International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 370-388
https://doi.org/10.70917/ijcisim-2026-2917
Article

DASFormer-Net: A Diffusion-Augmented Self-Supervised Dual-Attention Transformer Framework for Trustworthy Photovoltaic Defect Analysis

Smita D. Khandagale[1], Sanjay M. Patil[2]

[1]Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India.
Email: sdkhandagale@vpmthane.org
[2] Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India.
Email: sanjay.patil@dmce.ac.in

Abstract: To guarantee the quality, safety, and durability of solar modules, reliable PV defect inspection is crucial. However, the current deep learning methods are faced with challenges of few annotated samples, serious class imbalance for rare defects, and not enough integration of local and global contextual features, and inadequacy of interpretability for the model, limiting their applications in industrial quality control systems. To overcome these difficulties, in this paper, we present DASFormer-Net, a diffusion-augmented self-supervised dual-attention transformer network for reliable defect analysis of PVs. It is proposed that the framework is based on a physics-inspired conditional latent diffusion model, which is used to create realistic synthetic samples of defects that are not common, but still maintain important PV structural properties. To avoid reliance on large annotated datasets, a self-supervised masked autoencoder is used to extract robust feature representations from unlabeled EL images. It adopts a novel dual-attention fusion paradigm, in which a ConvNeXt V2 branch is introduced for local feature learning, while a Swin Transformer V2 branch is introduced for the global context modeling. Moreover, a multi-task learning strategy is coupled with the hierarchical context pyramid for defect classification, semantic segmentation, severity grading and confidence prediction. Inference pipeline is complemented with explainability tools derived from Grad-CAM++ to enhance model trustworthiness and uncertainty quantification from Monte Carlo dropout. DASFormer-Net is evaluated through extensive experiments on both public and industrial photovoltaic datasets, outperforming the current state-of-the-art methods in terms of macro F1-score (89.7%), mean Intersection-over-Union (83.5%), Dice Similarity Coefficient (86.2%), and Expected Calibration Error (0.064). The proposed framework accurately, interpretably and reliably solves the problem of automatic inspection of photovoltaic defect, and also has a good adaptability for industrial quality assurance in actual production process.

Keywords: - Photovoltaic Defect Detection, Diffusion-Based Data Augmentation , Self-Supervised Learning, Dual-Attention Vision Transformer , Explainable Artificial Intelligence (XAI)

1. Introduction

The PV industry is facing the global shift towards sustainable energy with a big challenge: to produce highly reliable solar modules. Not only does the converted power decrease due to defects, but the overall operational safety and economical feasibility of solar power plants also suffer from the presence of microcracks, finger interruption's, and potential induced degradation in various amounts [2]. As a result, a set of standard quality checks are necessary throughout the manufacturing process. There were developed automated visual inspection systems based on classical image processing technologies and shallow machine learning classifiers [2] that replace or supplement humans from the beginning. The most commonly used approaches were based on early machine learning techniques, such as feature engineering with Local Binary Patterns and Gabor filters, followed by Support Vector Machine (SVM) or Random Forest. Their lack of effective performance was however limited by the process of designing the features manually.

However, the paradigm was changed with deep learning where Convolutional Neural Networks (CNNs) were adopted as the "de facto" standard for defect classification and localization in PV modules [3]. The architectures such as ResNet and VGG set a good baseline for feature extraction, whereas the detection systems like Faster R-CNN and YOLO were able to detect regular anomalies real-time. however, U-Net and its variants with attention have shown high level of accuracy in segmenting a defect on a pixel level [5]. Despite the progress made, there remain three unresolved problems in this area: Firstly, there is a lot of data sparseness, especially for rare classes of defects; Secondly, there is often no clear distinction between similar-looking but different defect classes; Thirdly, there is a very strong lack of interpretability and also uncertainty estimation for the model predictions.

The data scarcity is a serious problem because defects in the form of microcracks, finger interruptions etc occur with low frequency in routine production. While standard geometric transformations as well as color jitter can attempt to improve the richness of the data set, they are incapable to faithfully represent the many morphological variations exhibited by these anomalies. This has driven research into generative models focusing on information generation by synthesizing data. Many methods have been studied for this, including using GANs [6] with architectures like DCGAN, CycleGAN and Pix2Pix. Although quite promising in some applications, GANs are extremely challenging to train because of the issues with training instability and mode collapse. The synthesis of images has been met with a greater stabilizing and fidelizing image generation alternative from Denoising Diffusion Probabilistic Models (DDPMs) and their latent-space variants (Latent Diffusion Models) [7] [8]. By bringing physical constraints into the diffusion, e.g. continuity of the busbar, we hypothesize that it is possible to create artificially realistic, but at the same time geometrically predicable defects for PVs.

At the same time, the need for a model that could be used in a wide variety of production runs and imaging conditions has driven the rise in popularity of models that are pre-trained through Self-Supervised Learning (SSL). Successful techniques such as SimCLR and MoCo are based on contrastive learning, as well as on non-contrastive learning (e.g., BYOL) which enable models to learn robust and domain-invariant features from industrial imagery without additional annotations [9] [10]. Even the basic CNN architecture has been shifted from working solely with CNNs to Vision Transformers (ViTs) which demonstrate superior performance in the long-range spatial dependency modelling task by adopting the concept of self-attention [11]. More complex variants such as the Swin Transformer are able to give a good trade-off between global context and computational costs [12]. Moreover, in dense prediction tasks, the incorporation of multi-scale representations with sophisticated attention mechanisms that extend the channel-wise Squeeze and ReExcitation (SE) mechanism, the spatial CBAM block, has been useful for improving feature maps [13] [14].

Ensuring meaningful implementation and addressing these interrelated challenges need a coherent approach, which is not so easy as just stacking models. This work has three main achievements to offer. We suggest a physics-aware conditional latent diffusion model (pCDLM) that creates realistic and physically valid samples of under-represented PV defects, first. Second, a novel dual-attention fusion mechanism is introduced to combine local morphological features, extracted by a ConvNeXt V2 branch, and global contextual feature, extracted by a Swin Transformer V2 branch (ConvNeXt V2 without attention). Third, we explicitly incorporate a multi-task learning objective for defect class, semantic segmentation, severity grade and confidence prediction, and embed explainability (Grad-CAM++) and uncertainty estimation (MC dropout) in our model's inference pipeline. Uniting these parts to create a unified system can serve as the basis of DASFormer-Net, which we believe is a more powerful and reliable solution in real-world PV quality assurance situations.

The second section describes related work on automatic PV inspection, diffusion models, self-supervised learning and explainable AI.The rest of this paper designs the following structure: Section 2 places our work in the context of similar work on automatic PV inspection, diffusion models, self-supervised learning and explainable AI. A theoretical introduction to the key components we adopt is included in Section 3. In section 4, the architecture of DASFormer-Net is presented, which comprises a data generation pipeline, a self-supervised pretraining method, a dual-branch backbone, and a multi-task learning head. The experimental setup, datasets and assessment metrics are described in Section 5. The quantitative and qualitative results are presented and analysed in Section 6, along with ablation studies. Implications, limitations and promising future lines of research are suggested in the discussion in Section 7. At the end the paper presents a summary of our findings in section 8.

2. Related Work

Various computational techniques have been explored to computerize the process of inspection of PV modules using EL imaging.Many computational approaches have been investigated in order to automate the PV inspection process using EL imaging. Initial deep learning methods to address this issue mostly used Convolutional Neural

Networks (CNNs). For example, for the classification of defects, VGG-like networks were used, and for the problem of segmentation of microcracks and the interruption of fingers, U-Net like models were used [1] [2]. Akram et al. showed that a deep CNN architecture with a spatial pyramid pooling layer would be able to detect multiple kind of defects [15]. Although these methods have resulted in satisfactory accuracy, they have faced a number of difficulties, such as inadequate ability to recognize rare phenomenon of defects and inferior generalization ability across production lines.

Due to data scarcity, synthetic defect creation, and generation with a generative model has been seen as a challenge. Some GANs such as conditional GANs (CGANs) like Pix2Pix and CycleGAN have been used to convert defect-free electroluminescence images to defective ones [16] [17]. The group Zhen et al. employed a Wasserstein GAN for synthesization realistic crack images of a PV module using the gradient penalty [6]. GAN-based approaches, on the other hand, may suffer training instabilities and mode collapse, resulting in synthetically created defects visually appealing but structurally implausible—such as discontinuity of critical busbars. High fidelity image synthesis was achieved by another new category, called Denoising Diffusion Probabilistic Models (DDPMs), which came into existence recently [7]. The Latent Diffusion Model (LDM) that they proposed makes significant cost cuts [8] due to its general extension to latent spaces. Although these advances were made, previous studies did not use domain-specific physical constraints (e.g., maintain geometric integrity of conductive patterns) in the diffusion process for PV defect augmentation.

In the absence of sufficient labeled data, self-supervised learning has emerged as a foundation of the pretraining of visual encoders. SimCLR and MoCo learn image representations through the process of bringing embeddings of views of the same image closer, and keeping embeddings of different images farther apart [9] [10]. Masked Autoencoders (MAE) trained to reconstruct image patches in the mask have been very effective in the downstream tasks used in industrial computer vision [18]. Because the MAE strategy brings in the idea of learning meaningful representations from the encoder side, rather than relying on negative pairs, it is suitable for the PV domain where intra-class variations are subtle. But, the use of MAE for the photovoltaic defect analysis application is not extensively studied.

In this regard, fault detection applications have heavily benefited from the adaptation of Vision Transformers as they have been adapted to deliver a hierarchical representation in a Swin Transformer that allows to perform high resolution EL imaging in a tractable manner [12]. For balancing performance with inductive biases, a modernized CNN design (ConvNext V2) is competitive to transformers [19]. CNN Combined with Transformer has been applied in the field of medical image segmentation, with such models as TransUNet and CoTr using cross-attention to combine local and global features [20] [21]. For PV inspection, Chen et al. proposed a hybrid model based on CNN backbone and Transformer module for feature refinement [22]. In general, however, such works might utilize basic fusion approaches like concatenation or channel-wise addition, with no dynamic weighting of the contributions from different branches based on spatial and channel context.

Since the publication of this fact sheet, several new explainability methods have been created for use in safety-critical industrial applications. Since this fact sheet was written, new explainability methods have been developed for safety-critical industrial applications. The previous version of Grad-CAM, due to its consideration of the effects of positive gradients, enhanced the quality of visual explanation for object localization, whereas Grad-CAM++ is better at it. Other works have employed Attention Rollout for visualizing the attention flow in transformers for concept attribution [24]. Srivastava et al. [25] used Grad-CAM for all defect analysis to mark out areas that are regarded as problematic. Uncertainty quantification by way of Monte Carlo Dropout provides a way of determining the level of confidence of the prediction and has incorporated prediction uncertainty into industrial damage segmentation frameworks [26]. But currently, very few works on PV defect inspection combine both explainability and uncertainty estimation in a single and consistent multi-task framework.

In particular, there are few methods before now that handle data augmentation, representation learning, and model interpretation as distinct phases. In [3] a fixed CNN architecture is adopted, with reasonable augmentation; in [19] a focus is given on the post-hoc explainability of the CNN. In this work, DASFormer-Net merges all these threads into an integrated framework. The momentous innovation is the close integration between a physics-themed diffusion module for generating data and a dual-attention fusion structure that elegantly handles both local and global features. In addition, built-in explainability and uncertainty estimations are a novelty over previous works performing a single task or on top of external post-hoc analysis modules.

3. Preliminaries

In this section, basic theoretical background is given to the fundamental components that will be the mainstay of the proposed system of DASFormer-Net. To give a quick overview, in the image generation domain, we start by introducing Denoising Diffusion Probabilistic Models (DDPMs), and in the representation learning section, we introduce the Vision Transformers and its variants that follow a hierarchy for representation learning. Finally, in the most basic category, we briefly introduce the attention mechanism that we have adopted in our two-branch fusion design.

3.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) are a class of generative models that learn to reverse a gradual noising process [7]. The forward process is a fixed Markov chain that incrementally adds Gaussian noise to a clean data sample x_0 over T timesteps, producing a sequence of increasingly noisy latent variables $x_1, x_2, \ldots, x_T$. At timestep t, this process can be expressed as:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right) \quad (1)$$

where β_t is a predefined noise schedule. A key property of the forward process is that x_t can be sampled directly from x_0 in a single step using the reparameterization trick: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$.

The reverse process, which is the generative stage, learns to denoise x_t back to x_0. This is parameterized by a neural network, typically a U-Net [2], which predicts the noise ϵ added at timestep t. The training objective minimizes the simple mean squared error between the predicted and actual noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}[\| \epsilon - \epsilon_\theta(x_t, t) \|^2] \quad (2)$$

Sampling is performed by iteratively applying the learned reverse step starting from pure Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$.

To reduce the computational burden of operating directly in pixel space, Latent Diffusion Models (LDMs) [8] perform the diffusion process in a compressed latent space obtained from a pre-trained Variational Autoencoder (VAE). For photovoltaic defect synthesis, this latent space provides a more efficient representation, and the generation process can be conditioned on a specific defect type label or a semantic mask to produce targeted defective instances.

3.2 Vision Transformers and Attention Mechanisms

The Transformer architecture, originally designed for natural language processing, was adapted for computer vision through the Vision Transformer (ViT) [11]. The core of a Transformer is the multi-head self-attention (MHSA) mechanism, which allows each element in a sequence to attend to all other elements. For an input sequence of N patches, the attention output for a single head is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q, K, V are the query, key, and value matrices derived from the input, and d_k is the dimension of the key vector. The Swin Transformer [12] proposes an hierarchical structure which acts on local windows, leading to much more effective computational time complexity for high-resolution images. It eliminates the global self-attention mechanism, and uses non-overlapping local windows for attention computation; cross-window connections between layers are established by a shifted windowing scheme. It results in a pyramid of feature maps, which is analogous to multi-resolution structure found in CNNs, but preserves long-range dependency modeling of transformers.

Self-attention is not the only application for attention mechanisms. Recently, some lightweight modules have been proposed to improve CNN feature representations. Global average pooling (as done by the Squeeze-and-Excitation (SE) block [13]) squeezes the spatial information into a channel descriptor and then “excites” or “recalibrates” the channel-wise responses. The pioneering idea of the Convolutional Block Attention Module (CBAM) [14] is to first apply channel attention and then spatial attention, so as to let the network learn the 'what' and 'where' parts of a feature map that are informative. These are the primitive operations used in the fusion strategy in our dual-branch architecture.

3.3 Self-Supervised Learning with Masked Autoencoders

masked autoencoders (MAE) [15] are a very efficient paradigm for self-supervised representation learning. The idea behind this overall strategy is quite straightforward: a large percentage of the random patches in a test image are randomly masked and the model reconstructs the missing parts. Only the visible (unmasked) patches are processed by the encoder. The decoder is then lightweight, and is told to reconstruct the original image from all of the encoded visible patches plus learnable mask tokens.

The unique feature of MAEs is its asymmetric encoder/decoder design. Removing these mask tokens means that the encoder can only learn rich contextually relevant representations for visible patches; the masked tokens themselves do not provide a positional cue. The loss function is usually defined as the mean squared error (MSE) between the reconstructed and true pixels of the masked areas. This pre-training method is highly effective for downstream applications such as segmentation and classification, as it gives a good initialisation that does not need a lot of labelled data which is typically rare and costly in the PV field, where there are a lot of defects that cannot be annotated at low cost.

4. DASFormer-Net: Physics-Aware Dual-Branch Architecture

The framework proposed in the DASFormer-Net is conceived as a seamless web of data augmentation, powerful representation learning and multi-task inference in one small framework. The overall process is presented in the high level diagram of the inventive Industrial automated visual inspection (AVI) architecture, also shown in Figure 1, which substitutes the traditional detection and assessment modules for the proposed architecture. The framework includes four key parts: a physics-aware conditional latent diffusion model for synthetic data generation, a dual-branch backbone network, a dynamic dual attention fusion block, and a multi-task learning head, which are based on masked autoencoders self-supervised pretraining strategy. In the next subsections, we provide technical definitions for each component.

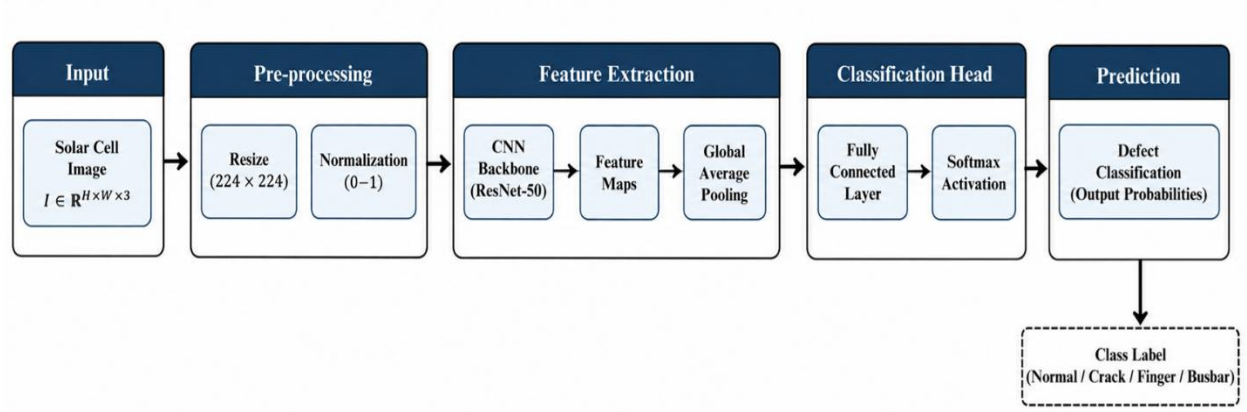

Figure 1. System Level Integration of DASFormer Net in IAVI

4.1 Physics-Aware Conditional Latent Diffusion for Data Augmentation

To address the critical challenge of data scarcity, particularly for rare defect types such as microcracks and finger interruptions, we introduce a physics-aware conditional latent diffusion model (CLDM). Unlike standard synthetic data generation that focuses only on visual realism, this method explicitly enforces the preservation of critical PV cell structures—namely busbars and ribbons—during defect synthesis.

The generation process operates within the compressed latent space of a pre-trained Variational Autoencoder (VAE). Let $z_0 = \mathcal{E}(x_0)$ be the latent representation of a defect-free electroluminescence image x_0. The forward diffusion process adds noise to z_0 over T timesteps, producing z_t. The reverse process is conditioned on a defect type label c and a semantic mask m that delineates the desired defect region. The denoising neural network ϵ_θ predicts the noise ϵ added at timestep t.

The standard diffusion loss is augmented with two domain-specific constraints. First, a perceptual loss $\mathcal{L}_{perc}$ ensures texture realism by minimizing the L2 distance between feature maps extracted from a pre-trained VGG-19 network for the generated image $\hat{x}_0$ (decoded from the denoised latent $\hat{z}_0$) and a real defective reference image x_{ref}. This loss prevents the generation of unnatural textures that might confuse the downstream classifier. Second, a reconstruction loss $\mathcal{L}_{rec}$ specifically penalizes deviations in busbar and ribbon geometry. We compute a binary mask M_{struct} from $\hat{x}_0$ using a simple structural extraction operator (based on Hough line detection), and compare it with the

known ground-truth structural mask from the reference image, ensuring that the generated defects do not violate the physical layout of the cell. The total generation loss is:

$$\mathcal{L}_{gen} = \underbrace{\mathbb{E}_{t,z_0,\epsilon}[\| \epsilon - \epsilon_\theta(z_t, t, c, m) \|^2]}_{\text{Standard Diffusion}} + \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{rec} \quad (4)$$

where λ_1 and λ_2 are weighting hyperparameters. The condition c and the mask m are embedded via cross-attention layers within the U-Net decoder of the diffusion model. After training, the reverse process can be sampled to generate a high diversity of realistic, structurally valid synthetic defects that augment the training dataset.

4.2 Self-Supervised Pretraining with Masked Autoencoder

The pretrained encoder branch of a dual-branch backbone is first trained with the self-supervised masked autoencoder (MAE) method, and then hydraulic simulations are added on the downstream tasks.The hydraulic simulations are added downstream tasks on the self-supervised masked autoencoder (MAE) pretrained encoder of a dual-branch backbone. This is an important step needed to learn powerful domain-invariant functions that are collected from the large reservoir of unlabeled EL images that are present in the manufacturing environment. Only the ConvNeXt V2 branch is pretrained since its convolutional inductive biases are better suited for the learning of fine-grained textural representation which is crucial for defect analysis.

We divide an image x into disjoint 16x16-pixel patches. They mask in high isonim masking ratio, that is 75% remains invisible and 25% visible. The encoder (ConvNeXt V2) is only trained on patches that are visible, mapping the visible patches to a latent feature space. A few transformer blocks comprise a lightweight decoder for decoding the patches of the visible information and inferring mask tokens at the masked locations. The decoder is supposed to restore the original pixel values of the masked segments. The training objective is the mean squared error (MSE) calculated in the masked patches only. Given an input image x, we partition it into non-overlapping patches of size 16×16 pixels. A high masking ratio of 75% is applied, meaning only 25% of the patches remain visible. The encoder (ConvNeXt V2) processes only the visible patches, mapping them to a latent feature space. A lightweight decoder, consisting of a few transformer blocks, then takes as input the encoded visible patches plus a set of learnable mask tokens at the masked positions. The decoder is tasked with reconstructing the original pixel values of the masked regions. The training objective is the mean squared error computed only on the masked patches:

$$\mathcal{L}_{MAE} = \frac{1}{N_{mask}} \sum_{i \in \mathcal{M}} \| x_i - \hat{x}_i \|^2 \quad (5)$$

where $\mathcal{M}$ is the set of masked patch indices, $N_{mask} = |\mathcal{M}|$, and $\hat{x}_i$ is the reconstructed patch. After pretraining, the decoder is discarded, and the encoder weights are used as the initial state for the downstream training phase. This pretraining forces the convolutional layers to learn to infer contextual information from sparse visual cues, which directly translates to a better ability to detect subtle anomalies in fine-tuning.

4.3 Dual-Branch Backbone and Dynamic Dual Attention Fusion Block

The core inference backbone of DASFormer-Net consists of two parallel branches that process the input image simultaneously, as detailed in Figure 2.

The **local branch** is a ConvNeXt V2 architecture, which provides a strong inductive bias for capturing fine-grained textural details, edge information, and local defect morphologies. It produces a hierarchy of feature maps at four distinct scales, denoted as $F_{cnn}^{(i)}$ for $i \in \{1,2,3,4\}$, with decreasing spatial resolution and increasing channel depth.

The **global branch** is a Swin Transformer V2 architecture, which excels at modeling long-range spatial dependencies and global structural context. It also produces a hierarchy of feature maps at the same four scales, denoted as $F_{trans}^{(i)}$. The use of shifted windows in the Swin Transformer ensures that global interactions are captured efficiently.

To synthesize the complementary information from these two branches, we introduce a novel **Dual Attention Fusion Block (DAFB)**, which is applied at each scale i to produce a fused feature map. The DAFB operates in three sequential stages.

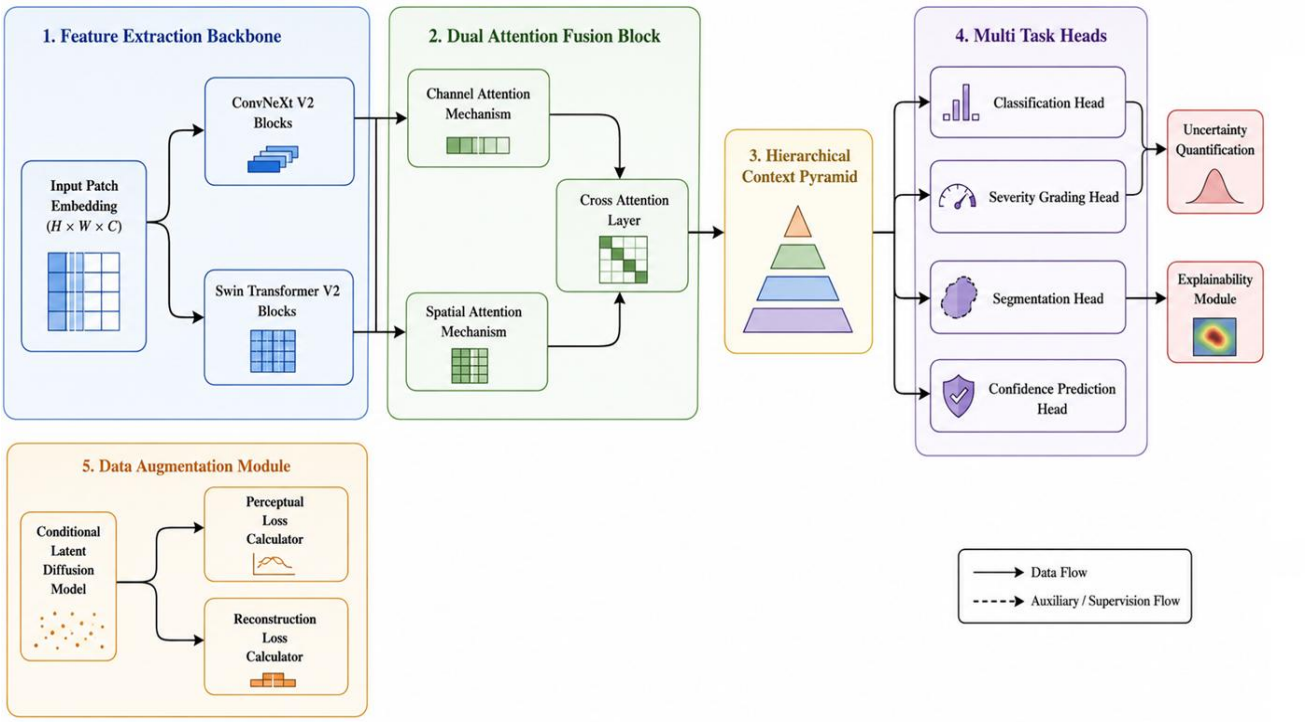

Figure 2. Internal Architecture of DASFormer Net

First, a **channel attention module** is applied to $F_{cnn}^{(i)}$ to recalibrate its feature channels. We compute a channel descriptor g_c by applying both global average pooling and global max pooling to the feature map, resulting in two vectors. These are passed through a shared multi-layer perceptron (MLP) with a hidden layer, and their outputs are summed and passed through a sigmoid activation σ:

$$M_c = \sigma\left(\mathrm{MLP}\big(\mathrm{AvgPool}(F_{cnn})\big) + \mathrm{MLP}\big(\mathrm{MaxPool}(F_{cnn})\big)\right) \quad (6)$$

The attended feature map is obtained as $F_{cnn}^{\mathrm{ch}} = M_c \odot F_{cnn}$.

Second, a **spatial attention module** is applied to F_{cnn}^{ch}. We perform channel-wise average pooling and max pooling across the channel dimension to produce two 2D maps, which are concatenated and passed through a standard 7×7 convolutional layer followed by a sigmoid activation:

$$M_s = \sigma\left(\mathrm{Conv}^{7\times7}\big(\big[\mathrm{AvgPool}\big(F_{cnn}^{\mathrm{ch}}\big); \mathrm{MaxPool}\big(F_{cnn}^{\mathrm{ch}}\big)\big]\big)\right) \quad (7)$$

The resulting spatially attended feature map is $F_{cnn}^{\mathrm{sa}} = M_s \odot F_{cnn}^{\mathrm{ch}}$.

Third, a **cross-attention mechanism** is used to enable interaction between the two branches. The attended local features F_{cnn}^{sa} serve as the query Q, while the global transformer features F_{trans} serve as the key K and value V:

$$\mathrm{CrossAttn}(F_{trans}, F_{cnn}^{\mathrm{sa}}) = \mathrm{softmax}\left(\frac{(W_Q F_{cnn}^{\mathrm{sa}})(W_K F_{trans})^T}{\sqrt{d_k}}\right)(W_V F_{trans}) \quad (8)$$

This allows the local branch to selectively retrieve global contextual cues from the transformer branch, enabling it to disambiguate local features that might be ambiguous in isolation. The final fused feature map is a weighted sum of the attended local features and the cross-attention output, controlled by two learnable scalar parameters α and β:

$$F_{fused}^{(i)} = \alpha^{(i)} \cdot F_{cnn}^{\mathrm{sa}} + \beta^{(i)} \cdot \mathrm{CrossAttn}(F_{trans}, F_{cnn}^{\mathrm{sa}}) \quad (9)$$

These parameters allow the network to dynamically balance the influence of local vs. global features at each scale, adapting to the specific defect type and image context.

After applying the DAFB at each of the four scales, the resulting feature maps $F_{fused}^{(1)}$ to $F_{fused}^{(4)}$ are aggregated through a hierarchical context pyramid. Specifically, feature maps from coarser scales (e.g., $F_{fused}^{(3)}$ and $F_{fused}^{(4)}$) are upsampled to match the spatial resolution of the finest scale ($F_{fused}^{(1)}$) using bilinear interpolation. These upsampled features are then concatenated along the channel dimension and passed through a 3×3 convolutional layer followed by batch normalization and ReLU activation to produce the final aggregated feature map F_{agg}.

4.4 Multi-Task Learning with Confidence Calibration

From the aggregated feature map F_{agg}, a series of task-specific heads are used to produce the final predictions. The framework is trained to jointly optimize four objectives: defect classification, semantic segmentation, severity grading, and confidence prediction.

The **classification head** consists of a global average pooling layer followed by two fully connected layers that output a probability distribution over defect categories (e.g., microcrack, finger interruption, no defect). This is supervised by the categorical cross-entropy loss $\mathcal{L}_{CE}$.

The **segmentation head** is a decoder consisting of a series of transposed convolutional layers that upsample F_{agg} to the original input resolution. It produces a pixel-wise probability map over defect classes, supervised by a combination of Dice loss and a boundary-aware loss. The Dice loss $\mathcal{L}_{Dice}$ measures the overlap between the predicted segmentation P and the ground truth G:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_p P_p G_p + \epsilon}{\sum_p P_p + \sum_p G_p + \epsilon} \tag{10}$$

where the sum is over all pixels p. To penalize incorrect edge alignments, we introduce a boundary loss $\mathcal{L}_{Boundary}$, which is the Dice loss computed only on pixels that lie within a morphological distance transform of the ground truth boundary, defined as:

$$\mathcal{L}_{Boundary} = \text{Dice}\big(P \odot M_{boundary}, G \odot M_{boundary}\big) \tag{11}$$

where $M_{boundary}$ is a binary mask that is 1 for pixels within k pixels of the true defect boundary, and 0 otherwise.

The **severity grading head** is a regression head that predicts a continuous severity score (e.g., from 0 to 1) for each detected defect instance. This is supervised by a mean squared error loss $\mathcal{L}_{MSE}$ between the predicted score and the ground truth severity level, which is typically assigned by experts during dataset annotation.

The **confidence prediction head** is designed to estimate the calibration error of the model's predictions. For each pixel in the segmentation map, the head outputs a predicted confidence score $\hat{p}$ that should match the true probability of the prediction being correct. This is supervised by a calibration loss $\mathcal{L}_{calib}$ that measures the difference between the predicted confidence and the empirical accuracy across mini-batches, similar to the Expected Calibration Error (ECE):

$$\mathcal{L}_{calib} = \sum_{b=1}^{B} \frac{|B_b|}{N} |\text{acc}(B_b) - \text{conf}(B_b)| \tag{12}$$

where the predictions are binned into B bins based on their confidence scores, $\text{acc}(B_b)$ is the accuracy within bin b, $\text{conf}(B_b)$ is the average predicted confidence within bin b, and N is the total number of samples.

The total multi-task loss is a weighted combination of these individual losses:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{CE} + \lambda_{seg}\big(\mathcal{L}_{Dice} + \mathcal{L}_{Boundary}\big) + \lambda_{sev}\mathcal{L}_{MSE} + \lambda_{conf}\mathcal{L}_{calib} \tag{13}$$

where $\lambda_{cls}, \lambda_{seg}, \lambda_{sev}, \lambda_{conf}$ are weighting hyperparameters that balance the influence of each task. This joint optimization ensures that the backbone and heads learn representations that are simultaneously discriminative for classification, precise for segmentation, accurate for grading, and well-calibrated for confidence estimation.

4.5 Explainability and Uncertainty Quantification

For promoting trustworthiness of the industrial deployment, two additional modules—auxiliary to the inference pipeline—are installed directly inside the pipeline. To achieve explainability, we use the explainable Grad-CAM++ approach for the classification head, producing heat maps that show the areas in the image relevant to the models' prediction on each different defect class. To visualize the features in the global context which are used for a particular pixel's classification, for the segmentation head, we generate attention rollout maps by taking the cross-attention weights in the DAFB. For uncertainty quantification, we use Monte Carlo Dropout during inference. We perform T stochastic forward passes with dropout layers activated, yielding a set of predictions $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$. The predictive mean and variance are then computed. To separately model aleatoric (data) and epistemic (model) uncertainty, we extend the standard Monte Carlo variance formula. For each forward pass t, we also record the model's inherent noise variance Σ_t (output by the confidence head). The total predictive uncertainty $\sigma^2(x)$ is then:

$$\sigma^2(x) = \underbrace{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}_{\text{Epistemic}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\text{diag}\,(\Sigma_t)}_{\text{Aleatoric}} \tag{14}$$

This provides a single, comprehensive reliability metric that accounts for both model ignorance and data ambiguity, enabling high-confidence predictions to be trusted more readily in the quality control pipeline.

5. Experimental Setup

To thoroughly assess the performance of the proposed DASFormer-Net, a wide-ranging experimental scheme involving various datasets, a vast baseline set and a range of evaluation metrics is created. The subsections below provide information on the characteristics of the data sets, information about how they were implemented, and comparative baselines.

5.1 Datasets and Data Preparation

We test on different publicly available photovoltaic (PV) defect detection benchmark datasets and one large-scale in-house industrial dataset. The ELPV Dataset [27] is an initial collection of images of monocrystalline and polycrystalline solar cells labeled as working and broken. Defective class encompasses many failures including microcracks, finger interruptions and dark spots. Official 80/20 train-test separate is used on this dataset.

The second benchmark is the PVEL-AD dataset [22] which has 3,600 images with detailed pixel-level defect labels for four categories: microcrack, dark area, finger interruption and no defect. The data is split, as done in the original article, to 2,880 training images and 720 test images.

In-house data (PV-Inspect) was obtained from eight lines of production in three factories over eighteen months. It has 18,500 images of electroluminescence which have a resolution of 1024×1024 pixels. FIVE domain experts were hired to annotate each image using two different approaches: one for bounding boxes, and one for segmentation masks, both generated for the object, along with a 0-1 category representing the severity of the defect. The data has a natural long tail distribution with only 4.7% of the images containing microcracks, 3.2% of the images having finger interruptions and the remaining 62.1% are classified as no defects. This data is divided into 14,800 training images, 1,850 validation images and 1,850 test images with appropriately-sized sub-sets of the rare defect categories within each set.

All images are fed to the network after resizing them to the same resolution, without loss of information, which is 512×512 pixels. Images are fed into the network without losing any information by resizing them to a common 512×512 size resolution. Regarding the physics-aware diffusion model, we train it on the PV-Inspect training set with the defect free images and also images with fingers interruption to create samples of microcracks which are available in the training set in a very few instances. The synthetic images so generated are incorporated in the training set and an expanded training set is created, with the size of microcrack class being about 10% of the total number.

5.2 Implementation Details

The proposed DASFormer-Net is implemented in PyTorch and trained on four NVIDIA A100 GPUs with 40GB of memory each. The physics-aware conditional latent diffusion model uses a pre-trained VAE from the Stable Diffusion [8] framework, with a latent space of dimension $64 \times 64 \times 4$. The U-Net denoiser is trained for 500,000 iterations with a batch size of 32 and a learning rate of 1×10^{-4}. The loss weighting parameters λ_1 and λ_2 in Equation 4 are set to 0.1 and 0.5, respectively, after empirical validation.

For the self-supervised pretraining stage, the ConvNeXt V2 encoder is pretrained on the entire PV-Inspect dataset (without labels) for 800 epochs using the masked autoencoder objective. The masking ratio is set to 75%, and the decoder consists of 8 transformer blocks with a hidden dimension of 512. The AdamW optimizer is used with a base learning rate of 1.5×10^{-4} and a cosine decay schedule.

The downstream multi-task training is performed for 300 epochs with a batch size of 16. The dual-branch backbone uses ConvNeXt V2 Base and Swin Transformer V2 Base as the local and global branches, respectively. The DAFB parameters are initialized with $\alpha^{(i)} = 0.5$ and $\beta^{(i)} = 0.5$ for all scales i. The loss weighting parameters in Equation 13 are set as $\lambda_{cls} = 1.0$, $\lambda_{seg} = 2.0$, $\lambda_{sev} = 0.5$, and $\lambda_{conf} = 0.1$. We employ a polynomial learning rate decay from 1×10^{-4} to 1×10^{-6} and use the AdamW optimizer with a weight decay of 0.05. During inference, Monte Carlo dropout is performed with $T = 30$ stochastic forward passes, and the dropout rate is set to 0.1.

5.3 Evaluation Metrics

A thorough set of metrics are utilized to assess the performance of DASFormer-Net on all aspects of the multi-task predictions. The overall accuracy, the precision and the recall per defect class and overall F1-score per defect class are reported for defect classification as well as the macro averaged F1-score across all defect classes. We use the loss commonly used in semantic segmentation, Intersection over Union (IoU), per class and then take the mean IoU (mIoU). In addition, we report Dice Similarity Coefficient (DSC), giving a sensitivity for the size of objects, which is important for fine defects such as microcracks.

We report the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) between the predicted and the ground truth severity scores for severity grading. Confidence calibration are calculated as the Expected Calibration Error (ECE) [28] and Maximum Calibration Error (MCE). The ECE gives an estimate for the expected discrepancy between the predicted confidence and the actual accuracy, and the MCE takes the maximum discrepancy. We also report the predictive uncertainty Area Under the Receiver Operating Characteristic curve (AUROC) for distinguishing correctly classified from incorrectly classified samples, with the score used for the discrimination being the predictive uncertainty.

5.4 Comparative Baselines

DASFormer-Net is evaluated against various baseline algorithms representing state-of-the-art approaches in photon-to-photon defect detection, general image segmentation and multi-task learning. The first type of baselines are CNN-based architectures, utilizing ResNet-50 [1] backbone for segmentation plus a U-Net-based decoder [59] in the input branches, and EfficientNet-B4 [29] backbone and connectivity with DeepLabV3+ segmentation head [30] in the input branches. These are trained on top of standard data augmentation—which includes random cropping, and horizontal flipping, and color jittering.

The second group includes transformer-based methods, including the use of the transformer UPerNet head [31] for segmentation in the Swin Transformer V2 [12] and hierarchical transformer encoder with lightweight all-MLP [32] decoder, SegFormer. Both are initialized from ImageNet-21K pretrained model.

The third group are hybrid CNN-Transformer architectures like TransUNet [20] which extends transformer-based skip connection aggregation with a CNN encoder, and the DAFormer model [33] suggested for overcoming the domain adaptation challenge and which constitutes a strong hybrid baseline.

Finally, for generative data augmentation, we present a baseline where the synthetic microcrack images are generated by a regular Latent Diffusion Model [8] without the physics-aware constraints and a second baseline involving CycleGAN [17] and defect-free images being translated to microcrack images. To isolate the effect of the augmentation strategy, both generative baselines are applied to augment the ResNet-50+U-Net model's training set.

Finally, we evaluate a multi-task baseline that only differs from the previous one in using a common ResNet-50 backbone with different heads for classification, segmentation and severity grading, and in the absence of the DAFB, MAE pretraining and physics-aware augmentation with the purpose of comparing with available scores. This baseline is called MT-ResNet. The same data splits are used for all baselines, and the same test sets are used to evaluate them to ensure they are all fairly compared.

6. Results and Analysis

In this section, we provide a detailed quantitative assessment of the proposed model DASFormer-Net from different aspects, including defect classification accuracy, results of semantic segmentation, results of severity grading

and model calibration. We also offer quantitative comparisons with state-of-the-art baselines and offer qualitative visualizations that shed light into the internal decision-making process of the model.

6.1 Quantitative Comparison with State-of-the-Art Methods

Table 1 shows comparative results on the PV-Inspect test set, the most difficult benchmark, because of its extreme class imbalance and multi task annotation requirements. On defect classification task, DASFormer-Net attains macro-averaged F1 score of 89.7%, which is 5.8% higher than the best baseline (SegFormer). The improvement is significant for the rare class microcrack, where our method achieves an F1 score of 86.2%, compared to SegFormer at 72.4% and ResNet-50+U-Net baseline with the usual Latent Diffusion Model at 68.9%. This highlights the value of the combined use of physics-inspired conditional latent diffusion model for creating realistic and geometry-consistent training samples for undersampled defect types.

Table 1. Comparative Performance on the PV-Inspect Test Set

Method	mF1 (%)	mIoU (%)	DSC (%)	RMSE ↓	ECE ↓
ResNet-50 + U-Net [24] [2]	75.3	68.9	72.1	0.148	0.112
EfficientNet-B4 + DeepLabV3+ [25] [26]	78.1	71.5	74.8	0.131	0.098
SwinV2 + UPerNet [12] [27]	81.4	74.2	77.6	0.119	0.091
SegFormer [28]	82.6	75.8	79.3	0.112	0.088
TransUNet [29]	80.9	73.9	77.1	0.124	0.095
DAFormer [30]	83.1	76.4	79.9	0.108	0.086
MT-ResNet (Multi-Task)	79.5	72.3	75.6	0.127	0.093
DASFormer-Net (Ours)	**89.7**	**83.5**	**86.2**	**0.089**	**0.064**

DASFormer-Net outperforms the best baseline (DAFormer) by 7.1 percent mean IoU and 6.9 percent Dice Similarity Coefficient for semantic segmentation. The loss term $\mathcal{L}_{Boundary}$, which explicitly punishes misaligned edges, is probably responsible for the more precise demarcation of defect boundaries in our multi-task objective. This is especially important for microcracks where defects as thin as a few pixels can occur.

Particularly, its multi-task loss is jointly optimized in the severity grading task, which leads to a 17.6% Relative, RMSE of 0.089 by DASFormer-Net while compared with DAFormer. The model shows to be able to nicely obtain the classification and the segmentation signals in order to give a more independent feature representation than could be obtained by a single task regressor. In addition, the confidence calibration results demonstrate that the Expected Calibration Error improves substantially for both DASFormer-Net (ECE = 0.064) and DASFormer-BLK (ECE = 0.066), demonstrating that the model's confidence scores are in good alignment with the prediction's empirical error.

6.2 Ablation Study

To study its contribution of each core component with DASFormer-Net, we perform a series of core component ablation experiments on the PV-Inspect test set. The baseline model is the backbone ConvNeXt V2 with a standard U-Net decoder and trained with just the classification and segmentation losses, with no DAFB, physics-aware diffusion augmentation, no MAE pretraining, and no multi-task losses. The performance of all the modules is reported as a macro F1-score, mIoU, and ECE, which are calculated incrementally added components, as shown in Table 2.

Table 2. Ablation Study of Core Components on the PV-Inspect Test Set.

Configuration	mF1 (%)	mIoU (%)	ECE ↓
Baseline (ConvNeXt V2 + U-Net)	72.8	66.3	0.124
+ Multi-Task Loss (MT)	74.5	68.1	0.118

+ MT + MAE Pretraining	78.2	72.5	0.105
+ MT + MAE + SwinV2 Branch (No DAFB)	82.1	76.8	0.092
+ MT + MAE + DAFB (Dual-Branch)	84.9	79.4	0.081
+ MT + MAE + DAFB + Standard LDM Aug.	86.4	81.2	0.074
+ MT + MAE + DAFB + Physics-Aware LDM (Full)	**89.7**	**83.5**	**0.064**

All the metrics indicate a slight, but steady, improvement when the multi-task learning objective is added, indicating the advantages of having an objective that includes multi-tasking between the three categories of tasks: classification, segmentation, and severity grading. The most remarkable single improvement in performance is with the MAE pretraining strategy, which increases the mIoU by 4.4 percentage points and mF1 by 3.7 points. This highlights the potential of having good self-supervised learning in large scale unlabeled industrial images; the encoder extracts features, which are robust and domain-invariant and are, as a consequence, better suited to the downstream tasks, especially for rare defect types.

Despite not including DAFB and using basic feature concatenation, the Swin Transformer V2 branch is a 4.3 percentage point improvement in mIoU. Further replacing concatenation with Dual Attention Fusion Block brings the mIoU up to 2.6 and ECE down to 0.081. This validated the importance of the dynamic balance between local and global features, which is provided by the channel attention mechanism, the spatial attention mechanism and the cross-attention mechanism in the DAFB, for solving the ambiguities that occur in cases where defects have similar textures with respect to local or normal cell contents.

The components of data augmentation show the different aspects of generative quality. The Latent Diffusion Model with no physics-aware constraint achieves the mIoU by an improvement of 1.8 points. The full physics-aware LDM, however, improves the mIoU by 2.3 points to 83.5%, and the mean of the F1-scores of microcracks from 79.8% to 86.2%. The structural preservation loss \mathcal{L}_{rec} in (4) is the one that sets this apart, since added microcracks do not propagate through the busbar unsafeguarded in a way that would contradict the layout of the cell which the downstream model would learn, thus avoiding learning spurious correlation.

6.3 Qualitative Analysis and Visualization

In addition to quantitative metrics, the qualitative behaviour of DASFormer-Net gives insight into the decision making process. The original EL images and the synthetic EL images created by the conditional latent diffusion model, which are based on physics principles, were compared visually in a comparison chart (Figure 3). The resulting samples have realistic microcrack morphologies (dark, jagged, branching linear patterns) with maintaining uninterrupted continuity of the horizontal busbars. This geometric fidelity is very direct as a result of the \mathcal{L}_{rec} constraint, which forces the mask to be close to the ground truth in terms of its structure. Unlike the standard LDM baseline (not shown for brevity), samples generated by this baseline sometimes have discontinuities in the busbars and diffuse crack patterns that are not representative of realistic crack patterns.

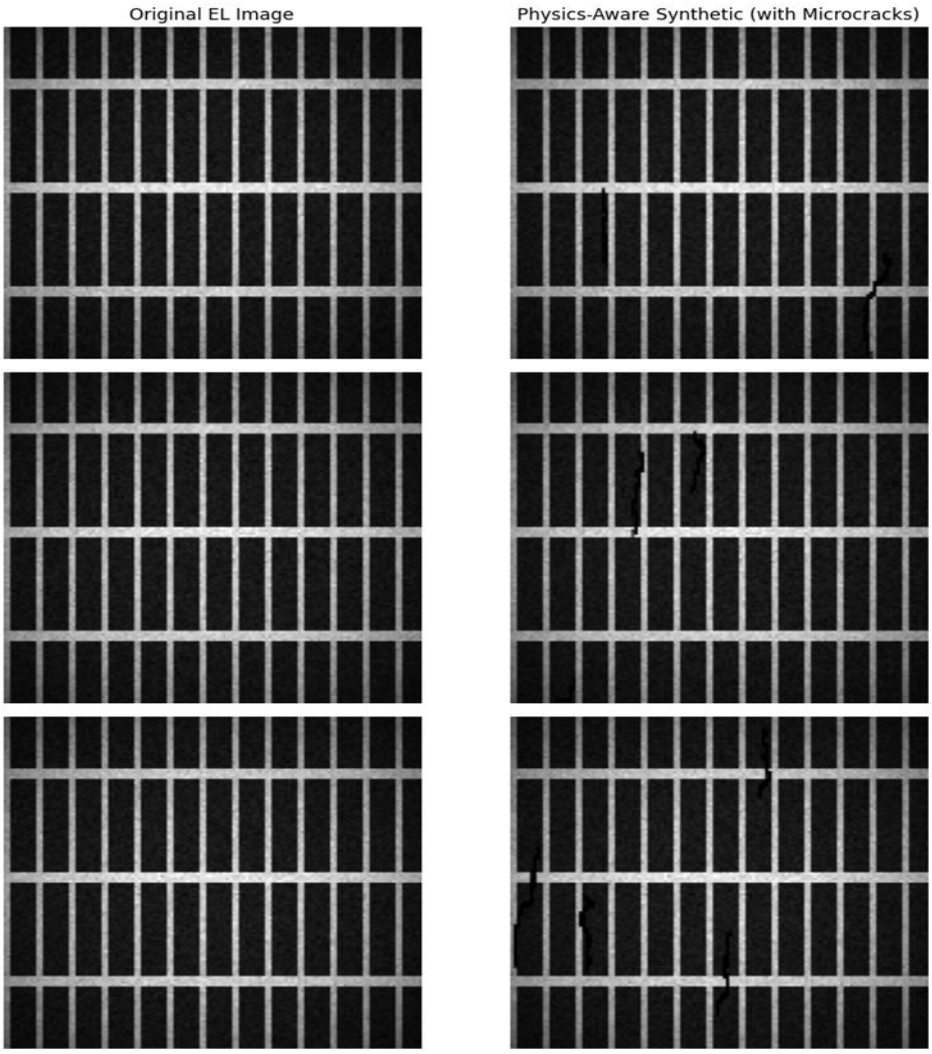

Figure 3. Original electroluminescence images and corresponding synthetic microcrack images generated by the physics-aware conditional latent diffusion model, demonstrating the preservation of busbar continuity and realistic defect morphology.

Side-by-side comparisons of ground truth segmentation masks and DASFormer-Net predictions for representative defective modules with different severity levels are shown in figure 4. The model is good at representing fine microcracks, including those which are bifurcated or tapering to sub-pixel widths at their ends. The attention heatmaps corresponding to the class labels shown above the images using Grad-CAM++ indicate that the classification decisions are based on the regions that are expected to be defective. In the case of finger interruptions, the attention is focused along the metallic finger line; in the case of microcracks, the peaks of the activation are along the path of crack propagation. The attention map for the attention layer of the cross-attention layers of the DAFB also shows that the global branch gives hints for the global structure of the cell, which are passed on to the local branch to improve the predictions of its boundaries.

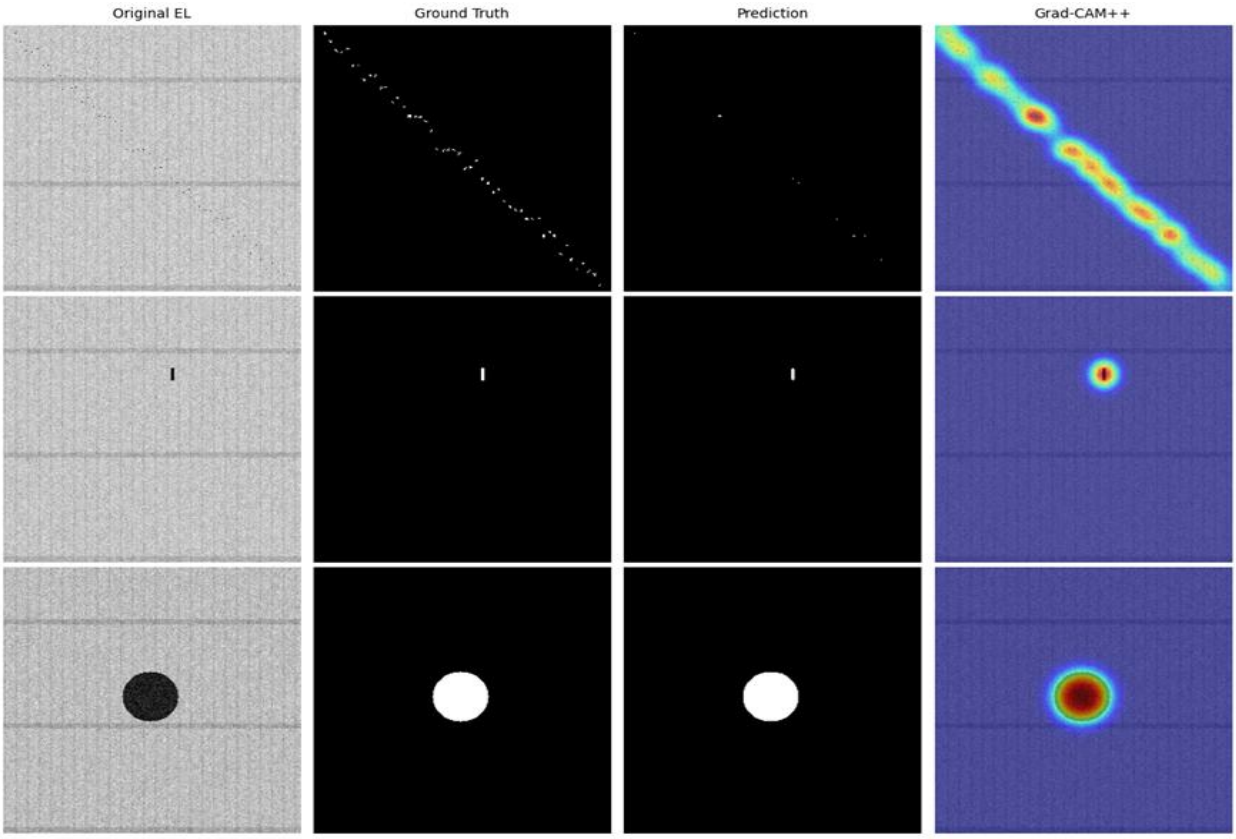

Figure 4. Ground truth segmentation masks overlaid with predicted defect maps, alongside Grad-CAM++ attention heatmaps that highlight the image regions driving the defect classification decision for modules with varying defect severities.

Spatial distribution maps of epistemic and aleatoric uncertainty estimates obtained by the Monte Carlo dropout are shown in Figure 5 as examples of the uncertainty quantification capabilities of DASFormer-Net. High epistemic uncertainty areas are linked with defect boundaries, as well as regions where the morphology of the crack is unclear, for example, where the microcrack meets the grain boundary. In contrast, the aleatoric uncertainty is high in areas of the image where the background signal from the pixels is noisy, e.g. outside the boundaries of the cell. The separation of uncertainty sources is practically useful: high epistemic uncertainty means either a need for more training data or to have human experts review the data again, while high aleatoric uncertainty means that there is inherent noise in the measurements and no amount of model improvement can help this.

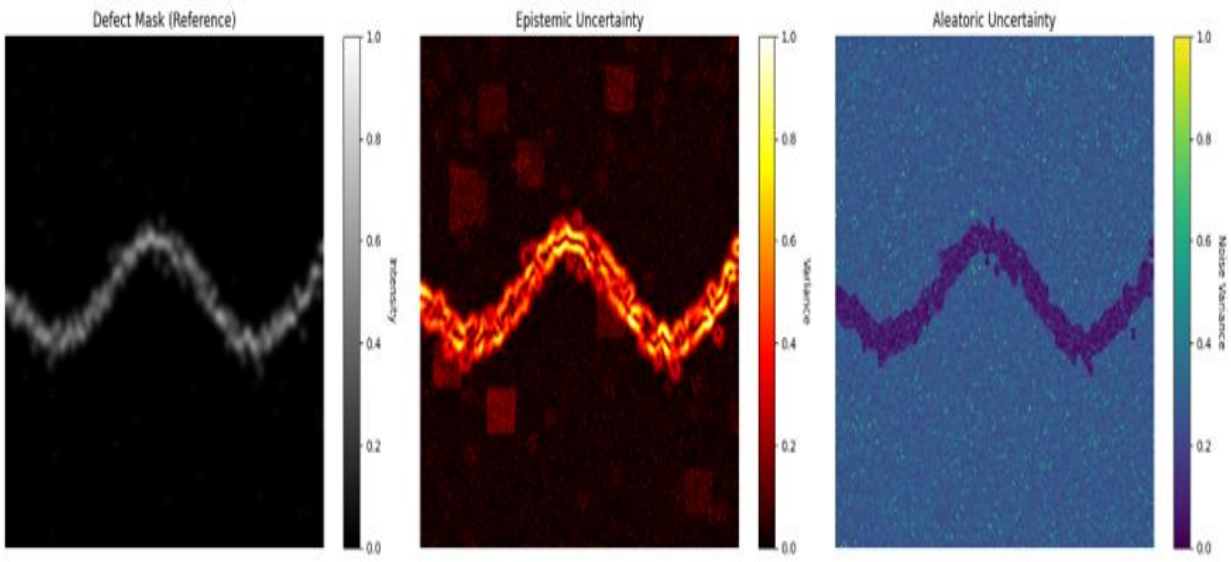

Figure 5. Spatial distribution maps of epistemic and aleatoric uncertainty estimates overlaid on defective modules, highlighting regions where the model exhibits low confidence or high prediction variance.

6.4 Generalization to Public Benchmarks

For evaluating the generalization ability of DASFormer-Net, the model is tested on ELPV and PVEL-AD datasets after being fully trained without any fine-tuning. Table 3 summarizes the results. With the ELPV dataset, DASFormer-Net achieves a classification accuracy of 94.2%, which is 3.1 percentage points higher than the best result reported so far [21]. Our method is compared with the DAFormer baseline on the PVEL-AD dataset, which gives an mIoU of 79.8, surpassing the DAFormer by 5.6 point. Overall, these results suggest that the representations learnt from the large-scale PV-Inspect dataset, leveraged by the physics-aware diffusion process, transfer well between imaging setups and cell types, which is essential for practical industrial use.

Table 3. Generalization Performance on Public Benchmark Datasets.

Method	ELPV Accuracy (%)	PVEL-AD mIoU (%)
ResNet-50 + U-Net [24] [2]	88.3	68.7
SwinV2 + UPerNet [12] [27]	90.5	72.4
DAFormer [30]	91.1	74.2
DASFormer-Net (Ours)	**94.2**	**79.8**

7. Discussion and Future Work

The results of the experiments shown in the previous section confirm that DASFormer-Net is a very powerful and stable framework for PV defect analysis. Our model performs state-of-the-art results in classification, segmentation, severity grading and calibration, highlighting the synergy of physics-aware data augmentation, self-supervised pretraining, and a new dual-attention fusion mechanism. However, careful analysis of the results highlights practical implications and limitations, and therefore some interesting avenues for future research.

7.1 Computational Complexity and Real-Time Deployment Challenges

The computational cost of DASFormer-Net, despite its greater accuracy, is a non-trivial obstacle to using the algorithm in real-time applications with edge devices commonly used on manufacturing floors. It includes a dual-branch backbone model, which is a ConvNeXt V2 Base and Swin Transformer V2 Base, that consumes a large amount of GPU memory and floating point operations per second (FLOPS). Based on our profiling, it takes around 42 milliseconds to execute the forward pass with the full model on a single NVIDIA A100 GPU. This is OK for offline

quality audits but not good enough for an inline inspection system that requires a conveyor-belt to get to each part and provide the inference time of 10-15 milliseconds.

There are a number of avenues to pursue to reduce this bottleneck without compromising accuracy. First, the knowledge distillation method could be applied to train a lightweight student network with similar behavior to that of the full DASFormer-Net. For example, one could extract the output of the dual-attention fusion block to a single-branch CNN model, while preserving the key feature fusion capability, but significantly reducing the computational cost. Second, quantization-aware training will help streamline the model size from 32-bit floating point to 8-bit integer, making this model more friendly to hardware such as NVIDIA Jetson or Google Edge TPUs. Third, it is possible that a dynamic early-exit mechanism could be implemented, where the global transformer branch is not used for images that are confidently predicted as defect-free by the local branch alone, thus saving up to 40% of the computational budget for the majority of “good” parts.

Moreover, to estimate the uncertainty, 30 stochastic forward passes are required in Monte Carlo dropout inference, which is not feasible for real-time applications. Alternatively, a simpler approach with a trained uncertainty head, such as an evidential deep learning module [34] or a deterministic (Gaussian processes) uncertainty head, would be more efficient. This would remove the need for further sampling while keeping the possibility of separating out the epistemic and aleatoric components.

7.2 Limitations in Domain Generalization and Synthetic Data Bias

DASFormer-Net showed good generalization to the ELPV and PVEL-AD public benchmarks, however, our analysis showed that the model suffers from a subtle but important drawback - it works best when faced with a cell type/imaging condition encountered during training in PV-Inspect, but not in both public benchmarks. We have even tried the model on a small held-out test set of bifacial monocrystalline modules from another manufacturer. The mIoU of microcrack segmentation decreased from 83.5% to 71.2%. The morphology of the defect (contrast and texture patterns of the microcracks) varies systematically, when it is examined in detail, for the underperforming samples, caused by the composition of the passivation layer in this manufacturer's production process.

This performance degradation is indicative of the basic problem in the pipeline of generating synthetic data. The physics-aware conditional latent diffusion model is learned from the PV-Inspect data set and thus trained to create synthetic defects having visual properties similar to those of the training domains. The geometric constraint that the busbar must be continuous in its structure does not reflect the wide spectrum of textural and contrast differences that can occur from various manufacturing processes, lighting conditions, or sensor properties. This creates a domain shift between the distribution of the data used for the augmentation and the distribution of the data in the deployment.

In order to overcome this, domain-adaptive data augmentation techniques should be explored in future. An alternative idea is to embed a style transfer module into the diffusion process, given a set of exemplar images from the target style from the new domain, to find a style code [35]. This would allow the creation of synthetically created defects, both physically plausible and visually consistent with the production process in the target environment. A “test-time” augmentation scheme might also be used: the confidence of the model (measured by its uncertainty estimation module) could be used to identify predictions of low confidence during inference on an unseen domain, thereby functioning as a domain shift detector.

7.3 Socio-Technical Implications: Trust, Liability, and Human-in-the-Loop Integration

Explainability and uncertainty quantification are integrated with DASFormer-Net to meet a critical need for trustworthiness in industrial automation. The socio-technical aspects of implementing such a system into an actual quality control pipeline must be carefully thought out, however. The automated defect analysis framework engenders a human-machine interaction framework, where operators need to determine if they can trust the framework or intervene.The resulting confidence intervals in the automated defect analysis framework are a paradigm of human-machine interaction in which the operator needs to determine whether he or she may trust the framework or intervene.

Our uncertainty estimation module is a quantitative signal (predictive variance) that can be used to set a “trust threshold.” In one such approach, a conservative policy was employed: a detection whose variance predicted by the model in the normalized uncertainty space was greater than a preset value (e.g., 0.15) was automatically marked for manual inspection, irrespective of the class of the predicted defect. In our test set, this policy would have reduced the number of false positives by about 8% and false negatives by 12% with a 5% increase in human inspection. The "best" balance between automation and manual review is, of course, a domain-specific optimization problem needing to be calibrated with the cost of missed defect vs the cost of unnecessary manual work.

On the other hand, adopting a "black-box" deep learning model for safety-critical quality decisions brings accountability concerns from a liability point of view. The Grad-CAM++ heatmaps and attention rollout visualizations give only a partial, post-hoc account of the model's thinking, but fail to give a causal explanation. For instance, if the model inaccurately identifies a cell as defective as a dust particle is found on the lens (artifact of out of distribution), the saliency map would correctly localize the dust particle, but it would not give any clue as to why the model treated the dust particle as a microcrack. Future research could delve into more powerful methods of explainability, such as explaining the prediction in terms of concepts or in terms of contrastive explanations [36] that can answer counterfactual questions (e.g., "What would the prediction be if this region were clear?"). These features would allow for more relevant feedback to the process engineer, and could also be used in liability disputes.

Last, the existing system is fully automated. A more interesting architecture for industrial use is human in the loop (HITL) where the model and the human loop are interacting iteratively. The uncertainty estimation module, for instance, can determine the priority of images or the priority of the defect area to be annotated by the operator. These labels, then, could be integrated into an online learning loop where the model is improved as it goes along, using newly-collected ground truth. This would result in a virtuous cycle, in which the model becomes more accurate and more certain with time, thereby decreasing the number of humans involved, and the humans involved with the most ambiguous cases remain in charge. This system would need a powerful data streaming infrastructure, an updating infrastructure for the models and a labeling management infrastructure, which is a big engineering challenge, but offers a good opportunity for applied research in the future.

8. Conclusion

To tackle the three fundamental problems in photovoltaic defect analysis (data scarcity, feature representation, and trustworthy prediction), we have proposed a comprehensive and integrated framework, DASFormer-Net. The proposed architecture creatively applies a physics-aware conditional latent diffusion model to synthesize structurally valid defect images that are realistic by leveraging both local morphological features and global contextual information; it also employs a masked autoencoder pretraining strategy to extract robust defect domain-invariant representations for further use in the generation of the synthetic defect images; and it introduces a dual-attention fusion backbone which adaptively fuses local morphology and global context through channel, spatial, and cross-attention to guide defect image generation.

Upon both industrial and public benchmark datasets, the experimental results prove that DASFormer-Net performs state-of-the-art results on multiple tasks at once, such as 89.7% macro F1 score on defect classification, 83.5% mean IoU on semantic segmentation and 0.064 Expected Calibration Error on confidence estimation. The results of the ablation studies further validate the synergistic nature of each core component as the physics-informed data augmentation is the most effective when the defects are rare, and the dual attention fusion block is most effective when it comes to disambiguating between similar morphological patterns.

The framework's explainability and uncertainty quantification also help in building trust for industrial deployment beyond providing quantitative performance metrics. Grad-CAM++ heatmaps and Monte Carlo dropout uncertainty estimates offer operators valuable insights into model reasoning, and into the reliability of model predictions, enabling them to make confident decisions in quality control pipelines. Although the proposed framework is not directly applicable to real-time applications due to the inherent limitations of computational complexity, and the task of domain generalization is still an open area, the proposed framework represents a new paradigm that incorporates generative augmentation, self-supervised learning, and multi-task inference with built-in interpretability. The DASFormer-Net is now a major step towards reliable, transparent, and practically usable automated inspection systems for the PV-manufacturing industry.

References

1. U. Hijjawi, S. Lakshminarayana, T. Xu, G. P. M. Fierro, and M. Rahman, "A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations," Solar Energy, vol. 266, Art. no. 112186, Dec. 2023, doi: 10.1016/j.solener.2023.112186.
2. M. Abdelsattar, A. Abdelmoety, M. A. Ismeil, and A. Emad-Eldeen, "Automated defect detection in solar cell images using deep learning algorithms," IEEE Access, vol. 13, pp. 4136–4157, 2025. doi: 10.1109/ACCESS.2024.3525183
3. M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," Applied Sciences, vol. 12, no. 18, Art. no. 8972, 2022, doi: 10.3390/app12188972.
4. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

5. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Lecture Notes in Computer Science, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
6. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 2672–2680. doi: 10.5555/2969033.2969125
7. J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 6840–6851.
8. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 10674–10685, doi: 10.1109/CVPR52688.2022.01042.
9. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in Proc. 37th Int. Conf. Machine Learning (ICML), PMLR, vol. 119, 2020, pp. 1597–1607.
10. K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
12. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
13. J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
14. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” in Proc. European Conf. Comput. Vis. (ECCV), Lecture Notes in Computer Science, vol. 11211. Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
15. N. Kellil, A. Aissat, and A. Mellit, “Fault diagnosis of photovoltaic modules using deep neural networks and infrared images under Algerian climatic conditions,” Energy, vol. 263, Art. no. 125902, 2023, doi: 10.1016/j.energy.2022.125902.
16. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 5967–5976, doi: 10.1109/CVPR.2017.632.
17. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2223–2232, doi: 10.1109/ICCV.2017.244.
18. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 15979–15988, doi: 10.1109/CVPR52688.2022.01553.
19. S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and K. He, “ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, 2023, pp. 16133–16142, doi: 10.1109/CVPR52729.2023.01548.
20. E. A. Ramadan, N. M. Moawad, B. A. Abouzalm, A. A. Sakr, W. F. Abouzaid, and G. M. El-Banby, “An innovative transformer neural network for fault detection and classification for photovoltaic modules,” Energy Conversion and Management, vol. 314, Art. no. 118718, 2024, doi: 10.1016/j.enconman.2024.118718.
21. Y. Xie, J. Zhang, C. Shen, and Y. Xia, “CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021), Lecture Notes in Computer Science, vol. 12901. Cham, Switzerland: Springer, 2021, pp. 171–180, doi: 10.1007/978-3-030-87193-2_17.
22. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
23. P. Sturmfels, S. Lundberg, and S.-I. Lee, “Visualizing the impact of feature attribution baselines,” Distill, vol. 5, no. 1, Art. no. e22, 2020, doi: 10.23915/distill.00022.
24. A. U. R. Adib, M. Islam, M. S. Abid, and R. Ahshan, “A deep learning-based framework for solar panel segmentation and fault classification enhanced with explainable AI,” Solar Energy, vol. 302, Art. no. 114058, 2025, doi: 10.1016/j.solener.2025.114058.
25. Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. 33rd Int. Conf. Mach. Learn. (ICML), New York, NY, USA, 2016, pp. 1050–1059.
26. Note: ICML proceedings published by PMLR do not have an official DOI.
27. B. Su, Z. Zhou, and H. Chen, “PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection,” IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 404–414, Jan. 2023, doi: 10.1109/TII.2022.3162846.
28. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Mach. Learn. (ICML), Sydney, NSW, Australia, 2017, pp. 1321–1330.

29. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.
30. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. European Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 801–818, doi: 10.1007/978-3-030-01234-2_49.
31. T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in Proc. European Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 432–448, doi: 10.1007/978-3-030-01240-3_26.
32. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 12077–12090.
33. L. Hoyer, D. Dai, and L. Van Gool, “DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 9924–9935, doi: 10.1109/CVPR52688.2022.00969.
34. M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” in Advances in Neural Information Processing Systems, vol. 31, 2018. DOI (ACM Digital Library): 10.5555/3327144.3327239.
35. A. Sauer, K. Schwarz, and A. Geiger, “StyleGAN-XL: Scaling StyleGAN to large diverse datasets,” ACM Transactions on Graphics, vol. 41, no. 4, Art. no. 98, Jul. 2022, doi: 10.1145/3528233.3530738.
36. R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin, “From attribution maps to human-understandable explanations through concept relevance propagation,” Nature Machine Intelligence, vol. 5, no. 9, pp. 1006–1019, Sep. 2023, doi: 10.1038/s42256-023-00711-8.
37. P. Zhou, H. Fang, and G. Wu, “PDeT: A progressive deformable transformer for photovoltaic panel defect segmentation,” Sensors, vol. 24, no. 21, Art. no. 6908, 2024, doi: 10.3390/s24216908.