Received 19 March 2024, accepted 21 May 2024, date of publication 28 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406551

RESEARCH ARTICLE

Combined Hybrid Neural Networks and Swarm Intelligence Optimization Algorithms for Photovoltaic Panel Segmentation From Remote Sensing Images

XIAOQING ZHANG[1], QINGQING QI[1,2,3,4], AND WEIKE LIU[5]

[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China
[2]Extended Energy Big Data and Strategy Research Center, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266104, China
[3]Shandong Energy Institute, Qingdao 266101, China
[4]Qingdao New Energy Shandong Laboratory, Qingdao 266101, China
[5]Center of Information and Network, Shandong University of Science and Technology, Qingdao, Shandong 266590, China

Corresponding authors: Qingqing Qi (qingqingqi@sdust.edu.cn) and Weike Liu (lwk@sdust.edu.cn)

ABSTRACT In the context of traditional energy shortage and climate warming, the development of solar energy, as a clean and renewable energy, is crucial. As an effective way to utilize solar energy resources, photovoltaic (PV) power generation technology has been widely used around the world. Using remote sensing images to extract PV panel information, including location, area, has a positive effect on understanding the development status, planning and construction of regional PV new energy. In this study, a semantic segmentation network called HCT-Net, combined with the hybrid neural networks and the swarm intelligence optimization algorithms, is designed to segment solar PV panels from remote sensing images automatically and accurately. To address the problem of inconsistent segmentation within PV regions, a hybrid encoder, which combines a convolutional neural network and a Transformer, is designed to extract local features with rich detail information and global features with global context dependencies, resulting in enhanced feature representations. The foreground relation module is designed to solve the problem of mis-segmentation of the background into PVs. This module strengthens the model's focus on the target object and suppresses the feature representations of non-PVs by explicitly learning the similarity relationship between the global PV feature representation and the feature representations of other objects, and by adaptively assigning weights according to the similarity. The swarm intelligence optimization algorithm is applied to adjust the learning rate and the balance coefficient of the composite loss function of HCT-Net during training. Experimental results show that compared with the current mainstream semantic segmentation network, the method in this study effectively alleviates the problem of inconsistent segmentation within PV regions and mis-segmentation and has advantages in the complete and accurate extraction of PV panels.

INDEX TERMS Photovoltaic panel extraction, remote sensing image, semantic segmentation, swarm intelligence optimization algorithm, CNN, transformer.

I. INTRODUCTION

As a novel form of clean energy, solar photovoltaic (PV) power generation is being vigorously promoted and developed. According to a report by the National Energy Administration of China, by the end of 2022, China had achieved a cumulative grid-connected PV capacity of 392.04 GW. The increasingly widespread application of solar PV systems has brought great challenges in statistics and planning management. Obtaining regional PV distribution information solely through the existing relevant statistical departments is costly and inefficient. With the advancement of remote sensing technology, the rapid and accurate extraction of PV information over large areas through high-resolution remote

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang.

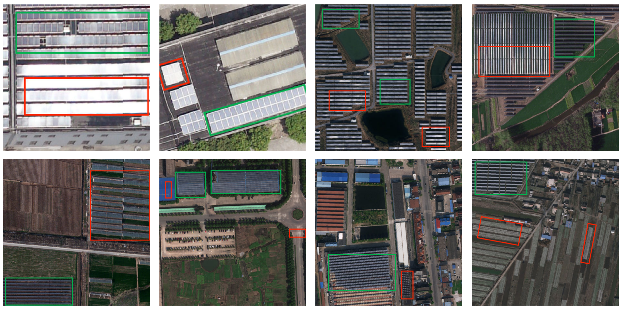

FIGURE 1. Description of the PV panel segmentation problems. First row: examples of the large intraclass feature variations, where the green and red boxes represent PV panels but exhibit different visual appearances. Second row: examples of the complex and diverse environment, where the green boxes represent PV panels, and the red boxes indicate confusing background objects similar to PV panels.

sensing images to observe the development trends of PV energy has become a hot research topic. As shown in Figure 1, the automatic extraction of PV panels from remote sensing images faces the following challenges:

(1) PV panels considerably vary in terms of intraclass features. Influenced by factors, such as installation angles and uneven lighting, PV panels exhibit different visual appearances. This phenomenon can lead to the problem of inconsistent segmentation within PV regions. That is, some PV panels may be incorrectly segmented as background or may have holes in internal areas or breaks at boundaries.

(2) The background environment of PV panels is complex and diverse. In the natural environment, various surface objects that share similar image features with PV panels (e.g., farmland and greenhouses) may appear. These objects are prone to mis-segmentation. That is, background areas are incorrectly segmented as PV panels.

Aiming at the problem of inconsistent segmentation within objects, Yu et al. [1] proposed the Smooth Network, which utilizes a U-shaped architecture, global average pooling, and channel attention blocks to capture multi-scale and global contextual features. Yeung and Lam [2] introduced a boundary-aware spatial attention module to capture the spatial interdependencies between the positions of the boundary features and the context features, thereby improving the consistency of defect features of the same class. Zheng et al. [3] proposed a large kernel pyramid pooling module to capture rich multiscale context with strong continuous feature relations, preserving semantic coherence within objects. Abdollahi et al. [4] proposed a shape-and-connectivity-preserving, deep learning based road identification architecture called SC-RoadDeepNet to overcome the discontinuous results and the quality of road shape and connectivity.

Aiming at the problem of objects of other categories being mis-segmented as target objects, some researchers [5], [6] used attention mechanism to establish relationships between different object feature representations, thereby enhancing feature discriminability among easily confused objects. Zhong et al. [7] designed an interference attenuation module, which effectively alleviates the problem of oversegmentation caused by nonlake objects by deeply mining the feature differences between lake water bodies and other ground objects. Zheng et al. [8] designed a foreground-scene relation module to learn the symbiotic relationship between the foreground and the scene, thus reducing false alarms. Zhou et al. [9] introduced a P2O subnetwork, which utilizes self-attention to model the pixel-to-object relation to offer valuable semantic information of the object.

Inspired by the aforementioned studies, we design a semantic segmentation network (HCT-Net) to enhance the accuracy of extracting PV panels from remote sensing images. HCT-Net uses the encoder-decoder structure as its skeleton. First, a hybrid encoder (HE) combining convolutional neural network (CNN) and Transformer is designed. The CNN part is used to extract local features to preserve low-level details, whereas the Transformer part is used to model long-range contextual dependencies. These two components are integrated to generate more robust local-to-global feature representations that encode rich contextual information. The HE effectively improves the consistency of PV panel segmentation. Second, a foreground relation module (FRM) is designed. This module calculates the similarity between the feature representation of each pixel and the global PV feature representation and then uses the similarity as weights to enhance the feature map, thus suppressing the feature representations of non-PV objects. The FRM effectively alleviates the problem of background objects being mis-segmented as PV panels. In addition, to further improve the segmentation accuracy, three swarm intelligence optimization algorithms, including the Particle Swarm Algorithm (PSO) [10], the Whale Optimization Algorithm (WOA) [11], and the Gannet Optimization Algorithm (GOA) [12], are applied to search for the optimal hyperparameters of the HCT-Net in the training phase, including the learning rate and the balance coefficient of the composite loss function.

The remainder of this article is organized as follows: Related work is reviewed in Section II. Section III introduces the detailed structure of the HCT-Net and the process of swarm intelligence algorithms applied for hyperparameter search. Section IV illustrates the implementation steps of the experiment and analyzes the experimental results in detail. Finally, Section V summarizes the work presented in this study.

II. RELATED WORK

A. PV PANEL SEGMENTATION

Given the advantages of data-driven, automated feature learning and extraction, deep learning-based semantic segmentation methods are widely applied in remote sensing PV panel extraction tasks. Jie et al. [13] introduced a gated fusion module on the encoder-decoder structure to alleviate the problem of difficult recognition of small PV panels and used an edge detection network to finely extract the boundaries of PV panels. Costa et al. [14] used Sentinel-2 multispectral images to explore the PV segmentation experiments of

16 semantic segmentation models with four architectures and four backbone network combinations, concluding that U-Net with EfficientNet-b7 as the encoder is the optimal PV segmentation model. Jianxun et al. [15] proposed PVNet consisting of a coarse prediction module and a fine optimization module to extract the complete region of a single PV panel and optimize its boundary. Zhu et al. [16] developed a deep solar PV refiner to improve the ability to segment small PV panels and refine boundaries. Zhuang et al. [17] proposed a cross-learning-driven U-Net and its extension — adaptive crossnet, to optimize the training process of automatic rooftop PV segmentation and explore better parameter models. However, the inconsistent segmentation of PV panels and the mis-segmentation of background objects have not been sufficiently addressed in the above work.

B. ENCODER-DECODER

Encoder-decoder architectures have been successfully applied to many computer vision tasks. Typically, encoder-decoder networks contain an encoder subnetwork that gradually reduces the feature maps and captures higher semantic information and a decoder subnetwork that gradually recovers the spatial information. Ronneberger et al. [18] proposed the U-Net, which uses a fully symmetric encoder-decoder structure and uses skip connections to concatenate shallow-layer features with corresponding deep-layer features during decoding. Based on U-Net, SegNet [19] recorded pooling indices during encoding and used these recorded pooling indices to supervise decoding. RefineNet [20] introduced numerous refinement blocks to improve the ability of hierarchical feature maps to capture semantic information. Based on U-Net structure, TransUNet [21] improved the encoder part by combining ResNet and Vision Transformer (ViT). Motivated by the success of the above work, this study adopts this structure as the model skeleton, which uses an encoder to generate features at different levels and a decoder to achieve feature map resolution recovery and to fuse the features at different levels to obtain a refined segmentation.

C. CNN AND TRANSFORMER HYBRID METHOD

CNNs have limitations in global context feature extraction, whereas Transformers can leverage self-attention mechanism to capture long-range contextual information. However, Transformer ignores the details of local features. To maximize the advantages of CNN and Transformer, some works proposed to combine CNN with Transformer for semantic segmentation work. Zhang et al. [22] proposed to use a combination of Swin Transformer and atrous spatial pyramid pooling module as the encoder and to use CNN with the addition of skip connections as the decoder to improve feature fusion. TransUNet [21] sequentially used ResNet to extract local features and ViT to extract global features in the encoder part. Xiao et al. [23] built a CNN-Transformer two-branch backbone network. Inspired by these works, this study designs an HE based on ResNet [24] and Mix Transformer (MiT) [25] to generate local detailed features and global semantic features with rich contextual information.

D. ATTENTION MECHANISM IN COMPUTER VISION

Attention mechanism in computer vision can be categorized into two types: self-attention mechanism and scaling attention mechanism. In this study, we focus on the scaling attention mechanism, the idea of which is to learn the weights of different channels or spatial locations adaptively and then use the learned weights to weight the original feature map. Hu et al. [26] proposed a squeeze-and-excitation (SE) block, which uses global pooling to generate channel-wise attention. The selective kernel unit [27], efficient channel attention module [28], and and coordinate attention [29] further boost the performance of the SE block. The convolutional block attention module [30] integrates two attentions, where the channel attention is to enhance the feature representation of different channels and the spatial attention is to extract the key information at different locations in the space. The FRM proposed in this study augments the original feature maps by using the similarity between the foreground feature representation and the feature representations of other objects as weights, thus enhancing the model's focus on the target objects and suppressing the representations of non-PV features.

E. APPLICATION OF SWARM INTELLIGENCE OPTIMIZATION ALGORITHMS IN DEEP NEURAL NETWORKS

Some studies have applied swarm intelligence optimization algorithms (SIO) to deep neural networks (DNN) related fields, such as neural architecture search and hyperparameter tuning. Wang et al. [31] proposed a novel deep architecture generation model based on Aquila optimization and genetic algorithm for efficient CNN architecture search. YOLOv4 [32] used genetic algorithm for selecting the optimal hyperparameters during network training on the first 10% of time periods, including the learning rate, the momentum, the IoU threshold for assigning ground truth and the loss normalizer. Dang et al. [33] introduced a medical image segmentation method based on a comprehensive learning PSO, which combines k segmenters based on integrated learning to make a final decision, and finds the combination weights using a comprehensive learning PSO. Zhang and Lim [34] proposed a PSO-enhanced ensemble deep neural network for optic disc segmentation in retinal images. This approach is based on an improved PSO and transfer learning strategy to search for the optimal hyperparameters, including learning rate and momentum. Due to the advantages of simple structure and global search of SIO, this study applies and compares three swarm intelligence optimization algorithms to the designed semantic segmentation network to search for

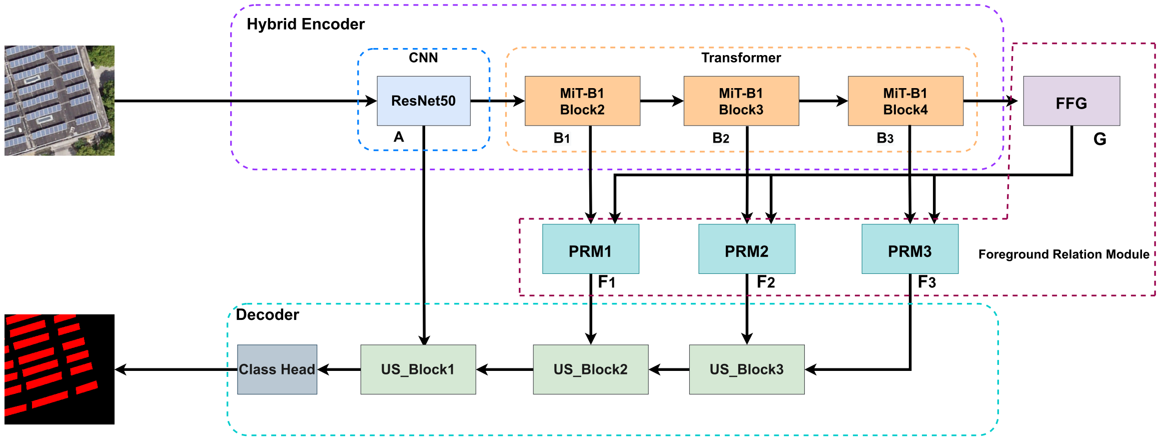

FIGURE 2. Structure of HCT-Net.

the optimal hyperparameters in the training phase, thereby further improving segmentation accuracy.

III. METHOD

A. OVERVIEW OF HCT-NET

This study designs HCT-Net, a semantic segmentation network for PV panel extraction. The network architecture is based on an encoder-decoder structure, maximizing CNN's ability to capture local features and Transformer's capability to model long-range contextual dependencies. HCT-Net consists of an HE, an FRM, and a decoder. Its structure is shown in Figure 2.

(1) HE: The HE designed in this study consists of CNN and Transformer in series. The aim is to generate sequentially local features with rich detailed information and global features with global context dependencies. These features can help the model better understand the characteristics of the PV panels themselves and their surrounding environment, thus improving the consistency of PV panel segmentation. For an input image $X \in \mathbb{R}^{C\times H\times W}$, where H, W, and C represent the height, width, and number of channels of the image, respectively, the process starts with the CNN part, which extracts local features, resulting in the feature map $A \in \mathbb{R}^{C_0\times(H/4)\times(W/4)}$. The CNN part consists of the 7×7 convolutional layer and stage 1 of ResNet50. Then, the feature map A is sequentially passed through three Transformer blocks to generate feature map $B_i \in \mathbb{R}^{C_i\times(H/2^{i+2})\times(W/2^{i+2})}$, where $i = 1, 2, 3$. The Transformer part uses block2-block4 of the MiT B1 version. Correspondingly, the channel numbers of C_i are 128, 320, and 512, respectively.

(2) FRM: The FRM is added between the HE and the decoder. Its purpose is to explicitly model the foreground relation of the feature map output by the HE to enhance further the discriminative ability of the features, thereby reducing the negative influence of non-PV objects. The detailed structure of this module is described in Section III-B.

(3) Decoder: The role of the decoder is to gradually restore the high-level feature map to the resolution of the original input image and to fuse the upsampled feature maps with the corresponding scale feature map output from the HE/FRMs using concatenation and 3×3 convolution to achieve finer semantic segmentation.

B. FRM

As shown in Figure 2, the FRM consists of two parts: the foreground feature generator (FFG) and the point-wise relation module (PRM). The FFG is used to generate the global PV feature representation. By contrast, the PRM is used to calculate the similarity relationship between the feature representation of each pixel and the global PV feature representation and to enhance the input feature map with the similarity as weights. Considering that the deeper feature maps contain stronger semantic information, while the shallower feature maps contain more local detailed information [35]. The detailed features may not be sufficient to help segment the foreground from the background accurately. Therefore, only feature maps B_1, B_2, B_3 are processed by the PRM.

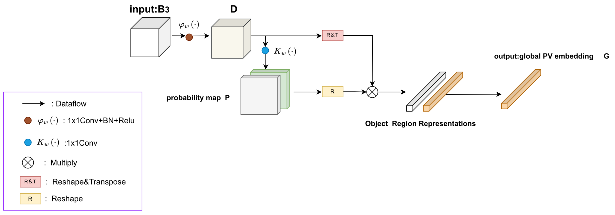

FIGURE 3. Process diagram of the FFG.

1) FFG

Given that B_3 is the deepest-level feature map, it contains the richest semantic information. Therefore, the global PV feature representation of the image is generated by the FFG using B_3. The structure of the FFG is shown in Figure 3.

First, a CBR (1×1 convolution, BN, and ReLU) is used to perform a channel dimension mapping of the feature map B_3 to form $D \in \mathbb{R}^{U\times(H/32)\times(W/32)}$. A 1×1 convolution is

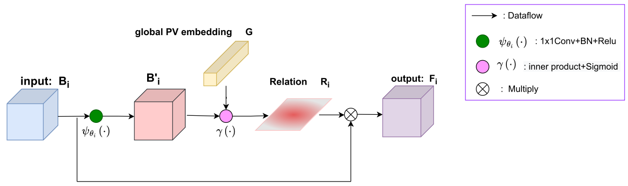

FIGURE 4. Computation detail of relation modeling for the pyramid level i in the PRM.

then used to map D to $P \in \mathbb{R}^{2\times(H/32)\times(W/32)}$. P represents the category probability distribution in D, which is used to learn the global PV feature representation under the supervision of the ground truth. Next, we reshape P to $\mathbb{R}^{2\times N}$ and reshape and transpose D to $\mathbb{R}^{N\times U}$, where $N = (H/32) \times (W/32)$, U represents the number of channels in the global PV feature representation. Afterward, we perform matrix multiplication between P and D to obtain object region representations $\in \mathbb{R}^{2\times U}$. Finally, the global PV feature representation $G \in \mathbb{R}^{U}$ is formed by taking the feature vector at the corresponding position in the object region representations.

2) PRM

After obtaining the global PV feature representation G, we combine the feature maps B_i (i = 1,2,3) and G into (B_1, G), (B_2, G), and (B_3, G) and send them to different PRMs for foreground correlation modeling, respectively. The structure of the PRM is shown in Figure 4. For a given input (B_i,G), a $\Psi_{\theta_i}(\cdot)$ (1×1 convolution, BN, and ReLU) is first used to the map B_i to the B'_i with the same number of channels as the feature vector G. The global PV feature representation G is shared for each pyramid level because the PV semantics is scale invariant across all pyramid levels. Then, a similarity relation is calculated for B'_i and G to generate the relation map R_i, which can be formulated by:

$$R_i = \gamma(B'_i, G) \tag{1}$$

where γ denotes the similarity estimation function, which is implemented by an element-wise inner product followed by a sigmoid function. Finally, the original feature map B_i is multiplied with R_i to produce the enhanced feature map F_i as the output.

C. COMBINATION OF SIO AND DNN

In this study, it is a joint loss to optimize the model parameters, defined by the following formula:

$$L = L_m + \alpha L_a \tag{2}$$

where L_m is the main loss function, which is calculated from the final output of the network with ground truth; L_a is the auxiliary loss for supervising the learning of the global PV feature representation mentioned in Section III-B1. L_m and L_a are cross-entropy loss functions. α is the hyperparameter that balances L_m and L_a.

HCT-Net has two hyperparameters that need to be adjusted during training: the learning rate (Lr) and the loss balance coefficient (α). Manual hyperparameter tuning often relies on the experience of the researcher, which may result in only locally optimal solutions. SIO automatically searches the entire hyperparameter space, helping to find optimal or near-optimal hyperparameter combinations. In this study, SIO is combined with the proposed HCT-Net to search for the optimal Lr and α during training.

Since the iterations of SIO are typically hundreds of times, and the number of epochs for training the basic neural network scenario is also typically hundreds of times, SIO applied to neural networks for hyperparameter search is typically expensive and time-consuming, possibly requiring hundreds or thousands of GPU hours. The time for a one-time conventional training mode is estimated in the following equation.

$$T = T_{ter} \times P \times T_{base} \tag{3}$$

where T is the total time, T_{ter} is the number of iterations of the SIO, P is the number of individuals in the population, and T_{base} is the training time for the basic scenario of the neural network. Assuming T_{ter}=100, P=3, T_{base}=4h, then T=1200h.

To reduce the cost of SIO hyperparameter search, we introduce the fine-tuning strategy of transfer learning. The entire training process is divided into a scratch training phase and a fine-tuning based SIO hyperparameter search phase. Specifically, the HCT-Net is first trained from scratch using a set of empirically based hyperparameters. Then, the model parameters are initialized using the weights obtained in the above stages, and the search of Lr and α is performed using SIO based on the fine-tuning of a small number of epochs.

The entire process mentioned above is denoted as SIONN, and its pseudo-code description is given in Algorithm 1.

IV. EXPERIMENTS

A. DATASET DESCRIPTION

The dataset [36] used in the experiment consists of three parts: PV01, PV03, and PV08. Among them, PV01 consists of 645 UAV images with a spatial resolution of 256 × 256 and a ground sampling distance (GSD) of 0.1 m, focusing on fine-grained rooftop PV panels (including flat concrete, steel tile, and brick roofs); PV03 consists of 2308 aerial images with a spatial resolution of 1024 × 1024 and a GSD of 0.3 m, focusing on PV panels of ground scenes (including shrublands, grasslands, farmlands, etc.); PV08 consists of 763 satellite images with a spatial resolution of 1024 × 1024 and a GSD of 0.8 m, focusing on large rooftop and ground-mounted centralized PV arrays. Given the lower resolution of PV08, it cannot meet the requirements for fine segmentation. In our experiments, we only use PV01 and PV03. The images in PV01 and PV03 are first randomly divided into training set, validation set, and test set in a ratio of 7:1:2, respectively, and then merged. The final training set contains 2065 images, the validation set contains 296 images,

Algorithm 1 Pseudo-Code of the SIONN

```
Input: $E_p$: empirical parameters, $D_t$: training set, $D_v$: validation set, $E_{ts}$: number of epochs for which HCT-Net is trained from scratch, $E_{fts}$: number of epochs for fine-tuning training, $N$: population size, $D$: problem dimension, $T_{max}$: maximum number of iterations for SIO
Output: GP: global best position, $\mathrm{GW_{sio}}$: global best weight in SIO search
Initialize: GbestIoU = 0, $\mathrm{GW_{ts}}$ = None, Gbestfit = $+\infty$, GP = None, $\mathrm{GW_{sio}}$ = None, Randomly initialize the population $X$
Stage of training from scratch
for ($i = 1$ to $E_{ts}$) do
    $\mathrm{IoU}_i$, $w_i$ = training_validation($D_t$, $D_v$, $E_p$)
    if ($\mathrm{IoU}_i$ > GbestIoU) then
        GbestIoU = $\mathrm{IoU}_i$, $\mathrm{GW_{ts}} = w_i$
    end if
end for
Stage of SIO to search the optimal hyperparameters
Initialization of model parameters with $\mathrm{GW_{ts}}$
while ($t < T_{max}$) do
    for ($i = 1$ to $N$) do
        $f_{t,i}$,$w_{t,i}$ = EvaluateFitness($x_i$, $E_{fts}$, $D_t$, $D_v$)
        if ($f_{t,i}$ < Gbestfit) then
            Gbestfit = $f_{t,i}$, GP = $x_i$, $\mathrm{GW_{sio}} = w_{t,i}$
        end if
    end for
    Updating the location of individuals
end while
return GP, $\mathrm{GW_{sio}}$
```

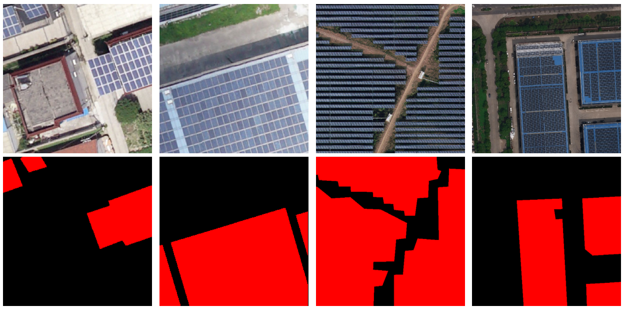

FIGURE 5. Samples from dataset. The first row is the original image and the second row is the label.

and the test set contains 573 images. Figure 5 shows some sample images and labels from the dataset.

B. IMPLEMENTATION DETAILS

In this study, IoU, precision, recall, and F1 are used as evaluation metrics.

IoU represents the degree of overlap between the segmentation result and the ground truth:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

Precision represents the proportion of samples that are actually positive among the samples classified as positive:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

Recall represents the proportion of samples correctly classified as positive among all samples that are actually positive:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

F1 represents the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{7}$$

where TP (True Positive) is the number of positive pixels correctly predicted, FP (False Positive) is the number of pixels predicted by the model to be positive samples but labeled as negative samples, and FN (False Negative) is the number of pixels predicted to be negative samples but labeled as positive samples.

Experiments are implemented on the basis of the PyTorch1.10 deep learning framework, and the hardware environment is a single NVIDIA A100 GPU.

1) DETAILS OF TRAINING FROM SCRATCH

We use a poly learning rate policy where the initial learning rate is multiplied by $(1 - \frac{\mathrm{epoch}}{\mathrm{max\_epoch}})^{0.9}$ after each epoch. Moreover, we utilize the Adam optimizer with an initial learning rate of 0.001 for training. The batch size is set to 8, and a total of 100 epochs are trained. For weight initialization, we initialize the CNN and Transformer part of the HE with the ResNet50 and MiT-B1 weights from ImageNet pretraining, respectively, and use Kaiming initialization [37] for the rest of our model. α is set to 0.4 following the literature [38], [39] at this stage. Figure 6 illustrates the changes of loss and IoU values during the scratch training of HCT-Net.

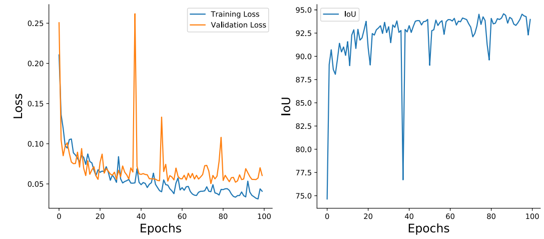

FIGURE 6. Changes of loss and IoU values during the scratch training of HCT-Net.

2) DETAILS OF SIO TO SEARCH THE OPTIMAL HYPERPARAMETERS

We compare the PSO, the GOA, and the WOA. The number of individuals in the population of each algorithm is three with dimension two, where the first dimension is the Lr with upper and lower bounds of [1e-8,1e-3], the second dimension is the α with upper and lower bounds of [0.1,1]. The number

TABLE 1. Parameter settings for each related algorithm.

Algorithms	Parameters	Values
PSO	c_1 and c_2	2
	w	Random numbers between 0 and 1
WOA	a	Linearly decreases from 2 to 0
	a_2	Linearly decreases from -1 to -2
	b	1
GOA	r_1 to r_6	Random numbers between 0 and 1
	c	0.2
	β	1.5

TABLE 2. Results of different encoders.

Encoders	IoU(%)	Precision(%)	Recall(%)	F1(%)
ResNet50	92.86	**97.52**	95.11	96.31
MiT-B1	92.35	96.54	95.12	95.82
HE	**93.63**	96.73	**96.69**	**96.71**

of iterations is 100 and the number of fine-tuning epochs is one. The IoU on the validation set is used as the fitness score. The settings of other parameters can be found in Table 1. The settings of these parameters refer to the corresponding parameter settings in the literature [12].

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) ABLATION STUDIES

We conducted ablation experiments to provide a more intuitive assessment of the effectiveness of the HE and the FRM. All ablation experiments were performed using the same encoder-decoder structure, and the decoder was kept consistent to ensure a fair comparison.

(1) Influence of different encoders. Table 2 shows that compared with using the pure CNN encoder (ResNet) and the pure Transformer encoder (MiT-B1), the IoU is improved by 0.77%, 1.28% and the recall is improved by 1.58%, 1.57%, respectively, when using the HE. The visualization results in Figure 7 show that some PV panels exhibit abrupt changes in texture and color due to uneven lighting conditions, especially the examples in the third and fourth rows. When using a pure CNN encoder or a pure Transformer encoder, such PV panels cannot be completely segmented (partial or complete absence, cavities inside, fracture at the boundaries). However, the above problem is effectively alleviated when using HE. These results verify that the use of HE has a reliable performance advantage in improving the consistent segmentation of PV panels.

(2) Influence of the FRM. We use the pure CNN encoder, the pure Transformer encoder, and the HE as baselines. We add the FRM to each of them to evaluate its effect on network performance. As shown in Table 3, when compared with Table 2, the IoU values and F1 values of the three baselines improved after using the FRM. The visualization results in Figure 8 show that some objects in the background are similar to the PV panels in terms of texture or color, such as farmland and rooftop structures. Without using the FRM, all three baseline networks show varying degrees of mis-segmentation. However, this problem is effectively alleviated after using the FRM. The network with ResNet50 as the encoder decreases in precision with the addition of the FRM, but the comprehensive metrics IoU and F1 are improved. Combining (d) and (g) in Figure 8, we analyze that, influenced by the specific structure, although the addition of FRM in the ResNet50 baseline alleviates the mis-segmentation of objects similar to PV panels to some extent, the increase in the recall value shows that the precision value may be decreased due to the introduction of other background information (e.g., edges are transitionally segmented). However, under the baseline based on the pure Transformer and the HE designed in this study, the FRM plays a positive role, and both comprehensive metrics and precision are improved. Furthermore, when the HE is combined with the FRM, the comprehensive metrics reach the highest. The above results show that the FRM can improve the anti-interference ability of the network and indicate that the HCT-Net (HE+FRM) designed in this study is the best combination scheme.

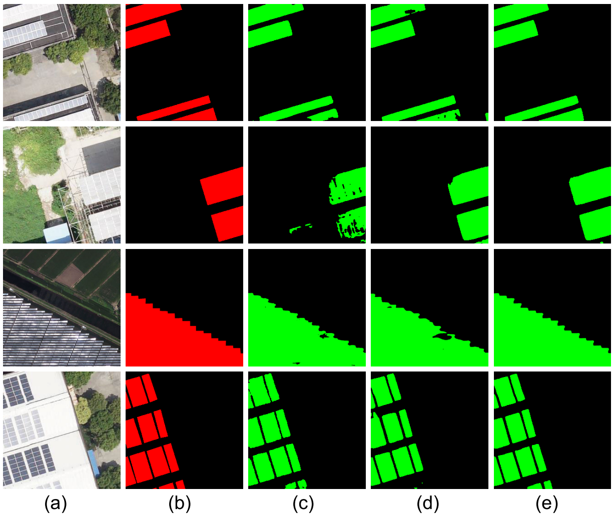

FIGURE 7. Experiments with different encoder effects: (a) Images; (b) GT; (c) MiT-B1; (d) ResNet50; (e) HE.

TABLE 3. Results of the FRM.

Models	IoU(%)	Precision(%)	Recall(%)	F1(%)
ResNet50+FRM	93.69	96.55	**96.93**	96.74
MiT-B1+FRM	93.33	96.57	96.53	96.55
HE+FRM	**94.01**	**97.14**	96.68	**96.91**

2) THE HYPERPARAMETER OPTIMIZATION RESULTS FOR DIFFERENT SWARM INTELLIGENCE ALGORITHMS

Figure 9 shows the change process of the global optimal individuals (i.e., global optimal Lr and global optimal α) of the three algorithms during the hyperparameter search process based on fine-tuning. It can be seen from Figure 9

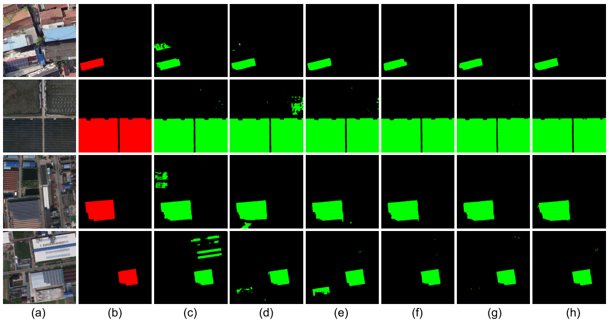

FIGURE 8. Effect of FRM: (a) Images; (b) GT; (c) MiT-B1; (d) ResNet50; (e) HE; (f) MiT-B1+FRM; (g) ResNet50+FRM; (h) HE +FRM (HCT-Net).

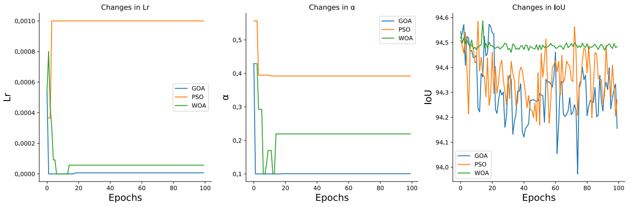

FIGURE 9. Changes in optimal Lr, α, and IoU during the iterations of three optimization algorithms.

TABLE 4. Results of the three optimization algorithms.

Algorithm	Best Lr	Best α	IoU(%)	Precision(%)	Recall(%)	F1(%)
PSO	1.00e-3	0.3922	94.00	97.05	96.77	96.91
GOA	7.01e-6	0.1006	94.03	97.04	**96.81**	96.92
WOA	5.68e-5	0.2194	**94.11**	**97.22**	96.71	**96.97**

that the optimal Lr and optimal α gradually converge during the iterative search of all three algorithms.

Table 4 shows the final values of Lr and α searched by each algorithm, as well as the results of the four metrics tested on the test set with the corresponding weights. We can find that compared to the other two algorithms, the WOA achieves the best results in the IoU, precision and F1 metrics. The corresponding Lr is 5.68e-5 and α is 0.2194. In addition, all metrics are improved after HCT-Net using WOA and GOA parameter search strategies compared to before. Although there is a phenomenon that PSO causes the metrics to decrease, the difference is not significant.

The above results show that the SIO parameter search method based on the fine-tuning strategy used in this study is mostly effective and can further find better models than the empirical parameters. However, since different algorithms have different preferences, advantages, and disadvantages, it is necessary to try different algorithms more in order to find the one best suited to the problem.

3) COMPARED WITH OTHER SEMANTIC SEGMENTATION NETWORKS

We compare HCT-Net with some CNN-based methods (including U-Net [18], DeepLabv3+ [40], PSPNet [41]) and some Transformer-based methods (including SETR [42], Swin-UNet [43], TransUNet [21]) to demonstrate the superiority of the method in this study. In the CNN-based models, all networks use ResNet50 as the backbone, except for U-Net, which follows the original design. In the Transformer-based models, we choose the corresponding open-source pretrained Transformer with a size similar to ResNet50 to ensure a fair comparison.

TABLE 5. Comparison results of HCT-Net with semantic segmentation networks based on CNN and Transformer structures.

NetWork	Backbone	IoU(%)	Precision(%)	Recall(%)	F1(%)
U-Net	-	93.05	96.97	95.84	96.40
PSPNet	ResNet50	92.91	96.73	95.92	96.32
DeepLabv3+	ResNet50	93.05	97.11	95.69	96.40
SETR	*ViT_Base*	79.91	89.12	88.31	88.70
TransUNet	*R50-ViT_Base*	92.47	96.58	95.60	96.08
Swin-UNet	*Swin_Tiny*	89.63	94.02	95.04	94.53
Ours(HCT-Net)	*R50-Mit_B1*	**94.11**	**97.22**	**96.71**	**96.97**

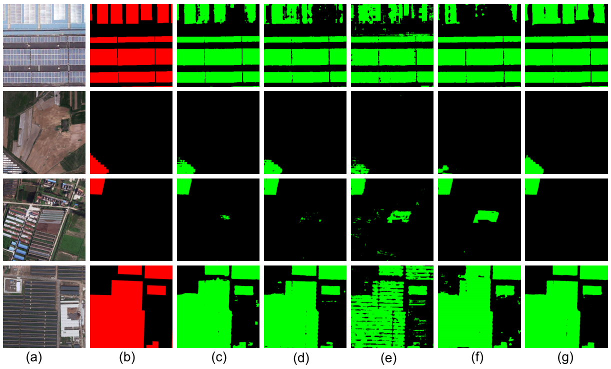

FIGURE 10. Example visualization of HCT-Net compared with CNN and Transformer-based semantic segmentation models: (a) Images; (b) GT; (c) U-Net; (d) DeepLabv3+; (e) Swin-UNet; (f) TransUNet; (g) HCT-Net.

As shown in Table 5, HCT-Net achieves the best IoU and F1 values. Compared with DeepLabv3+, the best-performing CNN-based model, its IoU value is 1.06% higher, and its F1 value is 0.57% higher. It has a 1.64% higher IoU and 0.89% higher F1 than the best-performing TransUNet model based on the Transformer. The visualization comparison in Figure 10 reveals that HCT-Net outperforms other models not only in completely segmenting PV panels, but also in accurately segmenting PV panels and background objects in complex background environments. In detail, in the examples of the first and second rows, comparing HCT-Net, the other methods clearly show inconsistent segmentation, i.e., the PV panels of the exposure transition are lost. In the examples of the third and fourth rows, HCT-Net successfully segments objects similar to PV panels (e.g., sheds and roof structures) into the background. The above results benefit from the fact that the HE generates local features with rich spatial detail information and global features with global context dependencies, and the FRM suppresses the interference of background information and improves the model's focus on the target object.

TABLE 6. Comparison of model size and computational complexity.

Method	FLOPs(Gbps)	Params(Mb)	IoU(%)
DeepLabv3+	173.40	39.76	93.05
U-Net	160.76	17.26	93.05
PSPNet	184.73	46.71	92.91
SETR	89.55	87.39	79.91
Swin-UNet	30.87	27.15	89.63
TransUNet	129.29	93.23	92.47
Ours	**20.15**	**16.76**	**94.11**

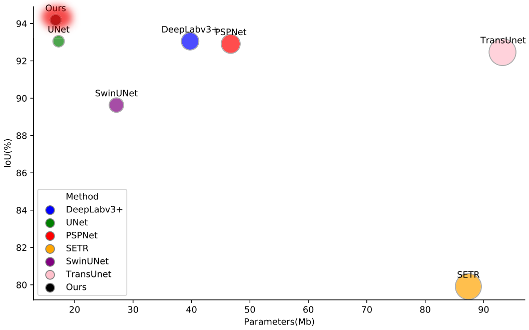

FIGURE 11. Trade-off between model size and accuracy.

In order to demonstrate that the excellent performance of HCT-Net is due to its efficient structural design rather than relying on the huge parameter size, we conduct a comparative analysis with other semantic segmentation networks in terms of number of parameters, computational complexity, and IoU. The detailed results are shown in Table 6. All figures are obtained by inference with an input size of [1×3×512×512] on a single NVIDIA A100 GPU running CUDA 11.0. As shown in Table 6, our HCT-Net requires the fewest parameters and the lowest FLOPs while achieving the highest IoU value. Notably, the number of parameters of our model is similar to that of U-Net. However, the computational complexity is nearly eight times lower than that of U-Net, and the IoU is also higher. Figure 11 shows in a more intuitive form that HCT-Net achieves a good trade-off between model complexity and segmentation accuracy. The above results can prove that the excellent performance of HCT-Net mainly depends on the effectiveness of the network structure design.

V. CONCLUSION

In this study, a semantic segmentation network called HCT-Net, which is based on an encoder-decoder structure, a hybrid of CNN and Transformer and combined with swarm intelligence optimization algorithms, is designed to accurately extract PV panels from remote sensing images. To improve the consistency of PV panel segmentation, we design an HE that combines CNN and Transformer to extract local detailed features and global semantic features. These features are fused in the decoder to generate a more robust feature representation. An FRM is designed to explicitly model the relation between PV feature representations and other object feature representations and thus enhance the feature discrimination ability, thereby alleviating the problem of background objects being mis-segmented as PV panels. Three swarm intelligence optimization algorithms, including PSO, GOA, and WOA, are introduced and combined with the fine-tuning strategy of transfer learning to adjust the hyperparameters of HCT-Net in the training phase, including the learning rate and the balance coefficient of the composite loss function. We verify the effectiveness of the HE and the FRM by performing ablation experiments on a publicly available PV semantic segmentation dataset. In addition, hyperparametric search using SIO further improves the segmentation accuracy. Compared with other semantic segmentation networks, HCT-Net not only achieves better segmentation accuracy, but also has fewer parameters and less computational complexity. In future work, we will try to extend HCT-Net to make it suitable for semi-supervised semantic segmentation to reduce the cost of manual label production, and apply it to PV information extraction from remote sensing images in real regions.

REFERENCES

[1] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1857–1866.

[2] C.-C. Yeung and K.-M. Lam, "Attentive boundary-aware fusion for defect semantic segmentation using transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[3] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 15–28, Dec. 2020.

[4] A. Abdollahi, B. Pradhan, and A. Alamri, "SC-RoadDeepNet: A new shape and connectivity-preserving road extraction deep learning-based network from remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617815.

[5] J. Chen, H. Wang, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Strengthen the feature distinguishability of geo-object details in the semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2327–2340, 2021.

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[7] H.-F. Zhong, Q. Sun, H.-M. Sun, and R.-S. Jia, "NT-Net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627513.

[8] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4096–4105.

[9] F. Zhou, R. Hang, H. Shuai, and Q. Liu, "Hierarchical context network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407612.

[10] Y. Shi, "Particle swarm optimization," *IEEE Connections*, vol. 2, no. 1, pp. 8–13, Feb. 2004.

[11] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, May 2016.

[12] J.-S. Pan, L.-G. Zhang, R.-B. Wang, V. Snášel, and S.-C. Chu, "Gannet optimization algorithm: A new metaheuristic algorithm for solving engineering optimization problems," *Math. Comput. Simul.*, vol. 202, pp. 343–373, Dec. 2022.

[13] Y. Jie, X. Ji, A. Yue, J. Chen, Y. Deng, J. Chen, and Y. Zhang, "Combined multi-layer feature fusion and edge detection method for distributed photovoltaic power station identification," *Energies*, vol. 13, no. 24, p. 6742, Dec. 2020.

[14] M. V. C. V. D. Costa, O. L. F. D. Carvalho, A. G. Orlandi, I. Hirata, A. O. D. Albuquerque, F. V. E. Silva, R. F. Guimarães, R. A. T. Gomes, and O. A. D. C. Júnior, "Remote sensing for monitoring photovoltaic solar plants in Brazil using deep semantic segmentation," *Energies*, vol. 14, no. 10, p. 2960, May 2021.

[15] W. Jianxun, C. Xin, J. Weicheng, H. Li, L. Junyi, and S. Haigang, "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, May 2023, Art. no. 103309.

[16] R. Zhu, D. Guo, M. S. Wong, Z. Qian, M. Chen, B. Yang, B. Chen, H. Zhang, L. You, J. Heo, and J. Yan, "Deep solar PV refiner: A detail-oriented deep learning network for refined segmentation of photovoltaic areas from satellite imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103134.

[17] L. Zhuang, Z. Zhang, and L. Wang, "The automatic segmentation of residential solar panels based on satellite images: A cross learning driven U-Net method," *Appl. Soft Comput.*, vol. 92, Jul. 2020, Art. no. 106283.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," 2015, *arXiv:1511.00561*.

[20] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5168–5177.

[21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[22] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[23] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605116.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[25] 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. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[27] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[29] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

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

[31] T.-T. Wang, S.-C. Chu, C.-C. Hu, H.-D. Jia, and J.-S. Pan, "Efficient network architecture search using hybrid optimizer," *Entropy*, vol. 24, no. 5, p. 656, 2022.

[32] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[33] T. Dang, T. T. Nguyen, C. F. Moreno-García, E. Elyan, and J. McCall, "Weighted ensemble of deep learning models based on comprehensive learning particle swarm optimization for medical image segmentation," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2021, pp. 744–751.

[34] L. Zhang and C. P. Lim, "Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models," *Appl. Soft Comput.*, vol. 92, Jul. 2020, Art. no. 106328.

[35] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6191–6201.

[36] H. Jiang, L. Yao, N. Lu, J. Qin, T. Liu, Y. Liu, and C. Zhou, "Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery," *Earth Syst. Sci. Data*, vol. 13, no. 11, pp. 5389–5401, Nov. 2021.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[38] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7169–7178.

[39] Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," 2019, *arXiv:1909.11065*.

[40] 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. (ECCV)*, 2018, pp. 801–818.

[41] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[42] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[43] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 205–218.

XIAOQING ZHANG received the Ph.D. degree from the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China. Her research interests include artificial intelligence, computer vision, and remote-sensing images.

QINGQING QI was born in Shandong, China, in 1998. She is currently pursuing the M.S. degree with Shandong University of Science and Technology, Qingdao, China. Her research interests include image processing and deep learning.

WEIKE LIU received the Ph.D. degree in geodesy and surveying engineering from Shandong University of Science and Technology, Qingdao, China, in 2013. His research interests include data mining, machine learning, and computer vision.