Semantic-Aware Region Loss for Land-Cover Classification

Xianwei Zheng, *Member, IEEE*, Qiyuan Ma, Linxi Huan, Xiao Xie, Hanjiang Xiong, and Jianya Gong

***Abstract*—Integrating superpixel segmentation into convolutional neural networks is known to be effective in enhancing the accuracy of land-cover classification. However, most of existing methods accomplish such integration by focusing on the development of new network architectures, which suffer from several flaws: conflicts between general superpixels and semantic labels introduce noise into the training, especially at object boundaries; absence of training guidance for superpixels leads to ineffective regional feature learning; and unnecessary superpixel segmentation in the testing stage not only increases the computational burden but also incurs jagged edges. In this study, we propose a novel semantic-aware region (SARI) loss to guide the effective learning of regional features with superpixels for accurate land-cover classification. The key idea of the proposed method is to reduce the feature variance inside and between homogeneous superpixels while enlarging feature discrepancy between heterogeneous ones. The SARI loss is thus designed with three subparts, including superpixel variance loss, intraclass similarity loss and interclass distance loss. We also develop semantic superpixels to assist in the network training with SARI loss while overcoming the limitations of general superpixels. Extensive experiments on two challenging datasets demonstrate that the SARI loss can facilitate regional feature learning, achieving state-of-the-art performance with mIoU scores of around 97.11% and 73.99% on Gaofen Image dataset and DeepGlobe dataset, respectively.**

***Index Terms*—Deep learning, land-cover classification, region loss, remote sensing.**

I. Introduction

LAND-COVER classification of remote sensing images is of considerable significance to a wide range of applications, such as precision agriculture [1], [2], urban planning [1], [3], and environmental monitoring [4]. With the rapid development of new sensors and data acquisition technologies, the spatial resolution of remote sensing images has significantly improved, opening up new opportunities to obtain fine-level land covers. However, the huge details contained in a remote sensing image hamper the extraction of useful information relevant to land-cover classification [5]. Meanwhile, remote sensing images commonly cover a large spatial extent, where objects from different areas may have considerable diversities, bringing extra difficulties to accurate and efficient land-cover classification [6].

Land-cover classification is typically regarded as a problem of semantic segmentation, with the goal to assign each pixel in an image with predefined land-cover categories, such as forest and river. In recent years, the performance of semantic segmentation has been significantly improved through deep learning-based methods, benefiting from the powerful hierarchical feature representation ability of *convolutional neural networks* (ConvNets/CNNs). In particular, following the first end-to-end fully convolutional network (FCN) [7], many FCN models with newly developed learning tools have been presented to further promote the performance of semantic segmentation. For example, multiscale feature fusion modules, such as atrous spatial pyramid pooling (ASPP) [8] and denseASPP [9], address the problem of large scale variation of objects in both natural and remote sensing images, and highly contribute to the improvement of segmentation accuracy. However, the huge volume of details and complex object spectrum of remote sensing images bring more intraclass variances and interclass similarities than natural images. These factors confuse the general models devised for natural image semantic segmentation, resulting in unexpected land-cover classification results. Therefore, numerous methods have attempted to enhance the network's ability in intra-class unification and interclass discrimination for remote sensing images.

It is widely regarded that visually consistent objects are easy to classify while visually inconsistent objects (e.g., different parts of an object with diverse textures) tend to mislead the network's recognition. However, the fact is that misclassification is also frequently found in the regions with high visual consistency. Specifically, two image patches that have similar appearance/texture and belong to the same object or land-cover category can derive very different convolutional features and are thus incorrectly segmented into different classes. This phenomenon indicates that CNNs have limitations in preserving region consistency during feature extraction and propagation. In addressing this technical hurdle of CNNs, a natural choice

Manuscript received 20 February 2023; revised 16 March 2023; accepted 29 March 2023. Date of publication 7 April 2023; date of current version 4 May 2023. This work was supported in part by the National Natural Science Foundation of China Project under Grant 42071370 and in part by the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001. *(Corresponding author: Xiao Xie; Hanjiang Xiong.)*

Xianwei Zheng, Qiyuan Ma, Linxi Huan, and Hanjiang Xiong are with the State Key Laboratory Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zhengxw@whu.edu.cn; qiyuanma@whu.edu.cn; whu_hlx@whu.edu.cn; xionghanjiang@whu.edu.cn).

Xiao Xie is with the Key Lab for Environmental Computation and Sustainability of Liaoning Province, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: xiexiao@iae.ac.cn).

Jianya Gong is with the State Key Laboratory Information Engineering in Surveying, Mapping and Remote Sensing, and the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: gongjy@whu.edu.cn).

Code is available at: https://github.com/geovsion/SARI.

Digital Object Identifier 10.1109/JSTARS.2023.3265365

is to use the superpixel segmentation, which can group visually similar pixels into regions. Early methods mainly use the superpixel segmentation to simplify the pixel-wise classification [10], [11]. Modern end-to-end CNN-based methods prefer to deploy superpixel segmentation sub-networks to obtain the similarity between pixels, and thus, enhance the learning of contextual information for coherent semantic prediction [12], [13], [14]. In such a network architecture, superpixel segmentation must be performed in both training and prediction stages. However, we argue that superpixel segmentation is not necessary in the prediction stage, as it not only reduces the computation efficiency but also leads to jagged edges.

In the search for more powerful feature representations that can improve semantic segmentation, extensive studies focused on the development of new neural network architectures. Following this trend, superpixel segmentation was widely adopted and devised as a part of the network structure. However, despite improvements observed in land-cover classification, the lack of task-specific loss function for superpixel learning still limits the performance and efficiency of existing models. Loss functions guide the network learning during training. Well-designed network structures can bring benefits to the information extraction, while the loss function determines whether and how the information are learned and used. With the aim to address certain issues that cannot be circumvented by simply changing the structure of networks, task-specific loss functions have gained interest. For example, the weighted cross entropy and focal loss [15] are developed to alleviate the effect of extreme class imbalance on model training, and the tracing loss [16] and edge-aware (EA) loss [17] are presented to guide the network in distinguishing the edge and nonedge pixels. These loss functions are demonstrated to be highly effective in boosting the performance of a given model. Unfortunately, most of existing models presented for the land-cover classification are still trained by the commonly used cross entropy (CE) loss, which treats each pixel independently (i.e., calculates loss pixel by pixel). The CE loss is thus ineffective in guiding the network to learn the relationship between neighboring pixels [18].

On the basis of the aforementioned observation, we are motivated to study a different aspect of the network design for superpixel-enhanced land-cover classification, that is, the semantic-aware region (SARI) loss guidance. Our goal is to reduce regional representation variance in semantic patches under the training guidance of SARI loss. To define the semantic patches, we derive semantic superpixels from training images with the assistance of corresponding semantic annotations instead of adding an extra superpixel learning branch. In this way, compared with general superpixels, semantic ones fit closer to the boundaries of ground objects with higher internal semantic uniformity, and are thus more compatible with the semantic segmentation for improving local prediction coherence. In the training stage, the semantic superpixels serve as supervision signals that guide a model to reduce the representation variance of features in each superpixel. The aim is to enhance the regional representation coherence and obtain consistent classification prediction for visually similar pixels that belong to the same object or category. Such strategy of strengthening regional consistency via a task-specific loss function frees the model from the reliance on additional superpixels at the testing phase, and therefore improves the accuracy and efficiency of the inference. Apart from serving as local consistency constraints, semantic superpixels can further benefit long-range information learning from the following two aspects:

1) Improving feature similarity between superpixels that share the same categories (homogeneous superpixels);
2) Enlarging feature discrepancy between superpixels that fall into different classes (heterogeneous superpixels).

Supported by the intra- and intersuperpixel constraints, the proposed SARI loss can teach a model with a comprehensive consideration of both regional and long-range relationships for high inference coherence in land-cover classification.

The main contributions of this study are summarized as follows:

1) We propose a SARI loss based on tailored semantic superpixels to improve the local classification consistency by reducing the representation variance in each superpixel, and further enhance the long-range consistency and discrimination between semantic regions by imposing intersuperpixel constraints.
2) We implement the proposed method with a superpixel-supervised encode–decoder network (SPSNet) that effectively aggregates multiscale features to exert the power of the SARI loss in guiding the learning of representative and semantically consistent features in land-cover classification.
3) Under the guidance of SARI loss, the SPSNet achieves state-of-the-art performance on two challenging land-cover classification benchmarks, Gaofen Image dataset (GID) [5] and DeepGlobe dataset [19], with mIoU scores of around 97.11% and 73.99%, respectively.

II. Related Work

A. Land-Cover Classification

The goal of land-cover classification is to assign land-cover categories to each pixel in a remote sensing image. Initially, land-cover classifiers are dominated by traditional supervised machine learning methods, including parametric classifiers (e.g., maximum-likelihood classifiers [20]), nonparametric classifiers (e.g., decision trees [21], support vector machines (SVMs) [22], and artificial neural networks (ANNs) [23]), and ensemble methods (e.g., random forests [24] and boosting [25]). Traditional classifiers mainly utilize the low-level spectral, textural and/or shape features extracted from local pixels to interpret land-cover types in a remote sensing image [11]. In addition, the Markov random fields (MRFs) and conditional random fields (CRFs) are also used to refine classification results [11]. However, traditional methods relying on low-level cues have difficulties in capturing the contextual information and spatial relationship of ground objects [26], and thus, their performance is severely limited [27].

Recently, deep learning has made progress in semantic segmentation (i.e., pixel-wise classification) tasks, and many deep CNNs have reported impressive results on natural image parsing.

Compared with traditional handcrafted feature-based methods, CNNs are more capable of hierarchical feature abstraction and high-level semantic information extraction. In particular, the development of the first end-to-end FCN brings semantic segmentation into a new era. Compared with traditional methods, the FCN can generate pixel-wise semantic labeling in an end-to-end manner, without the utilization of additional classifiers. However, the initial version of the FCN suffers from two problems, namely, the unbalanced segmentation quality for multiscale objects (due to the fixed receptive field) and significant detail loss (due to down-sampling). As a result, many learning techniques have been continuously developed to tackle the aforementioned two problems. For example, U-Net [28] introduces the encoder–decoder structure with skip connections to merge low-level details and high-level semantic information, and thus, multiscale features expressed in multiple convolution layers are fused to improve the multiscale object segmentation. For large objects, better segmentation is achieved by using dilated convolution that enlarges the receptive field in DeepLabv1 [29] and dilated residual network (DRN) [30]. Later, the spatial pyramid pooling (SPP) module [31] was proposed to provide multiple effective receptive fields, and embedding its variants into the encoder–decoder framework becomes a common solution for multiscale context capturing [32], [33]. For refining boundary details during segmentation, the discriminative feature network (DFN) [34] combines boundary detection subnetworks with semantic segmentation to amplify the distinction of features and EaNet [17] distinguishes the edge and nonedge pixels under the guidance of an edge-aware (EA) loss.

Along with the rapid development of deep learning-based methods for semantic segmentation of natural images, interest is growing in designing deep models for land-cover classification of remote sensing images. However, the complex spectrum and irregular boundaries of ground objects in remote sensing images pose a considerable challenge to the robust learning of ground objects for land-cover classification [35]. Contextual information refers to the relationship between pixels and objects [36], which serves as a significant cue for correct recognition of ground objects and coherent labeling of land-cover categories. Many methods attempted to excavate the rich contextual information to build relationships between pixels with local and long-range dependencies [37]. For instance, Zhao et al. [26] applied the object-based CRF to strengthen the contextual information of the raw semantic predictions acquired by a CNN. ScasNet [38] aggregates global-to-local contexts captured by the CNN in a self-cascaded manner. The ERN [39] introduces the spatial boundary context to alleviate the ambiguity resulting from the interclass similarity and shadows. RA-FCN [37] utilizes spatial and channel relation modules to learn and reason the global relationships between similar objects. To enrich contextual information in HRNet [40], Zhang et al. [41] modeled the long-range spatial correlations among the low-resolution features by using a spatial reasoning module and aggregated local contexts based on high-resolution features via an adaptive spatial pooling module. For the same purpose, HRCNet [42] obtains global contextual information through a light-weight dual attention module and fuses the multiscale contextual information by a feature enhancement pyramid structure.

B. Superpixel-Enhanced Semantic Segmentation

Superpixels are oversegmentation of images, which are generated by simply utilizing low-level image features to group pixels into perceptually meaningful regions [43], [44]. By combining the advantages of perceptual uniformity [45] and contour adherence [46], superpixels offer a more natural representation than individual image pixels [47]. The merits of superpixels have been extensively explored in diverse vision tasks, such as object detection [48], object tracking [49], optical flow estimation [50], and 3-D reconstruction [51].

In the semantic segmentation, many early methods directly replaced pixel-wise classification with a superpixel-based one to simplify and accelerate the classification [10], [11]. For instance, for superpixel-wise classification, the zoom-out network [52] first groups the generated superpixels into regions at different levels, and then, extracts multiscale features from local to global zoom-out regions. This approach avoids complex and expensive pixel-wise inference. Gadde et al. [53] adopted a bilateral inception (BI) module to perform convolution over superpixels and implements a direct long-range edge-aware inference between superpixels. The BI module provides an efficient integration of superpixels into semantic segmentation, and is thus widely adopted for a convenient superpixel-wise semantic segmentation, like in SEAL [54] and SSN [44].

Recently, the efficiency and accuracy of semantic segmentation has been significantly promoted by deep learning. The modern FCNs can realize an efficient end-to-end pixel-wise semantic prediction without using superpixels as the basic segmentation units. However, the local nature of the convolution of deep neural networks causes difficulties in capturing contextual information across large regions, leading to unexpected segmentation results (e.g., fragmented segmentation inside an object and blurry boundary). In remedying this deficiency of deep neural networks, superpixel segmentation has an unparalleled merit of semantic consistency preservation for visually consistent image regions. Thus, superpixel segmentation is often combined with deep neural networks for enhanced semantic segmentation. For instance, SDNF [12] uses an extra network branch to generate superpixels, which are then used to refine edges and enhance the classification consistency within classes. Inspired by SEAL [54], SDNF additionally designs a loss function to simultaneously guide the training of superpixel segmentation and superpixel-enhanced semantic segmentation. However, their loss function only pays attention to the pixels inside each superpixel and on the boundaries. Xu et al. [13] refined segmentation edges by using superpixels fine-tuned from pretrained models and adopted the logit consistency module to guarantee the classification consistency. Ouyang and Li [14] appended a graph convolutional neural subnetwork (GCN) to a CNN to construct object-level relationships between superpixels, bringing richer contextual information and finer edge details. In these methods, superpixels are used in subnetworks

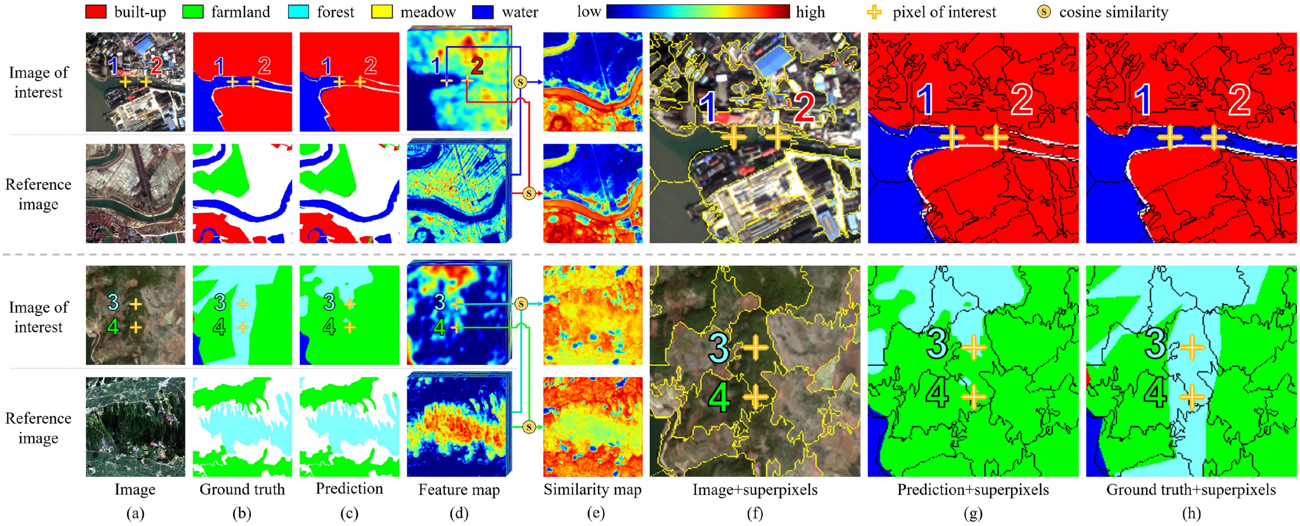


Fig. 1. Feature comparison between congeneric pixels with visual consistency but different predictions. In the similarity maps in column (e), the warmer the color, the more similar are the features. Column (e) reveals the feature difference between the pixels of interest, which provides an observation on that why misclassification happens. Columns (f)–(h) indicate that superpixels can group visually similar pixels into regions and have the potential to improve the land-cover classification results.

and play an indispensable role in the testing stage. Thus, both the generation and exploitation of superpixels in the prediction phase reduce the inference efficiency, and the poorly performed superpixels possibly create jagged edges in predictions during edge refinement. Meanwhile, given that limited semantic information is considered during superpixel segmentation, the superpixels obtained from complex remote sensing images using these methods are often unsatisfactory and adversely affect both the training and the testing stage. To overcome such weakness, we propose the high-quality semantic superpixels and adopt them only in the loss function to effectively guide the network training rather than in a subnetwork. As a result, our network is freed from reliance on superpixels in the prediction stage, accelerating the inference and preventing classification noises caused by low-quality superpixels.

III. Methodology

A. Our Observation and Motivation

In land-cover classification, most of existing methods mainly focus on tackling segmentation fragments caused by texture diversity of large ground objects [26], [36], [38], or reducing the misclassification resulting from the confusing appearance of different categories [12], [36], [37]. However, misclassification is also frequently observed in visually consistent regions that are intuitively easy to interpret, as illustrated in Fig. 1.

As shown by the cross marks in the first row of Fig. 1(c), pixels 1 and 2 are inside a region with a highly consistent texture reflecting water category, but pixel 2 is incorrectly predicted as a built-up area by a commonly used semantic segmentation network, i.e., DeepLabv3+ [33]. Similar cases can also be found in the third row of Fig. 1(c), where pixel 3 is also exceptionally misclassified. To explore the reason for the misclassifications, we select two reference images with good prediction results, as shown in the second and fourth rows of Fig. 1. We visualize the feature maps (segmentation feature maps) derived from the final convolution layer for all the input images, as shown in Fig. 1(d). We then generate the feature similarity maps for the pixels of interest, as shown in Fig. 1(e), by computing a cosine similarity between the pixels of interest (pixels 1, 2, 3, and 4) and each pixel in the feature maps of the corresponding reference image. The computation of cosine similarity can be seen in [9]. In the similarity maps, the warmer the color, the higher the feature similarity. For example, in the first row of Fig. 1(e), the warm color in the bottom of the similarity map indicates that the feature on pixel 1 has a high similarity with those extracted from the pixels of the built-up area. However, in the second row of Fig. 1(e), the similarity map shows a warmer color in the bottom part than that in the first row, indicating a greater similarity to the features extracted from the built-up area. As a result, the network is confused by the extracted features, leading to unstable prediction. From the similarity map in the fourth row of Fig. 1(e), the situation for pixel 4 is similar to that of pixel 2. The extracted feature on pixel 4 has a very high similarity with those extracted from the farmland and is thus incorrectly predicted as farmland.

Compared with the CNN-based segmentation, the superpixel segmentation relying on only low-level features is more capable of grouping visually similar pixels into regions, as depicted in Fig. 1(f)–(h). In particular, from the overlapped results in Fig. 1(g) and (h), if the feature consistency can be preserved for pixels inside a superpixel (visually consistent region), such as pixels 1 and 2, and also 3 and 4, the prediction correctness likewise improves. Previous works have tried to incorporate superpixels into semantic segmentation networks to enhance their regional feature learning ability. However, Fig. 1(h) shows that the boundaries of superpixels and ground-truth classes may have a misalignment, which can mislead the network training. More importantly, the commonly used CE loss is unable to provide the essential training guidance for learning superpixels.

The CE loss is defined as follows:

$$L_{\mathrm{CE}}(y,\hat{y}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log\hat{y}_{n,c} \tag{1}$$

where N denotes the number of pixels and C is the number of classes, $y \in \{0,1\}$ and $\hat{y} \in [0,1]$ represent the ground-truth label and prediction probability, respectively. If the nth pixel belongs to the cth class, then $y_{n,c}$ equals to 1 and otherwise 0. $\hat{y}_{n,c}$ is the probability of predicting the nth pixel as the cth category.

The aforementioned formulation clearly shows that the CE loss is designed and computed upon individual pixels without considering the relations between pixels inside a region (e.g., inside a superpixel). The relations between object regions are ignored. The CE loss encourages a high prediction score of the ground-truth category for every pixel, implying that it separately teaches the network to learn features for each individual pixel, which lacks training guidance for the extraction of contextual information [18]. In the absence of contextual constraints, networks are likely to encode very detailed but insignificant high-frequency signals into high-dimensional feature space, causing the locally similar appearances to produce very different feature representations [55]. In Fig. 1, the features generated from the pixels of interest provide such evidence.

According to the aforementioned analysis, to fully combine the merits of superpixel segmentation and pixel-wise semantic segmentation, it is essential to develop a specific region-level loss to guide the exploitation of rich contextual information implied in superpixels. According to Fig. 4, if the generated superpixels are semantic-aware (each superpixel belongs to one known class), then at least three kinds of contextual constraints can be excavated from the segmentation feature maps:

1) pixels inside a superpixel should have similar feature representation to guarantee the *intra-class consistency in a local region*;
2) superpixels belonging to the same class should have similar feature representations to guarantee the *intraclass consistency across a large region*;
3) superpixels belonging to different classes should have diverse global feature representations to preserve *interclass discrimination*.

With these considerations, we are ready to develop an SARI loss function that can reduce the feature variance between pixels inside a superpixel, or between superpixels of a same class, while increasing the feature distance between those of different classes. Moreover, to assist in training with SARI loss, the superpixels can be generated for only the training images and their semantics can be obtained from ground-truth labels. In the following, we detail the generation of semantic superpixels and the formulation of SARI loss.

B. Semantic Superpixel Pregeneration

As mentioned in Section II-B, existing methods deploy an extra superpixel segmentation subnetwork for generating superpixels directly from the input images. As a result, the superpixel segmentation has to be repeatedly performed during training and testing, bringing a heavy computation burden. Moreover, the generated superpixels only rely on low-level visual cues without considering the semantics offered by ground-truth labels, which incurs the following two problems:

1) the superpixel boundary can deviate far from the object class boundary;
2) adjacent objects of different classes with similar appearances are probably mixed in one superpixel.

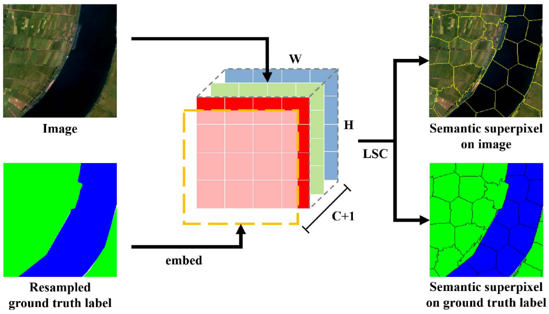


Image
Resampled ground truth label
W
H
C+1
LSC
embed
Semantic superpixel on image
Semantic superpixel on ground truth label

Fig. 2. Pregeneration of semantic superpixels. The ground-truth label is embedded as a new channel into the original image, and then, segmented into semantic superpixels by the LSC algorithm [45]. The generated semantic superpixels have high interior visual and semantic consistency, and a strong adherence to category boundaries.

These contradictions between superpixel segmentation and semantic segmentation can mislead the training of semantic segmentation networks, resulting in incoherent predictions.

To address the aforementioned problems, we propose to pregenerate semantic superpixels from training images with the assistance of corresponding semantic annotations, instead of redundantly producing low-quality general superpixels with an extra subnetwork. The generated semantic superpixels can maintain high consistency with the semantic labels, and are thus more compatible with the semantic segmentation task for improving the network training. Fig. 2 shows the pregeneration of semantic superpixels.

As shown in Fig. 2, given a pair of training image and its ground-truth label, we first append the latter to its corresponding image as a new channel to integrate semantic information. Then, the enhanced images are segmented into semantic superpixels by the classic LSC [45] algorithm. Specifically, to better separate different categories, the class difference in the ground-truth labels is magnified by resampling the labels to values ranging 0–255 with unified intervals before insertion into the images. For the same purpose, the resampled one-hot encoding labels can serve as an alternative scheme, and the encoding presents a superiority of an identical Euclidean distance between any two classes. However, the one-hot encoding introduces excessive dimensions into the image and significantly impairs the visual consistency of the generated superpixels. Thus, this study adopts the one-dimension resampled labels to enhance the semantic information. Some examples of semantic superpixels are illustrated in Fig. 3.

Compared with the general superpixels in Fig. 3(c), the semantic superpixels displayed in Fig. 3(b) manifest a high interior semantic and visual coherence, and a strong adherence

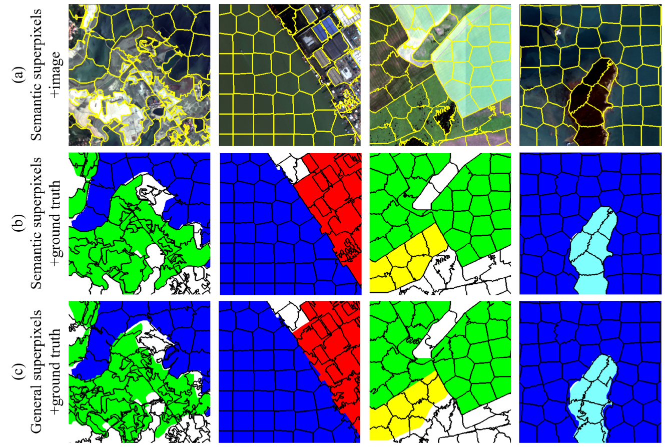

(a) Semantic superpixels +image
(b) Semantic superpixels +ground truth
(c) General superpixels +ground truth

Fig. 3. Comparison between semantic superpixels and general superpixels. Semantic superpixels in (b) present higher interior semantic consistency and stronger adherence to object boundaries than the general superpixels in (c).

to category boundaries. To effectively unify the superpixel supervision with semantic segmentation tasks, it is essential to generate high-quality superpixels that can satisfactorily assist in the training of a general semantic segmentation network with our SARI loss.

C. *SARI Loss*

In this section, we elaborate our proposed SARI loss that fully exploits the contextual constraints inside and between the semantic superpixels without introducing extra computational burden in the testing stage. Fig. 4 summarizes the principles of the SARI loss, which imposes additional constraints on the learned feature maps with three parts: a superpixel variance loss L_{var}, an intraclass similarity loss L_{intra} and an interclass distance loss L_{inter}. L_{var} strengthens region consistency by reducing feature variance between pixels inside each semantic superpixel, while L_{intra} and L_{inter} further use the long-range relationship between superpixels. Specifically, L_{intra} encourages the superpixels within the same class to have similar average features, and L_{inter} enlarges the discrepancy between the average features of superpixels from different classes.

1) Superpixel Variance Loss: Taking advantage of the semantic superpixels that group visually consistent pixels, we design a superpixel variance loss L_{var} to narrow the gap between features of regionally consistent pixels. This loss is formulated as follows:

$$L_{\text{var}} = \sqrt{\sum\nolimits_{s=1}^{S}\left[\frac{1}{N_s}\sum\nolimits_{d=1}^{D}\sum\nolimits_{n=1}^{N_s}\left(F_{s,n}^{d}-\overline{F}_{s}^{d}\right)^{2}\right]} \tag{2}$$

where $F_{s,n}^{d}$ and $\overline{F}_{s}^{d}$ refer to the individual and average feature values of pixels in the sth semantic superpixel on the dth channel, respectively. S, N_s, and D denote the number of superpixels, pixels in the sth superpixel, and channels of the segmentation feature map $\mathbf{F}$, respectively. Notably, the feature variance of each channel is calculated independently to avoid feature mixing across channels.

By minimizing L_{var}, the features inside each superpixel are forced to be similar, which enhances regional feature consistency and restrains the encoding of useless high-frequency signals from images into the feature space. In practice, $\overline{F}_{s}^{d}$ is excluded from the gradient descent during training to prevent gradient vanishing.

2) Intraclass Similarity Loss: Although regional consistency enforced by semantic superpixels enhances semantic segmentation within a superpixel, long-range contextual information is also indispensable to make accurate predictions for land covers with complex textures and multiple superpixels. However, the local nature of convolution hinders the capture of contextual information across large regions. Moreover, the appearance of land-cover belonging to the same class can sometimes highly differ, causing difficulties to extract representative features for each category.

To address the issue, we propose an intraclass similarity loss to impose longer-range constraints on features. The intraclass similarity loss encourages higher feature similarities between semantic superpixels of the same class. We define this loss as

$$L_{\text{intra}} = \frac{1}{D}\sum_{d=1}^{D}\sum_{i=1}^{C}\sum_{s_i}\sum_{s'_i\neq s_i}\left|\overline{F}_{s_i}^{d}-\overline{F}_{s'_i}^{d}\right| \tag{3}$$

where $\overline{F}_{s_i}^{d}$ and $\overline{F}_{s'_i}^{d}$ denote average pixel feature values of two semantic superpixels from the same ith category on the dth channel. D and C refer to the numbers of segmentation feature map channels and classification categories, respectively.

By minimizing L_{intra}, the features of superpixels belonging to the same class are forced to be similar and thus enhancing long-range intraclass consistency.

3) Interclass Distance Loss: In addition to intraclass similarity, weak interclass discrimination of different superpixels is also a major challenge of land-cover classification. Misclassification can be caused by the common issue that pixels from different categories have very similar appearances (e.g., forest and farmland), as mentioned in Section III-A.

To tackle this problem, we design an interclass distance loss to widen the difference of superpixels from different classes. This loss is defined as follows:

$$L_{\text{inter}} = -\log\left(\frac{1}{D}\sum_{d=1}^{D}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\sum_{s_i}\sum_{s_j}\left|\overline{F}_{s_i}^{d}-\overline{F}_{s_j}^{d}\right|\right) \tag{4}$$

where $\overline{F}_{s_i}^{d}$ and $\overline{F}_{s_j}^{d}$ represent the feature values of semantic superpixels that belong to the ith and jth categories on the dth channel, respectively. Each superpixel feature value is derived from the average one of pixels inside each superpixel. D and C are the numbers of feature map channels and classification categories, respectively.

As revealed in (4), minimizing L_{inter} encourages large channel-wise feature discrepancies between different classes, which largely alleviates the confusion of categories with similar appearances. Moreover, considering the inevitably imperfect annotations in the training subset, L_{inter} can also mitigate the

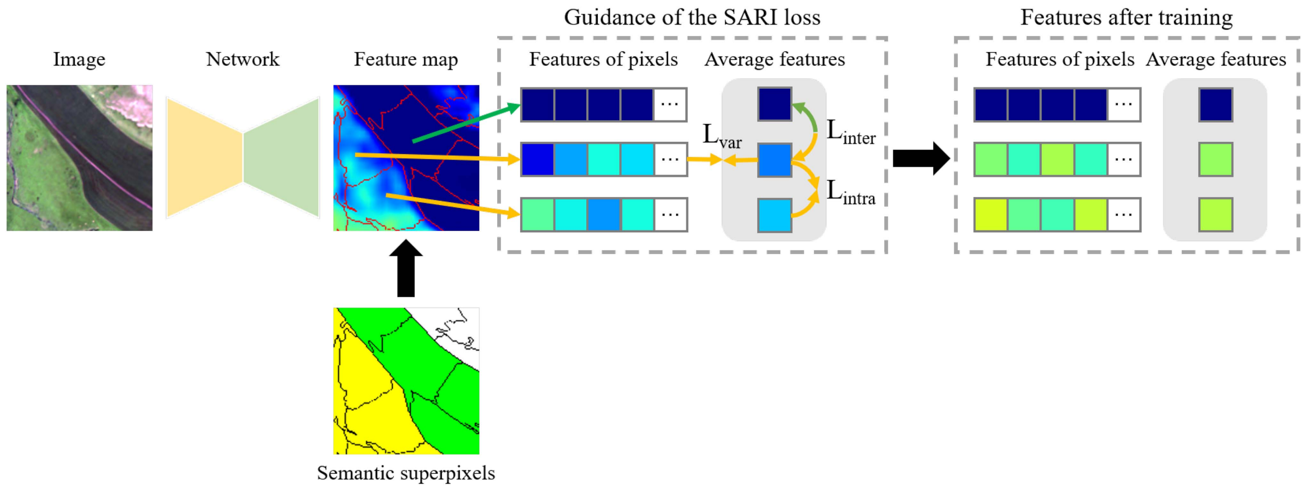


Fig. 4. Components of the SARI loss and their different functions.

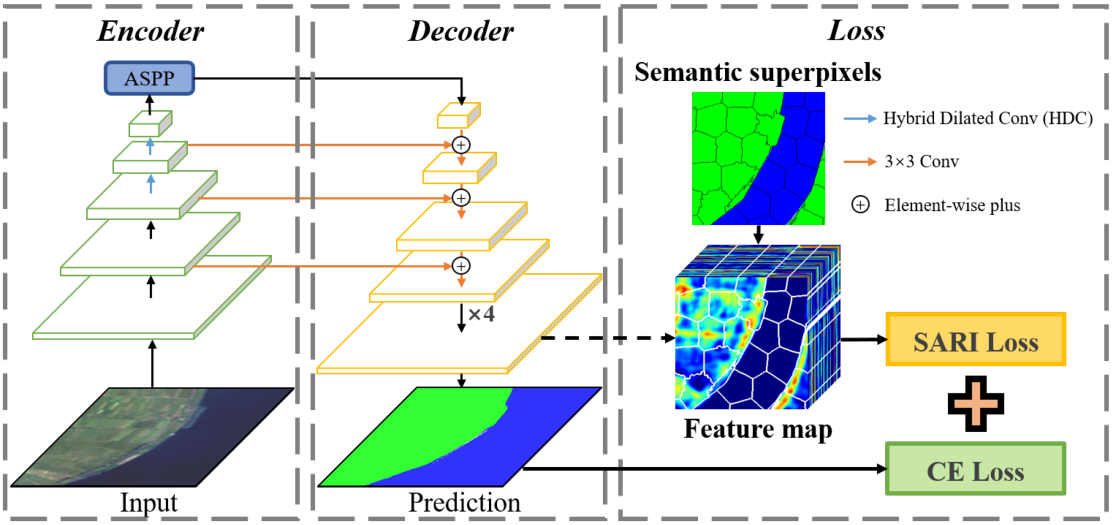


Fig. 5. Overall architecture of the SPSNet.

mixing effect of L_{var} on incorrectly annotated pixels within a superpixel.

4) Final Loss Function: Finally, the overall loss function combines the aforementioned SARI loss (superpixel variance loss L_{var}, intraclass similarity loss L_{intra}, and interclass distance loss L_{inter}) with the original CE loss [see (1)] for classification:

$$L = L_{\text{CE}} + \lambda_1 L_{\text{var}} + \lambda_2 L_{\text{inter}} + \lambda_3 L_{\text{intra}} \tag{5}$$

where weights λ_1, λ_2, and λ_3 are set to 0.1, 0.1, and 50 in practice.

Benefiting from both the intra- and intersuperpixel constraints, our SARI loss can fully exploit both regional and long-range relationships for stronger intraclass coherence and interclass discrimination, and thus, benefit the inference in land-cover classification.

D. Network Architecture

To verify the effectiveness of the proposed SARI loss for superpixel supervised land-cover classification, we implement it with a commonly used semantic segmentation network, which is similar to the classical DeepLabv3+ [33]. For convenience, we term the network as superpixel supervised network (SPSNet) and Fig. 5 shows its network architecture. The basic network is deployed as an encoder–decoder structure with embedded skip connections to reuse the low-level encoded features. In the encoder of many classical networks, e.g., DeepLabv3+, dilated convolutions are widely adopted to expand the receptive field for observing large objects. However, the holes in standard dilated convolutions cause disconnection and impairment of relations between neighbouring features, which incur misclassification or "gridding effect" for large objects [56]. The problem is especially severe in classifying land covers, most of which usually span large spatial areas (can be considered as extremely large objects in general semantic segmentation). To alleviate this problem, we replace the standard dilated convolution with hybrid dilated convolution (HDC) [56] in the backbone of SPSNet, as indicated by the blue lines in Fig. 5.

SPSNet takes an optical remote sensing image as input and outputs a semantic prediction map describing the pixel-wise land-cover classification result. In addition, the pregenerated semantic superpixels are input as a guidance only in the training stage. The feature maps derived from the last layer of the decoder (segmentation feature maps) are used for computing the SARI loss, which then provides guidance for superpixel-enhanced semantic learning of land-cover categories. By comparison, the traditional CE loss offers supervision for the final semantic prediction. Compared with existing superpixel-enhanced semantic segmentation networks, the architecture of SPSNet is free from the influence of low-quality superpixel segmentation during training and testing. Moreover, SPSNet has no structures tailored for the proposed SARI loss, which can be easily applied to other networks for superpixel-enhanced semantic segmentation.

IV. Experiments

To validate the effectiveness of the proposed SPSNet and SARI loss, we conducted extensive experiments on two widely used remote sensing land-cover datasets, that is, GID [5] and DeepGlobe Land Cover Classification Challenge Dataset [19]. In the following sections, we detail the experimental settings, results, and analysis.

A. Dataset and Implementation Details

1) Datasets: The following two widely used remote sensing land-cover datasets are used.

a) Gaofen Image dataset (GID): The GID [5] is a large-scale benchmark dataset for land-cover classification evaluation. GID consists of 150 images with a spatial resolution of 4 m acquired by Gaofen-2 (GF-2) satellite. Each image contains 6800×7200 pixels covering 506 km^2, and four bands covering the spectral range of blue, green, red, and near-infrared. The dataset is well annotated by five categories, namely, built-up, farmland, forest, meadow, and water. Pixels belonging to other categories or clutter regions are labeled as background and excluded for both training and evaluation. Following previous literature [57], we crop the images into patches of 1024×1024 pixels without overlaps. After excluding images without valid annotation, fivefold cross validation is applied for training and accuracy assessment [57]. In detail, the GID dataset is partitioned into five equally sized subsets, for which five models are trained on different combinations. For each model, four subsets are used for training and the remainder is used for evaluation.

b) DeepGlobe dataset: DeepGlobe [19] is an RGB dataset consisting of 803 high-resolution satellite images. Each image contains 2448×2448 pixels labeled with seven land-cover categories, namely, urban land, agriculture land, rangeland, forest land, water, barren land, and unknown. The last category is not considered in the assessment but is learned by our network during experiments. We adopt the same train/validation/test split as [58] and [59] with 454, 207, and 142 images for training, validation, and testing, respectively. In the training subset, images are split into 768×768 pixels with overlap.

Both GID and DeepGlobe datasets cover large areas with various geographic distributions. GID and DeepGlobe datasets collect images from both urban and rural areas, and cover areas of over 5000 km^2 and 1716.9 km^2, respectively. The rich ground object diversities in spectral responses and morphological structures present challenges in the feature generalization capacity of networks.

2) Implementation Details: The proposed SPSNet is implemented using the Pytorch framework and all the models are trained and evaluated with two NVIDIA GTX 2080Ti GPUs, each with a training batch size of four for both datasets. The number of input channels is equal to the count of the image bands, that is, four for the GID and three for the DeepGlobe dataset. The base learning rate is set to 3e-4, and is then updated with a cosine annealing policy for each batch following [60]. The learning rate restarts to the base at 5, 15, 35, and 75 epochs. In each restart cycle, the learning rate drops from the base to 0 following the cosine curve. The models are trained by 225 and 125 epochs on the GID and DeepGlobe dataset, respectively, by using the adaptive gradient optimizer AdamW with a momentum of 0.9 and a weight decay of 5e-4. In the training stage, commonly used data augmentation techniques, including random vertical and horizontal flipping and anticlockwise rotating, are applied on both datasets. In addition, a random cropping with 512×512 size is applied on the GID dataset. In the testing stage, the images are cropped into patches with overlap for inference and the results are spliced by averaging the predicted probability maps on the overlapped regions. Specifically, the images are cropped into patches of 512×512 with an overlap width of 256 pixels for GID and 768×768 and 208 pixels for DeepGlobe. The multi-scale inference is applied for the GID dataset by averaging probability maps predicted at multiple scales, including 0.75, 1.0 and 1.25. The test time augmentation (TTA) of flipping and rotating is applied for DeepGlobe.

The semantic superpixels are pregenerated by LSC [45] with OpenCV implementation. Ground-truth labels are appended as a new channel to the corresponding training images as mentioned in Section III-B. Empirically, the mean size of the superpixels is set to 50×50. Superpixels that are smaller than half of the mean size are merged into adjacent ones through automatic postprocessing to control their appropriate number and sizes. Afterwards, the semantic superpixels are segmented by ground-truth annotations into semantic patches to ensure interior class uniformity.

3) Evaluation Metrics: Following the standard evaluation protocol in the task of land-cover classification, the performances of our SPSNet are evaluated by mean intersection over union (mIoU) and pixel accuracy (PA) on different datasets. The metrics are formulated as follows:

$$\mathrm{PA} = \frac{\sum_{i=1}^{C} x_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} x_{ij}};$$

TABLE I
ABLATION STUDY ON GID DATASET FOR THE DIFFERENT MODULES

Method	CE loss	HDC	L_{var}	L_{intra}	L_{inter}	PA(%)	mIoU(%)
Baseline	✓					97.22±0.40	92.63±1.38
SPSNet	✓	✓				97.55±0.23	93.88±1.03
SPSNet	✓	✓	✓			98.53±0.12	96.40±0.32
SPSNet	✓	✓	✓	✓		98.58±0.11	96.51±0.13
SPSNet	✓	✓			✓	98.54±0.14	96.07±0.61
SPSNet	✓	✓	✓	✓	✓	**98.75±0.14**	**97.11±0.25**

$$\text{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{x_{ii}}{\sum_{j=1}^{C}(x_{ij}+x_{ji})-x_{ii}}, \tag{6}$$

where C represents the category numbers and x_{ij} denotes the number of pixels that belong to class i while are predicted as class j. Especially, x_{ii} means the number of true positive pixels of class i.

For the GID dataset, models obtained from the fivefold cross validation are evaluated by both mIoU and PA metrics, following [57]. For the DeepGlobe dataset, models are assessed through mIoU on the test subset as that of [58] and [59].

B. Ablation Study

To testify the effectiveness of the proposed SARI loss, and the necessity of using HDC block, we carry out ablation experiments on the GID dataset. DeepLabv3+ [33] trained with the standard CE loss is chosen as the baseline network and the ResNet-101 [61] is the chosen backbone. Table I lists the quantitative results of applying different network configurations.

As illustrated in Table I, by replacing standard dilated convolution with the HDC block, SPSNet yields a mIoU of around 93.88%, outperforming the baseline network (i.e., Deeplabv3+) by 1.25%. The possible reason is that the use of HDC block mitigates the "gridding" problem of dilated convolution, and thus, prevents the large-area land covers frequently being segmented into pieces. By applying the superpixel variance loss L_{var}, SPSNet yields a mIoU score of around 96.40%, surpassing the SPSNet trained with the CE loss alone by 2.52%. Moreover, the mIoU floating range also witnesses an apparent decrease when training SPSNet with L_{var}. The results reveal that enhancing regional feature consistency (inside superpixels) by L_{var} boosts not only the classification performance but also the training stability. After joining the intraclass similarity loss L_{intra}, the performance of SPSNet is further promoted with a mIoU score of 96.51%, indicating that the L_{var} and L_{intra} can cooperate to improve the feature learning ability and training stability of SPSNet. From Table I, incorporating the interclass distance loss L_{inter} alone with the CE loss obtains a performance gain of 2.19% in mIoU compared with SPSNet merely trained with the CE loss. The improvement can be attributed to that by training with L_{inter}, SPSNet learns to enlarge the feature discrepancy between semantic superpixels of different classes, which enhances the interclass discrimination ability, and thereby, reduces the misclassification of the confusing land covers. Furthermore, the complete SARI loss that integrates the L_{var}, L_{intra}, and L_{inter} achieves a highest mIoU score of 97.11% and a satisfactory mIoU floating range (0.25%), demonstrating the joint effect of the three subparts of the SARI loss. Figs. 6 and 7 show the visualization results for qualitative analysis of the HDC block and the SARI loss.

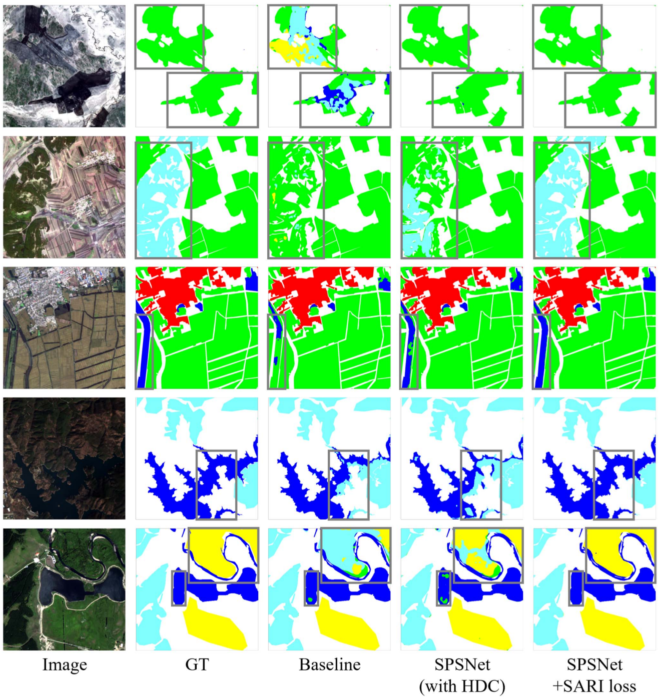


Fig. 6. Visualized results on GID dataset. The label includes five categories: built-up (red), farmland (green), forest (cyan), meadow (yellow), and water (blue). White pixels represent the background and are excluded in the accuracy assessment.

Fig. 6 first shows the classification results of the baseline network, the SPSNet with HDC block alone, and the complete SPSNet (with both HDC and SARI loss). As mentioned in Section III-D, in general semantic segmentation models, the dilated convolution is widely used to extend the network's receptive field for observing large objects. However, land-cover classification is different from the general semantic segmentation, as land covers usually span an extremely large spatial extent and have very complex geometric structures. In this case, the holes in standard dilated convolutions can cause a "gridding effect" that segments a complete land cover region into several pieces, or severe misclassification that recognizes a land cover region as a wrong category. The evidence can be found in the visualization results of the baseline network. For example, the

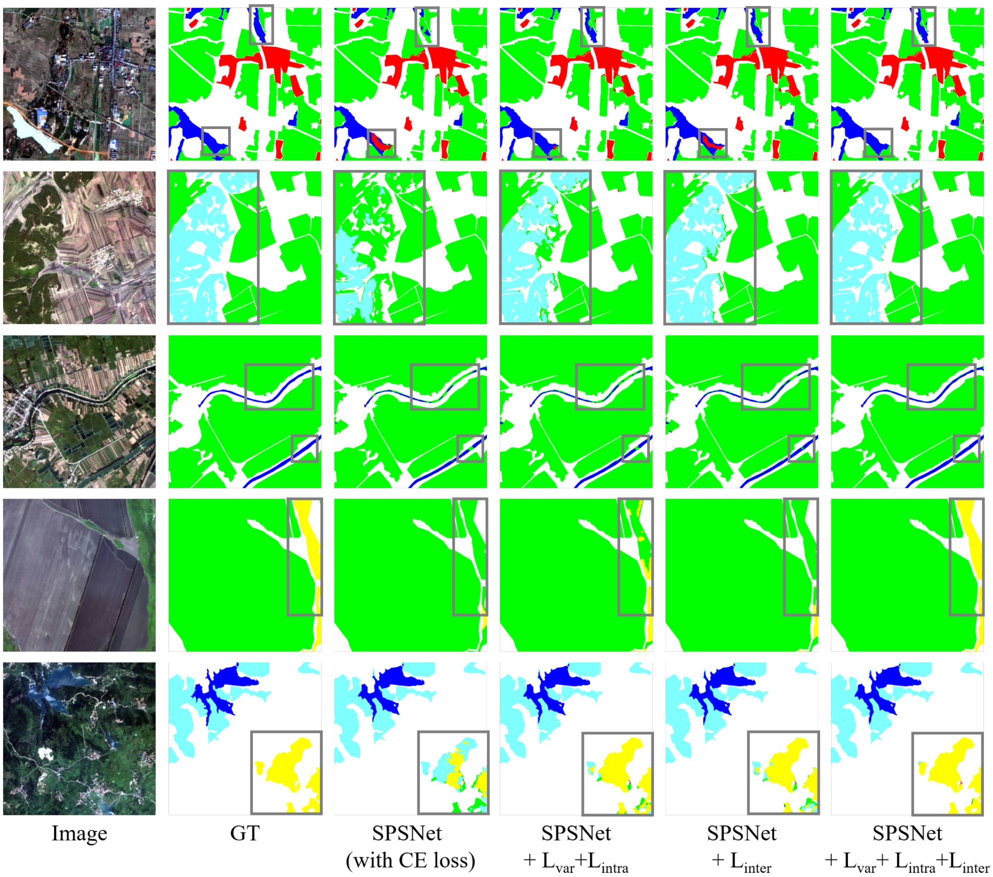


Image
GT
SPSNet (with CE loss)
SPSNet + L_{var}+L_{intra}
SPSNet + L_{inter}
SPSNet + L_{var}+ L_{intra}+L_{inter}

Fig. 7. Visualization results from the ablation study of the SARI loss on GID dataset. The label includes five categories: built-up (red), farmland (green), forest (cyan), meadow (yellow), and water (blue).

first and third rows clearly show the fragmented segmentation results of the farmland and water category, whereas the second row has a large area misclassification. Considering the specific challenge of land-cover classification, we add the HDC block into the baseline network to overcome the limitations of standard dilated convolution. From the results in the fourth column of Fig. 6, applying the HDC block alone can alleviate the "gridding effect" and the large area misclassification. The reason is that the HDC block is more capable of building continuous relations between neighboring features that are disconnected by standard dilated convolution. Thus, the basic feature learning ability of the network improves.

However, misclassification can still be found in regions with complex textures, as highlighted by the rectangles in the fourth column of Fig. 6. As a land-cover region usually contains multiple ground objects, which may show different visual appearances. Land-cover regions of different categories contrarily may have similar visual appearances (e.g., farmland and forest). The use of only HDC is inadequate to exploit the homogeneity within a land-cover region and heterogeneity between semantically different regions. The results in the last column of Fig. 6 convincingly demonstrate the effectiveness of the proposed SARI loss in improving the classification consistency of different land covers. By training with the SARI loss, SPSNet can fully utilize the context constraints within and between superpixels, thereby enhancing the regional feature learning ability for discriminating the confusing land covers.

In Fig. 7, the results of SPSNet trained with different sub-parts of the SARI loss are also visualized to further analyze the efficacy of the method. In the third column, the misclassfication can be frequently found in the results of basic SPSNet that is trained without the SARI loss. In the fourth column, by introducing the superpixel variance loss L_{var} and the intraclass similarity loss L_{intra}, the segmentation consistency inside different land cover regions significantly improves, especially for the forest and meadow as highlighted by the rectangle in the second and fifth rows. The reason is that in these land cover regions, the texture contains insignificant high-frequency details. The basic SPSNet trained with the CE loss tends to encode these details into high-dimension feature space and generate very different representations from the surrounding regions, which confuse the final prediction of the classifier. Applying the superpixel-enhanced training with L_{var} and L_{intra} can increase the consistency of feature representations inside and between homogeneous superpixels, which facilitates the network to correctly recognize the land covers within a same category. In the fifth column of Fig. 7, the interclass distance loss L_{inter} alone also reduces the misclassifications of the basic SPSNet, suggesting that enlarging the feature discrepancy between heterogeneous superpixels is another means to improve the discrimination

TABLE II
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON GID DATASET

Method	PA(%)	mIoU(%)
FCN-8s [7]	93.87±0.59	83.84±1.98
U-Net [28]	94.76±0.48	73.06±3.07
DeepLabv3 [62]	96.54±0.16	92.03±0.68
PSPNet [32]	96.76±0.14	92.24±0.54
DeepLabv3+ [33]	97.22±0.40	92.63±1.38
RSNet(GID) [57]	97.19±0.11	93.54±0.32
Trained with SARI loss		
FCN-8s	97.12±0.20	92.73±0.44
U-Net	98.27±0.18	95.02±1.13
DeepLabv3	97.19±0.25	93.23±0.22
PSPNet	97.73±0.14	94.36±0.34
DeepLabv3+	98.10±0.34	95.17±0.83
SPSNet(ours)	**98.75±0.14**	**97.11±0.25**

TABLE III
QUANTITATIVE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON DEEPGLOBE DATASET

Method	mIoU(%)
ICNet [63]	40.20
U-Net [28]	50.11
PSPNet [32]	56.60
FCN-8s [7]	62.43
SegNet [64]	68.40
DeepLab v3+ [33]	69.69
FPN [65]	70.98
FPN+DenseCRF [66]	70.36
FPN+DGF [67]	70.38
GLNet [58]	71.60
FPN+PointRend [68]	71.78
MagNet [59]	72.10
SPSNet(ours)	**73.99**

ability of the network. Compared with the classical CE loss, L_{var}, L_{intra}, and L_{inter} can provide training guidance for the extraction of different contextual information with semantic superpixels. As a result, each of these losses brings benefits to the regional feature learning and thus improves the segmentation of large-area land covers. However, misclassifications can still be found in the fourth and fifth columns, indicating that each part of the SARI loss alone cannot offer enough training guidance for comprehensively learning the contextual information within homogeneous superpixels and between heterogeneous superpixels. The SPSNet trained with the complete SARI loss (combining all of L_{var}, L_{intra}, and L_{inter}) shows a stronger learning ability for different land cover categories, which achieves the best segmentation quality for different input images.

C. Comparison With State-of-the-Art Methods

In this section, the proposed SPSNet was compared with other state-of-the-art methods on two benchmark land-cover classification evaluation datasets, GID [5] and DeepGlobe [19].

1) Results on GID Dataset: For a comprehensive evaluation of the proposed SPSNet, we compare it with several classical methods on the GID dataset. Moreover, we also trained the different comparison methods (with available codes) with our SARI loss to further testify its effectiveness and applicability. The quantitative results are reported in Table II.

As shown in Table II, all the comparison methods seem to yield acceptable results in terms of PA but manifest a relatively unstable performance on the mIoU metric, especially for the FCN-8s and U-Net. According to the definition, PA reveals the overall classification precision of pixels regardless of object classes, and a high PA score can be obtained if large objects that occupy most image pixels are correctly classified. In contrast, mIoU presents a more comprehensive performance assessment for multiclass semantic segmentation by averaging the IoU score of every class, which reveals both the classification precision and the recall ratio of target objects. Therefore, in land-cover classification, mIoU can better reveal the learning ability of a network than PA. From the comparison results in Table II, our proposed SPSNet achieves the top performances in terms of both PA and mIoU. SPSNet outperforms the existing methods by a large margin, which improves the previous state-of-the-art method (i.e., RSNet) by 3.57% in mIoU. The comparison results indicate that enhancing the regional feature learning with our proposed SARI loss is effective in improving the classification consistency of different land-cover categories. In Table II, results also demonstrate the effectiveness of SARI loss as it improves the performances of all the comparison methods. Results of PSPNet, U-Net, and DeepLabv3+ outperform RSNet in both PA and mIoU by replacing CE loss with SARI loss. The interclass and intraclass constraints imposed by SARI loss enhance the learning ability for different categories and improve performances over all classes.

2) Results on DeepGlobe Dataset: We carry out experiments on the DeepGlobe land cover classification dataset to further verify the effectiveness of the SPSNet and our SARI loss. A number of methods with different architectures are chosen for a comprehensive comparison, which include the following:

1) typical networks designed for general semantic segmentation, including ICNet [63], U-Net [28], PSPNet [32], FCN [7], SegNet [64], DeepLabv3+ [33], and FPN [65] with different settings;
2) state-of-the-art networks especially designed for semantic segmentation of high-resolution images, including GLNet [58] and MagNet [59].

Table III lists the numeric results of different models.

The overall situation in Table III shows that those methods designed for high-resolution images yield better performance than general semantic segmentation networks. As land covers usually span large spatial extent with very irregular boundaries, precise classification requires rich contextual information to preserve intraclass consistency while spatial details for guaranteeing the interclass discrimination. Among all the general semantic segmentation methods, the FPN combines contextually rich features and spatially detailed features to enhance the multilevel feature learning, which achieves a comparable performance with the networks designed for high-resolution images. By integrating the local fine details into global contexts in a more effective way, GLNet and MagNet further improve the classification performances of the general segmentation models by maximally

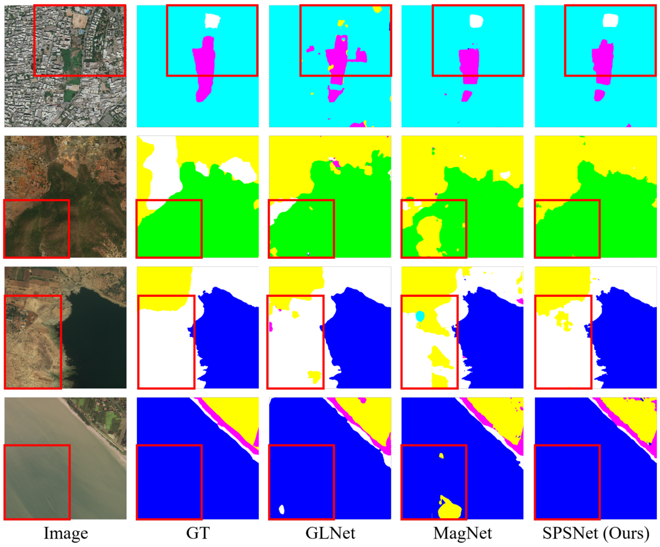


Fig. 8. Qualitative comparison results on DeepGlobe dataset. The label includes six categories: urban land (cyan), agriculture land (yellow), rangeland (magenta), forest land (green), water (blue), and barren land (white).

31.9% in mIoU. However, all these methods mainly focus on designing new network architectures to strengthen the extraction and fusion of multi-level features. Most of them are trained only with the classical CE loss, which lacks training guidance for effectively learning the local and long-range relationships between features. Basing on the semantic-aware superpixel segmentation, our proposed SARI loss can teach a network to fully exploit the contextual constraints inside or between superpixels. SARI loss works in a pull-and-push manner, which pulls regional features within a same class closer (increases feature similarity) while pushing regional features of different classes farther (enlarging feature discrepancy). As a result, our SPSNet trained with the SARI loss yields the top performance with a mIoU of 73.99%, revealing that SPSNet is more capable of addressing the challenges specific to land-cover classification in high-resolution remote sensing images. For a better visual inspection, we also visualize the classification results of our SPSNet and two recent state-of-the-art methods in Fig. 8.

In Fig. 8, we select four images with different texture complexity as examples to analyze the regional feature learning abilities of different networks. In the first row, textures of urban, range, and barren lands are confusing due to their high complexity. Consequently, GLNet delivers severely fragmented segmentation, while MagNet presents large-area misclassification (such as rangeland) as highlighted by the red rectangle. In the second and third rows, land covers (such as forest land and barren land) in the red rectangles have relatively simple and consistent texture/appearance. However, GLNet and MagNet still suffer from segmentation fragments and large-area misclassification. In the last row, the water region highlighted in the rectangle presents high visual consistency. It is surprising that GLNet and MagNet fail to produce coherent labeling results inside such region. By contrast, under the guidance of the SARI loss, our SPSNet can distinguish between these land-cover categories more correctly, and thus, obtains clean segmentation results. These visualized results further convincingly validate the effectiveness of our SARI loss.

V. Conclusion

This study presents an SARI loss to guide the effective learning of regional features with semantic superpixels for the accurate land-cover classification. To exert the power of the SARI loss, we also develop a SPSNet for land-cover classification. The quantitative and qualitative results in the ablation experiments show that the SARI loss can offer sufficient training guidance for the comprehensive learning of the contextual information within homogeneous superpixels and between heterogeneous superpixels, which significantly improves the segmentation consistency of different land-cover categories. Specifically, the independent ablation studies for the different parts of SARI loss verify that the superpixel variance loss L_{var} and the intraclass similarity loss L_{intra} can improve feature similarities between superpixels in the same category, while interclass distance loss L_{inter} can enlarge feature discrepancy between superpixels of different categories. The comparison with state-of-the-art methods on two benchmark land-cover classification datasets (i.e., GID and DeepGlobe) also convincingly demonstrates that the SARI loss can teach a model to effectively learn both regional and long-range relationships for coherent semantic labeling in land-cover classification. In the future work, we would like to implement SARI loss with more advanced networks for high-accuracy global-scale land-cover classification.

References

[1] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[2] H. Sheng, X. Chen, J. Su, R. Rajagopal, and A. Ng, "Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 60–61.

[3] M. M. Nielsen, "Remote sensing for urban planning and management: The use of window-independent context segmentation to extract urban features in stockholm," *Comput., Environ. Urban Syst.*, vol. 52, pp. 1–9, 2015.

[4] J. C. Tilton, W. T. Lawrence, and A. J. Plaza, "Utilizing hierarchical segmentation to generate water and snow masks to facilitate monitoring change with remotely sensed image data," *GISci. Remote Sens.*, vol. 43, no. 1, pp. 39–66, 2006.

[5] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[6] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2587–2600, Sep. 2006.

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

[8] 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.

[9] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[10] S. Gould, J. Zhao, X. He, and Y. Zhang, "Superpixel graph label transfer with learned distance metric," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 632–647.

[11] M. Volpi and V. Ferrari, "Structured prediction for urban scene semantic segmentation with geographic context," in *Proc. Joint Urban Remote Sens. Event*, 2015, pp. 1–4.

[12] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 140–152, 2020.

[13] Z. Xu, T. Ajanthan, and R. Hartley, "Refining semantic segmentation with superpixel by transparent initialization and sparse encoder," 2020, *arXiv:2010.04363*.

[14] S. Ouyang and Y. Li, "Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 13, no. 1, 2021, Art. no. 119.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[16] L. Huan, N. Xue, X. Zheng, W. He, J. Gong, and G.-S. Xia, "Unmixing convolutional features for crisp edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6602–6609, Oct. 2022.

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

[18] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region mutual information loss for semantic segmentation," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 11115–11125, 2019.

[19] I. Demir et al., "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 172–181.

[20] J. D. Paola and R. A. Schowengerdt, "A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 4, pp. 981–996, Jul. 1995.

[21] M. C. Hansen, R. S. DeFries, J. R. Townshend, and R. Sohlberg, "Global land cover classification at 1 km spatial resolution using a classification tree approach," *Int. J. Remote Sens.*, vol. 21, no. 6-7, pp. 1331–1364, 2000.

[22] C. Huang, L. Davis, and J. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, 2002.

[23] D. L. Civco, "Artificial neural networks for land-cover classification and mapping," *Int. J. Geographical Inf. Sci.*, vol. 7, no. 2, pp. 173–186, 1993.

[24] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 67, pp. 93–104, 2012.

[25] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sens. Environ.*, vol. 86, no. 4, pp. 554–565, 2003.

[26] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS J. Photogrammetry Remote Sens.*, vol. 132, pp. 48–60, 2017.

[27] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 149, pp. 188–199, 2019.

[28] 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.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

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

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[32] 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. 2881–2890.

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

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

[35] L. Shan and W. Wang, "DenseNet-based land cover classification network with deep fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[36] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022.

[37] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[38] 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. Photogrammetry Remote Sens.*, vol. 145, pp. 78–95, 2018.

[39] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "Ern: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1339.

[40] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[41] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 12, no. 4, 2020, Art. no. 701.

[42] Z. Xu, W. Zhang, T. Zhang, and J. Li, "HRCnet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 1, 2021, Art. no. 71.

[43] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 1, pp. 10–17.

[44] V. Jampani, D. Sun, M.-Y. Liu, M.-H. Yang, and J. Kautz, "Superpixel sampling networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 352–368.

[45] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1356–1363.

[46] L. Zhu et al., "Learning the superpixel in a non-iterative and lifelong manner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1225–1234.

[47] D. Stutz, A. Hermans, and B. Leibe, "Superpixels: An evaluation of the state-of-the-art," *Comput. Vis. Image Understanding*, vol. 166, pp. 1–27, 2018.

[48] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li, "Object detection by labeling superpixels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5107–5116.

[49] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1323–1330.

[50] J. Lu, H. Yang, D. Min, and M. N. Do, "Patch match filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1854–1861.

[51] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," in *Proc. ACM SIGGRAPH Papers*, 2005, pp. 577–584.

[52] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3376–3385.

[53] R. Gadde, V. Jampani, M. Kiefel, D. Kappler, and P. V. Gehler, "Superpixel convolutional networks using bilateral inceptions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[54] W.-C. Tu et al., "Learning superpixels with segmentation-aware affinity loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 568–576.

[55] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8684–8694.

[56] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.

[57] J. Wang, Y. Zhong, Z. Zheng, A. Ma, and L. Zhang, "RSNet: The search for remote sensing deep neural networks in recognition tasks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2520–2534, Mar. 2021.

[58] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8924–8933.

[59] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16755–16764.

[60] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[61] 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.

[62] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv:1706.05587*.

[63] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, “ICNet for real-time semantic segmentation on high-resolution images,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 405–420.

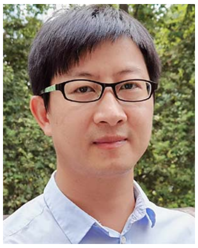
[64] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SEGNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[65] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[66] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected CRFs with Gaussian edge potentials,” *Adv. Neural Inf. Process. Syst.*, vol. 24, pp. 109–117, 2011.

[67] H. Wu, S. Zheng, J. Zhang, and K. Huang, “Fast end-to-end trainable guided filter,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1838–1847.

[68] A. Kirillov, Y. Wu, K. He, and R. Girshick, “Pointrend: Image segmentation as rendering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9799–9808.



Xianwei Zheng (Member, IEEE) received the M.S. and Ph.D. degrees in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010 and 2015, respectively .

He is currently working as an Associate Professor in computer vision and 3-D geographical information science with Wuhan University. His research interests include indoor and outdoor scene parsing, 3-D computer vision and reconstruction, and geovisualization.

Qiyuan Ma received the B.S. degree in geographic information science from the School of Geography, Nanjing Normal University, Nanjing, China, in 2020. He is currently working toward the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include land-cover classification and scene parsing.

Linxi Huan received the B.S. degree in mathematics and applied mathematics in 2018 from the School of Mathematics and Statistics, Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

Her research interests include machine learning, scene parsing, and 3-D reconstruction.

Xiao Xie received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She has been a Research Fellow with the Department of Cartography, Technical University of Munich, Munich, Germany, from 2014 to 2016 and now serving as a Senior Engineer with the Key Lab of Environmental Computing and Sustainability, Liaoning province, as well as an Assistant Professor in urban and environmental computation with the Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China. She is also a Postdoctoral Researcher with the School of Geodesy and Geomatics, Wuhan University. Her research interests include 3-D geographical information science and smart cities.

Hanjiang Xiong received the B.S. degree in photogrammetry and remote sensing from the School of Remote Sensing and Engineering, Wuhan University of Surveying and Mapping, Wuhan, China, in 1995, and the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, in 2002.

He was working as a Visiting Scholar with the Queensland University of Technology for three months in 2011. He is currently working as a Full Professor in 3-D geographical information science (GIS), Wuhan University. His current research interests include geospatial data management, 3-D visualization, augmented reality, and indoor and outdoor GIS.

Jianya Gong received the Ph.D. degree in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1992.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. He is an Academician with the Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image processing, spatial data infrastructure, and geospatial data sharing and interoperability.

Dr. Gong is the President of the Commission VI of the International Society for Photogrammetry and Remote Sensing.