energies

MDPI

Article

Enhancing Rooftop Photovoltaic Segmentation Using Spatial Feature Reconstruction and Multi-Scale Feature Aggregation

Yu Xiao [1], Long Lin [1], Jun Ma [1] and Maoqiang Bi [2,*]

[1] School of Electric Power Engineering, Chongqing Water Resources and Electric Engineering College, Chongqing 404155, China
[2] School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 404155, China
* Correspondence: bimaoqiang@cqut.edu.cn

Abstract: Amidst the dual challenges of energy shortages and global warming, photovoltaic (PV) power generation has emerged as a critical technology due to its efficient utilization of solar energy. Rooftops, as underutilized spaces, are ideal locations for installing solar panels, avoiding the need for additional land. However, the accurate and generalized segmentation of large-scale PV panel images remains a technical challenge, primarily due to varying image resolutions, large image scales, and the significant imbalance between foreground and background categories. To address these challenges, this paper proposes a novel model based on the Res2Net architecture, an enhanced version of the classic ResNet optimized for multi-scale feature extraction. The model integrates Spatial Feature Reconstruction and multi-scale feature aggregation modules, enabling effective extraction of multi-scale data features and precise reconstruction of spatial features. These improvements are particularly designed to handle the small proportion of PV panels in images, effectively distinguishing target features from redundant ones and improving recognition accuracy. Comparative experiments conducted on a publicly available rooftop PV dataset demonstrate that the proposed method achieves superior performance compared to mainstream techniques, showcasing its effectiveness in precise PV panel segmentation.

Keywords: spatial feature reconstruction; multi-scale feature fusion; PV detection; Res2Net

check for updates

Received: 29 November 2024
Revised: 22 December 2024
Accepted: 27 December 2024
Published: 31 December 2024

Citation: Xiao, Y.; Lin, L.; Ma, J.; Bi, M. Enhancing Rooftop Photovoltaic Segmentation Using Spatial Feature Reconstruction and Multi-Scale Feature Aggregation. *Energies* **2025**, *18*, 119. https://doi.org/10.3390/en18010119

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

1. Introduction

In the context of global energy resource depletion and climate warming, photovoltaic (PV) power generation has become an efficient solar energy utilization solution. According to data from China's National Energy Administration [1], as of June 2023, China's grid-connected photovoltaic capacity reached 470 gigawatts, with distributed systems accounting for 42%. Rooftop solar energy not only offers low costs and rapid construction but also provides additional income through electricity sales, with China's rooftop PV installation capacity reaching 51.1 gigawatts in 2022, accounting for 54% of the global total. Despite the rapid development of rooftop PV technology, some regions have not fully utilized its potential, while others have experienced electricity wastage due to excessive development and improper layout [2]. Therefore, a detailed assessment of the development of rooftop PV systems is particularly important for the effective management and planning of regional distributed PV energy systems. With the rapid development of remote sensing technology and deep learning, significant progress has been made in recent years in automatically obtaining information on solar panel installations in specific areas. This technology combines the efficient collection of high-resolution terrestrial feature images

with the powerful feature extraction capabilities of deep learning algorithms, becoming an important means of extracting rooftop and photovoltaic (PV) panel information [3–6]. Through remote sensing image analysis and the prediction of ground object information, it not only provides a viable alternative to laborious field surveys and traditional reporting but also demonstrates great potential in detecting and assessing solar panels regionally. In the development and application of PV systems, numerous studies focus on improving system efficiency and monitoring technology. Firstly, Li et al. [7] proposed an innovative joint task learning framework that utilizes a scale-adaptive module and a location guidance module, significantly improving the edge detection accuracy of domestic rooftop PV systems, thereby optimizing the planning and management of PV development. In terms of monitoring technology, Qi et al. [2] effectively monitored and assessed the current state and future potential of rooftop PV installations through a remote sensing image-based dual-branch deep learning framework and MANet network. Additionally, addressing partial shading issues, Giovanni et al. [8] introduced a solution to enhance the power generation efficiency of Building-Applied Photovoltaic (BAPV) systems. This solution optimizes the array's output through electrical reconfiguration, enhancing performance under adverse conditions. In forecasting and fault detection, Satpathy et al. [9] utilized a neural network model for the short-term prediction of rooftop PV power generation, demonstrating the model's strengths in data capture and energy management. Concurrently, Aljafari et al. [10] developed a fault detection and online monitoring system, integrating Internet of Things (IoT) technology and a 1D-CNN model, significantly enhancing the reliability and efficiency of fault detection in grid-connected PV systems. Lastly, Sulaiman et al. [11] combined operational data and Building Information Modeling (BIM) to improve and predict the energy output of BAPV systems, providing new insights for sustainable energy management and promoting continuous improvement in system design. However, despite the successful application of this approach in many studies, there remain some limitations that need further resolution: (1) In practical scenarios, spatial resolution plays a key role in semantic segmentation, as high-resolution images provide more information for training deep learning models. Currently, segmentation accuracy at different resolutions presents challenges in effectively utilizing spatial information across various scales. (2) Moreover, accurately and universally segmenting various PV categories and complex terrestrial features in remote sensing images remains a significant challenge. Consequently, training datasets may face multiple imbalances, particularly regarding resolution, size, and foreground–background categories. These imbalances complicate model training, leading to poor accuracy and generalization capabilities. As illustrated in Figure 1, from left to right, the spatial resolutions are 0.1 m, 0.3 m, and 0.8 m, respectively, with high-resolution scenarios being more challenging to recognize.

To address these issues, this paper proposes a novel segmentation network, summarized as follows:

1. This study introduces a new Multi-scale Feature Aggregation Network designed to better capture PV panel information in remote sensing imagery through feature extraction at different scales. The architecture of this network effectively integrates features from multiple levels, fully utilizing the rich information in images, thereby enhancing the model's ability to recognize PV panels of varying sizes and shapes. This multi-scale design not only increases the model's flexibility, adapting to more diverse ground situations, but also provides more comprehensive data support for subsequent feature analysis.
2. Furthermore, this study proposes a new Spatial Feature Reconstruction method that effectively captures global contextual information in horizontal and vertical directions by reconstructing spatial features based on the unique characteristics of PV panels.

In this process, the study focuses on modeling the rectangular key regions in the imagery, allowing the model to more accurately locate the positions and shape features of PV panels. This reconstruction technique not only improves the accuracy of PV panel segmentation but also provides robust support for feature understanding in complex scenes.

3. Finally, on the publicly available rooftop PV segmentation dataset, the proposed method demonstrates significant advantages in performance compared to other comparative methods. Systematic experiments validate that the combination of the new Multi-scale Feature Aggregation Network and Spatial Feature Reconstruction approach leads to substantial improvements in both the accuracy and robustness of the model. This result not only verifies the effectiveness of the proposed methods but also lays a solid foundation for future applications in the automatic recognition and monitoring of PV panels.

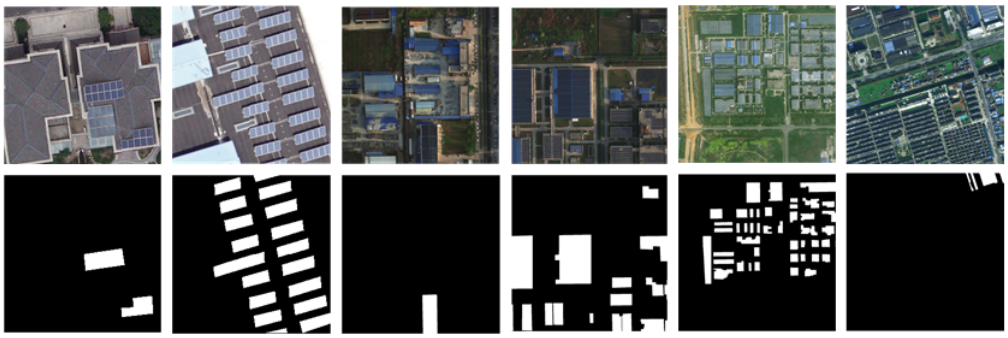


Figure 1. An illustrative diagram depicting rooftop photovoltaic panels with different spatial resolutions. The top row shows the original images, while the bottom row displays the corresponding segmentation labels.

The structure of the remaining sections in this paper is as follows: In Section 2, we present a summary of related work on rooftop photovoltaic segmentation. In Section 3, we provide a detailed description of our proposed network. Section 4 contains details about the dataset, evaluation metrics, implementation specifics, and a series of experiments aimed at evaluating the performance of our crack detection method in tunnel lining structures. Lastly, in Section 5, we conclude the paper and offer insights into future research directions.

2. Related Work

The detection, segmentation, and optimization of photovoltaic (PV) systems have garnered significant attention in recent years, fueled by advancements in remote sensing and deep learning technologies. Early contributions laid the foundation for PV panel recognition but faced limitations in accuracy, efficiency, and application breadth.

In 2015, Malof et al. [12] pioneered the use of support vector machine (SVM) models to analyze high-resolution remote sensing data from the US Geological Survey. While their approach successfully identified whether an image contained PV panels, it was unable to measure the exact area occupied by these panels. Building upon this, a research group [13] employed deep convolutional neural networks (CNNs) for large-scale PV detection, enabling pixel-level segmentation. Meanwhile, Golovko et al. [14] demonstrated the feasibility of using low-resolution Google satellite imagery for PV panel detection, though accuracy and computational efficiency remained significant challenges.

Recent efforts have focused on integrating deep learning with remote sensing to assess and optimize rooftop PV potential. Yan et al. [15] developed a deep learning framework to construct 3D building models from high-resolution satellite imagery, enabling precise PV potential estimation. Nasrallah et al. [16] further refined this process by using instance

segmentation to extract building boundaries and proposing an algorithm for optimal PV panel arrangement, considering rooftop morphology to maximize energy generation in Lebanon. Similarly, Cui et al. [17] applied an enhanced SegNeXt model to map rooftops in rural China, combining geographical and technical factors with solar radiation data to calculate PV potential based on national design codes. Liu et al. [18] employed Mask R-CNN to identify rooftops and obstacles in Changsha, using EnergyPlus models to predict energy generation under various scenarios. Lin et al. [19] used DeepLabv3+ for rooftop extraction in Beijing, integrating meteorological data and geometric transformations to analyze long-term carbon reduction.

The extraction of building rooftops, a critical step for PV placement, has been the focus of several studies. Chen et al. [20] proposed a contour-guided encoder–decoder network to improve rooftop boundary extraction, while Khan et al. [21] introduced an encoder–decoder framework for automatic rooftop contour detection. Shao et al. [22] presented the Building Residual Refinement Network, combining feature prediction and residual refinement modules to improve segmentation accuracy. Chen et al. [23] designed an Adaptive Feature Screening Network, utilizing a spatial upsampling module to extract multi-scale building features effectively. In addition to rooftop extraction, recent work has highlighted the importance of deep learning in optimizing PV placement. For example, Li et al. [7] introduced a joint task learning framework with scale-adaptive and location guidance modules, significantly improving edge detection and improving PV system planning. Qi et al. [2] used a dual-branch deep learning framework with MANet to monitor rooftop photovoltaic systems, providing insight into optimal module placement.

Recent studies have also explored segmentation techniques to improve the performance of the photovoltaic system. Jie et al. [24] incorporated a gated fusion module into EfficientUNet to identify distributed photovoltaic power stations. Hou et al. [25] combined U-Net with EMANet to enhance the segmentation accuracy of centralized photovoltaic systems. Costa et al. [26] evaluated 16 semantic segmentation networks, concluding that U-Net with EfficientNet-b7 offered the best performance for PV segmentation. Wang et al. [27] proposed PVNet, which combines coarse predictions with refinement modules to optimize segmentation boundaries. Beyond segmentation, some studies focus on improving system efficiency. Giovanni et al. [8] addressed the partial shading problem in Building-Applied Photovoltaic (BAPV) Systems by optimizing electrical reconfiguration, enhancing energy output under suboptimal conditions. Satpathy et al. [9] demonstrated the use of neural networks for the short-term prediction of rooftop photovoltaic power, which can also guide module placement. Sulaiman et al. [11] integrated operational data with Building Information Modeling (BIM) to improve the design of the photovoltaic system and predict energy output, providing valuable tools for sustainable energy management.

Although significant advances have been made in PV panel detection, rooftop extraction, and performance optimization, existing methods still face challenges in accurately segmenting small-scale PV modules, addressing data imbalance, and optimizing placements in complex environments. These gaps underscore the need for robust methods that can integrate multi-scale processing, spatial context reconstruction, and adaptive optimization to achieve more practical and efficient PV deployment.

3. Methods

3.1. Overview of the Proposed Method

This section provides a detailed explanation of the network's overall architecture, focusing on the Spatial Feature Reconstruction (SFR) module, Multi-scale Feature Aggregation (MFA), and the Res2Net module. The improved network, illustrated in Figure 2, consists of three main components: (1) a feature extraction backbone built on Res2Net

blocks; (2) a feature enhancement mechanism leveraging the SFR and MFA modules; and (3) an efficient feature fusion and output stage. The network adopts a U-Net framework, structured with four stages in both the encoder and decoder. In the encoder, the channel dimensions at each stage are 64, 128, 256, and 512, while in the decoder, they are reversed as 512, 256, 128, and 64, aligning with the U-Net design principles.

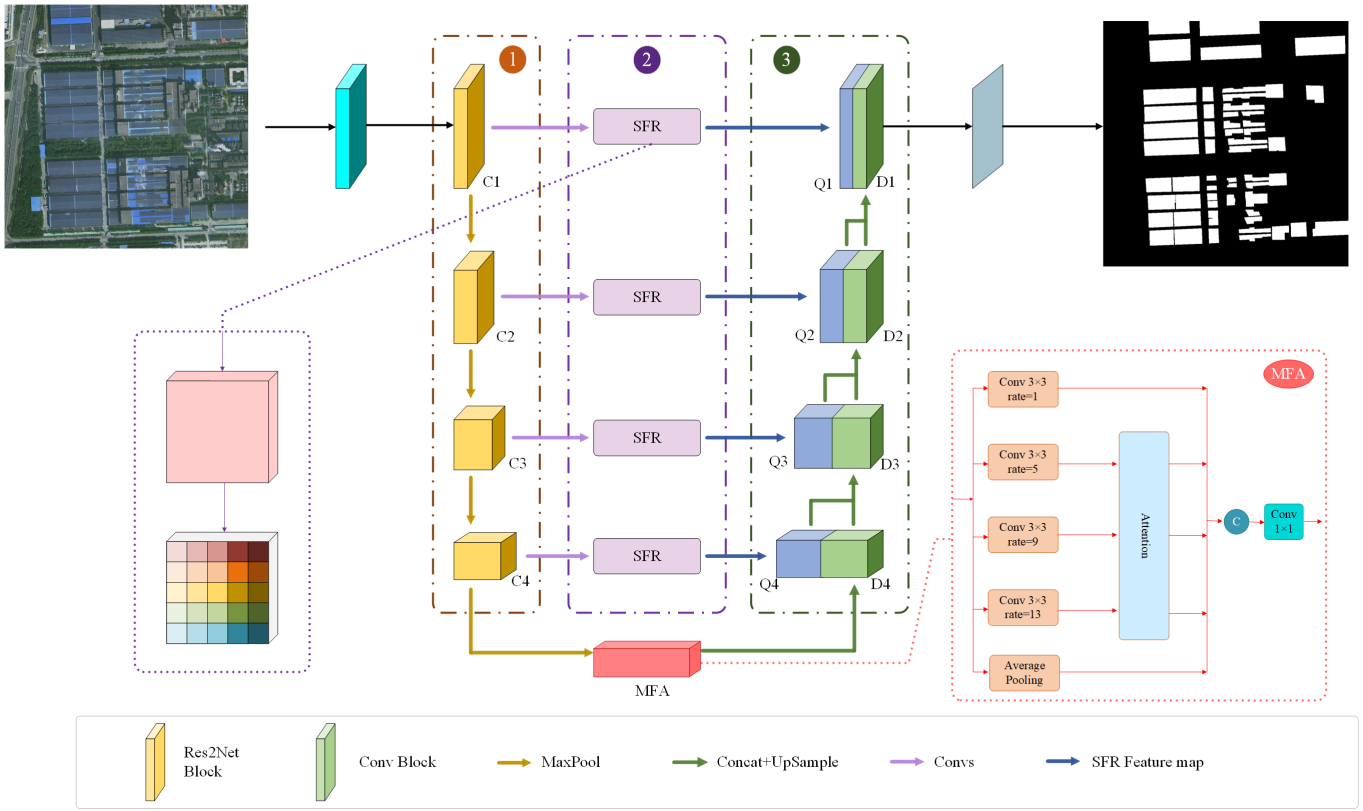


Figure 2. The proposed network consists of three key components: the encoding phase, skip connections, and the decoding phase. The skip connection segment incorporates the Spatial Feature Reconstruction (SFR) module to strengthen feature extraction. To enlarge the receptive field, Multi-scale Feature Aggregation (MFA) is applied in the lower layers using parallel dilated convolutions. During the decoding phase, high-level semantic features from the MFA module are merged with low-level features to improve feature fusion.

The SFR module is introduced as a replacement for the traditional skip connection layer, adopting a separate-and-reconstruct strategy to capture richer spatial information. In the middle section of the network, the original structure is replaced by the MFA module, which utilizes dilated convolutions with varying rates to capture a wider range of semantic information. This module dynamically aggregates multi-scale features through an attention mechanism, enhancing the network's ability to represent complex patterns. In the decoder, high-level semantic features from the MFA module are fused with low-level features to improve feature integration. This design streamlines feature processing while boosting the model's performance and accuracy.

3.2. Res2Net Backbone

As shown in Figure 3, Res2Net [28] divides the channels in each residual block into several groups and performs independent convolution within each group. Suppose the number of channels for a feature map X is C. This feature map will be divided into G

groups, each containing C/G channels. For each group i, the calculation can be expressed as follows:

$$Y_i = F_i(Y_{i-1} \oplus X_i) \tag{1}$$

Here, F_i denotes the convolution operation applied to the i-th group, Y_i represents the output feature map of the i-th group, and X_i corresponds to the i-th channel of the input feature map X. The symbol $\oplus$ indicates the concatenation of feature maps. Finally, the outputs Y_i from all groups are concatenated to produce the final output Y of the residual block, which is mathematically expressed as follows:

$$Y = Y_1 \oplus Y_2 \oplus \ldots \oplus Y_G \tag{2}$$

The Res2Net architecture integrated into our method improves the model's capacity for multi-scale feature extraction by embedding this capability directly within a single residual block. This design enables the model to capture both fine-grained details and broader contextual information, making it highly effective for handling images with rooftops of varying sizes and scales.

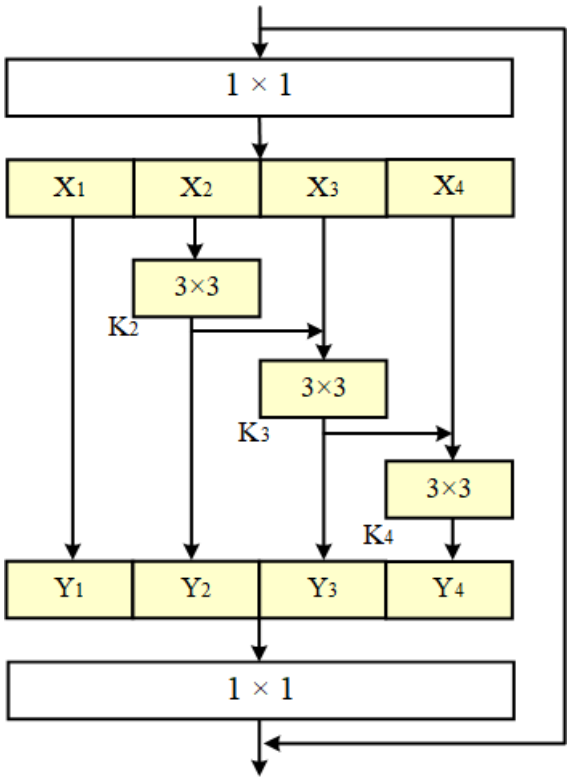


Figure 3. The structural diagram of the Res2Net module.

3.3. Spatial Feature Reconstruction

To effectively address the spatial redundancy in features, we introduced the Spatial Feature Reconstruction (SFR) [29] module depicted in Figure 4. Specifically, unlike existing spatial reconstruction methods that typically rely on complex attention mechanisms to extract features and assign corresponding weights to the original feature maps, the SFR module adopts an innovative design based on feature separation and cross-reconstruction, enabling more fine-grained selection and reorganization of spatial features. This approach effectively reduces redundant information and improves the efficiency of information flow, thereby further optimizing feature representation capabilities and significantly enhancing model performance. The main operation of SFR is divided into two stages: separation and reconstruction.

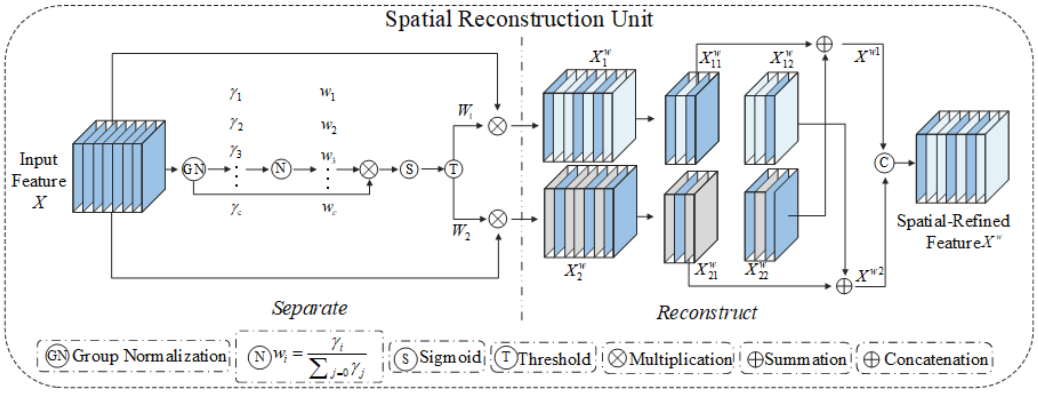


Figure 4. The structural diagram of the proposed Spatial Feature Reconstruction module.

Separation Operation: This process aims to distinguish information-rich feature maps from less informative ones based on their spatial content. To evaluate the information content, trainable parameters γ within the Group Normalization (GN) layer are utilized. Initially, the input feature map X undergoes normalization, followed by the application of a sigmoid function to scale the weight values into the range (0, 1). Weights exceeding a predefined threshold are assigned a value of 1, while those below the threshold are set to 0. This results in the separation of information-rich weights W_1 from less informative weights.

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C}\gamma_j}, \quad i, j = 1, 2, \ldots, C \tag{3}$$

$$W = \text{Gate}(\text{Sigmoid}(W_\gamma(\text{GN}(X)))) \tag{4}$$

Finally, element-wise multiplication produces weighted features X^w, where X_1^w contains information-rich features and X_2^w represents redundant features. Reconstruction Operation: To reduce spatial redundancy, a cross-reconstruction operation is proposed, which combines the information-rich features with the less informative ones to enhance the information flow. After cross-reconstructing these two parts, they are concatenated, resulting in a spatially refined feature map X^w.

$$\begin{cases} X_1^w = W_1 \otimes X, \\ X_2^w = W_2 \otimes X, \\ X_{11}^w \oplus X_{22}^w = X_1^w, \\ X_{21}^w \oplus X_{12}^w = X_2^w, \\ X_1^w \cup X_2^w = X^w \end{cases} \tag{5}$$

The Spatial Feature Reconstruction module is designed to improve spatial localization, ensuring that even small-scale PV panels are accurately detected despite their limited presence in high-resolution images.

3.4. Multi-Scale Feature Aggregation

The network proposed in this paper effectively addresses the scale variations in rooftop photovoltaic systems across different scenes through the MFA (Multi-scale Feature Aggregation) module. It captures multi-scale feature information, avoiding the conventional method of directly concatenating fixed receptive field feature maps. Direct aggregation of features at multiple scales could diminish the performance of features in high-resolution scenes due to misleading background information. Therefore, the MFA does not uniformly process features across all scales but instead uses a dynamic feature selection module to

independently handle key features, enhancing the efficiency of feature processing. Detailed structural information can be found in Figure 2.

During the feature extraction phase, the network first transmits the final output of the feature pyramid encoder, denoted as $F_{C\times H\times W}$, to four atrous convolutional branches with varying convolution rates, achieving multi-scale feature extraction. The output maps from each branch are then fed into their respective dynamic feature selection modules for precise extraction of crucial feature information. Additionally, $F_{C\times H\times W}$ is simultaneously sent to a 1×1 convolution branch and a global average pooling (GAP) branch to obtain their output feature maps.

In contrast, the MFA module addresses the limitations of existing multi-scale feature fusion methods. Currently, mainstream methods often use dilated convolutions with different dilation rates to achieve multi-scale feature fusion but lack dynamic regulation of the importance of different scale features. Building on this, we introduce an attention mechanism to effectively allocate and aggregate weights for multi-scale features. This improvement allows the model to more precisely capture contextual information, significantly enhancing its representation capability in complex scenarios, thereby improving the robustness and accuracy of the model.

To address the weakening of target area representation caused by direct Multi-scale Feature Aggregation, a dynamic feature selection mechanism is introduced. This mechanism enables the model to concentrate on extracting key information from multi-scale features and is inspired by the findings in study [30].

The dynamic feature selection mechanism comprises two key components: spatial attention and channel attention mechanisms.

The spatial attention mechanism takes the input feature map $P \in \mathbb{R}^{C\times H\times W}$ and processes it through a convolutional layer to generate two feature maps, K and V. These are reshaped, fused using matrix multiplication, and passed through a softmax layer to compute the spatial relationships between features. Meanwhile, the channel attention mechanism processes the input feature map C with a 1×1 convolution, resulting in three feature maps, Q, K, and V. These maps are reshaped, combined through matrix multiplication, and refined with a softmax layer to capture the dependencies between channels. Together, these attention mechanisms optimize feature selection and information extraction, contributing to improved model performance and accuracy.

The multi-scale feature aggregation module integrates features from different levels of the network, enabling the model to retain and utilize information across varying scales, which is critical for distinguishing PV panels from complex rooftop environments.

3.5. Loss Function

In rooftop photovoltaic detection applications, the commonly used Binary Cross-Entropy (BCE) loss function faces challenges due to the significantly smaller number of photovoltaic area pixels compared to the background pixels in high-resolution scenarios within the dataset. Using the standard BCE loss during training can cause the model to disproportionately focus on background pixels, hindering its ability to learn features specific to the target area. This imbalance may reduce the model's effectiveness in rooftop photovoltaic detection. To address this limitation, we adopt a weighted BCE loss function. By introducing weight coefficients, this loss function adjusts the emphasis between the photovoltaic regions and the background, ensuring a more balanced learning process. The weighted BCE loss is defined as follows:

$$\begin{aligned} L_{wbce}(W) = &- w_0 \sum_{j\in Y_+} \log \Pr\Big(y_j = 1 \mid X, W\Big) \\ &- w_1 \sum_{j\in Y_-} \log \Pr\Big(y_j = 0 \mid X, W\Big). \end{aligned} \tag{6}$$

Specifically, the weight $w_0 = \frac{|Y_-|}{|Y_+|}$ is defined for the photovoltaic area pixels, while the weight for the background area w_1 is set to 1. In this context, $|Y_+|$ and $|Y_-|$ denote the total number of photovoltaic and non-photovoltaic pixels across the entire training dataset, respectively.

4. Experimental Results and Analysis

4.1. Implementation Details

The model was developed with the assistance of MMSegmentation 0.25.0, a specialized toolkit for semantic segmentation that is publicly available and based on PyTorch, developed by Meta Platforms, Inc., located in Menlo Park, CA, USA.

The testing environment comprised a server outfitted with a single NVIDIA RTX 4090 GPU, manufactured by NVIDIA Corporation located in Santa Clara, CA, USA, which is equipped with 24 GB of memory. A detailed summary of the experimental settings and configurations is presented in Table 1. In addition to these general settings, the following experimental details are provided: Batch size—the batch size was set to 8 to balance computational efficiency and memory usage. Number of iterations—the model was trained for 160,000 iterations, with an early stopping strategy applied if no parameter updates occurred within 500 iterations to prevent overfitting. Learning rate schedule—the initial learning rate was decayed by 90% every 500 iterations, enabling a gradual convergence to the optimal parameters. Data augmentation—during training, random data augmentation strategies were employed to improve model robustness. These strategies included random flipping and random rotation of the training images, which helped mitigate overfitting and enhance generalization.

Table 1. Experiments settings.

Category	Item	Configuration
Hardware	GPU	RTX 4090 × 1
	Python	3.8.16
	CUDA	11.3
Environment Config.	Pytorch	1.11.0
	MMSegmentation	0.25.0
	Learning rate	0.0001
	Weight decay	0.01
	Loss function	BCE loss function
	Optimizer	AdamW

4.2. Datasets

The dataset used in the experimental section was sourced from Zenodo [26] and included various samples of photovoltaic panels. These samples were primarily composed of images taken by the GaoFen-2 and Beijing-2 satellites, as well as drones located in Jiangsu Province, China. The dataset was divided into three subsets corresponding to spatial resolutions of 0.8 m, 0.3 m, and 0.1 m, named PV08, PV03, and PV01, respectively.

In the PV08 subset, the samples were further categorized into two main classes based on the installation location of the photovoltaic panels: rooftop- and ground-mounted. The PV03 subset was more detailed, including classifications such as rooftop photovoltaic, shrubland photovoltaic, grassland photovoltaic, farmland photovoltaic, saline–alkali photovoltaic, and water surface photovoltaic. The PV01 subset classified panels based on the material of the rooftop (such as concrete, steel tiles, and bricks).

Since this study focuses primarily on detecting rooftop photovoltaics, the rooftop photovoltaic scenarios from the three datasets were extracted and merged during the experimental process. After merging the scenes from the three resolutions, a random dataset

split was performed, resulting in 776 training images and 195 testing images. Therefore, we abbreviate the dataset as rooftop pv. The image resolutions for both PV03 and PV08 were 1024 × 1024 pixels, while the resolution for the PV01 subset was 256 × 256 pixels. Additionally, to match the network input requirements, all images were uniformly resized to 256 × 256 pixels.

4.3. Comparison Methods

(1) U-Net [31]: U-Net is a deep learning framework widely used for image segmentation. It features an encoder–decoder structure with skip connections, which effectively combines high-resolution spatial features with low-resolution contextual information. This design makes it highly effective in applications such as medical image segmentation.
(2) DeepLabV3+ [32]: DeepLabV3+ is an image segmentation model that leverages atrous convolution and pyramid pooling to extract multi-scale contextual features. This approach enhances segmentation accuracy by capturing fine details and broader contextual relationships.
(3) HRNet [33]: The High-Resolution Network (HRNet) maintains high-resolution representations throughout its architecture, enabling precise multi-scale feature integration. It has demonstrated strong performance in image segmentation and object detection tasks.
(4) PSPnet [34]: The Pyramid Scene Parsing Network (PSPNet) introduces a pyramid pooling module to effectively capture global context information at various scales. This capability enhances its segmentation performance in complex and diverse scenes.
(5) Segformer [35]: Segformer is an efficient image segmentation model that integrates transformer-based and convolutional approaches. Its hybrid architecture balances accuracy and computational efficiency, achieving high performance across various segmentation tasks.
(6) Beit [36]: Beit is a vision transformer-based model designed for image segmentation and other vision tasks, leveraging masked image modeling pretraining to achieve high accuracy, particularly in large-scale datasets, though at the cost of increased computational complexity.
(7) Maskformer [37]: Maskformer is a transformer-based model that unifies segmentation tasks by predicting a set of binary masks with associated labels, effectively bridging the gap between instance, semantic, and panoptic segmentation, while maintaining high accuracy and adaptability across diverse segmentation scenarios.
(8) Mask2former [38]: Mask2Former is a versatile segmentation framework that leverages a transformer architecture to generate masks for both object-level and pixel-level segmentation tasks in a unified manner, achieving high performance on various benchmark datasets.

4.4. Evaluation Metrics

In this study, we utilize three widely recognized metrics—*Precision* (P), *Recall* (R), and the *F*1-*score* (F_1)—to evaluate the performance of the models. *Precision* represents the proportion of correctly identified positive predictions out of all positive predictions made by the model, reflecting its accuracy in distinguishing positive cases. *Recall*, in contrast, measures the proportion of correctly identified positive instances relative to the total number of actual positives in the dataset, highlighting the model's ability to capture all relevant instances. The *F*1-*score*, calculated as the harmonic mean of precision and recall, balances the trade-off between these two metrics, offering a comprehensive measure of the model's overall performance.

$$Precision\ (P) = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall\ (R) = \frac{TP}{TP + FN}, \tag{8}$$

$$F_1 = \frac{2 \times P \times R}{P + R}. \tag{9}$$

In this context, TP represents true positives, FP denotes false positives, and FN corresponds to false negatives. This study formulates photovoltaic (PV) detection as a binary semantic segmentation task, aiming to differentiate PV regions from the background. To assess the performance of the models in this task, we utilize three commonly used semantic segmentation metrics: the mean Dice coefficient ($mDice$), mean accuracy ($mAcc$), and mean intersection over union ($mIoU$).

$$mDice = \frac{2TP}{2TP + FP + FN}, \tag{10}$$

$$mAcc = \frac{1}{K+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij}}, \tag{11}$$

$$mIoU = \frac{1}{K+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{i=0}^{k} P_{ij} + \sum_{i=0}^{k} P_{ji} - P_{ii}}, \tag{12}$$

Here, K denotes the total number of classes (in this case, $K = 2$ for crack and non-crack), and p_{ij} represents the number of pixels from class i that are predicted to belong to class j.

4.5. Experimental Results

Figure 5 illustrates the training loss comparison across various methods over the training steps, highlighting differences in convergence speed and optimization efficiency. Among all methods, our proposed method ("Ours") demonstrates a significantly faster convergence rate, achieving stable training loss at approximately 8000 steps, whereas other methods, such as U-Net, DeeplabV3plus, and PSPNet, require approximately 20,000 to 40,000 steps to reach similar levels of convergence. This faster convergence indicates that our method efficiently learns meaningful features early in the training process, likely attributed to its enhanced architectural design and robust optimization strategies.

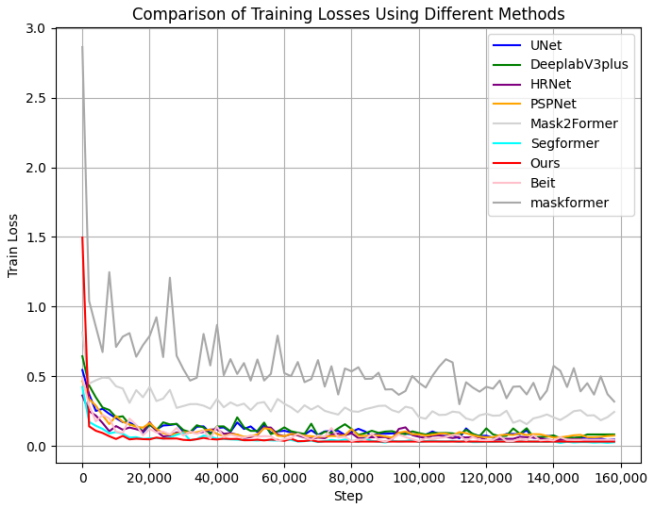


Figure 5. The training loss values of different methods are presented.

While methods like U-Net, DeeplabV3plus, and PSPNet show competitive performance with relatively low losses, their slower convergence makes them less efficient compared to our method. Mask2Former and Beit exhibit higher and more fluctuating losses throughout training, suggesting their increased complexity may hinder stable optimization. Segformer demonstrates moderate stability during training but with consistently higher final loss values compared to our method.

Overall, these results highlight the superiority of our method in terms of convergence speed and training efficiency, requiring fewer steps to achieve lower loss while maintaining robust performance, making it a more reliable and efficient approach for the task.

The combined Figure 6 compares the performance of different methods across three key metrics: mAcc, mFscore, and mIoU, highlighting the strengths of our proposed method ("Ours"). Across all metrics, our method demonstrates faster convergence, higher stability, and superior performance.

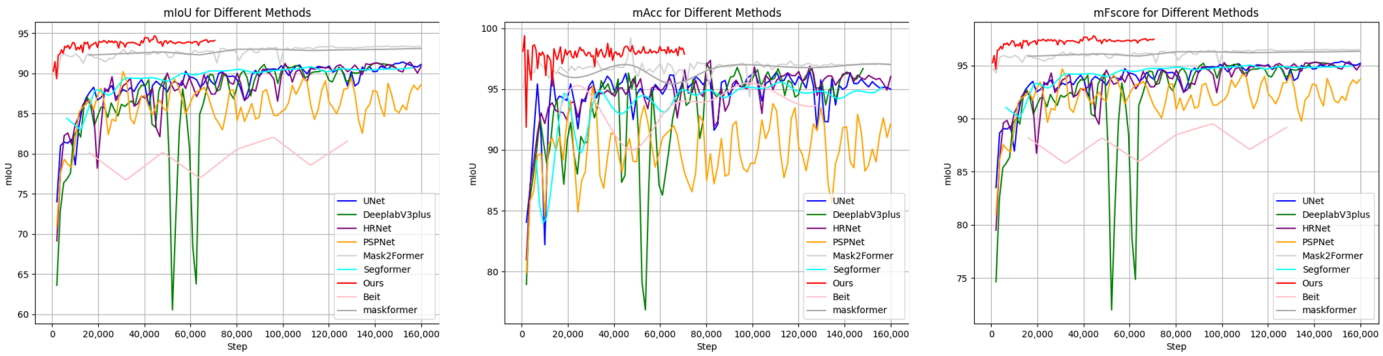


Figure 6. The mIoU, mAcc, and mFscore performance metrics across different methods.

For mAcc, our method rapidly achieves the highest accuracy and maintains it consistently, outperforming U-Net, DeeplabV3plus, and HRNet, while PSPNet and transformer-based models like Mask2Former and Beit exhibit significant fluctuations and lower performance. In terms of mFscore, our method achieves a strong balance between precision and recall, leading other methods with faster convergence and higher stability. Lastly, in mIoU, our method achieves higher IoU values early in training and sustains them better than competitors. PSPNet and Mask2Former show erratic fluctuations, while Beit and Segformer struggle to stabilize. Overall, the results clearly demonstrate that our method excels in efficiency, robustness, and overall performance compared to traditional and transformer-based models, making it well suited for practical applications.

As illustrated in Table 2, we evaluated various deep learning models on PV rooftop detection using key metrics: Precision, Recall, F1-score, mAcc, mIoU, and mDice. Our proposed model achieved the best performance across all metrics, with a Precision of 96.56%, Recall of 97.10%, F1-score of 97.75%, mAcc of 98.39%, mIoU of 94.15%, and mDice of 96.91%. Notably, it outperformed all competing methods in F1-score, mAcc, and mDice, demonstrating its robustness, accuracy, and superior spatial prediction capabilities.

The strong performance of our method is attributed to the use of the multi-scale Res2Net architecture and the innovative Spatial Feature Reconstruction (SFR) and Multi-scale Feature Aggregation (MFA) modules. The Res2Net design enables enhanced multi-scale feature representation, capturing both fine-grained details and global context. The SFR module effectively reduces redundant information, improving Spatial Feature Reconstruction, while the MFA module dynamically fuses multi-scale features with attention mechanisms, leading to precise contextual understanding and better segmentation results. Among competing models, HRNet and SwinLarge performed well in Recall (both exceeding 96%), showing their ability to minimize false negatives, which is essential for sensitive tasks

like medical imaging. HRNet also achieved a strong F1-score of 95.54%, demonstrating a balanced trade-off between Precision and Recall. PSPNet led in mIoU (92.39%), excelling in spatial overlap accuracy, but its overall performance fluctuated, likely due to its less robust handling of complex spatial features. MaskFormer showed consistency with a high mDice of 96.29%, reflecting its ability to handle positive class predictions, though it fell short in other metrics. SegFormer, while stable, achieved moderate results, and U-Net and DeepLabV3+ struggled to surpass 93% in any metric, highlighting limitations in capturing multi-scale and contextual information. In conclusion, our method's integration of advanced multi-scale feature representation and fusion techniques ensures faster convergence, higher accuracy, and superior spatial consistency, making it highly suitable for practical PV rooftop detection applications. These results underscore the importance of robust architectures and innovative modules in achieving state-of-the-art performance.

Table 2. Performance metrics of various models for PV detection.

Model	Precision	Recall	F1-Score	mAcc	mIoU	mDice
U-Net [31]	92.96	93.26	93.11	93.26	87.62	93.11
DeepLabv3+ [32]	94.81	95.99	95.39	95.99	91.43	95.39
HRNet [33]	94.77	96.36	95.54	96.36	91.70	95.54
PSPNet [34]	96.12	95.76	95.94	95.76	92.39	95.94
SegFormer [35]	93.58	93.70	93.64	93.70	88.49	93.64
Beit [36]	81.17	93.52	85.93	93.52	76.94	85.93
MaskFormer [37]	96.10	96.50	96.30	96.50	93.02	96.30
Mask2Former [38]	95.52	97.10	96.29	97.10	93.01	96.29
Ours	96.56	97.10	97.75	98.39	94.15	96.91

As shown in Table 3, our improved model is based on the Res2Net architecture, incorporating Spatial Feature Reconstruction and multi-scale feature aggregation modules. It achieves excellent real-time performance with 21.24 FPS while maintaining moderate computational complexity (42.18 G Flops) and parameter count (34.25 M). Compared to lightweight methods (e.g., U-Net and Segformer), our model significantly enhances feature representation and accuracy while maintaining high efficiency. Compared to complex methods (e.g., Mask2Former and BEiT), it greatly reduces computational cost and resource requirements, achieving a balance between efficiency and accuracy, making it suitable for tasks that require both real-time performance and precision.

Table 3. Comparison of Flops, parameters, and FPS across methods.

Method	Flops	Params	FPS
U-Net [31]	38.35 G	28.99 M	28.35
DeepLabv3+ [32]	48.67 G	60.21 M	15.42
HRNet [33]	17.94 G	65.85 M	12.09
PSPNet [34]	34.23 G	46.61 M	9.48
Segformer [35]	13.52 G	37.16 M	17.97
Mask2former [38]	125.43 G	184.52 M	6.05
Beit [36]	96.27 G	161.38 M	6.89
Maskformer [37]	79.41 G	148.54 M	12.61
Ours	42.18 G	34.25 M	21.24

To highlight the superiority of our proposed method, we present the results of different methods at various resolutions in Figure 7. Compared to the other evaluated approaches, our method demonstrates a more effective ability to extract target features.

Overall, the model introduced in this paper significantly enhances performance in the domain of rooftop PV detection through targeted improvements specific to rooftop

PV features, demonstrating its strong potential and practical value in handling complex detection tasks.

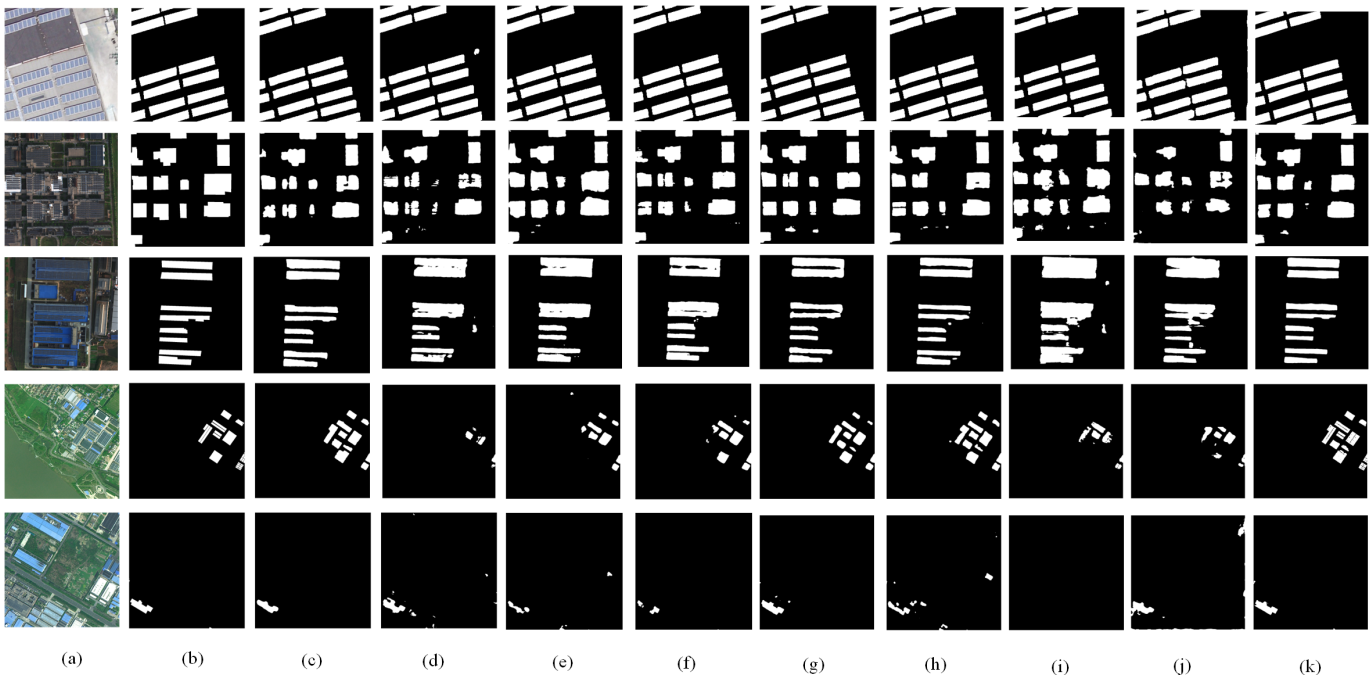


Figure 7. Visualization results obtained by various methods selected from the Rooftop PV dataset. From left to right column: (**a**) raw image; (**b**) ground truth; (**c**) Ours (**d**); U-Net [31]; (**e**) DeepLabv3+ [32]; (**f**) HRnet [33]; (**g**) Mask2Former [38]; (**h**) PSPNet [34]; (**i**) SegFormer [35]. (**j**) Beit [36]; (**k**) MaskFormer [37].

Ablation Studies

To verify the effectiveness of the proposed modules, we performed ablation experiments to evaluate the performance improvements introduced by each component. Using the U-Net architecture as the baseline model (denoted as Baseline, B), we incrementally integrated the enhanced backbone network Res2Net, the Multi-scale Feature Aggregation (MFA) module, and the Spatial Feature Reconstruction (SFR) module to assess their individual contributions. As shown in Table 4, the baseline model achieved the lowest scores across the three key metrics: F1-score, mAcc, and mIoU, recording values of 93.11, 93.26, and 87.62, respectively. By incrementally introducing the aforementioned modules, the experimental data demonstrated that each of the three modules significantly enhanced the performance of the baseline model. Notably, the MFA module exhibited the largest improvement, with the F1-score, mAcc, and mIoU scores improving to 94.99, 96.75, and 92.89, respectively. This confirms that MFA effectively enhances the model's capability to learn features across different scales. Additionally, the proposed Spatial Feature Reconstruction module also significantly improved the detection accuracy of the baseline network.

To further investigate the synergistic effects of combining different modules, we conducted experiments with various pairwise combinations, demonstrating that the modules could complement and enhance each other's performance. Notably, the combination of SFR and MFA yielded the most substantial improvements, achieving F1-score, mAcc, and mIoU values of 96.77, 98.09, and 93.97, respectively. Finally, by integrating all three modules into the baseline model to form the complete framework proposed in this study, the results showed the highest performance across all metrics. The comprehensive model achieved an F1-score of 97.75, mAcc of 98.39, and mIoU of 94.15. These results validate the significant impact and practical utility of the proposed modules in enhancing rooftop photovoltaic detection accuracy.

Table 4. Ablation analysis for the proposed architecture on rooftop PV datasets.

Methods	PV Dataset					
	Precision	**Recall**	**F1-Score**	**mAcc**	**mIoU**	**mDice**
BaseLine (B) [31]	92.96	93.26	93.11	93.26	87.62	93.11
B+Res2Net	93.45	93.79	93.62	94.64	92.95	93.78
B+SFR	94.06	94.67	94.36	95.42	92.38	94.24
B+MFA	94.67	95.32	94.99	96.75	92.89	94.76
B+Res2Net+SFR	95.38	95.94	95.66	97.34	93.47	95.38
B+Res2Net+MFA	95.77	96.14	95.95	97.72	93.58	95.78
B+MFA+SFR	96.12	97.42	96.77	98.09	93.97	96.42
Ours	97.56	97.10	97.75	98.39	94.15	96.91

As shown in Figure 8, the results clearly demonstrate the effectiveness of the proposed modules in improving the segmentation of rooftop photovoltaic (PV) panels. The baseline model struggles with feature extraction, showing poor focus and significant background noise. The addition of Res2Net enhances multi-scale feature extraction, while MFA improves feature aggregation and SFR refines spatial reconstruction, each addressing specific challenges. Combining MFA and SFR shows further improvement by synergistically enhancing feature focus and reducing background interference. The integration of all modules (Res2Net, MFA, and SFR) achieves the best results, delivering precise and robust segmentation by effectively capturing multi-scale details and reconstructing spatial features. This highlights the complementary strengths of the proposed framework in addressing resolution, scale, and foreground–background imbalance issues.

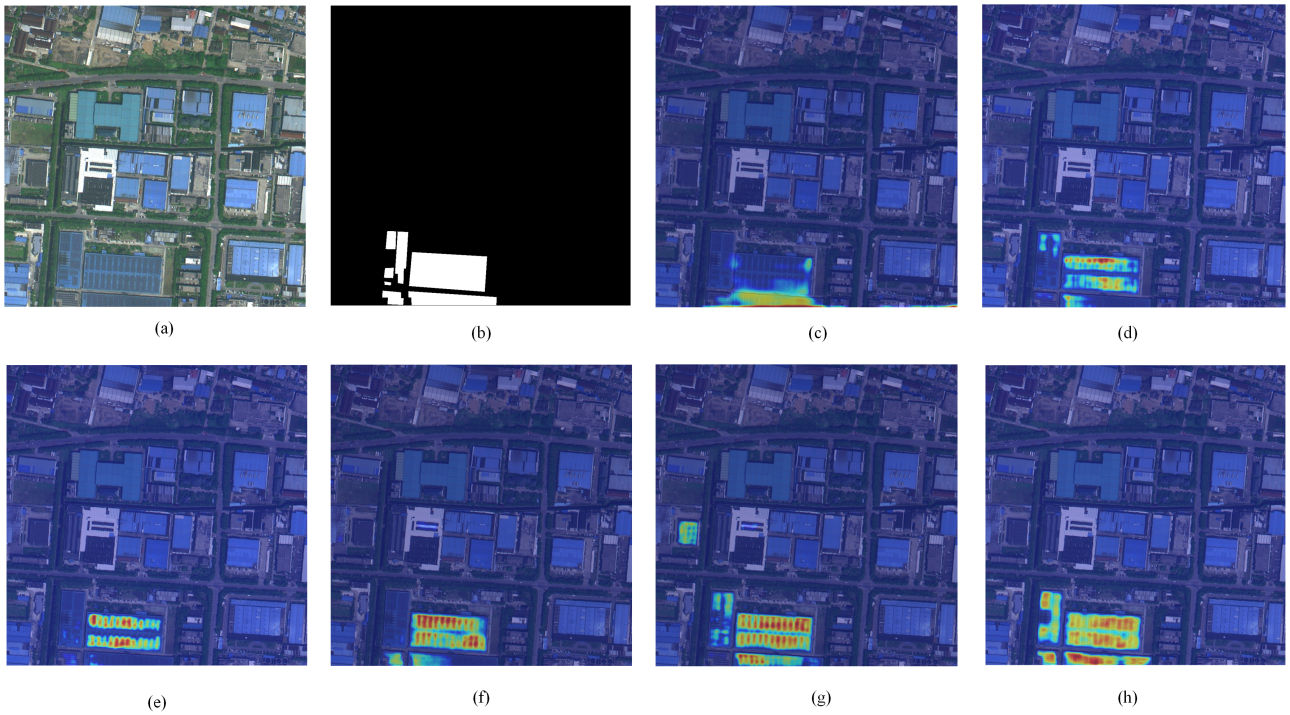


Figure 8. The visualized features map of the proposed different modules are shown, where (**a**,**b**) represent the original image and the label map, respectively. (**c**) represents the visualized feature map of the Baseline. (**d**–**f**) represent the visualized feature maps after applying Res2Net, MFA, and SFR, respectively. (**g**) represents the visualized feature map after combining MFA and SFR, and (**h**) represents the visualized feature map with all modules combined.

5. Failure Cases and Limitations

While the proposed method demonstrates strong performance in segmenting rooftop photovoltaic (PV) panels, some limitations remain, particularly in high-resolution images with small-scale PV targets. As shown in Figure 9, failure cases often occur in scenarios involving small PV panels occupying a minor fraction of the image. These panels are frequently missed or only partially segmented due to challenges such as scale imbalance, loss of fine-grained details during feature extraction, and insufficient spatial context. Additionally, in some instances, false positives are observed in regions with textures or reflectance similar to PV panels, such as metallic or glossy surfaces on rooftops.

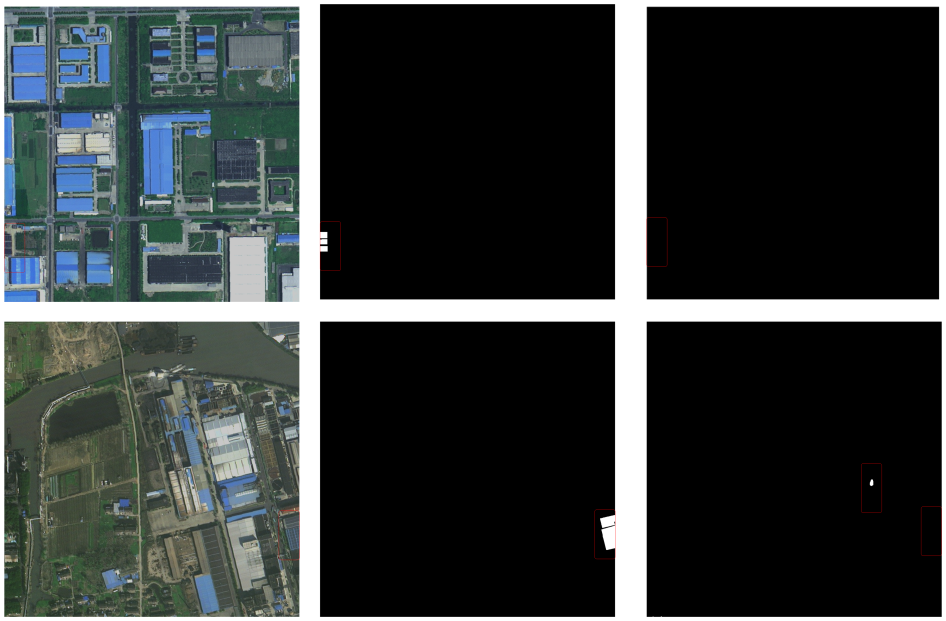


Figure 9. Failure cases of our method in high-resolution scenarios.

These limitations primarily arise from the foreground–background imbalance, where small PV panels are underrepresented during training, and the resolution loss caused by down-sampling in the network. To address these issues, potential solutions include enhancing multi-scale feature extraction using attention mechanisms, adopting balanced loss functions (e.g., focal loss) to better handle small targets, and augmenting the dataset with diverse examples of small or occluded PV panels. These improvements could further enhance the model's robustness and performance in challenging scenarios.

Limitation Statement: Due to resource and time constraints, this study was limited to a single dataset for evaluating the proposed method. While the results demonstrate the effectiveness of our approach, its generalizability to datasets with diverse geographical conditions, rooftop materials, and environmental backgrounds remains to be validated. Additionally, a detailed sensitivity analysis of key hyperparameters was not conducted in this study, which may affect the understanding of the model's robustness and stability under different configurations. Future research will aim to address these limitations by testing the method on more diverse datasets and conducting comprehensive sensitivity analyses to further assess its applicability and reliability in real-world scenarios.

6. Conclusions

This paper presents a novel model tailored for detecting rooftop photovoltaic (PV) panels across multiple resolutions. Initially, we employ a more robust backbone network, ResNet, and introduce a Spatial Feature Reconstruction (SFR) module. This module effectively filters out spatially redundant features and refines key characteristics of rooftop

PV panels. Additionally, we integrate a Multi-scale Feature Aggregation (MFA) module into the middle layers of the network. This module effectively addresses scale variations in rooftop PV panels at different resolutions, overcoming the limitations of traditional methods that directly concatenate fixed receptive field feature maps. By independently processing key features through a dynamic feature selection module, we significantly enhance the efficiency and effectiveness of feature aggregation. Our evaluation on a rooftop PV panel dataset demonstrates the effectiveness of our improved approach, highlighting its important contribution to the management and planning of regional distributed PV energy systems.

In future work, we plan to explore lightweight transformer-based models, such as MobileViT or TinyTransformer, to enhance computational efficiency and enable deployment on resource-constrained devices. Additionally, we aim to integrate more data modalities, including thermal imagery for low-light scenarios, LiDAR for high-precision depth information, and other modalities like acoustic or multispectral data to expand applicability. We will also investigate advanced cross-modal fusion methods, such as self-supervised learning and multimodal Transformers, to maximize the potential of multimodal data. Lastly, we intend to validate the improved model through extensive real-world experiments under diverse conditions.

Author Contributions: Conceptualization, Y.X. and M.B.; methodology, Y.X. and L.L.; validation, Y.X. and J.M.; formal analysis, Y.X.; investigation, J.M.; data curation, Y.X. and L.L.; writing—original draft, Y.X.; writing—reviewing and editing, Y.X. and M.B.; writing—original draft, Y.X. and L.L.; visualization, Y.X. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Research Project of Chongqing Education Commission KJQN202303812.

Data Availability Statement: The dataset that supports the findings of this study is openly available at https://zenodo.org/records/5171712 (accessed on 13 August 2024). The authors confirm that the data supporting the findings of this study are available from the corresponding author [Maoqiang Bi], upon reasonable request, and the data can also be obtained from reference materials.

Acknowledgments: The authors wish to thank the associate editors and anonymous reviewers for their valuable comments and suggestions on this paper.

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

References

1. SolarPower Europe. *Global Market Outlook for Solar Power 2023–2027*; Technique Report; SolarPower Europe: Brussels, Belgium, 2023.
2. Qi, Q.; Zhao, J.; Tan, Z.; Tao, K.; Zhang, X.; Tian, Y. Development assessment of regional rooftop photovoltaics based on remote sensing and deep learning. *Appl. Energy* **2024**, *375*, 124172. [CrossRef]
3. 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. [CrossRef]
4. Zhong, Q.; Nelson, J.R.; Tong, D.; Grubesic, T.H. A spatial optimization approach to increase the accuracy of rooftop solar energy assessments. *Appl. Energy* **2023**, *316*, 119128. [CrossRef]
5. Lu, R.; Wang, N.; Zhang, Y.; Lin, Y.; Wu, W.; Shi, Z. Extraction of agricultural fields via DASFNet with dual attention mechanism and multi-scale feature fusion in South Xinjiang, China. *Remote Sens.* **2022**, *14*, 2253. [CrossRef]
6. Qian, Z.; Chen, M.; Zhong, T.; Zhang, F.; Zhu, R.; Zhang, Z.; Zhang, K.; Sun, Z.; Lü, G. Deep roof refiner: Adetail-oriented deep learning network for refined delineation of roof structure lines using satellite imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102680. [CrossRef]
7. Li, L.; Lu, N.; Qin, J. Joint-task learning framework with scale adaptive and position guidance modules for improved household rooftop photovoltaic segmentation in remote sensing image. *Appl. Energy* **2025**, *377*, 124521. [CrossRef]
8. Di Giovanni, G.; Rotilio, M.; Giusti, L.; Ehtsham, M. Exploiting building information modeling and machine learning for optimizing rooftop photovoltaic systems. *Energy Build.* **2024**, *313*, 114250. [CrossRef]

9. Satpathy, P.R.; Ramacharamurthy, V.K.; Roslan, M.F.; Motahhir, S. An adaptive architecture for strategic Enhancement of energy yield in shading sensitive Building-Applied Photovoltaic systems under Real-Time environments. *Energy Build.* **2024**, *324*, 114877. [CrossRef]
10. Aljafari, B.; Satpathy, P.R.; Thanikanti, S.B.; Nwulu, N. Supervised classification and fault detection in grid-connected PV systems using 1D-CNN: Simulation and real-time validation. *Energy Rep.* **2024**, *12*, 2156–2178. [CrossRef]
11. Sulaiman, M.H.; Jadin, M.S.; Mustaffa, Z.; Daniyal, H.; Azlan, M.N.M. Short-term forecasting of rooftop retrofitted photovoltaic power generation using machine learning. *J. Build. Eng.* **2024**, *94*, 109948. [CrossRef]
12. Malof, J.M.; Hou, R.; Collins, L.M.; Bradbury, K.; Newell, R. Automatic solar photovoltaic panel detection in satellite imagery. In Proceedings of the 2015 International Conference on Renewable Energy Research and Applications (ICRERA), Palermo, Italy, 22–25 November 2015; pp. 1428–1431.
13. 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.
14. Golovko, V.; Bezobrazov, S.; Kroshchanka, A.; Sachenko, A.; Komar, M.; Karachka, A. Convolutional neural network based solar photovoltaic panel detection in satellite photos. In Proceedings of the 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, Romania, 21–23 September 2017; pp. 14–19.
15. Yan, L.; Zhu, R.; Kwan, M.P.; Luo, W.; Wang, D.; Zhang, S. Estimation of urban-scale photovoltaic potential: A deep learningbased approach for constructing three-dimensional building models from optical remote sensing imagery. *Sustain. Cities Soc.* **2023**, *93*, 104515. [CrossRef]
16. Nasrallah, H.; Samhat, A.E.; Shi, Y.; Zhu, X.X.; Faour, G.; Ghandour, A.J. Lebanon solar rooftop potential assessment using buildings segmentation from aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 18. [CrossRef]
17. Cui, W.; Peng, X.; Yang, J.; Yuan, H.; Lai, L.L. Evaluation of rooftop photovoltaic power generation potential based on deep learning and high-definition map image. *Energies* **2023**, *16*, 6563. [CrossRef]
18. Krapf, S.; Kemmerzell, N.; Khawaja Haseeb Uddin, S.; Hack Vazquez, M.; Netzler, F.; Lienkamp, M. Towards scalable economic photovoltaic potential analysis using aerial images and deep learning. *Energies* **2021**, *14*, 3800. [CrossRef]
19. Lin, S.; Zhang, C.; Ding, L.; Zhang, J.; Liu, X.; Chen, G.; Wang, S.; Chai, J. Accurate recognition of building rooftops and assessment of long-term carbon emission reduction from rooftop solar photovoltaic systems fusing GF-2 and multi-source data. *Remote Sens.* **2022**, *14*, 3144. [CrossRef]
20. Chen, S.; Shi, W.; Zhou, M.; Zhang, M.; Xuan, Z. CGSANet: A contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, 15, 1526–1542. [CrossRef]
21. Khan, S.D.; Alarabi, L.; Basalamah, S. An encoder–decoder deep learning framework for building footprints extraction from aerial imagery. *Arab. J. Sci. Eng.* **2023**, *48*, 1273–1284. [CrossRef]
22. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
23. Chen, J.; Jiang, Y.; Luo, L.; Gong, W. ASF-Net: Adaptive screening feature network for building footprint extraction from remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
24. Mei, J.; Li, R.J.; Gao, W.; Cheng, M.M. CoANet: Connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* **2021**, *30*, 8540–8552. [CrossRef] [PubMed]
25. 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.
26. Costa, M.V.C.V.D.; Carvalho, O.L.F.D.; Orlandi, A.G.; Hirata, I.; Albuquerque, A.O.D.; Silva, F.V.E.; Guimarães, R.F.; Gomes, R.A.T.; Júnior, O.A.D.C. Remote sensing for monitoring photovoltaic solar plants in Brazil using deep semantic segmentation. *Energies* **2021**, *14*, 2960. [CrossRef]
27. Wang, J.; Chen, X.; Jiang, W.; Hua, L.; Liu, J.; Sui, H. PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *119*, 103309. [CrossRef]
28. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
29. Li, J.; Wen, Y.; He, L. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6153–6162.
30. Zhou, Q.; Qu, Z.; Li, Y.X. Tunnel crack detection with linear seam based on mixed attention and multiscale feature fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–11. [CrossRef]

31. 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: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
32. 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. 801–808.
33. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *40*, 3349–3364. [CrossRef] [PubMed]
34. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
35. Xie, E.; Wang, W.; Yu, Z.; An kumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
36. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
37. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; pp. 17864–17875.
38. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.

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