Automatic detection of photovoltaic facilities from Sentinel-2 observations by the enhanced U-Net method

Zixuan Dui,[a,b] Yongjian Huang,[a] Jiuping Jin,[a] and Qianrong Gu[a,*]
[a]Chinese Academy of Sciences, Shanghai Advanced Research Institute, Shanghai Carbon Data Research Center, Key Laboratory of Low-Carbon Conversion Science & Engineering, Pudong, China
[b]University of Chinese Academy of Sciences, Beijing, China

Abstract. With the enactment of supportive government policies and the increasing maturity of solar photovoltaic (PV) technologies, solar PV energy has become the most cost-effective new energy resource worldwide. Geospatial information on existing solar PV power systems is necessary to manage and optimize the deployment of new PV facilities. In this study, we propose a new deep-learning network, named the enhanced U-Net (E-UNET), to detect PV facilities from Sentinel-2 multi-spectral remote sensing data. Our E-UNET features an enhanced encoder–decoder structure that can efficiently extract spectral and spatial features simultaneously by combining a multi-spectral three-dimensional convolution path and a multi-scale pooling block. We compare the performance of the E-UNET with other semantic segmentation deep-learning networks and a pixel-based random forest classifier. The experimental results show that the E-UNET performs better than the other methods. It achieves an overall accuracy, Matthews correlation coefficient, $F1$, kappa coefficient, and recall of 0.989, 0.862, 0.869, 0.934, and 0.875, respectively. The experimental results also indicate that the E-UNET accurately detects PV facilities from various complex environments with high accuracy in terms of PV integrity and details. [DOI: 10.1117/1.JRS.17.014516]

Keywords: enhanced U-Net; semantic segmentation; Sentinel-2; photovoltaic; remote sensing.

Paper 220236G received Apr. 20, 2022; accepted for publication Feb. 8, 2023; published online Mar. 6, 2023.

1 Introduction

The International Energy Agency's State Policies Scenario predicts that global solar photovoltaic (PV) capacity will grow at an average rate of 12% per annum, reaching a capacity of 2764 TWh in 2030.[1] The ability to efficiently census geospatial information on solar PV energy systems is highly important for countries to formulate strategies in accordance with their commitment to achieving carbon neutrality by 2050, as well as for system operators and market analysts to quantify and optimize the efficiency of PV facility deployments.

In recent years, a substantial amount of work has been undertaken to detect the spatial distribution of PV facilities from satellite remote sensing data by means of computer vision. Using computer vision techniques can overcome the incomplete, time-consuming, and labor-intensive problems associated with manual counting and mapping of PV facilities. In 2015, Malof et al.[2] first proposed using a support vector machine approach to locate PV facilities from satellite and aerial images. The feature extraction process of Malof et al.[2] has some limitations requiring manual adjustment of the image feature descriptors. With the development of convolutional neural network (CNN), many researchers begin to apply CNN to detecting PV facilities from satellite images.[3,4] The CNN approaches enable automatic representation learning and have the advantage of examining more complex spatial patterns that cannot be captured by shallow classifiers.[5] Therefore, they can significantly improve the accuracy of location and contour detection of PV facilities. In 2018, Hou et al.[6] proposed the SolarNet deep-learning framework, which

*Address all correspondence to Qianrong Gu, guqr@sari.ac.cn

combined a full convolutional network (FCN) and an expectation-maximization attention module to locate and estimate the surface area of solar PV facilities in China. Yu et al.[7] applied a semi-supervised object localization and segmentation method to generate class activation maps based on the Inception-v3 framework and built a database of PV facilities in the United States.

Although the above studies have achieved remarkable accuracy in detecting PV facilities, they were carried out only on red, green, blue (RGB) satellite images. Numerous multi-spectral images have become available with the rapid development of remote sensing technologies.[8] For cases in which PV facilities and backgrounds are visually similar or the scene is blurred, using multi-spectral information instead of only RGB information can further improve the detection accuracy.[9] In 2019, Kruitwagen et al.[10] used multi-spectral remote sensing images from Sentinel-2[11] (12 bands) and SPOT-6/7 (4 bands) to conduct a global survey of utility-scale (installed capacity larger than 10 kW) solar PV facilities by a double-branch machine learning pipeline method.

For PV detection with segmentation methods, accurate segmentation of multi-spectral satellite remote sensing images using end-to-end deep learning methods remains a challenge. The classical semantic segmentation model U-Net[12] has proven to be advantageous in multi-spectral satellite image segmentation and has been widely used in applications, such as road segmentation,[13] burned area mapping,[14,15] and cloud masking.[16] In this study, we propose the E-UNET network structure enhanced from the classical U-Net[12] structure to detect PV facilities from Sentinel-2[11] multi-spectral remote sensing images. The E-UNET is based on an encoder–decoder structure that extracts spatial-spectral features through a multi-spectral three-dimensional (3D) convolution (MSD) path. Its multi-scale pooling (MSP) block encodes contextual information from multiple scales. Therefore, the E-UNET effectively extracts and integrate spectral and spatial features at different scales to achieve fine-grained and better overall segmentation accuracy than the classical U-Net.[12] We use experiments to demonstrate and analyze the effectiveness of our E-UNET in detecting PV facilities from Sentinel-2[11] multi-spectral images. Furthermore, we experimentally compare the E-UNET approach with several state-of-the-art methods, and the experimental results show that the E-UNET achieves the best PV detection performance.

The remainder of the manuscript is organized as follows: Sec. 2 introduces the multi-spectral images used in this study, Sec. 3 describes the proposed E-UNET in detail, Sec. 4 describes the experimental setup, Sec. 5 presents the experimental results and discussion, and finally, the conclusions are given in Sec. 6.

2 Data

In this study, we use Sentinel-2[11] satellite remote sensing images to detect PV facilities. The Sentinel-2 mission comprises twin polar-orbiting satellites launched in 2015 and 2017, respectively.[11] Both the Sentinel-2 satellites carry a multi-spectral payload capable of acquiring observations in 13 spectral bands with spatial resolutions of 10, 20, and 60 m.[11]

As shown in Fig. 1, we collect 41 Sentinel-2 Level-2A[17] multi-spectral scenes containing large-scale, non-residential PV facilities. These scenes cover deserts, mountains, lakes, and coastal areas with different seasons, latitudes, longitudes, and topographies, representing different environmental disturbances to the PV detection task.

Because the smallest downloadable scenes of Sentinel-2 Level-2A products cover 100×100 km^2 and PV facilities typically occupy only a small portion of the scene, we visually crop 137 images (see Fig. 2) containing PV facilities from the 41 downloaded scenes using ENVI version 5.3 software. These cropped images range in size from 260×260 pixels to 1500×1500 pixels.

In addition, we use the Sen2Res[18] tool provided by the sentinel application platform to fuse the 20- and 60-m resolution bands of the cropped multi-spectral images with the corresponding 10-m resolution band. The Sen2Res[18] uses a super-resolution method to fuse a low-resolution band into a high-resolution band while keeping its reflectance value unchanged.[18–20] The super-resolution method explores geometric detail information among adjacent pixel contents shared between the low- and high-resolution bands to keep the local reflectance consistency of adjacent

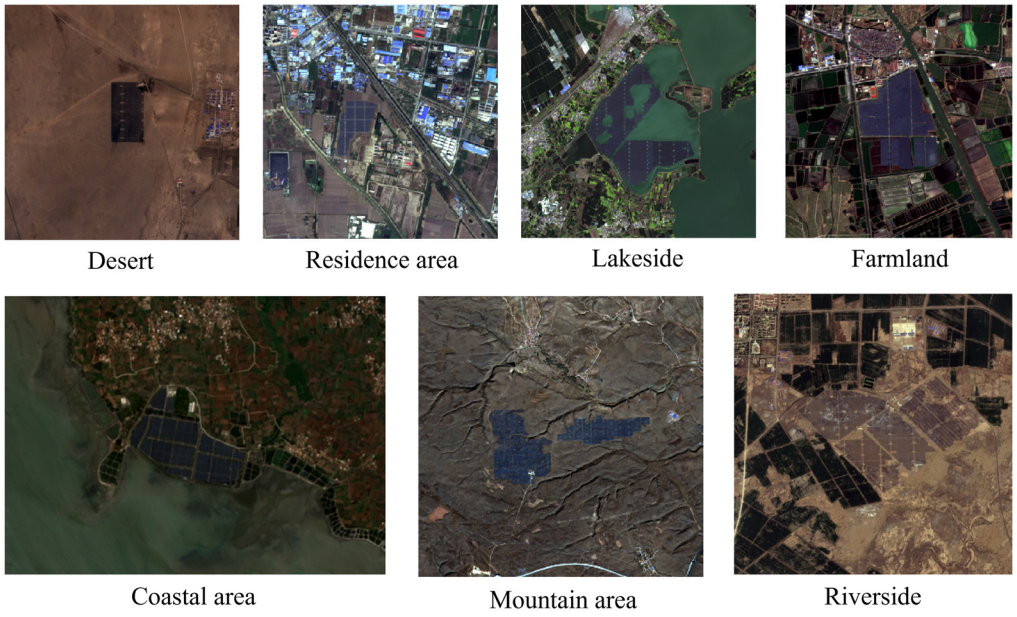

Fig. 1 RGB images synthesized from Sentinel-2 multi-spectral data.

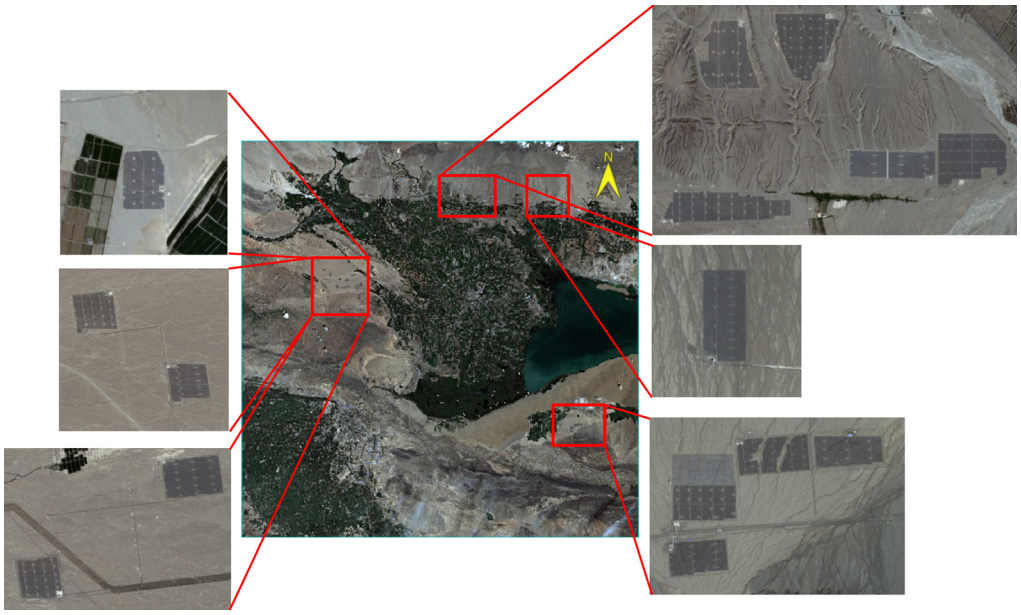

Fig. 2 Multi-spectral images containing PV facilities visually cropped from Sentinel-2 level 2A scenes.

pixels in the low-resolution band unchanged, as well as to keep the geometric details of sub-pixel components in the low-resolution band consistent with those in the high-resolution band.[18–20] Band 10 in the cropped images is discarded because it is generally used to detect cirrus clouds.[21]

3 E-UNET Method

As shown in Fig. 3, the E-UNET has an end-to-end CNN architecture modified from the classical U-Net[12] to improve the segmentation performance of multi-spectral remote sensing satellite

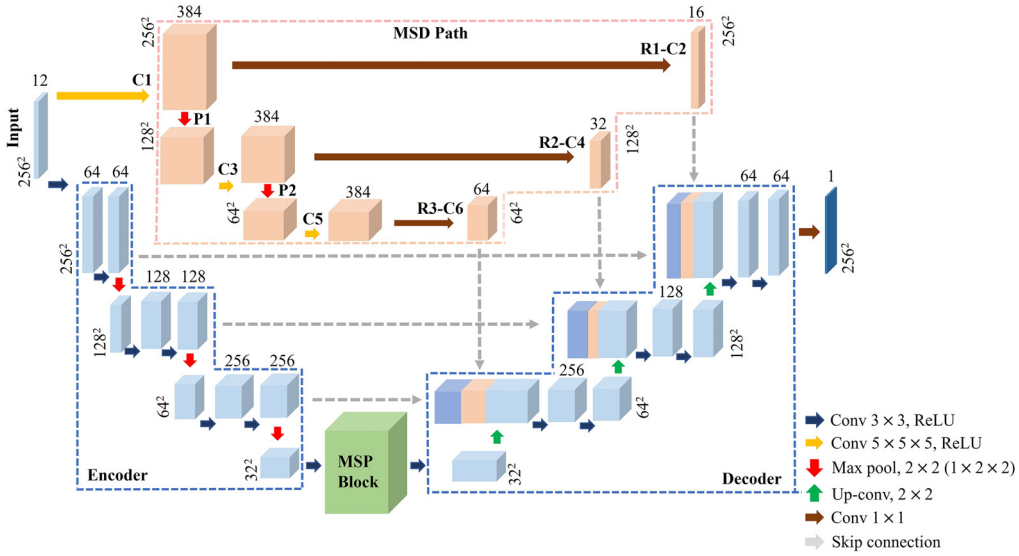

Fig. 3 E-UNET architecture. (**C**, **P**, and **R** stand for convolution, max-pooling, and reshaping filters, respectively.)

images. It consists of three key components: the feature encoder-decoder module, the MSP block, and the MSD path module.

3.1 *Feature Encoder–Decoder Module*

The feature encoder and decoder form a symmetrical U-shaped structure, which is the backbone of the E-UNET. The spatial feature encoder is divided into three layers, each layer contains two convolution filters, the second convolution filter is followed by a max-pooling kernel for down-sampling operations. As shown by the gray dashed arrows in Fig. 3, the output of each encoder layer and each MSD path is connected to the corresponding decoder input via a skip connection. During the feature decoding process, three cascade up-sampling operations are carried out to restore the size of the merged spatial and spectral feature maps to the same size as the input image. At the bottom of the U-shaped structure, an MSP block is embedded to improve the segmentation performance by including global context information.

3.2 *Multi-spectral 3D Convolution Path Module*

Although the classical U-Net[12] can also handle multi-spectral images, its two-dimensional (2D) convolution filters can only use features extracted from the spatial dimensions of each band.[22] Therefore, we add an MSD path module to capture the nonlinear relationships of adjacent pixels between different spectral bands, which are neglected by the 2D convolution filters in the classical U-Net.[12] Table 1 shows the size and number of filters in the 3D convolution layers (from C1 to R3-C6 in Fig. 3) and in the max-pooling layers (from P1 to P2 in Fig. 3), as well as the output size of each MSD path. The size of the 3D convolution filters is $5 \times 5 \times 5$. The max-pooling kernels down-sample the output spectral features of each 3D convolution filter, so the spectral feature maps are aligned in the cross-sectional direction with the spatial feature maps extracted by the encoder layers in the U-shaped structure. The spectral feature maps of each size are then sent to the corresponding decoders in the U-shaped structure through the skip connections.

To balance the weights between the spectral and spatial features and prevent the network from giving too much weight to the spectral features, a 1×1 sized convolution filter is added after each MSD path to reduce the dimensionality and computational cost of the spectral features. With the help of the MSD path module, the E-UNET automatically extracts the spectral features from adjacent pixels by the 3D convolution filters and combines them with the spatial features

Table 1 Details of the MSD path module of E-UNET. (C, P, and R stand for convolution, max-pooling, and reshaping filters, respectively).

Layer	Filter size, number	Output shape	Connected to
C1	$(5\times5\times5,32)$	$(256\times256\times12)\times32$	Input
R1	—	$256\times256\times384$	C1
C2	$(1\times1,16)$	$256\times256\times16$	R1
P1	$2\times2\times2$	$(128\times128\times6)\times32$	C1
C3	$(5\times5\times5,64)$	$(128\times128\times6)\times64$	P1
R2	—	$128\times128\times384$	C3
C4	$(1\times1,32)$	$128\times128\times32$	R2
P2	$2\times2\times2$	$(64\times64\times3)\times64$	C3
C5	$(5\times5\times5,128)$	$(64\times64\times3)\times128$	P2
R3	—	$64\times64\times384$	C5
C6	$(1\times1,64)$	$64\times64\times64$	R3

extracted by the U-shaped encoder–decoder module through the skip connections. Thus, E-UNET effectively improves the accuracy of multi-spectral image segmentation.

3.3 *Multi-scale Pooling Block*

In the PV semantic segmentation task, it is a big challenge to cope with the substantial variation in sizes of different PV facilities. In the Sentinel-2 multi-spectral images with the finest resolution of 10 m used in this study, the typical width of the total outline of large and continuously aligned PV facilities is about 100 to 200 pixels. Meanwhile, the outline of small and scattered PV facilities or the gap between PV panels is only a few to tens of pixels wide. In the classical U-Net[12] deep network, the maximum pooling uses only one fixed-size pooling kernel. Therefore, the classical U-Net[12] only perceives the context within a fixed-size receptive field and does not fully integrate important multi-scale spatial information.

Inspired by the pyramid pooling structure,[23] we add an MSP block at the bottom of the U-shaped structure, i.e., below the third encoder layer. As shown in Fig. 4, the MSP block uses

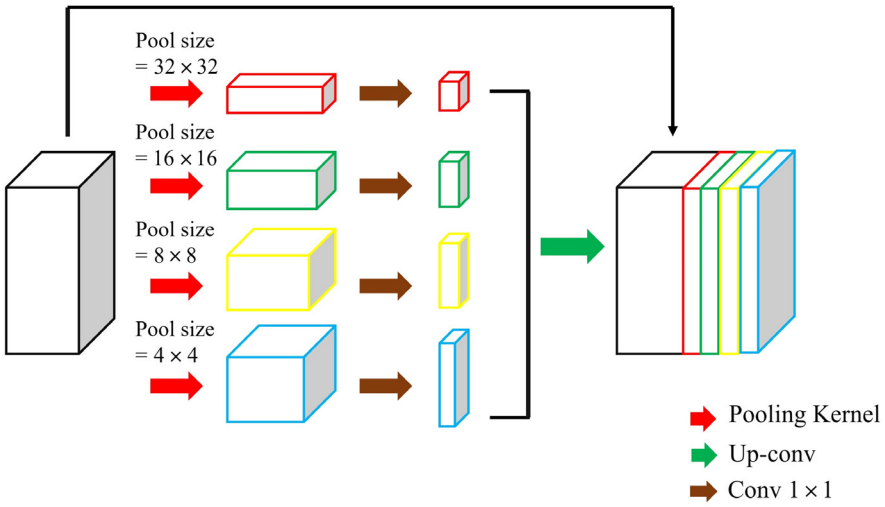

Fig. 4 Diagram of the MSP block.

four sizes of pooling kernels, namely, 32×32, 16×16, 8×8, and 4×4, to divide the spatial feature map into sub-regions of different sizes to perceive contextual relationships and information at different spatial scales.[24] To balance the number of features from different pooling kernels and reduce the computational cost, we add a 1×1 sized convolution filter after each pooling kernel to reduce the dimensionality of the spatial features extracted by each pooling kernel to $1/N$ of its original dimensionality. We then use up-sampling operations to map the spatial features at different sub-region scales back to the same size as the original spatial features. Finally, the spatial features at different sub-regional scales are cascaded to form a feature pyramid of the MSP block, as shown in Fig. 4.

4 Experimental Setup

4.1 *Dataset*

We use sliding cropping to cut the 137 images containing PV facilities into patches with a repetition rate of 0.01. To meet the E-UNET's requirements for input data, the size of each patch is set to 256×256 pixels. In addition, we perform data augmentation by flipping the patches vertically and horizontally to expand the dataset and prevent over-fitting of the training model. Then, we divide the entire dataset into training, validation, and test sets in a ratio of roughly 8:1:1. Therefore, we use 1746, 262, and 230 patches to train, validate, and test our E-UNET, respectively.

4.2 *Experimental Environment*

Keras 2.2.0 with a Tensorflow-gpu 1.7.0 backend is used as the deep learning framework in the experiments. The experiments are conducted on a server with two Intel(R) Xeon(R) Gold 5218R CPUs @ 2.10 GHz with a total of 40 cores, 125 GB RAM, and an NVIDIA Tesla T4 graphics card. The server's operating system is Ubuntu 18.04.5 LTS.

4.3 *Experimental Design*

The experiments are divided into two types: architecture ablation experiments and performance comparison experiments. We first optimize the architecture and parameters of the E-UNET through the architecture ablation experiments; we then analyze and evaluate the performance of the E-UNET in PV semantic segmentation task through the comparative experiments.

4.3.1 *Design of the architecture ablation experiments*

We modify the classical U-Net model to make it capable of processing 12-band Sentinel-2 multi-spectral images, which is referred to as U-Net+. To analyze the contribution of the MSP and MSD modules to the model segmentation performance and the rationality of MSD module parameter selection, we use the U-Net+ as the baseline model and conduct experiments to compare the segmentation performance of different model architectures and parameter selections.

The experiment of only adding the MSP module to the U-Net+ architecture is referred to as U-Net-MSP. The experiments of only adding the MSD module with 3D convolution filters of size $1 \times 1 \times 5$ and $5 \times 5 \times 5$ to the U-Net+ architecture are referred to as U-Net-MSD-k1 and U-Net-MSD-k5, respectively. The experiments of adding both the MSP module and the MSD module with 3D convolution filters of size $1 \times 1 \times 5$ and $5 \times 5 \times 5$ to the U-Net+ architecture are referred to as U-Net-MSP-MSD-k1 and U-Net-MSP-MSD-k5, respectively.

We use the Adam optimization algorithm[25] to train these models with an initial learning rate = 0.001, default hyperparameters $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and batch size = 2. We use the validation set to evaluate the training process and adjust the learning rate values. The learning rate is reduced by a factor of 0.5 if the validation loss does not improve within three epochs. When the validation loss does not improve within 20 epochs, the model training is stopped to

prevent overfitting,[26] and the model with the least validation loss during the training is selected as the training result.

4.3.2 *Design of the comparative experiments*

We select the model architecture and parameters with the best segmentation performance in the architecture ablation experiments as the E-UNET model. In the comparative experiments, we first compare the E-UNET with the U-Net+ and U-Net (only using the RGB bands of the Sentinel-2 data) to illustrate the necessity of using multi-spectral images in the PV detection and the effectiveness of the E-UNET in improving PV detection performance. We then use the RGB bands of the Sentinel-2 data to analyze the segmentation capability of the E-UNET by comparing it with other semantic segmentation networks such as SegNet,[27] FCN,[28] HRNet,[29] and PSPNet.[23]

In addition, we also compare the E-UNET with a pixel-based random forest (RF) classifier,[30] which is widely used in PV detection tasks.[31,32] To balance the computational cost and detection performance of the RF method, we set the number of trees in the forest to 100 and the maximum depth of the forest to 20. According to the conventional setting of RF,[30] the size of the feature subset extracted from each tree node is set to $\sqrt{N}$, where N is the dimensionality of the feature.

4.4 *Evaluation Metrics*

We use five metrics, namely, overall accuracy (OA), recall rate, $F1$, Matthews correlation coefficient (MCC),[33] and kappa coefficient,[34] defined by Eqs. (2)–(6), to evaluate the PV detection performance of each model. These evaluation metrics are calculated from a confusion matrix constructed from the number of pixels that are false negative (FN), false positive (FP), true positive (TP), and true negative (TN).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precison} + \text{Recall}}, \tag{4}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{5}$$

$$\text{Kappa coefficient} = \frac{\text{N}(\text{TP} + \text{TN}) - [(\text{TP} + \text{FP})(\text{TP} + \text{FN}) + (\text{FN} + \text{TN})(\text{FP} + \text{TN})]}{\text{N}^2 - [(\text{TP} + \text{FP})(\text{TP} + \text{FN}) + (\text{FN} + \text{TN})(\text{FP} + \text{TN})]}. \tag{6}$$

4.5 *Uncertainty Analysis*

We randomly generate five image sets for training, validation, and test from the entire image dataset in a ratio of roughly 8:1:1. We use each of these five image sets to train and evaluate the PV detection performance of all models. Following the strategy of Gu et al.,[24] we use the mean and variance of the PV detection performance of each model on these five image sets to analyze the uncertainty of the models.

Table 2 PV detection performance and uncertainty of each model in the architecture ablation experiments. The best-performing model and its metrics are highlighted in bold.

Model	OA	MCC	F1	Kappa coefficient	Recall	Training time
U-Net+	0.987 ± 0.001	0.847 ± 0.024	0.855 ± 0.024	0.919 ± 0.009	0.856 ± 0.023	3 h
U-Net-MSP	0.988 ± 0.000	0.857 ± 0.021	0.865 ± 0.022	0.928 ± 0.004	0.869 ± 0.023	3 h
U-Net-MSD-k1	0.987 ± 0.001	0.851 ± 0.022	0.859 ± 0.023	0.919 ± 0.011	0.862 ± 0.028	4 h
U-Net-MSD-k5	0.988 ± 0.000	0.859 ± 0.022	0.866 ± 0.022	0.931 ± 0.004	0.872 ± 0.021	4 h
U-Net-MSP-MSD-k1	0.988 ± 0.000	0.859 ± 0.021	0.866 ± 0.021	0.930 ± 0.007	0.869 ± 0.021	9 h
U-Net-MSP-MSD-k5	**0.989 ± 0.000**	**0.862 ± 0.022**	**0.869 ± 0.022**	**0.934 ± 0.008**	**0.875 ± 0.023**	**9 h**

5 Results

5.1 *Results of the Architecture Ablation Experiments*

Table 2 lists the PV detection performance and uncertainty of the six models in the architecture ablation experiments. Experimental results show that the U-Net-MSP-MSD-k5 model architecture formed by adding the MSP module and the MSP module with 3D convolution filters of size $5 \times 5 \times 5$ to the U-Net+ structure has the best PV detection performance.

The comparison of experimental results of the U-Net+ and the U-Net-MSP confirms that adding the MSP module capable of aggregating spatial information at different scales to the U-Net+ structure improves the PV detection performance.

The comparison of experimental results of the U-Net+, the U-Net-MSD-k1, and the U-Net-MSD-k5 confirms that adding the MSD module capable of extracting spectral features from multi-spectral images to the U-Net+ structure improves the PV detection performance. The comparison also shows that the U-Net-MSD-k5 model, which extracts spectral features from five adjacent spectral bands of each pixel, improves the PV detection performance more than the U-Net-MSD-k1 model, which only extracts spectral features from a single spectral band of each pixel.

As shown in Fig. 5, for images with similar spectral features in PV and background areas, a good PV detection performance cannot be achieved by adding only the MSD module or only the MSP module to the U-Net+ structure. The red and blue boxes in Fig. 5(a) indicate the area with PV panels installed and the background area without PV panels, respectively. The average spectral values of the pixels in the red and blue boxes are shown in Fig. 5(g). The average spectral values of these two regions are very similar. Figure 5(b) shows the manual labeling results of the PV pixels in Fig. 5(a), which are used as the true values to evaluate the PV detection performance of the models.

Figures 5(c)–5(f) show the PV detection results of the four models, namely, U-Net-MSP-MSD-k5, U-Net-MSD-k5, U-Net-MSP, and U-Net+, respectively. Compared with the true values in Fig. 5(b), it is obvious that the U-Net-MSD-k5, the U-Net-MSP, and the U-Net+ do not fully and accurately detect the PV panels in and around the area indicated by the red box in Fig. 5(a). The U-Net+ misses a large area of PV panels as indicated by the green box in Fig. 5(f). Because the U-Net-MSD-k5 and the U-Net-MSP add the MSD module for sensing spectral features and the MSP module for sensing spatial features at different scales to the U-Net+ structure, respectively, their PV detection results for the same area are much better than that of the U-Net+. As shown in Fig. 5(c), only the U-Net-MSP-MSD-k5, which is formed by adding both the MSD and the MSP modules to the U-Net+ structure, nearly completely detects the PV panels in the area indicated by the green box.

The experimental results shown in Fig. 5 confirm that the simultaneous use of spectral and spatial features extracted at different scales effectively improves the accuracy of PV detection from multi-spectral images. Therefore, we select the U-Net-MSP-MSD-k5 as the final E-UNET model.

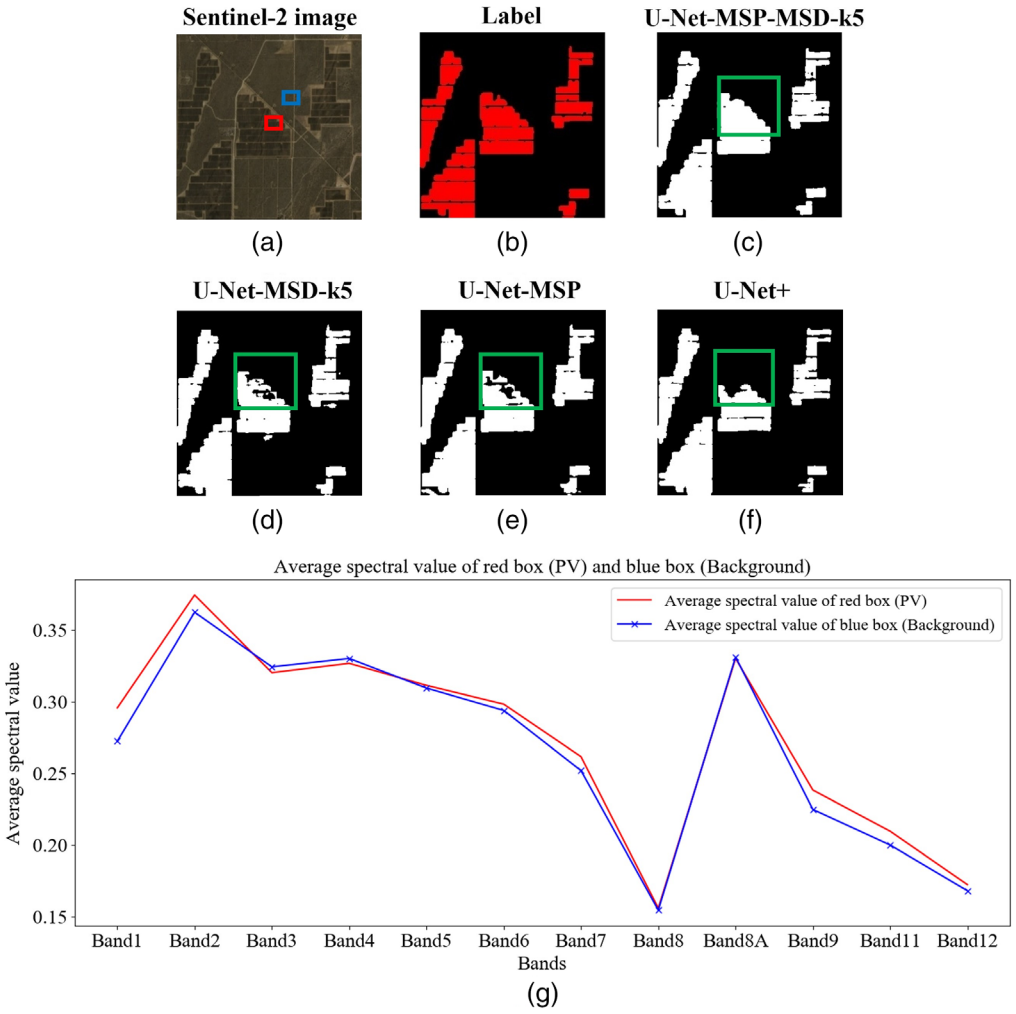

Fig. 5 PV detection results of an image containing PV areas and background areas with similar spectral features: (a) RGB image synthesized from Sentinel-2 multi-spectral image; (b) Manual labeling results of PV panels in (a); (c) U-Net-MSP-MSD-k5; (d) U-Net-MSD-k5; (e) U-Net-MSP; and (f) U-Net+; and (g) average spectral values of the pixels in the red and blue boxes of (a).

5.2 *Results of the Comparative Experiments*

5.2.1 *E-UNET versus other deep learning models*

Table 3 lists the PV detection performance and uncertainty of the E-UNET, U-Net+, U-Net,[12] and four state-of-the-art deep-learning models, namely SegNet,[27] FCN,[28] HRNet,[29] and PSPNet,[23] in the comparative experiments.

The experimental results show that the E-UNET achieves the highest values in all five detection performance evaluation metrics. Compared with the U-Net+, which has the second-best detection performance, the E-UNET's OA, MCC, $F1$, kappa coefficient, and recall metrics improve by 0.2%, 1.5%, 1.4%, 1.5%, and 1.9%, respectively.

Figure 6 shows boxplots of the recall rate of all seven models in the five uncertainty experiments with different sets of training, validation, and test images. The E-UNET has the highest median recall rate and the smallest boxplot height, indicating that the E-UNET achieves the best PV detection performance with the least performance fluctuations in the experiments.

Figure 7 shows the PV detection results of the seven models for images containing dark backgrounds, roads, and vegetation with texture features similar to PV facilities.

Table 3 PV detection performance and uncertainty of the seven models in the comparative experiments. The best-performing model and its metrics are highlighted in bold.

Model	OA	MCC	F1	Kappa coefficient	Recall	Training time
E-UNET	**0.989 ± 0.000**	**0.862 ± 0.022**	**0.869 ± 0.022**	**0.934 ± 0.008**	**0.875 ± 0.023**	**9 h**
U-Net+	0.987 ± 0.001	0.847 ± 0.024	0.855 ± 0.024	0.919 ± 0.009	0.856 ± 0.023	3 h
U-Net	0.982 ± 0.001	0.822 ± 0.026	0.830 ± 0.027	0.870 ± 0.016	0.839 ± 0.024	2.5 h
SegNet	0.973 ± 0.001	0.788 ± 0.022	0.803 ± 0.023	0.845 ± 0.009	0.811 ± 0.030	2 h
FCN	0.976 ± 0.001	0.801 ± 0.025	0.815 ± 0.026	0.860 ± 0.011	0.822 ± 0.026	5 h
PSPNet	0.981 ± 0.001	0.831 ± 0.022	0.843 ± 0.024	0.896 ± 0.008	0.845 ± 0.023	5.5 h
HRNet	0.986 ± 0.001	0.842 ± 0.023	0.850 ± 0.023	0.910 ± 0.008	0.848 ± 0.024	4 h

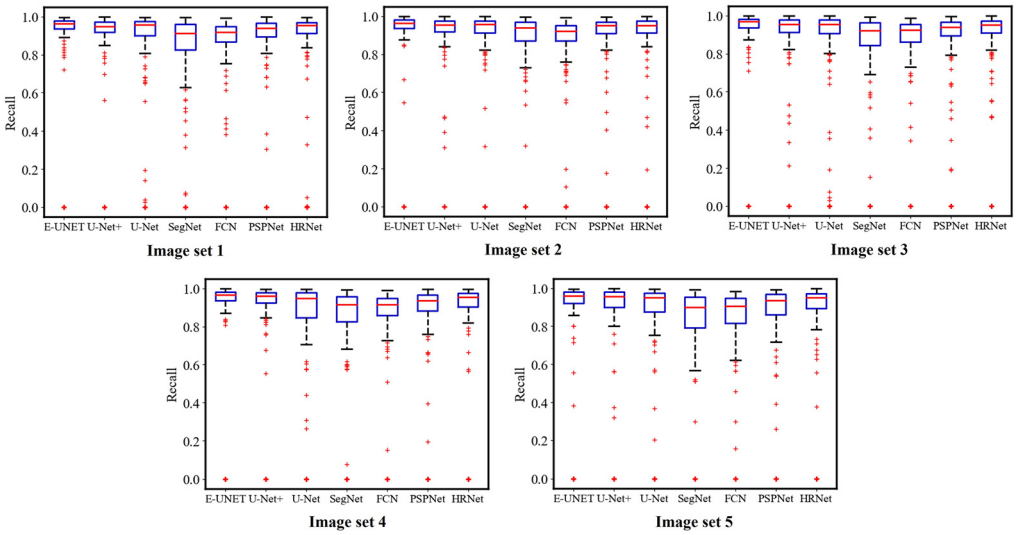

Fig. 6 Boxplots of the recall rate of E-UNET, U-Net+, U-Net, SegNet, FCN, PSPNet, and HRNet on five sets of test images. The outliers are indicated by +. The solid lines within the boxes represent the median value. The lower and upper whiskers represent the minimum and maximum values, respectively. The lower and upper edges of the boxes represent the first (Q1) and third quartiles (Q3), respectively.

The SegNet[27] and the PSPNet[23] have many incorrect detections for images containing vegetation or dark backgrounds. In the detection results of FCN,[28] there are many irregular burrs of different sizes at the edges of the PV panels. Both the U-Net[12] and the HRNet[29] obtain relatively good PV detection results for the images in Figs. 7(a)–7(c), but for the image containing dark backgrounds in Fig. 7(d), they both mis-detect a large area of the background as PV panels.

The U-Net+ obtains relatively complete PV detection results, but the PV panel edges and the gaps between PV panels in its detection results are more blurred than those in the detection results of the E-UNET. It also does not detect the PV panels within the area marked by the red box in Fig. 7(d), whereas the E-UNET accurately detects them.

The experimental results also indicate that the E-UNET outperforms other networks in detecting the overall contour and edge details of PV panels from multi-spectral images. The multiple connections and the complementary spatial-spectral information at different scales between the encoder, the decoder, the MSD path module, and the MSP block in the E-UNET prevent the problem of irregular PV contours when the decoder recovers the image size from partial features that have lost some detailed information in the multiple down-sampling

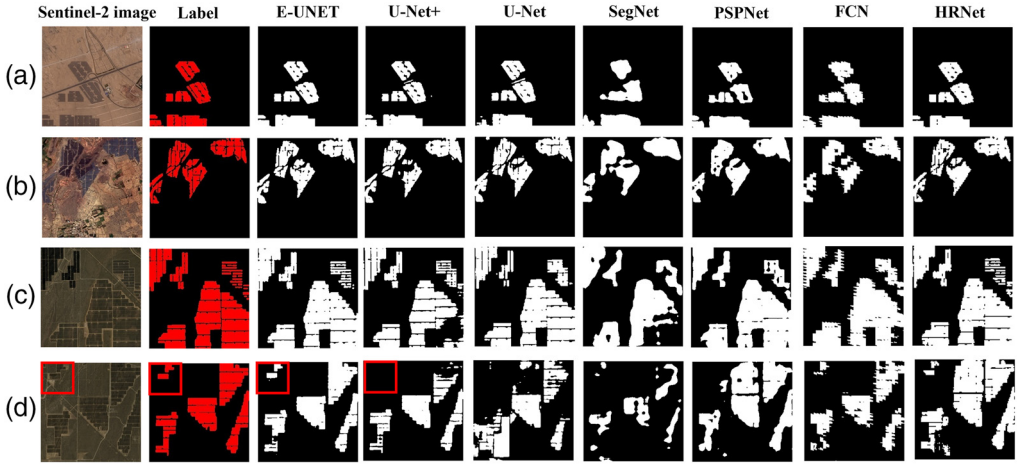

Fig. 7 Four examples of the PV detection results of E-UNET, U-Net +, U-Net, SegNet, PSPNet, FCN, and HRNet. (a) Containing roads; (b) containing vegetation; and (c) and (d) dark backgrounds.

operations. In addition, the MSP module in the E-UNET utilizes multi-scale spatial features captured by multiple receptive fields at different scales to enable the detection of super-large PV facilities and achieve more stable and reliable PV segmentation results.

5.2.2 *E-UNET versus the pixel-based RF classifier*

Table 4 lists the PV detection performance and uncertainty of the E-UNET and the pixel-based RF classifier[30] in the comparative experiments. The experimental results indicate that the E-UNET performs better than the pixel-based RF classifier.[30] Some examples of the PV detection results of these two models are shown in Fig. 8.

Although the pixel-based RF classifier[30] may be more accurate in detecting some small-scale details of PV panels, it does not take advantage of the information provided by neighboring pixels, which is usually strongly correlated, resulting in a lot of scattered and fragmented PV panels and backgrounds in its detection results, as shown in Figs. 8(a), 8(c), and 8(d).

Table 4 PV detection performance and uncertainty of the E-UNET and the pixel-based RF classifier in the comparative experiments. The best-performing model and its metrics are highlighted in bold.

Model	OA	MCC	F1	Kappa coefficient	Recall	Training time
E-UNET	**0.989 ± 0.000**	**0.862 ± 0.022**	**0.869 ± 0.022**	**0.934 ± 0.008**	**0.875 ± 0.023**	**9 h**
RF	0.981 ± 0.001	0.827 ± 0.025	0.834 ± 0.026	0.879 ± 0.019	0.822 ± 0.026	4 h

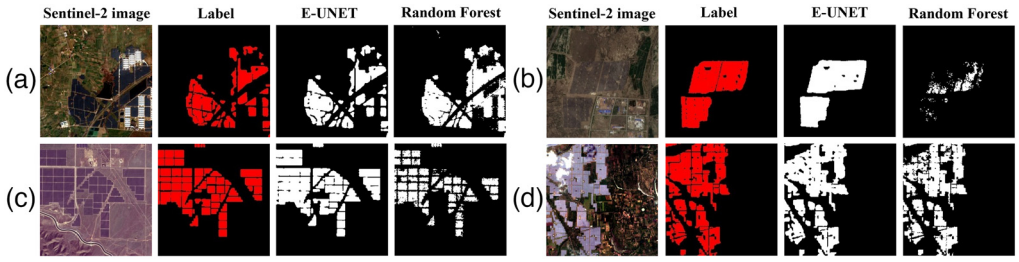

Fig. 8 Four examples of the PV detection results of the E-UNET and the pixel-based RF classifier. (a) Containing vegetation; (b) Dark backgrounds; (c)–(d) Photovoltaic array with many texture details.

Furthermore, as shown in Fig. 8(b), for images in which the backgrounds and PV panels have similar spectral or spatial texture features, the pixel-based RF classifier[30] is more likely to mis-detect PV panels as backgrounds than the E-UNET.

6 Conclusion

In this study, we proposed an end-to-end deep learning framework named the E-UNET to detect PV facilities from Sentinel-2 multi-spectral observation data. The E-UNET was improved from the classical U-Net[12] model by adding a multi-spectral 3D convolution (MSD) path and an MSP block to its U-shaped encoder-decoder structure. Therefore, the E-UNET effectively extracts and integrates spectral and spatial features at different scales to achieve fine-grained and better overall segmentation accuracy. We experimentally compared the PV detection performance of the E-UNET with the pixel-based RF classifier[30] and other deep-learning models of U-Net+, U-Net,[12] SegNet,[27] FCN,[28] HRNet,[29] and PSPNet.[23] The experimental results indicate that the E-UNET achieved the best results in all five performance evaluation metrics, i.e., OA, recall, $F1$, MCC,[33] and kappa coefficient.[34] The experimental results also confirmed that the E-UNET obtained good PV detection performance for images with different topographies and backgrounds. Our future work will involve using the E-UNET to survey larger PV facilities around the world from Sentinel-2 multi-spectral observation data.

7 Appendix A

For convenience, acronyms and abbreviations are given in Table 5.

Table 5 Acronyms and abbreviations used in the paper.

Acronym/Abbreviations	Description
PV	Photovoltaic
E-UNET	Enhanced U-Net
MSD	Multi-spectral 3D convolution
MSP	Multi-scale pooling
OA	Overall accuracy
MCC	Matthews correlation coefficient
CNN	Convolutional neural network
FCN	Full convolutional network
RF	Random forest
FN	False negative
FP	False positive
TN	True negative
TP	True positive

Acknowledgments

This study is supported by the National Natural Science Foundation of China, Urban Agglomeration Planning Evaluation Model for Carbon Peaking based on the Multiple Data (Project No. 52178060), and the International Partnership Program of the Chinese Academy of Sciences (Grant No. 131211KYSB20180002). The authors wish to thank the ESA/Copernicus

for the freely available data (Ref. 35, last accessed: November 22, 2022), as well as the Tensorflow (Ref. 36, last accessed: November 22, 2022) and Keras (Ref. 37, last accessed: November 22, 2022) deep learning development platforms used in this study. No potential conflicts of interest are reported by the authors.

References

1. IEA, "World Energy Outlook 2020," 2020, https://www.iea.org/reports/world-energy-outlook-2020.
2. J. M. Malof et al., "Automatic solar photovoltaic panel detection in satellite imagery," in *Int. Conf. Renew. Energy Res. Appl. (ICRERA)* (2016).
3. J. Camilo et al., "Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery," arXiv:1801.04018 (2018).
4. J. M. Malof, L. M. Collins, and K. Bradbury, "A deep convolutional neural network, with pre-training, for solar photovoltaic array detection in aerial imagery," in *IGARSS 2017 – IEEE Int. Geosci. Remote Sens. Symp.* (2017).
5. X. X. Zhu et al., "Deep learning in remote sensing: a comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.* **5**(4), 8–36 (2018).
6. X. Hou et al., "SolarNet: a deep learning framework to map solar power plants in China from satellite imagery," arXiv:1912.03685 (2019).
7. J. Yu et al., "DeepSolar: a machine learning framework to efficiently construct a solar deployment database in the United States," *Joule* **2**(12), 2605–2617 (2018).
8. C. Li et al., "The first all-season sample set for mapping global land cover with Landsat-8 data," *Sci. Bull.* **62**(7), 508–515 (2017).
9. N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.* **140**, 20–32 (2018).
10. L. Kruitwagen et al., "A global inventory of photovoltaic solar energy generating units," *Nature* **598**(7882), 604–610 (2021).
11. M. Drusch et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.* **120**, 25–36 (2012).
12. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
13. R. A. D. Sofla, T. Alipour-Fard, and H. Arefi, "Road extraction from satellite and aerial image using SE-UNet," *J. Appl. Remote Sens.* **15**(1), 014512 (2021).
14. A. Farasin, L. Colomba, and P. Garza, "Double-step U-Net: a deep learning-based approach for the estimation of wildfire damage severity through Sentinel-2 satellite data," *Appl. Sci.* **10**(12), 4332 (2020).
15. L. Knopp et al., "A deep learning approach for burned area segmentation with Sentinel-2 data," *Remote Sens.* **12**(15), 2422 (2020).
16. J. H. Jeppesen et al., "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.* **229**, 247–259 (2019).
17. U. Muller-Wilm et al., "Sentinel-2 level 2A prototype processor: architecture, algorithms and first results," in *Proc. ESA Living Planet Symp.*, pp. 9–13 (2013).
18. N. Brodu, "Super-resolving multiresolution images with band-independent geometry of multispectral pixels," *IEEE Trans. Geosci. Remote Sens.* **55**(8), 4610–4617 (2017).
19. H. Wang et al., "An automated snow mapper powered by machine learning," *Remote Sens.* **13**(23), 4826 (2021).
20. S. Stankevich et al., "Satellite imagery spectral bands subpixel equalization based on ground classes' topology," in *Int. Conf. Inf. Digit. Technol. (IDT)*, pp. 424–427 (2019).
21. H. Li et al., "Mapping salt marsh along coastal South Carolina using U-Net," *ISPRS J. Photogramm. Remote Sens.* **179**, 121–132 (2021).
22. Y. Chen et al., "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.* **54**(10), 6232–6251 (2016).

23. H. Zhao et al., "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2881–2890 (2017).
24. Z. Gu et al., "CE-Net: context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imaging* **38**(10), 2281–2292 (2019).
25. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
26. R. Naushad, T. Kaur, and E. Ghaderpour, "Deep transfer learning for land use and land cover classification: a comparative study," *Sensors* **21**(23), 8083 (2021).
27. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017).
28. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3431–3440 (2015).
29. K. Sun et al., "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pp. 5693–5703 (2019).
30. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
31. J. M. Malof et al., "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," *Appl. Energy* **183**, 229–240 (2016).
32. J. M. Malof et al., "A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery," in *IEEE Int. Conf. Renew. Energy Res. Appl. (ICRERA)*, pp. 650–654 (2016).
33. B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta Protein Struct.* **405**(2), 442–451 (1975).
34. M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Med.* **22**(3), 276–282 (2012).
35. ESA, "Copernicus open access hub," 2013, https://scihub.copernicus.eu/dhus.
36. M. Abadi et al., "TensorFlow," 2015, https://www.tensorflow.org/.
37. F. Chollet et al., "Keras," 2015, https://keras.io.

Zixuan Dui received her BS degree in computer science and technology from Zhengzhou University in 2021. Currently, she is pursuing an MS degree in electronic information at Shanghai Advanced Research Institute, Chinese Academy of Sciences. Her research interests include machine learning and computer vision.

Yongjian Huang received his MS degree in software engineering from Northeastern University, Shenyang, China, in 2013. Currently, he is pursuing a PhD in information processing at the University of Chinese Academy of Sciences. He also works in the Shanghai Advanced Research Institute, Chinese Academy of Sciences as an engineer. His research interests include natural language processing and data assimulation system.

Jiuping Jin is an engineer at Shanghai Advanced Research Institute, Chinese Academy of Sciences. Her research interests include machine learning and data assimilation system.

Qianrong Gu is a research scientist and senior software engineer at the Carbon Data Research Center of Shanghai Advanced Research Institute, Chinese Academy of Sciences. His research interests include machine learning techniques for remote sensing and Earth observation applications and inversion of global and regional CO_2 fluxes from satellite observations by data assimilation methods.