Remote Sensing Image Retrieval in the Past Decade: Achievements, Challenges, and Future Directions

Weixun Zhou, *Member, IEEE*, Haiyan Guan, *Senior Member, IEEE*, Ziyu Li, Zhenfeng Shao, and Mahmoud R. Delavar

***Abstract*—Remote sensing image retrieval (RSIR) aims to search and retrieve the images of interest from a large remote sensing image archive, which has remained to be a hot topic over the past decade. Benefited from the advent and progress of deep learning, RSIR has been promoted by developing novel approaches, constructing new datasets, and exploring potential applications. To the best of our knowledge, there lacks a comprehensive review of RSIR achievements, including systematic and hierarchical categorization of RSIR methods and benchmark datasets over the past decade. This article, therefore, provides a systematic survey of the recently published RSIR methods and benchmarks by reviewing more than 200 papers. To be specific, in terms of image source, label, and modality, we first group the RSIR methods into some hierarchical categories, each of which is reviewed in detail. Following the categorization of the RSIR methods, we list the benchmark datasets publicly available for performance evaluation and present our newly collected RSIR dataset. Moreover, some of the existing RSIR methods are selected and evaluated on the representative benchmark datasets. The results demonstrate that deep learning-based methods are currently the dominant RSIR approaches and outperform handcrafted feature-based methods by a significant margin. Finally, we discuss the main challenges of RSIR and point out some potential directions for the future RSIR research.**

***Index Terms*—Deep learning, feature extraction, literature review, remote sensing image retrieval (RSIR), similarity measure.**

I. Introduction

OVER the past decades, remote sensing (RS) earth observation has reached an unprecedented level, and the available RS data have grown exponentially; however, we are overwhelmed by the massive data with too much useless information due to the limitation of data processing techniques. Therefore, in the era of RS big data, how to efficiently organize and manage large RS archives and quickly search and retrieve the data of interest remains to be a significant challenge in the RS community.

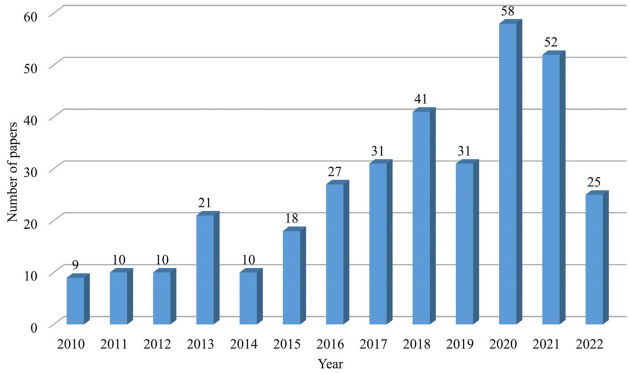


Fig. 1. Number of open publications on RSIR from 2010 to 2022. Data are collected by the advanced search of Google Scholar (all in title: "RS" OR aerial OR satellite "image retrieval").

Remote sensing image retrieval (RSIR), which aims to search and retrieve images of interest from a large RS image archive, is an effective technique to solve the problems mentioned above [1]. In the early years, many RSIR methods are derived from the text-based image retrieval and particularly content-based image retrieval (CBIR) in computer vision (CV) field [2], [3], [4]. Although RSIR can be regarded as the application of CBIR in the RS field, RSIR is a more challenging task than CBIR due to the high complexity of RS images, including multiscale objects, varied resolutions, different imaging modes, and so forth. To advance RSIR, RS literature has invested significant effort to develop RS image-specific methods, making RSIR an active research topic, as shown in Fig. 1. The number of publications on RSIR has dramatically increased over the past decade, especially from the year of 2012. On the one hand, deep learning achieved remarkable performance on the large-scale ImageNet in 2012, which has drawn much attention from the RS community since then. On the other hand, RS benchmark datasets have been increasingly constructed and publicly available, making it possible for developing deep learning based RSIR methods.

RSIR in early years is performed based on the metadata (thereby also called metadata or text-based RSIR), such as

Manuscript received 16 November 2022; revised 27 December 2022; accepted 8 January 2023. Date of publication 12 January 2023; date of current version 27 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42001285 and in part by the Natural Science Foundation of Jiangsu Province, China under Grant BK20200813. *(Corresponding author: Weixun Zhou.)*

Weixun Zhou and Haiyan Guan are with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: zhouwx@nuist.edu.cn; guanhy.nj@nuist.edu.cn).

Ziyu Li is with the School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China (e-mail: 221309020008@hhu.edu.cn).

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

Mahmoud R. Delavar is with the Center of Excellence in Geomatic Engineering in Disaster Management, and Land Administration in Smart City Lab., School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran 14174, Iran (e-mail: mdelavar@ut.ac.ir).

Digital Object Identifier 10.1109/JSTARS.2023.3236662

image resolution, geographic coordinate, sensor type, and image acquisition time; however, it does not take image content (i.e., buildings, roads, and rivers) into consideration [5]. Therefore, content-based RSIR (CBRSIR) that performs search and retrieval of RS images using low-level visual features [6], [7], [8], [9], [10], [11], [12], [13], [14] has gained much attention. Nevertheless, CBRSIR faces two significant challenges for large RS image archives: First, RS images usually contain a few object classes with varied scales; thus, low-level features are not available for the accurate representation of image content. Therefore, RSIR usually obtains unsatisfactory performance based on single low-level feature (e.g., spectral feature) or combined low-level features (e.g., spectral feature and texture feature); second, low-level features, known as handcrafted features, require laborious efforts. Moreover, it is not feasible to develop an effective feature representation suitable for RS images with different resolutions, different object types, and different image complexities. Thus, for CBRSIR, it is necessary to draw our attention from previously hand-engineered features to currently learned features. From here on, RSIR is termed for CBRSIR, unless otherwise stated.

Since 2012, deep learning has gradually developed as a dominant technique for feature extraction due to its remarkable performance on recognition tasks [15]. Accordingly, as an alternative to handcrafted features, deep learning has also explored by the RS community [16], [17], [18], [19], [20]. Yuan et al. [16] analyzed the potential of deep learning for environmental RS tasks (e.g., land cover mapping, environmental parameter retrieval, data fusion and downscaling, as well as information reconstruction and prediction). Ball et al. [17] gave a detailed survey of deep learning used for RS tasks in theories, tools, and challenges. Zhang et al. [18] provided a technical tutorial on the state-of-the-art (SOTA) deep learning techniques for RS big data from the four perspectives of image processing, pixel-based classification, target recognition, and scene understanding. Zhu et al. [19] analyzed the challenges of using deep learning for RS data analysis, reviewed the recent advances, and provided resources, attempting to ridiculously simple deep learning in the RS domain. Ma et al. [20] summarized several main subfields of deep learning used in RS and conducted a deep review to describe and discuss those techniques in all of these subfields.

These works demonstrate that deep learning has been one of the dominant techniques for RS tasks. Driven by deep learning, a great number of RSIR methods have been presented. The readers are referred to the reviews on RSIR [21], [22], [23], [24], [25]. Sudha and Aji [21] conducted a systematic study on the existing RSIR methods to guide the new researchers in the RS domain to choose effective methods for performance improvement of the RSIR system in different schemes. Gu et al. [22] comprehensively reviewed deep learning based methods for RS image understanding and pointed out some future directions and potential applications. Sudha and Aji [23] concentrated on the advancements and current trends related to deep learning based RSIR and analyzed how to use deep learning techniques and frameworks to address the challenges. Tong et al. [24] focused on three core issues of RSIR, i.e., feature extraction, similarity metric, and relevance feedback, and systematically investigated deep features for RSIR. In the recent work, Li et al. [25] systematically reviewed the emerging achievements of RSIR and discussed its applications, including fusion-oriented RS image processing, geolocalization, and disaster rescue. To date, it seems to be the most systematic and comprehensive review on RSIR. However, the existing RSIR methods were coarsely categorized into CBRSIR, hash-based RSIR, cross-modal RSIR (CMRSIR), and interactive RSIR, which was not a reasonable division of RSIR methods. For instance, in most studies, CBRSIR, as one kind of RSIR methods, actually contains hash-based RSIR, CMRSIR, and interactive RSIR. While in [25], they were categorized as four paralleled methods. Furthermore, the existing benchmark datasets were also coarsely categorized into single-modality data and multimodality data.

TABLE I
COMPARISONS BETWEEN THE EXISTING SURVEY WORKS AND OURS

Existing Reviews	Comparison with Ours
Sudha and Aji [21], [23]	The two works focused on RSIR models, deep learning techniques in RSIR challenges, and RSIR computational challenges, while ours focus on different types of RSIR methods and the corresponding benchmark datasets.
Gu et al. [22]	The authors introduced retrieval by distance measures, graph models, and hashing learning, which is pretty different from our work.
Tong et al. [24]	This work focused on feature extraction (low-level, mid-level, and high-level deep features) and extracted deep features, including fully connected and convolutional features for RSIR performance evaluation, While our work categorized the existing RSIR methods into five main groups and introduced the benchmark datasets for each kind of RSIR method.
Li et al. [25]	The main differences between our work and Li et al.'s lie in that we categorized the existing RSIR methods in a hierarchical category, and particularly, the methods were categorized into five groups. Additionally, our work reported results of representative methods of each group and presented the challenges and future directions for RSIR from a different perspective.

We, therefore, provide a comprehensive review of RSIR achievements, including RSIR methods and benchmark datasets over the past decade in this article. In addition, we also release a new dataset and present an RSIR method evaluated on the new released dataset. Our work categorizes the existing RSIR methods into a hierarchical category, and the benchmark datasets are categorized accordingly. To the best of our knowledge, it is the most sophisticated categorization of RSIR methods and benchmark datasets, and is complementary to the existing reviews. The comparisons between the several existing review works and ours are summarized in Table I.

The rest of this article is organized as follows. Section II surveys the conventional and deep learning RSIR methods of different categories. Section III introduces the benchmark datasets for the performance evaluation of RSIR. The performance metrics and results of RSIR methods are presented in Section IV. Section V discusses the current challenges and potential solutions for RSIR. Section VI points out some future directions of RSIR. Finally, Section VII concludes this article.

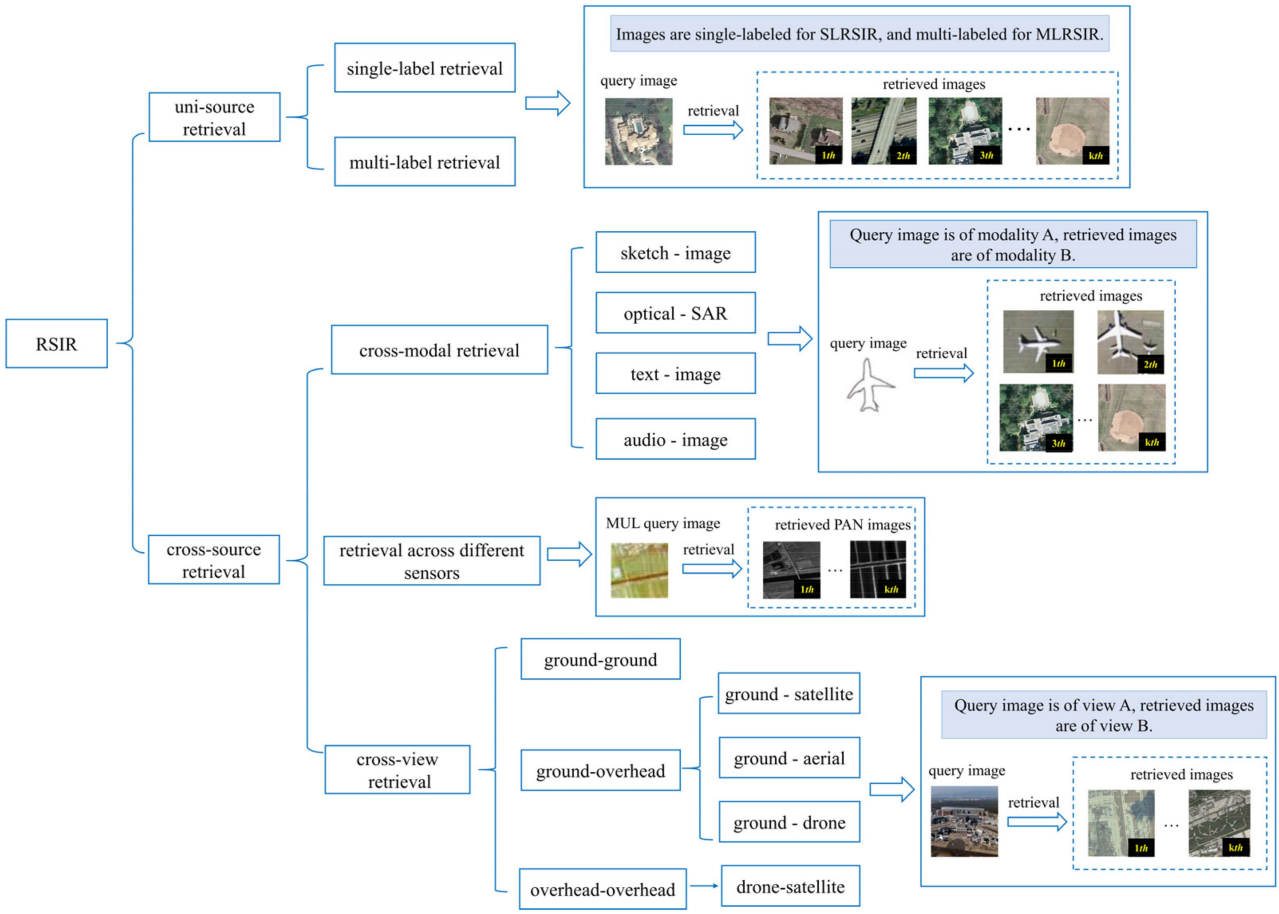


Fig. 2. Hierarchical category network of the existing RSIR methods.

II. Conventional and Deep Learning Methods for RSIR

Over the past decade, the RS community has witnessed the rapidly developed RSIR methods, including the handcrafted feature-based approaches and the recent deep learning based ones. To detail these RSIR methods, we organize the existing RSIR methods by a four-level hierarchical category network, in terms of data, as shown in Fig. 2. To our best of knowledge, the hierarchical categorization network provides a sophisticated organization of RSIR methods.

Similar to the scheme in [25], our hierarchical RSIR category network is composed of unisource retrieval and cross-source retrieval at the first level. The categorization criteria are that whether the query image and the retrieved images are from the same source or not. To be specific, for unisource retrieval, both the query image and the retrieved images are from the same source, while for cross-source retrieval, the query image and the retrieved images are from different sources, and generally two sources. In the second level, unisource retrieval is further categorized into single-label RSIR (SLRSIR) and multilabel RSIR (MLRSIR), depending on whether an image is associated with one label or multiple labels. Regarding cross-source retrieval, it is categorized into three subcategories, including CMRSIR, retrieval across different sensors (RASRSIR), and cross-view RSIR (CVRSIR), depending on whether the images are from the same modality, sensors, view or not. It is notable that for CMRSIR and CVRSIR, we also provide their subcategories to cover the existing RSIR methods. Specifically, CMRSIR consists of retrieval between sketch and image, retrieval between optical and SAR images, retrieval between text and images, as well as retrieval between audio and images. CVRSIR contains retrieval between ground images, retrieval between ground and overhead images (e.g., ground–satellite, ground–aerial, and ground–drone), as well as retrieval between overhead and overhead images (e.g., drone–satellite). Table II presents the categorization criteria and detailed description for each RSIR subcategory.

In the following part, we review the existing SLRSIR, MLRSIR, RASRSIR, CMRSIR, and CVRSIR methods, respectively, over the past decade. Because of the remarkable performance of deep learning for RSIR, we mainly focus on deep learning based RSIR methods.

A. SLRSIR Methods

SLRSIR is to perform retrieval with single-label images and has been the dominant RSIR methods. In this scenario, each of the query and retrieved images belongs to only one image

TABLE II
DETAILED CATEGORIZATION CRITERIA AND DESCRIPTION OF RSIR METHODS

RSIR Methods	Sub-Methods	Criteria	Description
single-label retrieval	-	Whether an image is associated with one label or multiple labels.	Each of the query and retrieved images is associated with one label.
multilabel retrieval	-		Each of the query and retrieved images is associated with at least one label.
cross-modal retrieval	sketch–image	Which group of data modality is used?	Retrieval between sketch and image.
	optical–SAR		Retrieval between optical and SAR images.
	text–image		Retrieval between text and image.
	audio–image		Retrieval between audio and image.
retrieval across different sensors	-	Whether the query is between images of different sensors.	The query and retrieved images are captured by different sensors. Particularly, retrieval between optical and SAR images is typically categorized to cross-modal retrieval.
cross-view retrieval	ground–ground	Which group of view are the images acquired?	Retrieval between ground- and ground-view images.
	ground–overhead		Retrieval between ground- and overhead-view images (i.e. drone, aerial, and satellite images).
	overhead–overhead		Retrieval between overhead- and overhead-view images, and generally drone and satellite images.

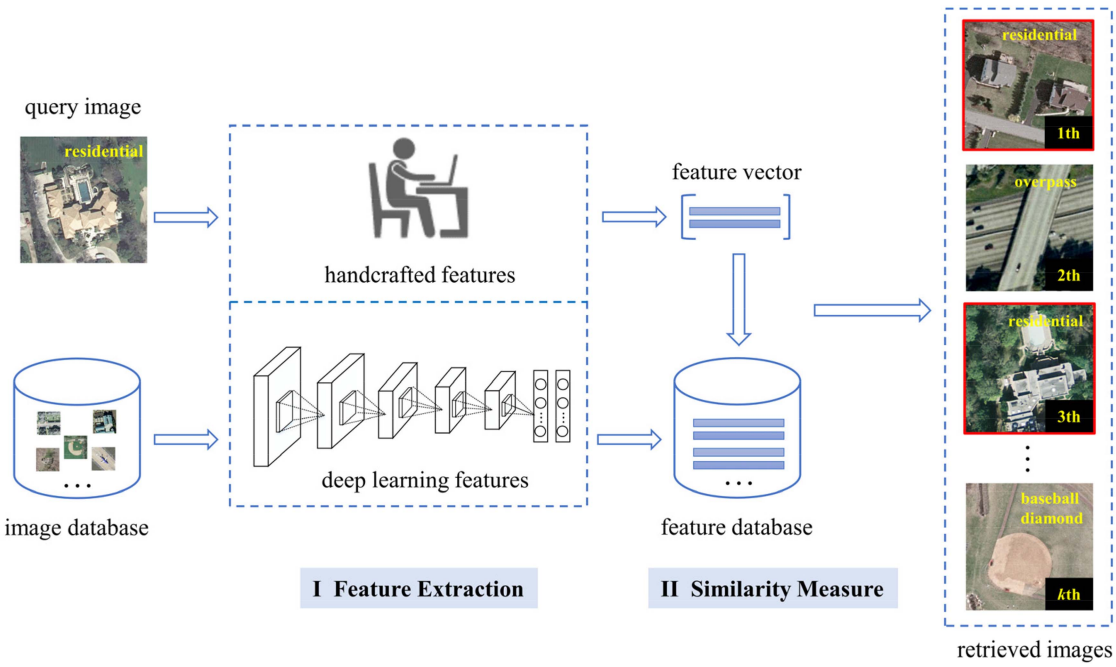


Fig. 3. Flowchart of SLRSIR method. The retrieved images with red rectangles stand for images that are correctly retrieved.

category. Generally, there are three modules in a typical RSIR system, including feature extraction, feature indexing, and similarity measures [25]. In practice, feature extraction and similarity measures are two indispensable parts because feature indexing is mostly used for large-scale RS image archives. Fig. 3 illustrates the flowchart of the SLRSIR method, which mainly contains two steps, including feature extraction and similarity measures. For feature extraction, the handcrafted features and deep learning features are available for the representation of the query image and images in the database. In the second step, the query

image is compared with each of the images in the database by calculating their feature similarity. Then, these similarity values are subsequently sorted in descending order to return the top k similar images. The kth image is determined as a correct query if it has the same category as the query image. Moreover, for one query, the higher ranks of the correctly retrieved images indicate the better performance of the SLRSIR method. The authors can be referred to Section IV for SLRSIR performance metrics in detail.

The literature has committed to SLRSIR over the past decade, and a number of advanced methods have been proposed, especially the methods driven by deep learning. Before the advent of deep learning in the RS community, SLRSIR mainly relies on handcrafted features, such as spectral features, texture features, and even combined features. Shao et al. [26] have proposed improved color texture descriptors for RSIR by taking color information into consideration. Bosilj et al. [27] have presented pattern spectra descriptors, which are computationally efficient histogram-like structures and described the global distribution of arbitrarily defined attributes of connected image components. Aptoula [28] exploited global morphological texture descriptors for RSIR, which outperforms the best-known retrieval scores, despite its shorter feature vector length. Chen et al. [29] have proposed a radar RSIR algorithm to solve the time-consuming problem. Kavitha and Vidhya Saraswathi [30] have proposed a fuzzy multicharacteristic clustering technique to provide retrieval outcomes with elevated retrieval accuracy. Sunitha and Sivarani [31] have proposed an efficient RSIR system utilizing weighted Brownian motion-based monarch butterfly optimizations to improve the retrieval accuracy along with computational intricacy. Ben-Ahmed et al. [32] have focused on the most relevant channels and studied spectral sensitivity functions in constructing discriminative representations for hyperspectral image retrieval. Tekeste and Demir [33] have introduced local binary patterns (LBP) variants for the first time in the framework of RSIR problems and presented a comparative study to analyze and compare advanced LBP variants for RSIR. Zhang et al. [34] have proposed a hyperspectral RSIR system based on spectral and texture features. Du et al. [35] have considered the topological structure of local features and proposed a new method to represent by taking the structural information of local features into consideration. Sukhia et al. [36] have utilized local ternary pattern to obtain upper and lower texture images and divided them into dense patches to build a final histogram representation. Yang et al. [37] have proposed a simple method to improve recognition performance of the typical Bag of Words (BoW) framework by representing images with local features extracted from base images in a large-scale image database. To further improve performance of RSIR, the combined features are also explored and exploited. For instance, Chaudhuri et al. [38] have introduced an unsupervised graph-theoretic approach for region-based retrieval by using intensity, texture, and shape features extracted from the regions to describe the node attributes of the graph. Ye et al. [39] have proposed an RSIR method based on the query-adaptive feature weights to fuse features and utilized two image similarities to improve retrieval performance. The rest of the works have been focused on RSIR in compressed image archive [40], image reranking [41], feature hashing [42], and relevance feedback [43].

Although handcrafted features have been demonstrated their capacity for RSIR, it is difficult to further improve retrieval performance due to the limitations of low-level features. The popularity and success of deep learning, particularly convolutional neural network (CNN), in RS community [16], [17], [18], [19], [20] have promoted the literature to develop a great number of deep learning driven methods for RSIR. The existing RSIR methods can be coarsely divided into five groups, including feature extraction-based methods, novel network-based methods, attention-based methods, metric learning-based methods, and hashing-based methods.

1) Feature Extraction-Based Methods: Feature extraction-based RSIR is to treat the pretrained deep networks as feature extractors or fine-tuning these networks to extract image features. To ensure high efficiency and accuracy, Cheng et al. [44] have proposed a distributed system architecture for high-resolution satellite image retrieval by combining deep and traditional handcrafted features. Ye et al. [45] have exploited a CNN regression model to develop a query-adaptive feature fusion method to alleviate the huge variation in retrieval performance among different image queries. In another work, they calculated the fuzzy class membership of images to reduce the overall search time [46]. Fan et al. [47] have used CNN to extract the effective coverage information of images and presented an automatic accurate high-resolution RSIR method. Vharkate and Musande [48] have proposed a hybrid visual geometry group network by integrating dimensionality reduction, feature extraction, loss function optimization, matching process, and relevance feedback for the appropriate retrieval of RS images. Zhuo and Zhou [49] have focused on feature dimension reduction and extracted low-dimensional, representative, and discriminative features from fully connected layers of CNN by using an extended method. Similarly, Hou et al. [50] have extracted low-dimensional features from the fully connected layers by fine-tuning the pretrained MobileNets, Sadeghi-Tehran et al. [51] have derived feature representations via a CNN feature extractor, and Ye et al. [52] have fine-tuned the pretrained CNNs to extract features.

Most of the aforementioned works regard CNNs as a feature extractor and extract features from fully connected layers. Actually, CNNs are also capable of extracting deep local features from convolutional layers, and the process is similar to that of scale-invariant feature transform [53]. Ge et al. [54] have aggregated the outputs of midlevel layers by means of average pooling with different pooling regions to extract CNN features for high-resolution RSIR. Imbriaco et al. [55] have presented a pipeline that used attentive, local convolutional features and aggregated them using the vector of locally aggregated descriptors (VLAD) to produce a global descriptor. Tang et al. [56] have conducted a similar work. Specifically, they have proposed an unsupervised deep learning method using deep convolutional autoencoder. The learned features were aggregated using BoW framework to obtain the final feature vector. Hu et al. [57] have provided a comparative study on deep representations extracted from either full-connected or convolutional layers. Napoletano [58]

has compared a few handcrafted features with CNN features on the two benchmark datasets. The obtained results indicate that CNN features achieve overall better performance. The rest of related works can be found in [59], [60], [61], and [62].

Treating the off-the-shelf networks as feature extractors often performs well on small-scale target datasets and particularly those similar to the source dataset on which the networks are pretrained. However, fine-tuning the off-the-shelf networks is often used to further improve performance, especially for datasets that have limited labeled images.

2) Novel Network-Based Methods: The novel network-based methods focus on designing new CNN architectures trained from scratch for learning powerful features. Zhou et al. [63] have compared the performance of various CNN features and proposed a low-dimensional CNN (LDCNN) for high-resolution RSIR, which outperforms the fine-tuned CNNs. Boualleg et al. have combined LDCNN model [63] with the triplet loss and proposed a triplet LDCNN [64]. Zhang et al. [65] constructed a triplet nonlocal neural network with dual-anchor triplet loss for high-resolution RSIR. Zhuo and Zhou [66] proposed an RSIR method for high-resolution RS images with Gabor-CA-ResNet and split-based deep feature transform network. Wu et al. [67] have developed and investigated two new rotation-aware CNN-based RSIR methods to learn rotation-aware representation. Liu et al. [68] have introduced an easy way to organize semantic relationship among classes as a category tree and proposed a tree-triplet-classification network. Wang et al. [69] have proposed a learnable joint spatial and spectral transformation model composed of parameter generation network, spatial conversion module, and spectral conversion module for RSIR. Sumbul and Demir [70] have proposed a novel plasticity–stability preserving multitask learning approach to ensure the plasticity and the stability conditions of the whole learning procedure independently of the number and type of tasks.

The works surveyed above focus on the conventional CNN architectures that take images as input. There have been other works conducted on image graphs. Wang et al. [71] have developed a graph-based learning method for effectively retrieving RS images. The method utilized a three-layer framework that integrates the strengths of query expansion and fusion of holistic and local features. Chaudhuri et al. [72] have argued the effectiveness of region adjacency graph-based image representations for very high resolution (VHR) RS images in terms of localized region and proposed a Siamese CNN architecture for assessing the similarity between a pair of graphs. Compared with image, graph is capable of capturing contextual information and, thus, is possible to improve RSIR performance.

Training from scratch with novel architectures tends to achieve more remarkable performance compared with feature extraction-based methods that use pretrained or fine-tuned networks for feature extraction. However, a large number of labeled samples are often required to train a successful CNN. Moreover, it is laborious to design a powerful CNN architecture even if the pretrained CNNs are taken as the backbones in the new architecture.

3) Attention-Based Methods: Generally, the attention-based methods are networks integrating attention modules in the architecture to learn more discriminative features. There have been several representative works related to this topic [73], [74], [75], [76], [77]. Wang et al. [73] have proposed a multiattention fusion network with dilated convolution and label smoothing to force the network to learn discriminative features of important objects. Wang et al. [74] have presented a wide-context attention network by leveraging two attention modules to adaptively learn local features correlated in the spatial and channel dimensions. Wang et al. [75] have introduced a second-order pooling named compact bilinear pooling into CNN containing three stages, i.e., pretraining, fine-tuning, and retrieval. Xiong et al. [76] have proposed two effective schemes for generating discriminative features for RSIR, where in the first scheme, the attention mechanism and a new attention module were introduced to the CNN architecture. In the second scheme, a multitask learning network structure was proposed to force the features to be more discriminative. Unlike the above works that focus on the spatial or channel attention to learn discriminative features, other works focus on edge and node attention to highlight important image context features by using image graphs as input. For example, Chaudhuri et al. [77] have proposed an attention-driven graph CNN for RSIR by attending over the edge matrix to highlight the interactions among meaningful regions and exploiting this edge attention mechanism together with node attention to highlight essential image context.

Integrating attention module in CNN network provides the literature a new manner to extract more discriminative image features. However, the existing attention-based methods mainly focus on extracting powerful features that ignore similarity measure, another indispensable part in an RSIR system. Therefore, there is still room for performance improvement when more sophisticated attention modules that take both feature extraction and similarity measure into account.

4) Metric Learning-Based Methods: Metric learning learns a distance metric for the input space of data from a given collection of pair of points [78] and is able to combine with loss functions, such as contrastive loss [79], to improve classification performance [80]. Considering similarity measure is an indispensable part for RSIR system; thus, there have been a great number of metrics learning-based methods developed for RSIR, which can be coarsely categorized into two groups. The first is integrating metric learning in a CNN network to improve performance. Zhao et al. [81] have proposed a global-aware ranking deep metric learning with intraclass space sample mining and cost-sensitive loss. Cao et al. [82] have developed a deep metric learning approach with generative adversarial network (GAN) regularization, aiming to obtain more accurate retrieval performance with small training samples. Cao et al. [83] have constructed a triplet network with metric learning to extract representative features in a semantic space where images from the same class are close to each other, while those from different classes are far apart to enhance RSIR. The second is combining metric learning and attention mechanism to learn discriminative features and, thus, achieving better performance. Cheng et al. [84] have proposed

an ensemble architecture of residual attention-based deep metric learning for RSIR to improve feature distinguishability and retrieval efficiency. Chung et al. [85] have introduced a method for retrieving aerial images by merging group convolution with attention mechanism and metric learning, resulting in robustness to rotational variations. Previous research on RSIR has ignored the advantages of joint optimization of RSIR and scene classification. To overcome this limitation, Liu et al. [86] have presented an eagle-eyed multitask CNN integrating three tasks, i.e., center-metric learning, similarity distribution learning, and aerial scene classification in a network. The extensive experiments over four public aerial image sets demonstrate its better performance than all of the existing methods. Other existing metric learning-based methods have been focused on defining novel losses [87], [88], [89], similarity learning [90], and reranking [91] to improve RSIR performance.

RS images usually have different resolutions, different object types, and different image complexities. Therefore, the data-driven metric learning is suitable for deep learning-based RSIR methods since both the features and similarity measure are learned from data. Moreover, RSIR performance is possible to be further improved when the network is trained in an end-to-end manner.

5) Hashing-Based Methods: For RSIR with large-scale archives, the storage cost and the retrieval efficiency are two factors to be considered. Hashing- based methods aim to perform RSIR with short binary codes, which have low storage cost and high retrieval efficiency. The existing hashing-based methods can be divided into unsupervised hashing and supervised hashing [25].

The unsupervised hashing methods rely on unlabeled data to generate binary hash codes. There have been few unsupervised hashing methods in RS community. Chen and Lu [92] have proposed an unsupervised multispectral RSIR method, making use of the unsupervised representation learning ability of GAN. Reato et al. [93] have presented a simple yet effective unsupervised RSIR method that represented each image with primitive-cluster-sensitive multihash codes. Lukac et al. [94] have improved the well-known kernelized locality-sensitive hashing method using graphical processing units to make it feasible for parallelization, and thus performing fast parallel image retrieval. Kong et al. [95] have proposed a low-rank hypergraph hashing to accomplish for the large-scale RSIR. To improve the performance of unsupervised hashing methods, self-supervised methods, semi-supervised methods, and methods relying on pseudolabel have been explored. As an example, Tan et al. [96] have proposed a deep contrastive self-supervised hashing, which uses unlabeled images to learn accurate hash codes. Tang et al. [97], [98] have proposed a semi-supervised deep hashing method based on the adversarial autoencoder network for RSIR. Sun et al. [99] have proposed a soft-pseudolabel-based unsupervised deep hashing method to well reflect the semantic distance between intercluster images.

Although unsupervised hashing methods are simple and effective for generating binary codes for large-scale RSIR, their performance improvement is limited due to the lack of supervised information. In contrast, supervised hashing methods often achieves better performance than unsupervised hashing methods. To address the problem that deep hashing networks tends to be highly expensive in terms of storage space and computing resources, Li et al. [100] have developed a quantized deep learning to hash framework for large-scale RSIR. Song et al. [101] have proposed an asymmetric hash code learning for RSIR, attempting to improve the conventional learning one hash function for both the query and database samples in a symmetric way. Motivated by the residual net, Han et al. [102] have developed a cohesion intensive deep hashing model for RSIR. Liu et al. [103] have presented a deep supervised hashing model for RSIR in the framework of GANs, named GAN-assist hashing. Li et al. [104] proposed a large-scale RSIR method based on deep hashing neural network. Ye et al. [105] have investigated multiple feature hashing learning for large-scale RSIR. Tang et al. [106] have developed a new supervised hash learning method for the large-scale high-resolution RSIR task based on metalearning. Wang et al. [107] have proposed a novel triplet ordinal cross-entropy hashing method to fix the problem that most of the existing hashing algorithms only emphasized preserving pointwise or pairwise similarity. Shan et al. [108] have presented a proxy-based hash retrieval method, called deep hashing using proxy loss, which combines hash code learning with proxy-based metric learning in a CNN. Liu et al. proposed a new RSIR method named feature and hash learning, which consists of a deep feature learning model and an adversarial hash learning model [109]. For most existing hashing methods, the hash functions are learned once for all and kept fixed all the time, which are not suitable for the ever-growing new RS images. Li et al. [110], therefore, proposed a new online hashing method, learning and adapting hashing functions with respect to the newly incoming RS images.

The above supervised hashing methods provide new ways of performing large-scale RSIR. However, a large number of labeled images are needed to train a successful network. Besides, it is also time-consuming to generate binary hash codes.

B. MLRSIR Methods

For SLRSIR, both the query images and other images in the database are single labeled. The assumption is that each image is annotated by a single label representing the most significant semantic content of the image. However, in practical scenarios, RS images might contain multiple classes (e.g., buildings, roads, trees, etc.). MLRSIR, a more challenging task than SLRSIR, assumes each image is associated with multiple labels (also known as primitive class), and thus is suitable for addressing the above problem. As illustrated in Fig. 4, the process of MLRSIR is similar to that of SLRSIR but is a bit different in feature extraction and performance evaluation. To be specific, for extraction, we need to extract the features of each primitive classes contained in the image. With respect to performance evaluation, there have been no ground truth images for each multilabel query images. The retrieved images are ranked according to the similarity scores between query image and images in the database. Thus, the metrics for SLRSIR are not available for evaluating MLRSIR methods.

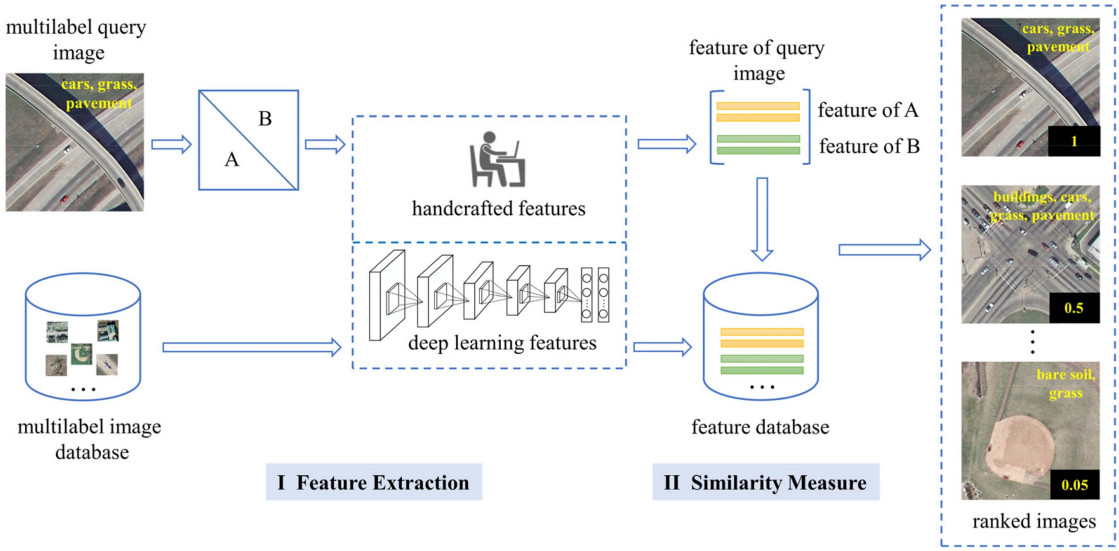


Fig. 4. Flowchart of MLRSIR method. The numbers in the ranked images indicate the similarity values between the query image and images in the database.

The literature has committed efforts to developing MLRSIR methods, such as handcrafted feature-based methods [111], [112], [113], [114] and deep learning-based ones [115], [116], [117], [118], [119], [120], [121]. For handcrafted methods, Chaudhuri et al. [111] have introduced a semi-supervised graph-theoretic method in the framework of MLRSIR problems. Dai et al. have presented a novel hyperspectral RSIR system consisting of a spatial and spectral image description scheme and a sparsity-based supervised retrieval method [112], which was improved in their later work [113]. Shao et al. [114] have conducted a comparative performance evaluation of SLRSIR and MLRSIR methods on a newly collected multilabeled dataset termed DLRSD, providing the literature a benchmark dataset along with the baseline results for MLRSIR research.

Regarding deep learning-based methods, Kang et al. [115] have proposed a new graph relation network to model the relations between samples by using a graph structure for multilabel RS scene categorization and retrieval. Hua et al. [122] took advantage of pairwise label relations to infer multiple object labels of a high-resolution aerial image and proposed an innovative inference network. Sumbul and Demir [117] have proposed a novel graph-theoretic deep representation learning method in the framework of MLRSIR problems, aiming to extract and exploit multilabel co-occurrence relationships associated to each RS image. In their another work [116], a novel triplet sampling method in the framework of deep neural networks defined for MLRSIR was proposed to obtain informative and representative triplet selection, which is an improved version of the previous work [114]. To increase retrieval efficiency and reduce feature storage while preserving semantic information, Cheng et al. [119] have presented a new semantic-preserving deep hashing model for MLRSIR. Shao et al. [121] have proposed a novel MLRSIR approach based on the fully convolutional network, where the single-scale and multiscale region features were extracted to perform region-based MLRSIR. Although these methods are able to achieve satisfactory performance, they mainly focus on extracting powerful features. The similarity measure as well as evaluation metric for MLRSIR are not considered. As an alternative, Imbriaco et al. [120] have defined protocols for performance evaluation using new metrics and studied the impact of commonly used losses as well as reranking methods for MLRSIR. It provides a direction for similarity measure for multilabel images.

In contrast to SLRSIR, MLRSIR is still a new topic, and most of the existing works focus on feature extraction. However, as stated above, there are no ground truth images for each query image. Therefore, more attention should be drawn to define protocols for similarity measures and performance evaluations.

C. RASRSIR Methods

RASRSIR performs RSIR between images captured by different sensors and, thus, having different resolutions (e.g., multispectral and panchromatic images, and multispectral and hyperspectral images). The special case is retrieval between optical and SAR images, which is generally categorized into CMRSIR, as illustrated in Table I. Given a multispectral (MUL) image as the query image and panchromatic images as the database images, Fig. 5 illustrates the flowchart of RASRSIR. The process is also similar to that of SLRSIR but different in feature extraction. Specifically, to avoid the difference caused by image resolution, the features of MUL and panchromatic (PAN) images are fed into the same feature space before conducting similarity measure. There is a continually increasing interest to develop RASRSIR methods with RS images from different sensors. Li et al. [123] have proposed a source-invariant deep hashing CNN for RASRSIR between MUL and PAN images, which were optimized in an end-to-end manner using a series of well-designed optimization constraints. To maintain the source discrepancy at the classifier level, Ma et al. [124] have presented teacher-ensemble learning with the knowledge distillation method. In [125], a discriminative distillation network was also proposed to address

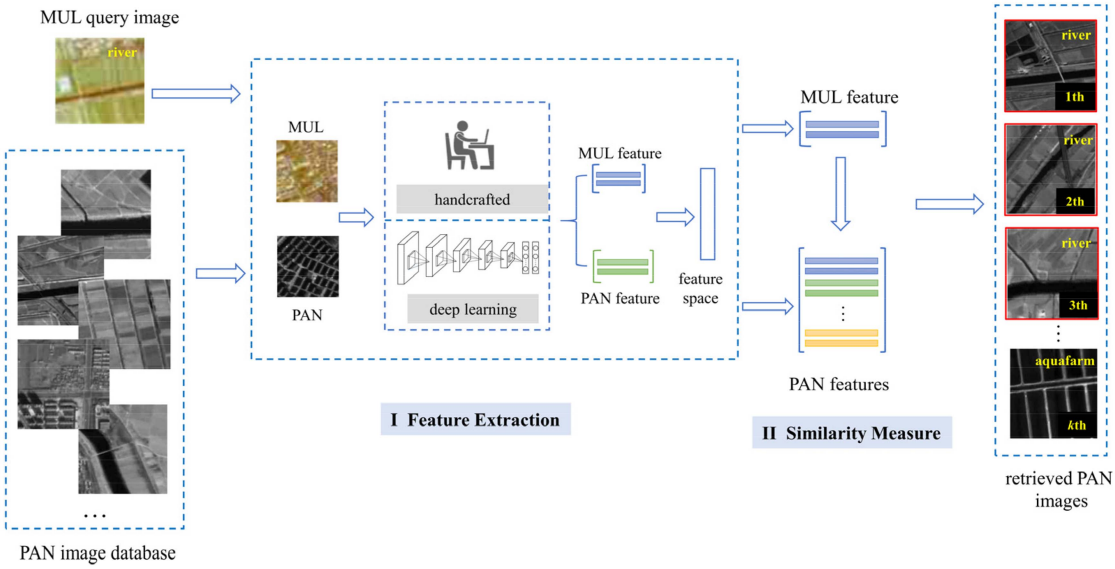


Fig. 5. Flowchart of RASRSIR method. The retrieved images with red rectangles stand for images that are correctly retrieved.

the inconsistency between different image sources. Gao et al. [126] have proposed a dubbed multiview graph convolutional hashing method to fuse multisource RS images. Xiong et al. [127] have explored to explicitly address a data drift problem by mapping the source domain to the target domain in an image translation-based framework. Ma et al. [128] have developed a dual-modality collaborative learning model to fully explore the specific information from diverse RS images.

RASRSIR is a promising direction for RSIR, and the core problem is to alleviate the effect of inconsistency between different image sources.

D. CMRSIR Methods

CMRSIR is similar to RASRSIR in terms of the retrieval process, and the single difference lies in data modality. For RASRSIR, the query image and the images in database have the same modality, which is not the case for CMRSIR. The literature focuses attention on four modality groups: sketch–image [129], [130], [131], optical–SAR [132], [133], [134], text–image[135], [136], [137], [138], [139], [140], and audio–image [141], [142], [143], [144], [145], [146], [147], [148].

The existing CMRSIR methods for sketch–image and optical–SAR are fewer than that of text–image and audio–image. Jiang et al. [129] have taken free-hand sketches into account and addressed the problem of sketch-based aerial image retrieval. Chaudhuri et al. [130] have exploited the data modality comprising more spatial information (sketch) to extract other modality features (image) with cross-attention networks. Xu et al. [131] have developed a sketch-based RSIR model to learn a deep joint embedding space with discriminative losses, which was then evaluated on a sketch RS image dataset. Regarding optical–SAR CMRSIR, Xiong et al. [132] have addressed the prominent modality discrepancy caused by different imaging mechanisms in a deep cross-modality hashing network. To effectively deal with the discrepancies, Sun et al. [133] have conducted a similar work. They have proposed a multisensor fusion and explicit semantic-preserving-based deep hashing method. Sumbul et al. [134] have collected the multimodal BigEarthNet (BigEarthNet-MM) benchmark archive containing pairs of Sentinel-1 and Sentinel-2 image patches.

For text–image and audio–image CMRSIR, there have been a great number of methods developed in recent years. Regarding text–image CMRSIR, to bridge the modality gap, Lv et al. [135] have proposed a fusion-based correlation learning model for text–image retrieval. Yuan et al. [136] have presented a text–image retrieval framework based on global and local information and designed a multilevel information dynamic fusion module to efficiently integrate features of different levels. In their other works, an asymmetric multimodal feature matching (FM) network was developed to explore a fine-grained multiscale method for text–image retrieval in [137]. Besides, a concise but effective cross-modal retrieval model was designed by considering the characteristics of multiscale and target redundancy in RS [138]. Cheng et al. [139] have established the direct relationship between RS images and paired text data. To address the limitation that the existing approaches require a high number of labeled training samples, Mikriukov et al. [140] have proposed an unsupervised cross-modal contrastive hashing method for text–image retrieval. With respect to audio–image CMRSIR, there have been a large number of methods developed in recent years. Chen et al. [141], [142], [143] have proposed a few audio–image retrieval methods. Guo et al. have proposed a CMRSIR method for RS image and spoken audio [144], which was improved in a later work [145]. Existing methods for RS image–voice retrieval rely primarily on the pairwise relationship. To overcome this limitation, Ning et al. [146] have proposed a semantics-consistent representation learning method for image–voice retrieval. Yang et al. [147] have proposed a cross-modal feature fusion retrieval model, which provides a more optimized cross-modal common

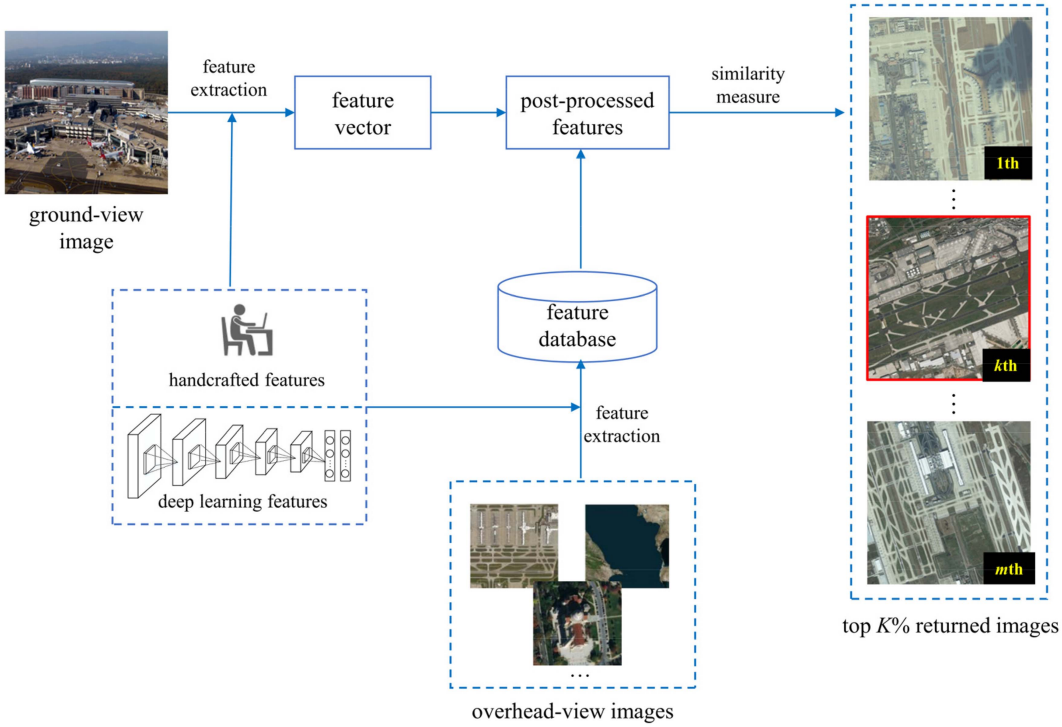


Fig. 6. Flowchart of CVRSIR method. The retrieved images with red rectangles stand for images that are correctly retrieved.

feature space than the previous models and, thus, optimizes the retrieval performance. In a later work [148], they have presented the multifusion method.

As an emerging research topic, a great number of CMRSIR methods have been developed to perform sketch–image, optical–SAR, text–image, and audio–image retrieval. The performance of CMRSIR methods would be further improved as long as the effect of modality difference is well alleviated.

E. CVRSIR Methods

CVRSIR is known as image geolocalization [149] in the literature, aiming to determine the geographic information of an image (e.g., ground-view image) by referencing to a geotagged image of another view (e.g., overhead-view image). Therefore, image geolocalization is essentially one kind of RSIR applications. Fig. 6 illustrates the basic flowchart of CVRSIR between ground-view image and overhead-view images. It is notable that CVRSIR is similar to SLRSIR in terms of the retrieval process, and the differences lie in which the query image and other images in the database are captured from different views, and for each query image, there is generally one ground truth retrieved image.

The existing CVRSIR methods consist of ground–ground, ground–overhead, and overhead–overhead, as listed in Table I. Most CVRSIR methods are overhead–overhead since they are able to obtain an overhead image with a random location on the earth. Zhang et al. [150] have proposed a deep network that embeds spatial configuration of the scenes into feature representation. Zemene et al. [151] have used image matching in a structured database of city-wide reference images with known GPS coordinates. Inspired by the human visual system, Lin et al. [152] have proposed a framework to jointly learn the discriminative representation and detect salient key points with a single network. Rodrigues and Tani [153] have retrieved corresponding aerial views from a large database of geotagged aerial imagery. Hu et al. [154] have leveraged on the recent success of deep learning and proposed a cross-view matching network for the ground-to-aerial geolocalization task. Ground–overhead geolocalization is the most challenging geolocalization task due to the large variation of viewpoint and irrelevant content. To address this issue, Zeng et al. [155] have taken drone-view information as a bridge between ground-view and satellite-view domains, and proposed a peer learning and cross diffusion framework. The rest of the works can be found in [156], [157], [158], [159], [160], [161], and [162].

The aforementioned CVRSIR works mainly focus on ground–satellite and ground–aerial geolocalization; we here introduce a novel similarity learning based on CVRSIR (SL-CVRSIR) method for ground–drone geolocalization. To evaluate the performance of SL-CVRSIR, we also collect a new dataset named CVGD, which will be opened later for uncommercial purposes.

Fig. 7 illustrates the architecture of SL-CVRSIR, which has two identical subnetworks (i.e., CNN1 and CNN2) without shared weights, and are designed to extract the features of overhead-view images and ground-view images, respectively. SL-CVRSIR takes the positive and negative image pairs as input, where positive pairs are composed of images from the same location, while negative pairs are composed of images from different locations. The output of the fully connected layer (i.e., F_o and F_g) from each subnetwork is then combined through subtraction, and the result is passed through a fully connected (Fc) layer with a single output. The sigmoid operation is used

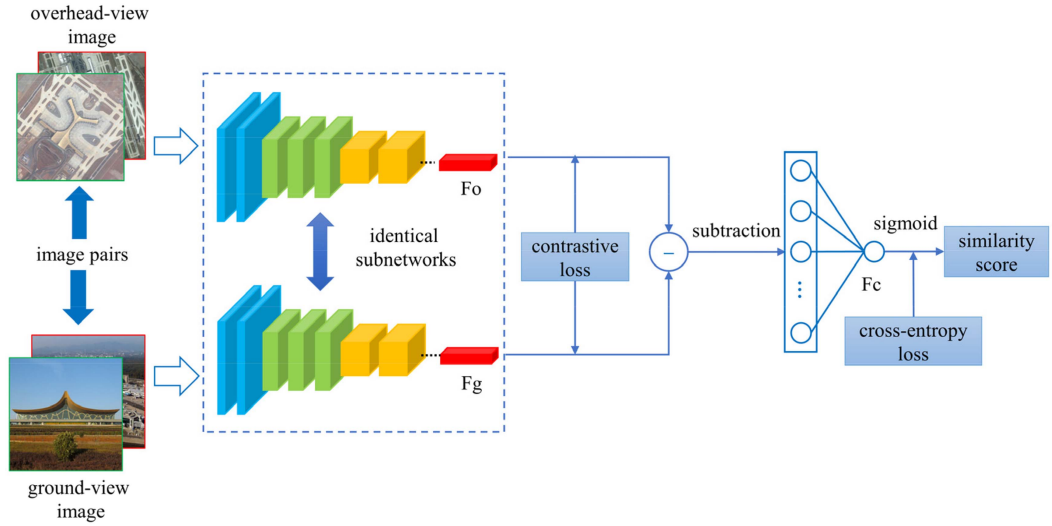


Fig. 7. Flowchart of the proposed CVRSIR method based on similarity learning.

to convert the output of Fc layer to a similarity score between 0 and 1, indicating whether the overhead-view and ground-view images in the image pair are from the same location or not.

To train SL-CVRSIR, the combined loss L is defined as follows:

$$L = L_c + L_{ce} \tag{1}$$

$$L_c = \frac{1}{2}(1-y)\, d_{og}^2 + \frac{1}{2} y \{\max(0, m - d_{og})\}^2 \tag{2}$$

$$L_{ce} = -q \log(p) - (1-q) \log(1-p) \tag{3}$$

where L_c is the contrastive loss [79], aiming to compare the similarity between overhead-view and ground-view images, L_{ce} is the cross-entropy loss, aiming to measure the difference between the label and similarity score. y and q are the labels of image pair with 1 for positive pair and 0 for negative pair, and p is the similarity score. d_{og} is the Euclidean distance between overhead-view image and ground-view image in the image pair, and m is the margin.

Once SL-CVRSIR is trained, the similarity score of each image pair is extracted to perform cross-view retrieval. Specifically, given one ground-view image (i.e., query image), the overhead-view images are sorted in descending order by the similarity values between the query image and each of the overhead-view images. The query is regarded as a correct query if the overhead-view image from the same location as the ground-view image is within the top *K*% retrieved images.

III. Benchmark Datasets for Performance Evaluation of RSIR

Benchmark datasets are indispensable for advancing RSIR approaches and further performance evaluation. As the development of RS technology, the literature has witnessed the remarkable progress on constructing publicly available datasets for RSIR. These datasets are collected for developing different RSIR methods and, thus, can be divided into different categories, including SLRSIR datasets, MLRSIR datasets, RASRSIR datasets, CMRSIR datasets, and CVRSIR datasets. We, therefore, survey the publicly available benchmark datasets presented for different RSIR methods in recent years, as shown in Table III. In the table, we list the basic characteristics of each dataset for simple comparison. The readers are referred to corresponding datasets for more details. It is worth noting that not all of these datasets are originally collected for RSIR. In the following section, we select and detail several representative datasets for each RSIR method category, i.e., SLRSIR, MLRSIR, RASRSIR, CMRSIR, and CVRSIR. These representative benchmark datasets will be further used for performance evaluation in Section IV.

A. SLRSIR Datasets

In the early years, SLRSIR datasets are commonly used for RSIR because most of the RSIR works focus on SLRSIR at that time. The accessible SLRSIR datasets in the literature include UC Merced [163], WHU-RS19 [164], RSSCN7 [165], AID [166], PatternNet [167], RSI-CB [168], SIRI-WHU [169], and NWPU-45 [170]. Among these datasets, PatternNet is originally collected for RSIR while the other datasets are originally collected for scene classification.

We select three representative datasets, i.e., UC Merced, WHU-RS19, and PatternNet, and introduce them in detail.

1) UC Merced: The UC Merced [163] dataset is originally collected for land use/land cover with 21 categories, including agricultural, airplane, baseball diamond, beach, building, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tank, and tennis court. Each category contains 100 images cropped from the United States Geological Survey (USGS) aerial images. Each image in the UC Merced dataset has the size of 256×256 pixels and the spatial resolution of 0.3 m. As the first

TABLE III
PUBLIC AVAILABLE BENCHMARK DATASETS FOR RSIR

	Dataset	Number of Classes (#)	Number of Images (#)	Image Size (pixels)	Resolution (m)	Image Source	Year
SLRSIR	UC Merced [163]	21	2100	256×256	0.3	USGS	2010
	WHU-RS19 [164]	19	1005	600×600	up to 0.5	Google Earth	2012
	RSSCN7 [165]	7	2800	400×400	N/A	Google Earth	2015
	AID [166]	30	10 000	600×600	8~0.5	Google Earth	2017
	PatternNet [167]	38	30 400	256×256	4.69~0.06	Google Earth & Google Map	2018
	RSI-CB [168]	35	24 000	256×256	3~0.22	Google Earth & Bing Map	2020
	SIRI-WHU [169]	12	2400	200×200	2	Google Earth	2016
	NWPU-45 [170]	45	31 500	256×256	about 30~0.2	Google Earth	2017
	Dataset	**Number of labels**	**Number of Images**	**Image Size**	**Resolution (m)**	**Image Source**	**Year**
MLRSIR	DLRSD [114]	17	2100	256×256	0.3	UC Merced	2018
	WHDLD [121]	6	4940	256×256	2	Gaofen-1 and Ziyuan-3	2020
	MLRSNet [171]	60	109 161	256×256	about 10–0.1	Google Earth	2020
	ML-AID [172]	17	3000	600×600	8–0.5	AID	2020
	MultiScene [173]	36	100 000	512×512	0.6–0.3	Google Earth	2022
	BigEarthNet [174]	43	590 326	20×20/60×60/ 120×120	60/20/10	Sentinel-2	2019
	Dataset	**Number of Classes**	**Number of pairs**	**Image Size**	**Resolution (m)**	**Image Source**	**Year**
RASRSIR	DSRSID [123]	8	80,000	256×256/ 64×64	2/8	Gaofen-1 panchromatic/multispectral	2018
	Dataset	**Data Modality**	**Description**				**Year**
CMRSIR	SODMRSID [132]	optical and SAR images	SODMRSID consists of 12 classes with 1000 SAR–optical image pairs of each class. The SAR and optical images are from Sentinel-1 and 2 with the resolution of 10m and the image size 256×256.				2020
	BigEarthNet-MM [134]	optical and SAR images	BigEarthNet-MM consists of 590 326 pairs of Sentinel-1 and 2 image patches. Each image patch is annotated by at least one of 19 multilabels.				2021
	RSketch [131]	image and sketch	RSKetch contains 20 categories, and each category has 200 RS images with the size of 256×256 and 45 sketches.				2020
	UCM-/Syndey-Captions [175]	text and image	UCM-Captions have 21 categories containing 2100 images with the size of 256×256 and 10500 captions, and Syndey-Captions have 7 categories containing 613 images with the size of 500×500 and 3065 captions.				2016
	RSICD [176]	text and image	RSICD has 30 categories consisting of 10 921 images with the size of 224×224. Each image is annotated with five sentences.				2018
	UCM-/Syndey-/RSICD-audio [144]	image and audio	UCM-/Syndey-/RSICD-audio is based on UCM-/Syndey-Captions and RSICD, respectively. Each image is annotated with five spoken audios.				2018
	TextRS [177]	text and image	TextRS is composed of 2144 images randomly selected from AID, UC Merced, PatternNet, and NWPU-RESISC45, and each image is annotated with five sentences.				2020
	RSITMD [137]	text and image	The images are collected from RSICD and Google Earth, and there have a total of 23 715 captions for 4773 images.				2022
	CBRSIR_VS [133]	optical and SAR images	CBRSIR_VS contains 10 class labels and 26 901 pairs of optical (with the size of 256×256 and the resolution of 1 m) and SAR images (with the size of 64×64 and the resolution of 10 m).				2022
	Dataset	**Data View**	**Description**				**Year**
CVRSIR	University-1652 [178]	ground and drone and satellite views	University-1652 training set contains images from ground, drone, and satellite views, with 71.64 images (54 drone, 16.64 ground, and 1 satellite images) per location on average.				2020
	CVACT [179]	ground and satellite views	CVACT consists of 35 532 training pairs and 8884 validation pairs. Besides, it also contains 92 802 testing pairs.				2019
	AiRound/CvBrCT [180]	ground and aerial views	AiRound is composed of 11 753 images distributed in 11 classes, and CvBrCT comprises about 24K pairs of images categorized in 9 classes.				2021
	CVUSA subset [181]	ground and aerial views	CVUSA subset is from the original CVUSA, containing 35 532 training and 8884 testing pairs, respectively.				2017
	VIGOR [182]	ground and aerial views	VIGOR consists of 90 618 aerial images and 238 696 panoramas, respectively.				2021
	Vo and Hays [183]	ground and aerial views	The dataset contains more than 1 million image pairs collected from 11 U.S. cities (eight cities for training, and the rest three cities for testing).				2016
	Two cities dataset [184]	ground and aerial views	The aerial images are collected from two cities and aligned with OSM road maps.				2016
	CVGD	ground and drone views	The images in CVGD are collected from 100 locations in the university NUIST. There are 2–6 drone images and 2–7 ground images per location.				2022

publicly available high-resolution RS evaluation dataset, it has been regarded as a benchmark to develop novel RSIR methods.

2) WHU-RS19: The WHU-RS19 [164] dataset is collected from the google earth and then categorize into 19 classes, including airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking, pond, port, railway station, residential area, river, and viaduct. WHU-RS19 contains 1005 images with the size of 600×600 pixels and the spatial resolution is up to 0.5 m. It is a more challenging dataset for RSIR, compared with the UC Merced dataset in terms of its varied spatial resolutions.

3) PatternNet: The PatternNet [167] dataset is a large-scale dataset collected from U.S. cities via google earth and google map API, developing RSIR methods and particularly deep learning based ones. It contains a total of 30 400 images evenly divided into 38 categories, including airplane, baseball field, basketball court, beach, bridge, cemetery, chaparral, Christmas tree farm, closed road, coastal mansion, crosswalk, dense residential, ferry terminal, football field, forest, freeway, golf course, harbor, intersection, mobile home park, nursing home, oil gas field, oil well, overpass, parking lot, parking space, railway, river, runway, runway marking, shipping yard, solar panel, sparse residential, storage tank, swimming pool, tennis court, transformer station, and wastewater treatment plant. The images in the dataset measure 256×256 pixels with the spatial resolution ranges from 4.69 to 0.06 m. The release of PatternNet is to overcome the limitations that the existing datasets are small scale and their images contain a large amount of background and, thus, might distract accurate retrieval.

B. MLRSIR Datasets

It is time-consuming and laborious to construct a multilabel archive for RSIR since each image in the MLRSIR dataset is associated with at least one primitive class (i.e., label). Thanks to the literature's efforts, several publicly available datasets available for MLRSIR have been collected and opened over the past five years. The existing MLRSIR datasets include DLRSD [114], WHDLD [121], MLRSNet [171], ML-AID [172], MultiScene [173], and BigEarthNet [174].

Among these datasets, three representative archives, i.e., DLRSD, WHDLD, and MLRSNet, are selected and introduced in detail.

1) DLRSD: The DLRSD [114] dataset is labeled based on the UC Merced archive [163], and therefore, it also consists of 2100 images with the size of 256×256 pixels and the spatial resolution of 0.3 m. DLRSD is a dense labeling dataset, where the pixels of each image in the UC Merced dataset are annotated with one of the 17 primitive classes (labels), including airplane, bare soil, building, car, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tank, tree, and water. It is an improved version of the multilabel archive [111] and is available for not only image-level tasks, such as SLRSIR and MLRSIR, but also pixel-level task, such as semantic segmentation.

2) WHDLD: Similar to DLRSD, the WHDLD [121] dataset is also a pixel-level dense labeling dataset, where the images are cropped from a large mosaic image with the images acquired by Gaofen-1 and Ziyuan-3 satellites. The pixels of each image in WHDLD are annotated with the following six primitive classes (labels), including building, road, pavement, vegetation, bare soil, and water. WHDLD contains 4940 images with the size of 256×56 pixels and the spatial resolution of 2 m. WHDLD is different from DLRSD in terms of the number of images, image labels, and the spatial resolution of images. Therefore, it is treated as a complementary dataset to DLRSD for both image-level and pixel-level tasks.

3) MLRSNet: The MLRSNet [171] dataset is composed of 109 161 images with the size of 256×256 pixels, and the spatial resolution ranges from 10 to 0.1 m. The images are divided into 46 broad categories, and the number of images in each category varies from 1500 to 3000. In addition, there are 60 predefined primitive classes (labels), containing airplane, airport, bare soil, baseball diamond, basketball court, beach, bridge, buildings, cars, chaparral, cloud, containers, cross walk, dense residential area, desert, dock, factory, field, football field, forest, freeway, golf course, grass, greenhouse, gully, harbor, intersection, island, lake, mobile home, mountain, overpass, park, parking lot, parkway, pavement, railway, railway station, river, road, roundabout, runway, sand, sea, ships, snow, snowberg, sparse residential area, stadium, swimming pool, tanks, tennis court, terrace, track, trail, transmission tower, trees, water, wetland, and wind turbine. Each image in MLRSNet is associated with at least one of the 60 labels. In contrast to DLRSD and WHDLD, MLRSNet has the characteristics of hierarchy, large scale, and high diversity. However, as an image-level multilabel dataset, it cannot be used for pixel-level tasks.

C. RASRSIR Dataset

The existing public datasets for RASRSIR are not as many as that for SLRSIR and MLRSIR, and the DSRSID [123] dataset is one of the open datasets. The images in DSRSID are titled from panchromatic images and multispectral images acquired by the Gaofen-1 optical satellite and are grouped into image pairs where each pair is a combination of one panchromatic image and one multispectral image. The one-channel panchromatic image has the size of 256×256 pixels and the spatial resolution of 2 m, and the four-channel multispectral image has the size of 64×64 pixels with the spatial resolution of 8 m. Additionally, DSRSID consists of eight classes, i.e., aquafarm, cloud, forest, high building, low building, farm land, river, and water, where each class contains 10 000 panchromatic and multispectral image pairs.

D. CMRSIR Datasets

CMRSIR is to perform RSIR between different data modalities, such as optical–SAR images, audio—image, and text–image. There has been a large number of benchmark archives constructed for CMRSIR: SODMRSID [132], BigEarthNet-MM [134], RSketch [131], UCM/Syndey-Captions [175], RSICD [176], UCM/Syndey/RSICD-audio [144], TextRS [177], and CBRSIR_VS [133].

In the following section, we select four representative archives, i.e., RSketch, UCM/Syndey/RSICD-audio, TextRS, and CBRSIR_VS.

1) RSketch: The RSketch [131] dataset is collected for CMRSIR between RS image and sketch. It is composed of 20 categories, including airplane, baseball diamond, basketball court, beach, bridge, closed road, crosswalk, football field, golf course, intersection, oil gas field, overpass, railway, river, runway, runway marking, storage tank, swimming pool, tennis court, and wastewater treatment plant, and each category contains 200 RS images and 45 sketches. Both the size of RS image and sketch are fixed to 256×256 pixels. The RS images are collected from the existing datasets, such as UC Merced [163], WHU-RS19 [164], AID [166], and PatternNet [167].

2) UCM-/Syndey-/RSICD-Audio: The UCM-/Syndey-/RSICD-audio [144] is collected based on the existing UCM-/Syndey-Captions [175] and RSICD [176], respectively. To construct the dataset, each image is given five sentences, and each sentence of an image is generated by five different speakers. The spoken audios have varied length ranging from 1 to 15 s, which make the dataset challenging.

3) TextRS: The images in the TextRS [177] dataset are collected from the four existing datasets containing UC Merced [163], AID [166], PatternNet [167], and NWPU-45 [170]. TextRS contains 2144 images randomly selected from the four datasets, and each image is then annotated with five sentences generated by five different people to guarantee diversity.

4) CBRSIR_VS: The CBRSIR_VS [133] dataset is an optical and SAR dual-modality RS image dataset. It consists of ten class labels and 26 901 pairs of optical and SAR images. The optical images are VHR images with the size of 256×256 and the resolution of 1 m, and the SAR images are from Sentinel-1 imagery with the size of 64×64 and the resolution of 10 m.

E. CVRSIR Datasets

The images in a CVRSIR dataset are usually collected from ground–aerial view and ground–satellite view. The two view images captured from the same location are aligned to form image pairs. CVRSIR is a hot topic in recent years, and thus, the literature has constructed benchmarks of different viewpoints to advance CVRSIR research, including University-1652 [178], CVACT [179], AiRound/CvBrCT [180], CVUSA subset [181], VIGOR [182], Vo and Hays [183], and two cities dataset [184]. Besides, we also release a novel dataset termed cross-view between ground and drone (CVGD) to perform CVRSIR between ground-view and drone-view images. The readers are referred to the following section for more details on CVGD.

We select University-1652, CVACT, CVUSA subset, and the newly constructed CVGD as representative benchmarks and introduce them in detail.

1) University-1652: The University-1652 [178] dataset is a multiview multisource benchmark for drone-based geolocalization. It has ground-view, drone-view, and satellite-view images collected from 1652 buildings of 72 universities. The training set contains 701 buildings of 33 universities, and the testing set contains 701 building of the rest 39 universities. It should be noted that the training set has 71.64 images on average per location, while the existing datasets generally contain two images per location. University-1652 has the characteristics of multisource, multiview, and more images per class.

2) CVACT: The CVACT [179] dataset is a city-scale fully GPS-tagged cross-view dataset consisting of ground-view panoramas collected via google street view API and the corresponding satellite-view images. The image resolution of ground-view image is 1664×832 pixels and is 1200×1200 pixels for satellite image. Regarding the training set and testing set, there are 35 532 and 92 802 image pairs, respectively. Besides, it also provides a validation set with 8884 image pairs. Note that the training set is from the CVUSA subset [181].

3) CVUSA Subset: The CVUSA subset [181] dataset, a small version of the original CVUSA [158], is a much larger dataset. Specifically, the panoramas of CVUSA are selected to form a CVUSA subset as ground-view images, for each of which the aerial images at zoom level 19 are downloaded from bing map in the same geographic area. The panoramas with unavailable corresponding aerial images are filtered out, and finally 35 532 training image pairs and 8884 testing image pairs are obtained. The image resolution of the ground-view image and the satellite image are 1232×224 pixels and 750×750 pixels, respectively.

4) CVGD: Most existing datasets of CVRSIR are ground- and aerial-view or ground- and satellite-view. No ground- and drone-view images are included except for the University-1652 dataset. However, the drone images in University-1652 are actually simulated drone images collected from google earth. Moreover, all the ground–drone images in University-1652 are building images without any other objects, and thus, the literature needs a dataset with high diversity, i.e., images containing different objects. To this end, we collect a ground- and drone-view dataset, named CVGD. The images in CVGD are collected from 100 locations in a university. Considering the fact that, in a real CVRSIR task, there are possibly more than one ground and drone image pairs indicating the same location but captured different viewpoints, we collect 2–6 drone images and 2–7 ground images per location, as shown in Fig. 8.

Fig. 9 illustrates nine examples of ground and drone pairs, and it can be observed that the images contain different types of objects, such as building, tree, road, grass, and lake; thus, it is more challenging than University-1652.

IV. Performance Evaluation of RSIR Methods on Benchmark Datasets

In this section, we first introduce the performance metrics for different RSIR methods and then present the results of representative RSIR methods.

A. Performance Metrics for RSIR

Performance metrics are crucial for performance evaluation of RSIR methods. Considering the fact that different RSIR methods need their own measures, we categorize the existing measures into SLRSIR, MLRSIR, RASRSIR, CMRSIR, and CVRSIR, and then introduce them in detail.

1) Metrics for SLRSIR: There are several metrics that are commonly used for performance evaluation of SLRSIR, which are average normalized modified retrieval rank (ANMRR), mean

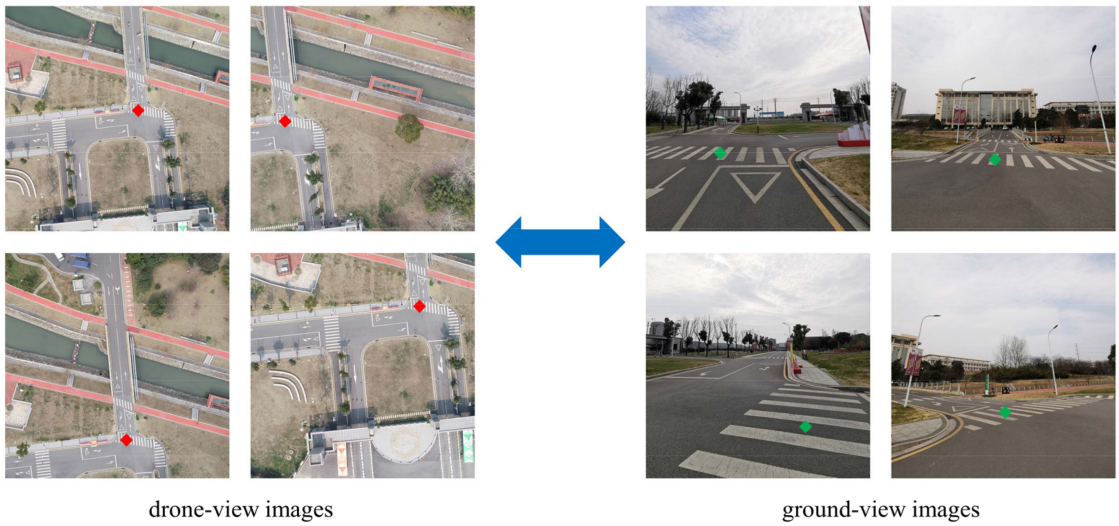


Fig. 8. Process of how the ground- and drone-view images are collected for each location. The red and green points indicate the same location on the ground.

Fig. 9. Example image pairs of nine locations in CVGD dataset.

average precision (MAP), precision at k (P@K), and precision–recall (PR) curve [58]. The definitions of these metrics are presented in the following text.

ANMRR is a metric that takes the rank of each returned image into account and is possibly the most widely used measure for SLRSIR. Let q be one query image, and $ng(q)$ is the number of its similar images. ANMRR is defined by

$$\text{ANMRR} = \frac{1}{nq}\sum_{q=1}^{nq}\text{NMRR}(q) \tag{4}$$

where nq is the query times, and $\text{NMRR}(q)$ is defined as follows:

$$\text{NMRR}(q) = \frac{ar(q) - 0.5\,[1 + ng(q)]}{1.25k(q) - 0.5\,[1 + ng(q)]} \tag{5}$$

where $ar(q) = \frac{1}{ng(q)}\sum_{k=1}^{ng(q)} r(k)$ is the average rank, and $r(k)$ is the retrieved rank of the kth image, which is defined as follows:

$$r(k) = \begin{cases} r(k), r(k) \le k(q) \\ 1.25k(q), r(k) > k(q) \end{cases} \tag{6}$$

where $k(q) = 2ng(q)$. ANMRR ranges between zero and one, and the lower the value, the better the performance.

To explain the other three metrics clearly, we here introduce precision and recall first. Precision is the ratio of the images retrieved that are similar to the query image, while recall is the ratio of the images that are similar to the query image that are successfully retrieved.

MAP is a commonly used performance metric and is defined as follows:

$$\text{MAP} = \frac{\sum_{q=1}^{nq} \text{AveP}(q)}{nq} \tag{7}$$

where AveP is the average precision. $P(k)$ is the precision at cutoff k, i.e., the metric *P@K*. rel(k) is the indicator function with the value of 1 if the image at rank is relevant and 0 if otherwise.

PR curve can be obtained by plotting precision and recall values. In practice, the 11-interpolated PR curve is usually selected, which is achieved by plotting the interpolated precision measure at 11 recall levels (i.e., 0, 0.1, 0.2, …, 1). The interpolated precision P_{inter} at recall level k is defined as the highest precision for any recall level $k' \geq k$

$$P_{\text{inter}}(k) = \max(P(k')). \tag{8}$$

2) Metrics for MLRSIR: The existing evaluation measures for SLRSIR are not suitable for MLRSIR since there has no ground truth (i.e., similar images) for each multilabel image in the MLRSIR dataset. To leverage this limitation, the measures designed for multilabel classification are used for performance of MLRSIR [111]. These measures are accuracy, precision, recall, hamming loss (HL), and F1-measure (F1), and are defined as follows:

$$p_{\text{Accuacy}} = \frac{1}{m}\sum_{i=1}^{m}\frac{|l_q \cap l_{r_i}|}{|l_q \cup l_{r_i}|} \tag{9}$$

$$p_{\text{Precision}} = \frac{1}{m}\sum_{i=1}^{m}\frac{|l_q \cap l_{r_i}|}{|l_{r_i}|} \tag{10}$$

$$p_{\text{Recall}} = \frac{1}{m}\sum_{i=1}^{m}\frac{|l_q \cap l_{r_i}|}{|l_q|} \tag{11}$$

$$p_{\text{HL}} = \frac{1}{m}\sum_{i=1}^{m}\frac{|l_q \oplus l_{r_i}|}{n} \tag{12}$$

$$p_{\text{F1}} = \frac{1}{m}\sum_{i=1}^{m}\frac{2\,|l_q \cap l_{r_i}|}{|l_q| + |l_{r_i}|} \tag{13}$$

where $\cap$, $\cup$, and $\oplus$ are the logical AND, logical OR, and logical XOR operations, respectively. $|\cdot|$ is the number of nonzeros, l_q is the multilabel vector of query image, l_{r_i} is the ith retrieved image, n is the number of labels (i.e., primitive classes), and m is the number of retrieved images.

Although the above-mentioned measures can be used for performance evaluation, they are originally designed for multilabel classification and, thus, are not well suited for MLRSIR. To circumvent this limitation, Imbriaco et al. [120] presented a novel and effective metric for MLRSIR.

3) Metrics for RASRSIR and CMRSIR: Regarding RASRIR and CMRSIR, there have been ground truth for each query image; therefore, the metrics for SLRSIR can be used for performance evaluation. It is worth noting that for text–image retrieval in CMRSIR, recall at k (*R@K*, $K = 1$, 5, and 10) is the widely used measure [177], which is defined as the fraction of the top *K* results that are relevant to the query.

4) Metrics for CVRSIR: The most commonly used metric for evaluating the performance of CVRSIR is recall at *K* (*R@K*, *K* = 1, 2, 3, …) [179]. For the metric *R@K*, the special case is which *K* equals $1\% \times N$, where N is the number of images. *R@K* then becomes *R*@top 1%, another widely used metric for performance evaluation.

B. Results of RSIR Methods

This section presents the performance comparisons of representative RSIR methods of each category on the corresponding benchmark datasets. It is notable that these results (except for the results on CVGD dataset) are collected from the published works since it is impossible to reimplement all of these algorithms that are not publicly available. To give a fair comparison, all the results are obtained on the same retrieval dataset. Besides, for RASRSIR and CMRSIR methods, there are two retrieval scenarios, i.e., "A→B" (Retrieve B with A as the query) and "B→A" (Retrieve A with B as the query).

1) Results of SLRSIR: Table IV presents the performance of some representative SLRSIR methods, including handcrafted feature-based methods (i.e., local features, VLAD-PQ, morphological texture, IRMFRCAMF, BoVW, GIST, LBP, gabor texture, and VLAD) and deep learning based methods (i.e., VGGM_P, VGGM_F, LDCNN, UFL, and ResNet50_P) on UC Merced, WHU-RS19, and PatternNet dataset.

As can be observed from Table IV, deep learning based methods greatly improve the performance of handcrafted feature-based methods by a significant margin on the three datasets. However, UFL achieves comparative performance with handcrafted feature-based methods on the PatternNet dataset. The result makes sense since UFL is an unsupervised deep learning method. For the deep learning-based methods, LDCNN (belonging to the novel network-based methods) achieves remarkable performance. Specifically, LDCNN outperforms CaffeRef_F and ResNet50_P (belonging to the feature extraction-based methods) on WHU-RS19 and PatternNet datasets, respectively, indicating that training novel CNNs from scratch are capable of learning more discriminative features. For UC Merced dataset, LDCNN obtains slightly worse performance than VGGM_P and VGGM_F. This is because LDCNN is trained using the AID [166] dataset, which has large variation from the UC Merced dataset. Therefore, the performance of LDCNN could be further improved when training using images that are similar to the target dataset, and in particular using the target dataset [163]. The results demonstrate that CNN features improve SLRSIR performance by a significant margin compared with handcrafted features. Among the CNN-based methods, the LDCNN network achieves the best performance on both WHU-RS19 and PatternNet dataset, and is potential to perform the best on the UC Merced dataset when the target dataset is used for training. Additionally, ResNet often performs better for RSIR than other pretrained CNNs and would be a potential CNN network for practical application scenarios.

2) Results of MLRSIR: The performance of MLRSIR methods is evaluated on DLRSD and WHDLD datasets. The results are shown in Table V. It is obvious that deep learning based

TABLE IV
PERFORMANCE COMPARISONS OF SLRSIR METHODS ON UC MERCED, WHU-RS19, AND PATTERNNET DATASETS

Method	Metric			Description
	ANMRR	MAP	P@100	
UC Merced				
Local Features [185]	0.591	-	-	The VGGM_F and LDCNN networks are trained using the AID dataset, and then transferred to UC Merced dataset. Besides, each image is used as the query image to perform retrieval; therefore, these results are averaged over all the queries.
VLAD-PQ [186]	0.451	-	-	
Morphological Texture [28]	0.575	-	-	
IRMFRCAMF [62]	0.715	-	-	
VGGM_P [187]	0.375	-	-	
VGGM_F [187]	**0.329**			
LDCNN [63]	0.439			
WHU-RS19				
BoVW [188]	0.497	-	-	The CaffeRef_F are LDCNN networks that are trained using the AID dataset, and then transferred to WHU-RS19. Besides, each image is used as the query image to perform retrieval; therefore, these results are averaged over all the queries.
GIST [189]	0.725	-	-	
LBP [190]	0.688	-	-	
VGGS_P [187]	0.279	-	-	
CaffeRef_F [191]	0.036	-	-	
LDCNN [63]	**0.019**	-	-	
PatternNet				
Gabor Texture [192]	0.6439	0.2773	0.3552	The PatternNet dataset is split into training set and testing set with the ratio of 80% and 20%, respectively. LDCNN is trained using the training set, and all the results are achieved on the testing dataset. Besides, each image is used as the query image to perform retrieval; therefore, these results are averaged over all the queries.
VLAD [193]	0.5677	0.3410	0.4111	
UFL [194]	0.6584	0.2535	0.3192	
ResNet50_P [195]	0.2584	0.6823	**0.7493**	
LDCNN [63]	**0.2416**	**0.6917**	0.6880	

Bold values indicate the best performance.

methods (i.e., SSRCF and MSRCF) outperform handcrafted feature-based methods (i.e., BoVW, LBP, gabor texture, and MLIR) in terms of HL, accuracy, precision, recall, and F1 metrics. For deep learning based methods, the multiscale region feature (i.e., MSRCF) performs better than the single-scale region feature (i.e., SSRCF) as expected; thus, MLRSIR could benefit from combining CNN features extracted from different layers. Specifically, MSCRF achieves about 3% and 5% improvement on DLRSD and WHDLD, respectively, in terms of accuracy values. Overall, for handcrafted features, BoVW obtains the best performance, with the accuracy value of 0.5454 for DLRSD and 0.7013 for WHDLD.

According to the results in Table V, deep learning has demonstrated its capacity for promoting MLRSIR. It is notable that both SSRCF and MSRCF features are extracted based on FCN network [196]. Thus, the performance is possible to be further improved if a more sophisticated network is exploited. The results indicate that CNN features improve MLRSIR performance by a significant margin compared with handcrafted features. Among the two CNN-based methods (i.e., SSRCF and MSRCF), MSRCF combines features of different layers to obtain multiscale features and thus achieves better performance. It would be a potential method for practical application scenarios.

3) Results of RASRSIR: The performance comparisons of RASRSIR methods for retrieval between MUL and PAN images on DSRSID dataset are summarized in Table VI. As the results shown, SIDHCNNs achieve remarkable performance for both "PAN→MUL" and "MUL→PAN" retrieval scenarios, and outperforms the other methods by a large margin in terms of MAP values.

To be specific, SIDHCNNs improves the worst performing method CCA by 80% for "PAN→MUL" and 81% for "MUL→PAN." The worst performance is because CCA works in an unsupervised way. In contrast, the supervised method SCM performs better than CCA, which achieves about 22% and 23% improvement for "PAN→MUL" and "MUL→PAN," respectively. Regarding DCHM, it is the second-best performing method benefiting from deep learning, indicating that deep learning can contribute to develop more effective RASRSIR methods. The remarkable performance of SIDHCNNs indicates that CNN integrating with hash learning is a potential solution

TABLE V
PERFORMANCE COMPARISONS OF MLRSIR METHODS ON DLRSD AND WHDLD DATASETS

Method	Metric					Description
	HL	Accuracy	Precision	Recall	F1	
DLRSD						
BoVW [188]	0.1718	0.5454	0.6952	0.6603	0.6459	The DLRSD dataset is split into training set and testing set with the ratio of 80% and 20%, respectively. SSRCF and MSRCF are trained using the training set, and the testing set is used for performance evaluation. Besides, there are 420 queries in total, and the results are averaged over all queries.
LBP [188]	0.2087	0.4904	0.6213	0.6492	0.5983	
Gabor Texture [190]	0.1931	0.5129	0.6484	0.6527	0.6167	
MLIR [112]	0.2017	0.5440	0.6095	0.7717	0.6539	
SSRCF [119]	0.1103	0.6939	**0.8191**	0.8246	0.7943	
MSRCF [119]	**0.1051**	**0.7193**	0.7996	**0.8830**	**0.8150**	
WHDLD						
BoVW [188]	0.2513	0.7013	0.8216	0.8328	0.7991	The WHDLD dataset is split into training set and testing set with the ratio of 80% and 20%, respectively. SSRCF and MSRCF are trained using the training set, and the testing set is used for performance evaluation. Besides, there are 988 queries in total, and the results are averaged over all queries.
LBP [190]	0.2261	0.7245	0.8477	0.8473	0.8202	
Gabor Texture [192]	0.2271	0.7338	0.8507	0.8478	0.8268	
MLIR [114]	0.2651	0.6974	0.7881	0.8614	0.7963	
SSRCF [121]	0.1561	0.8119	0.8818	0.9195	0.8842	
MSRCF [121]	**0.1151**	**0.8628**	**0.8912**	**0.9684**	**0.9172**	

Bold values indicate the best performance.

TABLE VI
PERFORMANCE COMPARISONS OF RASRSIR METHODS ON DSRSID DATASET

Method	Query →Retrieval (MAP)		Description
	PAN→MUL	MUL→PAN	
SIDHCNNs[123]	**0.9643**	**0.9789**	Training set is used for training and the searching set. Testing set is taken as the query set.
CCA [197]	0.1593	0.1594	
SCM [198]	0.3767	0.3871	
DCMH [199]	0.8509	0.8527	

Bold values indicate the best performance.

for retrieval between PAN and MUL images even with short hash codes.

4) Results of CMRSIR: Four kinds of CMRSIR methods, including sketch–image, optical–SAR, audio–image, and text–image, are evaluated on four cross-modal datasets, which are RSketch, CBRSIR_VS, UCM/Syndey/RSICD-audio, and TextRS, respectively. The results are presented in Table VII.

For sketch–image CMRSIR methods, SBRSIR performs the best on RSketch dataset. Specifically, the performances are 0.9091 and 0.5008 for seen and unseen categories (Seen categories mean the samples are used for training, while unseen categories mean the opposite), respectively, in terms of MAP values. Sketch-a-Net performs poorly on both seen and unseen categories due to the shallow architecture of network. Although DSM improves the performance of Sketch-a-Net, the improvement is limited. As the second-best method, LDF-CLS achieves a bit worse performance than SBRSIR but improves the performance of Sketch-a-Net and DSM by a remarkable margin. The results indicate that the deep hash-based method (i.e., SBRSIR) is a potential solution for sketch–image retrieval whether the categories are seen or unseen category. Regarding CMRSIR between VHR and SAR images, MsEspH outperforms other three methods for both "VHR→SAR" and "SAR→VHR" retrieval scenarios. For example, compared with the second-best method DSMHN, MsEspH achieves about 4% improvement and would be a potential solution for retrieval between VHR and SAR images.

For CMRSIR between audio (A) and image (I), CNN+M, CNN+ΔM, and CNN+Δ^2M obtain comparative performance for "I→A" on UCM-audio, Sydney-audio, and RSICD-audio datasets. For "A→I", overall, CNN+M is the best performing method. It is notable that CNN+SPEC performs the worst for both "A→I" and "I→A" retrieval scenarios; therefore, the discriminative features for audio and image are crucial for obtaining better performance. With respect to CMRSIR between text and image, overall, DBTN_EfficientNet outperforms other method in terms of $R@1$, $R@5$, and $R@10$ values. These results indicate that the deeper CNNs (i.e., ResNet50, Inception_v3, and VGG16) do not achieve remarkable performance when combined with DBTN.

5) Results of CVRSIR: Table VIII presents the results of CVRSIR on three benchmark datasets, i.e., University-1652, CVACT, and CVUSA subset. For the University-1652 dataset, the method proposed by Wang et al. outperforms other methods by a significant margin for "Drone → Satellite" and "Satellite → Drone" retrieval scenarios in terms of $R@1$ value. Besides, it can be observed that each method performs better for "Satellite → Drone" than "Drone → Satellite" scenario. This is because for

TABLE VII
PERFORMANCE COMPARISONS OF CMRSIR METHODS ON FOUR KINDS OF CROSS-MODAL DATASETS

Method	RSketch (MAP)					
	Seen			Unseen		
Sketch-a-Net [200]	0.3139			0.2573		
DSM [201]	0.5680			0.1929		
LDF-CLS [202]	0.8214			0.3815		
SBRSIR [131]	**0.9091**			**0.5008**		
	CBRSIR_VS (MAP)					
	VHR→SAR			SAR→VHR		
DCMH [199]	0.7947			0.7563		
SIDHCN [123]	0.8396			0.8051		
DSMHN [203]	0.8966			0.8058		
MsEspH [133]	**0.9398**			**0.8458**		
	UCM/Syndey/RSICD-audio (MAP)					
	A→I			I→A		
CNN+SPEC [145]	0.22	0.36	0.10	0.26	0.48	0.13
CNN+SPEC [145]	**0.32**	**0.64**	**0.16**	0.37	**0.72**	0.16
CNN+SPEC [145]	0.24	0.63	0.15	0.37	0.71	0.17
CNN+SPEC [145]	0.26	**0.64**	0.15	**0.38**	0.71	**0.18**
	TextRS					
	R@1	R@5	R@10			
DBTN_EfficientNet [177]	**0.172**	**0.514**	**0.730**			
DBTN_ResNet50 [177]	0.137	0.509	0.691			
DBTN_Inceptionv3 [177]	0.140	0.467	0.674			
DBTN_VGG16 [177]	0.119	0.444	0.637			

Bold values indicate the best performance.

"Satellite → Drone," there are multiple true-matched drone-view images for query satellite image. The success of Wang et al.'s method is due to the integration of local patterns in image, which are important for cross-view localization. Therefore, in practical application scenario, it is recommended to take the local features into account. Regarding the CVACT and CVUSA subset, SAFA achieves better performance than other methods and, in particular, CVM-Net in terms of $R@1$, $R@5$, $R@10$, and $R@$ Top 1% values. To be specific, the $R@1$ values of CVM-Net are 0.2015 and 0.1880 for CVACT and CVUSA subsets, respectively. In contrast, SAFA improves the performance of CVM-Net by 61% for CVACT and 71% for CVUSA subset. The results indicate that spatial features are important for cross-view localization. SAFA would be a potential solution for geolocalization.

In addition to the benchmark datasets mentioned above, we also report the performance of SL-CVRSIR on our CVGD dataset, as shown in Table IX. The presented SL-CVRSIR is compared with two methods, including FM and improved feature matching (IFM). For the FM method, we extract the features of overhead-view images and ground-view images from the first fully connected layer of the subnetwork pretrained on ImageNet and perform CVRSIR following the workflow, as illustrated in Fig. 6. While for the IFM method, it is similar to FM and the only difference is that the subnetwork in IFM is from SL-CVRSIR. To train SL-CVRSIR, the 100 locations in CVGD are randomly split into training set, validation set, and testing set with the ratio of 6:2:2, and then the images of each location constitute the positive and negative image pairs. Considering the fact that CVGD is a small-scale dataset, we increase the training set by flip and rotation, and the weights of convolutional layers of the subnetwork are kept frozen. The initial learning rate is set to 3e-5 and is decreased to 0.9 times of the former learning rate every ten epochs, and the batch size is set to 128 with 64 positive pairs and 64 negative pairs.

It can be observed that SL-CVRSIR outperforms FM and IFM and particularly FM in terms of the recall at top $K\%$ ($K = 1, 5, 10, 15, 20$) metric, indicating that the proposed similarity learning-based method is an effective approach for CVRSIR. Regarding FM and IFM, it is notable that the trained subnetwork performs better than the pretrained subnetwork as expected. Furthermore, the subnetwork VGG16 achieves overall better performance than AlexNet for all of the three methods; therefore, it is possible that the performance could be further improved when using a deeper subnetwork.

V. CHALLENGES AND POTENTIAL SOLUTIONS FOR RSIR

Over the past decade, RS community has witnessed the significant progress of RSIR on developing novel methods and constructing new benchmark datasets. However, RSIR has been facing some challenges, which need to be addressed to further promote RSIR research. Here, we present five main challenges for current RSIR: lack of large-scale RSIR datasets, large difference between RS images, difficulty in reproducing results of existing methods, inconsistent dataset split and evaluation protocol, and semantic gap in CBRSIR system.

A. Lack of Large-Scale RSIR Datasets

RSIR datasets are crucial for developing RSIR algorithms, especially data-driven methods (i.e., deep learning). The literature has committed to constructing new datasets for RSIR methods of different categories over the past decade, and there have been a few publicly available benchmark datasets, as shown in Table II. Unfortunately, the existing datasets have the following limitations. First, the commonly used datasets collected in early years are often small scale on which the results have gradually been saturated. Moreover, these datasets are too small to develop deep learning based approaches. Second, some existing datasets, such as WHU-RS19 [164], AID [166], and NWPU-45 [170], are originally collected for scene classification rather than RSIR. The problem is that their images contain large amounts of background unrelated to image category, and thus may distract RSIR. Third, the existing large-scale datasets are not large enough for

TABLE VIII
PERFORMANCE COMPARISONS OF CVRSIR METHODS ON UNIVERSITY-1652, CVACT, AND CVUSA SUBSET DATASETS

Method	Metric				Evaluation Detail
	R@1	R@5	R@10	R@ Top 1%	
University-1652 (Drone → Satellite/Satellite → Drone)					
Zheng et al. [178]	0.5849/0.7118	-	-	-	Drone→Satellite: There is only one true-matched satellite-view image for each query drone image. Satellite→Drone: There are multiple true-matched drone-view images for each query satellite image.
Wang et al. [204]	**0.7593/0.8645**	-	-	-	
Lin et al. [159]	0.5239/0.6391	-	-	-	
Chechik et al. [205]	0.5518/0.6362	-	-	-	
CVM-Net [154]	0.5321/0.6562	-	-	-	
CVACT					
Wang et al. [204]	0.7999	0.9063	0.9256	0.9703	The results are achieved on the validation set of CVACT. A query image only has one true-matched image.
SAFA [206]	**0.8103**	**0.9280**	**0.9484**	**0.9817**	
CVFT [207]	0.6105	0.8133	0.8652	0.9593	
Zheng et al. [178]	0.3120	0.5364	0.6300	0.8527	
CVM-Net [154]	0.2015	0.4500	0.5687	0.8757	
CVUSA subsect					
Wang et al. [204]	0.8579	0.9538	0.9698	0.9941	It contains 35,532 ground-and-satellite image pairs for training and 8884 image pairs for testing.
SAFA [206]	**0.8984**	**0.9693**	**0.9814**	**0.9964**	
CVFT [207]	0.6143	0.8469	0.9094	0.9902	
Zheng et al. [178]	0.4391	0.6638	0.7458	0.9178	
CVM-Net [154]	0.1880	0.4442	0.5747	0.9154	

Bold values indicate the best performance.

TABLE IX
PERFORMANCE OF SL-CVRSIR ON THE CVGD DATASET

Method	Subnetwork	Recall at Top K%				
		1	5	10	15	20
FM	AlexNet	8.16	26.53	38.78	48.98	56.12
	VGG16	11.22	32.65	45.92	58.16	63.27
IFM	AlexNet	11.22	33.67	46.94	67.35	78.57
	VGG16	14.29	46.94	62.24	70.41	76.53
SL-CVRSIR	AlexNet	17.35	43.88	65.31	**77.55**	**84.69**
	VGG16	**18.37**	**51.02**	**67.35**	75.51	82.65

Bold values indicate the best performance.

training deep learning networks from scratch, although the literature has collected some large-scale datasets. In practice, transfer learning is still a commonly accepted strategy to overcome this limitation.

To address the dataset related issues, much larger RSIR datasets are required. Besides, we should keep in mind the differences between different categories of RSIR methods when creating datasets and follow the guidance on creating benchmark datasets for RS image interpretation [208].

B. Large Difference Between RS Images

With the rapid development of RS technology, more and more RS images have become available. Generally, these images have varied resolutions, complexities, and even modalities, presenting great challenges for RSIR and in particular the cross-source RSIR. To be specific, for unisource RSIR (i.e., SLRSIR and MLRSIR), although the query images and images to be retrieved are from the same source, they might be different in terms of image size, resolution, scale, etc., leading to the problem of big intraclass diversity and high interclass similarity. For cross-source RSIR (i.e., RASRSIR, CMRSIR, and CVRSIR), retrieval is performed between two sources. The large differences between RS images may also degrade the performance. For example, RASRSIR is generally to perform retrieval between images captured by two different sensors (e.g., multispectral and panchromatic sensors of one satellite), which is also the case for the existing DSRSID dataset [123]. However, RASRSIR will become more challenging when the multispectral and panchromatic images are from more than one satellite.

Considering the large difference between RS images, there are two potential solutions: First, constructing a large dataset containing images with varied resolutions, complexities, and scales. Second, developing novel RSIR algorithms and particularly deep learning based approaches to learn more powerful and discriminative features.

C. Difficulty in Reproducing Results of Existing Methods

A large number of RSIR methods have been developed over the past decade, and some of them even achieved SOTA performance. Some of these methods, however, are unable to reimplement because of the lack of open implementations. Besides, it is difficult to replicate the same results presented in the published works because the necessary implementation details were not provided. For example, to train a successful CNN, some tricks, such as data augmentation and dropout, are often exploited. It is impossible to replicate exactly the same results without these details.

RS community may learn from CV domain, where the open-source implementation and necessary details are often provided.

D. Inconsistent Dataset Split and Evaluation Protocol

A developed algorithm needs to be compared with the existing methods to demonstrate its performance. To this end, we need to either reimplement these methods or collect the results from published works. The former is a challenge, as discussed above. While for the latter, the results presented in related works may not be feasible due to the inconsistent dataset split and evaluation protocol. For example, most existing RSIR datasets do not provide dataset splits, such as training set, validation set, and testing set. In practice, these dataset splits are often obtained by randomly dividing the dataset into different parts, resulting in different dataset splits. Besides, the evaluation protocol may also be different. For example, before evaluating an RSIR method, we need to select the query images and performance metrics. It is not fair to compare two RSIR methods with totally different query images. Moreover, the methods may be evaluated using different metrics, making it impossible to compare them directly.

The above-mentioned issues could be overcome as long as we provide the dataset splits and evaluation protocol when constructing new RSIR datasets, as CV domain does.

E. Existence of Semantic Gap

Most of the existing RSIR methods were performed in the feature level, relying on visual features to compare similarity between RS images. However, the results cannot well reflect the users' real query intensions due to the "semantic gap" between the visual features and, in particular, the low-level features and the richness of human semantics [209]. As the current mainstream technique for RSIR, deep learning is able to extract high-level features containing some semantic information, but the "semantic gap" problem is still not well explored.

To reduce the effect of "semantic gap," a potential solution is to combine deep learning with the traditional techniques, such as relevance feedback to learn users' intention.

VI. Future Directions for RSIR

RSIR is an effective technique for organizing and managing large RS image archive, and the RS community has committed great efforts to promote RSIR over the past decade. Thanks to deep learning, the literature has achieved significant progress in terms of new RSIR methods and benchmark datasets. However, some RSIR issues are still required to be addressed. In this section, we, therefore, point out several potential directions for RSIR.

A. Constructing More Large and Challenging Datasets

Deep learning has been the mainstream technique for RSIR. To train a successful CNN, a large volume of labeled images is required. However, the existing RSIR datasets (as shown in Table II) are still not large enough to train CNNs from scratch. Besides, RS images generally have varied resolutions, complexities, and modalities in real RSIR scenarios. Therefore, these factors should be taken into account when constructing RSIR datasets to meet the real RSIR scenarios.

B. Few- and Zero-Shot Learning (FSL and ZSL) for RSIR

Current RSIR methods focus the attention on developing novel supervised algorithms, in particular CNN-based methods. The prerequisites for these methods are large-scale labeled image archives. However, it is time-consuming and laborious to annotate a huge volume of RS images. FSL is a type of machine learning method where the training set contains limited labeled samples, and ZSL even requires no labeled samples, which have been two commonly used techniques for RS task, such as scene classification [210], [211], [212]. FSL and ZSL provide RS community a potential direction for developing effective RSIR methods without large-scale labeled images.

C. Developing Novel MLRSIR and CMRSIR Methods

MLRSIR and CMRSIR are still new topics and have some advantages over SLRSIR. For example, MLRSIR is able to perform fine-grained retrieval for users, which is, however, not available by using SLRSIR. By assuming such a situation, one intends to search an outdoor basketball court where there is a parking lot and restaurant nearby. MLRSIR is to conduct RSIR between multilabel images, thus is feasible for the above task. Regarding CMRSIR, the text–image and audio–image CMRSIR methods are friendly for users with no expert knowledge. Therefore, developing novel MLRSIR and CMRSIR methods will promote the application of RSIR.

D. Incremental Learning for RSIR

Current RSIR methods are trained and evaluated using static RS datasets and, thus, is not suited for incremental scenarios [213]. Specifically, most of the RSIR methods assume that the trained model has seen all the image categories, which is, however, not the case in real-world applications as new RS images are constantly emerging. This is also the reason why RSIR methods generally achieve worse performance when

transferred to unseen images. Therefore, one potential direction is to develop incremental learning methods for RSIR that can deal with incremental streams of new RS image.

E. Hashing Methods for Large-Scale RSIR

In practical RSIR applications, we are facing large amounts of RS images. RSIR in a large-scale scenario is challenging mainly in two aspects. On the one hand, more storage space is needed to store RS images and features. On the other hand, it is time-consuming to perform one query in a large-scale RS archive. Hashing methods are able to generate compact binary codes for RSIR, which can not only save storage cost but also have high retrieval efficiency, providing a potential direction for large-scale RSIR. The key is to balance the tradeoff between efficiency and accuracy.

VII. Conclusion

As one of the research topics in RS community, RSIR has obtained great improvements in terms of methods and benchmark datasets over the past decade. We, therefore, in this article provide a comprehensive and systematic survey of the recent achievements of RSIR and discuss its challenges and potential future directions. To be specific, we first group the existing RSIR methods in a hierarchical category and review the related works from five RSIR category, including SLRSIR, MLRSIR, RASRSIR, CMRSIR, and CVRSIR. Then, we present the benchmark datasets for each RSIR category. To promote CVRSIR, we proposed an effective method based on the similarity learning and constructed a new ground–drone dataset for performance evaluation. Besides, we compared the performance of representative RSIR methods of each category on some benchmark datasets. Finally, we discussed the challenges and potential directions for RSIR.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments to improve this work.

References

[1] P. Agouris, J. Carswell, and A. Stefanidis, "An environment for content-based image retrieval from large spatial databases," *ISPRS J. Photogramm. Remote Sens.*, vol. 54, no. 4, pp. 263–272, 1999.

[2] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, 2008, Art. no. 5.

[5] D. Espinoza-Molina and M. Datcu, "Earth-observation image retrieval based on content, semantics, and metadata," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 11, pp. 5145–5159, Nov. 2013.

[6] Y. N. Mamatha and A. G. Ananth, "Content based image retrieval of satellite imageris using soft query based color composite techniques," *Int. J. Comput. Appl.*, vol. 7, no. 5, pp. 40–45, 2010.

[7] C. Ma, Q. Dai, J. Liu, S. Liu, and J. Yang, "An improved SVM model for relevance feedback in remote sensing image retrieval," *Int. J. Digit. Earth*, vol. 7, no. 9, pp. 725–745, 2014.

[8] J. A. Piedra-Fernandez, G. Ortega, J. Z. Wang, and M. Canton-Garbin, "Fuzzy content-based image retrieval for oceanic remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5422–5431, Sep. 2014.

[9] J. Li and R. M. Narayanan, "Integrated spectral and spatial information mining in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 673–685, Mar. 2004.

[10] S. Bhagavathy and B. S. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3706–3715, Dec. 2006.

[11] M. Wang and T. Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2874–2886, May 2013.

[12] Z. F. Shao, W. X. Zhou, and Q. M. Cheng, "Remote sensing image retrieval with combined features of salient region," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 40, no. 6, pp. 83–88, 2014.

[13] N. Laban, M. ElSaban, A. Nasr, and H. Onsi, "System refinement for content based satellite image retrieval," *Egyptian J. Remote Sens. Space Sci.*, vol. 15, no. 1, pp. 91–97, 2012.

[14] M. Wang, Q. M. Wan, L. B. Gu, and T. Y. Song, "Remote-sensing image retrieval by combining image visual and semantic features," *Int. J. Remote Sens.*, vol. 34, no. 12, pp. 4200–4223, 2013.

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

[16] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[17] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042609.

[18] L. Zhang, Lefei Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[19] 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.

[20] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.

[21] S. K. Sudha and S. Aji, "A review on recent advances in remote sensing image retrieval techniques," *J. Indian Soc. Remote Sens.*, vol. 47, no. 12, pp. 2129–2139, 2019.

[22] 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, 2019, Art. no. 2110.

[23] S. K. Sudha and S. Aji, "An analysis on deep learning approaches: Addressing the challenges in remote sensing image retrieval," *Int. J. Remote Sens.*, vol. 42, no. 24, pp. 9405–9441, 2021.

[24] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 507–521, Sep. 2020.

[25] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, 2021.

[26] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *J. Appl. Remote Sens.*, vol. 8, no. 1, 2014, Art. no. 083584.

[27] P. Bosilj, E. Aptoula, S. Lefèvre, and E. Kijak, "Retrieval of remote sensing images with pattern spectra descriptors," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 12, 2016, Art. no. 228.

[28] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2014.

[29] G. Chen, Z. Jiang, and M. M. Kamruzzaman, "Radar remote sensing image retrieval algorithm based on improved Sobel operator," *J. Vis. Commun. Image Represent.*, vol. 71, 2020, Art. no. 102720.

[30] P. K. Kavitha and P. Vidhya Saraswathi, "Content based satellite image retrieval system using fuzzy clustering," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 5, pp. 5541–5552, 2021.

[31] T. Sunitha and T. S. Sivarani, "An efficient content-based satellite image retrieval system for big data utilizing threshold based checking method," *Earth Sci. Inform.*, vol. 14, no. 4, pp. 1847–1859, 2021.

[32] O. Ben-Ahmed, T. Urruty, N. Richard, and C. Fernandez-Maloigne, "Toward content-based hyperspectral remote sensing image retrieval (CB-HRSIR): A preliminary study based on spectral sensitivity functions," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 600.

[33] I. Tekeste and B. Demir, "Advanced local binary patterns for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6855–6858.

[34] J. Zhang, W. Geng, X. Liang, J. Li, L. Zhuo, and Q. Zhou, "Hyperspectral remote sensing image retrieval system using spectral and texture features," *Appl. Opt.*, vol. 56, no. 16, pp. 4785–4796, 2017.

[35] Z. Du, X. Li, and Xiaoqiang Lu, "Local structure learning in high resolution remote sensing image retrieval," *Neurocomputing*, vol. 207, pp. 813–822, 2016.

[36] K. N. Sukhia, M. M. Riaz, A. Ghafoor, and S. S. Ali, "Content-based remote sensing image retrieval using multi-scale local ternary pattern," *Digit. Signal Process.*, vol. 104, 2020, Art. no. 102765.

[37] J. Yang, J. Liu, and Q. Dai, "An improved bag-of-words framework for remote sensing image retrieval in large-scale image databases," *Int. J. Digit. Earth*, vol. 8, no. 4, pp. 273–292, 2015.

[38] B. Chaudhuri, B. Demir, L. Bruzzone, and S. Chaudhuri, "Region-based retrieval of remote sensing images using an unsupervised graph-theoretic approach," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 987–991, Jul. 2016.

[39] F. Ye, X. Zhao, W. Luo, D. Li, and W. Min, "Query-adaptive remote sensing image retrieval based on image rank similarity and image-to-query class similarity," *IEEE Access*, vol. 8, pp. 116824–116839, 2020.

[40] A. P. Byju, B. Demir, and L. Bruzzone, "A progressive content-based image retrieval in JPEG 2000 compressed remote sensing archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5739–5751, Aug. 2020.

[41] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5798–5817, Oct. 2017.

[42] R. Fernandez-Beltran, B. Demir, F. Pla, and A. Plaza, "Unsupervised remote sensing image retrieval using probabilistic latent semantic hashing," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 2, pp. 256–260, Feb. 2021.

[43] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2323–2334, May 2015.

[44] Q. Cheng, Z. Shao, K. Shao, C. Li, J. Li, and S. Li, "A distributed system architecture for high-resolution remote sensing image retrieval by combining deep and traditional features," *Image Signal Process. Remote Sens.*, vol. 10789, pp. 413–432, 2018.

[45] F. Ye, S. Chen, X. Meng, and J. Xin, "Query-adaptive feature fusion base on convolutional neural networks for remote sensing image retrieval," in *Proc. Int. Conf. Digit. Soc. Intell. Syst.*, 2021, pp. 148–151.

[46] F. Ye, W. Luo, M. Dong, D. Li, and W. Min, "Content-based remote sensing image retrieval based on fuzzy rules and a fuzzy distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8002505.

[47] Z. Fan, W. Zhang, D. Zhang, and L. Meng, "An automatic accurate high-resolution satellite image retrieval method," *Remote Sens.*, vol. 9, no. 11, 2017, Art. no. 1092.

[48] M. N. Vharkate and V. B. Musande, "Remote sensing image retrieval using hybrid visual geometry group network with relevance feedback," *Int. J. Remote Sens.*, vol. 42, no. 14, pp. 5540–5567, 2021.

[49] Z. Zhuo and Z. Zhou, "Low dimensional discriminative representation of fully connected layer features using extended Largevis method for high-resolution remote sensing image retrieval," *Sensors*, vol. 20, no. 17, 2020, Art. no. 4718.

[50] D. Hou, Z. Miao, H. Xing, and H. Wu, "Exploiting low dimensional features from the MobileNets for remote sensing image retrieval," *Earth Sci. Inform.*, vol. 13, no. 4, pp. 1437–1443, 2020.

[51] P. Sadeghi-Tehran, P. Angelov, N. Virlet, and M. J. Hawkesford, "Scalable database indexing and fast image retrieval based on deep learning and hierarchically nested structure applied to remote sensing and plant biology," *J. Imag.*, vol. 5, no. 3, 2019, Art. no. 33.

[52] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018.

[53] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[54] Y. Ge, S. Jiang, Q. Xu, C. Jiang, and F. Ye, "Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval," *Multimedia Tools Appl.*, vol. 77, no. 13, pp. 17489–17515, 2018.

[55] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. N. de With, "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 493.

[56] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1243.

[57] F. Hu, X. Tong, G.-S. Xia, and L. Zhang, "Delving into deep representations for remote sensing image retrieval," in *Proc. IEEE 13th Int. Conf. Signal Process.*, 2016, pp. 198–203.

[58] P. Napoletano, "Visual descriptors for content-based retrieval of remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 5, pp. 1343–1376, 2018.

[59] C. Ma, F. Chen, J. Yang, J. Liu, W. Xia, and X. Li, "A remote-sensing image-retrieval model based on an ensemble neural networks," *Big Earth Data*, vol. 2, no. 4, pp. 351–367, 2018.

[60] Y. Boualleg and M. Farah, "Enhanced interactive remote sensing image retrieval with scene classification convolutional neural networks model," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4748–4751.

[61] H. Cevikalp, G. G. Dordinejad, and M. Elmas, "Feature extraction with convolutional neural networks for aerial image retrieval," in *Proc. 25th Signal Process. Commun. Appl. Conf.*, 2017, pp. 1–4.

[62] Y. Li, Y. Zhang, C. Tao, and H. Zhu, "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion," *Remote Sens.*, vol. 8, no. 9, 2016, Art. no. 709.

[63] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 489.

[64] Y. Boualleg, M. Farah, and I. R. Farah, "TLDCNN: A triplet low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," in *Proc. Mediterranean Middle-East Geosci. Remote Sens. Symp.*, 2020, pp. 13–16.

[65] M. Zhang, Q. Cheng, F. Luo, and L. Ye, "A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2711–2723, Feb. 2021.

[66] Z. Zhuo and Z. Zhou, "Remote sensing image retrieval with gabor-CA-ResNet and split-based deep feature transform network," *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 869.

[67] Z.-Z. Wu, C. Zou, Y. Wang, M. Tan, and T. Weise, "Rotation-aware representation learning for remote sensing image retrieval," *Inf. Sci.*, vol. 572, pp. 404–423, 2021.

[68] Y. Liu, Yingbin Liu, C. Chen, and L. Ding, "Remote-sensing image retrieval with tree-triplet-classification networks," *Neurocomputing*, vol. 405, pp. 48–61, 2020.

[69] Y. Wang, S. Ji, and Y. Zhang, "A learnable joint spatial and spectral transformation for high resolution remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8100–8112, Aug. 2021.

[70] G. Sumbul and B. Demir, "Plasticity-stability preserving multi-task learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5620116.

[71] Y. Wang et al., "A three-layered graph-based learning approach for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6020–6034, Oct. 2016.

[72] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Comput. Vis. Image Understanding*, vol. 184, pp. 22–30, 2019.

[73] S. Wang, D. Hou, and H. Xing, "A novel multi-attention fusion network with dilated convolution and label smoothing for remote sensing image retrieval," *Int. J. Remote Sens.*, vol. 43, no. 4, pp. 1306–1322, 2022.

[74] H. Wang, Z. Zhou, H. Zong, and L. Miao, "Wide-context attention network for remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2082–2086, Dec. 2021.

[75] Y. Wang, S. Ji, M. Lu, and Y. Zhang, "Attention boosted bilinear pooling for remote sensing image retrieval," *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2704–2724, 2020.

[76] W. Xiong, Y. Lv, Y. Cui, X. Zhang, and X. Gu, "A discriminative feature learning approach for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 281.

[77] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "Attention-driven graph convolution network for remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 8019705.

[78] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," Michigan State Univ., East Lansing, MI, USA, 2006.

[79] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1735–1742.

[80] 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.

[81] H. Zhao, L. Yuan, H. Zhao, and Z. Wang, "Global-aware ranking deep metric learning for remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 8008505.

[82] Y. Cao et al., "DML-GANR: Deep metric learning with generative adversarial network regularization for high spatial resolution remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8888–8904, Dec. 2020.

[83] R. Cao et al., "Enhancing remote sensing image retrieval using a triplet deep metric learning network," *Int. J. Remote Sens.*, vol. 41, no. 2, pp. 740–751, 2019.

[84] Q. Cheng, D. Gan, P. Fu, H. Huang, and Y. Zhou, "A novel ensemble architecture of residual attention-based deep metric learning for remote sensing image retrieval," *Remote Sens.*, vol. 13, no. 17, 2021, Art. no. 3445.

[85] H. Chung, W.-J. Nam, and S.-W. Lee, "Rotation invariant aerial image retrieval with group convolutional metric learning," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 6431–6438.

[86] Y. Liu, Z. Han, C. Chen, L. Ding, and Y. Liu, "Eagle-eyed multitask CNNs for aerial image retrieval and scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6699–6721, Sep. 2020.

[87] L. Fan, H. Zhao, and Haoyu Zhao, "Global optimization: Combining local loss with result ranking loss in remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 7011–7026, Aug. 2021.

[88] H. Zhao, L. Yuan, and H. Zhao, "Similarity retention loss (SRL) based on deep metric learning for remote sensing image retrieval," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 2, 2020, Art. no. 61.

[89] L. Fan, H. Zhao, and Haoyu Zhao, "Distribution consistency loss for large-scale remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 175.

[90] Y. Liu, L. Ding, C. Chen, and Y. Liu, "Similarity-based unsupervised deep transfer learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7872–7889, Nov. 2020.

[91] F. Ye, M. Dong, W. Luo, X. Chen, and W. Min, "A new re-ranking method based on convolutional neural network and two image-to-class distances for remote sensing image retrieval," *IEEE Access*, vol. 7, pp. 141498–141507, 2019.

[92] X. Chen and C. Lu, "An end-to-end adversarial hashing method for unsupervised multispectral remote sensing image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 1536–1540.

[93] T. Reato, B. Demir, and L. Bruzzone, "An unsupervised multicode hashing method for accurate and scalable remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 276–280, Feb. 2019.

[94] N. Lukac, B. Zalik, S. Cui, and M. Datcu, "GPU-based kernelized locality-sensitive hashing for satellite image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1468–1471.

[95] J. Kong, Q. Sun, M. Mukherjee, and J. Lloret, "Low-rank hypergraph hashing for large-scale remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1164.

[96] X. Tan, Y. Zou, Z. Guo, K. Zhou, and Q. Yuan, "Deep contrastive self-supervised hashing for remote sensing image retrieval," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3643.

[97] X. Tang, C. Liu, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Large-scale remote sensing image retrieval based on semi-supervised adversarial hashing," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 2055.

[98] X. Tang, C. Liu, X. Zhang, J. Ma, C. Jiao, and L. Jiao, "Remote sensing image retrieval based on semi-supervised deep hashing learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 879–882.

[99] Y. Sun et al., "Unsupervised deep hashing through learning soft pseudo label for remote sensing image retrieval," *Knowl.-Based Syst.,* vol. 239, 2022, Art. no. 107807.

[100] P. Li et al., "Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7331–7345, Oct. 2020.

[101] W. Song, Z. Gao, R. Dian, P. Ghamisi, Y. Zhang, and J. A. Benediktsson, "Asymmetric hash code learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5617514.

[102] L. Han, P. Li, X. Bai, C. Grecos, X. Zhang, and P. Ren, "Cohesion intensive deep hashing for remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 101.

[103] C. Liu, J. Ma, X. Tang, X. Zhang, and L. Jiao, "Adversarial hash-code learning for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4324–4327.

[104] 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. 2018.

[105] D. Ye, Y. Li, C. Tao, X. Xie, and X. Wang, "Multiple feature hashing learning for large-scale remote sensing image retrieval," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 11, 2017, Art. no. 364.

[106] X. Tang et al., "Meta-hashing for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5615419.

[107] Z. Wang, N. Wu, X. Yang, B. Yan, and P. Liu, "Deep learning triplet ordinal relation preserving binary code for remote sensing image retrieval task," *Remote Sens.*, vol. 13, no. 23, 2021, Art. no. 4786.

[108] X. Shan, P. Liu, Y. Wang, Q. Zhou, and Z. Wang, "Deep hashing using proxy loss on remote sensing image retrieval," *Remote Sens.*, vol. 13, no. 15, 2021, Art. no. 2924.

[109] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Deep hash learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3420–3443, Apr. 2021.

[110] P. Li, X. Zhang, X. Zhu, and P. Ren, "Online hashing for scalable remote sensing image retrieval," *Remote Sens.,* vol. 10, no. 5, 2018, Art. no. 709.

[111] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[112] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content based retrieval of multi-label remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 1744–1747.

[113] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[114] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 964.

[115] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.

[116] G. Sumbul, M. Ravanbakhsh, and B. Demir, "A relevant, hard and diverse triplet sampling method for multi-label remote sensing image retrieval," in *Proc. IEEE Mediterranean Middle-East Geosci. Remote Sens. Symp.*, 2022, pp. 5–8.

[117] G. Sumbul and B. Demir, "A novel graph-theoretic deep representation learning method for multi-label remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 266–269.

[118] G. Sumbul, M. Ravanbakhsh, and B. Demir, "Informative and representative triplet selection for multilabel remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5405811.

[119] Q. Cheng, H. Huang, L. Ye, P. Fu, D. Gan, and Y. Zhou, "A semantic-preserving deep hashing model for multi-label remote sensing image retrieval," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 4965.

[120] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. N. de With, "Toward multilabel image retrieval for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 4703214.

[121] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, Jan. 2020.

[122] Y. Hua, L. Mou, and X. X. Zhu, "Label relation inference for multi-label aerial image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5244–5247.

[123] 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.

[124] J. Ma, D. Shi, X. Tang, X. Zhang, X. Han, and L. Jiao, "Cross-source image retrieval based on ensemble learning and knowledge distillation for remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2803–2806.

[125] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1234–1247, Mar. 2020.

[126] J. Gao, X. Shen, P. Fu, Z. Ji, and T. Wang, "Multiview graph convolutional hashing for multisource remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 8015305.

[127] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4860–4874, Jul. 2020.

[128] J. Ma, D. Shi, X. Tang, X. Zhang, and L. Jiao, "Dual modality collaborative learning for cross-source remote sensing retrieval," *Remote Sens.*, vol. 14, no. 6, 2022, Art. no. 1319.

[129] T. Jiang, G.-S. Xia, and Q. Lu, "Sketch-based aerial image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3690–3694.

[130] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "Attention-driven cross-modal remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4783–4786.

[131] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G.-S. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7801–7814, Nov. 2020.

[132] W. Xiong, Z. Xiong, Y. Zhang, Y. Cui, and X. Gu, "A deep cross-modality hashing network for SAR and optical remote sensing images retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5284–5296, Sep. 2020.

[133] Y. Sun et al., "Multisensor fusion and explicit semantic preserving-based deep hashing for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5219614.

[134] G. Sumbul et al., "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021.

[135] Y. Lv, W. Xiong, X. Zhang, and Y. Cui, "Fusion-based correlation learning model for cross-modal remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 6503205.

[136] Z. Yuan et al., "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5620616.

[137] Z. Yuan et al., "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 4404119.

[138] Z. Yuan et al., "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5612819.

[139] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4284–4297, Apr. 2021.

[140] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Unsupervised contrastive hashing for cross-modal retrieval in remote sensing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, 2022, pp. 4463–4467.

[141] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image–voice retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7049–7061, Oct. 2020.

[142] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 84.

[143] Y. Chen, S. Xiong, L. Mou, and X. X. Zhu, "Deep quadruple-based hashing for remote sensing image-sound retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 4705814.

[144] M. Guo, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–7.

[145] M. Guo, C. Zhou, and J. Liu, "Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 11, pp. 4644–4654, Nov. 2019.

[146] H. Ning, B. Zhao, and Y. Yuan, "Semantics-consistent representation learning for remote sensing image–voice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 4700614.

[147] R. Yang et al., "Cross-modal feature fusion retrieval for remote sensing image-voice retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2855–2858.

[148] R. Yang et al., "Multimodal fusion remote sensing image—audio retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6220–6235, Aug. 2022.

[149] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[150] X. Zhang et al., "SSA-Net: Spatial scale attention network for image-based geo-localization," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 8022905.

[151] E. Zemene, Y. T. Tesfaye, H. Idrees, A. Prati, M. Pelillo, and M. Shah, "Large-scale image geo-localization using dominant sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 148–161, Jan. 2019.

[152] J. Lin et al., "Joint representation learning and keypoint detection for cross-view geo-localization," *IEEE Trans. Image Process.*, vol. 31, pp. 3780–3792, May 2022.

[153] R. Rodrigues and M. Tani, "Global assists local: Effective aerial representations for field of view constrained image geo-localization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2694–2702.

[154] S. Hu, S. Feng, R. M. H. Nguyen, and G. H. Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7258–7267.

[155] Z. Zeng, Z. Wang, F. Yang, and S. Satoh, "Geo-localization via ground-to-satellite cross-view image retrieval," *IEEE Trans. Multimed.*, to be published, doi: 10.1109/TMM.2022.3144066.

[156] L. Cheng et al., "Crowd-sourced pictures geo-localization method based on street view images and 3D reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 141, pp. 72–85, 2018.

[157] Y. Tian, X. Deng, Y. Zhu, and S. Newsam, "Cross-time and orientation-invariant overhead image geolocalization using deep local features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2501–2509.

[158] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3961–3969.

[159] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5007–5015.

[160] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 891–898.

[161] N. Khurshid, T. Hanif, M. Tharani, and M. Taj, "Cross-view image retrieval-ground to aerial image retrieval through deep learning," in *Proc. 26th Int. Conf. Neural Inf. Process.*, 2019, pp. 210–221.

[162] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16989–16999.

[163] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM Int. Symp. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.

[164] G. Xia et al., "Structural high-resolution satellite image indexing to cite this version," in *Proc. ISPRS TC VII Symp.-100 Years ISPRS*, 2010, pp. 298–303.

[165] 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.

[166] 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.

[167] 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, 2018.

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

[169] 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. 2016.

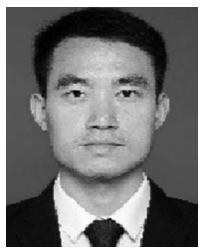
[170] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[171] 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, 2020.

[172] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[173] Y. Hua, L. Mou, P. Jin, and X. X. Zhu, "MultiScene: A large-scale dataset and benchmark for multiscene recognition in single aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5610213.

[174] 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.*, 2019, pp. 5901–5904.
[175] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst.*, 2016, pp. 1–5.
[176] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.
[177] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, "TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 405.
[178] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1395–1403.
[179] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5617–5626.
[180] G. Machado et al., "AiRound and CV-BrCT: Novel multiview datasets for scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 488–503, Oct. 2021.
[181] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4132–4140.
[182] S. Zhu, T. Yang, and C. Chen, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3640–3649.
[183] N. P. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 494–509.
[184] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 118.1–118.12.
[185] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.
[186] S. Özkan, T. Ateş, E. Tola, M. Soysal, and E. Esen, "Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1996–2000, Nov. 2014.
[187] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014.
[188] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 1470–1477.
[189] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[190] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[191] Y. Jia et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[192] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[193] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.
[194] W. Zhou, Z. Shao, C. Diao, and Q. Cheng, "High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder," *Remote Sens. Lett.*, vol. 6, no. 10, pp. 775–783, 2015.
[195] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[196] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[197] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.
[198] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, vol. 3, no. 1, pp. 2177–2183.
[199] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3270–3278.
[200] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5552–5561.
[201] F. Radenovic, G. Tolias, and O. Chum, "Deep shape matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 774–791.
[202] T.-B. Jiang, G.-S. Xia, Q.-K. Lu, and W.-M. Shen, "Retrieving aerial scene images with learned deep image-sketch features," *J. Comput. Sci. Technol.*, vol. 32, no. 4, pp. 726–737, 2017.
[203] L. Jin, Z. Li, and J. Tang, "Deep semantic multimodal hashing network for scalable image-text and video-text retrievals," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2020.2997020.
[204] T. Wang et al., "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 867–879, Feb. 2022.
[205] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, 2009, pp. 11–14.
[206] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for cross-view image based geo-localization," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 10090–10100.
[207] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11990–11997.
[208] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, Apr. 2021.
[209] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.
[210] Y. Li, Z. Zhu, J.-G. Yu, and Y. Zhang, "Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590–10603, Dec. 2021.
[211] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, 2021.
[212] G. Cheng et al., "SPNet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5608011.
[213] W. Chen et al., "Deep learning for instance retrieval: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2022.3218591.

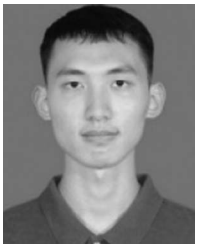


Weixun Zhou (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019.

He is currently a Lecturer with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include urban remote sensing and intelligent processing of remote sensing images.

Haiyan Guan (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009, and the Ph.D. degree in geomatics from the University of Waterloo, Waterloo, ON, Canada, in 2014.

She is currently a Professor with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China. Her current research interests include information extraction from LiDAR point clouds and from earth observation images. She has authored or coauthored more than 50 research papers in refereed journals, books, and proceedings, including the IEEE TGRS, IEEE-TITS, IEEE GRSL, *ISPRS Journal of Photogrammetry and Remote Sensing*, and IGARSS and ISPRS proceedings.



Ziyu Li received the B.S. degree in remote sensing science and technology from the Nanjing University of Information Science and Technology, Nanjing, China, in 2022. He is currently working toward the M.E. degree in surveying and mapping engineering with Hohai University, Nanjing, China.

His research interests include cross-view image geolocalization and hyperspectral image classification.

Zhenfeng Shao received the Ph.D. degree in aerial photogrammetry from Wuhan University, Wuhan, China, in 2004.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. His research interests include remote sensing and data mining.

Mahmoud R. Delavar received the B.Sc. degree in civil engineering and surveying from Khajeh Nasir Toosi University, Tehran, Iran, in 1988, the M.Sc. degree in civil engineering photogrammetry and remote sensing from the University of Roorkee (currently IIT Roorkee), Roorkee, India, in 1992, and the Ph.D. degree in geomatic engineering and geographic information system (GIS) from the University of New South Wales, Sydney, NSW, Australia, in 1997.

In 1998, he joined the College of Engineering, University of Tehran, Tehran, Iran, where he is currently a Full Professor of GIS and the Director of Center of Excellence in Geomatic Engineering in Disaster Management, School of Surveying and Geospatial Engineering. He has founded the Iranian Society of Surveying and Geomatic Engineering in 2001. He has been the national representative of the International Society of Urban Data Management since 2006, the Scientific Secretary of the International Society for Photogrammetry and Remote Sensing WG II/4 (uncertainty modeling and quality control for spatial data), in 2008–2012, and chairing ISPRS WG IV/3 (spatial data analysis, statistics, and uncertainty modeling), in 2016–2022. He is the representative of the University of Tehran in the International Geomatic network for networks. He is in the editorial board of a number of national and international scientific journals, such as *ISPRS International Journal of GeoInformation*, *Spatial Statistics* (Ex. Editorial Board), and *International Journal of Geo-spatial Information Science*. He has authored or coauthored more than 373 papers in national and international conferences and scientific journals. He has supervised 95 M.Sc. and 13 Ph.D. students, and 1 Postdoctoral research so far. His research interests include spatial data quality and uncertainty modeling, spatiotemporal GIS, disaster management, smart cities, land administration, SDI, spatial data fusion, spatiotemporal data mining, spatial data science, spatial big data, urban growth modeling, land use and land cover change modeling, remote sensing, and GIS integration.