Extracting Photovoltaic Panels From Heterogeneous Remote Sensing Images With Spatial and Spectral Differences

Zhiyu Zhao ⓘ, Yunhao Chen ⓘ, Kangning Li ⓘ, Weizhen Ji ⓘ, and Hao Sun ⓘ

Abstract—The accurate extraction of the installation area of the photovoltaic power station is an important basis for the management of the photovoltaic power generation system. Deep learning has proven to be a powerful tool for rapidly detecting the distribution of photovoltaic panels in remote sensing images. The wealth of information from various remote sensing sensors aids in distinguishing photovoltaic pixels within complex backgrounds. However, the distinct imaging characteristics of different sensors present challenges for deep learning models. In this article, we propose a deep learning extraction method for photovoltaic panels that effectively improves the spatial and spectral differences inherent in remote sensing images. Considering the characteristics of different sensors, two attention modules and a feature fusion module are applied to suppress the inconsistency of spatial resolution and spectral resolution. Based on the Unet model, we implement the photovoltaic power station identification method and compare it with several commonly used semantic segmentation models. Qualitative and quantitative accuracy assessments show that the PV-Unet method can effectively overcome the spatial and spectral differences of remote sensing images. It achieves 98.04% F1 score and 96.15% IoU on the test dataset, verifying the superiority of this method. PV-Unet method has the potential for identifying photovoltaic panels from multisource remote sensing data.

Index Terms—Deep learning, feature extraction, photovoltaic (PV) panels, remote sensing, spatial and spectral differences.

I. INTRODUCTION

I N RECENT years, the global demand for renewable energy has increased significantly due to concerns about climate change and the depletion of traditional fossil fuels. The large-scale use of fossil fuels has led to the increase of greenhouse gas emissions, the rise of global temperature, and a series of environmental problems. To this end, many countries have formulated relevant policies to achieve carbon neutrality goals, trying to limit global warming to within 2 °C [1], [2], [3]. Among various renewable energy technologies, solar photovoltaic (PV) power stations have attracted wide attention for their ability to utilize solar energy and convert it into electric energy. With the reduction of PV equipment cost and the improvement of power generation efficiency, a large number of PV devices have been deployed, and the installed capacity of solar PV has increased rapidly [4], [5], [6].

The widespread adoption of PV panels requires effective methods for identification, monitoring and management. The estimation of power generation and carbon emission reduction depends on the accurate assessment of the area of PV devices. Obtaining the precise spatial distribution of PV devices is the basis for solving the impact of large-scale PV panels on the regional environment and ecosystem [7], [8]. The traditional method for monitoring PV panels requires manual inspection, which is time-consuming, labor-intensive, and often restricted by ground roads. Remote sensing images, with their comprehensive coverage and rapid imaging advantages, have become a powerful tool for identifying and analyzing PV panels. The data obtained by remote sensing technology is very large, whereas manual inspection requires a high cost of time and manpower. These limitations have prompted the introduction of advanced technologies, especially deep learning, to automatically identify PV panels from remote sensing images and improve efficiency [9].

Deep learning is a subset of machine learning, which has achieved remarkable success in various computer vision tasks, such as object detection, segmentation and classification. Convolutional neural networks are good at extracting hierarchical features from image data. These models can learn complex patterns and representations directly from raw image pixels, enabling them to accurately identify and classify objects in images. Fully convolutional networks can perform pixel-level classification on images, becoming one of the main methods for semantic segmentation tasks [10]. In addition, Transformer models based on self-attention have also achieved excellent results in the field of computer vision. However, Transformer models such as ViT

Zhiyu Zhao and Weizhen Ji are with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: zhaozhiyu@mail.bnu.edu.cn; wison@mail.bnu.edu.cn).

Yunhao Chen is with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China, and also with the Beijing Key Laboratory of Environment Remote Sensing and Digital Cities, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: cyh@bnu.edu.cn).

Kangning Li and Hao Sun are with the College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China (e-mail: kangning@cumtb.edu.cn; sunhao@cumtb.edu.cn).

(Vision Transformer) perform better than CNN on large datasets, but worse than CNN-based methods on small datasets [11], [12]. Therefore, for specific tasks with limited data, CNN is still a more practical method [13]. Semantic segmentation of remote sensing images is one of the hot topics in remote sensing image analysis. For example, Nong et al. proposed a boundary attention module based on the Unet architecture, which showed excellent performance on remote sensing image datasets [14]. Hou et al. proposed an enhanced semantic segmentation network to address the issues of complex background, large-scale variation, numerous small objects, and extreme foreground-background imbalance in remote sensing images [15]. Pixel-level classification of semantic segmentation is very effective for determining the location, boundary and area of PV panels [16]. For example, models such as SegNet, U-net, FPN, DeepLabV3+ have been proven to be applicable for PV power station extraction [13], [17], [18], [19], [20], [21]. Tan et al. proposed a texture-enhanced deep learning network to automatically extract solar panels in mining areas with high groundwater levels, achieving excellent performance [22]. Nasrallah et al. used deep learning-based case segmentation to extract a building's footprint from satellite images to assess rooftop solar potential [23].

These general deep learning frameworks are mainly designed to solve natural image recognition problems and have achieved good results in the field of computer vision, but lack the analysis of PV remote sensing image features. In previous studies, high spatial resolution remote sensing images were mainly relied on to achieve accurate identification of PV panels [24], and most of them used three-channel RGB images, which are visible light images that conform to the human eye's ability to distinguish target features [25]. Remote sensing sensors can acquire multispectral information, and these remote sensing images can also be used for target detection. The study in [20] shows that multisource sensors may provide more effective information, which helps to accurately identify the location and area of PV panels. The response of PV panels to multispectral, background features, especially the vegetation changes, are the unique characteristics of PV panels, and multispectral images are very important for distinguishing PV from background [7]. This information is difficult to obtain from single-source images, but it is valuable for deep learning models. By integrating images captured at different times and from various sensors, accurate identification of PV panels can be achieved [20], [26]. At present, obtaining high spatial resolution from multispectral satellite remote sensing data is challenging. Insufficient spatial resolution can result in a significant decline in segmentation accuracy [27]. Despite the progress in imaging equipment, satellite remote sensing images have higher and higher spatial resolution, temporal resolution, and spectral resolution. However, limited by the signal-to-noise ratio, sensors have difficulty in providing data with high spatial and spectral resolution simultaneously. To capture more spatial and spectral features of PV panels simultaneously, it is common to utilize satellite remote sensing data from various sources, such as Gaofen-2 and Sentinel-2. These sources can individually provide spatial and spectral features of PV panels. Therefore, a deep learning model is needed to accurately extract both the texture and spectral features of PV

panels from remote sensing images originating from different sources. However, due to differences in imaging conditions and variations in spatial resolution and spectral characteristics among images from different sources, challenges arise [28]. Considering the characteristics of different sensors, to obtain more texture and spectral information, a deep learning model that can overcome the spatial and spectral differences of remote sensing images is needed, to achieve accurate identification of PV panels.

To address the aforementioned challenges, this study leverages the attention mechanism to enhance both texture and spectral features. Our approach involves the fusion of diverse sensor data, aimed at mitigating the impact of heterogeneous image distribution disparities. Consequently, we propose an effective method for identifying PV panels, which adeptly amalgamates heterogeneous images. By synergizing the high-resolution Gaofen-2 imagery with the multispectral channels of Sentinel-2, we integrate spatial and spectral features pertaining to PV power plants. This integration serves to enhance the segmentation efficacy of our deep learning model. Building upon this refined model, we achieve precise segmentation of pixels corresponding to PV panels and their background counterparts through the input of multisource data.

II. MATERIALS AND METHOD

A. Materials

1) Data Sources: The remote sensing images are from Gaofen-2 (GF-2) and Sentinel-2. GF-2 belongs to the "China High-resolution Earth Observation System," and was launched in 2014. The satellite carries two high-resolution panchromatic cameras and multispectral cameras, which can obtain 1-m panchromatic and 4-m multispectral images respectively [29]. The multispectral camera can obtain images of four bands: blue, green, red, and near-infrared. Sentinel-2 belongs to the European Space Agency Copernicus program, which includes Sentinel-2A and Sentinel-2B satellites. The Sentinel-2 satellite carries a multispectral imager, which can obtain images of 13 bands, with spatial resolutions of 10 m for 4 bands, 20 m for 6 bands, and 60 m for 3 bands. It covers the spectral range of visible light, near-infrared and short-wave infrared.

The images used in this study are located in Ordos City, Inner Mongolia Autonomous Region, China, and the satellite data were taken in 2022. The city covers an area of about 87 000 km^2, belonging to the northern temperate semiarid continental climate zone, and is relatively dry [30]. The main parts of the region are the Kubuqi Desert and the Mu Us Desert, which have abundant PV resources. The GF-2 images used in the study were mainly taken from October to December, and the Sentinel-2 images were taken in July. The GF-2 images are from the China Center for Resources Satellite Data and Application. Cloud-free images with good quality were selected, and the panchromatic images were radiometrically calibrated and orthorectified; the multispectral images were radiometrically calibrated and atmospherically corrected and orthorectified. Then, the panchromatic and multispectral data were fused to obtain 1m resolution images containing four bands: blue, green, red and near-infrared. The

TABLE I
PARAMETERS FOR REMOTE SENSING IMAGERY

	Spectral number	Wavelength	Resolution
GF-2 (Pan Sharpening)	B1	450–520 nm	1 m
	B2	520–590 nm	1 m
	B3	630–690 nm	1 m
	B4	770–890 nm	1 m
Sentinel-2 (L2A)	B1	443.9 nm (S2A)/ 442.3 nm (S2B)	60 m
	B2	496.6 nm (S2A)/ 492.1 nm (S2B)	10 m
	B3	560 nm (S2A)/ 559 nm (S2B)	10 m
	B4	664.5 nm (S2A)/ 665 nm (S2B)	10 m
	B5	703.9 nm (S2A)/ 703.8 nm (S2B)	20 m
	B6	740.2 nm (S2A)/ 739.1 nm (S2B)	20 m
	B7	782.5 nm (S2A)/ 779.7 nm (S2B)	20 m
	B8	835.1 nm (S2A)/ 833 nm (S2B)	10 m
	B8A	864.8 nm (S2A)/ 864 nm (S2B)	20 m
	B9	945 nm (S2A)/ 943.2 nm (S2B)	60 m
	B11	1613.7 nm (S2A)/ 1610.4 nm (S2B)	20 m
	B12	2202.4 nm (S2A)/ 2185.7 nm (S2B)	20 m

Fig. 1. Part of the dataset displays in true color. (a) GF-2. (b) Sentinel-2. (c) Label.

Sentinel-2 images are L2A level data obtained from Google Earth Engine (GEE) and preprocessed. Compared with L1 data, there is no B10 band, so there are a total of 12 bands of data. The Sentinel-2 multispectral data of Ordos City in July 2022 were obtained through GEE, and cloud masking and mosaic processing were performed. The final remote sensing images are shown in Table I.

2) PV Sample Production: Before labeling the samples, the original images were checked and the ones with poor quality, which contained a lot of clouds or noise, were removed. In addition, each Gaofen image and Sentinel image was manually georegistered to minimize image offset. When labeling the samples, the PV power station range on the Gaofen and Sentinel images was confirmed to be consistent. When annotating the boundaries, mainly annotate on the Gaofen images with a higher spatial resolution to ensure clearer and more accurate boundaries. After the samples were annotated, the boundaries of the PV panels were checked and they were verified to be available on the Sentinel-2 images, ensuring the consistency of the samples on the two images. Finally, after resampling and cropping, sample images with a size of 512×512 were obtained. Part of the data visualization is shown in Fig. 1. The leftmost picture in the figure is from GF-2, the middle column is from Sentinel-2, and the rightmost column shows the annotation results. The remote sensing images are displayed in true color.
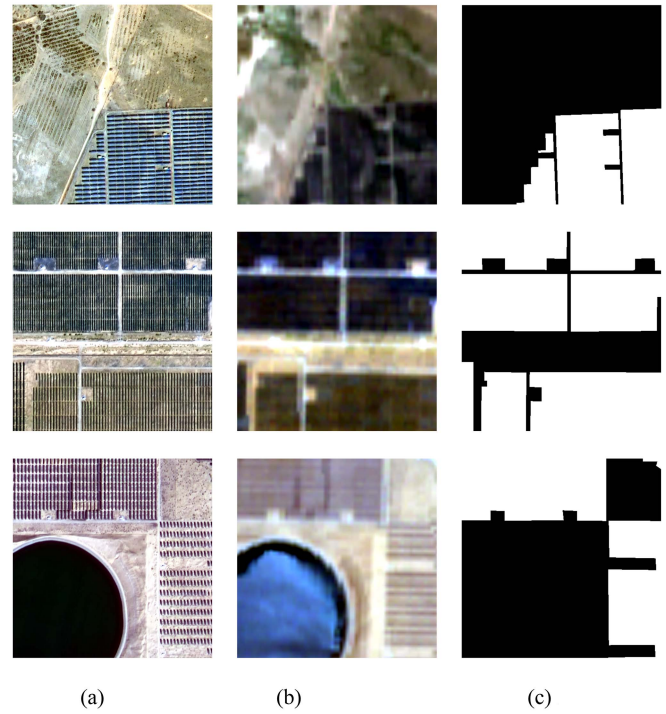
B. Method

1) Photovoltaic Power Station Extraction Framework: To leverage the distinctive features of GF-2 and Sentinel-2 images, a deep learning segmentation method has been proposed for the identification of PV panels. This method aims to overcome spatial and spectral differences present in remote sensing images from various sources. The flowchart of the proposed method is illustrated in Fig. 2.

The first step is the preprocessing of remote sensing data. High spatial resolution and multispectral remote sensing data were used. The data sources were explained in the previous section. The remote sensing data were processed by radiometric calibration, atmospheric correction, orthorectification, pansharpening, and cropping. The description of each band is shown in Table I.

The second step is the dataset preparation for model training. It includes georeferencing and sample annotation. The detailed annotation method was explained in the previous section.

The third step is the deep learning model for segmenting PV panels. The deep learning segmentation approach incorporates two feature extraction branches. The first branch captures texture features of PV panels from GF-2 images, whereas the second branch focuses on extracting spectral features from Sentinel-2 images. These features are efficiently fused using a dedicated feature fusion module (FFM). Built upon the architecture of a classic semantic segmentation model, specifically the Unet model, the proposed PV power station identification method is implemented.

2) Semantic Segmentation Model of Photovoltaic Power Plants: Unet model is a deep learning network, widely used for semantic segmentation. Unet was proposed in 2015, and since
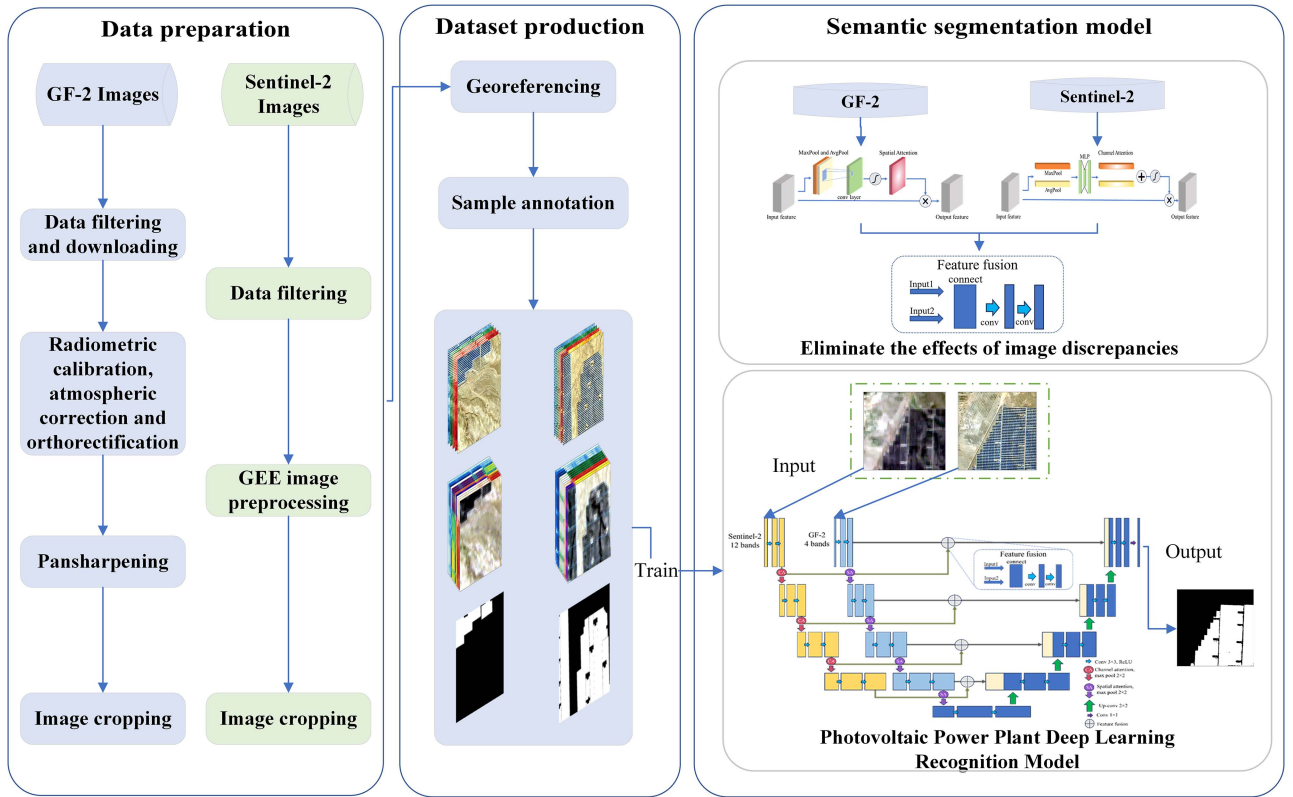
Fig. 2. Photovoltaic power station identification method to overcome the spatial and spectral differences of heterologous remote sensing images.

then it has been widely applied and achieved good results in various fields. Based on it, many improved models have been developed [31], [32], [33]. Unet is a deep learning network with an encoder-decoder structure. The encoder part extracts high-level semantic features from the input image through multiple convolutional blocks and outputs a feature map with low resolution and high channel number. The decoder part consists of multiple deconvolutional blocks, which restore the size of the feature map by upsampling and concatenate the feature map of the corresponding level of the encoder by skip connections, fusing features of different levels. Finally, it outputs a feature map with the same size as the original image, and reduces the channel number of the feature map to the number of target classes by a 1×1 convolution in the last layer [19], [31]. Based on this model, we proposed a PV power station segmentation model PV-Unet, which consists of spatial feature and spectral feature extraction modules and a FFM.

For the GF-2 image with higher spatial resolution, we are more concerned about its spatial relationship, whereas for the Sentinel-2 image with more channels and lower resolution, we should pay more attention to its channel features. Therefore, the attention mechanism is introduced into the model. Attention has been widely used in CNN models, by using the attention mechanism to focus on the main features in the CNN model and suppress unnecessary features, which helps to identify objects in cluttered and complex backgrounds [34], [35].

Convolutional block attention module (CBAM) [34] is a lightweight and general module that can be easily integrated into various CNN models. In our method, we inserted the spatial attention module into the high spatial resolution image feature extraction part and the channel attention module into the multispectral image feature extraction part. In CBAM, channel attention and spatial attention are performed sequentially. In our model, considering that the GF-2 image has rich spatial features, whereas the Sentinel-2 image has large differences among different channels, we divide CBAM into two parts, i.e., spatial attention module and channel attention module, and apply them to the extraction of texture features and spectral features of GF-2 and Sentinel-2 images respectively.

The spatial attention module is used to extract the high spatial resolution image features. The spatial attention map is generated by using the spatial relationship among the features. As shown in Fig. 3(a), in the spatial attention module, the input features are max-pooled and average-pooled along the channel axis. Then these results are concatenated to create an effective feature descriptor. Then, the spatial attention map is obtained by using a convolutional layer and a sigmoid activation function. The final spatial attention map $M_s(F)$ can be expressed by the following equation:

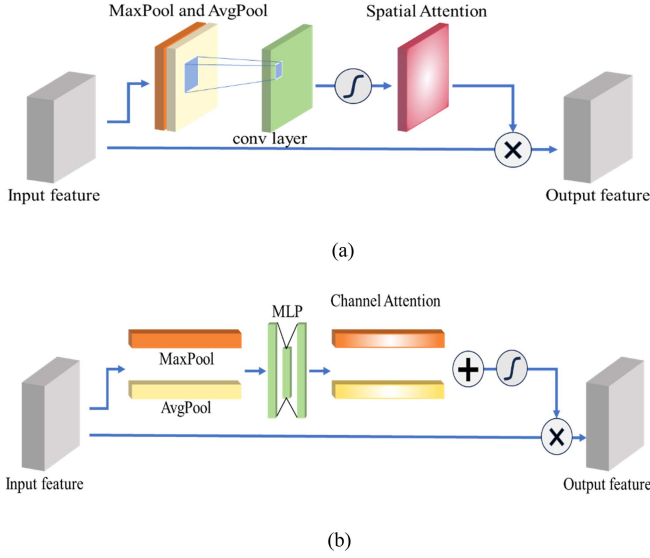$$M_s(F) = \sigma \left(f^{7\times7} \left([\text{AvgPool}(F); \text{MaxPool}(F)] \right) \right) \quad (1)$$

(a)

(b)

Fig. 3. Attention module. (a) Spatial attention module. (b) Channel attention module.

Fig. 4. PV-Unet model structure diagram.

where σ denotes the sigmoid function, $f^{7\times7}$ denotes the convolution operation with a kernel size of 7×7, AvgPool and MaxPool denote the average pooling and max pooling operations respectively, and F is the given input feature map. In each spatial attention module, the spatial attention map is multiplied by the weights of each channel of the input.

The channel attention module is used to extract features from the multispectral image. As shown in Fig. 3(b), first, the spatial features of the input feature map are aggregated by the max pooling layer and the average pooling layer respectively. Then, they are fed into a multilayer perceptron (MLP) with a hidden layer. Finally, the output feature vectors are merged by elementwise addition, and the channel attention weights are obtained by the sigmoid activation function. The channel attention $M_c(F)$ is calculated by the following method:

$$M_c\left(F\right) = \sigma\left(\text{MLP}\left(\text{AvgPool}\left(F\right)\right) + \text{MLP}\left(\text{MaxPool}\left(F\right)\right)\right) \tag{2}$$

where σ denotes the sigmoid function, MLP denotes the multilayer perceptron with a hidden layer, AvgPool and MaxPool denote the average pooling and max pooling operations respectively, and F is the given input feature map. These channel attention weights are broadcasted to the same size as the input feature map and multiplied elementwise with the input feature map.

Based on the Unet model, combined with channel attention and spatial attention mechanisms, a PV power station identification method PV-Unet that fuses high-resolution data and multispectral data is proposed. In the encoder part, we used the GF-2 image to extract the main features and used two convolution blocks with 3×3 convolution kernels for downsampling, with a batch normalization (BN) layer and a rectified linear unit layer after each convolution. In the decoder part, we restored the size of the feature map by deconvolution and connected it with the feature map output by the corresponding layer of the encoder.
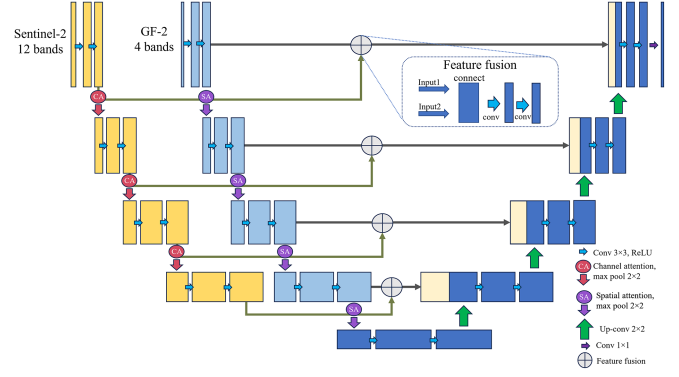
At the same time, we used the multispectral image of Sentinel-2 extracted features through another encoder branch concatenated the obtained feature map with the feature map of the GF-2 image, and input it into an FFM. The FFM optimized the feature fusion effect through two convolution blocks and input the result into the decoder part. Finally, to extract spatial and channel information more effectively, we inserted a spatial attention module after each GF-2 downsampling block and a channel attention module after each Sentinel-2 downsampling block. Fig. 4 shows the schematic diagram of the model structure.

3) Model Training and Evaluation Metrics: The model was implemented on the PyTorch framework. A total of 747 groups of images were used for training (randomly extracting 20% of them as the validation set), and 107 groups of images were used for testing the model. Each group of images included the GF-2 image and the Sentinel-2 image of the same area. To increase the scale of the training data, data augmentation methods were also used, including random horizontal flip, random vertical flip, and random rotation. The BCEWithLogitsLoss was used in the model training process. The loss function is implemented through PyTorch's function. The model was trained with the Adam optimizer, with a learning rate of 0.001, and used an early stopping mechanism to prevent overfitting. The final segmentation results were evaluated with five metrics, including accuracy, precision, recall, F1 score, and intersection over union (IoU). We compared PV-Unet with some other semantic segmentation models and calculated their accuracy, precision, recall, F1 score, and IoU. The calculation methods of each evaluation metric are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

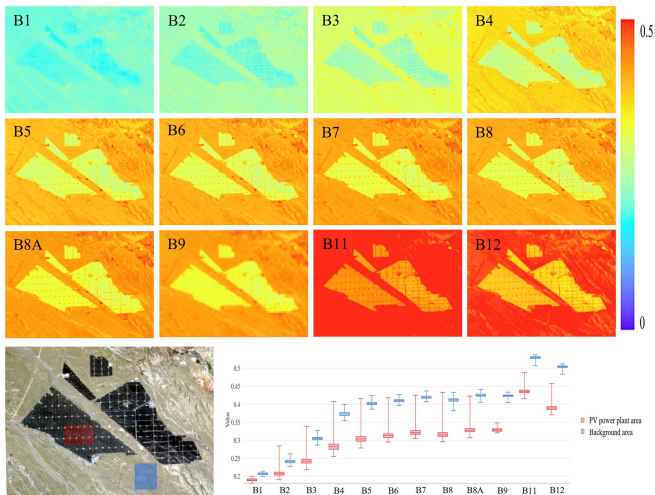$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{7}$$

Fig. 5. Images of Sentinel 2 in different bands. The box plot is the reflectance value of the specified area in the lower left image. Red is the photovoltaic panel area, and blue is the background area.

In the formula, TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively.

III. RESULTS

A. Impact of Multispectral Images

The reflectance of different types of land cover to electromagnetic waves of different wavelengths may vary significantly, so the difference between PV panels and background in images of different spectral ranges is different. As shown in Fig. 5, it can be clearly seen that the difference between PV panels and background varies in images of different bands. For example, in the Band1 image, the distinction between the PV panel area and background area is very low, whereas in the Band5–Band9 images, the PV panel area can be clearly distinguished. The box plots indicate that the spectral distribution difference between solar panels and background land cover is small in the first three bands, whereas the spectral distribution difference is large in the near-infrared and short-wave infrared bands. Therefore, the identification of PV panels should not be limited to the visible light band, the short-wave infrared and near-infrared bands may have more significant features, and more spectral channels can help distinguish PV panels and background pixels.

Due to the meteorological conditions at the time of shooting and the differences between different satellite sensors, images from different sources are not suitable for sharing weights. To show the difference between different feature extraction branches and the effect of feature fusion, we read the parameter distribution before and after the network feature fusion part. Include the convolutional layer and BN layer in the first down-sampling block model and the convolutional layer as well as the BN layer in the FFM of the trained PV-Unet. The parameter distribution is shown in Fig. 6, where conv.weight is the parameter distribution of the convolution layer, bn.mean is the shift parameter distribution of the BN layer, and bn.var is the scale parameter distribution of the BN layer. It is obvious that the

TABLE II
RESULTS ON DIFFERENT IMAGE DATASETS FOR GF-2 AND SENTINEL-2

Method	Accuracy	Precision	Recall	F1 score	IoU
Unet+GF-2+Sentinel-2 (16bands)	96.65%	98.97%	95.63%	97.27%	94.69%
Unet+GF-2+Sentinel-2 (RGB)	87.56%	92.69%	87.05%	89.78%	81.46%
Unet+GF-2 (4bands)	95.58%	99.28%	93.78%	96.45%	93.14%
Unet+Sentinel-2 (12bands)	90.98%	94.27%	88.14%	91.34%	84.06%

parameter distributions of images from different sources have a large difference, which means that it is unreasonable to directly mix images from different sources, and different branches learn different information. In our method, the model tries to extract valuable features from both images.

To further verify the effectiveness of adding Sentinel-2 images, as a comparison, we also trained different datasets with the Unet model, including the following:

1) 12 bands Sentinel-2 images;
2) 4 bands GF-2 images;
3) true color composite images of GF-2 and Sentinel-2 images;
4) 16 bands images obtained by overlaying GF-2 and Sentinel-2 images.

They are marked as Unet+Sentinel-2 (12 bands), Unet+GF-2 (4 bands), Unet+GF-2+Sentinel-2 (RGB), and Unet+GF-2+Sentinel-2 (16 bands), respectively. For the methods that only used one type of sensor data, the test set only has one type of data. For the methods that used two types of sensor data, the 16-band method merged each GF-2 and Sentinel-2 image into a 16-band image. The RGB-band method used the true color composite images of GF-2 and Sentinel-2 images and mixed the images of two types of sensors in the test set. The results are shown in Table II.

Obviously, the limited spatial resolution of Sentinel-2 makes it difficult to accurately identify PV panels. The results in the table show that the model trained with the Sentinel-2 image dataset has lower evaluation metrics than the model trained with the GF-2 image dataset. This indicates that the spatial resolution of the image is very important for the accurate identification of PV power station pixels. Even if the Sentinel-2 image has richer spectral information, due to its limited spatial resolution, the segmentation effect using only Sentinel-2 image is not as good as the GF-2 image. However, simply mixing GF-2 and Sentinel-2 images (the dataset contains both GF-2 and Sentinel-2 images) does not improve the identification accuracy, but may lead to a decrease in accuracy.

Since the GF-2 and Sentinel-2 images have different numbers of channels, we converted both images to RGB three-channel images here, and the number of images in the dataset is twice that of a single type of image (i.e., GF-2 and Sentinel-2 images each account for half). On this mixed dataset, the Accuracy is 87.56%,
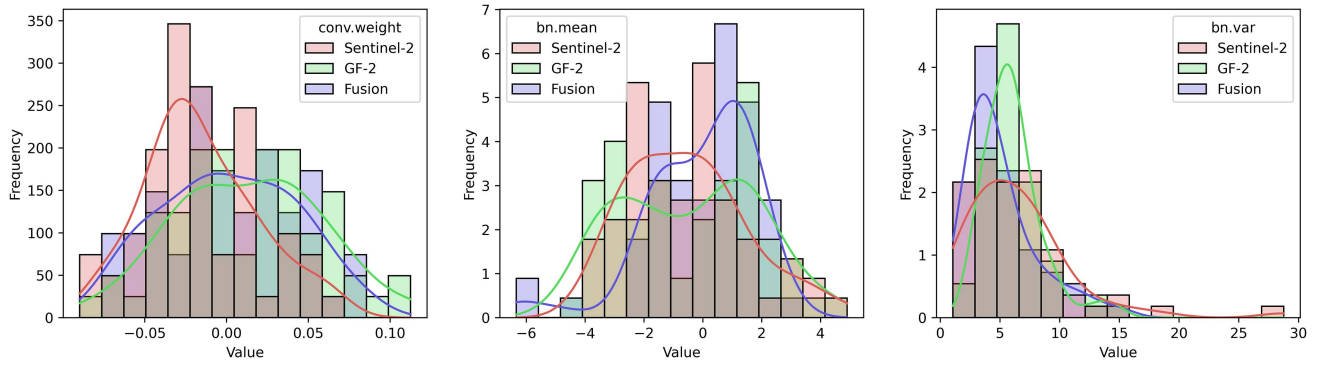
Fig. 6. Weight distribution of different convolutional layers.

the Precision is 92.69%, the Recall is 87.05%, the F1 score is 89.78%, and the IoU is 81.46%, which are 8.02%, 6.59%, 6.73%, 6.67%, and 12.67% lower than using only GF-2 images, respectively. Compared with using only Sentinel-2 images, they also have some decline.

This may be due to the large difference between the images, which makes it difficult for the model to learn the appropriate parameters. Remarkably, concatenating the two images into a 16-channel image yields the most favorable outcomes among the four datasets. This achieves 96.65% accuracy, 98.97% precision, 95.63% recall, 97.27% F1 score, and 94.69% IoU, respectively. These results underscore the positive impact of incorporating Sentinel-2 images on segmentation effectiveness. Nevertheless, it is crucial to acknowledge that this approach does not address the issue of image dissimilarity. The model may struggle to attain the most suitable parameters, particularly when confronted with significant differences in image resolution.

B. Accuracy Assessments

Fig. 7 shows the segmentation result of a PV power station. We compare it with some of the most advanced semantic segmentation models. These include Unet [31], DeepLabV3+ [36], SegNet [21], SegFormerB5 [37], PIDNet-M [38], and DDRNet-23 [39]. It is worth noting that, for the sake of fair comparison, all models were trained with identical hyperparameters. Unet and PV-Unet have better edge segmentation. SegNet misclassified some background pixels as PV pixels, DeepLabV3+ had unclear edge segmentation, and they both did not distinguish the narrow roads clearly. PIDNet failed to distinguish smaller patches and was not good at extracting the straight boundaries of PV panels. The prediction results of SegFormerB5 and DDRNet-23 were similar to those of DeepLabV3+, but they had less false recognition of tiny patches outside the power station range. The segmentation result of PV-Unet is similar to Unet, and can better maintain the regularity of the PV panel boundary. Unet identified more narrow roads inside the power station, whereas PV-Unet had fewer misidentifications.

In addition, we selected some representative segmentation results for visualization, and the results are shown in Fig. 8. The results showed that DeepLabV3+, SegFormerB5, PIDNet-M, and DDRNet-23 had poor road distinction effects between PV
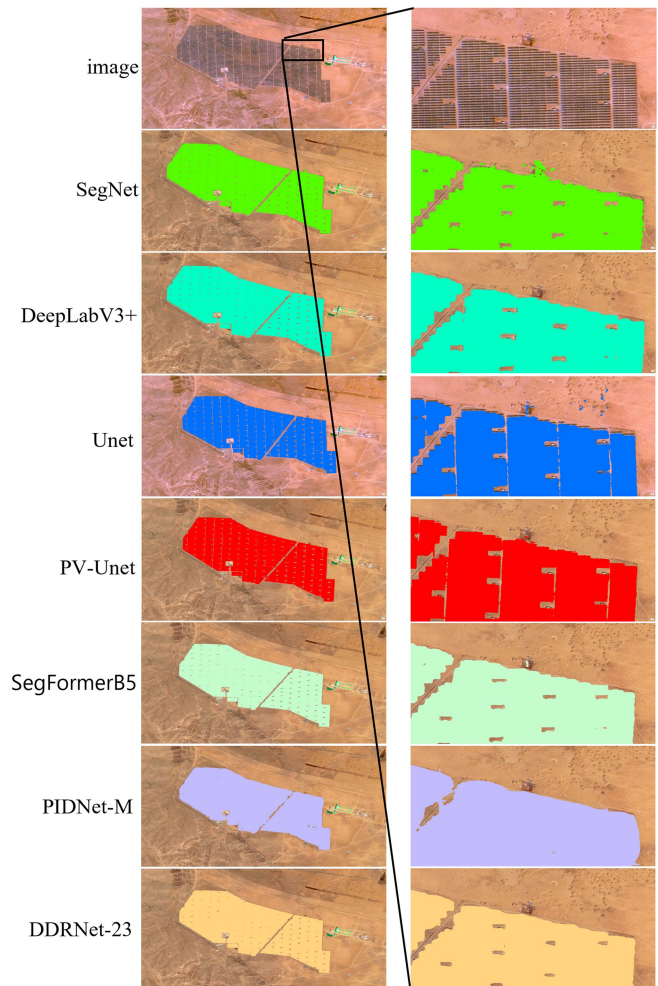

Fig. 7. Identification results of photovoltaic power plants.

panels. PIDNet-M and DDRNet-23 extracted the edges of PV panels with blur, which did not match their straight characteristics. In the second row of images, Unet and SegNet both have omissions.

We compared the PV-Unet model with some of the latest segmentation methods. The evaluation metrics for the trained model on the test set are summarized in Table III. The best
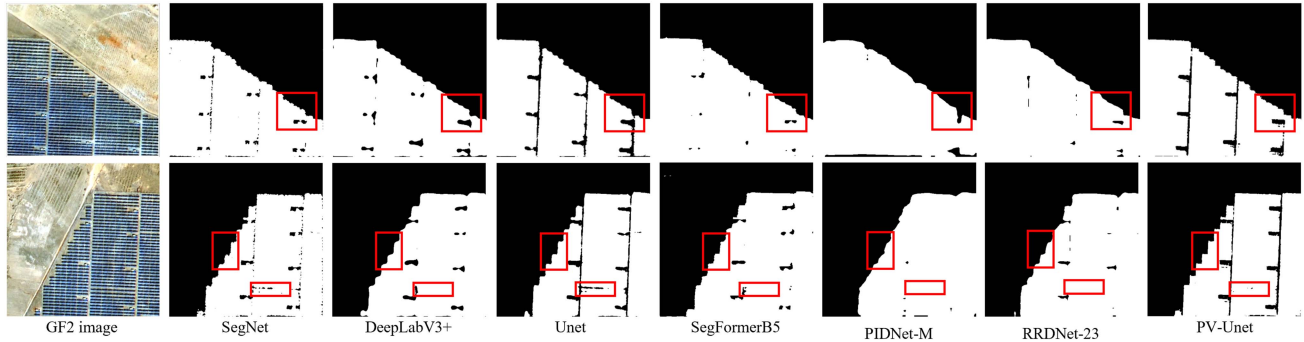
Fig. 8. Segmentation results of different models.

TABLE III
PERFORMANCE OF THE MODEL ON THE TEST SET

Model	Accuracy	Precision	Recall	F1 score	IoU
SegNet	96.52%	98.54%	95.78%	97.14%	94.44%
DeepLabV3+	96.77%	98.77%	95.98%	97.35%	94.84%
Unet	95.58%	**99.28%**	93.78%	96.45%	93.14%
SegFormerB5	96.73%	97.66%	95.84%	97.32%	94.78%
PIDNet-M	97.29%	98.75%	96.80%	97.76%	95.63%
DDRNet-23	96.74%	99.12%	95.60%	97.33%	94.80%
PV-Unet	**97.61%**	98.98%	**97.12%**	**98.04%**	**96.15%**

The bold values indicate the best metrics.

TABLE IV
EVALUATION RESULTS OF THE ATTENTION MODULE

Model	Accuracy	Precision	Recall	F1 score	IoU
PV-Unet (without AM)	96.47%	97.29%	96.88%	97.09%	94.34%
PV-Unet (SA)	96.73%	97.44%	**97.14%**	97.29%	94.73%
PV-Unet (CA)	95.59%	97.48%	95.32%	96.39%	93.03%
PV-Unet	**97.61%**	**98.98%**	97.12%	**98.04%**	**96.15%**

The bold values indicate the best metrics.

metrics in Table III are shown in bold, and the underline indicates the second-highest metrics. The results show that PV-Unet performs better on most metrics, Unet has the highest Precision, but the lowest Recall, which also leads to a lower F1 score for Unet. PIDNet-M achieves the second-highest results on most metrics. Compared with the other semantic segmentation models, PV-Unet has the best overall effect.

IV. DISCUSSION

A. Effectiveness of the Attention Module

To further investigate the effectiveness of the PV-Unet method, we conducted an ablation experiment on the attention modules. These modules are instrumental in enhancing the feature extraction capabilities of each branch. Given that the GF-2 image carries richer spatial information, and the Sentinel-2 image encompasses more spectral information, the attention modules are designed to guide the model's focus on spatial features from the GF-2 image and spectral information from the Sentinel-2 image. For the ablation experiments, we considered four configurations: one without any attention modules, one with only spatial attention modules, another with only channel attention modules, and finally, a setup incorporating both spatial and channel attention modules.

Fig. 9 illustrates the visualization results of the experiment. A clear observation from Fig. 9 is that the attention modules play

a pivotal role in enhancing the model's accuracy in target identification. Specifically, the spatial attention module contributes to the smoothness and continuity of the segmentation results, thereby reducing instances of omission. On the other hand, the channel attention module effectively mitigates false positives, although it may compromise segmentation integrity. Notably, the simultaneous use of both attention modules manages to circumvent certain misidentifications while preserving overall segmentation quality. For instance, in the first row, the PV-Unet without attention modules exhibits a blurred identification of the PV panel's edge. The model employing only the channel attention module improves edge segmentation but introduces some holes. Combining both attention modules reduces the holes induced by channel attention while maintaining superior edge segmentation. In the second and third rows, the attention modules are observed to suppress erroneously identified pixels.

Table IV presents the results of the evaluation metrics, where PV-Unet (without AM) indicates the absence of attention modules, PV-Unet (SA) signifies the inclusion of only the spatial attention module, and PV-Unet (CA) represents the configuration with only the channel attention module. The outcomes show that the spatial attention module primarily enhances recall, whereas the channel attention module predominantly improves precision. Combining both spatial attention and channel attention concurrently results in improvements in Accuracy, F1 score, and IoU, showcasing increments of 1.14%, 0.95%, and 1.81%, respectively, compared to scenarios without attention modules. In summary, the incorporation of both spatial and channel attention modules yields the most favorable outcomes. Both qualitative and quantitative results affirm the efficacy of the attention modules.
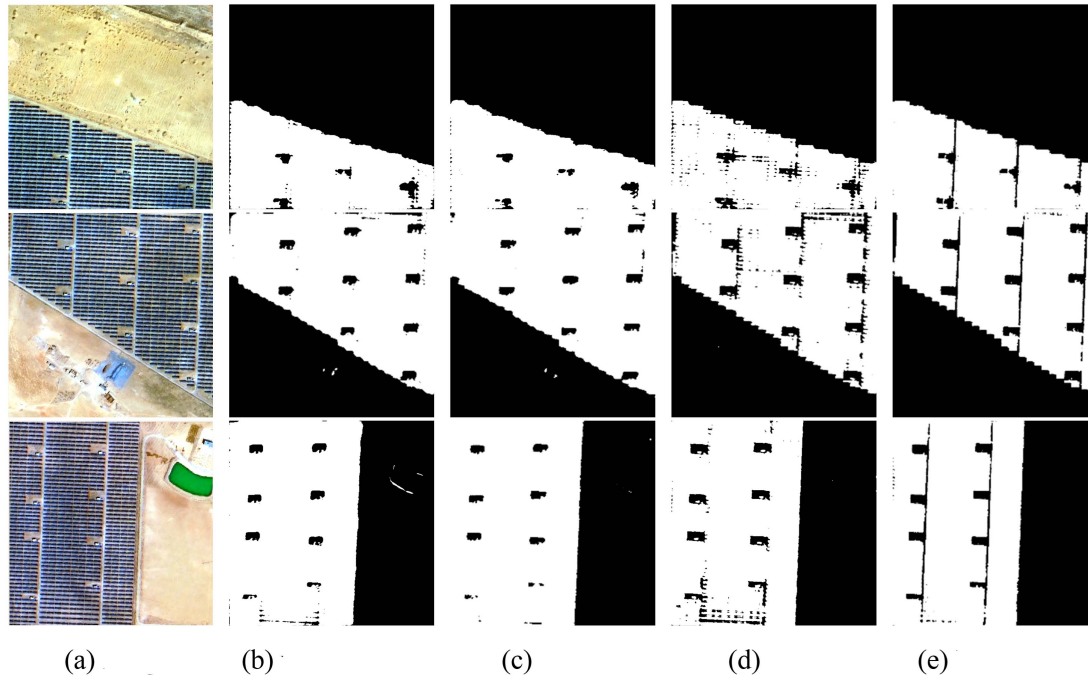
Fig. 9. Compare the results with or without the addition of attention modules. (a) Is the original remote sensing image. Columns (b)–(e), respectively, include not adding attention modules, only adding spatial attention modules, only adding channel attention modules, and adding both types of attention modules together. (a) Image. (b) Without AM. (c) SA. (d) CA. (e) PV-Unet.

TABLE V
EVALUATION RESULTS OF THE FEATURE FUSION MODULE

Model	Accuracy	Precision	Recall	F1 score	IoU
PV-Unet (without FFM)	95.77%	**99.09%**	94.20%	96.59%	93.40%
PV-Unet (with FFM)	**97.61%**	98.98%	**97.12%**	**98.04%**	**96.15%**

The bold values indicate the best metrics.

B. Effectiveness of the Feature Fusion Module

The role of the FFM is to iteratively optimize the feature fusion effect of high-resolution images and multispectral images. To analyze the effectiveness of the FFM, we designed an ablation experiment without the FFM. In the experiment without the FFM, the feature maps from different branches were directly stacked and then fed into the corresponding upsampling module. Fig. 10 shows the visualization results with and without the FFM. It can be seen from the examples that the FFM segments the target and background areas more completely.

The quantitative analysis presented in Table V reveals that although omitting the FFM results in higher precision, it comes at the expense of lower recall. Incorporating the FFM leads to improvements in Accuracy, Recall, F1 score, and IoU to varying degrees, with increases of 1.84%, 2.92%, 1.45%, and 2.75%, respectively. Both qualitative and quantitative analyses underscore that, when dealing with images from diverse sensors, direct concatenation of features falls short of achieving optimal results due to substantial image differences. The introduction of the FFM allows the model to derive more suitable fusion outcomes, and passing the optimized feature maps through the decoder subsequently leads to superior segmentation results.
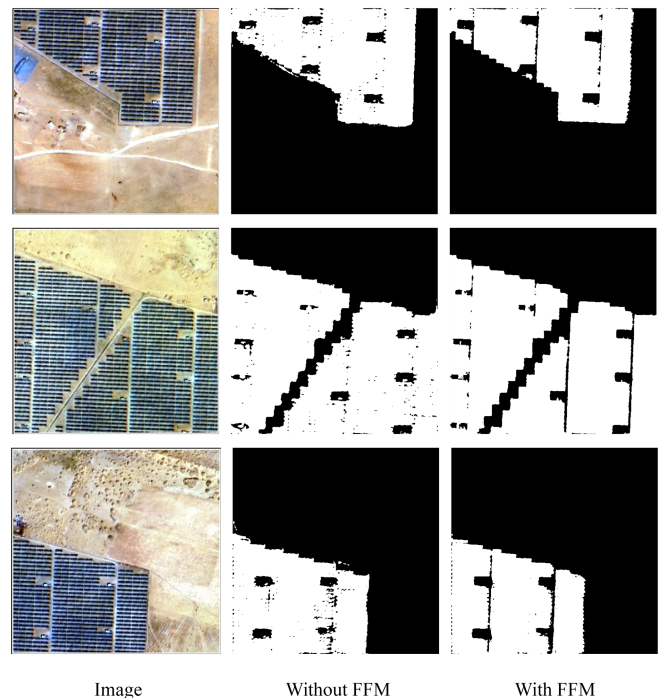

Image Without FFM With FFM

Fig. 10. Results with or without FFM.

C. Number of Parameters and Time Cost

We tested and compared the parameters and inference speed of all models. For a fair performance comparison, all reported FPS results were obtained at an input resolution of 512×512. The inference speed of all models was tested on a single NVIDIA

Model	Resolution	#Params	FPS
Unet	512×512	29.6060M	378.104
DeepLabV3+	512×512	56.5932M	73.756
SegNet	512×512	28.0801M	238.340
SegFormerB5	512×512	80.6782M	21.808
PIDNet-M	512×512	35.5774M	70.679
DDRNet-23	512×512	29.4176M	144.124
PV-Unet	512×512	56.5897M	184.074

RTX4070. All models are implemented using PyTorch. The experimental results are shown in Table VI. Where #Params is the number of parameters and FPS is the number of images processed per second. Compared with the Unet model, PV-Unet has increased computational cost but achieves better accuracy. Compared with SegFormerB5, the proposed model has a faster inference speed. Overall, PV-Unet balances accuracy and computational cost, and is an effective method for extracting PV panels.

V. CONCLUSION

This article proposes a semantic segmentation method for PV panels, fusing high spatial resolution remote sensing data and multispectral data. By extracting features from GF-2 and Sentinel-2 images through dedicated branches, this approach effectively captures the distinctive texture and spectral features of PV panels. The fusion of high spatial resolution and multispectral image features is accomplished through the strategic use of skip connections and FFMs. Spatial attention modules and channel attention modules are seamlessly incorporated into the two types of images, facilitating the extraction of texture and spectral features. This inclusion enhances the overall extraction effectiveness of the corresponding features. The main contributions of this article are as follows:

1) The proposed method successfully mitigates the impact of diverse spatial and spectral resolutions inherent to different sensors, resulting in the accurate identification of PV panels.

2) Compared to other segmentation models, the method successfully mitigated the effects of different spatial and spectral resolutions inherent in different sensors, resulting in the accurate identification of PV panels. The obtained results showcase the superior segmentation performance of the proposed method compared to other tested models, achieving 98.04% F1 score and 96.15% IoU on the test data.

3) A comprehensive ablation experiment analysis highlights the role of attention modules in balancing the extraction of texture and spectral features, ensuring segmentation result integrity, and effectively distinguishing background pixels. Additionally, FFMs play a crucial role in alleviating the influence of image differences stemming from various sensors, optimizing the feature fusion effect, and achieving

improvements of 1.45% in F1 score and 2.75% in IoU on the test data.

4) The proposed method achieves a trade-off between computational cost and accuracy. On the test device, PV-Unet achieved 184.074 FPS and 56.5897M parameters while achieving the best accuracy.

This study provides an effective method for accurate segmentation of PV panels using remote sensing images, which can fully utilize the advantages of satellite remote sensing technology, such as fast imaging speed, rich spectral channels, and long-term observation. It is expected to provide convenience for the identification of large-scale PV panels.

ACKNOWLEDGMENT

REFERENCES

[1] L. Chen et al., "Strategies to achieve a carbon neutral society: A review," *Environ. Chem. Lett.*, vol. 20, no. 4, pp. 2277–2310, 2022, doi: 10.1007/s10311-022-01435-8.

[2] M. Meinshausen et al., "Greenhouse-gas emission targets for limiting global warming to 2 °C," *Nature*, vol. 458, no. 7242, pp. 1158–1162, 2009, doi: 10.1038/nature08017.

[3] S. I. Seneviratne, M. G. Donat, A. J. Pitman, R. Knutti, and R. L. Wilby, "Allowable CO_2 emissions based on regional and impact-related climate targets," *Nature*, vol. 529, no. 7587, pp. 477–483, 2016, doi: 10.1038/nature16542.

[4] N. M. Haegel et al., "Terawatt-scale photovoltaics: Trajectories and challenges," *Science*, vol. 356, no. 6334, pp. 141–143, 2017, doi: 10.1126/science.aal1288.

[5] H. Mao et al., "Advances and prospects on estimating solar photovoltaic installation capacity and potential based on satellite and aerial images," *Renewable Sustain. Energy Rev.*, vol. 179, 2023, Art. no. 113276, doi: 10.1016/j.rser.2023.113276.

[6] H. E. Murdock et al., "Renewables 2021 - Global status report," International Atomic Energy Agency, Vienna, Austria, France, Rep. INIS-FR–21-0788, 2021. [Online]. Available: http://inis.iaea.org/search/search.aspx?orig_q=RN:52059346

[7] Z. Chen, Y. Kang, Z. Sun, F. Wu, and Q. Zhang, "Extraction of photovoltaic plants using machine learning methods: A case study of the pilot energy City of Golmud, China," *Remote Sens.*, vol. 14, no. 11, 2022, Art. no. 2697, [Online]. Available: https://www.mdpi.com/2072-4292/14/11/2697

[8] M. Hosenuzzaman, N. A. Rahim, J. Selvaraj, M. Hasanuzzaman, A. B. M. A. Malek, and A. Nahar, "Global prospects, progress, policies, and environmental impact of solar photovoltaic power generation," *Renewable Sustain. Energy Rev.*, vol. 41, pp. 284–297, 2015, doi: 10.1016/j.rser.2014.08.046.

[9] S. Ren et al., "Automated extraction of energy systems information from remotely sensed data: A review and analysis," *Appl. Energy*, vol. 326, 2022, Art. no. 119876, doi: 10.1016/j.apenergy.2022.119876.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[11] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. Nadai, "Efficient training of visual transformers with small datasets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 23818–23830.

[12] D. Zhang et al., "Deep learning for medical image segmentation: Tricks, challenges and future directions," 2022, *arXiv:2209.10307*.

[13] D. Jurakuziev, S. Jumaboev, and M. Lee, "A framework to estimate generating capacities of PV systems using satellite imagery segmentation," *Eng. Appl. Artif. Intell.*, vol. 123, 2023, Art. no. 106186, doi: 10.1016/j.engappai.2023.106186.

[14] Z. Nong, X. Su, Y. Liu, Z. Zhan, and Q. Yuan, "Boundary-aware dual-stream network for VHR remote sensing images semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5260–5268, 2021, doi: 10.1109/JSTARS.2021.3076035.

[15] J. Hou, Z. Guo, Y. Feng, Y. Wu, and W. Diao, "SPANet: Spatial adaptive convolution based content-aware network for aerial image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2192–2204, 2023, doi: 10.1109/JSTARS.2023.3244207.

[16] L. Li, N. Lu, H. Jiang, and J. Qin, "Impact of deep convolutional neural network structure on photovoltaic array extraction from high spatial resolution remote sensing images," *Remote Sens.*, vol. 15, no. 18, 2023, Art. no. 4554, [Online]. Available: https://www.mdpi.com/2072-4292/15/18/4554

[17] L. Du et al., "Photovoltaic power station extraction from high-resolution satellite images based on deep learning method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 3351–3354, doi: 10.1109/IGARSS46834.2022.9883970.

[18] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States," *Joule*, vol. 2, no. 12, pp. 2605–2617, 2018, doi: 10.1016/j.joule.2018.11.021.

[19] X. Hou, B. Wang, W. Hu, L. Yin, and H. Wu, "SolarNet: A deep learning framework to map solar power plants in China from satellite imagery," 2019, *arXiv:1912.03685*.

[20] L. Kruitwagen, K. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of photovoltaic solar energy generating units," *Nature*, vol. 598, no. 7882, pp. 604–610, 2021.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[22] M. Tan, W. Luo, J. Li, and M. Hao, "TEMCA-Net: A texture-enhanced deep learning network for automatic solar panel extraction in high groundwater table mining areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2838–2848, 2024, doi: 10.1109/JSTARS.2023.3347572.

[23] H. Nasrallah, A. E. Samhat, Y. Shi, X. X. Zhu, G. Faour, and A. J. Ghandour, "Lebanon solar rooftop potential assessment using buildings segmentation from aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4909–4918, 2022, doi: 10.1109/JSTARS.2022.3181446.

[24] J. Wang, X. Chen, W. Jiang, L. Hua, J. Liu, and H. Sui, "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 Observ. Geoinf.*, vol. 119, 2023, Art. no. 103309, doi: 10.1016/j.jag.2023.103309.

[25] H. Jiang et al., "Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery," *Earth Syst. Sci. Data*, vol. 13, no. 11, pp. 5389–5401, 2021.

[26] T. Sun, M. Shan, X. Rong, and X. Yang, "Estimating the spatial distribution of solar photovoltaic power generation potential on different types of rural rooftops using a deep learning network applied to satellite images," *Appl. Energy*, vol. 315, 2022, Art. no. 119025, doi: 10.1016/j.apenergy.2022.119025.

[27] P. Li et al., "Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning," *Adv. Appl. Energy*, vol. 4, 2021, Art. no. 100057, doi: 10.1016/j.adapen.2021.100057.

[28] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sens. Environ.*, vol. 277, 2022, Art. no. 113058, doi: 10.1016/j.rse.2022.113058.

[29] B. Ren et al., "A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102896, doi: 10.1016/j.jag.2022.102896.

[30] Q. Guo, B. Fu, P. Shi, T. Cudahy, J. Zhang, and H. Xu, "Satellite monitoring the spatial-temporal dynamics of desertification in response to climate change and human activities across the Ordos Plateau, China," *Remote Sens.*, vol. 9, no. 6, 2017, Art. no. 525, [Online]. Available: https://www.mdpi.com/2072-4292/9/6/525

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[32] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal. Int. Workshop Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.

[33] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *IS-PRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[35] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, 2020, doi: 10.1007/s10462-020-09825-6.

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.

[38] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19529–19539.

[39] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023.
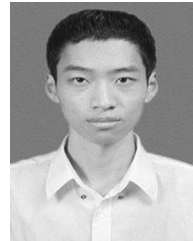
Zhiyu Zhao received the B.S. degree in remote sensing science and technology and the M.S. degree in photogrammetry and remote sensing from the College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing, China, in 2020 and 2023, respectively. He is currently working toward the Ph.D. degree with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China.

His research interests include thermal remote sensing of urban environments and remote sensing of ecological environments.

Yunhao Chen received the B.S. and M.S. degrees in resource management from the Anhui University of Science and Technology, Huainan, China, in 1994 and 1997, respectively, and the Ph.D. degree in geodetic engineering from the China University of Mining and Technology-Beijing, Beijing, China, in 1999.

From 2000 to 2001, he was a Postdoctoral Researcher with Beijing Normal University, Beijing, China. Since 2001, he has been with the Faculty of Geographical Science, Beijing Normal University, where he is currently a Professor with the State Key Laboratory of Remote Sensing Science. His research interests include thermal remote sensing of urban environment and applications of remote sensing in ecology.

Kangning Li received the B.S. degree in geographic information system from the College of Urban and Environmental Science, Northwest University, Xi'an, China, in 2016, and the Ph.D. degree in cartography and geographic information systems from the Faculty of Geographical Science, Beijing Normal University, Beijing, China, in 2022.

Her research interests include thermal remote sensing of urban environment.

Weizhen Ji received the B.S. degree in geographic information system from the School of Municipal and Geomatics Engineering, Hunan City University, Yiyang, China, in 2018, and the M.S. degree in cartography and geographic information system from the School of Civil and Surveying and Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou, China, in 2021. He is currently working toward the Ph.D. degree with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China.

His research interests include thermal remote sensing of urban environment and spatiotemporal downscaling of land surface temperature.

Hao Sun received the Ph.D. degree in cartography and geographic information system from Beijing Normal University, Beijing, China, in 2014.

He is currently an Associate Professor with the China University of Mining and Technology-Beijing, Beijing, China. His research interests include quantitative remote sensing methods and their applications in soil moisture, vegetation drought, and ecological environment.