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263–7271).

Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).

Redmon, Joseph, Divvala, Santosh, Girshick, Ross, & Farhadi, Ali (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Reis, Dillon, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. "Real-time flying object detection with YOLOv8." *arXiv preprint arXiv:2305.09972* (2023).

Su, Binyi, Chen, Haiyong, Zhu, Yifan, Liu, Weipeng, & Liu, Kun (2019). Classification of manufacturing defects in multicrystalline solar cells with novel feature descriptor. *IEEE Transactions on Instrumentation and Measurement, 68*(12), 4675–4688.

Su, Binyi, Chen, Haiyong, Chen, Peng, Bian, Guibin, Liu, Kun, & Liu, Weipeng (2020). Deep learning-based solar-cell manufacturing defect detection with complementary attention network. *IEEE Transactions on Industrial informatics, 17*(6), 4084–4095.

Su, Binyi, Chen, Haiyong, & Zhou, Zhong (2021). BAF-detector: An efficient CNN-based detector for photovoltaic cell defect detection. *IEEE Transactions on Industrial Electronics, 69*(3), 3161–3171.

Su, Binyi, Zhou, Zhong, & Chen, Haiyong (2022). PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection. *IEEE Transactions on Industrial Informatics, 19*(1), 404–413.

Su, Binyi, Zhong Zhou, Haiyong Chen, and Xiaochun Cao. "SIGAN: A novel image generation method for solar cell defect segmentation and augmentation." *arXiv preprint arXiv:2104.04953* (2021).

Tan, Mingxing, Pang, Ruoming, & Le, Quoc V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790).

Tsai, Du-Ming, Wu, Shih-Chieh, & Chiu, Wei-Yao (2012). Defect detection in solar modules using ICA basis images. *IEEE Transactions on Industrial Informatics, 9*(1), 122–131.

Wang, Chien-Yao, Bochkovskiy, Alexey, & Liao, Hong-Yuan Mark (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7464–7475).

Wei, Yi-Ming, Chen, Kaiyuan, Kang, Jia-Ning, Chen, Weiming, Wang, Xiang-Yu, & Zhang, Xiaoye (2022). Policy and management of carbon peaking and carbon neutrality: A literature review. *Engineering, 14*, 52–63.

Wu, Bichen, Wan, Alvin, Yue, Xiangyu, Jin, Peter, Zhao, Sicheng, Golmant, Noah, et al. (2018). Shift: A zero flop, zero parameter alternative to spatial convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9127–9135).

Xiao, Meng, Yang, Bo, Wang, Shilong, Mo, Fan, He, Yan, & Gao, Yifan (2023). GRA-Net: Global receptive attention network for surface defect detection. *Knowledge-Based Systems, 280*, Article 111066.

Yu, Xuyi, Lyu, Wentao, Wang, Chengqun, Guo, Qing, Zhou, Di, & Xu, Weiqiang (2023). Progressive refined redistribution pyramid network for defect detection in complex scenarios. *Knowledge-Based Systems, 260*, Article 110176.

Zhang, Zhixin, Chen, Min, Zhong, Teng, Zhu, Rui, Qian, Zhen, Zhang, Fan, et al. (2023). Carbon mitigation potential afforded by rooftop photovoltaic in China. *Nature Communications, 14*(1), 2347.

Zhang, Na, Yang, Gang, Wang, Dawei, Hu, Fan, Yu, Hua, & Fan, Jingjing (2024). A Defect Detection Method for Substation Equipment Based on Image Data Generation and Deep Learning. *IEEE Access : Practical Innovations, Open Solutions*.