AI Engineering
https://www.sciltp.com/journals/aieng

Article

Dynamic Attention and Context-Aware Feature Fusion for Multi-Scale Solar Panel Defect Detection

Ali Farajzadeh Bavil [1], Mahdi Khodayar [1], Jacob Regan [1] and Mohammad E. Khodayar [2,*]

[1] Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104, USA
[2] Department of Electrical and Computer Engineering, Southern Methodist University, Dallas, TX 75205, USA
* Correspondence: mkhodayar@smu.edu

How To Cite: Farajzadeh Bavil, A.; Khodayar, M.; Regan, J.; et al. Dynamic Attention and Context-Aware Feature Fusion for Multi-Scale Solar Panel Defect Detection. *AI Engineering* **2025**, *1*(1), 6. https://doi.org/10.53941/aieng.2025.100006

Received: 21 June 2025
Revised: 18 August 2025
Accepted: 16 October 2025
Published: 18 November 2025

Abstract: Ensuring the accurate detection of surface flaws in PV panels is vital for preserving energy efficiency and minimizing future repair expenses. Nevertheless, the diverse nature of these defects in terms of size, shape, and visibility presents significant localization challenges. Conventional models typically rely on fixed feature hierarchies, uniform spatial weighting, and static fusion strategies. These limitations restrict their ability to capture defects across scales, emphasize relevant regions, and integrate semantic information effectively under visually complex conditions. To overcome these challenges, we propose a Multi-Scale Attention-based Convolutional Neural Network (MSA-CNN), which is a compact detection framework composed of three specialized modules. The multi-scale feature extraction module first captures spatial patterns at varying resolutions through parallel convolutional branches, addressing scale-related limitations. These features are then refined by the dynamic attention module, which adaptively emphasizes defect-relevant regions based on spatial context. Finally, the context-aware fusion module integrates the attention-enhanced features by selectively combining multi-level information, producing semantically consistent representations for accurate detection. Experimental results on the PV Multi-Defect dataset show that MSA-CNN outperforms a range of state-of-the-art methods across all key detection metrics by achieving higher accuracy across all metrics and defect categories, with notable improvements in detecting small, low-contrast, and structurally irregular faults.

Keywords: solar panel defect detection; dynamic attention; context-aware feature fusion

1. Introduction

As photovoltaic (PV) systems continue to expand globally as a key pillar of sustainable energy, ensuring their reliability and efficiency has become increasingly critical. Broken cells, hot spots, inactive regions, and surface scratches can significantly degrade energy output, shorten module lifespan, and introduce safety risks [1]. Manual inspection methods, while once standard, are now insufficient due to their subjectivity, labor requirements, and lack of scalability, particularly in utility-scale solar farms [2]. In response, automated inspection systems based on solar panel imagery have become a central research focus. By extracting and interpreting visual cues from modalities such as infrared and electroluminescence imaging, these systems aim to detect defects accurately and at scale, addressing the limitations of conventional inspection methods [3].

Early PV defect detection techniques commonly followed a two-stage pipeline: extracting handcrafted visual features from electroluminescence images, followed by classification using traditional machine learning models. During the feature extraction stage, the spatial and structural properties of defects were quantified using descriptors such as the gray-level co-occurrence matrix, histogram of oriented gradients, and wavelet-based texture features. These

handcrafted descriptors were subsequently classified using traditional algorithms, including Kirsch operator, Support Vector Machines (SVMs), K-Nearest Neighbors, and decision trees [4–7]. The shift toward deep learning [8–10] introduced end-to-end models capable of extracting hierarchical visual features directly from the raw data, thereby eliminating the necessity of manual feature design [11]. Automatic feature learning by deep neural networks enables the model to discover complex, hierarchical patterns directly from raw image data. This ability allows them to outperform hand-crafted feature extraction methods by adapting to the specific details of each image processing task. Early work primarily relied on convolutional neural networks (CNNs) for image-level classification [12–15]. The study in [16] employed DenseNet169 as a feature extractor to classify defect types in solar panel images. To further enhance representation learning under challenging conditions, such as small sample sizes and diverse defect types, the work in [17] introduced a Siamese convolutional architecture augmented with depth-wise separable convolutions and information bottleneck regularization. While CNN-based methods were effective at capturing complex defect patterns, they could not localize them spatially [18,19]. To overcome this limitation, object detection frameworks such as YOLO models gained prominence for their ability to simultaneously localize and classify multiple defect instances in real time [20]. A notable early advancement was the multi-stage model built on YOLOv3, which first isolated PV modules before performing fine-grained defect detection within each panel. This hierarchical structure reduced background interference and improved defect focus, which set the foundation for task-specific refinements [21].

However, the fixed-scale architecture of YOLOv3 limited its sensitivity to small and irregular faults. To address this, YOLOv5-based models introduced higher architectural flexibility and scale-awareness. The study in [1] proposed GBH-YOLOv5, which replaced standard convolutions with ghost convolutions for efficiency and added cross-stage partial bottleneck modules to improve feature diversity. A dedicated small-object head was also introduced to better detect tiny defects such as cracks and hotspots. Building on this, the study in [22] further enhanced YOLOv5 by integrating deformable convolutions into the backbone to accommodate variations in defect shapes and by applying Efficient Channel Attention to suppress irrelevant background features. While YOLOv5 variants improved scale sensitivity, limitations remained in capturing dense and subtle defects. To address this, the work in [23] enhanced YOLOv7 by introducing three key components: Funnel Rectified Linear Unit (ReLU) for better spatial modeling, a lightweight pointwise convolution bottleneck to retain fine features, and a semantic enhancement attention module to suppress background noise and sharpen defect focus. Expanding on this, authors in [24] proposed YOLOv7-GX with a lightweight feature pyramid for efficient feature fusion and an embedded global attention mechanism to strengthen focus on defect-prone regions. In the study [25], YOLOv8 was enhanced using coordinate attention to better guide focus across channels. The model's neck was also replaced with a bidirectional feature pyramid network to improve resolution-aware feature propagation, particularly for small or low-contrast defects.

Despite advances in YOLO-based and segmentation-driven models, current approaches still exhibit key limitations in accurately detecting PV panel defects across varied scales and conditions: (1) Most existing detectors rely on fixed receptive fields and shallow multi-resolution hierarchies, limiting their ability to simultaneously model small-scale patterns such as hairline cracks and large-scale anomalies like discoloration. Although feature pyramids and multi-head architectures have been adopted, they often fuse features in a static or shallow manner, resulting in incomplete scale integration and reduced sensitivity to subtle defects; (2) Feature maps are typically processed with uniform spatial weighting, ignoring the saliency differences between defect regions and irrelevant textures like panel frames or surface glare. Even attention-based models tend to use fixed or non-adaptive weights, which restricts their flexibility in dynamic imaging conditions and leads to mislocalizations under occlusion or low contrast; (3) Fusion operations in state-of-the-art models, including summation, concatenation, or simple residual links, lack contextual adaptivity. These methods do not adjust their integration strategy based on the semantic importance of incoming features, which leads to feature imbalance and underutilization of high-level contexts that are critical for distinguishing visually similar defect types.

To address these limitations, we propose a novel Multi-Scale Attention-based Convolutional Neural Network (MSA-CNN) for accurate and efficient solar panel defect detection. The proposed work has the following contributions: (1) A multi-scale feature extraction module is devised by combining standard, dilated, and residual convolutions with diverse kernel sizes, enabling the model to represent defect patterns of varying spatial extents within a unified feature space; (2) A dynamic attention mechanism is designed to assign image-dependent weights to spatial regions, allowing the network to focus on defect-relevant areas while suppressing background clutter under diverse imaging conditions; (3) A context-aware feature fusion module is developed to integrate multi-scale and attention-enhanced features through a learnable gating mechanism, ensuring that semantic relevance guides the fusion process. This enhances both boundary localization and inter-class discrimination; (4) The overall architecture is constructed to be lightweight and computation-efficient by limiting network depth and incorporating

efficient convolutional designs, making it deployable on resource-constrained platforms embedded with solar monitoring systems.

2. Problem Formulation

Solar panel defect detection entails the localization and classification of faults within PV panel images using a structured image dataset. For a dataset D containing N images, each image I_i has a resolution of $H \times W \times C$, where H and W are the height and width, respectively, and C is typically 3 for RGB color channels. The objective is to predict a collection of bounding boxes and corresponding defect classes for each I_i. A bounding box b_{ij} is defined by its center coordinates x_{ij}, y_{ij}, along with width w_{ij} and height h_{ij}. The set of bounding boxes for image I_i is written as $B_i = \{b_{i1}, b_{i2}, \dots, b_{in_i}\}$, with n_i indicating the total number of defects identified.

Along with the bounding boxes, the model also predicts a set of class labels $L_i = \{l_{i1}, l_{i2}, \dots, l_{in_i}\}$, where each label l_{ij} corresponds to the class of the defect detected within the bounding box b_{ij}. The task is to accurately identify and localize defects in each solar panel image by developing a model that takes each image I_i from the dataset D and outputs the set of bounding boxes B_i and their corresponding class labels L_i. This formulation provides a structured approach to addressing the challenges of defect detection and localization across a dataset of solar panel images.

3. Proposed Method

This section introduces the MSA-CNN framework developed for defect detection in PV panels. The framework is composed of three core modules: Multi-Scale Feature Extraction, Dynamic Attention Mechanism, and Context-Aware Feature Fusion. Each component is designed to enhance the network's performance and is detailed in the following subsections.

Given the wide variability in defect shape, scale, and visual complexity, traditional single-scale detection models often fall short in solar panel applications. MSA-CNN addresses this by incorporating a Multi-Scale Feature Extraction module that processes image features through multiple convolutional layers with varying kernel sizes. This configuration enables the network to simultaneously capture both fine-grained visual textures and larger image structural patterns, making it well-suited for identifying defects of various types and dimensions.

To improve spatial selectivity and the ability to focus on specific image regions, the framework integrates a Dynamic Attention Mechanism that dynamically highlights regions of interest based on their defect relevance. Given the unpredictable distribution of defects across the panel surface, this mechanism adjusts attention weights at the image level. It emphasizes areas that are likely to contain defects and suppresses background noise. This adaptiveness improves the model's focus and contributes to higher detection accuracy.

Finally, the Context-Aware Feature Fusion module combines outputs from previous modules by assigning learnable weights to features based on their contextual importance. This allows the network to prioritize the most informative representations during the inference phase. By intelligently merging multi-scale and attention-refined features, the model strengthens its ability to detect and localize defects under challenging visual conditions.

3.1. Multi-Scale Feature Extraction

The Multi-Scale Feature Extraction module is designed to capture defects of varying sizes by extracting features at multiple scales. This is achieved by using M parallel convolutional branches with different kernel sizes. Let the set of selected kernel sizes be denoted by $K = \{k_1, k_2, \cdots, k_M\}$. The feature map extracted using a sequence of n_c convolutional layers with kernel size $k_i \in K$ is represented as $F_{k_i}(I)$, where $I \in \mathcal{R}^{H \times W \times C}$ is the input image. The outputs from these branches are then concatenated to form the multi-scale feature representation for an input image I:

$$F_{multi}(I) = Concat(F_{k_1}(I), F_{k_2}(I), \cdots, F_{k_M}(I)) \quad (1)$$

where $Concat$ denotes the concatenation operation.

To further enhance the feature extraction process, we introduce dilated convolutions. The feature map extracted by a dilated convolution with a kernel size $k_i \in K$ and a specific dilation rate $d_i \in \{d_1, d_2, \cdots, d_M\}$, is denoted as $D_{k_i,d_i}(I)$. The multi-scale feature extraction with dilation can be represented as:

$$F_{multi-dilated}(I) = Concat(D_{k_1,d_1}(I), D_{k_2,d_2}(I), \cdots, D_{k_M,d_M}(I)) \quad (2)$$

Additionally, we incorporate residual connections to preserve spatial information across different scales. Let $R_k(I)$ represent the residual feature map extracted using a residual convolutional layer with a kernel size $k \in K$. The multi-scale feature extraction with residual connections can be represented as:

$$F_{multi-residual}(I) = Concat(R_{k_1}(I), R_{k_2}(I), \cdots, R_{k_M}(I)) \quad (3)$$

By combining standard, dilated, and residual convolutions, the final multi-scale feature extraction module can be represented as:

$$F_{final}(I) = Concat(F_{multi}(I), F_{multi-dilated}(I), F_{multi-residual}(I)) \quad (4)$$

To formalize the operations within each branch, let X represent the input feature map to a given layer. A single standard convolutional operation, f_k, can be represented as:

$$f_k(X) = \sigma(W_k * X + b_k) \quad (5)$$

where W_k is the weight matrix, b_k is the bias term, $*$ denotes the convolution operation, and σ is the activation function. One can use the ReLU activation function for this formulation. By applying this convolutional operation, f_k, sequentially n_c times on the input Image I, the feature map $F_k(I)$ is formed. For dilated convolutions, the operation can be represented as:

$$\delta_{k,d}(X) = \sigma(W_{k,d} * X + b_{k,d}) \quad (6)$$

where $W_{k,d}$ is the weight matrix, $b_{k,d}$ is the bias term, and d is the dilation rate. By applying this dilated convolutional operation, $\delta_{k,d}$, sequentially n_c times on the input Image I, the dilated feature map $D_{k_i,d_i}(I)$ is formed. For residual connections, the operation can be represented as:

$$r_k(X) = \sigma(W_k * X + b_k) + X \quad (7)$$

where W_k is the weight matrix, b_k is the bias term, and X is the input feature. By applying this residual convolutional operation, r_k, sequentially for n_c times on the input Image I, the residual feature map $R_k(I)$ is formed. By leveraging multiple operations, the final multi-scale feature extraction module extracts fine visual characteristics and broader spatial image patterns to effectively identify defects of varying sizes. This is achieved by processing an image through parallel branches, each specialized for different scales. Small kernels capture fine details, while larger or dilated kernels identify broad patterns without the information loss associated with standard pooling. This parallel design provides a high tolerance to scale variance, which ensures reliable defect detection even when pixel sizes change due to varying physical panel dimensions. This is a key advantage for real-world applications. This design ensures reliable detection even when a defect's pixel dimensions change. This adaptability across different defect types improves the model's accuracy and reliability.

3.2. Dynamic Attention Mechanism

In solar panel imagery, defects can appear in dispersed locations, with some regions exhibiting more prominent faults than others. The Dynamic Attention Mechanism is designed to improve defect detection accuracy by directing the model's focus toward the most relevant areas of the input image. This mechanism refines the extracted multi-scale features by enhancing their spatial relevance to the defect detection task. To overcome the inefficiency of standard convolutions that apply uniform importance to all spatial regions, this module introduces spatial selectivity. Spatial selectivity enables the model to dynamically focus on the most informative locations within a feature map by computing a data-driven saliency map that assigns importance scores to each spatial region. These scores are then used to adjust the input features, boosting defect-related signals and reducing background noise. Let $F_{multi}(I)$ be the multi-scale feature map obtained from the Multi-Scale Feature Extraction module. The attention mechanism generates an attention map $A(I)$ from $F_{multi}(I)$. This can be represented as: $A(I) = \sigma(W_a * F_{multi}(I) + b_a)$ where W_a is the weight matrix, and b_a is the bias term. The refined feature map $F_{refined}(I)$ is obtained by element-wise multiplication of the attention map $A(I)$ with the multi-scale feature map $F_{multi}(I)$:

$$F_{refined}(I) = F_{multi}(I) \odot A(I) \quad (8)$$

where $\odot$ denotes element-wise multiplication. To further enhance the attention mechanism, we introduce a dynamic weighting scheme. The dynamically weighted attention map $A_{dynamic}(I)$ is given by:

$$A_{dynamic}(I) = \alpha \cdot A(I) + (1 - \alpha) \cdot F_{multi}(I) \quad (9)$$

Here, α is a learnable parameter which is initialized to a neutral value (e.g., 0.5) and then updated automatically during backpropagation along with all other model parameters. This process allows the network to dynamically weigh the contribution of the attention-guided features against the original features. The final refined feature map $F_{final-refined}(I)$ is obtained by element-wise multiplication of the dynamically weighted attention map $A_{dynamic}(I)$ with the multi-scale feature map $F_{multi}(I)$ using $F_{final-refined}(I) = F_{multi}(I) \odot A_{dynamic}(I)$.

By adapting attention weights in response to each input image, the model is able to accommodate varsying defect distributions. This dynamic attention mechanism enhances the model's ability to emphasize regions with a higher likelihood of defects, thereby improving detection accuracy.

3.3. Context-Aware Feature Fusion

The Context-Aware Feature Fusion module adaptively integrates features from different spatial scales and attention layers, guided by their relevance to the detection task. The core principle is to dynamically control the contribution of each feature stream, such as multi-scale versus attention-refined features, to the final fused representation. This ensures that the most useful information for a given input has the greatest influence on the fused feature map. Structurally, this is implemented through a learnable gating mechanism that computes input-dependent coefficients to scale each feature stream before they are aggregated. Let $F_{multi}(I)$ be the multi-scale feature map obtained from the Multi-Scale Feature Extraction module, and $F_{final-refined}(I)$ be the refined feature map obtained from the Dynamic Attention Mechanism. The Context-Aware Feature Fusion module integrates these feature maps to produce a final fused feature map. The fusion process can be represented as:

$$F_{fused}(I) = W_s \cdot F_{multi}(I) + W_a \cdot F_{final-refined}(I) \tag{10}$$

where W_s and W_a are learnable weights for the scale and attention features, respectively.

To ensure that the fusion process is context-aware, we introduce a gating mechanism that dynamically adjusts the weights based on the input image. Let $G(I)$ be the gating function that generates gating weights g_s and g_a for the scale and attention features, respectively. The gating function can be represented as:

$$G(I) = \sigma(W_g * I + b_g) \tag{11}$$

where W_g is the weight matrix, and b_g is the bias term. The gating weights g_s and g_a are obtained as $g_s = G_s(I)$ and $g_a = G_a(I)$ where $G_s(I)$ and $G_a(I)$ are the outputs of the gating function for the scale and attention features, respectively. The final fused feature map $F_{context-fused}(I)$ is computed as:

$$F_{context-fused}(I) = g_s \cdot F_{multi}(I) + g_a \cdot F_{final-refined}(I) \tag{12}$$

By integrating features from multiple scales and attention maps, the Context-Aware Feature Fusion module enhances the model's ability to detect and localize defects accurately, even in challenging conditions. This novel contribution ensures that the model adapts to the most relevant features.

3.4. Defect Detection and Localization

As shown in Figure 1, the final fused feature map $F_{context-fused}(I)$ is used to predict the bounding boxes and class labels for detected defects. The bounding boxes $B = \{b_1, b_2, \ldots, b_n\}$ are characterized by their coordinates (x_i, y_i, w_i, h_i), where x_i and y_i are the center coordinates, and w_i and h_i are the width, and height of the box, respectively. The class labels $L = \{l_1, l_2, \ldots, l_n\}$ represent the types of detected defects within each bounding box. The outputs of the framework are the set of bounding boxes B and their class labels L, which identify and localize defects in the input image.

3.5. Training Algorithm

The training procedure for the proposed MSA-CNN is formalized in Algorithm 1. This end-to-end optimization framework reflects the model's modular design, where multi-scale convolutional branches, dynamic spatial attention, and context-aware feature fusion are jointly trained. At each iteration, input samples from the PV Multi-Defect dataset are propagated through the network to produce bounding box coordinates and class logits. The model is trained using a combined loss that includes localization, object presence, and classification errors. This helps improve both accuracy and precision. Gradient-based updates are applied across all layers, which facilitates the coordinated learning of scale-adaptive features, attention maps, and fusion weights for robust defect detection.

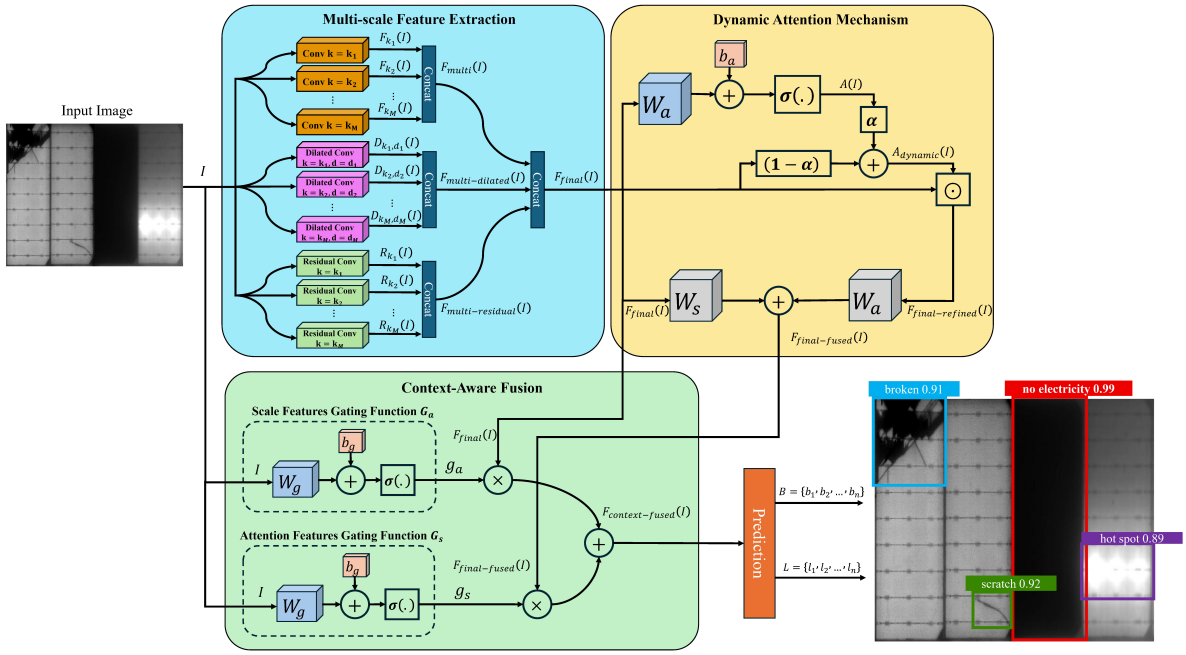


Figure 1. Overview of the proposed MSA-CNN framework for PV panel defect detection.

Algorithm 1 MSA-CNN Training for Solar Panel Defect Detection

Input: Training dataset $D = \{(I_i, B_i, L_i)\}_{i=1}^{N}$, where $I_i \in R^{H \times W \times 3}$ is an input image, $B_i = \{(x_j, y_j, w_j, h_j)\}_{j=1}^{n}$ are ground-truth bounding boxes, and $L_i = \{l_{i,j}\}_{j=1}^{n}$ are class labels for defects; model parameters θ (weights for convolutions, attention, fusion), and Hyperparameters: learning rate η, batch size m, epochs T.

Output: Trained model parameters θ^*

Initialization: Initialize all parameters θ randomly, Preprocess images (resize to 512×512, normalize pixels to [0,1])

Training Loop:

For $epoch = 1$ to T:

  Shuffle D and split into batches $\{B_1, \ldots, B_K\}$ of size m

  For each batch B_k:

    Forward Pass:

      Multi-Scale Feature Extraction:

$$F_{multi}(I) = Concat(F_{k_1}(I), F_{k_1}(I), \cdots, F_{k_M}(I))$$

      Dynamic Attention:

$$A(I_i) = \sigma(W_a * F_{multi}(I_i) + b_a)$$

$$F_{refined}(I_i) = F_{multi}(I_i) \odot (\alpha \cdot A(I_i) + (1-\alpha) \cdot F_{multi}(I_i))$$

      Context-Aware Fusion:

$$F_{context-fused}(I_i) = g_s \cdot F_{multi}(I_i) + g_a \cdot F_{final-refined}(I_i)$$

      Detection Head: Predict boxes $\hat{B}_i$ and labels $\hat{L}_i$

    Loss Computation:

      Localization Loss (Smooth L1):

$$\text{if } (\hat{B}_{i,j}^{c} - B_{i,j}^{c}) < 1$$

$$L_{box} = \Sigma_{j=1}^{n} \Sigma_c \, 0.5(\hat{B}_{i,j}^{c} - B_{i,j}^{c})^2$$

$$\text{else: } |\hat{B}_{i,j}^{c} - B_{i,j}^{c}| - 0.5$$

      Classification Loss (Cross-Entropy):

$$L_{class} = -\Sigma_{j=1}^{n} \Sigma_{l_{i,j}} \, l_{i,j} log(\hat{l}_{i,j})$$

      Total Loss: $L = \lambda_1 L_{box} + \lambda_2 L_{class}$

    Backward Pass:

      Compute gradients of parameter $\nabla_\theta L$

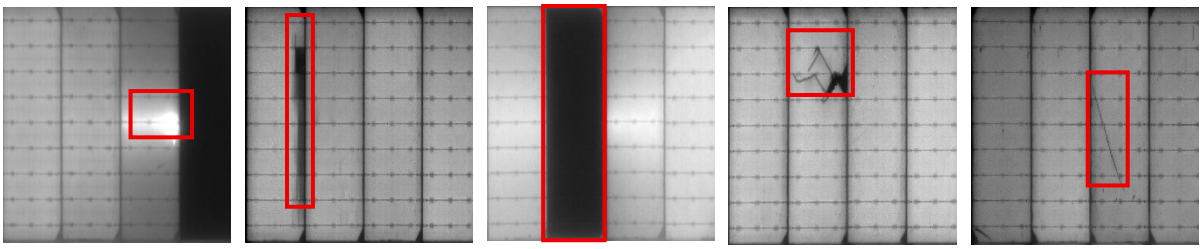
      Update: $\theta \leftarrow \theta - \eta \nabla_\theta L$

4. Numerical Results

4.1. Dataset

The PV Multi-Defect dataset [1] presents a detailed compilation of images tailored for the detection of surface anomalies on PV panels. This resource includes 1108 images, all standardized at a resolution of 600 × 600 pixels. For training, images were resized to 512 × 512. As our architecture relies on operations such as standard and dilated convolutions, this power-of-two input size speeds up the training process. These images were captured from PV modules with physical dimensions of 1.65 m by 0.991 m and consisting of 60 cells each. Preprocessing techniques, including grayscale transformation and cropping, were applied to center attention on defective areas. As illustrated in Figure 2, the dataset encompasses five primary defect classes: structural damage (broken cells), thermal faults (hot spots), edge degradation (black borders), surface abrasions (scratches), and electrical failure zones (no electricity). Each image includes annotations created with LabelImg [26], formatted according to the VOC2007 structure, identifying defect types and locations.

The dataset uses a 70-10-20 random split for training, validation, and testing, respectively, to promote comprehensive learning and robust evaluation. This distribution allows models to be trained on a large portion of the data while preserving unseen samples for validation and final assessment. The validation dataset is used for hyperparameter selection of the proposed method and benchmarks.



(**a**) hot spot (**b**) black border (**c**) no-electricity (**d**) broken (**e**) scratch

Figure 2. Examples of the five annotated defect categories in the PV Multi-Defect dataset**.**

4.2. Experimental Settings

The multi-scale solar panel defect detection framework was developed using Python 3.8 and implemented with the PyTorch deep learning framework. Experimental runs were conducted on a workstation equipped with an AMD Ryzen 9 7950X processor and an NVIDIA GeForce RTX 3090 Ti GPU featuring 24 GB of VRAM. This robust hardware setup significantly accelerated model training and allowed efficient processing of large-scale solar panel datasets.

4.3. Performance Metrics

The evaluation framework employs a dual-threshold approach to comprehensively assess model performance for photovoltaic defect detection. Following established practices in solar panel inspection, we utilize mAP@0.5 as our primary evaluation metric. This corresponds to mean average precision at 50% bounding box overlap, as determined by intersection-over-union (IoU), and reflects real-world operational requirements where identifying approximate defect locations is sufficient for maintenance purposes. This is complemented by precision ($P = TP/(TP + FP)$) and recall ($R = TP/(TP + FN)$) metrics, which quantify the model's ability to minimize false positives (FPs) and false negatives (FNs), respectively, with their harmonic mean represented through the F1-score ($F1 = 2 \cdot P \cdot R/(P + R)$). Here TP shows the number of true positives. For defects requiring precise localization, such as cracked cells where exact boundary delineation impacts repair decisions, we additionally report mAP@0.75 and corresponding average IoU values computed as the mean IoU between matched predicted and ground-truth bounding boxes across the test set. The average IoU provides direct measurement of localization accuracy, with the dual-threshold system offering both practical evaluation for general inspections (IoU = 0.5) and stringent quality control standards for critical applications (IoU = 0.75). This multi-faceted assessment ensures robust characterization of the model's detection capabilities across varying operational requirements, from rapid field surveys to detailed engineering analyses.

4.4. Hyperparameter Selection

The optimal combination of the number of convolutional layers (L), kernel size (K), dilation rate (D), and loss coefficients λ_1 and λ_2 are determined by conducting an exhaustive grid search. The hyperparameter search ranges were selected based on task-specific requirements and empirical design considerations. Kernel sizes from 2 to 9 were chosen to balance local texture sensitivity with broader contextual coverage. Smaller kernels (e.g., 3 × 3) are effective for capturing fine details such as hairline cracks or scratches, while larger kernels (e.g., 9 × 9) allow the network to model broader structures like black borders or cell-level anomalies. Dilation rates between 1 and 5 were included to expand the receptive field without significantly increasing model depth or computational load. Higher dilation values beyond 5 were excluded due to observed degradation in spatial coherence and diminishing returns in accuracy. Preliminary experiments indicated that deeper networks beyond 6 layers resulted in minimal performance gains while increasing the risk of overfitting, especially given the moderate dataset size. The search range for the loss coefficients, λ_1 and λ_2, was set to [0,2] with a step size of 0.1 to explore the full spectrum of balance between the localization (L_{box}) and classification (L_{class}) losses. An upper bound of 2 enables strong objective prioritization, allowing for up to a 2-fold weighting, as preliminary experiments confirmed that higher values led to training instability. Each configuration was evaluated using mean Average Precision (mAP@0.5) on the validation set. The training was conducted for 50 epochs using a fixed batch size of 16 and a learning rate of 0.001 with the Adam optimizer.

The results consistently identified L = 4, K = 5, D = 3, $\lambda_1 = 0.6$, and $\lambda_2 = 0.7$ as the optimal configuration, achieving the highest mAP@0.5 of 0.91 across all evaluated combinations. This configuration balanced model depth, receptive field size, and feature extraction capability effectively. Deeper networks beyond L = 4 showed diminishing returns, while larger kernel sizes (K > 5) or higher dilation rates (D > 3) did not provide additional improvements in detection accuracy. The selected hyperparameters ensure robust performance for multi-scale defect detection while maintaining computational efficiency.

The results presented in this paper demonstrate the consistency of the optimal configuration across different hyperparameter pairings, reinforcing the reliability of the chosen values. Further refinement could explore the interaction of these hyperparameters with other training parameters, such as learning rate schedules or alternative optimization strategies, to potentially achieve marginal gains in performance. However, the current selection provides a strong baseline for model architecture.

4.5. Performance Comparison

We evaluate the proposed MSA-CNN framework against six benchmark methods, including SVM [4], Faster R-CNN [15], Mask R-CNN [12], YOLO-v3 [21], Single Shot MultiBox Detector (SSD) [27], and GBH-YOLO-v5 [1] for PV panel defect detection using the PV Multi-Defect dataset. Table 1 presents a comparative summary using standard object detection metrics: mAP@0.5, mAP@0.75, precision, recall, F1-score, and average IoU. The methods are arranged in ascending order of mAP@0.5 to reflect the progression of algorithmic advancements. The baseline SVM model shows limited capability, with an mAP@0.5 of 45.3%, due to its reliance on handcrafted features lacking spatial and contextual modeling. The introduction of deep learning in Faster R-CNN yields a 24% absolute gain in mAP@0.5 and an 18.5% improvement in average IoU, highlighting the benefits of hierarchical feature extraction and region-based localization. Mask R-CNN further improves boundary accuracy with instance segmentation, increasing mAP@0.75 by 2.9% over Faster R-CNN, though at the cost of higher complexity. YOLOv3 introduces a one-stage grid-based detection pipeline that enhances recall and speeds up inference, achieving a 6.9% mAP@0.5 improvement over Mask R-CNN. SSD leverages multi-scale feature maps to better detect smaller defects, with a 4.3% mAP@0.5 increase over YOLOv3 and a 4.6% recall improvement. GBH-YOLOv5 significantly advances detection accuracy and efficiency by integrating Ghost Convolution and BottleneckCSP modules, offering an 11.8% gain over SSD while maintaining a lightweight architecture. Our proposed MSA-CNN outperforms all baselines, achieving 97.8% mAP@0.5 and 93.5% mAP@0.75. The observed 3.6% mAP@0.5 gain over GBH-YOLOv5 reflects the framework's superior ability to capture and integrate spatial patterns across a wide range of defect scales. This is particularly evident in its strong performance on both small, localized anomalies and broader, irregular defect regions, highlighting the effectiveness of the architecture's enhanced receptive field design. The 98.2% precision indicates a substantial reduction in false positives, suggesting that the model successfully isolates defect-relevant regions while suppressing visually similar noise, such as edges or texture inconsistencies. The high recall (96.7%) further confirms that subtle and low-contrast defects are consistently detected, due to improved semantic consistency across feature levels. Additionally, the average IoU of 0.912 demonstrates precise localization boundaries, especially for defects with non-uniform shapes, where previous models typically underperform on. Together, these results affirm that MSA-CNN's integrated

enhancements in multi-scale representation, adaptive attention, and contextual feature aggregation translate into concrete and robust improvements in PV detection accuracy.

Table 2 provides a detailed comparison of MSA-CNN and GBH-YOLOv5 across all defect categories, revealing consistent and meaningful performance gains. For broken cells, which often exhibit irregular boundaries and low contrast, MSA-CNN demonstrates superior localization accuracy, reflected in a higher average IoU and a well-balanced precision–recall profile. This indicates the model's effectiveness in resolving subtle structural patterns without overfitting to noise. In the case of hot spots, where variations in intensity and lighting conditions challenge consistent detection, MSA-CNN achieves marginal yet significant improvements, particularly in reducing false positives, which is critical for operational reliability in real-world settings. Detection of black borders benefits from MSA-CNN's attention-guided processing, with a 4.8% mAP@0.5 improvement over the baseline. This defect type is prone to misclassification due to low visual saliency and its proximity to panel edges, yet MSA-CNN's enhanced focus and semantic fusion yield sharper localization and more confident classification.

Table 1. Comprehensive performance comparison of defect detection methods.

Method	mAP@0.5	mAP@0.75	Precision	Recall	F1-Score	Avg IoU
SVM	45.3%	32.1%	51.2%	42.8%	0.466	0.481
Faster R-CNN	69.3%	58.7%	73.5%	68.2%	0.707	0.623
Mask R-CNN	71.2%	62.4%	76.8%	70.5%	0.735	0.654
YOLOv3	78.1%	69.8%	82.4%	77.3%	0.798	0.712
SSD	82.4%	74.6%	86.7%	81.9%	0.843	0.758
GBH-YOLOv5	94.2%	88.3%	95.1%	93.4%	0.942	0.873
MSA-CNN (Ours)	97.8%	93.5%	98.2%	96.7%	0.974	0.912

Table 2. Per-class detection performance comparison.

Defect Category	Method	mAP@0.5	Recall	Precision	F1-Score	Avg IoU
Broken Cells	GBH-YOLOv5	95.1%	93.8%	96.4%	0.951	0.892
	MSA-CNN	98.3%	97.1%	98.9%	0.980	0.928
Hot Spots	GBH-YOLOv5	96.7%	95.2%	97.3%	0.962	0.901
	MSA-CNN	98.6%	97.8%	99.1%	0.984	0.935
Black Borders	GBH-YOLOv5	91.4%	89.3%	92.6%	0.909	0.842
	MSA-CNN	96.2%	94.7%	97.5%	0.961	0.896
Scratches	GBH-YOLOv5	90.8%	88.1%	91.5%	0.897	0.831
	MSA-CNN	95.9%	94.3%	96.8%	0.955	0.893
No Electricity	GBH-YOLOv5	97.5%	96.2%	98.1%	0.971	0.918
	MSA-CNN	99.1%	98.4%	99.6%	0.990	0.949

The most substantial gains are observed in scratch detection, a category defined by small size, low contrast, and frequent overlap with other features. Here, MSA-CNN achieves notable increases in both recall and mAP, demonstrating its ability to capture fine-grained, low-saliency patterns with high spatial precision. Even in the no-electricity (dark area) category, where both models already perform well, MSA-CNN delivers marginal gains in precision and IoU, confirming its robustness across varying defect scales and contrast conditions. Overall, the model's improvements across all categories reflect the strength of its multiscale representation, adaptive attention, and context-aware fusion, offering high reliability in both common and challenging PV defect scenarios.

5. Conclusions

In this work, we addressed key limitations in photovoltaic panel defect detection by developing MSA-CNN, a compact and modular architecture designed to improve accuracy under diverse visual conditions. The proposed framework incorporates three core components: multi-scale feature extraction, dynamic attention, and context-aware fusion to enhance scale adaptability, spatial focus, and semantic integration, respectively. Extensive evaluations on the PV Multi-Defect dataset demonstrate that MSA-CNN consistently outperforms state-of-the-art methods. For instance, our model achieved an mAP@0.5 of 97.8%, representing a 3.6% improvement over the next-best method, GBH-YOLOv5. The performance gain was even more pronounced under stricter localization criteria, where MSA-CNN achieved a 5.2% higher mAP@0.75. Notably, it yields substantial improvements in identifying small and low-contrast defects, such as scratches, where it surpassed the strongest baseline by 5.1% in mAP@0.5. These results validate the framework's effectiveness and its potential for deployment in real-world PV inspection systems, including embedded platforms. Future work will address data scarcity and class imbalance

using a more data-efficient framework. We will explore unsupervised learning to build robust feature representations from unlabeled data and employ generative models to learn the underlying patterns of rare defects. This strategy reduces the need for manual annotation, enabling a more autonomous and economically viable system for industrial applications.

Author Contributions

A.F.B.: writing—original draft preparation, investigation, validation, formal analysis, data curation, conceptualization; M.K.: writing—review & editing, writing—original draft preparation, validation, supervision, software, resources, investigation, project administration, methodology, formal analysis; J.R.: methodology, software, formal analysis, investigation, data curation; M.E.K.: writing—review & editing, supervision.

Funding

This research received no external funding.

Data Availability Statement

The data that support the findings of this study are available within the article.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Li, L.; Wang, Z.; Zhang, T. Gbh-yolov5: Ghost convolution with bottleneckcsp and tiny target prediction head incorporating yolov5 for pv panel defect detection. *Electronics* **2023**, *12*, 561.
2. Tang, W.; Yang, Q.; Dai, Z.; et al. Module defect detection and diagnosis for intelligent maintenance of solar photovoltaic plants: Techniques, systems and perspectives. *Energy* **2024**, *297*, 131222.
3. Hijjawi, U.; Lakshminarayana, S.; Xu, T.; et al. A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations. *Sol. Energy* **2023**, *266*, 112186.
4. Juan, R.O.S.; Kim, J. Photovoltaic Cell Defect Detection Model Based-On Extracted Electroluminescence Images Using SVM s. In Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 19–21 February 2020; pp. 578–582.
5. Patel, A.V.; McLauchlan, L.; Mehrubeoglu, M. Defect Detection in PV Arrays Using Image Processing. In Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Vegas, NV, USA, 16–18 December 2020; pp. 1653–1657.
6. Guan, Y.; Wu, G.; Huang, W.; et al. Gray Level Co-Occurrence Matrix-Based Defect Detection Method for Photovoltaic Power Plant Panels. In Proceedings of the 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), Ottawa, ON, Canada, 26–28 August 2023; pp. 703–707.
7. Bordihn, S.; Fladung, A.; Schlipf, J.; et al. Machine Learning Based Identification and Classification of Field-Operation Caused Solar Panel Failures Observed in Electroluminescence Images. *IEEE J. Photovolt.* **2022**, *12*, 827–832.
8. Dolatyabi, P.; Regan, J.; Khodayar, M. Deep Learning for Traffic Scene Understanding: A Review. *IEEE Access* **2025**, *13*, 13187–13237.
9. Regan, J.; Khodayar, M. A triplet graph convolutional network with attention and similarity-driven dictionary learning for remote sensing image retrieval. *Expert Syst. Appl.* **2023**, *232*, 120579.
10. Saffari, M.; Khodayar, M. Low-Rank Sparse Generative Adversarial Unsupervised Domain Adaptation for Multitarget Traffic Scene Semantic Segmentation. *IEEE Trans. Ind.Inform.* **2024**, *20*, 2564–2576.
11. Masita, K.; Hasan, A.; Shongwe, T.; et al. Deep Learning in Defects detection of PV modules: A Review. *Sol. Energy Adv.* **2025**, *5*, 100090.
12. Rocha, D.; Alves, J.; Lopes, V.; et al. Multidefect detection tool for large-scale PV plants: Segmentation and classification. *IEEE J. Photovolt.* **2023**, *13*, 291–295.
13. Tang, W.; Yang, Q.; Xiong, K.; et al. Deep learning based automatic defect identification of photovoltaic module using electroluminescence images. *Sol. Energy* **2020**, *201*, 453–460.

14. Zyout, I.; Oatawneh, A. Detection of PV Solar Panel Surface Defects Using Transfer Learning of the Deep Convolutional Neural Networks. In Proceedings of the 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 4 February–9 April 2020; pp. 1–4.
15. Zhang, Z.; Cao, Y.; Yang, Q. Defect Detection and Classification of Photovoltaic Modules Based on Image Fusion Analysis. In Proceedings of the 2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2), Chengdu, China, 11–13 November 2022; pp. 2460–2465.
16. Nagar, S.; Mishra, M.K.; Rai, P.K. Using Densenet169 for Image-Based Classification of Solar Panel Defects. In Proceedings of the 7th International Conference on Contemporary Computing and Informatics (IC3I), Greater Noida, India, 18–20 September 2024; pp. 479–484.
17. Ma, W.; Chen, B.; Wang, B.; et al. Photovoltaic Panel Defect Detection via Multi-scale Siamese Convolutional Fusion Network with Information Bottleneck Theory. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5030815.
18. Zhou, P.; Fang, H.; Wu, G. PDeT: A Progressive Deformable Transformer for Photovoltaic Panel Defect Segmentation. *Sensors* **2024**, *24*, 6908.
19. Saffari, M.; Khodayar, M.; Jalali, S.M.J. Sparse Adversarial Unsupervised Domain Adaptation with Deep Dictionary Learning for Traffic Scene Classification. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 1139–1150.
20. Hussain, M.; Khanam, R. In-depth review of yolov1 to yolov10 variants for enhanced photovoltaic defect detection. *Solar* **2024**, *4*, 351–386.
21. Di Tommaso, A.; Betti, A.; Fontanelli, G.; et al. A multi-stage model based on YOLOv3 for defect detection in PV panels based on IR and visible imaging by unmanned aerial vehicle. *Renew. Energy* **2022**, *193*, 941–962.
22. Zhang, M.; Yin, L. Solar cell surface defect detection based on improved YOLO v5. *IEEE Access* **2022**, *10*, 80804–80815.
23. Liu, H.; Zhang, F. A Photovoltaic Panel Defect Detection Method Based on the Improved Yolov7. In Proceedings of the 2024 5th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM), Nanjing, China, 26–28 April 2024; pp. 359–362.
24. Wang, Y.; Zhao, J.; Yan, Y.; et al. Pushing the boundaries of solar panel inspection: Elevated defect detection with yolov7-gx technology. *Electronics* **2024**, *13*, 1467.
25. Hu, H.; Li, Y.; Wang, J.; et al. Aerial Photovoltaic Panel Infrared Image Defect Detection Method Based on Improved YOLOv8. In Proceedings of the 2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP), Xian, China, 19–21 April 2024; pp. 1777–1780.
26. Tzutalin. LabelImg. Available online: https://github.com/tzutalin/labelImg (accessed on 23 October 2025).
27. Kong, X.; Xu, W.; Xu, B.; et al. Defect detection of photovoltaic modules based on improved SSD algorithm. In Proceedings of the 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Chengdu, China, 3–5 November 2023; pp. 1063–1066.