Fine-Grained Abandoned Cropland Mapping in Southern China Using Pixel Attention Contrastive Learning

Haoyang Li, Haomei Lin, Junshen Luo, Teng Wang, Hao Chen, Qiuting Xu, and Xinchang Zhang

***Abstract*—Cropland abandonment has multifaceted and controversial impacts on the natural environment and socioeconomic development. Utilizing remote sensing data offers the potential for comprehensive coverage and large-scale insights into automated abandoned cropland identification. However, accurately capturing small abandoned cropland, particularly in regions, such as southern China, with fragmentized farmland, poses a significant challenge using the traditional optical image-based mapping methods due to their coarse spatial resolution. In addition, irregular and chaotic textures of abandoned cropland further complicate the accurate prediction using very high resolution (VHR) data. In this article, we propose a novel deep learning network termed pixel attention contrastive network (PACnet) to map fine-grained abandoned cropland based on VHR data. Cross-image pixel contrast learning is introduced to discern distinctive features distinguishing abandoned cropland from other land types across various inter-images. Moreover, a criss-cross attention module is embedded to enhance the contrasting characteristics within individual intra-images. Experimental outcomes validate the efficacy of PACnet, showcasing the highest accuracy (OA = 93.8% and mIOU = 71.7%) when compared with classical semantic segmentation networks. Our proposal not only underscores the potency of VHR remote sensing data in finely delineating abandoned cropland but also carries significant implications for cropland abandonment impact analysis and informed policy formulation.**

***Index Terms*—Abandoned cropland, contrastive learning, deep learning (DL), very high resolution (VHR).**

Manuscript received 12 July 2023; revised 31 August 2023 and 22 October 2023; accepted 24 November 2023. Date of publication 1 December 2023; date of current version 4 January 2024. 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 42071441. *(Corresponding author: Teng Wang.)*

Haoyang Li is 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: lihy256@mail2.sysu.edu.cn).

Haomei Lin and Junshen Luo are with the School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: linhm23@mail2.sysu.edu.cn; luojsh7@mail2.sysu.edu.cn).

Teng Wang, Hao Chen, and Qiuting Xu are with the Guangdong Department of Land and Resources, Institute of Surveying and Mapping, 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 Guangdong Natural Resources Science and Technology Collaborative Innovation Center, Guangzhou 510663, China (e-mail: 5765008@qq.com; a635503656@qq.com; 510612269@qq.com).

Xinchang Zhang is with the School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China, and also with the College of Environment and Planning, Henan University, Kaifeng 475004, China (e-mail: eeszxc@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3338454

I. Introduction

ABANDONED cropland represents a form of land use characterized by marginalization arising from inadequate suitability and economic viability [1], [2], [3]. The abandonment of arable land has multifaceted and profound effects on factors, such as soil erosion, biodiversity, carbon storage, and the development of the agricultural economy [4], [5], [6], [7]. In China, the matter of abandoned cropland has garnered significant attention, primarily driven by concerns over food security, particularly in the economically developed regions of southern China [3], [8], [9]. Precise mapping of abandoned cropland is essential for analyzing the driving factors contributing to its occurrence and understanding its impact on the natural environment and socioeconomic aspects. In contrast to less efficient field-based research, remote sensing (RS) technology provides a more convenient and expeditious means of mapping large-scale land use types, and it can also be applied effectively for monitoring abandoned cropland [10].

Numerous studies have developed algorithms for mapping abandoned cropland, with the majority utilizing time series of optical imagery and depending on temporal feature analysis [8], [10], [11], [12], [13], [14], [15], [16], [17], [18]. Cultivated cropland exhibits rapid changes in RS observations due to activities, such as planting, harvesting, and swift phenological transitions. Abandoned cropland exhibits lower variability in optical time-series curves when compared with cultivated areas [10]. Consequently, these studies investigate the disparities in temporal fluctuations between abandoned cropland and various other land categories [13], [14]. Researchers create suitable spectral indices as indicators, employing machine learning classifiers (e.g., random forest [10], [11], [16]) and temporal change detection algorithms (e.g., LandTrendr [17], [18]) to map abandoned cropland. In addition to directly mapping abandoned cropland as described above, certain studies attempt an indirect approach. They map multiple land cover types over an extended time series and analyze land conversion patterns to identify abandoned cropland [8], [9]. In terms of spatial resolution, the studies mentioned above utilizing MODIS (250 m) [12], [13], [14], [15], Landsat (30 m) [11], [16], [17], [18], and Sentinel-2 (10 m) [10] operate at a relatively coarse scale.

Existing mapping methods employing coarse images (>= 10 m) frequently encounter challenges posed by mixed pixels at the edge of small parcels, rendering the accurate

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

depiction of small abandoned cropland particularly challenging, notably in southern China. Owing to physical geographical conditions and historical factors, cropland parcels in Southern China exhibit fragmentation and marginalization [19]. In contrast, very high resolution (VHR) images (<= 1 m) can more effectively identify small and fragmented parcels. Therefore, the utilization of VHR images is imperative for acquiring finely detailed maps of abandoned land. Constrained by data availability, VHR images with limited temporal series density pose challenges for conducting temporal and spectral analysis. Thus, the key to mapping abandoned cropland at the submeter level lies in the profound exploration of fine-grained information and visual features within a single-phase VHR image.

Recently, methods based on deep learning (DL) have demonstrated their efficacy in VHR image interpretation [20], [21]. In contrast to traditional texture modeling methods, DL networks exhibit a superior capacity to harness spatial-context information in VHR images, offering an enhanced depiction of surface details and intricate spatial information [22], [23]. The extraction of deeper texture features and semantic information through the layers of deep neural networks effectively captures the visual target information within VHR images [24]. Numerous DL networks have been proposed for the fine-grained extraction of diverse land cover types from VHR images [25], [26]. An increasing number of DL frameworks are under development to enhance the capability to perceive complex VHR semantic information for specific tasks, including change detection [27], building extraction [28], tree crown mapping [29], etc. DL methods can yield improved results by analyzing specific objectives and adjusting DL modules. Critical directions for DL network design and improvement include feature enhancement [30], [31] and feature fusion [22], [32]. In addition, incorporating graph structures constitutes a novel approach for mining topological information [33] and distilling contextual information [34]. The powerful learning capability of DL in the context of fine-grained high-resolution landscapes offers a viable avenue for mapping VHR abandoned cropland.

Abandoned cropland exhibits distinct visual discriminative features in VHR images, yet its fuzzy and amorphous characteristics present a significant challenge for DL networks. Unlike cultivated farmland, abandoned cropland in VHR images exhibits distinct fuzzy texture features. It lacks signs of artificial cultivation and typically appears as grassland and shrub characteristics. Contrasted with the orderly patterns found in neighboring farmland and orchards, abandoned cropland displays conspicuous chaotic and disordered texture characteristics, as illustrated in Fig. 1. This forms a clear visual foundation for identifying abandoned cropland. DL techniques can effectively train and predict by extracting the distinct visual textures of the target areas.

Nonetheless, the fuzzy and uncertain textures of abandoned cropland continue to present challenges for accurate identification by DL architectures. Indeed, DL algorithms excel in recognizing objects with well-defined boundaries and textures (e.g., buildings and roads) rather than natural amorphous regions characterized by fuzzy edges and intraclass variations

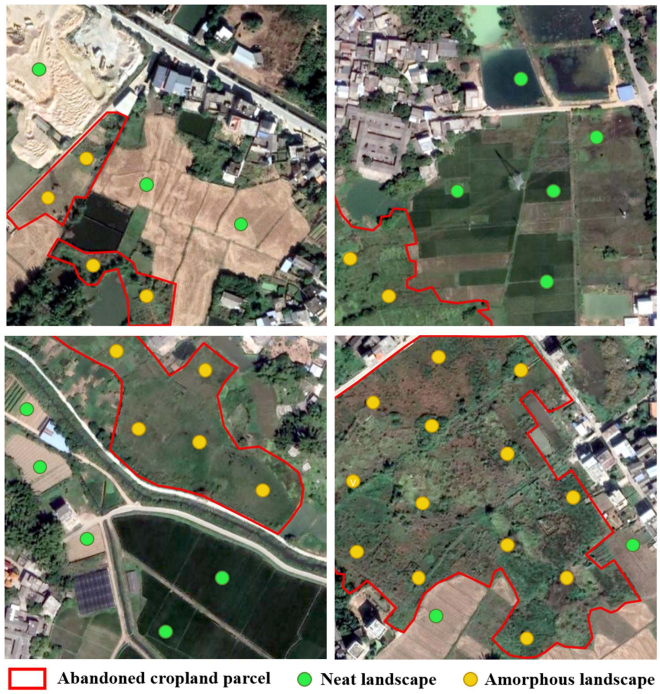

Fig. 1. Abandoned cropland in VHR images. Compared with the surrounding neat landscape, abandoned cropland presents amorphous and disorderly characteristics in vision.

(e.g., agricultural areas) [35], [36]. Abandoned cropland exhibits more pronounced amorphous characteristics than typical natural features, posing significant challenges for fine-grained mapping. Shen et al. [37] conducted experimental work to map abandoned cropland in VHR images, introducing a neural network with a texture calculation module. Although texture enhancement learning can bolster the context-awareness capability of DL, the capacity for perceiving textures in abandoned cropland remains inadequate. Consequently, tackling the challenge of perceiving the amorphous features of abandoned cropland remains a paramount concern.

Driven by the pronounced distinctions between abandoned cropland and well-maintained cropland, our approach centers on contrasting the heterogeneity between cultivated and abandoned areas rather than directly identifying the unclear features. To accentuate the distinctive features, we have developed a pixel attention contrastive network (PACnet) designed to capture the differentiated features of both inter and intraimages. In PACnet, we employ cross-image pixel contrastive learning (CPCL) to analyze distinctive features among interimages. Subsequently, a criss-cross attention module (CCAM) is applied to bolster the capacity to capture disparities among neighboring regions and highlight the contrast between abandoned cropland and its nearby surroundings.

In a nutshell, our contributions can be listed as follows.

1) We construct a VHR abandoned cropland dataset (VACD) exceeding 14 000 samples for DL network training and propose a DL network called PACnet designed explicitly for extracting fine-grained details of abandoned cropland from VHR (0.5 m) RS images.

2) Confronted with abandoned cropland's ambiguous and disordered visual attributes, we introduce CPCL and CCAM to characterize contrasting features between abandoned cropland and other land categories.
3) Our proposed approach yields competitive results on the VACD dataset, affirming the viability and substantial potential of VHR abandoned cropland mapping.

II. Related Work

A. Semantic Segmentation in RS

Semantic segmentation is the foundational task in the computer vision field, which refers to labeling all the pixels in images. The proposal of fully convolutional networks (FCNs) [38] marked a significant milestone in image segmentation. FCN, employing deconvolution in place of a fully connected layer, enables processing input images of any size and predicting every pixel within them. Recent research endeavors [39], [40], [41], [42], [43], [44], [45] in the domain of semantic segmentation fall into two primary categories. One aims to expand the receptive field and facilitate multiscale context extraction, while the other focuses on the incorporation of attention modules. Unet [39] exemplifies the former category, characterized by its distinctive asymmetric "U" shape structure. U-Net merges low- and high-level features via skip connections, retaining some edge characteristics.

Furthermore, the utilization of atrous convolution [40] constitutes another representative approach. By incorporating a fully connected conditional random field (CRF) and atrous spatial pyramid pooling (ASPP), DeepLab series [41] attains enhanced representation capability and more precise multiscale object segmentation. The introduction of attention mechanisms represents another crucial technique for improving DL performance. These mechanisms excel in extracting contextual importance by calculating correlations among instances. The convolutional block attention module [42] and dual attention module [43] serve as typical examples of such attention modules. In addition, leveraging attention mechanisms, the transformer-based architecture has introduced a new perspective to the computer vision field. Milestone works in this regard include the vision transformer [44] and the segmentation transformer [45], which have significantly advanced the utilization of transformers in semantic segmentation.

The rapid advancements in DL offer a novel perspective for RS image classification, akin to the principles underlying semantic segmentation in the computer vision field. In contrast to natural images, RS images captured by satellites and aircraft are susceptible to various factors, including lighting and photography angles [46]. Hence, incorporating spatially contextual information is essential in RS image segmentation. In recent years, extensive research efforts [18], [47], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60] have focused on enhancing pixel-level accuracy in semantic segmentation of RS images. These efforts can be categorized into three primary strategies: multiscale contextual information extraction, information fusion, and postprocessing techniques [24]. Zhao and Du [47] employed a multiscale convolutional neural network to extract deep spatial information from hyperspectral images. ScasNet was developed to capture multiscale contexts on the encoder output [48]. MC-FCN [47] applied additional constraints to intermediate layers, thereby enhancing its multiscale feature representations and improving building segmentation accuracy. MGSNet [58] extracted background information surrounding the target to improve sample distinguishability. Li et al. [59] proposed multiscale split attention to acquire more detailed representations through grouping. ACAHNet [55] utilizes the asymmetric multiheaded cross-attention module to enhance the contextual features extracted by both CNN and transformer network. Information fusion is another integral aspect of the DL network. Marmanis et al. [49] extracted spectral and digital elevation model information from two channels, and a convolution layer combines the results from both channels. Wang et al. [53] introduced a gated convolutional neural network for selecting adaptive features during the fusion of different layer features. AERNet [56] employed a contextual feature aggregation module to fuse information from different context features. SSPN [57] applied multiscale interfusion to enrich the extracted features and improve the sensitivity of the spectral–spatial information. Wang et al. [60] designed a global dependence fusion module to fuse features extracted from hyperspectral and SAR images. Postprocessing techniques, such as simple linear iterative clustering superpixel segmentation [50] and CRFs [51], are commonly applied to refine the RS image segmentation results. Additional postprocessing techniques, such as the integration of point clouds and high-resolution images, mitigate the salt-and-pepper noise in classification results [54].

B. Contrastive Learning

Categorized by label availability, contrastive learning can be grouped into unsupervised and supervised contrastive forms. In the realm of unsupervised contrastive learning (UCL), pioneering studies [61], [62] laid the foundation by introducing pretext tasks and defining positive/negative samples. Wu et al. [61] introduced instance discrimination as the pretext task for UCL. Ye et al. [62] defined positives as varying augmentation outcomes from a single image, considering other images and their augmentations in the dataset as negatives. Nearly all studies [63], [64] find that the size of the negative sample collection dramatically influences the performance of UCL. However, the challenge of designing sample collections that balance computational efficiency and UCL performance persists. Subsequently, milestone work—MoCo [63] and simCLR [64]—were proposed to solve the above problem. MoCo introduced a queue structure and momentum encoder to create a comprehensive and coherent sample collection. SimCLR, on the other hand, discarded conventional data containers in favor of memory banks and raised projection heads to outperform prior self-supervised methods significantly. Most UCL serves as a pretraining step for the downstream task, especially for the classification task. The performance of UCL pretraining on dense work, such as semantic segmentation and object detection, is unsatisfactory [65]. DenseCL [66] and VICRegL [67] were developed to address this issue. However, while they demonstrate

effectiveness on natural image datasets, their performance on RS image datasets warrants further improvement. Concerning supervised contrastive learning (SCL), it primarily incorporates the concept of CL to amplify representational capacity and regularize the embedding space. Zhao et al. [68] defined pixels belonging to the same class in other images as additional positive samples. The introduction of these more challenging positives directed the network to cluster pixels of the same class. The model was initially trained using pixelwise label-based contrastive loss and, subsequently, fine-tuned with pixelwise cross-entropy loss for semantic segmentation. Chaitanya et al. [69] employed the global contrastive loss to enhance image-level representative capacity and the local contrastive loss to distinguish adjacent regions. These studies leverage both global and local context at the image level while striving to extract distinctive pixel-level features.

III. Methodology

This section introduces the dataset prepared for experiments and details the proposed PACnet. The dataset prepared for experiments is presented in Section III-A. Then, we briefly introduce the PACnet framework in Section III-B. In Section III-C, CPCL is proposed for enhancing contrasts interimages, and in Section III-D, CCAM is embedded for exploring contrasts intraimages. Finally, we introduce the loss function of PACnet.

A. Dataset Preparation

The VACD is annotated on Google Earth VHR images (0.5 m) obtained in 2022 in Guangdong Province, China. We label the abandoned cropland through human visual interpretation. As shown in Fig. 1, abandoned cropland and cultivated farmland are similar in spectral as vegetation but quite different in texture information. The cultivated fields have neater and more regular textures, while the textures of abandoned cropland are significantly irregular and messy. Abandoned cropland filled with ruderal is often located in monticules and depressions. Since weeds often overgrow shrubs, the surrounding shrubs hinder the extraction of abandoned farmland.

We crop the complete scene of images into 512×512 patches and randomly divide them into a train collection of 10 608 patches, a validation collection of 2653 patches, and a test collection of 1474 patches. Some patches of the detailed VACD are shown in Fig. 2.

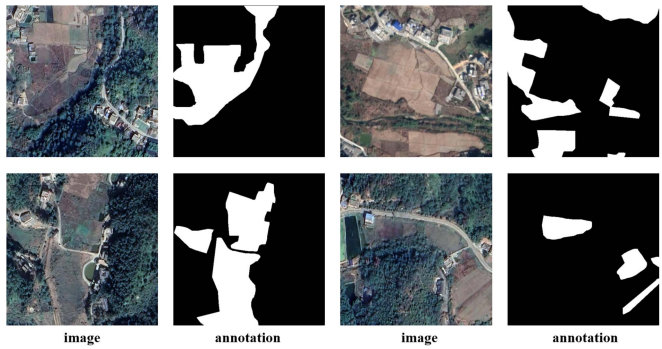

Fig. 2. Details of VACD dataset.

B. PACnet Architecture

According to the characteristics of abandoned cropland and the surrounding surface features, we propose PACnet as Fig. 3 to enhance the comparable features from interimages and intraimages. We introduce CPCL to focus on the overall and global feature contrast. We cast CPCL as a dictionary query task [52]. The target pixel for prediction is seen as a query, and the search range containing samples (positive and negative) is similar to a dictionary with keys. CPCL calculates the contrastive loss between the selected pixel(query) embeddings and other pixel and region embeddings(keys) we sample from the memory bank. As for CCAM, we attach importance to the local difference of the pixels in the current image from two essential directions. And the powerful feature representative capacity by applying CCAM can guide the CPCL network to have perfect performance on contextual information extraction in return.

The network architecture is represented in Fig. 3. For an input image P, it passes through the encoder ResNet50 and is mapped into dense features $\boldsymbol{P}$ with a spatial size of $H \times W \times D$. D denotes the number of dimensions here. Then, $\boldsymbol{P}$ is fed into two branches. The one is to apply a CCAM in capturing contextual importance from both lateral and longitudinal orientation to intensify the pixelwise representative ability. The output embeddings $\boldsymbol{P}' \in \mathbb{R}^{H\times W\times D}$ of CCAM are then projected into DeepLabV3 decoder, where $\boldsymbol{P}'$ is transformed into a score map $\boldsymbol{S} \in \mathbb{R}^{H\times W\times|\mathcal{C}|}$. $|\mathcal{C}|$ denotes the number of classes in the dataset. And finally, in this branch, we calculate the segmentation loss between $\boldsymbol{S}$ and the label of the input image.

The other branch, CPCL branch, is to pass $\boldsymbol{P}$ through a projection head, which is composed of two 1×1 convolution layers with ReLU. The projection head maps every high-level pixel feature $\boldsymbol{p} \in \boldsymbol{P}$ into a 256-dimension ℓ_2-normalized feature vector, making preparations for the calculation of contrastive loss. The projection head applied here is only complemented in the training process and is eliminated in the inference section. The contrastive loss is later computed between the query and keys selected from the memory bank. The memory bank contains the pixel and region embeddings, and the region embeddings are calculated by image projected feature $\boldsymbol{P}$ and corresponding labels.

C. Cross-Image Pixel Contrastive Learning (PCL)

This section provides a detailed introduction to CPCL in PACnet.

1) Pixel Contrastive Learning: Unlike classical image contrastive learning (ICL), CPCL is a kind of PCL, a supervision algorithm. The brief frameworks of ICL and PCL are shown in Fig. 4. ICL conducts CL by using different data augmentation of one image and finally implementing the features from the output of projections. In contrast, PCL performs contrastive feature mining at the pixel level and mines fine-grained features. For RS images, the details in a scene are often too complex to clarify

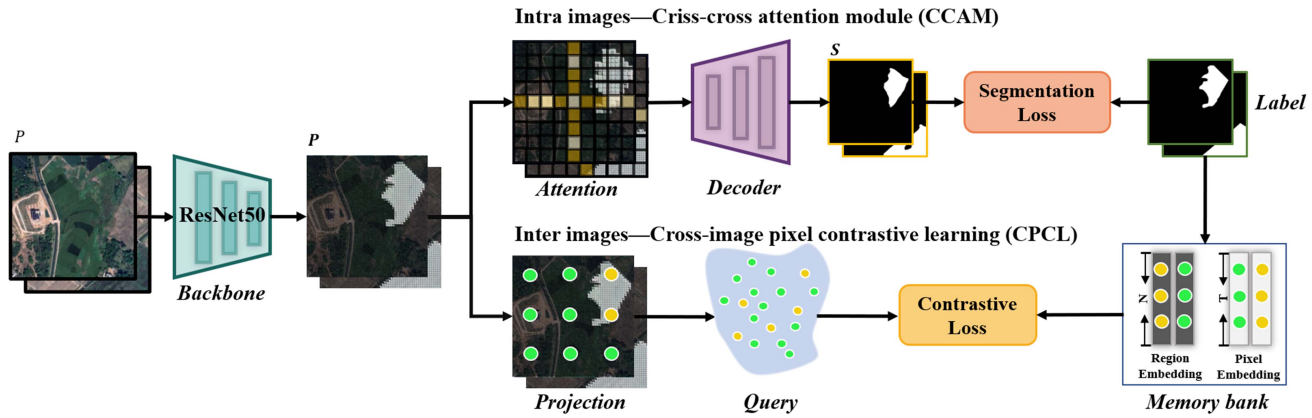

Fig. 3. Framework of the proposed PACnet.

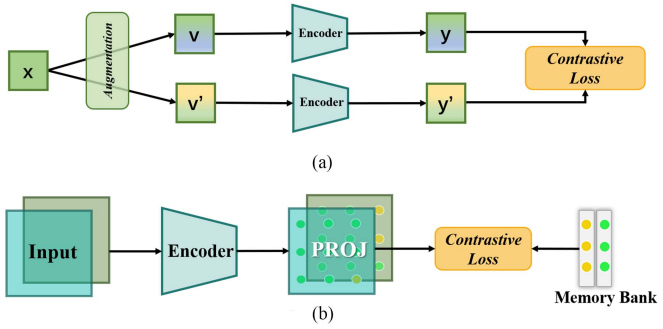

Fig. 4. Comparison of the structures of ICL and PCL. (a) Image contrastive learning. (b) Pixel contrastive learning.

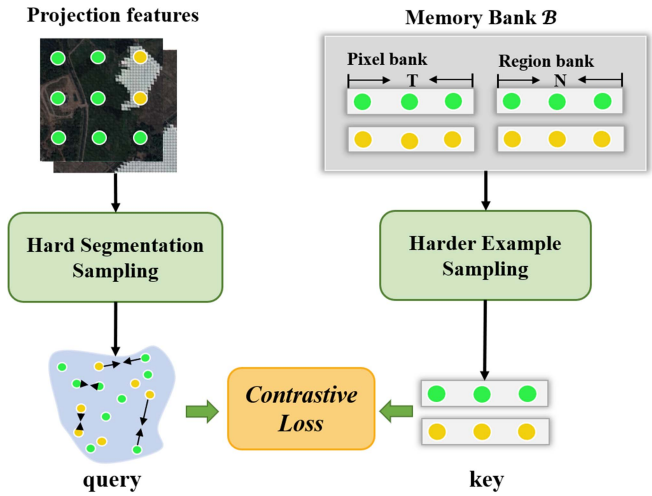

Fig. 5. Structure of CPCL.

the semantics for ICL, making pixel-level contrast as PCL more meaningful.

2) Pixel-to-Pixel and Pixel-to-Region Contrast: CPCL is introduced to explore the significantly different texture information between abandoned cropland and other land types. Through pixel-to-pixel and pixel-to-region contrasting, CPCL regularizes the embedding space by shortening the distance between the same class features while lengthening the different class features' distance. Both pixel embeddings and region embeddings are stored in a memory bank $\mathcal{B}$. The details of CPCL are shown in Fig. 5.

As for pixel-to-pixel contrast, given that pixel p in training images is the query with the semantic label $\bar{c}$, then the positive samples here are other pixels with the same label, while the negative samples are the pixels not belonging to $\bar{c}$. The positive and negative samples mentioned above as keys are not restricted to being selected from the same image.

For pixel-to-region contrast, it is proposed to supplement the image content information lost during the downsample process. Concerning pixel p labeled $\bar{c}$ as a query, the positive samples are the $\bar{c}$ class semantic regions in all images and the negative ones are the $\mathcal{C}\backslash\bar{c}$ classes semantic regions in the dataset.

During training, we select queries by the "hard segmentation sampling" strategy [70] and keys by the "harder example sampling" strategy [70], [71], [72], [73]. For the former, half of the queries are chosen randomly, and half are sampled from the harder queries. The harder queries here are the pixels with the wrong prediction in the segmentation task (*i.e.,* $c \neq \bar{c}$). This strategy guides the CPCL to focus on the pixels that make it difficult for the network to predict and intensify the critical feature generation. As for key selection, we use the "harder example sampling" strategy. For each query embedding $\boldsymbol{p}$, we select the top 10% harder negatives from memory bank $\mathcal{B}$ as negative collection, and positives are the same. The definition of "harder" here relates to the computation of the designed contrastive loss, and we will further explain it in Section III-C4. Then, we randomly sample K negative/positive embeddings from the respective collection to compute the designed contrastive loss $\mathcal{L}^{\text{NCE}}$. K denotes the number of samples here.

3) Memory Bank: Our designed memory bank $\mathcal{B}$ aims to balance training efficiency and representative capacity. The memory bank contains pixel and region embeddings. For pixel embeddings, a pixel queue with size T is stored for each category. The pixel embeddings are contained in $\mathcal{B}$ with a size of $|\mathcal{C}| \times T \times D$ and part of them (V/T) are dynamically updated by the recent batch. That is, during training, only a few pixels (i.e., $V, T \gg V$) are selected from the images in the latest batch

and pulled into the queue. The above design guarantees pixel embeddings' consistency and time efficiency in the memory bank. For region embeddings, providing that we have a segmentation dataset with N images and $|\mathcal{C}|$ classes, the keys for pixel-to-region contrast are region embeddings with size $|\mathcal{C}| \times N \times D$, where D is the dimension of the pixel embeddings. The $(\bar{c}, n)$th element of the region embeddings is the feature vector calculated by the average pooling of every pixel embedding with the same $\bar{c}$ class in the nth image. Therefore, the total size of the memory bank $\mathcal{B}$ is $|\mathcal{C}| \times (T + N) \times D$.

4) Loss Function of CPCL: The InfoCE loss is widely used in UCL, and it can be represented as

$$\mathcal{L}_P^{\text{NCE}} = -\log \frac{\exp\left(\boldsymbol{v} \cdot \boldsymbol{v}^+/\tau\right)}{\exp\left(\boldsymbol{v} \cdot \boldsymbol{v}^+/\tau + \sum_{\boldsymbol{v}^- \in \mathcal{N}_P} \exp\left(\boldsymbol{v} \cdot \boldsymbol{v}^-/\tau\right)\right)} \tag{1}$$

where $\boldsymbol{v}^+(\boldsymbol{v}^-)$ is the embedding of the positive(negative) sample for image P, $\mathcal{N}_P$ stores the embeddings of negative samples, "$\cdot$" denotes the dot product, and $\tau > 0$ is a temperature hyperparameter. All the embeddings here are ℓ_2-normalized.

CPCL extends (1) to the supervised dense prediction task to practice the pixel-to-pixel and pixel-to-region contrast mentioned above. It can be defined as

$$\mathcal{L}_p^{\text{NCE}} = \frac{1}{|\mathcal{P}_p|} \boldsymbol{X}. \tag{2}$$

And $\boldsymbol{X}$ is defined as follows:

$$\boldsymbol{X} = \sum_{\boldsymbol{e}^+ \in \mathcal{P}_p} -\log \frac{\exp\left(\boldsymbol{p} \cdot \boldsymbol{e}^+/\tau\right)}{\exp\left(\boldsymbol{p} \cdot \boldsymbol{e}^+/\tau\right) + \sum_{\boldsymbol{e}^- \in \mathcal{N}_p} \exp\left(\boldsymbol{p} \cdot \boldsymbol{e}^-/\tau\right)} \tag{3}$$

where $\mathcal{P}_p$and $\mathcal{N}_p$ denote the positive and negative sample collections stored in the memory bank $\mathcal{B}$ for pixel p, $\boldsymbol{e}^+$ and $\boldsymbol{e}^-$ are the embeddings of positives and negatives, respectively, and $\boldsymbol{p}$ represents the pixel embedding of the query pixel p.

The discernibility of training samples is vital in the segmentation task. In our work, the derivation of the contrastive loss (2) with respect to the query embedding $\boldsymbol{p}$ can be given as follows:

$$\frac{\partial \mathcal{L}_p^{\text{NCE}}}{\partial \boldsymbol{p}} = -\frac{1}{\tau |\mathcal{P}_p|} \boldsymbol{Y}. \tag{4}$$

And $\boldsymbol{Y}$ is defined as follows:

$$\boldsymbol{Y} = \sum_{\boldsymbol{e}^+ \in \mathcal{P}_p} \left((1 - m_{p+}) \cdot \boldsymbol{e}^+ - \sum_{\boldsymbol{e}^- \in \mathcal{N}_p} m_{p-} \cdot \boldsymbol{e}^- \right) \tag{5}$$

where $m_{p+/-} \in [0, 1]$ is the matching probability between the key $\boldsymbol{e}^+/\boldsymbol{e}^-$ and the query $\boldsymbol{p}$, the computation of the probability can be represented as follows:

$$m_{p+/-} = \frac{\exp\left(\boldsymbol{p} \cdot \boldsymbol{e}^{+/-}/\tau\right)}{\sum_{\boldsymbol{e}' \in \mathcal{P}_p \cup \mathcal{N}_p} \exp\left(\boldsymbol{p} \cdot \boldsymbol{e}'/\tau\right)}. \tag{6}$$

The dot product of query $\boldsymbol{p}$ and negative $\boldsymbol{e}^-$ with a value closer to 1 is deemed to be a sign of a harder negative sample. i.e., the negative key is similar to the query $\boldsymbol{p}$. Meanwhile, the positive $\boldsymbol{e}^+$ with a value closer to -1 is regarded as a harder positive, i.e., the positive key is dissimilar to the query $\boldsymbol{p}$.

TABLE I
EXPERIMENTAL RESULTS WITH OTHER NETWORKS

Method	F1	mIoU	OA	Precision	Recall
Unet++	63.9	69.8	93.2	64.8	63.0
DeepLabV3+	65.1	70.4	92.9	62.0	68.5
PSPnet	65.1	70.4	93.0	62.5	67.9
OCRnet	66.1	71.0	93.2	63.9	68.4
Segformer	67.2	**71.7**	93.3	63.0	**72.2**
PACnet	**68.3**	**71.7**	**93.8**	**70.7**	66.0

The bold font represents the highest performance of different methods.

TABLE II
EXPERIMENTAL RESULTS OF ABLATION STUDIES

Method	F1	mIoU	OA	Precision	Recall
Base	63.9	69.9	93.3	66.6	61.3
+CCAM	63.2	69.7	93.5	70.1	57.6
+CPCL	66.1	71.3	93.8	69.7	62.8
PACnet	**68.3**	**71.7**	**93.8**	**70.7**	**66.0**

The bold font represents the highest performance of different methods.

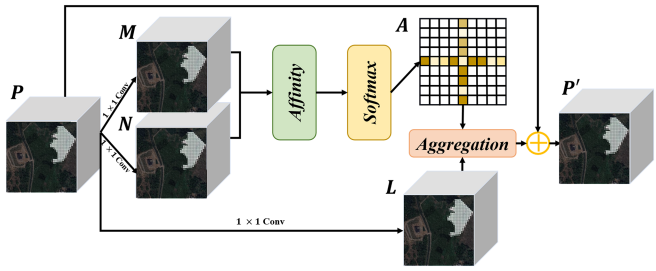

Fig. 6. Details of CCAM.

D. Criss-Cross Attention Module

The nearest surface features (e.g., cultivated farmland and shrub) in the current scene of an image contain massive and abundant content information, so we use CCAM to intensify the extraction of contextual importance and local feature representative capacity from two orientations within individual images. And better representative capacity in intraimages can help improve the performance of PACnet. Unlike nonlocal attention modules [74] that calculate all pixel weights directly, CCAM focuses on the pixels in essential directions and dramatically reduces the computed quantity.

The detailed structure of the CCAM is represented in Fig. 6. The CCAM captures contextual information from both lateral and longitudinal directions. For a feature map $\boldsymbol{P}$ with a spatial size of $H \times W \times D$, it first passes through two branches with 1×1 convolution and is transformed into $\boldsymbol{M} \in \mathbb{R}^{H \times W \times D'}$ and $\boldsymbol{N} \in \mathbb{R}^{H \times W \times D'}(D > D')$, respectively. Via the affinity computation of $\boldsymbol{M}$ and $\boldsymbol{N}$, we generate the affinity matrix, that is, the attention map $\boldsymbol{A}$ with a spatial size of $(H \times W - 1) \times (W \times H)$. For each pixel p in the feature $\boldsymbol{M}$, we can obtain $\boldsymbol{M}_p \in \mathbb{R}^{D'}$. By extracting the feature vector in the same row/column as p in the feature $\boldsymbol{N}$, we can acquire $\boldsymbol{\Omega}_p \in \mathbb{R}^{(H \times W - 1) \times D'}$. The affinity matrix is computed as follows:

$$z_{i,p} = \boldsymbol{M}_p \, \boldsymbol{\Omega}_{i,p}^T \tag{7}$$

where $\boldsymbol{\Omega}_{i,p} \in \mathbb{R}^{D'}$ is the ith element of $\boldsymbol{\Omega}_p (i = [1,\ldots,H+W-1])$, $z_{i,p} \in \boldsymbol{Z}$ is the correlation between $\boldsymbol{M}_p$ and $\boldsymbol{\Omega}_{i,p}$, and $\boldsymbol{Z} \in \mathbb{R}^{(H+W-1)\times(W\times H)}$.

After that, a softmax operation is applied on $\boldsymbol{Z}$ to calculate the final attention map $\boldsymbol{A}$. Another 1×1 filter used on $\boldsymbol{P}$ generates $\boldsymbol{L} \in \mathbb{R}^{H\times W\times D}$ for feature adaptation. For each pixel p in the feature $\boldsymbol{L}$, we can acquire a feature vector $\boldsymbol{L}_p \in \mathbb{R}^D$ and a collection of feature vectors $\boldsymbol{\Phi}_p \in \mathbb{R}^{(H+W-1)\times D}$ whose position is in the same row/column as p. The horizon and vertical contextual information of p are obtained by aggregation operation represented as follows:

$$\boldsymbol{P}'_p = \sum_{i=0}^{H+W-1} \boldsymbol{A}_{i,p}\, \boldsymbol{\Phi}_{i,p} + \boldsymbol{P}_p \tag{8}$$

where $\boldsymbol{P}'_p$ is a feature vector with a spatial size of $H \times W \times D$ for pixel p and $\boldsymbol{A}_{i,p}$ is the correlation value at channel i and pixel p in the attention map $\boldsymbol{A}$.

In (8), the contextual importance is joined to feature $\boldsymbol{P}$ to enhance the representative capacity. With a broader context extraction perspective and richer context aggregation from attention map $\boldsymbol{A}$, the PCAnet achieves significant progress and is more robust for the segmentation task.

E. Total Loss Function

Our loss function contains classical segmentation loss and the designed contrastive loss we put forward above. The former allows PACnet to study the discriminative features essential for abandoned cropland classification, and the latter enhances the contrast between abandoned farmland and surrounding ground features (e.g., cultivated cropland and shrubs) by explicitly exploring global semantics between pixel and region samples.

The segmentation loss we use in PACnet is the cross-entropy loss. Given pixel p in the image P is classified into a semantic class $\bar{c} \in \mathcal{C}$. The cross-entropy loss $\mathcal{L}^{\text{CE}}$ can be computed as follows:

$$\mathcal{L}_p^{\text{CE}} = -\mathbf{1}_{\bar{c}}^T \log\left(\text{softmax}\left(\boldsymbol{s}\right)\right) \tag{9}$$

where $\mathbf{1}_{\bar{c}}^T$ denotes the one-hot encoding of $\bar{c}$, $\bar{c} \in \mathcal{C}$ represents the label of pixel p, $\boldsymbol{s} = [s_1, s_2, \ldots, s_{|\mathcal{C}|}] \in \mathbb{R}^{|\mathcal{C}|}$is the unnormalized score vector for pixel p, and $\boldsymbol{s} \in \boldsymbol{S}$. For the softmax optimization, that is

$$\text{softmax}\ (s_c) = \frac{\exp\left(s_c\right)}{\sum_{c'=1}^{|\mathcal{C}|} \exp\left(s_{c'}\right)}. \tag{10}$$

The contrastive loss $\mathcal{L}^{\text{NCE}}$ is computed as (2) between the query embeddings and the key embeddings from the memory bank $\mathcal{B}$. The hard segmentation sampling strategy selects the former, and the harder example sampling strategy samples the latter. Then, the ultimate training loss $\mathcal{L}^{\text{Overall}}$ is computed as follows:

$$\mathcal{L}^{\text{Overall}} = \sum_p \left(\mathcal{L}_p^{\text{CE}} + \lambda \mathcal{L}_p^{\text{NCE}}\right) \tag{11}$$

where $\lambda > 0$ is the weight of contrastive loss.

IV. Experiments and Results Analysis

A. Experimental Setting

Our VACD dataset contains 14 735 samples with a size of 512×512. The size of the train set is 10 608, that of the validation set is 2653, and the test set is 1474. We train the model on the train set for 100 epochs with a batch size of 64. We use the stochastic gradient descent optimizer to optimize the parameters in the model. The initial learning rate is 0.01, the weight decay is 0.0001, and the momentum is 0.9. The temperature τ in (3) is set as 0.1. The weight λ of the contrastive loss $\mathcal{L}_p^{\text{NCE}}$ is 1. The learning rate decay strategy is LambdaLR with a step size of 100 and gamma of 0.5. Random horizontal flips and brightness are used to intensify the model's generalization for data augmentation. The probability of the image being flipped is 0.5. All training images are brightened, and the shift value is 10. The validation and test datasets do not have any augmentation operations. Our models and experiments are implemented by the open-source DL framework Pytorch. We train the model by the Distributed DataParallel strategy. 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. Evaluation Metrics

In this study, we use overall accuracy (OA), intersection over union (IoU), recall rate, precision rate, and $F1$ score to evaluate the effectiveness of all models. In binary classification, true positive (TP) represents the positive pixels in the label correctly classified as positive pixels. True negative (TN) means the negative pixels in the label correctly classified as negative pixels. False positive (FP) represents the negative pixels in the label, which are incorrectly classified as positive pixels. False negative (FN) means the label's positive pixels, which are incorrectly classified as negative pixels. TP, TN, FP, and FN are used to calculate the evaluation metrics.

IoU is a widely used metric in semantic segmentation, which calculates the intersection of label and prediction over the union of label and prediction, indicating the effectiveness of a model at pixel level by the overlap of label and prediction. mIoU is the average of the IoU of every class i. OA shows the overall prediction accuracy. Recall rate indicates the proportion of positive pixels identified in the label, while precision rate suggests the accuracy of all positive pixels in the prediction. $F1$ score weighs the recall rate and precision rate to represent the overall performance to avoid bias due to sample imbalance. All these five metrics can be calculated as follows:

$$\text{OA} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{12}$$

$$\text{IoU} = \frac{\text{TP}}{\text{FN}+\text{TP}+\text{FP}} \tag{13}$$

$$\text{mIoU} = \frac{\sum_{i=1}^{n} \text{IoU}_i}{n} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{15}$$

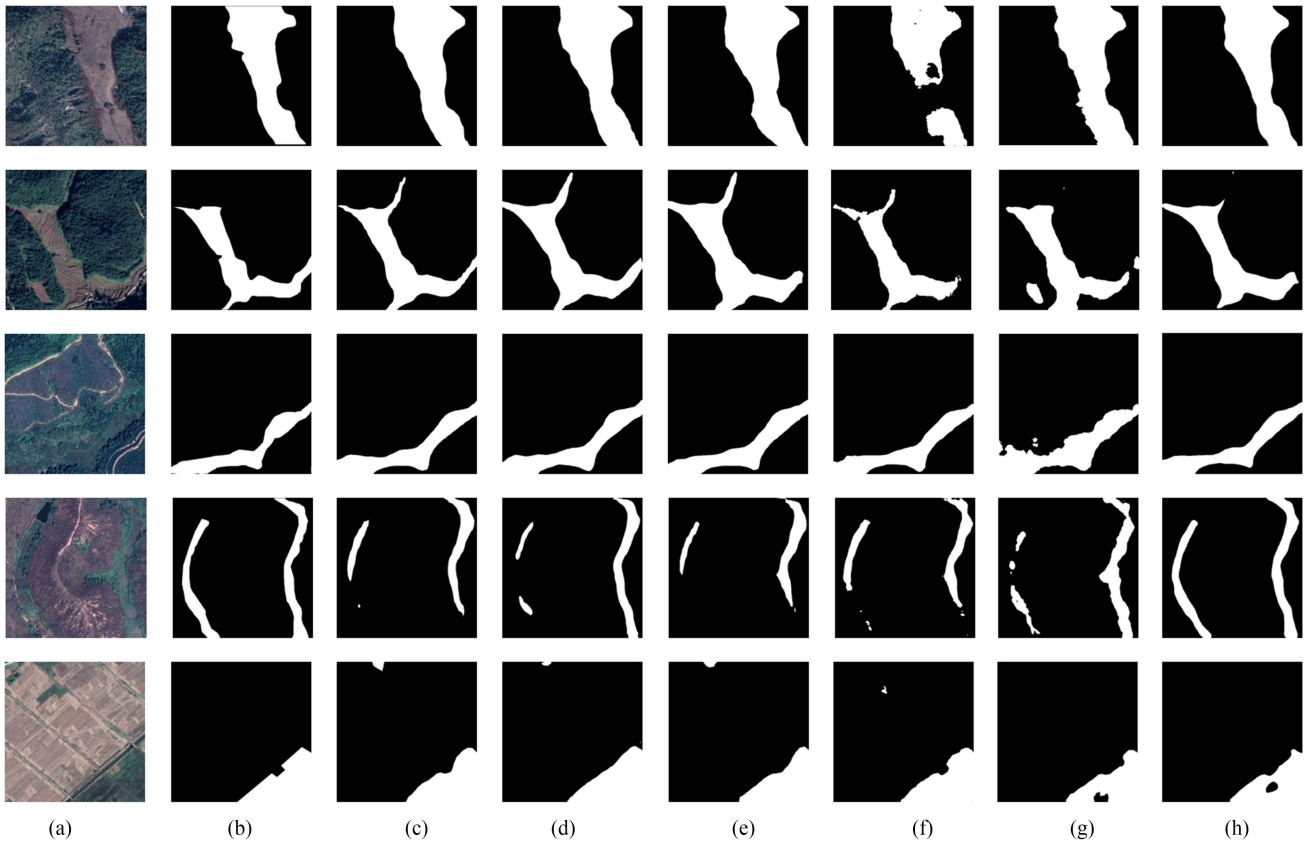

Fig. 7. Visualization results of comparisons with other methods. (a) Original images. (b) Ground truth labels. (c) DeepLabV3+. (d) OCRnet. (e) PSPnet. (f) Unet++. (g) Segformer. (h) Our PACnet.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$F1\ \text{Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{17}$$

C. Comparisons With Other Methods

To better demonstrate the effectiveness of PCAnet, we select the following mainstream segmentation networks for comparison on the VACD dataset and quantify the evaluation results.

Unet++: Unet++ [75] is a semantic segmentation network developed from Unet. It nests encoder and decoder subnetworks to Unet and redesigns the skip-connection module in Unet. By adding the deep supervision mechanism, Unet++ achieves faster model convergence.

DeepLabV3+: DeepLabV3+ [41] is one of the DL networks in the DeepLab series. DeepLabV3+ uses a typical encoder–decoder network structure. The encoder can extract multiple-resolution features, and by introducing ASPP, DeepLabV3+ expands the receptive field and enhances the representative capacity.

Pyramid scene parsing network (*PSPnet*): The critical module of the PSPnet [76] is the pyramid pooling module. It can combine contextual information from diverse regions and obtain a more potent global representative capacity. To some extent, PSPnet solves the problem of mismatch of pixel context, confused semantic labels, and difficulty in small class prediction.

OCRnet: OCRnet [77] is often paired with HRnet as a backbone to obtain high-quality context importance and maintain high-resolution features. OCRnet implements a coarse-to-fine strategy to get a pixel-enhanced object-contextual representation.

Segformer: Segformer [78] is a simple and efficient semantic segmentation network with a transformer framework. Segformer extracts multiscale features by using a hierarchically structured transformer decoder.

It can be seen from Table I that our proposed PACnet achieves the highest OA, mIoU, Precision rate, and $F1$ score of 93.8%, 71.7%, 70.7%, and 68.3%, respectively. Segformer obtains the highest recall rate of 72.2% and 6.2% higher than our proposed PACnet due to its transformer framework. Although the recall rate of our proposed PACnet is lower than some other competitive models, the precision rate is much higher due to a specific inhibitory effect on noise labels, which will be discussed and analyzed in Section V. Among all these models, our proposed PACnet obtains the best result considering all accuracy metrics comprehensively.

From Fig. 7, we could find that the extraction result of our proposed PACnet is smooth and precise. From the images in Row 4, it is evident that other models cannot extract the target abandoned cropland completely but PACnet does. Besides, the

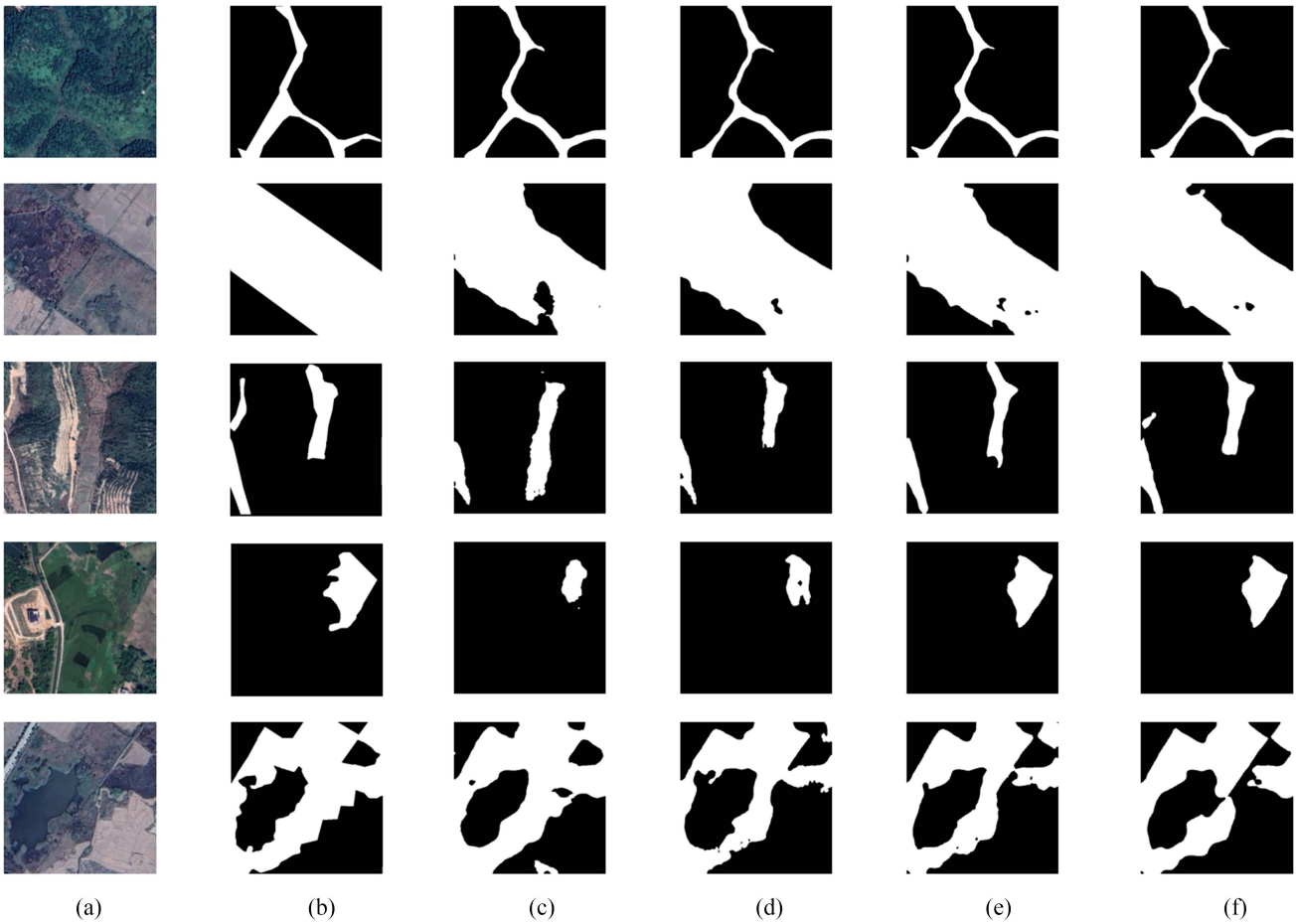

Fig. 8. Visualization results of ablation experiments. (a) Original images. (b) Ground truth labels. (c) DeepLabV3. (d) DeepLabV3+CCAM. (e) DeepLabV3+CPCL. (f) Our PACnet.

prediction of PACnet is more in line with the actual surface of the label in Fig. 7(b) because of its noise resistance. In conclusion, our proposed PACnet with its promising ability to capture texture information and hidden key features makes fewer mistakes than other models and can extract the complete abandoned cropland in a more complex scene.

D. Ablation Experiments

To represent the influence of CPCL and CCAM, we conduct ablation experiments on the VACD dataset and quantify the evaluation results. First, we carry out the baseline experiment of the initial network with ResNet50 as the encoder and DeepLabV3 as the decoder. Then, we add the CCAM into the baseline model to better extract the intraimage features from different directions. In like manner, we introduce CPCL to the baseline to enhance the interimage feature extraction. Finally, the experiment of the baseline with CPCL and CCAM is conducted.

Table II presents the metric results of our ablation experiments. The "Base" model is the baseline model without any tricks. "+CCAM" represents the base model with CCAM. "+CPCL" means the base model with CPCL. From Table II, we can find that adding CCAM into the baseline improves by 0.2% in OA and 3.5% in precision rate, which indicates that CCAM pays more attention to contextual importance and local feature representative capacity from two orientations within images to lower the possibility of mistakenly classifying. Moreover, the involvement of CPCL improves the performance of the baseline remarkably in all metrics. Therefore, PACnet with CCAM and CPCL at the same time obtains further improvement compared with the baseline with only a single module.

Fig. 8 further demonstrates the function of CCAM and CPCL. We can find in Fig. 8(c) and (d) that the baseline makes some mistakes in organizing the background as abandoned cropland, but the baseline with CCAM does not. By comparing Fig. 8(d) and (e), we can find that CPCL makes fewer mistakes and tends to extract the abandoned cropland more completely. The prediction results in Fig. 8(f) are significantly closer to the accurate label, which proves the effectiveness of our proposed CCAM and CPCL. We believe that using CCAM and CPCL simultaneously can enhance PACnet's ability to extract comprehensive texture

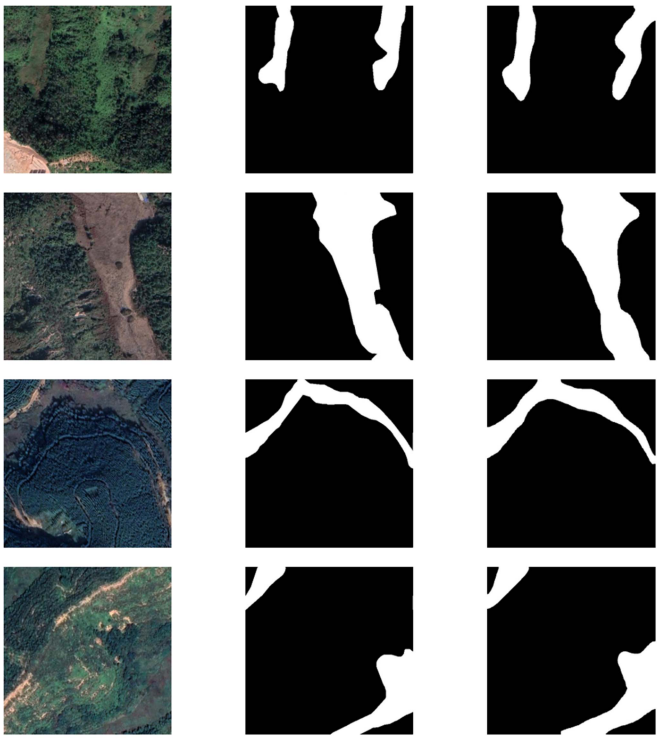

Fig. 9. Visualization details of the inhibitory effect of CPCL on noise labels. (a) Image. (b) Label. (c) PACnet.

information and essential deep features, leading to improved performance.

V. Discussion

A. Noise Suppression Ability of PACnet

Due to the difficulty in ensuring absolute accuracy through manual annotation, labels used for training often have some noise. Indeed, some studies have shown that contrastive learning has a particular antinoise performance for labels with noise [79], [80]. The experimental results indicate that the CPCL used in this article has a specific inhibitory effect on noise labels because CPCL smooths out erroneous information in segmentation loss by continuously comparing the similarities and differences of pixels. As shown in Table I, the method proposed in this article has a relatively low recall rate but the highest precision rate. This suggests that CPCL can identify more typical abandoned labels, implying a certain level of noise resistance against incorrect labels. As shown in Fig. 9, for labels that were not accurately annotated during the training process, predicted results could be closer to the actual surface textures. The above findings further demonstrate the inhibitory effect of CPCL on noise labels, which alleviates the considerable cost of fine labeling and is worth further exploration.

B. Pros and Cons of PACnet

As we analyze in Section IV, PACnet can accurately extract fine-grained abandoned cropland from single-time-phase VHR data. Our qualitative and quantitative assessments substantiate that PACnet outperforms mainstream segmentation networks. This success can be attributed to the incorporation of CPCL and CCAM. PACnet distinguishes itself by emphasizing capturing intra- and interimage contrastive features, a critical aspect often overlooked by classical models. This differentiation is particularly significant, given the inherent complexity of directly modeling amorphous abandoned cropland. The integration of CPCL into PACnet enhances its proficiency in discerning different characteristics of abandoned cropland and other land types at both pixel and semantic region levels. Furthermore, introducing CCAM enriches the network's representative capacity, contributing to its superior performance.

Nevertheless, it is imperative to acknowledge a limitation in PACnet's performance. Our sampled experimental area contains a substantial expanse equivalent to that of a province. This coverage confirms PACnet's proficiency in mapping fine-grained abandoned cropland across southern China. However, we remain aware that its efficacy might not be universally consistent when applied to regions characterized by distinct topography and cropland attributes. Addressing this potential deficiency constitutes a crucial direction for our future work, where we intend to prioritize the advancement of PACnet's transfer learning capabilities and will further mine the abundant information of time-series data.

VI. Conclusion

In this article, faced with the problems of farmland fragmentation and amorphous characteristics of abandoned cropland in southern China, we proposed a new fine-grained abandoned cropland mapping method (PACnet) based on the pixel-level contrast learning. By integrating CPCL and CCAM, our proposal enhances the comparative characteristics between abandoned land and other land features from inter- and intraimages. The experimental results show that PACnet has the highest accuracy (OA = 93.8% and mIOU = 71.7%) in mapping abandoned cropland compared with classical DL algorithms. We can find that CPCL has a specific inhibitory effect and antinoise performance on inaccurate labels. Our proposed method has vital reference significance for VHR abandoned cropland mapping and analysis research. In the future, we will continue to explore the synergistic use of time-series features and VHR images to map abandoned cropland more accurately.

References

[1] T. Lasanta, J. Arnáez, N. Pascual, P. Ruiz-Flaño, M. Errea, and N. Lana-Renault, "Space–time process and drivers of land abandonment in Europe," *Catena*, vol. 149, pp. 810–823, 2017.

[2] D. MacDonald et al., "Agricultural abandonment in mountain areas of Europe: Environmental consequences and policy response," *J. Environ. Manage.*, vol. 59, no. 1, pp. 47–69, 2000.

[3] C. Ren et al., "Ageing threatens sustainability of smallholder farming in China," *Nature*, vol. 616, no. 7955, pp. 96–103, 2023.

[4] C. Queiroz, R. Beilin, C. Folke, and R. Lindborg, "Farmland abandonment: Threat or opportunity for biodiversity conservation? A global review," *Front. Ecol. Environ.*, vol. 12, no. 5, pp. 288–296, 2014.
[5] D. K. Munroe, D. B. van Berkel, P. H. Verburg, and J. L. Olson, "Alternative trajectories of land abandonment: Causes, consequences and research challenges," *Curr. Opin. Environ. Sustain.*, vol. 5, no. 5, pp. 471–476, 2013.
[6] N. Vuichard, P. Ciais, L. Belelli, P. Smith, and R. Valentini, "Carbon sequestration due to the abandonment of agriculture in the former USSR since 1990," *Glob. Biogeochem. Cycles*, vol. 22, no. 4, 2008, Art. no. 2154.
[7] M. Baumann et al., "Patterns and drivers of post-socialist farmland abandonment in western Ukraine," *Land Use Policy*, vol. 28, no. 3, pp. 552–562, 2011.
[8] D. Hou, F. Meng, and A. V. Prishchepov, "How is urbanization shaping agricultural land-use? Unraveling the nexus between farmland abandonment and urbanization in China," *Landscape Urban Plan.*, vol. 214, 2021, Art. no. 104170.
[9] M. Zhang et al., "Reveal the severe spatial and temporal patterns of abandoned cropland in China over the past 30 years," *Sci. Total Environ.*, vol. 857, 2023, Art. no. 159591.
[10] B. Liu and W. Song, "Mapping abandoned cropland using within-year sentinel-2 time series," *Catena*, vol. 223, 2023, Art. no. 106924.
[11] H. Yin et al., "Monitoring cropland abandonment with Landsat time series," *Remote Sens. Environ.*, vol. 246, 2020, Art. no. 111873.
[12] X. Zhao, T. Wu, S. Wang, K. Liu, and J. Yang, "Cropland abandonment mapping at sub-pixel scales using crop phenological information and MODIS time-series images," *Comput. Electron. Agriculture*, vol. 208, 2023, Art. no. 107763.
[13] S. Estel, T. Kuemmerle, C. Alcántara, C. Levers, A. Prishchepov, and P. Hostert, "Mapping farmland abandonment and recultivation across Europe using MODIS NDVI time series," *Remote Sens. Environ.*, vol. 163, pp. 312–325, 2015.
[14] C. Alcantara et al., "Mapping the extent of abandoned farmland in central and eastern Europe using MODIS time series satellite data," *Environ. Res. Lett.*, vol. 8, no. 3, 2013, Art. no. 035035.
[15] O. Dubovyk, G. Menz, and A. Khamzina, "Trend analysis of MODIS time-series using different vegetation indices for monitoring of cropland degradation and abandonment in Central Asia," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 6589–6592.
[16] Z. Zhao, J. Wang, L. Wang, X. Rao, W. Ran, and C. Xu, "Monitoring and analysis of abandoned cropland in the Karst plateau of eastern Yunnan, China based on Landsat time series images," *Ecol. Indicators*, vol. 146, 2023, Art. no. 109828.
[17] 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.
[18] 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.
[19] L. Xia, J. Luo, Y. Sun, and H. Yang, "Deep extraction of cropland parcels from very high-resolution remotely sensed imagery," in *Proc. Int. Conf. Agro-Geoinform., Agro-Geoinform.*, 2018, pp. 1–5.
[20] J. Song, L. Miao, Q. Ming, Z. Zhou, and Y. Dong, "Fine-grained object detection in remote sensing images via adaptive label assignment and refined-balanced feature pyramid network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 71–82, Nov. 2022.
[21] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.
[22] S. Liu et al., "A shallow-to-deep feature fusion network for VHR remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5410213.
[23] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2019.
[24] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, 2021, Art. no. 114417.
[25] M. Luo and S. Ji, "Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 105–128, 2022.
[26] X. Dong, C. Zhang, L. Fang, and Y. Yan, "A deep learning based framework for remote sensing image ground object segmentation," *Appl. Softw. Comput.*, vol. 130, 2022, Art. no. 109695.
[27] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.
[28] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan, "CGSANet: A contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1526–1542, Dec. 2021.
[29] J. P. Ardila, W. Bijker, V. A. Tolpekin, and A. Stein, "Context-sensitive extraction of tree crown objects in urban areas using VHR satellite images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 15, pp. 57–69, 2012.
[30] S. Dong, Y. Zhuang, Z. Yang, L. Pang, H. Chen, and T. Long, "Land cover classification from VHR optical remote sensing images by feature ensemble deep learning network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1396–1400, Aug. 2020.
[31] J. Wang, W. Li, Y. Wang, R. Tao, and Q. Du, "Representation-enhanced status replay network for multisource remote-sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, Jun. 2023, doi: 10.1109/TNNLS.2023.3286422.
[32] Y. Liu, L. Wang, J. Cheng, and X. Chen, "Multiscale feature interactive network for multifocus image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, Oct. 2021, Art. no. 5019316.
[33] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5526914.
[34] J. Wang, F. Gao, J. Dong, S. Zhang, and Q. Du, "Change detection from synthetic aperture radar images via graph-based knowledge supplement network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1823–1836, Feb. 2022.
[35] Q. Liu, M. Kampffmeyer, R. Jessen, and A. B. Salberg, "Dense dilated convolutions merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.
[36] 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.
[37] Q. Shen, H. Deng, X. Wen, Z. Chen, and H. Xu, "Statistical texture learning method for monitoring abandoned suburban cropland based on high-resolution remote sensing and deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3060–3069, Mar. 2023.
[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.
[40] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv.1706.05587*.
[41] 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.
[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[43] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3141–3149.
[44] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.
[45] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6877–6886.
[46] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2021, Art. no. 4404512.
[47] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[48] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.
[49] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNNS," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. III-3, pp. 473–480, Jun. 2016.
[50] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[51] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[52] G. Wu et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 407.
[53] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sens.*, vol. 9, no. 5, May 2017, Art. no. 446.
[54] Y. Liu, D. M. Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-shapeNetwork based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, May 2017, Art. no. 522.
[55] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 2000415.
[56] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Aug. 2023, Art. no. 5617116.
[57] J. Zhou, S. Zeng, G. Gao, Y. Chen, and Y. Tang, "A novel spatial-spectral pyramid network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Aug. 2023, Art. no. 5519314.
[58] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-sensing scene classification via multistage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jul. 2023, Art. no. 5615312.
[59] M. Li et al., "Remote sensing object detection based on strong feature extraction and prescreening network," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jan. 2023, Art. no. 8000505.
[60] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Hyperspectral and SAR image classification via multiscale interactive fusion network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10823–10837, Dec. 2023, doi: 10.1109/TNNLS.2022.3171572.
[61] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.
[62] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6203–6212.
[63] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[64] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv.2002.05709*.
[65] K. He, R. Girshick, and P. Dollar, "Rethinking imagenet pre-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4917–4926.
[66] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," *Vis. Inform.*, vol. 7, no. 1, pp. 30–40, Mar. 2023.
[67] A. Bardes, J. Ponce, and Y. LeCun, "VICRegL: Self-supervised learning of local visual features," 2022, *arXiv.2210.01571*.
[68] X. Zhao et al., "Contrastive learning for label efficient semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10603–10613.
[69] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," 2020, *arXiv.2006.10511*.
[70] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7283–7293.
[71] J. Xie, X. Zhan, Z. Liu, Y.-S. Ong, and C. C. Loy, "Delving into inter-image invariance for unsupervised visual representations," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 2994–3013, Dec. 2022.
[72] T. T. Cai, J. Frankle, D. J. Schwab, and A. S. Morcos, "Are all negatives created equal in contrastive instance discrimination?," 2020, *arXiv.2010.06682*.
[73] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[74] 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.
[75] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov Ed. Cham, Switzerland: Springer, 2018, pp. 3–11.
[76] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6230–6239.
[77] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.
[78] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.
[79] J. Li, C. Xiong, and S. C. H. Hoi, "Learning from noisy data with robust representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9465–9474.
[80] S. Li, X. Xia, S. Ge, and T. Liu, "Selective-supervised contrastive learning with noisy labels," 2022, *arXiv:2203.04181*.

Haoyang Li received the B.S. degree in geographic information science in 2022 from Sun Yat-sen University, Guangzhou, China, where he is currently working toward the Ph.D. degree in cartography and geographic information system with the School of Geography and Planning.

His research interests include VHR images LULC mapping, agricultural remote sensing, and multimodal data fusion.

Haomei Lin is currently working toward the B.S. degree in geographic information science with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China.

Her research interests include remote sensing processing, deep learning, and geospatial simulation.

Junshen Luo is currently working toward the B.S. degree in geographic information science with Sun Yat-sen University, Guangzhou, China.

His research interests include an intelligent understanding of remote sensing images, machine learning, and deep learning.

Teng Wang received the M.S. degree in communication engineering from the University of Technology Sydney, Ultimo, NSW, USA, in 2014.

His research interests include remote sensing image classification.

Hao Chen received the B.S. degree in computer science and technology from Northeast Petroleum University, Daqing, China, in 2018.

His research interests include intelligent remote sensing mapping.

Qiuting Xu received the B.S. degree in land resource management from South China Agricultural University, Guangzhou, China, in 2016.

Her research interests include agricultural remote sensing mapping.

Xinchang Zhang received the B.S. degree in cartography from the Wuhan Institute of Surveying and Mapping, Wuhan, China, in 1982, the M.S. degree in cartography from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1994, and the Ph.D. degree in resources and environmental sciences from Wuhan University, Wuhan, China, in 2004.

His research interests include spatial database updating, spatial data integration, and smart cities.