Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification

Article *in* IEEE Transactions on Cybernetics · May 2020

DOI: 10.1109/TCYB.2020.2989241

3 authors, including:

Yansheng Li
Wuhan University
111 PUBLICATIONS **3,869** CITATIONS

SEE PROFILE

Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification

Yansheng Li, Yongjun Zhang, *Member, IEEE*, and Zhihui Zhu, *Member, IEEE*

***Abstract*—Due to its various application potentials, the remote sensing image scene classification (RSSC) has attracted a broad range of interests. While the deep convolutional neural network (CNN) has recently achieved tremendous success in RSSC, its superior performances highly depend on a large number of accurately labeled samples which require lots of time and manpower to generate for a large-scale remote sensing image scene dataset. In contrast, it is not only relatively easy to collect coarse and noisy labels but also inevitable to introduce label noise when collecting large-scale annotated data in the remote sensing scenario. Therefore, it is of great practical importance to robustly learn a superior CNN-based classification model from the remote sensing image scene dataset containing non-negligible or even significant error labels. To this end, this article proposes a new RSSC-oriented error-tolerant deep learning (RSSC-ETDL) approach to mitigate the adverse effect of incorrect labels of the remote sensing image scene dataset. In our proposed RSSC-ETDL method, learning multiview CNNs and correcting error labels are alternatively conducted in an iterative manner. It is noted that to make the alternative scheme work effectively, we propose a novel adaptive multifeature collaborative representation classifier (AMF-CRC) that benefits from adaptively combining multiple features of CNNs to correct the labels of uncertain samples. To quantitatively evaluate the performance of error-tolerant methods in the remote sensing domain, we construct remote sensing image scene datasets with: 1) simulated noisy labels by corrupting the open datasets with varying error rates and 2) real noisy labels by deploying the greedy annotation strategies that are practically used to accelerate the process of annotating remote sensing image scene datasets. Extensive experiments on these datasets demonstrate that our proposed RSSC-ETDL approach outperforms the state-of-the-art approaches.**

***Index Terms*—Adaptive multifeature collaborative representation classifier (AMF-CRC), corrupted labels, remote sensing image scene classification (RSSC), RSSC-oriented error-tolerant deep learning (RSSC-ETDL).**

Manuscript received December 4, 2019; accepted April 15, 2020. Date of publication May 14, 2020; date of current version March 17, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505003, in part by the National Natural Science Foundation of China under Grant 41971284, in part by the China Postdoctoral Science Foundation under Grant 2016M590716 and Grant 2017T100581, and in part by the Hubei Provincial Natural Science Foundation of China under Grant 2018CFB501. This article was recommended by Associate Editor M. Han. *(Corresponding author: Yongjun Zhang.)*

Yansheng Li and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; zhangyj@whu.edu.cn).

Zhihui Zhu is with the Electrical and Computer Engineering, University of Denver, Denver, CO 80208 USA (e-mail: zhihui.zhu@du.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2020.2989241.

Digital Object Identifier 10.1109/TCYB.2020.2989241

I. Introduction

DIFFERENT from the traditional pixel-level remote sensing imagery classification [1]–[5], the remote sensing image scene classification (RSSC) [6]–[10], which aims at predicting the semantic category of one scene (i.e., one image block) through perceiving the objects in the scene and their spatial topology, benefits many applications, such as geospatial object detection [11], [12]; content-based remote sensing image retrieval [13], [14]; and so forth. In the literature, lots of approaches have been proposed to cope with RSSC, such as the classical methods [6], [7] and the deep-learning-based methods [8]–[10]. In particular, as one representative of deep learning [10], [15], [16], the deep convolutional neural network (CNN) not only achieves tremendous success in computer vision [17]–[19], speech recognition [20], and natural language processing [21], but also dramatically improves the performance of the RSSC task [10]. Because of the fact that CNN obviously outperforms the traditional RSSC methods based on the combination of traditional handcrafted features and shallow classifiers, this article mainly focuses on discussing the deep-learning-based RSSC technique.

In the remote sensing big data era, it is quite easy to collect the remote sensing data itself, but labeling these data from scratch becomes the real challenge as the state-of-the-art RSSC methods (e.g., deep-learning-based methods) often need a large-scale labeled dataset. To accelerate the process of annotating remote sensing image scene datasets, two categories of greedy annotation approaches [22]–[25] have been developed. In the former category of approaches [22], [23], a large amount of remote sensing image scenes in the original dataset is first aggregated into a small number of clusters by unsupervised algorithms, and then the dataset is manually labeled one cluster by one cluster instead of one scene by one scene that is adopted by the traditional manual annotation strategy. With the aid of the geographic coordinate registration between the remote sensing imagery and geospatial data, the latter category of methods [24], [25] utilizes the crowdsourcing geospatial semantic information (e.g., the semantic tags in the Google Map and the point of interests (POI) in the OpenStreetMap) or the accumulated geodatabases (e.g., the released global land cover (GLC) product) to label the remote sensing image scenes. Both of these two categories of methods indeed save lots of manual annotation labor, but inevitably bring in some error labels. Throughout this article, a noisy label or error label refers to a wrong label in the sense that the corresponding scene is labeled to other class instead of the true one. Unfortunately, the value of the greedily collected

remote sensing image scene datasets with label noise has not been realized and mined.

As an open problem, the label errors in the classification dataset would inevitably degenerate all kinds of supervised learning classifiers, including shallow classifiers and deep learning models. As is well known, deep learning models often contain a large number of hyperparameters to be learned from data [10], [15]. As a consequence, when the labels of the classification dataset are corrupted, the performance degeneration issue of deep learning models becomes more severe than the shallow classifiers [26]. With this consideration, this article mainly aims to pursue the excellent performance of deep learning models even under the supervision of remote sensing image scene datasets with corrupted labels.

In literature, pioneers in the computer vision domain have developed many error-tolerant deep learning (ETDL) methods to alleviate the adverse effect of inaccurate labels of the natural image object recognition datasets from the Web resource. Specifically, the existing ETDL methods can be coarsely divided into two major categories: 1) label-noise-minimization methods [27]–[32] and 2) label-noise-correction methods [33], [34]. However, there was very little research work on the classification of remote sensing image scenes under error labels. Compared with natural images, remote sensing images often present additional challenging characteristics [35], including large resolution variations, arbitrary orientations, and dense structures. In addition, compared with object detection, scene classification is more difficult because it not only needs to recognize the objects in the scene but also requires perceiving the spatial layout among objects. Hence, it is inadequate to address the RSSC problem under the influence of error labels by directly applying the existing ETDL methods for natural image object recognition [27]–[34]. Thus, it becomes crucial to develop the RSSC-oriented ETDL (RSSC-ETDL) technique to fully exploit the potential value of the low-cost remote sensing image scene datasets with noisy labels.

With these aforementioned considerations, this article proposes a novel RSSC-ETDL approach to robustly learn a superior RSSC model from the corrupted remote sensing image scene dataset. More specifically, the proposed RSSC-ETDL approach is conducted in an iterative scheme and each iteration step consists of learning multiview CNNs using multiple nonoverlapped subdatasets which are randomly sampled from the original dataset (or the iteratively refined dataset) and correcting the potential error labels of the original dataset by employing the learned CNNs. To pursue the high effectiveness of the error correction process, a novel adaptive multifeature collaborative representation classifier (AMF-CRC), which takes the intermediate features of the learned CNNs as the input, is proposed to cleanse the corrupted dataset. In return, with more data correctly labeled, we can learn better CNNs which then help to lift the performance of the error correction process as the input of AMF-CRC is improved. It is worth noting that our proposed AMF-CRC can adaptively employ the intermediate features of CNNs based on their relative importance. Compared with the existing multiview ETDL methods (e.g., iterative cross learning (ICL) in [34]), the recommended robust error correction module in our RSSC-ETDL approach is the key to lift the overall performance. Extensive experiments on the remote sensing image scene datasets with noisy labels show that the proposed RSSC-ETDL approach significantly outperforms the existing methods. As an auxiliary output of our RSSC-ETDL approach, the corrected training dataset could be taken as the fuel of all supervised learning methods. The main contributions of this article can be summarized as follows.

1) This article proposes a novel RSSC-ETDL framework, which adopts a popular multiview manner to monitor the uncertain labels and recommends a relearning module to correct the uncertain labels where the relearning module is highly flexible and can be implemented by any shallow feature classifier based on the deep feature representations of the learned multiview CNNs.
2) To improve the error correction performance, we propose a new AMF-CRC that can adaptively combine multiple intermediate features of CNNs with optimally learning the combination coefficients. In addition, we give a strict mathematical proof of the optimization convergence of the proposed AMF-CRC.
3) The effectiveness of our RSSC-ETDL approach has been verified on multiple remote sensing image scene datasets with kinds of error labels, including simulated and real noisy labels.

The remainder of this article is organized as follows. Section II specifically introduces the related work in the literature. Section III presents the RSSC-ETDL approach for RSSC when the labels of the training dataset are corrupted by various reasons and depicts the details of the proposed AMF-CRC algorithm. Section IV depicts the experimental results in detail. Finally, Section V gives the conclusions of this article.

II. Related Work

In this section, we review the related work from two aspects: 1) the ETDL methods in the computer vision field and 2) the error-tolerant classification approaches in the remote sensing domain.

A. ETDL Methods in the Computer Vision Field

In the computer vision domain, lots of ETDL methods have been proposed to learn deep networks from natural image datasets containing some error labels. Generally speaking, the existing ETDL methods can be coarsely divided into two major categories: 1) label-noise-minimization methods and 2) label-noise-correction methods. The former adopts the bootstrapping strategy [27], the noise modeling [28], and the dropout regularization [29], which aims at minimizing the adverse effect of label errors to train deep networks. In recent years, Ghosh *et al.* [30] theoretically showed the mean absolute error (MAE) can be robust against the uniform and asymmetric label noise. To improve the robustness under label noise and decrease the convergence time, Zhang and Sabuncu proposed a generalized cross-entropy loss (i.e., the Lq loss in [31]). To minimize the label noise memory of deep networks under the Lq loss, a cross-training method [32] was proposed

to gradually learn deep networks from data with error labels. In contrast, the latter tries to correct the potential label errors and the refined dataset is further adopted to train deep networks. For example, the sentiments of adjective–noun pairs and tags are used to refine the labels of noisy datasets from social networks, and the deep learning model is further trained on the refined dataset [33]. In an iterative manner, multiple deep networks are first trained using nonoverlapped subdatasets and then used to cleanse the original dataset [34]. As a whole, these ETDL methods have achieved a certain extent of success, but they still cannot well address the learning problem from the remote sensing image dataset with label noise as the remote sensing imagery shows a more complex structure compared with natural images. Consequently, robustly learning deep networks from remote sensing image datasets with noisy labels requires specific exploitation.

B. Error-Tolerant Classification Approaches in the Remote Sensing Domain

As we move into the era of remote sensing big data, big challenges arise on data acquisition and analysis [36]. As one distinctive characteristic (i.e., veracity) of the remote sensing big data, remote sensing image datasets with noisy labels grow at an alarming rate. Without any special modification, CNN is used to segment objects from remote sensing images under the noisy supervision (i.e., the crowd-sourced maps) [37]. Even though CNN is directly trained under the supervision of noisy labels, it also performs much better than many baselines that fully reveals the great value of remote sensing image datasets with noisy labels. But on the other hand, the improved performance can be rationally expected if CNN is tuned by some error-tolerant skills. Afterward, researchers analyze the effect of noisy labels on classification performance on satellite image time series [38] and hyperspectral imagery [39]–[46]. To alleviate the adverse effect of noisy labels on RSSC, Jian *et al.* [47] and Damodaran *et al.* [48] proposed loss functions to learn the improved classification model. In summary, all of the existing error-tolerant methods [47], [48] for RSSC are designed from the label-noise-minimization perspective. Generally speaking, the label-noise-minimization method is often designed for a specific model so that this kind of algorithm lacks universality. Hence, combining the label-noise-correction strategy and deep learning is a promising way to address RSSC under noisy labels and deserves much more exploitation.

III. Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification

To benefit clarifying the RSSC-ETDL method, we first depict the AMF-CRC in Section III-A. Based on the AMF-CRC, Section III-B further introduces the RSSC-ETDL approach in detail. Finally, Section III-C introduces the RSSC process based on the RSSC-ETDL approach.

A. Adaptive Multifeature Collaborative Representation Classifier

Supposing that we have owned the feature extraction function (e.g., the handcrafted feature descriptor or the fully connected layer output of one pretrained deep network), the RSSC task predigests into the feature classification problem. Naturally, feature classification can be addressed by the classic support vector machine (SVM) [49], which is famous for its stable performance even under the supervision of a small set of labeled samples. However, when the number of labeled samples further decreases, the performance of SVM may dramatically degenerate. Considering the intensive demand (i.e., learning a robust feature classifier under the supervision of a limited number of labeled samples) of the proposed RSSC-ETDL framework in this article, we need to exploit a more privileged feature classifier. From the representation perspective, the collaborative representation classifier (CRC) [50], [51] is proposed to cope with the more challenging small-sample-size problem. To pursue a faithful representation of the testing sample, CRC encourages samples from all classes to collaboratively represent the testing sample. However, the existing CRC [50], [51] mainly considers the single feature case. In many practical applications, one sample can be depicted by multiple heterogeneous features. To exploit the effectiveness of multiple features, a multifeature dictionary learning-based CRC (MDLCRC) approach was presented in [52]. In MDLCRC, all the features are treated equally. However, in our case, the features from different layers of CNNs instead can have different importance in classification. As a consequence, MDLCRC may not fully exploit the complementary effectiveness of different features. Different from MDLCRC, our AMF-CRC can automatically estimate the contribution rates of different features. Intuitively, the advanced performance of CRC can be rationally expected by adaptively combining multiple heterogeneous features. In the following, we introduce AMF-CRC in detail.

Assume we have M feature generators. The training dataset includes N remote sensing image scenes with C classes, where each remote sensing image scene can be represented by types of features. Let $X = \{X^1, X^2, \ldots, X^M\}$ stand for the feature set of the remote sensing image scenes in the training dataset, where $X^v \in \mathbb{R}^{d_v \times N}$ denotes the feature matrix of the training remote sensing image scenes using the vth kind of feature, d_v stands for the feature dimension of the vth kind of feature, and N is the number of the training remote sensing image scenes. In addition, each kind of feature matrix $X^v \in \mathbb{R}^{d_v \times N}$ also follows the arrangement that $X^v = [X_1^v, X_2^v, \ldots, X_C^v]$, where X_i^v means the vth kind of feature matrix of the remote sensing image scenes of the ith class, and each column of X_i^v stands for the vth kind of feature vector of one remote sensing image scene of the ith class.

Let $y = \{y_1, y_2, \ldots, y_M\}$ denote the feature set of one testing remote sensing image scene, where $y_v \in \mathbb{R}^{d_v \times 1}$ denotes the vth kind of feature and d_v stands for the feature dimension of the vth kind of feature. The proposed AMF-CRC can recover the label of $y = \{y_1, y_2, \ldots, y_M\}$ based on $X = \{X^1, X^2, \ldots, X^M\}$ by the following three steps.

1) Calculating the Representation Coefficient Vector: Given one testing remote sensing image scene, based on M feature generators, we can calculate its corresponding feature set $y = \{y_1, y_2, \ldots, y_M\}$. If we want to obtain the label of the testing feature set $y = \{y_1, y_2, \ldots, y_M\}$, we can calculate its

representation coefficient vector $\rho \in \mathbb{R}^{N\times 1}$ along the training feature set $X = \{X^1, X^2, \ldots, X^M\}$ by optimizing the following loss function:

$$\max_{\rho,w} f(\rho, w): \sum_{v=1}^{M} w_v \frac{\|y_v - X^v\rho\|_2^2}{d_v} + \alpha\|w\|_2^2 + \beta\|\rho\|_2^2$$
$$\text{s.t. } 0 \le w_v \le 1, \sum_{v=1}^{M} w_v = 1 \quad (1)$$

where $w = [w_1, w_2, \ldots, w_M]$ stands for the weight vector of different features. α and β are regularization constants.

As depicted in (1), the feature weight vector $\boldsymbol{w}$ and representation coefficient vector $\boldsymbol{\rho}$ are jointly optimized, leading to a nonconvex problem which does not have simple closed-form solution. Fortunately, note that once the feature weight vector $\boldsymbol{w}$ in (1) is determined, a closed-form solution can be obtained to calculate the representation coefficient vector. After the representation coefficient vector $\boldsymbol{\rho}$ is updated and determined, the objective function in terms of the feature weight vector is a quadratic programming which can be efficiently optimized by the classic constrained quadratic optimization algorithm. With this consideration, we adopt an iteratively alternating optimization method to determine the feature weight vector and representation coefficient vector. In particular, in the kth step, we update the representation coefficient vector by

$$\begin{aligned}\rho(k) &= \arg\min_{\rho} f(\rho, w(k-1)) \\ &= \left(\sum_{v=1}^{M} w_v(k-1)\frac{(X^v)^T X^v}{d_v} + \beta I\right)^{-1} \\ &\quad \times \left(\sum_{v=1}^{M} w_v(k-1)\frac{(X^v)^T y^v}{d_v}\right) \end{aligned} \quad (2)$$

where I denotes the identity matrix.

The feature weight vector is then updated by

$$\begin{aligned} w(k) &= \arg\min f(\rho(k), w) \\ \text{s.t.} \quad & 0 \le w_v \le 1, \sum_{v=1}^{M} w_v = 1. \end{aligned} \quad (3)$$

It is noted that (3) can be efficiently solved by using off-the-shelf quadratic programming solvers. The full procedure is depicted in Algorithm 1. The following theorem establishes the convergence guarantee for Algorithm 1.

Theorem 1 (Convergence of Algorithm 1): Let $(\rho(k), w(k))$ be the sequence generated by Algorithm 1. Then, the sequence $(\rho(k), w(k))$ satisfies the following properties.

1) The sequence is regular and obeys sufficient decrease

$$\begin{aligned} &f(\rho(k), w(k)) - f(\rho(k+1), w(k+1)) \\ &\quad \ge \alpha\|w(k) - w(k+1)\|^2 + \beta\|\rho(k) - \rho(k+1)\|^2. \end{aligned} \quad (4)$$

2) The sequence is bounded and converges to a stationary point of (1).

The proof of Theorem 1 is in the Appendix. In words, Theorem 1 guarantees that the alternating minimization algorithm (i.e., Algorithm 1) finds a stationary point of (1). We empirically observe that this alternative optimization algorithm has rapid convergence speed, with a few iterations giving reasonably good feature weight vector and representation coefficient vector. More specifically, the number of iterations (i.e., K in Algorithm 1) is empirically set to 3 in our implementation.

Algorithm 1 Alternating Minimization for Solving (1)

Input: The training feature set $X = \{X^1, X^2, \cdots, X^M\}$, the testing feature set $y = \{y_1, y_2, \cdots, y_M\}$

Output: The optimized the feature weight vector $\overline{w}$, the optimized representation coefficient vector $\overline{\rho}$.

Initialization: $w(0) = [w_1(0), w_2(0), \ldots, w_M(0)]$ (e.g., $w_1(0) = \cdots = w_M(0) = 1/M$)

for $k = 1 : K$ **do**

- Update the representation coefficient vector by $\rho(k) = \arg\min_{\rho} f(\rho, w(k-1))$.
- Update the feature weight vector by $w(k) = \arg\min_{w} f(\rho(k), w)$.

end for

$\overline{w} = w(K), \overline{\rho} = \rho(K)$

2) Calculating the Reconstruction Residuals: Based on the optimized feature weight vector $\overline{w}$ and representation coefficient vector $\overline{\rho}$, the class-specific reconstruction residual of the testing feature set can be formulated as

$$R_i = \frac{\sum_{v=1}^{M} \overline{w}_v \cdot \left(\|y^v - X_i^v \cdot \overline{\rho}_i\|_2^2 / d_v\right)}{\|\overline{\rho}_i\|_2^2} \quad (5)$$

where i denotes the class index and $i = 1, 2, \ldots, C$; and $\overline{\rho}_i$ denotes the representation coefficient subvector with respect to remote sensing image scenes of the ith class.

3) Predicting the Label of the Testing Remote Sensing Image Scene: The label of the testing remote sensing image scene can be inferred from the class-specific reconstruction residuals

$$t(y) = \arg\min_{i} \{R_i\} \quad (6)$$

where $t(y)$ denotes the label of the testing remote sensing image scene.

B. RSSC-Oriented Error-Tolerant Deep Learning Approach

Let $\Gamma_R = \{(I_1, O_1), (I_2, O_2), \ldots, (I_r, O_r)\}$ denote the original training remote sensing image scene dataset, where r denotes the number of remote sensing image scenes in the original training remote sensing image scene dataset, I stands for the remote sensing image scene, and O denotes the remote sensing image scene label which may be incorrect. To robustly learn high-quality deep networks under the supervision of the training remote sensing image scene dataset with noisy labels, we propose a new RSSC-ETDL framework, which is visually illustrated in Fig. 1. Overall, our proposed RSSC-ETDL method is designed based on the assumption that the deep network can learn useful information even with noisy labels [34]. As depicted in Fig. 1, our proposed RSSC-ETDL is conducted via an iteration manner where each iteration step

includes two alternative modules: 1) learning multiview deep networks and 2) correcting potential error labels. In addition, these two modules are detailed in the following.

1) Learning Multiview Deep Networks: To benefit perceiving and correcting error labels, this substep aims at learning multiview deep networks which could discriminate remote sensing image scenes from different perspectives. First, the original training dataset is randomly partitioned into Z nonoverlapped subdatasets $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_Z\}$, where $\Gamma_R = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_Z$ and $\Gamma_i \cap \Gamma_j = \varnothing; i = 1, 2, \ldots, Z; j = 1, 2, \ldots, Z$. Second, we learn Z different CNN models on Z subdatasets, respectively. Here, the architecture of Z CNN models follows the same style and is visually illustrated in Fig. 2. In addition, we conduct the learning process on each subdataset just like the normal deep learning case and the hyperparameters of Z CNN models are represented by $\{\Phi_1, \Phi_2, \ldots, \Phi_Z\}$.

2) Correcting Potential Error Labels: To correct potential error labels, we first use the learned Z CNN models to identify the samples with certain labels (i.e., the labels of these samples are correct with a high probability) from the original training RS image scene dataset $\Gamma_R = \{(I_1, O_1), (I_2, O_2), \ldots, (I_{\text{tr}}, O_{\text{tr}})\}$. More specifically, if one sample in the original training dataset is predicted with the same label by all of Z CNN models, this sample seems to have a certain label and is moved to the strong dataset along with its label; otherwise, the label of this sample is probably incorrect and this sample is moved to the weak dataset along with its label. The strong dataset is depicted by $\Gamma_S = \{(I_1, O_1), (I_2, O_2), \ldots, (I_{sn}, O_{sn})\}$ with sn samples and the weak dataset is denoted by $\Gamma_W = \{(I_1, O_1), (I_2, O_2), \ldots, (I_{wn}, O_{wn})\}$ with wn samples, where $\Gamma_R = \Gamma_s \cup \Gamma_w$ and $r = sn + wn$.

Here, we have obtained Z trained CNN models, the strong dataset, and the weak dataset. That means we have feature function (i.e., the fully connected layer output of CNN models) and the label supervision as the samples in the strong dataset are assumed to have the correct labels. Furthermore, we train feature classifier under the supervision of the strong dataset to predict the label of samples in the weak dataset. More specifically, as depicted in Fig. 2, the adopted deep network architecture has three fully connected layers (i.e., FC1, FC2, and FC3 in Fig. 2). Hence, on the basis of CNN models, each remote sensing image scene can be represented by $3 \times Z$ feature vectors. Based on the AMF-CRC introduced in Section III-A, the label of samples in the weak dataset is recovered via a classification way instead of inheriting the original noisy labels. As depicted in Fig. 1, the union of the strong dataset with original labels and the weak dataset with predicted labels is taken as the dataset with corrected labels, which is utilized to train deep networks in the next iteration.

To facilitate understanding, we summarize our proposed RSSC-ETDL approach in Algorithm 2. As depicted in Algorithm 2, the strong dataset Γ_S is fixed after the first round iteration which mainly aims at avoiding the error propagation. In the RSSC-ETDL method, the number of iterations is quantitatively analyzed in Section IV. Given one remote sensing image scene dataset Γ_R with noisy labels, the proposed RSSC-ETDL approach outputs Z high-quality CNN models with hyperparameters $\{\Phi_1, \Phi_2, \ldots, \Phi_Z\}$ as well as the dataset Γ_C with refined labels.

Algorithm 2 RSSC-ETDL Approach Based on AMF-CRC

Input: The original training remote sensing image scene dataset $\Gamma_R = \{(I_1, O_1), (I_2, O_2), \ldots, (I_r, O_r)\}$

Output: The hyper-parameters $\{\Phi_1, \Phi_2, \ldots, \Phi_Z\}$ of Z CNN models

Initialization: The corrected remote sensing image scene dataset $\Gamma_C = \Gamma_R$; The strong dataset $\Gamma_S = \varnothing$; The weak dataset $\Gamma_W = \varnothing$

for *iterID* $= 1$: *maxIter* **do**

- Randomly split the corrected dataset Γ_C into Z sub-datasets $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_Z\}$
- **for** *viewID=1* : Z **do**
  Learn the hyper-parameter Φ_{viewID} of CNN on the sub-dataset Γ_{viewID}.
  end for
- **if** iterID==1
  Split the original dataset Γ_R into the strong dataset Γ_S and the weak dataset Γ_W via the vote of the learned Z CNN models with the corresponding hyper-parameters $\{\Phi_1, \Phi_2, \ldots, \Phi_Z\}$
  end
- Utilize AMF-CRC to correct the label of each sample in Γ_W under the supervision of Γ_S where the feature extraction function is conducted by the fully connected layer output of the learned Z CNN models with the corresponding hyper-parameters $\{\Phi_1, \Phi_2, \ldots, \Phi_Z\}$.
- Update the remote sensing image scene dataset $\Gamma_C = \Gamma_S \cup \Gamma_W^{'}$ where $\Gamma_W^{'}$ denotes the refined one of Γ_W.

end for

C. Learning Deep Networks Under Noisy Labels for Remote Sensing Image Scene Classification

As aforementioned, the proposed RSSC-ETDL approach can automatically learn high-quality CNN models from the remote sensing image scene dataset with noisy labels. It is assumed that the RSSC-ETDL approach has learned Z CNN models with hyperparameters $\{\Phi_1, \Phi_2, \ldots, \Phi_Z\}$. As depicted in Algorithm 2, these Z CNN models are trained under the supervision of subdatasets which are not overlapped with each other. As a general deduction, these Z CNN models should own the complementary prediction performance. With this consideration, the label of one testing remote sensing image scene I can be predicted by the vote of multiview complementary CNN models

$$t = \arg\max_{c} \left(\sum_{d=1}^{Z} V_d^c \right) \tag{7}$$

where c and d stand for the category component and the CNN model index; and $V_d = \Psi(I; \Phi_d) \in \mathbb{R}^{T \times 1}$ denotes the softmax layer output of the testing remote sensing image scene I using

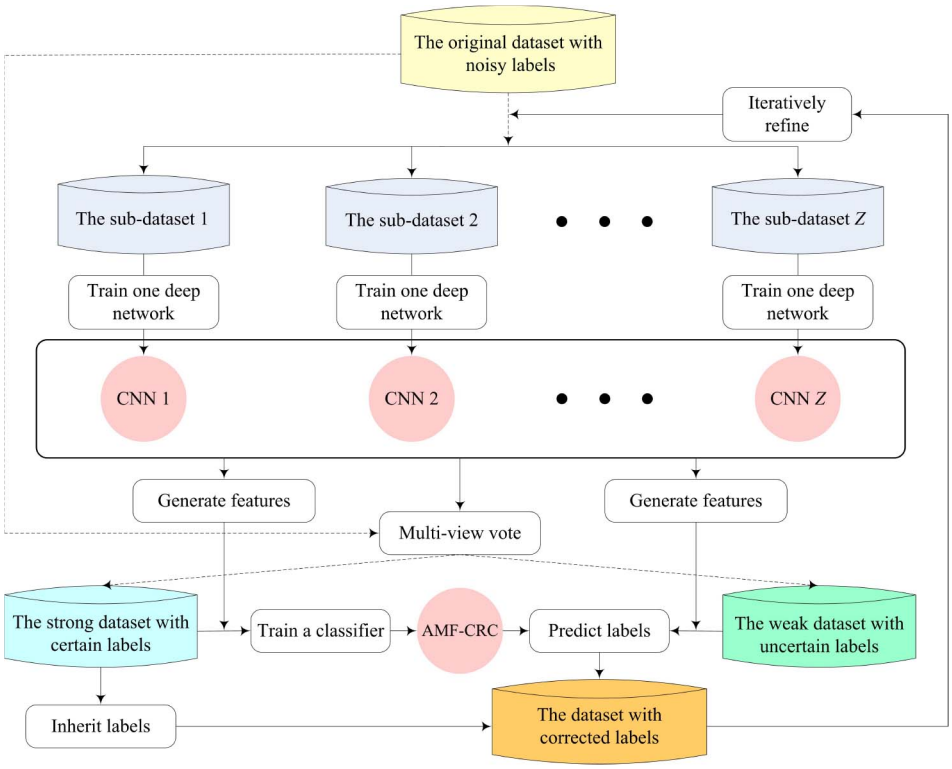


Fig. 1. Workflow of the proposed RSSC-ETDL framework. The dashed arrow means that the process just occurs in the first iteration and does not repeat in the following iterations.

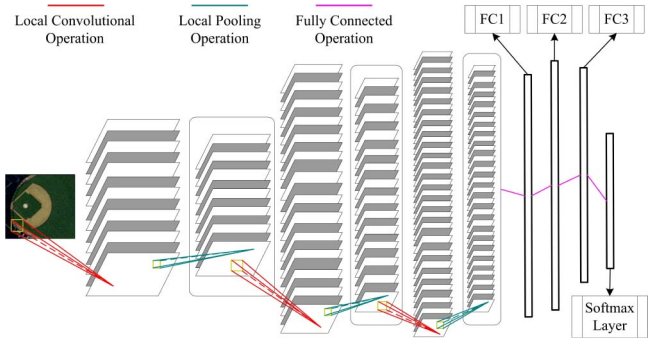


Fig. 2. Overall architecture of the adopted CNN.

TABLE I
CONFIGURATION OF CNN

Layer1	Configuration
Conv1	filter:64 × 11 × 11 × 3, stride1:4 × 4 pool:3 × 3, stride2:2 × 2
Conv2	filter:256 × 5 × 5 × 64, stride1:1 × 1 pool:3 × 3, stride2:2 × 2
Conv3	filter:256 × 3 × 113 × 256, stride1:1 × 1
Conv4	filter:256 × 3 × 3 × 256, stride1:1 × 1
Conv5	filter:256 × 3 × 3 × 256, stride1:1 × 1 pool:3 × 3, stride2:2 × 2
FC1	4096
FC2	4096
FC3	1000
Softmax Layer	the number of categories

the dth CNN model with the hyperparameter Φ_d and T denotes the number of categories.

IV. Experimental Results

In this section, we conduct extensive experiments to verify the effectiveness of the proposed approach for RSSC with noisy labels. Section IV-A depicts the experimental setting and the evaluation criteria. In Section IV-B, we conduct the experiments on two remote sensing image scene datasets with simulated label noise. In addition, Section IV-C reports and discusses the experimental results on two remote sensing image scene datasets with real noisy labels.

A. Experimental Setting and Evaluation Criteria

Because of its wide application and superior performance in RSSC, VGG [53] is taken as the architecture of CNN in this experiment. The specific configuration of the adopted CNN is shown in Table I, and the CNN can process an input image of 244 × 244 × 3. In Table I, "filter" specifies the number of filters, the size of a field, and the dimensions of input data, and it can be formulated as num × size × size × dim. The "stride1" denotes the sliding step of the convolution operation, "pool" denotes the downsampling factor, and "stride2" denotes the sliding step of the local pooling operation. As shown in Table I, the adopted CNN has five convolutional layers, three fully connected layers, and one softmax classification layer. Once the CNN is trained, each remote sensing image scene can be represented by three feature vectors using the corresponding three fully connected layers.

This article is implemented by MATLAB and conducted on a Dell station with 8 Intel Core i7-6700 processors, 32 GB of RAM, and the NVIDIA GeForce GTX 745.

We train the model on the training dataset with noisy labels. In addition, we test the performance of one trained model on the testing dataset with accurate labels using the widely adopted overall accuracy (OA) indicator.

B. Experimental Results on Remote Sensing Datasets With Simulated Noisy Labels

1) Collection of Remote Sensing Datasets With Simulated Noisy Labels: In the following experiments, we evaluate the methods on two publicly open large-scale remote sensing image scene datasets, including RSI-CB256 [24] and PatternNet [54]. More specifically, RSI-CB256 includes 35 land cover categories and has a total of 24 000 remote sensing image scenes where each remote sensing image scene is with a size of 256 × 256. In addition, PatternNet includes 38 land cover categories and has a total of 30 400 RS image scenes where each remote sensing image scene is with a size of 256 × 256. In both remote sensing image scene datasets, 20% of the original dataset is randomly selected as the training dataset and the rest is taken as the testing dataset. To quantitatively evaluate the error-tolerant learning methods, we corrupt the labels of the training dataset with different error rates (i.e., eRate = 0.4, eRate = 0.6, and eRate = 0.8 in this experiment) based on the existing noisy simulation methods [29], [34].

2) Sensitivity Analysis of the Critical Parameters: Fixing the error rate (i.e., eRate = 0.8), we evaluate the performance of our proposed RSSC-ETDL approach under different regularization parameters α and β on RSI-CB256. More specifically, the corresponding results are summarized in Table II. As depicted in Table II, the performance of our proposed RSSC-ETDL approach obviously changes along with the variation of α and β. This phenomenon fully verifies the effectiveness of our proposed approach from two aspects: 1) the idea to adaptively combine multiple features and 2) the strategy to borrow information from other classes to represent the testing sample. As a tradeoff, our proposed RSSC-ETDL approach can achieve the best performance when $\alpha = 3.0 \times 10^3$ and $\beta = 1.0 \times 10^3$. To reduce the computational expense, we follow this setting in the following. Naturally, much better performance can be rationally expected if we tune the parameter setting again in the new data environment.

Furthermore, we evaluate the performance of our proposed RSSC-ETDL approach under different view numbers (i.e., changing the view number in Algorithm 1) and the corresponding results are summarized in Table III. Generally, the increase of the view number would lift the accuracy of samples in the strong dataset, but inevitably reduce the volume of the strong dataset. As a whole, the increase of the view number does not benefit in improving the performance. Hence, the view number is set to 2 in the following.

3) Convergence Analysis of Our RSSC-ETDL Approach: To conduct the convergence analysis, we summarize the performance of our proposed RSSC-ETDL approach under different iterations on RSI-CB256 in Fig. 3(a). As shown in Fig. 3(a), more iterations indeed help to improve the performance of our proposed approach especially when the error rate is high (e.g., eRate = 0.6 and eRate = 0.8). In addition, the performance improvement seems to be small after three iterations. Obviously, the training complexity of our proposed RSSC-ETDL approach is proportional to the number of iterations. To obtain a tradeoff between the training complexity and classification performance, the number of iterations is set to 3 in this experiment. Like the experiment on RSI-CB256, we also conduct the convergence analysis of our RSSC-ETDL approach on PatternNet and the corresponding results are summarized in Fig. 3(b). As shown in Fig. 3(b), the performance of our RSSC-ETDL approach on PatternNet can obtain a similar variation trend on RSI-CB256. More specifically, our RSSC-ETDL approach can reach the saturation condition with three iterations. With this consideration, the number of iterations is also set to 3 on PatternNet.

As depicted in Section III-B, our RSSC-ETDL approach works in an iterative manner where each round of iteration includes two main modules (i.e., learning multiview deep networks and correcting potential error labels). Considering that each round of iteration of our RSSC-ETDL approach costs a fixed time, we report the separate running time of two main modules in the first iteration. More specifically, the running time of our RSSC-ETDL approach on RSI-CB256 and PatternNet under different error rates is summarized in Tables IV and V. The potential readers can directly infer the entire running time of our RSSC-ETDL approach once the number of iterations is determined.

4) Comparison With the State-of-the-Art Approaches: To verify the superiority of our proposed RSSC-ETDL approach, we compare our method with five recently published methods, including the ETDL method via the MAE loss [30], the ETDL approach via the Lq loss [31], the ETDL method via bootstrapping [27], the ETDL approach via dropout [29], and the ETDL algorithm via ICL [34]. In addition, to verify the effectiveness of the proposed AMF-CRC classifier in our RSSC-ETDL approach, we consider two more baselines by extending our RSSC-ETDL approach. By naively aggregating multiple features into one feature vector, the AMF-CRC in our RSSC-ETDL approach can be replaced by some traditional classifiers, such as SVM and CRC. More specifically, RSSC-ETDL-SVM stands for our RSSC-ETDL

TABLE II
OA VALUES OF OUR RSSC-ETDL APPROACH UNDER DIFFERENT REGULARIZATION PARAMETERS ON RSI-CB256

	$\beta = 0.5 \times 10^3$	$\beta = 1.0 \times 10^3$	$\beta = 2.0 \times 10^3$
$\alpha = 1.0 \times 10^3$	0.7757	0.7603	0.7243
$\alpha = 2.0 \times 10^3$	0.7062	0.7755	0.7358
$\alpha = 3.0 \times 10^3$	0.7737	0.8333	0.7773
$\alpha = 4.0 \times 10^3$	0.6789	0.8025	0.7693
$\alpha = 5.0 \times 10^3$	0.6587	0.7627	0.6578

TABLE III
OA VALUES OF OUR RSSC-ETDL APPROACH UNDER DIFFERENT REGULARIZATION PARAMETERS ON RSI-CB256

	maxIter=1	*maxIter*=2	*maxIter*=3	*maxIter*=4	*maxIter*=5
Z=2	0.6264	0.8157	0.8333	0.8337	0.8284
Z=3	0.5955	0.7164	0.7274	0.7234	0.7253
Z=4	0.5754	0.4659	0.4797	0.4798	0.4786

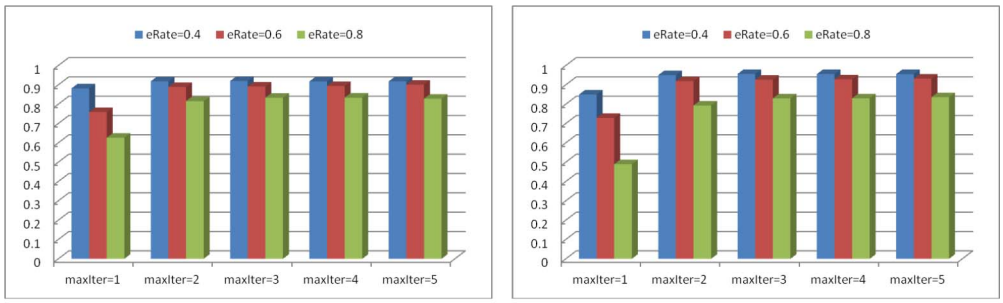


Fig. 3. OA values of our RSSC-ETDL approach under different iterations on two remote sensing image scene datasets (i.e., AID-GA and BUD-GLC) with real noisy labels. (a) Results on RSI-CB256 under different error rates. (b) Results on PatternNet under different error rates.

TABLE IV
RUNNING TIME OF OUR RSSC-ETDL APPROACH ON RSI-CB256 UNDER DIFFERENT ERROR RATES

	RSI-CB256 (eRate = 0.4)		RSI-CB256 (eRate = 0.6)		RSI-CB256 (eRate = 0.8)	
	Learning deep networks	Correcting image labels	Learning deep networks	Correcting image labels	Learning deep networks	Correcting image labels
Time (minutes)	46.35	20.48	46.28	19.99	46.32	17.03

TABLE V
RUNNING TIME OF OUR RSSC-ETDL APPROACH ON PATTERNNET UNDER DIFFERENT ERROR RATES

	PatternNet (eRate = 0.4)		PatternNet (eRate = 0.6)		PatternNet (eRate = 0.8)	
	Learning deep networks	Correcting image labels	Learning deep networks	Correcting image labels	Learning deep networks	Correcting image labels
Time (minutes)	57.44	35.39	57.12	31.06	57.56	15.93

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART APPROACHES ON RSI-CB256 UNDER DIFFERENT ERROR RATES

	MAE in [30]	Lq in [31]	Bootstrapping in [27]	Dropout in [29]	ICL in [34]	Our RSSC-ETDL-SVM	Our RSSC-ETDL-CRC	Our RSSC-ETDL
eRate=0.4	0.5496	0.8932	0.8891	0.5097	0.7282	0.9086	0.9104	0.9185
eRate=0.6	0.3656	0.8251	0.8415	0.4718	0.6006	0.8650	0.8950	0.8915
eRate=0.8	0.3737	0.6373	0.6430	0.2162	0.3986	0.7127	0.7298	0.8333

approach equipped with the SVM classifier, RSSC-ETDL-CRC denotes our RSSC-ETDL approach equipped with the CRC classifier, and RSSC-ETDL denotes our RSSC-ETDL approach equipped with the AMF-CRC classifier.

To verify the superiority of our RSSC-ETDL approach, we evaluate the aforementioned methods on RSI-CB256 and summarize the quantitative evaluation results in Table VI. As shown in Table VI, our RSSC-ETDL and its variants can dramatically outperform the state-of-the-art approaches [27], [29]–[31], [34] which fully reflect the effectiveness of our proposed error-tolerant learning framework. When eRate = 0.4 and eRate = 0.6, our RSSC-ETDL approach obtains a similar performance level compared with RSSC-ETDL-CRC and performs better than RSSC-ETDL-SVM. However, when eRate = 0.8, our proposed RSSC-ETDL approach outperforms RSSC-ETDL-SVM and RSSC-ETDL-CRC with a large margin. When the labels are corrupted by heavy noise, the number of samples in the strong dataset is very small which makes the label recovery of the weak dataset become a classic small-sample-size classification problem. The significant performance improvement sufficiently verifies the superiority of the presented AMF-CRC compared with the traditional classifiers, including SVM and CRC.

To check the universality of our RSSC-ETDL method, we evaluate our method on another publicly open RS image scene dataset (i.e., PatternNet). As depicted in Table VII, our RSSC-ETDL approach and its variants obviously outperform the recently published ETDL methods [27], [29]–[31], [34]. Under the condition with a small error rate (i.e., eRate = 0.4 and eRate = 0.6), the performance of our RSSC-ETDL method obtains close to its two variants, including RSSC-ETDL-SVM and RSSC-ETDL-CRC. When eRate = 0.8, our RS-ETDL method obviously performs better than RSSC-ETDL-SVM and RSSC-ETDL-CRC which reflects the robustness of the proposed AMF-CRC classifier.

C. Experimental Results on Remote Sensing Datasets With Real Noisy Labels

1) Collection of Remote Sensing Datasets With Real Noisy Labels: As analyzed in Section I, the real label noise of remote sensing image scenes mainly comes from the greedy annotation process. To fully verify the validity of our proposed RSSC-ETDL approach, we construct two remote sensing image scene datasets with real label noise by adopting two different greedy remote sensing image annotation strategies, which can cover the main types of greedy remote sensing

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART APPROACHES ON PATTERNNET UNDER DIFFERENT ERROR RATES

	MAE in [30]	Lq in [31]	Bootstrapping in [27]	Dropout in [29]	ICL in [34]	Our RSSC-ETDL-SVM	Our RSSC-ETDL-CRC	Our RSSC-ETDL
eRate=0.4	0.4638	0.8963	0.8916	0.4947	0.6241	0.9528	0.9548	0.9564
eRate=0.6	0.2556	0.8656	0.8671	0.4525	0.5607	0.9215	0.9244	0.9264
eRate=0.8	0.1955	0.7097	0.6868	0.3703	0.4182	0.7634	0.7003	0.8297

image scene annotation methods mentioned in Section I. In the following, we detail the construction process of two remote sensing image scene datasets with real label noise.

In the construction process of the first remote sensing image scene dataset with real label noise, we deploy the greedy annotation algorithm in [23]. In this annotation algorithm, the image scenes in the original dataset are first aggregated into a limited cluster by unsupervised methods, and then we label the original dataset on the cluster level for accelerating the annotation process and saving the labor cost. Here, we take the publicly available dataset (i.e., AID in [55]) as the source data. Specifically, AID includes 30 land cover categories and has a total of 10 000 remote sensing image scenes, where each remote sensing image scene is with a size of 600×600. We randomly select 50% of the original dataset as the training dataset and the rest is taken as the testing dataset. Different from the noise simulation process in Section IV-B1, we totally throw the labels of the training dataset of AID and label it by the greedy annotation algorithm in [23]. In detail, we set the greedy annotation rate as 5%, which means that the annotation process is accelerated by 20 times. To facilitate clarifying, we call the greedily annotated training dataset of AID and the testing dataset of AID as AID-GA in the following.

We consider another greedy annotation strategy to construct the second remote sensing image scene dataset with real label noise. As examined in Section I, the existing geodatabase can be taken as the semantic layer to automatically annotate the RS imagery. As a first and primary attempt, we take the detection of built-up (BU) areas as a case study. In many existing techniques [12], [56], the detection of BU areas is taken as a binary scene classification task (i.e., one scene is classified to BU or non-BU) where the annotation of the BU scene dataset often needs massive labor cost. Here, we take the recently released GLC product (i.e., FROM-GLC10 in [57]) as the semantic reference map at a 10-m spatial resolution, and the Google Earth imagery at a 0.5-m spatial resolution is taken as the source imagery. After the geographic coordinate registration between the Google Earth imagery and the FROM-GLC10 product, the impervious layer in FROM-GLC10 is used to generate the scenes of the BU category, and the other layers are adopted to generate the scenes of the non-BU category where each image scene is with a size of 256×256. We collect 5000 image scenes as the training dataset. Considering that the FROM-GLC10 product is generated by some interpretation methods and the accuracy of the impervious layer is around 72% as reported in [57], the training dataset inevitably contains a certain degree of error labels. In addition, we manually refine a testing dataset containing 5000 image scenes, which does not have any overlap with the training part. To facilitate clarifying, we call this dataset (i.e., the combination of the noisy training dataset and the clean testing dataset) BUD-GLC in the following.

TABLE VIII
RUNNING TIME OF OUR RSSC-ETDL APPROACH ON AID-GA AND BUD-GLC

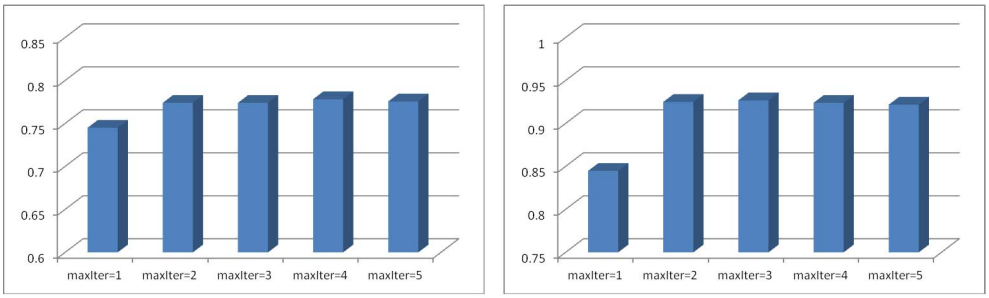
	AID-GA		BUD-GLC	
	Learning deep networks	Correcting image labels	Learning deep networks	Correcting image labels
Time (minutes)	46.20	21.28	46.23	11.73

2) Convergence Analysis of Our RSSC-ETDL Approach: Fixing the setting of hyperparameters, discussed in Section IV-B2, we conduct the convergence analysis of our RSSC-ETDL approach on two remote sensing image scene datasets with real noisy labels. More specifically, we report the quantitative evaluation results (i.e., OA) of our RSSC-ETDL approach under different iterations on AID-GA and BUD-GLC in Fig. 4(a) and (b), respectively. As shown in Fig. 4(a) and (b), our RSSC-ETDL approach with two iterations can dramatically outperform the approach with only one round of iteration. In addition, our RSSC-ETDL approach seems to converge to a stable state after three iterations. Because of this empirical characteristic, the number of iterations is also set to 3 on both datasets.

Similar to Section IV-B3, we also report the running time of the first iteration of our RSSC-ETDL approach on AID-GA and BUD-GLC. More specifically, the separate running time of two main modules (i.e., learning deep networks and correcting potential error labels) of the first iteration is summarized in Table VIII.

3) Comparison With the State-of-the-Art Approaches: Similar to the comparison setting in Section IV-B3, we also compare our RSSC-ETDL approach with five recently published methods, including the ETDL method via the MAE loss [30], the ETDL approach via the Lq loss [31], the ETDL method via bootstrapping [27], the ETDL approach via dropout [29], and the ETDL algorithm via ICL [34]. In addition, to verify the effectiveness of the proposed AMF-CRC classifier in our RSSC-ETDL approach, we consider two more baselines (i.e., RSSC-ETDL-SVM and RSSC-ETDL-CRC) by extending our RSSC-ETDL framework.

As shown in Tables IX and X, extensive experiments on AID-GA and BUD-GLC show that our RSSC-ETDL and its two variants (i.e., RSSC-ETDL-SVM and RSSC-ETDL-CRC) can obviously outperform the state-of-the-art approaches [27], [29]–[31], [34] which fully reflects the effectiveness of our proposed error-tolerant learning framework. In addition, an obvious performance improvement also verifies the superiority



(a) (b)

Fig. 4. OA values of our RSSC-ETDL approach under different iterations on two remote sensing image scene datasets (i.e., AID-GA and BUD-GLC) with real noisy labels. (a) Results on AID-GA. (b) Results on BUD-GLC.

TABLE IX
COMPARISON WITH THE STATE-OF-THE-ART APPROACHES ON AID-GA

	MAE in [30]	Lq in [31]	Bootstrapping in [27]	Dropout in [29]	ICL in [34]	Our RSSC-ETDL-SVM	Our RSSC-ETDL-CRC	Our RSSC-ETDL
OA	0.3046	0.7100	0.7246	0.4210	0.5998	0.7482	0.7592	0.7740

TABLE X
COMPARISON WITH THE STATE-OF-THE-ART APPROACHES ON BUD-GLC

	MAE in [30]	Lq in [31]	Bootstrapping in [27]	Dropout in [29]	ICL in [34]	Our RSSC-ETDL-SVM	Our RSSC-ETDL-CRC	Our RSSC-ETDL
OA	0.8322	0.5568	0.7910	0.8224	0.7346	0.9124	0.9112	0.9269

of the presented AMF-CRC compared with the traditional classifiers, including SVM and CRC.

V. CONCLUSION

Along with the development of information technology, imaging technology, and manufacturing, we have been in the era of remote sensing big data. As is well known, veracity is one of the most distinctive characteristics of remote sensing big data. How to mine the intrinsic knowledge from the remote sensing data with noisy labels becomes a new learning paradigm in the era of remote sensing big data. Driven by this intensive demand, this article proposes a new RSSC-ETDL approach for RSSC. The proposed RSSC-ETDL involves an alternating procedure that learns multiview deep networks and corrects potential error labels. To correct the error labels, we proposed a novel AMF-CRC classifier which can adaptively combine multiple features to improve the classification accuracy. To fully show the superiority of our RSSC-ETDL method, we constructed two remote sensing image scene datasets with simulated noisy labels by randomly corrupting the existing open datasets under varying error rates, and collect two remote sensing image scene datasets with real noisy labels by employing the existing greedy annotation strategies. Extensive experiments on multiple remote sensing image scene datasets with varying kinds of error labels demonstrate that our proposed RSSC-ETDL approach can dramatically outperform the state-of-the-art approaches. In addition, through comparison with some traditional classifiers, including SVM and CRC, the effectiveness of the proposed AMF-CRC classifier is also verified. It is noted that as a general classifier, the proposed AMF-CRC classifier may benefit more applications in the computer vision domain.

APPENDIX

In this section, we first give out some necessary definitions. Based on these definitions, we further give the proof of the properties of Theorem 1, respectively.

Definition 1: Let $\psi : \mathbb{R}^N \to \mathbb{R} \cup \infty$ be a proper lower semicontinuous function.

1) The domain of ψ is defined by $\operatorname{dom}\psi : \{x \in \mathbb{R}^N : \psi(x) < \infty\}$.
2) For any $x \in \operatorname{dom}\psi$, the subdifferential $\partial\psi$ is defined by

$$\partial\psi(x) = \left\{ z : \liminf_{y \to x} \frac{\psi(y) - \psi(x) - \langle z, y - x\rangle}{\|x - y\|} \geq 0 \right\}$$

and $\partial\psi(x) = \emptyset$ if $x \notin \operatorname{dom}\psi$.

3) We say $x \in \operatorname{dom}\psi$ a stationary point if $0 \in \partial\psi(x)$.

To establish the convergence, one important property regarding the objective function is the strong convexity of in (1) in terms of w or ρ, namely

$$\begin{aligned} f(\rho, w) - f(\rho', w) &\geq (\rho - \rho')^T \nabla_\rho f(\rho', w) + \beta \|\rho - \rho'\|^2 \\ f(\rho, w) - f(\rho, w') &\geq (w - w')^T \nabla_w f(\rho, w') + \alpha \|w - w'\|^2. \end{aligned} \tag{8}$$

Denote by

$$\mathbb{W} := \left\{ w \in \mathbb{R}^M : 0 \leq w_v \leq 1, \sum_{v=1}^{M} w_v = 1 \right\}.$$

Another useful property is that f is Lipschitz smooth for bounded (ρ, w), that is, there exists L such that $\|\nabla^2 f(\rho, w)\| \leq L$ for any $w \in \mathbb{W}$ and bounded ρ.

A. Proof 1) of Theorem 1

By the definition of (2), we have

$$f(\rho(k), w(k)) - f(\rho(k+1), w(k)) \geq \beta \|\rho(k) - \rho(k+1)\|_2^2 \tag{9}$$

where the inequality follows because of strong convexity of subproblem $f(\rho, w(k))$ as in (8). Similarly, it follows from the definition of (3), strong convexity of subproblem $f(\rho(k+1), w)$ as in (8), and convexity of the set $\mathbb{W}$, that:

$$f(\rho(k+1), w(k)) - f(\rho(k+1), w(k+1)) \geq \alpha \|w(k) - w(k+1)\|_2^2 \tag{10}$$

which together with (9) gives

$$f(\rho(k), w(k)) - f(\rho(k+1), w(k+1)) \geq \alpha \|w(k) - w(k+1)\|_2^2 + \beta \|\rho(k) - \rho(k+1)\|_2^2. \tag{11}$$

Due to the fact that $f(\rho, w) \geq 0$, the above equation implies that the sequence $\{f(\rho(k), w(k))\}$ is decreasing hence is convergent. Summing (11) for all k from zero to infinty gives

$$\sum_{k=0}^{\infty} \|w(k) - w(k+1)\|^2 + \|\rho(k) - \rho(k+1)\|^2 \leq \frac{1}{\min\{\alpha, \beta\}} f(\rho(0), w(0)). \tag{12}$$

Furthermore, (12) implies that $(\rho(k), w(k))$ is regular

$$\lim_{k \to \infty} \|w(k) - w(k+1)\|^2 + \|\rho(k) - \rho(k+1)\|^2 = 0. \tag{13}$$

B. Proof 2) of Theorem 1

We first show that the sequence $\{f(\rho(k), w(k))\}$ is bounded. It is clear that $w(k)$ is always bounded since $w(k) \in \mathbb{W}$. Noting that $f(\rho(0), w(0)) \geq f(\rho(k), w(k)) \geq \beta \|\rho(k)\|_2^2$ we also have $\|\rho(k)\|_2^2 \leq f(\rho(0), w(0))/\beta$.

Since the sequence $\{f(\rho(k), w(k))\}$ is bounded, by the Bolzano–Weiestrass theorem, we know this sequence has at least one convergent subsequence. Let $\{f(\rho(*), w(*))\}$ be the limit point of any convergent subsequence $\{f(\rho(k_k), w(k_k))\}$. To show $\{f(\rho(*), w(*))\}$ is a stationary point, we first transfer the constrained problem in (1) into the following equivalent form without any constraints to simplify the notation of subdifferential:

$$g(\rho, w) := f(\rho, w) + \delta_{\mathbb{W}}(w) \tag{14}$$

where f is defined in (1) and $\delta_{\mathbb{W}}$ is the indicator function of the set $\mathbb{W}$, that is, $\delta_{\mathbb{W}}(w) = 0$ if $w \in \mathbb{W}$ and $\delta_{\mathbb{W}}(w) = \infty$ if $w \notin \mathbb{W}$.

By the optimality of (2), we have

$$\nabla_p f(\rho(k+1), w(k)) = 0 \tag{15}$$

which together with the Lipschitz continuity of $\nabla f(\rho, w)$ (i.e., $\|\nabla^2 f(\rho, w)\| \leq L$) that

$$\left\| \nabla_p f(\rho(k+1), w(k+1)) \right\| \leq L \|w(k) - w(k+1)\|. \tag{16}$$

Recalling (13), we further have

$$\lim_{k \to \infty} \nabla_p f(\rho(k+1), w(k+1)) = 0. \tag{17}$$

Similarly, by the optimality of (3), it always holds that

$$0 \in \nabla_w g(\rho(k+1), w(k+1)) \nabla_w f(\rho(k+1), w(k+1)) + \partial \delta_{\mathbb{W}}.$$

This together with the above equation implies that

$$0 \in \nabla_\rho g(\rho(*), w(*)) + \nabla_w g(\rho(*), w(*)). \tag{18}$$

Thus, $(\rho(*), w(*))$ is a stationary point of g, that is, the problem in (1). Finally, by noting that g obeys the so-called Kurdyka–Lojasiewicz (KL) property [58], we obtain that $(\rho(*), w(*))$ is the only limiting point of the sequence, that is, $(\rho(k), w(k))$ itself converges to $(\rho(*), w(*))$. This completes the proof of Theorem 1.

REFERENCES

[1] H. Yuan and Y. Y. Tang, "Spectral–spatial shared linear regression for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 934–945, Apr. 2017.

[2] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial–spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[3] H. Li, G. Xiao, T. Xia, Y. Y. Tang, and L. Li, "Hyperspectral image classification using functional data analysis," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1544–1555, Sep. 2013.

[4] S. Jia, L. Shen, J. Zhu, and Q. Li, "A 3-D Gabor phase-based coding and matching framework for hyperspectral imagery classification," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1176–1188, Apr. 2018.

[5] Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016.

[6] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1465–1472.

[7] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

[8] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 157–161, Feb. 2016.

[9] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.

[10] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[11] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.

[12] Y. Tan, S. Xiong, and Y. Li, "Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3988–4004, Nov. 2018.

[13] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[14] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2017.

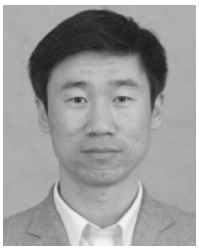
[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] Z. Zhu, D. Soudry, Y. C. Eldar, and M. B. Wakin, "The global optimization geometry of shallow linear neural networks," *J. Math. Imag. Vis.*, vol. 62, pp. 1–14, May 2019.

[17] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.
[18] Y. Lu, G. Lu, J. Li, Y. Xu, Z. Zhang, and D. Zhang, "Multiscale conditional regularization for convolutional neural networks," *IEEE Trans. Cybern.*, early access, Apr. 2, 2020, doi: 10.1109/TCYB.2020.2979968.
[19] L. Wu, Y. Wang, X. Li, and J. Gao, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1791–1802, May 2019.
[20] J. Pu, Y. Panagakis, S. Petridis, J. Shen, and M. Pantic, "Blind audio–visual localization and separation via low-rank and sparsity," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2288–2301, May 2020.
[21] M. T. Mills and N. G. Bourbakis, "Graph-based methods for natural language processing and understanding—A survey and analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 1, pp. 59–71, Jan. 2014.
[22] G.-S. Xia, Z. Wang, C. Xiong, and L. Zhang, "Accurate annotation of remote sensing images via active spectral clustering with little expert knowledge," *Remote Sens.*, vol. 7, no. 11, pp. 15014–15045, 2015.
[23] Y. Li and D. Ye, "Greedy annotation of remote sensing image scenes based on automatic aggregation via hierarchical similarity diffusion," *IEEE Access*, vol. 6, pp. 57376–57388, 2018.
[24] H. Li, C. Tao, Z. Wu, J. Chen, J. Gong, and M. Deng, "RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data," 2017. [Online]. Available: arXiv:1705.10450.
[25] P. Jin, G.-S. Xia, F. Hu, Q. Lu, and L. Zhang, "AID++: An updated version of aid on scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2018, pp. 4721–4724.
[26] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," 2017. [Online]. Available: arXiv:1712.05055.
[27] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," 2014. [Online]. Available: arXiv:1412.6596.
[28] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2691–2699.
[29] I. Jindal, M. Nokleby, and X. Chen, "Learning deep networks from noisy labels with dropout regularization," in *Proc. IEEE 16th Int. Conf. Data Min. (ICDM)*, 2016, pp. 967–972.
[30] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1919–1925.
[31] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8778–8788.
[32] Z. Zhang, J. Yang, Z. Zhang, and Y. Li, "Cross-training deep neural networks for learning from label noise," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 4100–4104.
[33] L. Wu, S. Liu, M. Jian, J. Luo, X. Zhang, and M. Qi, "Reducing noisy labels in weakly labeled data for visual sentiment analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 1322–1326.
[34] B. Yuan, J. Chen, W. Zhang, H.-S. Tai, and S. McMains, "Iterative cross learning on noisy labels," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 757–765.
[35] Y. Li, Y. Tan, J. Deng, Q. Wen, and J. Tian, "Cauchy graph embedding optimization for built-up areas detection from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2078–2096, May 2015.
[36] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.
[37] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.
[38] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. M. Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sens.*, vol. 9, no. 2, p. 173, 2017.
[39] X. Kang, P. Duan, X. Xiang, S. Li, and J. A. Benediktsson, "Detection and correction of mislabeled training samples for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5673–5686, Oct. 2018.
[40] B. Tu, C. Zhou, W. Kuang, L. Guo, and X. Ou, "Hyperspectral imagery noisy label detection by spectral angle local outlier factor," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1417–1421, Sep. 2018.
[41] F. Condessa, J. Bioucas-Dias, and J. Kovać, "Supervised hyperspectral image classification with rejection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2321–2332, Jun. 2016.
[42] 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. 2019.
[43] B. Tu, X. Zhang, X. Kang, G. Zhang, and S. Li, "Density peak-based noisy label detection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1573–1584, Mar. 2019.
[44] J. Peng, W. Sun, L. Ma, and Q. Du, "Discriminative transfer joint matching for domain adaptation in hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 972–976, Jun. 2019.
[45] J. Peng, W. Sun, and Q. Du, "Self-paced joint sparse representation for the classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1183–1194, Feb. 2019.
[46] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, p. 884, 2019.
[47] L. Jian, F. Gao, P. Ren, Y. Song, and S. Luo, "A noise-resilient online learning algorithm for scene classification," *Remote Sens.*, vol. 10, no. 11, p. 1836, 2018.
[48] B. B. Damodaran, R. Flamary, V. Seguy, and N. Courty, "An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images," *Comput. Vis. Image Understand.*, vol. 191, Feb. 2020, Art. no. 102863.
[49] C.-C. Chang, "lIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, Feb. 2011. [Online]. Available: http://www.csie.ntu.edu.tw/čjlin/libsvm
[50] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 471–478.
[51] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," 2012. [Online]. Available: arXiv:1204.2358.
[52] 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.
[53] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. BMVC*, vol. 2, 2011, p. 8.
[54] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
[55] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[56] Y. Li, Y. Tan, Y. Li, S. Qi, and J. Tian, "Built-up area detection from satellite images using multikernel learning, multifield integrating, and multihypothesis voting," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1190–1194, Jun. 2015.
[57] P. Gong *et al.*, "Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017," *Sci. Bull.*, vol. 64, no. 6, pp. 370–373, 2019.
[58] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, pp. 459–494, Aug. 2014.

Yansheng Li received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

He is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, where he was employed as an Assistant Professor in 2015, and became an Associate Research Fellow in 2017. From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He has authored more than 30 peer-reviewed articles (SCI papers) in international journals from multiple domains, such as remote sensing and computer vision. His research interests mainly lay in the field of computer vision, machine learning, and their applications in remote sensing big data analysis.



Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently the Vice Dean of the School of Remote Sensing and Information Engineering, Wuhan University, where he has been a Full Professor with the School of Remote Sensing and Information Engineering since 2006. From 2014 to 2015, he was a Senior Visiting Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. From 2015 to 2018, he was a Senior Scientist with Environmental Systems Research Institute, Inc. (Esri), Redlands, CA, USA. He has published more than 150 research articles and one book. He holds 23 Chinese patents and 26 copyright registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multi-source datasets, integration of LiDAR point clouds and images, and 3-D city reconstruction.

Zhihui Zhu (Member, IEEE) received the B.Eng. degree in communication engineering from the Zhejiang University of Technology, Hangzhou, China, in 2012, and the Ph.D. degree in electrical engineering from the Colorado School of Mines, Golden, CO, USA, in 2017.

He was a Postdoctoral Fellow with the Mathematical Institute for Data Science and the Center for Imaging Science, Johns Hopkins University, Baltimore, MD, USA, from 2018 to 2019. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO, USA. His research interests include exploiting inherent structures and applying optimization methods with guaranteed performance for signal processing and machine learning.