Accurate Label Refinement From Multiannotator of Remote Sensing Data

Xiangyu Wang, Lyuzhou Chen, Taiyu Ban, Derui Lyu, Yifeng Guan, Xingyu Wu, Xiren Zhou, and Huanhuan Chen, *Senior Member, IEEE*

***Abstract*— The remote sensing (RS) field has an increasing research interest in using deep learning (DL) models to recognize kinds of RS data, leading to a great demand for training data annotation. Due to the high cost of expertise, using nonexperts to label data has become an important way to improve labeling efficiency. Commonly, a single data sample is labeled by multiple annotators and the most voted label is accepted to promise accuracy. But in the RS context, the widely admitted strategy could lose effect. Usually RS data involve considerable classes on account of the complexity of surface environments, which is prone to interclass similarity difficult to distinguish. Annotators without expertise probably make mistakes on these indistinguishable classes, thus causing error voted labels. Although classification of different characteristics in RS data has been widely documented, the nonexpert annotators are unfamiliar with these expertise, and it is difficult to force them to handle specialized labeling skills. To address the issues, this article bases multiannotator label selection on the investigation of annotators' own ability in distinguishing similar classes of images. A quality evaluation process is designed which weights the labels from capable annotators higher than those from weak ones. By a multi-round quality evaluation algorithm, correct labels could outcompete the wrong ones even disadvantaged in numbers. Experimental results demonstrate the advance of the proposed method on the RS datasets.**

***Index Terms*— Crowdsourcing, data processing, ground penetrating radar (GPR), remote sensing (RS) data annotation.**

I. Introduction

REVOLUTIONS of the deep learning (DL) technique have recently hit the world of remote sensing (RS) data processing [1], [2], resulting in an increasing demand for the annotation of training data [3], [4]. As a result, high-quality labeled dataset construction becomes an essential step for DL-based RS data classification [5], [6]. However, constructing high-quality labels is challenging for RS data, which requires considerable expertise [7], [8], [9]. As an observation of Earth, RS data usually involve a number of classes, where interclass similarity and intraclass diversity are well-known issues bothering its annotation and analysis [10], [11]. In this context, no-experts annotating RS data could be extremely incredible, but pure expert annotation is expensive and usually not achievable [12], [13]. Therefore, it is necessary to improve the quality of RS data annotation with acceptable cost.

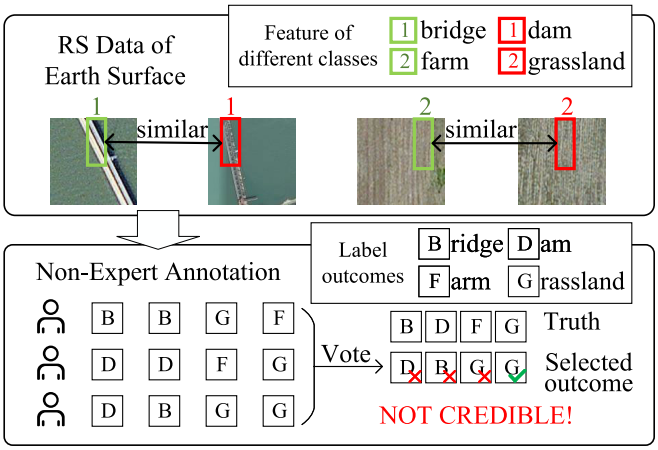


Fig. 1. Failure of the voting approach to select true labels of RS data that are similar but belong to different classes. The examples are chosen from the dataset large scale remote sensing image classification benchmark via crowdsource datapairwise (RSI-CB). Class "farm" is referred to as "dry farm" and "grassland" as "natural grassland" in RSI-CB.

Considering the balance between quality and cost, many existing studies rely on multiple nonexpert annotators (crowdsourcing method) to repeatedly label the same data [14], [15], [16]. The most occurring label by different annotators on data is considered to have high quality and taken as the true label. However, for RS data annotation, there could be dozens of classes in a dataset [10], [17], [18], [19], [20]. Distinguishing such a number of classes[1] is already exhausting, and the existing similar classes further increase the labeling difficulty. Despite that the classification of similar RS data characteristic may be well-documented, it is not feasible to force the used nonexperts to handle these expertise. When labeling the indistinguishable classes of data, annotators could easily make errors [21], [22], [23]. This could lead to more wrong repeated labels than the correct ones, therefore breaking the criterion by which the existing methods select accurate labels, i.e., more votes are more credible, as shown in Fig. 1.

Manuscript received 7 November 2022; revised 27 December 2022; accepted 29 January 2023. Date of publication 1 February 2023; date of current version 9 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0111700; in part by the National Nature Science Foundation of China under Grant 62206261, Grant 62137002, and Grant 62176245; in part by the Key Research and Development Program of Anhui Province under Grant 202104a05020011; in part by the Key Science and Technology Special Project of Anhui Province under Grant 202103a07020002; and in part by the Fundamental Research Funds for the Central Universities under Grant WK2150110026. *(Corresponding authors: Xiren Zhou; Huanhuan Chen.)*

The authors are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: zhou0612@ustc.edu.cn; hchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3241402

[1]This article focuses on the task of single label classification, where a sample of RS data is categorized into one specific class.

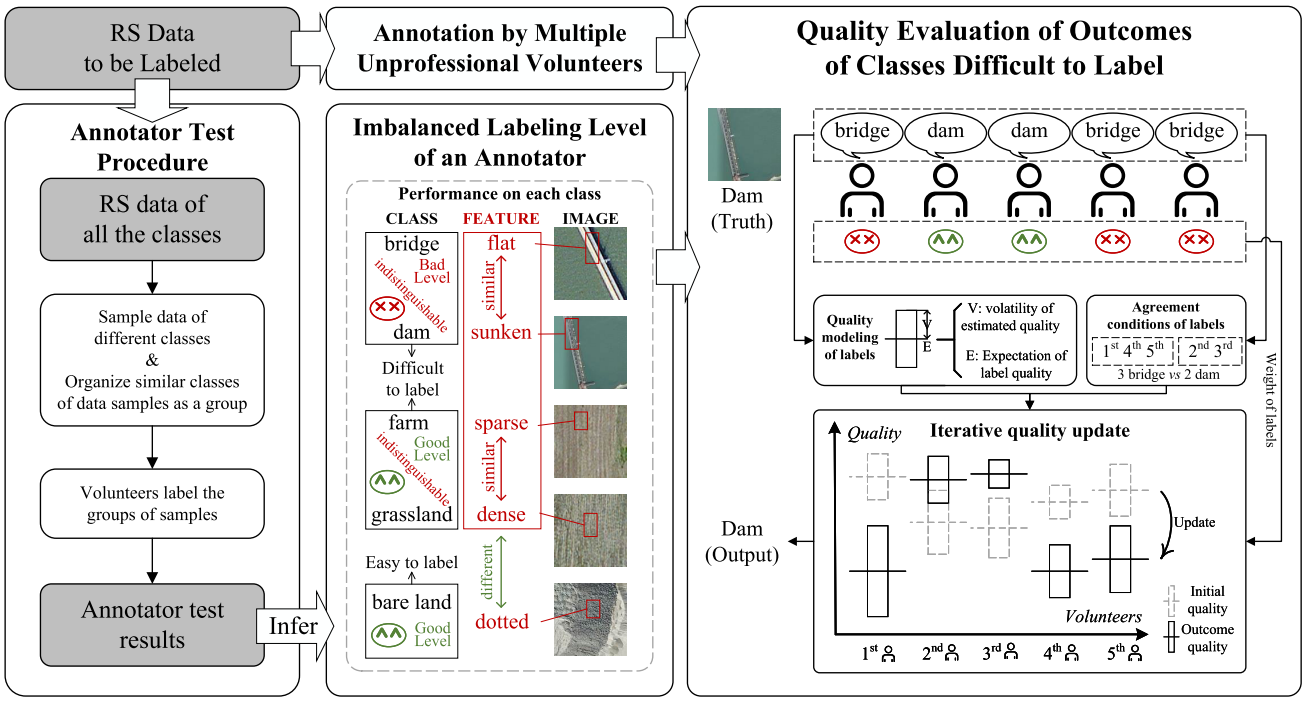


Fig. 2. Proposed crowdsourcing labeling framework for RS data annotation. 1. Similar classes of RS data are first sampled to test the annotators. 2. And their distinguishing capability on the similar classes is subsequently inferred from the test results. 3. A quality evaluation algorithm is designed to weight the multiannotator label and estimate the label quality based on the inferred distinguishing capability. Note that data sampling in "Annotator test procedure" contains another step to test annotators' overall labeling level, which is omitted here and introduced in Part 1, Section III-A. "Good Level" and "Bad Level" of an annotator denote his/her own distinguishing capability on different groups of classes, which is inferred by a process introduced in Part 2, Section III-A. The rectangle with midline in "Iterative quality update" represents the quality of multiannotator label outcomes, where the midline is quality expectation and the height of rectangle is quality volatility. The reason to model quality in such a way is introduced in Section III-B.

A common idea to refine the voting approach is to weight the outcome results with the annotator level in data annotation [24], [25], [26], [27], which could be somehow assessed in practice. For example, partial data is sampled to test annotators, and their labeling accuracy could be taken as the corresponding labeling level. However, only using such single indicator to estimate the level of an annotator could still fail in quality evaluation in RS data labels [28]. This is caused by the imbalanced labeling level of different annotators on RS data of different classes. For data in classes that an annotator cannot distinguish, the annotator may make more mistakes than other classes, and the real labeling level may be much lower than assessed. Therefore, single indicator could produce error estimation of annotator levels on certain classes of RS data, thus causing wrong voted results by the weighted approach.

To address the above issues, this article proposes a crowdsourcing labeling framework for credible label selection (presented in Fig. 2), which is evaluated in several kinds of RS data. An annotator test procedure is designed to estimate their ability to distinguish different classes of images, where images of each class are sampled to be labeled by annotators. Inferring from the annotators' outcomes, the imbalanced labeling level of each annotator is obtained, which provides a proper estimation of prior quality of the annotator's labels on different classes. On this basis, a quality evaluation process is designed that exploits the prior quality as the weight of labels. For the class of images difficult to label, labels from annotators that perform better on them have higher weights. In this way, the correct labels are prone to outcompete the wrong ones, even in a smaller number.

The contributions of this article are summarized as follows:

1) Difficulties in RS data annotation are fully considered and analyzed in this article. On this basis, a multiannotator label selection framework is proposed to infer accurate results from the multiannotator outcomes;
2) Imbalanced annotator performance on a large number of RS data classes is addressed by a test procedure. By gathering similar classes of data as samples, the annotators' distinguishing ability is assessed, thus providing a credible prior quality of the labels;
3) An iterative quality evaluation algorithm is designed that weights the labels with the annotator's performance on the corresponding class, thus obtaining more proper quality of labels difficult to annotate.

The rest of this article is organized as follows. After introducing the background in Section II, the proposed method is described in Section III. The experimental results and discussions are presented in Sections IV and V, respectively. Finally, Section VI concludes this article.

II. Related Work

A. Developments and Revolutions in RS Data Classification

RS data classification in a broad sense contains pixel-level, object-level, and scene-level tasks [11]. Pixel-level classification is known as semantic segmentation, whose emergence can be traced back to 1970s when the first Landset satellite was launched [29]. In earlier stage, the spatial resolution of RS data is low and the objects of interest are almost as large in size as one or several pixels [30]. Therefore, researchers focus on labeling each pixel with a semantic class to analyze

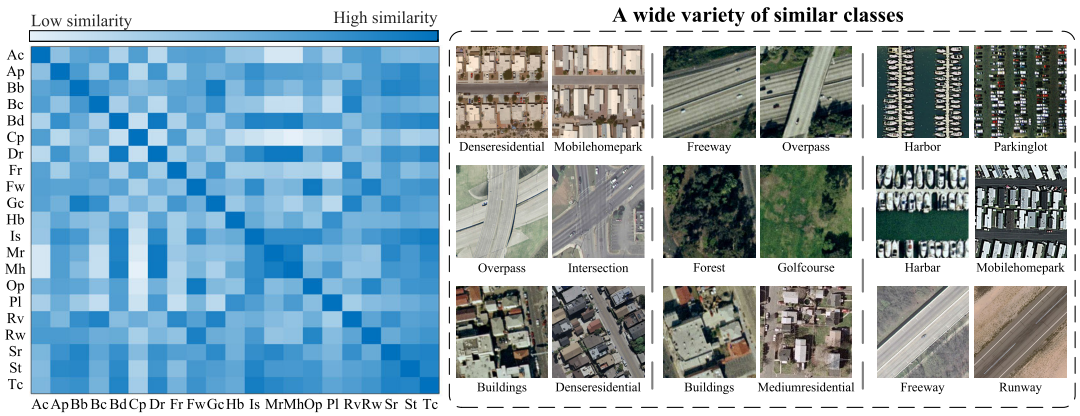


Fig. 3. Similarity between different classes of the UC-Merced data and typical similar image from different classes. *Ac, Ap, Bb, Bc, Bd, Cp, Dr, Fr, Fw, Gc, Hb, Is, Mr, Mh, Op, Pl, Rv, Rw, Sr, St, and Tc represent agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court, respectively.

RS data [31]. Object-level classification (object detection) of RS data is proposed in 2000s [32]. As the resolution becomes increasingly finer, a single pixel cannot represent a semantic object any longer, and researchers use "object" to denote a meaningful scene unit and detect them for RS data analysis [33]. Time to 2010s, RS image scene classification is proposed, which aims to classify a patch[2] of RS data with a scene [34]. The terminology "scene" becomes the interested unit that gives each observed surface area a semantic meaning. Annotation for the former two tasks involves region and boundary labeling in addition to giving class labels. This work focuses on label refinement of RS image scene classification that only requires class labeling.

In the past decade, DL technology has set off a revolution in the world of RS data classification. DL models such as Transformer and deep convolutional networks (deep CNNs) have shown powerful capability to automatically encode the features of an image [35], [36]. They simplify the handcrafted feature engineering of RS data, such as scale-invariant feature transformation (SIFT) [37] and color histogram (CH) [38]. All the needed manual work is to give each training RS data sample a correct scene label, which is an essential foundation for DL models to learn the features of each scene and to properly recognize the scene. Error labels in training data could directly harm the performance of the DL model [6], [39], and therefore, accurate label refinement is necessary for promising the effect of DL-based RS data analysis.

B. Difficulties in RS Data Annotation

The RS technologies observe all kinds of environments such as subsurface, oceans, and atmosphere to document our world [40], [41], [42], [43]. Due to the complexity of natural environments, the observed data on an area involve considerable classes of objects. Therefore, the classification/recognition task of RS data usually requires annotation on much more classes of data than that in the context of common sense [17], [18]. For example, the classical RS image scene classification dataset, UC-Merced, consists of up to 21 classes of urban buildings and natural landscapes [19]. Cheng et al. [10] provide a 45-scene-class dataset (known as NWPU-RESISC45), which contains richer meaningful scene classes than the UC-Merced dataset. Li et al. [20] construct the RSI-CB dataset using the crowdsoursing method, which contains six major classes with 35 subclasses. The large number of classes in RS data leads to great difficulty in annotating them with the right labels, especially for manual work [44].

In addition, RS data may contain visually similar classes, which may be caused by fine-grained classification criteria or the detection sensors [45], [46]. For example, RS data of residential buildings in the UC-Merced dataset [19] are divided into three fine-grained classes based on the density of structures, where the class delimitation of some samples is extremely ambiguous [18]. The RSI-CB dataset [20] organizes the RS data into six major classes and 35 subclasses, where several subclasses belong to a major class and have similar features [47]. Distinguishing RS data with multiple similar classes may be easy for experts; however, it is often difficult for non-professors [23], [48]. Some typical examples of the great similarity between different classes of RS data are presented in Fig. 3.

The difficulty of RS data annotation introduces labeling errors frequently, making it difficult for DL-based RS data analysis. To address error RS labels, some researches attempt to achieve error tolerance in DL models by reducing the label noise [13], [49], [50], which are often designed for specialized models and lacks generalization. Generalized error-tolerant learning approaches are researched as well, which alleviate the affect of error labels by correcting the label noise in the learning stage [51], [52]. However, the field of error-tolerant learning still faces plenty of unresolved challenges and lacks credibility in eliminating the harm of error labels [53]. In addition, another crucial solution is to correct the wrong labels before learning, where the most direct method is to examine the annotations by domain experts [54]. Such approaches are restricted by expensive expert resources and limited by labeling efficiency in large-scale annotation tasks. As a result, crowdsourcing approaches are usually adopted to reduce the dependence on expertise, which is introduced as following.

[2]A patch of RS data is cropped from a very large RS image.

C. Crowdsourced Annotation

The annotation of RS data could be labor-intensive, requiring work of multiple annotators, a.k.a. the crowdsourced way. It is observed that crowdsourced annotation of RS data is an error-prone task, especially for nonexperts, from aforementioned introductions. To promise label quality in the crowdsourced context, the existing researches 1) improve the accuracy of the unprofessional annotators or 2) select the most credible labels from the existing ones.

For the first strategy, the ESP game [55] was first proposed in the crowdsourced image labeling task to encourage annotators to label accurately. It will reward the annotator if the annotator and another annotator enter identical label of an image. Inspired by this strategy, many variants of ESP game are developed and show effectiveness to improve label outcomes [56], [57], [58], [59]. In addition, filtering annotators to improve the overall level of annotators is also a common approach [60], [61], [62]. In a word, these methods are mainly based on the concept of "improving the overall annotating ability of unprofessional annotators." However, this way of improving label quality is not stable enough and still suffers from annotator errors on the images that are hard to distinguish.

Another idea for improving the label quality of unprofessional annotators is to evaluate the quality of labels and choose high-quality ones. A research direction is leveraging the collaborative aspect of crowdsourcing, which enables manual checks or assists validation on labeling results. Revolt is such a framework for nonexperts, which uses the workflow label–check–modify [63]. The annotators work together through the following steps: vote (selecting an label for a image), explain (providing justifications if selecting a different label from others), and categorize (reviewing others' explanations and tagging conflicting labels). Pairwise hyperlink-induced topic search (HITS) [64] estimates the label quality by making evaluators compare a pair of conflicting labels for an image and choose a better one. Wallace et al. [65] allow workers to specify their low confidence when labeling an image, which is subsequently routed to other annotators to get a more accurate label.

In addition to leveraging crowd wisdom, many researchers try to evaluate label quality by building analytical algorithms. An earliest algorithm of this idea may be [66], which estimates the correct and error possibility of each label based on an iterative maximum likelihood method. On this basis, some typical algorithms are developed to improve their efficiency and utility to noisy label results, typically using Bayesian estimation [67], [68]. These methods usually use the statistics of the labeling results to evaluate label quality and do not rely on other information, which is concise but may lose accuracy in the case of real-world data. Some methods include more information to obtain an accurate performance estimation of annotators, such as annotator behavior. For instance, Mao et al. [69] construct statistical model which predicts signals of the attention and effort that workers allocate to tasks by learning from the data of user behavior like the disengagement of annotators from the work or the consumed time on each image. This kind of method is specialized with the scenarios of domains or tasks, whose efficiency is improved but generalization could be limited. The proposed method aims to include additional information to efficiently evaluate the label quality remaining good utility in common label selection scenarios.

In view of the characteristics of RS data annotation and the limitations of current crowdsourced annotation methods, this article aims to explore a standardized RS data annotation framework, which could be roughly described as two processes. First, annotators' ability to recognize different classes of RS data was estimated, and annotators with stronger recognition ability were more trusted, similar to the first strategy that improves the annotation level. Next, the proposed method selects the most credible annotations for each data by referring to the indicative information of the annotators' recognition ability and the annotation results. The methodology is elaborated in Section III.

III. Method

This section introduces the proposed labeling framework for selecting the most credible labels from the crowdsourced labels[3] of RS data, which mainly consists of two modules, imbalanced annotator labeling-level inference and quality evaluation of crowdsourced labels. The first module tests annotators by sampling similar data that belong to different classes and infers their individual labeling level on each class. On this basis, the module of quality evaluation uses the inferred annotator levels to weight their outcome labels and update the quality of labels with an iterative algorithm. The details are introduced as follows.

A. Imbalanced Labeling Level Inference of Annotators

To assess the annotator level on labeling different classes of RS data, images with similar visual features but belonging to different classes are sampled to test the annotators. The label outcomes of the test procedures are subsequently analyzed to infer the labeling levels of annotators on each class. The processes are introduced as follows.

1) Image Organization for Annotator Test: The images are chosen and organized as groups to test annotators. Each group consists of similar images in different classes. The similarity between images is assessed by the cosine similarity of their visual representations. The visual representations are obtained using a pretrained momentum contrast (MoCov2) model [70], which is an effective visual representation learning method that can learn visual features with large-scale image datasets in an unsupervised way.

To effectively choose images for the annotator test, the similarity between classes is first calculated based on the mean visual representation of each class' images. On this basis, the most similar classes are detected, and then the images of similar classes are sampled according to image similarity. Take partial classes of the dataset RSI-CB as an example, the similarity between these classes is presented in Fig. 4. We can observe that there are two groups of similar classes, "city road" & "fork road" and "dry farm" & "natural grass." Then images are sampled and grouped by these similar classes. An example of the annotator test images is

[3]Crowdsourced labels refer to a collection of labels provided by different annotators for a sample to be annotated.

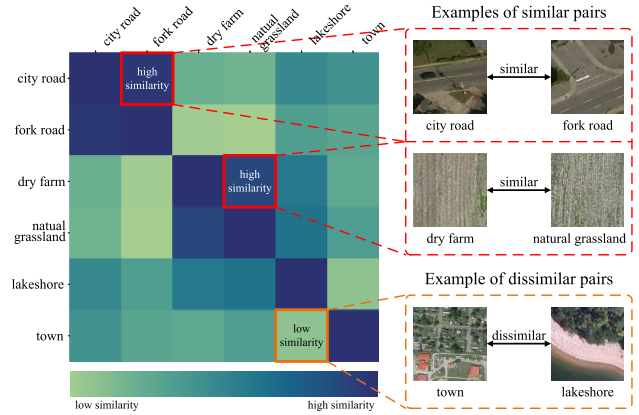


Fig. 4. Example of the similarity condition between different classes of RS data.

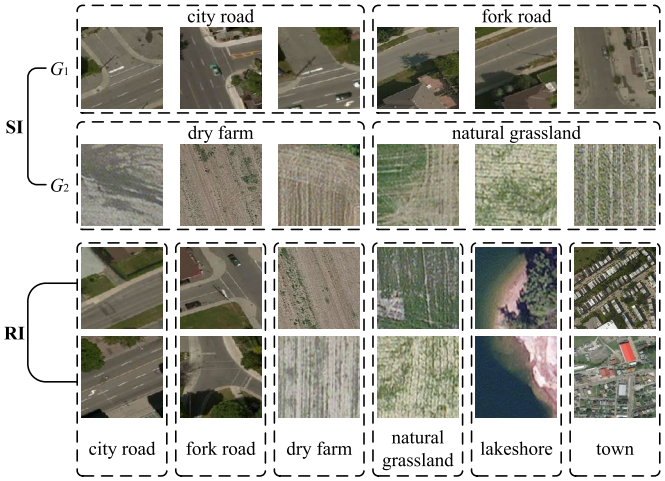


Fig. 5. Example of sampled data in different groups for the annotator test procedure.

presented in Fig. 5. The chosen images could be formalized as $SI = \{G_k\}, k \in \{1, 2, \ldots, m\}$, where G_k is a group of sampled images containing similar images from different classes. In addition, there are a group of images randomly chosen from all the classes, denoted as RI. The outcome labels by annotators on SI and RI are subsequently analyzed to infer their imbalanced labeling levels.

2) Annotator Labeling Level Inference: The labeling level of an annotator is modeled by two factors, the overall accuracy and the distinguishing ability on similar classes. The overall accuracy of an annotator is calculated by the accuracy of annotations on the random images RI, and that of the ith annotator is denoted as p_i

$$p_i = \mathrm{Acc}_i(RI) \tag{1}$$

where $\mathrm{Acc}_i(RI)$ is the accuracy of the outcome labels of the ith annotator on the images of RI. It is used as the initial weight of labels of the annotators in the quality evaluation algorithm introduced in Section III-B. As for the distinguishing ability on similar classes of an annotator, it is assessed according to its performance on similar images SI. This ability is denoted as the set of similar classes that the annotator cannot distinguish, i.e., the set of indistinguishable labels, expressed as

$$L_i = \{\mathrm{Class}(G_k) | \mathrm{Acc}_i(G_k) \le \lambda\}, \quad k \in \{1, 2, \ldots, m\} \tag{2}$$

where $\mathrm{Class}(G_k)$ represents similar classes whose images are sampled in the group G_k. $\mathrm{Acc}_i(G_k)$ is the accuracy of the outcome labels of the ith annotator on the images of G_k. λ is the threshold to decide whether an annotator can distinguish similar classes.

B. Quality Evaluation of MultiAnnotator Labels

To select the most credible label from the outcomes of multiple annotators, an iterative algorithm is designed to evaluate the quality of the outcome labels on each image, and the one with the highest quality is taken as the final output. Instead of directly voting for the most credible label, the algorithm simulates a process of multi-round "debate." In a round, each annotator "debates" with others, increasing its "confidence" in its own label if the labeling results are the same, and decreases if not. This "confidence" on a label corresponds to its quality, and the change in the "confidence" in the multi-round "debate" corresponds to the iterative update of label quality. The following discussions show the debate and the corresponding evaluation process for different annotators on a certain image. Note that the symbol representing the image is omitted since the label quality evaluation process on each image is identical and independent.

The label quality of annotator a_i is expressed as

$$Q_i = (E_i, V_i), \quad E_i \in [0, 1], \ V_i \in [0, \infty) \tag{3}$$

where E_i and V_i represent the expectation and the variance of the estimated quality, respectively. As for how to update the quality, the algorithm integrates the annotator level as a weight to decide the update amount of its label quality. A main idea motivating this algorithm is that the annotator with high confidence and high labeling level tends to believe its outcome and not to be affected by others, whose label quality is thus changed less, and vise versa. To initialize the confidence of annotators on the labels, their overall accuracy is used, shown as follows:

$$E_i = p_i, V_i = 0, \quad i = 1, 2, \ldots, n \tag{4}$$

where p_i is calculated by (1). On this basis, the label quality of each annotator is updated according to the agreement of its outcome labels with others'. The overall update equation is presented as follows:

$$\widetilde{E_i} = S\big(S^{-1}(E_i) + f(E_i, L_i; Y, \{L_k\}, \{E_k\})\big)$$
$$i, k \in \{1, 2, \ldots, M\}, i \ne k, Y = \{y_1, y_2, \ldots, y_M\} \tag{5}$$

where y_i is the label of the ith annotator. $\widetilde{E_i}$ is the updated quality of label y_i. L_i and L_k are the indistinguishable classes of annotators, calculated by (2). E_i and E_k are the label quality before update. $f(\cdot)$ is the update amount based on the agreement condition (Y), the confidence of annotators (E_i, E_k), and the annotator level on the class (L_i, L_k). S^{-1} is a function that maps the quality space to the update amount space, and S is the inverse function of S^{-1}. S and S^{-1} are needed since the update process could make E_i out of its range if without them, expressed as

$$S(x) = \frac{1}{1 + e^{-x}}, S^{-1}(x) = \ln\left(\frac{x}{1 - x}\right). \tag{6}$$

The update amount $f(\cdot)$ in (5) is defined as follows:

$$f(E_i, L_i; Y, \{L_k\}, \{E_k\}) = \frac{1}{M}\sum_{k=1}^{M} c_{i,k} g(E_i, E_k; L_k) \quad (7)$$

$$c_{i,k} = \begin{cases} 1, & y_k = y_i \\ -1, & y_k \neq y_i \end{cases} \quad (8)$$

$$g(E_i, E_k; L_k) = \frac{1}{|L_k^{y_k}|} E_k(1 - E_i) \quad (9)$$

$$L_k^{y_k} = \{C | C \in L_k, \text{truth} \in C\} \quad (10)$$

where $c_{i,k}$ indicates whether annotator a_k agrees with annotator a_i, and $g(E_i, E_k; L_k)$ is the corresponding update amount of the quality of y_i. The symbol "truth" is the true class of the target image. Equation 8 indicates that if a_k agrees/disagrees a_i, the quality of y_i has a positive/negative update. Equation 9 suggests that the update amount is positively correlated with E_k and is negatively correlated with E_i. It is also negatively correlated with $|L_k^{y_k}|$, which represents the number of similar classes that a_k cannot distinguish with the class of y_k (including the class of y_k itself).

The design of (7)–(10) is motivated by the following situations. 1) The confidence of an annotator increases if the label is agreed by others and decreases if not. 2) The change in an annotator's confidence is greater when it debates with one having great confidence (vise versa), and it is less when the annotator itself has strong confidence (vise versa). 3) The ability of an annotator to distinguish the class of the target image affects the change in other annotators' confidence during debating.

The quality of outcome labels from all the annotators is updated through the aforementioned process in a round, and the outcome quality $\widetilde{E_i}$ is taken as the initial quality E_i in the next round. This iterative process repeats T times to obtain the final outcome quality. In comparison to quality evaluation by directly voting, this iterative way highlights the power of labeling ability. By multiple rounds of updates, the annotators that hold the truth but are in a small number could finally outcompete those making errors, whereas direct voting may still miss the truth with only one round. Algorithm 1 clarifies the calculation process of update expectation in detail.

Specifically, the variance V is calculated as follows:

$$V_i = P_i^2 + N_i^2 \quad (11)$$

where the expressions of P_i and N_i are

$$P_i = S\left(S^{-1}(E_i) + \sum_{c_{i,k}=1} c_{i,k} g(E_i, E_k; L_k)\right) - E_i \quad (12)$$

$$N_i = E_i - S\left(S^{-1}(E_i) + \sum_{c_{i,k}=-1} c_{i,k} g(E_i, E_k; L_k)\right). \quad (13)$$

The disparity between support and opposition for E_i indicates the controversy over the credibility of target labeling. When the conflict of opinions is intense, the degree of support and opposition is significantly greater than 0, and then the obtained variance is large. If most of the annotators agree, the support or opposition is weak and the variance is small. In particular, both P_i and N_i tend to 0 if the update of E_i tends to be stable. At this time, it could be considered that the estimation of E_i is relatively accurate. The specific calculation process of V_i could be referred to Algorithm 2.

Algorithm 1 Quality Expectation Evaluation Based on MultiAnnotator

Input: Accuracy of annotators $p_1, p_2, \ldots, p_M$;
Corresponding label set of target RS data $Y = \{y_1, y_2, \ldots, y_M\}$;
The set of indistinguishable label $L_1, L_2, \ldots, L_M$;
Number of total iterations T.
Output: Quality expectation of annotators $E_1, E_2, \ldots, E_M$.

for i=1:M **do**
  $E_i = p_i$; // Initialization
end
for $t = 1{:}T$ **do**
  // In the t^{th} iteration
  for $i = 1{:}M$ **do**
    // Quality update of label y_i
    for $k = 1{:}M$ **do**
      $L_k^{y_k} = \{C | C \in L_k, y_k \in C\}$; // (10)
      $g(E_i, E_k; L_k) = \frac{1}{|L_k^{y_k}|} E_k(1 - E_i)$; // (9)
      if $y_k = y_i$ **then**
        $c_{i,k} = 1$
      else
        $c_{i,k} = -1$
      end
    end
    $f(E_i, L_i; Y, \{L_k\}, \{E_k\}) = \frac{1}{M}\sum_k c_{i,k} g(E_i, E_k; L_k)$; // (7)
    $\widetilde{E_i} = S(S^{-1}(E_i) + f(E_i, L_i; Y, \{L_k\}, \{E_k\}))$; // (5)
  end
  for $i = 1{:}M$ **do**
    $E_i \longleftarrow \widetilde{E_i}$
  end
end

Algorithm 2 Quality Variance Evaluation Based on MultiAnnotator

Input: Quality expectation of annotators $E_1, E_2, \ldots, E_M$;
$g(E_i, E_k; L_k)$ and $c_{i,k}$ calculated in the last iteration of Alg. 1
Output: Variance of annotation credibility $V_1, V_2, \ldots, V_M$.

for i=1:M **do**
  // (12,13)
  $P_i = S(S^{-1}(E_i) + \sum_{c_{i,k}=1} c_{i,k} g(E_i, E_k; L_k)) - E_i$
  $N_i = E_i - S(S^{-1}(E_i) + \sum_{c_{i,k}=-1} c_{i,k} g(E_i, E_k; L_k))$
  $V_i = P_i^2 + N_i^2$; // (11)
end
return $V_1, V_2, \ldots, V_M$;

Finally, we end this section by introducing how to select credible labels from the outcomes of annotators through E and V. As aforementioned, E is the expectation of label quality, while V is the variance of label quality. Intuitively, when V of different label quality is identical, we should select the label with the highest E. While E is identical, higher V indicates higher volatility, i.e., the real quality could be lower than E with a higher possibility. In this case, we would choose the label with the lowest V that corresponds to the most confident E. From this perspective, the indicator $E - \sqrt{V}$ is used as the quality estimation result, and the label with the highest estimation of $E - \sqrt{V}$ is selected as the final result.

The design of the quality indicator is motivated by the "empirical rule" of the Gaussian distribution.

TABLE I
DETAILED INFORMATION OF DATASETS

Dataset	Number of classes	Number of images per class	Size of training set	Size of test set
UC-Merced [19]	21	100	1890	210
RSI-CB [20]	45	100	4050	450
Pipeline	6	154/189/201/161/157/147	807	202
Tianjin [76]	2	1500	2000	1000

As introduced in (3), the quality of labels is modeled by (E, V) instead of a single value. We subsequently discuss its meaning from the perspective of the Gaussian distribution. Gaussian distribution is the most universal in different scenarios, since it has the largest entropy among all the distributions that satisfy the mean and variance [71]. Therefore, in this article, it is assumed that the probability distribution of the label quality is a Gaussian distribution $N(\mu, \sigma^2)$. For each label, E and V are the estimation of the mean μ and variance σ^2 of its quality distribution. According to the characteristics of the Gaussian distribution, $E - \sqrt{V}$ $(\mu - \sigma)$ actually calculates a value which the real quality exceeds with a probability 0.6826. This indicator actually represents a confident lower bound of quality estimation [72], [73], [74], [75].

IV. EXPERIMENTAL RESULTS

A. Datasets

The used datasets are (A) UC-Merced [19], (B) RSI-CB [20], (C) Pipeline, and (D) Tianjin. The detailed information of the datasets and the split between the training and test sets is concluded in Table I. Note that for any dataset, the proportion of each class used for testing and training is the same.

1) UC-Merced: Published by the United States Geological Survey (USGS) in 2010, UC-Merced is a classical satellite imagery dataset for RS image scene classification. It includes 2100 images sized 256 × 256 pixels with 0.3 m/pixel spatial resolution.

2) RSI-CB: RSI-CB is an RS image classification benchmark published in 2020, which is constructed using crowdsourced data. RSI-CB contains 45 classes and has more than 36 000 images sized 128 × 128 pixels with a spatial resolution ranging from 0.22 to 3 m/pixel. Experiments were performed on partial images of RSI-CB.

3) Pipeline: Pipeline is the dataset originally constructed. The source data were collected in urban areas using the ground-penetrating radar (GPR). After processing the obtained GPR B-scan images, the dataset contains a total of more than 1000 segmented B-scan images. In this article, the size of each segmented image is set to 321 × 321 pixels, where the first 321 indicates the distance traveled (the horizontal width of one pixel corresponds to 1.41 cm), and the second 321 indicates the depth beneath the ground. The measured data may contain objects including pipeline or subsurface anomalies (as shown in Fig. 5), where subsurface anomalies are fine-grained into void, crack, loose, and cavity. In addition, the remaining images that do not contain any objects or subsurface anomalies are annotated as normal.

4) Tianjin: The source data of Tianjin is obtained by visible light RS with satellites, with 0.2 m/pixel spatial resolution, and covering an area of 38.4 km^2 in the urban area of Tianjin. Buildings in RS images for object detection are manually labeled with ArcGIS [77].

In this article, the source dataset is modified and used for a binary classification task of whether the area in the image contains buildings. To enhance the joint recognition quality of multiple buildings at close range, we performed morphological expansion operations three times with a width of 10 pixels on the boxes with buildings, and the obtained target expansion results were used in the reuse method frame out. Finally, we take the target smaller than 512 × 512 in the source data as the positive samples in this article, while the negative ones are randomly sampled from the image using the same size boxes. The negative region does not contain the part that intersects with the positive region.

B. Experimental Setup

Datasets (A) and (B) are used to simulate annotation results that come from multiple annotators. Original labels of the images in Datasets (A) and (B) are taken as true labels, and annotations with wrong labels are generated based on two parameters. 1) Annotator's labeling level p controls the proportion of images that the annotator correctly labels. The label from an annotator with level p is set to a wrong label with a probability $1 - p$. 2) The labeling difficulty setting d indicates the types of labels that an annotator is short in distinguishing. For example, if indistinguishable label of an annotator under d is $(L1, L2, L3)$, images with labels $L1$, $L2$, and $L3$ are randomly set to label $L1$, $L2$, or $L3$. Images could be wrongly labeled as they are similar to some types of labels. On this basis, it is assumed that images of each label could only be mislabeled as its most similar label. However, they could be mislabeled as different labels by certain annotators that could not distinguish the labels, which depends on the setting of d.

Datasets (C) and (D) are labeled by six real annotators. They are used for the evaluation of the proposed method in real-world scenarios. Various models are trained with the labels obtained from different label selection methods. In addition, a pretrained MoCov2 model on the ImageNet [70] dataset is used to sample similar images. The encoder of the pretrained MoCov2 model is fine-tuned with the used data to obtain the vector representation of 1000 dimensions of GPR images. Moreover, four DL models are used, whose main parameter settings are: 1) total epochs of training: 500; 2) training batch size: 64; 3) learning rate: 0.001; 4) weight decay ($L2$ loss on parameters): 0.001; and 5) input images are resized as 128 × 128. All the experiments are run on a dedicated server with OS: Ubuntu 20.04, CPU: Intel Core i5-11500, and GPU: Nvidia GeForce RTX 3090.

C. Results and Analysis

This section conducts experiments based on both the simulated data and real-world data, which answers the following research questions accordingly.

RQ1: How does the proposed method compete with other label selection methods under various simulation cases?

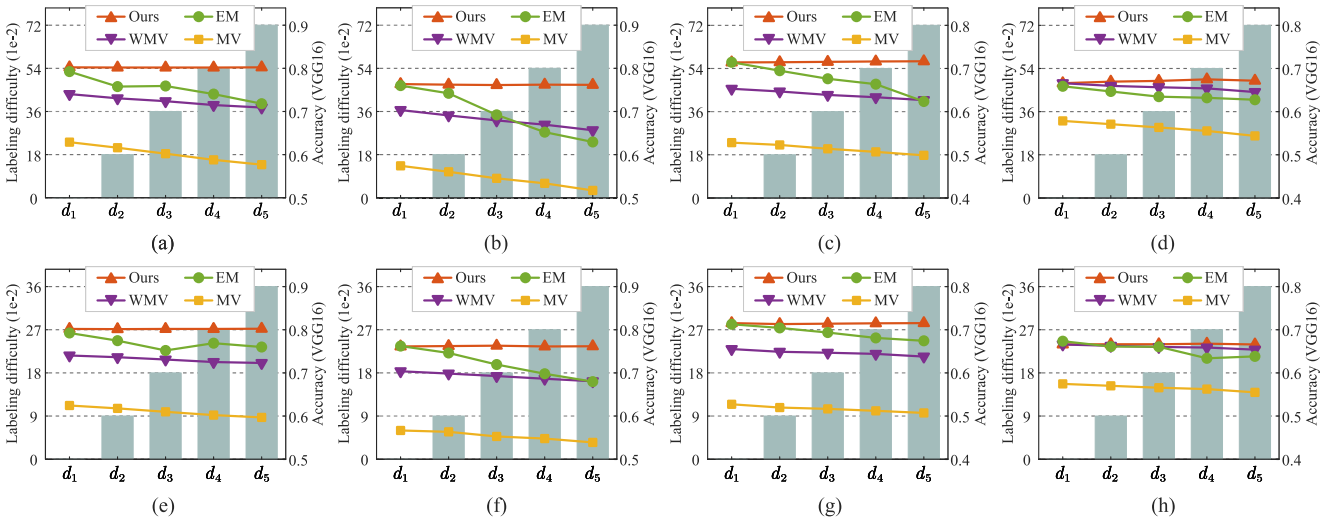


Fig. 6. Accuracy of VGG16 trained on the selected labels by the four methods under different conditions of labeling difficulty and annotator levels. (a) Dataset (A), $p1$. (b) Dataset (A) $p2$. (c) Dataset (A) $p3$. (d) Dataset (A) $p4$. (e) Dataset (B), $p1$. (f) Dataset (B), $p2$. (g) Dataset (B), $p3$. (h) Dataset (B), $p4$.

RQ2: Under various simulated conditions, how much could the labels selected by the proposed method improve the classifier's performance?

RQ3: What are the benefits of the procedure of annotator test and the mechanism of distrusting indistinguishable labels?

RQ4: How does the proposed method perform under different conditions of annotator numbers?

RQ5: How does the proposed method perform on datasets labeled by multiple real annotators?

RQ6: How does the accuracy of image labels influence the performance of models trained on the images?

1) Comparison Experiments (RQ1 & RQ3): The proposed method and three other methods are performed on generated labels based on Datasets (A) and (B), including majority vote (MV), weighted MV (WMV), and expectation–maximization (EM) algorithm. Experiments are conducted in different p and d to show the performances of these methods in various cases. The settings of d and p are presented in Tables II and III, respectively. The accuracy of the methods is reported in Table IV. Note that from d_1 to d_5, the indistinguishable labels of annotators become more, which represents that more images are hard to label correctly (no indistinguishable label under d_1). Similarly, from p_1 to p_4, the labeling level of annotators becomes lower, representing that more annotators are bad at labeling the images.

From the results, it is observed that the proposed method significantly outperforms other methods under d_2 to d_5, which justifies the advantage of the proposed method in the case that partial images are hard to distinguish. Other methods' performance decreases as the labeling difficulty increases. In comparison, the proposed method keeps in good performance even under the most difficult labeling setting d_5, which demonstrates the effectiveness of the investigation of annotators' indistinguishable label. When an annotator labels some images as one of its indistinguishable labels, the proposed method decreases the confidence of the results and produces a more correct label selection.

2) Influence on Classifier Performance (RQ2): To investigate the benefits of our method on the performance of classifiers, VGG16 is trained on the selected labels by four comparative methods (our method, EM, WMV, and MV). Its accuracy under various labeling difficulty settings d and labeling level of annotators p is presented in Fig. 6. Note that the labeling difficulty is quantified as the average proportion of labels that an annotator cannot distinguish.

It is observed that the proposed method significantly improves the accuracy of VGG16 in most cases, especially under higher labeling difficulty and higher annotator level. Note that the EM algorithm is close to our method in

TABLE II

DETAILED SETTINGS OF d_1–d_5

–		L1 L2 L3	L4 L5 L6	L7 L8 L9	L10 L11 L12	L13 L14 L15	L16 L17 L18	L19 L20 L21	L15 L18 L21
d_2	A1								
	A2	x							
	A3		x						
	A4	x	x						
	A5	x	x						
d_3	A1	x							
	A2		x	x					
	A3		x		x				
	A4	x		x	x				
	A5	x	x	x	x				
d_4	A1		x	x					
	A2	x		x	x	x			
	A3	x	x		x	x	x		
	A4			x			x		
	A5	x	x		x	x	x		
d_5	A1		x	x					
	A2	x	x	x	x	x			
	A3	x			x	x	x	x	x
	A4		x			x	x	x	x
	A5	x		x	x		x	x	x

*A1-A5 represent the five annotators. L1-L21 represent 21 types of labels. The 'x' in the table indicates that the annotator in its row can not distinguish the labels in its column. Note that no labels are indistinguishable for each annotator under d_1, so a detailed representation of d_1 is omitted.

TABLE III

DETAILED SETTINGS OF p_1–p_4

–	A1	A2	A3	A4	A5
p_1	0.99	0.5	0.9	0.3	0.7
p_2	0.9	0.3	0.7	0.6	0.7
p_3	0.8	0.3	0.6	0.5	0.7
p_4	0.7	0.6	0.6	0.5	0.6

TABLE IV
ACCURACY OF THE PROPOSED METHOD AND OTHER THREE METHODS IN DIFFERENT CONDITIONS OF p AND d

Labeling level		p_1					p_2					p_3					p_4				
Labeling difficulty		d_1	d_2	d_3	d_4	d_5	d_1	d_2	d_3	d_4	d_5	d_1	d_2	d_3	d_4	d_5	d_1	d_2	d_3	d_4	d_5
Dataset (A)	Our method	0.991	0.990	0.990	0.990	0.991	0.904	0.901	0.899	0.901	0.900	0.793	0.794	0.796	0.798	0.799	0.685	0.693	0.697	0.705	0.698
	MV	0.602	0.573	0.542	0.510	0.485	0.479	0.448	0.414	0.388	0.351	0.376	0.364	0.344	0.328	0.310	0.489	0.472	0.455	0.437	0.411
	WMV	0.852	0.830	0.815	0.795	0.782	0.769	0.741	0.715	0.693	0.664	0.656	0.642	0.624	0.612	0.597	0.685	0.671	0.664	0.658	0.639
	EM	0.969	0.891	0.894	0.852	0.802	0.896	0.856	0.744	0.654	0.603	0.794	0.750	0.708	0.680	0.590	0.669	0.642	0.615	0.609	0.599
Dataset (B)	Our method	0.990	0.989	0.990	0.990	0.991	0.899	0.901	0.903	0.899	0.900	0.796	0.791	0.793	0.795	0.796	0.685	0.686	0.686	0.689	0.685
	MV	0.592	0.577	0.559	0.542	0.530	0.462	0.455	0.431	0.420	0.400	0.374	0.357	0.350	0.340	0.329	0.480	0.470	0.460	0.453	0.436
	WMV	0.852	0.843	0.831	0.818	0.813	0.770	0.758	0.745	0.731	0.718	0.661	0.647	0.642	0.636	0.622	0.685	0.678	0.672	0.669	0.658
	EM	0.969	0.929	0.878	0.916	0.896	0.901	0.865	0.805	0.757	0.716	0.790	0.771	0.747	0.719	0.704	0.702	0.674	0.673	0.613	0.623

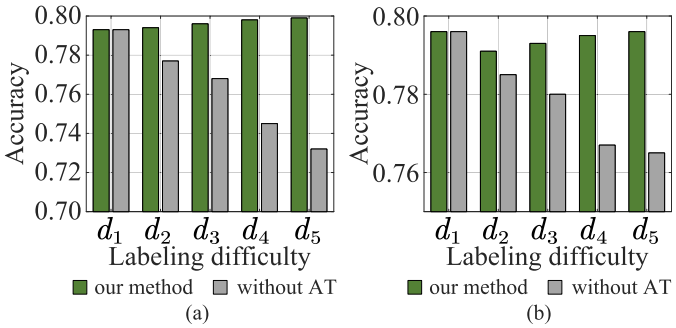


Fig. 7. Ablation study of the proposed method. (a) Dataset (A). (b) Dataset (B).

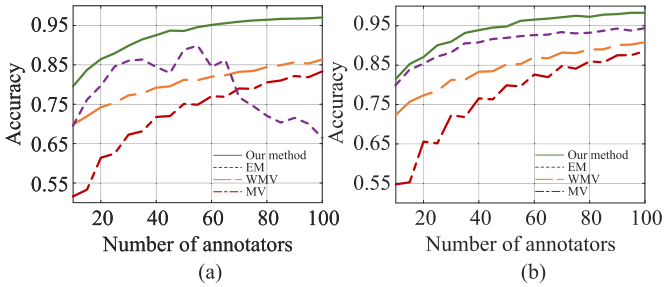


Fig. 8. Performance of the four methods under different number of annotators. (a) Dataset (A). (b) Dataset (B).

performance under d_1. However, d_1 represents an ideal case that all the images are equally difficult for labeling and the wrong labels are in a completely random distribution. In fact, in real-world scenarios, there are always partial images more difficult to label, which are more likely to be wrongly labeled than others. Therefore, d_2–d_5 better simulate the real-world situation, and the proposed method performs well in this case by the procedure of annotator test and the mechanism of distrusting indistinguishable labels.

3) Ablation Experiments (RQ3): To further reveal the benefits of investigating annotators' distinguishing ability on similar classes, an ablation study is performed by removing the procedure of annotator test (denoted as "without AT") to observe the performance of the procedure. Parameters p_3 and d_1–d_5 are selected for this ablation experiment. The results are reported in Fig. 7.

From the experimental result, the proposed method with annotator test outperforms that without annotator test under the existence of indistinguishable labels (d_2–d_5). As the labeling difficulty increases, the performance improvement also increases, which indicates that the annotator test and the indistinguishable label distrusting mechanism perform well to select correct labels under the high labeling difficulty of images. The indistinguishable labels of annotators are detected by the annotator test procedure, whose confidence is decreased in the label selection process to alleviate their interference to true label inference.

4) Influence of Annotator Number (RQ4): In real-world scenarios, the number of annotators participating in labeling may have a wide range of values. To investigate the effect of the proposed method under different numbers of annotators, it is performed with three comparative methods with numbers of annotators ranging from 10 to 100. The way to increase the number of annotators is to add five annotators with parameters p_3 and d_3 at a time. The results are presented in Fig. 8. It indicates that the proposed method outperforms other methods under the same number of annotators. In other words, the proposed method could produce as good labels as other methods in need of fewer annotators.

5) Study on Real-World Datasets (RQ5 & RQ6): Experiments are conducted on the images labeled by real annotators to investigate the performance of the proposed method in real-world scenarios. Four DL models, including VGG16, ResNet18, DenseNet12, and AlexNet, are trained on images of Datasets (C) and (D). The results labeled by six real annotators (each annotator labels all the images) are separately taken as training data for the image classification task. Moreover, the labels selected by our method are also taken as training data of the four models. Each experiment is repeated ten times and the best accuracy is taken. The accuracy of the four models is presented in Table V. It is observed that the proposed method significantly improves the model accuracy in most cases for real-world scenarios.

In addition, the detailed labeling accuracy of each annotators and its correlation with model accuracy are shown in Fig. 9. It is suggested that the models tend to perform better when trained on more accuracy labels. This result confirms the basic assumption of this article that the accuracy of training labels greatly affects the performance of models. Moreover, it is observed that the real points are mostly far from the fit line when the labeling accuracy is low, which indicates that the model performance is unstable in this case. When the labeling accuracy is high (i.e., greater than 90%), the real points are mostly near the fit line, in which case the models are prone to perform stably.

V. DISCUSSION

This section discusses some characteristics of the proposed method and the potential future work of this article.

TABLE V
ACCURACY OF FOUR DL MODELS TRAINED ON IMAGES LABELED BY SIX REAL ANNOTATORS AND OUR METHOD

-	Dataset (C)					Dataset (D)				
	Annotation	VGG16	Resnet18	Densenet121	Alexnet	Annotation	VGG16	Resnet18	Densenet121	Alexnet
A1	0.837	0.884(-0.053)	0.911(-0.026)	0.858(-0.053)	0.847(-0.058)	0.702	0.748(-0.092)	0.714(-0.136)	0.778(-0.084)	0.747(-0.102)
A2	0.777	0.868(-0.069)	0.895(-0.042)	0.837(-0.074)	0.821(-0.084)	0.781	0.758(-0.082)	0.739(-0.111)	0.777(-0.085)	0.756(-0.093)
A3	0.815	0.863(-0.074)	0.884(-0.053)	0.879(-0.032)	0.842(-0.063)	0.887	0.799(-0.041)	0.816(-0.034)	0.836(-0.026)	0.804(-0.045)
A4	0.831	0.853(-0.084)	**0.932(-0.005)**	0.874(-0.037)	0.811(-0.094)	0.727	0.739(-0.101)	0.736(-0.114)	0.77(-0.092)	0.813(-0.036)
A5	0.918	0.863(-0.074)	0.921(-0.016)	0.884(-0.027)	**0.905(0.0)**	0.807	0.764(-0.076)	0.754(-0.096)	0.779(-0.083)	0.715(-0.134)
A6	0.736	0.842(-0.095)	0.863(-0.074)	0.805(-0.106)	0.789(-0.116)	0.910	0.812(-0.028)	0.824(-0.026)	0.836(-0.026)	0.829(-0.02)
Ours	**0.957**	**0.911(-0.026)**	**0.932(-0.005)**	**0.9(-0.011)**	0.884(-0.121)	**0.963**	**0.840(-0.004)**	**0.845(-0.005)**	**0.859(-0.003)**	**0.838(-0.011)**
Truth	**1.000**	**0.937**	**0.937**	**0.911**	**0.905**	**1.000**	**0.844**	**0.850**	**0.862**	**0.849**

*Rows "A1-A6" refer to the label outcomes of six real annotators. The column "Annotation" is the label accuracy. Values in the parentheses represent the relative accuracy to the model trained with ground truth.

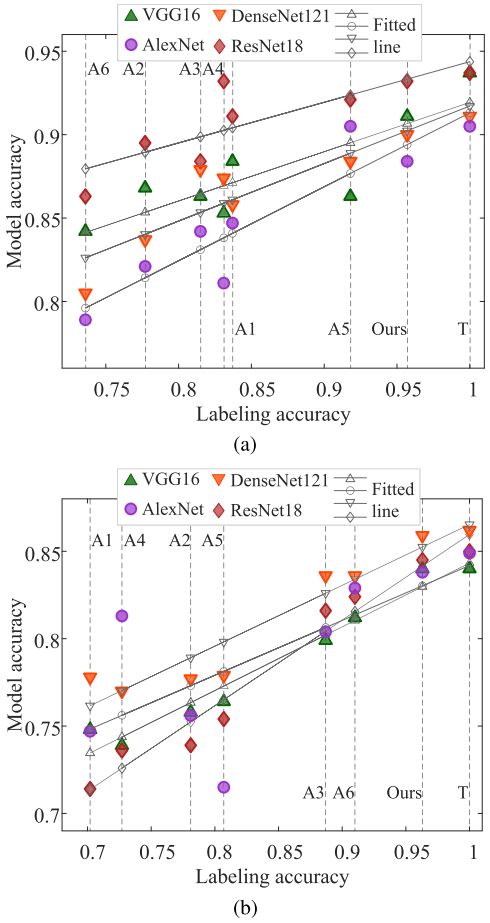


Fig. 9. Influence of the accuracy of labels on the performance of various models. (a) Dataset (C). (b) Dataset (D).

A. Application Scenarios in RS Data Annotation

The proposed RS data annotation framework aims to use the available nonexpert labor to label large-scale RS data with a satisfying accuracy. It serves the scenario that there is no enough expertise to afford the labeling work. Especially, the proposed framework focuses on the scenario that the scene classes are much more than two, in which case the estimation of distinguishing capability on similar classes has a sense. This is actually a common case in the RS field. We present 10 well-known researches publishing RS datasets and report their number of scene classes in Table VI. It suggests that all of them contain sufficient scene classes for the proposed labeling framework to work effectively. Actually, the label quality evaluation algorithm is specially designed to address the large number of scene classes. As for the further specialized design for RS data, it is discussed as follows.

TABLE VI
NUMBER OF SCENE CLASSES AND IMAGES] OF TEN WELL-KNOWN RS DATASETS

Researches	Images per class	Scene classes	Total images
Yang *et al.* [19]	100	21	2100
Xia *et al.* [78]	~50	19	1005
Zou *et al.* [79]	400	7	2800
Zhao *et al.* [80]	~100	11	1232
Zhao *et al.* [81]	200	12	2400
Xia *et al.* [18]	200~400	30	10000
Cheng *et al.* [10]	700	45	31500
Zhou *et al.* [82]	800	38	30400
Li *et al.* [83]	~800	45	36707
Long *et al.* [84]	500~3000	46	117000

B. Specialized Weights to Address Interclass Similarity

The specialized design of the proposed quality evaluation method is to build the weight based on the investigation of annotators' indistinguishable classes. This mechanism is motivated by the common issue of interclass similarity in RS data classification. For the issue, various detailed classification approaches of similar characteristics in RS data are widely documented in related researches. Despite this, it is difficult to force annotators to understand the differences in similar classes since our application case is about unprofessional and possibly irresponsible annotators. This work focuses on how to leverage the advantage of crowd's power instead of how to produce elite annotators. Therefore, the capability of multiple annotators on distinguishing the groups of similar classes is investigated to reweight the confidence of label outcomes. It discovers the strength of each individual on distinguishing partial similar classes. With sufficient individuals as a crowd, most interclass similarity could be hopefully addressed.

C. Implementation Flexibility

The proposed method designs an annotator test procedure to investigate the labeling ability of the annotators. It is logically separated from quality evaluation in the proposed framework,

requiring an additional step except for multiannotator labeling. In fact, it could be integrated into the quality evaluation process in implementation without an extra step. The data for known labels could be mixed with those to be labeled, which are both given to annotators to label. The annotator levels could be inferred by observing the accuracy of their label outcomes on the sampled images for the annotator test procedure.

D. Potential Adaption to Boundary Labeling

The proposed method serves the annotation of RS image scenes. In addition, object detection is also a mainstream task for RS data analysis. Except for object class annotation, it requires annotators to label the boundary of the interested objects as well. In this case, the core issue is to determine the most accurate boundary for each object of interest in an RS image. Since the labeled boundary is a type of continuous data rather than finite discrete values, the agreement condition between different annotators is not explicitly known. Whereas if certain measurement can be applied to judge whether two labeled boundaries are consistent, the proposed method could be processed on the consistent boundaries to select a credible label of object class. However, how to specify a most accurate boundary still needs more exploration. A possibly feasible solution could use certain measurement of object detection, such as intersection-over-union (IoU), to detect consistent boundaries and to refine the output boundary accordingly.

VI. Conclusion

This work proposes a nonexpert-driven accurate labeling refinement framework to construct high-quality labeled RS scene data. The common issue of interclass similarity in RS data is addressed in the proposed label quality evaluation algorithm, through a reweight mechanism based on the estimation of annotators' distinguishing capability on similar classes. The idea is to leverage the crowd's power to reach a satisfying annotation accuracy on large-scale RS data when lacking expertise. As a crowdsourcing-based annotation framework, the proposed framework eliminates extra manual interactions with annotators, which uses statistical analysis to infer annotators' imbalanced labeling capability for more credible quality evaluation. Thus, it provides a diagram of observation-based label selection to produce high-quality label that addresses the annotation issue brought by interclass similarity in RS data. Without complex interaction procedure or any expertise requirement, the proposed framework is expected to be able to motivate more generalized, low-cost, and high-quality RS annotation system researches.

References

[1] E. Mohan, A. Rajesh, G. Sunitha, R. M. Konduru, J. Avanija, and L. G. Babu, "A deep neural network learning-based speckle noise removal technique for enhancing the quality of synthetic-aperture radar images," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 13, p. e6239, Jul. 2021.

[2] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[3] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigearthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 5901–5904.

[4] 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)*, Jul. 2018, pp. 4721–4724.

[5] I.-F. Chen, B. King, and J. Droppo, "Investigation of training label error impact on RNN-T," 2021, *arXiv:2112.00350*.

[6] Y. Li, Y. Zhang, and Z. Zhu, "Learning deep networks under noisy labels for remote sensing image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3025–3028.

[7] T. Li, Z. Wang, and J. Liu, "Evaluation method for impact of jamming on radar based on expert knowledge and data mining," *IET Radar, Sonar Navigat.*, vol. 14, no. 9, pp. 1441–1450, Sep. 2020.

[8] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11621–11631.

[9] W. Zhang, L. Jiao, F. Liu, J. Liu, and Z. Cui, "LHNet: Laplacian convolutional block for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5626513.

[10] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[11] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3735–3756, Jun. 2020.

[12] P. Bota, J. Silva, D. Folgado, and H. Gamboa, "A semi-automatic annotation approach for human activity recognition," *Sensors*, vol. 19, no. 3, p. 501, Jan. 2019.

[13] D. Malmgren-Hansen and M. Nobel-Jorgensen, "Convolutional neural networks for SAR image segmentation," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2015, pp. 231–236.

[14] E. Salalioglu and O. Gungor, "Crowdsourcing in remote sensing: A review of applications and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 89–110, Dec. 2020.

[15] B. A. Johnson, K. Iizuka, M. A. Bragais, I. Endo, and D. B. Magcale-Macandog, "Employing crowdsourced geographic data and multi-temporal/multi-sensor satellite imagery to monitor land cover change: A case study in an urbanizing region of the Philippines," *Comput., Environ. Urban Syst.*, vol. 64, pp. 184–193, Jul. 2017.

[16] S. Xiong, C. Wang, C. Chen, B. Zhang, and Q. Li, "InSAR crowdsourcing annotation system with volunteers uploaded photographs: Toward a hazard alerting system," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[17] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[18] 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, Apr. 2017.

[19] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2010, pp. 270–279.

[20] H. Li et al., "RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, p. 1594, 2020.

[21] B. Li, C. Zhang, W. Pei, and L. Shen, "Design of ISAR image annotation system based on deep learning," in *Proc. Int. Conf. Comput. Eng. Netw.* Singapore: Springer, 2020, pp. 283–288.

[22] N. Scheiner, N. Appenrodt, J. Dickmann, and B. Sick, "Automated ground truth estimation of vulnerable road users in automotive radar data using GNSS," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Apr. 2019, pp. 1–5.

[23] J. Chen, D. Wang, I. Xie, and Q. Lu, "Image annotation tactics: Transitions, strategies and efficiency," *Inf. Process. Manage.*, vol. 54, no. 6, pp. 985–1001, Nov. 2018.

[24] C. Qiu, A. Squicciarini, D. R. Khare, B. Carminati, and J. Caverlee, "Crowdeval: A cost-efficient strategy to evaluate crowdsourced worker's reliability," in *Proc. 17th Int. Conf. Auto. Agents MultiAgent Syst.*, 2018, pp. 1486–1494.

[25] S. Ørting et al., "A survey of crowdsourcing in medical image analysis," 2019, *arXiv:1902.09159*.

[26] T. Ban, X. Wang, L. Chen, X. Wu, Q. Chen, and H. Chen, "Quality evaluation of triples in knowledge graph by incorporating internal with external consistency," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 4, 2022, doi: 10.1109/TNNLS.2022.3186033.

[27] X. Wang, T. Ban, L. Chen, X. Wu, D. Lyu, and H. Chen, "Knowledge verification from data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 9, 2022, doi: 10.1109/TNNLS.2022.3202244.

[28] V. C. Raykar et al., "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, no. 4, pp. 1297–1322, 2010.

[29] M. Ji and J. R. Jensen, "Effectiveness of subpixel analysis in detecting and quantifying urban imperviousness from Landsat thematic mapper imagery," *Geocarto Int.*, vol. 14, no. 4, pp. 33–41, Dec. 1999.

[30] L. L. F. Janssen and H. Middelkoop, "Knowledge-based crop classification of a landsat thematic mapper image," *Int. J. Remote Sens.*, vol. 13, no. 15, pp. 2827–2837, Oct. 1992.

[31] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[32] T. Blaschke and J. Strobl, "What's wrong with pixels? Some recent developments interfacing remote sensing and GIS," *Zeitschrift für Geoinformationssysteme*, vol. 14, pp. 12–17, Jun. 2001.

[33] T. Blaschke, "Object-based contextual image classification built on image segmentation," in *Proc. IEEE Workshop Adv. Techn. Anal. Remotely Sensed Data*, Oct. 2003, pp. 113–119.

[34] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[35] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sep. 2019.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[38] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, Nov. 1991.

[39] L. Liao, L. Du, and Y. Guo, "Semi-supervised SAR target detection based on an improved faster R-CNN," *Remote Sens.*, vol. 14, no. 1, p. 143, Dec. 2021.

[40] S. Khorram, F. H. Koch, C. F. van der Wiele, and S. A. Nelson, *Remote Sensing*. New York, NY, USA: Springer, 2012.

[41] H.-C. Li, W.-S. Hu, W. Li, J. Li, Q. Du, and A. Plaza, "A3CLNN: Spatial, spectral and multiscale attention ConvLSTM neural network for multisource remote sensing data classification," 2022, *arXiv:2204.04462*.

[42] S. Mei, X. Chen, Y. Zhang, J. Li, and A. Plaza, "Accelerating convolutional neural network-based hyperspectral image classification by step activation quantization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.

[43] L. Jiao, J. Gao, X. Liu, F. Liu, S. Yang, and B. Hou, "Multi-scale representation learning for image classification: A survey," *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 23–43, Feb. 2021.

[44] B. C. Benato, J. F. Gomes, A. C. Telea, and A. X. Falcão, "Semi-automatic data annotation guided by feature space projection," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107612.

[45] H. Li, N. Li, R. Wu, H. Wang, Z. Gui, and D. Song, "GPR-RCNN: An algorithm of subsurface defect detection for airport runway based on GPR," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3001–3008, Apr. 2021.

[46] V. Kafedziski, S. Pecov, and D. Tanevski, "Detection and classification of land mines from ground penetrating radar data using faster R-CNN," in *Proc. 26th Telecommun. Forum (TELFOR)*, Nov. 2018, pp. 1–4.

[47] X. Qi et al., "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.

[48] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 399–407.

[49] Y. Gu, Y. Wang, and Y. Li, "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection," *Appl. Sci.*, vol. 9, no. 10, p. 2110, 2019.

[50] A. K. Aksoy, M. Ravanbakhsh, and B. Demir, "Multi-label noise robust collaborative learning for remote sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 20, 2022, doi: 10.1109/TNNLS.2022.3209992.

[51] 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.

[52] Q. Li, Y. Chen, and P. Ghamisi, "Complementary learning-based scene classification of remote sensing images with noisy labels," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[53] C. Torres-Huitzil and B. Girau, "Fault and error tolerance in neural networks: A review," *IEEE Access*, vol. 5, pp. 17322–17341, 2017.

[54] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 184–199, Mar. 2020.

[55] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2004, pp. 319–326.

[56] P. Upchurch, D. Sedra, A. Mullen, H. Hirsh, and K. Bala, "Interactive consensus agreement games for labeling images," in *Proc. AAAI Conf. Human Comput. Crowdsourcing*, vol. 4, 2016, pp. 239–248.

[57] C.-W. Lin, K.-T. Chen, L.-J. Chen, I. King, and J. H. Lee, "An analytical approach to optimizing the utility of ESP games," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, vol. 1, Dec. 2008, pp. 184–187.

[58] S.-W. Huang and W.-T. Fu, "Enhancing reliability using peer consistency evaluation in human computation," in *Proc. Conf. Comput. Supported Cooperat. Work*, Feb. 2013, pp. 639–648.

[59] C.-M. Chang, C.-H. Lee, and T. Igarashi, "Spatial labeling: Leveraging spatial layout for improving label quality in non-expert image annotation," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–12.

[60] H. Irshad et al., "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd," in *Pacific Symp. Biocomputing Co-Chairs*, 2014, pp. 294–305.

[61] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "OpenSurfaces: A richly annotated catalog of surface appearance," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–17, Jul. 2013.

[62] D. Gurari et al., "How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 1169–1176.

[63] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proc. Comput.-Hum. Interact. Conf. Hum. Factors Comput. Syst.*, 2017, pp. 2334–2346.

[64] T. Sunahase, Y. Baba, and H. Kashima, "Pairwise hits: Quality estimation from pairwise comparisons in creator-evaluator crowdsourcing process," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 977–984.

[65] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Who should label what? Instance allocation in multiple expert active learning," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2011, pp. 176–187.

[66] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Statist. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[67] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver, "How to grade a test without knowing the answers—A Bayesian graphical model for adaptive crowdsourcing and aptitude testing," 2012, *arXiv:1206.6386*.

[68] P. Welinder, S. Branson, P. Perona, and S. Belongie, "The multidimensional wisdom of crowds," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, 2010, pp. 1–9.

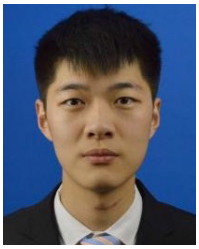
[69] A. Mao, E. Kamar, and E. Horvitz, "Why stop now? Predicting worker engagement in online crowdsourcing," in *Proc. 1st AAAI Conf. Hum. Comput. Crowdsourcing*, 2013, pp. 1–9.

[70] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[71] S. Y. Park and A. K. Bera, "Maximum entropy autoregressive conditional heteroskedasticity model," *J. Econometrics*, vol. 150, no. 2, pp. 219–230, 2009.

[72] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychol. Assessment*, vol. 6, no. 4, p. 284, Dec. 1994.

[73] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 73–79, 2011.
[74] L. Bornmann and W. Marx, "How good is research really?: Measuring the citation impact of publications with percentiles increases correct assessments and fair comparisons," *Eur. Mol. Biol. Org. Rep.*, vol. 14, no. 3, pp. 226–230, Mar. 2013.
[75] L. Bornmann and W. Marx, "Distributions instead of single numbers: Percentiles and beam plots for the assessment of single researchers," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 1, pp. 206–208, Jan. 2014.
[76] Y. Zhu, B. Huang, J. Gao, E. Huang, and H. Chen, "Adaptive polygon generation algorithm for automatic building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
[77] L. M. Scott and M. V. Janikas, "Spatial statistics in ArcGIS," in *Handbook of Applied Spatial Analysis*. Berlin, Germany: Springer, 2010, pp. 27–41.
[78] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC 7th Symp.*, vol. 38, Sep. 2010, pp. 298–303.
[79] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
[80] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, p. 035004, 2016.
[81] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2015.
[82] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS-J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
[83] H. Li et al., "RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data," 2017, *arXiv:1705.10450.*
[84] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.



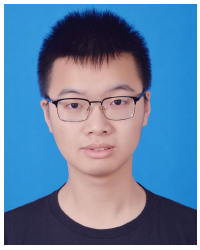
Xiangyu Wang received the B.Sc. degree from Donghua University, Shanghai, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Data Science, University of Science and Technology of China, Hefei, China.

His research interests include knowledge engineering and machine learning.

Lyuzhou Chen received the B.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Data Science. His research interests include ensemble learning, knowledge engineering, and causal learning.

Taiyu Ban received the B.Sc. degree in computer science and technology from the School of the Gifted Young, University of Science and Technology of China, Hefei, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology.

His research interests include machine learning and knowledge engineering.



Derui Lyu received the B.Sc. degree in intelligence science and technology from the School of Artificial Intelligence, Xidian University, Xi'an, China, in 2021. He is currently pursuing the M.Sc. degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China.

His research interests include machine learning and knowledge engineering.

Yifeng Guan received the B.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2021, where he is currently pursuing the M.Sc. degree with the School of Computer Science and Technology.

His research interests include knowledge engineering and causal learning.

Xingyu Wu received the B.Sc. degree from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China.

His research interests include causality-based machine learning, feature selection, causal learning, and causal inference.

Xiren Zhou received the B.Sc. degree from Shandong University, Jinan, China, in 2014, and the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2019.

He is currently an Associate Researcher with the School of Computer Science and Technology, USTC. His research interests include machine learning, ground-penetrating radar, and multisensor data fusion.

Huanhuan Chen (Senior Member, IEEE) received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004, and the Ph.D. degree in computer science from the University of Birmingham, Birmingham, U.K., in 2008.

He is currently a Full Professor with the School of Computer Science and Technology, USTC. His research interests include neural networks, Bayesian inference, and evolutionary computation.

Dr. Chen received the International Neural Network Society Young Investigator Award in 2015, the IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award in 2012, the IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award (bestowed in 2011 and only one paper in 2009), and the British Computer Society Distinguished Dissertations Award in 2009. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEM and the IEEE TRANSITION ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.