

Received August 7, 2021, accepted August 31, 2021, date of publication September 3, 2021, date of current version April 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3109989

Robust Semisupervised Land-Use Classification Using Remote Sensing Data With Weak Labels

RUI WANG¹ AND MAN-ON PUN^{1,2,3}, (Senior Member, IEEE)

¹School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, Guangdong 518172, China

²Shenzhen Research Institute of Big Data, Shenzhen, Guangdong 518172, China

³Shenzhen Key Laboratory of IoT Intelligent Systems and Wireless Network Technology, Shenzhen, Guangdong 518172, China

Corresponding author: Man-On Pun (simonpun@cuhk.edu.cn)

This work was supported in part by Shenzhen Science and Technology Innovation Committee under Grant ZDSYS20170725140921348 and Grant JCYJ20190813170803617, and in part by Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS) under Grant AC01202005001.

ABSTRACT This work develops robust semisupervised classifiers to tackle the three most challenging problems in land-use classification using remote sensing data, namely, information imbalance, label noise, and image uncertainty. Limited by technology and cost, collecting clean labels for remote sensing images is difficult and often impractical. The change of environment and time also increases the uncertainty of remote sensing images. To overcome the obstacles incurred by the mixed pixels and weak labels, this work proposes dividing the pixels in remote sensing images into two groups, namely, pixels with accurate labels and those with weak labels, before processing the weakly labeled pixels using a nuclear norm-based cost function. To address the imbalanced data problem in pixels with accurate labels, an improved cross-entropy-based cost function is proposed to weigh the contributions from data of different classes based on their importance by exploiting the term frequency-inverse document frequency (TF-IDF) algorithm. Finally, an artificial class called “unknown” is proposed to cope with the interference caused by weakly labeled data with unrepresentative spatial features. Extensive experiments validate the effectiveness of the proposed semisupervised classifier.

INDEX TERMS Land-use classification, nuclear norm, semisupervised learning, weak labels.


I. INTRODUCTION

In the history of earth observation, land-use information has been considered a key factor in observing human development. Reliable and accurate land-use information is critical for understanding historical land use and planning for future land use. The advent of remote sensing technology has enabled the accurate and dynamic monitoring of land-use changes and global resource distributions in a periodic and timely manner [1], [2]. In particular, high-resolution remote sensing images can provide detailed land-use information, which enables us to perform a thorough study of the changes in land resource distributions. As a result, deep learning technology has been widely adopted in remote sensing. Deep learning-based algorithms are highly efficient in processing large-scale high-resolution remote sensing images to reveal hidden spatial features, which helps improve our understanding of remote sensing images. For instance, the seminal work

on fully convolutional networks (FCNs) for semantic segmentation proposed in [3] has inspired tremendous research interest, and deep learning technology has been widely adopted in remote sensing [4]–[9]. However, it has been observed that the generalization capabilities of these deep learning-based algorithms are unsatisfactory [10], which hinders the adoption of these deep learning-based algorithms for the automatic processing of large-scale remote sensing data.

The generalization ability of a classifier has a solid logical relationship with the quality of the dataset [11]. However, remote sensing data are different from traditional computer vision (CV) data. The differences include the information imbalance in natural image data, and natural image datasets also contain considerable noise in their labeling systems [12]. The common errors in these datasets are shown in Fig. 1. This research will focus on the three characteristics of remote sensing datasets: data imbalance, noise, and uncertainty [13], [14].

Imbalanced dataset information is an essential factor that leads to the degradation of classifier performance.

The associate editor coordinating the review of this manuscript and approving it for publication was Khin Wee Lai .

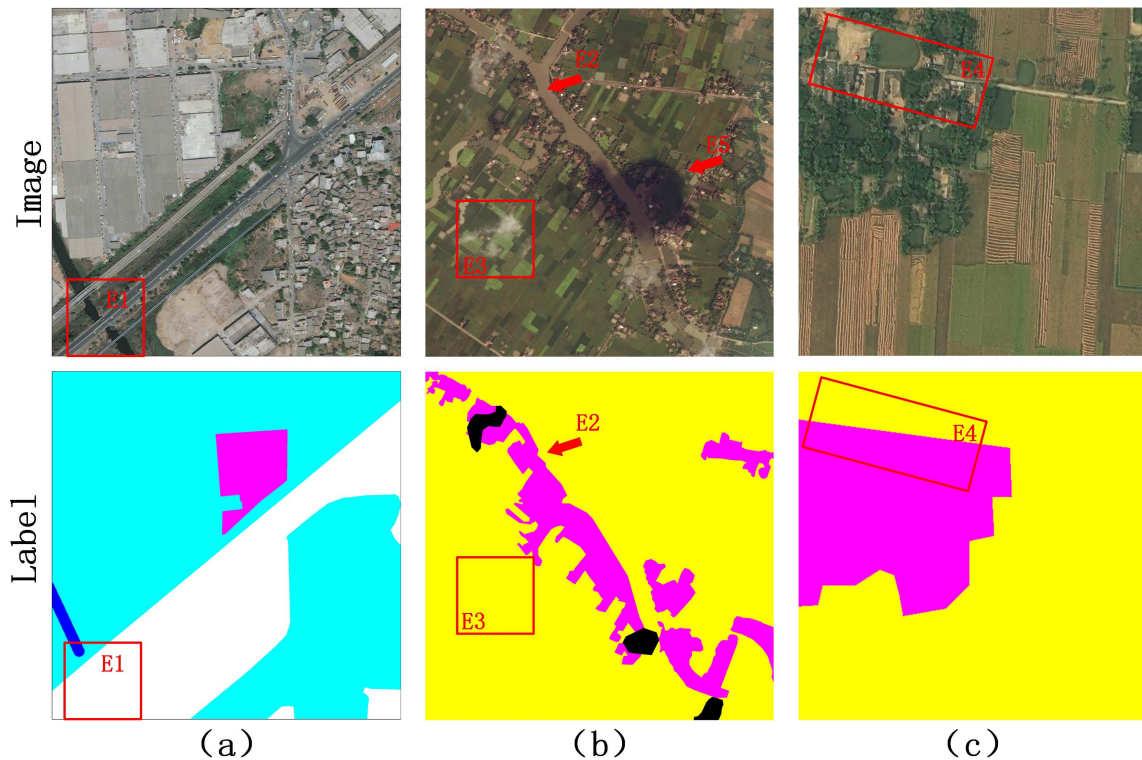


FIGURE 1. Errors in a remote sensing dataset. The red numbers in the figure represent different types of errors. For example, the labels of E1 E2 and E3 are erroneous; this appears to be the result of an unprofessional drafter. E4 is an error caused by the complexity of the land cover. E5 is not an error, but it represents a feature anomaly caused by cloud shadows. All samples come from the DeepGlobe LandCover CVPR2018 dataset.

The number of samples may vary dramatically across different land-use classes owing to the uneven spatial distribution of land resources. Therefore, such a problem of sample imbalance restricts the classification accuracy of small classes, which reduces the average classification accuracy of image segmentation. In the field of machine learning, the problem of imbalanced learning has been a topic of great interest [15], and learning the decision boundaries between different classes can be a very difficult task [16], [17].

In addition to information imbalance, label noise is a common problem in remote sensing datasets. The problem of label noise might be pervasive for the following reasons [18]: First of all, there is a high probability of label errors when the land cover in remote sensing images is highly complex or the information provided to drafters is minimal. Furthermore, the credibility is significantly reduced when an automatic label system or unprofessional drafters are used to cut costs. In addition, experts in different fields have different identification standards for the same land, which eventually leads to inconsistent labeling results. Finally, various noise interferes with remote sensing images when capturing and transferring data. When the classification datasets are corrupted, the performance degeneration issue of deep learning models becomes more severe than that of shallow classifiers. Therefore, researchers have developed techniques to combat data noise. Although there are

some studies on the robustness of remote sensing classifiers, only a few pieces of research focus on label noise land-use classification. However, in the practical application of remote sensing, label noise is an urgent and inevitable problem.

Imbalanced data and noise are explained in the previous paragraphs. Uncertainty refers to the random abnormal features in remote sensing images. Compared with natural images, remote sensing images have higher robustness to uncertainty [19]–[21]. Because sensors collect cloud shadows and other irrelevant information, remote sensing images may contain many invalid features. If there is no effective mechanism to deal with these features without classification significance, then these invalid features will be uncontrollably distributed to different land-use classifications, which will cause the overfitting of classifiers.

This study uses a scheme similar to semisupervised learning and proposes a loss function composed of two components. The first component computes the distance between the label and the corresponding prediction matrix using an improved cross-entropy (ICE) approach. In addition, a new weight representing the importance of sample information has been added into the cross-entropy function. The second component is designed to maximize the rank of the prediction matrix by exploiting the nuclear norm. An increase in the rank of the prediction matrix means a decrease in

redundant information. More specifically, the contributions of this research can be summarized as follows.

- To circumvent the imbalanced data problem, an effective solution for remote sensor image classification in the presence of noisy labels is provided. It is very general and can be seamlessly applied to current neural networks.
- To circumvent the imbalanced data problem, the term frequency-inverse document frequency (TF-IDF) [22] is introduced. An algorithm initially developed for document search and information retrieval is utilized to weigh the loss function based on the sample size of each class, and the weight is added into the cross-entropy computation;
- There is an additional component of the nuclear norm in the loss function. The information redundancy in the prediction matrix is reduced by maximizing the nuclear norm. This is similar to minimizing information entropy but maximizing the kernel norm can avoid the performance degradation of the classifiers due to information imbalances.
- A new classification called the “unknown” class will be added to the classifier. None of the information in the dataset is about the “unknown” class. This class does not have any labels, so it cannot participate in the cross-entropy computation. However, it will have significant implications in nuclear norm maximization. In addition, the “unknown” class collects anomalous features to prevent overlearning of the classifier.

Extensive computer experiments were performed to show that the resulting semisupervised classifier is highly robust against the mixed pixel, weak label, and imbalanced data problems by exploiting a smaller amount of weakly labeled data. The proposed classifier is particularly attractive because it can make use of weak data, such as historical data of the same area accumulated over the years. The remainder of this paper is organized as follows. Sec. II introduces the classical techniques for improving classifier performance. Sec. IV elaborates on the proposed semisupervised classifier, and the extensive simulation results are presented in Sec. V. Finally, the conclusions are presented in Sec. VI.

Notation: Vectors and matrices are denoted by boldface letters. $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_v$ denote the Frobenius and nuclear norms of \mathbf{A} , respectively. Furthermore, $[\mathbf{A}]_{i,j}$ denotes the i -th row and the j -th column element of \mathbf{A} . $\text{rank}(\mathbf{A})$ and $\text{trace}(\mathbf{A})$ represent the rank and trace of \mathbf{A} , respectively. In addition, \mathbf{A}^T and \mathbf{A}^H are the transpose and conjugate transpose of \mathbf{A} , respectively. Finally, sets are represented by calligraphic letters, while $|\mathcal{X}|$ represents the cardinality of the set \mathcal{X} .

II. RELATED WORK

A. INFORMATION IMBALANCE IN DEEP LEARNING

Imbalanced information is a traditional and common problem, and research on this problem has drawn extensive attention. For this problem, traditional solutions include

resampling and reweighting. Chawla *et al.* [23] proposed a scheme called the synthetic minority oversampling technique to increase the importance of unusual samples. He and Garcia [24] explained how to process unbalanced data and explored the relationship between different resampling methods and classifier performance. Recently, Byrd *et al.* [25] discussed the relationship between the training samples' position and the classifier's performance. They think that when the samples are sufficient, the information balance can be better achieved by resampling. These methods help us understand the relationship between samples and classifier performance from the perspective of resampling. The representative scheme is reweighting. Khan *et al.* [26] proposed a cost-sensitive (CoSen) deep neural network, which can automatically learn robust feature representations for both the majority and minority classes. Cui *et al.* [27] were convinced that datasets contain information overlap, so they proposed a novel theoretical framework to characterize data overlap, and a class-balanced reweighting term that is inversely proportional to the adequate number of samples was added to the loss function. Cao *et al.* [28] alternatively studied the minimum margin per class and designed a label-distribution-aware loss function that encourages a model to have the optimal trade-off between per-class margins. Tan *et al.* [29] proposed equalization loss to tackle the problem of rare long-tailed categories by ignoring the gradients for rare categories. In recent years, all of these methods have become popular reweighting methods. Researchers have not stopped exploring the imbalanced information problem. Kang *et al.* [30] compared jointly learning a representation and classifier to many straightforward decoupled methods and found that instance-balanced sampling gives more generalizable representations that can achieve state-of-the-art performance after properly rebalancing the classifiers. Zhou *et al.* [31] proposed a new model consisting of two branches, termed the “conventional learning branch” and the “rebalancing branch,” to simultaneously address both representation learning and classifier learning. These methods have also received more attention in recent years, although their methods increase the computational costs.

B. LABELS NOISE IN DATASETS

The high cost of acquiring satellite image labels is a well-known problem in the field of remote sensing. Almost all data sets related to semantic segmentation are faced with label noise. Label noise was first considered by pioneers in CV, and these pioneers have produced many exciting and significant research results. Angulin and Laird [32] asked the following question: how can a learning algorithm cope with incorrect training examples?. Since then, label noise has been the focus of researchers. Lawrence and Schölkopf [33] proposed an algorithm for constructing a kernel Fisher discriminant (KFD) from training examples with noisy labels. Natarajan *et al.* [34] theoretically studied binary classification in the presence of random classification noise and provided two approaches to suitably modify any given surrogate

loss function. Liu and Tao [35] presented a necessary reweighting framework for classification in the presence of label noise. Theoretical analyses were provided to assure that the learned classifier will converge to the optimal noise-free sample. Applying these methods to natural images is successful, but their performance degrades when directly applied to remote sensing images. The specificity of remote sensing images causes this. Li *et al.* directed the label noise problem of remote sensing data based on the multifeature dictionary learning-based collaborative representation classifier (MDLCRC) [36], and a new RSSC-oriented error-tolerant deep learning (RSSC-ETDL) approach to mitigate the adverse effect of incorrect labels in a remote sensing image scene dataset was proposed [37]. Kang *et al.* [38] used the newly defined robust normalized softmax loss (RNSL). In the same year, they proposed a new deep metric learning loss function, termed noise-tolerant deep neighborhood embedding (NTDNE), which can accurately capture the semantic relations among remote sensing scenes in a feature space [39]. These results show that the label noise problem has become a focus in the remote sensing field.

C. UNCERTAINTY OF REMOTE SENSING IMAGE

The natural surface of the Earth is composed of a uniform material. As a result, many pixels in remote sensing images may cover multiple substances with different spectral properties [40], [41]. In addition, each pixel in remote sensing images can exhibit spatial characteristics belonging to one or more classes, which may interfere with land-use classification. One naive solution to the mixed pixel problem is to decompose the multiclass classification problem into multiple independent single-class classification problems while ignoring the cross-class correlation. Unfortunately, mixed pixels usually demonstrate nonlinear mixing of different classes, particularly in high-resolution remote sensing images [42]. As reported by Stubenrauch *et al.* [43], on average, more than 50% of the Earth’s surface is covered by clouds every day. Clouds and “cloud shadows” are symbiotic in remote sensing images. Arguably, any classifier faces the challenge of “clouds” and “cloud shadows” when it is used. In general, some commonly used methods, including band grouping/thresholding methods [44]–[46], traditional image segmentation methods [47]–[49], and deep learning-based segmentation methods [50]–[52] can lower the interference of these factors but cannot be eliminated. Notably, these features are complex and cannot be comprehensively characterized with accurate labeling. Many classifiers are designed without considering the uncertainty of remote sensing images. Therefore, more robust classifiers need to be designed to overcome the uncertainty of remote sensing images.

III. PROBLEM FORMULATION AND ASSUMPTIONS

Given a set of K equal-size remote sensing images with N_{total} pixels each, the task of remote sensing image semantic segmentation is to develop a classifier to produce a prediction matrix $\hat{A} \in \mathbb{R}^{N_{total} \times N_C}$ for an input image, where N_C is

the total number of classes. Furthermore, each element $[\hat{A}]_{i,j} \geq 0$ represents the probability of the i -th pixel of the input image belonging to the j -th class with

$$\sum_{j=1}^{C_N} [\hat{A}]_{i,j} = 1, \tag{1}$$

for $i = 1, 2, \dots, N_{total}$ and $j = 1, 2, \dots, C_N$.

The conventional supervised learning approach constructs \hat{A} by training on a large set of data samples with correct labels. This data requirement can be an issue of concern in practice when correctly labeled samples are not available. This study designs a robust semisupervised classifier by exploiting imbalanced remote sensing datasets with both accurate and weak labels. To facilitate the development of the semisupervised classifier, we propose dividing the pixels of the k -th image into two sets, namely, $\mathcal{X}_k^{(c)}$ for those pixels falling within the core area of the cluster with well-defined labels and $\mathcal{X}_k^{(b)}$ for those pixels with weak labels for $k = 1, 2, \dots, K$. Pixels in $\mathcal{X}_k^{(b)}$ are primarily in the boundary area and potentially belong to multiple classes. Fig. 2 illustrates a hypothetical example of three land-use classes. The pixels were divided into pixels with well-defined and weak labels. Note that pixels with weak labels are defined along each boundary line between any two classes.

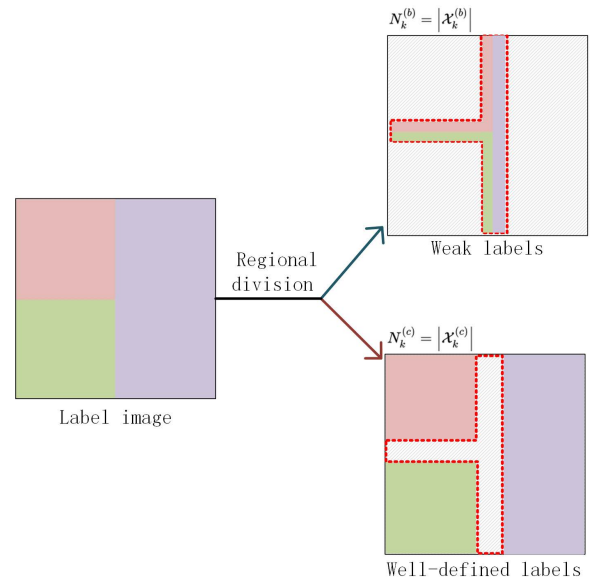


FIGURE 2. Dividing pixels with meaningful labels from those with weak labels.

Before elaborating on our proposed classifier, we first state the three assumptions necessary for establishing valid semisupervised learning models [53].

- Smoothness assumption: Two geographically close pixels in a high-density region should have a strong spatial correlation and subsequently, similar classification labels of high probability [54].

- Cluster assumption: If two pixels are in the same cluster, they belong to the same class with a high probability. Furthermore, if the spectral characteristics of two pixels are similar, the probability of these two pixels possessing identical classification labels should be high [55]–[57].
- Manifold assumption: Remote sensing data reside roughly in a low-dimensional manifold. In other words, samples are assumed to have similar spatial characteristics in a small local proximity and therefore belong to similar classes [58]–[60].

IV. PROPOSED SEMISUPERVISED CLASSIFIER

In this section, we propose a semisupervised classifier to perform robust land-use classification by effectively exploiting weakly labeled and imbalanced remote sensing data with the inherent mixed pixel problem.

A. TF-IDF-BASED WEIGHTING

We begin with the pixels with accurate labels in $\mathcal{X}_k^{(c)}$ and address the imbalanced data problem. Conventionally, cross entropy is employed as the cost function to measure the discrepancy between the true labels and the predicted values in machine learning-based applications [61], [62]. For a given pair of prediction matrices $\hat{A}_k^{(c)}$ and its corresponding ground truth $A_k^{(c)}$ generated with true labels, the cross entropy of $\hat{A}_k^{(c)}$ and $A_k^{(c)}$ is given by

$$H_{CE}(\hat{A}_k^{(c)}, A_k^{(c)}) = - \sum_{i=1}^{N_{\text{total}}} \sum_{j=1}^{C_N} [A_k^{(c)}]_{i,j} \cdot \log [\hat{A}_k^{(c)}]_{i,j}. \quad (2)$$

However, imbalanced training data will negatively impact the classification decision boundary, and a strong bias toward the more populated classes will exist. To address this problem, we propose a weighted loss function by exploiting the TF-IDF algorithm originally developed for document search and information retrieval [63]. Fig. 3 illustrates the decision boundary before and after applying the weight adjustment in a hypothetical example. As depicted in Fig. 3, the weighted loss function usually focuses on the important data samples while shrinking the decision boundary toward the center of gravity of each class. Furthermore, the TF-IDF algorithm assigns different weights to the contributions from different classes in its loss function based on the frequency and importance of the classes [64]. More specifically, the weighting coefficient of a word in a set of files (also known as a corpus) in TF-IDF is positively proportional to the frequency of its appearance in one file but inversely proportional to the number of files containing the word in the corpus. Thus, the TF-IDF algorithm generates a larger weighting coefficient for a given word if it appears frequently in one file but rarely in other files. Inspired by the TF-IDF algorithm, we treat each *class* in a set of remote sensing images as one word in a corpus. If pixels corresponding to one class appear more frequently in one image but rarely in other images, then a larger weight is assigned to their contribution to the loss function. $d_{k,j}^{(c)}$ denotes

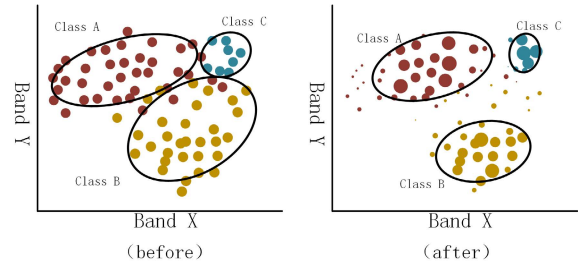


FIGURE 3. The influence of adjusting weights for the decision boundary.

the total number of pixels in $\mathcal{X}_k^{(c)}$ that belong to the j -th class; that is ,

$$N_k^{(c)} = \sum_{j=1}^{C_N} d_{k,j}^{(c)}, \quad (3)$$

where $N_k^{(c)} = |\mathcal{X}_k^{(c)}|$ is the total number of pixels with well-defined labels in the k th image.

Now, we define the importance of the samples that belong to the j th class in the k -th image as follows:

$$\tilde{\omega}_{k,j} = \frac{d_{k,j}^{(c)}}{N_k^{(c)}} \log \left(\frac{\sum_{k=1}^K N_k^{(c)}}{\sum_{k=1}^K d_{k,j}^{(c)}} \right). \quad (4)$$

After normalizing $\tilde{\omega}_{k,j}$, the normalized weighting coefficient for the samples that belong to the j -th class in the k -th image can be expressed as

$$\omega_{k,j} = \frac{\tilde{\omega}_{k,j}}{\sum_{j=1}^{C_N} \sum_{k=1}^K \tilde{\omega}_{k,j}}. \quad (5)$$

Finally, we propose the following ICE approach as the cost function for $\mathcal{X}_k^{(c)}$ using the following TF-IDF-based weighting coefficients:

$$H_{ICE}(\hat{A}^{(c)}, A^{(c)}, W) = - \sum_{k=1}^K \sum_{i=1}^{N_k^{(c)}} \sum_{j=1}^{C_N} \omega_{k,j} [A_k^{(c)}]_{i,j} \log [\hat{A}_k^{(c)}]_{i,j}, \quad (6)$$

where

$$\hat{A}^{(c)} = \{\hat{A}_1^{(c)}, \hat{A}_2^{(c)}, \dots, \hat{A}_K^{(c)}\}, \quad (7)$$

$$A^{(c)} = \{A_1^{(c)}, A_2^{(c)}, \dots, A_K^{(c)}\}, \quad (8)$$

$$W = \{\omega_{k,j}\}. \quad \forall k, j \quad (9)$$

In the following, H_{ICE} , which is defined in Eq. (6), is referred to as ICE. It is worth noting that the contributions from pixels associated with the less populated classes, such as “Urban land”, are more heavily weighted in ICE compared to those pixels from the more populated classes, such

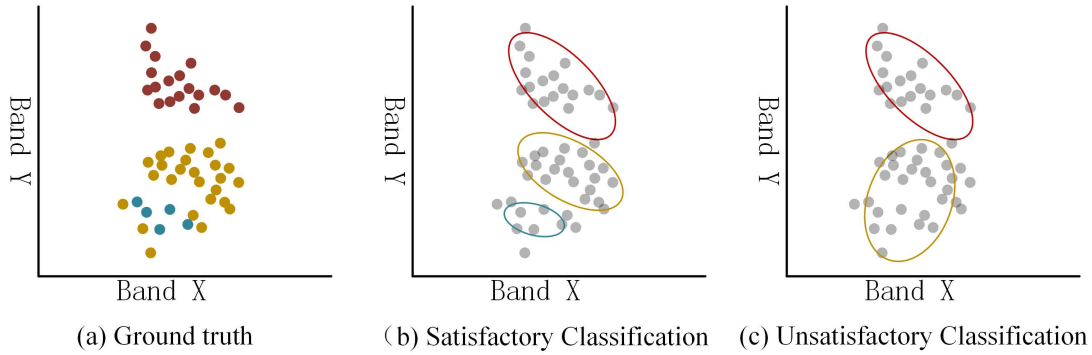


FIGURE 4. Schematic diagram of the decline in classifier diversity in unsupervised learning.

as “Agriculture land”. Furthermore, ICE degenerates to the conventional cross entropy if $\omega_{k,j} = 1$ for $j = 1, 2, \dots, N_C$ and $k = 1, 2, \dots, K$.

B. NUCLEAR NORM-BASED COST FUNCTION

Next, we concentrate on the pixels with weak labels $\mathcal{X}_k^{(b)}$. For data with accurate labels, the minimization of $H_{CE}(\hat{A}_k^{(c)}, A_k^{(c)})$ can effectively improve the classification performance by reducing the discrepancy between $\hat{A}_k^{(c)}$ and $A_k^{(c)}$. However, H_{CE} is not a good performance metric for data with weak labels, as its corresponding $A_k^{(b)}$ is prone to errors. Fig. 4(a) illustrates a hypothetical example with three classes of weakly labeled data. Fig. 4(b) shows the decision boundary if the correct data labels are used. In contrast, if the classifier is trained to minimize H_{CE} , then the resulting classifier may mistakenly categorize the lower two classes of data samples into one, as shown in Fig. 4(c)].

Inspired by the observation that inconsistent labels arise owing to the mixed spectral characteristics of several land-use classes, we propose to maximize the rank of the resulting prediction matrix $\hat{A}_k^{(b)}$. For instance, we consider the following two prediction matrices denoted by $\hat{A}_1^{(b)}$ and $\hat{A}_2^{(b)}$, which are given in Eq. (10):

$$\hat{A}_1^{(b)} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \hat{A}_2^{(b)} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.8 & 0.1 & 0.1 \\ 0.8 & 0.1 & 0.1 \end{bmatrix} \quad (10)$$

Despite their similar appearances, their ranks are different, that is, $\text{rank}(\hat{A}_1^{(b)}) = 3$, whereas $\text{rank}(\hat{A}_2^{(b)}) = 1$. As a result, the land-use classifier outputting $\hat{A}_2^{(b)}$ can only identify the first land-use class, whereas $\hat{A}_1^{(b)}$ is more advantageous as a prediction matrix.

Unfortunately, the maximization of $\text{rank}(\hat{A}_k^{(b)})$ is nonconvex. Thus, it is nontrivial to directly maximize the rank of the prediction matrix. To address this problem, we propose to maximize the nuclear norm of \hat{A} as follows:

$$\|\hat{A}_k^{(b)}\|_v = \text{trace} \left\{ \sqrt{\hat{A}_k^{(b)H} \hat{A}_k^{(b)}} \right\}. \quad (11)$$

It is worth noting that the nuclear norm is essentially the convex envelope of the matrix rank [65]. Nuclear norm-based optimization has been used for matrix completion and robust principal component analysis (PCA) [66]–[68]. Recall that the nuclear norm of $\hat{A}_k^{(b)}$ is the sum of its singular values, and we can consider $\|\hat{A}_k^{(b)}\|_v$ to be the approximation of $\text{rank}(\hat{A}_k^{(b)})$. Thus, the maximization of the nuclear norm of $\hat{A}_k^{(b)}$ can effectively increase the number of predicted classes that can be identified in the remote sensing data, which can be translated into classification performance improvement. Furthermore, it has been shown that [69]

$$\frac{1}{\sqrt{Q}} \|\hat{A}_k^{(b)}\|_v \leq \|\hat{A}_k^{(b)}\|_F \leq \|\hat{A}_k^{(b)}\|_v, \quad (12)$$

where $Q = \min(N_k^{(b)}, N_C)$ and $N_k^{(b)} = |\mathcal{X}_k^{(b)}|$ is the cardinality of $\mathcal{X}_k^{(b)}$. Furthermore, $\|\hat{A}_k^{(b)}\|_F$ is the Frobenius norm of $\hat{A}_k^{(b)}$ and is defined as follows:

$$\|\hat{A}_k^{(b)}\|_F = \sqrt{\sum_{i=1}^{N_k^{(b)}} \sum_{j=1}^{N_C} \left| [\hat{A}_k^{(b)}]_{i,j} \right|^2}. \quad (13)$$

Thus, the maximization of $\|\hat{A}_k^{(b)}\|_v$ effectively increases the upper and lower bounds of $\|\hat{A}_k^{(b)}\|_F$, as shown in Eq. (12). We recall that $\|\hat{A}_k^{(b)}\|_F$ is inversely related to the entropy of $\hat{A}_k^{(b)}$, which is given by

$$H_E(\hat{A}_k^{(b)}) = - \sum_{i=1}^{N_k^{(b)}} \sum_{j=1}^{N_C} [\hat{A}_k^{(b)}]_{i,j} \log [\hat{A}_k^{(b)}]_{i,j}. \quad (14)$$

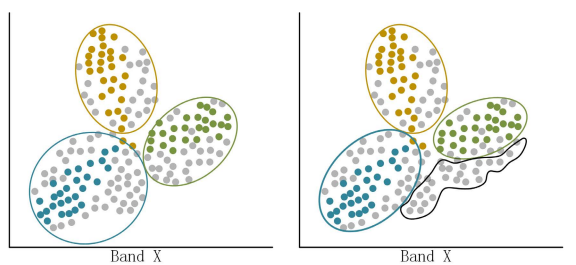
Therefore, an increase in $\|\hat{A}_k^{(b)}\|_F$ leads to a reduction in $H_E(\hat{A}_k^{(b)})$, i.e. the uncertainty of the prediction matrix $\hat{A}_k^{(b)}$ is reduced. Thus, maximizing $\|\hat{A}_k^{(b)}\|_v$ improves the diversity of $\hat{A}_k^{(b)}$ while reducing the uncertainty of $\hat{A}_k^{(b)}$, which contributes to the improvement in the classification accuracy.

C. THE “UNKNOWN” CLASS

Conventional classifiers are designed to adjust their decision boundaries to accommodate all pixels regardless of the

confidence levels of the data labels. As a result, conventional classifiers suffer from poor generalization capabilities, as they are forced to accommodate data with noisy features. Motivated by this observation, we propose the creation of an additional artificial class called the “unknown” class to handle data with weak labels, that is, $\mathcal{X}_k^{(b)}$. Therefore, weakly labeled data with atypical spatial features can be classified into this new class without overfitting the classifier, as shown in Fig. 5. As shown in the later experimental results, the new “unknown” class can help expedite the training process by preventing overfitting. With the additional “unknown” class, the nuclear norm of $\hat{A}_k^{(b)}$ takes the following form:

$$\|\hat{A}_k^{(b)}\|_v = \text{trace} \left\{ \sqrt{\hat{A}_k^{(b)H} \hat{A}_k^{(b)}} \right\}, \quad (15)$$



a. All the samples were forcibly classified b. Appropriate samples are classified correctly.

FIGURE 5. Introduction of the “unknown” class for data with high uncertainty.

where $\hat{A}_k^{(b)} \in \mathbb{R}^{N_k^{(b)} \times (N_C + 1)}$ is the prediction matrix for the $N_C + 1$ classes. In the following, we use the nuclear norm of $\hat{A}_k^{(b)}$ defined in Eq. (15) as the loss function for the weakly labeled data in our proposed classifier.

D. SEMISUPERVISED LEARNING

In recent years, semisupervised learning has attracted wide attention from scholars in the field of CV. The core idea of semi-supervision is to use a small labeled data set to define features and then use unlabeled data to enhance the classifier’s ability to understand features. Many semisupervised learning methods are proposed based on intelligent data enhancement strategies such as RandAugment [70] or AutoAugment [71], such as MixMatch [72] method, and Unsupervised Data Augmentation [73]. Recently, there has been widespread concern about the use of pseudo-marking and consistent regularization. FixMatch [74] has achieved state-of-the-art results on four benchmark data sets. The above research has a guiding role in the application of semisupervised learning in remote sensing. Semisupervised learning is ideal for land use classification because of the low cost of acquiring remote sensing images. The application of semisupervised learning technology in satellite remote sensing land classification is still in the development stage. Experiments are only performed on some simple datasets and have not been applied to actual scenes on a large scale, such as image

classification [75]–[79] and information extraction [80]–[83]. In large-scale scenarios, three assumptions of semisupervised learning cannot be satisfied if unlabeled samples are added blindly.

We used the logic of semisupervised learning to cope with the noise of remote sensing data. We pre-circle out untrusted regions in the data whose pixels will not be involved in the cross-entropy calculation of the loss function but enter into an unsupervised computational process. Our scheme satisfies the three assumptions of semisupervised learning because the information involved in semisupervised learning comes from the same sample. The pseudocode is as shown in Algorithm 1.

Algorithm 1 Land Use Classifier Based on Semisupervised Learning

- Require:** The dataset images \mathcal{X} into two sets, namely $\mathcal{X}^{(c)}$ with well-defined labels $A^{(c)}$ and $\mathcal{X}^{(b)}$ for those pixels with weakly labels.
- Require:** Learning rate ϵ and initial parameter θ
- Require:** C_N is the number of dataset classes
- 1: **for** $t = 1, \dots, T$ **do**
 - 2: Sample a minibatch of S_t examples from the training set $\{\mathcal{X}_1, \dots, \mathcal{X}_t\}$
 - 3: The maximum number of categories is $C_N + 1$: $\hat{A}_t \leftarrow f(\mathcal{X}_t; \theta)$
 - 4: Divide the prediction matrix into two sets: $\hat{A}_t = \hat{A}_t^{(c)} + \hat{A}_t^{(b)}$
 - 5: Remove Unknown class: $\hat{A}_t^{(c)} \leftarrow \hat{A}_t^{(c)}$
 - 6: $\mathcal{W}_t \leftarrow \text{WEIGHT}(\mathcal{X}_t^{(c)})$, via Eq.5
 - 7: $\mathcal{L}_{\text{ICE}} \leftarrow H_{\text{ICE}}(\hat{A}_t^{(c)}, \mathcal{A}_t^{(c)}, \mathcal{W}_t)$, via Eq.6
 - 8: $\mathcal{L}_{\text{NuN}} \leftarrow \sum_{k=1}^K \left[1 - \frac{1}{\sqrt{N_t^{(b)}}} \|\hat{A}_t^{(b)}\|_v \right]$, $N_t^{(b)}$ is the Num. of pixels with weakly labels in batch t .
 - 9: $\mathcal{L} = \mathcal{L}_{\text{ICE}} + \lambda \mathcal{L}_{\text{NuN}}$
 - 10: Computr gradient estimate: $\hat{g} \leftarrow +\nabla_{\theta} \mathcal{L}$
 - 11: Apply update: $\theta \leftarrow \theta - \epsilon \hat{g}$
 - 12: **end for**

Fig. 6 shows a flowchart of the proposed semisupervised classification framework. More specifically, the proposed classification framework can be divided into three components: inference, training, and data preprocessing. During the inference process, the backbone of the proposed classifier is trained with well-known semantic segmentation models, such as FCN and DeepLabV3+, using our proposed cost functions. The supervised and semisupervised learning modules share this backbone in the proposed classifier. For data preprocessing, the proposed classifier defines the areas of large uncertainty around each cluster with a width of m pixels, as inaccurate labeling mainly occurs in the boundary areas of different land-use clusters. Note that the parameter m can be adjusted according to the noise level of the given dataset. The detailed structure of the training process is shown in Fig. 7. For the data with accurate labels in $\mathcal{X}_k^{(c)}$, the following

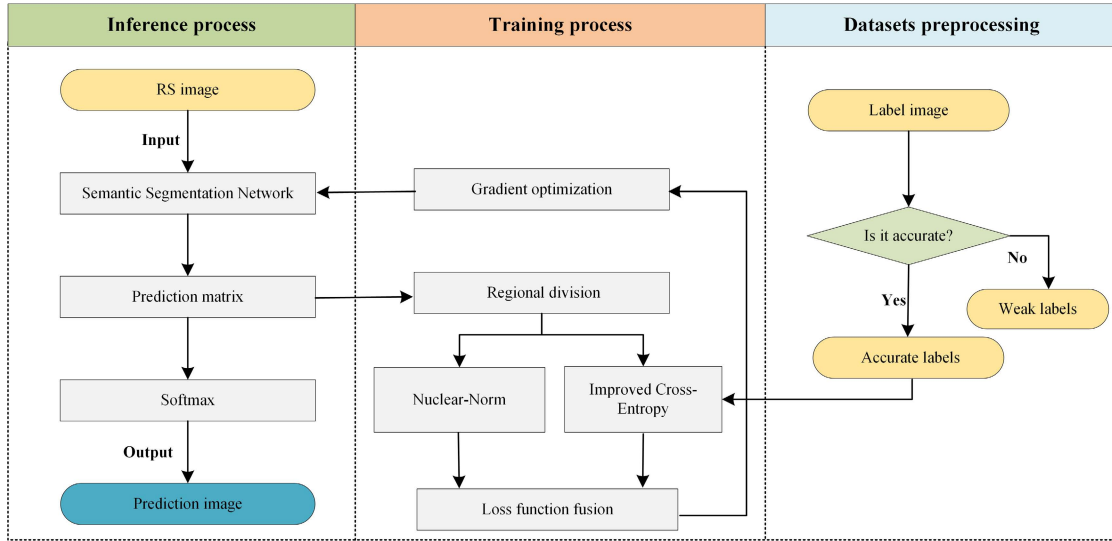


FIGURE 6. Flowchart for the proposed semisupervised classification framework.

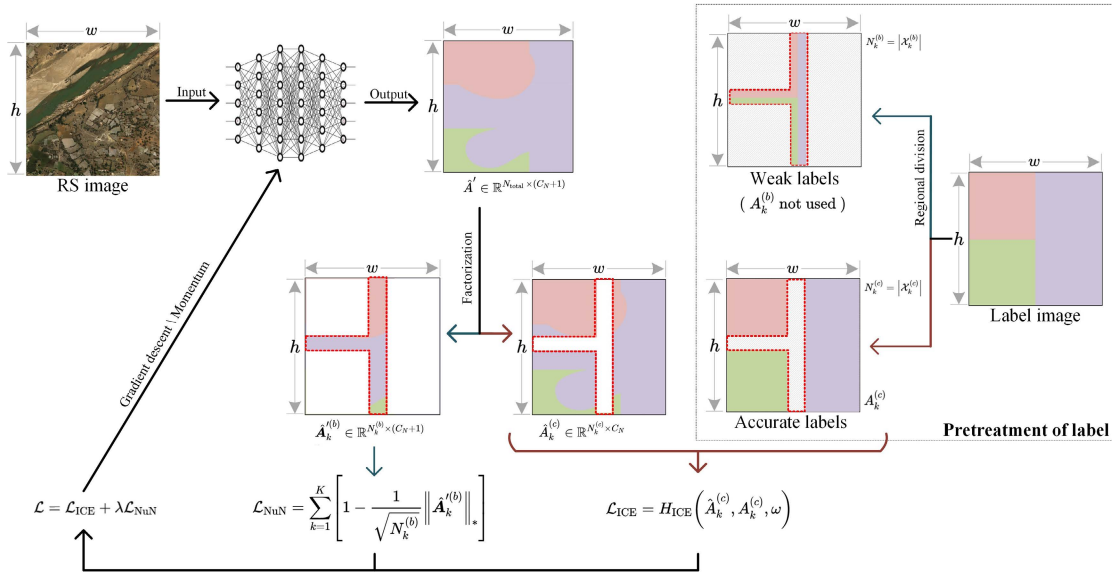


FIGURE 7. Loss function design scheme and learning logic for semisupervised classification.

ICE-based cost function is used to evaluate the prediction performance, as shown in Eq. (6):

$$\mathcal{L}_{ICE} = H_{ICE}(\hat{A}^{(c)}, A^{(c)}, W). \quad (16)$$

In contrast, for the data with weak labels in $\mathcal{X}_k^{(b)}$, the following nuclear norm-based cost function is utilized:

$$\mathcal{L}_{NuN} = \sum_{k=1}^K \left[1 - \frac{1}{\sqrt{N_k^{(b)}}} \|\hat{A}_k^{(b)}\|_* \right]. \quad (17)$$

Note that the label information for weakly labeled data is discarded in Eq. (17). Therefore, Eq. (17) represents a cost function for unsupervised learning. In summary, the cost function proposed by combining Eq. (16) and Eq. (17) can be

expressed as

$$\mathcal{L} = \mathcal{L}_{ICE} + \lambda \mathcal{L}_{NuN}, \quad (18)$$

where λ is a parameter designed to adjust the contribution of \mathcal{L}_{NuNorm} to the cost function. The proposed cost function is formulated to enhance the generalization capability of the classifier by minimizing the negative impact due to the weakly labeled data while expediting the training process by adding an “unknown” class to include pixels of high uncertainty.

V. RESULTS AND DISCUSSIONS

In this section, we show the effect of our scheme on the classifier through different experiments. Experiment 1 compares the prediction results of different classifiers on cloud-cover

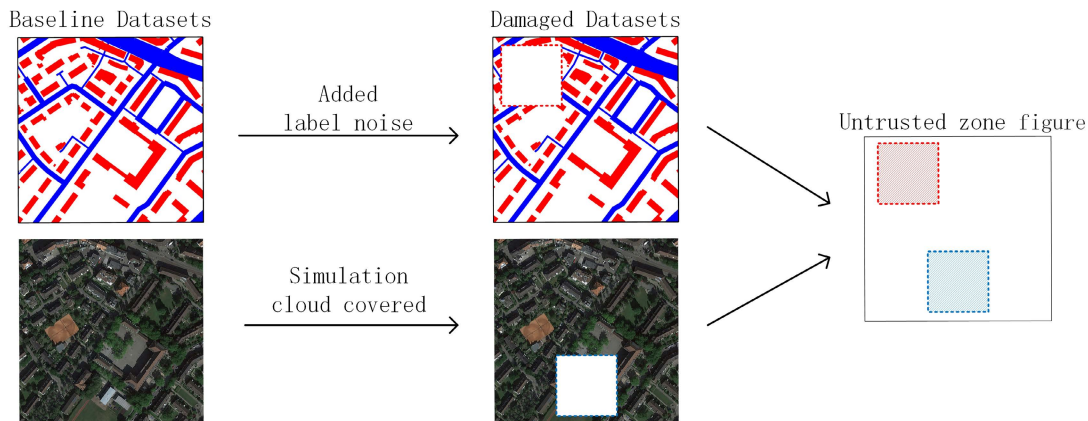


FIGURE 8. Two approaches to add noise to the training set.

images, and demonstrates the improvement in terms of classification robustness. In Experiment 2, we consider a chaotic dataset whose quality is closer to that of datasets in industrial applications. Our experimental results prove that our scheme can significantly improve the classifier's forecast accuracy and generalization ability.

A. EXPERIMENTS ON THE AIS DATASET

Experiment 1 is designed to show the improvement in terms of classification robustness. The aerial image segmentation (AIS) dataset was used as the baseline dataset, and a "Damaged dataset" was created. Three classifiers are trained in this experiment:

- When "deepLabeV3plus+cross-entropy" is used to train on the baseline dataset, it is called the "Baseline" classifier.
- When "deepLabeV3plus+cross-entropy" is used to train on the damaged dataset, it is called the "Damaged" classifier.
- When "deepLabeV3plus+our loss function" is used to train on the damaged dataset, it is called "Our" classifier.

Implementation details: We use gray patches of size 512×512 as inputs. Furthermore, we utilize the Adam optimizer with parameters of $\alpha = 0.0001$ and $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The training procedure follows the minibatch strategy, and the batch size is 8. All the networks in the experiments are implemented using the PyTorch platform and trained with an NVIDIA GeForce RTX 3080TI GPU.

1) DATASET

The AIS dataset contains labels for buildings and roads in Berlin, Chicago, Paris, Potsdam, and Zurich. This experiment used the Zurich data as the baseline dataset. The Zurich AIS dataset contains 364 samples, and we downsampled them to 512×512 pixels. Finally, 14000 reliable samples were adopted in this experiment. The 14000 samples were divided as follows: 10000 samples were used for training while 4000 for validation.

We altered the 40% training set to create the "Damaged" set. There are two ways to add noise to the training set, as shown in Fig. 8. First, we randomly "damaged" the labels of 170×170 pixels to simulate label noise. Second, we randomly broke 170×170 pixels in the image to simulate cloud cover. Fig. 9 shows the percentage of error samples in the training set. There are 1300 samples with noisy labels, 1300 samples with noisy pixels, and 1400 samples with both kinds of noise. The remaining samples are reliable.

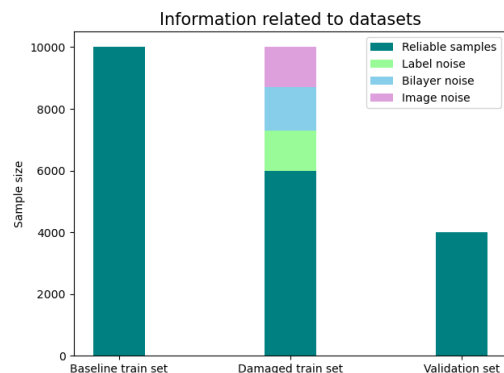


FIGURE 9. The proportion of different types of samples in the dataset.

2) RESULTS

Twenty training epochs were conducted for each classifier. Fig. 10 shows that the mean intersection over union (mIoU) of the validation set on the "Baseline" classifier is 0.6159 ± 0.0052 , but that of the "Damaged" classifier is only 0.4328 ± 0.0089 . This indicates that the degradation in classifier performance is caused by noise. Our scheme refined the mIoU to 0.5974 ± 0.0093 , and it is shown that our scheme can provide a better quality classifier. Note that the accurate positioning of the error pixels is the key to the good results. Fig. 11 shows the image classification results on three classifiers. The classification effect of "Our" classifier is similar to that of the "Baseline" classifier, while the classification effect of "Damaged" classifier is the worst. These classification effects were obtained on images without noise information.

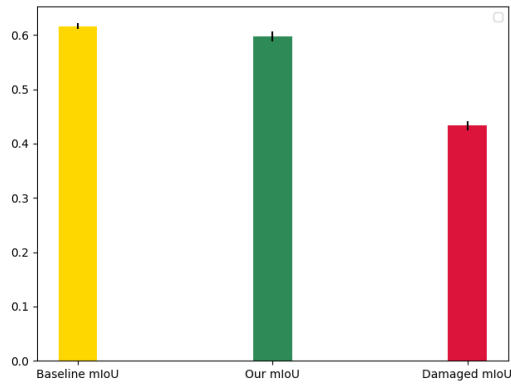


FIGURE 10. The application performance of the three classifiers on the validation set.

Some interesting classification results are shown in Fig. 12 in which noisy images are input into the three classifiers to evaluate their robustness. The “Baseline” classifier and “Damaged” classifier could not correctly process the abnormal features as they had to classify the abnormal pixels into the building, background, and road categories. In contrast,

our scheme provides an “unknown” class that can be selected for abnormal pixels. As a result, our scheme demonstrated significantly improved robustness in classification.

B. EXPERIMENTS ON THE DeepGlobe DATASET

In this section, we validate our proposed semisupervised classifier through extensive simulations using the DeepGlobe Land Cover Classification Challenge dataset. We compare the classification performance of the four classifiers discussed above.

- 1) **Supervised CE:** The conventional supervised classifier based on the cross-entropy function proposed in [61], [62]. Note that we use this classifier as the baseline to benchmark our proposed classifiers.
- 2) **Supervised ICE:** The supervised classifier based on the proposed ICE function.
- 3) **Semisupervised CE+NuN:** The semisupervised classifier based on the cross-entropy function and the nuclear norm.
- 4) **Semisupervised ICE+NuN:** The semisupervised classifier based on the proposed ICE function and the nuclear norm.

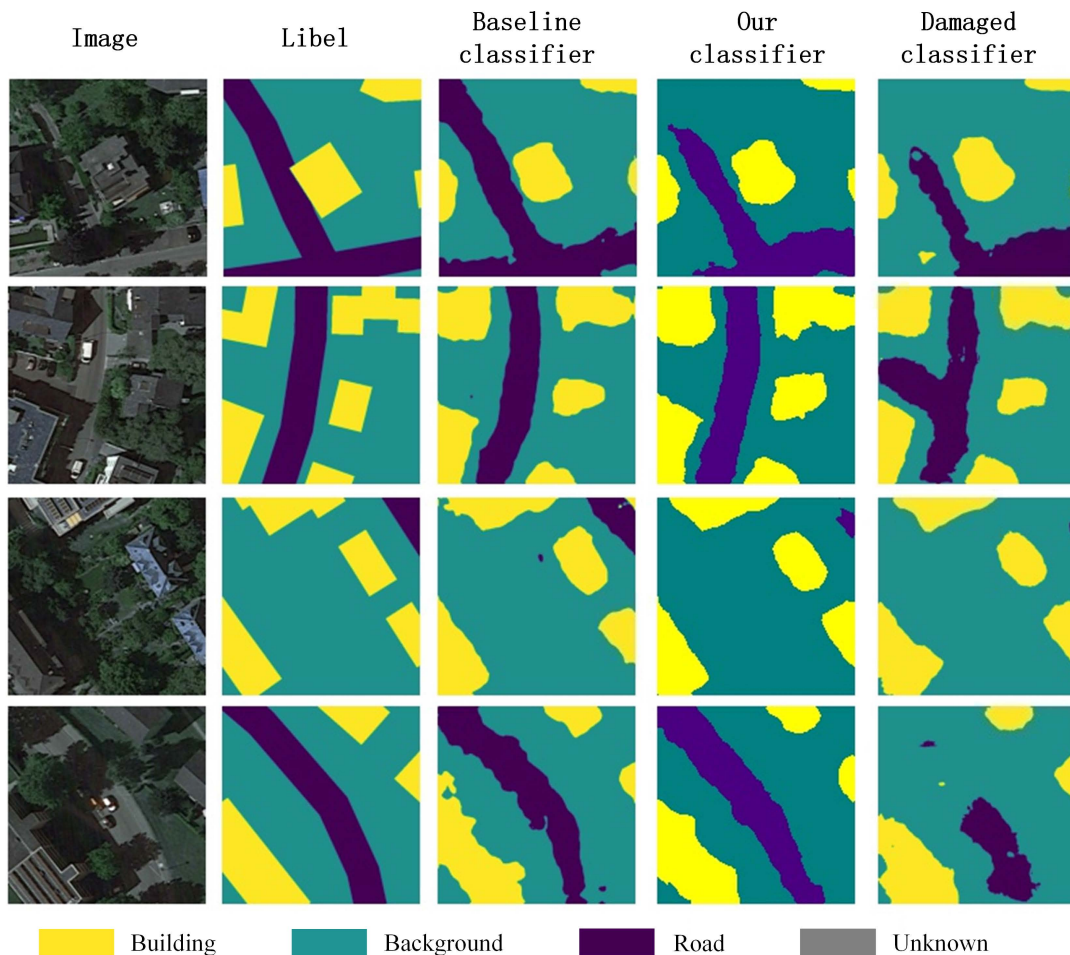


FIGURE 11. Prediction results with noise-free images in the classifier.

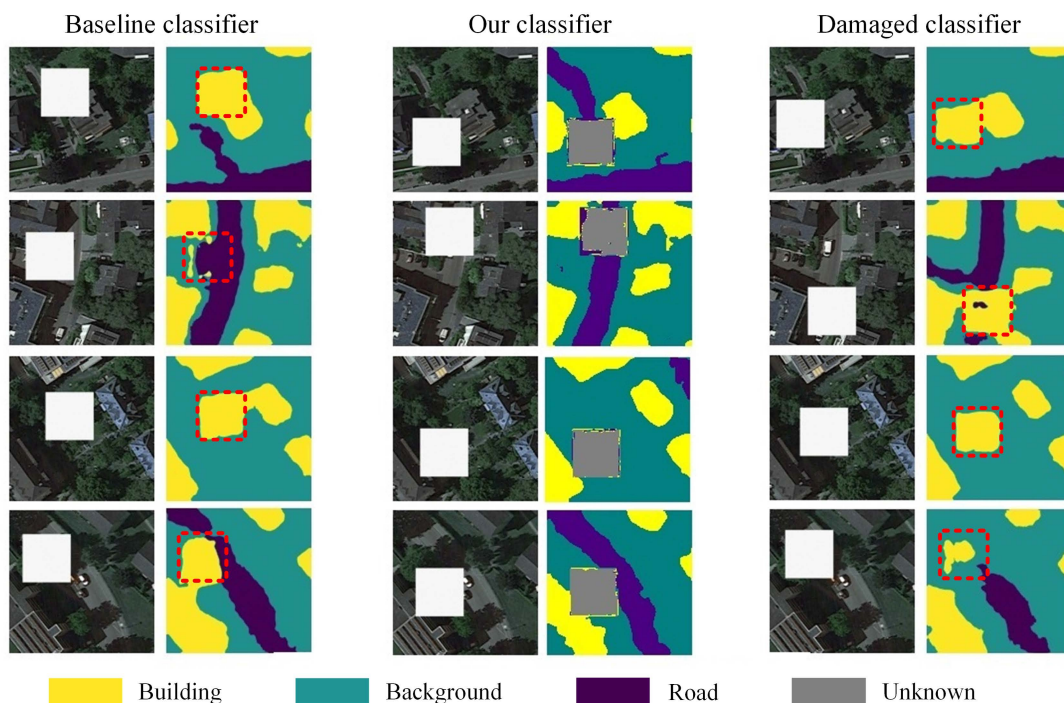


FIGURE 12. Prediction results with noisy images in the classifier.

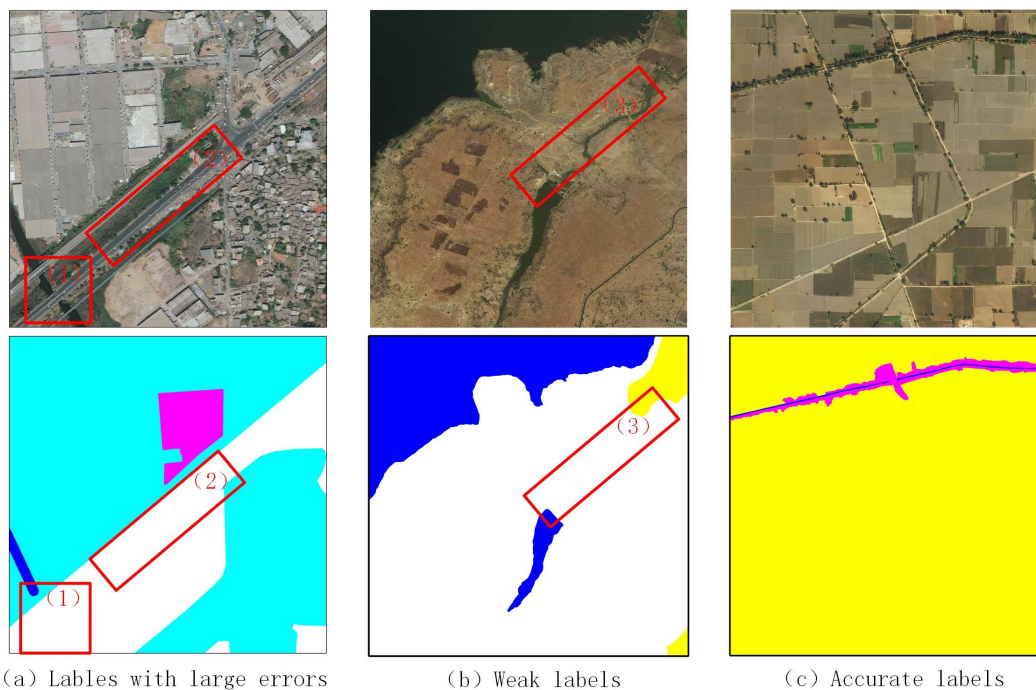


FIGURE 13. Illustration of the different levels of labeling accuracy.

The following experiments were implemented using the TensorFlow deep learning framework and performed on a computer equipped with GeForce RTX™ 2080 Ti. DeepLabV3+ was adopted as the backbone deep network, while minibatch gradient descent (MBGD) was employed as the optimization method with a batch size of 10 and a learning

rate of 0.0001. Finally, 20 training epochs were conducted for each experiment.

1) DATASET

The dataset was originally designed for a multiclass segmentation task to detect cities, agriculture areas, pastures, forests,

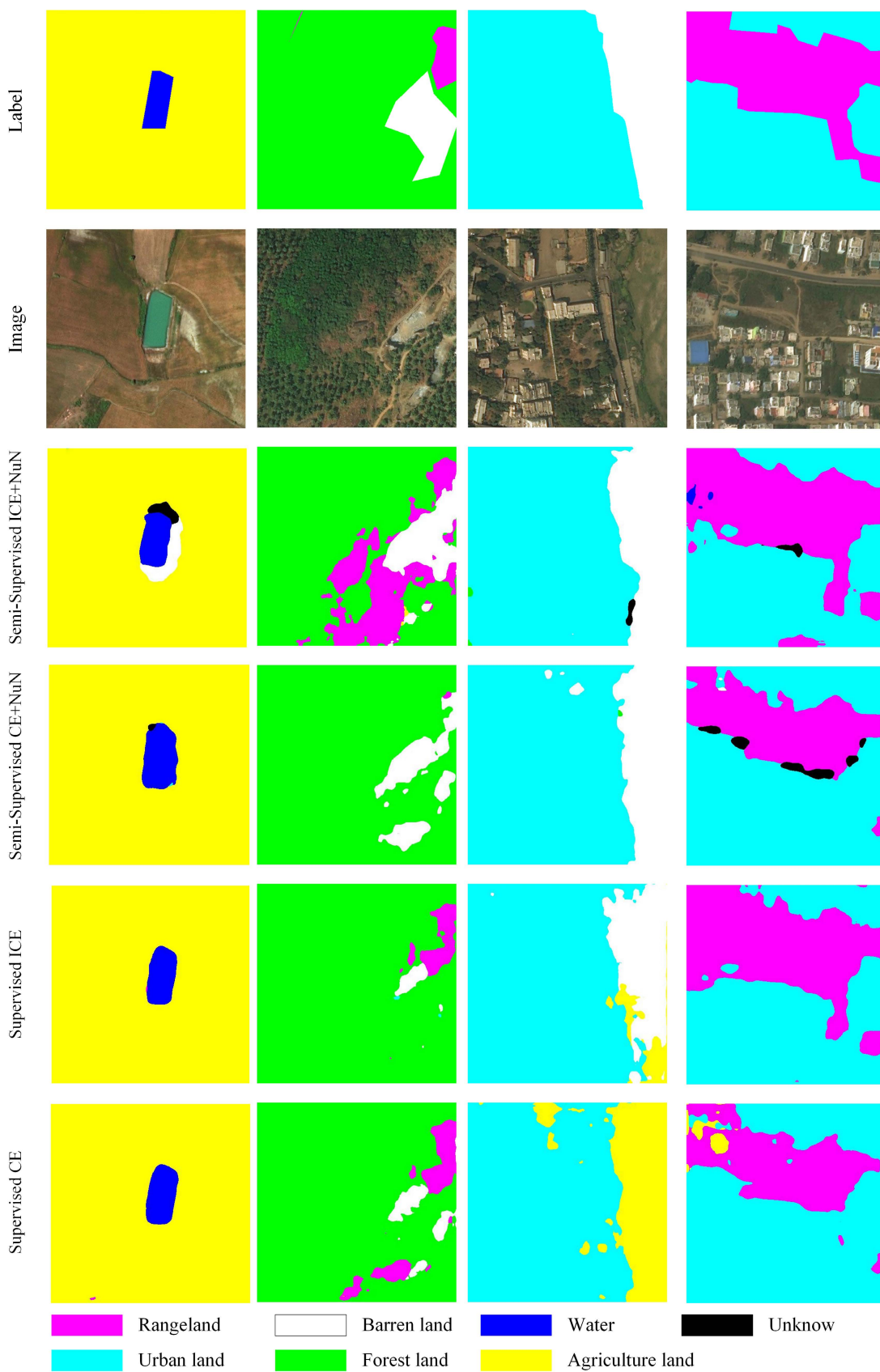


FIGURE 14. Performance comparison of different methods.

water sources, barren areas, and unknown areas. Similar to all other remote sensing datasets, DeepGlobe contains a large number of weakly labeled data. We preprocessed the DeepGlobe data by first downsampling its original image of size 2448×2448 to 512×512 pixels.

Next, we discuss the selection of 7000 downsampled images to create our training and test datasets. As illustrated in Fig. 13(a), some images suffer from large labeling errors. For instance, even though the pixels within the two boxed areas in Fig. 13(a) have similar attributes, they were divided and classified into two different classes, namely, “Water” and “Rangeland”, in the corresponding labels. Because dealing with large labeling errors is beyond the scope of this work, we excluded such images with large labeling errors from our datasets. In contrast, the weak labels in the boxed areas in Fig. 13(b) are more negligible, while the annotation in Fig. 13(c) is accurate. We included such images with either weak or accurate labels in our data sets. More specifically, we select 6500 images of 512×512 pixels and use 5000 for training and 1500 for testing. Note that these 6500 images may contain weakly labeled pixels. In addition, we manually chose 500 images with accurate labels to form another test dataset for performance analysis. In the sequel, this 500-image test set is referred to as the accurate-label test set, whereas the 1500-image test set is referred to as the weak-label test set. It should be emphasized that the classes contained in these 5000 training images are highly imbalanced, as shown in Table 1, which shows the percentages of all land-use classes in the pixels in the selected training data sets. Clearly, the “Agricultural land” class substantially outnumbers the other classes.

TABLE 1. Percentages of land-use classes in pixels.

| Classes | Percentage |
|------------------|------------|
| Urban land | 11.3% |
| Agriculture land | 58.2% |
| Rangeland | 8.7% |
| Forest land | 11.6% |
| Water | 2.7% |
| Barren land | 7.5% |

2) RESULTS

Next, we first investigate the classification results using the weak-label test set. Fig. 14 shows that the performance of “supervised CE” is not satisfactory, as it failed to distinguish “Rangeland”, which is colored pink, from “Agricultural land”, which is colored yellow; this may be caused by the imbalanced data between these two land-use classes, as shown in Table 1. In contrast, the proposed “supervised ICE” can significantly improve the classification accuracy of “Rangeland” by considering the imbalanced data problem. However, “supervised ICE” cannot properly classify the boundary regions between two adjacent classes. This shortcoming was overcome by the proposed semisupervised learning method based on the nuclear norm. An inspection of Fig. 14 suggests that “semisupervised CE+NuN” can

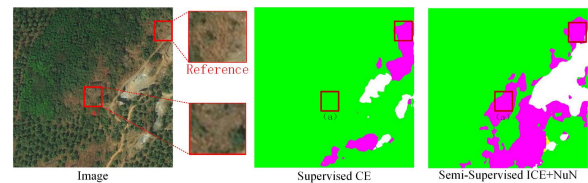


FIGURE 15. Magnified images for close-up inspection on the classification of “Barren land”.

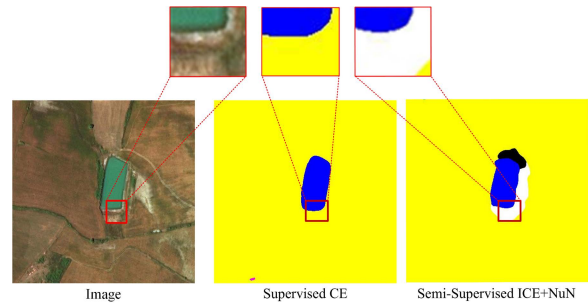


FIGURE 16. Magnified images for a close-up inspection of the classification of “Rangeland”.

better handle the boundary regions with the “unknown” class colored in black. Finally, Fig. 14 reveals that “semisupervised ICE+NuN” can further improve the classification accuracy. Because such boundary regions randomly appear in different land classes, their unrepresentative spatial characteristics confuse the learning process and slow down the convergence process. Therefore, the proposed semisupervised learning method can reduce the interference caused by these pixels by classifying them into one new land class.

Fig. 15 magnifies the bottom area near the water body of the image shown in the first column of Fig. 14. From the magnified image, we can see that the area immediately below the water body should *not* be classified as “Agricultural land.” Interestingly, the proposed “semisupervised ICE+NuN” method classified this area as “Barren land,” whereas “supervised CE” classified it as “Agricultural land.” We believe the classification of “Barren land” is more accurate, as vastly different spatial features can be observed between this area and its neighboring “Agriculture land” even by a visual inspection.

Furthermore, we can observe from the results presented in the second column of Fig. 14 that the proposed “semisupervised ICE+NuN” method has identified substantially more “Rangeland” areas than “supervised CE.” To validate this observation, we magnified the center part of the image, as shown in Fig. 16. First, we observe from the boxed area labeled “Reference” that both “Supervised CE” and “Semisupervised ICE+NuN” classified this area as “Rangeland”. This classification result also agrees well with the label, as shown in Fig. 14. In contrast, a visual inspection suggests that the boxed area labeled “(a)” actually contains very similar spatial features. However, “Supervised CE” classified this area as “Forest land”, while “Semisupervised ICE+NuN” classified this area as “Rangeland.” With

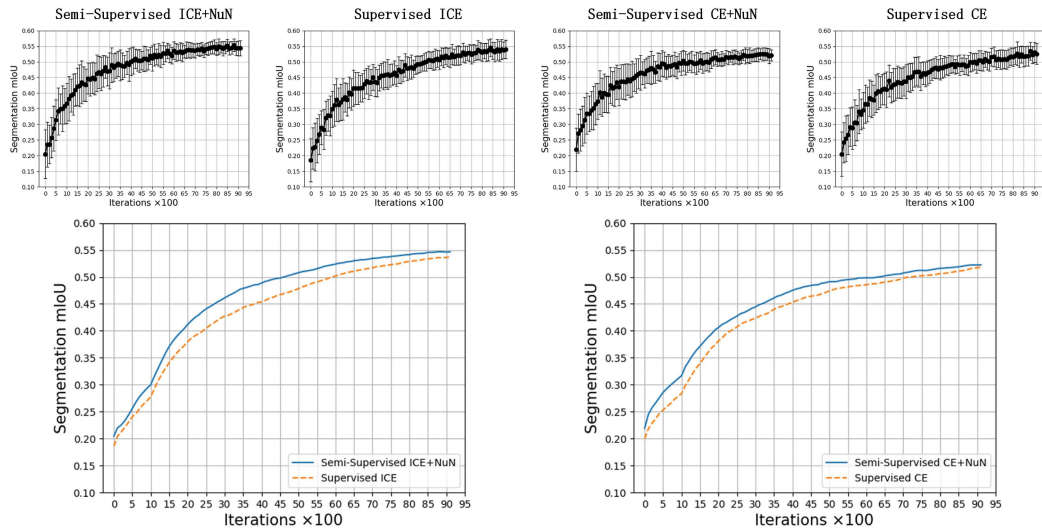


FIGURE 17. Comparison of convergence rates.

TABLE 2. Object segmentation mIoU(%) on Data-set.

| | Supervised CE | Supervised ICE | SemiSupervised CE+NuN | SemiSupervised ICE+NuN |
|------------------|---------------|----------------|-----------------------|------------------------|
| Urban land | 61.86±2.79 | 63.07±1.17 | 62.07±2.96 | 63.73±1.67 |
| Agriculture land | 72.74±1.79 | 72.51±1.62 | 72.36±4.12 | 72.49±1.77 |
| Rangeland | 34.63±4.27 | 43.46±2.72 | 37.26±3.21 | 44.12±2.23 |
| Forest land | 52.61±3.17 | 54.73±2.33 | 53.12±4.93 | 55.18±2.36 |
| Water | 40.93±3.25 | 46.88±1.11 | 41.07±3.86 | 49.3±1.83 |
| Barren land | 45.01±4.33 | 55.92±1.88 | 46.48±3.17 | 58.22±1.23 |

the information in the boxed area labeled “Reference,” we believe “Rangeland” is a more appropriate land-use classification for the boxed area labeled “(a)”. Similar observations can be made in the other figures.

Table 2 shows the mIoU performance of the four classifiers on the accurate-label test set. Thus, classifiers with ICE can more effectively lessen the adverse effects caused by imbalanced data than their counterparts with conventional CE. For instance, the mIoU for “Barren land” was improved from 43.13 (“Supervised CE”) to 56.74 (“Supervised ICE”) by exploiting ICE. Similar observations can be obtained for “semisupervised CE + NuN” and “semisupervised ICE + NuN.” Furthermore, the proposed semisupervised learning method helped further improve the recognition accuracy of the supervised learning-based classifiers. In particular, compared to the conventional “supervised CE” method, the proposed “semisupervised ICE+NuN” demonstrated impressive performance gains of the order of 10% for the three least represented land-use classes, namely, “Rangeland”, “Water”, and “Barren land”. In addition, the mIoU performance for the three most represented classes is comparable for the four classifiers on the test dataset with more accurate labels.

Fig. 17 shows the mIoU performance as a function of the iteration number for the four classifiers under consideration. An inspection of Fig. 17 reveals that the two proposed semisupervised classifiers achieved faster convergence, as the

nuclear norm can remove the interference caused by the unrepresentative features from the mixed pixels, particularly the ambiguous features arising from the junction of multiple classes. Using the unsupervised learning technique, the proposed classifier can spend less time learning invalid or even incorrect features, which shortens the training time without overfitting the proposed classifier.

Finally, we compared the performance difference of using different convolutional neural network (CNN) models, including U-Net, FCN-8s, DeepLabv3, FPN, and DeepLab3+. The quantitative results shown in Table 3 indicate that DeepLab3+ generally exhibits the best performance. In addition, regardless of the CNN model, the proposed “semisupervised ICE+NuN” model outperformed the conventional “supervised CE” model by 3% – 5%. As shown in the experiments, the proposed “NuN” cost function worked well with any existing CNN model..

C. DISCUSSIONS AND FUTURE WORK

It has been a long-standing problem that remote sensing data suffer from much larger uncertainty than data in other research areas, such as CV, which has become a major challenge for researchers applying machine learning techniques to remote sensing data. In this work, we have made an initial attempt to open a new avenue for handling uncertainty by recognizing that most pixels in remote sensing images may exhibit characteristics of multiple land-use classes.

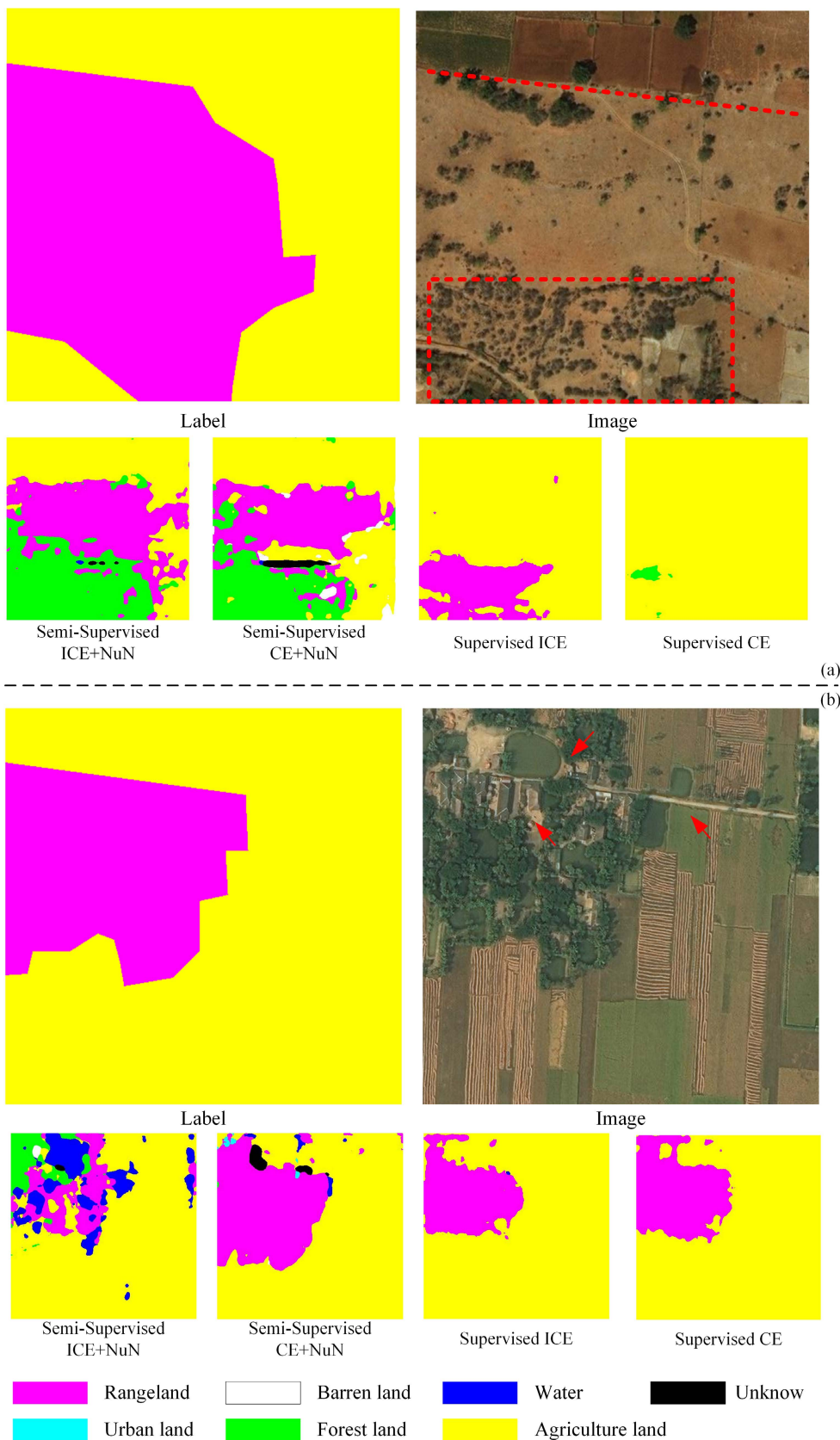


FIGURE 18. Classification by the proposed semisupervised classifier in the presence of large labeling errors.

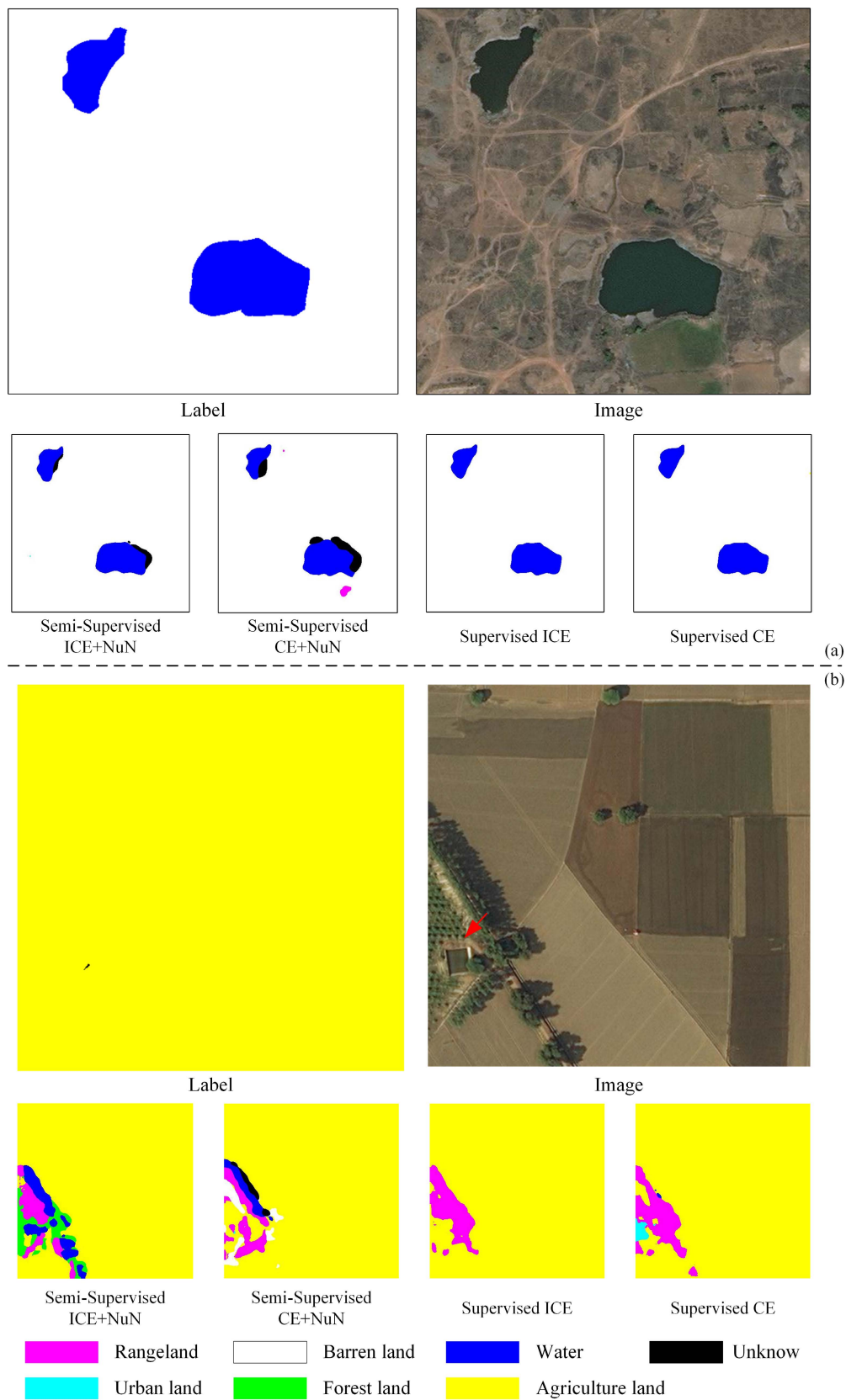


FIGURE 19. The classifier has high classification accuracy for “water,” but its generalization is not good.

TABLE 3. The application performance of the new loss function on different neural networks.

| | Method | U-Net | FCN | FPN | Deeplabv3 | Deeplabv3+ |
|------------------|---------|--------------|--------------|--------------|--------------|--------------|
| mIoU(%) | CE | 37.45± 2.346 | 38.88± 2.683 | 42.84± 2.206 | 49.02± 1.904 | 52.16± 3.089 |
| | ICE+NuN | 44.09± 1.781 | 45.95± 1.059 | 47.41± 1.137 | 51.82± 1.321 | 55.72± 1.629 |
| Urban land | CE | 47.23± 2.483 | 49.43± 1.912 | 48.43± 2.109 | 55.12± 2.039 | 61.86± 2.785 |
| | ICE+NuN | 52.62± 1.185 | 53.76± 0.880 | 50.01± 0.559 | 57.34± 1.810 | 63.72± 1.669 |
| Agriculture land | CE | 43.06± 1.973 | 46.19± 3.008 | 47.38± 1.832 | 64.51± 1.782 | 72.74± 1.791 |
| | ICE+NuN | 46.22± 1.128 | 48.62± 1.055 | 50.93± 1.807 | 67.33± 1.213 | 72.49± 1.767 |
| Rangeland | CE | 22.84± 2.872 | 23.17± 3.652 | 27.69± 2.837 | 31.52± 3.804 | 34.63± 4.274 |
| | ICE+NuN | 24.07± 2.009 | 25.23± 1.034 | 28.33± 2.019 | 37.48± 1.744 | 44.11± 2.233 |
| Forest land | CE | 43.94± 3.398 | 48.04± 2.182 | 44.27± 1.115 | 49.71± 2.551 | 52.61± 3.168 |
| | ICE+NuN | 48.74± 1.860 | 52.96± 1.267 | 49.38± 0.913 | 51.26± 1.133 | 55.18± 2.364 |
| Water | CE | 22.78± 1.937 | 24.92± 3.111 | 31.73± 1.807 | 32.48± 1.679 | 40.93± 3.247 |
| | ICE+NuN | 27.63± 2.299 | 27.6± 1.124 | 34.81± 1.628 | 37.81± 1.661 | 49.83± 1.829 |
| Barren land | CE | 33.79± 1.509 | 35.25± 1.827 | 41.52± 2.315 | 42.09± 1.204 | 45.01± 4.332 |
| | ICE+NuN | 39.16± 1.522 | 38.96± 1.646 | 47.07± 1.239 | 55.72± 0.908 | 58.22± 1.232 |

Thus, in lieu of forcibly classifying the pixel into one specific land-use class, it is more appropriate to classify the mixed pixels into multiple classes using the proposed unsupervised approach. Furthermore, if the pixels show unrepresentative characteristics, we propose classifying the pixels to an “unknown” class to accommodate these indistinguishable pixels. Thus, the proposed semisupervised classifier analyzes the uncertainty associated with each pixel before applying unsupervised learning to pixels with high uncertainty. As a result, our proposed semisupervised learning approach has a better generalization capability, more robustness, and faster convergence.

As discussed before, remote sensing images with large labeling errors are beyond the scope of this work. In the future, we plan to extend the current work to these images. Furthermore, because our proposed semisupervised learning approach can relax the stringent requirements for accurately labeled data, it may be possible for the proposed approach to further reduce its dependence on accurate labels. Fig. 18 shows some interesting observations derived from our experiments on images with large labeling errors.

More specifically, the pixels in Fig. 18(a) belong to “Agriculture land,” “Rangeland,” and “Forest land.” However, the labels corresponding to “Rangeland” and “Forest land” were largely mistaken. Similarly, the labels shown in Fig. 18(b) also exhibit large errors because the features of water, trees, and houses were ignored in the labels. These large labeling errors may mislead the training of classifiers, especially for “supervised learning,” if they are contained in a training dataset. Interestingly, even though the labels were largely mistaken, the proposed “semisupervised ICE+NuN” method was able to accurately identify the “water” pixels colored in blue and the details in the top image. Furthermore, for the natural environment shown in the middle image, transitional areas between “Forest land” and “Rangeland” were correctly recognized. This suggests that the nuclear norm can prevent the proposed classifier from overfitting, particularly when the training dataset has a large domain difference and imbalanced data.

Another interesting observation about the proposed semisupervised classifier is shown in Fig. 19, in which the class of “Water” was clearly recognized. Because the features of “Water” are vastly different from those of other land-use classes, the proposed classifier could accurately identify “Water” even with limited information provided by the dataset. However, Fig. 19 also shows that the proposed classifier was not quite able to distinguish “Shadow” from “Water”, as these two classes demonstrate very similar spatial features that are difficult to differentiate even by visual observation. This issue can be an interesting extension of this study and can be further explored.

VI. CONCLUSION

In this study, we developed a semisupervised classifier using a small set of remote sensing data with accurate labels and remote sensing data with weak labels. A weighted cross entropy-based cost function was proposed to circumvent the imbalanced data problem by utilizing the term frequency-inverse document frequency (TF-IDF) algorithm to weigh the contributions from imbalanced data of different classes. In addition, a nuclear norm-based cost function was developed to maximize the rank of the prediction matrix derived from the weakly labeled data without requiring data labels. Furthermore, an artificial class called “unknown” was created to alleviate the interference caused by weakly labeled data with unrepresentative spatial features. Extensive experiments were performed using the DeepGlobe Land Cover Classification Challenge dataset and the AIS dataset. The experimental results confirm the effectiveness of the proposed semisupervised classifier.

REFERENCES

- [1] J. R. Jensen and D. C. Cowen, “Remote sensing of urban/suburban infrastructure and socio-economic attributes,” *Photogramm. Eng. Remote Sens.*, vol. 65, pp. 611–622, May 1999.
- [2] J.-P. Donnay, M. J. Barnsley, and P. A. Longley, *Remote Sensing and Urban Analysis: GISDATA 9*. Boca Raton, FL, USA: CRC Press, 2000.
- [3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

- [4] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [5] Q. Wang, X. Zhang, G. Chen, F. Dai, Y. Gong, and K. Zhu, "Change detection based on faster R-CNN for high-resolution remote sensing images," *Remote Sens. Lett.*, vol. 9, no. 10, pp. 923–932, Oct. 2018.
- [6] A. Rakhlin, A. Davydov, and S. Nikolenko, "Land cover classification from satellite imagery with U-Net and Lovász-softmax loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 262–266.
- [7] S. Seferbekov, V. Igloukov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 272–275.
- [8] V. Mazzia, A. Khaliq, and M. Chiaberge, "Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN)," *Appl. Sci.*, vol. 10, no. 1, p. 238, Dec. 2019.
- [9] N. Venugopal, "Automatic semantic segmentation with DeepLab dilated learning network for change detection in remote sensing images," *Neural Process. Lett.*, vol. 51, no. 3, pp. 2355–2377, Jun. 2020.
- [10] V. Mithal, G. Nayak, A. Khandelwal, V. Kumar, R. Nemani, and N. Oza, "Mapping burned areas in tropical forests using a novel machine learning framework," *Remote Sens.*, vol. 10, no. 2, p. 69, Jan. 2018.
- [11] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Comput.*, vol. 8, no. 3, pp. 643–674, Apr. 1996.
- [12] C. E. Woodcock, "Uncertainty in remote sensing," *Uncertainty Remote Sens. GIS*, pp. 19–24, Dec. 2002.
- [13] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. Hoboken, NJ, USA: Wiley, 2009.
- [14] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1–9.
- [15] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 1–6, Jun. 2004.
- [16] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognit.*, vol. 57, pp. 164–178, Sep. 2016.
- [17] G. Douzas, F. Bacao, J. Fonseca, and M. Khudinyan, "Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm," *Remote Sens.*, vol. 11, no. 24, p. 3040, Dec. 2019.
- [18] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, Feb. 2018.
- [19] J. A. Benediktsson, J. Chanussot, and W. M. Moon, "Advances in very-high-resolution remote sensing," *Proc. IEEE*, vol. 101, no. 3, pp. 566–569, Mar. 2013.
- [20] R. Dong, C. Li, H. Fu, J. Wang, W. Li, Y. Yao, L. Gan, L. Yu, and P. Gong, "Improving 3-m resolution land cover mapping through efficient learning from an imperfect 10-m resolution map," *Remote Sens.*, vol. 12, no. 9, p. 1418, 2020.
- [21] X. Luo, X. Tong, Z. Hu, and G. Wu, "Improving urban land cover/use mapping by integrating a hybrid convolutional neural network and an automatic training sample expanding strategy," *Remote Sens.*, vol. 12, no. 14, p. 2292, Jul. 2020.
- [22] A. Aizawa, "An information-theoretic perspective of TF-IDF measures," *Inf. Process. Manage.*, vol. 39, no. 1, pp. 45–65, 2003.
- [23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
- [24] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
- [25] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 872–881.
- [26] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2017.
- [27] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9268–9277.
- [28] K. Cao, C. Wei, A. Gaidon, N. Arcehiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," 2019, *arXiv:1906.07413*. [Online]. Available: <http://arxiv.org/abs/1906.07413>
- [29] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan, "Equalization loss for long-tailed object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11662–11671.
- [30] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," 2019, *arXiv:1910.09217*. [Online]. Available: <http://arxiv.org/abs/1910.09217>
- [31] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9719–9728.
- [32] D. Angluin and P. Laird, "Learning from noisy examples," *Mach. Learn.*, vol. 2, no. 4, pp. 343–370, 1988.
- [33] N. Lawrence and B. Schölkopf, "Estimating a kernel Fisher discriminant in the presence of label noise," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*. San Mateo, CA, USA: Morgan Kaufmann, 2001, p. 306.
- [34] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1196–1204.
- [35] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2015.
- [36] H. Su, B. Zhao, Q. Du, P. Du, and Z. Xue, "Multifeature dictionary learning for collaborative representation classification of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2467–2484, Apr. 2018.
- [37] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2020.
- [38] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, early access, Dec. 16, 2020, doi: [10.1109/TGRS.2020.3042607](https://doi.org/10.1109/TGRS.2020.3042607).
- [39] J. Kang, R. Fernandez-Beltran, X. Kang, J. Ni, and A. Plaza, "Noise-tolerant deep neighborhood embedding for remotely sensed images with label noise," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2551–2562, 2021.
- [40] Z. Na, "Scale issues in ecology: Concepts of scale and scale analysis," *Acta Ecol. Sinica*, vol. 26, no. 7, pp. 2340–2355, 2006.
- [41] C. Song, C. Cheng, and P. Shi, "Geography complexity: New connotations of geography in the new era," *Acta Geographica Sinica*, vol. 73, no. 7, pp. 1204–1213, 2018.
- [42] P. Townsend, "A quantitative fuzzy approach to assess mapped vegetation classifications for ecological applications," *Remote Sens. Environ.*, vol. 72, no. 3, pp. 253–267, Jun. 2000.
- [43] C. J. Stubenrauch, W. B. Rossow, S. Kinne, S. Ackerman, G. Cesana, H. Chepfer, L. D. Girolamo, B. Getzewich, A. Guignard, A. Heidinger, and B. C. Maddux, "Assessment of global cloud datasets from satellites: Project and database initiated by the GEWEX radiation panel," *Bull. Amer. Meteorol. Soc.*, vol. 94, no. 7, pp. 1031–1049, Jul. 2013.
- [44] Q. Zhang and C. Xiao, "Cloud detection of RGB color aerial photographs by progressive refinement scheme," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7264–7275, Nov. 2014.
- [45] Z. Zhu, S. Wang, and C. E. Woodcock, "Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images," *Remote Sens. Environ.*, vol. 159, pp. 269–277, Mar. 2015.
- [46] Z. Li, H. Shen, H. Li, G. Xia, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery," *Remote Sens. Environ.*, vol. 191, pp. 342–358, Mar. 2017.
- [47] Z. An and Z. Shi, "Scene learning for cloud detection on remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4206–4222, Aug. 2015.
- [48] P. Li, L. Dong, H. Xiao, and M. Xu, "A cloud image detection method based on SVM vector machine," *Neurocomputing*, vol. 169, no. 2, pp. 34–42, Dec. 2015.

- [49] N. K. Greeshma, M. Baburaj, and S. N. George, "Reconstruction of cloud-contaminated satellite remote sensing images using kernel PCA-based image modelling," *Arabian J. Geosci.*, vol. 9, no. 3, p. 239, Mar. 2016.
- [50] A. Francis, P. Sidiropoulos, and J.-P. Müller, "CloudFCN: Accurate and robust cloud detection for satellite imagery with deep learning," *Remote Sens.*, vol. 11, no. 19, p. 2312, Oct. 2019.
- [51] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftgaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, May 2019.
- [52] K. Xu, K. Guan, J. Peng, Y. Luo, and S. Wang, "DeepMask: An algorithm for cloud and cloud shadow detection in optical satellite remote sensing images using deep residual network," 2019, *arXiv:1911.03607*. [Online]. Available: <http://arxiv.org/abs/1911.03607>
- [53] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning), vol. 14. San Rafael, CA, USA: Morgan & Claypool, 2009.
- [54] R. L. King and N. H. Younan, "Pixel unmixing via information of neighboring pixels," *GISci. Remote Sens.*, vol. 43, no. 4, pp. 310–322, Dec. 2006.
- [55] T. N. Carlson and D. A. Ripley, "On the relation between NDVI, fractional vegetation cover, and leaf area index," *Remote Sens. Environ.*, vol. 62, no. 3, pp. 241–252, 1997.
- [56] S. K. Mcfeeter, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *Int. J. Remote Sens.*, vol. 17, no. 7, pp. 1425–1432, 1996.
- [57] Y. Zhang, I. O. A. Odeh, and C. Han, "Bi-temporal characterization of land surface temperature in relation to impervious surface area, NDVI and NDBI, using a sub-pixel image analysis," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 11, no. 4, pp. 256–264, Aug. 2009.
- [58] X. Li, J. Wang, and A. H. Strahler, "Apparent reciprocity failure in directional reflectance of structured surfaces," *Prog. Natural Sci.-Beijing*, vol. 9, pp. 747–752, Oct. 1999.
- [59] X. W. Li, "Retrospect prospect and innovation in quantitative remote sensing," *J. Henan Univ., Natural Sci.*, vol. 35, no. 4, pp. 49–56, 2005.
- [60] B. Sun and Q. Zhou, "Expressing the spatio-temporal pattern of farmland change in arid lands using landscape metrics," *J. Arid Environ.*, vol. 124, pp. 118–127, Jan. 2016.
- [61] Y. Xu, P. Cao, Y. Kong, and Y. Wang, "L_DMI: A novel information-theoretic loss function for training deep nets to label noise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6225–6236.
- [62] A. Bahri, S. G. Majelan, S. Mohammadi, M. Noori, and K. Mohammadi, "Remote sensing image classification via improved cross-entropy loss and transfer learning strategy based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1087–1091, Jun. 2020.
- [63] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.
- [64] 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. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [65] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. Amer. Control Conf.*, vol. 6, Jun. 2001, pp. 4734–4739.
- [66] R. He, Z. Sun, T. Tan, and W.-S. Zheng, "Recovery of corrupted low-rank matrices via half-quadratic based nonconvex minimization," in *Proc. CVPR*, Jun. 2011, pp. 2889–2896.
- [67] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
- [68] Y. Liu, F. Nie, and Q. Gao, "Nuclear-norm based semi-supervised multiple labels learning," *Neurocomputing*, vol. 275, pp. 940–947, Jan. 2018.
- [69] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3941–3950.
- [70] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.
- [71] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*. [Online]. Available: <http://arxiv.org/abs/1805.09501>
- [72] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," 2019, *arXiv:1905.02249*. [Online]. Available: <http://arxiv.org/abs/1905.02249>
- [73] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2019, *arXiv:1904.12848*. [Online]. Available: <http://arxiv.org/abs/1904.12848>
- [74] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," 2020, *arXiv:2001.07685*. [Online]. Available: <http://arxiv.org/abs/2001.07685>
- [75] S. Roy, E. Sangineto, N. Sebe, and B. Demir, "Semantic-fusion gans for semi-supervised satellite image classification," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 684–688.
- [76] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 23–43, Nov. 2018.
- [77] X. Dai, X. Wu, B. Wang, and L. Zhang, "Semisupervised scene classification for remote sensing images: A method based on convolutional neural networks and ensemble learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 869–873, Jun. 2019.
- [78] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, early access, Dec. 7, 2021, doi: [10.1109/LGRS.2020.3038420](https://doi.org/10.1109/LGRS.2020.3038420).
- [79] K. Zhang and H. Yang, "Semi-supervised multi-spectral land cover classification with multi-attention and adaptive kernel," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1881–1885.
- [80] T. W. Cenggoro, S. M. Isa, G. P. Kusuma, and B. Pardamean, "Classification of imbalanced land-use/land-cover data using variational semi-supervised learning," in *Proc. Int. Conf. Innov. Creative Inf. Technol. (ICITech)*, Nov. 2017, pp. 1–6.
- [81] S. Fang, D. Quan, S. Wang, L. Zhang, and L. Zhou, "A two-branch network with semi-supervised learning for hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 3860–3863.
- [82] J. Liu, Q. Feng, Y. Wang, B. Batsaikhan, J. Gong, Y. Li, C. Liu, and Y. Ma, "Urban green plastic cover mapping based on VHR remote sensing images and a deep semi-supervised learning framework," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 9, p. 527, Sep. 2020.
- [83] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, p. 371, Jan. 2021.



RUI WANG received the Ph.D. degree from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. His research interests include the use of big data and remote sensing technology to time-series reservoir water changes, remote sensing data management methods for computational analysis, and the application of deep learning techniques in water monitoring.



MAN-ON PUN (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, USA, in 2006.

He was a Postdoctoral Research Associate at Princeton University, from 2006 to 2008. Currently, he is an Associate Professor with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen (CUHKSZ). Prior to joining CUHKSZ, in 2015, he held research positions at Huawei, USA; Mitsubishi Electric Research Laboratories (MERL), Boston; and Sony, Tokyo, Japan. His research interests include the AI Internet of Things (AIoT) and applications of machine learning in communications and satellite remote sensing.

...