DPiT: Detecting defects of Photovoltaic Solar Cells with Image Transformers

XIANGYING XIE\textsuperscript{1,2}, HU LIU\textsuperscript{3}, ZHIXIONG NA\textsuperscript{2,*}, XIN LUO\textsuperscript{2}, DONG WANG\textsuperscript{2}, and BIAO LENG\textsuperscript{1}

\textsuperscript{1}School of Computer Science & Engineering, Beihang University, Beijing 100191, China
\textsuperscript{2}State Grid e-commerce Co., Ltd, Beijing 100053, China
\textsuperscript{3}State Grid Corporation of China, Beijing 100031, China

Corresponding author: Zhixiong NA (e-mail: hellonazx@163.com).

This work was supported by National Key Research and Development Program of China under Grant 2018YFB1500800 and Technology Project of State Grid Corporation of China under Grant SGTJDK00DYJS2000148.

ABSTRACT Solar energy is one of the most important resources that can be a clean and renewable alternative to traditional fuels. The collection process of solar energy mainly rely on the photovoltaic solar cells. The defects, such as microcracks and finger interruption on the photovoltaic solar cells can reduce its efficiency a lot. To solve this problem, defects detection of solar cells have attracted attention from many researchers. In this paper, we propose a transformer based network to detect defects on solar cells efficiently and effectively. First, we introduce convolutions into the transformer to enable the input embeddings of the transformer, positional information of patches and spatial context more accurate and precise. Secondly, cross window based multi-head self-attention (CW-MSA) is proposed to enlarge the window relation modeling capacity via the strong attention mechanism and can be an effective alternative to the original counterpart. Finally, we propose a multi-scale aggregation block to merge the low-level features into deep semantically strong features by attention to obtain accurate geometry information. Extensive experiments on the elpv dataset demonstrate DPiT can consistently bring significant improvements over its strong baseline Swin Transformer with subtle extra computational overhead. The visualization results show that the proposed DPiT is able to detect various complex defects correctly. In particular, DPiT can achieve impressive 91.7 top-1 accuracy and greatly outperforms other competitive counterparts.

INDEX TERMS Solar energy, defects detection, transformer, multi-head self-attention, multi-scale aggregation

I. INTRODUCTION

With the rapid development of technology, people are paying more attention to the environment protection than ever before. Clean and renewable energy will become progressively more important as time goes on. As a good alternative to traditional energy, solar energy is one of the most important resource to solve the energy issue. People mainly make use of solar energy by solar cells. Solar cells can be divided into monocrystalline cells and polycrystalline cells. Polycrystalline solar cells are the most widely used solar cells in the world, whose cracks detection is much more harder to be handled than monocrystalline. As shown in Fig. 1, photovoltaic solar cells may get damaged during the solar energy collection process. For example, thermal stress and improper operations is able to bring the photovoltaic solar cells cracks and cause cell breakages. These defects are the main reason why the power efficiency of photovoltaic solar cells decline gradually. To prevent the solar cells from getting terrible electrical power loss, we need to detect the defects in the cells in due course to enable the production process stable.

As one of the most challenging problems, defects inspection of solar cells have attracted much attention of many researchers in the past few years. Previous works mainly tend to detect defects of solar cells by physical methods. The work [1] introduces current-voltage (IV) characteristics for silicon cell characterization. Besides, thermography which is only implemented on infrared ray images is adopted for defects detection [2]–[4]. However, the work [2] points out that infrared ray images are at a lower resolution that leads to make the microcracks detection much more difficult. Electroluminescence imaging [5], [6] is an improved alternative to thermography. Actually, electroluminescence (EL) imaging is one of the important technologies for defects visualiza-
tion of photovoltaic solar cells [7]. Besides, The automatic processing of EL owns the ability to analyze the defects of high-resolution solar cell images. The work [5] introduces the Fourier image reconstruction into the field of defects detection in EL data. Tseng et al. [8] first list some regions with high confidence scores and find the target regions by binary feature clustering. Some previous works [9] detect defects of solar cells by hand-crafted methods. The work [9] proposes the dimensional matched filters to inspect the cracks on solar cells.

Methods base on deep learning have dominated image classification and detection in the past few years. Convolutional Neural Network (CNN) has been widely used in computer vision and plays an important role in image classification, object detection and other tasks, which shows that CNN has led the development of computer vision. For example, the pioneering work AlexNet [10] is proposed to prove CNN can achieve better performance than traditional image classification methods. He et al. [11] introduces ResNet to shown residual learning is able to considerably ease the optimization of deep convolutional neural networks. As for object detection, Fast R-CNN [12] and Faster-RCNN [13] first obtain multiple regions of interest and extract region features. These regions are further fed into the detection head for classification and bounding box regression. Inspired by these milestone works, some works [14]–[16] have been proposed to prove that CNN also achieves impressive performance on defects detection. The work [17] adopts a novel method called normalized lp norm to solve the problem of weakly supervised segmentation of cracks on solar cell. Recently, U-net with a region proposal network [18] is proposed to accurately localize and segment the silicon polycrystalline defects. Besides, the method adopts dilated convolution to enlarge the receptive field to enable defects segmentation more precise.

However, convolutional operation only processes a local neighborhood and fails to capture the long-range dependencies among pixels [19]. Different from CNN, the transformer is able to model the global relations [20]. The transformer architecture was first proposed for machine translation in the filed of natural language processing [21]. The encoder network in transformer is composed of self-attention operation in conjunction with multi-layer perceptron (MLP), obtaining remarkable results in capturing the long-range dependencies. Transformers have since become the state-of-the-art method in many NLP tasks [21]. The success of the transformer in NLP leads to many attempts at adapting the transformer architecture to images. Recently, many visual transformers [22], [23], [25] are proposed to achieve state-of-the-art performance in the field of image classification. ViT [22] first splits an image into many non-overlapping patches, and model the relations among these patches by a transformer encoder. DeiT [23] points out that ViT only outperforms the state-of-the-art CNN methods when the training data is sufficient and introduces better hyperparameters settings and distillation strategy to train a data-efficient ViT. CPVT [24] proposes a dynamic positional encodings that are implemented by convolution with zero paddings to enable the transformer is able to receive images with different resolutions. The attention map will be very similar in the deep layers of transformer, which may lead to attention collapse. To solve this problem, DeepViT [25] introduces re-attention mechanism to combine attention maps of different heads to enhance the diversities of attention maps. Other work, such as T2T-ViT [26] and PVT [27], exploit efficient transformer designing principles. To reduce the computation cost of transformer, Swin Transformer [28] introduces Window based MSA (W-MSA) and Shifted Window based MSA (SW-MSA). This hierarchical transformer can serve as a general-purpose backbone and achieve the state-of-the-art performance on image classification and other dense recognition tasks.

However, there still exists gap between the data of polycrystalline solar cells and general images. For example, semantically strong features in deep layers are generally utilized for classification and detection in current frameworks. Low-level features are usually deemed rich in containing geometric patterns and stripes, which is contrary to high-level features. As for the defects detection of polycrystalline cells, the defects including microcracks and finger interruption are mainly contained in the low-level features. It is intuitive that the stripes in the image of polycrystalline solar cell own rich discriminative information, which is beneficial for detecting the defects in the image. Inspired by this, we propose DPiT, which builds upon the Swin Transformer to effectively and efficiently detect the defects of polycrystalline solar cells. The patch embedding layer in Swin Transformer directly split the image into non-overlapping windows, failing to model more precise information. We add another fully convolutional network to enhance feature extraction. As for self-attention layer, we introduce Cross Window based MSA (CW-MSA) to replace the SW-MSA to enlarge the window relation modeling capacity. We also model spatial context in the MLP which only applied on channel dimension. Dy-
namic positional encodings based on convolution with zero paddings are added before every stage of the transformer. Besides, we introduce a bottom-up pathway which is parallel to the backbone forwarding to gradually integrate both high-level and low-level features by attention mechanism. To demonstrate its robustness and effectiveness, we test our proposed DPiT on the elpv dataset. The results show that DPiT consistently outperforms the original Swin Transformer and other CNN based network by a large margin.

The main contributions of this paper are as follows,

- We introduce convolutions into the patch embeddings, positional encodings and MLP to enable the input features of the transformer, positional information of patches and spatial context more accurate and precise.
- Cross Window based MSA (CW-MSA) is proposed to replace the original SW-MSA to enlarge the window relation modeling capacity via the strong attention mechanism.
- We propose a multi-scale aggregation block to merge the low-level features into deep semantically strong features by attention before classification to obtain accurate geometry information.
- Our proposed DPiT is able to outperform the baseline Swin Transformer and achieve very impressive performance on the elpv dataset.

II. METHODOLOGY

A. OVERALL FRAMEWORK

As shown in Fig. 2, our proposed DPiT builds upon the Swin Transformer architecture. The first step of classification pipeline is to flatten the input 2D image with resolution \( H \times W \) into a sequence of 2D non-overlapping patches. Then this feature sequence of patches is fed into a patch embedding block to be projected to a dimension 48. Following the default settings in Swin Transformer, patches with size of \( 4 \times 4 \) are merged as the inputs of the transformer. The hierarchical backbone of DPiT is composed of 12 DPiT blocks and the patches are downsampled via a patch merging layer. We split the transformer into 4 stages as shown in Figure. Stage 1 receives the output of linear embedding layer with shape \( 48 \times \frac{W}{4} \times \frac{W}{4} \) and generates the extracted feature with the same resolution but larger channel \( C \). A patch merging layer is inserted into the bottom of “Stage 2”, which merges the features of groups of \( 2 \times 2 \) adjacent patches. The downsampled features are further fed into transformer blocks after neighboring feature merging. In a similar way, “Stage 3” and “Stage 4” receive inputs with resolution \( \frac{H}{16} \times \frac{W}{16} \) and \( \frac{H}{32} \times \frac{W}{32} \) and downsample them by \( 2 \times \). Finally, a global average pooling layer and a linear classifier are applied on the output feature of “Stage 4”. We further specially incorporate four effective components for the defects detection task of polycrystalline solar cells.

B. PATCH EMBEDDING

The original patch embedding block in the Swin Transformer is adopted to partition the input 2D image into some non-overlapping patches. We reshape the input image \( x \in \mathbb{R}^{3 \times H \times W} \) to obtain the patch grids \( x_p \in \mathbb{R}^{48 \times \frac{W}{4} \times \frac{W}{4}} \). A linear embedding layer is applied to enable the transformer use the constant latent dimension \( C \) as the feature channel through all of its blocks. This patch partition process is very simple but cannot capture low-level information in the earlier layers and ignores the relation among all the pixels near the border of all patches. To effectively learn representations of the low-level information in the image, we additionally introduce convolutions into the patch embedding block to perform resolution reduction due to their strong inductive biases. As shown in Fig. 3, we inserted 4 layers of \( 3 \times 3 \) convolutions followed by a BN (Batch Normalization) [29] layer and GELU activation [30] function to directly extract...
features. To avoid breaking the initial behavior of pre-trained model, we design these 4 convolution layers as another extraction path and initialize the weight of the last BN layer as zero. The extraction process of patch embedding block can be illustrated as:

\[
    z_0 = F_{\text{linear}}(x_p) + F_{\text{conv}}(x_p),
\]

where \(z_0\) denotes the patch embeddings, \(F_{\text{linear}}\) and \(F_{\text{conv}}\) represents the original linear patch embedding and convolution patch embedding, respectively. We do not add position embeddings to the patch embeddings in this block. The patch embeddings \(z_0\) are the input to the transformer encoder.

C. TRANSFORMER BLOCK

The transformer block proposed in the transformer encoder consists of a multi-head self-attention (MSA) layer and a MLP. Unlike the standard multi-head self-attention layer in other visual transformers, a novel window based multi-head self-attention (W-MSA) layer is proposed in the baseline model. This window based attention mechanism proposes a shifted window partitioning approach to introduce cross-window connections. The cyclic shift method still can be improved since only a few pixels of each window are shifted during window based self-attention computation. The MLP in transformer block is composed of 2 fully-connected layers and GELU non-linearity in between, which lacks the ability to model spatial relations. To solve these problems, we introduce cross-window multi-head self-attention (CW-MSA) to capture salient window relations and MLP with spatial context module (SC-MLP) to promote the spatial modeling power of MLP in Fig. 4. The forward process of the transformer block can be illustrated as:

\[
    \hat{z}_l = \text{W-MSA}(\text{LN}(z_{l-1})) + z_{l-1},
\]

\[
    z_l = \text{SC-MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l,
\]

\[
    \hat{z}_{l+1} = \text{CW-MSA}(\text{LN}(z_l)) + z_l,
\]

\[
    z_{l+1} = \text{SC-MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1},
\]

where \(l\), \(\hat{z}\) and \(z\) denote the index of transformer block, the output of self-attention layer and the output of MLP. W-MSA, CW-MSA and SC-MLP denote the window based multi-head self-attention layer, cross-window multi-head self-attention layer and MLP with spatial context augmentation.

D. WINDOW BASED SELF-ATTENTION

In our implementation, an image is partitioned into \(\frac{h}{M} \times \frac{w}{M}\) non-overlapping windows. We first split a windows into \(M \times M\) patches and compute the self-attention matrix of all patches within the local window in the window based self-attention layer. The computation process of W-MSA is shown as:

\[
    \text{W-MSA}(Q_1, K_1, V_1) = \text{SoftMax}(\frac{Q_1 K_1^T}{\sqrt{d}} + B_1)V_1,
\]
where \( Q_1, K_1, V_1 \) denote query, key and value with shape \( M^2 \times d \), \( d \) is the channel dimension of feature, \( B_1 \) is the position embedding bias. W-MSA is able to model the global relation among all the patches in a local window. Through computing dot-product, the similarity between different patches in the window is calculated, resulting in long-range and global attention. To further enhance the capacity of global context, we design the CW-MSA to capture cross-window relation. The forwarding of CW-MSA is similar to W-MSA:

\[
\text{CW-MSA}(Q_2, K_2, V_2) = \text{SoftMax}\left(\frac{Q_2K_2^T}{\sqrt{d}} + B_2\right)V_2, \tag{7}
\]

where \( Q_2, K_2, V_2 \in \mathbb{R}^{h_w \times w \times d} \) denote query, key and value of non-overlapping windows. \( B_2 \) is the relative position bias of all windows, which leads to significant improvements. Our proposed CW-MSA directly model the relation among all windows, achieving better performance compared to the original shifted window based self-attention (SW-MSA) in the experiments.

**E. MLP WITH SPATIAL CONTEXT MODULE**

![Diagram](image)

The MLP (multi-layer perceptron) which adds linear transformation to the features plays a crucial role in almost all transformer-based architectures. MLP performs point-wise operations and are applied to each token separately. It is composed of two fully-connected layers and non-linear activation in between. Given a input feature \( z \) of self-attention layer, the MLP transforms the feature as follows:

\[
\text{MLP}(z) = \sigma(zW_1 + b_1)W_2 + b_2, \tag{8}
\]

where \( W_1 \in \mathbb{R}^{C \times K} \) and \( W_2 \in \mathbb{R}^{K \times C} \) denote the weight of the first layer and the second layer, respectively. In our implementation, we set \( K \) as \( 4C \). \( b_1 \) and \( b_2 \) are the biases. And \( \sigma \) is the non-linear activation, which is GELU in most visual transformers.

Obviously, MLP only performs linear and non-linear transformation on each window, enhancing the representation ability of all windows. However, the spatial information of windows are ignored. In this paper, we propose a spatial context module into MLP to form SC-MLP to extract the spatial representations and merge them into the semantic information. As shown in Figure., we first adopt a convolution layer \( W^k \) with kernel size of \( 3 \times 3 \) and softmax function to obtain the spatial attention weights \( A_l \). The attention weights \( A_l \) is further multiplied by the input features \( \hat{z}_l \). Then a \( 1 \times 1 \) convolution \( W^p_l \) is used to linearly transform the spatial information to semantic information. Finally, we normalize the global context to channel attention weights by softmax function. The output \( z_l \) is obtained via scaling the features by channel attention weights. The computation process of SC-MLP can be shown as following:

\[
A_p = \text{Max}(W^k_l\hat{z}_l) \tag{9}
\]

\[
z_l = F_{\text{scale}}(\text{Max}(A_p\hat{z}_lW^p_l), \hat{z}_l) \tag{10}
\]

where \( F_{\text{scale}} \) is the channel-wise multiplication between the attention map and features, \( z_l \) denotes the output features of SC-MLP. We also have tried different kernel sizes of \( W^k \) such as 5 and 7, convolution with kernel size of 3 achieves the best speed-accuracy trade-off. Therefore, we adopt convolution with kernel size of 3 as \( W^k \) by default.

**F. DYNAMIC POSITIONAL ENCODINGS**

![Diagram](image)

In Swin Transformers, the attention maps in self-attention layer are added with the same number of learnable absolute positional encoding vectors. The shape of the positional encodings is fixed, which prevents the model from handling input sequences are shorter or longer than the original sequences. To solve this problem, many visual transformers resize the shape of fixed positional encoding vectors by bicubic interpolation, which still hurts the performance of
the model. Recent works have observed that the tokens on the borders can be aware of their absolute positions if we use zero paddings as illustrated in Fig. 6. Inspired by this, we introduce dynamic positional embedding into features of windows before every stage of the transformer. As shown in Figure., we adopt a very simple depth-wise convolution with kernel size of $k (k \geq 3)$ and zero paddings $\frac{k-1}{2}$ to dynamically assign positional information. The dynamic positional embedding is added back to the main-stream features. In our experiments, we have tested depth-wise convolution with $k = 3, 5, 7$ and regular convolution with $k = 3$.

**G. MULTI-SCALE AGGREGATION BLOCK**

![Multi-scale Aggregation Block](image)

**FIGURE 7.** The architecture of our proposed multi-scale aggregation block (MAB). “Q”, “K” and “V” denote query, key and value of attention mechanism. The downsampling process of $x_{i-1}$ is ommited in the figure.

It is intuitive that the stripes in the image of polycrystalline solar cell own rich discriminative information, which is beneficial for detecting the defects in the image. In the image classification network, the final output high-level features are semantically strong, but lacks precise low-level information. Low-level features are usually deemed rich in containing geometric patterns and stripes, which is contrary to high-level features. Inspired by this, we propose a multi-scale aggregation block (MAB) to fuse as much as possible low-level information to enhance the defects detection ability of the transformer.

Unlike some previous works [31], [32] integrate multi-scale features by convolutions, we propose to facilitate multi-scale feature fusion by cross-attention in this paper. It is worth noting that we are the first to adopt attention to aggregate multi-level features. As shown in Fig. 7, we first resize the low-level feature $x_{i-1}$ of stage $i$ by $3 \times 3$ convolution with stride 2. Then we adopt attention mechanism to fuse downsampled low-level feature $x_{i-1}$ and feature $x_i$ at higher level. An attention function can be described as mapping a query and a set of key-value pairs to an output. Therefore, we treat the low-level feature $x_{i-1}$ as the query and high-level feature $x_i$ as key and value in the attention function. Our motivation is simple: treating the low-level feature as the query to select the most similar features which contains rich geometric stripes information to help defects detection. We apply a linear layer on $x_{i-1}$ to get the query and two linear layers on $x_i$ to get the key and value. Different from aforementioned W-MSA and CW-MSA, we model the relationships among all patches in the image to get precise and accurate low-level information. A residual connection is added to ease optimization. We insert two MAB at Stage 3 and Stage 4 to construct a bottom-up pathway which is parallel to the transformer. Here we show the forwarding process of MAB:

\[
Q_i = x_{i-1} W_i^q
\]

\[
K_i = x_i W_i^{key}
\]

\[
V_i = x_i W_i^v
\]

\[
\hat{x}_i = x_i + \text{Softmax}\left(\frac{Q_i K_i}{\sqrt{d_i}}\right)V_i,
\]

where $\hat{x}_i$ denotes the output of MAB and $d_i$ is the channel dimension of feature at stage $i$. $W_i^q$, $W_i^{key}$ and $W_i^v$ denote the linear layers in the MAB.

**H. TRAINING LOSS**

Assuming the data loaded from the dataset is composed of a sequence of image-label pairs \{(X_1, Y_2), ..., (X_N, Y_N)\}, where $X_i$ and $Y_i$ denote the input image and the corresponding defect label, $N$ denotes the number of images. We use a general cross-entropy loss to optimize the transformer during training stage as follows:

\[
L = \frac{1}{N} \sum_{i=1}^{N} -(Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)),
\]

where $p_i$ denotes the probability of $Y_i$ is 1. We only consider the situation that $Y_i$ is 0 and $Y_i$ is 1 since there are only two classes in defects detection.

**III. EXPERIMENT RESULTS AND ANALYSIS**

**A. DATASETS**

We conduct experiments on elpv dataset [33]–[35] of solar cell images extracted from high-resolution electroluminescence images of photovoltaic modules. The elpv dataset consists of 2624 samples of 8-bit grayscale images of functional and defective solar cells with varying degree of degradations. The defects in the images are known to reduce the power efficiency of solar modules. The training set that consists of 2298 images from 2 classes is utilized for training. We report the top-1 accuracy on a single crop by default on the validation set which has 326 images. The labels of the dataset are the defect likelihood created by experts. Labels are categorized as four types. A label of 0 percent denotes that the expert ensure the solar cell is correctly labelled as not defective. A label of 33 percent denotes that the expert...
is not sure about that the solar cell is not defective. A label of 67 percent represents that the expert is not sure about that the solar cell is defective. Finally, a percent of 100 denotes that the expert ensure the solar cell is correctly labelled as defective. The detailed information of labels is shown in Table 1.

B. EVALUATION METRICS
We evaluate the performance of our proposed model by following the general evaluation metrics, including accuracy, precision, recall and F1 score. The accuracy, precision, recall and F1 score are calculated as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN},
\]

\[
\text{Precision} = \frac{TP}{TP + FN},
\]

\[
\text{Recall} = \frac{TP}{TP + FP},
\]

\[
F_1 \text{ score} = \frac{2TP}{2TP + FP + FN}.
\]

where TP and TN denote the number of images with and without defects are correctly classified, FP and FN denote the number of images with and without defects are incorrectly classified. Accuracy represents the probability of correct classification generated by our proposed transformer. Precision represents the proportion of detected defective images to all the defective images. Recall denotes the proportion of defective images that are correctly detected by our model. F1 score denotes the harmonic mean of precision and recall.

C. IMPLEMENTATION DETAILS
We train DPiT on 8 GeForce GTX 1080Ti GPUs for 200 epochs (10 epochs for linear warm-up) and adopt the cosine decay learning rate scheduler. AdamW optimizer is utilized for training. The batch size is set as 256 when we train DPiT. If not specifically noted, we use an initial learning rate of 0.0001 and a weight decay of 0.05. Some data augmentation strategies, including color jitter, rotation, flip, etc. are adopted during training. As for the initialization of our proposed model, we follow the kaiming initialization for all convolution layers.

D. MAIN RESULTS
In this subsection, we show the main results on the elpv dataset in Table 2. We first perform a comprehensive component-wise ablation study on Swin-Small, where different components are gradually inserted to thoroughly analyze the effect of each component. As shown in Table 1, by adding another fully-convolutional patch embedding to the transformer, DPiT-A achieves 90.5 top-1 accuracy, which leads to an absolute 0.7 points gain. This dramatic improvements prove that the low-level information captured by the fully convolutional networks is very essential to defects detection. DPiT-B still makes significant improvements (0.3 points) with Cross-Window MSA instead of vanilla Shifted Windows based MSA involving. We also find that the spatial attention and channel attention captured in SC-MLP benefit the classification performance of the DPiT-C by 0.4 points. Therefore, our proposed SC-MLP owns stronger capacity for both spatial and semantic modeling. Compared to DPiT-C, DPiT-D with dynamic positional encodings boosts the accuracy of the model by 0.3 points, which is a dramatic improvement. Finally, the insertion of MAB fully enables the deep features contain more low-level features to obtain 0.3 points improvement.

Besides, we also evaluate the precision, recall and F1 score of our proposed models in Table 3. Obviously, the gradual insertion of our modules results in that all models (from Swin-Small to DPiT) are arranged in the increasing order of all metrics, which also verifies the effectiveness of DPiT.

E. COMPARISON WITH OTHER METHODS
We report the results of DPiT on the elpv dataset and compare it with other recent competitive photovoltaic defects detection methods. We only focus on the performance of accuracy since precision, recall and F1 score ommitted in some papers. It is worth noting that the resolution of input image in our model is only $224 \times 224$, which is significantly smaller than the resolution in other methods. It is obvious that the baseline Swin-Small has surpassed all these CNN-based approaches with input resolution $224 \times 224$, proving the powerful global relation modeling ability of transformer. By augmenting another the fully convolutional patch embedding path, the accuracy is boosted to 90.5, which is dramatically improved to 91.7, which is superior than other CNN-based methods. The previous work reveals that large scale training trumps inductive bias in transformer. This means that transformer is able to achieve much better performance than CNN when large scale data is utilized for training. However, we only use 2624 EL images to train our DPiT. We speculate our proposed DPiT can achieve better performance if more data is used for training.

F. ABLATION STUDIES ON PATCH EMBEDDING
As shown in Table 5, We conduct additional experiments to find the optimal setting of our proposed fully convolution patch embedding. We first use two $3 \times 3$ convolutions with stride 2 to downsample the input image. Furthermore, $3 \times 3$ convolution layers with stride of 1 are used for feature refinement. “$3 \times 3$ Conv stride 1” in Table 5 allows us to vary the capacity and computational cost of the patch embedding block. To investigate the trade-off between accuracy and parameters, we conduct experiments on DPiT with various
numbers of $3 \times 3$ convolution. It is obvious that more $3 \times 3$ convolutions is able to lead better classification performance but bring additional parameters and computation cost. To achieve a good balance between accuracy and complexity, we use two $3 \times 3$ convolutions to downsample and two $3 \times 3$ convolutions for refinement.

### G. ABLATION STUDIES ON SC-MLP

As shown in Table 6, We conduct additional experiments to delve into the settings of our proposed SC-MLP. “add” denotes the addition fusion without the last softmax operation. “mul” denotes the feature is multiplied by the normalized channel attention weights. It is obvious that the kernel size of 1 achieves the worst performance since the convolution lacks the ability to capture spatial information. Besides, the top-1 accuracy of “mul” is higher than “add”, proving that the performance is robust to different kernel sizes. The classification performance reaches the peak by setting the kernel size as 4. We also find DPiT earns the maximum accuracy increase when the kernel size is 11. This relatively improved performance is a natural result of large receptive field.

### I. ABLATION STUDIES ON MULTI-SCALE AGGREGATION BLOCK

The multi-scale fusion strategy of multi-scale aggregation block (MAB) is similar to the milestone work FPN except for the information flow of the module. The MAB and FPN both merge the precise low-level features and high-level features with coarser resolutions. However, FPN propagates the high-level features at the top-most level to the lower levels, which aims to enhance the semantic information contained in the low-level features with accurate localization information. The top-down information flow of FPN is reverse to the bottom-up information flow of backbone. Different from the design principle of FPN, the motivation we introduce MAB is to make the deep features more discriminative by merging them with low-level features which contains accurate geometry stripes information. Another difference between FPN and MAB is the fusion operation of feature integration. FPN adopts the addition fusion followed by regular $3 \times 3$ convolution to aggregate multi-scale features. However, vanilla $3 \times 3$ convolution only processes the local neighborhood, ignoring the global relation. Besides, the relations among pixels from different scales are important for feature fusion. As shown in Table 8, we compare some fusion methods to find the best setting for MAB. It is surprising that “Conv-Conv”, “Conv-MSA” and “MSA-Conv” achieve the same performance. The top-1 accuracy of “MSA-MSA” is slightly higher than other 3 methods. This is a natural result of the global relation modeling ability of cross-attention.
TABLE 4. Comparing our proposed DPiT against other CNN-based approaches on the elpv dataset. Results are arranged in the increasing order of accuracy.

| Method            | Architecture | Epochs | Resolution | Accuracy |
|-------------------|--------------|--------|------------|----------|
| ADI [38]          | CNN          | 1000   | 300 x 300  | 83.0     |
| HFCNN8 [39]       | CNN          | 100    | 300 x 300  | 85.3     |
| HFCNN7 [39]       | CNN          | 100    | 300 x 300  | 86.5     |
| VGG-16 [40]       | CNN          | 100    | 300 x 300  | 86.5     |
| VGG-19 [40]       | CNN          | 100    | 300 x 300  | 87.1     |
| HFCNN10 [39]      | CNN          | 100    | 300 x 300  | 87.4     |
| HFCNN9 [39]       | CNN          | 100    | 300 x 300  | 88.4     |
| AC-CNN [35]       | CNN          | 100    | 300 x 300  | 88.4     |
| GoogleNet [41]    | CNN          | 200    | 300 x 300  | 88.8     |
| Swin-Small        | Transformer  | 200    | 224 x 224  | 89.8     |
| DPiT-A            | Transformer  | 200    | 224 x 224  | 90.5     |
| DPiT-B            | Transformer  | 200    | 224 x 224  | 90.8     |
| DPiT-C            | Transformer  | 200    | 224 x 224  | 91.2     |
| DPiT-D            | Transformer  | 200    | 224 x 224  | 91.5     |
| DPiT              | Transformer  | 200    | 224 x 224  | 91.7     |

TABLE 5. Performance of various fully convolutional patch embedding in DPiT. Results are evaluated on the elpv dataset. "k3 Conv s2" indicates the number of 3 × 3 convolution layers used for downsampling. "k3 Conv s1" denotes the number of 3 × 3 convolutions with stride 1.

| Method            | k3 Conv s2 | k3 Conv s1 | GFLOPs | Accuracy |
|-------------------|------------|------------|--------|----------|
| Swin-Small        | 0          | 0          | 8.75   | 89.8     |
| DPiT-A            | 2          | 0          | 8.75   | 90.2     |
| DPiT-A            | 2          | 1          | 8.76   | 90.3     |
| DPiT-A            | 2          | 2          | 8.77   | 90.5     |
| DPiT-A            | 2          | 3          | 8.78   | 90.6     |
| DPiT-A            | 2          | 4          | 8.79   | 90.6     |
| DPiT-C            | 3          | add        | 90.9   |          |
| DPiT-C            | 3          | mul        | 91.1   |          |
| DPiT-C            | 3          | mul        | 91.2   |          |

TABLE 6. Ablation study on the architecture of SC-MLP. Results evaluated on the elpv dataset are reported. "Spatial kernel" denotes the kernel size of the first convolution in the SC-MLP. "Fusion" denotes the approach to fuse the output of MLP and the transformed feature.

| Method            | Spatial kernel | Fusion | Accuracy |
|-------------------|----------------|--------|----------|
| Swin-Small        | N/A            | N/A    | 89.8     |
| DPiT-C            | add            |        | 90.9     |
| DPiT-C            | add            | mul    | 91.1     |
| DPiT-C            | add            | mul    | 91.2     |

J. VISUALIZATION OF DETECTION RESULTS

We visualize several detection results of DPiT on elpv in Figure 8. These results demonstrate that our proposed model is able to focus on the defects areas in the solar cells.

K. VISUALIZATION OF TRAINING PROCESS

In this subsection, we analyze the different behaviors of Swin-Small and our proposed DPiT during training. The accuracy diagrams of Swin-Small and DPiT are shown in Fig. 9. It is obvious that both two transformer-based methods is able to reach around 90% accuracy quickly. This is a natural result of small scale data training. Specially, the red curve of DPiT is generally higher than the blue curve of Swin-Small, proving its superior classification performance. After around 20 epochs, the accuracy curves become stable and stay around 90%. As can be seen in Fig. 10, both transformers converge very quickly. The loss of both two models fluctuate after around 20 epochs. The scale of the elpv dataset is small for the transformer architecture, which may lead to fast convergence and loss fluctuation.

IV. CONCLUSION

In this paper, we propose a novel transformer based network to detect the defects of photovoltaic solar cells effectively. We first analyze the limitations of the Swin Transformer for photovoltaic defects detection. To alleviate these limitations, we introduce DPiT, which includes fully convolutional embedding, Cross-Window MSA (CW-MSA), SC-MLP, dynamic positional encodings and Multi-scale Aggregation Block for photovoltaic defects detection. To prove the effectiveness of DPiT, we conduct additional experiments on the elpv dataset. The results on this dataset demonstrate DPiT can consistently
FIGURE 8. Visualizations of detection results on elpv dataset. We observe that the high confidence images are are correctly classified.

REFERENCES

[1] P. Tománek, P. Škarvada, R. Mack, and L. Grmela, “Detection and localization of defects in monocrystalline silicon solar cell.” Advances in Optical Technologies, 2010.

[2] M. Paggi, I. Berardone, A. Infuso, and M. Corrado, “Fatigue degradation and electric recovery in silicon solar cells embedded in photovoltaic modules,” scientific reports, vol. 4, no. 1, pp. 1–7, 2014.

[3] R. Ebner, S. Zamini, and G. Újvári, “Defect analysis in different photovoltaic modules using electroluminescence (el) and infrared (ir)-thermography,” in 25th European Photovoltaic Solar Energy Conference and Exhibition, 2010, pp. 333–336.

[4] U. Jahn, M. Herz, M. Köntges, D. Parlevliet, M. Paggi, and I. Tsanakas, Review on infrared and electroluminescence imaging for PV field applications: International Energy Agency Photovoltaic Power Systems Programme: IEA PVPS Task 13, Subtask 3.3: report IEA-PVPS T13-12: 2018. International Energy Agency, 2018.

[5] D.-M. Tsai, S.-C. Wu, and W.-Y. Chiu, “Defect detection in solar modules using ica basis images,” IEEE Transactions on Industrial Informatics, vol. 9, no. 1, pp. 122–131, 2012.

[6] K. Bedrich, M. Bokalčič, M. Bliss, M. Topić, T. R. Betts, and R. Gottschalg, “Electroluminescence imaging of pv devices: advanced vignetting calibration,” IEEE Journal of Photovoltaics, vol. 8, no. 5, pp. 1297–1304, 2018.

[7] Y. Jiang, C. Zhao, W. Ding, L. Hong, and Q. Shen, “Attention m-net for automatic pixel-level micro-crack detection of photovoltaic module cells in electroluminescence images,” in 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, 2020, pp. 1415–1421.

[8] D.-C. Tseng, Y.-S. Liu, and C.-M. Chou, “Automatic finger interruption detection in electroluminescence images of multicrystalline solar cells,” Mathematical problems in engineering, vol. 2015, 2015.

[9] S. Spataru, P. Hacke, and D. Sera, “Automatic detection and evaluation of solar cell micro-cracks in electroluminescence images using matched filters,” in 2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC). IEEE, 2016, pp. 1602–1607.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
FIGURE 9. Visualizations of top-1 accuracy on elpv dataset. The curves of Swin-Small and DPiT are shown in this figure.

FIGURE 10. Visualizations of training loss on elpv dataset. The curves of Swin-Small and DPiT are shown in this figure.

[12] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, vol. 28, pp. 91–99, 2015.

[14] S. Tian, W. Li, S. Li, G. Tian, L. Sun, and X. Ning, “Image defect detection and segmentation algorithm of solar cell based on convolutional neural network,” in 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), IEEE, 2021, pp. 154–157.

[15] M. W. Akram, G. Li, Y. Jin, X. Chen, C. Zhu, X. Zhao, A. Khalig, M. Faheem, and A. Ahmad, “Cnn based automatic detection of photovoltaic cell defects in electroluminescence images,” Energy, vol. 189, p. 116319, 2019.

[16] R. Perdicca, M. Paolanti, A. Felicetti, F. Piccinini, and P. Zingaretti, “Automatic faults detection of photovoltaic farms: solaris, a deep learning-based system for thermal images,” Energies, vol. 13, no. 24, p. 6496, 2020.

[17] M. Mayr, M. Hoffmann, A. Maier, and V. Christlein, “Weakly supervised segmentation of cracks on solar cells using normalized l p norm,” in 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1885–1889.

[18] J. Balzategui, L. Eciolaza, and N. Arana-Arexolaleiba, “Defect detection on polycrystalline solar cells using electroluminescence and fully convolutional neural networks,” in 2020 IEEE/SICE International Symposium on System Integration (SII). IEEE, 2020, pp. 949–953.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in International Conference on Machine Learning. PMLR, 2021, pp. 10 347–10 357.

[24] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, “Conditional positional encodings for vision transformers,” arXiv preprint arXiv:2102.10882, 2021.

[25] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, “Deepvit: Towards deeper vision transformer,” arXiv preprint arXiv:2103.11886, 2021.

[26] Y. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, “Tokens-to-token vit: Training vision transformers from scratch on imagenet,” arXiv preprint arXiv:2101.11896, 2021.

[27] D. Zhang, J. Wang, Z. Li, M. Yang, and Y. Wei, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” arXiv preprint arXiv:2102.12122, 2021.

[28] 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,” arXiv preprint arXiv:2103.14030, 2021.

[29] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International conference on machine learning. PMLR, 2015, pp. 448–456.

[30] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” arXiv preprint arXiv:1606.08415, 2016.

[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[32] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 781–10 790.

[33] C. Buerhop-Lutz, S. Deitsch, A. Maier, F. Gallwitz, S. Berger, B. Doll, J. Hauch, C. Camus, and C. J. Brabc, “A benchmark for visual identification of defective solar cells in electroluminescence imagery,” in European PV Solar Energy Conference and Exhibition (EU PVSEC), 2018.

[34] S. Deitsch, C. Buerhop-Lutz, E. Sovetkin, A. Steland, A. Maier, F. Gallwitz, and C. Riess, “Segmentation of photovoltaic module cells in uncalibrated electroluminescence images,” vol. 32, no. 4.

[35] S. Deitsch, V. Christlein, S. Berger, C. Buerhop-Lutz, A. Maier, F. Gallwitz, and C. Riess, “Automatic classification of defective photovoltaic module cells in electroluminescence images,” Solar Energy, vol. 185, pp. 455–468, Jun. 2019.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.

[37] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.

[38] W. Tang, Q. Yang, K. Xiong, and W. Yan, “Deep learning based automatic defect identification of photovoltaic cells using electroluminescence images,” Solar Energy, vol. 201, pp. 453–460, 2020.

[39] C. Ge, Z. Liu, L. Fang, H. Ling, A. Zhang, and C. Yin, “A hybrid fuzzy convolutional neural network based mechanism for photovoltaic cell defect detection with electroluminescence images,” IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 7, pp. 1653–1664, 2020.

[40] K. Simonoy and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2015.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
XIANGYING XIE received the B.S. degree in computer science and technology from Wuhan University of Science and Technology in 2003, and the M.S. degree in software engineering from Sichuan University. He is pursuing a PhD degree in Beihang University. He is currently a deputy general manager of photovoltaic cloud (e-commerce poverty alleviation) Business Department of State Grid e-commerce Co., Ltd. and general manager of State Grid new energy cloud Technology Co., Ltd. His research interests include the new energy industry Internet. He won the honor of “outstanding individual” in SASAC central enterprise innovation achievement exhibition.

HU LIU received the B.S. degree in mechanical engineering and automation, computer science and technology, and the M.S. degree in mechanical and electrical engineering from North China Electric Power University in 2003 and 2006. He is currently a deputy director of the Internet Department of State Grid Corporation of China. His research interests include power marketing, power grid production and energy services. He won the first prize of geographic information technology progress of China Geographic Information Industry Association, the special award of scientific progress of State Grid Corporation of China, and the first prize of China electric power innovation.

ZHIXIONG NA received the B.S. degree in thermal engineering from Changzhou University in 1996, the M.S. degree in computer application from University of Science and Technology of China in 2001, and the M.B.A degree from University of International Business and Economics in 2009. He is currently a Professor with State Grid e-commerce Co., Ltd. His research interests include industrial automation, new energy industrial Internet, etc.

XIN LUO received the B.S. degree from University of Science and Technology of China and the Master of electrical engineering, North China Electric Power University. He is currently Senior Engineer with State Grid e-commerce Co., Ltd. He has participated in a number of State Grid science and technology projects, National 863 plan, national key R & D plan and other projects. His research interests include industrial Internet, distributed generation technology application, etc.

DONG WANG received the M.S. degree from Beijing University of Posts and Telecommunications. He is currently Senior Engineer and the deputy director of Digital Technology Department of State Grid e-commerce Co., Ltd., Deputy Secretary General of central enterprise blockchain cooperation and innovation platform, and chairman of IEEE PES power information and communication blockchain subcommittee. His research interests include network security, intelligent operation and maintenance, blockchain, etc. He has won the second prize of Beijing Science and technology progress award, electric power Innovation Award (Special Award), etc.

BIAO LENG received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China in 2009, and the B.S degree in computer science and technology from National University of Defense Technology, Changsha, China in 2004. He is currently a Professor at the School of Computer Science and Engineering, Beihang University. He visited the Pennsylvania State University during 2011-2012, and the Hong Kong University of Science and Technology in 2012. His research interest includes artificial intelligent, and industrial internet.

***