sensors

MDPI

Article

PDeT: A Progressive Deformable Transformer for Photovoltaic Panel Defect Segmentation

Peng Zhou [1,2], Hong Fang [3] and Gaochang Wu [1,*]

[1] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China
[2] School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300130, China
[3] School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
* Correspondence: wugc@mail.neu.edu.cn

Abstract: Defects in photovoltaic (PV) panels can significantly reduce the power generation efficiency of the system and may cause localized overheating due to uneven current distribution. Therefore, adopting precise pixel-level defect detection, i.e., defect segmentation, technology is essential to ensuring stable operation. However, for effective defect segmentation, the feature extractor must adaptively determine the appropriate scale or receptive field for accurate defect localization, while the decoder must seamlessly fuse coarse-level semantics with fine-grained features to enhance high-level representations. In this paper, we propose a Progressive Deformable Transformer (PDeT) for defect segmentation in PV cells. This approach effectively learns spatial sampling offsets and refines features progressively through coarse-level semantic attention. Specifically, the network adaptively captures spatial offset positions and computes self-attention, expanding the model's receptive field and enabling feature extraction across objects of various shapes. Furthermore, we introduce a semantic aggregation module to refine semantic information, converting the fused feature map into a scale space and balancing contextual information. Extensive experiments demonstrate the effectiveness of our method, achieving an mIoU of 88.41% on our solar cell dataset, outperforming other methods. Additionally, to validate the PDeT's applicability across different domains, we trained and tested it on the MVTec-AD dataset. The experimental results demonstrate that the PDeT exhibits excellent recognition performance in various other scenarios as well.

Keywords: photovoltaic panel defects; defect segmentation; deformable attention; feature aggregation

check for updates

Citation: Zhou, P.; Fang, H.; Wu, G. PDeT: A Progressive Deformable Transformer for Photovoltaic Panel Defect Segmentation. *Sensors* **2024**, *24*, 6908. https://doi.org/10.3390/s24216908

Academic Editor: Hugo Aguas

Received: 28 September 2024
Revised: 23 October 2024
Accepted: 25 October 2024
Published: 28 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Defects in photovoltaic (PV) cell substrates can reduce photoelectric conversion efficiency, leading to a decrease in system power generation [1]. These defects may also cause uneven current distribution in localized areas, resulting in hotspot effects [2]. This abnormal temperature increase in defect regions not only further decreases power generation efficiency but can also damage PV panels and even pose safety hazards, such as fire risks. Therefore, incorporating defect detection technology in PV quality control processes is crucial for identifying and preventing such defects in advance, ensuring the efficient and safe operation of PV systems. Defect segmentation (pixel-level defect detection) [3] provides detailed information about defects, including their precise location, shape, and size. With accurate segmentation data, PV manufacturers can detect potential issues in the panels early, preventing further problems and losses during the system's operational phase [4]. This not only boosts the economic benefits for enterprises but also contributes to the sustainable development of clean energy.

Recent research has utilized deep learning-based segmentation models [5–15] to segment defects in solar panels. Compared to manual inspection, deep learning techniques offer significant advantages. First, deep learning can automatically process and analyze

large amounts of data, greatly improving inspection efficiency, especially in large-scale production settings. Second, deep learning models exhibit high consistency and precision in identifying complex features, avoiding fatigue and subjective errors that are common in manual inspections. Additionally, deep learning can continuously optimize and improve through data training, adapting to new types of defects and changing environments, which manual inspection cannot match in terms of adaptability and continuous improvement.

However, current defect segmentation algorithms still struggle with detecting elongated cracks and tend to miss or falsely detect small flaws around grid lines. During the feature extraction stage (encoding phase), Convolutional Neural Networks (CNNs) are limited by their fixed position sampling of input features. This constraint makes it difficult for models to handle fine-grained recognition tasks like detecting elongated cracks, leading to confusion or the loss of important features [3]. Although Vision Transformers (ViTs) [16] have been proven to effectively capture long-range feature relationships, the dense attention mechanism in ViTs results in high memory and computational costs. In the feature fusion and upsampling stage (decoding phase), the transmission of high-level semantic information to shallow layers gets disrupted and diluted by a large number of local patterns in the shallow layers [17].

Therefore, learning-based methods for PV defect segmentation need to address the following challenges. First, in the feature extraction phase, the encoder needs to adaptively select feature sampling locations. Second, in the feature fusion and upsampling phase, the decoder must refine features based on coarse-level semantics and prevent the dilution of fine-grained information.

In this paper, we propose a novel Progressive Deformable Transformer, dubbed PDeT, to effectively learn the offset of spatial sampling positions and progressively refine features based on coarse-level semantics for enhancing defect segmentation in PV panels. Specifically, we employ a deformable self-attention module in the encoder, which generates reference points as a unified network and takes query features as input for learning offsets, generating corresponding offsets for all reference points. In this way, the candidate keys/values can focus on defect regions and other crucial areas, enabling precise localization of defect features in complex backgrounds. This module is more flexible and efficient, allowing the dynamic adjustment of attention regions, thereby capturing the edges and details of defects more effectively and improving the accuracy of segmentation. In the decoder, we introduce a semantic aggregation module that combines feature differences across four different scales, reassembling coarse-level features with fine-grained ones. This approach effectively distinguishes between background and defect features in PV panels, providing clear and detailed spatial boundaries without losing semantic information. Our PDeT is capable of flexibly adjusting spatial sampling positions and progressively refining feature extraction using coarse-level semantic information, thereby improving the accuracy and reliability of PV defect detection.

The main contributions of this paper are summarized as follows:

- A progressive deformable Transformer is proposed to achieve high-quality segmentation of PV panel defects, significantly enhancing the ability to detect complex defect features.
- A deformable self-attention module is introduced to adaptively learn spatial feature sampling locations. This module adjusts sampling positions based on the shape and structure of input features, flexibly capturing irregular or deformed features.
- A semantic aggregation module is designed to ensure the retention and integration of both coarse and fine-grained information, effectively balancing these features while incorporating contextual information for better segmentation accuracy.

We conducted extensive experiments to validate the effectiveness of the introduced modules. On the solar cell defect dataset, our model achieved 88.41% mIoU. Additionally, we also conducted training and evaluation on the MVTec-AD [18] dataset. The experimental results indicate that the PDeT demonstrates exceptional recognition capabilities across different scenarios.

2. Related Work

This paper focuses on photovoltaic defect segmentation using a Vision Transformer backbone. To provide context, we present a brief overview of relevant research in this area.

2.1. Photovoltaic Defect Segmentation

The photovoltaic industry has extensively adopted deep learning-based methods for defect detection, offering significant improvements in both detection efficiency and accuracy compared to traditional approaches. By analyzing features from large-scale datasets, deep learning enables the precise identification of defects in photovoltaic panels. Tang et al. [15], for instance, used limited samples to generate high-resolution electroluminescence images and employed CNNs to automatically classify defects. Jiang et al. [13] proposed an m-shaped architecture to extract and fuse shallow and deep features in segmentation networks, incorporating an attention module to suppress the photovoltaic background and improve segmentation accuracy. Xie et al. [12] embedded a domain discriminator into the CNN to distinguish between data domains, allowing the feature extractor to learn domain-invariant features. Pratt et al. [19] evaluated four deep learning models (U-Net, PSPNet, and DeepLabv3+) for detecting cracks, inactive areas, and grid line defects in photovoltaic panels. Jha et al. [11] introduced a semi-supervised semantic segmentation method for defect detection with limited labeled data, significantly reducing manual annotation costs. Kaligambe et al. [10] developed a lightweight CNN model and compared its performance against a fine-tuned VGG16 model. Fioresi et al. [8] used ResNet50 and DeepLabv3 as backbone networks and segmentation decoders to identify defects such as cracks, contact interruptions, cell interconnection failures, and contact corrosion in both polycrystalline and monocrystalline silicon cells. Chen et al. [7] fine-tuned a U-Net model with a pre-trained VGG16 encoder, achieving excellent segmentation results for cracks and busbars. Similarly, Zhang et al. [6] refined CNN-extracted features using a global pairwise similarity module and a connection saliency module, enhancing defect detection in photovoltaic images.

2.2. Vision Transformer

The Vision Transformer (ViT) [16] has revolutionized computer vision tasks by leveraging the self-attention mechanism originally designed for natural language processing. By treating images as sequences of patches, ViT captures global context more effectively than traditional convolutional neural networks (CNNs). However, the high computational complexity and large data requirements have led to the development of more efficient Transformer variants. One such model is the Data-Efficient Image Transformer (DeiT), introduced by Touvron et al. [20], which improves the training process of ViT through data augmentation and advanced techniques, making it more suitable for smaller datasets.

In addition to image classification, ViT has been successfully applied to tasks like object detection and segmentation [21,22]. To enhance Transformer-based architectures further, Liu et al. [23] developed the Swin Transformer, which utilizes a hierarchical structure and local window self-attention, significantly improving computational efficiency and achieving state-of-the-art performance across multiple benchmarks.

Beyond these advancements, the neighborhood attention Transformer [24] offers an alternative approach by refining local attention mechanisms. Unlike the standard self-attention mechanism that emphasizes global information, NAT focuses on attention within local neighborhoods, making it particularly effective for tasks requiring fine-grained feature extraction. In summary, while ViTs laid the foundation for Transformer-based models in computer vision, optimizations in models like the DeiT, Swin Transformer, and neighborhood attention Transformer have addressed challenges such as computational complexity and fine feature extraction, broadening their applicability to a wider range of visual tasks.

3. Methodology

In this section, we delve into the specific architecture of the proposed Progressive Deformable Transformer (PDeT) designed for photovoltaic panel defect segmentation. The ViT [16] was the first to demonstrate the viability of Transformer mechanisms in computer vision. Unlike CNNs, Transformer-based models, with their Multi-Head Self-Attention (MHSA) mechanism, excel at capturing long-range dependencies, making them particularly effective for target localization. Building on this foundation, we introduce a deformable sampling operation before the MHSA in each Transformer encoder. This operation enables the network to focus more effectively on relevant regions. In the decoder, we propose a semantic aggregation module, which enhances the fusion of multi-stage features, improving overall segmentation accuracy. The following subsections will provide a detailed explanation of the network framework, the Deformable Transformer Block, and the Semantic Aggregation Block.

3.1. PDeT Overall Architecture

Our PDeT follows the widely used encoder–decoder structure. The encoder extracts feature information from the input image, while the decoder progressively restores the spatial dimensions to generate pixel-level segmentation results. As illustrated in Figure 1, the input image of shape $H \times W \times 3$ is first processed through two convolutional and normalization embedding layers, producing patch embeddings of size $H/4 \times W/4 \times C$ (where $C = 128$). Each convolution employs a kernel size of 3×3, with a stride of 2 and padding of 1, expanding output channels from 3 to 64 and then to 128. To construct a multi-level feature pyramid, our backbone is organized into four stages, each with a progressively increasing stride. Each stage comprises N stacked neighborhood attention Transformer blocks (NATs) [24] and Deformable Transformer Blocks (DTBs). In our implementation, $\{N_1, N_2, N_3, N_4\}$ are set to $\{1, 2, 9, 1\}$, respectively. The feature maps extracted at each stage are downsampled by a factor of 4, 8, 16, and 32 times the input size, denoted as $f = \{f_i \mid 1 \leq i \leq 4\}$.

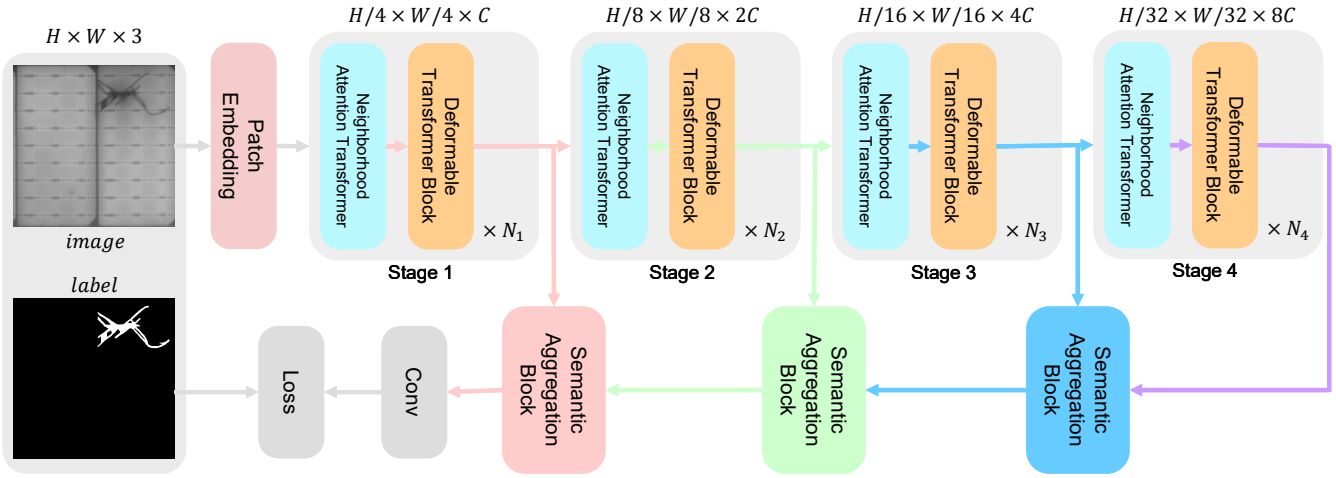


Figure 1. Overall architecture of the proposed PDeT for photovoltaic defect segmentation.

Next, we aim to seamlessly fuse and upsample the multi-level features in the decoder stage. We employ the Semantic Aggregation Block (SAB) to fuse adjacent features $\{f_i, f_{i+1}\}$, with the output serving as the input to the subsequent SAB. This process reduces inconsistencies between feature levels, allowing coarse-level semantic information to effectively refine finer features, thereby achieving a balance between global semantics and local details. The output from the encoder is processed through two rounds of bilinear interpolation and convolution operations to generate the predicted output.

3.2. Deformable Transformer Block

Before introducing deformable attention into the Transformer, we will first review the vanilla MHSA mechanism in the Transformer. Given an input feature map $x \in R^{H\times W\times C}$, the MHSA mechanism with M attention heads is applied, where each attention head independently learns different relationships. The outputs from these attention heads are then concatenated. The vanilla MHSA mechanism can be formulated as follows:

$$\begin{aligned} q = xW_q, k = xW_k, v = xW_v, \\ y^{(m)} = \text{softmax}\left(\frac{q^{(m)}k^{(m)T}}{\sqrt{d}}\right)v^{(m)}, m \in \{1,\ldots,M\}, \\ y = \text{concat}(y^{(1)},\ldots,y^{(M)})W_o, \end{aligned} \tag{1}$$

where $d = \frac{C}{M}$ represents the dimension of each attention head, and $q^{(m)}, k^{(m)}, v^{(m)}$ denote the query, key, and value embeddings, respectively. The output of the m-th attention head is represented as $y^{(m)}$. The learnable parameters $W_q, W_k, W_v, W_o \in R^{C\times C}$ are the projection matrices for the feature embeddings. The formulation for the i-th Transformer block is given as follows:

$$\begin{aligned} y'_i &= \text{MHSA}(\text{LN}(y_{i-1})) + y_{i-1}, \\ y_i &= \text{MLP}(\text{LN}(y'_i)) + y'_i, \end{aligned} \tag{2}$$

where $LN(\cdot)$ represents Layer Normalization [25], $\text{MHSA}(\cdot)$ denotes the MHSA mechanism in Equation (1), and $\text{MLP}(\cdot)$ is the MLP with two linear layers.

Despite the vanilla MHSA mechanism's effectiveness in capturing global information, it has certain limitations when processing local features and irregular shapes. To address these issues, we incorporate deformable operations into our PDeT. Deformable operations allow flexible attention adjustments, enabling the model to better focus on critical information. This is especially effective for handling complex or irregular features in PV defect segmentation. Next, we will provide a detailed explanation of how deformable operations are integrated into MHSA to enhance the performance of Transformer models.

As shown in Figure 2, our goal is to enable the key–value pairs to adaptively sample feature information based on offsets provided by the offset network. First, given the input feature $x \in R^{H\times W\times C}$, a uniform grid of reference points $r \in R^{H_r\times W_r\times 2}$ is generated. This grid is downsampled by a factor of relative to the size of the input feature, where $H_r \in \frac{H}{r}$ and $W_r \in \frac{W}{r}$. The values of the reference points correspond to 2D coordinates $(0,0),\ldots(H_r-1, W_r-1)$. The ranges of the two dimensions, $[0,(H_r-1)]$ and $[0,(W_r-1)]$, are then normalized to $[-1,+1]$, where $(-1,-1)$ represents the reference point at the top-left corner and $(+1,+1)$ represents the reference point at the bottom-right corner.

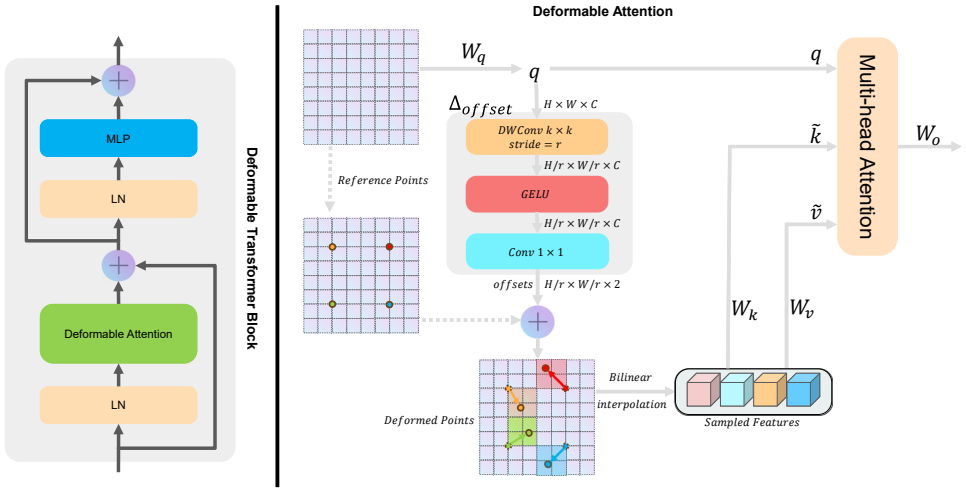


Figure 2. Detailed structure of the Deformable Transformer Block (DTB).

Next, we need to obtain the offsets for each reference point. We start by using W_q to generate the query tokens q, which are then input into the Δ_{offset} module to learn the offsets for the reference points. This module consists of a depth-wise convolution, a GELU activation function, and 1×1 convolution, resulting in an output with the same shape as the reference points. The calculated offsets are then added to the reference points to derive the positions of the deformable points. Sampling is performed at these deformable points, and the keys and values are computed using the projection matrices. At this stage, we obtain the deformable keys $\tilde{k}$ and values $\tilde{v}$, which are combined with q to compute the MHSA. Consequently, Equation (1) is updated to:

$$\begin{aligned} q &= xW_q, \tilde{k} = \tilde{x}W_k, \tilde{v} = \tilde{x}W_v \\ \Delta_{\mathbf{p}} &= \Delta_{offset}(q), \\ \tilde{x} &= \varphi(x; \mathbf{p} + \Delta_{\mathbf{p}}), \end{aligned} \tag{3}$$

where $\Delta_{\mathbf{p}}$ represents the offset of the reference point, $\mathbf{p}$ represents the position of the reference point, and $\varphi(a; \mathbf{b})$ denotes the sampling result at location $\mathbf{b}$ in the feature map a.

This deformable operation enables the Transformer structure to adaptively sample based on the query features, allowing the model to flexibly adjust the shape and position of the self-attention, particularly when handling features that span large scales. Moreover, by integrating the deformable operation, the only component that introduces a small number of learnable parameters is the Δ_{offset} module. This is one reason why we employ lightweight depth-wise convolution, ensuring that the model remains efficient while maintaining performance.

3.3. Semantic Aggregation Block

The encoder of our PDeT incorporates multiple stages to generate feature maps at different hierarchical levels. To progressively fuse the feature maps extracted from each stage, we integrate the Semantic Aggregation Block (SAB), which extracts and merges features across these levels. This block effectively preserves coarse semantic information while retaining fine-grained details, ultimately leading to high-quality segmentation results. Figure 3 illustrates the detailed structure of the SAB.

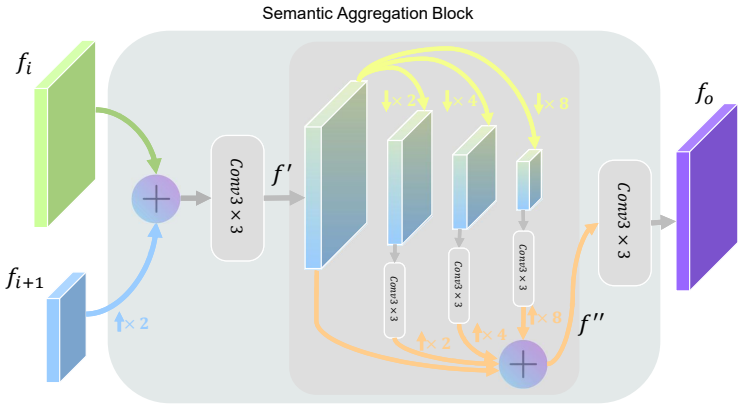


Figure 3. Detailed structure of the Semantic Aggregation Block (SAB).

The SAB processes feature maps from two different stages, f_i and f_{i+1}, where $i \in \{1, 2, 3\}$. We first align the channel dimensions of the feature maps into $C = 256$ via a 1×1 convolution. Then, a bilinear interpolation is applied to upsample the higher-level semantic feature f_{i+1}, resizing it to match the dimensions of the lower-level feature f_i. These operations lay the groundwork for subsequent feature fusion. The interpolated feature is then added to the lower-level feature, followed by a 3×3 convolution layer for

initial feature extraction. This step not only enhances the fused feature representation but also captures richer detail information. The process is formulated as:

$$f' = \mathrm{Conv}_{3\times3}(f_i + \phi_{bi}(f_{i+1}; 2)), \tag{4}$$

where $\phi_{bi}(\cdot; 2)$ indicates the 2×2 bilinear upsampling operation, $\mathrm{Conv}_{3\times3}$ represents the 3×3 convolution layer, and f' is the intermediate feature map within the SAB.

Next, the SAB applies an average pooling operation at multiple scales, transforming the feature map into various spatial resolutions. This multi-scale processing allows the model to capture features comprehensively, from local details to global context. It enhances the model's ability to perceive defect regions at different scales and improves its adaptability to diverse scenes and variations. After convolutional processing, the features from each sub-branch are upsampled to the original size and combined with the input feature map through residual connections. These connections ensure that original feature information is not lost during multi-scale transformations, preserving both fine-grained details and global semantic information, thereby enhancing the stability and expressiveness of the fusion. The process can be formulated as:

$$f'' = f' + \sum_{j\in\{2,4,8\}} \phi_{bi}\big(\mathrm{Conv}_{3\times3}\big(\mathrm{Pool}(f'; j)\big); j\big), \tag{5}$$

where $\mathrm{Pool}(\cdot; j)$ indicates the average pooling operation at scale $j \in \{2, 4, 8\}$.

Finally, a 3×3 convolution layer is applied to the fused features for further refinement and smoothing, generating the output of the SAB as $f_o = \mathrm{Conv}_{3\times3}(f'')$. This step enhances the features after multi-scale processing, improving the model's contextual awareness and leading to more robust feature representations.

Overall, the proposed SAB achieves a balance between coarse and fine-grained information through multi-scale processing, enabling the model to generalize effectively when dealing with targets of varying scales, shapes, and distributions, thereby improving its adaptability and expressive power in complex photovoltaic environments.

4. Experiments

We conducted ablation and comparison experiments on the photovoltaic dataset to validate the performance of the PDeT network. Additionally, to assess the applicability of the PDeT network in other industrial scenarios, we selected four industrial scenes from the public MVTec-AD [18] dataset for further evaluation.

4.1. Implementation

All experiments were conducted using an NVIDIA GeForce RTX 3060 (12GB) and an Intel Core i5-12490F processor, both sourced from Santa Clara, CA, USA. The model was built and trained using the PyTorch 1.12.0 framework. During training, the AdamW [26] optimizer was employed with an initial learning rate of 0.0001 and a weight decay coefficient of 0.0001. The exponential decay rates for computing the first- and second-moment estimates of gradients were set to 0.9 and 0.999, respectively. A polynomial learning rate schedule (with a power of 0.9) was adopted, allowing dynamic adjustment of the learning rate throughout training. The maximum number of iterations was set to 160,000, with model evaluations performed every 8000 iterations using the mIoU as the evaluation metric. Cross-entropy loss was used to calculate the loss, ensuring efficient model convergence and accurate classification.

The photovoltaic dataset used in our experiments contains a total of 1024 images, which were split into training and validation sets at a 2:1 ratio. Each image has a resolution of 640×640, and the segmentation results are classified into two categories: defect and background. During training, we applied random cropping as a data augmentation technique to both the input images and their corresponding labels.

4.2. Ablation Study

We conducted ablation experiments using the photovoltaic panel dataset. To validate the effectiveness of the deformable convolution, we sequentially replaced the NAT Block in the feature extraction stage with the DTB. In the overall network structure described in Section 3, we stack N_i Transformer blocks in each stage. Each Transformer block originally consisted of two NATs in series as the baseline, and then, we replaced the second NAT in each Transformer with the DTB. The results, as shown in Table 1, indicate that while adding the DTB only in the fourth stage slightly decreases model performance, incorporating the DTB in the third stage yields an improvement of about one percentage point. When all stages utilize the DTB, the model performance increases by approximately three percentage points compared to the baseline.

Table 1. Ablation study on the application of deformable attention at different stages. ✗ and ✓ indicate whether DTB is used in the stage.

Stage #/Deformable Attention				mIoU
Stage 1	Stage 2	Stage 3	Stage 4	
✗	✗	✗	✗	85.27
✗	✗	✗	✓	85.05
✗	✗	✓	✓	88.22
✗	✓	✓	✓	86.17
✓	✓	✓	✓	88.41

As shown in Figure 1, we replaced the SAB with standard convolution and bilinear interpolation to establish a comparison baseline, effectively omitting the transition from f' to f'' . We then sequentially integrated multi-scale methods into the different stages of the decoder. As shown in Table 2, using SAB in stages 1 and 2 resulted in a performance improvement of 0.62%, while adding SAB from stages 2 to 3 further increased performance by 1.12%.

Table 2. The results of the ablation experiment for the SAB module in the encoding structure. The check mark (✓) signifies the utilization of the SAB, whereas the cross (✗) indicates its exclusion.

Stage #/SAB			mIoU
Stage 1–Stage 2	Stage 2–Stage 3	Stage 3–Stage 4	
✗	✗	✗	87.25
✓	✗	✗	87.87
✓	✓	✗	88.37
✓	✓	✓	88.41

To evaluate the impact of different Decoder heads on model performance while keeping the encoder architecture unchanged, we conducted comparative experiments. This experiment utilized the encoder proposed in this paper as the baseline architecture and introduced various Decoder heads to observe their performance in specific tasks. The experimental results, as shown in Table 3, indicate that the IoU for the background class remained consistent across all Decoder heads. However, significant differences were observed in the IoU for the defect class among the various Decoder heads.

Specifically, the IoU for the defect class with the FPN [27] head is 74.52%, resulting in an overall mIoU of 87.25%. The UPer [28] head shows a slight improvement, with a defect class IoU of 75.36% and an overall mIoU of 87.67%. The most significant enhancement is observed with the proposed PDeT model I, achieving a defect class IoU of 76.85% and an overall mIoU of 88.41%. This indicates that the PDeT demonstrates advantages in fine-grained feature aggregation when processing defect features, significantly improving the model's performance in defect detection tasks.

Table 3. A comparison of the experimental results with other Decoder heads.

Decoder Head	IoU		mIoU
	Background	Defect	
FPN head [27]	99.98	74.52	87.25
UPer head [28]	99.98	75.36	87.67
PDeT head (ours)	99.98	76.85	88.41

Figure 4 illustrates the training process of the PDeT model. During the initial phase of training, the loss fluctuates significantly, indicating that the model is still learning and adjusting its parameters. However, as training progresses, the model gradually learns to capture the characteristics of the data, leading to a steady improvement in both the overall mIoU and mAcc metrics. Notably, the best mIoU is achieved at the 300th epoch, reaching 88.41%, which reflects the model's superior performance at this stage. Meanwhile, the loss value gradually converges and stabilizes at a lower level, indicating that the model effectively reduced its error rate during training.

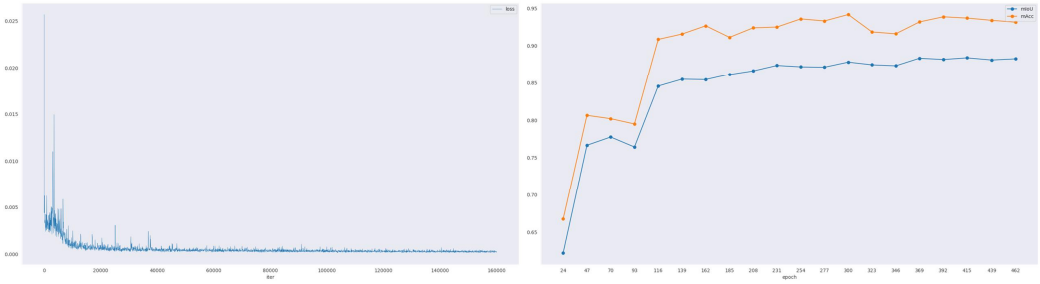


Figure 4. The training process of the PDeT is assessed using three metrics: loss, mIoU, and mAcc. These metrics offer valuable insights into the model's performance and effectiveness during training. The left panel displays the loss at each iteration, while the right panel presents the validation results throughout the training process.

4.3. Comparison with Other Segmentation Networks

In this section, we compare our method with existing semantic segmentation networks, including EMANet [29], STDC [30], DDRNet [31], K-Net [32], DNLNet [33], CCNet [34], ANN [35], DMNet [36], PIDNet [37], ISANet [38], GCNet [39], and SIIF [3].

Table 4 provides a detailed comparison of various models in the task of defect detection on photovoltaic panels, including metrics such as model parameters (Params), floating-point operations per second (FLOPs), frames per second (FPS), precision, recall, F1 score, and mIoU. Precision represents the proportion of true positive samples among those predicted as positive by the model. High precision indicates that most of the detected defects are genuine. Recall, on the other hand, measures the proportion of correctly identified positive samples out of all actual positive cases. High recall signifies that the majority of actual defects have been successfully detected. Additionally, mIoU reflects the model's ability to distinguish between defect areas and the background. A higher mIoU ensures greater detection accuracy and fewer missed defects.

Analysis reveals that the PDeT model outperforms others by at least four percentage points in mIoU, demonstrating its exceptional performance in photovoltaic defect segmentation. This advantage is primarily due to PDeT's feature extraction network, which employs a Transformer architecture. However, this comes at the cost of a relatively large number of parameters, leading to increased computational overhead. Thus, there is still room for optimization in the inference efficiency of the PDeT. In contrast, models like DDRNet-S and STDC, while having fewer parameters and higher computational efficiency, generally perform worse in the actual segmentation results, showing a significant gap compared to the PDeT. Despite their successes in lightweight design, these models exhibit relatively

limited feature extraction capabilities when dealing with complex defect backgrounds, making it challenging to accurately capture the details of defect regions.

Table 4. Comparison of experimental results of various indicators with other segmentation networks.

Method	Params	FLOPs	FPS	Precision	Recall	F1	mIoU
EMANet [29]	42.08M	263.17G	15.37	34.03	40.53	37.00	61.78
STDC [30]	8.57M	13.24G	78.05	64.88	51.99	57.72	70.02
DDRNet-S [31]	7.72M	9.267G	143.24	68.20	58.62	63.50	73.00
K-Net [32]	62.29M	285.92G	5.99	33.69	82.87	47.91	74.65
DNLNet [33]	50.1M	312.25G	5.75	58.91	79.89	67.82	75.82
CCNet [34]	49.81M	313.09G	12.97	65.29	79.81	71.83	78.15
ANN [35]	46.22M	289.28G	14.00	72.78	78.24	75.41	80.20
DMNet [36]	53.16M	305.64G	13.15	73.95	78.65	76.23	80.61
PIDNet-L [37]	36.93M	53.83G	26.23	71.70	81.40	76.25	80.79
ISANet [38]	37.69M	233.42G	15.37	81.46	75.67	78.46	81.84
GCNet [39]	49.62M	308.89G	13.85	76.93	80.35	78.60	82.51
SIIF [3]	58.72M	75.74G	25.09	79.80	83.03	81.38	84.30
PDeT (our)	104.15M	200.22G	10.49	90.12	89.37	89.66	88.41

Figure 5 presents a detailed comparison of different networks in the segmentation of defects on photovoltaic panels. By examining the results, we can clearly observe the differences among models when addressing specific types of defects, particularly in the segmentation of elongated cracks, where the discrepancies are especially pronounced. From the segmentation results in the first and third rows, it is evident that many models struggle with elongated cracks, often producing coarse results. These defects typically exhibit fine, elongated features, and conventional segmentation networks tend to overlook edge information when dealing with such detail-rich defects, leading to imprecise segmentation outcomes. For instance, while K-Net, a model with a larger number of parameters, has strong global perception capabilities during feature extraction, allowing it to capture the general outline of defects, the edges of the cracks in its output still appear blurred, failing to accurately segment the detailed parts of elongated defects.

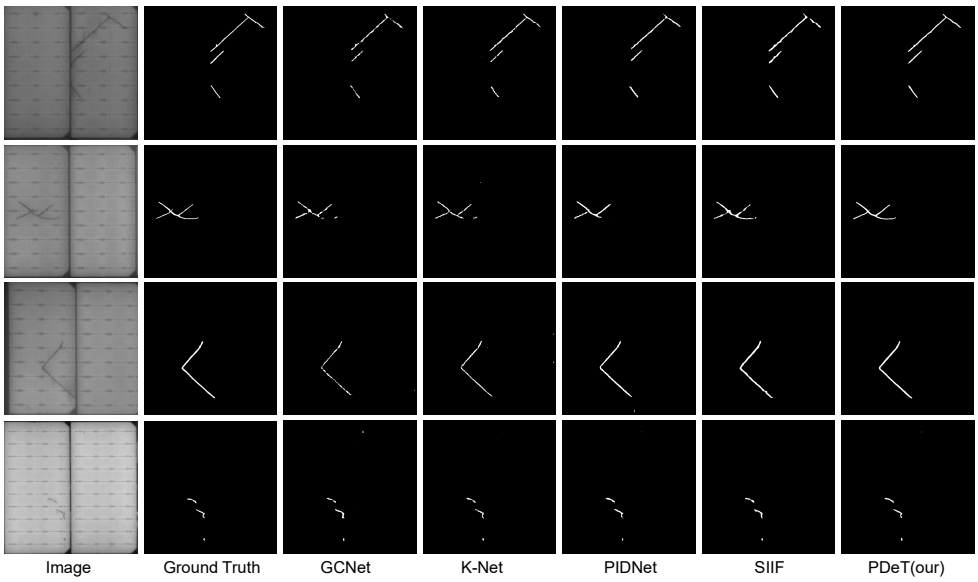


Figure 5. Quantitative comparison with other segmentation networks.

In contrast, the PDeT demonstrates superior capability in handling such defects. The model employs deformable self-attention, which allows for dynamic adjustment of sampling point positions, enabling a more flexible focus on the edges of cracks and, thus, accurately capturing these details. Moreover, during the decoding process, the SAB module

plays a crucial role in refining segmentation. This module progressively aggregates coarse and fine feature information by fusing multi-scale features, allowing for precise segmentation of defect edges without losing semantic information. Due to the need for further optimization of the model's parameter count and computational load, we will consider employing pruning and quantization methods suitable for Transformer architectures in the future. These approaches aim to reduce the model's computational cost and memory usage. Figure 5 clearly illustrates that the PDeT's segmentation results outperform other models in detailing and edge treatment of cracks. This refined segmentation ability enables the PDeT to maintain global perception while flexibly adjusting its focus areas and enhancing the feature aggregation process, ultimately improving defect detection accuracy.

4.4. Applicability in Other Industrial Fields

To validate the application potential of the PDeT photovoltaic defect detection model in other industrial scenarios, we selected four typical scenes from the MVTec-AD [18] anomaly detection dataset for experimentation. These scenes include Wood, Metal Nut, Tile, and Hazelnut, representing common complex backgrounds and diverse defect types in industrial manufacturing, thus providing high representativeness. The selection of these scenes aims to ensure that the PDeT model demonstrates good adaptability across different materials and defect patterns, further validating its broad potential for practical industrial applications.

The MVTec-AD dataset is primarily used for anomaly detection and contains both normal and abnormal data. In our experimental design, we performed stratified sampling for each scene, dividing the normal and abnormal data into training and testing sets at a 2:1 ratio. This approach simulates real industrial defect detection tasks and aligns more closely with actual application requirements. The evaluation metric for the model is mIoU, which is widely used in semantic segmentation tasks and provides a comprehensive assessment of the model's recognition performance across various categories.

In this experiment, we compared the performance of K-Net, UPerNet, GCNet, and SIIF models across four scenes in the MVTec-AD dataset. The experimental results are presented in Table 5, with each column representing the mIoU for the respective scenes.

Table 5. Comparative experimental results of the model across the four scenes: Hazelnut, Metal Nut, Tile, and Wood. All values shown in the table represent the mIoU for assessing the model's recognition performance in each scene.

Method	Hazelnut	Metal Nut	Tile	Wood
K-Net [32]	76.06	95.29	92.81	74.17
UPerNet [28]	93.57	96.6	95.44	88.35
GCNet [39]	93.27	95.32	95.07	87.93
SIIF [3]	94.13	95.22	92.95	87.86
PDeT (our)	94.34	95.32	92.46	88.54

The performance of K-Net [32] across the four scenes is relatively low, with an mIoU of 76.06% for Hazelnut and 74.17% for Wood, indicating its limited recognition capability in complex backgrounds. In contrast, UPerNet [28] performed exceptionally well in all scenes, achieving mIoU of 93.57% and 96.6% for Hazelnut and Metal Nut, respectively, demonstrating its superior feature extraction ability. The performances of GCNet [39] and SIIF [3] are comparable to UPerNet, maintaining high mIoU values across the scenes. Notably, in the Metal Nut scene, GCNet [39] achieved an mIoU of 95.32%, indicating a certain advantage in detecting defects in metal nuts. It is noteworthy that the proposed PDeT model performed well across all four scenes: Hazelnut, Metal Nut, Tile, and Wood. The mIoU for the Hazelnut scene was 94.34%, while for the Metal Nut scene, it was 95.32%. This indicates that the PDeT model demonstrates strong adaptability and stability across different materials and defect patterns.

Figure 6 presents a quantitative display of the segmentation results of the comparative models across four different scenes. In the first row, under the wood scene, the model effectively segments the liquid and combined areas, while the K-Net model shows relatively poor segmentation results. In the tile scene (third row), the model struggles with segmenting fine cracks, resulting in discontinuities; in contrast, our PDeT model performs significantly better. In the hazelnut scene (fifth and sixth rows), the results indicate that K-Net's segmentation is subpar, while other models demonstrate notably superior performance. In the Metal Nut scene, all models are capable of effectively identifying anomalous defects, showcasing their robust segmentation capabilities.

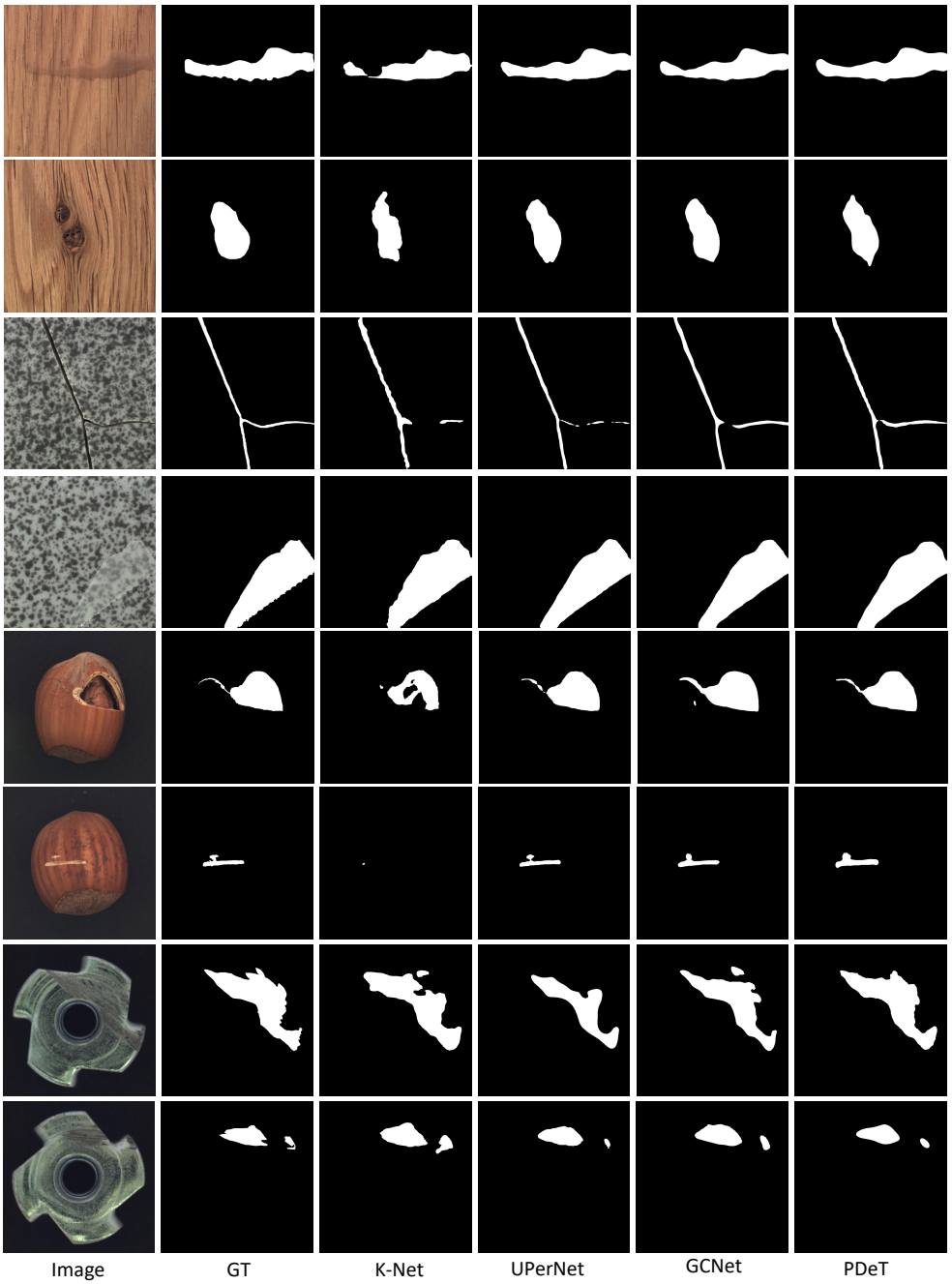


Figure 6. Quantitative display of segmentation results from multiple models across four distinct scenes, with each scene presenting two quantitative outcomes. The columns represent the segmentation results of the model.

These results underscore the practical applicability of our model, especially in diverse industrial contexts. By incorporating deformable self-attention and semantic aggregation modules, we not only enhance the geometric distinction capabilities but also refine the representation of semantic features. The introduction of these modules facilitates efficient collaboration between the encoder and decoder, ensuring precise feature transmission and gradual reconstruction, thereby allowing the PDeT to exhibit greater adaptability and stability in real-world applications.

5. Conclusions

Our research leverages image sensors for defect detection in photovoltaic panels. Compared to traditional electrical sensor-based methods, such as open-circuit voltage and short-circuit current measurements, we employ electroluminescence (EL) imaging to capture high-resolution, two-dimensional signals, significantly enhancing defect detection accuracy. Additionally, the deep learning techniques incorporated in our study not only improve the perception of defects in photovoltaic panels but also pave the way for advancements in intelligent recognition systems. To further advance this field, we have successfully proposed a Progressive Deformable Transformer for photovoltaic panel defect segmentation, which enhances the segmentation of defects in solar panels. By incorporating deformable self-attention and a semantic aggregation module, we not only improved the ability to differentiate geometric shapes but also refined the representation of semantic features. Furthermore, the introduction of these modules facilitates efficient collaboration between the encoder and decoder, ensuring precise feature transfer and progressive reconstruction. Researchers can flexibly adjust the model based on practical application needs to optimize its performance in specific scenarios. This achievement has significant implications not only for photovoltaic defect segmentation but also provides insights and exploration avenues for research in other industrial defect segmentation fields. The comprehensive experimental results indicate that our method achieved an mIoU of 88.41% on the photovoltaic dataset, outperforming existing segmentation algorithms. Additionally, the network demonstrated comprehensive and high-precision recognition capabilities across four scenarios in the MVTec-AD industrial dataset. By assisting companies in implementing rigorous quality inspections, our approach enhances product quality and economic benefits, promoting the development of sustainable energy systems. By leveraging defect information, photovoltaic manufacturers can optimize production processes, reduce resource waste, and transition to more sustainable manufacturing.

Author Contributions: Conceptualization, G.W.; methodology, G.W. and P.Z.; software, P.Z. and H.F.; validation, P.Z.; formal analysis, G.W. and P.Z.; investigation, P.Z.; resources, G.W.; data curation, P.Z.; writing—original draft preparation, P.Z.; writing—review and editing, G.W. and P.Z.; visualization, P.Z.; supervision, G.W.; project administration, G.W.; funding acquisition, G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China (NSFC; grant numbers: 62103092 and U20A20189), in part by the Major Program of National NSFC (grant numbers: 61991404 and 61991401), and in part by the Research Program of the Liaoning Liaohe Laboratory (No.: LLL23ZZ-05-01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: The authors appreciate the constructive criticism and suggestions from the editor(s) and all anonymous reviewers.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

mIoU	Mean Intersection over Union
IoU	Intersection over Union
mAcc	Mean Accuracy

References

1. Sun, C.; Zou, Y.; Qin, C.; Zhang, B.; Wu, X. Temperature effect of photovoltaic cells: A review. *Adv. Compos. Hybrid Mater.* **2022**, *5*, 2675–2699. [CrossRef]
2. Garcìa, M.C.; Herrmann, W.; Böhmer, W.; Proisy, B. Thermal and electrical effects caused by outdoor hot-spot testing in associations of photovoltaic cells. *Prog. Photovolt. Res. Appl.* **2003**, *11*, 293–307. [CrossRef]
3. Zhou, P.; Wang, R.; Wang, C.; Chen, H.; Liu, K. SIIF: Semantic information interactive fusion network for photovoltaic defect segmentation. *Appl. Energy* **2024**, *371*, 123643. [CrossRef]
4. Shaik, A.; Balasundaram, A.; Kakarla, L.S.; Murugan, N. Deep Learning-Based Detection and Segmentation of Damage in Solar Panels. *Automation* **2024**, *5*, 128–150. [CrossRef]
5. Hou, D.; Ma, J.; Huang, S.; Zhang, J.; Zhu, X.T. Classification of defective photovoltaic modules in ImageNet-trained networks using transfer learning. In Proceedings of the 2021 IEEE 12th Energy Conversion Congress & Exposition-Asia (ECCE-Asia), Singapore, 24–27 May 2021; pp. 2127–2132.
6. Zhang, J.; Shen, Y.; Jiang, J.; Fang, S.; Chen, L.; Yan, T.; Li, Z.; Zhang, K.; Wei, H.; Guo, W. Automatic detection of defective solar cells in electroluminescence images via global similarity and concatenated saliency guided network. *IEEE Trans. Ind. Inform.* **2022**, *19*, 7335–7345. [CrossRef]
7. Chen, X.; Karin, T.; Libby, C.; Deceglie, M.; Hacke, P.; Silverman, T.J.; Jain, A. Automatic crack segmentation and feature extraction in electroluminescence images of solar modules. *IEEE J. Photovolt.* **2023**, *13*, 334–342. [CrossRef]
8. Fioresi, J.; Colvin, D.J.; Frota, R.; Gupta, R.; Li, M.; Seigneur, H.P.; Vyas, S.; Oliveira, S.; Shah, M.; Davis, K.O. Automated defect detection and localization in photovoltaic cells using semantic segmentation of electroluminescence images. *IEEE J. Photovolt.* **2021**, *12*, 53–61. [CrossRef]
9. Wang, C.; Chen, H.; Zhao, S.; Wang, Y.; Cao, Z. A Low-Cost Defect Segmentation System Based On IoT for Large-Scale Photovoltaic Manufacturing. *IEEE Internet Things J.* **2024**, *11*, 16928–16940. [CrossRef]
10. Kaligambe, A.; Fujita, G. A Deep Learning-Based Framework for Automatic Detection of Defective Solar Photovoltaic Cells in Electroluminescence Images Using Transfer Learning. In Proceedings of the 2023 4th International Conference on High Voltage Engineering and Power Systems (ICHVEPS), Bali, Indonesia, 6–10 August 2023; pp. 81–85.
11. Jha, A.; Rawat, Y.; Vyas, S. PV-S3: Advancing Automatic Photovoltaic Defect Detection using Semi-Supervised Semantic Segmentation of Electroluminescence Images. *arXiv* **2024**, arXiv:2404.13693.
12. Xie, X.; Lai, G.; You, M.; Liang, J.; Leng, B. Effective transfer learning of defect detection for photovoltaic module cells in electroluminescence images. *Sol. Energy* **2023**, *250*, 312–323. [CrossRef]
13. Jiang, Y.; Zhao, C. Attention classification-and-segmentation network for micro-crack anomaly detection of photovoltaic module cells. *Sol. Energy* **2022**, *238*, 291–304. [CrossRef]
14. Otamendi, U.; Martinez, I.; Quartulli, M.; Olaizola, I.G.; Viles, E.; Cambarau, W. Segmentation of cell-level anomalies in electroluminescence images of photovoltaic modules. *Sol. Energy* **2021**, *220*, 914–926. [CrossRef]
15. Tang, W.; Yang, Q.; Xiong, K.; Yan, W. Deep learning based automatic defect identification of photovoltaic module using electroluminescence images. *Sol. Energy* **2020**, *201*, 453–460. [CrossRef]
16. Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv: 2010.11929.
17. Liu, J.J.; Hou, Q.; Cheng, M.M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3917–3926.
18. Bergmann, P.; Fauser, M. A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9592–9600.
19. Pratt, L.; Mattheus, J.; Klein, R. A benchmark dataset for defect detection and classification in electroluminescence images of PV modules using semantic segmentation. *Syst. Soft Comput.* **2023**, *5*, 200048. [CrossRef]
20. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10347–10357.
21. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
22. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10 –21 October 2021; pp. 10012–10022.

24. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood attention transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6185–6194.
25. Xu, J.; Sun, X.; Zhang, Z.; Zhao, G. Layer normalization. *arXiv* **2016**, arXiv: 1607.06450.
26. Loshchilov, I. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
27. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
28. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
29. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.
30. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
31. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
32. Zhang, W.; Pang, J.; Chen, K.; Loy, C.C. K-net: Towards unified image segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10326–10338.
33. Yin, M.; Yao, Z.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; Hu, H. Disentangled non-local neural networks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 191–207.
34. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
35. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 593–602.
36. He, J.; Deng, Z.; Qiao, Y. Dynamic multi-scale filters for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3562–3572.
37. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19529–19539.
38. Huang, L.; Yuan, Y.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Interlaced sparse self-attention for semantic segmentation. *arXiv* **2021**, arXiv:1907.12273.
39. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.