remote sensing

MDPI

Article

DSFA-SwinNet: A Multi-Scale Attention Fusion Network for Photovoltaic Areas Detection

Shaofu Lin [1], Yang Yang [1], Xiliang Liu [1,*] and Li Tian [2]

[1] College of Computer Science, Beijing University of Technology, Chaoyang District, Beijing 100124, China; linshaofu@bjut.edu.cn (S.L.); yangyang@emails.bjut.edu.cn (Y.Y.)
[2] The Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing 100101, China; tianli@igsnrr.ac.cn
* Correspondence: liuxl@bjut.edu.cn

Abstract: Precise statistics on the spatial distribution of photovoltaics (PV) are essential for advancing the PV industry, and integrating remote sensing with artificial intelligence technologies offers a robust solution for accurate identification. Currently, numerous studies focus on the detection of single-type PV installations through aerial or satellite imagery. However, due to the variability in scale and shape of PV installations in complex environments, the detection results often fail to capture detailed information and struggle to scale for multi-scale PV systems. To tackle these challenges, a detection method known as Dynamic Spatial-Frequency Attention SwinNet (DSFA-SwinNet) for multi-scale PV areas is proposed. First, this study proposes the Dynamic Spatial-Frequency Attention (DSFA) mechanism, the Pyramid Attention Refinement (PAR) bottleneck structure, and optimizes the feature propagation method to achieve dynamic decoupling of the spatial and frequency domains in multi-scale representation learning. Secondly, a hybrid loss function has been developed with weights optimized employing the Bayesian Optimization algorithm to provide a strategic method for parameter tuning in similar research. Lastly, the fixed window size of Swin-Transformer is dynamically adjusted to enhance computational efficiency and maintain accuracy. The results on two PV datasets demonstrate that DSFA-SwinNet significantly enhances detection accuracy and scalability for multi-scale PV areas.

Keywords: high-resolution images; photovoltaic; swin-transformer; dynamic spatial-frequency attention

Academic Editor: Gabriele Bitelli

Received: 7 November 2024
Revised: 6 January 2025
Accepted: 17 January 2025
Published: 18 January 2025

Citation: Lin, S.; Yang, Y.; Liu, X.; Tian, L. DSFA-SwinNet: A Multi-Scale Attention Fusion Network for Photovoltaic Areas Detection. *Remote Sens.* **2025**, *17*, 332. https://doi.org/10.3390/rs17020332

1. Introduction

As global demand for renewable energy rises, solar photovoltaic (PV) technology, a pivotal form of renewable energy, has experienced remarkable growth worldwide. The National Energy Administration (NEA) of China reports in its latest construction update for 2023 that newly installed PV capacity of China hits a record 216.88 GW by December 2023, representing a 148% increase compared to the previous year [1]. These achievements herald the swift progression of the PV industry in the renewable energy sector and establish a robust foundation for the sustainable utilization of clean energy in the future.

However, the accelerated growth of photovoltaic technology, while a key catalyst in the evolution of the clean and renewable energy sector, has unintentionally given rise to a variety of issues, including land degradation, environmental pollution, and the

encroachment on arable land [2–4]. Driven by the goal of enhancing the environmental compatibility of PV technologies and reducing their potential detrimental effects on socio-ecological systems, the imperative for precise detection of PV areas has escalated. Currently, this process is beset by formidable challenges:

- The geographical distribution of PV installations exhibits significant unevenness, indicating a lack of real-time coordination strategies. This results in an inability to integrate fine-grained data effectively, which constrains the maintenance of the facilities and the assessment of their eco-efficiency;
- The morphological complexity and textural diversity of PV areas pose a significant challenge to accurately identify their detailed features;
- Existing methods focus on single spatial forms, overlooking the diverse scales of PV installations.

Conventional visual detection methods for PV areas frequently demand substantial resources, struggle to meet the need for rapid response within a short period, and face challenges in accurately detecting extensive areas. Consequently, the automatic and precise segmentation of PV areas through high-resolution remote sensing imagery (HRSI) is particularly urgent and has garnered broad interest from both industrial and academic researchers [5–7]. Current methods for large-scale PV areas detection predominantly rely on machine learning [8–10] and deep learning [11,12].

Machine learning methods primarily focus on the extraction of morphological features or artificially designed features [13]. These morphological features include shape, scale, spectral, and texture, which can be obtained through morphological operations such as binarization, expansion, and closed operations, etc. [14,15]. Following the feature extraction process, researchers construct shallow models such as RF, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) using manually designed features [16]. Malof et al. are the first to deploy automated remote sensing to assess distributed solar PV [17]. Chen et al. use raw spectral features, photovoltaic extraction indices, and topographic features as XGBoost classification features to extract time-series data of large-scale photovoltaic power plants for the first time from medium-resolution imagery [18]. Li et al. design the SolarFinder system, which combines the advantages of SVM and CNN with a linear regression method to achieve better true positives and true negatives in the PV array detection task [19]. Despite the success of these methods, manual feature design limits the generalization of these methods due to challenges in capturing imaging conditions and PV material properties [20]. Advancing deep learning in remote sensing provides adaptive feature extraction for improved large-scale PV areas detection.

Current deep learning detection methods for PV areas can be broadly categorized into four types, namely CNN [21,22], FCN [23,24], models derived from FCN (e.g., DeepLabV3+, SegNet, Unet, FPN, etc.) [25–27], and Visual Transformer [28].

CNN-based methods provide a technical foundation for the application of deep learning in the field of PV detection. Yu et al. develop DeepSolar, a CNN-based deep learning framework, and successfully create a high-fidelity solar deployment database covering the contiguous U.S. [29]. Castello et al. propose a CNN-based pixel-level image segmentation method for obtaining the location and size information of rooftop PV panels [30]. Following the emergence of FCN, it adopts convolutional layers instead of fully connected layers and generates feature maps with the same size as the input image through upsampling, supporting inputs of arbitrary size and enabling end-to-end pixel-level prediction. Yuan et al. are pioneers in proposing FCN model for the identification of distributed PV power plants [31]. Based on this, Sizkouhi et al. use the Mask-RCNN architecture to enhance the detection accuracy of PV plant boundaries [32]. While FCN improves the recognition accuracy of PV areas, it faces issues such as the loss of details in the feature maps and the inability of a fixed receptive field to capture the global contextual

relationships when dealing with small objects in complex scenes. To curtail detail loss and broaden the receptive field, researchers have turned to FCN-derived semantic segmentation models including DeepLabV3+ [33], SegNet [34], Unet [35], etc. Though the previous studies have provided insights into detecting PV areas, the limited receptive field of convolutional methods, constrained by their kernel size, impacts the capture of large-scale feature relationships [36–38]. In 2021, Dosovitskiy et al. initially propose Visual Transformer (ViT) model [39], which enables Transformer to process image data. Subsequently, a range of remote sensing methods integrating CNN and Transformer have emerged [40,41]. Chen et al. use a Trans-UNet-based architecture to analyze multi-year satellite data for PV variations in China [42]. Guo et al. propose the TransPV model by combining Unet and Visual Transformer, which further enhances the ability of the model to comprehend the global context [43].

On the other hand, to further improve detection performance and overcome the challenges posed by the heterogeneous textures and color features of PV areas, related studies have proposed various deep learning model enhancement strategies, including perceptual enhancement, positive and negative sample balancing, and multi-scale feature optimization.

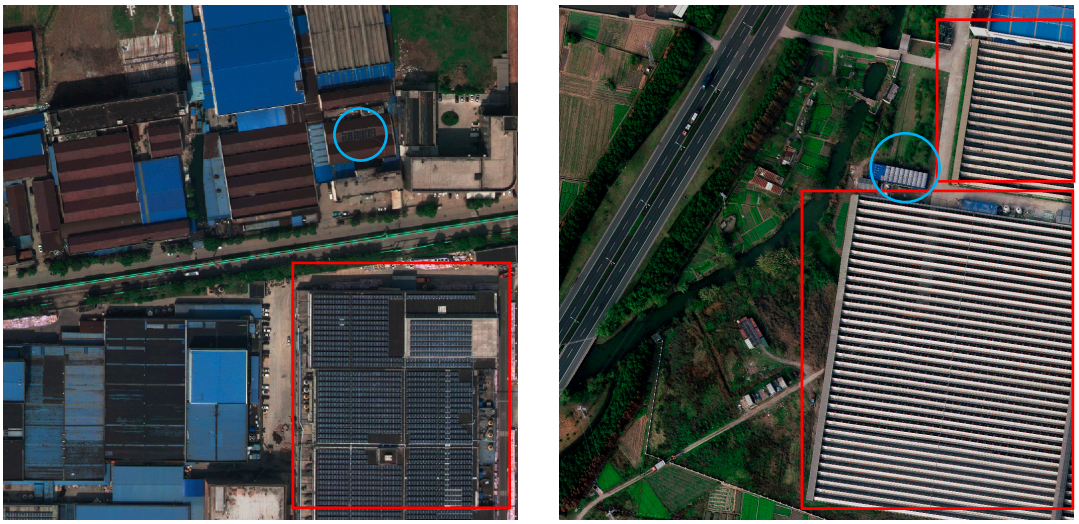
In the realm of perceptual enhancement, beyond the aforementioned explorations ranging from CNN to FCN-derived semantic segmentation models and extending to Visual Transformer approaches, several studies have explored how attention mechanisms influence the extraction of PV features. Hou et al. develop an Expectation Maximization Attention (EMA) module that utilizes clustering to enhance spatial feature capture [44]. Zhu et al. incorporate a Dual-Attention Module (DAM) to dynamically merge local features with their global dependencies [45].

To manage the balance between positive and negative samples, many studies have adjusted the loss function. Guo et al. introduce Focal loss for mining hard samples [43], while Zhu et al. add IoU loss to address the imbalance between foreground and background in positive and negative samples [45]. However, most hybrid loss functions are constructed without discussing the weight relationship of each loss, and their default weights may not be truly applicable to PV features.

Regarding multi-scale feature optimization, numerous studies have leveraged the Atrous Spatial Pyramid Pooling (ASPP) concept [46]. Tan et al. enhance the ASPP structure for the horizontal features of PV arrays, focusing on processing the contextual features output from the encoder [47]. Moreover, other researchers have explored ways to augment the multi-scale feature extraction capabilities of the model at the data and model training levels. Kleebauer et al. propose hyper-parameter tuning of the PV detection model to adapt it to images of various resolutions [48].

Despite the rich technical support provided by the aforementioned studies for PV areas detection, existing methods still exhibit limitations in global feature extraction, multi-scale feature modeling, and computational efficiency.

- As depicted in Figure 1, there is a significant scale difference between PV arrays and panels in images. While Swin-Transformer [49] mitigates computational complexity with windowing and hierarchies, its fixed window size limits the capture of multi-scale features, leading to an absence of internal multi-scale information within the model [50–52];
- In feature learning, existing enhancement strategies do not sufficiently account for the specific characteristics of different PV areas, making it difficult to dynamically adapt to the multi-scale and multi-textural nature of PV features. Furthermore, they lack interpretability regarding hyperparameter tuning;
- Challenges arise as higher image resolutions lead to longer feature sequences, slowing down attention computations and reducing efficiency.



(**a**) (**b**)

Figure 1. Examples of scale differences between PV arrays and PV panels are illustrated. Figure 1 (**a**,**b**) show size differences between PV arrays (rectangles) and panels (circles).

To address the issues mentioned above, this study proposes a multi-scale attention fusion network called Dynamic Spatial-Frequency Attention Swin-Transformer Network (DSFA-SwinNet). The contributions of this study are summarized as follows:

Firstly, to tackle the issue of high intra-class variation and inter-class resemblance in PV areas in HRSI, this study proposes a dynamic window size adjustment mechanism (DWA). This mechanism dynamically adjusts the feature window size based on the long-term token dependency mining mechanism of Swin-Transformer, and effectively captures both the local and global dependencies of multi-scale features in HRSI.

Secondly, this study introduces the Dynamic Spatial-Frequency Attention (DSFA) mechanism and the multi-scale feature refinement bottleneck structure Pyramid Attention Refinement (PAR). These methods significantly bolster the performance of the model by dynamically decoupling the spatial and frequency domains within multi-scale representation learning.

Lastly, this study puts forward the refined skip connection strategy and the depth-supervision-based Multi-Level Upsampling Head (MLUH) module, aiming to augment the representation learning capabilities for PV areas by refining the feature propagation mechanism.

The structure of this study is organized as follows: Section 2 details the datasets, explains the theoretical architecture, and outlines the step-by-step development process of DSFA-SwinNet; Section 3 details the datasets and evaluation criteria employed to assess the performance of the model and presents the experimental results; Section 4 addresses the limitations of the model and proposes potential future enhancements; Section 5 outlines the conclusions.

2. Materials and Methods

2.1. Materials

2.1.1. BDAPPV Dataset

This study employs the Google subset of the BDAPPV dataset [53], as shown in Figure 2. This subset contains 13,303 images from Google Earth with a spatial resolution of 0.1 and a resolution of 400 × 400 pixels.

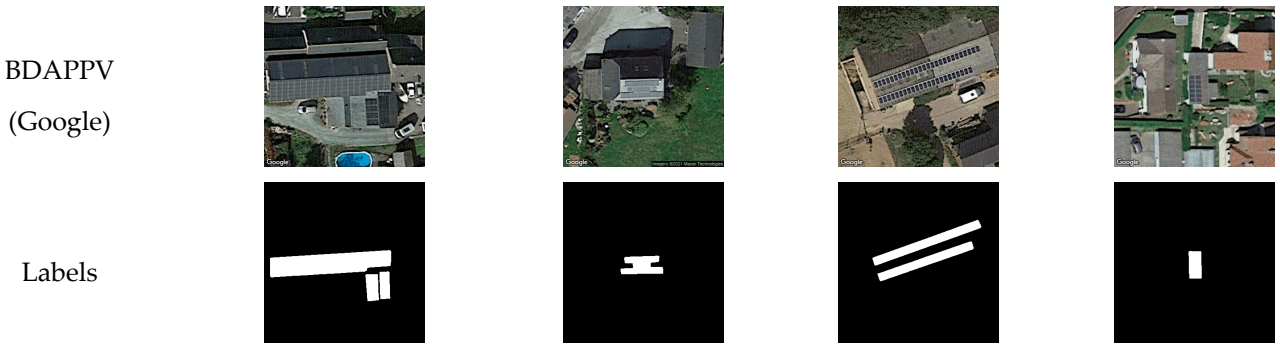


Figure 2. BDAPPV (Google) subset.

2.1.2. Jiangsu PV Dataset

To enhance PV sample diversity, this study employs the PV03 subset of a multi-resolution PV Dataset from Jiangsu Province, China [54], containing 2308 satellite and aerial images, as shown in Figure 3. With a spatial resolution of 0.3 m and image size of 1024 × 1024 pixels, the PV03 subset includes both rooftop and ground-based PV systems in areas like shrubland, grassland, cropland, saline-alkali land, and water surfaces.

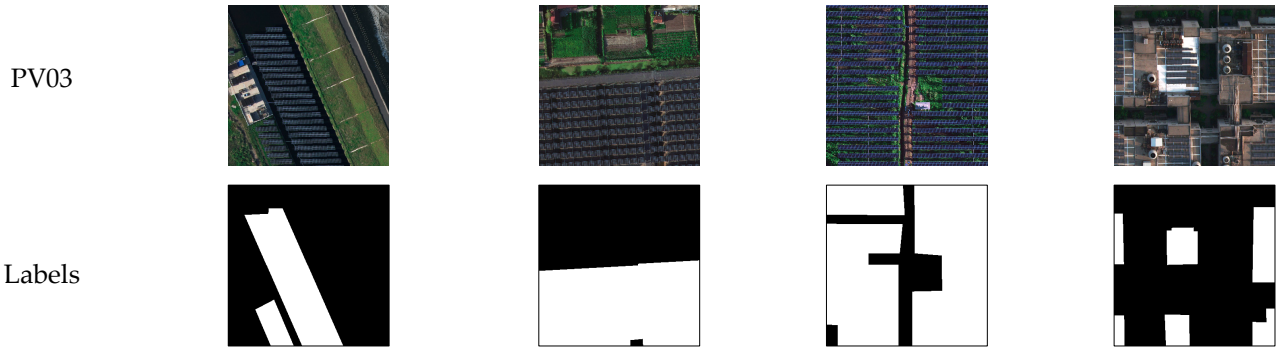


Figure 3. PV03 subset.

2.2. DSFA-SwinNet

As shown in Figure 4, the architecture of DSFA-SwinNet is outlined in this section, encompassing the encoder-decoder U-structure architecture, the DSFA mechanism, the PAR bottleneck structure, the refined skip connection strategy, and the hybrid loss function for training.

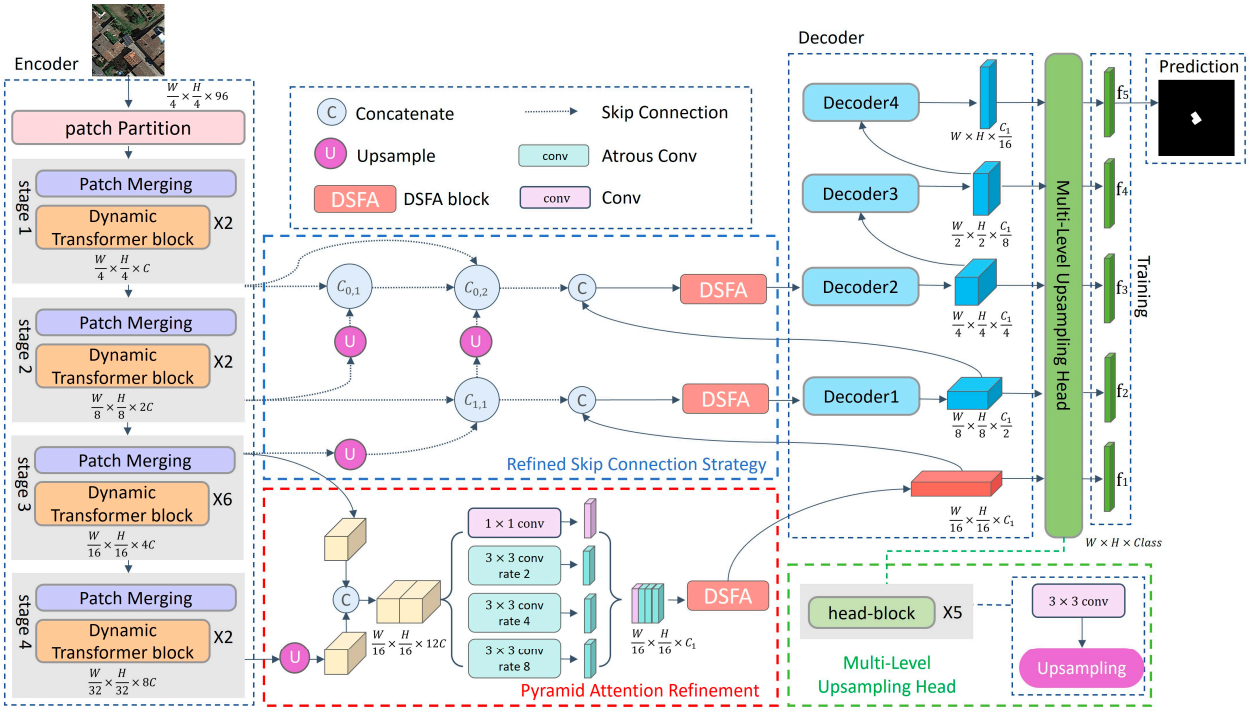


Figure 4. DSFA-SwinNet structure.

2.2.1. Swin-Transformer Based U-Model Architecture

DSFA-SwinNet employs Swin-Transformer as the encoder backbone and refers to the ideas of Ren et al. [55] to construct a dynamic window size adjustment mechanism that applies different windows to different channels to achieve multi-scale feature extraction at different levels.

The structure of the dynamic window size adjustment mechanism is depicted in Figure 5. The mechanism constructs a pyramid structure that divides the input feature maps and computes window attention in a distributed manner by utilizing multiple parallel branches, each equipped with distinct window sizes along the channel dimension. This algorithm effectively mitigates the constraints imposed by a fixed window size on the extraction of representational information in traditional vision Transformers, without incurring a substantial increase in computational load. Moreover, it notably enhances the performance of DSFA-SwinNet in the multi-scale representation learning of PV areas features.

The initial size of the input tensor of DSFA-SwinNet is noted as (H, W, C), and multiple patches of size ($\frac{H}{4}$, $\frac{W}{4}$, 96) are generated after window splitting. Each patch is considered as a token and fed into the downsampling layer of Swin-Transformer to deduce feature representations. DSFA-SwinNet utilizes each stage of Swin-Transformer as a layer in Unet encoder, the input tensor with shape (H_{cur}, W_{cur}, C_{cur}) will be downsampled to ($\frac{H_{cur}}{2}$, $\frac{W_{cur}}{2}$, $2C_{cur}$) after traversal through each layer, with the final encoder outputting feature maps with the shape ($\frac{H}{32}$, $\frac{W}{32}$, 8C). Furthermore, to ensure that DSFA-SwinNet can acquire multi-scale features, the output of each stage is stored in a cache, which is then leveraged in the decoder to execute the refined skip connection strategy.

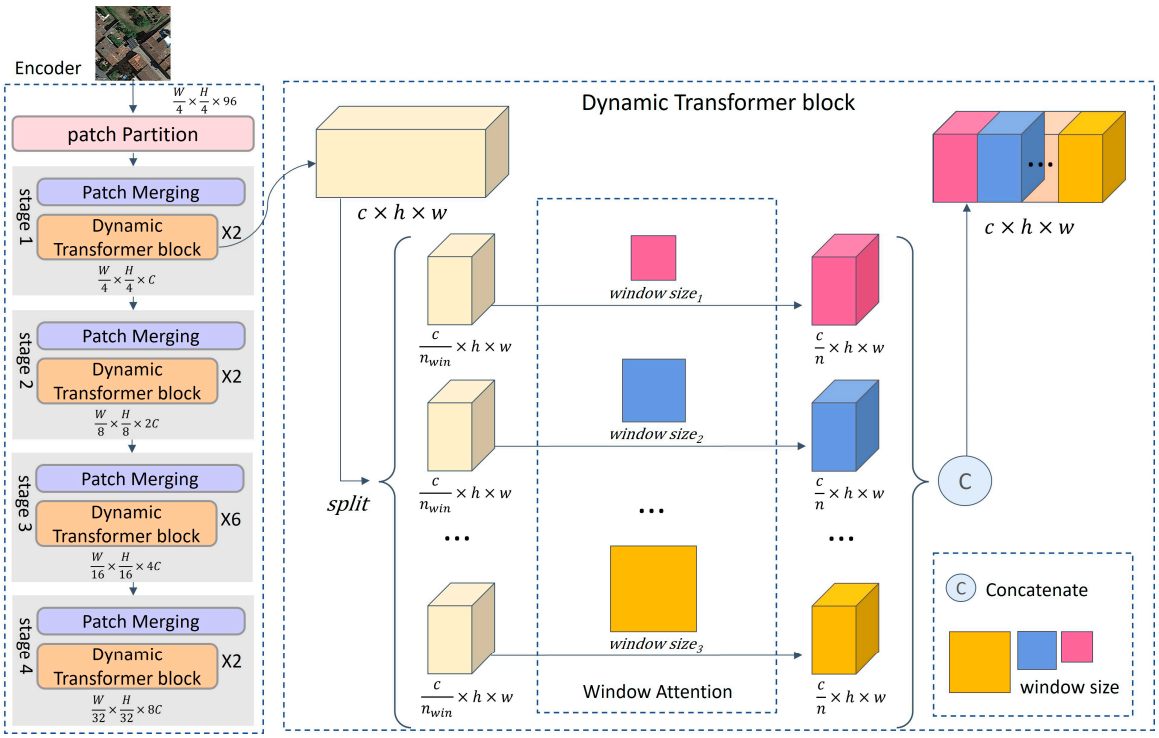


Figure 5. Structure of DWA mechanism.

In the decoder, the DecoderBlock is employed to progressively restore the spatial dimensions of the feature maps through upsampling. Each DecoderBlock is composed of one upsampling layer followed by two convolutional layers. After processing by DecoderBlock module, the input feature maps with shape ($\frac{H_{cur}}{2}$, $\frac{W_{cur}}{2}$, $2C_{cur}$) will be upsampled to (H_{cur}, W_{cur}, C_{cur}). Within the refined skip connection strategy, the output feature maps are integrated with the skip-connection outcomes at matching resolutions. These are then subjected to the DSFA mechanism to augment the multi-scale information embedded within the feature maps. The output feature maps from both the bottleneck structure and each decoder layer are cached, as indicated by the box called Multi-Level Upsampling Head in Figure 4. These feature maps are converted to feature maps of shape (H_{cur}, W_{cur}, $Class$) by 3 × 3 convolutional layers in the MLUH module, and then upsampled to the original image resolution. During training, the deeply supervised approach is adopted, employing output feature maps from the bottleneck structure and the last three decoder layers for auxiliary loss and the top layer for the primary prediction task. During inference, only output feature maps from the top decoder layer are utilized for the main loss. With this multi-scale and deep supervised learning framework, DSFA-SwinNet is more proficient in capturing feature information at various scales.

2.2.2. Dynamic Spatial-Frequency Attention

As the network depth increases, original feature maps evolve from capturing low-level to high-level features, potentially leading to the loss of details. To overcome this issue, this study has engineered the DSFA mechanism to achieve fine-grained PV areas detection by dynamically decoupling multi-scale representation learning in the spatial and frequency domains. The architecture of the DSFA mechanism is delineated in Figure 6. Within the DSFA mechanism, the input feature maps X_i, shaped as (C_{cur}, H_{cur}, W_{cur}), are initially directed into the concurrent pathways of channel-space attention and channel-frequency attention.

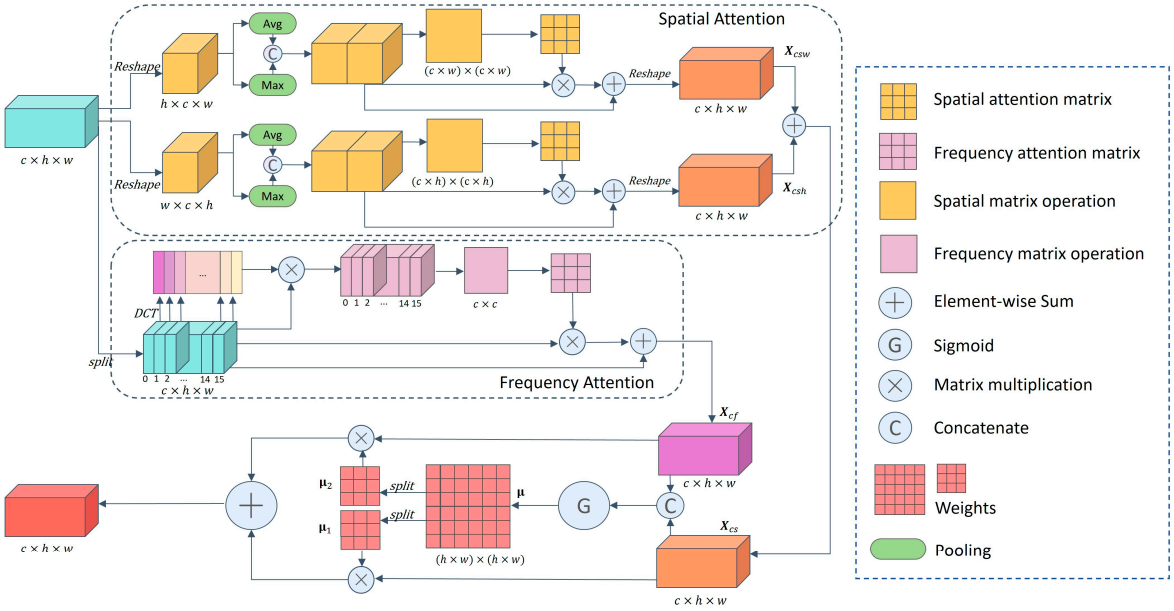


Figure 6. The structure of the DSFA mechanism.

In the channel-space attention path, features from the H_{cur} and W_{cur} dimensions are independently reconstructed, the processing results of the maximum pooling and average pooling operations are concatenated, and spatial attention weight maps are dynamically generated through the 3 × 3 convolutional layer combined with the Sigmoid function. These maps are then multiplied and summed with the reconstructed features to learn the importance of the feature maps at different spatial locations. Finally, the feature maps are reshaped into the shape of (C_{cur}, H_{cur}, W_{cur}) to obtain the outputs X_{csh} and X_{csw} for this path.

In the channel-frequency attention path, the Discrete Cosine Transform (DCT) is harnessed to craft a frequency attention mechanism. Figure 7 illustrates the application process of DCT. According to the experimental findings of Qin et al. [56], a predefined set of 16-band frequency position indices is established to direct the creation of the DCT filter, as represented by $freq_x$ and $freq_y$ in Figure 7. The input feature maps are split along the channel dimension into 16 parts. DCT is applied according to each frequency index, resulting in a corresponding set of DCT filters. These filters are sized to match the spatial dimensions of the input features, ensuring comprehensive representation of each feature dimension across different frequency bands. The input features are then translated into the frequency domain through an element-wise multiplication with the DCT filters. The mapping relationship from the spatial domain to the frequency domain is as follows:

$$F_i[u_i, v_i] = \lambda(u_i)\lambda(v_i) \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} T_i[x, y] \cos[\frac{\pi}{H}(x + \frac{1}{2})u_i] \cos[\frac{\pi}{W}(y + \frac{1}{2})v_i] \quad (1)$$

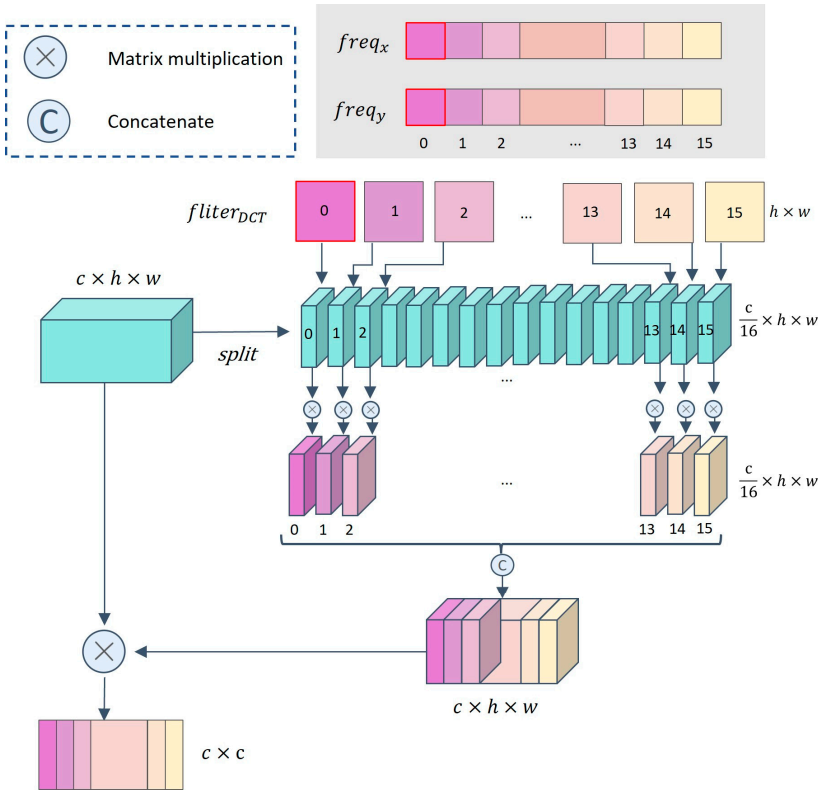


Figure 7. DCT Application Process.

For the *ith* tensor slice X_i (i = 0, 1, 2, …, 15), $F_i[u_i, v_i]$ represents the 2D frequency space coefficients, and $T_i[x, y]$ represents the tensor values in the spatial domain. H and W are the resolutions of the input tensor, while u_i and v_i are the frequency coordinates, defined as follows:

$$u_i = freq_x^i \tag{2}$$

$$v_i = freq_y^i \tag{3}$$

$\lambda(u_i)$ and $\lambda(v_i)$ are the regularization coefficients, and their definitions are shown in Equations (4) to (5).

$$\lambda(u_i) = \begin{cases} \sqrt{\frac{1}{H}}, u_i = 0 \\ \sqrt{\frac{2}{H}}, \text{else} \end{cases} \tag{4}$$

$$\lambda(v_i) = \begin{cases} \sqrt{\frac{1}{W}}, v_i = 0 \\ \sqrt{\frac{2}{W}}, \text{else} \end{cases} \tag{5}$$

Subsequently, the 3 × 3 convolutional layer combined with the Sigmoid function dynamically produces the frequency-attention weight maps. They are then employed to combine and aggregate with the reconstructed feature maps, culminating in the path outputs X_{cf}.

The process is advanced by utilizing weight fusion to integrate the outputs from both attention paths. X_{csh} and X_{csw} are concatenated as X_{cs}, and then the Sigmoid function

is employed to adjust features of X_{cs} and X_{cf} to derive the new weights of attention μ after the space-frequency interaction. The recalibrated weights μ are then split into weights μ_1 and μ_2 corresponding to the channel-space and channel-frequency, and a dot-multiplication operation is performed with the feature maps X_{cs}, X_{cf} by element, respectively. The resultant weighted feature maps X_i' are produced. The feature maps that have been subjected to the DSFA mechanism exhibit an enhanced representation of multi-scale information. This multidimensional feature fusion capability allows DSFA-SwinNet to concentrate on and capitalize on key features more effectively, while filtering out irrelevant background noise.

2.2.3. Pyramid Attention Refinement

Previous research [57] frequently connects the output from the full Swin-Transformer encoder directly to the decoder, neglecting the integration of intermediate features. This may cause disruptions in the continuous alignment of PV array features or lead to smaller PV panels being overlooked in the imagery.

To bolster the capacity of the model to assimilate the intrinsic details within feature maps, as depicted by the box called Pyramid Attention Refinement in Figure 4, this study incorporates the PAR bottleneck structure, which draws inspiration from the ASPP module. The feature fusion technology of the PAR bottleneck structure is elucidated in Figure 8, showcasing the integration of the output feature maps from the final two layers of the encoder ($\frac{H}{16}\times\frac{W}{16}$, $\frac{H}{32}\times\frac{W}{32}$) via upsampling and concatenation. Pyramidal parallel pathways are established by employing 1 × 1 convolutional layers intertwined with 3 × 3 convolutional layers possessing diverse dilation rates [2,4,8], thereby encompassing a spectrum of receptive field sizes to capture multi-scale contextual insights of the input features. Thereafter, the conjoined features are steered into the DSFA mechanism designed to distill the intrinsic spatial and frequency bi-dimensional information. The PAR bottleneck structure mitigates the erosion of profound feature information and guarantees the accuracy of PV area extraction throughout the sampling phase on DSFA-SwinNet.

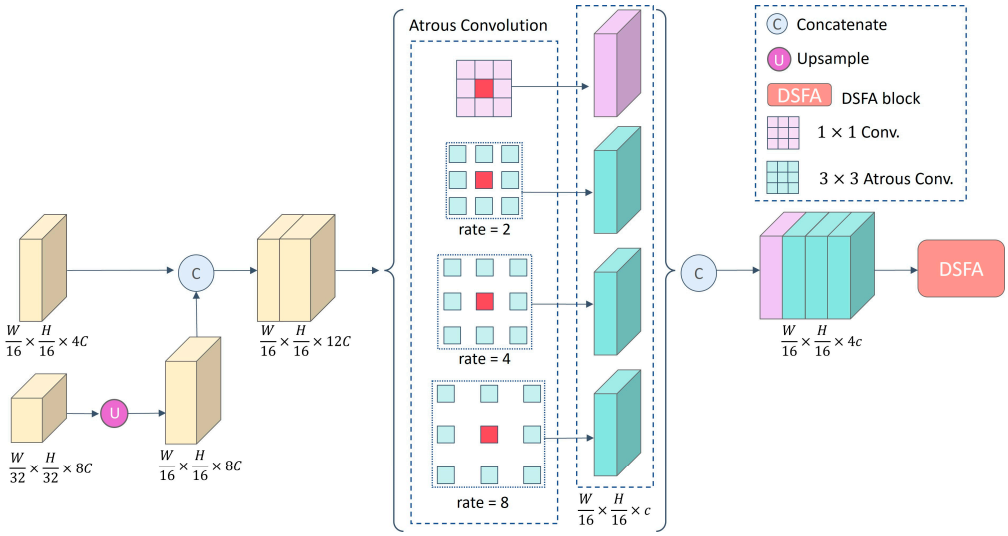


Figure 8. Structure of PAR bottleneck structure.

2.2.4. Refined Skip Connection Strategy

As delineated by the box called Refined Skip Connection Strategy in Figure 4, three intermediary units are introduced into the network architecture, designed to nurture the interplay and conveyance of information between multi-scale features via dense

interconnections, including not only the traditional skip connections but also the integration of cross-layer features.

Within the encoder, the output feature maps ($\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$) of each downsampled layer except the last layer of the encoder are temporarily stored, which are denoted as $X_{0,0}$, $X_{1,0}$, and $X_{2,0}$, respectively, and the intermediate feature maps, $X_{0,1}$, $X_{1,1}$, and $X_{0,2}$, are acquired through the dense connectivity paths of the intermediate units. For the feature maps emanating from the deeper layers, an initial upsampling operation is requisite, followed by a serial connection of feature maps of equivalent resolution. The connectivity (6)–(8) for the intermediate feature maps are as follows:

$$X_{0,1} = X_{0,0} \oplus \text{Upsample}(X_{1,0}) \tag{6}$$

$$X_{1,1} = X_{1,0} \oplus \text{Upsample}(X_{2,0}) \tag{7}$$

$$X_{0,2} = X_{0,1} \oplus \text{Upsample}(X_{1,1}) \tag{8}$$

The output feature maps processed by the PAR bottleneck structure are denoted as $X'_{3,0}$, and the skip-connections (9)–(11) of the decoder are as follows:

$$X_{2,1} = X_{2,0} \oplus X'_{3,0} \tag{9}$$

$$X_{1,2} = X_{1,1} \oplus X_{1,0} \oplus X'_{2,1} \tag{10}$$

$$X_{0,3} = X_{0,0} \oplus X_{0,1} \oplus X_{0,2} \oplus X'_{1,2} \tag{11}$$

where $X'_{2,1}$, $X'_{1,2}$ are the feature maps of $X_{2,1}$, $X_{1,2}$, respectively after processing by the sampling module on the decoder.

2.2.5. Loss Function

In DSFA-SwinNet, the PV area detection task is framed as a binary classification problem, with the goal of categorizing each pixel in the image as either "PV area" or "non-PV area". However, this task is characterized by a significant imbalance in the proportion of positive to negative samples within the labeling. Table 1 presents the sample ratios for the BDAPPV Rooftop PV dataset and the Jiangsu PV dataset.

Table 1. Ratio of PV area pixels to non-PV area pixels in the experimental dataset.

Dataset	Ratio (PV:Not PV)
BDAPPV (Google)	1:23.09
BDAPPV (IGN)	1:75.43
PV01	1:1.32
PV03	1:0.89

To tackle the issues arising from imbalanced samples, this study comprehensively considers three different loss functions: the Weighted Binary Cross-Entropy loss (WBCE), the Dice loss, and the Lovasz-Softmax loss [58]. The optimal weights for these functions are determined through theoretical analysis and experiments.

The function equations of WBCE loss are shown in (12)–(14):

$$L_{WBCE} = -\frac{1}{N}\sum_{n=1}^{N}(\omega_0 y_n \log \hat{y}_n + \omega_1(1-y_n)\log(1-\hat{y}_n)) \tag{12}$$

where y_n is the actual value of the *nth* pixel, $\hat{y}_n$ is the predicted value of the *nth* pixel, N is the total number of pixels, ω_0 and ω_1 are the weights of the positive and negative classes, respectively, and the calculation expressions are as follows:

$$\omega_0 = \frac{N}{y_n} \tag{13}$$

$$\omega_1 = \frac{N}{N - y_n} \tag{14}$$

The function equation of Dice loss is shown in (15):

$$L_{Dice} = 1 - \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{15}$$

where y represents the actual value of the *nth* pixel and $\hat{y}$ denotes the predicted value of the nth pixel.

For the feature map S_i (i = 1, 2, 3, 4, 5) produced by the bottleneck structure and the decoder, the loss function is expressed as:

$$Loss_i = \alpha L_{WBCE} + \beta L_{Dice} + \gamma L_{LS} \tag{16}$$

where α, β, and γ represent the weight parameters for L_{WBCE}, L_{Dice}, and L_{Ls}, respectively.

The overall loss function for DSFA-SwinNet training is as follows:

$$Loss = \sum W_i \times Loss_i \tag{17}$$

where W_i (i = 1,2,3,4,5) corresponds to f_i (i = 1,2,3,4,5) in Figure 4, respectively.

3. Results

3.1. Experimental Setup

All experiments are conducted on a server equipped with 32 GB of RAM and an NVIDIA Quadro P6000 GPU with 16 GB of memory.

3.2. Evaluation Metrics

This study utilizes Recall, Precision, F1 Score, and IoU as evaluation metrics, which are defined as shown in Equations (18) to (21) below.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FN + FP} \tag{20}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{21}$$

where *TP* represents the proportion of PV area pixels that are accurately detected, *TN* indicates the proportion of background pixels that are accurately classified, *FP* corresponds to the proportion of background pixels mistakenly identified as PV area, and *FN* indicates the proportion of PV area pixels misclassified as background.

3.3. Preprocessing and Parameterization

3.3.1. Data Preprocessing

The experimental models accept inputs of sizes that are powers of 2. Considering the original data sizes of the two datasets, a 256 × 256 resolution is chosen as the image input size. The HRSI data preprocessing involves standardizing the PV area labels, cropping the images, and filtering out samples with background pixel counts below 10% or above 90%. For the Google subset, images are directly cropped at the center point. For the PV03 subset, 1024 × 1024 images are cropped into 16 non-overlapping patches. This study applies horizontal and vertical flips for data augmentation and divides the datasets into training, validation, and testing sets in the ratios of 60%, 20%, and 20%, respectively. Detailed information is given in Table 2.

Table 2. Division of the datasets.

Dataset	Input Size	Original Data	Processed Data	Training Set	Validation Set	Test Set
BDAPPV (Google)	256 × 256	13,303	13,302	7981	2660	2661
PV03	256 × 256	2308	11,593	6955	2318	2320

3.3.2. Hyperparameter Optimization

This study employs the Ray Tune automated tuning tool [59], combined with Grid Search and Bayesian Optimization, to hyperparameter tune DSFA-SwinNet. Based on the initial values, the hyperparameters are fine-tuned, including training batch (*batchsize*), learning rate (*lr*, *momentum*), and loss function weights (*flooding*, α, β, γ, and loss weights of the feature maps produced by the bottleneck structure and each layer of the decoder [W_1, W_2, W_3, W_4, W_5]) in turn. Among them, flooding comes from the study of Ishida et al. [60], which aims to prevent training overfitting.

The search space for hyperparameters is shown in Table 3. A total of 1250 instances are randomly selected from the training set of the Google subset to create a small dataset for hyperparameter search experiments. The dataset is subsequently split into training and validation sets in a 75%/25% ratio, resulting in 1000 instances for training and 250 for validation. To ensure the adequacy of parameter exploration and optimization effect, the number of rounds of each search is set to 30 epochs based on the results of the pre-experimental runs of the experimental dataset. The convergence speed and IoU are comprehensively considered to select the optimal hyperparameters, and the initial values are updated.

Table 3. Hyperparameter search initial values, algorithms, and search space definitions. Grid search is denoted as GS and Bayesian optimization is denoted as BO.

Param	Search Space	Algorithm	Initial Values
batchsize	{2,4,8,16}	GS	4
flooding	[0,1]	BO	0.4
[W_1, W_2, W_3, W_4, W_5]	[0,1]	BO	{1,1,1,1,1}
α, β, γ	[0,1]	BO	{1,1,1}
lr	[0.0001,0.1]	BO	0.0001
momentum	[0.1,0.9]	BO	0.9

Figure 9 presents the correlation between different values of hyperparameters and IoU during the training process. Additionally, variance analysis evaluates the sensitivity of IoU to each hyperparameter, with the results shown in Figure 10. Among the 12 hyperparameters of interest, W_4, W_1, *lr*, and W_5 are ranked by their significant impact on IoU, indicating that the deeply supervised training method effectively enhances the ability

of the model to learn multi-scale features. Figure 9d–g further demonstrate that although *flooding*, α, β, and γ do not have a significant direct impact on IoU, their adjustments help accelerate the convergence speed of the model, and in some cases, there are local optimal intervals.

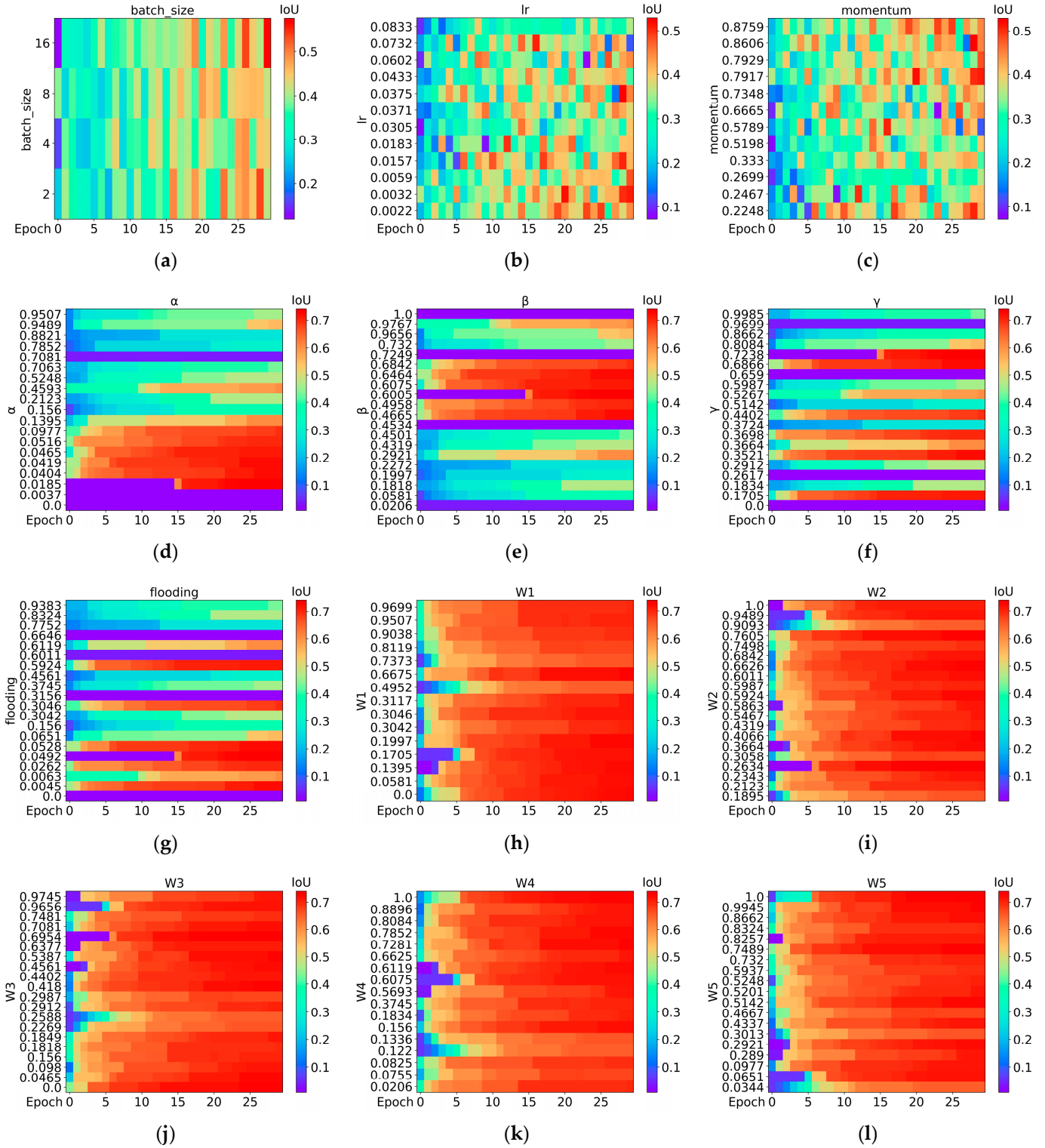


Figure 9. Heat map of hyperparameter optimization experiment. The horizontal axis represents epoch, the vertical axis represents the hyperparameter value, and the color bar from purple to red represents the size of IoU. (**a**) training batch (*batchsize*), (**b**,**c**) learning rate (*lr*, *momentum*), (**d**–**g**) loss function weights (α, β, γ, *flooding*), and (**h**–**l**) loss weights of the feature maps (W_1, W_2, W_3, W_4, W_5).

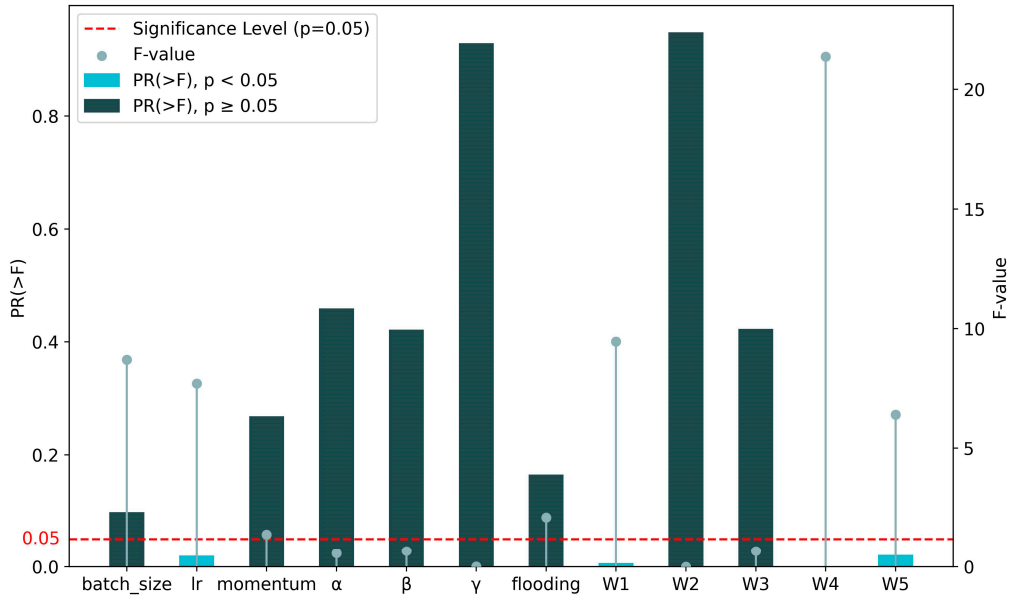


Figure 10. Sensitivity of IoU to Each Hyperparameter Based on Variance Analysis. The left vertical axis represents PR(>*F*), which is the *p*-value indicating the probability of obtaining an *F*-statistic at least as extreme as the one observed, under the null hypothesis. A lower PR(>*F*) value suggests a stronger relationship between the hyperparameter and IoU. The right vertical axis represents the corresponding *F* values, which measure the ratio of variance explained by each hyperparameter relative to the residual variance. The red dashed line marks PR(>*F*) = 0.05, serving as a threshold for statistical significance. Values below this threshold indicate that the corresponding hyperparameter has a significant impact on IoU.

3.3.3. Experimental Parameter Settings

During training, the Stochastic Gradient Descent (SGD) optimizer is utilized along with Cosine Annealing and Label Smoothing techniques, and the model is trained for 100 epochs to obtain the best parameters. Table 4 shows the hyperparameters of the training process.

Table 4. Training hyperparameters.

Param	Value
batchsize	16
lr	0.0030
flooding	0.0528
momentum	0.8000
weight_decay	0.0001
α	0.0419
β	0.6464
γ	0.6866
W_1	0.0000
W_2	0.6626
W_3	0.4180
W_4	1.0000
W_5	1.0000

3.4. Comparison Evaluation

To further validate the effectiveness and practicality of DSFA-SwinNet for PV area extraction, this study compares it with several well-known segmentation methods, including Swin-Unet [61], TransUnet [62], ACCoNet [63], Unet [35], MP-ResNet [64], DeepLabV3+ [33], MAResU-Net [65], CMTFNet [66], LETNet [67], PIDNet-L [68] and BASNet [69]. Both DSFA-SwinNet and Swin-Unet utilize the Swin Transformer as their backbone encoder, and they are similar to TransUnet, which combines Transformer and CNN. All three architectures can effectively handle long-term dependencies. Unet and DeepLabV3+ have been employed for architectural segmentation in HRSI, and their performance and robustness have been well proven in complex environments. MP-ResNet utilizes parallel multiscale branches to capture semantic context effectively, significantly expanding its valid receptive fields and enhancing the embedding of locally discriminative features. ACCoNet and MAResU-Net enhance the segmentation performance of the model in complex remote sensing scenes by integrating an attention mechanism to capture contextual information. CMTFNet specializes in leveraging multiscale features and graph structures to improve segmentation accuracy while maintaining high computational efficiency. BASNet improves the accuracy of edge segmentation by designing a bilateral self-attention module to effectively capture global contextual information. LETNet constructs the Lightweight Dilated Bottleneck (LDB) module and the Feature Enhancement (FE) module to enhance the ability to capture local feature details. PIDNet-L is a novel three-branch network that efficiently parses detailed, contextual, and boundary information, leveraging boundary attention to guide the fusion of detailed and context branches. Through comparison with these established methods, the performance and potential of DSFA-SwinNet in extracting PV areas from HRSI will be further assessed, while offering valuable insights and guidance for both research and practical applications in this domain. To ensure a fair comparison, each model is configured with identical training hyperparameters and the same loss function as DSFA-SwinNet. Table 5 provides the specific parameter configurations for each model.

Table 5. Model parameter setting. Number of Classification Head refers to the number of prediction results from different layers or branches of the model that are involved in the loss function calculation during training.

Model	Input Size	Number of Classification Head
Swin-Unet	256 × 256	1
TransUnet	256 × 256	1
ACCoNet	256 × 256	5
Unet	256 × 256	1
MP-ResNet	256 × 256	2
DeepLabV3+	256 × 256	2
MAResU-Net	256 × 256	1
CMTFNet	256 × 256	1
BASNet	256 × 256	3
LETNet	256 × 256	1
PIDNet-L	256 × 256	3
DSFA-SwinNet	256 × 256	5

Table 6 displays the quantitative comparison results on the Google subset. BASNet achieves the highest Precision (86.99%) with its multi-scale residual refinement module (RRM) and ranks second in F1 (90.32%) and IoU (83.41%). Compared to BASNet, DSFA-SwinNet has higher F1 and IoU values of 0.12% and 0.09%, respectively. ACCoNet, despite recording the top Recall value of 98.47% through its adjacent context coordination

modules (ACCoM), lags behind the other models in Precision (44.34%), F1 (58.48%), and IoU (43.95%). Compared to other models, DSFA-SwinNet excels in PV area segmentation, ranking as the best performer in F1 (90.44%) and IoU (83.50%), and is only 1.91% lower than ACCoNet in Recall (96.56%).

Figure 11 provides an example extraction on the Google subset. In Figure 11a, under various lighting conditions, Swin-Unet, TransUnet, ACCoNet, MP-ResNet, LETNet, and CMTFNet struggle to precisely delineate the PV areas within shadows. When the color of the PV areas is similar to the background, in Figure 11b, all models except DeepLabV3+ and DSFA-SwinNet exhibit varying degrees of missed detections. In Figure 11c–f, where the PV areas are darker than the surroundings, Swin-Unet, ACCoNet, and CMTFNet frequently misclassify roofs, walls, and glass as PV areas. Moreover, when addressing the scenario in Figure 11f with narrow gaps, DSFA-SwinNet displays superior boundary segmentation capabilities compared to the comparatively high-performing TransUnet, DeepLabV3+, and MAResU-Net. Overall, DSFA-SwinNet shows better robustness and accuracy in segmenting diverse PV shapes.

Table 6. Quantitative comparison (%) of the BDAPPV (Google) dataset.

Model	Precision	Recall	F1	IoU
Swin-Unet	72.91	89.23	78.64	67.05
TransUnet	<u>86.73</u> [1]	94.67	89.8	82.81
ACCoNet	44.34	**98.47**	58.48	43.95
Unet	82.31	95.45	87.62	79.47
MP-ResNet	81.52	94.48	86.62	78.03
DeepLabV3+	86.24	95.82	90.28	83.25
MAResU-Net	84.91	95.42	89.86	82.55
CMTFNet	80.80	94.56	86.04	77.33
BASNet	**86.99** [2]	95.26	<u>90.32</u>	<u>83.41</u>
LETNet	82.50	95.45	87.73	79.40
PIDNet-L	78.69	94.58	84.83	75.25
DSFA-SwinNet	85.97	<u>96.56</u>	**90.44**	**83.50**

[1] Underlined text highlights the second-best value in each column. [2] Bold text marks the highest value in each column.

Table 7 displays the quantitative comparison results on the PV03 subset. On the PV03 subset with more balanced positive and negative samples, DSFA-SwinNet ranks the top metrics for F1 (95.57%) and IoU (92.00%), places second for Recall (96.42%), and is third for Precision (95.24%). In addition, compared to the quantitative results on the BDAPPV dataset, ACCoNet shows a significant improvement in performance on the PV03 subset, which contains a balanced combination of positive and negative samples. This suggests that the impact of sample proportion on the performance of ACCoNet may be more pronounced than that on the performance of DSFA-SwinNet.

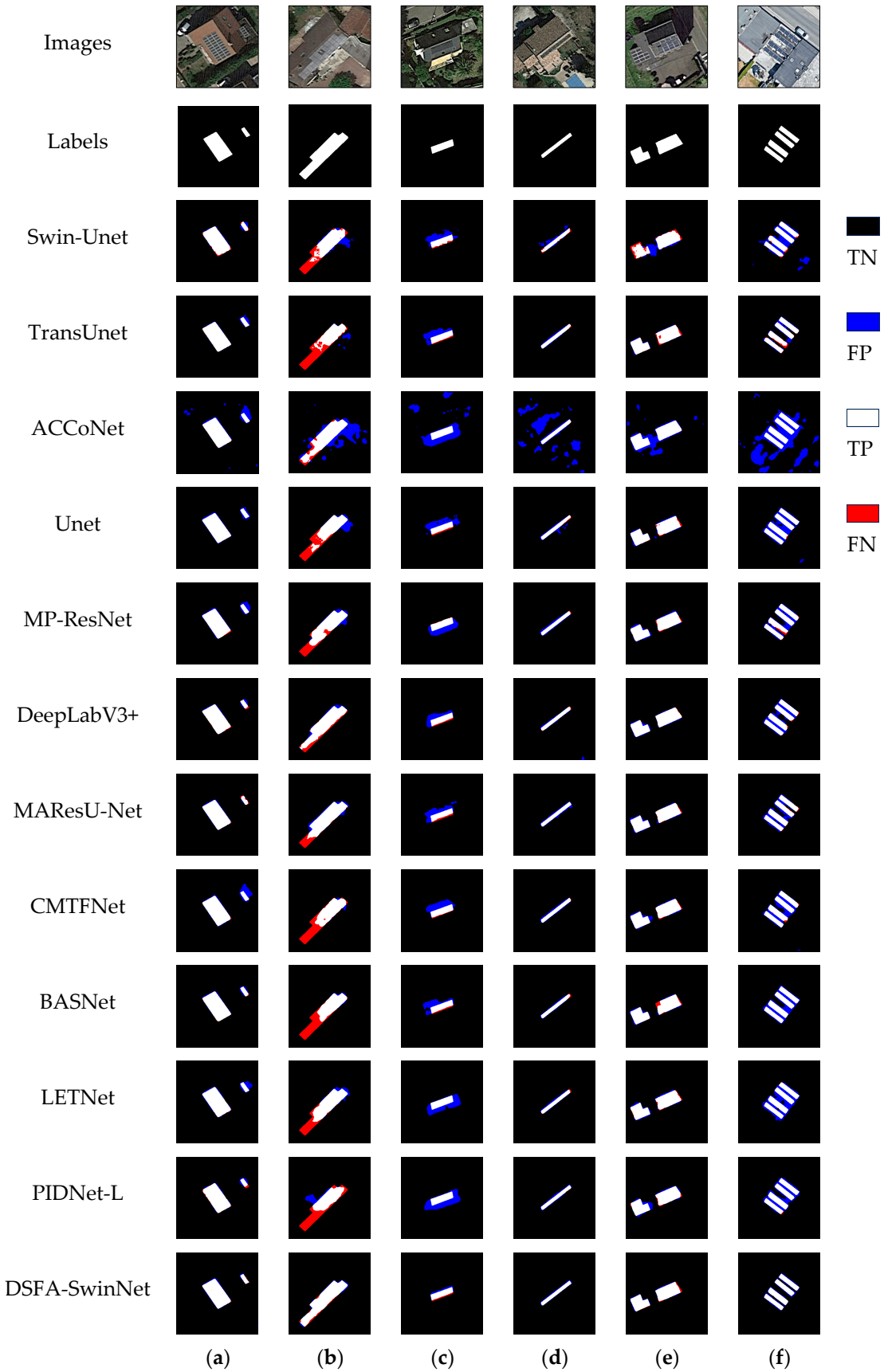


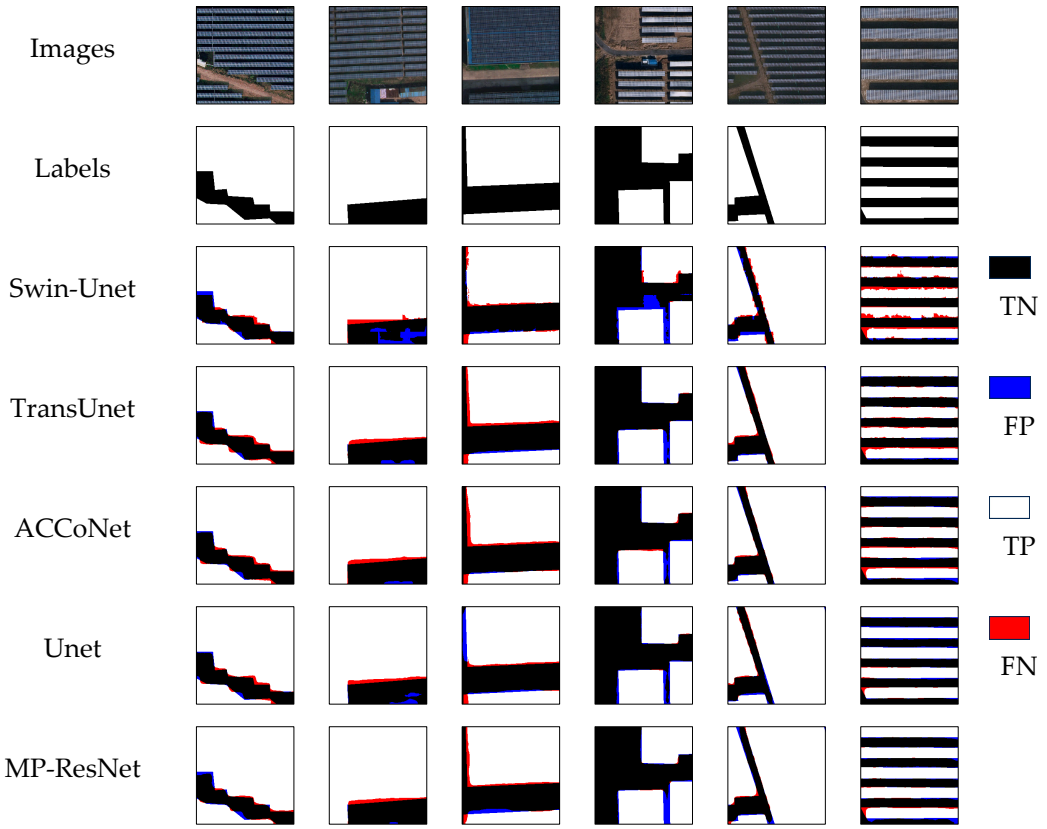
Figure 11. Example of BDAPPV (Google) subset detection. False positives and false negatives are shown in red and blue, respectively. (**a**–**f**) represent the results of different test cases.

Table 7. Quantitative comparison (%) of the Jiangsu PV dataset (PV03).

Model	Precision	Recall	F1	IoU
Swin-Unet	94.10	94.67	94.00	89.43
TransUnet	95.15	96.36	95.52	91.84
ACCoNet	<u>95.74</u> [1]	95.65	95.43	91.72
Unet	95.17	96.14	95.37	91.64
MP-ResNet	94.86	96.12	95.48	91.76
DeepLabV3+	94.69	**96.81**	95.52	91.81
MAResU-Net	95.27	96.25	95.55	91.89
CMTFNet	94.57	96.05	94.95	91.02
BASNet	**95.83** [2]	95.69	<u>95.56</u>	<u>91.92</u>
LETNet	94.85	95.94	95.39	91.62
PIDNet-L	94.52	95.62	94.68	90.65
DSFA-SwinNet	95.24	<u>96.42</u>	**95.57**	**92.00**

[1] Underlined text highlights the second-best value in each column. [2] Bold text marks the highest value in each column.

Figure 12 provides an example of the extraction on the PV03 subset. In Figure 12b, all models except DeepLabV3+, MAResU-Net, and DSFA-SwinNet incorrectly classify plants and roofs with colors similar to the PV areas as PV panels. In Figure 12d, DSFA-SwinNet is the only one, apart from the others, that accurately identifies the PV gaps between the panels. Furthermore, DSFA-SwinNet also excels over other models in segmenting boundaries, as demonstrated in Figure 12a,c,e. These results collectively indicate that DSFA-SwinNet possesses higher accuracy and refinement in the task of segmenting diverse types of PV samples.



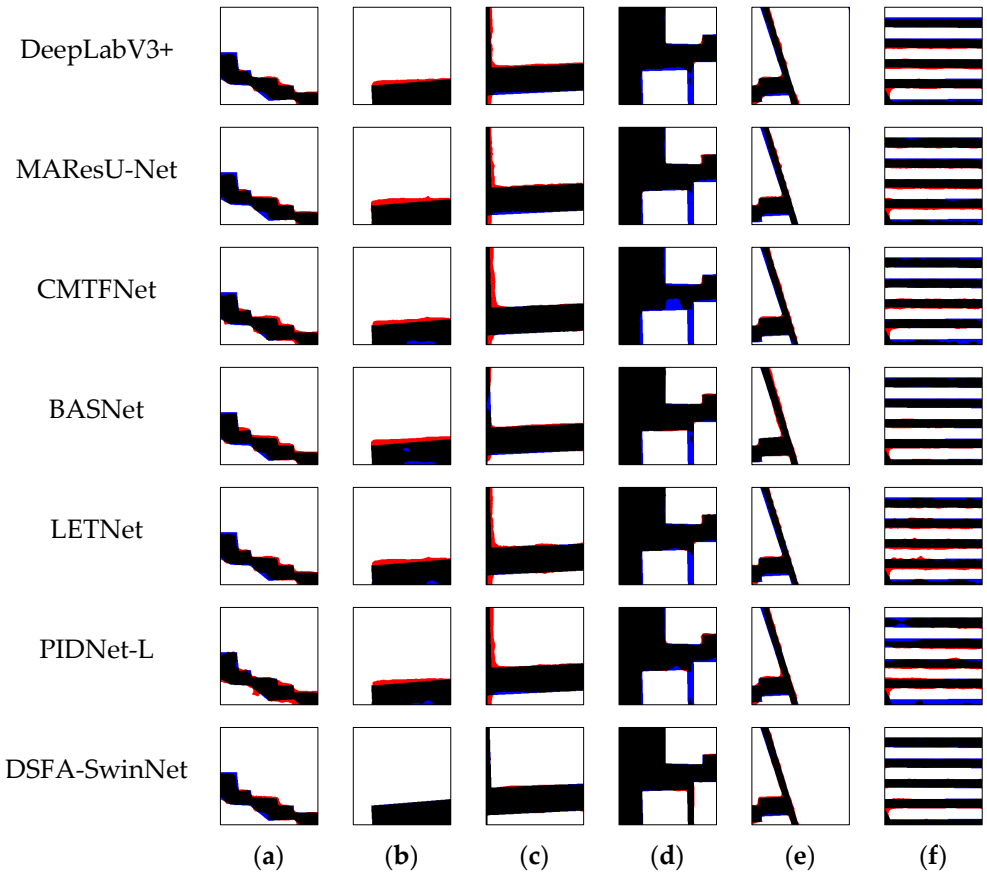


Figure 12. Example of PV03 subset detection. False positives and false negatives are shown in red and blue, respectively. (**a**–**f**) represent the results of different test cases.

3.5. Complexity Analysis

The time complexity of DSFA-SwinNet depends on the time complexity of the Window Multi-Head Self-Attention in the backbone network. After incorporating the DWA mechanism, assuming the input tensor format is (*C*, *H*, *W*), where M_i (i = 0, 1, …, K) represents the size of the *ith* window, the time complexity of self-attention for each window size in each encoder layer is $O(M_i^2 \frac{C^2}{K})$. Therefore, the time complexity within each window is:

$$\sum_{i=0}^{K-1} O\left(\left(\frac{H}{M_i} \times \frac{W}{M_i}\right) \times M_i^2 \frac{C^2}{K}\right) = \sum_{i=0}^{K-1} O\left(\frac{HW}{M_i^2} \times M_i^2 \frac{C^2}{K}\right) = O(HWC^2) \tag{22}$$

The time complexity for cross-window computation is:

$$\sum_{i=0}^{K-1} O((\frac{H}{M_i} \times \frac{W}{M_i}) \times M_i^4 \frac{C^2}{K}) = \sum_{i=0}^{K-1} O(\frac{HW}{M_i^2} \times M_i^4 \frac{C^2}{K}) = \sum_{i=0}^{K-1} O(M_i^2 \frac{HWC^2}{K}) \tag{23}$$

Therefore, the overall time complexity of DSFA-SwinNet is:

$$O = 4O(HWC^2) + 2\sum_{i=0}^{K-1} O\left(M_i^2 \frac{HWC^2}{K}\right) = 4HWC^2 + 2\frac{HWC^2}{K}\sum_{0}^{K-1} M_i^2 \tag{24}$$

In the complexity analysis experiments, this study assesses key performance indicators including training duration, inference efficiency, parameters, and the count of floating-point operations (FLOPs). Training duration is the average time required to complete a full training epoch on the specific dataset. Inference efficiency indicates the average time

required for the model to generate predictions for the given dataset. FLOPs are determined based on the model input tensor dimensions (1,3,256,256) to gauge the computational intensity of the model. The PV03 dataset is selected as the experimental dataset and all experiments are executed on identical hardware to ensure comparable and reliable outcomes.

As depicted in Table 8, Swin-Unet possesses the lowest FLOPs value (0.031Gps). LETNet has the lowest number of parameters (0.95 M). BASNet is the fastest to train, while UNet has the fastest inference. Compared to the baseline model, DSFA-SwinNet utilizes the vision Transformer property to obtain the second highest FLOPs value (0.99 Gps) after Swin-Unet without a substantial increase in the parameter count (25.88 M). On the PV03 dataset, DSFA-SwinNet places second in training time after BASNet and third in inference speed after UNet and BASNet. Compared with Swin-Unet, which employs Swin-Transformer with a fixed window as the backbone network, DSFA-SwinNet improves the training time and inference speed by 67.60 s and 5.86 s, respectively, despite incorporating additional convolutional operations.

Table 8. Complexity study on the PV03 dataset.

Model	Training Duration (s)	Inference Efficiency (s)	Parameters (M)	Flops (Gps)
Swin-Unet	236.38	114.24	41.39	**0.031**
TransUnet	317.59	135.06	105.32	3.88
ACCoNet	2498.21	139.02	127.02	13.30
Unet	235.00	**83.50**	13.40	10.68
MP-ResNet	299.2	104.29	55.03	8.12
DeepLabV3+	195.92	112.28	46.62	3.99
MAResU-Net	300.00	121.90	26.28	2.61
CMTFNet	206.70	116.21	30.07	2.56
BASNet	**145.71** [1]	<u>83.66</u>	<u>12. 57</u>	1.18
LETNet	295.8	300.04	**0.95**	1.08
PIDNet-L	224.46	309.87	37.30	2.16
DSFA-SwinNet	<u>168.78</u> [2]	108.38	25.88	<u>0.99</u>

[1] Bold text highlights the smallest value in each column. [2] Underlined text denotes the second smallest value in each column.

3.6. Ablation Experiments

3.6.1. Ablation Experiments on Model Components

To assess the efficacy of DSFA-SwinNet, this section examines how various components influence its performance, including the PAR bottleneck structure, DSFA mechanism, refined skip connection approach, MLUH module, and DWA mechanism, through ablation experiments and analyses conducted on the Google subset.

Table 9 presents the quantitative evaluation results from the ablation experiments of DSFA-SwinNet. Tests 1 through 5 utilize Swin-Transformer with the window size of 7 as the encoder. The model in Test 1 ranks the lowest in Precision (76.23%), Recall (89.04%), F1 (80.95%), and IoU (69.58%) metrics. Compared to Test 1, Test 2 achieves an improvement of 3.18%, 4.98%, 3.92%, and 5.81% in Precision, Recall, F1, and IoU, respectively, attributed to the cross-tier feature fusion of the refined skip connection strategy. In comparison with Test 2, Test 3 fuses the concatenation results of skip-connections from both spatial and frequency dimensions by incorporating the DSFA mechanism into the refined skip connection strategy, leading to model enhancements of 3.61%, 0.04%, 3.32%, and 3.68% in Precision, Recall, F1, and IoU, respectively. The PAR introduced in Test 4 enriches the internal information of the encoder output features by merging the last two

layers of encoder features through a pyramidal structure, resulting in model improvements of 0.35%, 0.65%, 0.48%, and 1.36% in Precision, Recall, F1, and IoU, respectively, over Test 3. Test 4 exhibits a more gradual ascent in Precision and F1 compared to the other tests, primarily due to the inherently small size of high-level features when extracted at multiple scales, which restricts the potential for enhancing model accuracy. Considering that the output feature dimension of the final encoder layer is 8 × 8 based on the experimental dataset, the accuracy improvement for the model is relatively modest. In Test 5, DSFA-SwinNet with the MLUH module enhances PV areas extraction precision by integrating decoder feature maps through the deeply supervised approach during training. Contrasted with Test 4, Test 5 realizes improvements of 1.17%, 0.66%, 0.96%, and 2.26% in Precision, Recall, F1, and IoU, respectively, with notable advancements in Precision, F1, and IoU. Test 6 broadens the scope of single window constraints on PV feature extraction based on Test 5 through the DWA mechanism. Compared with Test 5, the Precision, Recall, F1 and IoU metrics of Test 6 are improved by 1.43%, 1.19%, 0.81% and 0.81%, respectively.

Table 9. Ablation experiment results on model components for DSFA-SwinNet on the Google subset.

Test	Skip Connection	DSFA	PAR	MLUH	DWA	Precision	Recall	F1	IoU
1	-	-	-	-	-	76.23	89.04	80.95	69.58
2	√	-	-	-	-	79.41	94.02	84.87	75.39
3	√	√	-	-	-	83.02	94.06	88.19	79.07
4	√	√	√	-	-	83.37	94.71	88.67	80.43
5	√	√	√	√	-	84.54	95.37	89.63	82.69
6	√	√	√	√	√	**85.97** [1]	**96.56**	**90.44**	**83.50**

[1] Bold text marks the highest value in each column.

The results of the sample ablation experiment are shown in Figure 13. Since Swin-Transformer is originally designed for medical imaging tasks, the detection of PV areas fails to outline the boundary accurately, and the phenomenon of misdetection of other objects into PV areas occurs in example (c). The incorporation of the refined skip connection strategy notably ameliorates segmentation precision, yet instances of misdetection persist, such as in example (b). Subsequently, as shown in example (c), the integration of the DSFA mechanism renders the feature capturing ability of the model. However, the high-level features from the encoder are not optimally leveraged. To address this, the PAR bottleneck structure is introduced, as evident in example (e), which mitigates the missed detection problem. In example (a), Tests 1–5 inaccurately identify roof glazing as a PV areas. In example (b), Tests 1–5 mistakenly interpret the background roofs as PV areas to varying extents due to the elongated shape and color similarity with the background. The addition of the DWA mechanism bolsters the capability of the model to learn multi-scale representations of PV, leading to marked enhancements in resolving these misclassifications.

To more intuitively observe the learning process of PV features by DSFA-SwinNet, this study visualizes the features of a PV panel sample from the Google subset and a PV array sample from the PV03 subset based on Test 6. These samples represent two typical spatial distribution forms of PV scenes, with the results presented in Figures 14 and 15, respectively.

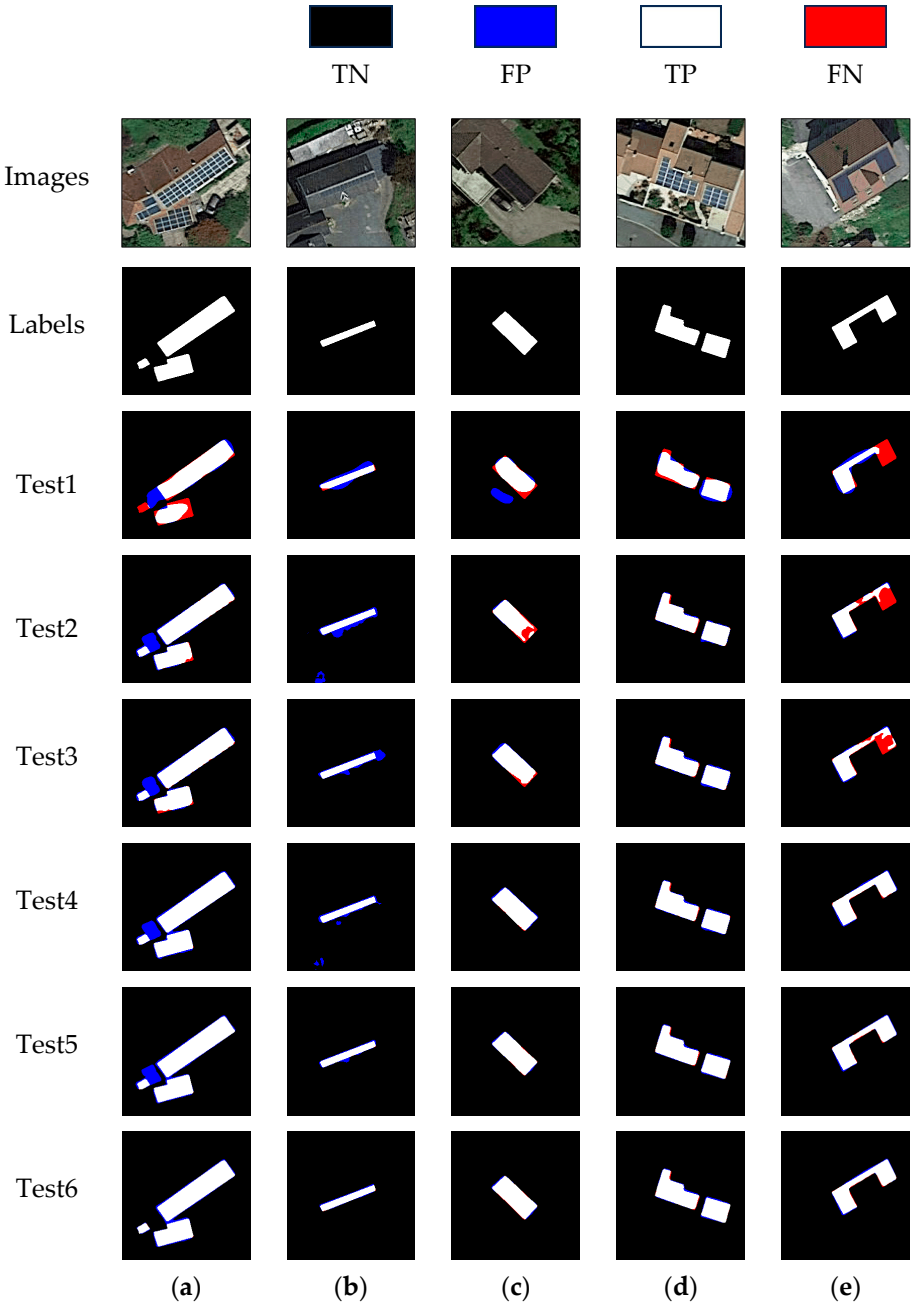


Figure 13. Example of ablation experiment. False positives and false negatives are shown in red and blue, respectively. (**a**–**e**) represent the results of different test cases in the Google subset.

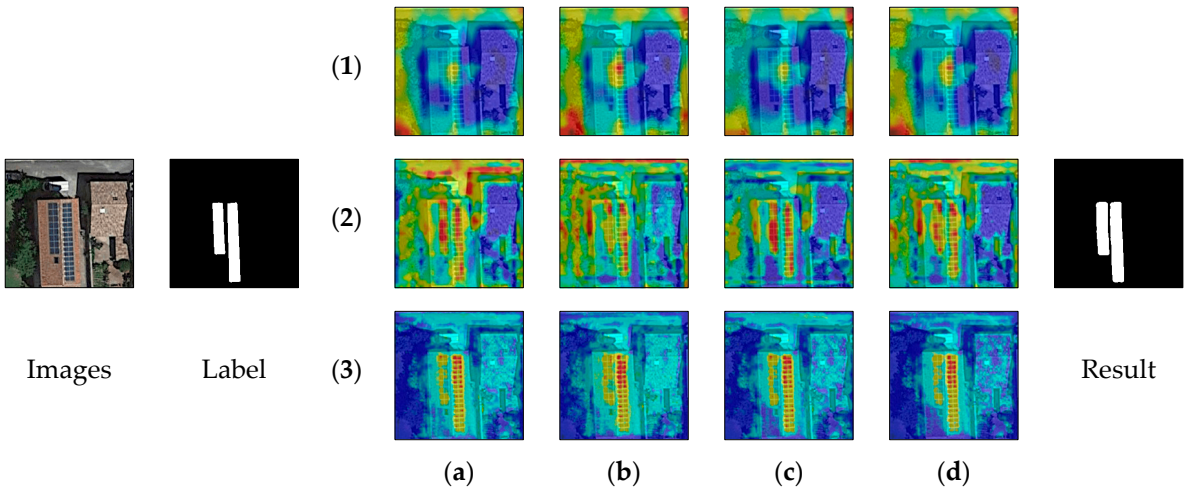


Figure 14. Visualization of PV panel sample from the Google subset. The colors from red to blue represent the probability of being predicted as PV pixels. The (1) row presents feature maps with a resolution of $(\frac{H}{16}, \frac{W}{16})$; the (2) row corresponds to a resolution of $(\frac{H}{32}, \frac{W}{32})$; and the (3) row corresponds to a resolution of $(\frac{H}{64}, \frac{W}{32})$. The (**a**) column shows the feature maps input to the DSFA mechanism, where (1, a) represents the output feature map of the PAR bottleneck structure. The (**b**) column presents the feature maps processed by the spatial attention within the DSFA mechanism. The (**c**) column shows the feature maps processed by the frequency attention within the DSFA mechanism. The (**d**) column presents the feature maps generated by dynamically fusing spatial and frequency attention within the DSFA mechanism, representing the output feature maps of the DSFA mechanism.

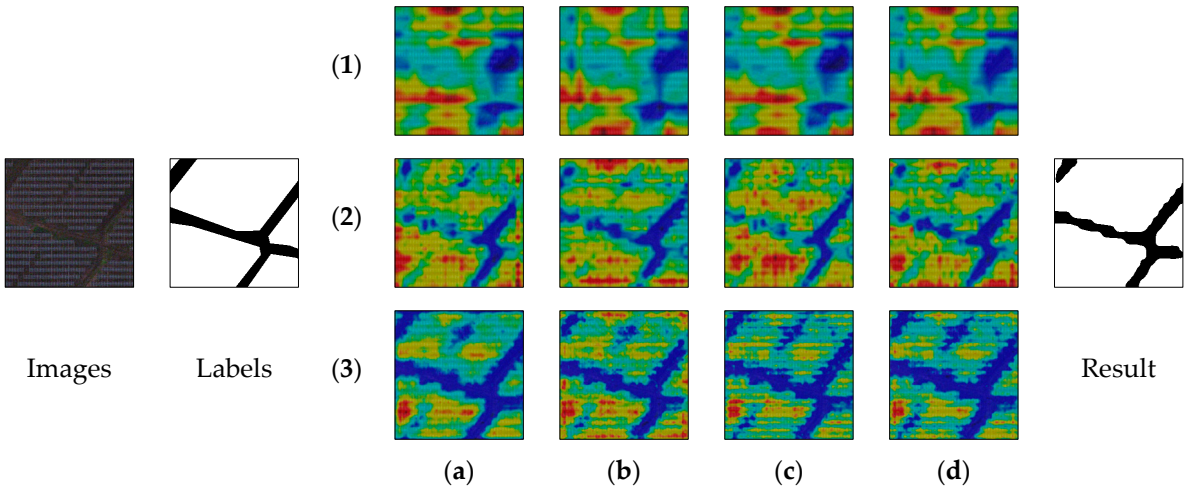


Figure 15. Visualization of PV array samples from the PV03 subset. The colors from red to blue represent the probability of being predicted as PV pixels. The (1) row presents feature maps with a resolution of $(\frac{H}{16}, \frac{W}{16})$; the (2) row corresponds to a resolution of $(\frac{H}{32}, \frac{W}{32})$; and the (3) row corresponds to a resolution of $(\frac{H}{64}, \frac{W}{32})$. The (**a**) column shows the feature maps input to the DSFA mechanism, where (1, a) represents the output feature map of the PAR bottleneck structure. The (**b**) column presents the feature maps processed by the spatial attention within the DSFA mechanism. The (**c**) column shows the feature maps processed by the frequency attention within the DSFA mechanism. The (**d**) column presents the feature maps generated by dynamically fusing spatial and frequency attention within the DSFA mechanism, representing the output feature maps of the DSFA mechanism.

Vertically, the high-resolution feature maps emphasize fine-grained texture features, while the low-resolution feature maps capture broader spatial patterns related to PV structures. Horizontally, the attention regions of spatial attention and frequency attention differ, and dynamic weighted fusion effectively integrates both regions through pixel-wise adaptive weighting.

In the PV panel sample, there is a large background area, and narrow gaps exist between the PV panels. As shown in Figure 14 (1, a), the output feature map of the PAR bottleneck structure performs fusion and multi-scale processing of the high-dimensional features from the last two layers of the encoder, preliminarily defining the probability distribution of the PV areas. By analyzing Figure 14 (1, b) and (1, c), as well as Figure 14 (2, b) and (2, c), it can be observed that spatial attention significantly enhances the expansion of PV features, while frequency attention focuses more on removing background noise and clearly defining the boundaries of the PV areas. In the high-resolution feature maps, such as Figure 14 (3, b) and (3, c), both attention mechanisms capture subtle texture variations in the PV areas and complement the PV features that are not predicted. Comparing Figure 14a,d, after processing with the DSFA mechanism, PV detail features are further enhanced, and background noise is significantly reduced. Additionally, by comparing Figure 14 (1, d) and (2, a), as well as Figure 14 (2, d) and (3, a), it is evident that in the refined skip connection strategy, after fusion with the higher-level features from the encoder, the distribution of the PV areas becomes more centralized, and the PV texture features become clearer.

In the PV array sample, Figure 15 (1, b) and (1, c), as well as Figure 15 (2, b) and (2, c), show that the attention regions of spatial attention and frequency attention focus on different aspects. It is worth noting that the annotation data for this sample contains some errors, as the serrated boundaries of the PV array are not annotated in detail. However, from Figure 15 (2, b) and (3, b), as well as Figure 15 (2, c) and (3, c), it can be observed that the striped texture of the PV array becomes clearer. In Figure 15 (3, d) and the final prediction result, DSFA-SwinNet model successfully focuses on the gaps between the PV panels and the serrated boundaries in the PV array, demonstrating its excellent ability to recognize the fine details of the PV structure.

In conclusion, the PAR bottleneck structure, DSFA mechanism, refined skip connection strategy, MLUH module, and DWA mechanism collectively refine extraction precision, segmentation accuracy, and overall robustness of DSFA-SwinNet model for the PV areas extraction task. They diminish the interference of complex background with PV areas detection and empower DSFA-SwinNet to adeptly discern the PV areas within HRSI. DSFA-SwinNet is fully capable of meeting the stringent requirements for PV areas detection in HRSI.

3.6.2. Ablation Experiments on Loss Functions

To alleviate the imbalance between positive and negative samples, optimize different model objectives, and learn diverse features at different levels, this study designs a weighted combination of hybrid loss functions. For each feature map, the computation involves three loss functions: WBCE Loss, Dice Loss, and Lovasz-Softmax Loss. N denotes the number of pixels in the image. The time complexity for calculating each loss term is $O(N)$. Thus, the total time complexity for the three loss functions is:

$$O(3 \times N) = O(N) \tag{25}$$

Assuming there are k feature maps, the time complexity of the computation process is $O(k \times N)$. Since backpropagation requires the weighted values of all feature maps, let the time complexity of backpropagation be M, then the overall time complexity is $O(k \times N + M)$. In this study, the hybrid loss function contains 15 terms, but to make it

more focused on the key objectives and avoid unnecessary redundant terms, 8 hyperparameters (α, β, γ, and loss weights of the feature maps produced by the bottleneck structure and each layer of the decoder [W_1, W_2, W_3, W_4, W_5]) are tuned. In the resulting hyperparameter combination, W_1 is 0, with W_4 and W_5 are 1, implying that during the training process, the focus is placed on W_4 and W_5.

Table 10 presents the results of DSFA-SwinNet segmentation efficiency and performance on the Google subset for different loss functions and varying numbers of classification heads. Compared to Test 1, Test 3 exhibits an increase in training duration of 7.98 s after incorporating Dice Loss and Lovasz-Softmax Loss. However, this leads to significant improvements in Precision, Recall, F1, and IoU, which increase by 11.58%, 0.37%, 8.10%, and 11.40%, respectively. This result suggests that the weighted combination of these three losses effectively enhances ability of the model to capture PV features. Building on this, Test 3, by incorporating 5 classification heads in the loss function, results in an increase in training duration of 9.98 s compared to Test 2. The Precision, Recall, F1, and IoU improve by 1.39%, 0.71%, 1.32%, and 1.88%, respectively. This change implies that the use of deep supervision in the training process significantly enhances understanding of PV details by the model. The above results indicate that, in practical training, the increase in training time caused by the addition of different loss functions and classification heads is acceptable.

Table 10. Ablation experiment results on loss functions for DSFA-SwinNet on the Google subset. W denotes WBCE Loss, D denotes Dice Loss, and L denotes Lovasz-Softmax Loss, while α, β, and γ represent the corresponding weights. "Number of Classification Head = 1" indicates that the output feature maps from the top decoder layer are used in the loss calculation. Training duration is the average time required to complete a full training epoch.

Test	Loss Function	Number of Classification Head	Training Duration (s)	Precision	Recall	F1	IoU
1	αW	5	177.15	74.39	96.19	82.34	72.10
2	αW + βD + γL	1	**175.15** [1]	84.58	95.85	89.12	81.62
3	αW + βD + γL	5	185.13	**85.97**	**96.56**	**90.44**	**83.50**

[1] Bold text highlights the optimal value in each column.

3.7. Cross-Subset Testing

The Google subset is sampled from Europe, while the PV03 subset is sampled from Jiangsu, China. Therefore, two sets of tests were designed in this section:

- Training on the Google subset and predicting the test set from the PV03 subset;
- Training on the PV03 subset and predicting the test set from the Google subset.

Cross-testing on datasets sampled from different environments aims to evaluate the ability of DSFA-SwinNet model to understand PV features. The segmentation results are shown in Table 11, with sample results presented in Figure 16.

Table 11. Cross-Subset Testing Results.

Test	Training Set	Validation Set	Test Set	Precision	Recall	F1	IoU
1	Google	Google	PV03	74.32	81.73	74.87	67.04
2	PV03	PV03	Google	68.91	94.40	79.67	66.21

In Test 1, DSFA-SwinNet model trained on the Google subset demonstrates more precise detail capture of the PV panels. As shown in Figure 16b, although there is a gap

error in the sample label, the model focuses on the gaps between the PV arrays, indicating that the model segments more based on PV panel features. In Test 2, Figure 16d–f reveal that DSFA-SwinNet model trained on the PV03 subset accurately identifies PV panels of varying scales and shapes. The above experiments demonstrate that DSFA-SwinNet model can effectively cope with environmental interference and learn to extract PV features.

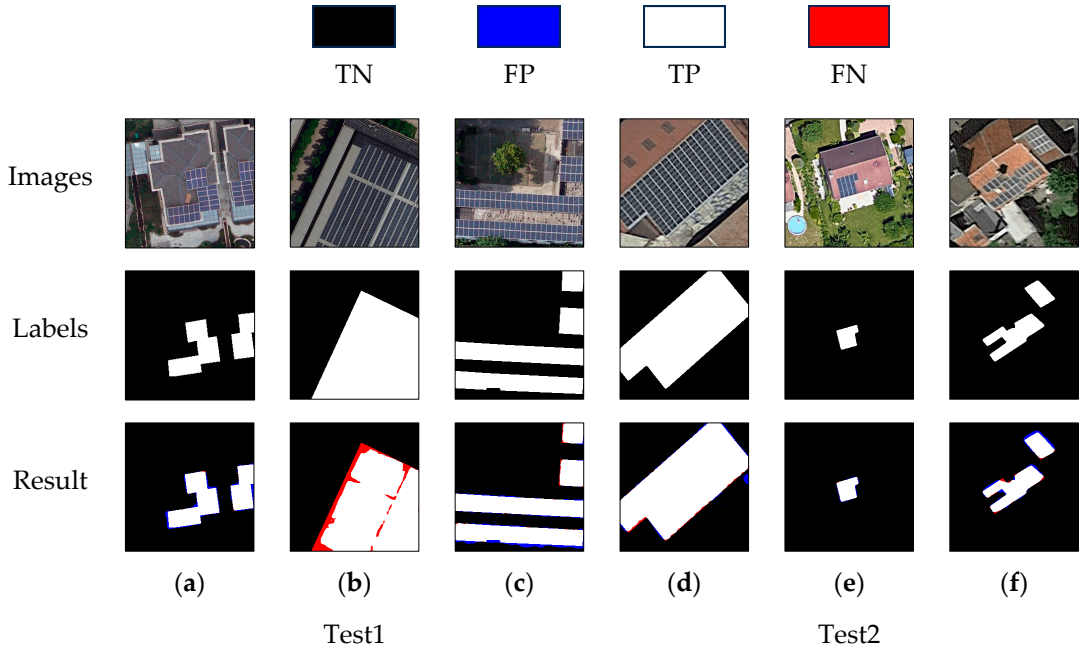


Figure 16. Samples of Cross-Subset Testing. False positives and false negatives are shown in red and blue, respectively. (**a**–**c**) represent the results of different test cases in the PV03 subset. (**d**–**f**) represent the results of different test cases in the Google subset.

4. Discussion

As highlighted in Table 1, this study employs two open-source PV datasets with significantly imbalanced positive-to-negative sample ratios for comparative experiments. The quantitative results presented in Tables 6 and 7 show that models such as Swin-Unet, ACCoNet, MP-ResNet, and PIDNet-L perform better on the PV03 dataset, where the positive-to-negative sample ratio is more balanced. This indicates that these models still face limitations when generalizing to multi-scale PV area extraction in HRSI.

The proposed DSFA-SwinNet excels in extracting both PV panels and arrays, achieving the best performance on two datasets. The DWA mechanism, by incorporating multiple window sizes within the window intervals for patch cutting, enables DSFA-SwinNet encoder to capture multi-scale contextual information at each layer, overcoming the limitations of single window approach of the traditional Swin-Transformer. The Refined Skip Connection Strategy and PAR bottleneck structure optimize the feature propagation paths, ensuring that features at different scales are uniformly captured. The combined features are fed into the DSFA mechanism, where spatial and frequency domain feature information is decoupled, and background noise is filtered through dynamically generated weights. As illustrated in the heatmaps of hyperparameters *flooding*, α, β, and γ in Figure 9, these hyperparameters exhibit an optimal selection range. The sensitivity analysis results in Figure 10 indicate that W_4, W_1, lr, and W_5 have a significant impact on the IoU metric. Researchers are encouraged to adjust hyperparameters with notable effects on IoU based on the characteristics of different datasets, while selectively tuning those influencing convergence speed of the model as needed. This integration of the training strategy

with hyperparameter optimization leads to substantial improvements in both model performance and training efficiency.

However, there are still failure cases in complex scenes within the PV areas detection task. As shown in Figure 17a, in samples with occlusions from trees and buildings, almost all models fail to detect the occluded photovoltaics. In Figure 17b, glass that is similar in color and texture to photovoltaics is misclassified as PV by all models. Although DSFA-SwinNet outperforms other models, failure may still occur in certain complex scenes. In the future, the synergistic relationship between PV textures and the surrounding environment will be further analyzed and refined, and an attempt will be made to integrate multi-source data fusion to improve the recognition of PV areas in such complex environments.

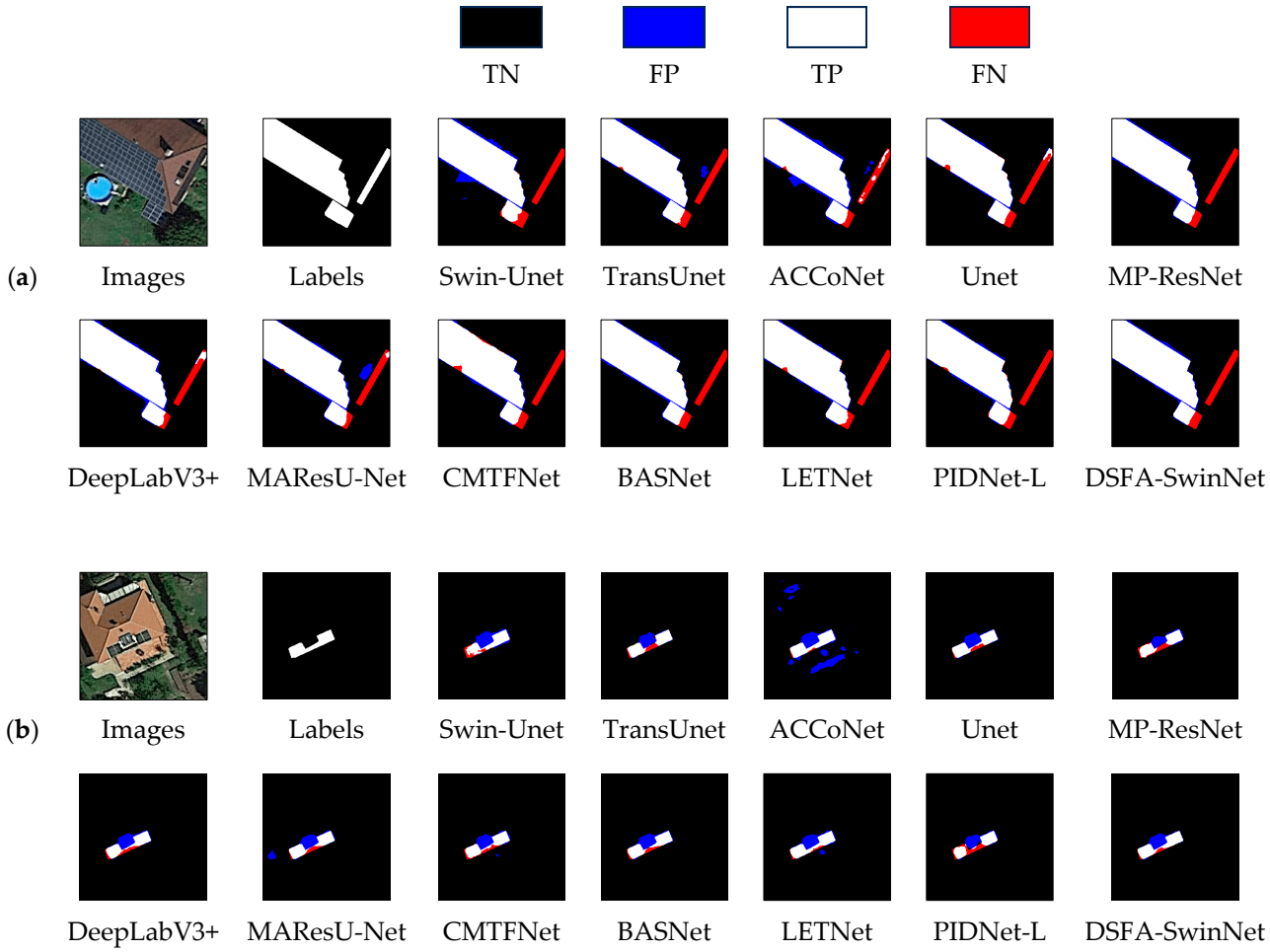


Figure 17. Failure cases analysis. False positives and false negatives are shown in red and blue, respectively. (**a**) represents a case where PV areas are covered by tree shade. (**b**) represents a case with glass similar to PV areas.

In the hyperparameter optimization experiment, Grid Search and Bayesian Optimization algorithms are used to select training batch, learning rate, and loss function weights. However, other hyperparameters, such as the size and stride of the convolution kernel, as well as the window size range selected by DWA, may also impact the final results, and their effects are not analyzed. Additionally, as shown in Table 8, the proposed DSFA-SwinNet still has potential for improvement in training duration and inference speed. Future work will focus on further refining hyperparameter tuning and considering the integration of advanced models like VMamba to enable optimal extraction of PV areas in a shorter time.

5. Conclusions

Detection of PV areas from HRSI is a crucial research focus within the domain of remote sensing image segmentation. However, the diversity of viewpoints of PV built-up areas poses a challenge for effective extraction of PV built-up areas.

In this study, a multi-scale PV areas extraction method (DSFA-SwinNet) is introduced that dynamically decouples the spatial and frequency domains. The refined skip connection strategy is designed that integrates the proposed DSFA mechanism, PAR bottleneck structure, and MLUH module to achieve fine-grained PV areas detection by performing multiscale representation learning from both spatial and frequency domain dimensions. In addition, the physical element constraints are combined with the dynamic window size adjustment mechanism to lift the restriction of fixed window size on Swin-Transformer backbone network, thereby enhancing computational efficiency and maintaining better extraction accuracy. To address the issue of extreme sample imbalance, a hybrid loss function is designed and an automated parameter tuning tool is applied to perform meticulous hyperparameter tuning by combining Grid Search and Bayesian Optimization, which provides a programmatic idea of the parameter ratio of the hybrid loss function for related research. Finally, rigorous and extensive performance evaluation experiments are conducted on the Google subset of the BDAPPV dataset and the PV03 subset of the Jiangsu PV dataset. DSFA-SwinNet performs well in PV segmentation while maintaining a low computational footprint (0.99 Gps) and a moderate number of parameters (25.88 M). Specifically, DSFA-SwinNet achieves 83.50% and 92.00% in IoU, as well as 90.44% and 95.57% in F1 Score, respectively, outperforming other models in the comparative experiments.

DSFA-SwinNet provides an efficient, cost-effective, and reliable method for dynamic PV areas detection based on HRSI providing more comprehensive data support for urban PV planning. Moving forward, the robustness of the model will be further enhanced, and lightweight detection networks for PV areas will be explored, contributing to the promotion of sustainable energy development.

Author Contributions: Methodology, X.L.; Resources, X.L.; Software, Y.Y.; Supervision, S.L.; Writing—original draft, Y.Y.; Writing—review and editing, X.L. and L.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Acknowledgments: The author expresses deep gratitude to the anonymous reviewers.

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

References

1. Construction of Photovoltaic Power Generation in 2023—National Energy Administration. Available online: http://www.nea.gov.cn/2024-02/28/c_1310765696.htm (accessed on 30 May 2024).
2. Hernandez, R.R.; Easter, S.B.; Murphy-Mariscal, M.L.; Maestre, F.T.; Tavassoli, M.; Allen, E.B.; Barrows, C.W.; Belnap, J.; Ochoa-Hueso, R.; Ravi, S.; et al. Tawalbeh. *Renew. Sustain. Energy Rev.* **2014**, *29*, 766–779. https://doi.org/10.1016/j.rser.2013.08.041.
3. Tawalbeh, M.; Al-Othman, A.; Kafiah, F.; Abdelsalam, E.; Almomani, F.; Alkasrawi, M. Environmental Impacts of Solar Photovoltaic Systems: A Critical Review of Recent Progress and Future Outlook. *Sci. Total Environ.* **2021**, *759*, 143528. https://doi.org/10.1016/j.scitotenv.2020.143528.
4. Levin, M.O.; Kalies, E.L.; Forester, E.; Jackson, E.L.A.; Levin, A.H.; Markus, C.; McKenzie, P.F.; Meek, J.B.; Hernandez, R.R. Solar Energy-Driven Land-Cover Change Could Alter Landscapes Critical to Animal Movement in the Continental United States. *Environ. Sci. Technol.* **2023**, *57*, 11499–11509. https://doi.org/10.1021/acs.est.3c00578.

5. Cheng, Y.; Wang, W.; Ren, Z.; Zhao, Y.; Liao, Y.; Ge, Y.; Wang, J.; He, J.; Gu, Y.; Wang, Y.; et al. Multi-Scale Feature Fusion and Transformer Network for Urban Green Space Segmentation from High-Resolution Remote Sensing Images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103514. https://doi.org/10.1016/j.jag.2023.103514.
6. Manso-Callejo, M.-Á.; Cira, C.-I.; Arranz-Justel, J.-J.; Sinde-González, I.; Sălăgean, T. Assessment of the Large-Scale Extraction of Photovoltaic (PV) Panels with a Workflow Based on Artificial Neural Networks and Algorithmic Postprocessing of Vectorization Results. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *125*, 103563. https://doi.org/10.1016/j.jag.2023.103563.
7. Yang, R.; He, G.; Yin, R.; Wang, G.; Zhang, Z.; Long, T.; Peng, Y.; Wang, J. A Novel Weakly-Supervised Method Based on the Segment Anything Model for Seamless Transition from Classification to Segmentation: A Case Study in Segmenting Latent Photovoltaic Locations. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103929. https://doi.org/10.1016/j.jag.2024.103929.
8. Kruitwagen, L.; Story, K.T.; Friedrich, J.; Byers, L.; Skillman, S.; Hepburn, C. A Global Inventory of Photovoltaic Solar Energy Generating Units. *Nature* **2021**, *598*, 604–610. https://doi.org/10.1038/s41586-021-03957-7.
9. Zhang, X.; Xu, M.; Wang, S.; Huang, Y.; Xie, Z. Mapping Photovoltaic Power Plants in China Using Landsat, Random Forest, and Google Earth Engine. *Earth Syst. Sci. Data* **2022**, *14*, 3743–3755. https://doi.org/10.5194/essd-14-3743-2022.
10. Jörges, C.; Vidal, H.S.; Hank, T.; Bach, H. Detection of Solar Photovoltaic Power Plants Using Satellite and Airborne Hyperspectral Imaging. *Remote Sens.* **2023**, *15*, 3403. https://doi.org/10.3390/rs15133403.
11. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. https://doi.org/10.1109/JSTARS.2018.2835377.
12. Li, F.; Dong, W.; Wu, W. A General Model for Comprehensive Electrical Characterization of Photovoltaics under Partial Shaded Conditions. *Adv. Appl. Energy* **2023**, *9*, 100118. https://doi.org/10.1016/j.adapen.2022.100118.
13. Lin, S.; Yao, X.; Liu, X.; Wang, S.; Chen, H.-M.; Ding, L.; Zhang, J.; Chen, G.; Mei, Q. MS-AGAN: Road Extraction via Multi-Scale Information Fusion and Asymmetric Generative Adversarial Networks from High-Resolution Remote Sensing Images under Complex Backgrounds. *Remote Sens.* **2023**, *15*, 3367. https://doi.org/10.3390/rs15133367.
14. Zhang, X.; Zeraatpisheh, M.; Rahman, M.M.; Wang, S.; Xu, M. Texture Is Important in Improving the Accuracy of Mapping Photovoltaic Power Plants: A Case Study of Ningxia Autonomous Region, China. *Remote Sens.* **2021**, *13*, 3909. https://doi.org/10.3390/rs13193909.
15. Xia, Z.; Li, Y.; Guo, X.; Chen, R. High-Resolution Mapping of Water Photovoltaic Development in China through Satellite Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102707. https://doi.org/10.1016/j.jag.2022.102707.
16. Plakman, V.; Rosier, J.; van Vliet, J. Solar Park Detection from Publicly Available Satellite Imagery. *GISci. Remote Sens.* **2022**, *59*, 462–481. https://doi.org/10.1080/15481603.2022.2036056.
17. Malof, J.M.; Bradbury, K.; Collins, L.M.; Newell, R.G. Automatic Detection of Solar Photovoltaic Arrays in High Resolution Aerial Imagery. *Appl. Energy* **2016**, *183*, 229–240. https://doi.org/10.1016/j.apenergy.2016.08.191.
18. Chen, Z.; Kang, Y.; Sun, Z.; Wu, F.; Zhang, Q. Extraction of Photovoltaic Plants Using Machine Learning Methods: A Case Study of the Pilot Energy City of Golmud, China. *Remote Sens.* **2022**, *14*, 2697. https://doi.org/10.3390/rs14112697.
19. Li, Q.; Feng, Y.; Leng, Y.; Chen, D. SolarFinder: Automatic Detection of Solar Photovoltaic Arrays. In Proceedings of the 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Sydney, NSW, Australia, 21–24 April 2020; pp. 193–204.
20. Yuan, W.; Xu, W. MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer. *Remote Sens.* **2021**, *13*, 4743. https://doi.org/10.3390/rs13234743.
21. Ying, Z.; Li, M.; Tong, W.; Haiyong, C. Automatic Detection of Photovoltaic Module Cells Using Multi-Channel Convolutional Neural Network. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 3571–3576.
22. He, K.; Zhang, L. Automatic Detection and Mapping of Solar Photovoltaic Arrays with Deep Convolutional Neural Networks in High Resolution Satellite Images. In Proceedings of the 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2), Wuhan, China, 30 October–1 November 2020; pp. 3068–3073.
23. Ishii, T.; Simo-Serra, E.; Iizuka, S.; Mochizuki, Y.; Sugimoto, A.; Ishikawa, H.; Nakamura, R. Detection by Classification of Buildings in Multispectral Satellite Imagery. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 3344–3349.
24. Shi, K.; Bai, L.; Wang, Z.; Tong, X.; Mulvenna, M.D.; Bond, R.R. Photovoltaic Installations Change Detection from Remote Sensing Images Using Deep Learning. In Proceedings of the IGARSS 2022—IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 3231–3234.

25. Parhar, P.; Sawasaki, R.; Todeschini, A.; Reed, C.; Vahabi, H.; Nusaputra, N.; Vergara, F. HyperionSolarNet: Solar Panel Detection from Aerial Images. *arXiv* **2022**, arXiv:2201.02107.
26. Jurakuziev, D.; Jumaboev, S.; Lee, M. A Framework to Estimate Generating Capacities of PV Systems Using Satellite Imagery Segmentation. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106186. https://doi.org/10.1016/j.engappai.2023.106186.
27. Zhao, Z.; Chen, Y.; Li, K.; Ji, W.; Sun, H. Extracting Photovoltaic Panels From Heterogeneous Remote Sensing Images With Spatial and Spectral Differences. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5553–5564. https://doi.org/10.1109/JSTARS.2024.3369660.
28. Gasparyan, H.A.; Davtyan, T.A.; Agaian, S.S. A Novel Framework for Solar Panel Segmentation From Remote Sensing Images: Utilizing Chebyshev Transformer and Hyperspectral Decomposition. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. https://doi.org/10.1109/TGRS.2024.3386402.
29. Yu, J.; Wang, Z.; Majumdar, A.; Rajagopal, R. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule* **2018**, *2*, 2605–2617. https://doi.org/10.1016/j.joule.2018.11.021.
30. Castello, R.; Roquette, S.; Esguerra, M.; Guerra, A.; Scartezzini, J.-L. Deep Learning in the Built Environment: Automatic Detection of Rooftop Solar Panels Using Convolutional Neural Networks. *J. Phys. Conf. Ser.* **2019**, *1343*, 012034. https://doi.org/10.1088/1742-6596/1343/1/012034.
31. Yuan, J.; Yang, H.-H.L.; Omitaomu, O.A.; Bhaduri, B.L. Large-Scale Solar Panel Mapping from Aerial Images Using Deep Convolutional Networks. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 2703–2708.
32. Moradi Sizkouhi, A.M.; Aghaei, M.; Esmailifar, S.M.; Mohammadi, M.R.; Grimaccia, F. Automatic Boundary Extraction of Large-Scale Photovoltaic Plants Using a Fully Convolutional Network on Aerial Imagery. *IEEE J. Photovolt.* **2020**, *10*, 1061–1067. https://doi.org/10.1109/JPHOTOV.2020.2992339.
33. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
34. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
36. Dong, Y.; Yang, Z.; Liu, Q.; Zuo, R.; Wang, Z. Fusion of GaoFen-5 and Sentinel-2B Data for Lithological Mapping Using Vision Transformer Dynamic Graph Convolutional Network. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103780. https://doi.org/10.1016/j.jag.2024.103780.
37. Liang, M.; Zhang, X.; Yu, X.; Yu, L.; Meng, Z.; Zhang, X.; Jiao, L. An Efficient Transformer with Neighborhood Contrastive Tokenization for Hyperspectral Images Classification. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103979. https://doi.org/10.1016/j.jag.2024.103979.
38. Zhang, J.; Lin, S.; Ding, L.; Bruzzone, L. Multi-Scale Context Aggregation for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 701. https://doi.org/10.3390/rs12040701.
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
40. Fu, W.; Xie, K.; Fang, L. Complementarity-Aware Local–Global Feature Fusion Network for Building Extraction in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3370714.
41. Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20. https://doi.org/10.1109/TGRS.2023.3286826.
42. Chen, Y.; Zhou, J.; Ge, Y.; Dong, J. Uncovering the Rapid Expansion of Photovoltaic Power Plants in China from 2010 to 2022 Using Satellite Data and Deep Learning. *Remote Sens. Environ.* **2024**, *305*, 114100. https://doi.org/10.1016/j.rse.2024.114100.
43. Guo, Z.; Lu, J.; Chen, Q.; Liu, Z.; Song, C.; Tan, H.; Zhang, H.; Yan, J. TransPV: Refining Photovoltaic Panel Detection Accuracy through a Vision Transformer-Based Deep Learning Model. *Appl. Energy* **2024**, *355*, 122282. https://doi.org/10.1016/j.apenergy.2023.122282.
44. Hou, X.; Wang, B.; Hu, W.; Yin, L.; Wu, H. SolarNet: A Deep Learning Framework to Map Solar Power Plants In China From Satellite Imagery. *arXiv* **2019**, arXiv:1912.03685.

45. Zhu, R.; Guo, D.; Wong, M.S.; Qian, Z.; Chen, M.; Yang, B.; Chen, B.; Zhang, H.; You, L.; Heo, J.; et al. Deep Solar PV Refiner: A Detail-Oriented Deep Learning Network for Refined Segmentation of Photovoltaic Areas from Satellite Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103134. https://doi.org/10.1016/j.jag.2022.103134.
46. Wang, J.; Chen, X.; Shi, W.; Jiang, W.; Zhang, X.; Hua, L.; Liu, J.; Sui, H. Rooftop PV Segmenter: A Size-Aware Network for Segmenting Rooftop Photovoltaic Systems from High-Resolution Imagery. *Remote Sens.* **2023**, *15*, 5232. https://doi.org/10.3390/rs15215232.
47. Tan, M.; Luo, W.; Li, J.; Hao, M. TEMCA-Net: A Texture-Enhanced Deep Learning Network for Automatic Solar Panel Extraction in High Groundwater Table Mining Areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2838–2848. https://doi.org/10.1109/JSTARS.2023.3347572.
48. Kleebauer, M.; Marz, C.; Reudenbach, C.; Braun, M. Multi-Resolution Segmentation of Solar Photovoltaic Systems Using Deep Learning. *Remote Sens.* **2023**, *15*, 5687. https://doi.org/10.3390/rs15245687.
49. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
50. Chen, Z.; Luo, Y.; Wang, J.; Li, J.; Wang, C.; Li, D. DPENet: Dual-Path Extraction Network Based on CNN and Transformer for Accurate Building and Road Extraction. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103510. https://doi.org/10.1016/j.jag.2023.103510.
51. Fan, J.; Shi, Z.; Ren, Z.; Zhou, Y.; Ji, M. DDPM-SegFormer: Highly Refined Feature Land Use and Land Cover Segmentation with a Fused Denoising Diffusion Probabilistic Model and Transformer. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104093. https://doi.org/10.1016/j.jag.2024.104093.
52. Liu, Y.; Gao, K.; Wang, H.; Yang, Z.; Wang, P.; Ji, S.; Huang, Y.; Zhu, Z.; Zhao, X. A Transformer-Based Multi-Modal Fusion Network for Semantic Segmentation of High-Resolution Remote Sensing Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104083. https://doi.org/10.1016/j.jag.2024.104083.
53. Kasmi, G.; Saint-Drenan, Y.-M.; Trebosc, D.; Jolivet, R.; Leloux, J.; Sarr, B.; Dubus, L. A Crowdsourced Dataset of Aerial Images with Annotated Solar Photovoltaic Arrays and Installation Metadata. *Sci. Data* **2023**, *10*, 59. https://doi.org/10.1038/s41597-023-01951-4.
54. Jiang, H.; Yao, L.; Lu, N.; Qin, J.; Liu, T.; Liu, Y.; Zhou, C. Multi-Resolution Dataset for Photovoltaic Panel Segmentation from Satellite and Aerial Imagery. *Earth Syst. Sci. Data* **2021**, *13*, 5389–5401. https://doi.org/10.5194/essd-13-5389-2021.
55. Ren, P.; Li, C.; Wang, G.; Xiao, Y.; Du, Q.; Liang, X.; Chang, X. Beyond Fixation: Dynamic Window Visual Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2022; pp. 11987–11997.
56. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency Channel Attention Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 783–792.
57. Chi, K.; Yuan, Y.; Wang, Q. Trinity-Net: Gradient-Guided Swin Transformer-Based Remote Sensing Image Dehazing and Beyond. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. https://doi.org/10.1109/TGRS.2023.3285228.
58. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4413–4421.
59. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv* **2018**, arXiv:1807.05118.
60. Ishida, T.; Yamane, I.; Sakai, T.; Niu, G.; Sugiyama, M. Do We Need Zero Training Loss After Achieving Zero Training Error? *arXiv* **2021**, arXiv:2002.08709.
61. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *Proceedings of the Computer Vision—ECCV 2022 Workshops*; Karlinsky, L., Michaeli, T., Nishino, K., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 205–218.
62. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
63. Li, G.; Liu, Z.; Zeng, D.; Lin, W.; Ling, H. Adjacent Context Coordination Network for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Cybern.* **2023**, *53*, 526–538. https://doi.org/10.1109/TCYB.2022.3162945.
64. Ding, L.; Zheng, K.; Lin, D.; Chen, Y.; Liu, B.; Li, J.; Bruzzone, L. MP-ResNet: Multipath Residual Network for the Semantic Segmentation of High-Resolution PolSAR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3079925.

65. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3063381.
66. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. https://doi.org/10.1109/TGRS.2023.3314641.
67. Xu, G.; Li, J.; Gao, G.; Lu, H.; Yang, J.; Yue, D. Lightweight Real-Time Semantic Segmentation Network with Efficient Transformer and CNN. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15897–15906.
68. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19529–19539.
69. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. BASNet: Boundary-Aware Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7479–7489.