Statistical Texture Learning Method for Monitoring Abandoned Suburban Cropland Based on High-Resolution Remote Sensing and Deep Learning

Qianhui Shen, Haojun Deng, Xinjian Wen, Zhanpeng Chen, and Hongfei Xu

***Abstract*—Cropland abandonment is crucial in agricultural management and has a profound impact on crop yield and food security. In recent years, many cropland abandonment identification methods based on remote sensing observation data have been proposed, but most of these methods are based on coarse-resolution images and use traditional machine learning methods for simple identification. To this end, we perform abandonment recognition on high-resolution remote sensing images. According to the texture features of the abandoned land, we combine the method of statistical texture learning and propose a new deep learning framework called pyramid scene parsing network-statistical texture learning (PSPNet-STL). The model integrates high-level semantic feature extraction and deep mining of low-level texture features to identify cropland abandonment. First, we labeled the abandoned cropland area and built the high-resolution abandoned cropland (HRAC) dataset, a high-resolution cropland abandonment dataset. Second, we improved PSPNet by fusing statistical texture learning modules to learn multiple texture information on low-level feature maps and combined high-level semantic features for cropland abandonment recognition. Experiments are performed on the HRAC dataset. Compared with other methods, the proposed model has the best performance on this dataset, both in terms of accuracy and visualization, proving that deep mining of low-level statistical texture features is beneficial for crop abandonment recognition.**

***Index Terms*—Cropland abandonment, deep learning (DL), remote sensing, statistical learning, very high resolution (VHR).**

Manuscript received 1 December 2022; revised 26 January 2023 and 24 February 2023; accepted 27 February 2023. Date of publication 10 March 2023; date of current version 31 March 2023. This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3903402, in part by the National Natural Science Foundation of China under Grant 42222106, and in part by the National Natural Science Foundation of China under Grant 61976234. *(Corresponding author: Xinjian Wen.)*

Qianhui Shen and Haojun Deng are with the Guangdong Provincial Key Laboratory for Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: sqhsysu98@163.com; denghj5@mail2.sysu.edu.cn).

Xinjian Wen, Zhanpeng Chen, and Hongfei Xu are with the Surveying and Mapping Institute Lands and Resource Department of Guangdong Province, Guangzhou 510663, China, also with the Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources, Guangzhou 510663, China, and also with the Guangdong Science and Technology Collaborative Innovation Center for Natural Resources, Guangzhou 510663, China (e-mail: 100401019@qq.com; 543283684@qq.com; 1161444387@qq.com).

Digital Object Identifier 10.1109/JSTARS.2023.3255541

I. Introduction

CROPLAND change is one of the major types of land use and land cover (LULC) change, which has an important impact on food production. Global food demand is estimated to grow by 100% over the next half century [1]. Further growth in agricultural production is critical to global political and social stability and equity. The development of agricultural production is highly dependent on cropland, which provides the foundation for the survival of human [2]. The total agricultural area has stabilized in some parts of the world in recent decades [3], but there is still an invasion of cropland, which hinders the growth in agricultural production. Under the rapid expansion of urbanization, growing problems, such as cropland abandonment, have been prominent. Suburban cropland abandonment, as a form of land marginalization, has led to a decrease in the utilization rate of cropland, which seriously affects food security and the adequate supply of agricultural products. With the development of urbanization, a large number of farmers have entered the city, resulting in the abandonment of large areas of suburban cropland.

Previous recognition of abandoned cropland relies on manual in situ investigation, which is usually applied to a very small region motivated by the local government. However, in situ investigation is such a labor-intensive and time-consuming task, causing delayed knowledge of the cropping status of agricultural land. Remote sensing observation technology has developed rapidly in recent years and has become one of the main relies on agricultural monitoring and intelligent perception. Some studies have monitored abandoned cropland using existing publicly available datasets from satellite observation programs [4], such as moderate resolution imaging spectroradiometer from Earth observation system and Landsat program. Yin et al. [5] used the entire Landsat time series to map the extent and timing of abandoned cultivated land and simplified per-pixel classification by generating multiyear training data that could be used for annual classification. With annual cropland maps, the abandoned cropland trajectories were recognized. In summary, there are two types of research methods for the existing abandonment identification. One is the extraction of abandoned cropland based on the time trajectory. In this way, it is essentially the classification of the LULC types in multiple years. Based on the classification

results, certain rules are designed to discriminate the cropland that has changed to other types as abandoned cropland [6]. In addition to hard classification, some studies have improved the time trajectory-based methods into soft classification, that is to perform change detection algorithms, such as LandTrendr [7] and CCDC [8], using the time curves of classification probability of cropland component [9]. The other kind of method is to perform machine learning (ML) classification using single-year vegetation indices, such as normalized difference vegetation index, under certain phenological constraints or climate zone constraints [3], [5], [10]. However, the effect of phenological constraints is also limited, and a single or a small set of vegetation indices is not capable of presenting the characteristics of cropland abandonment and learning the semantic feature. Both methods are mostly based on images of low or moderate resolution. Therefore, both methods are not sufficient to extract precise textual features from the abandoned cropland. In addition, there is a lack of abandonment labels, which hinders the further research on the problem of cropland abandonment identification.

ML has been a hot topic in remote sensing interpretation for a long time. Classical ML models are widely used and demonstrated on the google Earth engine cloud computing platform with their large-scale computing ability [11], [12], [13], [14]. However, there are still some shortcomings in ML technology, such as ignoring the spatial association of neighboring pixels, which will cause the loss of image texture information, while the spatial information is crucial in remote sensing interpretation. To better capture the spatial texture information, some statistical texture learning (STL) methods are used to improve the performance of ML algorithms. Iqbal et al. [15] extracted gray-level co-occurrence matrix (GLCM) based features to classify different types of crops. As deep convolution neural network (DCNN) continues to advance in computer vision tasks, it is also used in Earth system scientific problems [16]. Compared with the traditional ML methods, deep learning (DL) can automatically learn and extract more advanced deep-level features in images and has the advantages of high adaptability, stable background models, high robustness of extracted features, and real-time detection. Using convolution operations, DCNNs can acquire receptive fields in spatial dimension to capture information about pixels and the connection with their neighbors, which is useful for learning spatial texture information of images.

The interpretation of remote sensing images is still a pixel-based tasks, which is very suitable for semantic segmentation algorithms. With the development of satellite observation technology, very high resolution (VHR) images tend to be more abundant and accessible. In interpreting complicated scene from VHR images, the spatial texture information plays a more important role. Existing research articles have applied semantic segmentation to LULC mapping of remote sensing images. McGlinchy et al. [17] applied U-Net on impervious surfaces mapping to extract complex features at the pixel level from high-resolution satellite imagery. Zhang et al. [18] used pyramid scene parsing network (PSPNet) fused with shallow edge information to identify farmland on high-resolution remote sensing images and extracted farmland information with higher precision than the existing farmland products. Liu et al. [19] proposed a CNN-transformer architecture with multiscale context aggregation to identify the nonagricultural areas in high-resolution images. The semantic segmentation model is successful using an encoder–decoder structure, where the input image is encoded to extract low-level features, such as color, texture, and edges, and then decoded to process the low-level feature information to obtain high-level features rich in semantic information. But the ability of encoders to extract low-level texture feature is limited. Several studies have explored how to better learn low-dimensional information. Yu et al. [20] proposed the dilated residual network to inject holes into the standard convolution to expand the reception filed, consequently improving the capture of spatial texture information. Wang et al. [21] used the idea of nonlocal to extract spatial texture information and weight it in a relatively large search range. Most of these methods focus on expanding the receptive field and learning image spatial autocorrelation implicitly. However, it is difficult to take into account both high-level semantic information and low-level texture features in this way. At present, these methods have not solved the problem of cropland abandonment identification.

In this work, we construct an abandoned cropland dataset. A DL-based encoder–decoder architecture for abandoned cropland extraction is then proposed, which combines low-dimensional STL methods to dig out the high-dimensional and low-level features of abandoned farmland at multiple levels. The proposed network realizes the recognition of abandoned farmland on high-resolution remote sensing images and proves the effectiveness of combining low-dimensional texture statistical information in the monitoring task of abandoned cropland.

II. Related Work

A. Semantic Segmentation

Semantic segmentation, image classification, and object detection are the three major tasks of convolutional neural networks. In 2014, fully convolutional networks (FCNs) [42] were proposed, which expanded the original CNN structure and trained and predicted without a fully connected layer. This method has achieved good results in the field of semantic segmentation of images and can generate images of any size, laying a foundation for the development of semantic segmentation networks. The U-Net proposed in 2015 uses an encoder–decoder structure to restore image details and spatial dimensions through deconvolution operations [22], that is, a U-shaped structure that is downsampled by convolution and upsampled by deconvolution. Since then, most of the networks that have emerged in the field of semantic segmentation are based on the encoder–decoder structure and use Atrous convolution and conditional random field postprocessing techniques to improve performance. For example, the Deeplab series of networks proposed in 2016 uses Atrous convolution [23] and fully connected conditional random field, PSPNet [24], etc., use a pyramid-shaped hole pooling module in the spatial dimension to synthesize background information and improve the receptive field of convolution. In recent years, attention mechanism and transformer-based architecture also make great progress on semantic segmentation. Attention modules extract spatial or contextual importance to enhance the representation ability of features, such as convolutional

block attention module [25] and dual attention module [26]. Transformers are successfully transferred to visual tasks by effectively encoding image patches, such as the milestone work vision transformer [27] and Swin transformer [28]. Although the semantic segmentation algorithm is booming in the field of computer vision, its application in the field of LULC mapping is still worth studying.

B. DL in LULC Mapping

With computer vision and DL methods continuing to develop, their applications in LULC mapping with remote sensing images are extensive. Over the past several years, DCNNs play a considerable role and promote the development of visual intelligent understanding in the tasks of image processing and interpretation. Successful examples include AlexNet [29], GoogleNet [30], VGG [31], and ResNet [32].

Semantic segmentation belongs to pixel-level classification, that is, each pixel in an image is assigned a category. Since the characteristics of the semantic segmentation task are consistent with the traditional LULC classification task in the field of remote sensing, the application of semantic segmentation network in the field of remote sensing is more extensive than the other traditional networks, and it has achieved relatively good results in road extraction and building extraction in remote sensing [33], [34], [35], [36], [37], [38], [39], [40], [41], [42]. Other than single-type extraction, there are also researches on all types of LULC mapping in a large scale using DL algorithms [43], [44], [45]. In summary, remote sensing LULC mapping is essentially a pixel-based computer vision task, so the semantic segmentation method is very suitable for LULC mapping and has made great progress.

III. Materials and Method

In this section, we will introduce in detail the framework of the proposed semantic segmentation and the dataset we prepared for the experiments.

A. High-Resolution Abandoned Cropland (HRAC) Dataset

The remote sensing dataset is named HRAC dataset, containing a series of GF-2 VHR images obtained in 2020 with a spatial resolution of 2 m. In this dataset, we labeled the abandoned cropland area through human visual interpretation. An example of image scenes in the dataset is shown in Fig. 1. Abandoned cropland can be positioned in hills and depressions and cropland patches surrounded by grass. Compared with the normally cultivated cropland with a clear and regular texture, abandoned cropland is overgrown with weeds and shrubs and has a messy texture, which is easy to be confused with nearby shrubs.

In this study, considering the spatial isolation of the training area and the test area, we randomly divided every scene of GF-2 imagery obtained in different regions into two parts and then crop each image scene into patches. We crop every complete scene of the GF-2 image into several image patches with a size of 512 $\times$ 512, dividing them into a training set and a test set. The training set contains 6376 image patches, while the test set contains 1625 patches. And some examples of detailed abandoned cropland are shown in Fig. 2.

Fig. 1. Sample scenes of the HRAC dataset (satellite scenes).

Fig. 2. Details of abandoned cropland in the HRAC dataset.

B. PSPNet-STL

1) Overview: The PSP-STL cultivated land abandonment extraction network proposed in this study combines the STL module, where the STL is composed of texture enhance module (TEM) and pyramid texture feature extraction module (PTFEM), as shown in Fig. 3. The framework we proposed uses low-level features for STL and then combines the high-level features extracted by the backbone network to achieve end-to-end learning. Multiple layers of intermediate features are generated by the backbone ResNet50 with the input of the original image. To encode and obtain the high-level semantic features, the output of the last layer of the backbone is input into the PPM as high-level information for pyramid scene parsing. Meanwhile, the outputs of the first and second layers in ResNet50 are considered as low-level features. Then, the low-level features are fused as the input of the TEM module to enhance texture features from low-level features with large spatial sizes and

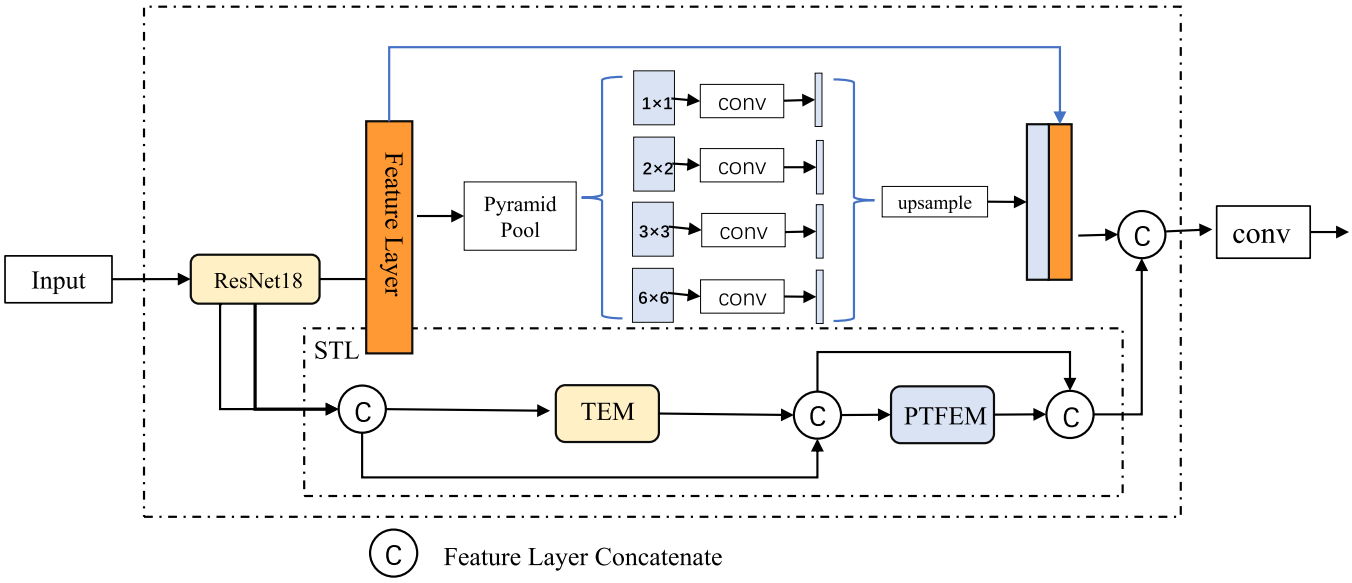


Fig. 3. Framework of proposed PSPNet-STL.

less loss of texture information. In this branch, we adopted the strategy of residual connection. The low-level features are again fused with the output of the TEM module as the input of the PTFEM. The features that flow through PTFEM module are concatenated with the output of PTFEM in a residual way. Finally, the high-level features generated by PPM are combined with the statistical texture features learned by the STL module to obtain the probability map of abandoned cropland. According to the proposed framework, high-level semantic information and low-level texture can be considered simultaneously, and thus improve the robustness of the model.

2) Pyramid Scene Parsing Network: PSPNet was proposed in 2017, aggregating context information based on different locations and scales with a good ability to extract global context information [24]. The network has good segmentation performance in multiple semantic segmentation datasets. The core of PSPNet is the spatial pyramid pooling (SPP) module. In SPP, different scale pooling operations are performed on the feature map, which expands the receptive field and can extract the information of different scale regions on the feature image. The features are concatenated and convoluted to fuse the global context information of the image. Compared with ordinary global pooling, pyramid pooling is more capable of extracting deep global information at different scales. In a deep neural network, the size of the receptive field determines the range of information that the network can use. Some scholars have shown that the actual receptive field of CNN is much smaller than the theoretical receptive field [46], especially in the deep features of the network. This makes many deep neural networks unable to fully integrate global scene information. Therefore, PPM contains multilayer parallel pooling and convolution operations of different scales to extract the global information of each scale, which fully utilize the global scene information.

We followed the original structure of PSPNet and used ResNet50 as the backbone. The feature map is input into the SPP module for pyramid pooling operation after being extracted by the backbone. There are four scales in the pyramid pooling operation in PPM, and the scales of the pooling are set to 1×1, 2×2, 3×3, and 6×6, respectively. Adaptive global average pooling is performed for each region separately and then activated through a 1×1 convolution kernel, batch normalization layer, and ReLU function.

3) STL Module: The texture feature is a kind of visual feature that reflects homogeneous information in an image. It is represented by the gray-scale value distribution of pixels and their spatial neighborhood pixels and performs statistical calculations on local or global information in the image.

STL module introduces statistical texture information into semantic segmentation to fully use texture information and achieve plug-in-and-play functionality. Most of the existing semantic segmentation research utilizes the CNN structure to extract contextual information in the deep semantic features. In the traditional CNN structure, the convolution kernel is good at extracting local texture features and local shape features, such as boundary shape, smoothness, and roughness. As a feature extraction operator, the convolution kernel is very suitable for image data mode. Although convolution operators have achieved great success, only extracting information and features with convolution is not enough. Another important information about texture features is statistical information, such as gray-scale histograms. Statistical texture information, as low-level information, is widely used in many traditional image process algorithms and has been proven to be important. Low-level features play a crucial role in improving semantic segmentation performance. Statistical texture features, such as frequency histograms, are difficult to extract by ordinary convolution operators.

To effectively describe the statistical texture in deep neural networks, a new feature encoding method is designed in STL: quantization and counting operator (QCO) [47]. QCO is similar to the convolution kernel in CNN, which is divided into three

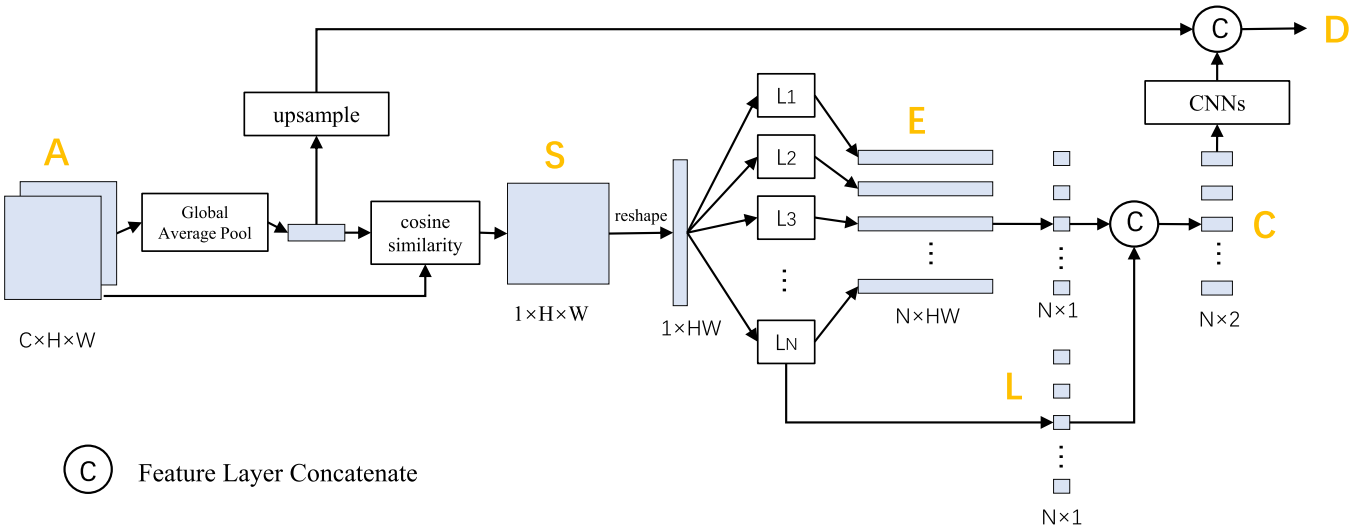


Fig. 4. Structure of 1D-QCO.

parts: quantization, counting, and average feature encoding. The QCO is divided into one-dimensional (1D-QCO) and 2D-QCO. The structure of 1D-QCO is shown in Fig. 4. The implementation is as follows. The first step is quantification. The size of the input feature map A is $C \times H \times W$, and the global average pooling of A is performed, and then the cosine similarity is calculated between g and the feature vector at each spatial position of A, and the feature map of $1 \times H \times W$ is obtained. The feature map is then converted to a 1-D vector S of size $(H \times W)$. S is quantized to obtain N-layer feature information L, and the number N can be set by users, and it is set to 128 in this article. L is then quantized to obtain an encoding matrix E.

The second feature of 1D-QCO includes E and D, where E and D represent the quantized coding map and statistical features, respectively. By summing and normalizing E into $N \times 1$ size, and then concatenating with L, after passing through a layer of the neural network, it is concatenated with variable D. The output of 1D-QCO reflects the distribution of features at various spatial locations.

4) TEM and PTFEM: In the STL module, the 1D-QCO and 2D-QCO operators are combined to construct a TEM and a PTFEM. How to effectively extract and utilize low-level features plays a crucial role in improving the performance of semantic segmentation. Simple multilevel feature addition or concatenation operations may lead to problems, such as feature dislocation, reducing the effectiveness of low-level features. The low-level features extracted from the backbone network are often of low quality, especially in the case of low contrast, the texture details are more blurred, and the extraction and utilization of low-level information are difficult to obtain ideal results. Therefore, as shown in Fig. 5 , TEM is specially designed to enhance the texture details of low-level features so that it is easier to capture texture-related feature information in later steps. The texture enhancement method is inspired by histogram quantization, a classic image quality enhancement method, where the horizontal and vertical axes of the histogram represent each gray level and its count value, respectively. TEM

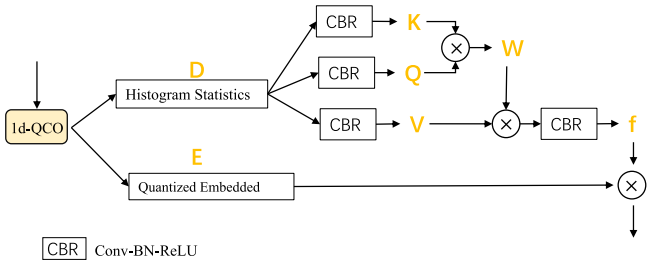


Fig. 5. Structure of TEM.

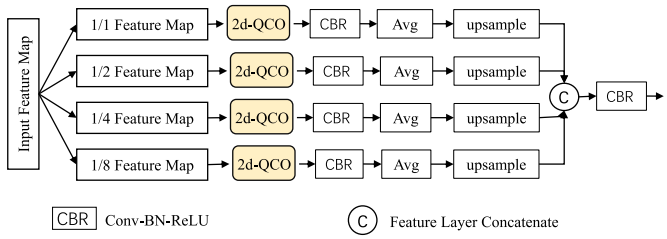


Fig. 6. Structure of PTFEM.

generates histograms of different levels and the corresponding quantized coding matrix through the 1D-QCO operator and calculates the weight of each corresponding quantized coding matrix according to the histogram through an operation similar to the attention mechanism. A weighting operation is performed on the quantized coding matrix by using the weight.

As shown in Fig. 6, the PTFEM aims to mine texture-related information from a multiscale feature map that contains rich texture details. Since the texture features are highly correlated with the statistical information of the spatial relationship between pixels, the method for extracting texture information in PTFEM draws on the GLCM. In GLCM, a co-occurrence matrix is first generated, and then the texture information of the region is represented by artificially setting statistical descriptors, such

as contrast and uniformity. A principle similar to GLCM is implemented in 2D-QCO to extract co-occurrence statistical features. Unlike the hand-designed statistical descriptors used in GLCM, 2D-QCO automatically learns effective statistical representations from samples through DL. Then, a multilayer perceptron is used to further extract texture features. In addition, a pyramid structure is also used in PTFEM to capture texture features at multiple scales to improve the performance and robustness of semantic segmentation. Four feature maps are obtained after parallel branches of four different scales, and they are upsampled after concatenating them together.

C. Loss Function

The logistics loss function is a commonly used loss function for neural networks. The cross entropy is calculated through the logarithmic function, reflecting the difference between the predicted probability distribution and the real probability distribution. The calculation formula is given as follows:

$$\text{Cross Entropy loss} \ = \ -y\log\hat{y} + (1-y)\log(1-\hat{y}) \quad (1)$$

where y represents the ground truth label and $\hat{y}$ represents the predicted label.

In addition to the commonly used logistic loss function, the loss function used in this study also combines the Dice loss function. In cropland abandonment monitoring, the target area where abandonment occurs often occupies a small area, which means that there are few positive samples, and the positive and negative samples are extremely unbalanced. In the past, sample weighting and other methods are used to balance. Despite this, the network is still easy to fall into the local minimum of the loss function during the learning process. This makes the prediction results of the trained network often have a strong tendency to predict the background, which will seriously affect the prediction effect and accuracy.

The Dice loss function [48] can automatically balance the relationship between background and foreground pixels. During the training process, more emphasis is placed on mining the foreground area, and it can still be better when there are few positive samples and the imbalance between positive and negative samples. The principle of Dice loss is equivalent to calculating IoU, which is the intersection of the real area and the predicted area divided by the union. The calculation formula is given as follows:

$$\text{Dice loss} \ = \ 1 - \frac{2 * \text{sum}(y \cdot \hat{y}) + \varepsilon}{\text{sum}(y) + \text{sum}(\hat{y}) + \varepsilon} \quad (2)$$

where y is the matrix of the ground truth label, represented by the value of 0 and 1, $\hat{y}$ is the matrix of the predicted label, and ε is an extremely small value set to prevent the denominator from being 0, which generally set to 10^{-6}.

Dice loss is prone to unstable training, and logistics loss has the function of guiding Dice loss. Therefore, in our experiment, a weighted loss function was introduced to achieve a better model training process. Dice loss and logistics loss were combined by, respectively, weight coefficient as follows:

$$L \ = w_D \ * \ L_D + w_L * \ L_L \quad (3)$$

where L refers to the total loss, w_D and w_L refer to the weight coefficients of the Dice loss and the logistic loss, L_D refers to the Dice loss, and L_L refers to the logistic loss.

After many experiments, the best weight ratio (w_D, w_L) is set to (0.6, 0.4).

IV. Experiments

A. Implementation Details

According to the abandoned labeling data, 512×512 samples were cut out from the area with abandoned labeling, and a small number of samples without abandoned labeling area were added. A total of 8000 samples were randomly divided into training set of 6376 samples and test set of 1625 samples. By training on the training set, we trained the model with learning rate decay strategy to automatically learn the best model and automatically stop when the loss comes to convergence. In the training process of this experiment, the Adam optimizer is used to optimize the parameters of the DL model. The initial learning rate is set to 0.0001, and the batch size is set to 16. Additional label-based data augmentation and mirror augmentation are also used. Our models and experiments are based on the open-source DL framework PyTorch. The experimental environment is Centos 7.5.1804. The GPU is GeForce RTX 2080ti. The CPU is Intel(R) Xeon(R) CPU E5-2680

B. Experiments and Analysis

1) Comparisons With Other Models: In order to show the effectiveness of the model proposed in this article, we select the following three semantic segmentation networks with better performance for comparison, which are briefly introduced as follows.

DeepLab v3+: The DeepLab v3+ semantic segmentation model incorporates atrous spatial pyramid pooling (ASPP) and encode–decoder (ED) structure [23]. Among them, the SPP module improves the multiscale global information utilization ability, and the encoder–decoder structure is connected with the lower level information during decoding, which can help restore image edge information. DeepLabv3+ uses Atrous convolution and depthwise separable convolution, which improves the receptive field of the convolution kernel and deepens the network depth while reducing the amount of computation. The model achieved the highest accuracy of semantic segmentation models at the time.

BiseNet: Bilateral Segmentation Network (BiSeNet) is a semantic segmentation model composed of a spatial branch and a context branch [49]. The model is designed to use feature fusion module (FFM) and refinement module (RM). The RM module in the context branch extracts features and combines them into the spatial branch and inputs the FFM module to output the final result. The spatial branch is mainly responsible for obtaining detailed spatial information, while the context branch extracts global contextual information through a lightweight module. Through the FFM module, the features extracted from the two branches are fused. Through the division between the above two

branches, not only the accuracy is improved but also the model operation speed is accelerated.

CCNet: Criss-cross network (CCNet) [50] uses the criss-cross attention module, which captures the global context information to solve the problems of traditional FCN, such as its fixed geometric structure, limited local receptive field with short-distance information, and the limitation of insufficient context information. CCNet is similar to a graph neural network, which regards each pixel in the convolutional feature map as a node and uses the relationship between nodes to extract high-level features.

2) Ablation Study: To better show the influence of the STL module and its components, we conducted the ablation studies on our HRAC dataset and quantified the results. First, we conducted the baseline experiment without any modules, which is the PSPNet with ResNet50. Then, we added the TEM branch on this baseline with low-level features generated by backbone. Similarly, we added the PTFEM branch to extract information from low-level features without TEM. Finally, the baseline experiment with the complete STL module was conducted.

3) Evaluation Metrics: In this study, we use overall accuracy (OA), intersection over union (IoU), recall ratio, precision, and $F1$ score to evaluate the result. In the binary classification problem, True Positive (TP) refers to the positive samples in the label, which are also correctly predicted as positive samples. True Negative (TN) refers to a negative sample in the label, and it is correctly predicted as a negative sample. False Positive (FP) refers to a negative sample in the label, but it is incorrectly predicted as a positive sample. False Negative (FN) refers to a positive sample in the label, but it is incorrectly predicted as a negative sample. The evaluation metrics are all calculated based on the above four indicators.

In the task of semantic segmentation, IoU is a commonly used evaluation metric. In this study, IoU is used as the area of IoU, the intersection of the real target area and the predicted target area divided by the area of the union. This indicator can reflect the degree of overlap between the real target area and the predicted area, and reflects the prediction ability of the model at the pixel level. OA reflects the overall prediction accuracy. Recall rate reflects the proportion of positive samples identified in the real positive samples. Precision rate refers to the proportion of samples with predicted value of positive and original value of positive in all samples with predicted value of positive. $F1$ score is calculated based on Recall and Precision. The F1 score weighs Recall and Precision to comprehensively reflect the overall performance, avoiding the bias caused by OA due to sample imbalance. All the metrics mentioned above are calculated as follows:

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

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

TABLE I
EXPERIMENTAL RESULTS WITH OTHER NETWORKS ON THE HRAC DATASET

Method	OA(%)	IoU(%)	Rec(%)	Pre(%)	F1(%)
CCNet	97.38	77.75	93.77	81.98	87.48
BiseNet	97.70	80.21	91.39	86.77	89.02
Deeplab v3+	98.31	83.37	**95.17**	88.85	91.90
Ours	**98.56**	**85.33**	94.55	**90.75**	**92.61**

The bold entitles indicate the best results for each indicator.

$$F1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{8}$$

C. Result Analysis

As can be seen from Table I, the proposed method PSPNet-STL outperforms the baseline models on the HRAC dataset, achieving the highest OA, IoU, Precision, and $F1$ score of 98.56%, 85.33%, 90.75%, and 92.61%. The second-ranked model Deeplab v3+ gets the OA of 98.31%, IoU of 83.37%, Precision of 88.85%, $F1$ score of 91.90%, and the best Recall of 95.17%, which is 0.62% higher than our proposed model of 94.55%. Another baseline model BiSeNet gets the OA of 97.70%, IoU of 80.21%, Recall of 91.39%, Precision of 86.77%, and $F1$ score of 89.02%. CCNet gets the worst OA, IoU, Precision, and $F1$ score of 97.38%, 77.75%, 81.98%, and 87.48%. From the results, we can see that, in the morphologically complex task of cropland abandonment recognition, a complicated network, such as Deeplab v3+, using an encoder–decoder structure, combined with various semantic segmentation tricks, such as ASPP and depthwise separable convolution, performs better than networks, such as BiSeNet and CCNet, that implicitly capture spatial and contextual information. However, we found that in this task, compared with simply extracting low-level features and using them directly, the explicit use of STL can make better use of the rich texture, structure, shape, and other information contained in the low-level features. It is also clear that the proposed model with STL module works better on the task of abandoned cropland extraction.

Fig. 7 further demonstrates the performance of different models on the HRAC dataset. On the mask of a given farmland range, we can avoid the high noise caused by other types of features and only consider the abandoned and nonabandoned situations on the farmland. From the images and labels, as shown in Fig. 7, we found that there are obvious differences in shape and texture between abandoned and nonabandoned cropland. The normal cropland parcels have clear texture, consistent shape, and obvious planting status. Abandoned farmland often has chaotic textures and fuzzy shapes. It can be clearly seen from the figure (row 2 of Fig. 7) that BiSeNet and CCNet have obvious misclassification, identifying the cultivated land in the center of the image as abandoned cropland. From row 3 of Fig. 7, it is also clear that BiSeNet has misclassified the cropland in the center-bottom part of image as abandoned cropland. It is obvious that Deeplab v3+ and CCNet have omitted the parcel of abandoned cropland on the left, while the model we proposed has neither omitted this part nor misclassified the cropland. According to row 4 of Fig. 7, the proposed model has also shown advantage in recognizing abandoned cropland. CCNet can only capture a part

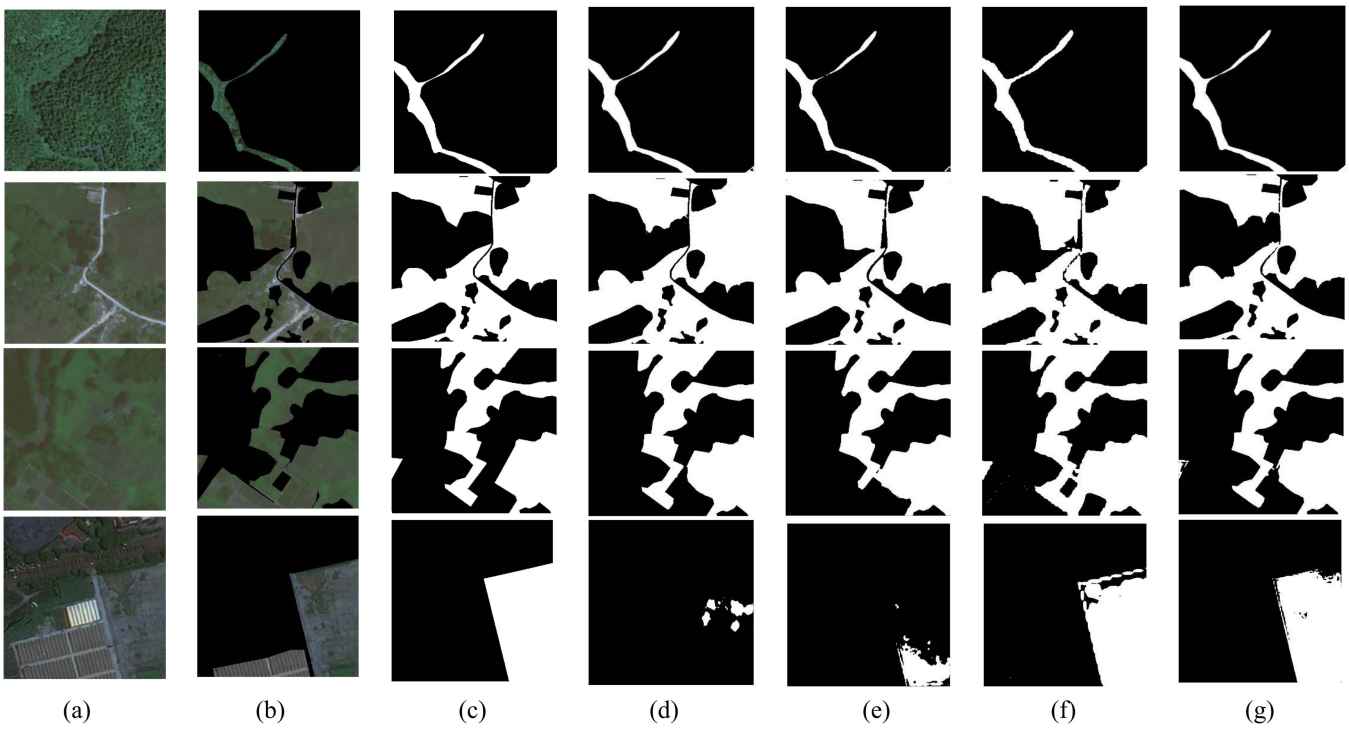


Fig. 7. Visualization results on the HRAC dataset. (a) Original images. (b) Images masked by cropland extent. (c) Ground truth label. (d)–(g) are the predicted labels of Deeplab v3+, CCNet, BiSeNet, and PSPNet-STL.

TABLE II
EXPERIMENTAL RESULT OF ABLATION STUDIES ON THE HRAC DATASET

Method	OA(%)	IoU(%)	Rec(%)	Pre(%)	F1(%)
w/o STL	97.95	81.46	91.94	86.96	89.38
w TEM	98.01	82.98	92.88	88.61	90.70
w PTFEM	98.03	83.22	92.51	89.22	90.84
proposed	**98.56**	**85.33**	**94.55**	**90.75**	**92.61**

The bold entitles indicate the best results for each indicator.

of abandonment information and partial boundaries. Compared with Deeplab v3+, which performs the worst in this image for omitting most of the abandoned cropland, BiSeNet demonstrates its advantage of its context path. However, BiSeNet still omits some target, causing strange holes presented in the image, while the proposed model has a smoother result. In general, our model performs the best among these models on the HRAC dataset. We believe that the low-level texture information is very important in the task of abandonment extraction, and our model can better process low-level information by first enhancing texture details and then extracting enhanced feature information, thus achieving better performance.

V. DISCUSSION

In the discussion part, we discuss the detailed model design of our proposed method and conduct some experiments to analyze the performance difference of the different parts of the model.

Numerical results of ablation study are shown in Table II. Compared with the baseline model without any STL module, the model with TEM or PTFEM performs better on all the evaluation metrics that we used. The two components of STL are added separately, and the performance of the model has its own advantages and disadvantages. The model with TEM only gets the better Recall of 92.88, while the model with PTFEM only gets the better OA, IoU, Precision, and $F1$ score of 98.03%, 83.22%, 89.22%, and 90.84%. It can be seen that the STL module, which is the combination of the TEM and the PTFEM, is significantly better than adding one of the components alone. It achieves a highest OA of 98.56%, IoU of 85.33%, Recall of 94.55%, Precision of 90.75, and $F1$ score of 92.61%. TEM is used to enhance the texture details of low-level features, while PTFEM extracts the information of feature maps containing rich texture details from multiple scales. Therefore, after using TEM to enhance the texture details, PTFEM is used to encode the TEM-enhanced features. The STL module uses these two groups of modules successively, which can better extract the information from low-dimensional features that is beneficial to identify the abandoned cropland.

VI. CONCLUSION

In this article, a new DL-based framework using STL called PSPNet-STL was proposed to effectively extract cropland abandonment from VHR imagery, aiming at the characteristics of focusing on local texture details in the extraction of cropland abandonment and the limitation of current methods relying on long time-series medium-resolution images. We also produced an HRAC dataset called the HRAC dataset based on single-temporal high-resolution GF-2 imagery, focusing on cropland abandonment within cropland extent. The feasibility

of abandoned cropland identification through single-temporal high-resolution remote sensing images is verified. The PSPNet-STL model uses the low-level feature map of the original input data to calculate the quantized texture statistical features and extracts high-level semantic features through pyramid pooling to simultaneously realize the maintenance of low-level texture features and the mining of deep texture statistics to achieve better capturing the morphological characteristics of abandoned farmland. Experimental results prove that the proposed PSPNet-STL model outperforms other models on the HRAC dataset. The proposed framework demonstrates the effectiveness of low-level texture statistics in cropland abandonment recognition. In the future, we will explore a more stable architecture to accurately identify abandoned cropland and improve recognition performance in more diverse scenarios.

REFERENCES

[1] D. Tilman, K. G. Cassman, P. A. Matson, R. Naylor, and S. Polasky, "Agricultural sustainability and intensive production practices," *Nature*, vol. 418, pp. 671–677, 2002.

[2] J. R. Porter et al., "Food security and food production systems," Cambridge Univ. Press, Cambridge, U.K., 2015.

[3] C. Alcantara, T. Kuemmerle, A. V. Prishchepov, and V. C. Radeloff, "Mapping abandoned agriculture with multi-temporal MODIS satellite data," *Remote Sens. Environ.*, vol. 124, pp. 334–347, 2012.

[4] F. Löw, E. Fliemann, I. Abdullaev, C. Conrad, and J. P. A. Lamers, "Mapping abandoned agricultural land in Kyzyl-Orda, Kazakhstan using satellite remote sensing," *Appl. Geogr.*, vol. 62, pp. 377–390, 2015.

[5] H. Yin, A. V. Prishchepov, T. Kuemmerle, B. Bleyhl, J. Buchner, and V. C. Radeloff, "Mapping agricultural land abandonment from spatial and temporal segmentation of Landsat time series," *Remote Sens. Environ.*, vol. 210, pp. 12–24, 2018.

[6] Z. Du, J. Yang, C. Ou, and T. Zhang, "Agricultural land abandonment and retirement mapping in the Northern China crop-pasture band using temporal consistency check and trajectory-based change detection approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4406712.

[7] R. E. Kennedy, Z. Yang, and W. B. Cohen, "Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr—Temporal segmentation algorithms," *Remote Sens. Environ.*, vol. 114, pp. 2897–2910, 2010.

[8] Z. Zhu and C. E. Woodcock, "Continuous change detection and classification of land cover using all available Landsat data," *Remote Sens. Environ.*, vol. 144, pp. 152–171, 2014.

[9] A. Dara et al., "Mapping the timing of cropland abandonment and recultivation in northern Kazakhstan using annual Landsat time series," *Remote Sens. Environ.*, vol. 213, pp. 49–60, 2018.

[10] X. Tong et al., "The forgotten land use class: Mapping of fallow fields across the Sahel using Sentinel-2," *Remote Sens. Environ.*, vol. 239, 2020, Art. no. 111598.

[11] B. Chen et al., "Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017," *Sci. Bull.*, vol. 64, pp. 370–373, 2019.

[12] A. R. Phalke et al., "Mapping croplands of Europe, Middle East, Russia, and Central Asia using Landsat, random forest, and Google Earth engine," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 104–122, 2020.

[13] P. S. Thenkabail et al., "Global cropland-extent product at 30-m resolution (GCEP30) derived from Landsat satellite time-series data for the year 2015 using multiple machine-learning algorithms on Google Earth Engine cloud," *US Geological Survey*, Reston, VA, USA, 2021.

[14] D. Wuyun et al., "Mapping fallow fields using Sentinel-1 and Sentinel-2 archives over farming-pastoral ecotone of Northern China with Google Earth Engine," *GIScience Remote Sens.*, vol. 59, pp. 333–353, 2022.

[15] N. Iqbal, R. Mumtaz, U. Shafi, and S. M. H. Zaidi, "Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms," *PeerJ Comput. Sci.*, vol. 7, 2021, Art. no. e536.

[16] J. Xu et al., "DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111946.

[17] J. McGlinchy, B. Johnson, B. Muller, M. Joseph, and J. Diaz, "Application of UNet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3915–3918.

[18] D. Zhang et al., "A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111912.

[19] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, May 2022.

[20] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 472–480.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

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

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

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

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

[26] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.

[27] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, 2017.

[30] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[31] K. Simonyan et al., "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

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

[33] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1743–1751.

[34] H. Liu et al., "DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, 2019, Art. no. 2380.

[35] A. Abdollahi, B. Pradhan, S. Gite, and A. Alamri, "Building footprint extraction from high resolution aerial images using generative adversarial network (GAN) architecture," *IEEE Access*, vol. 8, pp. 209517–209527, 2020.

[36] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, 2018, Art. no. 144.

[37] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, 2020, Art. no. 1050.

[38] Q. Hu, L. Zhen, Y. Mao, X. Zhou, and G. Zhou, "Automated building extraction using satellite remote sensing imagery," *Autom. Construction*, vol. 123, 2021, Art. no. 103509.

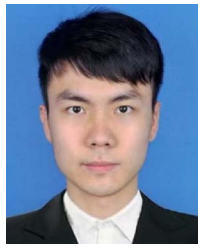
[39] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[40] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building

footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, 2021, Art. no. 112589.
[41] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, 2022.
[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 640–651, 2017.
[43] D. He, Q. Shi, X. Liu, Y. Zhong, G. Xia, and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using Landsat imagery," *GIScience Remote Sens.*, vol. 59, pp. 2036–2067, 2022.
[44] Y. Zhang, K. Liu, Y. Dong, K. Wu, and X. Hu, "Semisupervised classification based on SLIC segmentation for hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1440–1444, Aug. 2020.
[45] S.-H. Lee, K.-J. Han, K. Lee, K.-J. Lee, K.-Y. Oh, and M.-J. Lee, "Classification of landscape affected by deforestation using high-resolution remote sensing data and deep-learning techniques," *Remote Sens.*, vol. 12, 2020, Art. no. 3372.
[46] B. Zhou et al., "Object detectors emerge in deep scene CNNs," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7-9, 2015.
[47] L. Zhu, D. Ji, S. Zhu, W. Gan, W. Wu, and J. Yan, "Learning statistical texture for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12532–12541.
[48] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[49] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 334–349.
[50] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

Qianhui Shen received the B.S. degree in geographic information science from the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China, in 2021, where he is currently working toward the M.S. degree in cartography and geographic information system with the School of Geography and Planning.

His research interests include remote sensing image processing, deep learning, and agricultural application.



Haojun Deng received the B.S. degree in geographic information science and the M.S. degree in remote sensing and geographic information engineering from Sun Yat-sen University, Guangzhou, China, in 2020 and 2022, respectively.

His research interests include machine learning and deep learning in phenology and urbanization.

Xinjian Wen received the B.S. degree in surveying engineering from the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, in 2003.

He is currently working with Surveying and Mapping Institute Lands and Resource Department of Guangdong Province, Guangzhou, China. His research interests include interpretation of remote sensing images, surface classification, and geographic information system.

Zhanpeng Chen received the B.S. degree major in information engineering from Jinan University, Guangzhou, China, in 2011.

He is currently working with Surveying and Mapping Institute Lands and Resource Department of Guangdong Province, Guangzhou, China. His research interests include interpretation of remote sensing images and geographic information system.

Hongfei Xu received the B.S. degree in surveying engineering from the School of Surveying and Mapping, Wuhan University, Wuhan, China, in 2017.

He is currently working with the Surveying and Mapping Institute Lands and Resource Department of Guangdong Province, Guangzhou, China. His research interests include remote sensing image processing and applications in cultivated land protection.